UiT The Arctic University of Norway

Faculty of Health Sciences, Department of Clinical Medicine

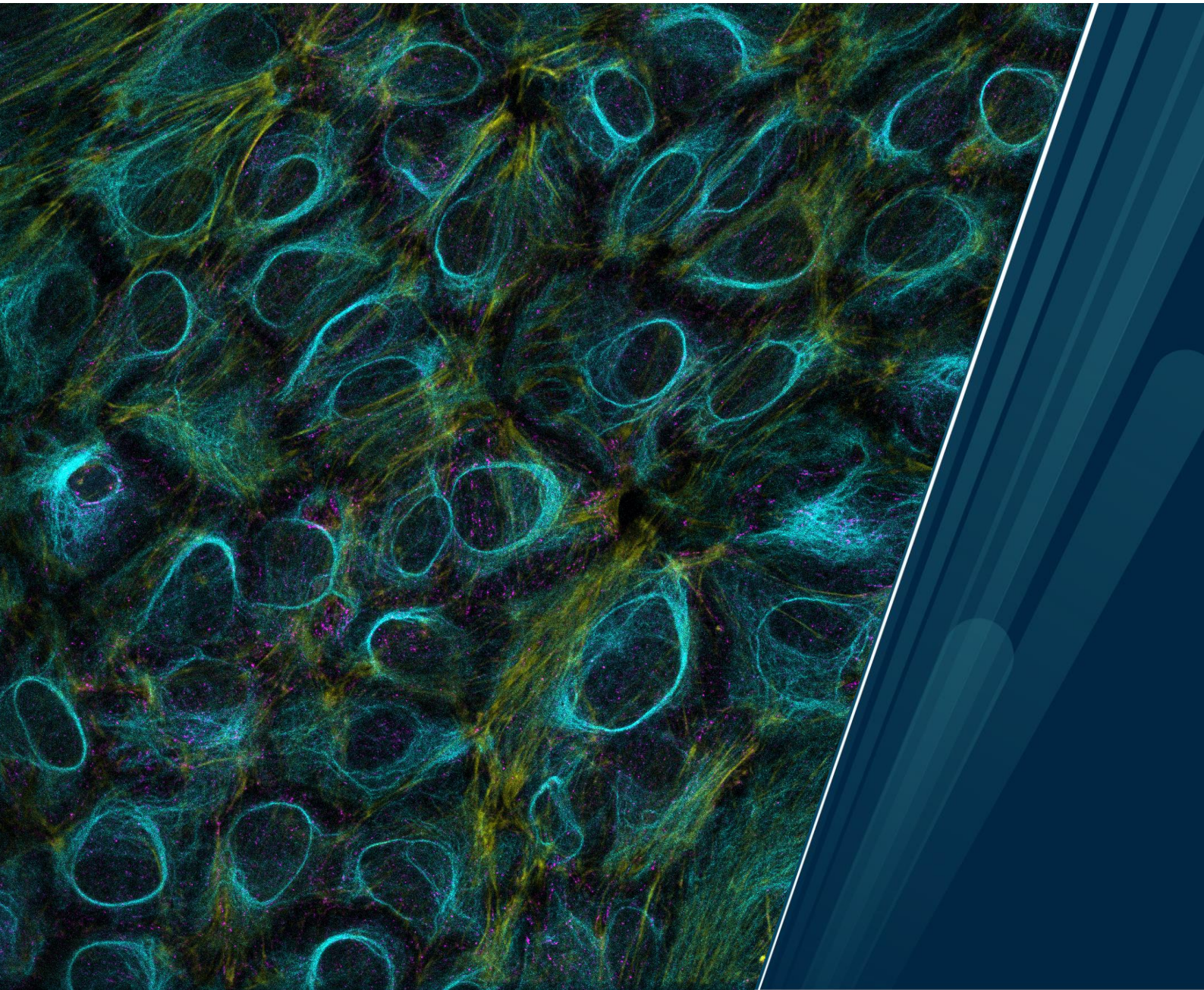
Deciphering human cell type enriched transcriptomes across tissue types and the functional study of the endothelial enriched protein KANK3

Cell type profiling using bulk RNA sequencing data to generate candidates for functional investigation

Sofia Maria Öling

A dissertation for the degree of Philosophiae Doctor

September 2023




THE ARCTIC UNIVERSITY OF NORWAY
UIT

Table of Contents

Acknowledgments

I would like to start with saying that I am truly grateful for this project opportunity together with the most amazing team and friends. This project was brought together under the supervision of Lynn Butler and Philip Dusart, in addition to the institutions and funding agencies that provided the financial means and resources. I want to give the most heartfelt thanks to the wonderful TVR/CAP group and UiT.

"Everything is awesome, everything is cool when you're part of a team, everything is awesome when you're living out the dream" – Tegan and Sara, the LEGO movie

Lynn, I am incredible grateful to you for this amazing opportunity, it has truly been life changing! You are a great leader and scientist that has accomplished the wonderful achievement of creating a small work-family group – the unwavering support, encouragement, patience, and empathy deserves a standing ovation!

Phil, I have learned so much from you and this thesis and journey would definitely not have been the same without you. Thank you for your constantly sharing your knowledge, for providing guidance and for your friendship. I have gone from being your master's student to now completing my PhD thesis. Also, thank you for not only teaching me about science but also about tea, and together we will always stand tall in support of team tea, we cannot let the coffee drinkers win!

Jacob, I would not have been here if it was not for your presentation about VEBIOS at KTH all those years ago. It was in a moment in time when I was truly curious about why I ('out of sooo many') had a pulmonary embolism, and thanks to the opportunity of being a master's student at CAP I have learned so much. Thank you!

Eike, you have definitely made this journey a real adventure, and most days should definitely start with a chocolate croissant! We have conducted cord extractions during a heatwave, without working ventilation and no water breaks – I don't think I have ever laughed so much, if someone would look up the definition of 'hilarious' they would see us doing that cord extraction. Thank you for being a true friend, for venting together,

for all the gossip, for all the ‘oopsies’ and for welcoming me into your life and family. Fun fact: Eike invented the color yellow. Lucifans for life!

Marthe, I am really grateful for getting to know you, from showing me Western blot, to trying to figure out that pesky PCR machine together, and to searching for reference transcripts. I also would never have attempted hiking if it wasn’t for you, where I learned that a ‘Norwegian easy’ hike is definitely not the same as ‘easy’, and to attempt skiing again – let’s just say that one of the adventures was easier than the other.

Clément, Maria, Jeong and Vera the TVR/CAP family would not be complete without you. Clément, thank you for your guidance and help in the lab, for pre-meeting meetings and impromptu coffee breaks. Maria, thank you for always being there to lend a helping hand, be it with diluting samples to running around the lab in hunt for yellow boxes. Your good mood brightens the day! Jeong, thank you for all the scientific discussions, and all the amazing Korean food! Vera, thank you for joining our team and making it complete, and for doing the tough job of keeping track of us in the lab.

Moving to another country to embark on a several years long PhD journey would not have been possible without friends to make a new country feel a bit more like home.

Larissa, you and Eike are truly amazing people. During the pandemic you opened your home and hearts, and together you made a really tough time feel joyous. We formed our own little cohort and had everything from movie nights to trips to Sommaröy and Senja.

Karolina and Christopher, thank you for all the laughter, the game nights and all the red soup, ice bears and Hammerfest.

Tine and Olivia, I am so thankful for getting to know you both and for showing the importance of standing up for what you believe in.

Casper, sharing office with you and Eike was an experience, to say the least. Some mornings I would walk past, thinking you two had a loud argument, but you were only discussing Swedish music. Thank you for teaching me all the dad jokes!

Kajsa-My, thank you for welcoming me into your life – you have made Sweden feel like home! Your sticktoitiveness, determination and knowledge are inspiring.

There is no doubt when I say that this journey would not have been possible without the constant support from my family.

Daniel, you are my love, my life, my person. This complete journey would not even have started if it was not for you believing in me, supporting me and motivating me. We both knew that embarking on this journey would mean that we would live apart again, but neither of us could have predicted that a pandemic would make traveling impossible – but we made it. They say "*home is where the heart is*", and you definitely got mine, you and our 'furbaby' Mochi.

My parents Peter and Christina, tack för all er kärlek, all ert stöd, för allt ni lärt mig, för allt ni visat mig. Jag älskar er!

My sisters Kajsa and Anna-Stina, de bästa systrar man kan tänka sig.

Sofia Öling, Stockholm 2023

List of papers

I. A human stomach cell type transcriptome atlas

S Öling, E Struck, MN Thorsen, M Zwahlen, K von Feilitzen, J Odeberg, F Pontén, C Lindskog, M Uhlén, P Dusart, LM Butler

Pre-print: doi: https://doi.org/10.1101/2023.01.10.520700

BMC Biology *Resubmitted following revision*

II. A human colon cell type transcriptome atlas

S Öling, E Struck, MN Thorsen, M Zwahlen, J Odeberg, F Pontén, M Uhlén, P Dusart, LM Butler

Manuscript

III. A tissue centric atlas of cell type transcriptome enrichment signatures

P Dusart, **S Öling**, E Struck, M Norreen-Thorsen, M Zwahlen, K von Feilitzen, P Oksvold, M Bosic, MJ Iglesias, T Renne, J Odeberg, F Pontén, C Lindskog, M Uhlén, LM Butler

Pre-print: doi: https://doi.org/10.1101/2023.01.10.520698

Manuscript

IV. KANK3 is a shear stress regulated endothelial protein with a role in cell migration and tissue factor regulation

E Struck, **S Öling**, P Dusart, MN Thorsen, J Eckel, L Kruse, CU Wahlund, J Odeberg, C Naudin, LM Butler

Manuscript

Abbreviations

A_{II}	Angiotensin-II
BMI	Body mass index
BSA	Bovine serum albumin
CAGE	Cap analysis of gene expression
CAT	Calibrated automated thrombinoscope
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
DNAse	Deoxyribonuclease
EC	Endothelial cell
ECL	Enterochromaffin-like
eNOS	Endothelial nitric oxide synthase
eQTL	Expression quantitative trait loci
ESAM	Endothelial cell-selective adhesion molecule
ESEL	E-selectin
EST	Expressed sequence tag
ET-1	Endothelin-1
F3	Coagulation factor III, Tissue factor
FACS	Flow-activated cell sorting
FCS	Foetal calf serum
FDR	False discovery rate
GI	Gastrointestinal
GIT	Gastrointestinal tract
GTEx	The Genotype Tissue Expression Project
H_2O_2	Hydrogen peroxide
HO	Hydroxyl radical
HUVEC	Human umbilical vein endothelial cell
ICAM1	Intracellular adhesion molecule 1
IL-1	Interleukin-1
IL-6	Interleukin-6
JAMs	Junctional adhesion molecules
KANK3	KN motif and ankyrin repeat domain-containing protein 3
LCM	Laser-capture microdissection
LDL	Low-density lipoprotein
LPS	Lipopolysaccharide
LSEC	Liver sinusoidal endothelial cell
MFI	Median fluorescence intensity
MPSS	Massively parallel signature sequencing
mRNA	Messenger RNA
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NGI	National genomics infrastructure Sweden
NO	Nitric oxide
O_2^-	Superoxide anion

PAI-1	Plasminogen activator inhibitor-1
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PECAM	Platelet endothelial cell adhesion molecule 1
PGI_2	Prostacyclin
qPCR	Quantitative PCR
Ref.T.	Reference transcript
RIN score	RNA integrity number
RIPA	Radioimmunoprecipitation assay buffer
RNA	Ribonucleic acid
RNAseq	RNA sequencing
ROS	Reactive oxygen species
rRNA	Ribosomal RNA
RT-PCR	Reverse-transcription PCR
SAGE	Serial analysis of gene expression
scRNAseq	Single cell RNA seq
SEL-P	P-selectin
SIM	Structure illumination microscopy
siRNA	Small/short interfering RNA
SNIC	Swedish national infrastructure for computing
SPEM	Spasmolytic polypeptide expressing metaplasia
spRNAseq	Spatial RNA sequencing
STRT-seq	Single-cell tagged reverse transcription sequencing
TCGA	The cancer genome atlas program
TF	Tissue factor
TFPI	Tissue factor pathway
TNFα	Tumor-necrosis factor α
tPA	Tissue plasminogen activator
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
VCAM1	Vascular cell adhesion molecule 1
VEGF	Vascular endothelial growth factor
VTE	Venous thromboembolism
vWF	von Willebrand factor
WGCNA	Weighted correlation network analysis

Summary

The identification of cell type-specific genes, which tend to have key cell type specific functions, and their modification under certain conditions is essential for our understanding of the human body. Single-cell RNA sequencing (scRNAseq) can be used to profile different cell types, but cell removal from tissue and processing can introduce artifacts, and some cell types are too fragile to analyze this way. The aim of this thesis was to elucidate cell type-specific transcriptomes within human organs, using unfractionated bulk RNA sequencing (RNAseq) data, and to use the data generated to select a candidate gene for functional analysis in endothelial cells (EC), which line the vasculature across tissues.

We used an integrative correlation-based method to analyze publicly available RNAseq data from the Genotype Tissue Expression Project (GTEx). The primary focus was on the gastrointestinal tract, where stomach (paper I) and colon (paper II) were analyzed. In both cases, we profiled cell-enriched transcriptomes of cell types that are absent from the major scRNAseq databases and identified several non-coding genes with cell type enriched profiles. A sex-based subset analysis revealed that cell profiles were broadly similar in males and females, but in both stomach and colon a small panel of genes were identified as cell type enriched in males only.

We extended our analysis to include 15 human organs (paper III). We identified co-enriched genes in related cell types, such as pancreatic alpha and beta cells, and stage-specific enrichment signatures during spermatogenesis. A cross-tissue analysis revealed common cell type enriched gene signatures between related cell types, such as those with motile cilia, or the same cell types found in different tissues, such as EC.

We subsequently selected one of the most consistently EC enriched genes across tissues for further study (paper IV), KN motif and ankyrin repeat domain-containing protein 3 (KANK3), which currently lacks any reports of function in this cell type. We found that KANK3 levels and its cellular distribution was regulated by shear stress. EC KANK3 depletion revealed a role in EC migration, and in the regulation of tissue factor (TF) expression, which is involved in blood coagulation.

These studies provide a roadmap for cell type enrichment profiles and demonstrate how such data can be used to select genes for functional investigation in a cell-type appropriate context. It also highlights the value of using alternative methods to extract information from previously published data.

1 Introduction

The first part of this thesis (papers I, II and III) uses an in-house developed method to extract cell type enrichment signatures from publicly available unfractionated bulk RNAseq tissue data, and the second (paper IV) uses this information to select a target gene for functional study in endothelial cells. The introduction covers methods used for cell profiling, information about the vascular endothelium, and a focus on the key organs profiled as part of this thesis.

1.1 Transcriptomics

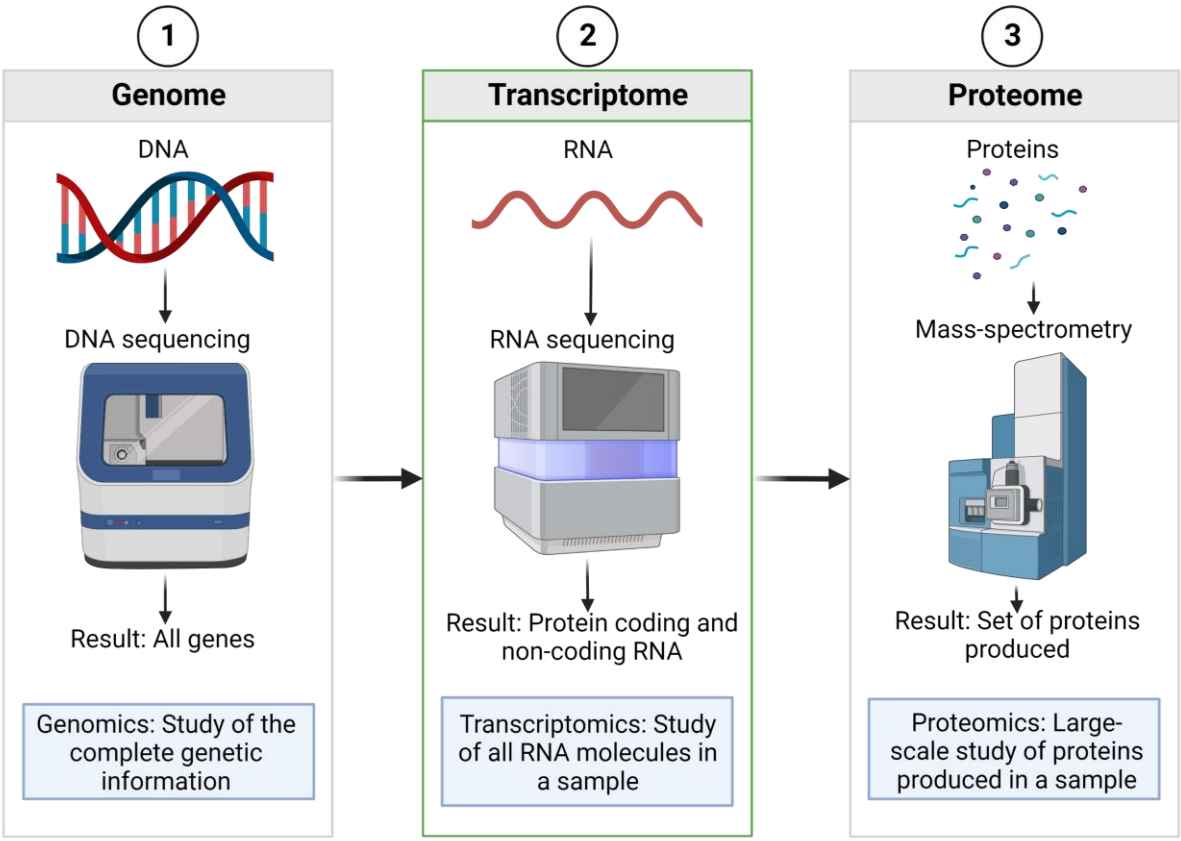


Figure 1: Schematic overview of the differences between the genome, transcriptome, and proteome as well as genomics, transcriptomics, and proteomics. (1) *Genomics is focused on the study of the genome – the complete genetic information.* ***(2)*** *Transcriptomics is centered around studying all RNA molecules in a sample – the transcriptome.* ***(3)*** *Proteomics studies the proteins produced in a sample – the proteome. Illustration created with Biorender.com.*

The entire human genome is constituted of approximately 20.000 protein-coding genes (accounting for approximately 1.5% of the genome), the rest is constituted of

non-coding genes [1–3]. While the genome is fairly constant within each individual (in all non-gamete cells), the transcriptome is extremely variable with each cell expressing different sets of genes depending on its tissue location, enabling them to perform distinct functions. The cellular transcriptome further varies depending on factors such as cell cycle state, disease, drug exposure and ageing. The generation of a transcriptome is built around the central dogma in which the DNA is transcribed by RNA polymerase to create complementary DNA (cDNA), the mature transcripts are obtained after intron removal by splicing. The study of the collection of transcripts is called transcriptomics. The understanding of the transcriptome is essential for interpreting the genome, understanding of cell mechanisms and the development of diseases. The transcriptome consists of both protein coding and non-coding RNA (earlier often referred to as “junk” DNA [4]), however recognition of the importance of the non-coding RNA is growing as recent studies have shown that they can perform highly distinct functions without coding for a protein [5,6].

After the development of automated DNA sequencing in 1980s [7], various methods using expressed sequence tags (ESTs) were developed with the possibility to rapidly sequence expressed genes (or parts of) [8,9], however cost and technical limitations prevented the identification of complete transcriptomes. Various tag-based methods, such as Serial Analysis of Gene Expression (SAGE) [10], Cap Analysis of Gene Expression (CAGE) [11] and Massively Parallel Signature Sequencing (MPSS) [12] were developed as a complement to the EST-approach. Using these methods, quantification of unique transcripts on a gene-level was made possible, however cost limitations prevented large-scale applications. Large-scale identification of transcripts was made possible with the development of hybridization-based microarrays in the late 1990 to early 2000 [13–15]. The principle behind hybridization-based assay is that RNA from a sample of interest is harvested, reversely transcribed, fragmented and labelled and allowed to hybridize (bind through Watson and Crick base pairing) to known sequence probes that have been fixed onto a microarray [16]. By comparing hybridization patterns between different samples one can identify mRNA sequences that have different abundances.

Unlike previous methods, hybridization-based assays made it possible to identify the different RNA transcript splicing isoforms [17]. The formation of various RNA splicing isoforms occurs during a process called alternative splicing, in which exons are combined in different formations and introns are removed. As the method

relies on hybridization, the output signal is often noisy as cross-hybridization and hybridization strength varies. Furthermore, identification of novel transcripts or isoforms is not possible as the hybridization sequence probe is pre-defined [18,19].

The complete characterization of the RNA transcripts produced by a cell was made possible with the development of RNA sequencing methods [20–22], the first sequencing method to offer high-throughput quantitative analysis of the complete transcriptome [23,24]. Unlike earlier methods, RNA sequencing is not limited to the detection of known transcripts, and can capture the entirety of transcribed RNA, which enabled the detection of novel transcripts and with the capacity to measure a wider range of expression levels with lower background noise.

Whereas RNA sequencing methods traditionally measure the transcripts from a large number of cells (bulk RNA sequencing), single cell sequencing allows for measurements of the transcript levels of a single cell [25]. This makes it possible to generate complete transcriptomic maps of individual human cells [1], and to understand individual cellular response to drugs and drug resistance in cancer treatment [26] or to study the differences in the immune cell population in healthy and diseased states [27]. However scRNAseq comes with some practical and technical challenges such as efficient cell isolation, material amplification, cost and data interpretation [28–30]. Additionally, the complex nature of scRNAseq can be very sensitive to the act of sample processing itself, and the changes it can cause on individual cell transcriptomes [31]. While both bulk and single cell sequencing methods are used to study tissue and cell populations on a transcript level, both methods lack spatial information. The development of spatial RNA sequencing (spRNAseq) methods links transcriptome data with complete spatial information, enabling the localization of gene expression events within tissues [32].

1.1.1 Genomics

Since the completion of the human genome sequence, efforts have been made to fully understand the complete information written in the DNA sequences. Several, extensive, genome-wide studies have been conducted to determine gene function, products, interactions and potential pathological implications [33–36]. Development of genomics approaches have made it possible to detect both coding and non-coding variants within the DNA sequences and have contributed to gain further insights into

the origin of species [37], the understanding of conserved genes [38] or to gain information about the gene function in coding and non-coding regions [39].

The applications for genomics-based approaches will not be covered in depth in this thesis, but it is mentioned to clarify the distinction from transcriptomics.

1.1.2 Proteomics

In contrast to genomics and transcriptomics, the proteomics field is focused on the functional aspects of gene expression by studying the individual proteins produced by a cell, organism, or tissue. It complements genomics and transcriptomics approaches by measuring the protein identity, protein structure or function. Proteins perform a wide range of intracellular functions within each organism and abnormal expression can disturb the normal cellular functions [40]. Studies have shown that changes in gene expression on an mRNA or DNA level does not necessarily implicate a change in the protein level, and that a change in protein level does not always lead to a change in DNA or mRNA level [41,42]. Furthermore, the proteome is under constant changes as it adapts to change in external stimuli and post-translational modifications [43]. There is a wide application of proteomics approaches in clinical settings to identify biomarkers that can distinguish between healthy and diseased subjects [44]. Proteomics has also been used to identify possible vaccine targets, for instance for the malaria parasite *Plasmodium falciparum* [45,46]. Proteomics approaches are commonly combined with immunohistochemistry (IHC), a technique in which antigen-targeting antibodies can be used to determine protein (antigen) localization within cells and tissues using microscopy methods [47]. Another example is the use of antibody microarrays, in which antigen-targeting antibodies are immobilized onto a suitable surface, following a blocking step, the sample is added to the array where the immobilized antibodies can bind the target proteins. By using fluorescent labelling or by a secondary detection antibody, the results can be measured by quantifying the signal intensity [48].

Further applications will not be covered in this thesis.

1.2 Bulk sequencing

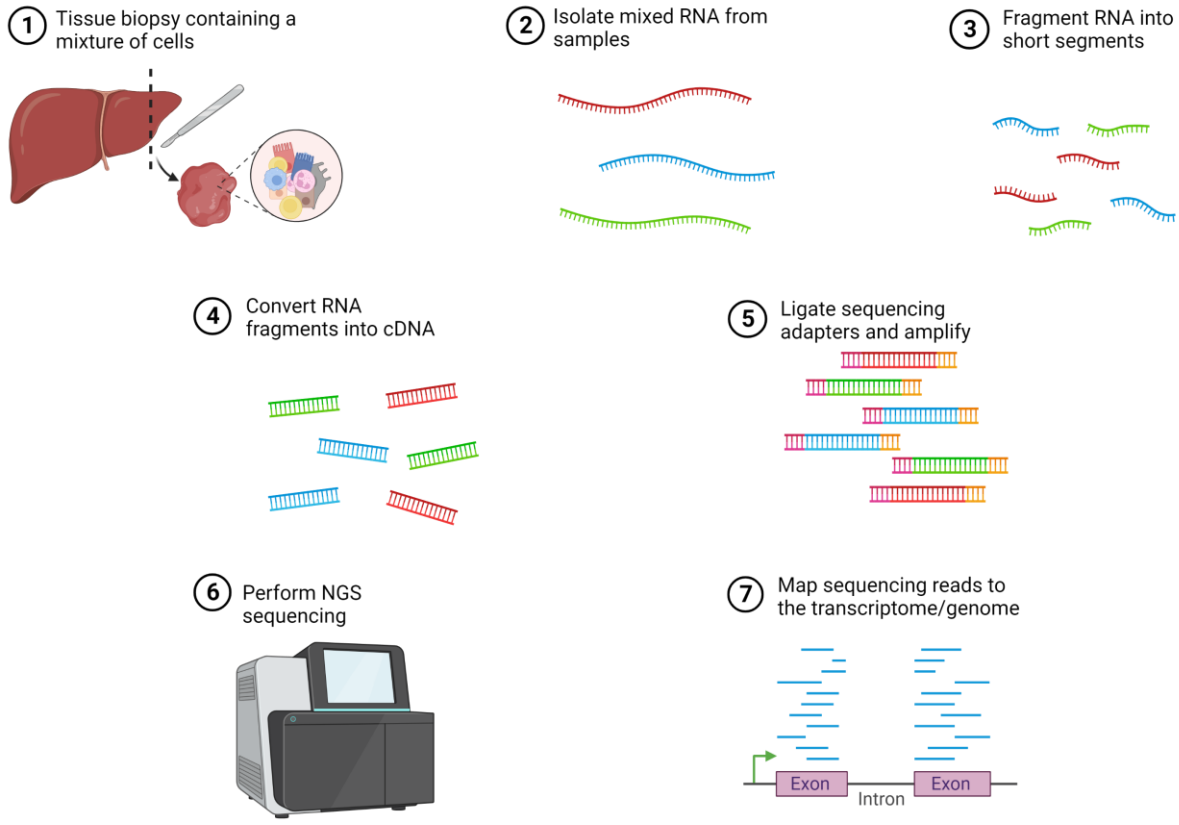


Figure 2: Simplified overview of the experimental steps in a bulk RNA sequencing protocol. (1) *First a tissue biopsy is obtained containing a mixture of all constituent cell types.* ***(2)*** *Mixed RNA sample is isolated and* ***(3)*** *fragmented, followed by* ***(4)*** *generation of complementary DNA library (cDNA). Following an* ***(5)*** *amplification step, the cDNA is* ***(6)*** *sequenced and* ***(7)*** *mapped against a reference transcriptome or genome. Adapted from "RNA sequencing", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

The first step in a standard RNAseq experiment is extraction and purification of the sample RNA e.g., from whole tissue, organism, organoid or cell culture. Bulk RNA sequencing analyses a mixture of RNA, meaning that the expression profiles are averaged across all cell types present in the sample. The sample RNA is usually fragmented, before being converted into a complementary DNA (cDNA) library. The cDNA fragments are then ligated with sequencing adapters and amplified, followed by analysis by a high-throughput sequencing technique (commonly used platforms are Illumina IG [21,22,49], Applied biosystems SOLiD [20] and Roche 454 Life Science [50,51]). Lastly, the reads can either be mapped to a reference genome/transcriptome or assembled *de novo*. As an example, the human reference genome is a template genome that has been assembled by comparing and combining the DNA of multiple

people to which sequencing data can be compared to identify similarities or differences. Contrastingly, *de novo* sequencing is conducted without the aid of a reference genome, instead the fragmented sequences are assembled by finding common overlapping shared regions until the entire genome is assembled.

Sequencing strategies for bulk RNA is typically divided into two categories; short-read only library (35-6000 bp) and long-read only library (1.000-10.000 bp) [52]. The short-read only libraries are often used for the identification and analysis of differential gene expression [20,22,53]. The wide use of this method for detection of transcriptomes is due to its robustness, cheaper cost and high-quality data output [54]. However, with the increased interest in isoform detection from longer transcripts, it is clear that short-read only libraries face difficulties in determining which isoform is present when given multiple options [55]. This drawback comes from the lack of protocol scalability to whole-transcriptome analysis [56,57]. The wide range of computational methods for data analysis of RNAseq data can also lead to bias [58–61]. To overcome this bias, the method of tagging the cDNA with unique molecular identifiers (UMIs) was developed [62,63]. The limitations of short-read libraries can be overcome with the use of a whole transcriptome library approach, or long-read cDNA sequencing. As long-read sequencing methods commonly deplete the rRNA, the method overcomes some of the limitations associated with short-read sequencing; such as sequencing mapping ambiguity, possible analysis of longer transcripts and reduced fraction of false-positive splice junctions [64]. As the method analyses longer transcripts, it makes it possible to confirm earlier gene predictions and to discover previously unannotated transcripts [65,66]. However, the method is still biased towards sequencing of short transcripts as they diffuse more quickly to the active surface than the longer transcripts [54], this can be overcome by modifying the sample loading conditions [67] or by the use of long-read direct RNA sequencing [68]. While long-read methods overcome the major limitation of short-read methods – namely the read length, the method does not offer the same read-depth as short-read methods, it also comes with a higher error rate [69]. It is therefore important to factor in whether read-depth or read-length is the more important factor in the RNA-seq analysis. Read-depth refers to the number of times that each base has been sequenced and read-length refers to the number of base pairs in the sequenced fragment.

One of the main advantages of RNA sequencing methods, compared to hybridization-based approaches, is the possibility to sequence and analyze novel

organisms and genes, which makes it possible to study non-model organisms at the gene-level and provide essential information in the field of biomarker discovery. RNA-seq techniques are frequently used in clinical oncology where the detection of cancer biomarkers has been instrumental in the disease diagnosis, prognosis and outcome prediction [70–73]. Additional advantages of RNAseq compared to DNA microarrays include detection of the precise intron-exon boundaries and single-nucleotide polymorphisms, reduced background signal, increased levels of quantification, high reproducibility and use of lower sample volumes [23].

RNAseq does come with several limitations, the first of which is related to library construction. Despite the development of long-read library methods, they cannot analyze infinite lengths and the fragmentation methods used can introduce transcript bias [21,22]. There are additional limitations related to the storage, retrieval and processing of the data, which has led to the development of several analytical approaches [60]. Another important consideration is the importance of sequence coverage, which has an additional cost implication. Sequencing coverage is defined as the number of reads that are aligned to a reference and therefore cover a known sequence. To detect rare or novel transcripts a higher sequencing coverage is needed, which comes with added cost. Lastly, higher coverage is needed to adequately cover large and complex transcriptomes.

1.3 Single-cell sequencing

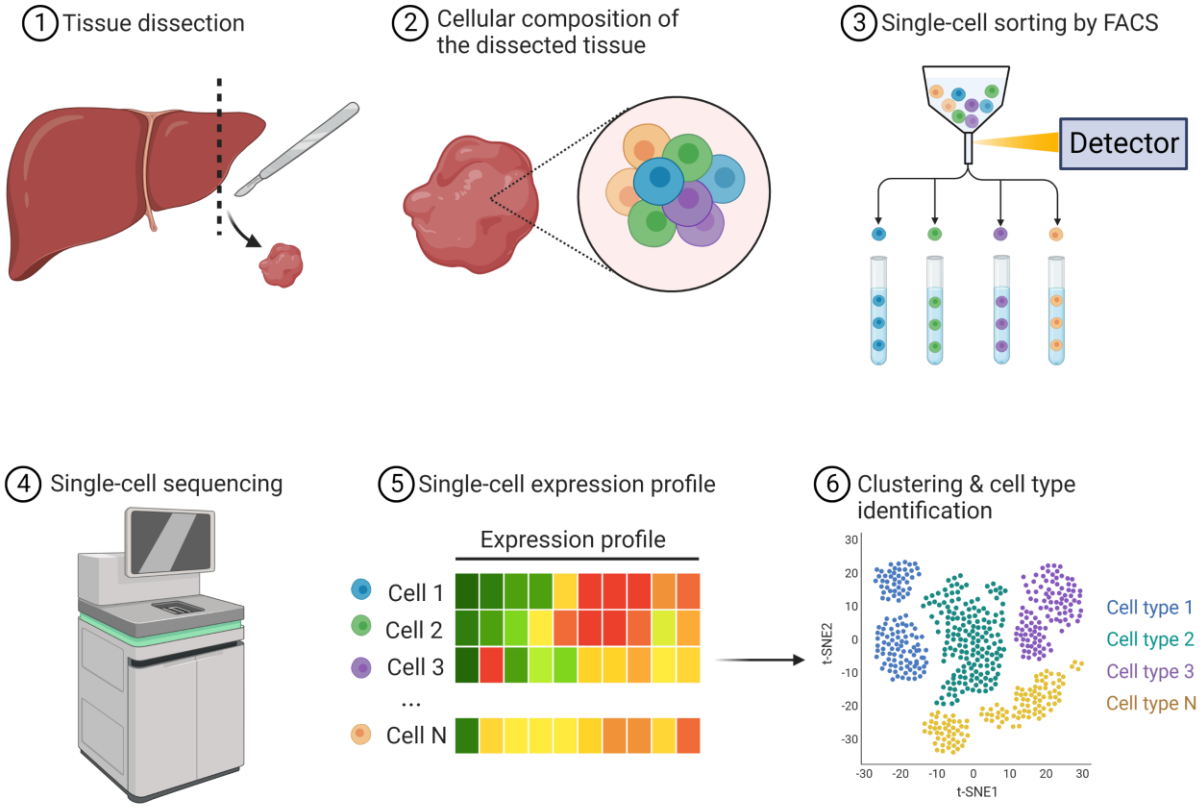


Figure 3: Simplified overview of the steps in a single-cell RNA sequencing protocol. *Following **(1)** tissue dissection, a **(2)** mixed cell sample is obtained. After **(2)** identification of markers for constituent cell types, **(3)** single cell separation can be accomplished by various methods, such as FACS. The individual cells are then **(4)** sequenced, generating **(5)** unique expression profiles for each sequenced cell. **(6)** Based on expression similarity the cells can be clustered and cell types identified. Reprinted from "Single-cell sequencing", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

The first instance of single-cell sequencing (scRNAseq) was reported in 2009 with the isolation and sequencing of individual oocytes [74]. In contrast to bulk RNAseq, which focuses on identifying global gene expression from a pool of multiple cells, single-cell sequencing investigates the transcriptome of each cell individually. While bulk RNAseq revolutionized the biological understanding of tissues, cells and organisms, and provided tremendous information about the pathology of diseases, the method lacks cellular resolution, and much complex spatial information is unresolved. scRNAseq allows for studies of the transcriptome of individual cells, overcoming one of the major limitations with bulk RNAseq. Since the first publication in 2009, further development of single-cell methods such as; tagged reverse transcription sequencing (STRT-seq) [62], inDrop [75], Drop-seq [76] and the launch of the 10x genomics

platform [77] has allowed for a wider adaption of the method and reduced the cost. Aspects common to all the methods are that they require solid tissue dissection, identification of the cellular composition, separation of the single cells and RNA labelling and amplification before sequencing.

Single cell RNAseq experiments start with dissection of the solid tissue sample to prepare the single-cell suspension, commonly with collagenase and DNase to produce a high yield of RNA [78]. The individual cells are then separated by various methods. The early methods of separation by limiting dilution and micromanipulation were time-consuming and very low throughput [79,80]. The development of the now popular flow-activated cell sorting (FACS) method enabled specific sorting of distinct cell populations at a high-throughput rate and is one of the most common strategies [81], however varying the method of sample preparation can yield very different results and complicate reproduction of results [82]. Microfluidic technology further revolutionized scRNAseq as it allows for low sample consumption, low analysis cost and greater fluid control [83]. Microfluidics offer different types of platform for single-cell isolation such as; capture on microfluidic chips [84], loading into nanowell systems [85] or capturing cells into individual droplets using inDrop [75] or Drop-seq [76].

Following isolation of single cells is the library preparation starting with cell lysis, reverse transcription, and cDNA amplification. One of the main challenges of scRNAseq methods is that due to low mRNA capture rates, only a low percentage (10-20%) of transcripts will be reverse transcribed [62,86]. In-depth analysis of full transcriptomes therefore further requires profiling of a large number of cells, increasing the method cost. Earlier methods overcame this issue by either focusing on the 5’ or 3’ transcript end [87,88]. More recently, the incorporation of unique molecular identifiers (UMIs) or barcodes (a short random basepair (4-8bp) sequence) in the reverse transcription step removed the PCR bias and improved accuracy as each read can be assigned to its original cell [76,87,88]. However, the UMI based methods are not suited for detection of transcript isoforms as they are limited to sequencing only either the 5’ or 3’ transcript end.

Enormous amounts of data are generated from scRNAseq compared to bulk RNA-seq, as thousands of individual cells are sequenced, leading to challenges with data handling and processing and the associated hardware and software requirements. Open-source software tools have been developed within the scRNAseq community to help overcome some of the issues related to data processing [89–91].

However, the issue still remains that scRNAseq data analysis requires complex bioinformatics knowledge and techniques [80], posing challenges with data interpretation [28,92]. Additional limitations are related to the potential modification of cell transcription in profiled cells by removing them from native tissues and exposure to the subsequent processing [93–95]. Cost restrictions typically lead to analysis of a limited number of biological replicates, further leading to an underestimation of biological and cellular variance which increases the likelihood of false discoveries [96,97].

The development of single-cell sequencing methods has made it possible to uncover rare cell populations within tissues, to study the effect of drugs and metabolites on individual cell types and study regulatory mechanisms between cell types. The application to novel biological questions contributes to rapid advances, making recent reviews outdated [98]. The power of scRNAseq to resolve cell-specific transcriptomes on a high-throughput basis is driving large-scale cell atlas projects, such as The Human Cell Atlas [1] and the NIH Brain Initiative [99].

1.4 Spatial RNA sequencing

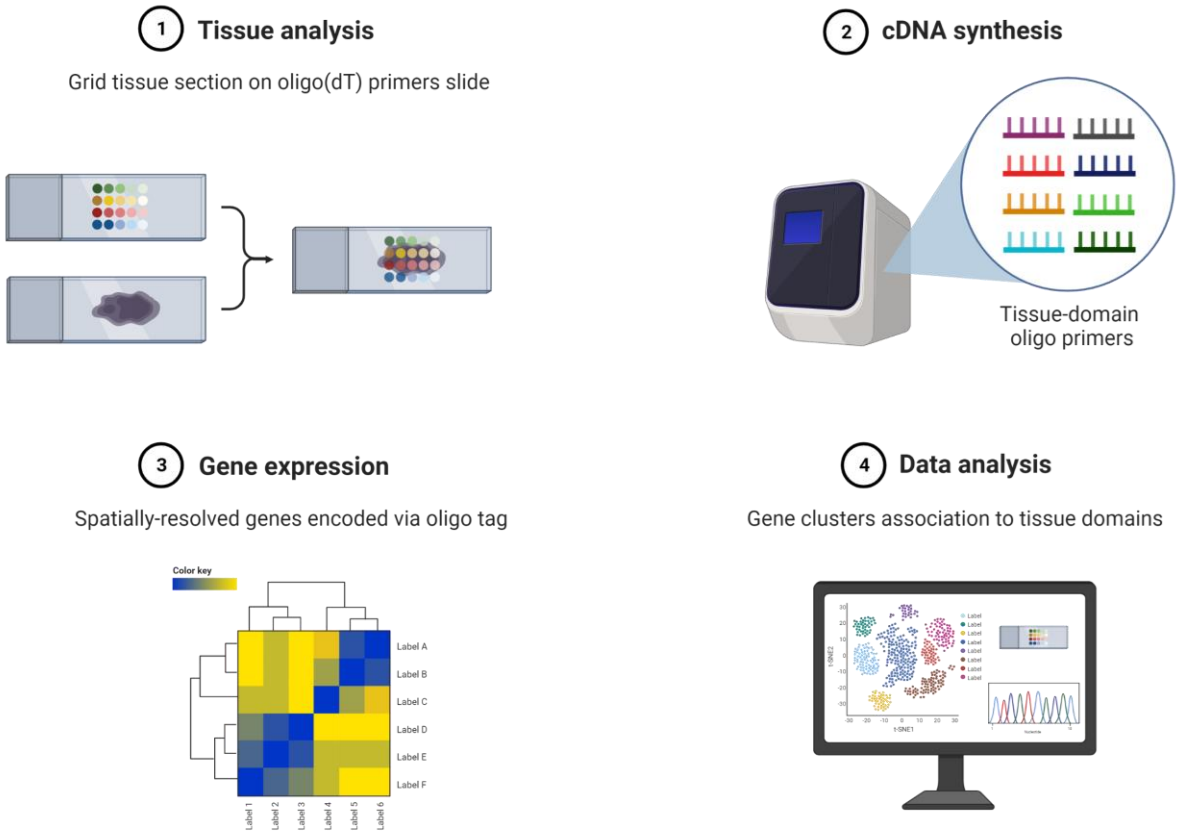


Figure 4: Simplified overview of the spatial encoding approach used for spatial RNA sequencing. (1) *The frozen tissue section is overlayed with the oligo-coated bead microarray. After* ***(2)*** *cDNA synthesis the* ***(3-4)*** *spatial information can be resolved using the oligo tag information. Reprinted from "Spatial transcriptomics", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

The recent development of spatial RNA sequencing (spRNAseq) provides whole transcriptomic data with spatial information. This is accomplished by combining techniques from bulk RNA-seq and *in situ* hybridization [32]. There are currently two different approaches for spRNAseq. First, the 'spatial encoding' methods can either record the spatial information of transcripts during library preparation by isolating spatially restricted cells by laser-capture microdissection (LCM) [100] or by barcoding techniques [101,102]. The second approach is built on '*in situ* transcriptomics' to generate sequencing information within tissue sections or by imaging RNA in cells [103,104].

The use of LCM techniques has been successful in the isolation and profiling of individual cells [100]. However, the method requires highly specialized equipment and

is difficult to scale. The different barcoding approaches [101,102] capture mRNA directly from frozen tissue sections by applying it to a oligo-coated bead microarray. The oligo-coated beads correspond to a specific barcode which can uniquely identify transcripts and their location. The sequences can then be tracked back to the slide coordinates to provide the spatial information. These barcoding methods do not require as much specialized equipment as LCM and are easier to scale. However, they can only be applied to fresh frozen tissue and the resolution is highly limited by the array size and the spacing of the oligo coated beads.

The alternative approach of '*in situ* transcriptomics' involves either *in situ* sequencing or imaging of transcripts visualized using single-molecule fluorescence hybridization [103,104]. In contrast to the earlier mentioned LCM techniques, this approach generates a much narrower transcriptome profile, however it instead allows for profiling of low-abundance transcripts [105] and provides subcellular information [106]. The limiting factor of these methods is the requirement for high-to-super-resolution microscopy methods and automated fluidics platforms.

The methods are still being developed and improvements are being made which have made it possibly to apply the technology to clinical samples [107], as well as to whole mouse embryo to track transcriptomic expression patterns during organogenesis [108]. spRNAseq is very likely to become adapted to the wider community if the technical limitations related to cost, resolution and lack of deep transcriptome data can be overcome.

1.5 Bulk RNAseq deconvolution methods

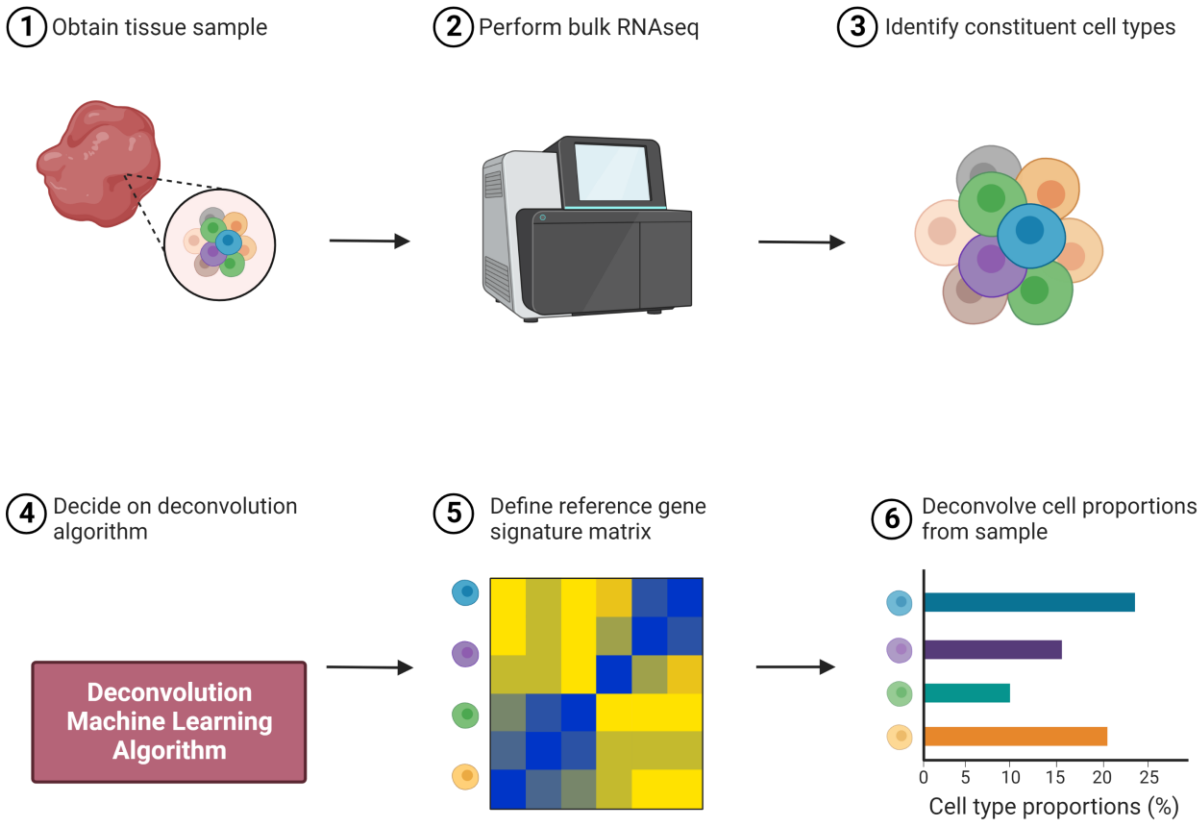


Figure 5: Simplified overview of the workflow for cellular deconvolution of bulk RNAseq data. (1) *First tissue samples are obtained, followed by* ***(2)*** *bulk RNAseq and* ***(3)*** *identification of constituent cell types.* ***(4)*** *After selection of deconvolution algorithm, (5) a reference matrix is defined.* ***(6)*** *Finally, constituent cell type proportions can be identified. Adapted from "Bulk RNA sequencing deconvolution", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

Most commonly used bulk RNAseq deconvolution methods are designed to estimate the proportions of the constituent cell types within samples using bulk RNAseq data e.g. CIBERSORT [109].

Tissue samples, both from healthy and disease tissue, are typically heterogenous as they contain a variable portion of cells and cell types. Moreover, bulk RNAseq results in tissue-averaged expression levels of each gene, thus the expression contribution of low abundant cell types within the sample can be masked as the RNA contribution by the more abundant cell types will be higher [110]. This limitation of bulk RNAseq has led to the development of multiple deconvolution methods to attempt to infer cell type abundance from bulk RNAseq data [111].

The mathematical problem that the various deconvolution methods try to solve can be defined as;" *in a heterogenous sample, the expression of an individual gene*

can be expressed as the linear combination of the contributing expression values from each constituent sample cell type, assuming that all cell types have similar gene expression levels across the samples" [111]. This mathematical problem can also be expressed using the matrix notation below.

$$T = C \cdot P$$

Where T = the measured expression values from a heterogenous sample; C = average expression values in the constituent cell types and P = relative composition of cell types in sample [111].

There are multiple available deconvolution algorithms to solve this problem such as; linear least square (LLS) [112], non-negative matrix factorization (NNMF) [113] and support vector regression approaches, for instance using CIBERSORT [109].

Following selection of a deconvolution algorithm, comes the identification of cell type-specific markers or expression profiles to define a reference gene signature matrix. A cell type-specific marker is considered as a gene whose expression is uniquely restricted to one specific cell type, with a stable expression across replicate samples [114]. However, this ideal definition must be modified to solve the deconvolution problem, as any given gene is rarely uniquely restricted to one specific cell type. The modified definition of a cell type-specific marker is therefore; a gene that is to a large extent expressed by a cell type than in others [111]. The selected maker genes will form a cell type-specific expression matrix, which will be used in combination with the selected deconvolution algorithm to identify the cellular composition of the sample.

There are several advantages of using deconvolution algorithms to resolve bulk RNAseq data, of which a major factor is that the methods are capable of taking advantage of the numerous large-scale transcriptome studies already available, and analyzing them to get cell type-specific resolution, rather than the need to generate new RNAseq data themselves [115]. Limitations include possible exclusion of cell types, as the methods rely on an assumption of sample constitution, as well as inability to identify novel cell types [111].

1.6 Integrative correlation-based analysis

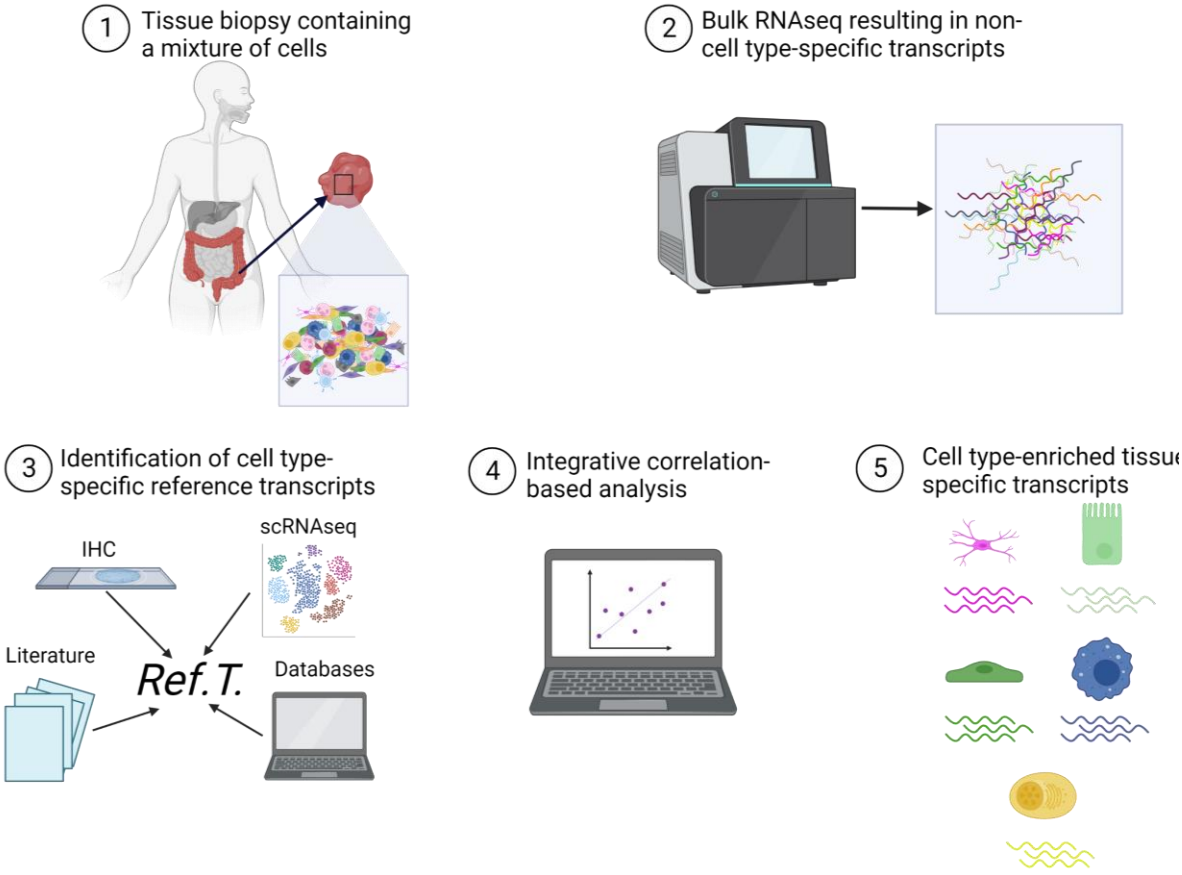


Figure 6: Simplified overview of the method behind integrated correlation-based analysis. *After* ***(1)*** *obtaining a tissue biopsy sample containing a mixture of cells, the sample is* ***(2)*** *sequenced for RNA. By* ***(3)*** *using cell type-specific reference transcripts in combination with* ***(4)*** *integrative correlation-based analysis,* ***(5)*** *cell type-enriched transcriptomes can be identified. Illustration created with Biorender.com.*

The integrative correlation-based method, which uses bulk RNAseq data to identify the transcriptomes of tissue specific cell types, was developed in our group and an early version was first published in 2016 [116]. Since then, our group has used modified versions of the method to identify cell type-enriched transcriptomes within individual organs, including the brain [117] and adipose tissue [118], both of which contain cell types that difficult to process for scRNAseq – neurons and adipocytes, respectively [119–122].

As the cells in the analyzed samples have not been removed from the tissue prior to sequencing, difficulties associated with the processing and artefacts associated with scRNAseq can be avoided [93–95].

Unlike scRNAseq and spRNAseq the integrated correlation-based analysis method does not require advanced bioinformatics expertise to resolve the constituent

cell type-enrichment profiles, and can be used on pre-existing RNAseq datasets [116–118]. In contrast to the aforementioned deconvolution methods, our integrative correlation-based method does not aim to calculate cell type proportions. Instead, the aim is to identify cell type-enriched transcripts. For more detailed information on analysis method see chapter 3.1 and 5.1.1-5.1.4.

1.7 Vascular endothelium

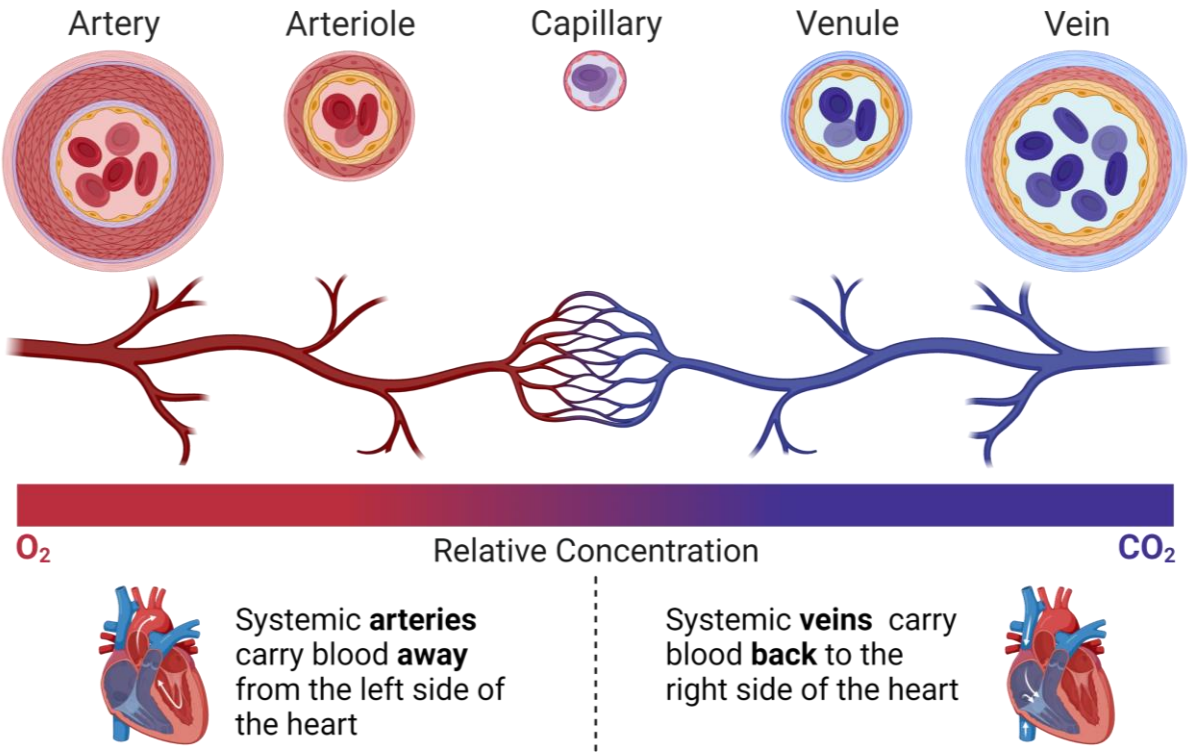


Figure 7: Illustration of the human vasculature. *Endothelial cells line the innermost surface of all the blood vessels in the body. Reprinted from "Systemic blood vessels", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

The vascular endothelium consists of a single layer of endothelial cells that separates the blood from the surrounding tissues [123]. ECs line all the blood vessels in the body where they have roles in several processes, such as regulation of hemostasis, thrombosis, and inflammation. They are known to be involved in cardiovascular diseases, where some endothelial-specific genes play a significant role, such as involvement in the coagulation cascade and thrombosis [116,124]. The

endothelial cells are connected to the basal lamina, below which is a layer of connective tissue and smooth muscle cells [125]. On the apical (or luminal plasma membrane) side of the ECs is the glycocalyx, a complex macromolecule network that provides ECs with a framework to interact with and bind plasma proteins [126].

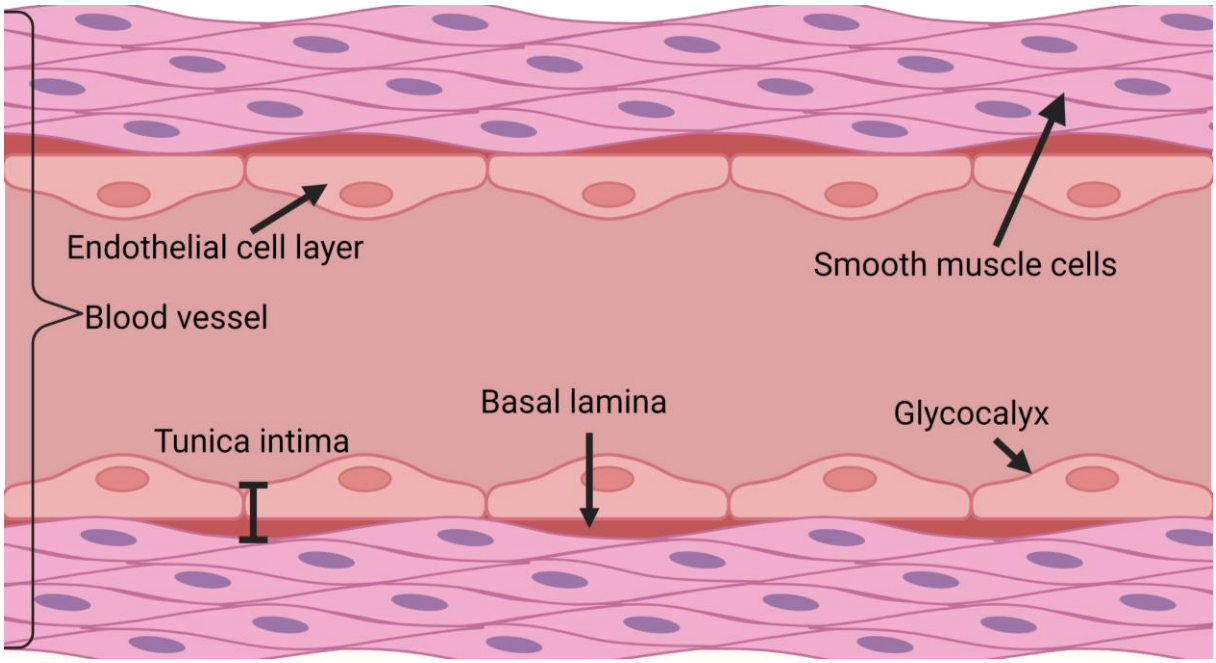


Figure 8: Illustration of a cross-section of the general vascular endothelium structure. *The endothelial cells that surround the blood vessel sit on top of the basal lamina and smooth muscle cells. The glycocalyx, located on the apical side of the endothelial cells, provides a macromolecule network for cell and plasma protein interactions. Illustration created with Biorender.com.*

Under normal conditions the endothelial surface is anti-thrombotic, inhibiting platelet attachment via production of nitric oxide and prostaglandins. Vascular injury can cause endothelial dysfunction and the vessel surface can become pro-coagulant [127], predisposing to thrombotic disease, such as venous thromboembolism (VTE).

1.7.1 Endothelial specific functions

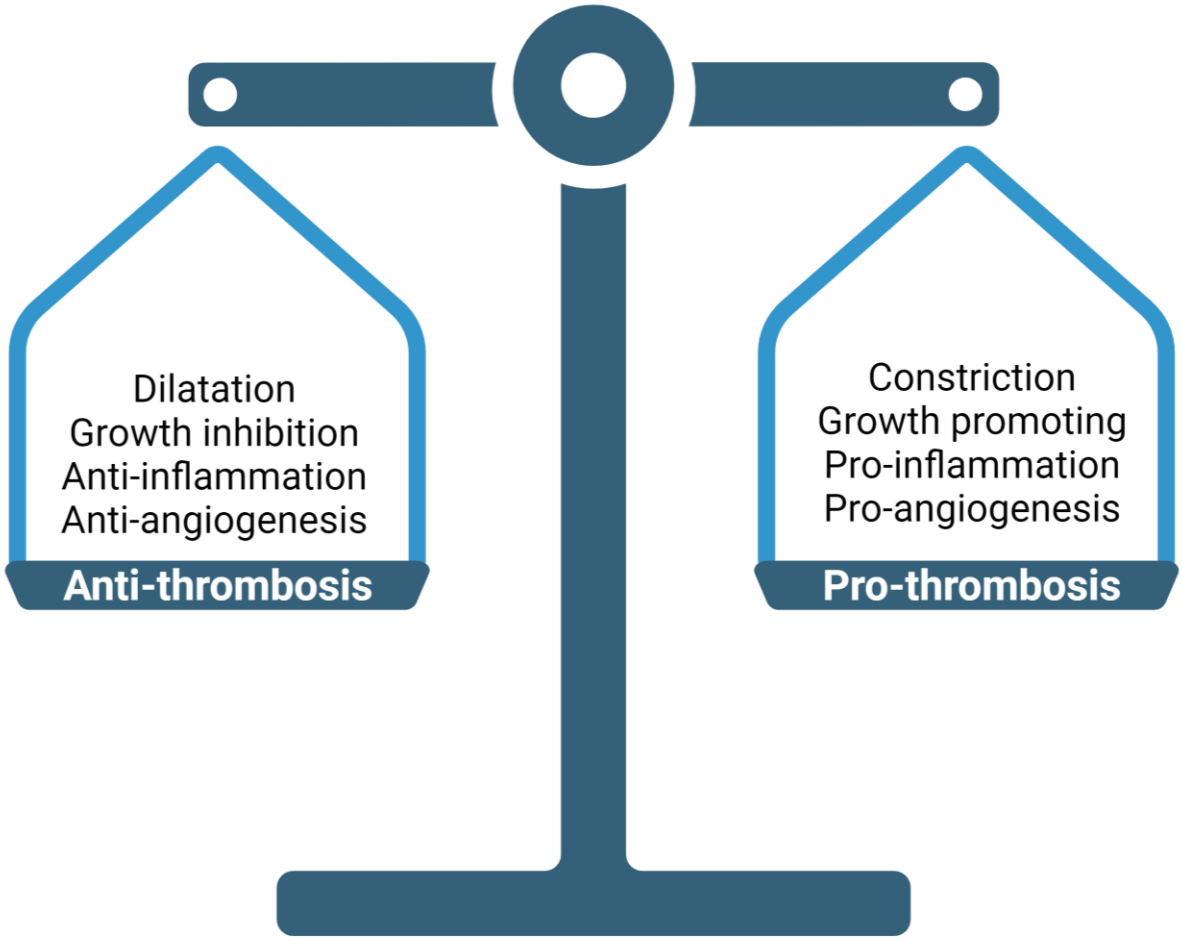


Figure 9: Summary of endothelial specific functions. *Endothelial cells maintain the hemostatic balance in the vasculature, such as the ratio of anti- and pro-thrombosis signals where excessive pro-thrombotic stimulus can lead to thrombosis. Illustration created with Biorender.com.*

The endothelium has a crucial function in the maintenance of hemostatic balance. The ECs are continuously exposed to the hemodynamic forces of the pulsatile blood flow. The cellular surface is mostly affected by shear stress forces, but the pulsatile changes create a cyclic strain on the entire vasculature by stretching the vessels. ECs change their morphology to adapt to the different hemodynamic forces and direction of blood flow. In steady-state blood flow, endothelial cells have an elongated morphology that align in the flow direction whereas a disturbed flow leads to a rounder shape with non-uniform cell orientation [128]. In addition to changing their morphology to adapt to the blood flow, ECs respond by altering their production of vasoactive substances. In response to increased shear stress, which contributes to

vasodilation, ECs increase the production of nitric oxide (NO) and endothelial nitric oxide synthase (eNOS) activity [129]. The increased eNOS activity mediates NO-driven vasodilation. Shear stress further induces altered expression of PGI_2 and ET-1 [129–132], which are important for regulating the vascular tone, by either inhibiting (NO, PGI_2) or promoting (ET-1) smooth muscle cell growth [133–135]. Furthermore, shear stress in the vessels contributes to the maintenance of the non-thrombogenic endothelial surface by stimulating the expression of thrombomodulin [136], heparin sulfate proteoglycans [137] and tissue plasminogen activator (tPA) [138].

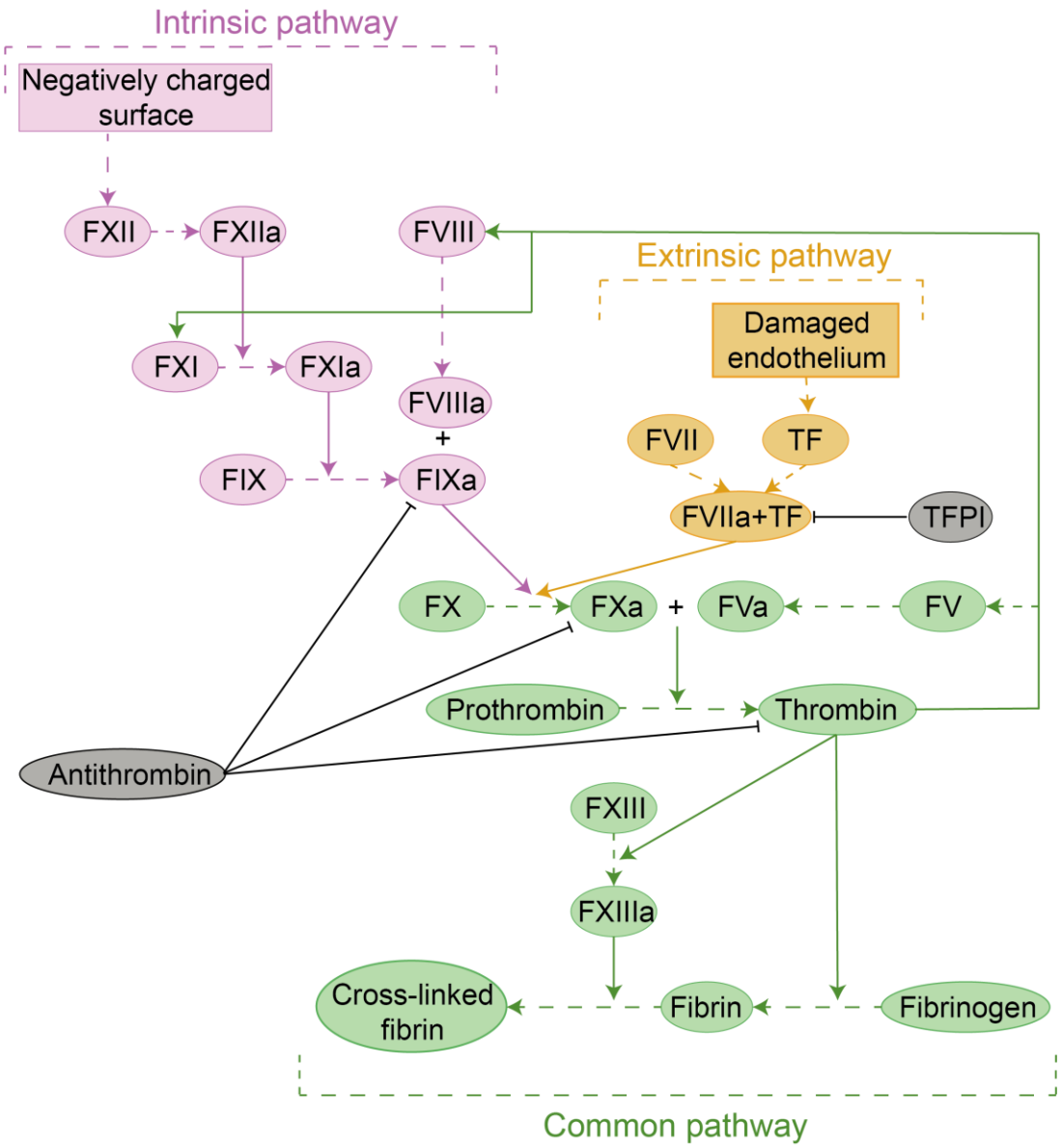


Figure 10: Diagram of the coagulation cascade. *Overview of the proteins involved in maintaining the hemostatic balance. The intrinsic and extrinsic pathway converge in the activation of FX, part of the common pathway, to generate thrombin as well as cross-linked fibrin.*

The vascular endothelium maintains a non-thrombogenic surface on the inner lining of the vessel wall, through the expression of inhibitors of platelet aggregation, NO, PGI_2, and heparin-like molecules (thrombomodulin, tPA) [139]. However, the endothelium can become pro-thrombotic through the expression of various co-factors for platelet adhesion, such as von Willebrand factor (vWF), fibronectin and thrombospondin, in addition to pro-coagulant factors like factor V [140]. Additionally, pathophysiologic stimuli of the endothelium can trigger the expression of EC tissue factor (TF), which in turn activates the coagulation cascade (Figure 15) [141,142]. The coagulation cascade can be described as three phases. During the first phase, the initiation, TF expressed by EC acts as a cofactor for factor (f) fVII and activates it into fVIIa which forms a complex with TF (TF/fVIIa). The complex then cleaves fX into its active form, fXa, which can then generate more thrombin [127]. During the following amplification phase, thrombin activates the adhered platelets to form a thrombus. Furthermore, thrombin can cleave fV into fVa, fVIII into fVIIIa, and fXI into fXIa[127]. In the third and final phase, the propagation phase, the activated platelet surface promotes additional thrombin formation. The prothrombinase complex is formed when fVa binds to fXa, and fVIIIa and fIXa form the intrinsic tenase complex. These two complexes contribute to increased thrombin, which generate fibrin from fibrinogen cleavage. In parallel, thrombin cleaves fXIII into fXIIIa which contributes to the formation of the protective fibrin mesh by cross-linking fibrin chains [127]. ECs can further reduce the rate of fibrin breakdown through expression of plasminogen activator inhibitor-1 (PAI-1), which is an inhibitor of the fibrinolytic pathway [143]. Once the vessel is healed, the fibrinolytic pathway will initiate thrombus dissolution by tissue plasminogen activator (tPA) or urokinase (uPA) generating plasmin from plasminogen on the thrombus surface [144].Thus, the vascular endothelium maintains the homeostasis between the pro-coagulant and anti-coagulant state of the blood vessels throughout the body.

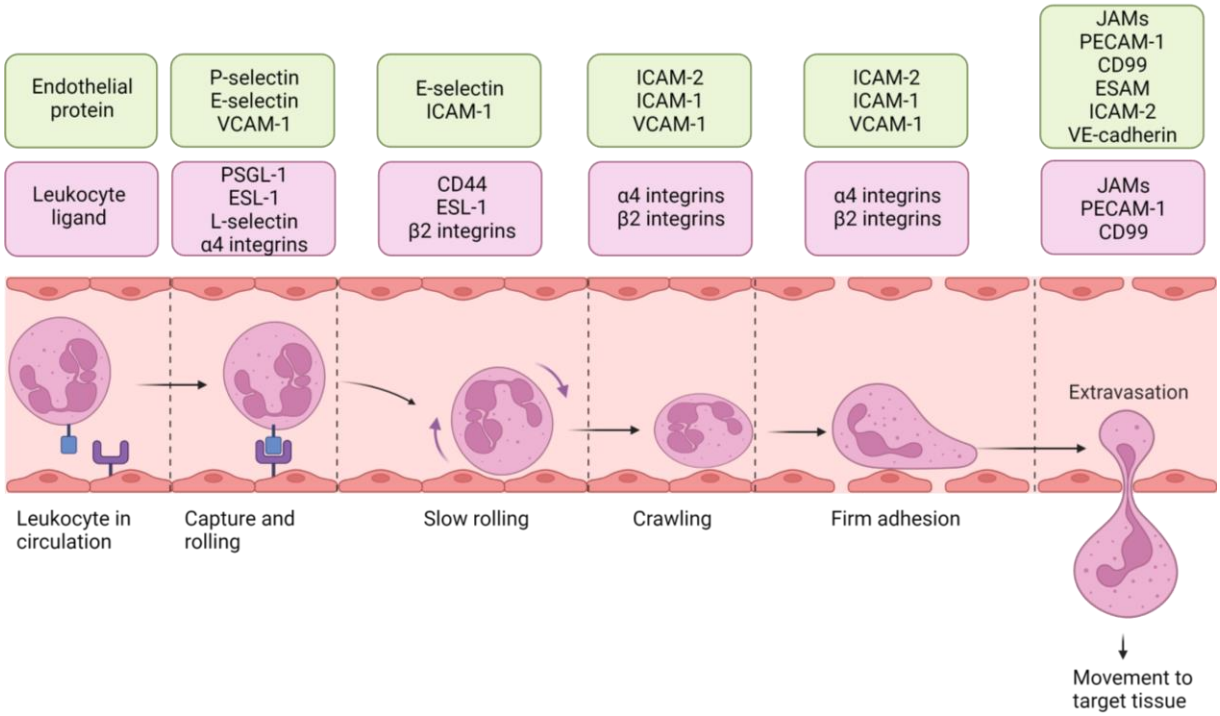


Figure 11: Illustration of leukocyte recruitment to site of inflammation by endothelial cells*. The free leukocytes present in the circulation are first captured to initiate the rolling and slow rolling process. By interaction with different endothelial specific proteins, the leukocytes crawl across the vascular surface to finally adhere and extravasate to the inflamed target tissue. Adapted from "Neutrophil recruitment pathway", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

Endothelial cells are involved in inflammation related processes, in which they capture leukocytes from the blood flow, activate and guide them to extravasation sites where they can pass through the vessel wall to reach the site of inflammation. Inflammation leads to expression of EC E-selectin and release of P-selectin from Weibel-Palade bodies, which mediate leukocyte capture [145,146]. The weak interaction between the leukocytes and EC selectins, binding to glycosylated structures on the leukocyte, allows them to roll along the EC surface. As the leukocytes roll across the endothelial surface, slow rolling and firm adhesion is achieved by further interactions with intracellular adhesion molecule 1 (ICAM1) and vascular cell adhesion molecule 1 (VCAM1) expressed by ECs, both binding to integrins on the leukocytes [147]. The leukocytes then transmigrate though the endothelial barrier using either paracellular or transcellular diapedesis, extravasation either through the space between two neighboring ECs or through the EC body, respectively. Several EC adhesion molecules are involved in the transmigration process, such as endothelial cell-selective adhesion molecule (ESAM) [148], CD99 [149], junctional adhesion

molecules (JAMs) [150] and platelet endothelial cell adhesion molecule 1 (PECAM1) [151]. These adhesion molecules are enriched at endothelial cell junction sites. Paracellular diapedesis has been confirmed as the major transmigration pathway for leukocytes [152,153].

Angiogenesis, the formation of blood vessels, occurs in two steps; during development angioblasts are differentiated from primitive mesodermal cells, followed by the formation of primitive blood vessels from the angioblasts. Angioblasts then differentiate into ECs during development. ECs will later cover the inner surface of all blood vessels in the body [154]. In adults, vascular endothelial growth factor (VEGF) has several functions in angiogenesis, where it targets ECs. VEGF promotes the growth of vascular ECs, acts as a survival factor by preventing EC apoptosis, and can induce both vascular leakage and vasodilation [155]. Vascular pruning, a process in which the vascular density is adapted to the needs of the tissue, is mediated by VEGF and by this process the vascular tone can be adjusted in order to match the supply of oxygen, where hyperoxia triggers vessel regression [156]. Hypoxia on the other hand induces VEGF leading to activated angiogenesis [157]. Additionally, VEGF controls the migration of endothelia at the tip of the angiogenic sprout and also the proliferation in the stalk [158].

1.7.2 Endothelial dysfunction

The endothelium separates the blood from the surrounding tissue and serves many functions, such as regulation of the vascular tone, leukocyte adhesion, angiogenesis and coagulation. Under normal conditions, the endothelial cells maintain a state of homeostasis in the vessels by producing various molecules that can tip the homeostatic balance in opposite directions, including vasodilators and vasoconstrictors, pro and anticoagulants, inflammatory and anti-inflammatory, oxidizing and antioxidizing and fibrinolytics and antifibrinolytics [159,160]. However, endothelial dysfunction can cause disturbance in the homeostasis and thus lead to a proinflammatory state, expression of pro-thrombotic peptides and reduced vasodilation. There are several potential causes for endothelial dysfunction including physical inactivity, smoking, diabetes and hypertension [161]. Endothelial dysfunction is associated with various diseases such as chronic heart failure, cancer and most forms of cardiovascular diseases [162].

Vasodilation is mostly stimulated by NO release from the EC caused by shear stress, and the vasodilation is proportional to the amount of NO released [163]. However, both low and high shear stress is associated with endothelial dysfunction, where a low shear stress leads to a pro-inflammatory state and high shear stress can cause endothelial erosion, plaque rupture and platelet aggregation. When ECs are exposed to a state of oxidative stress they produce the NO antagonist, angiotensin-II (A_{II}), that has a vasoconstrictive effect in addition to exhibiting a prothrombogenic, oxidizing and antifibrinolytic properties and furthermore, A_{II} increases leukocyte adhesion by upregulating the expression of adhesion molecules [164]. A_{II} synthesis can be triggered either directly by oxidative stress, or by stimulating NF-κB replication leading to production of TNFα, IL-1, IL-6 and adhesion molecules [165].

During oxidative stress, there exists an excess of reactive oxygen species (ROS) such as hydrogen peroxide (H_2O_2), hydroxyl radical (HO) or superoxide anion (O_2^-) [166]. The low-density lipoprotein (LDL)-cholesterol molecule is normally innocuous, but is easily oxidized to LDL-ox in states of oxidative stress which is highly immunogenic [167]. LDL-ox is known to activate the endothelium by causing a release of phospholipids [168]. Further, it leads to increased production of adhesion molecules, supporting leukocyte attraction and platelet aggregation [169]. Additionally, it increases the activity of proinflammatory genes, has a cytotoxic effect on the endothelium, provokes endothelial dysfunction, favors thrombogenesis and induces endothelial cell apoptosis [170,171].

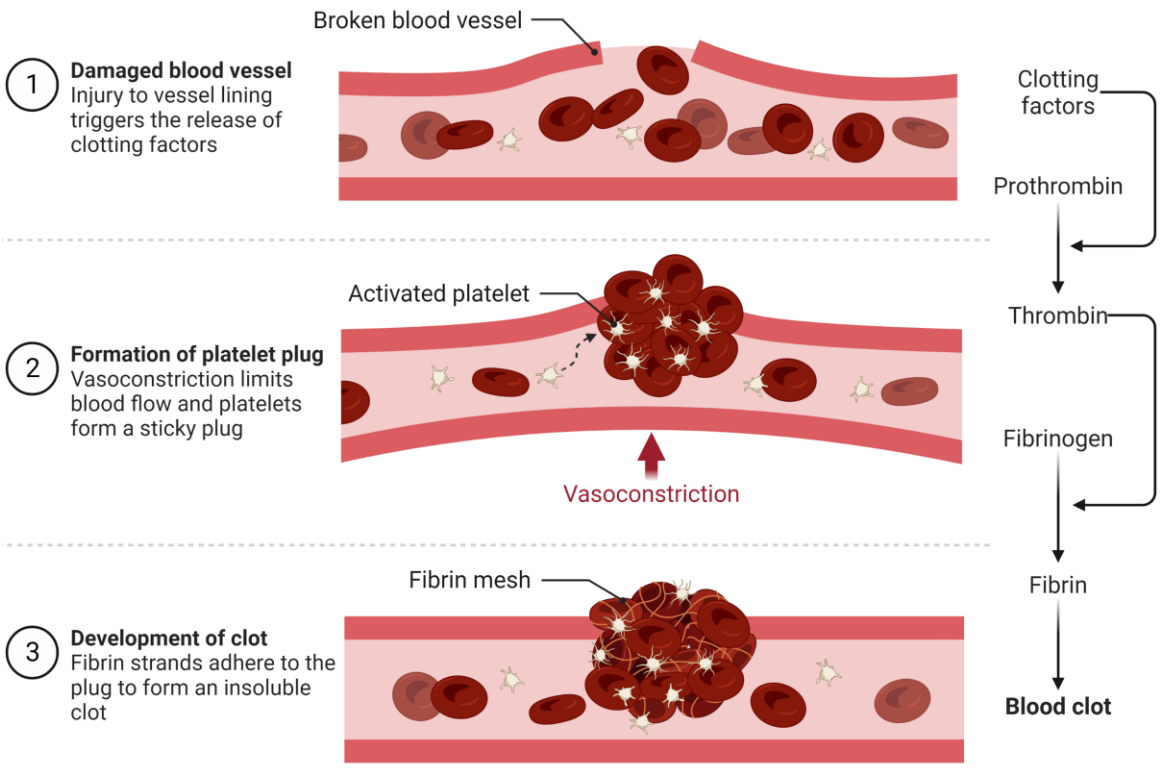


Figure 12: Schematic illustration over thrombus formation caused by damaged endothelium. (1) *Injury to the blood vessel lining triggers the release of clotting factors.* ***(2)*** *Thrombin activates the circulating platelets, which will start thrombus formation at the site of injury which leads to vasoconstriction.* ***(3)*** *Lastly, fibrin strands will adhere to the formed thrombus to create an insoluble clot. Reprinted from "Blood clot formation in broken vessel", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

Vascular injury causes endothelial dysfunction whereby cellular and protein material can gather at the site of injury, creating a blood clot [124]. Once the endothelium becomes inflamed, a two-step process of coagulation activation is initiated. Type I activation (stimulation) leads to elevated levels of Ca^{2+} causing increased blood flow and recruitment of leukocytes. Type II activation is mediated by tumor-necrosis factor α (TNFα) and interleukin-1 (IL-1) production, which in turn leads to an increase in blood flow, vascular leakage of plasma proteins and recruitment of leukocytes. Thrombus formation is important in inflammation since it separates the infected tissue from healthy tissue and prevents microbes from spreading [172]. However, excess coagulation can be detrimental, the regulation of blood coagulation is therefore important in the maintenance of a healthy endothelium [124].

1.7.3 Endothelial-enriched transcriptome

Many of the genes that are critical for the endothelial cell (EC) functions described in section 1.7. tend to have EC restricted expression profiles (endothelial-enriched genes). In earlier work from our group, Butler et al. [116] analyzed bulk RNAseq data from 124 unfractionated tissue samples, from 32 human organs, using methods as described in (section 1.6), to predict genes with endothelial enriched expression across tissue beds.

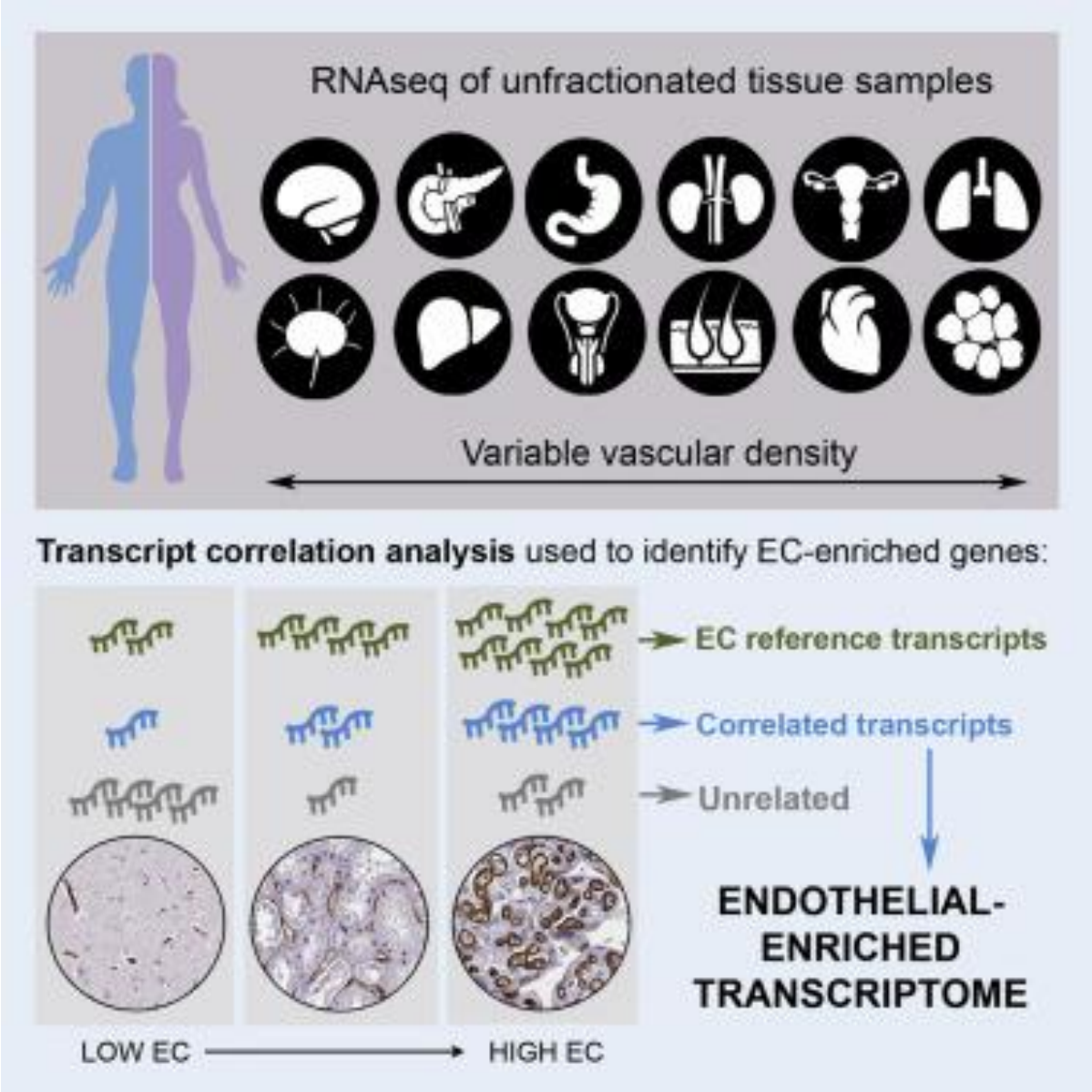


Figure 13: Method overview of the identification of the body-wide endothelial-enriched transcriptome by Butler et.al. in 2016 *[116].*

234 ‘EC-enriched genes’ were identified. 116 of these were previously reported as EC-expressed, 88 were not previously reported in EC, and 30 were totally uncharacterized. Comparison with RNAseq data from *in vitro* cultured EC indicated that some EC-enriched transcripts were expressed in tissue, but not in cultured cells, probably due to environmental changes.

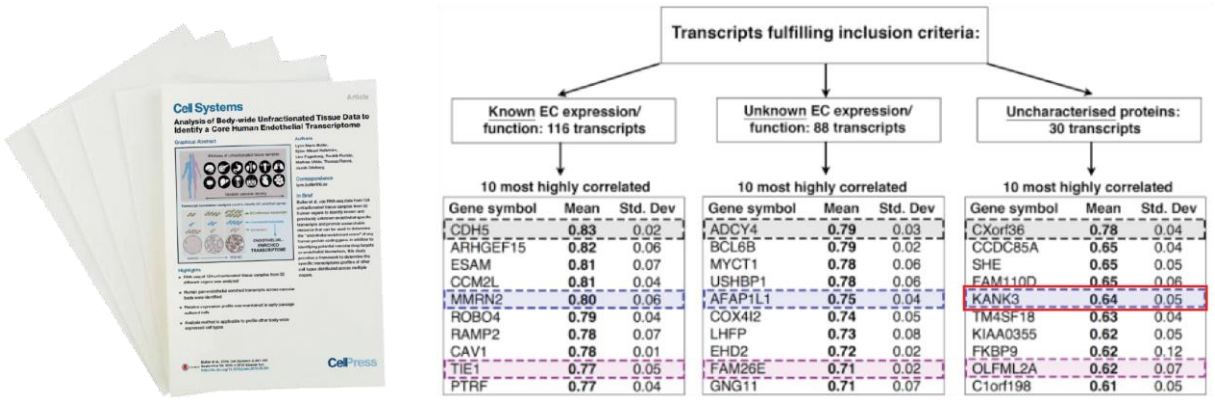


Figure 14: KANK3 is classified as an uncharacterized endothelial enriched protein. *Butler et.al. used bulk RNAseq data to identify both known endothelial enriched transcripts as well as unknown and uncharacterized transcripts as endothelial enriched [116].*

Whilst this study predicted body-wide EC enriched genes the total analyzed number of samples was relatively low (n=2-7 samples per organ) and individual tissue types were not analyzed in detail. It is well established that there is clear variability in EC gene expression profiles between vascular beds in different tissue types [173,174], for example in the gastrointestinal tract where EC has a role in controlling the passage of both antigens and commensal bacteria into the blood stream [173,174]. Genes that are consistently EC enriched across vascular beds are interesting candidates for functional investigation, as they likely have a key role in general EC function.

1.8 The gastrointestinal tract

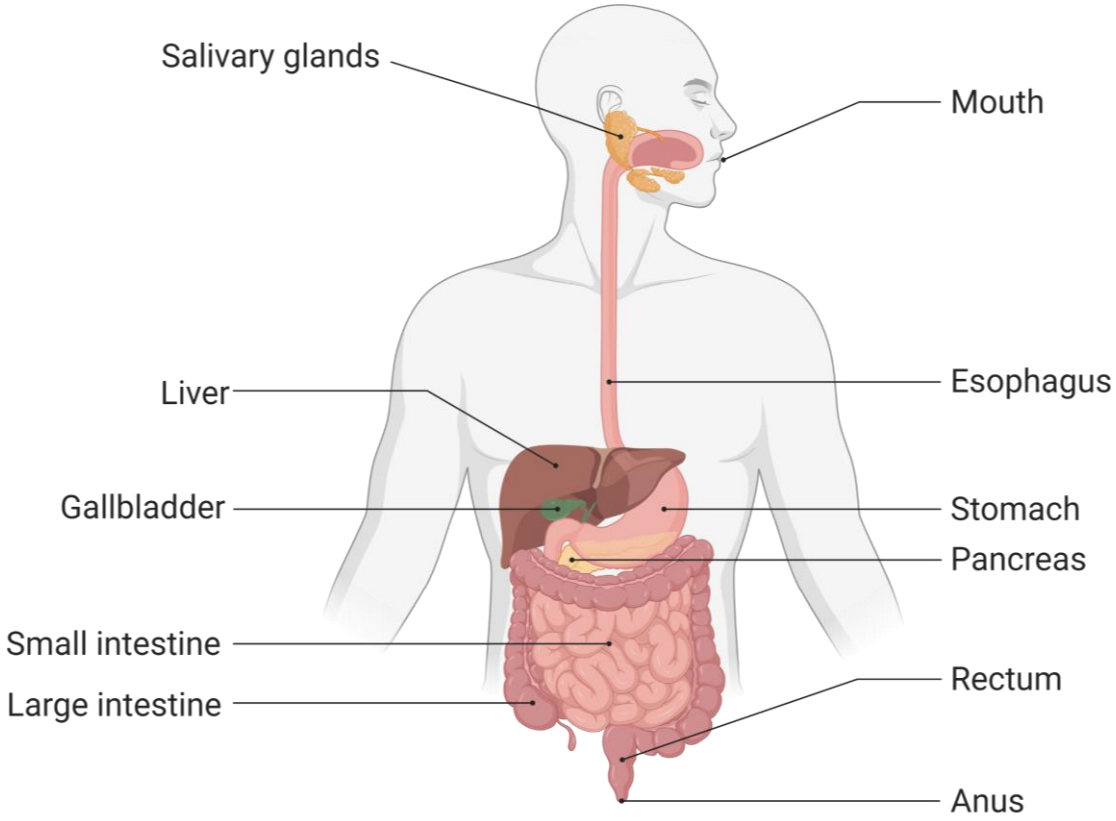


Figure 15: Schematic overview of the human gastrointestinal system. *The large organ system is divided into an upper and lower part and has accessory organs that aid in the digestive process. Adapted from "Digestive system", by Biorender.com (2023), retrieved from https://app.biorender.com/biorender-templates.*

The gastrointestinal tract is a large organ system that can be divided into an upper and lower part, the upper part consists of the mouth, esophagus, stomach, and small intestine while the large intestine (colon), rectum and anus constitute the lower part. The salivary glands, liver, gallbladder and pancreas are accessory organs that aid in the digestive processes [175]. The cellular characteristics, ratio and types of absorptive and secretory cells, change throughout the organ system, allowing the different organs to perform their distinct functions including absorption of nutrients, digestion and reabsorption of water [176–179]. These functionally different absorptive and secretory cell types constitute the gastrointestinal epithelial lining, forming a selective permeable barrier, preventing unwanted agents from entering the body while allowing nutrients to pass though [180].

1.8.1 Stomach

The stomach is a hollow muscular organ, located in the upper GI tract, and produces an array of acids and gastric enzymes as well as acting as a reservoir for the mechanical and chemical digestion of ingested food [181]. The constituent epithelial cell types of the stomach include parietal cells, chief cells, gastric mucous cells, gastric enteroendocrine cells and mitotic cells [2,182].

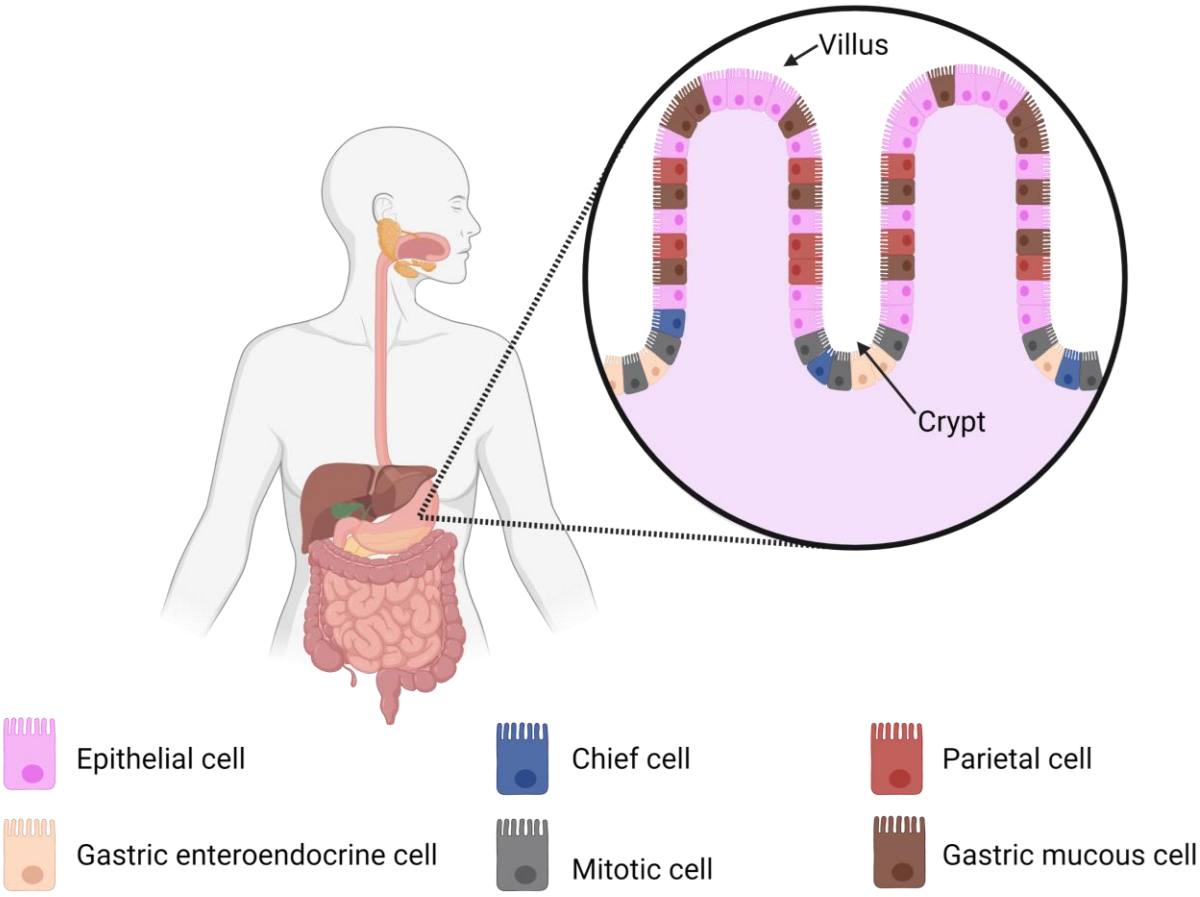


Figure 16: Illustrated schematic of the stomach epithelial lining. *Illustration is complete with the unique cell types: epithelial cells, gastric enteroendocrine cells, chief cells, mitotic cells, parietal cells and gastric mucous cells, and their location within the stomach villi. Illustration created with Biorender.com.*

Gastric mucous cells (or foveolar cells) constitute the majority of the gastric epithelial mucous lining, the secreted mucous forms a protective barrier against the corrosive gastric acid [183]. The gastric chief cells (or zymogenic cells) are located at the base of the gastric glands, or crypts, where they are characterised by pepsinogen secretion (the inactive form of pepsin) [184,185]. The chief cells also function as a reserve stem cell in the gastric epithelium that can be activated upon injury or disturbed

homeostasis [186,187]. Chief cells arise through transdifferentiation of mucous cells, a process that excludes cell division, and during loss of parietal cells the chief cell can further transdifferentiate into a mucous cell metaplasia (called spasmolytic polypeptide expressing metaplasia (SPEM)) [188]. Parietal cells (or oxyntic cells) are located in the neck (middle) region of the gastric glands, where they are responsible for gastric acid secretion, which is important for food digestion and mineral absorption, in addition to neutralising harmful food-derived bacteria [189]. The enteroendocrine cells (or neuroendocrine cells) are located in the base and neck of the gastric gland and classified into subtypes, based on the particular hormone or molecule the cell secretes [183,189]. G-cells are responsible for gastrin secretion, which stimulates gastric acid production by activating enterochromaffin-like cells (ECL-cell) and parietal cells. ECL-cells secrete histamine, after stimulation by gastrin, which stimulates the parietal cells to increase gastric acid production. D-cells secrete an inhibitory gastrin molecule, somatostatin, and they are activated when the stomach acidity reaches an upper level [183]. Additional subtypes include the serotonin secreting enterochromaffin cell, X-cells that produce ghrelin, and enteroendocrine cells that produce chromogranin A [176]. Mitotic cells (transient amplifying cells) constitute a population of undifferentiated epithelial cells that are responsible for epithelial cell replacement upon injury, whereby they differentiate into the suitable epithelial cell type [190].

In contrast to the better studied lower GI tract, descriptions of the cell type-enriched transcriptional landscape in the stomach are lacking, as the stomach is absent from several large scale single cell sequencing (scRNAseq) initiatives, such as Tabula Sapiens [191] and the Human Cell Atlas [1]. Where scRNAseq has been used to profile gene expression in the adult stomach, studies have typically focused on specific cell types, such as the epithelia [192,193], or in pathological states such as gastric cancer [194–197]. There are limited existing studies that focus on the gene expression profiles of EC in the healthy adult stomach.

1.8.2 Colon

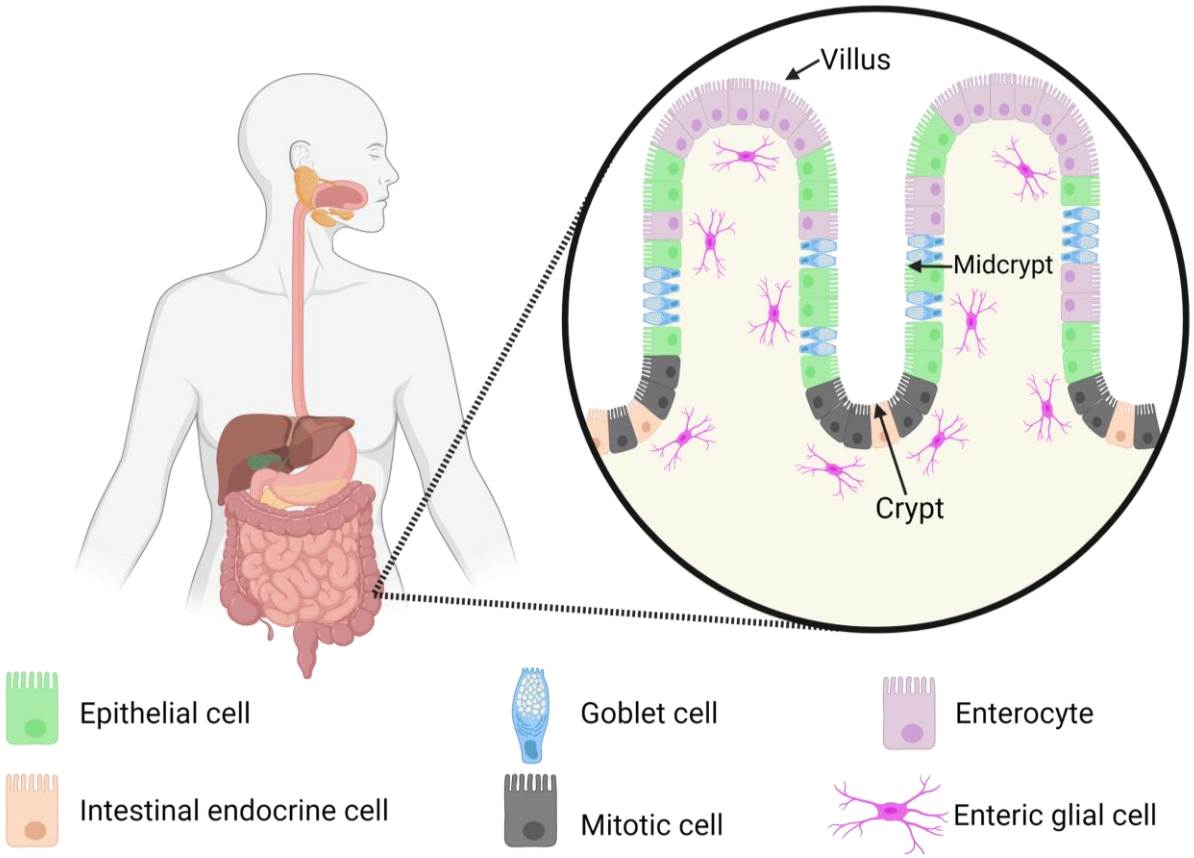


***Figure 17: Schematic illustration of the colon epithelial lining**. Illustration is complete with the unique cell types: epithelial cells, intestinal endocrine cells, goblet cells, mitotic cells, enterocytes and enteric glial cells, and their unique location within the colon villi. Illustration created with Biorender.com.*

The main function of the colon is to absorb water and salt, and the most abundant epithelial cell types in colon are enterocytes and goblet cells [177,198,199]. The goblet cells are mainly located in the midcrypt, located between the tip of the villus and the crypt, while the enterocytes are located at the villi surface, and the minority cell type intestinal endocrine cells are located at the base of the crypt together with proliferating mitotic cells [177].

The absorptive colonic enterocytes (also sometimes called colonocytes) constitute the majority of epithelial cells on the villi [200]. As absorptive cells, the enterocytes function in the uptake of various substances in the intestine such as water, ions, vitamins, nutrients and unconjugated bile salts [201]. Intestinal endocrine cells and goblet cells secret mucous and various hormones and are located both on the villi and in the crypts. Through acute notch inhibition and the transcription factor SPDEF,

all proliferative epithelial cells can convert into mucous secreting goblet cells [200]. There are various subtypes of intestinal endocrine cells, where the classification is based on the specific hormone they secrete [202]. Mitotic cells (or transient amplifying cells) are undifferentiated epithelial cells located within the crypts, and maintain the intestinal homeostasis by providing continuous replacement cells [203]. The mitotic cells also have an important role in modulating the ratio of secretory and absorptive cells in the intestinal epithelium [204]. Enteric glial cells constitute a specialized population of peripheral neuroglia cells that are associated with enteric neurons throughout the gastrointestinal system and maintain gastrointestinal homeostasis [205]. The enteric glial cells are located closely below the epithelial cell layer and share several structural and functional aspects with astrocytes. Enteric glial cells play an important role in gut epithelial integrity, mucosal barrier function, protection against bacterial invasion and they can regulate the epithelial cell transcriptome to shift towards increased cell adhesion and differentiation [206].

The human colon has been studied extensively in the context of colorectal cancer, which is the third most common cancer type worldwide [207]. Bulk sequencing studies of colorectal cancer have identified genetic and genomic alterations [208–210], however these studies do not identify cell-specific changes. Single-cell studies on healthy colon have focused on the epithelium [211] and changes in gene expression during inflammation in ulcerative colitis patients [212]. Various studies have focused on the involvement of individual cell types in the pathophysiology of colorectal cancer, such as: endothelial cells (ECs) [213–216], T-cells [217] and macrophages [218–220]. ECs are the major constituent cell type in the vasculature, and they have an important function in the initiation and progression of inflammatory bowel diseases (IBD) [221]. Furthermore, ECs provide a gut-vascular barrier that prevents the normal intestinal microbes (the microbiota) from entering the bloodstream [222]. However, there are limited studies that focus on the gene expression profiles of EC in the healthy adult sigmoid colon.

2 Aim of the thesis

Overall aim: To profile cell type enriched transcriptomes across human tissues and to select an uncharacterized endothelial cell enriched gene for functional investigation.

Aim 1:

- To generate a stomach cell type enriched protein- and non-coding gene atlas
    - Identification of expression profiles of rarely studied epithelial cell types in stomach tissue
    - Identification of sex specific enrichment signatures

Aim 2:

- To generate a colon cell type enriched protein- and non-coding gene atlas
    - Identification of expression profiles of cell types present in colon tissue
        - Identify expression profiles of cell types constituting the enteric nervous system
    - Identification of sex specific enrichment signatures

Aim 3:

- To generate multiple organ cell type protein-coding gene atlas
    - Perform cell type comparisons across tissue types to identify core cell type signatures, with a specific focus on endothelial cells

Aim 4:

- To use the data generated in Aims 1-3 to select a gene with highly endothelial specific expression across tissue types. To elucidate the role of this gene in endothelial cells by using *in vitro* cell culture systems to study the effects of gene silencing on endothelial cell specialized functions.

3 Methods and Methodology

3.1 Integrative correlation-based analysis

3.1.1 Dataset population

In papers I-III, publicly available bulk RNAseq data from the Genotype Tissue Expression Project (GTEx) [223] was used. The sequenced transcripts were categorized according to the biotype definitions in ENSEMBL, release 102 [224].

The GTEx project is a large-scale ongoing project with the aim to provide a comprehensive public resource to study human tissue-specific gene expressions and regulations. The project has had 4 data releases, in which more tissues and donor samples have been incorporated, for project I-III we have used release V8, which in total includes 54 distinct tissues and 948 donors.

Enrollment criteria stipulated that either sex from any ancestry group within the age of 21-70 can be included if the tissue sampling started within 24h of death. There are several exclusion criteria including diseases such as human immunodeficiency virus (HIV), viral hepatitis, metastatic cancer as well as whole-blood transfusion within the last 48h and a body mass index (BMI) >35 or <18.5. Specific protocols for tissue biopsy provide specific instructions and ensures common procedures between hospitals. Following sampling, the tissue biopsy samples were added to a stabilizing solution with ethanol and methanol, acetic acid and a fixating agent, however blood, brain and full-thickness skin samples remain unfixed. The samples are then sent for analysis where a section of each sample was stained for histological analysis – both to verify the organ source and for general tissue characterization. Following sample quality analysis, the DNA was genotyped for eQTL analysis, which in GTEx V8 was performed with whole genome sequencing technology (WGS), and RNAseq using a paired-end Illumina TruSeq RNA protocol which results in an average of 50 million reads per sample. The read depth of the sequencing was set to capture lowly-expressed transcripts, but is limited in detection of rare transcripts as well as splice variants [225].

For project I-III, we analyzed 15 out of the 54 sample sets included in GTEx.

3.1.2 Reference transcript selection

The bulk RNAseq data analyzed in paper I-III contains mixed cell types that are present in differing proportions (Fig 18 A). The integrative correlation-based method used in papers I-III is based on the selection of 3 cell type specific genes (‘reference transcripts´ [*Ref.T*]) for each constituent cell type found in the tissue, which act as a proxy for the proportion of that particular cell type within the sample (Fig 18 B). *Ref.T*. were identified using a mixture of older ‘none-omics’ studies [226], in-house protein profiling [2,182], single-cell sequencing data [192,227] or collated databases from multiple sources, e.g. Cell Marker [228] and PanglaoDB [229]. *Ref.T.* within each cell type panel were required to have a high correlation with each other (indicating cell type co-expression), a low correlation with *Ref.T.* representing the other cell types (indicating cell type specificity) and a normal expression distribution across the samples. As cell type expression and constitution varies from tissue to tissue, the *Ref.T*. were selected on a tissue-by-tissue basis. Spearman correlation coefficients between the selected *Ref.T.* and all other sequenced transcripts in the dataset (56200 total) were calculated in R using the corr.test function from the psych package (v 1.8.4). While the proportion of cell types varies between samples, the ratio between cell-enriched genes should remain relatively constant. Therefore, a high correlation coefficient of a given gene with only one *Ref.T*. panel indicates enrichment of that gene(s) in the corresponding cell type (Fig 18 Cii), whilst a lack of correlation indicates that the gene is <u>not</u> cell type-enriched (Fig 18 Ci). Genes were classified as cell type enriched (Fig 18 E) when the following criteria were fulfilled: (i) a mean correlation >0.50 (FDR <0.0001) with the *Ref.T.* panel representing that cell type <u>*and*</u> (ii) a minimum ´differential correlation´ between this value and the *next highest* mean correlation with any other *Ref.T.* panel (representing another cell type) >0.15 (Fig 18 C, D) <u>*and*</u> (iii) TPM expression <0.1 in over 50% of samples. In the case that the criteria were not fulfilled, the transcripts were classified as not cell type-enriched (Fig 18 F).

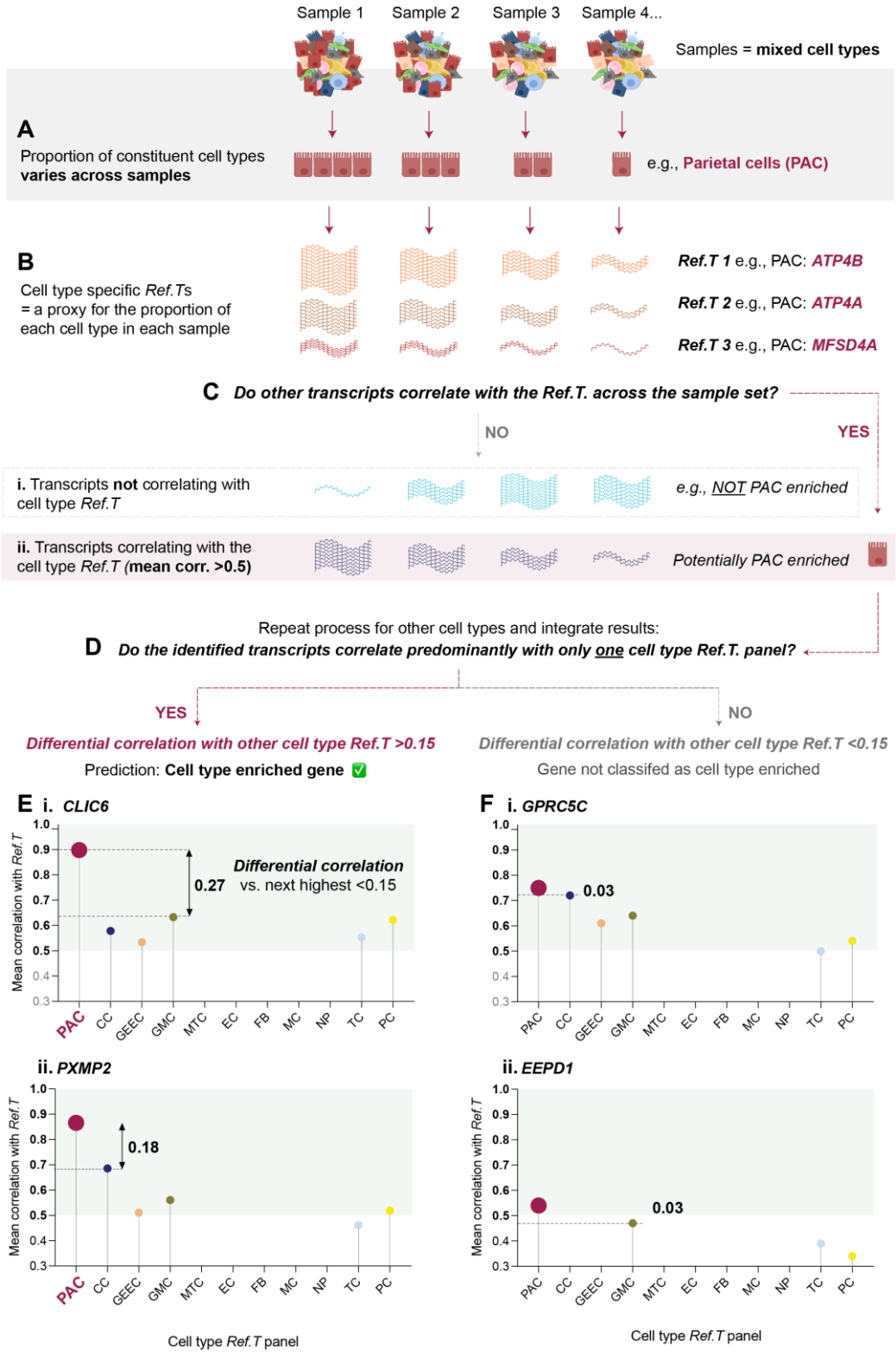


Figure 18: Detailed overview of the analysis methodology used in the integrative correlations-based method. (A) *Mixed cell types are present in differing proportions.* ***(B)*** *Ref.T. were selected as a proxy for cell proportions within sample.* ***(C)*** *Spearman correlation coefficients were calculated between Ref.T and all sequenced transcripts, transcripts were then classified as (i) not correlated or (ii) correlated.* ***(D)*** *The process was repeated for all Ref.T. represented cell types. Based on the results from Spearman correlation and differential correlation values, the transcripts were classified as* ***(E)*** *cell type enriched or* ***(F)*** *not cell type enriched.*

3.1.3 Verification using weighted gene correlation network analysis

As the analysis is based on manual selection of Ref.T. panels it can be subject to input bias. To verify results presented in paper I-III, we also used an alternative unbiased method that does not require any manual input – weighted gene correlation network analysis (WGCNA) [230]. WGCNA calculates correlation values between all transcripts across the sample set, before clustering transcripts together into groups, based on the degree of expression pattern similarity. Thus, transcripts are clustered together based on expression in a common cell type, or involvement in a common process. When genes we classified as cell type enriched appeared in the same WGCNA clusters it added weight to the accuracy to our classifications.

The R package WGCNA [230] was used to perform co-expression network analysis for gene clustering, on log2 expression TPM values. The analysis was performed according to recommended settings in the WGCNA manual. Transcripts with too many missing values were excluded using the goodSamplesGenes() function. The remaining genes were used to cluster the samples, and obvious outlier samples were excluded.

3.1.4 Verification using tissue profiling

Tissue-profiling for selected proteins expressed by predicted cell type-enriched transcripts was used to further verify results presented in papers I-III. Human tissue sections were stained, as previously described, as part of the Human Protein Atlas project [2,231]. Briefly, formalin fixed and paraffin embedded tissue samples were sectioned, de-paraffinised, hydrated and blocked for endogenous peroxidase in hydrogen peroxide solution. Antigen retrieval was done using a Decloaking chamber® (Biocare Medical, CA). Following boiling of the slides, primary antibodies and a dextran polymer visualization system (UltraVision LP HRP polymer®, Lab Vision) were added, and the slides were incubated and were developed using Diaminobenzidine (Lab Vision) as the chromogen. Slides were counterstained in Mayers hematoxylin (Histolab) and scanned using Scanscope XT (Aperio).

3.1.5 Single-cell verification

We sourced scRNAseq data from Tabula sapiens as an additional means of verification of cell type expression profiled for both protein-coding and non-coding enriched transcripts [191]. However, Tabula sapiens did not contain scRNAseq data for all tissues, in those cases gene expression in general cell type compartments (epithelial, endothelial, stromal, and immune) was assessed. scRNAseq data was downloaded and UMAP plots created using the Seurat package in R [232]. UMAP plots for selected protein-coding and non-coding transcripts were generated on a tissue or compartment basis.

3.2 Functional characterization of KANK3

For full method details, see methods section of paper IV.

3.2.1 Isolation and culture of primary endothelial cells

Ethical approval for endothelial cell isolation and subsequent experimentation was granted by *Regionala etikprövningsnämnden i Stockholm* (diarienummer 2015/1294-31/2).

Human umbilical vein endothelial cells (HUVECs) were isolated from anonymized human umbilical cords, collected from Karolinska Hospital (Stockholm, Sweden) as previously described [233]. Briefly, the umbilical cord was isolated and a glass cannula was inserted into the vein. Following rigorous washing, collagenase type II solution was inserted into the vein, after which and the cord was incubated. The cell suspension was collected, pelleted, and resuspended in Medium M199+ (M199+).

HUVEC were cultured in M199+, supplemented with 20 % fetal bovine serum (FBS), 10 ml/l Penicillin-Streptomycin, 2.5 mg/l Amphotericin B (all ThermoFisher, Gibco), 1 mg/l Hydrocortisone 1 µg/l and human Epidermal Growth Factor (hEGF) (both Merck). For some experiments, cells were cultured in 1 % FBS. For some experiments, HUVEC were purchased from Merck/Sigma Aldrich.

3.2.2 Gene knockdown and recombinant KANK3 protein expression

To investigate the function of KANK3, gene knockdown or over expression strategies were used. Briefly, KANK3 knockdown was achieved using siRNA

(short/small interfering RNA) targeting KANK3. Cells were transfected with siRNA in a lipofectamine solution. After incubation with the transfection solution, the cells were washed and continued culture in previously described M199+. Transcripts should be knocked down after 48-72 h, and knockdown efficiency was evaluated with RT-qPCR and Western blot.

Overexpression was induced by transfection of plasmids in Opti-MEM using Lipofectamine 3000 transfection reagent - according to manufacturer instructions. Medium was changed to standard cell culture medium after incubation with transfection solutions. Transfection efficiency was investigated using Western blot or immunofluorescence staining after 48 h.

3.2.3 Western blot

To evaluate KANK3 knockdown efficiency on protein expression level in HUVEC, cell samples were obtained and analyzed for KANK3 by Western blot. Briefly, cell lysates from HUVECs, obtained using RIPA buffer, were mixed with Laemmli buffer in reducing conditions. The samples were then heated before loading onto the SDS-PAGE gel. After electrophoreses, proteins were transferred to PVDF membranes. Membranes were then blocked and incubated with primary rabbit anti-KANK3 antibody overnight. After washing, the secondary antibody solution horseradish peroxidase – conjugated goat anti-rabbit antibody was applied. After incubation, the membrane was washed and ECL detection solution was added to the membrane and incubated for 5 min, after which they were imaged using iBright™ 1500 (thermofisher). Similar protocol was used to detect recombinant KANK3-eGFP in HEK293T cells.

3.2.4 RT-qPCR

To confirm KANK3 knockdown efficiency on RNA level RT-qPCR was performed according to TaqManTM Fast Cells-to-CT^{TM} Kit provided by Thermofisher Scientific [234]. Cultured cells were washed, and lysis solution was added to each sample. Following incubation, the samples were collected and added to the stop solution. The samples were incubated for 2 min and unless the RT-PCR were performed right after, the samples were stored in the freezer. To convert the obtained RNA to cDNA, RT-PCR was run. Briefly, the master mix was prepared for each reaction

and sample lysate were added to each well and thoroughly mixed. After thermocycler run, according to protocol, the cDNA was stored in a freezer at -20 °C. To quantify the effect of KANK3 knockdown on various genes, qPCR reactions were run using TaqMan target primers. The qPCR master mix were prepared and mixed with diluted cDNA. The qPCR was performed using a RealTime PCR Lightcycler 96 ® system (Roche Life Sciences), after which the results could be analyzed.

3.2.5 Cytokine stimulation

Endothelial activity in response to inflammation can be stimulated by addition of inflammatory mediators such as lipopolysaccharide (LPS) or tumor necrosis factor α (TNFα) at a concentration of 10 ng/ml to the culture media. The added cytokine provokes an inflammatory response in the endothelial cells and the response can be measured with RT-qPCR or flow cytometry [235].

3.2.6 Shear stress exposure

Shear stress exposure assays was used to mimic in vitro conditions on cultured HUVECs. HUVECs were cultured in flow chamber slides connected to an Ibidi pump system which applied laminar shear stress onto the cells by pumping M199+ at either 4dyn or 40dyn. The results could later be analyzed by qPCR, Western blot, or microscopy.

3.2.7 Microscopy

To determine the protein location of the gene of interest within cells, or the effect of loss of gene expression, several microscopy methods were used in paper IV.

Cells for confocal microscopy, used to study protein localization, were fixed, permeabilized and blocked. Cells were then incubated with KANK3 primary antibody, followed by FITC-conjugated anti-rabbit antibody and TRITC-conjugated phalloidin (targeting the primary antibody and enables visualization). Images were taken using a Leica TP5 SP5 confocal microscope and image analysis was performed in Fiji ImageJ2 graphics procession software.

Structure illumination microscopy (SIM), used to get a detailed image of protein localization within cells. Cells were plated subconfluently on glass coverslips and

cultivated for 1 h (HEK293 cells) or 4 h (HUVECs). Afterwards, cells were fixed, washed, permeabilized and blocked. Primary antibodies targeting either the gene of interest or potential localization partners were prepared and added to the sample, followed by incubation with secondary antibody. Images were taken in an OMX Blaze SIM microscope and reconstituted using SoftWoRx software (GE Healthcare). Image analysis was performed in Fiji ImageJ2 graphics procession software.

Live-cell microscopy for gap closing (wound healing assay), used to study the effect of KANK3 depletion on living cells, was begun 24 h after siRNA treatment. HUVEC were seeded into 24-well plates and cultured for a further 24 h after which medium was changed to either 20 % serum or 1 % serum for either model. After additional 24 h an artificial wound was created in the center of the wells by scratching with a pipette tip. The plate was placed into a stage incubator chamber and analyzed with an Olympus IXplore Live microscope in phase/contrast mode in 10x magnification. Gap size was measured every 6 h in Fiji using ImageJ2 graphics procession software.

3.2.8 RNA isolation and sequencing

The RNA isolation and purification were carried out using the RNAeasy mini kit by Qiagen and was used to determine the expression ratio of the various gene isoforms expressed by ECs. The concentration of RNA was determined with the help of a Nanodrop 2000 spectrophotometer, and its integrity was assessed using Agilent 2100 Bioanalyzer. Library preparation and RNA sequencing were performed by the National Genomics Infrastructure Sweden (NGI) using Illumina stranded TruSeq poly-A selection kit and Illumina NovaSeq6000S. The sequencing was done with four lanes, 2x 150bp reads, and included 2Xp kits. The data was processed using demultiplexing, and the storage and initial analyses were done using server-sided computation provided by the Swedish National Infrastructure for Computing (SNIC).

3.2.9 Calibrated automated thrombinoscope (CAT) assay

CAT assay was used to investigate the effect of KANK3 on coagulation pathway in ECs. HUVECs were seeded into flat-bottom 96 well plates. After TNFα stimulation, medium was removed, and the wells were blocked. Following washing, thrombin formation was initiated. As controls, tissue factor mouse monoclonal anti-TF antibody

or corn trypsin inhibitor were added 15 min prior adding fluorogenic substrate. Thrombin generation was quantified using the Thrombinoscope software package (Version 5.0.0.742) that reported means ± SD.

3.2.10 Flow cytometry

HUVECs were cultured, transfected, and stimulated with TNFα before harvesting. The supernatant was collected, and concentrated by centrifugation the pellet was then resuspended. Cells were treated with PE-conjugated anti-CD142 Clone NY2 and isotype-matched control mouse-IgG1 followed by incubation on ice and centrifugation, and the pellet resuspended in PBS. Flow cytometry was then performed using the Beckman Coulter CytoFLEX Flow Cytometer. Gating and data analysis was done using CytExpert for CytoFLEX Acquisition and Analysis Software and FlowJo™ v10.7.

4 Results

4.1 Results paper I

Objectives: The stomach is located in the upper gastrointestinal tract and provides an acidic environment that contributes to microbial defense and food processing. Despite its important functions, detailed descriptions of cell type gene enrichment of stomach tissue are absent from major single cell sequencing-based atlases. We aimed to identify stomach cell type-enriched transcriptomes of 11 different cell types, including endothelial cells.

Methods: We used the integrative correlation-based analysis method (Fig 19 A) to analyze unfractionated bulk RNAseq data from 359 human stomach tissue samples. Weighted network correlation analysis and protein profiling was used to verify the results. Available scRNAseq data from Tabula Sapiens, categorized into broad cell type compartments (epithelial, endothelial, immune and stromal), was used to provide supportive evidence for gene cell type enrichment classifications.

Results: We profiled the transcriptome of stomach-specific cell types; parietal, chief, gastric mucous and gastric enteroendocrine cells as well as core cell-types found in multiple tissues; mitotic, endothelial, fibroblast, macrophage, neutrophil, T-cell and plasma cells (Fig 19 B, C and D). We identified both protein coding and non-coding enriched cell-type specific signatures (Fig 19 D ii and iv), several of which are strongly associated with the progression of gastric cancer. Additionally, by conducting a sex-based subset analysis we identify a small panel of male-only enriched chief-cell genes (Fig 19 E).

Conclusions: We provide a transcriptomic atlas of cell-enriched gene signatures of the human stomach, which have been absent from major single-cell sequencing-based atlases.

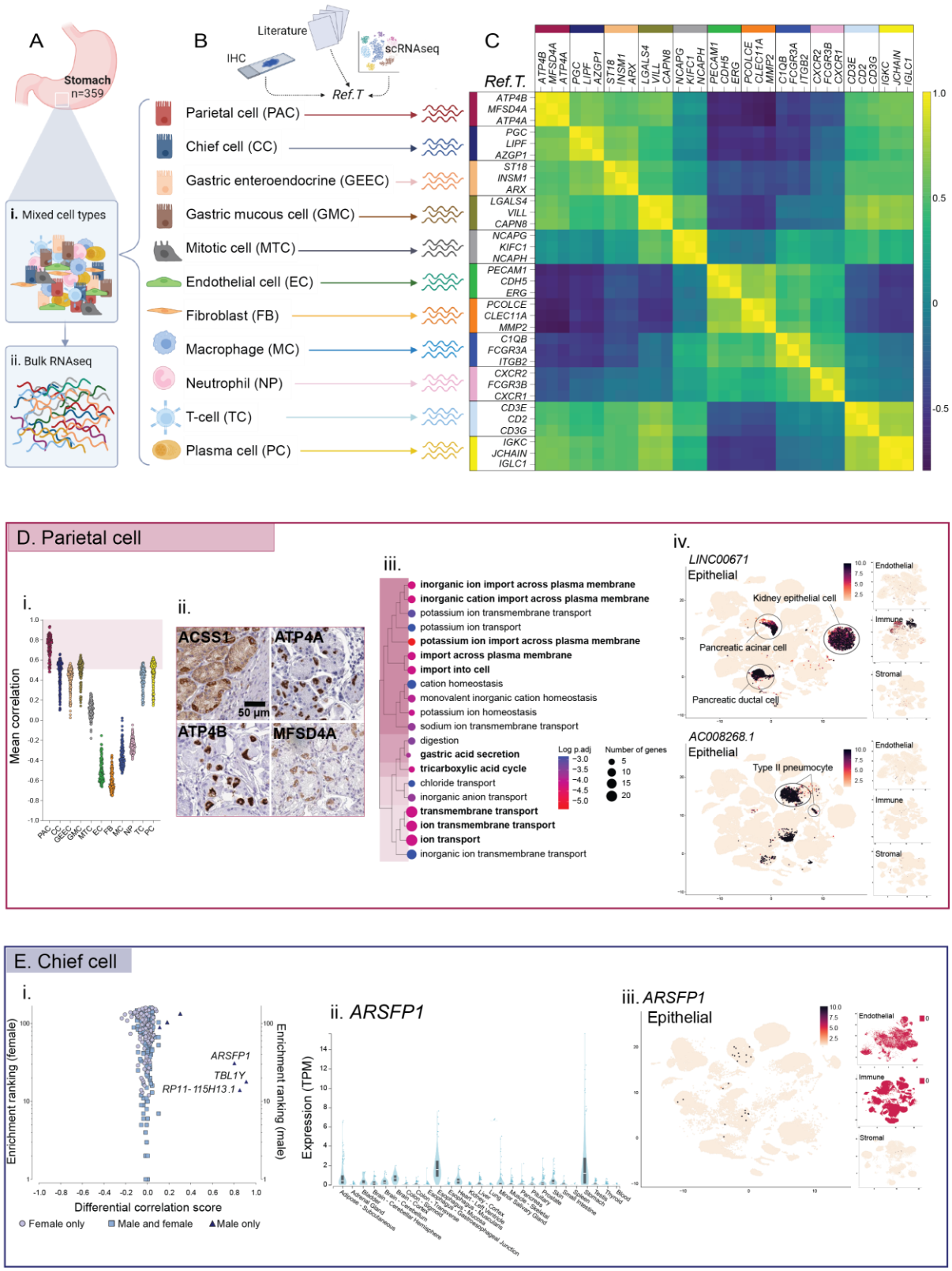


Figure 19: Summary of Paper I. (A) *Bulk RNAseq samples from stomach tissue using* ***(B-C)*** *Ref.T was used to identify the transcriptomes of 11 different cell types present in stomach tissue.* ***(Di)*** *Cell type-enriched protein coding genes were verified using (ii) protein profiling, (iii) enriched GO-terms and (iv) enriched non-coding genes were verified using available scRNAseq data.* ***(Ei)*** *Additionally, male enriched chief cell transcripts were identified with (ii) enriched expression in male stomach tissue and in (iii) epithelial cells using scRNAseq data.*

4.2 Results paper II

Objectives: The colon is part of the gastrointestinal tract, where it performs functions such as reabsorption of water. The cellular composition of the GI tract varies within the system, making it possible for the organs to perform their specific functions. The human colon has been studied extensively in the context of colorectal cancer, with focus on individual cell types involved in cancer progression or inflammation. We aimed to identify cell type-enriched transcriptomes of 12 different cell types, including endothelial cells, in healthy sigmoid colon.

Methods: We used the integrative correlation-based analysis method to analyze unfractionated bulk RNAseq data from human sigmoid colon tissue (Fig 20 A). Weighted network correlation analysis and protein profiling were used to verify the results. Available scRNAseq data from human large intestine Tabula Sapiens was used to verify tissue enriched cell type-specific non-coding genes.

Results: We identify cell type-specific gene enrichment profiles for 12 different human sigmoid colon cell types, including endothelial cells (Figure 20 B), in tissue, with a total of over 3000 cell type-enriched transcripts. We identify non-coding enriched cell-specific signatures (Fig 20 E i) as well as protein coding (Fig 20 E ii), several of which have previously been associated with colorectal cancer. We also identify enriched transcripts of cell types present in the enteric nervous system, such as enteric glial cells (Fig 20 D). Sex-based subset analysis also identified a couple of male-enriched genes (Fig 20 F).

Conclusions: Using publicly available bulk RNAseq data, we successfully identify cell-type specific transcriptome of human sigmoid colon.

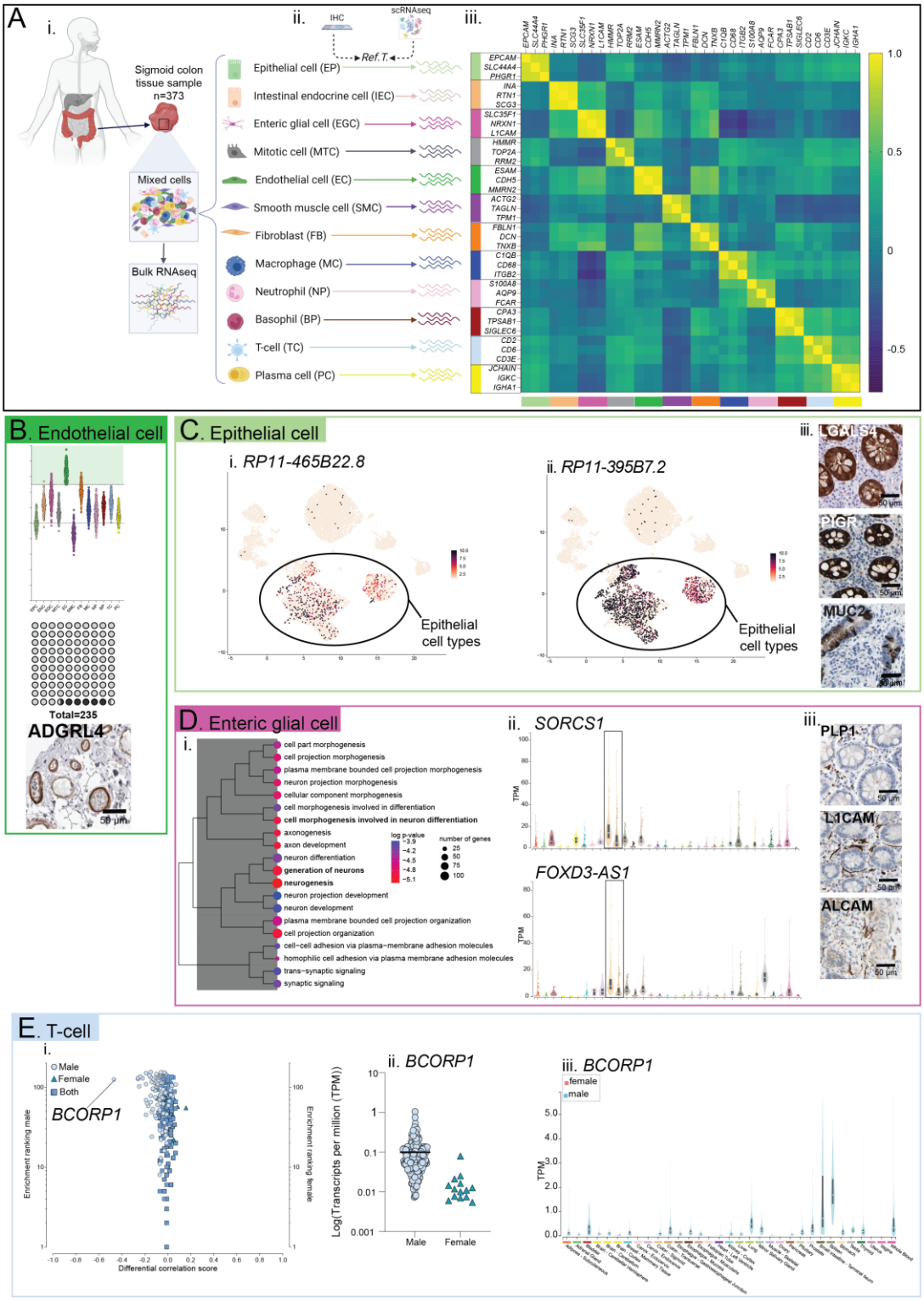


Figure 20: Summary of Paper II. (A) *Bulk RNAseq samples from colon tissue using (ii-iii) Ref.T was used to identify 12 cell type transcriptomes.* ***(B)*** *Enriched gene signatures were identified for all constituent cell types, including EC.* ***(Ci)*** *Non-coding enriched genes were verified using available scRNAseq data and (ii) protein coding enriched genes using protein profiling.* ***(D)*** *Enteric glial cell, constituting part of the enteric nervous system, enriched transcripts were identified and verified using enriched GO terms (i), expression profiling (ii) and protein profiling (iii).* ***(E)*** *Male enriched transcripts were identified (i) with enriched expression in male stomach tissue (ii-iii).*

4.3 Results paper III

Objectives: Single-cell RNA sequencing can be used to identify cellular expression on a single cell-level and is commonly used for biomarker discovery and to study the changes in expression in health and disease. However, scRNAseq has limitations such as artefactual changes of gene expression during cell processing. Furthermore, several cell types e.g., adipocytes are absent from major databases due to damaging isolation processes. We show in paper I-II, as well as our previous studies [117,117,118], that the integrative correlation-based method can be used to circumvent limitations related to scRNAseq. Therefore, the objectives were to use publicly available bulk RNAseq data to identify cell type-enriched transcriptomes of several human organs to generate a tissue-by-tissue enrichment prediction atlas of protein-coding transcripts.

Methods: We used the integrative correlation-based analysis method to analyze bulk RNAseq data from 15 human tissues (Fig 21 A, B). Weighted network correlation analysis, gene ontology (GO) term analysis and protein profiling was used to verify the results (Fig 21 C).

Results: We successfully profile all the major constituent cell types of 15 human tissues, including several cell types that are difficult to process and not included in existing scRNAseq databases. We were able to identify co-enriched gene panels between pancreatic alpha and beta cells (Fig 21 E), identify temporal gene changes during spermatogenesis (Fig 21 F). Comparing the transcriptome signatures of common cell types identified in multiple tissues enabled identification of core cell type identity profiles (Fig 21 G).

Conclusions: We provide a cell type gene enrichment atlas that has been generated independently of scRNAseq. Providing an additional tool to understand the human gene expression across intact tissues.

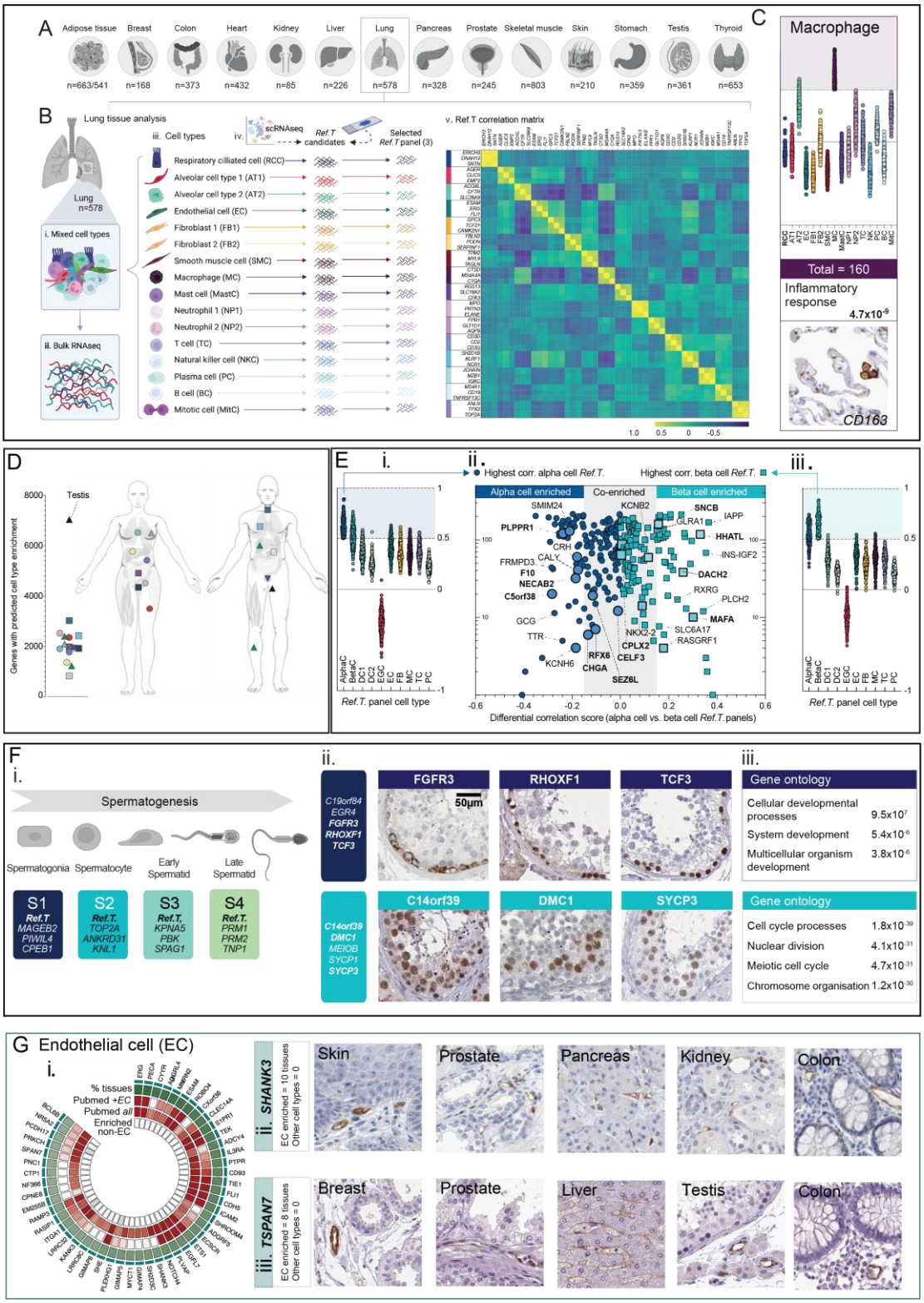


Figure 21: Summary of paper III*. **(A)** 15 human tissues were analyzed using **(B)** the integrative correlations-based approach to generate **(C)** cell type-enriched transcriptomic profiles. **(D)** Each profiled tissue contained different number of enriched genes. **(Ei and iii)** Cell type-enriched, as well as (ii)co-enriched genes were identified for some cell types. **(Fi)** In testis, the analysis showed distinct transcriptome signatures during spermatogenesis, which were verified using (ii) protein profiling and (iii) GO-enriched terms. **(Gi)** Identification of core-cell types in which selected transcripts were verified using protein profiling (ii and iii) across tissues.*

4.4 Results paper IV

Objectives: Endothelial cells perform important functions in several vascular processes, such as coagulation, inflammation, and angiogenesis. Proteins with endothelial restricted expression profiles are known to be important in these processes. In papers I-III, we identified KANK3 as an uncharacterized endothelial-enriched transcript across multiple tissue types. The objective of paper IV was to functionally characterize KANK3 in endothelial cells.

Methods: Human umbilical vein endothelial cells (HUVECs) were extracted, cultured, and transfected with siRNA targeting KANK3. Functional assays such as inflammatory cytokine stimulations, flow cytometry, wound healing assays, thrombin generation and culture under static/flow were carried out, and protein localization in HUVECs was analyzed using fluorescence microscopy.

Results: KANK3 was verified to be EC enriched by analyzing over-represented gene ontology terms for top KANK3-correlating genes (Fig 22 A i), scRNAseq data from Tabula Sapiens (Fig 22 A ii) and protein profiling (Fig 22 A iii). HUVEC cell culture under static and shear stress conditions showed increased KANK3 expression and distribution under shear stress (Fig 22 B). Wound healing assays showed increased cell motility in HUVECs after KANK3 knockdown in both high and low serum culture conditions (Fig 22 C). KANK3 knockdown increased levels of F3-expression (Fig 22 D i). Results were confirmed using flow cytometry (Fig 22 D ii), indicating an involvement in coagulation. These results were further verified using thrombin generation assay in which KANK3 depletion led to enhanced thrombin formation following TNF stimulation (Fig 22D iii).

Conclusions: The protein encoded by KANK3 has an endothelial specific expression profile. KANK3 is a shear stress regulated protein. EC KANK3 depletion also revealed a role in EC migration, and in regulation of tissue factor (TF) expression, which is involved in coagulation.

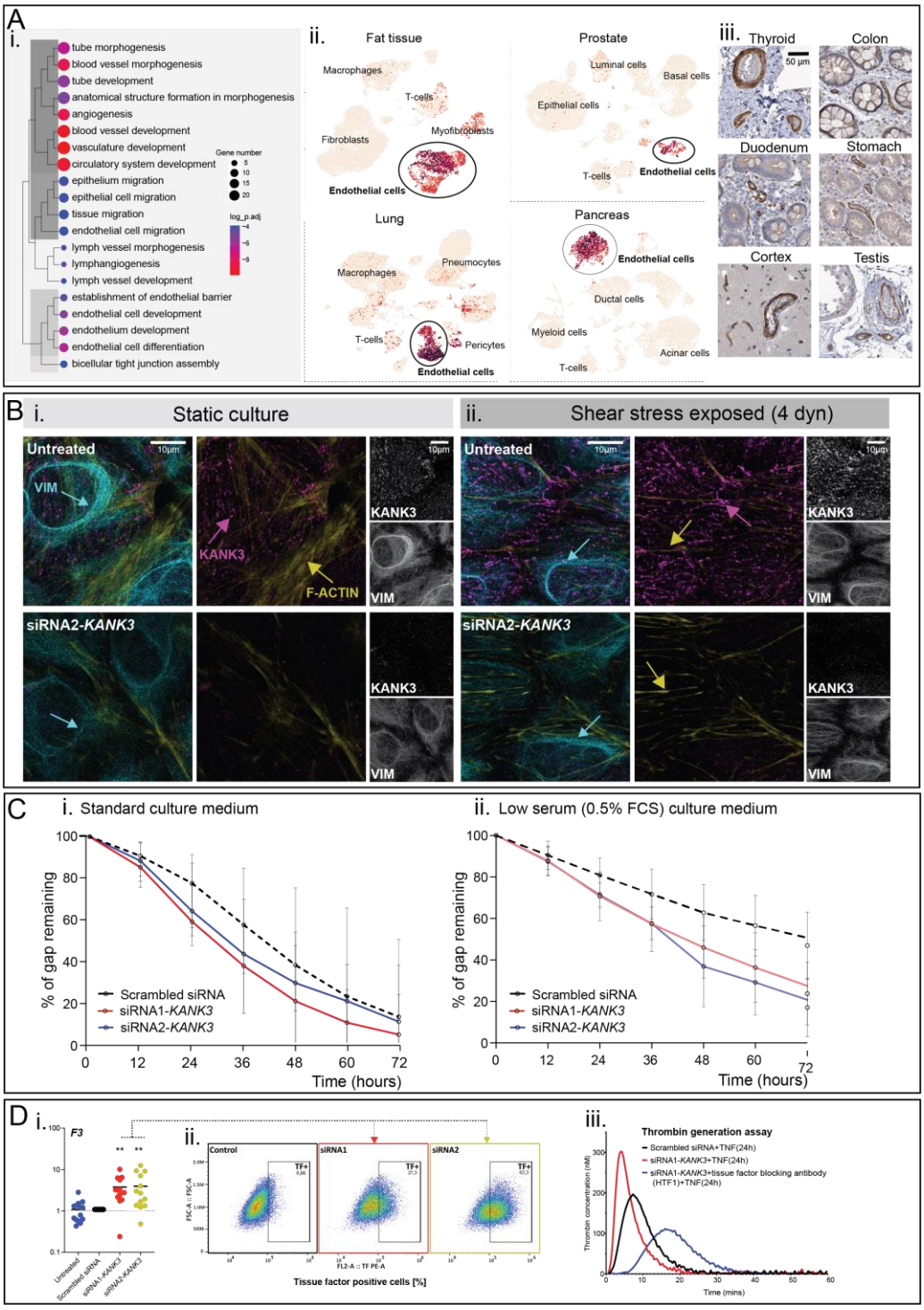


Figure 22: Summary of paper IV. (A) *KANK3 is endothelial enriched, where (i) over-represented Gene Ontology terms of top KANK3-correlating genes are associated with known EC functions, (ii) scRNAseq data from Tabula Sapiens [191] shows KANK3 enrichment in EC clusters and (iii) protein profiling from HPA [231] shows KANK3 protein expression in EC.* ***(B)*** *HUVEC culture under (i) static and (ii) shear stress conditions followed by IHC using antibodies targeting KANK3 (magenta), F-actin (yellow) or vimentin (VIM; cyan) shows elevated expression under shear stress.* ***(C)*** *Wound healing assay on HUVEC KANK3 depleted cells in (i) standard medium and (ii) low serum shows that KANK3 depletion increases EC migration in vitro.* ***(D)*** *KANK3 depletion increased F3 expression on (i) mRNA level and (ii) cell surface expression. (iii) Calibrated automated thrombogram (CAT) measurement on KANK3 depleted HUVEC shows increased thrombin formation.*

5 Discussion

5.1 Methodological considerations

5.1.1 Dataset selection

Papers I-III use publicly available RNAseq data to extract single-cell transcriptome signatures. There are multiple publicly available datasets, but to perform the analysis with sufficient statistical power the sample set should be as large as possible. GTEx offers over 17,000 samples from over 900 different donors taken from 54 different healthy tissues making it ideal for this analysis, as opposed to other multi-tissue datasets, such as The Cancer Genome Atlas program (TCGA) [236] which primarily contains cancer samples, with fewer healthy tissue samples. Additional criteria for study selection included criteria related to sampling, in which the tissue samples should be taken from the same organ location, this is to avoid false classifications due to varying sample location. This excluded multiple studies that focused on disease mechanisms, for instance Crohn's disease, as the sample location depend on where the disease is expressed, which also determines the location of the matching healthy sample [237].

The GTEx study used in paper I-III fulfills the above-mentioned criteria, and while the samples have been obtained at various hospitals, they have all followed the same biopsy and processing protocols. However, while the GTEx dataset contains hundreds of samples for each of the 54 tissues, there are a few limitations to consider while analyzing the data. First, all tissues contain both male and female samples, but there are far fewer female samples than male, potentially leading to enrichment classifications being driven by the male population. Secondly, the age of the donors are predominantly over 50 years of age, leading to enrichment classifications that could be limited to a certain age demographic.

5.1.2 Cell-type inclusion

In papers I-III, bulk RNAseq from GTEx was analyzed and while the dataset contains samples from multiple human tissues, the specific biopsy location or tissue processing has limited the inclusion of certain cell types [223,225]. In paper II we successfully identify several of the epithelial cell subtypes that are present in stomach mucosa, such as parietal and chief cells. In contrast, in sigmoid colon tissue (paper III)

we were unable to make such specific distinctions, as the mucous layer was removed from this tissue before sequencing, and so we could only identify epithelial cells as a general group. However, the epithelial cell enriched genes clustered together in the WGCNA. This processing limitation did not affect any of the other tissues included in paper III, as the tissue biopsies contained a cell type representative sample.

In paper I we included multiple epithelial cell subtypes in the analysis of tissue cell type-enriched genes, while 'gastric enteroendocrine cell' functioned as an umbrella term for the multiple enteroendocrine subtypes present in stomach tissue. Additionally, in paper III we merge basal and suprabasal keratinocytes into one group. Similar grouping of cell types into a broader category is frequently observed in scRNAseq studies, as cell subtype classification can be based on a higher expression of a limited number of specialized proteins or hormones [227,238–240], rather than expression of a large number of highly specific genes.

A further limitation of the method is the lack of potential to identify novel cell type transcriptomes, as input Ref.T. for such cell types would not be included in the analysis. Additionally, rare cell types, that only constitute a small percentage of the cell population present in the sample, are not included as it is difficult to detect the correlation due to background noise levels.

5.1.3 Cell type identification and classification

In papers I-III we have used our integrative correlation-based analysis method to identify cell type-enriched transcriptomes of various healthy human tissues. There are currently no other methods to extract such detailed single cell gene enrichment data from bulk RNAseq, other than scRNAseq. While it has been stated that bulk RNAseq data cannot be used to extract cell type gene enrichment profiles [241,242], we show in paper I-III that it is possible. Our analysis method does not require advanced level of bioinformatics to perform [111]. Further, the cell types have been processed in their natural environment, circumventing the risk of introduction of technical artifacts caused by cell extraction or processing related to scRNAseq techniques [93–95]. Moreover, this made it possible to profile cell types that would have been excluded from scRNAseq due to their sensitive nature [94,121]. Additionally, we are able to classify lowly expressed transcripts as cell type-enriched, which might only have been detected in a small minority of cells using scRNAseq,

possibly due to the method's limited read depth [243]. Transcriptomic deconvolution methods, such as CYBERSORT, might also not detect such lowly expressed genes as they classify cell proportions and the expression might be masked by more abundant cell types [110].

5.1.4 Input bias and misclassification

The integrative correlation-based analysis used in papers I-III can be subjected to user input bias during Ref.T. selection. This input bias is shared with deconvolution methods as they depend on input expression matrices of the cell type reference genes [111]. To overcome this potential bias in paper I-III, several criteria were included for Ref.T. selection to ensure correct cell type identification and classification. First, to prevent selection bias of the 'virtual markers' the marker selection process was made with multiple researchers working in a collaboration to avoid individual selection bias Secondly, the 'virtual markers' are selected for each tissue independently as gene expression might vary between tissues. Thirdly, genes that are highly expressed by multiple cell types were excluded as 'virtual markers' as they might correlate with multiple cell types. Furthermore, results are verified using multiple approaches, such as protein profiling for protein-coding genes [2] and verification using scRNAseq for several non-coding genes [191].

The method uses enrichment criteria to reduce the risk of false positive classifications. By ensuring that the cell type-enriched classified transcripts have the highest mean correlation with the corresponding Ref.T. panel, the analysis method captures true positive transcripts. Furthermore, false positives are excluded by ensuring a minimal differential correlation value between transcripts of >0.15. This criterion prevents misclassification by increasing the selectivity by which transcripts are classified. However, it is important to note that this criterion could potentially exclude true positive transcripts from being classified as cell type-enriched. Further, there is a likelihood of false negative classifications in the analysis as high thresholds are used for classification of genes as cell type enriched. Therefore, it is likely that some cell type enriched genes have been excluded due to not reaching the required threshold.

There is a balance between minimizing both false positives and false negatives in each analysis method. The integrative correlations-based method is more focused

on minimizing false positives by using high thresholds for differential correlation values. As there is a likelihood of false negatives, it is important to consider each enrichment classification on a transcript-by-transcript basis.

5.1.5 Primary cells

Paper IV used primary endothelial cells obtained from the umbilical vein (HUVEC) to study the function of an endothelial enriched transcript, and while studying an unclassified transcript it is of importance to choose a relevant model. Endothelial cells constitute the innermost layer of all blood vessels, making primary HUVECs a suitable model. However, primary cells tend to lose some of their characteristics as they reach higher passage (reportedly between 4-8 [244–246]) and enter senescence, for *in vitro* experiments it is therefore important to use freshly isolated cells of a low passage. In paper IV, HUVECs were used up to passage 4. An alternative to primary cells would be immortalized endothelial cell lines, however, studies have shown that there are significant differences in the phenotype of primary endothelial cells and several established endothelial cell lines [247,248]. Immortalized endothelial cell lines showed a lack of PECAM-1 expression, as well as inability to induce VCAM-1 and E-selectin in response to TNFα stimulation or MHC class II antigens in response to IFNγ [247]. Additionally, the immortalized cell lines differed significantly from primary endothelial cells in the expression of vWF, CD31 and CD34 as well as ICAM-1, IL-6 and IL-8, making them a poor substitute for primary endothelial cells, especially in regard to functional annotation and inflammatory response [248].

It could be noted that *in vitro* conditions do not necessarily reflect *in vivo* settings, which should be considered when translating and interpreting *in vitro* findings. For example, unlike cultured endothelial cells, *in vivo* ECs are in constant contact with the basement membrane as well as the pulsating blood flow and shear stress. Further, *in vitro* and *in vivo* approaches can be combined to provide complementary information [249]. For instance, further *in vivo* studies using animal models with KANK3 depletion could provide information about the role of KANK3 in vasculature.

5.1.6 Assays related to cell proliferation and migration

To study the role of KANK3 in cell migration and proliferation we have used one of the earliest developed live cell assays, published in 1965, the wound-healing assay (also commonly called gap-closing or scratch assay) [250]. The method is based on observing cells migrate and closing an artificially made wound in a monolayer of cells. It is not an exact replication of wound-healing *in vivo* (as there is a lack of supporting cell types to fully repair vascular damages (e.g., platelets) as well as lacking the inflammatory aspect of wound healing). However, the *in vitro* assay mimics the cell migration to the extent that the migration pattern and behavior will be similar [251,252]. Additionally, to determine if the cell movement is caused by migration or proliferation, we have used serum starvation to reduce cell proliferation [253]. However, we observed similar effects of KANK3 in both conditions, indicating that the impact of KANK3 depletion is on cell migration. Further, while we have used the assay to observe collective cell migration, similar observation techniques can be used to study individual cell migration by seeding cells at sub-confluence or by transfection of a fluorescent marker [252,253].

However, while the wound-healing assays gives an understanding of collective endothelial cell migration, they also only capture the movement of the cells under static conditions rather than mimic the conditions in the blood vessel. Additionally, it does not capture the effect of vessel sprouting. Recently there has been development of microfluidic assays to monitor endothelial cell migration under flow. For example, one microfluidic assay uses a three parallel fluid flow to create a wound inside the closed microfluidic channel. Two of the flows contain normal culture media, whereas the third flow contains trypsin – the cells that are exposed to this flow will detach and create a wound [254]. The same concept can also be combined with studying the effects of shear stress on cell migration [255]. However, this method requires access to highly specialized equipment, and it does not capture the 3-dimensional nature of vessel sprouting.

In *in vivo*, the endothelium is characterized by a unique tube-like formation of the blood vessels and the cells are in continuous contact with the basement membrane to maintain the tube-like structure [256]. Tube formation in endothelial cells can be studied by cultivating cells on basement membrane matrix, such as Matrigel[257]. The assay measures multiple steps involved in angiogenesis and is simple, quick, gives quantitative results and can be operated to generate high-throughput results [258,259].

Other methods include the assay developed by Nehls and Drenckhahn [260], in which beads are coated with endothelial cells and then embedded in a fibrin matrix. However, the method of generating the fibrin matrix by the use of fibrinogen is a sensitive process, the use of Matrigel offers a more stable assay that is more reproducible [261]. While the tube formation assays give an insight into the 3-dimensional landscape of vessel sprouting, the sprouts do not originate from an accessible parent vessel. To overcome some of the potential drawbacks by not mimicking the exact vasculature to study angiogenesis there has been a development of several microfluidic based assay options [262–264]. An angiogenesis assay studying endothelial cell sprouting and vessel formation from a parent vessel was only recently developed in a microfluidic concept [265]. The assay allows for observations of sprouting angiogenesis of endothelial cells from a quiescent parent vessel triggered by a chemokine gradient, mimicking *in vivo* conditions.

In the future, additional angiogenesis experiments using 3D-models mentioned above, could be used to provide additional information about KANK3 involvement in vessel sprouting and tube formation. As indicated by the scratch assay, KANK3 depletion leads to increased cell migration, in combination with KANK3 localization to cell-cell contact sites, it would be of interest to study vessel formation and stability in case KANK3 depletion leads to ‘leaky vessels’ caused by increased migration and unstable cell-cell interactions.

5.1.7 Assays related to inflammation

Paper IV used a cytokine stimulation assay to provoke an inflammatory response in cultured HUVECs, which was measured both using RT-qPCR and flow cytometry, providing complementary information as both mRNA and protein response are measured. We show that KANK3 depletion leads to an elevated response of F3 to TNFα. However, neither assay studies the functional interaction between leukocytes and HUVECs, which is an important factor in adequate immune response.

Alternative assays to study leukocyte-endothelium interactions include the use of specially designed flow perfusion chambers on top of microscopic slides, such as Ibidi Chambers [266,267]. Additionally, transmigration of neutrophils can be observed by culturing ECs on a layer of hydrated type I collagen [268], or on the upper surface of an amniotic stroma sheet [269] and monitored using e.g. phase contrast video

microscopy. These transmigration assays are static assays and do not capture some of the leukocyte-endothelial interactions that occur under flow [270,271]. Recently, a bioinspired assay has been developed which allows leukocyte interactions to be monitored in a miniaturized vascular network, mimicking in vivo conditions [272]. The assay allows for monitoring of all the steps involved in leukocyte recruitment, namely adhesion, rolling and transmigration. However, it requires access to technologies for microfluidic device fabrication.

5.1.8 Assays related to coagulation

Endothelial dysfunction is associated with increased coagulation and, in some cases, thrombus formation. As one of the end products of the coagulation cascade (and main agonist of a several coagulation feedback loops), thrombin generation is an important indicator for coagulation. EC expression of various molecules involved in coagulation, such as tissue factor, vWF (pro-coagulant) or TFPI (anti-coagulant), can modulate thrombin generation in plasma. In paper IV thrombin generation is measured on cultured endothelial cells using a calibrated automated thrombinogram (CAT) [273,274]. CAT is a technique commonly used in clinical evaluations to control coagulation capacities of the patient plasma. The assay consists of quantifying thrombin formation after addition of exogenous coagulation activators, by detecting the amount of specific fluorogenic substrate cleaved over time [275]. In our case, thrombin formation is trigger by F3 expressed directly by the ECs in the well after cytokine stimulation instead of commercial reagent containing recombinant F3, or other activators. Alternatively, coagulation can be measured using thromboelastometry [276], in which beads are coated with ECs and added to whole blood and the EC/bead incorporation into thrombi can be visualized with scanning electron microscopy (SEM). An advantage is that the method captures the entire coagulation process as it is conducted using whole blood samples, however it studies the incorporation of ECs coated beads into thrombi rather than the effect of EC dysfunction on thrombi formation.

5.2 Discussion of the main results

5.2.1 Identification of cell type-enriched transcripts

In paper III we identify a tissue-centric, cell type gene enrichment atlas of 15 human tissues, while in papers I and II we focus on individual organs of the gastrointestinal tract. Descriptions of the cell-enriched transcriptomic composition of stomach is commonly lacking from several large scale scRNAseq databases, or has been focused on specific cell types rather than complete transcriptomic profile or expression changes during pathological states [1,191,192,195]. Here we show in paper I that we are able to complement existing data with cell type-enriched transcriptomes of stomach tissue.

In papers I-III, as well as our previous studies [117,118], we show that it is possible to identify lowly expressed genes as cell type-enriched, which is a significant limitation of other methods, such as scRNAseq in which they might be excluded due to limited read depth [243]. We have also included cell types that are known to be difficult to process in scRNAseq, such as kidney podocytes [94], as well as epithelial cell subtypes in stomach tissue.

5.2.2 Identification of non-coding enriched transcripts

Paper I-II, as well as our previous study [118], show that it is possible to identify and correctly classify both protein coding and non-coding transcripts as cell type-enriched. This provides new and unique data for both tissues as there is currently no dataset of stomach enriched non-coding genes, or one with as much cell type detail in colon tissue. Additionally, there is a general lack of information regarding the function of non-coding genes in gastrointestinal healthy tissue as the focus has been on their involvement in cancer [277–279]. The identification of non-coding transcripts in cancerous tissues indicates that they have important functions, and our data could provide additional information about their role in healthy tissue or possible disease development as we identify their cell type-enriched expression profiles. For example, we identify the non-coding genes *LINC01133* and *FER1L4* as gastric mucous enriched, two non-coding genes that have been suggested to act as inhibitors of gastric cancer progression [280,281].

In general, the greatest number of non-coding genes were identified in tissue specific cell types, such as gastric mucous cells in stomach (paper I) and enteric glial

cells in colon (paper II). The expression values of non-coding genes expressed by cell types unique to gastrointestinal tissues, e.g., gastric mucous cells (paper I) and epithelial cells (paper II) were generally higher (based on mean TPM values) than in other cell types. This likely reflects the proportion of each given cell type within the samples, in addition to individual gene regulation. Further, in gastric mucous cells (paper I) and epithelial cells (paper II) we identified cell type-enriched antisense transcripts corresponding to cell enriched protein coding genes, which could suggest a local regulation of gene transcription in these cell types [282], similar to previous descriptions in adipose tissue [118].

The results presented were verified using the limited available non-coding scRNAseq data from tabula sapiens [191], which in paper I can only be used on a compartment basis (where cell types from all organs are broadly classified as endothelial, immune, stromal or epithelial in origin), as organ specific results are lacking, and in paper II can be used on 42% of the identified colon cell types. In both cases, our cell type-enriched classification corresponds well with the available data.

5.2.3 Identification sex-specific cell type-enriched transcripts

Despite reported differences in gastrointestinal function between males and females, such as gastric emptying [283], motility [284] and in both incidence and survival of gastric cancer [285,286], there is a lack of studies on the underlying differences in cell type gene expression between the sexes.

As the large GTEx datasets used in papers I and II contained over hundred samples for each sex (male and female), we conducted a sex-split subset analysis in both stomach (paper I) and colon tissue (paper II). In concordance with our previous study [118], we show that, in both tissues, the expression values are comparable between the sexes in all cell types and only identify a small subset of male-enriched transcripts. Furthermore, two of the identified male-enriched genes were non-coding pseudogenes. While it has often been assumed that pseudogenes lack specific functions, there is a recent amount of growing evidence that support a key function of pseudogenes, in roles as antisense, interference or competing endogenous transcript [287–289]. Our results could indicate that there is an additional role of pseudogenes in sex-specific gene expression.

The presented results could be used to provide further information about the observed gastrointestinal differences between the sexes. However, the genes identified as male-enriched in either stomach or colon tissue are lacking in information about their specific function, especially in gastrointestinal tissues. Further studies on their specific function in combination with our sex-specific cell type-enriched classifications could provide useful insights to the differences between the sexes.

5.2.4 Functional characterization of KANK3

In paper IV we present the functional characterization of the uncharacterized, endothelial enriched protein KANK3. KANK3 has previously been identified in vascular endothelial cells [290,291], and our results further support the critical role of KANK3 in vascular function as well as a potential involvement in thrombosis regulation.

KANK3 belongs to the KANK family, a protein family with four members in humans (KANK1-4), which are defined by their common and unique structure; a small N-terminal motif (“KN motif”), C-terminal coiled-coil domains and ankyrin repeats [292–294]. The KANK protein family is well conserved throughout evolution and has been identified as involved in actin cytoskeletal organization[295–297]. KANK1-3 have all been described as potential tumor suppression targets that either regulate cell migration or cell proliferation of hepatocellular carcinoma and lung adenocarcinoma [298,299]. Contrasting to KANK1 and KANK2, which are well studied, KANK3 has remained completely undescribed in a vascular context in vertebrates. KANK3 has been identified in vascular endothelial cells in a zebrafish homologue, as well as in vascular and lymphatic endothelial cells in human tissues (lung, pancreas and testis) [290,291].

Previous studies of the KANK protein family has suggested a role within actin cytoskeletal organization as focal adhesion proteins [295–297]. Focal adhesions are specialized protein structures that mediate cell-matrix interactions and have important role in several cellular functions such as migration, cell signaling and tissue development [300–303]. Further, live-cell microscopy studies of KANK3 knockdown ECs showed that KANK3 depletion results in an increased cell motility independent of cell proliferation. These results indicate an important role of KANK3 in addition to thrombosis regulation within modulating cellular migration. The results were further supported by immunohistochemistry studies were KANK3 was observed to

accumulate in cell-cell interaction sites. Additionally, shear stress studies resulted in an upregulation of KANK3, on both mRNA and protein level, indicating a potential role in anchoring the cell to the basal membrane. Furthermore, the shear stress studies show that KANK3 depletion leads to reduced expression of vimentin. Vimentin has important roles in several EC functions, such as cell migration, polarity and differentiation [304–307], as well as a vital role in cell adhesion and EC sprouting [308].

Our results show that, knockdown of KANK3 in ECs resulted in an upregulation of the prothrombotic tissue factor (TF/F3). These results indicate that loss of KANK3 expression can cause a shift towards a prothrombotic state, thereby indicating an important role of KANK3 in the maintenance of the balance between the pro- and antithrombotic factors. These results are further supported by an enhancement of thrombin formation triggered by TF after KANK3 depletion. Thrombin is a crucial enzyme involved in the coagulation cascade, thus the accelerated thrombin formation as result of KANK3 depletion suggest a potential prothrombotic phenotype.

Further studies are needed, such as 3D-angiogenesis assays or *in vivo* models, to fully elucidate the intricate function and precise molecular mechanisms of KANK3 within the vasculature, including its precise localization and its potential role as a therapeutic target for thrombotic disorders.

6 Conclusion

To shortly summarize the results of papers I-IV, in concordance with our previous studies, we have used the integrative correlation-based method on unfractionated bulk RNAseq data from multiple tissues to identify individual cell type enrichment signatures. Further application of functional assays on transcript with previously unidentified function can reveal important involvement in cell specific processes both on cell and tissue level.

- In paper I we use the integrative correlation-based method to identify cell enriched transcriptomes of stomach tissue, successfully identifying transcriptomes of several epithelial cell subtypes. Further, we identified both protein-coding and non-coding cell type enriched genes, which were either supported by available protein-profiling or scRNAseq data. Lastly, we identified a small panel of male-enriched chief cell transcripts.
- In paper II we use the integrative correlations-based method to identify cell enriched transcriptomes of sigmoid colon tissue, additionally we identify transcriptomes of two cell types constituting the enteric nervous system. We identify both protein-coding and non-coding genes as cell type-enriched, as well as a small panel of male-enriched genes.
- In paper III, we extended the study to incorporate 15 different human tissues. In addition to identifying cell-type enriched genes for the constituent cell types for all included tissues, we identified gene enrichment signature profiles for cell types found in all or most tissue types (referred to as core cell types), including endothelial cells.
- In paper IV, we show that the previously uncharacterized gene, KANK3, that we classified as endothelial-enriched in paper I-III, has an important function in endothelial cell specific functions such as coagulation and in cell migration. Additionally, we show that KANK3 is induced by shear stress and that KANK3 depletion leads to subcellular redistribution of vimentin.

7 Final remarks and future perspectives

The data and results presented in this thesis support the use of *in silico* methods to extract additional data, such as cell type-enriched transcriptomes, from previously published, large-scale, bulk RNAseq studies. We successfully identified cell type-enriched transcriptomes of 15 human tissues, and in the future even more tissues can be incorporated.

As mentioned, a limitation of deconvolution methods is the dependency of suitable reference transcripts for the constitutional cell types. However, the method is easily adjustable, and as new reference transcripts (that fulfill the criteria) are identified, more cell types can be included in the analysis. In paper I and II we analyzed two organs that are part of the gastrointestinal tract, in the future, this analysis could be expanded to include the additional gastrointestinal organs, such as the small intestine, to investigate the large-scale gastrointestinal transcriptome. Both stomach and colon, papers I-II, are well studied in the context of gastrointestinal cancer and inflammatory bowel disease, with bulk RNAseq data available. This data could be analyzed to identify disease specific cell type-enriched transcriptomes, using the integrative correlations-based method, which might aid in the disease treatment if a biomarker can be identified. In paper I-II, we identified multiple non-coding genes as enriched, these results could be expanded further if more data on non-coding genes become available. Further, paper III could be expanded to include non-coding genes, as well as additional tissues. The results presented in paper I-III could be used to select candidate genes for functional studies (in any of the analyzed cell types), highlighted by the results presented in paper IV.

In paper IV, we presented the functional annotation of a previously uncharacterized, endothelial-enriched gene KANK3. While the data presented shows promising results supporting an important function of KANK3 in endothelial cells, several of the experiments need more replicates to strengthen the data. In the near future, such validating experiments should be carried out using multiple HUVEC donors. It would also be of interest to further investigate the function of KANK3 in coagulation by running a fibrin formation assay, which could show if KANK3 has an additional role in thrombosis formation by fibrin accumulation on the endothelial surface. Additional experiments could include further studies on KANK3 localization within ECs, as well as in EC-EC contact sites and during EC migration.

8 References

[1] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife 2017;6:e27041. https://doi.org/10.7554/eLife.27041.

[2] Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Proteomics. Tissue-based map of the human proteome. Science 2015;347:1260419. https://doi.org/10.1126/science.1260419.

[3] Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187–97. https://doi.org/10.1038/nature09792.

[4] S O. So much “junk” DNA in our genome. In “Evolution of Genetic Systems.” Brookhaven Symp Biol 1972;23:366–70.

[5] Eddy SR. Non–coding RNA genes and the modern RNA world. Nat Rev Genet 2001;2:919–29. https://doi.org/10.1038/35103511.

[6] Mattick JS. The central role of RNA in human development and cognition. FEBS Lett 2011;585:1600–16. https://doi.org/10.1016/j.febslet.2011.05.001.

[7] Griffin HG, Griffin AM. DNA sequencing. Appl Biochem Biotechnol 1993;38:147–59. https://doi.org/10.1007/BF02916418.

[8] Adams MD, Kerlavage AR, Fields C, Venter JC. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 1993;4:256–67. https://doi.org/10.1038/ng0793-256.

[9] Adams MD, Kerlavage AR, Fleischmann RD, Fuldner RA, Bult CJ, Lee NH, et al. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 1995;377:3–174.

[10] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial Analysis of Gene Expression. Science 1995;270:484–7. https://doi.org/10.1126/science.270.5235.484.

[11] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci 2003;100:15776–81. https://doi.org/10.1073/pnas.2136655100.

[12] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000;18:630–4. https://doi.org/10.1038/76469.

[13] Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci 1997;94:13057–62. https://doi.org/10.1073/pnas.94.24.13057.

[14] Clark TA, Sugnet CW, Ares M. Genomewide Analysis of mRNA Processing in Yeast Using Splicing-Specific Microarrays. Science 2002;296:907–10. https://doi.org/10.1126/science.1069415.

[15] Schena M, Shalon D, Davis RW, Brown PO. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science 1995;270:467–70. https://doi.org/10.1126/science.270.5235.467.

[16] Murphy D. Gene expression studies using microarrays: principles, problems, and prospects. Adv Physiol Educ 2002;26:256–70. https://doi.org/10.1152/advan.00043.2002.

[17] Castle JC, Zhang C, Shah JK, Kulkarni AV, Kalsotra A, Cooper TA, et al. Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. Nat Genet 2008;40:1416–25. https://doi.org/10.1038/ng.264.

[18] Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 2006;7:276. https://doi.org/10.1186/1471-2105-7-276.

[19] Pan Q, Shai O, Misquitta C, Zhang W, Saltzman AL, Mohammad N, et al. Revealing Global Regulatory Features of Mammalian Alternative Splicing Using a Quantitative Microarray Platform. Mol Cell 2004;16:929–41. https://doi.org/10.1016/j.molcel.2004.12.004.

[20] Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 2008;5:613–9. https://doi.org/10.1038/nmeth.1223.

[21] Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science 2008;320:1344–9. https://doi.org/10.1126/science.1158441.

[22] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8. https://doi.org/10.1038/nmeth.1226.

[23] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63. https://doi.org/10.1038/nrg2484.

[24] Salzberg SL. Recent advances in RNA sequence analysis. F1000 Biol Rep 2010;2:64. https://doi.org/10.3410/B2-64.

[25] Kashima Y, Suzuki A, Suzuki Y. An Informative Approach to Single-Cell Sequencing Analysis. In: Suzuki Y, editor. Single Mol. Single Cell Seq., Singapore: Springer; 2019, p. 81–96. https://doi.org/10.1007/978-981-13-6037-4_6.

[26] Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature 2017;546:431–5. https://doi.org/10.1038/nature22794.

[27] Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 2013;498:236–40. https://doi.org/10.1038/nature12172.

[28] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet 2016;17:175–88. https://doi.org/10.1038/nrg.2015.16.

[29] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–30. https://doi.org/10.1038/nrg3542.

[30] Grün D, van Oudenaarden A. Design and Analysis of Single-Cell Sequencing Experiments. Cell 2015;163:799–810. https://doi.org/10.1016/j.cell.2015.10.039.

[31] O'Flanagan CH, Campbell KR, Zhang AW, Kabeer F, Lim JL, Biele J, et al. Dissociation of solid tumour tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. bioRxiv 2019:683227. https://doi.org/10.1101/683227.

[32] Salmén F, Ståhl PL, Mollbrink A, Navarro JF, Vickovic S, Frisén J, et al. Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections. Nat Protoc 2018;13:2501–34. https://doi.org/10.1038/s41596-018-0045-2.

[33] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931–45. https://doi.org/10.1038/nature03001.

[34] Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73. https://doi.org/10.1038/nature09534.

[35] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74. https://doi.org/10.1038/nature11247.

[36] Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 2013;132:1077–130. https://doi.org/10.1007/s00439-013-1331-2.

[37] Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, et al. Genomics and the origin of species. Nat Rev Genet 2014;15:176–92. https://doi.org/10.1038/nrg3644.

[38] Allendorf FW, Hohenlohe PA, Luikart G. Genomics and the future of conservation genetics. Nat Rev Genet 2010;11:697–709. https://doi.org/10.1038/nrg2844.

[39] Hardison RC. Comparative Genomics. PLOS Biol 2003;1:e58. https://doi.org/10.1371/journal.pbio.0000058.

[40] Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008;582:1977–86. https://doi.org/10.1016/j.febslet.2008.03.004.

[41] Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, et al. Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae*S. Mol Cell Proteomics 2002;1:323–33. https://doi.org/10.1074/mcp.M200001-MCP200.

[42] Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between Protein and mRNA Abundance in Yeast. Mol Cell Biol 1999;19:1720–30. https://doi.org/10.1128/MCB.19.3.1720.

[43] Krishna RG, Wold F. Post-Translational Modifications of Proteins. In: Imahori K, Sakiyama F, editors. Methods Protein Seq. Anal., Boston, MA: Springer US; 1993, p. 167–72. https://doi.org/10.1007/978-1-4899-1603-7_21.

[44] Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA. Clinical proteomics: translating benchside promise into bedside reality. Nat Rev Drug Discov 2002;1:683–95. https://doi.org/10.1038/nrd891.

[45] Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, et al. A proteomic view of the Plasmodium falciparum life cycle. Nature 2002;419:520–6. https://doi.org/10.1038/nature01107.

[46] Lasonder E, Ishihama Y, Andersen JS, Vermunt AMW, Pain A, Sauerwein RW, et al. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 2002;419:537–42. https://doi.org/10.1038/nature01111.

[47] Magaki S, Hojat SA, Wei B, So A, Yong WH. An Introduction to the Performance of Immunohistochemistry. Methods Mol Biol Clifton NJ 2019;1897:289–98. https://doi.org/10.1007/978-1-4939-8935-5_25.

[48] Chen Z, Dodig-Crnković T, Schwenk JM, Tao S. Current applications of antibody microarrays. Clin Proteomics 2018;15:7. https://doi.org/10.1186/s12014-018-9184-2.

[49] Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 2008;453:1239–43. https://doi.org/10.1038/nature07002.

[50] Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. Plant J 2007;51:910–8. https://doi.org/10.1111/j.1365-313X.2007.03193.x.

[51] Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res 2007;17:69–73. https://doi.org/10.1101/gr.5145806.

[52] Li X, Wang C-Y. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci 2021;13:1–6. https://doi.org/10.1038/s41368-021-00146-0.

[53] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 2008;133:523–36. https://doi.org/10.1016/j.cell.2008.03.029.

[54] Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet 2019;20:631–56. https://doi.org/10.1038/s41576-019-0150-2.

[55] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012;489:101–8. https://doi.org/10.1038/nature11233.

[56] Tilgner H, Jahanbani F, Gupta I, Collier P, Wei E, Rasmussen M, et al. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome Res 2018;28:231–42. https://doi.org/10.1101/gr.230516.117.

[57] Gazzoli I, Pulyakhina I, Verwey NE, Ariyurek Y, Laros JFJ, 't Hoen PAC, et al. Non-sequential and multi-step splicing of the dystrophin transcript. RNA Biol 2016;13:290–305. https://doi.org/10.1080/15476286.2015.1125074.

[58] Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. BMC Bioinformatics 2017;18:38. https://doi.org/10.1186/s12859-016-1457-z.

[59] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 2015;16:59–70. https://doi.org/10.1093/bib/bbt086.

[60] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17:13. https://doi.org/10.1186/s13059-016-0881-8.

[61] Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun 2017;8:59. https://doi.org/10.1038/s41467-017-00050-4.

[62] Islam S, Zeisel A, Joost S, La Manno G, Zajac P, Kasper M, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat Methods 2014;11:163–6. https://doi.org/10.1038/nmeth.2772.

[63] Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2012;9:72–4. https://doi.org/10.1038/nmeth.1778.

[64] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Rätsch G, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 2013;10:1185–91. https://doi.org/10.1038/nmeth.2722.

[65] Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol 2013;31:1009–14. https://doi.org/10.1038/nbt.2705.

[66] Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. PLOS ONE 2016;11:e0157779. https://doi.org/10.1371/journal.pone.0157779.

[67] Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res 2018;46:2159–68. https://doi.org/10.1093/nar/gky066.

[68] Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 2018;15:201–6. https://doi.org/10.1038/nmeth.4577.

[69] Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang X-J, et al. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research 2017;6:100. https://doi.org/10.12688/f1000research.10571.2.

[70] Chen P-L, Roh W, Reuben A, Cooper ZA, Spencer CN, Prieto PA, et al. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. Cancer Discov 2016;6:827–37. https://doi.org/10.1158/2159-8290.CD-15-1545.

[71] Shukla S, Evans JR, Malik R, Feng FY, Dhanasekaran SM, Cao X, et al. Development of a RNA-Seq Based Prognostic Signature in Lung Adenocarcinoma. JNCI J Natl Cancer Inst 2017;109:djw200. https://doi.org/10.1093/jnci/djw200.

[72] Zhou J-G, Liang B, Jin S-H, Liao H-L, Du G-B, Cheng L, et al. Development and Validation of an RNA-Seq-Based Prognostic Signature in Neuroblastoma. Front Oncol 2019;9.

[73] Han L, Li X, Cao M, Cao Y, Zhou L. Development and validation of an individualized diagnostic signature in thyroid cancer. Cancer Med 2018;7:1135–40. https://doi.org/10.1002/cam4.1397.

[74] Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 2009;6:377–82. https://doi.org/10.1038/nmeth.1315.

[75] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell 2015;161:1187–201. https://doi.org/10.1016/j.cell.2015.04.044.

[76] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell 2015;161:1202–14. https://doi.org/10.1016/j.cell.2015.05.002.

[77] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049. https://doi.org/10.1038/ncomms14049.

[78] Leelatian N, Doxie DB, Greenplate AR, Mobley BC, Lehman JM, Sinnaeve J, et al. Single cell analysis of human tissues and solid tumors with mass cytometry. Cytometry B Clin Cytom 2017;92:68–78. https://doi.org/10.1002/cyto.b.21481.

[79] Brehm-Stecher BF, Johnson EA. Single-Cell Microbiology: Tools, Technologies, and Applications. Microbiol Mol Biol Rev 2004;68:538–59. https://doi.org/10.1128/MMBR.68.3.538-559.2004.

[80] Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 2018;50:1–14. https://doi.org/10.1038/s12276-018-0071-8.

[81] Julius MH, Masuda T, Herzenberg LA. Demonstration That Antigen-Binding Cells Are Precursors of Antibody-Producing Cells After Purification with a Fluorescence-Activated Cell Sorter. Proc Natl Acad Sci 1972;69:1934–8. https://doi.org/10.1073/pnas.69.7.1934.

[82] Hines WC, Su Y, Kuhn I, Polyak K, Bissell MJ. Sorting Out the FACS: A Devil in the Details. Cell Rep 2014;6:779–81. https://doi.org/10.1016/j.celrep.2014.02.021.

[83] Whitesides GM. The origins and the future of microfluidics. Nature 2006;442:368–73. https://doi.org/10.1038/nature05058.

[84] Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods 2013;10:1093–5. https://doi.org/10.1038/nmeth.2645.

[85] Goldstein LD, Chen Y-JJ, Dunne J, Mir A, Hubschle H, Guillory J, et al. Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics 2017;18:519. https://doi.org/10.1186/s12864-017-3893-1.

[86] Pan Y, Landis JT, Moorad R, Wu D, Marron JS, Dittmer DP. The Poisson distribution model fits UMI-based single-cell RNA-sequencing data. BMC Bioinformatics 2023;24:256. https://doi.org/10.1186/s12859-023-05349-2.

[87] Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res 2011;21:1160–7. https://doi.org/10.1101/gr.110882.110.

[88] Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. Cell Rep 2012;2:666–73. https://doi.org/10.1016/j.celrep.2012.08.003.

[89] McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 2017;33:1179–86. https://doi.org/10.1093/bioinformatics/btw777.

[90] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36:411–20. https://doi.org/10.1038/nbt.4096.

[91] Senabouth A, Lukowski SW, Hernandez JA, Andersen SB, Mei X, Nguyen QH, et al. ascend: R package for analysis of single-cell RNA-seq data. GigaScience 2019;8:giz087. https://doi.org/10.1093/gigascience/giz087.

[92] Jiang R, Sun T, Song D, Li JJ. Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol 2022;23:31. https://doi.org/10.1186/s13059-022-02601-5.

[93] O'Flanagan CH, Campbell KR, Zhang AW, Kabeer F, Lim JLP, Biele J, et al. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. Genome Biol 2019;20:210. https://doi.org/10.1186/s13059-019-1830-0.

[94] Denisenko E, Guo BB, Jones M, Hou R, de Kock L, Lassmann T, et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol 2020;21:130. https://doi.org/10.1186/s13059-020-02048-6.

[95] Massoni-Badosa R, Iacono G, Moutinho C, Kulis M, Palau N, Marchese D, et al. Sampling time-dependent artifacts in single-cell genomics studies. Genome Biol 2020;21:112. https://doi.org/10.1186/s13059-020-02032-0.

[96] Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun 2021;12:5692. https://doi.org/10.1038/s41467-021-25960-2.

[97] Denninger JK, Walker LA, Chen X, Turkoglu A, Pan A, Tapp Z, et al. Robust Transcriptional Profiling and Identification of Differentially Expressed Genes With Low Input RNA Sequencing of Adult Hippocampal Neural Stem and Progenitor Populations. Front Mol Neurosci 2022;15:810722. https://doi.org/10.3389/fnmol.2022.810722.

[98] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 2018;13:599–604. https://doi.org/10.1038/nprot.2017.149.

[99] Insel TR, Landis SC, Collins FS. The NIH BRAIN Initiative. Science 2013;340:687–8. https://doi.org/10.1126/science.1239276.

[100] Chen J, Suo S, Tam PP, Han J-DJ, Peng G, Jing N. Spatial transcriptomic analysis of cryosectioned tissue samples with Geo-seq. Nat Protoc 2017;12:566–80. https://doi.org/10.1038/nprot.2017.003.

[101] Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 2016;353:78–82. https://doi.org/10.1126/science.aaf2403.

[102] Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science 2019;363:1463–7. https://doi.org/10.1126/science.aaw1219.

[103] Ke R, Mignardi M, Pacureanu A, Svedlund J, Botling J, Wählby C, et al. In situ sequencing for RNA analysis in preserved tissue and cells. Nat Methods 2013;10:857–60. https://doi.org/10.1038/nmeth.2563.

[104] Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, Cai L. Single-cell in situ RNA profiling by sequential hybridization. Nat Methods 2014;11:360–1. https://doi.org/10.1038/nmeth.2892.

[105] Wang X, Allen WE, Wright MA, Sylwestrak EL, Samusik N, Vesuna S, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 2018;361:eaat5691. https://doi.org/10.1126/science.aat5691.

[106] Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly Multiplexed Subcellular RNA Sequencing in Situ. Science 2014;343:1360–3. https://doi.org/10.1126/science.1250212.

[107] Maniatis S, Äijö T, Vickovic S, Braine C, Kang K, Mollbrink A, et al. Spatiotemporal dynamics of molecular pathology in amyotrophic lateral sclerosis. Science 2019;364:89–93. https://doi.org/10.1126/science.aav9776.

[108] Chen A, Liao S, Cheng M, Ma K, Wu L, Lai Y, et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. Cell 2022;185:1777-1792.e21. https://doi.org/10.1016/j.cell.2022.04.003.

[109] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods 2015;12:453–7. https://doi.org/10.1038/nmeth.3337.

[110] Kuhn A, Kumar A, Beilina A, Dillman A, Cookson MR, Singleton AB. Cell population-specific expression analysis of human cerebellum. BMC Genomics 2012;13:610. https://doi.org/10.1186/1471-2164-13-610.

[111] Avila Cobos F, Vandesompele J, Mestdagh P, De Preter K. Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics 2018;34:1969–79. https://doi.org/10.1093/bioinformatics/bty019.

[112] Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of Blood Microarray Data Identifies Cellular Activation Patterns in Systemic Lupus Erythematosus. PLOS ONE 2009;4:e6098. https://doi.org/10.1371/journal.pone.0006098.

[113] Gaujoux R, Seoighe C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study. Infect Genet Evol 2012;12:913–21. https://doi.org/10.1016/j.meegid.2011.08.014.

[114] Hoffmann M, Pohlers D, Koczan D, Thiesen H-J, Wölfl S, Kinne RW. Robust computational reconstitution – a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. BMC Bioinformatics 2006;7:369. https://doi.org/10.1186/1471-2105-7-369.

[115] Sutton GJ, Poppe D, Simmons RK, Walsh K, Nawaz U, Lister R, et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. Nat Commun 2022;13:1358. https://doi.org/10.1038/s41467-022-28655-4.

[116] Butler LM, Hallström BM, Fagerberg L, Pontén F, Uhlén M, Renné T, et al. Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome. Cell Syst 2016;3:287-301.e3. https://doi.org/10.1016/j.cels.2016.08.001.

[117] Dusart P, Hallstrom BM, Renne T, Odeberg J, Uhlen M, Butler L. A systems-based map of human brain cell-type enriched genes and malignancy-associated endothelial changes. bioRxiv 2019:528414. https://doi.org/10.1101/528414.

[118] Norreen-Thorsen M, Struck EC, Öling S, Zwahlen M, Von Feilitzen K, Odeberg J, et al. A human adipose tissue cell-type transcriptome atlas. Cell Rep 2022;40:111046. https://doi.org/10.1016/j.celrep.2022.111046.

[119] Yim AKY, Wang PL, Bermingham JR, Hackett A, Strickland A, Miller TM, et al. Disentangling glial diversity in peripheral nerves at single-nuclei resolution. Nat Neurosci 2022;25:238–51. https://doi.org/10.1038/s41593-021-01005-1.

[120] Piwecka M, Rajewsky N, Rybak-Wolf A. Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. Nat Rev Neurol 2023:1–17. https://doi.org/10.1038/s41582-023-00809-y.

[121] Rondini EA, Granneman JG. Single cell approaches to address adipose tissue stromal cell heterogeneity. Biochem J 2020;477:583–600. https://doi.org/10.1042/BCJ20190467.

[122] Viswanadha S, Londos C. Optimized conditions for measuring lipolysis in murine primary adipocytes. J Lipid Res 2006;47:1859–64. https://doi.org/10.1194/jlr.D600005-JLR200.

[123] Davies MG, Hagen PO. The vascular endothelium. A new horizon. Ann Surg 1993;218:593–609.

[124] Yau JW, Teoh H, Verma S. Endothelial cell control of thrombosis. BMC Cardiovasc Disord 2015;15:130. https://doi.org/10.1186/s12872-015-0124-z.

[125] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Blood Vessels and Endothelial Cells. Mol Biol Cell 4th Ed 2002.

[126] Krüger-Genge A, Blocki A, Franke R-P, Jung F. Vascular Endothelial Cell Biology: An Update. Int J Mol Sci 2019;20:4411. https://doi.org/10.3390/ijms20184411.

[127] Yau JW, Teoh H, Verma S. Endothelial cell control of thrombosis. BMC Cardiovasc Disord 2015;15:130. https://doi.org/10.1186/s12872-015-0124-z.

[128] Langille BL, Adamson SL. Relationship between blood flow direction and endothelial cell orientation at arterial branch sites in rabbits and mice. Circ Res 1981;48:481–8. https://doi.org/10.1161/01.RES.48.4.481.

[129] Redmond EM, Cahill PA, Sitzmann JV. Flow-Mediated Regulation of G-Protein Expression in Cocultured Vascular Smooth Muscle and Endothelial Cells. Arterioscler Thromb Vasc Biol 1998;18:75–83. https://doi.org/10.1161/01.ATV.18.1.75.

[130] Kuchan MJ, Frangos JA. Shear stress regulates endothelin-1 release via protein kinase C and cGMP in cultured endothelial cells. Am J Physiol-Heart Circ Physiol 1993;264:H150–6. https://doi.org/10.1152/ajpheart.1993.264.1.H150.

[131] Hendrickson RJ, Cappadona C, Yankah EN, Sitzmann JV, Cahill PA, Redmond EM. Sustained Pulsatile Flow Regulates Endothelial Nitric Oxide Synthase and Cyclooxygenase Expression in Co-Cultured Vascular Endothelial and Smooth Muscle Cells. J Mol Cell Cardiol 1999;31:619–29. https://doi.org/10.1006/jmcc.1998.0898.

[132] Bhagyalakshmi A, Frangos JA. Mechanism of shear-induced prostacyclin production in endothelial cells. Biochem Biophys Res Commun 1989;158:31–7. https://doi.org/10.1016/S0006-291X(89)80172-X.

[133] Garg UC, Hassid A. Nitric oxide-generating vasodilators and 8-bromo-cyclic guanosine monophosphate inhibit mitogenesis and proliferation of cultured rat vascular smooth muscle cells. J Clin Invest 1989;83:1774–7. https://doi.org/10.1172/JCI114081.

[134] Alberts GF, Peifley KA, Johns A, Kleha JF, Winkles JA. Constitutive endothelin-1 overexpression promotes smooth muscle cell proliferation via an external autocrine loop. J Biol Chem 1994;269:10112–8. https://doi.org/10.1016/S0021-9258(17)36997-1.

[135] Kohno M, Yokokawa K, Yasunari K, Kano H, Minami M, Yoshikawa J. Effect of the Endothelin Family of Peptides on Human Coronary Artery Smooth-Muscle Cell Migration. J Cardiovasc Pharmacol 1998;31:S84.

[136] Takada Y, Shinkai F, Kondo S, Yamamoto S, Tsuboi H, Korenaga R, et al. Fluid Shear Stress Increases the Expression of Thrombomodulin by Cultured Human Endothelial Cells. Biochem Biophys Res Commun 1994;205:1345–52. https://doi.org/10.1006/bbrc.1994.2813.

[137] Arisaka T, Mitsumata M, Kawasumi M, Tohjima T, Hirose S, Yoshida Y. Effects of shear stress on glycosaminoglycan synthesis in vascular endothelial cells. Ann N Y Acad Sci 1995;748:543–54. https://doi.org/10.1111/j.1749-6632.1994.tb17359.x.

[138] Diamond SL, Eskin SG, McIntire LV. Fluid Flow Stimulates Tissue Plasminogen Activator Secretion by Cultured Human Endothelial Cells. Science 1989;243:1483–5. https://doi.org/10.1126/science.2467379.

[139] JIN RC, VOETSCH B, LOSCALZO J. Endogenous Mechanisms of Inhibition of Platelet Function. Microcirculation 2005;12:247–58. https://doi.org/10.1080/10739680590925493.

[140] Tanaka KA, Key NS, Levy JH. Blood Coagulation: Hemostasis and Thrombin Regulation. Anesth Analg 2009;108:1433. https://doi.org/10.1213/ane.0b013e31819bcc9c.

[141] Mackman N. The many faces of tissue factor. J Thromb Haemost 2009;7:136–9. https://doi.org/10.1111/j.1538-7836.2009.03368.x.

[142] Giesen PLA, Fyfe BS, Fallon JT, Roque M, Mendlowitz M, Rossikhina M, et al. Intimal Tissue Factor Activity Is Released from the Arterial Wall after Injury. Thromb Haemost 2000;83:622–8. https://doi.org/10.1055/s-0037-1613874.

[143] Murata T, Nakashima Y, Yasunaga C, Maeda K, Sueishi K. Extracellular and cell-associated localizations of plasminogen activators and plasminogen activator inhibitor-1 in cultured endothelium. Exp Mol Pathol 1991;55:105–18. https://doi.org/10.1016/0014-4800(91)90046-Z.

[144] Chapin JC, Hajjar KA. Fibrinolysis and the control of blood coagulation. Blood Rev 2015;29:17–24. https://doi.org/10.1016/j.blre.2014.09.003.

[145] McEver RP. Selectins: initiators of leucocyte adhesion and signalling at the vascular wall. Cardiovasc Res 2015;107:331–9. https://doi.org/10.1093/cvr/cvv154.

[146] Vestweber D, Blanks JE. Mechanisms That Regulate the Function of the Selectins and Their Ligands. Physiol Rev 1999;79:181–213. https://doi.org/10.1152/physrev.1999.79.1.181.

[147] van Buul JD, Kanters E, Hordijk PL. Endothelial Signaling by Ig-Like Cell Adhesion Molecules. Arterioscler Thromb Vasc Biol 2007;27:1870–6. https://doi.org/10.1161/ATVBAHA.107.145821.

[148] Wegmann F, Petri B, Khandoga AG, Moser C, Khandoga A, Volkery S, et al. ESAM supports neutrophil extravasation, activation of Rho, and VEGF-induced vascular permeability. J Exp Med 2006;203:1671–7. https://doi.org/10.1084/jem.20060565.

[149] Schenkel AR, Mamdouh Z, Chen X, Liebman RM, Muller WA. CD99 plays a major role in the migration of monocytes through endothelial junctions. Nat Immunol 2002;3:143–50. https://doi.org/10.1038/ni749.

[150] Nourshargh S, Krombach F, Dejana E. The role of JAM-A and PECAM-1 in modulating leukocyte infiltration in inflamed and ischemic tissues. J Leukoc Biol 2006;80:714–8. https://doi.org/10.1189/jlb.1105645.

[151] Muller WA. The role of PECAM-1 (CD31) in leukocyte emigration: studies in vitro and in vivo. J Leukoc Biol 1995;57:523–8. https://doi.org/10.1002/jlb.57.4.523.

[152] Carman CV, Springer TA. A transmigratory cup in leukocyte diapedesis both through individual vascular endothelial cells and between them. J Cell Biol 2004;167:377–88. https://doi.org/10.1083/jcb.200404129.

[153] Woodfin A, Voisin M-B, Beyrau M, Colom B, Caille D, Diapouli F-M, et al. The junctional adhesion molecule JAM-C regulates polarized transendothelial migration of neutrophils in vivo. Nat Immunol 2011;12:761–9. https://doi.org/10.1038/ni.2062.

[154] Risau W, Flamme I. Vasculogenesis. Annu Rev Cell Dev Biol 1995;11:73–91. https://doi.org/10.1146/annurev.cb.11.110195.000445.

[155] Ferrara N, Gerber H-P, LeCouter J. The biology of VEGF and its receptors. Nat Med 2003;9:669–76. https://doi.org/10.1038/nm0603-669.

[156] Benjamin LE, Golijanin D, Itin A, Pode D, Keshet E. Selective ablation of immature blood vessels in established human tumors follows vascular endothelial growth factor withdrawal. J Clin Invest 1999;103:159–65.

[157] Shweiki D, Itin A, Soffer D, Keshet E. Vascular endothelial growth factor induced by hypoxia may mediate hypoxia-initiated angiogenesis. Nature 1992;359:843–5. https://doi.org/10.1038/359843a0.

[158] Gerhardt H, Golding M, Fruttiger M, Ruhrberg C, Lundkvist A, Abramsson A, et al. VEGF guides angiogenic sprouting utilizing endothelial tip cell filopodia. J Cell Biol 2003;161:1163–77. https://doi.org/10.1083/jcb.200302047.

[159] Vanhoutte PM. How to assess endothelial function in human blood vessels. J Hypertens 1999;17:1047.

[160] Rubanyi GM. The role of endothelium in cardiovascular homeostasis and diseases. J Cardiovasc Pharmacol 1993;22 Suppl 4:S1-14. https://doi.org/10.1097/00005344-199322004-00002.

[161] Ruilope LM, Redón J, Schmieder R. Cardiovascular risk reduction by reversing endothelial dysfunction:ARBs, ACE inhibitors, or both? Expectations from The ONTARGET Trial Programme. Vasc Health Risk Manag 2007;3:1–9. https://doi.org/10.2147/vhrm.s12187331.

[162] Rajendran P, Rengarajan T, Thangavel J, Nishigaki Y, Sakthisekaran D, Sethi G, et al. The Vascular Endothelium and Human Diseases. Int J Biol Sci 2013;9:1057. https://doi.org/10.7150/ijbs.7502.

[163] Cooke JP, Tsao PS. Go With the Flow. Circulation 2001;103:2773–5. https://doi.org/10.1161/01.CIR.103.23.2773.

[164] Kawano H, Do YS, Kawano Y, Starnes V, Barr M, Law RE, et al. Angiotensin II Has Multiple Profibrotic Effects in Human Cardiac Fibroblasts. Circulation 2000;101:1130–7. https://doi.org/10.1161/01.CIR.101.10.1130.

[165] Esper RJ, Nordaby RA, Vilariño JO, Paragano A, Cacharrón JL, Machado RA. Endothelial dysfunction: a comprehensive appraisal. Cardiovasc Diabetol 2006;5:4. https://doi.org/10.1186/1475-2840-5-4.

[166] Nickenig G, Harrison DG. The AT1-Type Angiotensin Receptor in Oxidative Stress and Atherogenesis. Circulation 2002;105:393–6. https://doi.org/10.1161/hc0302.102618.

[167] Peiser L, Mukhopadhyay S, Gordon S. Scavenger receptors in innate immunity. Curr Opin Immunol 2002;14:123–8. https://doi.org/10.1016/S0952-7915(01)00307-7.

[168] Leitinger N. Oxidized phospholipids as modulators of inflammation in atherosclerosis. Curr Opin Lipidol 2003;14:421.

[169] Li D, Mehta JL. Antisense to LOX-1 Inhibits Oxidized LDL–Mediated Upregulation of Monocyte Chemoattractant Protein-1 and Monocyte Adhesion to Human Coronary Artery Endothelial Cells. Circulation 2000;101:2889–95. https://doi.org/10.1161/01.CIR.101.25.2889.

[170] Tsimikas S, Witztum JL. Measuring Circulating Oxidized Low-Density Lipoprotein to Evaluate Coronary Risk. Circulation 2001;103:1930–2. https://doi.org/10.1161/01.CIR.103.15.1930.

[171] Li D, Mehta JL. Upregulation of Endothelial Receptor for Oxidized LDL (LOX-1) by Oxidized LDL and Implications in Apoptosis of Human Coronary Artery Endothelial Cells. Arterioscler Thromb Vasc Biol 2000;20:1116–22. https://doi.org/10.1161/01.ATV.20.4.1116.

[172] Pober JS, Sessa WC. Evolving functions of endothelial cells in inflammation. Nat Rev Immunol 2007;7:803–15. https://doi.org/10.1038/nri2171.

[173] Thomas H. Gut endothelial cells — another line of defence. Nat Rev Gastroenterol Hepatol 2016;13:4–4. https://doi.org/10.1038/nrgastro.2015.205.

[174] Heidemann J, Domschke W, Kucharzik T, Maaser C. Intestinal Microvascular Endothelium and Innate Immunity in Inflammatory Bowel Disease: a Second Line of Defense? Infect Immun 2006;74:5425–32. https://doi.org/10.1128/IAI.00248-06.

[175] Greenwood-Van Meerveld B, Johnson AC, Grundy D. Gastrointestinal Physiology and Function. In: Greenwood-Van Meerveld B, editor. Gastrointest. Pharmacol., Cham: Springer International Publishing; 2017, p. 1–16. https://doi.org/10.1007/164_2016_118.

[176] Choi E, Roland JT, Barlow BJ, O'Neal R, Rich AE, Nam KT, et al. Cell lineage distribution atlas of the human stomach reveals heterogeneous gland populations in the gastric antrum. Gut 2014;63:1711–20. https://doi.org/10.1136/gutjnl-2013-305964.

[177] de Santa Barbara P, van den Brink GR, Roberts DJ. Development and differentiation of the intestinal epithelium. Cell Mol Life Sci CMLS 2003;60:1322–32. https://doi.org/10.1007/s00018-003-2289-3.

[178] Thompson CA, DeLaForest A, Battle MA. Patterning the gastrointestinal epithelium to confer regional-specific functions. Dev Biol 2018;435:97–108. https://doi.org/10.1016/j.ydbio.2018.01.006.

[179] Kim Y, Pritts TA. The Gastrointestinal Tract. In: Luchette FA, Yelon JA, editors. Geriatr. Trauma Crit. Care, Cham: Springer International Publishing; 2017, p. 35–43. https://doi.org/10.1007/978-3-319-48687-1_5.

[180] Laukoetter MG, Nava P, Nusrat A. Role of the intestinal barrier in inflammatory bowel disease. World J Gastroenterol 2008;14:401. https://doi.org/10.3748/wjg.14.401.

[181] Kim TH, Shivdasani RA. Stomach development, stem cells and disease. Development 2016;143:554–65. https://doi.org/10.1242/dev.124891.

[182] Gremel G, Wanders A, Cedernaes J, Fagerberg L, Hallström B, Edlund K, et al. The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling. J Gastroenterol 2015;50:46–57. https://doi.org/10.1007/s00535-014-0958-7.

[183] Hsu M, Safadi AO, Lui F. Physiology, Stomach. StatPearls, Treasure Island (FL): StatPearls Publishing; 2023.

[184] Karam SM, Leblond CP. Dynamics of epithelial cells in the corpus of the mouse stomach. III. Inward migration of neck cells followed by progressive transformation into zymogenic cells. Anat Rec 1993;236:297–313. https://doi.org/10.1002/ar.1092360204.

[185] Raufman J-P. Gastric chief cells: Receptors and signal-transduction mechanisms. Gastroenterology 1992;102:699–710. https://doi.org/10.1016/0016-5085(92)90124-H.

[186] Leushacke M, Tan SH, Wong A, Swathi Y, Hajamohideen A, Tan LT, et al. Lgr5-expressing chief cells drive epithelial regeneration and cancer in the oxyntic stomach. Nat Cell Biol 2017;19:774–86. https://doi.org/10.1038/ncb3541.

[187] Stange DE, Koo B-K, Huch M, Sibbel G, Basak O, Lyubimova A, et al. Differentiated Troy+ Chief Cells Act as Reserve Stem Cells to Generate All Lineages of the Stomach Epithelium. Cell 2013;155:357–68. https://doi.org/10.1016/j.cell.2013.09.008.

[188] Goldenring JR, Nam KT, Mills JC. The origin of pre-neoplastic metaplasia in the stomach: Chief cells emerge from the Mist. Exp Cell Res 2011;317:2759–64. https://doi.org/10.1016/j.yexcr.2011.08.017.

[189] Engevik AC, Kaji I, Goldenring JR. The Physiology of the Gastric Parietal Cell. Physiol Rev 2020;100:573–602. https://doi.org/10.1152/physrev.00016.2019.

[190] Fagoonee S, Singh SR, Altruda F, Pellicano R. Surface markers of gastric cancer stem cells. Minerva Biotecnol 2015;27:225–33.

[191] Tabula Sapiens C, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR, et al. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science 2022;376:eabl4896. https://doi.org/10.1126/science.abl4896.

[192] Busslinger GA, Weusten BLA, Bogte A, Begthel H, Brosens LAA, Clevers H. Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution. Cell Rep 2021;34:108819. https://doi.org/10.1016/j.celrep.2021.108819.

[193] Tsubosaka A, Komura D, Katoh H, Kakiuchi M, Onoyama T, Yamamoto A, et al. Single-Cell Transcriptome Analyses Reveal the Cell Diversity and Developmental Features of Human Gastric and Metaplastic Mucosa 2022:2022.05.22.493006. https://doi.org/10.1101/2022.05.22.493006.

[194] Zhang P, Yang M, Zhang Y, Xiao S, Lai X, Tan A, et al. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep 2019;27:1934-1947.e5. https://doi.org/10.1016/j.celrep.2019.04.052.

[195] Sathe A, Grimes SM, Lau BT, Chen J, Suarez C, Huang RJ, et al. Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment. Clin Cancer Res 2020;26:2640–53. https://doi.org/10.1158/1078-0432.CCR-19-3231.

[196] Wang R, Dang M, Harada K, Han G, Wang F, Pool Pizzi M, et al. Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma. Nat Med 2021;27:141–51. https://doi.org/10.1038/s41591-020-1125-8.

[197] Kim J, Park C, Kim KH, Kim EH, Kim H, Woo JK, et al. Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral

heterogeneity. NPJ Precis Oncol 2022;6:9. https://doi.org/10.1038/s41698-022-00251-1.

[198] Noah TK, Donahue B, Shroyer NF. Intestinal development and differentiation. Exp Cell Res 2011;317:2702–10. https://doi.org/10.1016/j.yexcr.2011.09.006.

[199] May CL, Kaestner KH. Gut endocrine cell development. Mol Cell Endocrinol 2010;323:70–5. https://doi.org/10.1016/j.mce.2009.12.009.

[200] Clevers H. The Intestinal Crypt, A Prototype Stem Cell Compartment. Cell 2013;154:274–84. https://doi.org/10.1016/j.cell.2013.07.004.

[201] Miron N, Cristea V. Enterocytes: active cells in tolerance to food and microbial antigens in the gut. Clin Exp Immunol 2012;167:405–12. https://doi.org/10.1111/j.1365-2249.2011.04523.x.

[202] Buffa R, Capella C, Fontana P, Usellini L, Solcia E. Types of endocrine cells in the human colon and rectum. Cell Tissue Res 1978;192:227–40. https://doi.org/10.1007/BF00220741.

[203] Umar S. Intestinal Stem Cells. Curr Gastroenterol Rep 2010;12:340–8. https://doi.org/10.1007/s11894-010-0130-3.

[204] Sanman LE, Chen IW, Bieber JM, Steri V, Trentesaux C, Hann B, et al. Transit-amplifying cells coordinate changes in intestinal epithelial cell type composition. Dev Cell 2021;56:356-365.e9. https://doi.org/10.1016/j.devcel.2020.12.020.

[205] Seguella L, Gulbransen BD. Enteric glial biology, intercellular signalling and roles in gastrointestinal disease. Nat Rev Gastroenterol Hepatol 2021;18:571–87. https://doi.org/10.1038/s41575-021-00423-7.

[206] Yu Y-B, Li Y-Q. Enteric glial cells and their role in the intestinal epithelial barrier. World J Gastroenterol WJG 2014;20:11273–80. https://doi.org/10.3748/wjg.v20.i32.11273.

[207] Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol 2019;16:713–32. https://doi.org/10.1038/s41575-019-0189-8.

[208] Han S-W, Kim H-P, Shin J-Y, Jeong E-G, Lee W-C, Lee K-H, et al. Targeted Sequencing of Cancer-Related Genes in Colorectal Cancer Using Next-Generation Sequencing. PLoS ONE 2013;8. https://doi.org/10.1371/journal.pone.0064271.

[209] Yaeger R, Chatila WK, Lipsyc MD, Hechtman JF, Cercek A, Sanchez-Vega F, et al. Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. Cancer Cell 2018;33:125-136.e3. https://doi.org/10.1016/j.ccell.2017.12.004.

[210] Pira G, Uva P, Scanu AM, Rocca PC, Murgia L, Uleri E, et al. Landscape of transcriptome variations uncovering known and novel driver events in colorectal carcinoma. Sci Rep 2020;10:1–12. https://doi.org/10.1038/s41598-019-57311-z.

[211] Wang Y, Song W, Wang J, Wang T, Xiong X, Qi Z, et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. J Exp Med 2020;217. https://doi.org/10.1084/jem.20191130.

[212] Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, et al. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. Cell 2019;178:714-730.e22. https://doi.org/10.1016/j.cell.2019.06.029.

[213] Lu J, Ye X, Fan F, Xia L, Bhattacharya R, Bellister S, et al. Endothelial Cells Promote the Colorectal Cancer Stem Cell Phenotype through a Soluble Form of Jagged-1. Cancer Cell 2013;23:171–85. https://doi.org/10.1016/j.ccr.2012.12.021.

[214] Hong BS, Cho J-H, Kim H, Choi E-J, Rho S, Kim J, et al. Colorectal cancer cell-derived microvesicles are enriched in cell cycle-related mRNAs that promote proliferation of endothelial cells. BMC Genomics 2009;10:556. https://doi.org/10.1186/1471-2164-10-556.

[215] Teranishi N, Naito Z, Ishiwata T, Tanaka N, Furukawa K, Seya T, et al. Identification of neovasculature using nestin in colorectal cancer. Int J Oncol 2007;30:593–603. https://doi.org/10.3892/ijo.30.3.593.

[216] Ishigami S-I, Arii S, Furutani M, Niwano M, Harada T, Mizumoto M, et al. Predictive value of vascular endothelial growth factor (VEGF) in metastasis and prognosis of human colorectal cancer. Br J Cancer 1998;78:1379–84. https://doi.org/10.1038/bjc.1998.688.

[217] Di J, Liu M, Fan Y, Gao P, Wang Z, Jiang B, et al. Phenotype molding of T cells in colorectal cancer by single-cell analysis. Int J Cancer 2020;146:2281–95. https://doi.org/10.1002/ijc.32856.

[218] Bailey C, Negus R, Morris A, Ziprin P, Goldin R, Allavena P, et al. Chemokine expression is associated with the accumulation of tumour associated macrophages (TAMs) and progression in human colorectal cancer. Clin Exp Metastasis 2007;24:121–30. https://doi.org/10.1007/s10585-007-9060-3.

[219] Kang J-C, Chen J-S, Lee C-H, Chang J-J, Shieh Y-S. Intratumoral macrophage counts correlate with tumor progression in colorectal cancer. J Surg Oncol 2010;102:242–8. https://doi.org/10.1002/jso.21617.

[220] Erreni M, Mantovani A, Allavena P. Tumor-associated Macrophages (TAM) and Inflammation in Colorectal Cancer. Cancer Microenviron 2011;4:141–54. https://doi.org/10.1007/s12307-010-0052-5.

[221] Cromer WE, Mathis JM, Granger DN, Chaitanya GV, Alexander JS. Role of the endothelium in inflammatory bowel diseases. World J Gastroenterol WJG 2011;17:578–93. https://doi.org/10.3748/wjg.v17.i5.578.

[222] Spadoni I, Zagato E, Bertocchi A, Paolinelli R, Hot E, Di Sabatino A, et al. A gut-vascular barrier controls the systemic dissemination of bacteria. Science 2015;350:830–4. https://doi.org/10.1126/science.aad0135.

[223] Consortium GTe. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348:648–60. https://doi.org/10.1126/science.1262110.

[224] Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res 2020;48:D682–8. https://doi.org/10.1093/nar/gkz966.

[225] Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013;45:580–5. https://doi.org/10.1038/ng.2653.

[226] Hassan MI, Toor A, Ahmad F. Progastriscin: structure, function, and its role in tumor progression. J Mol Cell Biol 2010;2:118–27. https://doi.org/10.1093/jmcb/mjq001.

[227] Karlsson M, Zhang C, Mear L, Zhong W, Digre A, Katona B, et al. A single-cell type transcriptomics map of human tissues. Sci Adv 2021;7. https://doi.org/10.1126/sciadv.abh2169.

[228] Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res 2019;47:D721–8. https://doi.org/10.1093/nar/gky900.

[229] Franzen O, Gan LM, Bjorkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database Oxf 2019;2019. https://doi.org/10.1093/database/baz046.

[230] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9:559. https://doi.org/10.1186/1471-2105-9-559.

[231] Ponten F, Jirstrom K, Uhlen M. The Human Protein Atlas - a tool for pathology. J Pathol 2008;216:387–93. https://doi.org/10.1002/path.2440.

[232] Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell 2021;184:3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

[233] Cooke BM, Usami S, Perry I, Nash GB. A simplified method for culture of endothelial cells and analysis of adhesion of blood cells under conditions of flow. Microvasc Res 1993;45:33–45. https://doi.org/10.1006/mvre.1993.1004.

[234] Van Peer G, Mestdagh P, Vandesompele J. Accurate RT-qPCR gene expression analysis on cell culture lysates. Sci Rep 2012;2:222. https://doi.org/10.1038/srep00222.

[235] Higueras MÁ, Jiménez-García L, Herranz S, Hortelano S, Luque A. Screening Assays to Characterize Novel Endothelial Regulators Involved in the Inflammatory Response. J Vis Exp JoVE 2017:55824. https://doi.org/10.3791/55824.

[236] McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8. https://doi.org/10.1038/nature07385.

[237] Peters LA, Perrigoue J, Mortha A, Iuga A, Song W-M, Neiman EM, et al. A functional genomics predictive network model identifies regulators of inflammatory bowel disease. Nat Genet 2017;49:1437–49. https://doi.org/10.1038/ng.3947.

[238] Sjölund K, Sandén G, Håkanson R, Sundler F. Endocrine Cells in Human Intestine: An Immunocytochemical Study. Gastroenterology 1983;85:1120–30. https://doi.org/10.1016/S0016-5085(83)80080-8.

[239] Engelstoft MS, Egerod KL, Lund ML, Schwartz TW. Enteroendocrine cell types revisited. Curr Opin Pharmacol 2013;13:912–21. https://doi.org/10.1016/j.coph.2013.09.018.

[240] Gribble FM, Reimann F. Enteroendocrine Cells: Chemosensors in the Intestinal Epithelium. Annu Rev Physiol 2016;78:277–99. https://doi.org/10.1146/annurev-physiol-021115-105439.

[241] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 2018;9:997. https://doi.org/10.1038/s41467-018-03405-7.

[242] Mou T, Deng W, Gu F, Pawitan Y, Vu TN. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Front Genet 2020;10.

[243] Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 2018;19:562–78. https://doi.org/10.1093/biostatistics/kxx053.

[244] Bala K, Ambwani K, Gohil NK. Effect of different mitogens and serum concentration on HUVEC morphology and characteristics: Implication on use of higher passage cells. Tissue Cell 2011;43:216–22. https://doi.org/10.1016/j.tice.2011.03.004.

[245] Rodriguez-Morata A, Garzon I, Alaminos M, Garcia-Honduvilla N, Sanchez-Quevedo MC, Bujan J, et al. Cell Viability and Prostacyclin Release in Cultured Human Umbilical Vein Endothelial Cells. Ann Vasc Surg 2008;22:440–8. https://doi.org/10.1016/j.avsg.2008.03.004.

[246] Hätinen O-PA, Lähteenvuo JE, Korpela HJ, Pajula JJ, Ylä-Herttuala S. Isolation of fresh endothelial cells from porcine heart for cardiovascular studies: a new fast protocol suitable for genomic, transcriptomic and cell biology studies. BMC Mol Cell Biol 2019;20:32. https://doi.org/10.1186/s12860-019-0215-2.

[247] Lidington E, Moyes D, McCormack A, Rose M. A comparison of primary endothelial cells and endothelial cell lines for studies of immune interactions. Transpl Immunol 1999;7:239–46. https://doi.org/10.1016/S0966-3274(99)80008-2.

[248] Unger RE, Krump-Konvalinkova V, Peters K, Kirkpatrick CJ. In Vitro Expression of the Endothelial Phenotype: Comparative Study of Primary Isolated Cells and Cell Lines, Including the Novel Cell Line HPMEC-ST1.6R. Microvasc Res 2002;64:384–97. https://doi.org/10.1006/mvre.2002.2434.

[249] Khan A, Waqar K, Shafique A, Irfan R, Gul A. Chapter 18 - In Vitro and In Vivo Animal Models: The Engineering Towards Understanding Human Diseases and Therapeutic Interventions. In: Barh D, Azevedo V, editors. Omics Technol. Bio-Eng., Academic Press; 2018, p. 431–48. https://doi.org/10.1016/B978-0-12-804659-3.00018-X.

[250] Todaro GJ, Lazar GK, Green H. The initiation of cell division in a contact-inhibited mammalian cell line. J Cell Comp Physiol 1965;66:325–33. https://doi.org/10.1002/jcp.1030660310.

[251] Haudenschild CC, Schwartz SM. Endothelial regeneration. II. Restitution of endothelial continuity. Lab Investig J Tech Methods Pathol 1979;41:407–18.

[252] Rodriguez LG, Wu X, Guan J-L. Wound-Healing Assay. In: Guan J-L, editor. Cell Migr. Dev. Methods Protoc., Totowa, NJ: Humana Press; 2005, p. 23–9. https://doi.org/10.1385/1-59259-860-9:023.

[253] Reinhart-King CA. Chapter 3 Endothelial Cell Adhesion and Migration. Methods Enzymol., vol. 443, Academic Press; 2008, p. 45–64. https://doi.org/10.1016/S0076-6879(08)02003-X.

[254] Nie F-Q, Yamada M, Kobayashi J, Yamato M, Kikuchi A, Okano T. On-chip cell migration assay using microfluidic channels. Biomaterials 2007;28:4017–22. https://doi.org/10.1016/j.biomaterials.2007.05.037.

[255] van der Meer AD, Vermeul K, Poot AA, Feijen J, Vermes I. A microfluidic wound-healing assay for quantifying endothelial cell migration. Am J Physiol-Heart Circ Physiol 2010;298:H719–25. https://doi.org/10.1152/ajpheart.00933.2009.

[256] Kalluri R. Basement membranes: structure, assembly and role in tumour angiogenesis. Nat Rev Cancer 2003;3:422–33. https://doi.org/10.1038/nrc1094.

[257] Kleinman HK, Martin GR. Matrigel: Basement membrane matrix with biological activity. Semin Cancer Biol 2005;15:378–86. https://doi.org/10.1016/j.semcancer.2005.05.004.

[258] Arnaoutova I, George J, Kleinman HK, Benton G. The endothelial cell tube formation assay on basement membrane turns 20: state of the science and the art. Angiogenesis 2009;12:267–74. https://doi.org/10.1007/s10456-009-9146-4.

[259] Arnaoutova I, Kleinman HK. In vitro angiogenesis: endothelial cell tube formation on gelled basement membrane extract. Nat Protoc 2010;5:628–35. https://doi.org/10.1038/nprot.2010.6.

[260] Nehls V, Drenckhahn D. A Novel, Microcarrier-Based in Vitro Assay for Rapid and Reliable Quantification of Three-Dimensional Cell Migration and Angiogenesis. Microvasc Res 1995;50:311–22. https://doi.org/10.1006/mvre.1995.1061.

[261] Crabtree B, Subramanian V. Behavior of endothelial cells on Matrigel and development of a method for a rapid and reproducible in vitro angiogenesis assay. Vitro Cell Dev Biol - Anim 2007;43:87–94. https://doi.org/10.1007/s11626-007-9012-x.

[262] Song JW, Munn LL. Fluid forces control endothelial sprouting. Proc Natl Acad Sci 2011;108:15342–7. https://doi.org/10.1073/pnas.1105316108.

[263] Chung S, Sudo R, Mack PJ, Wan C-R, Vickerman V, Kamm RD. Cell migration into scaffolds under co-culture conditions in a microfluidic platform. Lab Chip 2009;9:269–75. https://doi.org/10.1039/B807585A.

[264] Bischel LL, Young EWK, Mader BR, Beebe DJ. Tubeless microfluidic angiogenesis assay with three-dimensional endothelial-lined microvessels. Biomaterials 2013;34:1471–7. https://doi.org/10.1016/j.biomaterials.2012.11.005.

[265] Wang WY, Lin D, Jarman EH, Polacheck WJ, Baker BM. Functional angiogenesis requires microenvironmental cues balancing endothelial cell migration and proliferation. Lab Chip 2020;20:1153–66. https://doi.org/10.1039/c9lc01170f.

[266] Vajen T, Heinzmann ACA, Dickhout A, Zhao Z, Nagy M, Heemskerk JWM, et al. Laminar Flow-based Assays to Investigate Leukocyte Recruitment on Cultured Vascular Cells and Adherent Platelets. J Vis Exp JoVE 2018:57009. https://doi.org/10.3791/57009.

[267] Ganguly A, Zhang H, Sharma R, Parsons S, Patel KD. Isolation of Human Umbilical Vein Endothelial Cells and Their Use in the Study of Neutrophil Transmigration Under Flow Conditions. J Vis Exp JoVE 2012:4032. https://doi.org/10.3791/4032.

[268] Muller WA, Weigl SA. Monocyte-selective transendothelial migration: dissection of the binding and transmigration phases by an in vitro assay. J Exp Med 1992;176:819–28. https://doi.org/10.1084/jem.176.3.819.

[269] Furie MB, Naprstek BL, Silverstein SC. Migration of neutrophils across monolayers of cultured microvascular endothelial cells. An in vitro model of leucocyte extravasation. J Cell Sci 1987;88:161–75. https://doi.org/10.1242/jcs.88.2.161.

[270] Finger EB, Purl KD, Alon R, Lawrence MB, von Andrian UH, Springer TA. Adhesion through L-selectin requires a threshold hydrodynamic shear. Nature 1996;379:266–9. https://doi.org/10.1038/379266a0.

[271] Lawrence MB, Kansas GS, Kunkel EJ, Ley K. Threshold Levels of Fluid Shear Promote Leukocyte Adhesion through Selectins (CD62L,P,E). J Cell Biol 1997;136:717–27. https://doi.org/10.1083/jcb.136.3.717.

[272] Lamberti G, Prabhakarpandian B, Garson C, Smith A, Pant K, Wang B, et al. Bioinspired Microfluidic Assay for In Vitro Modeling of Leukocyte–Endothelium Interactions. Anal Chem 2014;86:8344–51. https://doi.org/10.1021/ac5018716.

[273] Billoir P, Miranda S, Damian L, Richard V, Benhamou Y, Le Cam Duchez V. Development of a thrombin generation test in cultured endothelial cells: Evaluation of the prothrombotic effects of antiphospholipid antibodies. Thromb Res 2018;169:87–92. https://doi.org/10.1016/j.thromres.2018.07.021.

[274] Billoir P, Blandinières A, Gendron N, Chocron R, Gunther S, Philippe A, et al. Endothelial Colony-Forming Cells from Idiopathic Pulmonary Fibrosis Patients Have a High Procoagulant Potential. Stem Cell Rev Rep 2021;17:694–9. https://doi.org/10.1007/s12015-020-10043-4.

[275] Diamond SL. Systems Analysis of Thrombus Formation. Circ Res 2016;118:1348–62. https://doi.org/10.1161/CIRCRESAHA.115.306824.

[276] Zipperle J, Schlimp CJ, Holnthoner W, Husa A-M, Nürnberger S, Redl H, et al. A novel coagulation assay incorporating adherent endothelial cells in thromboelastometry. Thromb Haemost 2013;109:869–77. https://doi.org/10.1160/TH12-10-0767.

[277] Li P-F, Chen S-C, Xia T, Jiang X-M, Shao Y-F, Xiao B-X, et al. Non-coding RNAs and gastric cancer. World J Gastroenterol WJG 2014;20:5411–9. https://doi.org/10.3748/wjg.v20.i18.5411.

[278] Gao Y, Wang JW, Ren JY, Guo M, Guo CW, Ning SW, et al. Long noncoding RNAs in gastric cancer: From molecular dissection to clinical application. World J Gastroenterol 2020;26:3401–12. https://doi.org/10.3748/wjg.v26.i24.3401.

[279] Ghafouri-Fard S, Taheri M. Long non-coding RNA signature in gastric cancer. Exp Mol Pathol 2020;113:104365. https://doi.org/10.1016/j.yexmp.2019.104365.

[280] Xia T, Chen S, Jiang Z, Shao Y, Jiang X, Li P, et al. Long noncoding RNA FER1L4 suppresses cancer cell growth by acting as a competing endogenous RNA and regulating PTEN expression. Sci Rep 2015;5:13445. https://doi.org/10.1038/srep13445.

[281] Yang X-Z, Cheng T-T, He Q-J, Lei Z-Y, Chi J, Tang Z, et al. LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/β-catenin pathway. Mol Cancer 2018;17:126. https://doi.org/10.1186/s12943-018-0874-1.

[282] Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. Nat Rev Genet 2013;14:880–93. https://doi.org/10.1038/nrg3594.

[283] Datz FL, Christian PE, Moore J. Gender-related differences in gastric emptying. J Nucl Med Off Publ Soc Nucl Med 1987;28:1204–7.

[284] Al-Shboul O. The role of the RhoA/ROCK pathway in gender-dependent differences in gastric smooth muscle contraction. J Physiol Sci 2016;66:85–92. https://doi.org/10.1007/s12576-015-0400-9.

[285] Lou L, Wang L, Zhang Y, Chen G, Lin L, Jin X, et al. Sex difference in incidence of gastric cancer: an international comparative study based on the Global Burden of Disease Study 2017. BMJ Open 2020;10:e033323. https://doi.org/10.1136/bmjopen-2019-033323.

[286] Li H, Wei Z, Wang C, Chen W, He Y, Zhang C. Gender Differences in Gastric Cancer Survival: 99,922 Cases Based on the SEER Database. J Gastrointest Surg 2020;24:1747–57. https://doi.org/10.1007/s11605-019-04304-y.

[287] Pink RC, Wicks K, Caley DP, Punch EK, Jacobs L, Carter DRF. Pseudogenes: Pseudo-functional or key regulators in health and disease? RNA 2011;17:792–8. https://doi.org/10.1261/rna.2658311.

[288] Kovalenko TF, Patrushev LI. Pseudogenes as Functionally Significant Elements of the Genome. Biochem Mosc 2018;83:1332–49. https://doi.org/10.1134/S0006297918110044.

[289] Cheetham SW, Faulkner GJ, Dinger ME. Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat Rev Genet 2020;21:191–201. https://doi.org/10.1038/s41576-019-0196-1.

[290] Boggetti B, Jasik J, Takamiya M, Strähle U, Reugels AM, Campos-Ortega JA. NBP, a zebrafish homolog of human Kank3, is a novel Numb interactor essential for epidermal integrity and neurulation. Dev Biol 2012;365:164–74. https://doi.org/10.1016/j.ydbio.2012.02.021.

[291] Guo SS, Seiwert A, Szeto IYY, Fässler R. Tissue distribution and subcellular localization of the family of Kidney Ankyrin Repeat Domain (KANK) proteins. Exp Cell Res 2021;398:112391. https://doi.org/10.1016/j.yexcr.2020.112391.

[292] Kakinuma N, Zhu Y, Wang Y, Roy BC, Kiyama R. Kank proteins: structure, functions and diseases. Cell Mol Life Sci 2009;66:2651–9. https://doi.org/10.1007/s00018-009-0038-y.

[293] THE GTEX CONSORTIUM, Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 2015;348:648–60. https://doi.org/10.1126/science.1262110.

[294] Zhu Y, Kakinuma N, Wang Y, Kiyama R. Kank proteins: A new family of ankyrin-repeat domain-containing proteins. Biochim Biophys Acta BBA - Gen Subj 2008;1780:128–33. https://doi.org/10.1016/j.bbagen.2007.09.017.

[295] Bouchet BP, Gough RE, Ammon Y-C, van de Willige D, Post H, Jacquemet G, et al. Talin-KANK1 interaction controls the recruitment of cortical microtubule stabilizing complexes to focal adhesions. eLife 2016;5:e18124. https://doi.org/10.7554/eLife.18124.

[296] Sun Z, Tseng H-Y, Tan S, Senger F, Kurzawa L, Dedden D, et al. Kank2 activates talin, reduces force transduction across integrins and induces central adhesion formation. Nat Cell Biol 2016;18:941–53. https://doi.org/10.1038/ncb3402.

[297] Guo Q, Liao S, Zhu Z, Li Y, Li F, Xu C. Structural basis for the recognition of kinesin family member 21A (KIF21A) by the ankyrin domains of KANK1 and KANK2 proteins. J Biol Chem 2018;293:557–66. https://doi.org/10.1074/jbc.M117.817494.

[298] Sarkar S, Roy BC, Hatano N, Aoyagi T, Gohji K, Kiyama R. A Novel Ankyrin Repeat-containing Gene (Kank) Located at 9p24 Is a Growth Suppressor of Renal Cell Carcinoma *. J Biol Chem 2002;277:36585–91. https://doi.org/10.1074/jbc.M204244200.

[299] Dai Z, Xie B, Yang B, Chen X, Hu C, Chen Q. KANK3 mediates the p38 MAPK pathway to regulate the proliferation and invasion of lung adenocarcinoma cells. Tissue Cell 2023;80:101974. https://doi.org/10.1016/j.tice.2022.101974.

[300] Kim D-H, Wirtz D. Predicting how cells spread and migrate. Cell Adhes Migr 2013;7:293–6. https://doi.org/10.4161/cam.24804.

[301] Zhao X, Guan J-L. Focal adhesion kinase and its signaling pathways in cell migration and angiogenesis. Adv Drug Deliv Rev 2011;63:610–5. https://doi.org/10.1016/j.addr.2010.11.001.

[302] Sieg DJ, Hauck CR, Schlaepfer DD. Required role of focal adhesion kinase (FAK) for integrin-stimulated cell migration. J Cell Sci 1999;112:2677–91. https://doi.org/10.1242/jcs.112.16.2677.

[303] Cox BD, Natarajan M, Stettner MR, Gladson CL. New concepts regarding focal adhesion kinase promotion of cell migration and proliferation. J Cell Biochem 2006;99:35–52. https://doi.org/10.1002/jcb.20956.

[304] Pogoda K, Byfield F, Deptuła P, Cieśluk M, Suprewicz Ł, Skłodowski K, et al. Unique Role of Vimentin Networks in Compression Stiffening of Cells and Protection of Nuclei from Compressive Stress. Nano Lett 2022;22:4725–32. https://doi.org/10.1021/acs.nanolett.2c00736.

[305] Satelli A, Li S. Vimentin in cancer and its potential as a molecular target for cancer therapy. Cell Mol Life Sci CMLS 2011;68:3033–46. https://doi.org/10.1007/s00018-011-0735-1.

[306] Ridge KM, Eriksson JE, Pekny M, Goldman RD. Roles of vimentin in health and disease. Genes Dev 2022;36:391–407. https://doi.org/10.1101/gad.349358.122.

[307] Boraas LC, Ahsan T. Lack of vimentin impairs endothelial differentiation of embryonic stem cells. Sci Rep 2016;6:30814. https://doi.org/10.1038/srep30814.

[308] Dave JM, Bayless KJ. Vimentin as an integral regulator of cell adhesion and endothelial sprouting. Microcirc N Y N 1994 2014;21:333–44. https://doi.org/10.1111/micc.12111.

Paper I

A human stomach cell type transcriptome atlas

S Öling[1], E Struck[1], M Noreen-Thorsen[1], M Zwahlen[2], K von Feilitzen[2], J Odeberg[1,2,3,4], F Pontén[5], C Lindskog[5], M Uhlén[2], P Dusart[2,6], LM Butler [1,2,6*]

[1] Translational Vascular Research, Department of Clinical Medicine, The Arctic University of Norway, 9019 Tromsø, Norway

[2] Science for Life Laboratory, Department of Protein Science, Royal Institute of Technology (KTH), 171 21 Stockholm, Sweden

[3] The University Hospital of North Norway (UNN), 9019 Tromsø, Norway

[4] Coagulation Unit, Department of Haematology, Karolinska University Hospital, 171 76 Stockholm, Sweden

[5] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, 752 37 Uppsala, Sweden

[6] Clinical Chemistry and Blood Coagulation Research, Department of Molecular Medicine and Surgery, Karolinska Institute, 171 76 Stockholm, Sweden, *and* Clinical Chemistry, Karolinska University Laboratory, Karolinska University Hospital, 171 76 Stockholm, Sweden

[*] **Lead contact/corresponding author**:

Dr. L.M Butler, PhD

Email: Lynn.butler@ki.se or lynn.m.butler@uit.no

Key words: Cell profiling, gene enrichment, bulk RNAseq, stomach

SUMMARY

The identification of cell type-specific genes and their modification under different conditions is central to our understanding of human health and disease. The stomach, a hollow organ in the upper gastrointestinal tract, provides an acidic environment that contributes to microbial defence and facilitates the activity of secreted digestive enzymes to process food and nutrients into chyme. In contrast to other sections of the gastrointestinal tract, detailed descriptions of cell type gene enrichment profiles in the stomach are absent from the major single cell sequencing-based atlases. Here, we use an integrative correlation analysis method to predict human stomach cell type transcriptome signatures using unfractionated stomach RNAseq data from 359 individuals. We profile parietal, chief, gastric mucous, gastric enteroendocrine, mitotic, endothelial, fibroblast, macrophage, neutrophil, T-cell and plasma cells, identifying over 1600 cell type-enriched genes. We uncover the cell type expression profile of several non-coding genes strongly associated with the progression of gastric cancer and, using a sex-based subset analysis, uncover a panel of male-only chief cell-enriched genes. This study provides a roadmap to further understand human stomach biology.

INTRODUCTION

The gastrointestinal (GI) tract is a multiple organ system which can be divided into upper and lower parts, the physical properties and cellular characteristics of which reflect their different roles in digestion, absorption of nutrients, and excretion of waste products (de Santa Barbara, van den Brink and Roberts, 2003; Choi *et al.*, 2014; Thompson, DeLaForest and Battle, 2018). The stomach, a hollow muscular organ in the upper GI tract, produces an array of acids and gastric enzymes, acting as a reservoir for the mechanical and chemical digestion of ingested food (Kim and Shivdasani, 2016). The constituent cell types of the stomach include parietal cells, chief cells, gastric mucous cells, gastric enteroendocrine cells, mitotic cells, endothelial cells, fibroblasts, and various immune cells (Gremel *et al.*, 2015; Uhlen *et al.*, 2015). In contrast to lower sections of the GI tract, descriptions of the cellular transcriptional landscape in the stomach are lacking, with this organ absent from large scale single cell sequencing (scRNAseq) initiatives, such as Tabula Sapiens (Tabula Sapiens *et al.*, 2022) and the Human Cell Atlas (Regev *et al.*, 2017). Where scRNAseq has been used to profile gene expression in the adult stomach, studies have typically focused on specific cell types, such as the epithelia (Busslinger *et al.*, 2021; Tsubosaka *et al.*, 2022), or in pathological states such as gastric cancer (P. Zhang *et al.*, 2019; Sathe *et al.*, 2020; Wang *et al.*, 2021; Kim *et al.*, 2022). Whilst scRNAseq studies provide high resolution of individual cell (sub)type gene expression profiles, challenges remain, including artefactual modification of gene expression due to cell removal and processing (O'Flanagan *et al.*, 2019; Denisenko *et al.*, 2020; Massoni-Badosa *et al.*, 2020), compromised read depth, and difficulties with data interpretation (Gawad, Koh and Quake, 2016; Jiang *et al.*, 2022). As a limited number of biological replicates are typically analysed, underestimation of biological variance can increase the likelihood of potential false discoveries (Squair *et al.*, 2021; Denninger *et al.*, 2022).

Non-coding RNA is emerging as a novel, important class of molecules, involved in the maintenance of healthy stomach tissue, and the development and progression of gastric cancer (Gao *et al.*, 2020; Razavi and Katanforosh, 2022), but to date there is no overall description of stomach cell type enriched non-coding RNAs.

Here, we analysed 359 bulk RNAseq human stomach samples to identify over 1600 genes with cell type-enriched expression, using our previously developed integrative correlation analysis (Butler *et al.*, 2016; Dusart *et al.*, 2019; Norreen-Thorsen *et al.*, 2022). Gastric mucous cells had the highest number of predicted protein-coding and non-coding enriched genes and represented the primary site of expression of genes that were tissue enriched in stomach over other tissue types. Gastric enteroendocrine cells expressed a panel of non-coding genes that are also selectively expressed in pancreatic and intestinal endocrine cells, indicting a common function in these cell types. Several of the identified cell type enriched non-coding genes have previously been associated with the progression of gastric cancer, but until now the cell type site of expression had not been described. Sex subset analysis revealed a high global similarity in cell type transcriptomes between males and females, but a panel of chief cell enriched Y-linked genes were identified. Data is available through the Human Protein Atlas (HPA) portal (www.proteinatlas.org/humanproteome/tissue+cell+type/stomach).

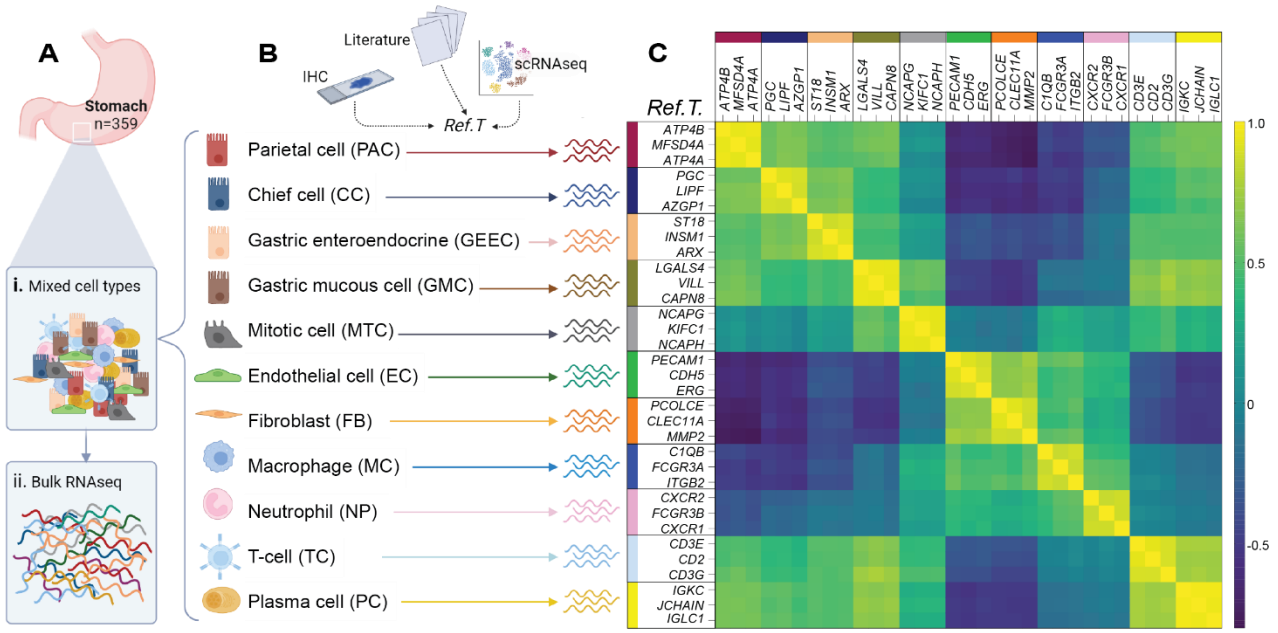
RESULTS

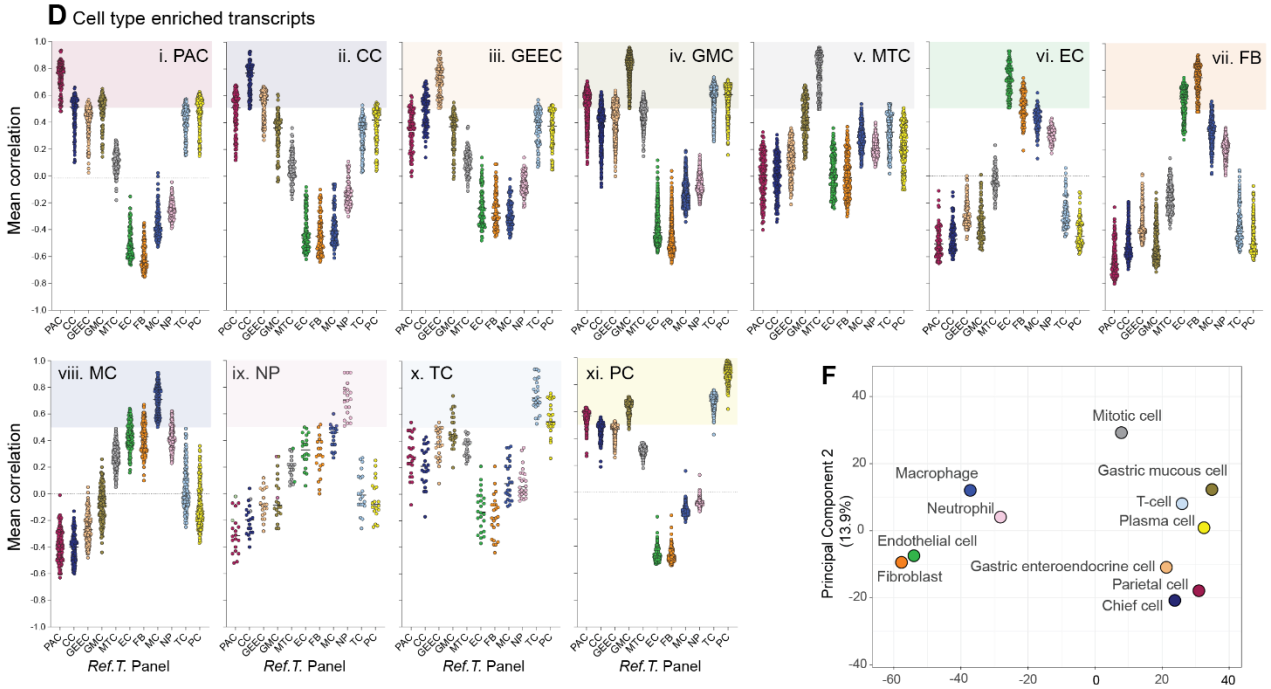
Identification of cell type transcriptome profiles in stomach

Cell type reference transcripts correlate across unfractionated RNAseq data

To identify stomach cell type-enriched transcriptome profiles, we conducted an analysis based on our previously developed method (Butler *et al.*, 2016; Dusart *et al.*, 2019; Norreen-Thorsen *et al.*, 2022), using human stomach bulk RNAseq data (N=359) from Genotype-Tissue Expression (GTEx) portal V8 (https://gtexportal.org) (Consortium, 2015) (see Figure S1 for method overview). Each sample was unfractionated and thus contained a mix of cell types (Figure 1 A.i), which contribute differing proportions of transcripts subsequently measured by RNAseq (Figure 1 A.ii) (Figure S1 A). For each major constituent stomach cell type, candidate cell type specific genes (termed ´reference transcripts´ [*Ref.T.*]) were selected based on: (**i**) our in-house proteomic profiling of stomach tissue (Gremel *et al.*, 2015; Uhlen *et al.*, 2015), (**ii**) older ´none-omics´ studies (Hassan, Toor and Ahmad, 2010), (**iii**) scRNAseq data were available (Busslinger *et al.*, 2021; Karlsson *et al.*, 2021) or (**iv**) databases collated from multiple sources, e.g. Cell Marker (X. Zhang *et al.*, 2019) and PanglaoDB (Franzen, Gan and Bjorkegren, 2019) (Figure 1 B and Figure S1 B). Three markers were selected for each cell type, based on the following criteria**:** (**i**) A high corr. (>0.85) between *Ref.T.* within each cell type panel (Figure 1 C and Table S1, Tab 1), indicating *cell type co-expression*: parietal cells (PAC) [*ATP4B*, *MFSD4A*, *ATP4A* mean corr. ± STD 0.94±0.013], chief cells (CC) [*PGC*, *LIPF*, *AZGP1*, 0.89±0.013], gastric enteroendocrine cells (GEEC) [*ST18*, *INSM1*, *ARX*, 0.89±0.021], gastric mucous cells (GMC) [*LGALS4*, *VILL*, *CAPN8*, 0.94±0.008], mitotic cells (MTC) [*NCAPG*, *KIFC1*, *NCAPH*, 0.93±0.009], endothelial cells (EC) [*PECAM1*, *CDH5*, *ERG*, 0.89±0.013], fibroblasts (FB) [*PCOLCE*, *CLEC11A*, *MMP2*, 0.87±0.027], macrophages (MC) [*C1QB*, *FCGR3A*, *ITGB2*, 0.86±0.015], neutrophils (NP) [*CXCR2*, *FCGR3B*, *CXCR1*, 0.86±0.009], T-cells (TC) [*CD3E*, *CD2*, *CD3G*, 0.9±0.019] and plasma cells (PC) [*IGKC*, *JCHAIN*, *IGLC1*, 0.97±0.009]. (**ii**) A low corr. between *Ref.T.* across the different cell type panels (Figure 1 C) (Table S1, Tab 1), indicating *cell type specificity* (mean inter-panel corr. ±

STD 0.08 ± 0.14) and (**iii**) a normal distribution of *Ref.T*. expression across the samples (Figure S2 A).





E Gene ontology terms for predicted cell-type enriched genes

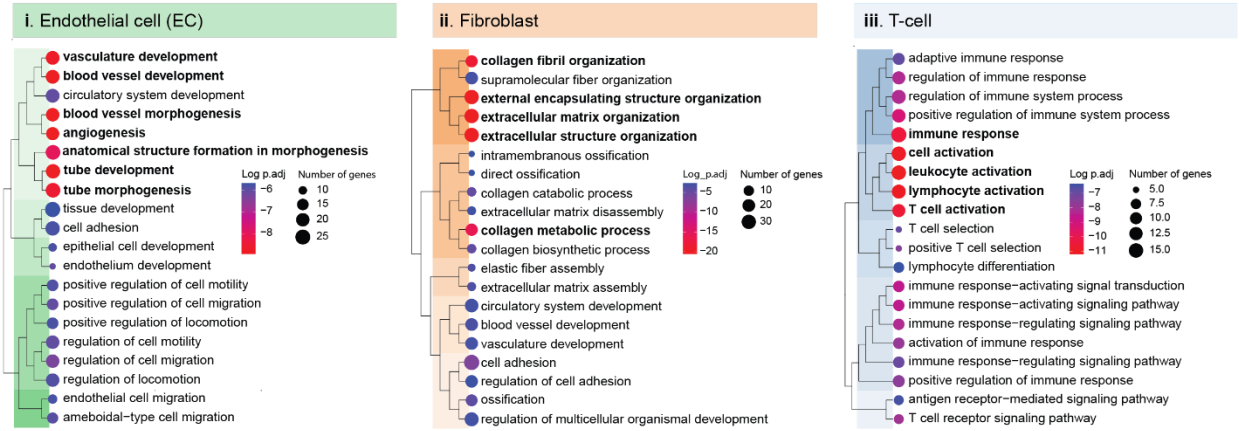


Figure 1. Integrative co-expression analysis can resolve constituent cell type identities from unfractionated human stomach tissue RNAseq data. (**A**) RNAseq data for 359 unfractionated human stomach samples were retrieved from GTEx V8. Each sample contained (i) mixed cell types, which contributed (ii) differing proportions of sequenced mRNA. (**B**) To profile cell type-enriched transcriptomes, constituent cell types were identified and candidate marker genes (´reference transcripts´ [*Ref.T.*]) for virtual tagging of each were selected, based on in house tissue protein profiling and/or existing literature and datasets. (**C**) Matrix of correlation coefficients between selected *Ref.T.* across the sample set. (**D**) Mean correlation coefficients of genes above designated thresholds for classification as cell-type enriched in stomach: (i) parietal cells [PC], (ii) chief cells [CC], (iii) gastric enteroendocrine cells [GEEC], (iv) gastric mucous cells [GMC], (v) mitotic cells [MTC], (vi) endothelial cells [EC], (vii) fibroblasts [FB], (viii) macrophages [MC], (ix) neutrophils [NP], (x) T-cells [TC], (xi) plasma cells [PC] with all *Ref.T.* panels. (**E**) Over-represented gene ontology terms among genes predicted to be: (i) endothelial cell, (ii) fibroblast or (iii) T-cell enriched. (**F**) Principal component analysis of correlation profiles of cell type enriched genes. See also Table S1 Tab 1 and 2 and Figure S1 for method overview

Using reference transcript analysis to identify cell type-enriched genes

Correlation coefficients (corr.) between each selected *Ref.T.* and all other sequenced transcripts (>56,000) were calculated across stomach RNAseq samples (Figure S1 C). The proportion of cell types represented in each sample varies, due to biological and sampling variability, but ratios should remain consistent between constitutively expressed cell-enriched genes. Thus, a high corr. of a given transcript with all *Ref.T.* in only one cell type panel is consistent with enrichment in the corresponding cell type. For each cell type, a list of enriched genes was generated (Figure 1 D.i-xi), with inclusion based on: (**i**) the gene having a mean corr. >0.50 with the *Ref.T.* panel representing the cell type (Figure S1 C.ii), and (**ii**) a *differential correlation* between this value and the maximum mean corr. with any other *Ref.T.* panel >0.15 (Figure S1 D-E). This excluded genes that were potentially co-enriched in two or more cell types, as we previously described (Norreen-Thorsen *et al.*, 2022) (all data in Table S1, Tab 2). For certain cell types, enriched genes were less well separated by corr. value that others, e.g., those most highly correlating with the fibroblast *Ref.T.* panel (Figure 1 D.vii) tended to show elevated corr. with the *Ref.T.* panel for endothelial cells, and *vice versa* (Figure 1 D.vi). However, all cell type enriched genes were well separated when the individual gene differential correlations vs. other *Ref.T.* panels were plotted (Figure S2 B) and gene ontology (GO) and reactome analysis (Ashburner *et al.*, 2000; Gene Ontology, 2021) revealed over represented terms for these cell types were consistent with known functions e.g., for endothelial cells most significantly enriched terms included ´*vascular development*´ and ´*angiogenesis*´ (Figure 1 E.i), for fibroblasts ´*extracellular matrix organisation*´ and ´*collagen fibril organization*´ (Figure 1 E.ii) and for T-cells ´*T-cell activation*´ and ´*immune response*´ (Figure 1 E.iii) (Table S1, Tab 8, 9 and 12). Principal component analysis of the corr. values of cell type-enriched genes (generated using (https://biit.cs.ut.ee/clustvis/) (Metsalu and Vilo, 2015) revealed the largest variance was between stomach specific cell types vs. stromal/vasculature related ones (Figure 1 F).

Stomach cell type enriched gene signatures

The majority of stomach cell type enriched genes are protein coding

1694 genes were predicted to be cell type-enriched (Figure 2 A and Table S1, Tab 2). Gastric mucous cells, plasma cells and fibroblasts had the highest number of predicted enriched genes (n=517, 214 and 186 respectively) (Figure 2 A.i, ii and iii). Of the other cell types found in all, or most, tissue types, mitotic cells and macrophages had the most enriched genes (n=171 and 158, respectively) (Figure 1 A.iv-v). Other stomach specialised cell types, parietal cells, chief cells and gastric enteroendocrine cells, had significantly fewer enriched genes (n=123, 103 and 86, respectively) (Figure 2 A.vi, vii and ix), and T-cells and neutrophils had the fewest overall (n=24 and 20, respectively) (Figure 2 A.x and xi). In all cases, the majority of cell type enriched genes were classified as protein coding (Yates *et al.*, 2020), with the exception of plasma cells, in which immunoglobulin (IG) gene was the most common classification (Figure 2 A.ii). lncRNA were the most common type of non-coding cell type enriched transcript, with the exception of plasma cells, where immunoglobulin (IG) pseudogene was the most common non-coding classification (Figure 2 A.ii).

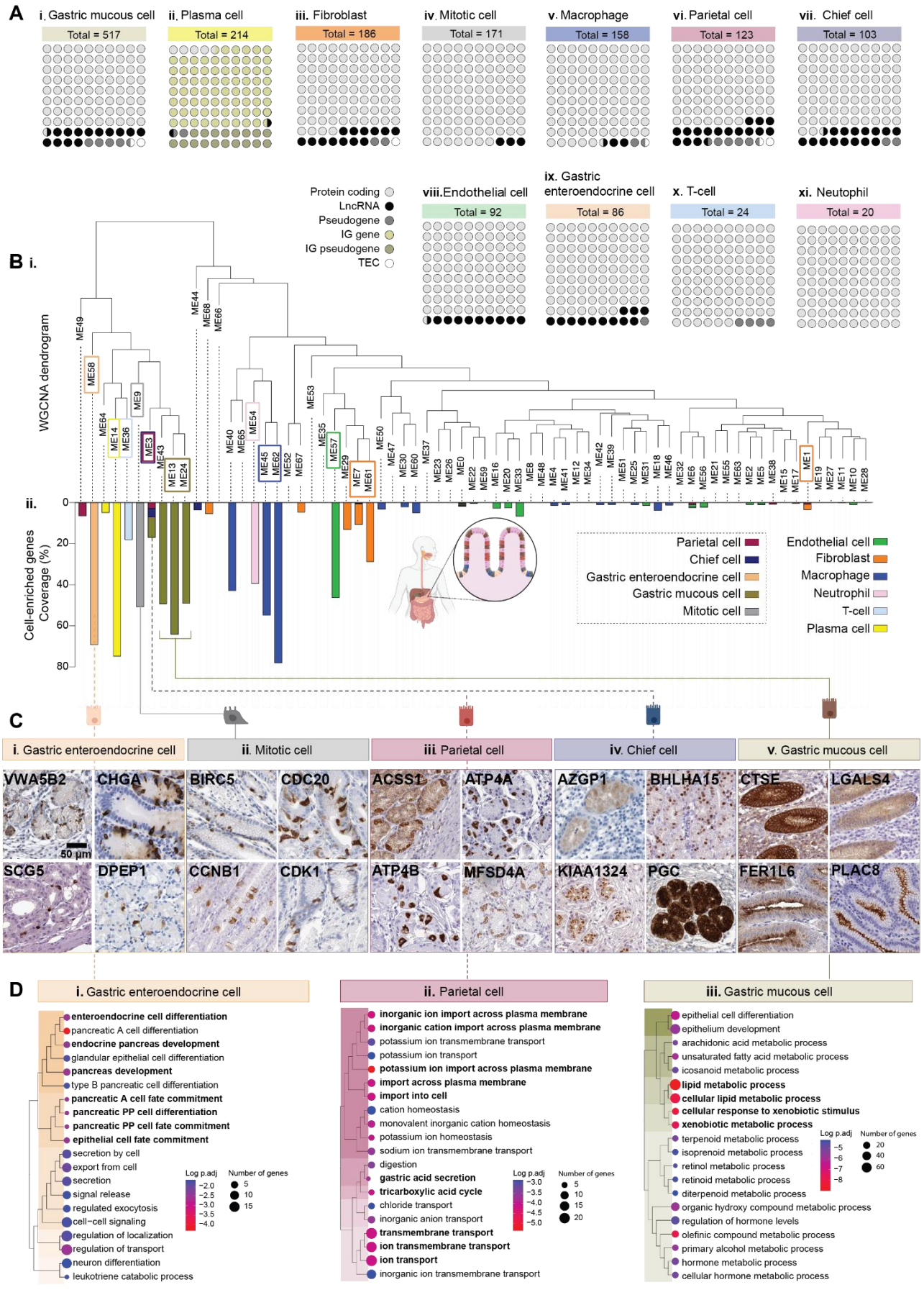

A
i. Gastric mucous cell
Total = 517
ii. Plasma cell
Total = 214
iii. Fibroblast
Total = 186
iv. Mitotic cell
Total = 171
v. Macrophage
Total = 158
vi. Parietal cell
Total = 123
vii. Chief cell
Total = 103
viii. Endothelial cell
Total = 92
ix. Gastric enteroendocrine cell
Total = 86
x. T-cell
Total = 24
xi. Neutophil
Total = 20
Protein coding
LncRNA
Pseudogene
IG gene
IG pseudogene
TEC
B i.
WGCNA dendrogram
ii.
Cell-enriched genes Coverage (%)
Parietal cell
Chief cell
Gastric enteroendocrine cell
Gastric mucous cell
Mitotic cell
Endothelial cell
Fibroblast
Macrophage
Neutrophil
T-cell
Plasma cell
C
i. Gastric enteroendocrine cell
ii. Mitotic cell
iii. Parietal cell
iv. Chief cell
v. Gastric mucous cell
VWA5B2
CHGA
BIRC5
CDC20
ACSS1
ATP4A
AZGP1
BHLHA15
CTSE
LGALS4
50 µm
SCG5
DPEP1
CCNB1
CDK1
ATP4B
MFSD4A
KIAA1324
PGC
FER1L6
PLAC8
D
i. Gastric enteroendocrine cell
enteroendocrine cell differentiation
pancreatic A cell differentiation
endocrine pancreas development
glandular epithelial cell differentiation
pancreas development
type B pancreatic cell differentiation
pancreatic A cell fate commitment
pancreatic PP cell differentiation
pancreatic PP cell fate commitment
epithelial cell fate commitment
secretion by cell
export from cell
secretion
signal release
regulated exocytosis
cell−cell signaling
regulation of localization
regulation of transport
neuron differentiation
leukotriene catabolic process
Log p.adj
Number of genes
ii. Parietal cell
inorganic ion import across plasma membrane
inorganic cation import across plasma membrane
potassium ion transmembrane transport
potassium ion transport
potassium ion import across plasma membrane
import across plasma membrane
import into cell
cation homeostasis
monovalent inorganic cation homeostasis
potassium ion homeostasis
sodium ion transmembrane transport
digestion
gastric acid secretion
tricarboxylic acid cycle
chloride transport
inorganic anion transport
transmembrane transport
ion transmembrane transport
ion transport
inorganic ion transmembrane transport
Log p.adj
Number of genes
iii. Gastric mucous cell
epithelial cell differentiation
epithelium development
arachidonic acid metabolic process
unsaturated fatty acid metabolic process
icosanoid metabolic process
lipid metabolic process
cellular lipid metabolic process
cellular response to xenobiotic stimulus
xenobiotic metabolic process
terpenoid metabolic process
isoprenoid metabolic process
retinol metabolic process
retinoid metabolic process
diterpenoid metabolic process
organic hydroxy compound metabolic process
regulation of hormone levels
olefinic compound metabolic process
primary alcohol metabolic process
hormone metabolic process
cellular hormone metabolic process
Log p.adj
Number of genes

Figure 2. Integrative co-expression analysis of unfractionated RNAseq reveals enriched genes in human stomach cell types. (**A**) Total number and proportional representation of class for cell type enriched genes in: (i) gastric mucous cells, (ii) plasma cells, (iii) fibroblasts, (iv) mitotic cells, (v) macrophages, (vi) parietal cells, (vii) chief cells, (viii) endothelial cells, (ix) gastric enteroendocrine cells, (x) T-cells and (xi) neutrophils. (**B**) RNAseq data for 359 unfractionated human stomach samples was subject to weighted correlation network analysis (WGCNA). (i) Coloured squares indicate cell type *Ref.T.* positions on resultant dendrogram. (ii) Coloured bars show distribution of protein coding genes classified as cell type-enriched across dendrogram groups. (**C**) Human stomach tissue profiling for proteins encoded by genes classified as: (i) gastric enteroendocrine cell, (ii) mitotic cell, (iii) parietal cell, (iv) chief cell or (v) gastric mucous cell enriched. (**D**) Over-represented gene ontology terms among genes predicted to be (i) gastric enteroendocrine cell, (ii) parietal cell or (iii) gastric mucous cell enriched. See also Table S1 Tab 2, 3, 5 and 6.

Alternative analysis and protein profiling support cell-type classifications

Unsupervised weighted network correlation analysis is consistent with Ref.T. analysis

As our analysis is based on manually selected *Ref.T.* panels, cell-type classification is subject to an input bias. As a comparison, we subjected the same GTEx RNAseq dataset to a weighted network correlation analysis (WGCNA) (Langfelder and Horvath, 2008), an unbiased method that does not require any manual input or marker gene selection. WGNCA generates corr. coefficients between all transcripts and subsequently clusters them into related groups, based on expression similarity (Figure 2 B). In general, *Ref.T*. belonging to the same cell type panel were found in the same WCGNA cluster (Figure 2 B.i, coloured boxes represent *Ref.T*. locations), e.g., gastric enteroendocrine cells (cluster 58) or adjacent clusters on the same branch, e.g., gastric mucous cells (clusters 13 and 24) and macrophages (clusters 45 and 62) (Figure 2 B.i). Protein coding genes that we predicted to be cell type enriched were predominantly clustered into the same WGCNA group as the corresponding *Ref.T.,* or into adjacent groups on the same branch, consistent with our classifications (Figure 2 B.ii). Most genes in the *Ref.T.* panels representing parietal and chief cells appeared in the same large group (cluster 3) (Figure 2 B.ii), as were the genes in the respective predicted enriched gene lists, despite clear separation in our *Ref.T* based method (Figure 1 C, D). Despite the lack of separation for the enriched gene signatures for parietal and chief cells by WGNCA, each contained several well described marker genes for the respective cell type, e.g., *GIF, SLC26A7* (parietal) and *PGA4, SLC1A2* (chief cell). Indeed, we have previously shown that *Ref.T.* based analysis can have a higher sensitivity than WGNCA for cell type gene enrichment analysis (Dusart *et al.*, 2019). Stomach tissue protein profiling revealed staining consistent with expression in the respective cell types for proteins encoded by genes predicted to be gastric enteroendocrine cell (Figure 2 C.i), mitotic cell (Figure 2 C.ii), parietal cell (Figure 2 C.iii), chief cell (Figure 2 C.iv) or gastric mucous cell (Figure 2 C.v) enriched. GO and reactome analysis (Ashburner *et al.*, 2000; Gene Ontology, 2021) revealed over represented terms for predicted stomach specialised cell type enriched genes were consistent with known cell functions e.g., for gastric enteroendocrine cells *´enteroendocrine cell differentiation´* (Figure 2 D.i), for parietal

cells ´*inorganic ion transport across the plasma membrane*´ and ´*gastric acid secretion*´ (Figure 2 D.ii) and for gastric mucous cells '*lipid metabolic processes*´ (Figure 2 D.iii), (for all cell types see Table S1, Tab 3-13).

Stomach cell type gene enrichment signatures

Figure 3 shows up to the top 50 most enriched protein coding enriched genes for each cell type, ranked by highest corr. with the relevant *Ref.T.* panel (Figure 3 A.i-K.i), with differential corr. values and expression levels in the bulk RNAseq dataset (mean TPM). Mean TPM levels were generally highest for genes predicted to be enriched in parietal cells (Figure 3 A.i), chief cells (Figure 3 B.i), gastric mucous cells (Figure 3 D.i), fibroblasts (Figure 3 G.i) and plasma cells (Figure 3 K.i), and lowest for those in mitotic cells (Figure 3 E.i), neutrophils (Figure 3 I.i) and T-cells (Figure 3 J.i). This likely reflects differing numbers of each given cell type with the samples, however, as a range of expression values are observed within each given cell type, there is likely also individual gene variation in factors such as regulation and transcript stability. The highest differential values, and thus relative uniqueness among the profiled cell types, was observed for mitotic cell enriched genes (Figure 3 E.i), most of which have well studied roles in the regulation of the cell cycle, such as *TOP2A* and *BUB1B*. For all other cell types, top enriched genes included both known cell type specific genes, together with those that have not been previously reported as such, e.g., *PECAM1* and *SHE* were both predicted to be endothelial cell enriched (Figure 3 F.i); *PECAM1* is a commonly used marker gene for this cell type, whilst there are no existing reports for the selective expression of *SHE* in this context. Tissue profiling for proteins encoded by representative cell type enriched genes showed expression consistent with our classifications (Figure 3 A.ii-K.ii).

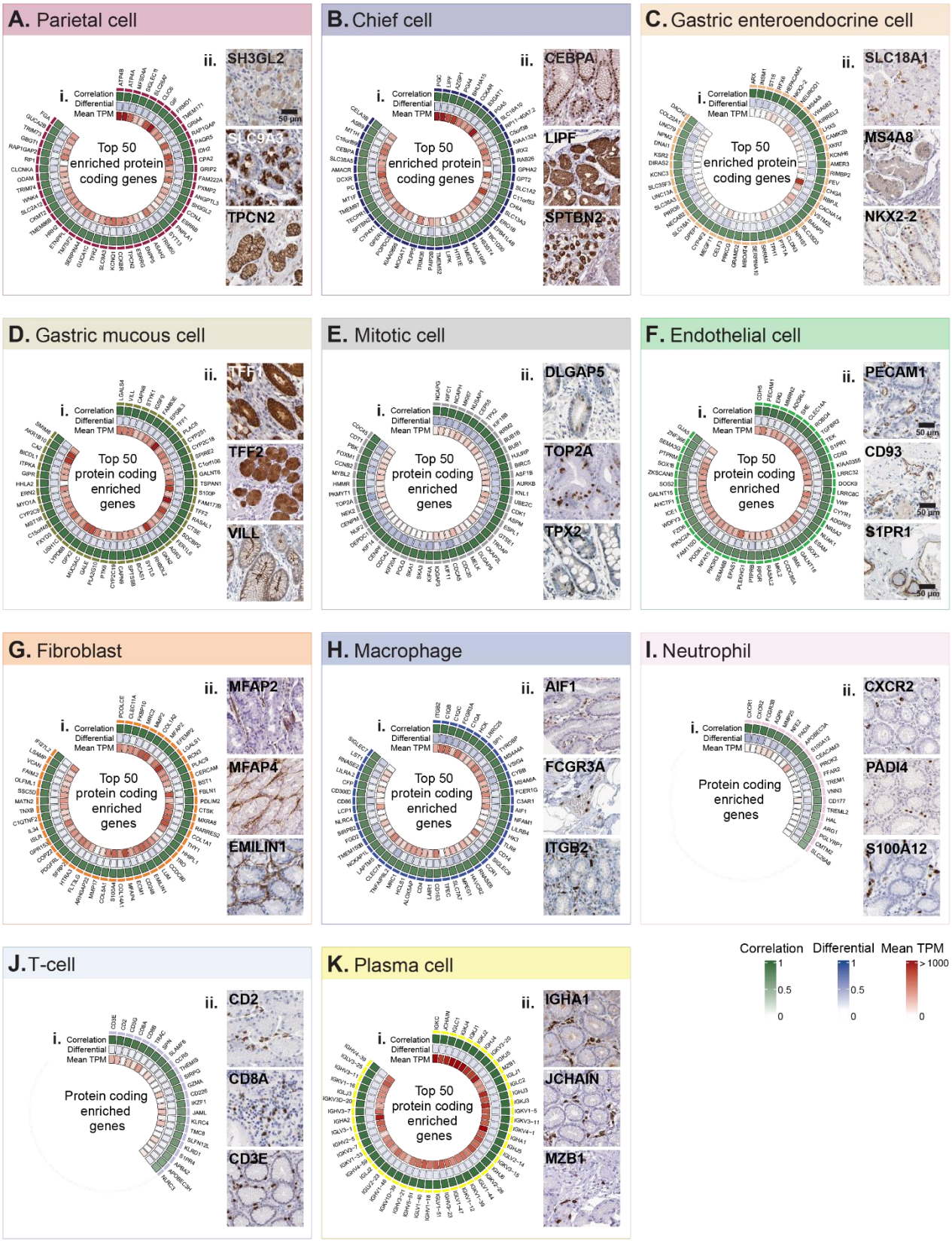

A. Parietal cell
i.
Correlation
Differential
Mean TPM
Top 50 enriched protein coding genes
ii.
SH3GL2
SLC9A3
TPCN2
50 µm
B. Chief cell
i.
Correlation
Differential
Mean TPM
Top 50 enriched protein coding genes
ii.
CEBPA
LIPF
SPTBN2
C. Gastric enteroendocrine cell
i.
Correlation
Differential
Mean TPM
Top 50 enriched protein coding genes
ii.
SLC18A1
MS4A8
NKX2-2
D. Gastric mucous cell
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
TFF1
TFF2
VILL
E. Mitotic cell
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
DLGAP5
TOP2A
TPX2
F. Endothelial cell
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
PECAM1
CD93
S1PR1
50 µm
G. Fibroblast
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
MFAP2
MFAP4
EMILIN1
H. Macrophage
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
AIF1
FCGR3A
ITGB2
I. Neutrophil
i.
Correlation
Differential
Mean TPM
Protein coding enriched genes
ii.
CXCR2
PADI4
S100A12
Correlation
1
0.5
0
Differential
1
0.5
0
Mean TPM
> 1000
0
J. T-cell
i.
Correlation
Differential
Mean TPM
Protein coding enriched genes
ii.
CD2
CD8A
CD3E
K. Plasma cell
i.
Correlation
Differential
Mean TPM
Top 50 protein coding enriched genes
ii.
IGHA1
JCHAIN
MZB1

Figure 3. Protein coding gene signatures of human stomach cell types. Cell type-enriched protein coding genes in: (**A**) parietal cells, (**B**) chief cells, (**C**) gastric enteroendocrine cells, (**D**) gastric mucous cells, (**E**) mitotic cells, (**F**) endothelial cells, (**G**) fibroblasts, (**H**) macrophages, (**I**) neutrophils (**J**) T-cells and (**K**) plasma cells, showing: (i) correlation coefficient with the cell type *Ref.T.* panel, differential correlation score (correlation with cell type *Ref.T.,* panel minus max correlation with any other *Ref.T.* panel) and mean expression in bulk RNAseq. (ii) Human stomach tissue protein profiling for selected cell type enriched genes. See also Table S1 Tab 2

Ref.T. analysis can predict source of stomach enriched protein coding genes

Genes with enriched expression in the human stomach vs. other tissue types can be identified by a comparative analysis of unfractionated tissue RNAseq data. We extracted the top 200 human stomach-enriched genes from the Human Protein Atlas (HPA) (Uhlen *et al.*, 2015) and GTEx project (Consortium, 2015), through the Harminozome database (Rouillard *et al.*, 2016) (Figure 4). Of the 78 genes classified as stomach-enriched in both datasets, 46/78 (59.0%) were classified as cell type enriched in our analysis; 28/46 (61.0%) in gastric mucous cells, 11/46 (24.0%) in parietal cells, 6/46 (13.0%) in chief cells, and 1/46 (2.2%) in gastric enteroendocrine cells (Figure 4 B.i and B.ii, respectively, large symbols). Of those not classified as cell type-enriched in our analysis (n=32), 11/32 (34.4%), only narrowly failed to reach one of the thresholds for classification as either parietal-, chief- or gastric mucous cell-enriched (Figure 4 B.i and B.ii, medium symbols). The majority of the remaining genes most highly correlated with *Ref.T.* panel representing one, or more, of the same cell types; parietal, chief or gastric mucous, but were excluded from the cell-type classifications due to shared enrichment. None of the stomach-enriched genes were predicted to be enriched in any cell type found across multiple tissue types, such as endothelial or immune cells, consistent with the lack of specificity of these cell type to the stomach. Thus, our analysis indicates that most stomach-tissue enriched genes are primarily expressed in gastric mucous, parietal or chief cells.

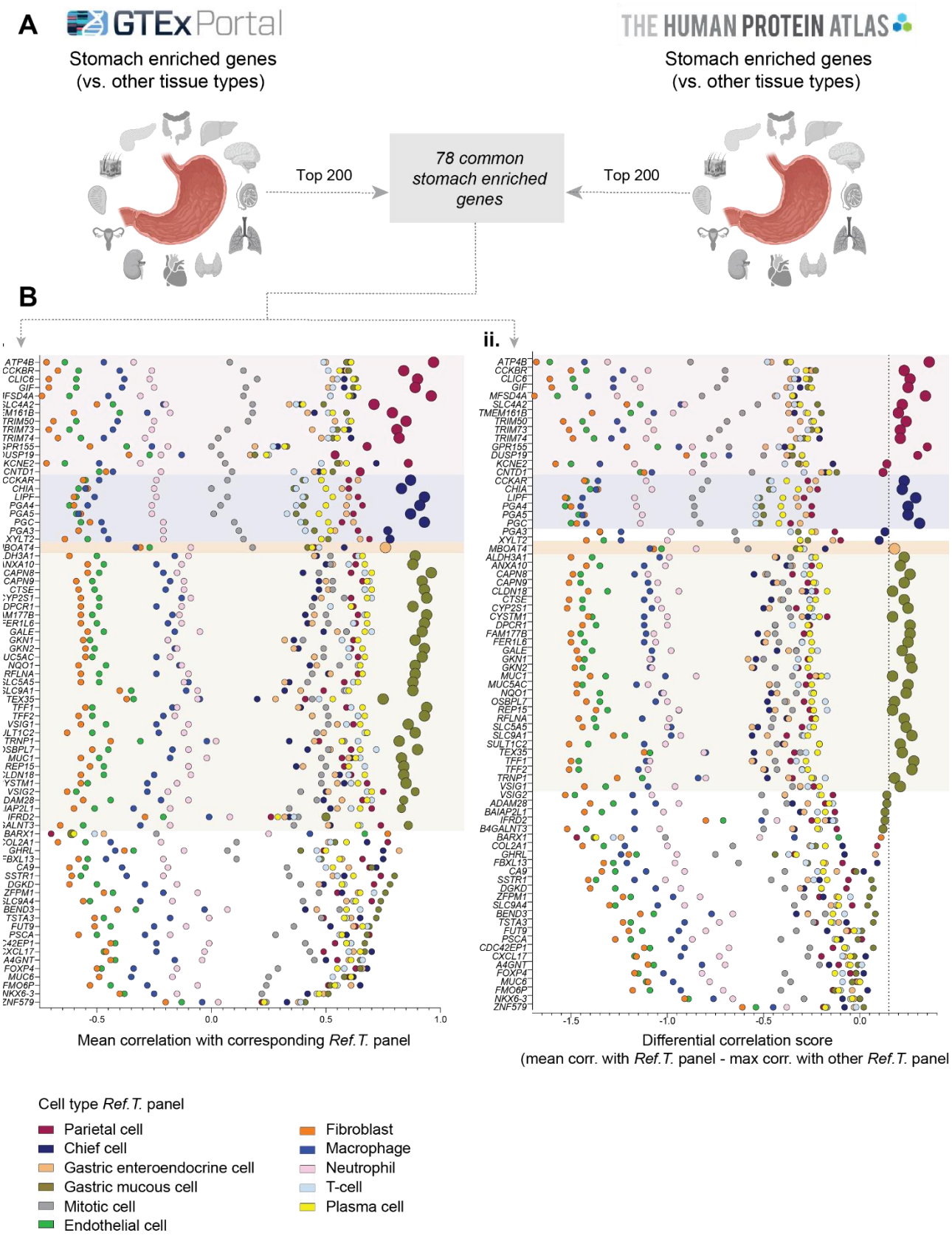

A
GTEx Portal
Stomach enriched genes
(vs. other tissue types)
THE HUMAN PROTEIN ATLAS
Stomach enriched genes
(vs. other tissue types)
Top 200
78 common stomach enriched genes
Top 200
B
ii.
-0.5
0.0
0.5
1.0
Mean correlation with corresponding Ref.T. panel
-1.5
-1.0
-0.5
0.0
Differential correlation score
(mean corr. with Ref.T. panel - max corr. with other Ref.T. panel
Cell type Ref.T. panel
Parietal cell
Chief cell
Gastric enteroendocrine cell
Gastric mucous cell
Mitotic cell
Endothelial cell
Fibroblast
Macrophage
Neutrophil
T-cell
Plasma cell

Figure 4. Gastric mucous cells, parietal cells and chief cells are the primary source of stomach tissue enriched genes. (**A**) The top 200 stomach enriched genes (vs. other tissue types) in RNAseq data from the GTEx Portal or Human Protein Atlas (HPA) were compared to identify genes common to both datasets (n=78). For each, the following was plotted: (**B**) (i) the mean correlation with each cell type *Ref.T.* panel, and (ii) the differential value vs. the next most highly correlating *Ref.T.* panel (dotted line indicates threshold for classification as cell type enriched). Enlarged circles represent genes with predicted cell type enrichment.

Cell-type enriched non-coding genes in stomach

A total of 252 non-coding genes were identified as cell-type enriched in the stomach (Figure 5 A), the greatest number of which were in gastric mucous cells, plasma cells, or fibroblasts (n=100, 44 and 30, respectively). When the sample set was analysed by WGNCA (Figure 5 B.i) non-coding genes that we predicted to be cell type enriched predominantly clustered into the same WGCNA group as the corresponding *Ref.T.,* or into adjacent groups on the same branch (Figure 5 B.ii). Up to the top 50 non-coding enriched genes in gastric enteroendocrine cells (Figure 5 C.i), gastric mucous cells (Figure 5 D.i), endothelial cell (Figure 5 E.i), parietal cells (Figure 6 A.i), chief cells (Figure 6 B.i), plasma cells (Figure 6 C.i), and fibroblasts (Figure 6 D.i), ranked by corr. with the relevant Ref.T panel, are displayed with differential corr. values vs. other profiled cell types, expression in the bulk RNAseq data (mean TMP) and transcript type. In all cell types, with the exception of plasma cells, where the most common type of enriched non-coding gene was IG pseudogene (Figure 6 C.i), long non-coding RNAs made up the majority of the predicted enriched genes. Generally, gastric mucous cell (Figure 5 D.i) and fibroblast (Figure 6 D.i) enriched non-coding genes were expressed at the highest levels in the stomach bulk RNAseq. This likely reflects the differing numbers of each given cell type within the samples, but the intra-cell type variation also indicates individual gene regulation.

There is currently no existing dataset of non-coding enriched genes in stomach cell types that could be used to validate our predictions. However, we sourced scRNAseq data from the analysis of 24 tissue types in Tabula sapiens (Tabula Sapiens *et al.*, 2022) (data for stomach was not available) that had been classified into endothelial, epithelial, immune and stromal cell functional compartments (for Tabula sapiens UMAP cell type classifications see Figure S3). We generated UMAP plots for each of these compartments to determine expression profiles for selected non-coding genes that we predicted to be cell type enriched. The predicted gastric enteroendocrine enriched genes *MIR7-3HG* and *RP5-984P4.6* were expressed only in the epithelial cell compartment, specifically in the clusters annotated as intestinal enteroendocrine and pancreatic alpha and beta cells (Figure 5 C.ii and iii), consistent with a specialised role in endocrine cells, not only in the stomach, but also in the pancreas and other parts of the GI

tract. The predicted gastric mucous cell enriched genes *CTD-2396E7.11* and *RP11-27G14.4* were widely expressed in the epithelial compartment, but not in the endothelial, immune, or stromal cell compartments (Figure 5 D.ii and iii). The predicted endothelial cell enriched genes *GATA2-AS1* and *AC007743.1* were expressed predominantly in the endothelial cell compartment (Figure 5 E.ii and iii), also consistent with our classifications. Genes predicted to be parietal cell enriched, *LINC00671* and *AC008268.1* (Figure 6 A.ii and iii), and chief cell enriched, *RP11-526I8.2* and *AZGP1P1* (Figure 6 B.ii and iii), were predominantly expressed in the epithelial compartment. The type of epithelial cell in which the genes were expressed varied, e.g., the chief cell enriched gene *AZGP1P1* (Figure 6 B.ii) was expressed predominantly in luminal cells of the prostate and hepatocytes; one could speculate that this gene indicates a shared secretory function between these specific cell types, whilst *RP11-526I8.2* was more generally expressed in the epithelial compartment (Figure 6 B.iii) perhaps indicating a more general role. The predicted plasma cell enriched genes *IGLV2-5* and *IGLVI-70* were expressed only in the immune cell compartment (Figure 6 C.ii and iii) in clusters annotated as either plasma cells or B-cells. The predicted fibroblast enriched genes *LINC01140* and *AC006007.1* were expressed predominantly in the stromal cell compartment (Figure 6 D.ii and iii), also consistent with our classifications. Thus, the Tabula sapiens scRNAseq data provides supportive evidence for our cell type classifications, despite the lack of stomach cell type analysis in this dataset.

Of those non-coding genes that we classified as cell type enriched, 17 had relatively high expression in the bulk RNAseq stomach samples (mean TPM >10) and were most frequently predicted to be gastric mucous cell enriched (Figure 6 E). To determine the expression profile of these genes in different organ types, we sourced data from bulk RNAseq of other tissues in GTEx (Figure 6 F). The most highly expressed parietal cell enriched non-coding genes, *LINC00982* and *PP7080* (mean TPM 99 and 49, respectively) both had high relative expression in stomach tissue (Figure 6 F.i and ii), consistent with a specialised function in this organ. *IGLC6*, the most highly expressed non-coding transcript we predicted to be enriched in plasma cells was highly expressed in spleen and salivary gland; tissues that contain high

numbers of plasma cells (Figure 6 F.iii). The most highly expressed non-coding genes we predicted to be enriched in gastric mucous cells, *FER1L4* and *RP11-363E7.4*, both had high relative expression in stomach and bladder (Figure 6 F.iv and v); one could speculate these genes have specific functions in the mucous cells found in these tissue types. *HSPA7*, the most highly expressed predicted fibroblast enriched gene had variable expression across tissue types (Figure 6 F.vi), consistent with the ubiquitous presence of this cell type across organs, whilst the chief cell enriched transcript, *C9orf147,* had high relative expression only in stomach tissue (Figure 6 F.vii). Thus, the most highly expressed non-coding genes predicted to be enriched in the stomach specialised cell types were detected at relatively high levels in stomach tissue (and in relatively few other tissue types), consistent with a specialised function here. Conversely, those predicted to be enriched in less specialised cell types, such as plasma cells, were more broadly expressed across tissue types, consistent with a common cell type function in multiple organs. All data for non-coding genes can be searched via the web portal https://cell-enrichment.shinyapps.io/noncoding_stomach/.

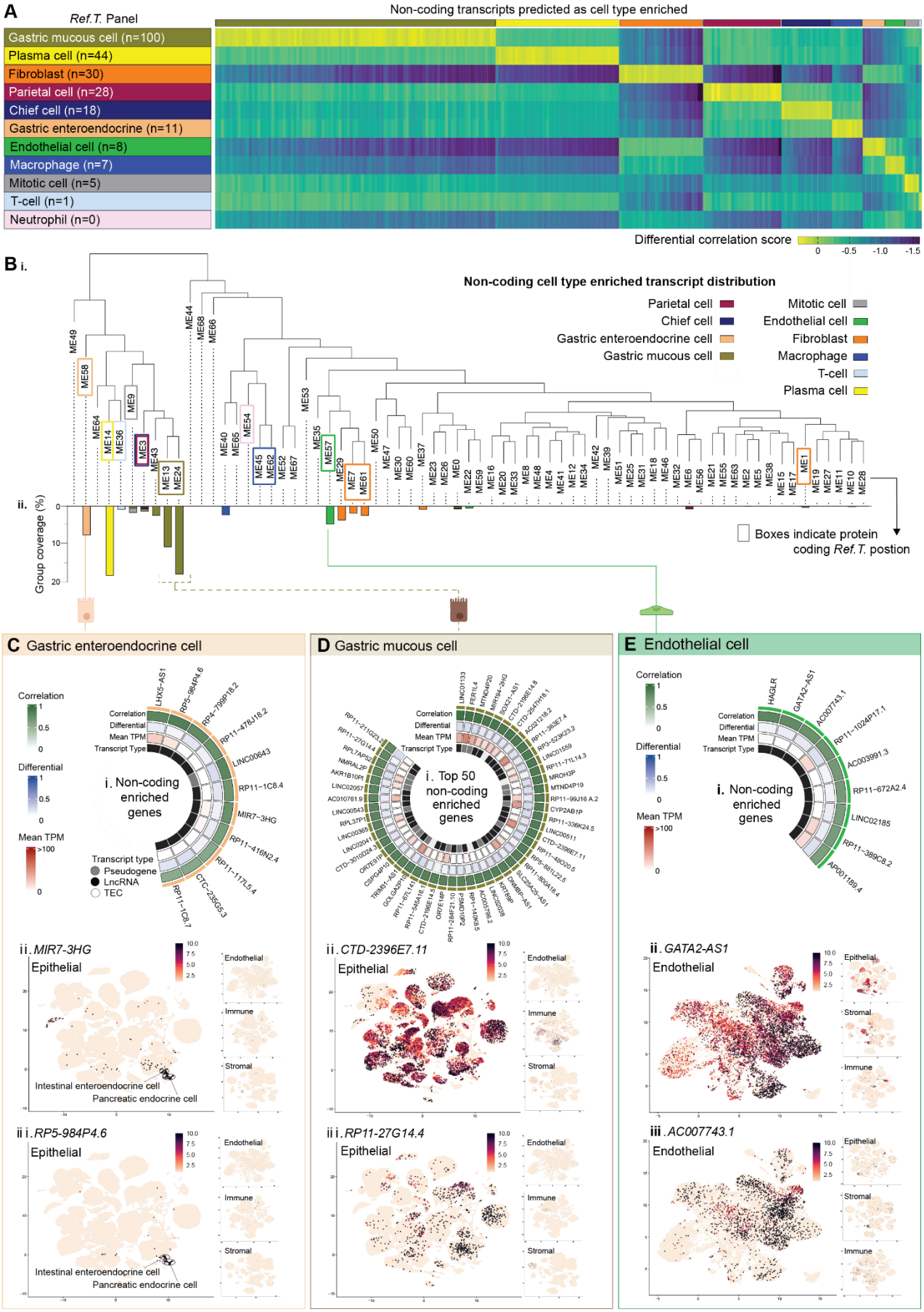

A
Ref.T. Panel
Gastric mucous cell (n=100)
Plasma cell (n=44)
Fibroblast (n=30)
Parietal cell (n=28)
Chief cell (n=18)
Gastric enteroendocrine (n=11)
Endothelial cell (n=8)
Macrophage (n=7)
Mitotic cell (n=5)
T-cell (n=1)
Neutrophil (n=0)
Non-coding transcripts predicted as cell type enriched
Differential correlation score
0
-0.5
-1.0
-1.5
B i.
Non-coding cell type enriched transcript distribution
Parietal cell
Chief cell
Gastric enteroendocrine cell
Gastric mucous cell
Mitotic cell
Endothelial cell
Fibroblast
Macrophage
T-cell
Plasma cell
ii.
Group coverage (%)
Boxes indicate protein coding Ref.T. postion
C Gastric enteroendocrine cell
Correlation
Differential
Mean TPM
Transcript Type
i. Non-coding enriched genes
Transcript type
Pseudogene
LncRNA
TEC
ii. MIR7-3HG
iii. RP5-984P4.6
Epithelial
Endothelial
Immune
Stromal
Intestinal enteroendocrine cell
Pancreatic endocrine cell
D Gastric mucous cell
i. Top 50 non-coding enriched genes
ii. CTD-2396E7.11
iii. RP11-27G14.4
E Endothelial cell
i. Non-coding enriched genes
ii. GATA2-AS1
iii. AC007743.1

Figure 5. Non-coding gene signatures of human stomach cell types. (A) Heat map of non-coding genes predicted to be cell type enriched, showing differential score between mean correlation coefficient with the corresponding *Ref.T.* panel vs. highest mean correlation coefficient amongst the other *Ref.T.* panels. (**B**) RNAseq data for 359 unfractionated human stomach samples was subject to weighted correlation network analysis (WGCNA). (i) Coloured squares indicate cell type *Ref.T.* positions on resultant dendrogram. (ii) Coloured bars show distribution of non-coding genes classified as cell type-enriched across dendrogram groups. Non-coding gene enrichment signatures for: (**C**) gastric enteroendocrine cells, (**D**) gastric mucous cells and (**E**) endothelial cells, detailing: (i) up to the top 50 cell type enriched non-coding genes, showing correlation coefficients with the *Ref.T.* panel, differential scores (correlation with corresponding cell type *Ref.T.,* panel minus max correlation with any other *Ref.T.* panel), mean expression in bulk RNAseq and transcript type. (ii and iii) scRNAseq data from analysis of epithelial, endothelial, immune or stromal cell compartments across 24 human tissues was sourced from Tabula Sapiens (Tabula Sapiens et al., 2022), and used to generate UMAP plots showing the expression profiles of example cell type enriched non-coding genes. The largest plot shows the compartment with the highest expression. See also Table S1 Tab 2 and Figure S3 (for all UMAP plot annotations).

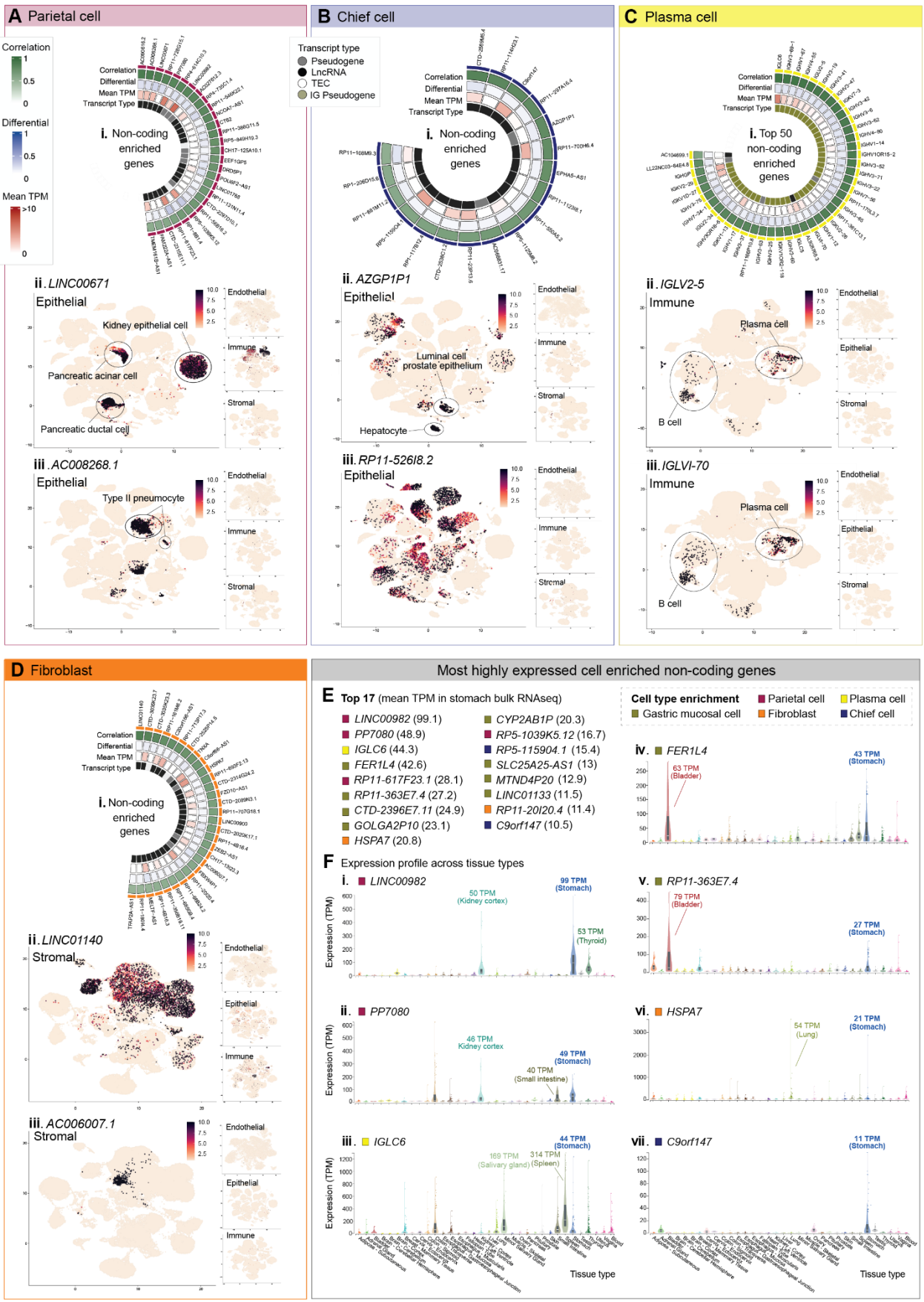
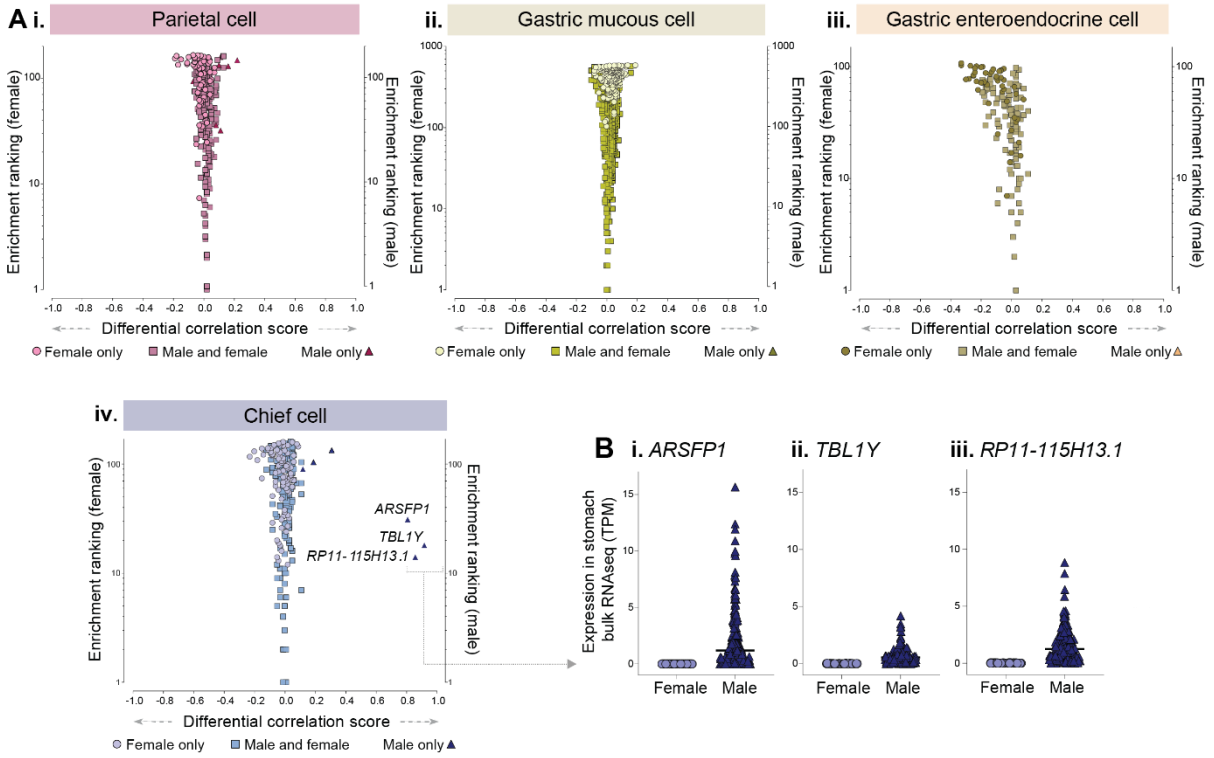

A Parietal cell
B Chief cell
C Plasma cell
D Fibroblast
Most highly expressed cell enriched non-coding genes
Correlation
Differential
Mean TPM
Transcript Type
Transcript type
Pseudogene
LncRNA
TEC
IG Pseudogene
i. Non-coding enriched genes
i. Top 50 non-coding enriched genes
ii. LINC00671
Epithelial
Kidney epithelial cell
Pancreatic acinar cell
Pancreatic ductal cell
ii. AZGP1P1
Luminal cell prostate epithelium
Hepatocyte
ii. IGLV2-5
Immune
Plasma cell
B cell
iii. AC008268.1
Type II pneumocyte
iii. RP11-526I8.2
iii. IGLVI-70
Endothelial
Immune
Epithelial
Stromal
ii. LINC01140
Stromal
iii. AC006007.1
E Top 17 (mean TPM in stomach bulk RNAseq)
LINC00982 (99.1)
PP7080 (48.9)
IGLC6 (44.3)
FER1L4 (42.6)
RP11-617F23.1 (28.1)
RP11-363E7.4 (27.2)
CTD-2396E7.11 (24.9)
GOLGA2P10 (23.1)
HSPA7 (20.8)
CYP2AB1P (20.3)
RP5-1039K5.12 (16.7)
RP5-115904.1 (15.4)
SLC25A25-AS1 (13)
MTND4P20 (12.9)
LINC01133 (11.5)
RP11-20I20.4 (11.4)
C9orf147 (10.5)
Cell type enrichment
Parietal cell
Plasma cell
Gastric mucosal cell
Fibroblast
Chief cell
F Expression profile across tissue types
iv. FER1L4
63 TPM (Bladder)
43 TPM (Stomach)
i. LINC00982
50 TPM (Kidney cortex)
99 TPM (Stomach)
53 TPM (Thyroid)
v. RP11-363E7.4
79 TPM (Bladder)
27 TPM (Stomach)
ii. PP7080
46 TPM Kidney cortex
49 TPM (Stomach)
40 TPM (Small intestine)
vi. HSPA7
54 TPM (Lung)
21 TPM (Stomach)
iii. IGLC6
169 TPM (Salivary gland)
314 TPM (Spleen)
44 TPM (Stomach)
vii. C9orf147
11 TPM (Stomach)
Expression (TPM)
Tissue type

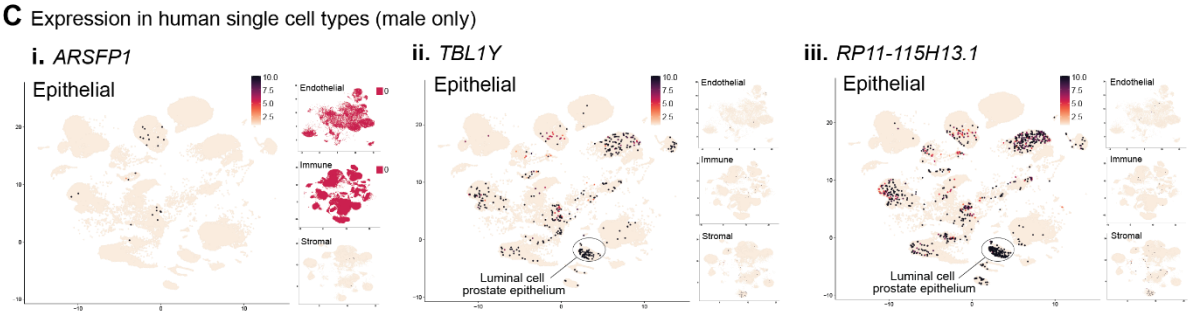
Figure 6. Core non-coding gene signatures of human stomach cell types and tissue distribution patterns. Non-coding gene enrichment signatures for: (**A**) parietal cells, (**B**) chief cells, (**C**) plasma cells and (**D**) endothelial cells, detailing (i) up to the top 50 cell type enriched non-coding genes, showing correlation coefficients with the *Ref.T.* panel, differential scores (correlation with corresponding cell type *Ref.T.,* panel minus max correlation with any other *Ref.T.* panel), mean expression in bulk RNAseq and gene type. (ii and iii) scRNAseq data from analysis of epithelial, endothelial, immune, or stromal cell compartments across 24 human tissues was sourced from Tabula Sapiens (Tabula Sapiens et al., 2022), and used to generate UMAP plots showing the expression profiles of example cell type enriched non-coding genes. The largest plot shows the compartment with the highest expression. (**E**) The top 50 most highly expressed cell type enriched non-coding genes in stomach bulk RNAseq. (**F**) Expression of genes classified as enriched in parietal cells: (i) *LINC00982* and (ii) *PP7080*, plasma cells: (iii) *IGLC6*, gastric mucous cells: (vi) *FER1L4* and (v) *RP11-363E7.4,* fibroblasts: (vi) *HSPA7* and chief cells: (vii) *C9orf147,* in bulk RNAseq of different human organs. Mean TMP expression is annotated for selected organs on each plot. See also Table S1 Tab 2 and Figure S2 (for all UMAP plot annotations).

Comparison of predicted sex-specific stomach cell type enriched genes

We performed a subset analysis of the stomach RNAseq dataset (male n=227, female n=132,), to identify sex-specific cell type enriched genes. Similar to the full dataset, intra-panel cell type *Ref.T.* correlated well in single-sex sample subsets (all >0.84) (Table S2, Tab 1, Table A and B). Cell type enriched genes were calculated as for the whole dataset. To compare gene enrichment profiles in males and females, the following was calculated for any gene that was classified as cell type enriched in either subset: (**i**) the '*differential correlation score*', defined as the difference between the mean corr. coefficient with the cell type *Ref.T*, in the male and female sample subsets, (**ii**) the '*enrichment ranking*', based on the mean corr. value with the *Ref.T.* panel (rank 1 = highest corr.). Cell profiles were mainly comparable between sexes, for both stomach specialised cell types (Figure 7 Ai-iv) and others (Figure S4 A-G) (genes enriched in *both* males and females represented by square symbols). For those genes classified as enriched *only* in males or females (represented by differently coloured triangle and circle symbols, respectively), most had differential corr. scores close to zero; indicating that they fell marginally below the designated threshold for classification as enriched in the other sex. A small number of distinct male-only enriched genes were identified in chief cells; *ARSFP1*, *TBL1Y* and *RP11-115H13.1* (Figure 7 A.iv), all of which were Y-linked, with expression levels above background level only in male samples (Figure 7 Bi-iii). As described above, we sourced scRNAseq data from Tabula sapiens (Tabula Sapiens *et al.*, 2022) for cells classified as endothelial, epithelial, immune or stromal (Figure S3). We generated UMAP plots (using cell data from male donors only) to show expression profiles of the male-only chief cell enriched genes. *ARSFP1* was detected only at low levels in the epithelial compartment (Figure 7 C.i), whilst *TBL1Y* (Figure 7 C.ii) and *RP11-115H13.1* (Figure 7 C.iii) had strikingly similar expression profiles, with the highest levels in both cases detected in prostate epithelial cells. All 3 male-only chief cell enrichened genes had low/no expression in the endothelial, immune or stromal compartments (Figure 7 Ci-iii). To determine the broad expression profile of the most highly expressed non-coding enriched genes across organs (from male donors), we sourced data from GTEx (Figure 7 D). *ARSFP1* had enhanced expression only in the stomach

and esophagus (Figure 7 D.i); both of which are tissue types not included in the Tabula sapiens dataset, consistent with the low detection observed there. *TBL1Y* and *RP11-115H13.1* had similar expression profiles across tissue types, with enhanced expression in thyroid (which was also absent from the Tabula Sapiens dataset) followed by prostate; in keeping with the high expression observed in prostate epithelial cells in the scRNAseq (Figure 7 D.ii-iii). Thus, one could speculate that male-only chief cell enriched gene *ARSFP1* has a stomach specific function, whilst *TBL1Y* and *RP11-115H13.1* appear to be co-expressed also in cell types outside the stomach, suggesting a broader function in multiple cell types.


A i. Parietal cell
ii. Gastric mucous cell
iii. Gastric enteroendocrine cell
iv. Chief cell
Enrichment ranking (female)
Enrichment ranking (male)
Differential correlation score
Female only
Male and female
Male only
ARSFP1
TBL1Y
RP11-115H13.1
B i. ARSFP1
ii. TBL1Y
iii. RP11-115H13.1
Expression in stomach bulk RNAseq (TPM)
Female
Male


C Expression in human single cell types (male only)
i. ARSFP1
ii. TBL1Y
iii. RP11-115H13.1
Epithelial
Endothelial
Immune
Stromal
Luminal cell
prostate epithelium

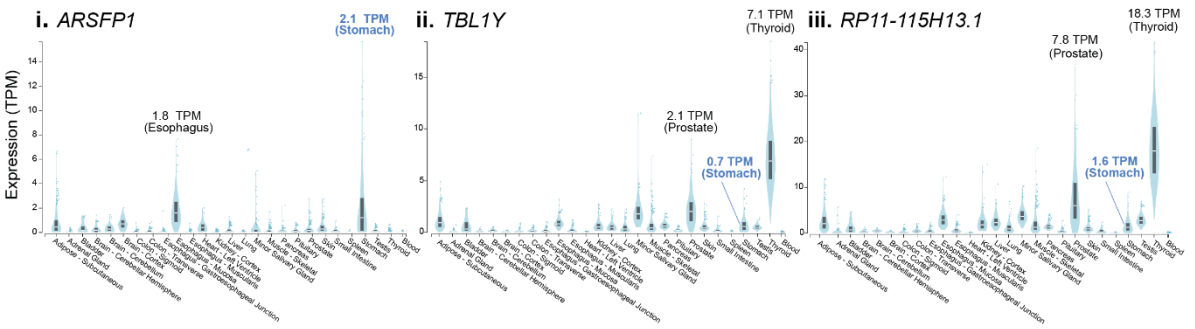

D Expression in human organs (male only)
i. ARSFP1
ii. TBL1Y
iii. RP11-115H13.1
Expression (TPM)
2.1 TPM (Stomach)
1.8 TPM (Esophagus)
7.1 TPM (Thyroid)
2.1 TPM (Prostate)
0.7 TPM (Stomach)
18.3 TPM (Thyroid)
7.8 TPM (Prostate)
1.6 TPM (Stomach)

Figure 7. Identification of sex-specific cell-enriched genes in human stomach tissue. (**A**) Human stomach RNAseq data (n=359 individuals) was retrieved from GTEx V8 and divided into female (n=132) and male (n=227) subgroups before classification of cell type-enriched genes. For genes classified as: (i) parietal, (ii) gastric mucous, (iii) gastric enteroendocrine or (vi) chief cell enriched in either sex, the *´sex differential corr. score'* (difference between mean corr. with the *Ref.T.* panel in females vs. males) was plotted vs. 'enrichment ranking' (position in each respective enriched list, highest corr. = rank 1). On each plot, genes enriched in *both* females and males are represented by common-coloured square symbols, and genes classified as enriched *only* in females or males are represented by differently coloured circle and triangle symbols, respectively. (**B**) Expression in female or male samples for genes classified as male-only enriched in chief cells: (i) *ARSFP1*, (iii) *TBL1Y* and (iii) *RP11-115H13.1*. (**C**) scRNAseq data from analysis of epithelial, endothelial, immune or stromal cell compartments across human tissues from male donors was sourced from Tabula Sapiens (Tabula Sapiens et al., 2022), and used to generate UMAP plots showing the expression profiles of: (i) *ARSFP1*, (iii) *TBL1Y* and (iii) *RP11-115H13.1*. (**D**) Expression of: (i) *ARSFP1*, (iii) *TBL1Y* and (iii) *RP11-115H13.1* in bulk RNAseq of different human organs from male donors. The largest plot shows the compartment with the highest expression. Mean expression is annotated for selected organs on each plot. See also Table S2 Tab 1, Figure S2 (for all UMAP plot annotations) and Figure S3.

DISCUSSION

Here, we present a genome wide cell type enriched transcriptome atlas for the human stomach, using our previously described method to resolve unfractionated tissue RNAseq data to the cell type level (Butler *et al.*, 2016; Dusart *et al.*, 2019; Norreen-Thorsen *et al.*, 2022). Our method circumvents some challenges associated with scRNAseq analysis, including issues associated with cell isolation, material amplification (Shapiro, Biezuner and Linnarsson, 2013; Grün and van Oudenaarden, 2015; Gawad, Koh and Quake, 2016) and induction of expression artefacts, due to loss of tissue specific cues or processing (O'Flanagan *et al.*, 2019). Our analysis incorporates a high number of biological replicates, reducing the impact of individual variation and allowing for well powered subgroup comparisons e.g., female vs. male. As data for gene enrichment signatures of stomach cell types are lacking in the existing literature, with this organ absent from large scale single cell sequencing (scRNAseq) initiatives, such as Tabula Sapiens (Tabula Sapiens *et al.*, 2022) and the Human Cell Atlas (Regev *et al.*, 2017) our study provides a useful resource, which can be searched on a gene-by-gene basis on the human protein atlas (www.proteinatlas.org/humanproteome/tissue+cell+type/stomach) or https://cell-enrichment.shinyapps.io/noncoding_stomach/, for protein coding and non-coding genes, respectively.

Of the 11 cell types we profiled in the stomach, gastric mucous cells had the highest number of predicted enriched genes, which included those encoding for proteins with known cell type specific functions, such as in mucosal defence, e.g., *CAPN8*, *CAPN9* (Hata *et al.*, 2010), *GKN1* (Choi *et al.*, 2013), *MUC13* (Ja *et al.*, 2020), *TFF1* and *TFF2* (Aihara, Engevik and Montrose, 2017) and lipid metabolism, e.g., *PLPP2* (Hooks, Ragan and Lynch, 1998), *PPARG* (Kang *et al.*, 2015) and *PLA2G10* (Hanasaki *et al.*, 2002). In addition, several genes we identified have no reported role in this cell type, including *FAM83E*, *CYP2S1* and *PLAC8*.

Predicted gastric enteroendocrine enriched genes also included those with known cell type function, such as *CAMK2B,* which is involved in intracellular calcium signalling (Tsakmaki *et al.*, 2020), and the neuroendocrine secretory protein *CHGA* (Goldspink, Reimann and Gribble, 2018). Other predicted gastric enteroendocrine enriched genes had not been described in

gastric enteroendocrine cells previously, such as *LHX5, SERPINA10* and *KCNH6*. *LHX5* has mainly been studied in the context of neuronal development (Zhao *et al.*, 1999; Pillai *et al.*, 2007), but in the GTEx database the only tissue type, outside the brain, where *LHX5* had elevated expression compared to others was the stomach (Consortium, 2015), thus, one could speculate that this gene also has a specific functional role here. *SERPINA10* was previously identified as a biomarker for gastrointestinal neuroendocrine carcinoma (Leja *et al.*, 2009), and *KCNH6* has a role in the regulation of insulin secretion in the pancreas (J.-K. Yang *et al.*, 2018); both consistent with our prediction that these genes have an endocrine cell enriched profile.

Many genes we predicted to be parietal cell enriched were well known markers of this cell type, such as *GIF* (Alpers and Russell-Jones, 2013) and *SLC26A7* (Petrovic *et al.*, 2003). However, others had no reported cell type specific expression or function, such as *ACSS1*, a mitochondrial matrix protein functioning as a catalyst of acetyl-CoA synthesis (Schwer *et al.*, 2006) and *MFSD4*, a marker for hepatic metastasis in gastric cancer (Shimizu, Kanda and Kodera, 2018). Our classifications were supported by a scRNAseq study that showed elevated expression of *ACSS1* and *MFSD4* in parietal cells vs. other stomach epithelial cells (Busslinger *et al.*, 2021). Other predicted enriched genes for which a function in parietal cells has not yet been described included *SLC12A3*, *ETNPPL*, *FNDC10*, *TUBA3C*, *TRIM73*, *TRIM74* and *CLCNKA*. Chief cell enriched genes included *BHLHA15*, a known chief cell marker (Lennerz *et al.*, 2010) and *KIAA1324*, which is required for chief cell secretory granule maturation (Cho, Park and Mills, 2022). Novel predicted chief cell enriched genes included the orphan receptor *GPR150*, a G-protein coupled receptor in which aberrant methylation has been linked to ovarian cancer (Cai *et al.*, 2007), *MOGAT1*, a monoacylglycerol acyltransferase that functions in the absorption of dietary fat in the intestine (Yen *et al.*, 2002) and *LIPK*, previously identified in the epidermis with a function in lipid metabolism (Toulza *et al.*, 2007).

Whilst there is no existing database of non-coding gene enrichment profiles in the cell types of the stomach, and a lack of information regarding the function of any such genes in normal tissue, increasing evidence of the involvement of non-coding genes in the development of gastric cancer (Li *et al.*, 2014; Gao *et al.*, 2020; Ghafouri-Fard and Taheri, 2020) and

associated drug resistance (Wei *et al.*, 2020) indicates that this transcript class has important functions in this tissue type. Of the stomach specialised cell types we profiled, gastric mucous cells had the highest number of predicted enriched non-coding genes, which included several antisense transcripts to corresponding gastric mucous cell enriched protein coding genes, such as *SOX21-AS1* and *TRIM31-AS1*, suggesting a local regulation of gene transcription. Many gastric mucous cell enriched non-coding genes were expressed at relatively high levels, compared to other non-coding genes in the same or other cell types, including *LINC01133*, *FER1L4, RP11-363E7.4* and *CTD-2396E7.11*. *LINC01133* and the pseudogene *FER1L4* are inhibitors of gastric cancer progression, with reduced expression associated with a more aggressive tumour phenotype (Xia *et al.*, 2015; X.-Z. Yang *et al.*, 2018). To date, there is a single publication on *RP11-363E7.4*, where a genome wide screen of gastric cancer samples identified it as a key regulator of disease progression, with higher expression associated with overall survival (Wang *et al.*, 2018). All the aforementioned studies were based on analysis of bulk RNAseq cancer samples, and the cell type in which these genes primarily function in healthy tissue is not reported; our data strongly indicates that this site is the mucous cell compartment. *CTD-2396E7.11* has not been described in the context of gastric cancer, but it was identified as one of four hub lncRNAs associated with reduced colon adenocarcinoma progression (Jiang, Tan and Zhang, 2019). As this tumour type also arises from the mucosa, on could speculate *CTD-2396E7.11* has a similar expression profile in healthy colon tissue. *LIN00982*, the highest expressed of all classified non-coding genes, was enriched in parietal cells and had, similar to those discussed above been shown to have a role in the inhibition of gastric cancer progression (Zheng *et al.*, 2021).

Examples of non-coding genes we predicted to have gastric enteroendocrine cell enriched expression included *MIR7-3HG* and *RP5-984P4.6*. The selective expression of these genes in pancreatic and intestinal endocrine cells (Tabula Sapiens *et al.*, 2022), is consistent with them having a conserved endocrine function. *MIR7-3HG* can act as an autophagy inhibitor (Capizzi *et al.*, 2017), but there are no reports of its function in an endocrine context. *RP5-984P4.6* is currently completely uncharacterised. Other gastric enteroendocrine cell enriched non-coding

genes included *LHX5-AS1*, an antisense transcript to the gastric enteroendocrine cell enriched corresponding protein coding gene.

Despite reported differences in stomach function between males and females, such as in speed of gastric emptying (Datz, Christian and Moore, 1987), gastrointestinal motility (Al-Shboul, 2016), incidence of gastric cancer (Lou *et al.*, 2020) and in gastric cancer survival (Li *et al.*, 2020), there are no studies of sex differences between stomach cell-type gene enrichment profiles. We found that global cell type gene enrichment signatures were similar between sexes, but we did identify 3 male-only chief cell enriched genes - *ARSFP1*, *RP11-115H13.1* and *TBL1Y*, all of which were Y-linked (Kirsch *et al.*, 2004; Yan *et al.*, 2005). In the GTEx database, the pseudogene *ARSFP1* was most highly expressed in male stomach samples, compared to the other 53 tissue types profiled from males (Consortium, 2015), supportive of a currently unknown sex and tissue specific role, and consistent with our predicted enrichment in a stomach-specific cell type in males. Although it is often assumed that pseudogenes lack function, recent studies have shown that they can have key roles, functioning as antisense, interference or competing endogenous transcripts (Pink *et al.*, 2011; Kovalenko and Patrushev, 2018; Cheetham, Faulkner and Dinger, 2020). *RP11-115H13.1* was one of only eight lncRNAs identified as associated with a high-risk of gastric cancer (Zhao *et al.*, 2022), but the dataset analysed in this study contained both male and female samples, meaning the prognostic value of *RP11-115H13.1* in male patients was likely underestimated. To our knowledge, there are no existing reports of the potential cellular function of *RP11-115H13.1* or *ARSFP1*. *TBL1Y* has been reported as involved in syndromic hearing loss (Di Stazio *et al.*, 2019) and cardiac differentiation (Meyfour *et al.*, 2017), but studies of its function in the stomach are lacking.

There are limitations in our study. We do not profile cell subtypes, such as those included under the umbrella term of ´gastric enteroendocrine cells´ including D-cells and G-cells, for which it was not possible to identify *Ref.T.* that fulfilled the required criteria. Our observations are consistent with these sub-cell types being typically defined by the expression of a limited number of specialised proteins (Sjölund *et al.*, 1983; Engelstoft *et al.*, 2013; Gribble and

Reimann, 2016), rather than large distinct gene signature panels. Gene expression in stomach can be modified by genetic or environmental factors, such as the individual variation in the gastrointestinal microbiome (Nichols and Davenport, 2021). Strongly regulated genes may therefore not correlate with the more constitutively expressed *Ref.T.* selected to represent the cell type in which they are primarily expressed, as variation across samples could be independent of cell type proportions. Thus, such genes could be false negatives in our analysis. Furthermore, we have used high thresholds for the classification of genes as cell type-enriched, which could lead to incorrect exclusion. For example, tissue profiling showed that proteins encoded by *MUC4* and *MUC5B* are selectively expressed in gastric mucous cells (Uhlen *et al.*, 2019), but they fall just below the threshold for classification as such in our analysis. In addition, exclusion of lowly expressed genes from the analysis many also result in false negative classifications for rarer cell types, for example *PAX6*, which controls endocrine cell differentiation (Beucher *et al.*, 2012), and proglucagon (Hill, Asa and Drucker, 1999) and gastric inhibitory polypeptide (Fujita *et al.*, 2008) production, was excluded from classification as a gastric enteroendocrine enriched gene only due to expression level below the designated cut off. However, in all cases the individual enrichment scores clearly indicate a cell-type enriched expression; thus, our classifications should be regarded as a guide, and the data should be considered on a gene-by-gene basis.

ACKNOWLEDGEMENTS

Funding granted to LMB from Hjärt Lungfonden (20170759, 20170537, 20200544) and Swedish Research Council (2019-01493), and to JO from Stockholm County Council (SLL 2017-0842). The Human Protein Atlas is funded by The Knut and Alice Wallenberg Foundation. **Data usage:** We used data from Genotype-Tissue Expression (GTEx) Project (gtexportal.org) (Consortium, 2015) supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

AUTHOR CONTRIBUTIONS Conceptualisation: LMB. Methodology: SÖ, ES, MNT, PD. Formal analysis: SÖ, PD, LMB. Investigation SÖ, PD, LMB, CL. Resources: MU, FP, JO, LB, CL. Writing – Original Draft: SÖ, LMB. Writing – Review & Editing: All, Visualisation: SÖ, LMB, PD, MZ, KVF. Supervision: LMB, PD. Funding Acquisition: LMB, JO**.**

DECLARATION OF INTERESTS The authors declare no competing interests.

METHODS AND RESOURCES

LEAD CONTACT

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact: Dr. Lynn Marie Butler. Email: Lynn.butler@ki.se

MATERIALS AVALIBILITY

This study did not generate new unique reagents.

DATA AND CODE AVAILABILITY

- This paper analyses existing, publicly available data from the Genotype-Tissue Expression(GTEx) project with accession number phs000424.v8.p2 (Consortium, 2015) and single cell RNAseq data from Tabula Sapiens (Tabula Sapiens *et al.*, 2022) retrieved on 2022/07/29.
- All original code has been deposited at GitHub and is publicly available as of the date of publication, link: https://github.com/PhilipDusart/cell-enrichment.
- No additional information should be required to reanalyse the data reported in this paper, but any necessary clarifications or queries can be directed towards the lead contact.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bulk RNAseq data analysed in this study was obtained from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) (Consortium, 2015) accessed on 2021/04/26 (dbGaP Accession phs000424.v8.p2). Transcript types were categorised according to Biotype definitions in ENSEMBL release 102 (Yates *et al.*, 2020). Human tissue protein profiling was performed in house as part of the Human Protein Atlas (HPA) project (Ponten, Jirstrom and Uhlen, 2008; Uhlen *et al.*, 2015, 2017) (www.proteinatlas.org). Human stomach tissue samples were obtained from the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden, as part of the Uppsala Biobank. Samples were handled in accordance with Swedish laws and regulations, with approval from the Uppsala Ethical Review Board (Uhlen et al., 2015).

METHOD DETAILS

Tissue Profiling: Human tissue sections

Stomach tissue sections were stained, as previously described (Ponten, Jirstrom and Uhlen, 2008; Uhlen *et al.*, 2015). Briefly, formalin fixed and paraffin embedded tissue samples were sectioned, de-paraffinised in xylene, hydrated in graded alcohols and blocked for endogenous peroxidase in 0.3% hydrogen peroxide diluted in 95% ethanol. For antigen retrieval, a Decloaking chamber® (Biocare Medical, CA) was used. Slides were boiled in Citrate buffer®, pH6 (Lab Vision, CA). Primary antibodies and a dextran polymer visualization system (UltraVision LP HRP polymer®, Lab Vision) were incubated for 30 min each at room temperature and slides were developed for 10 minutes using Diaminobenzidine (Lab Vision) as the chromogen. Slides were counterstained in Mayers hematoxylin (Histolab) and scanned using Scanscope XT (Aperio). Primary antibodies, source, target and identifier are as follows: Atlas Antibodies: ACSS1 (Cat#HPA043228, RRID:AB_2678372), ATP4A (Cat#HPA076684), ATP4B (Cat#HPA045400, RRID:AB_2679314), MFSD4A (Cat#055407), SH3GL2 (Cat#HPA026685, RRID:AB_1856817), SLC9A3 (Cat#HPA036493, RRID:AB_10673353), TPCN2 (Cat#HPA027080, RRID:AB_10600917), CEBPA (Cat#HPA065037, RRID:AB_2685410), LIPF (Cat#HPA045930, RRID:AB_10959518), SPTBN2 (Cat#HPA043529, RRID:AB_2678531), BHLHA15 (Cat#HPA047834, RRID:AB_2680172), KIAA1324 (Cat#HPA029869, RRID:AB_10794320), PGC (Cat#HPA031717, RRID:AB_10670130), CAMK2B (Cat#HPA053973, RRID:AB_2682328), SLC18A1 (Cat#HPA063797, RRID:AB_2685125), MS4A8 (Cat#HPA007319, RRID:AB_1854138), NKX2-2 (Cat#HPA003468, RRID:AB_1079490), TFF2 (Cat#HPA036705, RRID:AB_2675263), VILL (Cat#HPA035675, RRID:AB_10671223), CTSE (Cat#HPA012940, RRID:AB_2668773), FER1L6 (Cat#HPA054117, RRID:AB_2682387), LGALS4 (Cat#HPA031186, RRID:AB_2673778), PLAC8 (Cat#HPA040465, RRID:AB_10794875), CCNB1 (Cat#HPA061448, RRID:AB_2684522), DLGAP5 (Cat#HPA005546, RRID:AB_1078677), TPX2 (Cat#HPA005487, RRID:AB_1858223), PECAM1 (Cat#HPA004690, RRID:AB_1078462), CD93 (Cat#HPA009300, RRID:AB_1846342),

MFAP2 (Cat#HPA007354, RRID:AB_1079365), MFAP4 (Cat#HPA054097, RRID:AB_2682378) EMILIN1 (Cat#HPA002822, RRID:AB_1078738), AIF1 (Cat#HPA049234, RRID:AB_2680685), ITGB2 (Cat#HPA016894, RRID:AB_1846257), CXCR2 (Cat#HPA032017, RRID:AB_2674112), PADI4 (Cat#HPA017007, RRID:AB_1854921), S100A12 (Cat#HPA002881, RRID:AB_1848175), CD2 (Cat#HPA003883, RRID:AB_1846263), CD3E (Cat#HPA043955, RRID:AB_2678747), IGHA1 (Cat#HPA001217, RRID:AB_1079120), JCHAIN (Cat#HPA044132, RRID:AB_2678826) and MZB1 (Cat#HPA043745, RRID:AB_10960359) from Santa Cruz Biotechnology: AZGP1 (Cat#sc-13585, RRID:AB_667849), VWA5B2 (Atlas Antibodies Cat#HPA036823, RRID:AB_10672269), BIRC5 (Cat#sc-17779, RRID:AB_628302), CDC20 (Cat#sc-13162, RRID:AB_628089), S1PR1 (Cat#sc-48356, RRID:AB_2238920), FCGR3A (Cat#sc-20052, RRID:AB_626925) from Agilent: CD8A (Cat#M7103, RRID:AB_2075537) from Leica Biosystems: TOP2A (Cat#NCL-TOPOIIA, RRID:AB_564035), TFF1 (Cat#NCL-pS2, RRID:AB_563985) from Epitomics an AbCam company: CDK1 (Cat#1161-1, RRID:AB_344898) and from Roche: CHGA (Product name: 1199 021).

QUANTIFICATION AND STATISTICAL ANALYSIS

Reference transcript-based correlation analysis and criteria for cell type enrichment

This method was adapted and expanded from that previously developed to determine the cross-tissue pan-EC-enriched transcriptome (Butler *et al.*, 2016) and human brain and adipose tissue cell-enriched genes (Dusart *et al.*, 2019; Norreen-Thorsen *et al.*, 2022). Pairwise Spearman correlation coefficients were calculated between reference transcripts selected as proxy markers (´*Ref.T.* panels´) for: parietal cells [*ATP4B, MFSD4A, ATP4A*], chief cells [*PGC, LIPF, AZGP1*], gastric enteroendocrine cells [*ST18, INSM1, ARX*], gastric mucous cells [*LGALS4, VILL, CAPN8*], mitotic cells [*NCAPG, KIFC1, NCAPH*], endothelial cells [PECAM1, CDH5, ERG], fibroblasts [*PCOLCE, CLEC11A, MMP2*], macrophages [*C1QB, FCGR3A, ITGB2*], neutrophils [*CXCR2, FCGR3B, CXCR1*], T-cells [*CD3E, CD2, CD3G*] and plasma cells [*IGKC, JCHAIN, IGLC1*] and all other sequenced transcripts. Correlation coefficients were calculated in R using the *corr.test* function from the *psych* package (v 1.8.4) and False

Discovery Rate (FDR) adjusted p-values (using Bonferroni correction) and raw p-values were calculated. Genes were classified as cell type enriched when the following criteria were fulfilled: (i) a mean correlation >0.50 (FDR <0.0001) with the *Ref.T.* panel representing that cell type *and* (ii) a minimum ´differential correlation´ between this value and the *next highest* mean correlation with any other *Ref.T.* panel (representing another cell type) >0.15 *and* (iii) TPM expression <0.1 in over 50% of samples. See Figure S1 for method overview.

Weighted correlation network (WGCNA) analysis

The R package WGCNA (Langfelder and Horvath, 2008) was used to perform co-expression network analysis for gene clustering, on log2 expression TPM values. The analysis was performed according to recommendations in the WGCNA manual. Transcripts with too many missing values were excluded using the goodSamplesGenes() function. The remaining genes were used to cluster the samples, and obvious outlier samples were excluded.

Gene ontology and reactome analysis

The Gene Ontology Consortium (Ashburner *et al.*, 2000) and PANTHER classification resource (Mi *et al.*, 2013, 2016) were used to identify over represented terms (biological processes) in each set of predicted cell type enriched genes from the GO ontology (release date 2022-10-13) or reactome (Version 77, release date 2021-10-01) databases. Plots of GO terms were created using REVIGO (Supek et al., 2011) where stated.

Visualisation

Circular graphs were constructed using the R package *circlize* (Gu *et al.*, 2014). Principle component analysis plot was generated using https://biit.cs.ut.ee/clustvis/ (Metsalu and Vilo, 2015). Some figure sections were created with BioRender.com.

Additional datasets and analysis

Single cell RNAseq data from Tabula Sapiens (Tabula Sapiens et al., 2022) was downloaded and UMAP plots created using the Seurat package in R (Hao et al., 2021). Tissue enriched genes were downloaded from the Human Protein Atlas (HPA) tissue atlas (Uhlen et al., 2015) or GTEx database (Consortium, 2015), as collated in the Harminozome database (Rouillard et al., 2016).

ADDITIONAL RESOURCES

Analysed data for all protein coding genes is provided on the Human Protein Atlas website: (https://www.proteinatlas.org/humanproteome/tissue+cell+type/stomach). Data for non-coding genes is provided on https://cell-enrichment.shinyapps.io/noncoding_stomach/. The published article includes all datasets generated during this study (Tables S1 and S2).

SUPPLEMENTAL TABLE LEGENDS

Table S1. Reference transcript selection and analysis criteria.

(Tab **1**): Correlation coefficient values were calculated between selected *Ref.T.* to represent constituent stomach cell types. (Tab **2)**: Correlation coefficient values were calculated between selected *Ref.T.* and all other sequenced transcripts in GTEx stomach mRNAseq data (Table A) and the mean differential vs. all *Ref.T.* panels (Table B). Genes classified as enriched in: (Tab **3**) parietal cells, (Tab **4**) chief cells, (Tab **5**) gastric enteroendocrine cells, (Tab **6**) gastric mucous cells, (Tab **7**) mitotic cells, (Tab **8**) endothelial cells, (Tab **9**) fibroblasts, (Tab **10**) macrophages, (Tab **11**) neutrophils, (Tab **12**) T-cells and (Tab **13**) plasma cells were analysed to identify over-represented terms in the (Table **A**) gene ontology or (Table **B**). *Related to all Figures*.

Table S2. Sex stratified subset analysis of cell-enriched genes in human stomach. (Tab **1)**: Correlation coefficient values were calculated between selected *Ref.T.* to represent constituent stomach cell types in females (Table **A**) or males (Table **B**). (Tab **2**) Correlation coefficient values were calculated between selected *Ref.T.* and all other sequenced transcripts in stomach mRNAseq data (GTEx), subdivided into (Table **A**) female or (Table **B**) male only sample sets. See key for column details. *Related to Figure 7 and S4*.

REFERENCES

Aihara, E., Engevik, K.A. and Montrose, M.H. (2017) 'Trefoil Factor Peptides and Gastrointestinal Function', *Annual Review of Physiology*, 79, pp. 357–380. Available at: https://doi.org/10.1146/annurev-physiol-021115-105447.

Alpers, D.H. and Russell-Jones, G. (2013) 'Gastric intrinsic factor: The gastric and small intestinal stages of cobalamin absorption. A personal journey', *Biochimie*, 95(5), pp. 989–994. Available at: https://doi.org/10.1016/j.biochi.2012.12.006.

Al-Shboul, O. (2016) 'The role of the RhoA/ROCK pathway in gender-dependent differences in gastric smooth muscle contraction', *The Journal of Physiological Sciences*, 66(1), pp. 85–92. Available at: https://doi.org/10.1007/s12576-015-0400-9.

Ashburner, M. *et al.* (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, 25(1), pp. 25–29. Available at: https://doi.org/10.1038/75556.

Beucher, A. *et al.* (2012) 'The Homeodomain-Containing Transcription Factors Arx and Pax4 Control Enteroendocrine Subtype Specification in Mice', *PLOS ONE*, 7(5), p. e36449. Available at: https://doi.org/10.1371/journal.pone.0036449.

Busslinger, G.A. *et al.* (2021) 'Human gastrointestinal epithelia of the esophagus, stomach, and duodenum resolved at single-cell resolution', *Cell Reports*, 34(10), p. 108819. Available at: https://doi.org/10.1016/j.celrep.2021.108819.

Butler, L.M. *et al.* (2016) 'Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome', *Cell Systems*, 3(3), pp. 287-301.e3. Available at: https://doi.org/10.1016/j.cels.2016.08.001.

Cai, L. *et al.* (2007) 'Identification of PRTFDC1 silencing and aberrant promoter methylation of GPR150, ITGA8 and HOXD11 in ovarian cancers', *Life Sciences*, 80(16), pp. 1458–1465. Available at: https://doi.org/10.1016/j.lfs.2007.01.015.

Capizzi, M. *et al.* (2017) 'MIR7-3HG, a MYC-dependent modulator of cell proliferation, inhibits autophagy by a regulatory loop involving AMBRA1', *Autophagy*, 13(3), pp. 554–566. Available at: https://doi.org/10.1080/15548627.2016.1269989.

Cheetham, S.W., Faulkner, G.J. and Dinger, M.E. (2020) 'Overcoming challenges and dogmas to understand the functions of pseudogenes', *Nature Reviews Genetics*, 21(3), pp. 191–201. Available at: https://doi.org/10.1038/s41576-019-0196-1.

Cho, C.J., Park, D. and Mills, J.C. (2022) 'ELAPOR1 is a secretory granule maturation-promoting factor that is lost during paligenosis', *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 322(1), pp. G49–G65. Available at: https://doi.org/10.1152/ajpgi.00246.2021.

Choi, E. *et al.* (2014) 'Cell lineage distribution atlas of the human stomach reveals heterogeneous gland populations in the gastric antrum', *Gut*, 63(11), pp. 1711–1720. Available at: https://doi.org/10.1136/gutjnl-2013-305964.

Choi, W.S. *et al.* (2013) 'Gastrokine 1 expression in the human gastric mucosa is closely associated with the degree of gastritis and DNA methylation', *Journal of Gastric Cancer*, 13(4), pp. 232–241. Available at: https://doi.org/10.5230/jgc.2013.13.4.232.

Consortium, G.Te. (2015) ‘Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans’, *Science*, 348(6235), pp. 648–60. Available at: https://doi.org/10.1126/science.1262110.

Datz, F.L., Christian, P.E. and Moore, J. (1987) ‘Gender-related differences in gastric emptying’, *Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine*, 28(7), pp. 1204–1207.

Denisenko, E. *et al.* (2020) ‘Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows’, *Genome Biol*. 2020/06/04 edn, 21(1), p. 130. Available at: https://doi.org/10.1186/s13059-020-02048-6.

Denninger, J.K. *et al.* (2022) ‘Robust Transcriptional Profiling and Identification of Differentially Expressed Genes With Low Input RNA Sequencing of Adult Hippocampal Neural Stem and Progenitor Populations’, *Frontiers in Molecular Neuroscience*, 15, p. 810722. Available at: https://doi.org/10.3389/fnmol.2022.810722.

Di Stazio, M. *et al.* (2019) ‘TBL1Y: a new gene involved in syndromic hearing loss’, *European Journal of Human Genetics*, 27(3), pp. 466–474. Available at: https://doi.org/10.1038/s41431-018-0282-4.

Dusart, P. *et al.* (2019) ‘A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes’, *Cell Reports*, 29(6), pp. 1690-1706.e4. Available at: https://doi.org/10.1016/j.celrep.2019.09.088.

Engelstoft, M.S. *et al.* (2013) ‘Enteroendocrine cell types revisited’, *Current Opinion in Pharmacology*, 13(6), pp. 912–921. Available at: https://doi.org/10.1016/j.coph.2013.09.018.

Franzen, O., Gan, L.M. and Bjorkegren, J.L.M. (2019) ‘PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data’, *Database (Oxford)*. 2019/04/06 edn, 2019. Available at: https://doi.org/10.1093/database/baz046.

Fujita, Y. *et al.* (2008) ‘Pax6 and Pdx1 are required for production of glucose-dependent insulinotropic polypeptide in proglucagon-expressing L cells’, *American Journal of Physiology-Endocrinology and Metabolism*, 295(3), pp. E648–E657. Available at: https://doi.org/10.1152/ajpendo.90440.2008.

Gao, Y. *et al.* (2020) ‘Long noncoding RNAs in gastric cancer: From molecular dissection to clinical application’, *World J Gastroenterol*. 2020/07/14 edn, 26(24), pp. 3401–3412. Available at: https://doi.org/10.3748/wjg.v26.i24.3401.

Gawad, C., Koh, W. and Quake, S.R. (2016) ‘Single-cell genome sequencing: current state of the science’, *Nature Reviews Genetics*, 17(3), pp. 175–188. Available at: https://doi.org/10.1038/nrg.2015.16.

Gene Ontology, C. (2021) ‘The Gene Ontology resource: enriching a GOld mine’, *Nucleic Acids Res*. 2020/12/09 edn, 49(D1), pp. D325–D334. Available at: https://doi.org/10.1093/nar/gkaa1113.

Ghafouri-Fard, S. and Taheri, M. (2020) ‘Long non-coding RNA signature in gastric cancer’, *Experimental and Molecular Pathology*, 113, p. 104365. Available at: https://doi.org/10.1016/j.yexmp.2019.104365.

Goldspink, D.A., Reimann, F. and Gribble, F.M. (2018) ‘Models and Tools for Studying Enteroendocrine Cells’, *Endocrinology*, 159(12), pp. 3874–3884. Available at: https://doi.org/10.1210/en.2018-00672.

Gremel, G. *et al.* (2015) 'The human gastrointestinal tract-specific transcriptome and proteome as defined by RNA sequencing and antibody-based profiling', *Journal of Gastroenterology*, 50(1), pp. 46–57. Available at: https://doi.org/10.1007/s00535-014-0958-7.

Gribble, F.M. and Reimann, F. (2016) 'Enteroendocrine Cells: Chemosensors in the Intestinal Epithelium', *Annual Review of Physiology*, 78(1), pp. 277–299. Available at: https://doi.org/10.1146/annurev-physiol-021115-105439.

Grün, D. and van Oudenaarden, A. (2015) 'Design and Analysis of Single-Cell Sequencing Experiments', *Cell*, 163(4), pp. 799–810. Available at: https://doi.org/10.1016/j.cell.2015.10.039.

Gu, Z. *et al.* (2014) 'circlize Implements and enhances circular visualization in R', *Bioinformatics*. 2014/06/16 edn, 30(19), pp. 2811–2. Available at: https://doi.org/10.1093/bioinformatics/btu393.

Hanasaki, K. *et al.* (2002) 'Potent Modification of Low Density Lipoprotein by Group X Secretory Phospholipase A2 Is Linked to Macrophage Foam Cell Formation *', *Journal of Biological Chemistry*, 277(32), pp. 29116–29124. Available at: https://doi.org/10.1074/jbc.M202867200.

Hassan, M.I., Toor, A. and Ahmad, F. (2010) 'Progastriscin: structure, function, and its role in tumor progression', *J Mol Cell Biol*. 2010/03/17 edn, 2(3), pp. 118–27. Available at: https://doi.org/10.1093/jmcb/mjq001.

Hata, S. *et al.* (2010) 'Calpain 8/nCL-2 and calpain 9/nCL-4 constitute an active protease complex, G-calpain, involved in gastric mucosal defense', *PLoS genetics*, 6(7), p. e1001040. Available at: https://doi.org/10.1371/journal.pgen.1001040.

Hill, M.E., Asa, S.L. and Drucker, D.J. (1999) 'Essential Requirement for Pax6 in Control of Enteroendocrine Proglucagon Gene Transcription', *Molecular Endocrinology*, 13(9), pp. 1474–1486. Available at: https://doi.org/10.1210/mend.13.9.0340.

Hooks, S.B., Ragan, S.P. and Lynch, K.R. (1998) 'Identification of a novel human phosphatidic acid phosphatase type 2 isoform', *FEBS Letters*, 427(2), pp. 188–192. Available at: https://doi.org/10.1016/S0014-5793(98)00421-9.

Ja, G. *et al.* (2020) 'Mucins in Intestinal Mucosal Defense and Inflammation: Learning From Clinical and Experimental Studies', *Frontiers in immunology*, 11. Available at: https://doi.org/10.3389/fimmu.2020.02054.

Jiang, R. *et al.* (2022) 'Statistics or biology: the zero-inflation controversy about scRNA-seq data', *Genome Biol*. 2022/01/23 edn, 23(1), p. 31. Available at: https://doi.org/10.1186/s13059-022-02601-5.

Jiang, S., Tan, B. and Zhang, X. (2019) 'Identification of key lncRNAs in the carcinogenesis and progression of colon adenocarcinoma by co-expression network analysis', *Journal of Cellular Biochemistry*, 120(4), pp. 6490–6501. Available at: https://doi.org/10.1002/jcb.27940.

Kang, Y. *et al.* (2015) 'PPARG Modulated Lipid Accumulation in Dairy GMEC via Regulation of ADRP Gene', *Journal of Cellular Biochemistry*, 116(1), pp. 192–201. Available at: https://doi.org/10.1002/jcb.24958.

Karlsson, M. *et al.* (2021) 'A single-cell type transcriptomics map of human tissues', *Sci Adv*. 2021/07/30 edn, 7(31). Available at: https://doi.org/10.1126/sciadv.abh2169.

Kim, J. *et al.* (2022) 'Single-cell analysis of gastric pre-cancerous and cancer lesions reveals cell lineage diversity and intratumoral heterogeneity', *NPJ Precis Oncol*. 2022/01/29 edn, 6(1), p. 9. Available at: https://doi.org/10.1038/s41698-022-00251-1.

Kim, T.H. and Shivdasani, R.A. (2016) 'Stomach development, stem cells and disease', *Development*. 2016/02/18 edn, 143(4), pp. 554–65. Available at: https://doi.org/10.1242/dev.124891.

Kirsch, S. *et al.* (2004) 'Molecular and evolutionary analysis of the growth-controlling region on the human Y chromosome', *Human Genetics*, 114(2), pp. 173–181. Available at: https://doi.org/10.1007/s00439-003-1028-z.

Kovalenko, T.F. and Patrushev, L.I. (2018) 'Pseudogenes as Functionally Significant Elements of the Genome', *Biochemistry (Moscow)*, 83(11), pp. 1332–1349. Available at: https://doi.org/10.1134/S0006297918110044.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: an R package for weighted correlation network analysis', *BMC Bioinformatics*, 9(1), p. 559. Available at: https://doi.org/10.1186/1471-2105-9-559.

Leja, J. *et al.* (2009) 'Novel markers for enterochromaffin cells and gastrointestinal neuroendocrine carcinomas', *Modern Pathology: An Official Journal of the United States and Canadian Academy of Pathology, Inc*, 22(2), pp. 261–272. Available at: https://doi.org/10.1038/modpathol.2008.174.

Lennerz, J.K.M. *et al.* (2010) 'The Transcription Factor MIST1 Is a Novel Human Gastric Chief Cell Marker Whose Expression Is Lost in Metaplasia, Dysplasia, and Carcinoma', *The American Journal of Pathology*, 177(3), pp. 1514–1533. Available at: https://doi.org/10.2353/ajpath.2010.100328.

Li, H. *et al.* (2020) 'Gender Differences in Gastric Cancer Survival: 99,922 Cases Based on the SEER Database', *Journal of Gastrointestinal Surgery*, 24(8), pp. 1747–1757. Available at: https://doi.org/10.1007/s11605-019-04304-y.

Li, P.-F. *et al.* (2014) 'Non-coding RNAs and gastric cancer', *World Journal of Gastroenterology : WJG*, 20(18), pp. 5411–5419. Available at: https://doi.org/10.3748/wjg.v20.i18.5411.

Lou, L. *et al.* (2020) 'Sex difference in incidence of gastric cancer: an international comparative study based on the Global Burden of Disease Study 2017', *BMJ open*, 10(1), p. e033323. Available at: https://doi.org/10.1136/bmjopen-2019-033323.

Massoni-Badosa, R. *et al.* (2020) 'Sampling time-dependent artifacts in single-cell genomics studies', *Genome Biol*. 2020/05/13 edn, 21(1), p. 112. Available at: https://doi.org/10.1186/s13059-020-02032-0.

Meyfour, A. *et al.* (2017) 'Y Chromosome Missing Protein, TBL1Y, May Play an Important Role in Cardiac Differentiation', *Journal of Proteome Research*, 16(12), pp. 4391–4402. Available at: https://doi.org/10.1021/acs.jproteome.7b00391.

Mi, H. *et al.* (2013) 'Large-scale gene function analysis with the PANTHER classification system', *Nat Protoc*, 8(8), pp. 1551–66. Available at: https://doi.org/10.1038/nprot.2013.092.

Mi, H. *et al.* (2016) 'PANTHER version 10: expanded protein families and functions, and analysis tools', *Nucleic Acids Res*, 44(D1), pp. D336-42. Available at: https://doi.org/10.1093/nar/gkv1194.

Nichols, R.G. and Davenport, E.R. (2021) 'The relationship between the gut microbiome and host gene expression: a review', *Human Genetics*, 140(5), pp. 747–760. Available at: https://doi.org/10.1007/s00439-020-02237-0.

Norreen-Thorsen, M. *et al.* (2022) 'A human adipose tissue cell-type transcriptome atlas', *Cell Reports*, 40(2), p. 111046. Available at: https://doi.org/10.1016/j.celrep.2022.111046.

O'Flanagan, C.H. *et al.* (2019) 'Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses', *Genome Biol*. 2019/10/19 edn, 20(1), p. 210. Available at: https://doi.org/10.1186/s13059-019-1830-0.

Petrovic, S. *et al.* (2003) 'Identification of a basolateral Cl−/HCO 3 − exchanger specific to gastric parietal cells', *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 284(6), pp. G1093–G1103. Available at: https://doi.org/10.1152/ajpgi.00454.2002.

Pillai, A. *et al.* (2007) 'Lhx1 and Lhx5 maintain the inhibitory-neurotransmitter status of interneurons in the dorsal spinal cord', *Development*, 134(2), pp. 357–366. Available at: https://doi.org/10.1242/dev.02717.

Pink, R.C. *et al.* (2011) 'Pseudogenes: Pseudo-functional or key regulators in health and disease?', *RNA*, 17(5), pp. 792–798. Available at: https://doi.org/10.1261/rna.2658311.

Ponten, F., Jirstrom, K. and Uhlen, M. (2008) 'The Human Protein Atlas - a tool for pathology', *Journal of Pathology*, 216(4), pp. 387–393. Available at: https://doi.org/10.1002/path.2440.

Razavi, H. and Katanforosh, A. (2022) 'Identification of novel key regulatory lncRNAs in gastric adenocarcinoma', *BMC Genomics*. 2022/05/08 edn, 23(1), p. 352. Available at: https://doi.org/10.1186/s12864-022-08578-6.

Regev, A. *et al.* (2017) 'The Human Cell Atlas', *eLife*. Edited by T.R. Gingeras, 6, p. e27041. Available at: https://doi.org/10.7554/eLife.27041.

Rouillard, A.D. *et al.* (2016) 'The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins', *Database (Oxford)*. 2016/07/05 edn, 2016. Available at: https://doi.org/10.1093/database/baw100.

de Santa Barbara, P., van den Brink, G.R. and Roberts, D.J. (2003) 'Development and differentiation of the intestinal epithelium', *Cellular and Molecular Life Sciences (CMLS)*, 60(7), pp. 1322–1332. Available at: https://doi.org/10.1007/s00018-003-2289-3.

Sathe, A. *et al.* (2020) 'Single-Cell Genomic Characterization Reveals the Cellular Reprogramming of the Gastric Tumor Microenvironment', *Clinical Cancer Research*, 26(11), pp. 2640–2653. Available at: https://doi.org/10.1158/1078-0432.CCR-19-3231.

Schwer, B. *et al.* (2006) 'Reversible lysine acetylation controls the activity of the mitochondrial enzyme acetyl-CoA synthetase 2', *Proceedings of the National Academy of Sciences*, 103(27), pp. 10224–10229. Available at: https://doi.org/10.1073/pnas.0603968103.

Shapiro, E., Biezuner, T. and Linnarsson, S. (2013) 'Single-cell sequencing-based technologies will revolutionize whole-organism science', *Nature Reviews Genetics*, 14(9), pp. 618–630. Available at: https://doi.org/10.1038/nrg3542.

Shimizu, D., Kanda, M. and Kodera, Y. (2018) 'Emerging evidence of the molecular landscape specific for hematogenous metastasis from gastric cancer', *World Journal of*

Gastrointestinal Oncology, 10(6), pp. 124–136. Available at: https://doi.org/10.4251/wjgo.v10.i6.124.

Sjölund, K. *et al.* (1983) 'Endocrine Cells in Human Intestine: An Immunocytochemical Study', *Gastroenterology*, 85(5), pp. 1120–1130. Available at: https://doi.org/10.1016/S0016-5085(83)80080-8.

Squair, J.W. *et al.* (2021) 'Confronting false discoveries in single-cell differential expression', *Nature Communications*, 12(1), p. 5692. Available at: https://doi.org/10.1038/s41467-021-25960-2.

Tabula Sapiens, C. *et al.* (2022) 'The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans', *Science*. 2022/05/14 edn, 376(6594), p. eabl4896. Available at: https://doi.org/10.1126/science.abl4896.

Thompson, C.A., DeLaForest, A. and Battle, M.A. (2018) 'Patterning the gastrointestinal epithelium to confer regional-specific functions', *Developmental Biology*, 435(2), pp. 97–108. Available at: https://doi.org/10.1016/j.ydbio.2018.01.006.

Toulza, E. *et al.* (2007) 'Large-scale identification of human genes implicated in epidermal barrier function', *Genome Biology*, 8(6), p. R107. Available at: https://doi.org/10.1186/gb-2007-8-6-r107.

Tsakmaki, A. *et al.* (2020) 'ISX-9 manipulates endocrine progenitor fate revealing conserved intestinal lineages in mouse and human organoids', *Molecular Metabolism*, 34, pp. 157–173. Available at: https://doi.org/10.1016/j.molmet.2020.01.012.

Tsubosaka, A. *et al.* (2022) 'Single-Cell Transcriptome Analyses Reveal the Cell Diversity and Developmental Features of Human Gastric and Metaplastic Mucosa'. bioRxiv, p. 2022.05.22.493006. Available at: https://doi.org/10.1101/2022.05.22.493006.

Uhlen, M. *et al.* (2015) 'Proteomics. Tissue-based map of the human proteome', *Science*. 2015/01/24 edn, 347(6220), p. 1260419. Available at: https://doi.org/10.1126/science.1260419.

Uhlen, M. *et al.* (2017) 'A pathology atlas of the human cancer transcriptome', *Science*, 357(6352). Available at: https://doi.org/10.1126/science.aan2507.

Uhlen, M. *et al.* (2019) 'A genome-wide transcriptomic analysis of protein-coding genes in human blood cells', *Science*, 366(6472), p. eaax9198. Available at: https://doi.org/10.1126/science.aax9198.

Wang, P. *et al.* (2018) 'A Novel LncRNA-miRNA-mRNA Triple Network Identifies LncRNA RP11-363E7.4 as An Important Regulator of miRNA and Gene Expression in Gastric Cancer', *Cellular Physiology and Biochemistry*, 47(3), pp. 1025–1041. Available at: https://doi.org/10.1159/000490168.

Wang, R. *et al.* (2021) 'Single-cell dissection of intratumoral heterogeneity and lineage diversity in metastatic gastric adenocarcinoma', *Nature Medicine*, 27(1), pp. 141–151. Available at: https://doi.org/10.1038/s41591-020-1125-8.

Wei, L. *et al.* (2020) 'Noncoding RNAs in gastric cancer: implications for drug resistance', *Molecular Cancer*, 19(1), p. 62. Available at: https://doi.org/10.1186/s12943-020-01185-7.

Xia, T. *et al.* (2015) 'Long noncoding RNA FER1L4 suppresses cancer cell growth by acting as a competing endogenous RNA and regulating PTEN expression', *Scientific Reports*, 5(1), p. 13445. Available at: https://doi.org/10.1038/srep13445.

Yan, H.-T. *et al.* (2005) 'Molecular analysis of TBL1Y, a Y-linked homologue of TBL1X related with X-linked late-onset sensorineural deafness', *Journal of Human Genetics*, 50(4), pp. 175–181. Available at: https://doi.org/10.1007/s10038-005-0237-9.

Yang, J.-K. *et al.* (2018) 'From Hyper- to Hypoinsulinemia and Diabetes: Effect of KCNH6 on Insulin Secretion', *Cell Reports*, 25(13), pp. 3800-3810.e6. Available at: https://doi.org/10.1016/j.celrep.2018.12.005.

Yang, X.-Z. *et al.* (2018) 'LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/β-catenin pathway', *Molecular Cancer*, 17(1), p. 126. Available at: https://doi.org/10.1186/s12943-018-0874-1.

Yates, A.D. *et al.* (2020) 'Ensembl 2020', *Nucleic Acids Research*, 48(D1), pp. D682–D688. Available at: https://doi.org/10.1093/nar/gkz966.

Yen, C.-L.E. *et al.* (2002) 'Identification of a gene encoding MGAT1, a monoacylglycerol acyltransferase', *Proceedings of the National Academy of Sciences of the United States of America*, 99(13), pp. 8512–8517. Available at: https://doi.org/10.1073/pnas.132274899.

Zhang, P. *et al.* (2019) 'Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer', *Cell Reports*, 27(6), pp. 1934-1947.e5. Available at: https://doi.org/10.1016/j.celrep.2019.04.052.

Zhang, X. *et al.* (2019) 'CellMarker: a manually curated resource of cell markers in human and mouse', *Nucleic Acids Res*. 2018/10/06 edn, 47(D1), pp. D721–D728. Available at: https://doi.org/10.1093/nar/gky900.

Zhao, X. *et al.* (2022) 'An Immunity-Associated lncRNA Signature for Predicting Prognosis in Gastric Adenocarcinoma', *Journal of Healthcare Engineering*, 2022, p. e3035073. Available at: https://doi.org/10.1155/2022/3035073.

Zhao, Y. *et al.* (1999) 'Control of Hippocampal Morphogenesis and Neuronal Differentiation by the LIM Homeobox Gene Lhx5', *Science*, 284(5417), pp. 1155–1158. Available at: https://doi.org/10.1126/science.284.5417.1155.

Zheng, L. *et al.* (2021) 'Long noncoding RNA LINC00982 upregulates CTSF expression to inhibit gastric cancer progression via the transcription factor HEY1', *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 320(5), pp. G816–G828. Available at: https://doi.org/10.1152/ajpgi.00209.2020.

Consortium, G.T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648-660. 10.1126/science.1262110.
Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

Metsalu, T., and Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. Nucleic Acids Res *43*, W566-570. 10.1093/nar/gkv468.
Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) *2016*. 10.1093/database/baw100.
Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800. 10.1371/journal.pone.0021800.
Tabula Sapiens, C., Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science *376*, eabl4896. 10.1126/science.abl4896.
Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. 10.1126/science.1260419.

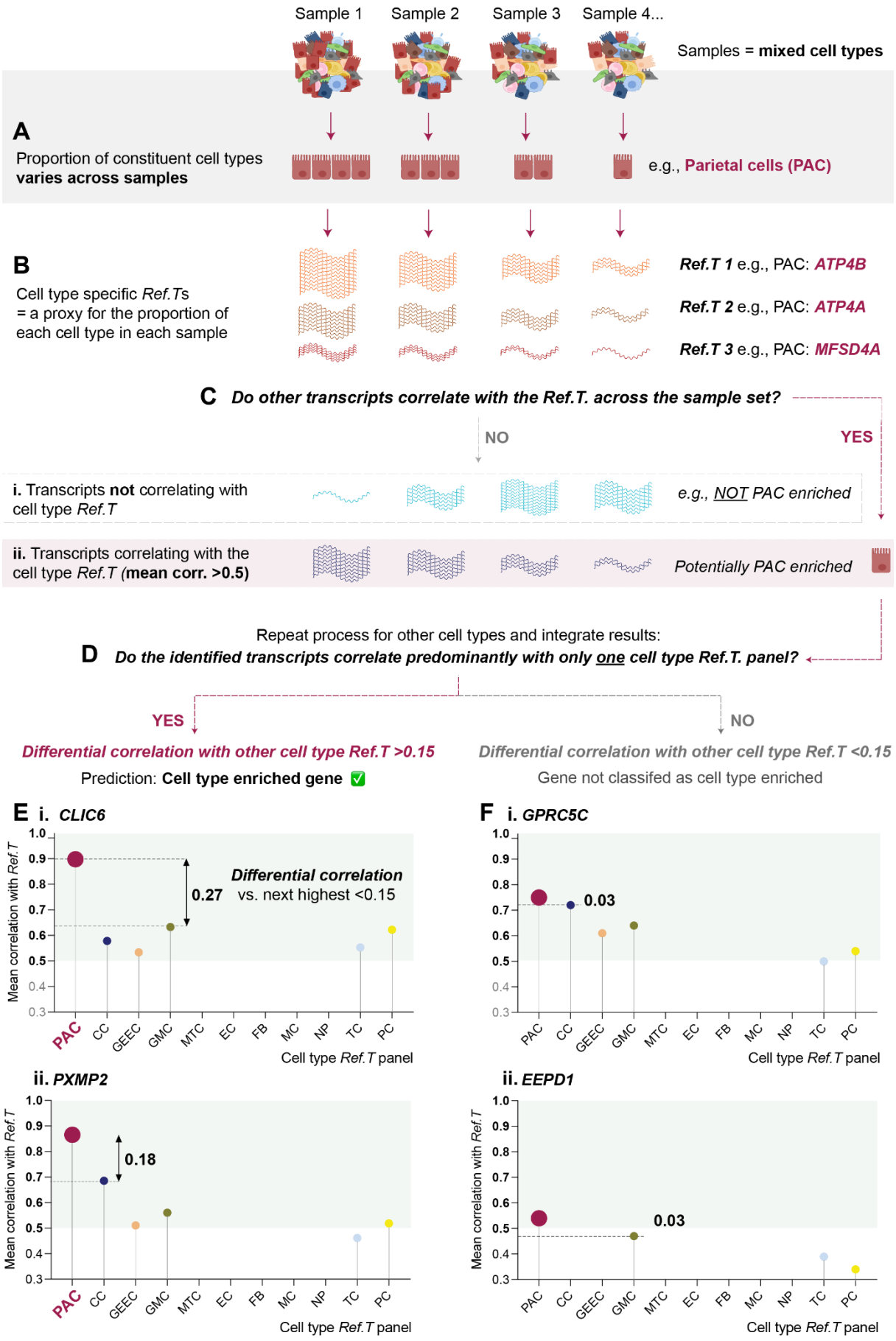

Sample 1
Sample 2
Sample 3
Sample 4...
Samples = mixed cell types
A
Proportion of constituent cell types varies across samples
e.g., Parietal cells (PAC)
B
Cell type specific Ref.Ts = a proxy for the proportion of each cell type in each sample
Ref.T 1 e.g., PAC: ATP4B
Ref.T 2 e.g., PAC: ATP4A
Ref.T 3 e.g., PAC: MFSD4A
C Do other transcripts correlate with the Ref.T. across the sample set?
NO
YES
i. Transcripts not correlating with cell type Ref.T
e.g., NOT PAC enriched
ii. Transcripts correlating with the cell type Ref.T (mean corr. >0.5)
Potentially PAC enriched
Repeat process for other cell types and integrate results:
D Do the identified transcripts correlate predominantly with only one cell type Ref.T. panel?
YES
NO
Differential correlation with other cell type Ref.T >0.15
Prediction: Cell type enriched gene
Differential correlation with other cell type Ref.T <0.15
Gene not classifed as cell type enriched
E i. CLIC6
Differential correlation
0.27 vs. next highest <0.15
Mean correlation with Ref.T
Cell type Ref.T panel
PAC
CC
GEEC
GMC
MTC
EC
FB
MC
NP
TC
PC
ii. PXMP2
0.18
F i. GPRC5C
0.03
ii. EEPD1
0.03

Figure S1. Overview of the analysis methodology. Related to all figures. RNAseq data for 359 unfractionated human stomach samples were retrieved from GTEx V8. (**A**) Each sequenced sample contained mixed cell types, present in differing proportions. (**B**) Cell type marker genes (´reference transcripts´ [*Ref.T.*]) were selected, based on in house tissue protein profiling and/or existing literature and datasets, as a proxy for the cell proportion within each sample (e.g., *ATP4B*, *ATP4A* and *MFSD4A* for parietal cells [PAC]). (**C**) Spearman correlation coefficients (corr.) between each selected *Ref.T.* and all other sequenced transcripts (>56,000) were calculated across samples and classified as: (i) not correlated, or (ii) correlated (corr. >0.50, p-value <0.00001). (**D**) This process was repeated for *Ref.T.* representing all cell types, and results integrated to identify genes that (**E**) correlated predominantly with only one cell type *Ref.T* panel (´differential corr.´ to next highest >0.15), which were classified as cell type enriched e.g., (i) *CLIC6* and (ii) *PXMP2,* or those that (**F**) did not selectively correlate with one cell type *Ref.T* panel, e.g., (i) *GPRC5C* and (ii) *EEPD1.*

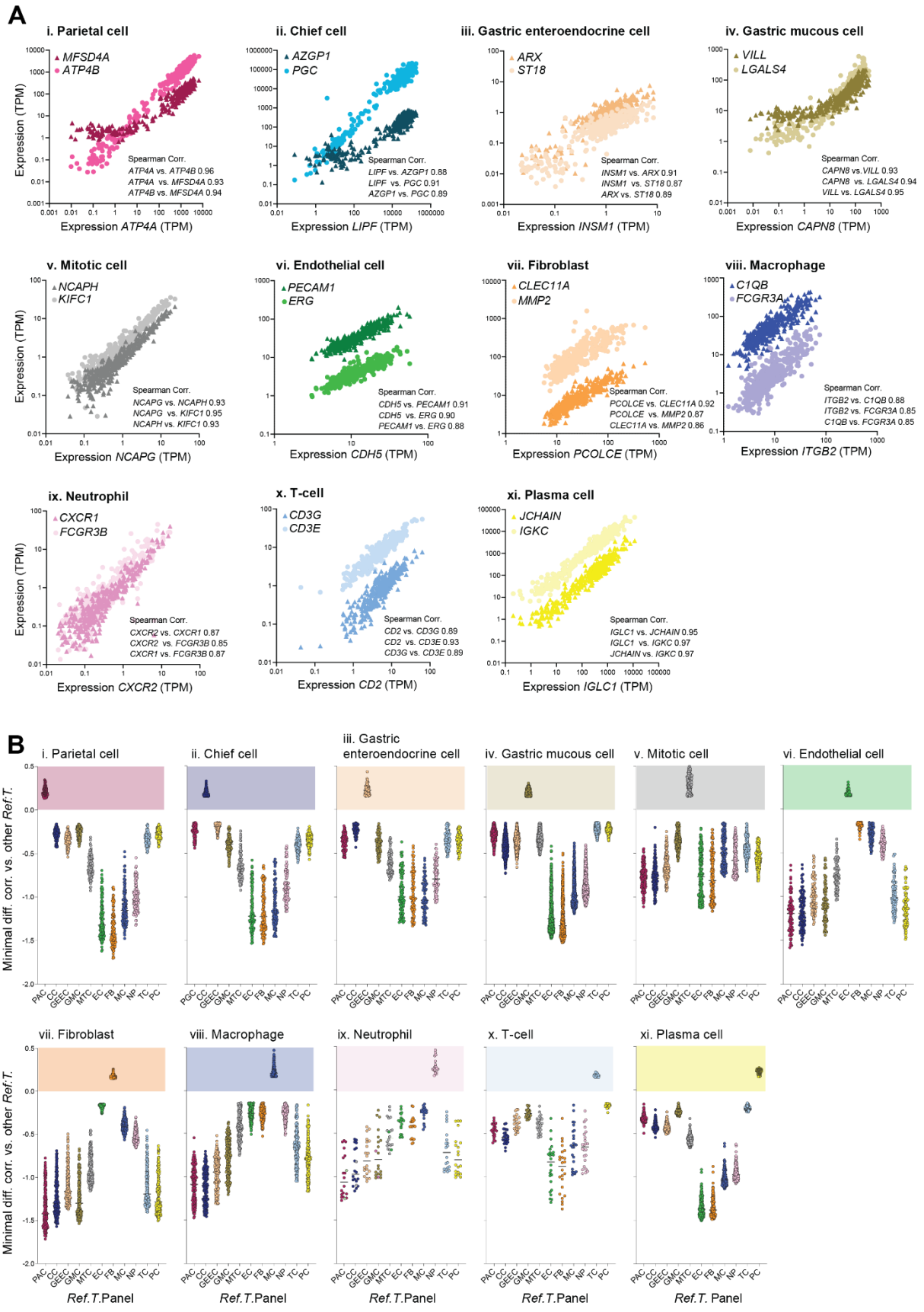

A
i. Parietal cell
MFSD4A
ATP4B
Spearman Corr.
ATP4A vs. ATP4B 0.96
ATP4A vs. MFSD4A 0.93
ATP4B vs. MFSD4A 0.94
Expression (TPM)
Expression ATP4A (TPM)
ii. Chief cell
AZGP1
PGC
Spearman Corr.
LIPF vs. AZGP1 0.88
LIPF vs. PGC 0.91
AZGP1 vs. PGC 0.89
Expression LIPF (TPM)
iii. Gastric enteroendocrine cell
ARX
ST18
Spearman Corr.
INSM1 vs. ARX 0.91
INSM1 vs. ST18 0.87
ARX vs. ST18 0.89
Expression INSM1 (TPM)
iv. Gastric mucous cell
VILL
LGALS4
Spearman Corr.
CAPN8 vs. VILL 0.93
CAPN8 vs. LGALS4 0.94
VILL vs. LGALS4 0.95
Expression CAPN8 (TPM)
v. Mitotic cell
NCAPH
KIFC1
Spearman Corr.
NCAPG vs. NCAPH 0.93
NCAPG vs. KIFC1 0.95
NCAPH vs. KIFC1 0.93
Expression NCAPG (TPM)
vi. Endothelial cell
PECAM1
ERG
Spearman Corr.
CDH5 vs. PECAM1 0.91
CDH5 vs. ERG 0.90
PECAM1 vs. ERG 0.88
Expression CDH5 (TPM)
vii. Fibroblast
CLEC11A
MMP2
Spearman Corr.
PCOLCE vs. CLEC11A 0.92
PCOLCE vs. MMP2 0.87
CLEC11A vs. MMP2 0.86
Expression PCOLCE (TPM)
viii. Macrophage
C1QB
FCGR3A
Spearman Corr.
ITGB2 vs. C1QB 0.88
ITGB2 vs. FCGR3A 0.85
C1QB vs. FCGR3A 0.85
Expression ITGB2 (TPM)
ix. Neutrophil
CXCR1
FCGR3B
Spearman Corr.
CXCR2 vs. CXCR1 0.87
CXCR2 vs. FCGR3B 0.85
CXCR1 vs. FCGR3B 0.87
Expression CXCR2 (TPM)
x. T-cell
CD3G
CD3E
Spearman Corr.
CD2 vs. CD3G 0.89
CD2 vs. CD3E 0.93
CD3G vs. CD3E 0.89
Expression CD2 (TPM)
xi. Plasma cell
JCHAIN
IGKC
Spearman Corr.
IGLC1 vs. JCHAIN 0.95
IGLC1 vs. IGKC 0.97
JCHAIN vs. IGKC 0.97
Expression IGLC1 (TPM)
B
i. Parietal cell
ii. Chief cell
iii. Gastric enteroendocrine cell
iv. Gastric mucous cell
v. Mitotic cell
vi. Endothelial cell
vii. Fibroblast
viii. Macrophage
ix. Neutrophil
x. T-cell
xi. Plasma cell
Minimal diff. corr. vs. other Ref.T.
PAC CC GEEC GMC MTC EC FB MC NP TC PC
Ref.T.Panel

Figure S2. Expression distribution and correlations between human stomach cell type reference transcripts; Related to Figure 1 and Table S1, Tab 1. (**A**) Expression of *Ref.T* selected to represent: (i) parietal cells, (ii) chief cells, (iii) gastric enteroendocrine cells, (iv) gastric mucous cells, (v) mitotic cells, (vi) endothelial cells, (vii) fibroblasts, (viii) macrophages, (ix) neutrophils, (x) T-cells and (xi) plasma cells. (**B**) Minimal differential correlations between mean correlation coefficients with corresponding *Ref.T.* panel for genes above designated thresholds for classification as cell type enriched in: (i) parietal cells, (ii) chief cells, (iii) gastric enteroendocrine cells, (iv) gastric mucous cells, (v) mitotic cells, (vi) endothelial cells, (vii) fibroblasts, (viii) macrophages, (ix) neutrophils, (x) T-cells and (xi) plasma cells, and the mean correlation coefficients with all other *Ref.T.* panels.

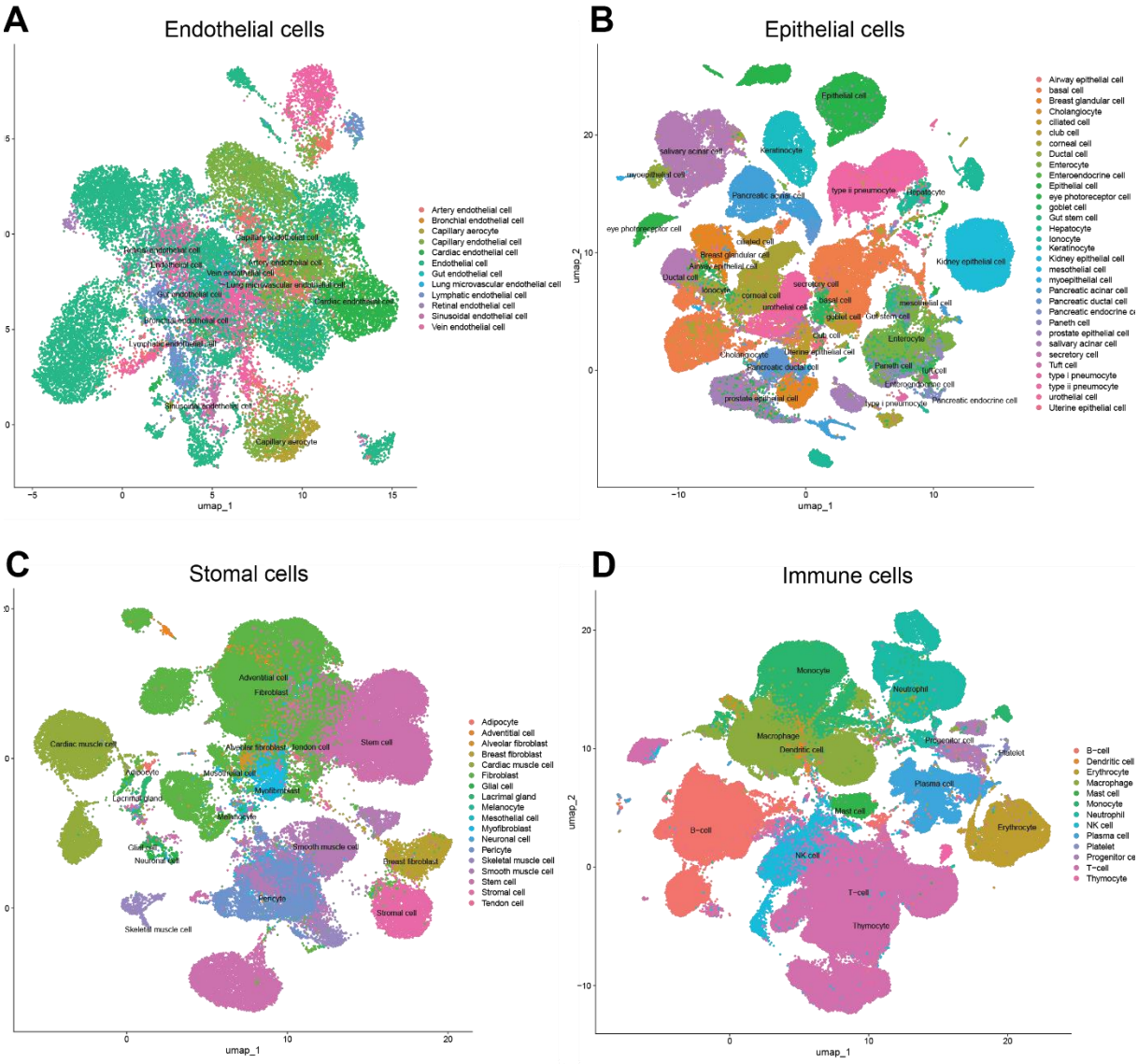


Figure S3. Single cell RNAseq (scRNAseq) annotations; Related to Figure 5, 6 and 7. scRNAseq data was sourced from Tabula Sapiens (Tabula Sapiens et al., 2022). UMAP plots showing original annotations of cell clusters designated as: (**A**) endothelial, (**B**) epithelial, (**C**) stromal or (**D**) immune.

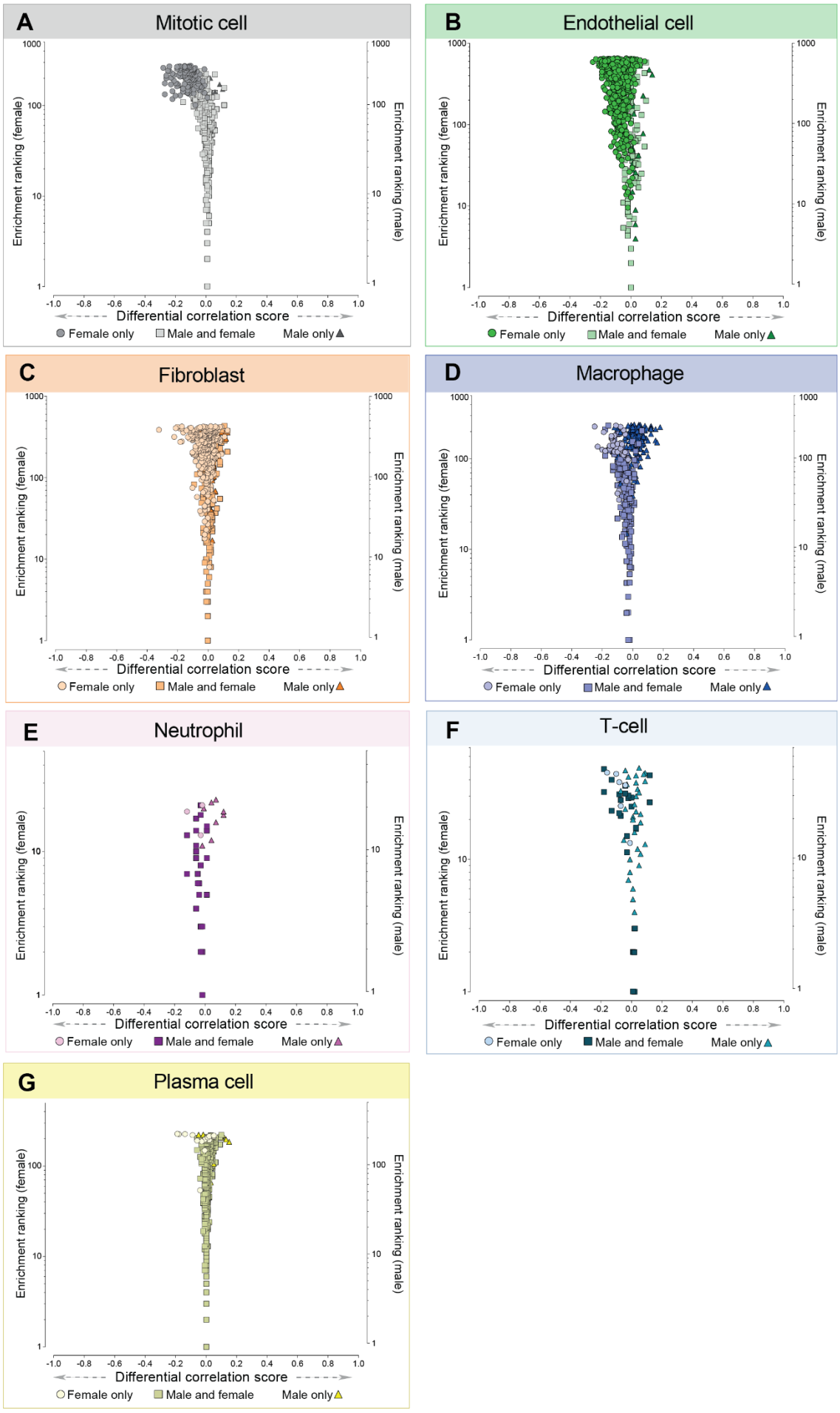

A
Mitotic cell
B
Endothelial cell
C
Fibroblast
D
Macrophage
E
Neutrophil
F
T-cell
G
Plasma cell
Enrichment ranking (female)
Enrichment ranking (male)
Differential correlation score
Female only
Male and female
Male only

Figure S4. Identification of sex-specific cell type enriched genes in human stomach; Related to Figure 7. Human stomach RNAseq data (n=359 individuals) was retrieved from GTEx V8 and divided into female (n=132) and male (n=227) subgroups before classification of cell type-enriched genes. For genes classified as: (**A**) mitotic cell, (**B**) endothelial cell, (**C**) fibroblast, (**D**) macrophage, (**E**) neutrophil, (**F**) T-cell and (**G**) plasma cell enriched, in either female or male subsets, the *´sex differential correlation score'* (difference between mean correlation with the *Ref.T* panel in females vs. males) was plotted vs. 'enrichment ranking' (position in each respective enriched list, highest correlation = rank 1). See also Table S2, Tab 1. On each plot, genes enriched in *both* females and males are represented by common-coloured square symbols, and genes classified as enriched *only* in females or males are represented by differently coloured circle and triangle symbols, respectively. See also Table S2, Tab 1.

Paper II

A human colon cell type transcriptome atlas

S Öling[1], E Struck[1], MN Thorsen[1], M Zwahlen[2], J Odeberg[1,2,3,4], F Pontén[5], M Uhlén[2], P Dusart[1,2,6], LM Butler [1,2,6*]

[1] Department of Clinical Medicine, The Arctic University of Norway, Tromsø, Norway

[2] Science for Life Laboratory, Department of Protein Science, School of Engineering Sciences in Chemistry, Biotechnology and Health, Kungliga Tekniska Högskolan (KTH) Royal Institute of Technology, Stockholm, Sweden

[3] The University Hospital of North Norway (UNN), Tromsø, Norway

[4] Coagulation Unit, Department of Haematology, Karolinska University Hospital, Stockholm, Sweden

[5] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

[6] Clinical Chemistry and Blood Coagulation Research, Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden, *and* Clinical Chemistry, Karolinska University Laboratory, Karolinska University Hospital, Stockholm, Sweden

[*] **Lead contact/corresponding author**:

Dr. L.M Butler, PhD

Email: Lynn.butler@ki.se or lynn.m.butler@uit.no

Key words: Colon, Cell profiling, bulk RNAseq, gene enrichment

SUMMARY

The identification of cell type-specific genes and their modification under different conditions is central to our understanding of human health and disease. The colon, a tubular organ in the lower gastrointestinal tract, absorbs water and electrolytes, produces and absorbs vitamins, and forms and propels feces towards the rectum. In contrast to other sections of the gastrointestinal tract, descriptions of cell type gene enrichment profiles in the colon tend to be lacking from the major single cell sequencing-based atlases. Here, we use an integrative correlation analysis method to predict human sigmoid colon cell type transcriptome signatures using unfractionated colon RNAseq data from 373 individuals. We profile epithelial, enteric neuron, enteric glial, mitotic, endothelial, smooth muscle, fibroblast, macrophage, neutrophil, basophil, T-cell and plasma cells, identifying more than 3000 cell type-enriched transcripts. We uncover the cell type expression profile of several non-coding genes associated with colorectal cancer. Using a sex-based subset analysis, we uncover a small panel of male-only enriched genes. This study contributes to further the understanding of the biology of the human colon.

INTRODUCTION

The colon is a part of the multiple organ system that constitutes the gastrointestinal (GI) tract. The GI tract can be divided into upper and lower sections that have various functions, such as the absorption of nutrients, digestion and reabsorption of water (Choi et al. 2014; de Santa Barbara, van den Brink, and Roberts 2003; Thompson, DeLaForest, and Battle 2018; Kim and Pritts 2017), reflected by the characteristics of the constituent cell types. GI tract epithelium predominantly consists of two functionally different cell types, absorptive and secretory, the specific types and relative abundance of which depend on the location within the system. These cell types constitute the gastrointestinal epithelial lining, forming a selective permeable barrier, preventing unwanted agents from entering the body while allowing nutrients to pass though (Laukoetter, Nava, and Nusrat 2008). The most abundant epithelial cell types in colon are colonocytes and goblet cells (Noah, Donahue, and Shroyer 2011; May and Kaestner 2010), while the entire GI tract has an extensive underlying stromal network including endothelial cells, smooth muscle cells and macrophages.

Characterisation of human organs, and their cell-type specific gene expression profiles, is a cornerstone in the understanding of their biological processes and involvement in disease development; a basis for both the Human Protein Atlas (Uhlén et al. 2015) and Human Cell Atlas (Regev et al. 2017). The Genotype-tissue expression (GTEx) project provides RNAseq data from unfractionated human normal and disease tissue (Consortium 2015). Single-cell RNA sequencing (scRNAseq) technology has made it possible to sequence individual cells, allowing for much finer resolution of gene expression within tissues, but practical and technical challenges exist, such as requirement for fresh tissue, sequencing depth, financial constraints and data interpretation (Gawad, Koh, and Quake 2016; Shapiro, Biezuner, and Linnarsson 2013; Grün and van Oudenaarden 2015). Additionally, artifacts due to sample isolation and processing can be problematic (O'Flanagan et al. 2019). Fragile cell types, such as neurons, are difficult to analyse using standard scRNAseq protocols and instead require nuclear sequencing to avoid aberrant transcription caused by heating or sample digestion (Lacar et al. 2016; Lake et al. 2016). Further, due to the limited number of biological replicates typically

analysed with scRNAseq, there is a risk of false discoveries due to underestimation of biological variance (Squair et al. 2021; Denninger et al. 2022).

The human colon has been studied extensively in the context of colorectal cancer, the third most common cancer type worldwide (Keum and Giovannucci 2019). Most studies have focused on individual cell types involved in colorectal cancer, such as: endothelial cells (Lu et al. 2013; Hong et al. 2009; Teranishi et al. 2007; Ishigami et al. 1998), T-cells (Di et al. 2020) and macrophages (Bailey et al. 2007; J.-C. Kang et al. 2010; Erreni, Mantovani, and Allavena 2011), or specific tissue regions such as sites of metastases (Leung et al. 2017; Yunbin Zhang et al. 2020) and mucosa (Díez-Obrero et al. 2021) as well as genomic alterations (Bian et al. 2018). Bulk sequencing studies of colorectal cancer have identified genetic and genomic alterations (Han et al. 2013; Yaeger et al. 2018; Pira et al. 2020), however these studies do not identify cell type specific changes. scRNAseq studies on colon have focused on the epithelium (Wang et al. 2020; Burclaff et al. 2022), macrophages (Domanska et al. 2022), neuron subtypes in colon (Hockley et al. 2019), changes in gene expression during inflammatory bowel disease (Smillie et al. 2019; Serigado et al. 2022; Kong et al. 2023; Kanke et al. 2022) and tumour profiles (Dalerba et al. 2011; Huipeng Li et al. 2017; Zhou, Guo, and Wang 2022).

Here, we analysed unfractionated RNAseq data from 373 human sigmoid colon samples to identify over 3000 cell type enriched genes, using our previously described integrative correlation analysis method (Norreen-Thorsen et al. 2022; Butler et al. 2016; Dusart et al. 2019). Enteric glial cells had the highest number of predicted protein-coding and non-coding enriched transcripts. We identified a high global similarity in cell type profiles between male and female samples, as well as a small panel of male-only cell type enriched genes. A number of non-coding transcripts with predicted cell type enrichment have been previoulsy associated with cancer progression.

RESULTS

Identification of cell type transcriptome profiles in sigmoid colon

Cell type reference transcripts correlate across unfractionated colon RNAseq data

Identification of colon cell type-enriched transcriptome profiles was performed using a method based on our previous work (Butler et al. 2016; Dusart et al. 2019; Norreen-Thorsen et al. 2022). Briefly, bulk RNAseq data from unfractionated sigmoid colon tissue (n=373 samples) was acquired from the Genotype-Tissue Expression (GTEx) portal V8 (https://gtexportal.org) (Consortium 2015) (Figure 1A). A list of candidate 'virtual markers' were sourced for each cell type based on: (i) in-house protein profiling (Gremel et al. 2015; Uhlen et al. 2015), (ii) single-cell sequencing data (Karlsson et al. 2021) and (iii) collated databases, e.g. Cell Marker (X. Zhang et al. 2019) and PanglaoDB (Franzen, Gan, and Bjorkegren 2019) (Figure 1B). We then selected a panel of three established cell-type specific markers as 'reference transcripts' (*Ref.T.*) for each major constituent cell type (Figure 1C). The panels were chosen based on the following criteria: (i) a high mean correlation within each cell type panel (indicating cell type co-expression) (Figure 1C and Figure S1A), (ii) a low correlation between *Ref.T.* across the different cell type panels (indicating cell type specificity) (Figure 1C) and (iii) a normal distribution of *Ref.T*. expression across the samples (Figure S1). The calculated mean correlations and standard deviations for each *Ref.T*. panel were; Epithelial cells (EP) [*EPCAM, SLC44A4, PHGR1* 0.88±0.005], enteric neuron cells (ENC) [*INA, RTN1, SCG3* 0.95±0.005], enteric glial cells (EGC) [*SLC35F1, NRXN1, L1CAM* 0.93±0.009], mitotic cells (MTC) [*HMMR, TOP2A, RMM2* 0.74±0.01], endothelial cells (EC) [*ESAM, CDH5, MMRN2* 0.92±0.005], smooth muscle cells (SMC) [*ACTG2, TAGLN, TPM1* 0.81±0.054], fibroblasts (FB) [*FBLN1, DCN, TNXB* 0.79±0.054], macrophages (MC) [*C1QB, CD68, ITGB2* 0.82±0.034], neutrophils (NP) *[S100A8, AQP9, FCAR* 0.73±0.007], basophils (BP) [*CPA3, TPSAB1, SIGLEC6* 0.89±0.043], T-cells (TC) [*CD2, CD6, CD3E* 0.82±0.032] and plasma cells (PC) [*JCHAIN, IGKC, IGHA1* 0.9±0.029] (Table S1).

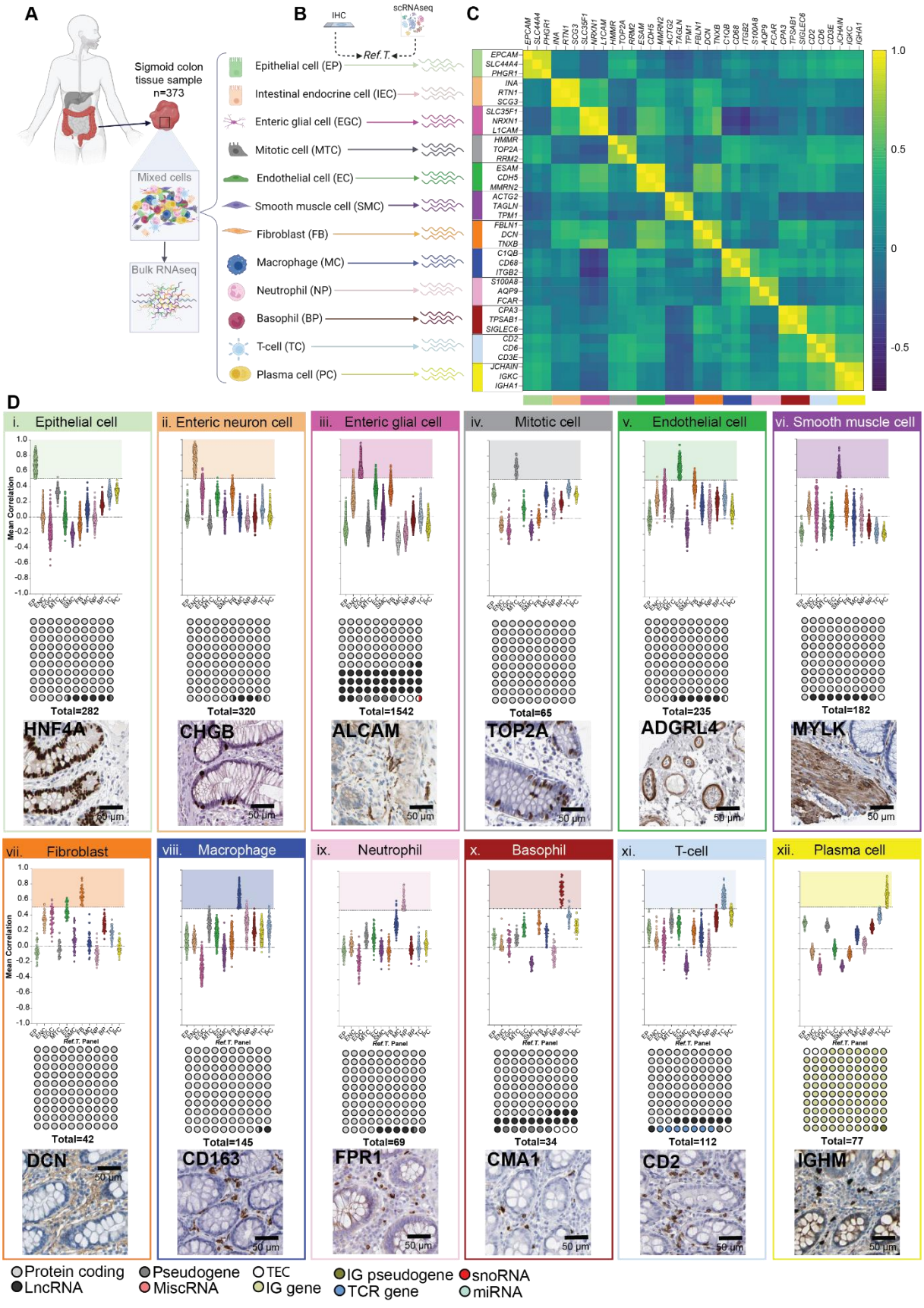

A
Sigmoid colon tissue sample n=373
Mixed cells
Bulk RNAseq
B
IHC
scRNAseq
Ref. T.
Epithelial cell (EP)
Intestinal endocrine cell (IEC)
Enteric glial cell (EGC)
Mitotic cell (MTC)
Endothelial cell (EC)
Smooth muscle cell (SMC)
Fibroblast (FB)
Macrophage (MC)
Neutrophil (NP)
Basophil (BP)
T-cell (TC)
Plasma cell (PC)
C
EPCAM
SLC44A4
PHGR1
INA
RTN1
SCG3
SLC35F1
NRXN1
L1CAM
HMMR
TOP2A
RRM2
ESAM
CDH5
MMRN2
ACTG2
TAGLN
TPM1
FBLN1
DCN
TNXB
C1QB
CD68
ITGB2
S100A8
AQP9
FCAR
CPA3
TPSAB1
SIGLEC6
CD2
CD6
CD3E
JCHAIN
IGKC
IGHA1
1.0
0.5
0
-0.5
D
i. Epithelial cell
ii. Enteric neuron cell
iii. Enteric glial cell
iv. Mitotic cell
v. Endothelial cell
vi. Smooth muscle cell
Mean Correlation
Total=282
Total=320
Total=1542
Total=65
Total=235
Total=182
HNF4A
CHGB
ALCAM
TOP2A
ADGRL4
MYLK
50 µm
vii. Fibroblast
viii. Macrophage
ix. Neutrophil
x. Basophil
xi. T-cell
xii. Plasma cell
Ref. T. Panel
Total=42
Total=145
Total=69
Total=34
Total=112
Total=77
DCN
CD163
FPR1
CMA1
CD2
IGHM
Protein coding
LncRNA
Pseudogene
MiscRNA
TEC
IG gene
IG pseudogene
TCR gene
snoRNA
miRNA

Figure 1. Resolving constituent cell type identities from unfractionated sigmoid colon tissue RNAseq data using integrative co-expression analysis. (A) bulk RNAseq data from 373 sigmoid colon tissue samples were retrieved from GTEx V8. Each tissue sample contained a mixture of the constituent cell types and contributed differing proportions of mRNA. **(B)** Candidate marker genes ('reference transcripts [Ref.T.]) were identified for each cell type, based on in house tissue protein profiling and available datasets. **(C)** Matrix of correlation coefficients between the selected Ref.T. across the sample set. **(D)** Mean correlation coefficients of genes above designated thresholds for classification as cell-type enriched, with dot plots showing enriched gene classifications and associated protein profiling in (i) epithelial cells [EP], (ii) intestinal endocrine cells [IEC], (iii) Enteric glial cells [EGC], (iv) Mitotic cells [MTC], (v) Endothelial cells [EC], (vi) smooth muscle cell [SMC], (vii) Fibroblasts [FB], (viii) Macrophages [MC], (ix) Neutrophils [NP], (x) Basophil [BP], (xi) T-cells [TC] and (xii) plasma cells [PC]. Scale bar 50 µm.

Reference transcript analysis to identify colon cell type-enriched genes

We performed a full reference transcript-based analysis on the colon RNAseq data to produce correlation values between each selected *Ref.T.* and all other sequenced transcripts within the GTEx data (n=56,200). Inter-sample cell type proportion varies, but the ratio between transcripts in the same cell type should be consistent, therefore a correlation coefficient with a high mean value with the *Ref.T.* panels for a specific cell type should indicate enrichment of the gene(s) in that cell type. A list of enriched genes was generated for each cell type (Figure 1D i-xii, top, Table S1, Tab 2) based on the following enrichment criteria: (i) the gene should have a mean corr. >0.5 with the cell-type *Ref.T.* panel – indicated by a dashed horizontal line (Figure 1D) and (ii) the differential between this mean corr. value and the maximum mean corr. value with any other *Ref.T.* panel should be >0.15. This excluded genes that were potentially co-enriched in two or more cell types, as we previously described (Norreen-Thorsen et al. 2022).

Colon cell type enriched gene signatures

A total of 3105 transcripts were predicted to be cell type-enriched in colon cell types. The individual cell types with the highest number of enriched genes were found in tissue specific cell types: epithelial cells (n=282), enteric neuron cells (n=320) and enteric glial cells (n=1542). The fewest enriched genes were found in neutrophils (n=69), fibroblasts (n=42) and basophils (n=34) (Figure 1D). In almost all cell-types the majority of enriched genes were protein-coding, with the exception of plasma cells in which immunoglobulin (IG) genes was most common (Figure 1D xii, middle) (Yates et al. 2020). Amongst the non-coding enriched genes, lncRNA was the most common classification, except in plasma cells where IG pseudogene was the most common (Figure 1D xii, middle). Protein profiling of selected proteins expressed by cell type-enriched transcripts showed consistent staining with cell type classification (Figure 1D, bottom).

Alternative analysis method and protein profiling support cell type-enriched classifications

Colon cell type enrichment signatures

We extracted up to the top 50 most enriched protein coding transcripts for each cell type, i.e., those with the highest correlation value with the corresponding *Ref.T.* panel, and plotted circular graphs of the mean corr. value, differential corr. value and expression level (mean TPM) in the bulk RNAseq dataset. Mitotic cell-enriched genes (Figure 2A i) had comparatively low TPM expression values, in comparison with epithelial cell (Figure 2B i), endothelial cell (Figure 2E i) and smooth muscle cell-enriched genes (Figure 2D i). Gene ontology analysis of the predicted cell type enriched genes (Ashburner et al. 2000; Gene Ontology 2021)(Table S1, Tab 3-14) revealed over represented terms consistent with known cell functions. For example, analysis of the mitotic-enriched genes resulted in GO terms consistent with cell cycle function (Figure 2A ii) such as: *'mitotic cell cycle process*' (FDR $1.78x10^{-46}$), '*mitotic nuclear division'* (FDR $9.38x10^{-33}$) and '*nuclear division*' (FDR $1.25x10^{-38}$) (Figure 2A ii), for epithelial cell genes most significantly enriched terms included *´epithelial cell differentiation´* (FDR $1.10x10^{-11}$), '*epithelium development'* (FDR $2.87x10^{-10}$) and '*epithelial cell development'* (FDR $8.33x10^{-8}$) (Figure 2B ii), and GO terms for endothelial cell genes were vasculature related: '*angiogenesis*' (FDR $5.35x10^{-21}$), '*vasculature development'* (FDR $9.18x10^{-20}$) and '*blood vessel development*' (FDR $1.30x10^{-19}$) (Figure 2C ii) and for smooth muscle cells most significantly enriched terms included '*actin filament-based process*' (FDR $4.45x10^{-20}$), '*cytoskeleton organization'* (FDR $1.18x10^{-18}$) and '*muscle structure development*' (FDR $1.54x10^{-10}$) (Figure 2D ii). Protein profiling showing staining consistent with the respective cell type enrichment predictions in mitotic cells (Figure 2A iii), epithelial cells (Figure 2B iii), endothelial cells (Figure 2C iii) and smooth muscle cells (Figure 2D iii).

Immune cell enriched genes (Figure 3 A-E) showed high expression variability, with T-cell enriched genes having the lowest (Figure 3D i) and plasma cell enriched genes having the highest TPM expression values (Figure 3E i). Indeed, within all cell types, the range of

expression values varies, reflecting individual gene variations in regulation and stability. GO analysis revealed over represented terms for predicted colon immune cell type enriched genes were also consistent with known cell functions e.g., for macrophages ´defence response´(FDR $2.30x10^{-29}$) and ´immune response' (FDR $1.50x10^{-29}$) (Figure 3A ii), for neutrophils ´leukocyte chemotaxis´(FDR $1.68x10^{-11}$) and ´neutrophil chemotaxis' (FDR $2.09x10^{-9}$) (Figure 3B ii), for basophils 'angiotensin maturation´(FDR $1.79x10^{-7}$) and 'regulation of leukocyte activation' (FDR $7.22x10^{-7}$) (Figure 3C ii), for T-cell ´lymphocyte activation' (FDR $8.69x10^{-39}$) and ´T-cell activation´(FDR $2.00x10^{-38}$) (Figure 3D ii), for plasma cells ´adaptive immune response' (FDR $1.45x10^{-50}$) and ´immunoglobulin production´(FDR $2.94x10^{-45}$) (Figure 3 E ii). Protein profiling of the different immune cell types showed staining consistent with predicted enrichment profiles in macrophages (Figure 3A iii), neutrophils (Figure 3B iii), basophils (Figure 3C iii), T-cells (Figure 3D iii), plasma cells (Figure 3E iii).

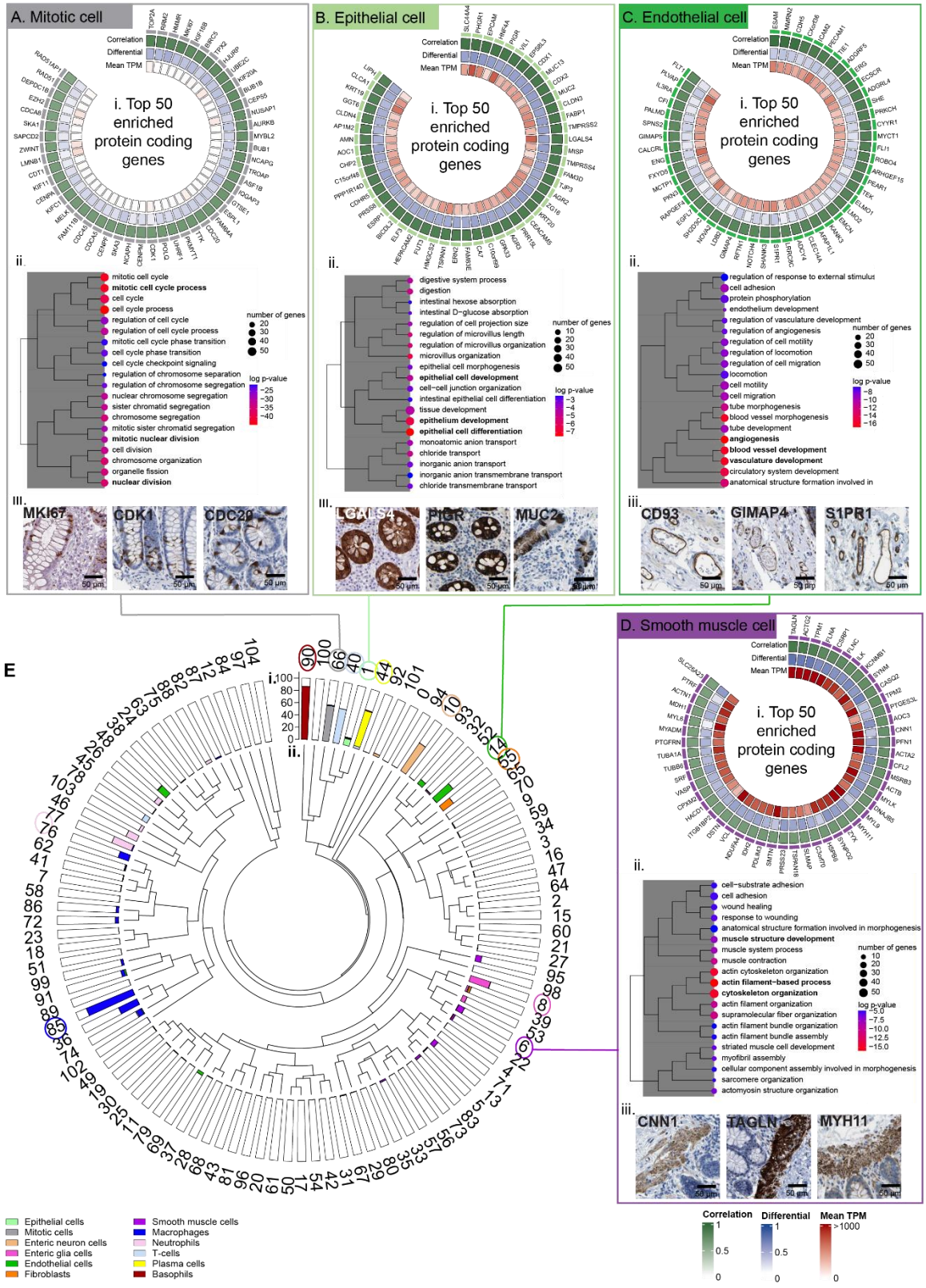

A. Mitotic cell
B. Epithelial cell
C. Endothelial cell
D. Smooth muscle cell
E
i. Top 50 enriched protein coding genes
Correlation
Differential
Mean TPM
ii.
number of genes
log p-value
mitotic cell cycle
mitotic cell cycle process
cell cycle
cell cycle process
regulation of cell cycle
regulation of cell cycle process
mitotic cell cycle phase transition
cell cycle phase transition
cell cycle checkpoint signaling
regulation of chromosome separation
regulation of chromosome segregation
nuclear chromosome segregation
sister chromatid segregation
chromosome segregation
mitotic sister chromatid segregation
mitotic nuclear division
cell division
chromosome organization
organelle fission
nuclear division
digestive system process
digestion
intestinal hexose absorption
intestinal D-glucose absorption
regulation of cell projection size
regulation of microvillus length
regulation of microvillus organization
microvillus organization
epithelial cell morphogenesis
epithelial cell development
cell–cell junction organization
intestinal epithelial cell differentiation
tissue development
epithelium development
epithelial cell differentiation
monoatomic anion transport
chloride transport
inorganic anion transport
inorganic anion transmembrane transport
chloride transmembrane transport
regulation of response to external stimulus
cell adhesion
protein phosphorylation
endothelium development
regulation of vasculature development
regulation of angiogenesis
regulation of cell motility
regulation of locomotion
regulation of cell migration
locomotion
cell motility
cell migration
tube morphogenesis
blood vessel morphogenesis
tube development
angiogenesis
blood vessel development
vasculature development
circulatory system development
anatomical structure formation involved in
cell–substrate adhesion
cell adhesion
wound healing
response to wounding
anatomical structure formation involved in morphogenesis
muscle structure development
muscle system process
muscle contraction
actin cytoskeleton organization
actin filament–based process
cytoskeleton organization
actin filament organization
supramolecular fiber organization
actin filament bundle organization
actin filament bundle assembly
striated muscle cell development
myofibril assembly
cellular component assembly involved in morphogenesis
sarcomere organization
actomyosin structure organization
iii.
MKI67
CDK1
CDC20
CA4
PIGR
MUC2
CD93
GIMAP4
S1PR1
CNN1
TAGLN
MYH11
50 µm
Epithelial cells
Mitotic cells
Enteric neuron cells
Enteric glia cells
Endothelial cells
Fibroblasts
Smooth muscle cells
Macrophages
Neutrophils
T-cells
Plasma cells
Basophils

Figure 2. Protein coding gene signatures of human sigmoid colon cell types. Protein coding gene enrichment signatures of human sigmoid colon samples for: **(A)** mitotic cells, **(B)** epithelial cells, **(C)** endothelial cells and **(D)** smooth muscle cells with plots of (i) top 50 enriched protein-coding transcripts showing the correlation coefficient with the cell type Ref.T., differential correlation value and mean expression in bulk RNAseq, (ii) over-represented gene ontology terms among genes predicted to be cell type-enriched and (iii) human colon tissue profiling for proteins encoded by genes classified as cell type-enriched. Scale bar 50 µm. **(E)** RNAseq data for 373 unfractionated human sigmoid colon samples were subjected to weighted correlation network analysis (WGCNA). Colored circles around clusters indicate corresponding cell type Ref.T. positions on dendrogram. The colored bars show distribution of transcripts predicted to be cell type-enriched across the dendrogram clusters.

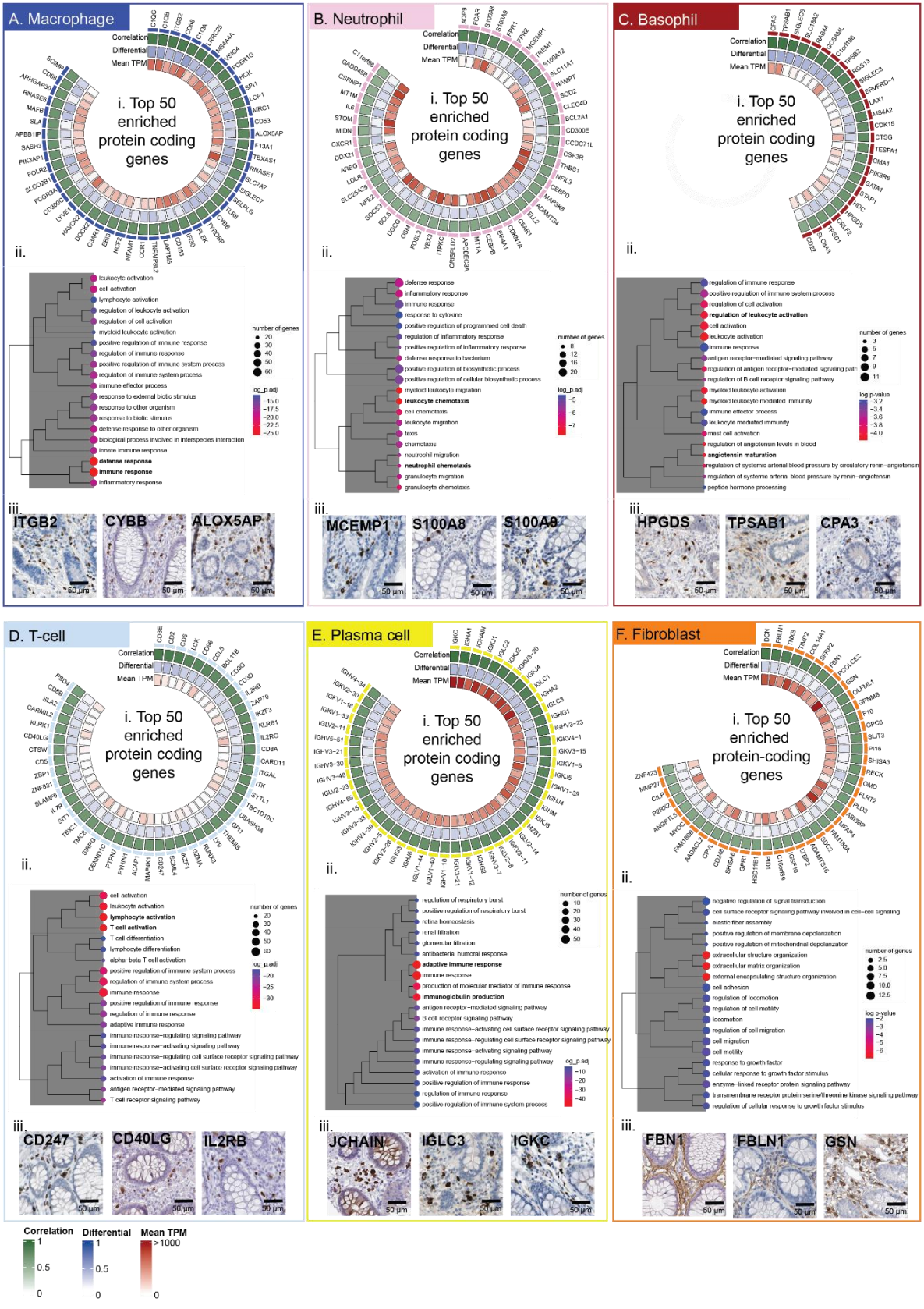

A. Macrophage
i. Top 50 enriched protein coding genes
Correlation
Differential
Mean TPM
ii.
iii.
ITGB2
CYBB
ALOX5AP
B. Neutrophil
i. Top 50 enriched protein coding genes
ii.
iii.
MCEMP1
S100A8
S100A9
C. Basophil
i. Top 50 enriched protein coding genes
ii.
iii.
HPGDS
TPSAB1
CPA3
D. T-cell
i. Top 50 enriched protein coding genes
ii.
iii.
CD247
CD40LG
IL2RB
E. Plasma cell
i. Top 50 enriched protein coding genes
ii.
iii.
JCHAIN
IGLC3
IGKC
F. Fibroblast
i. Top 50 enriched protein-coding genes
ii.
iii.
FBN1
FBLN1
GSN
Correlation
Differential
Mean TPM

Figure 3. Protein coding gene signatures of human sigmoid colon immune cell types. Cell type-enriched protein coding transcripts in: (A) macrophage, (B) neutrophil, (C) basophil, (D) T-cell, (E) plasma cell and (F) fibroblast with plots of (i) top 50 enriched protein-coding transcripts showing the correlation coefficient with the cell type Ref.T., differential correlation value and mean expression in bulk RNAseq, (ii) over-represented gene ontology terms among genes predicted to be cell type-enriched and (iii) human colon tissue profiling for proteins encoded by genes classified as cell type-enriched. Scale bar 50 µm.

Unsupervised weighted network correlation analysis of colon (WGCNA)

As the *Ref.T.* analysis is based on manual selection of *Ref.T.* panels it can be subject to input bias. To determine the reliability of our results we subjected the same GTEx colon RNAseq dataset to a different analysis method – weighted correlation network analysis (WGCNA) (Langfelder and Horvath 2008), a method that can compute relationships between genes with similar correlations, without manual input of target reference genes. WGCNA generates correlation values between all transcripts and then clusters similar transcripts together, based on expression similarity. WGCNA analysis of the colon RNAseq data resulted in 104 distinct clusters (Figure 2E*),* and the genes that we had selected as *Ref.T.* for specific cell types consistently found in the same cluster, e.g., mitotic cell *Ref.T.s* HMMR, TOP2A and RMM2 in cluster 66 and endothelial cell *Ref.T.s* ESAM, CDH5 and MMRN2 in cluster 14. The majority of genes identified as cell type enriched in our analysis were clustered into the same, or related groups in the WGNCA (Figure 2E).

Identification of an enteric glial cell-enriched transcriptome profile

Enteric glial cells (EGCs) are a key component of the enteric nervous system (ENS), with potential roles in neuron survival, immune system function, and the development of several immunological disorders, such as inflammatory bowel and celiac disease (Liu and Yang 2022)**.** Enteric glial cells had a large number of predicted cell type enriched protein-coding transcripts (n=904) and non-coding (n=638) transcripts. We extracted the top 50 most enriched protein coding and non-coding transcripts, as ranked by correlation value with the corresponding *Ref.T.* panel, and plotted mean corr. value, differential corr. value and expression level (mean TPM) in the bulk RNAseq dataset (Figure 4A). To compare the expression of these enteric glial cell enriched genes with expression levels in different human organs, we sourced GTEx bulk RNAseq expression data across multiple tissues (Figure 4B). Among the top enteric glial cell enriched protein-coding genes were *SORCS1*, *COL28A1* and *SHISA9,* all of which had elevated expression in colon, or other gastrointestinal organs vs. other tissue types (Figure 4B i-iii), consistent with a gastrointestinal-specific function. Additionally, *SHISA9* showed elevated expression in brain tissue (Figure 4B i-ii), indicating that this gene could have a general glial

cell function, beyond the gastrointestinal tract. Protein-coding and non-coding enriched enteric glial transcripts were analysed by gene ontology analysis (Ashburner et al. 2000; Gene Ontology 2021) (Figure 4C). The analysis revealed top enriched GO terms consistent with enteric glial cell functions such as: ´*neurogenesis*' (FDR $1.10x10^{-9}$), ´*generation of neurons*´ (FDR $5.34x10^{-9}$) and ´*cell morphogenesis involved in neuron differentiation'* (FDR $1.20x10^{-8}$). Protein profiling of selected enteric glial cell-enriched transcripts showed consistent cell-type staining in colon tissue (Figure 4D). Several of the non-coding transcripts also showed elevated expression in colon tissue (both sigmoid and transverse) compared to other tissues (Figure 4E i, ii, v, vi), consistent with a specialised function in this organ. Additionally, several transcripts showed elevated expression in brain regions (Figure 4E ii-vi), consistent with the neuronal maintenance function of enteric glial cells. Furthermore, some transcripts showed elevated expression in multiple organs (Figure 4E iv), which could indicate a broader function.

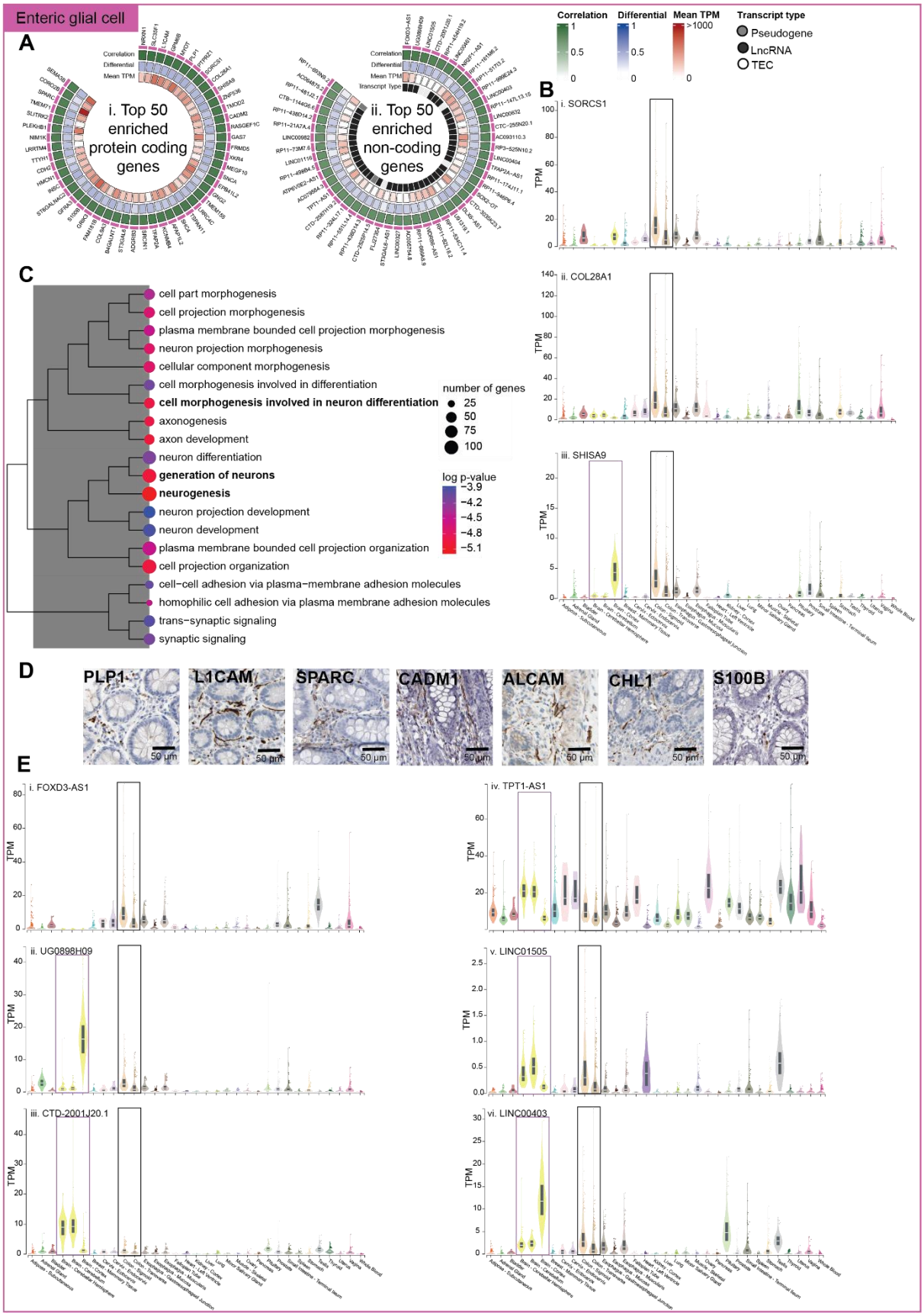

Enteric glial cell
A
i. Top 50 enriched protein coding genes
ii. Top 50 enriched non-coding genes
Correlation
Differential
Mean TPM
Transcript type
Pseudogene
LncRNA
TEC
B
i. SORCS1
ii. COL28A1
iii. SHISA9
TPM
C
cell part morphogenesis
cell projection morphogenesis
plasma membrane bounded cell projection morphogenesis
neuron projection morphogenesis
cellular component morphogenesis
cell morphogenesis involved in differentiation
cell morphogenesis involved in neuron differentiation
axonogenesis
axon development
neuron differentiation
generation of neurons
neurogenesis
neuron projection development
neuron development
plasma membrane bounded cell projection organization
cell projection organization
cell−cell adhesion via plasma−membrane adhesion molecules
homophilic cell adhesion via plasma membrane adhesion molecules
trans−synaptic signaling
synaptic signaling
number of genes
log p-value
D
PLP1
L1CAM
SPARC
CADM1
ALCAM
CHL1
S100B
50 µm
E
i. FOXD3-AS1
ii. UG0898H09
iii. CTD-2001J20.1
iv. TPT1-AS1
v. LINC01505
vi. LINC00403

Figure 4: Gene enrichment signatures for human sigmoid colon enteric glial cells. (A) Protein-coding (i) and non-coding (ii) gene enrichment signatures for enteric glial cells showing the correlation coefficient with the Ref.T. panel, differential score, mean expression value in bulk RNAseq and additionally in (ii) the non-coding transcript type. **(B)** Expression of protein-coding genes classified as enteric glial-enriched in bulk RNAseq in different human organs (i) SORCS1 (ii) COL25A1 and (iii) SHISHA9. **(C)** over-represented gene ontology terms among genes predicted to be cell type-enriched. **(D)** Human colon protein profiling of transcripts predicted to be cell-type enriched. Scale bar 50 µm. **(E)** Expression of non-coding genes classified as enteric glial-enriched (i) FOXD3-AS1, (ii) UG0898H09, (iii) CTD-2001J20.1, (iv) TPT1-AS1, (v) LINC01505 and (v) LINC00403.

Identification of enteric neuron cell-enriched transcriptome profile

Enteric neuron cells (ENC) constitute the elaborate network of neuron cells within the enteric nervous system (ENS) that, together with EGC, facilitate communication within the gastrointestinal tract (Spencer and Hu 2020; Schneider, Wright, and Heuckeroth 2019). We extracted the top 50 most enriched protein-coding transcripts from the 306 protein coding genes we defined as enriched within ENCs, and plotted mean corr. value, differential corr. value and expression level (mean TPM) (Figure 5A i). ENC classifications were supported by protein profiling in colon tissue (Figure 5A ii). Gene ontology analysis of the complete list of protein-coding and non-coding enriched ENC transcripts (Ashburner et al. 2000; Gene Ontology 2021) (Figure 5A iii) revealed top enriched GO terms consistent with ENC function such as: *´trans-synaptic signaling´* (FDR $2.82x10^{-38}$), *neuron development'* (FDR $6.24x10^{-21}$) and *'cell-cell signaling'* (FDR $1.93x10^{-28}$). Of the non-coding genes that were predicted to be ENC-enriched (n=14) (Figure 5B i), the majority were classified as lncRNA (n=10).

Numerous ENC enriched protein coding (Figure 5A iv) and non-coding (Figure 5B ii) genes were elevated in brain tissues, compared to other tissue types, consistent with neuronal functions.

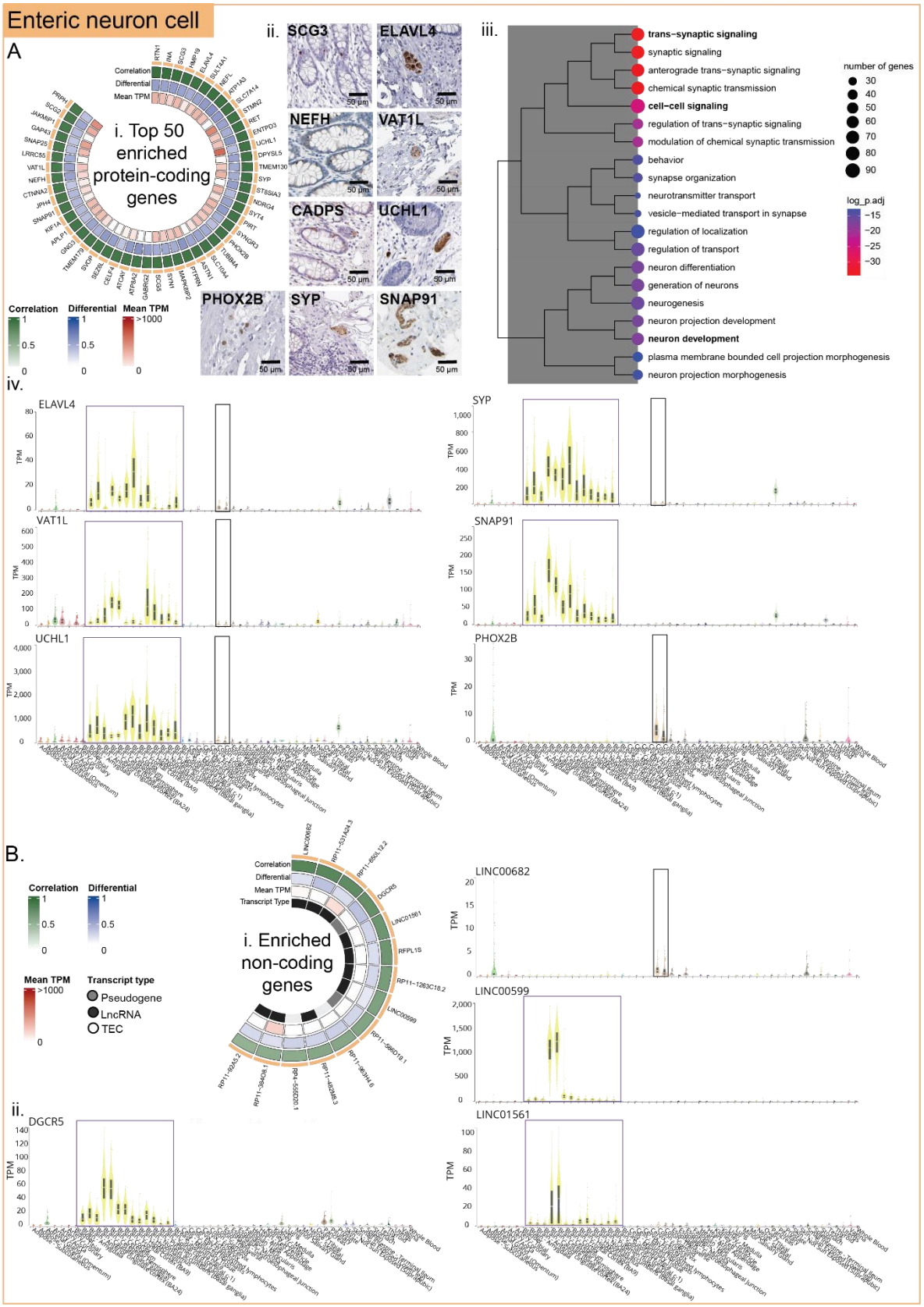

Enteric neuron cell
A
i. Top 50 enriched protein-coding genes
Correlation
Differential
Mean TPM
ii.
SCG3
ELAVL4
NEFH
VAT1L
CADPS
UCHL1
PHOX2B
SYP
SNAP91
50 μm
iii.
trans−synaptic signaling
synaptic signaling
anterograde trans−synaptic signaling
chemical synaptic transmission
cell−cell signaling
regulation of trans−synaptic signaling
modulation of chemical synaptic transmission
behavior
synapse organization
neurotransmitter transport
vesicle−mediated transport in synapse
regulation of localization
regulation of transport
neuron differentiation
generation of neurons
neurogenesis
neuron projection development
neuron development
plasma membrane bounded cell projection morphogenesis
neuron projection morphogenesis
number of genes
log_p.adj
iv.
ELAVL4
VAT1L
UCHL1
SYP
SNAP91
PHOX2B
TPM
B.
i. Enriched non-coding genes
Transcript type
Pseudogene
LncRNA
TEC
LINC00682
LINC00599
LINC01561
ii.
DGCR5

Figure 5: Enrichment signatures for human sigmoid colon enteric neuron cells. (A) Protein-coding (i) gene enrichment signatures for enteric neuron cells showing the correlation coefficient with the Ref.T. panel, differential score, mean expression value in bulk RNAseq, (ii) human colon protein profiling of transcripts predicted to be enteric neuron cell enriched, (iii) shows plots over-represented gene ontology terms among genes predicted to be cell type-enriched. (iv) Expression of protein-coding genes classified as enteric neuron enriched in various human organs ELAVL4, VAT1L, UCHL1, SYP, SNAP91 and PHOX2B. **(B)** (ii) gene enrichment signatures for transcripts classified as non-coding in enteric neuron cells showing the correlation coefficient with the Ref.T. panel, differential score, mean expression value in bulk RNAseq as well as non-coding transcript-type. (ii) Expression of non-coding genes classified as enteric neuron-enriched in bulk RNAseq in different human organs LINC00682, LINC00599, DGCR5 and LINC01561.

Cell-type enriched non-coding transcripts

In addition to the non-coding transcripts predicted to be enriched in EGC (n=638) and ENC (n=14) (Table S1, Tab 2), a further 68 were classified as cell-type enriched in the other cell types (Table S1, Tab 2), including epithelial cells (n=16) (Figure 6A i), endothelial cells (n=15) (Figure 6B i), T-cells (n=11) (Figure 6C i), basophils (n=8) (Figure 6D i) and plasma cells (n=1) (Figure 6E i). No non-coding transcripts were enriched in mitotic cells or fibroblasts. Expression of cell enriched non-coding transcripts was generally highest in epithelial cells (Figure 6A). Unlike protein-coding transcripts, it is not possible to verify cell type expression profile of non-coding transcripts with protein profiling. As an alternative, we used scRNAseq data from Tabula Sapiens (Tabula Sapiens et al. 2022) generated from large intestine. The Tabula Sapiens dataset lacked several cell types that we profiled, but verification of five cell types was possible (Figure 6). We generated UMAP plots for the Tabula Sapiens large intestine dataset to compare the cell expression profile of representative non-coding transcripts from each cell type present in both datasets (see Figure S2 for cell culture annotations). Non-coding epithelial cell enriched transcripts *RP11-465B22.8* and *RP11-395B7.2* (Figure 6A ii and iii), endothelial cell transcripts *SENCR* and *GATA2-AS1* (Figure 6B ii and iii), T-cell enriched transcripts *LINC00861* and *RP11-326C3.2* (Figure 6C ii and iii), basophil enriched transcripts *LINC01835* and *RP11-354E11.2* (Figure 6D ii and iii), and the plasma cell enriched transcript *IGHGP* (Figure 6D ii), all showed enrichment within corresponding cell type clusters in the Tabula Sapiens single cell data. Thus, the Tabula sapiens scRNAseq data provides supportive evidence for our non-coding cell type classifications.

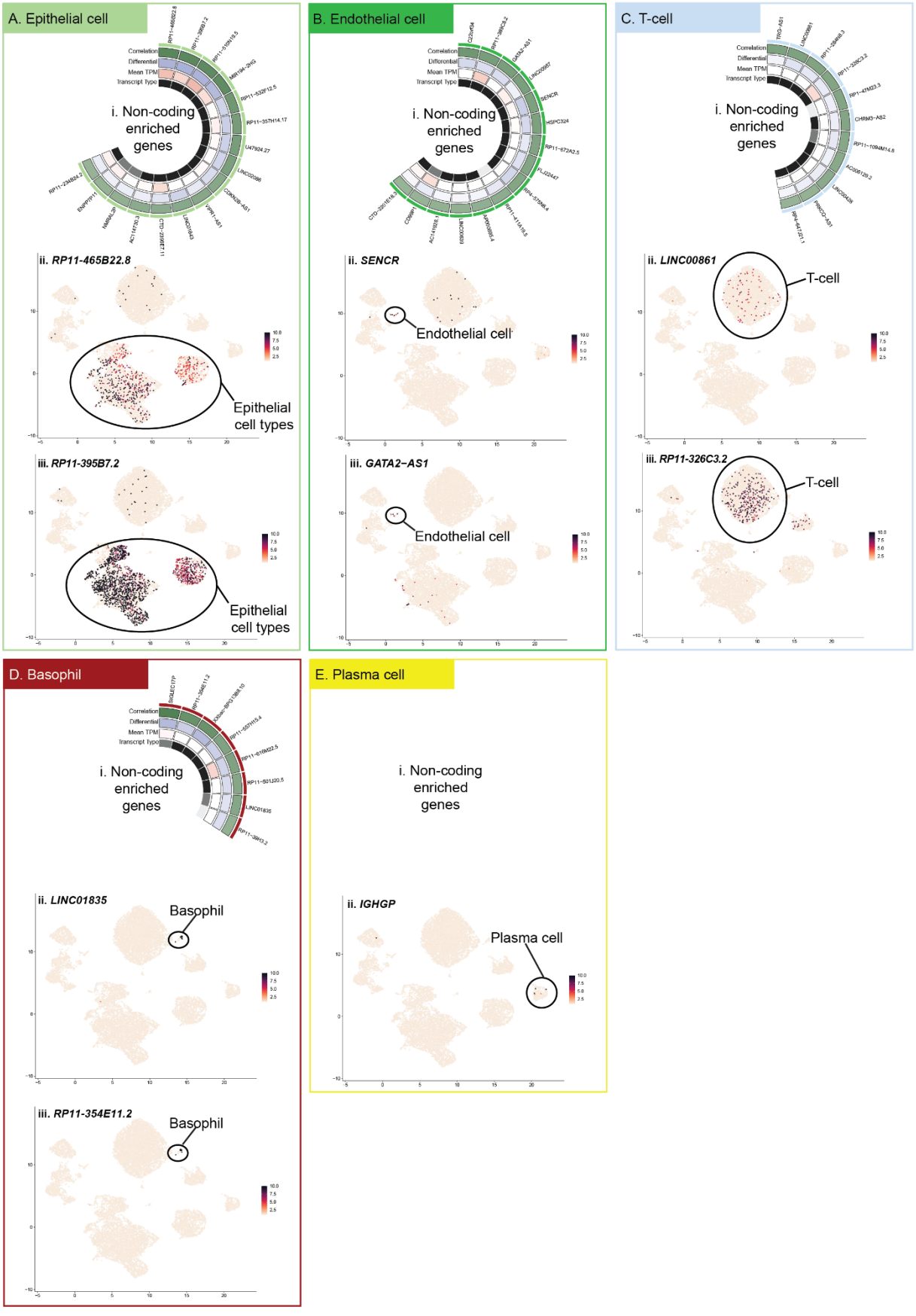

A. Epithelial cell
i. Non-coding enriched genes
ii. RP11-465B22.8
Epithelial cell types
iii. RP11-395B7.2
Epithelial cell types
B. Endothelial cell
i. Non-coding enriched genes
ii. SENCR
Endothelial cell
iii. GATA2-AS1
Endothelial cell
C. T-cell
i. Non-coding enriched genes
ii. LINC00861
T-cell
iii. RP11-326C3.2
T-cell
D. Basophil
i. Non-coding enriched genes
ii. LINC01835
Basophil
iii. RP11-354E11.2
Basophil
E. Plasma cell
i. Non-coding enriched genes
ii. IGHGP
Plasma cell

Figure 6: Non-coding gene signatures of human sigmoid colon cell types. Non-coding gene enrichment signatures for **(A)** epithelial cell, **(B)** endothelial cell, **(C)** T-cell, **(D)** Basophil and **(E)** plasma cell showing: (i) up to the top 50 cell type-enriched non-coding transcripts ranked by highest mean correlation to Ref.T., showing correlation coefficients with the Ref.T. panel, differential score (correlation with corresponding cell type Ref.T. minus maximum correlation with any other Ref.T. panel), mean expression value in bulk RNAseq and the transcript type. (ii and iii) scRNAseq data from analysis of the large intestine was sourced from Tabula Sapiens and used to generate UMAP plots showing the expression profiles of cell type-enriched non-coding genes. See figure SX for all UMAP plot annotations.

Sex-subset comparison

We performed a sex subset analysis of the sigmoid colon RNAseq dataset from GTEx (male n=240, female n=133). Male and female data subsets were analysed separately and enrichment profiles for each cell type calculated as for the whole dataset (Table S2). To compare gene enrichment profiles in males and females, the following was calculated for any gene that was classified as cell type enriched in either subset: (**i**) the '*differential correlation score*', defined as the difference between the mean corr. coefficient with the cell type *Ref.T*, in the male and female sample subsets, (**ii**) the '*enrichment ranking*', based on the mean corr. value with the *Ref.T.* panel (rank 1 = highest corr.). Generally, the cell type gene enrichment profiles were largely comparable between the sexes (Figure 7A and Figure S3, transcripts enriched in both male and female are represented by square symbols). Genes that were classified as only enriched in males or females (differently coloured circle or triangle symbols), mostly had a differential corr. score close to zero; indicating that they only fell slightly below the enrichment threshold in the other sex.

One distinct male-only enteric neuron enriched gene was identified, *RP4-555D20.4* (Figure 7B i), a transcript with a higher mean expression in males samples, compared to female ones (Figure 7B ii). *RP4-555D20.4* had elevated mean expression in multiple tissues, notably brain and thyroid, compared to other tissue types, potentially indicating a broader function beyond the colon. We also identified one male-only T-cell enriched gene *BCORP1* (Figure 7C i), a Y-chromosome gene with expression levels above background level only in male samples (Figure 7C ii). *BCORP1* had enhanced expression in colon tissue, and other gastrointestinal tissues such as small intestine and oesophagus, compared to other tissue types (Figure 7C

iii), indicating a potential sex-linked gastrointestinal-specific T-cell function.

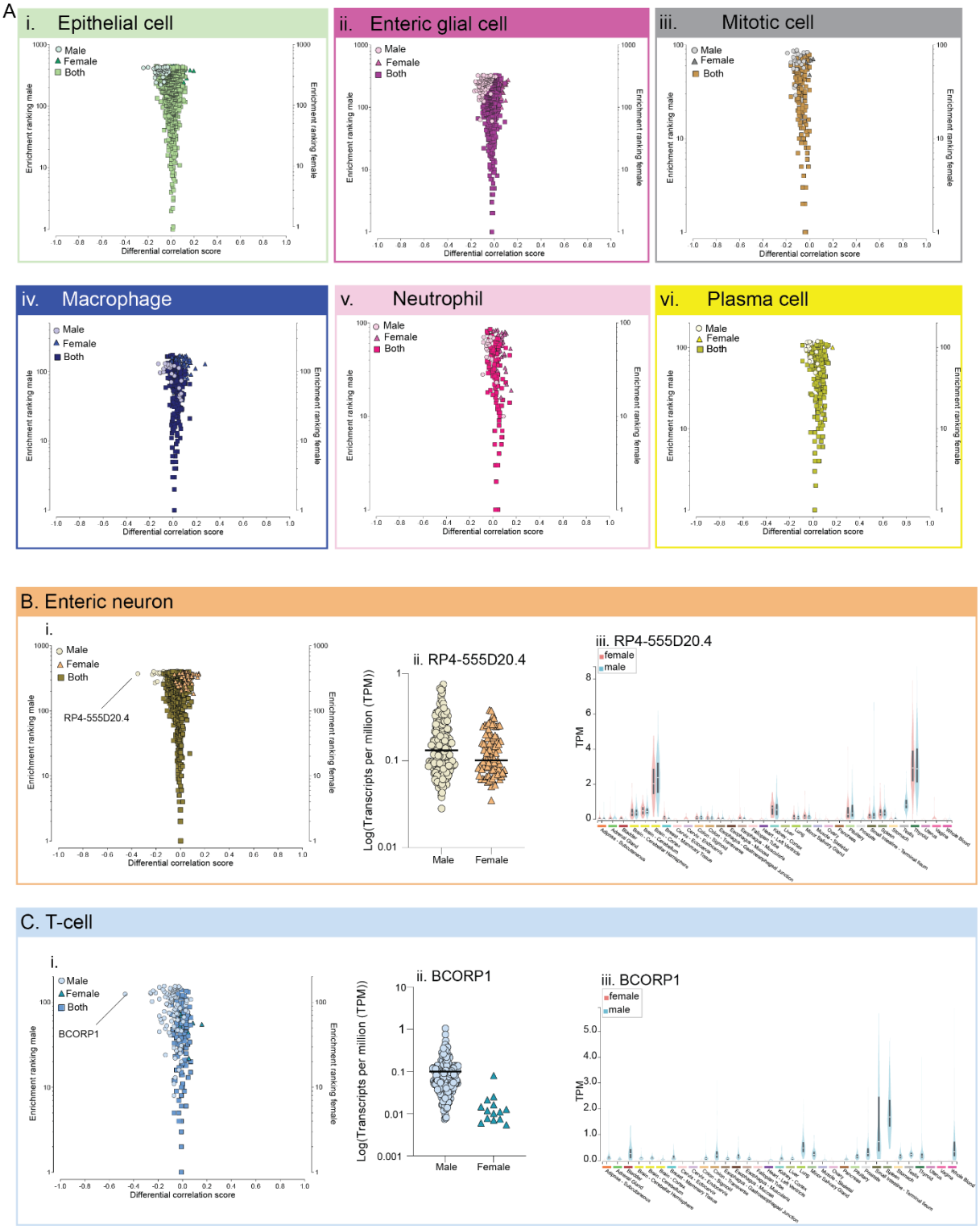


Figure 7. Identification of sex-specific cell type-enriched transcripts in human sigmoid colon tissue. (A) Human colon RNAseq data was retrieved from GTEx and divided into female (n=133) and male (n=240) subgroups identification of sex-specific cell type-enriched transcripts. For transcripts classified as (i) epithelial, (ii) enteric glial, (iii) mitotic, (iv) macrophage, (v) neutrophil and (vi) plasma cell enriched the 'sex differential corr. score' was plotted vs. 'enrichment ranking' was plotted. Transcripts classified as enriched in both sexes are represented by square symbols and transcripts that are classified as enriched in only male or female are represented by circle or triangle symbols, respectively. The sex-subset analysis for **(B)** Intestinal endocrine cell and **(C)** T-cell revealed male-enriched expression profiles (i). (iii) Expression in female or male samples for transcripts identified as male-only enriched and (iv) expression of male-enriched genes in bulk RNAseq of different human organs from male and female donors.

DISCUSSION

Here, we present a genome wide cell type enriched transcriptome atlas for the human sigmoid colon. We have previously demonstrated that using the reference transcript analysis method can be used to identify cell-enriched transcripts from unfractionated RNAseq data (Dusart et al. 2019; Norreen-Thorsen et al. 2022; Butler et al. 2016). Here we use this approach to identify several cell-specific protein coding and non-coding transcriptomes in the human colon. Data is available on the human protein atlas online resource, which can be searched on a gene-by-gene basis (https://www.proteinatlas.org/humanproteome/tissue+cell+type/colon).

Using our approach has some advantages and limitations compared to other types of transcriptome profiling, for example single cell sequencing (scRNAseq). scRNAseq has some challenges involving efficient cell isolation and material amplification (Gawad, Koh, and Quake 2016; Shapiro, Biezuner, and Linnarsson 2013; Grün and van Oudenaarden 2015). Our reference transcript-based method circumvents these issues by not requiring isolation of single cells from tissue to identify cell specific transcriptome profiles. Additionally, by not removing the cells of interest from the tissue we also avoid changing the acute gene expression due to loss of tissue specific cues. Sample processing of live cells in scRNAseq can also alter the transcriptome of individual cells (O'Flanagan et al. 2019), and therefore needs to be tightly controlled, and usually uses cells from only a handful of individuals. Our method extracts information from readily available bulk RNAseq, with a possibility to process hundreds of individual samples. By incorporating a larger number of biological replicates, it also enables the subgroup comparison between sexes and organs.

Of the 12 cell types profiled, epithelial cells were one of the cell types with a high number with enriched genes, likely constituting a number of epithelial cell subtypes present in colon tissue. These enriched genes included proteins with known cell type specific functions, such as in mucosal defense *MUC13, MUC4, MUC17* (Sheng et al. 2013; Grondin et al. 2020), lipid metabolism *FABP1* and *LIPH* (Jin et al. 2002; Rodriguez Sawicki et al. 2017), structural integrity *KRT18*, *KRT19, KRT20* (Coulombe and Omary 2002) and transcription factors *CDX1* and *CDX2* (Eda et al. 2002; Grainger, Hryniuk, and Lohnes 2013; Silberg et al. 2000). We also

identified genes with previously unknown functions in epithelial cells such as: *PRR15*, a low molecular weight nuclear protein previously associated with gastrointestinal tumors (Meunier et al. 2011). *PRR15* has previously shown increased expression in colon tissue (D.-H. Yu et al. 2013) and loss of *PRR15* mRNA in animal models causes embryonic lethality (Purcell et al. 2009). We also identified *PRR15L* as epithelial cell enriched, a protein with unknown functions in epithelial cells but previously associated specifically with sigmoid colon cancer (Mizuguchi et al. 2019) and gastric cancer (Wei et al. 2021). Additionally, *GPA33* , a cell-surface differentiation protein (Heath et al. 1997) was classified as epithelial cell enriched. The specific function of *GPA33* is unknown, but it has been linked to cell-cell adhesion (Frey et al. 2008) and is expressed in intestinal mucosa and in more than 95% of colorectal cancer tumors (GarinChesa et al. 1996).

Predicted enteric neuron cell enriched genes also included transcripts with known cell type functions, such as hormone production and maturation (*SCG5, SCGN* and *CHGB)* (Busslinger et al. 2021), bile acid transport (*SLC10A4)* (Claro da Silva, Polli, and Swaan 2013), amino acid transport (*SLC7A14)* (Fotiadis, Kanai, and Palacín 2013), insulin secretion (*PRPRN* and *PTPRN2)* (Atari et al. 2019) and neuronal proteins (*TUBB2B, PHOX2B* and *RET)* (Elmentaite et al. 2021). Novel enteric neuron cell genes identified included *GABRG2* and *SLC6A17*, previously identified as differentially expressed hub genes in breast cancer (Yuanyuan Zhang, Yang, and Jiao 2022), cell adhesion-related gene *ASTN1* (Tang et al. 2018) and genes related to colorectal cancer *GNG3* and *UCHL1* (Okochi-Takada et al. 2006; Y. Li et al. 2022).

Several genes we predicted to be enteric glial cell enriched were well known markers for this cell type, such as S100B and PLP1 (Boesmans et al. 2022) and GAS7 and SPARC (Elmentaite et al. 2021). However, others had no reported cell type specific function. Our classifications were supported by showing elevated expression in both colon and brain tissue, compared to other organs available on the GTEx portal V8 (https://gtexportal.org) (Consortium 2015), as well as protein profiling.

Currently there is no known database of non-coding gene enrichment profiles in the cell types of the human sigmoid colon, in addition to a general lack of information regarding the function

of any such genes in healthy tissue, while there is increasing amounts of evidence of the involvement of non-coding transcripts in the development of cancer (P.-F. Li et al. 2014; Gao et al. 2020; Ghafouri-Fard and Taheri 2020) indicating that these non-coding transcript classes have important functions.

SORCS1, a non-coding gene that we predicted to be enteric glial cell enriched, has been associated with gastrointestinal cancerous malignancies (Hua et al. 2017; Rademakers et al. 2021; Willnow, Petersen, and Nykjaer 2008). *SORCS1* functions as a NRXN-binding protein and is a critical regulator of trafficking neuronal receptors (Ribeiro et al. 2019; Savas et al. 2015) and has been identified in murine brain glial cells (Nielsen et al. 2008). Enteric glial enriched *TPT1-AS1* has been associated with colorectal cancer (CRC) progression where the expression is upregulated in cancerous tissue and further associated with a poor prognosis, functional analysis of *TPT1-AS1* suggest a pro-angiogenic and metastatic role (Yiyun Zhang et al. 2020). *FOXD3-AS1* has been implicated in the involvement of several cancer types (Z.-H. Chen et al. 2016; X. Chen et al. 2019; Guan et al. 2019; Ji et al. 2020; Yang et al. 2021) in addition to CRC where it has been linked to tumour growth (Q. Wu et al. 2019). *FOXD3-AS1* has previously been identified in Müller glial cells (Rochet et al. 2019). Other genes we identified as enteric glial cell enriched include *COL28A1,* which has previously been identified in satellite glial cells (Chu et al. 2023; Mapps et al. 2022), a cell type which surrounds the cell body of peripheral neuron cells, and *SHISA9,* which codes for an AMPA receptor auxiliary subunit that modifies AMPA receptor activity (Farrow et al. 2015; Khodosevich et al. 2014).

Of the profiled cell types, epithelial cells and enteric glial cells had the highest number of predicted non-coding enriched genes. Several of the epithelial cell non-coding enriched transcripts included antisense transcripts corresponding to epithelial cell enriched protein coding genes, such as *SATB2-AS1, VIPR1-AS1* and *TRIM31-AS1*, suggesting local gene regulation. Several of the smooth muscle cell enriched non-coding transcripts with higher TPM values had previously been mentioned in the context of cancer, and especially in relation to CRC. *LINC01278* has been shown to increase CRC progression (Xi, Ye, and Wang 2020) and *MBNL1-AS1* is downregulated during CRC but is involved in invasion, migration and

proliferation of cancer stem cells (K. Zhu et al. 2020). Additionally, *AF001548.6* has been identified as downregulated in stomach adenocarcinoma (Q. Li et al. 2020) and *RP11-92C4.6* was identified as a potential candidate for putative cancer driver mutation across whole cancer genomes (Rheinbay et al. 2017). Epithelial cell enriched *LINC00261* has numerous reports of abnormal expression in several human malignancies, where it mainly functions as a tumour suppressor regulating cellular functions such as apoptosis, proliferation and motility (H.-F. Zhang, Li, and Han 2018; Shi et al. 2019; Sha et al. 2017; B. Zhang, Li, and Sun 2018; Y. Yu et al. 2017; Yan et al. 2019; M. Zhang et al. 2021). *LINC00261* has also been shown to be important for maintaining a pro-epithelial state, which is associated with a favourable disease outcome, in pancreatic adenocarcinoma (Dorn et al. 2020). Additional reports establish *LINC00261* as an epithelial cell marker in lung (Dhamija et al. 2019). *CDKN2B-AS1* has been reported in several cancerous malignancies in addition to CRC (Ma et al. 2021; Dasgupta et al. 2020; L. Zhu et al. 2019; Huang et al. 2018), it has also been reported in glaucoma (Burdon et al. 2011; Pasquale et al. 2013) and atherosclerosis (Ou et al. 2020; Haocheng Li et al. 2019). Additionally, *CDKN2B-AS1* has shown involvement in gastrointestinal diseases such as inflammatory bowel disease, specifically in ulcerative colitis (Tian et al. 2020; Rankin et al. 2019), where the expression of *CDKN2B-AS1* was downregulated and negatively correlated with the level of inflammatory cytokines.

BCORP1 was identified as T-cell enriched in male colon tissue, and was previously identified as T-cell enriched in male visceral adipose tissue (Norreen-Thorsen et al. 2022). Overexpression of the Y-linked *BCORP1* has been reported during the differentiation of embryonic stem cells to cardiomyocytes (Meyfour et al. 2017), there are also reports that indicate that loss of chromosome Y in leukocytes, and CD4+ T-cells, is involved in an increased risk of diseases and cancer in males (Dumanski et al. 2021). The X-linked counterpart of *BCORP1, BCOR* (encoding for a BCL6 corepressor), has been identified as a potential tumour suppressor in T-cell linked malignancies (Tanaka et al. 2017; Dobashi et al. 2016; J. H. Kang et al. 2021). We also identified *RP4-555D20.4* as a male-enriched enteric neuron cell transcript. *RP4-555D20.4* is a long non-coding RNA (LncRNA) with little previous information

other than reports of downregulation in breast cancer cell lines after *SIRT7* knockdown (K.-L. Chen et al. 2017).

There are some limitations to our method; there might be incorrect classification of transcripts, especially between cell-types that are closely related or expressed at similar ratios across samples or with genes that are regulated by environmental factors. In the case for the sigmoid colon we were unable to identify transcriptomes for the epithelial subtypes, such as colonocytes and goblet cells. This could be due to the tissue sample processing, as there was extreme inconsistency with the mucous layer being removed before sequencing in many samples, which can be seen in a large portion of the epithelial cell enriched genes having large percentages of samples with TPM <0.1, with large standard deviations that were not seen in most non-epithelial cell types. This prevented identification of suitable *Ref.T* panels that fulfilled the required criteria for rarer epithelial cell subtype-enriched transcriptome identification, and so we used a more general epithelial cell definition.

ACKNOWLEDGEMENTS

Funding granted to LMB from Hjärt Lungfonden (20170759, 20170537, 20200544) and Swedish Research Council (2019-01493), and to JO from Stockholm County Council (SLL 2017-0842). The Human Protein Atlas is funded by The Knut and Alice Wallenberg Foundation. **Data usage:** We used data from Genotype-Tissue Expression (GTEx) Project (gtexportal.org) (Consortium 2015) supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

AUTHOR CONTRIBUTIONS Conceptualisation: LMB. Methodology: SÖ, ES, MNT, PD. Formal analysis: SÖ, PD, LMB. Investigation SÖ, PD, LMB, CL. Resources: MU, FP, JO, LB, CL. Writing – Original Draft: SÖ, LMB. Writing – Review & Editing: All, Visualisation: SÖ, LMB, PD, MZ, KVF. Supervision: LMB, PD. Funding Acquisition: LMB, JO**.**

DECLARATION OF INTERESTS The authors declare no competing interests.

METHODS AND RESOURCES

LEAD CONTACT

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact: Dr. Lynn Marie Butler. Email: Lynn.butler@ki.se

MATERIALS AVALIBILITY

This study did not generate new unique reagents.

DATA AND CODE AVAILABILITY

- This paper analyses existing, publicly available data from the Genotype-Tissue Expression (GTEx) project with accession number phs000424.v8.p2 (Consortium 2015) and single cell RNAseq data from Tabula Sapiens (Tabula Sapiens et al. 2022) retrieved on 2022/07/29.
- All original code has been deposited at GitHub and is publicly available as of the date of publication, link: https://github.com/PhilipDusart/cell-enrichment.
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bulk RNAseq data analysed in this study was obtained from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) (Consortium 2015) accessed on 2021/04/26 (dbGaP Accession phs000424.v8.p2). Transcript types were categorised according to Biotype definitions in ENSEMBL release 102 (Yates et al. 2020). Human tissue protein profiling was performed in house as part of the Human Protein Atlas (HPA) project (Ponten, Jirstrom, and Uhlen 2008; Uhlen et al. 2015; 2017) (www.proteinatlas.org). Human colon tissue samples were obtained from the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden, as part of the Uppsala Biobank. Samples were handled in accordance with Swedish laws and regulations, with approval from the Uppsala Ethical Review Board (Uhlen et al., 2015).

METHOD DETAILS

Tissue Profiling: Human tissue sections

Colon tissue sections were stained, as previously described (Ponten, Jirstrom, and Uhlen 2008; Uhlen et al. 2015). Briefly, formalin fixed and paraffin embedded tissue samples were sectioned, de-paraffinised in xylene, hydrated in graded alcohols and blocked for endogenous peroxidase in 0.3% hydrogen peroxide diluted in 95% ethanol. For antigen retrieval, a Decloaking chamber® (Biocare Medical, CA) was used. Slides were boiled in Citrate buffer®, pH6 (Lab Vision, CA). Primary antibodies and a dextran polymer visualization system (UltraVision LP HRP polymer®, Lab Vision) were incubated for 30 min each at room temperature and slides were developed for 10 minutes using Diaminobenzidine (Lab Vision) as the chromogen. Slides were counterstained in Mayers hematoxylin (Histolab) and scanned using Scanscope XT (Aperio). Primary antibodies, source, target and identifier are as follows: Atlas Antibodies: LGALS4 (Cat#HPA031184, RRID:AB_2673776), CHGB (Cat#HPA012602, RRID:AB_1846706), CADPS (Cat#HPA059328, RRID:AB_2683982), ALCAM (Cat#HPA010926, RRID:AB_1078449), PLP1 (Cat#HPA004128, RRID:AB_1079635), ANLN (Cat#HPA050556, RRID:AB_2681175), CCNB1 (Cat#HPA061448, RRID:AB_2684522), ADGRL4 (Cat#HPA025229, RRID:AB_10602493), CD93 (Cat#HPA012368, RRID:AB_1846341), PDIA5 (Cat#HPA030355, RRID:AB_10602200), PRKD2 (Cat#HPA021490, RRID:AB_1855708), MYLK (Cat#HPA031677, RRID:AB_10794617), CNN1 (Cat#HPA014263, RRID:AB_1847039), TAGLN (Cat#HPA019467, RRID:AB_1857245), FXYD2 (Cat#HPA068838, RRID:AB_2686047), FBN1 (Cat#HPA021057, RRID:AB_1848586), ITGB2 (Cat#HPA016894, RRID:AB_1846257), CYBB (Cat#HPA051227, RRID:AB_2681395), PGD (Cat#HPA031315, RRID:AB_2673825), FPR1 (Cat#HPA046550, RRID:AB_2679694), MCEMP1 (Cat#HPA014731, RRID:AB_1845619), CRISPLD2 (Cat#HPA030055, RRID:AB_10611821), CEACAM3 (Cat#HPA011041, RRID:AB_1078481), CD247 (Cat#HPA008750, RRID:AB_1857863), CD40LG (Cat#HPA045827, RRID:AB_10959606), IL2RB (Cat#HPA062657, RRID:AB_2684822), JCHAIN (Cat#HPA044132, RRID:AB_2678826), GIMAP4 (Cat#HPA019135,

RRID:AB_1849670), MYH11 (Cat#HPA014539, RRID:AB_1234906), ALOX5AP (Cat#HPA026592, RRID:AB_10601115), S100A8 (Cat#HPA024372, RRID:AB_1856536), HPGDS (Cat#HPA024035, RRID:AB_1855743), VAT1L (Cat#HPA061138), PHOX2B (Cat#HPA074325, RRID:AB_2686678), from Santa Cruz Biotechnology: PIGR (Cat#sc-20656, RRID:AB_2164819), MUC2 (Cat#sc-7314, RRID:AB_627970), SCG3 (Cat#sc-50289, RRID:AB_2302033), S100B (Cat#AMAb91038, RRID:AB_2665776), CDC20 (Cat#sc-13162, RRID:AB_628089), S1PR1 (Cat#sc-48356, RRID:AB_2238920), FBLN1 (Cat#sc-25281, RRID:AB_671972), NCF4 (Cat#sc-48388, RRID:AB_627989), S100A9 (Cat#sc-20173, RRID:AB_2184420), IGLC3 (Cat#sc-53344, RRID:AB_629719), IGKC (Cat#sc-52338, RRID:AB_2251264), ELAVL4 (Cat#sc-28299, RRID:AB_627765), SNAP91 (Cat#sc-25552, RRID:AB_2302221), from Leica Biosystems: SPARC (Cat#NCL-O-NECTIN, RRID:AB_563919), TOP2A (Cat#NCL-TOPOIIA, RRID:AB_564035), CD163 (Cat#NCL-CD163, RRID:AB_563510), CD2 (Cat#NCL-CD2-271, RRID:AB_442057), UCHL1 (Cat#NCL-L-PGP9.5, RRID:AB_563981), SYP(Cat#NCL-L-SYNAP-299, RRID:AB_564017), from Sigma-Aldrich: L1CAM (Cat#L4543, RRID:AB_609903), CADM1 (Cat#S4945, RRID:AB_532287), from R&D Systems: CHL1 (Product name: MAB2126), from AbCam: CDK1 (Cat#1161-1, RRID:AB_344898), CMA1 (Cat#ab2377, RRID:AB_2083616), NEFH (Cat#1707-1, RRID:AB_598179), from Origene: DCN (Product name: 2354), CPA3 (Product name: 3129.00.02), from Merck: CTSD (Product name: MAB422), TPSAB1 (Product name: MAB1222), from Agilent: LYZ (Cat#A0099, RRID:AB_2341230), MKI67 (Cat#M7240, RRID:AB_2142367),from AbFrontier: IGHM (Cat#LF-MA0164, RRID:AB_1617732), from NCI-CPTAC: GSN (Product name: CPTC-Gelsolin-1) and from Cell signaling technology, Inc: HNF4A (Cat#3113, RRID:AB_2295208).

QUANTIFICATION AND STATISTICAL ANALYSIS

Reference transcript-based correlation analysis

This method was adapted and expanded from that previously developed to determine the cross-tissue pan-EC-enriched transcriptome (Butler et al. 2016) and human brain and adipose

tissue cell-enriched genes (Dusart et al. 2019; Norreen-Thorsen et al. 2022). Pairwise Spearman correlation coefficients were calculated between reference transcripts selected as proxy markers for Epithelial cells [*EPCAM, SLC44A4, PHGR1*], enteric neuron cells [*INA, RTN1, SCG3*], enteric glial cells [*SLC35F1, NRXN1, L1CAM*], mitotic cells [*HMMR, TOP2A, RMM2*], endothelial cells [*ESAM, CDH5, MMRN2*], smooth muscle cells [*ACTG2, TAGLN, TPM1*], fibroblasts [*FBLN1, DCN, TNXB*], macrophages [*C1QB, CD68, ITGB2*], neutrophils *[S100A8, AQP9, FCAR* 0.73], basophils [*CPA3, TPSAB1, SIGLEC6*], T-cells [*CD2, CD6, CD3E*] and plasma cells [*JCHAIN, IGKC, IGHA1*] and all other sequenced transcripts. Transcripts with a TPM value <0.1 in more than 50% of samples were excluded from analysis (but are still included in data tables). See results section for full criteria required for transcript classification of transcripts as cell-type enriched. Correlation coefficients were calculated in R using the *corr.test* function from the *psych* package (v 1.8.4). In addition to correlation coefficients False Discovery Rate (FDR) adjusted p-values (using Bonferroni correction) and raw p-values were calculated. FDR <0.0001 for correlation was required for inclusion as cell type enriched, but no transcripts required exclusion due to this criterion.

Weighted correlation network (WGCNA) analysis

The R package WGCNA (Langfelder and Horvath 2008) was used to perform co-expression network analysis for gene clustering, on log2 expression TPM values. The analysis was performed according to recommendations in the WGCNA manual. Transcripts with too many missing values were excluded using the goodSamplesGenes() function. The remaining genes were used to cluster the samples, and obvious outlier samples were excluded.

Gene ontology analysis

The Gene Ontology Consortium (Ashburner et al. 2000) was used to identify over represented terms (biological processes) in the panel of identified cell-type-enriched transcripts from the GO ontology (release date 2022-03-22) database. Dendrogram plots showing over-represented GO terms were created using the R package ClusterProfiler (T. Wu et al. 2021).

Visualisation

Circular graphs were constructed using the R package *circlize* (Gu et al. 2014). Some figure sections were created with BioRender.com.

Additional datasets and analysis

Single cell RNAseq data from Tabula Sapiens (Tabula Sapiens et al., 2022) was downloaded and UMAP plots created using the Seurat package in R (Hao et al., 2021). Tissue enriched genes were downloaded from the Human Protein Atlas (HPA) tissue atlas (Uhlen et al., 2015) or GTEx database (Consortium, 2015), as collated in the Harminozome database (Rouillard et al., 2016).

ADDITIONAL RESOURCES

Analysed data for all protein coding genes is provided on the Human Protein Atlas website: (https://www.proteinatlas.org/humanproteome/tissue+cell+type/colon).

SUPPLEMENTAL TABLE LEGENDS

Table S1. Reference transcript selection and analysis criteria.

(Tab **1**): Correlation coefficient values were calculated between selected *Ref.T.* to represent constituent colon cell types. (Tab **2**): Correlation coefficient values were calculated between selected *Ref.T.* and all other sequenced transcripts in GTEx colon mRNAseq data (Table A) and the mean differential vs. all *Ref.T.* panels (Table B). Over represented gene ontology terms in transcripts classified as enriched in: (Tab **3**) epithelial cells, (Tab **4**) enteric neuron cells, (Tab **5**) enteric glial cells, (Tab **6**) mitotic cells, (Tab **7**) endothelial cells, (Tab **8**) smooth muscle cells, (Tab **9**) fibroblasts, (Tab **10**) macrophages, (Tab **11**) neutrophils, (Tab **12**) basophils, (Tab **13**) T-cells and (Tab **14**) plasma cells. *Related to all Figures*.

Table S2. Sex stratified subset analysis of cell-enriched transcripts in human colon.

(Tab **1)**: Correlation coefficient values were calculated between selected *Ref.T.* to represent constituent colon cell types in females (Table **A**) or males (Table **B**). (Tab **2**) Correlation coefficient values were calculated between selected *Ref.T.* and all other sequenced transcripts in colon mRNAseq data (GTEx), subdivided into (Table **A**) female or (Table **B**) male only sample sets. See key for column details. *Related to Figure 7 and S3*.

REFERENCES

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology.” *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Atari, Ealla, Mitchel C Perry, Pedro A Jose, and Sivarajan Kumarasamy. 2019. “Regulated Endocrine-Specific Protein-18, an Emerging Endocrine Protein in Physiology: A Literature Review.” *Endocrinology* 160 (9): 2093–2100. https://doi.org/10.1210/en.2019-00397.

Bailey, Charles, Rupert Negus, Alistair Morris, Paul Ziprin, Robert Goldin, Paola Allavena, David Peck, and Ara Darzi. 2007. “Chemokine Expression Is Associated with the Accumulation of Tumour Associated Macrophages (TAMs) and Progression in Human Colorectal Cancer.” *Clinical & Experimental Metastasis* 24 (2): 121–30. https://doi.org/10.1007/s10585-007-9060-3.

Bian, Shuhui, Yu Hou, Xin Zhou, Xianlong Li, Jun Yong, Yicheng Wang, Wendong Wang, et al. 2018. “Single-Cell Multiomics Sequencing and Analyses of Human Colorectal Cancer.” *Science* 362 (6418): 1060–63. https://doi.org/10.1126/science.aao3791.

Boesmans, Werend, Amelia Nash, Kinga R. Tasnády, Wendy Yang, Lincon A. Stamp, and Marlene M. Hao. 2022. “Development, Diversity, and Neurogenic Capacity of Enteric Glia.” *Frontiers in Cell and Developmental Biology* 9 (January): 775102. https://doi.org/10.3389/fcell.2021.775102.

Burclaff, Joseph, R. Jarrett Bliton, Keith A. Breau, Meryem T. Ok, Ismael Gomez-Martinez, Jolene S. Ranek, Aadra P. Bhatt, Jeremy E. Purvis, John T. Woosley, and Scott T. Magness. 2022. “A Proximal-to-Distal Survey of Healthy Adult Human Small Intestine and Colon Epithelium by Single-Cell Transcriptomics.” *Cellular and Molecular Gastroenterology and Hepatology* 13 (5): 1554–89. https://doi.org/10.1016/j.jcmgh.2022.02.007.

Burdon, Kathryn P., Stuart Macgregor, Alex W. Hewitt, Shiwani Sharma, Glyn Chidlow, Richard A. Mills, Patrick Danoy, et al. 2011. “Genome-Wide Association Study Identifies Susceptibility Loci for Open Angle Glaucoma at TMCO1 and CDKN2B-AS1.” *Nature Genetics* 43 (6): 574–78. https://doi.org/10.1038/ng.824.

Busslinger, Georg A., Bas L. A. Weusten, Auke Bogte, Harry Begthel, Lodewijk A. A. Brosens, and Hans Clevers. 2021. “Human Gastrointestinal Epithelia of the Esophagus, Stomach, and Duodenum Resolved at Single-Cell Resolution.” *Cell Reports* 34 (10): 108819. https://doi.org/10.1016/j.celrep.2021.108819.

Butler, Lynn Marie, Björn Mikael Hallström, Linn Fagerberg, Fredrik Pontén, Mathias Uhlén, Thomas Renné, and Jacob Odeberg. 2016. “Analysis of Body-Wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome.” *Cell Systems* 3 (3): 287-301.e3. https://doi.org/10.1016/j.cels.2016.08.001.

Chen, Kun-Lin, Lian Li, Yi-Ru Wang, Cheng-Min Li, Tarig Mohammed Badri, and Gen-Lin Wang. 2017. “Long Noncoding RNA and mRNA Profiling in MDA-MB-231 Cells Following RNAi-Mediated Knockdown of SIRT7.” *OncoTargets and Therapy* 10 (October): 5115–28. https://doi.org/10.2147/OTT.S149048.

Chen, Xige, Juan Gao, Yanhua Yu, Zhengjuan Zhao, and Yingli Pan. 2019. “LncRNA FOXD3-AS1 Promotes Proliferation, Invasion and Migration of Cutaneous Malignant Melanoma via Regulating miR-325/MAP3K2.” *Biomedicine & Pharmacotherapy* 120 (December): 109438. https://doi.org/10.1016/j.biopha.2019.109438.

Chen, Zhen-Hua, Hong-Kang Hu, Chen-Ran Zhang, Cheng-Yin Lu, Yi Bao, Zheng Cai, Yong-Xiang Zou, Guo-Han Hu, and Lei Jiang. 2016. “Down-Regulation of Long Non-Coding RNA FOXD3 Antisense RNA 1 (FOXD3-AS1) Inhibits Cell Proliferation, Migration, and Invasion in Malignant Glioma Cells.” *American Journal of Translational Research* 8 (10): 4106–19.

Choi, Eunyoung, Joseph T Roland, Brittney J Barlow, Ryan O’Neal, Amy E Rich, Ki Taek Nam, Chanjuan Shi, and James R Goldenring. 2014. “Cell Lineage Distribution Atlas of the

Human Stomach Reveals Heterogeneous Gland Populations in the Gastric Antrum." *Gut* 63 (11): 1711–20. https://doi.org/10.1136/gutjnl-2013-305964.

Chu, Yanhao, Shilin Jia, Ke Xu, Qing Liu, Lijia Mai, Jiawei Liu, Wenguo Fan, and Fang Huang. 2023. "Single-Cell Transcriptomic Profile of Satellite Glial Cells in Trigeminal Ganglion." *Frontiers in Molecular Neuroscience* 16 (February): 1117065. https://doi.org/10.3389/fnmol.2023.1117065.

Claro da Silva, Tatiana, James E. Polli, and Peter W. Swaan. 2013. "The Solute Carrier Family 10 (SLC10): Beyond Bile Acid Transport." *Molecular Aspects of Medicine*, The ABCs of membrane transporters in health and disease (SLC series), 34 (2): 252–69. https://doi.org/10.1016/j.mam.2012.07.004.

Consortium, G. TEx. 2015. "Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science* 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

Coulombe, Pierre A, and M. Bishr Omary. 2002. "'Hard' and 'Soft' Principles Defining the Structure, Function and Regulation of Keratin Intermediate Filaments." *Current Opinion in Cell Biology* 14 (1): 110–22. https://doi.org/10.1016/S0955-0674(01)00301-5.

Dalerba, Piero, Tomer Kalisky, Debashis Sahoo, Pradeep S Rajendran, Michael E Rothenberg, Anne A Leyrat, Sopheak Sim, et al. 2011. "Single-Cell Dissection of Transcriptional Heterogeneity in Human Colon Tumors." *Nature Biotechnology* 29 (12): 1120–27. https://doi.org/10.1038/nbt.2038.

Dasgupta, Pritha, Priyanka Kulkarni, Shahana Majid, Yutaka Hashimoto, Marisa Shiina, Varahram Shahryari, Nadeem S. Bhat, et al. 2020. "LncRNA CDKN2B-AS1/miR-141/Cyclin D Network Regulates Tumor Progression and Metastasis of Renal Cell Carcinoma." *Cell Death & Disease* 11 (8): 1–12. https://doi.org/10.1038/s41419-020-02877-0.

Denninger, Jiyeon K., Logan A. Walker, Xi Chen, Altan Turkoglu, Alex Pan, Zoe Tapp, Sakthi Senthilvelan, et al. 2022. "Robust Transcriptional Profiling and Identification of Differentially Expressed Genes With Low Input RNA Sequencing of Adult Hippocampal Neural Stem and Progenitor Populations." *Frontiers in Molecular Neuroscience* 15 (January): 810722. https://doi.org/10.3389/fnmol.2022.810722.

Dhamija, Sonam, Andrea C. Becker, Yogita Sharma, Ksenia Myacheva, Jeanette Seiler, and Sven Diederichs. 2019. "LINC00261 and the Adjacent Gene FOXA2 Are Epithelial Markers and Are Suppressed during Lung Cancer Tumorigenesis and Progression." *Non-Coding RNA* 5 (1): 2. https://doi.org/10.3390/ncrna5010002.

Di, Jiabo, Maoxing Liu, Yingcong Fan, Pin Gao, Zaozao Wang, Beihai Jiang, and Xiangqian Su. 2020. "Phenotype Molding of T Cells in Colorectal Cancer by Single-Cell Analysis." *International Journal of Cancer* 146 (8): 2281–95. https://doi.org/10.1002/ijc.32856.

Díez-Obrero, Virginia, Christopher H. Dampier, Ferran Moratalla-Navarro, Matthew Devall, Sarah J. Plummer, Anna Díez-Villanueva, Ulrike Peters, et al. 2021. "Genetic Effects on Transcriptome Profiles in Colon Epithelium Provide Functional Insights for Genetic Risk Loci." *Cellular and Molecular Gastroenterology and Hepatology* 12 (1): 181–97. https://doi.org/10.1016/j.jcmgh.2021.02.003.

Dobashi, Akito, Naoko Tsuyama, Reimi Asaka, Yuki Togashi, Kyoko Ueda, Seiji Sakata, Satoko Baba, Kana Sakamoto, Kiyohiko Hatake, and Kengo Takeuchi. 2016. "Frequent BCOR Aberrations in Extranodal NK/T-Cell Lymphoma, Nasal Type." *Genes, Chromosomes and Cancer* 55 (5): 460–71. https://doi.org/10.1002/gcc.22348.

Domanska, Diana, Umair Majid, Victoria T. Karlsen, Marianne A. Merok, Ann-Christin Røberg Beitnes, Sheraz Yaqub, Espen S. Bækkevold, and Frode L. Jahnsen. 2022. "Single-Cell Transcriptomic Analysis of Human Colonic Macrophages Reveals Niche-Specific Subsets." *Journal of Experimental Medicine* 219 (3): e20211846. https://doi.org/10.1084/jem.20211846.

Dorn, Agnes, Markus Glaß, Carolin T. Neu, Beate Heydel, Stefan Hüttelmaier, Tony Gutschner, and Monika Haemmerle. 2020. "LINC00261 Is Differentially Expressed in Pancreatic Cancer Subtypes and Regulates a Pro-Epithelial Cell Identity." *Cancers* 12 (5): 1227. https://doi.org/10.3390/cancers12051227.

Dumanski, Jan P., Jonatan Halvardson, Hanna Davies, Edyta Rychlicka-Buniowska, Jonas Mattisson, Behrooz Torabi Moghadam, Noemi Nagy, et al. 2021. “Immune Cells Lacking Y Chromosome Show Dysregulation of Autosomal Gene Expression.” *Cellular and Molecular Life Sciences* 78 (8): 4019–33. https://doi.org/10.1007/s00018-021-03822-w.

Dusart, Philip, Björn Mikael Hallström, Thomas Renné, Jacob Odeberg, Mathias Uhlén, and Lynn Marie Butler. 2019. “A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes.” *Cell Reports* 29 (6): 1690-1706.e4. https://doi.org/10.1016/j.celrep.2019.09.088.

Eda, Akashi, Hiroyuki Osawa, Ichiro Yanaka, Kiichi Satoh, Hiroyuki Mutoh, Ken Kihira, and Kentaro Sugano. 2002. “Expression of Homeobox Gene CDX2 Precedes That of CDX1 during the Progression of Intestinal Metaplasia.” *Journal of Gastroenterology* 37 (2): 94–100. https://doi.org/10.1007/s005350200002.

Elmentaite, Rasa, Natsuhiko Kumasaka, Kenny Roberts, Aaron Fleming, Emma Dann, Hamish W. King, Vitalii Kleshchevnikov, et al. 2021. “Cells of the Human Intestinal Tract Mapped across Space and Time.” *Nature* 597 (7875): 250–55. https://doi.org/10.1038/s41586-021-03852-1.

Erreni, Marco, Alberto Mantovani, and Paola Allavena. 2011. “Tumor-Associated Macrophages (TAM) and Inflammation in Colorectal Cancer.” *Cancer Microenvironment* 4 (2): 141–54. https://doi.org/10.1007/s12307-010-0052-5.

Farrow, Paul, Konstantin Khodosevich, Yechiam Sapir, Anton Schulmann, Muhammad Aslam, Yael Stern-Bach, Hannah Monyer, and Jakob von Engelhardt. 2015. “Auxiliary Subunits of the CKAMP Family Differentially Modulate AMPA Receptor Properties.” Edited by Marlene Bartos. *eLife* 4 (December): e09693. https://doi.org/10.7554/eLife.09693.

Fotiadis, Dimitrios, Yoshikatsu Kanai, and Manuel Palacín. 2013. “The SLC3 and SLC7 Families of Amino Acid Transporters.” *Molecular Aspects of Medicine*, The ABCs of membrane transporters in health and disease (SLC series), 34 (2): 139–58. https://doi.org/10.1016/j.mam.2012.10.007.

Franzen, O., L. M. Gan, and J. L. M. Bjorkegren. 2019. “PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data.” *Database (Oxford)* 2019 (January). https://doi.org/10.1093/database/baz046.

Frey, Dietmar, Vania Coelho, Ulf Petrausch, Michael Schaefer, Ulrich Keilholz, Eckhard Thiel, and P. Markus Deckert. 2008. “Surface Expression of gpA33 Is Dependent on Culture Density and Cell-Cycle Phase and Is Regulated by Intracellular Traffic Rather than Gene Transcription.” *Cancer Biotherapy and Radiopharmaceuticals* 23 (1): 65–73. https://doi.org/10.1089/cbr.2007.0407.

Gao, Y., J. W. Wang, J. Y. Ren, M. Guo, C. W. Guo, S. W. Ning, and S. Yu. 2020. “Long Noncoding RNAs in Gastric Cancer: From Molecular Dissection to Clinical Application.” *World J Gastroenterol* 26 (24): 3401–12. https://doi.org/10.3748/wjg.v26.i24.3401.

GarinChesa, P., J. Sakamoto, S. Welt, F. Real, W. Rettig, and L. Old. 1996. “Organ-Specific Expression of the Colon Cancer Antigen A33, a Cell Surface Target for Antibody-Based Therapy.” *International Journal of Oncology* 9 (3): 465–71. https://doi.org/10.3892/ijo.9.3.465.

Gawad, Charles, Winston Koh, and Stephen R. Quake. 2016. “Single-Cell Genome Sequencing: Current State of the Science.” *Nature Reviews Genetics* 17 (3): 175–88. https://doi.org/10.1038/nrg.2015.16.

Gene Ontology, Consortium. 2021. “The Gene Ontology Resource: Enriching a GOld Mine.” *Nucleic Acids Res* 49 (D1): D325–34. https://doi.org/10.1093/nar/gkaa1113.

Ghafouri-Fard, Soudeh, and Mohammad Taheri. 2020. “Long Non-Coding RNA Signature in Gastric Cancer.” *Experimental and Molecular Pathology* 113 (April): 104365. https://doi.org/10.1016/j.yexmp.2019.104365.

Grainger, Stephanie, Alexa Hryniuk, and David Lohnes. 2013. “Cdx1 and Cdx2 Exhibit Transcriptional Specificity in the Intestine.” *PLOS ONE* 8 (1): e54757. https://doi.org/10.1371/journal.pone.0054757.

Gremel, Gabriela, Alkwin Wanders, Jonathan Cedernaes, Linn Fagerberg, Björn Hallström, Karolina Edlund, Evelina Sjöstedt, Mathias Uhlén, and Fredrik Pontén. 2015. “The Human Gastrointestinal Tract-Specific Transcriptome and Proteome as Defined by RNA Sequencing and Antibody-Based Profiling.” *Journal of Gastroenterology* 50 (1): 46–57. https://doi.org/10.1007/s00535-014-0958-7.

Grondin, Jensine A., Yun Han Kwon, Parsa Mehraban Far, Sabah Haq, and Waliul I. Khan. 2020. “Mucins in Intestinal Mucosal Defense and Inflammation: Learning From Clinical and Experimental Studies.” *Frontiers in Immunology* 11. https://www.frontiersin.org/articles/10.3389/fimmu.2020.02054.

Grün, Dominic, and Alexander van Oudenaarden. 2015. “Design and Analysis of Single-Cell Sequencing Experiments.” *Cell* 163 (4): 799–810. https://doi.org/10.1016/j.cell.2015.10.039.

Gu, Z., L. Gu, R. Eils, M. Schlesner, and B. Brors. 2014. “Circlize Implements and Enhances Circular Visualization in R.” *Bioinformatics* 30 (19): 2811–12. https://doi.org/10.1093/bioinformatics/btu393.

Guan, Yaoyao, Adheesh Bhandari, Erjie Xia, Fan Yang, Jingjing Xiang, and Ouchen Wang. 2019. “lncRNA FOXD3-AS1 Is Associated with Clinical Progression and Regulates Cell Migration and Invasion in Breast Cancer.” *Cell Biochemistry and Function* 37 (4): 239–44. https://doi.org/10.1002/cbf.3393.

Han, Sae-Won, Hwang-Phill Kim, Jong-Yeon Shin, Eun-Goo Jeong, Won-Chul Lee, Kyung-Hun Lee, Jae-Kyung Won, et al. 2013. “Targeted Sequencing of Cancer-Related Genes in Colorectal Cancer Using Next-Generation Sequencing.” *PLoS ONE* 8 (5). https://doi.org/10.1371/journal.pone.0064271.

Heath, Joan K., Sara J. White, Cameron N. Johnstone, Bruno Catimel, Richard J. Simpson, Robert L. Moritz, Guo-Fen Tu, et al. 1997. “The Human A33 Antigen Is a Transmembrane Glycoprotein and a Novel Member of the Immunoglobulin Superfamily.” *Proceedings of the National Academy of Sciences* 94 (2): 469–74. https://doi.org/10.1073/pnas.94.2.469.

Hockley, James R F, Toni S Taylor, Gerard Callejo, Anna L Wilbrey, Alex Gutteridge, Karsten Bach, Wendy J Winchester, David C Bulmer, Gordon McMurray, and Ewan St John Smith. 2019. “Single-Cell RNAseq Reveals Seven Classes of Colonic Sensory Neuron.” *Gut* 68 (4): 633. https://doi.org/10.1136/gutjnl-2017-315631.

Hong, Bok Sil, Ji-Hoon Cho, Hyunjung Kim, Eun-Jeong Choi, Sangchul Rho, Jongmin Kim, Ji Hyun Kim, et al. 2009. “Colorectal Cancer Cell-Derived Microvesicles Are Enriched in Cell Cycle-Related mRNAs That Promote Proliferation of Endothelial Cells.” *BMC Genomics* 10 (1): 556. https://doi.org/10.1186/1471-2164-10-556.

Hua, Yang, Xiukun Ma, Xianglong Liu, Xiangfei Yuan, Hai Qin, and Xipeng Zhang. 2017. “Abnormal Expression of mRNA, microRNA Alteration and Aberrant DNA Methylation Patterns in Rectal Adenocarcinoma.” *PLOS ONE* 12 (3): e0174461. https://doi.org/10.1371/journal.pone.0174461.

Huang, Yuqi, Bo Xiang, Yuanhua Liu, Yu Wang, and Heping Kan. 2018. “LncRNA CDKN2B-AS1 Promotes Tumor Growth and Metastasis of Human Hepatocellular Carcinoma by Targeting Let-7c-5p/NAP1L1 Axis.” *Cancer Letters* 437 (November): 56–66. https://doi.org/10.1016/j.canlet.2018.08.024.

Ishigami, S.-I., S. Arii, M. Furutani, M. Niwano, T. Harada, M. Mizumoto, A. Mori, H. Onodera, and M. Imamura. 1998. “Predictive Value of Vascular Endothelial Growth Factor (VEGF) in Metastasis and Prognosis of Human Colorectal Cancer.” *British Journal of Cancer* 78 (10): 1379–84. https://doi.org/10.1038/bjc.1998.688.

Ji, Tao, Yanan Zhang, Zheng Wang, Zuoxu Hou, Xuhui Gao, and Xiaoming Zhang. 2020. “FOXD3-AS1 Suppresses the Progression of Non-Small Cell Lung Cancer by Regulating miR-150/SRCIN1axis.” *Cancer Biomarkers* 29 (3): 417–27. https://doi.org/10.3233/CBM-200059.

Jin, Weijun, Uli C. Broedl, Houshang Monajemi, Jane M. Glick, and Daniel J. Rader. 2002. “Lipase H, a New Member of the Triglyceride Lipase Family Synthesized by the Intestine.” *Genomics* 80 (3): 268–73. https://doi.org/10.1006/geno.2002.6837.

Kang, Jin Hyun, Seung Ho Lee, Jawon Lee, Murim Choi, Junhun Cho, Seok Jin Kim, Won Seog Kim, Young Hyeh Ko, and Hae Yong Yoo. 2021. “The Mutation of BCOR Is Highly Recurrent and Oncogenic in Mature T-Cell Lymphoma.” *BMC Cancer* 21 (1): 82. https://doi.org/10.1186/s12885-021-07806-8.

Kang, Jung-Cheng, Jin-Shuen Chen, Chien-Hsing Lee, Jao-Jen Chang, and Yi-Shing Shieh. 2010. “Intratumoral Macrophage Counts Correlate with Tumor Progression in Colorectal Cancer.” *Journal of Surgical Oncology* 102 (3): 242–48. https://doi.org/10.1002/jso.21617.

Kanke, Matt, Meaghan M. Kennedy Ng, Sean Connelly, Manvendra Singh, Matthew Schaner, Michael T. Shanahan, Elizabeth A. Wolber, et al. 2022. “Single-Cell Analysis Reveals Unexpected Cellular Changes and Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naïve Adult Crohn’s Disease Patients.” *Cellular and Molecular Gastroenterology and Hepatology* 13 (6): 1717–40. https://doi.org/10.1016/j.jcmgh.2022.02.005.

Karlsson, M., C. Zhang, L. Mear, W. Zhong, A. Digre, B. Katona, E. Sjostedt, et al. 2021. “A Single-Cell Type Transcriptomics Map of Human Tissues.” *Sci Adv* 7 (31). https://doi.org/10.1126/sciadv.abh2169.

Keum, NaNa, and Edward Giovannucci. 2019. “Global Burden of Colorectal Cancer: Emerging Trends, Risk Factors and Prevention Strategies.” *Nature Reviews Gastroenterology & Hepatology* 16 (12): 713–32. https://doi.org/10.1038/s41575-019-0189-8.

Khodosevich, Konstantin, Eric Jacobi, Paul Farrow, Anton Schulmann, Alexandru Rusu, Ling Zhang, Rolf Sprengel, Hannah Monyer, and Jakob von Engelhardt. 2014. “Coexpressed Auxiliary Subunits Exhibit Distinct Modulatory Profiles on AMPA Receptor Function.” *Neuron* 83 (3): 601–15. https://doi.org/10.1016/j.neuron.2014.07.004.

Kim, Young, and Timothy A. Pritts. 2017. “The Gastrointestinal Tract.” In *Geriatric Trauma and Critical Care*, edited by Fred A. Luchette and Jay A. Yelon, 35–43. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-48687-1_5.

Kong, Lingjia, Vladislav Pokatayev, Ariel Lefkovith, Grace T. Carter, Elizabeth A. Creasey, Chirag Krishna, Sathish Subramanian, et al. 2023. “The Landscape of Immune Dysregulation in Crohn’s Disease Revealed through Single-Cell Transcriptomic Profiling in the Ileum and Colon.” *Immunity* 56 (2): 444-458.e5. https://doi.org/10.1016/j.immuni.2023.01.002.

Lacar, Benjamin, Sara B. Linker, Baptiste N. Jaeger, Suguna Rani Krishnaswami, Jerika J. Barron, Martijn J. E. Kelder, Sarah L. Parylak, et al. 2016. “Nuclear RNA-Seq of Single Neurons Reveals Molecular Signatures of Activation.” *Nature Communications* 7 (1): 11022. https://doi.org/10.1038/ncomms11022.

Lake, Blue B., Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, et al. 2016. “Neuronal Subtypes and Diversity Revealed by Single-Nucleus RNA Sequencing of the Human Brain.” *Science* 352 (6293): 1586–90. https://doi.org/10.1126/science.aaf1204.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An R Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9 (1): 559. https://doi.org/10.1186/1471-2105-9-559.

Laukoetter, Mike G, Porfirio Nava, and Asma Nusrat. 2008. “Role of the Intestinal Barrier in Inflammatory Bowel Disease.” *World Journal of Gastroenterology* 14 (3): 401. https://doi.org/10.3748/wjg.14.401.

Leung, Marco L., Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E. Navin. 2017. “Single-Cell DNA Sequencing Reveals a Late-Dissemination Model in Metastatic Colorectal Cancer.” *Genome Research* 27 (8): 1287–99. https://doi.org/10.1101/gr.209973.116.

Li, Haocheng, Song Han, Qingfeng Sun, Ye Yao, Shiyong Li, Chao Yuan, Bo Zhang, et al. 2019. “Long Non-Coding RNA CDKN2B-AS1 Reduces Inflammatory Response and Promotes Cholesterol Efflux in Atherosclerosis by Inhibiting ADAM10 Expression.” *Aging (Albany NY)* 11 (6): 1695–1715. https://doi.org/10.18632/aging.101863.

Li, Huipeng, Elise T Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, et al. 2017. “Reference Component Analysis of Single-Cell Transcriptomes Elucidates Cellular Heterogeneity in Human Colorectal Tumors.” *Nature Genetics* 49 (5): 708–18. https://doi.org/10.1038/ng.3818.

Li, Pei-Fei, Sheng-Can Chen, Tian Xia, Xiao-Ming Jiang, Yong-Fu Shao, Bing-Xiu Xiao, and Jun-Ming Guo. 2014. “Non-Coding RNAs and Gastric Cancer.” *World Journal of Gastroenterology : WJG* 20 (18): 5411–19. https://doi.org/10.3748/wjg.v20.i18.5411.

Li, Qun, Xiaofeng Liu, Jia Gu, Jinming Zhu, Zhi Wei, and Hua Huang. 2020. “Screening lncRNAs with Diagnostic and Prognostic Value for Human Stomach Adenocarcinoma Based on Machine Learning and mRNA-lncRNA Co-Expression Network Analysis.” *Molecular Genetics & Genomic Medicine* 8 (11): e1512. https://doi.org/10.1002/mgg3.1512.

Li, Yijie, Feng Yuan, Zhiren Lin, and Yanling Pan. 2022. “Construction of Endogenous RNA Regulatory Network for Colorectal Cancer Based on Bioinformatics.” *Zhong Nan Da Xue Xue Bao Yi Xue Ban = Journal of Central South University. Medical Sciences* 47 (4): 416–30. https://doi.org/10.11817/j.issn.1672-7347.2022.210532.

Liu, Chang, and Jing Yang. 2022. “Enteric Glial Cells in Immunological Disorders of the Gut.” *Frontiers in Cellular Neuroscience* 16 (April): 895871. https://doi.org/10.3389/fncel.2022.895871.

Lu, Jia, Xiangcang Ye, Fan Fan, Ling Xia, Rajat Bhattacharya, Seth Bellister, Federico Tozzi, et al. 2013. “Endothelial Cells Promote the Colorectal Cancer Stem Cell Phenotype through a Soluble Form of Jagged-1.” *Cancer Cell* 23 (2): 171–85. https://doi.org/10.1016/j.ccr.2012.12.021.

Ma, Mei-Li, Hong-Yan Zhang, Shu-Yi Zhang, and Xiao-Li Yi. 2021. “LncRNA CDKN2B-AS1 Sponges miR-28-5p to Regulate Proliferation and Inhibit Apoptosis in Colorectal Cancer.” *Oncology Reports* 46 (4): 1–11. https://doi.org/10.3892/or.2021.8164.

Mapps, Aurelia A., Michael B. Thomsen, Erica Boehm, Haiqing Zhao, Samer Hattar, and Rejji Kuruvilla. 2022. “Diversity of Satellite Glia in Sympathetic and Sensory Ganglia.” *Cell Reports* 38 (5): 110328. https://doi.org/10.1016/j.celrep.2022.110328.

May, Catherine Lee, and Klaus H Kaestner. 2010. “Gut Endocrine Cell Development.” *Molecular and Cellular Endocrinology* 323 (1): 70–75. https://doi.org/10.1016/j.mce.2009.12.009.

Meunier, Dominique, Kalicharan Patra, Ron Smits, Andrea Hägebarth, Angela Lüttges, Rolf Jaussi, Matthew J. Wieduwilt, et al. 2011. “Expression Analysis of Proline Rich 15 (Prr15) in Mouse and Human Gastrointestinal Tumors.” *Molecular Carcinogenesis* 50 (1): 8–15. https://doi.org/10.1002/mc.20692.

Meyfour, Anna, Hassan Ansari, Sara Pahlavan, Shahab Mirshahvaladi, Mostafa Rezaei-Tavirani, Hamid Gourabi, Hossein Baharvand, and Ghasem Hosseini Salekdeh. 2017. “Y Chromosome Missing Protein, TBL1Y, May Play an Important Role in Cardiac Differentiation.” *Journal of Proteome Research* 16 (12): 4391–4402. https://doi.org/10.1021/acs.jproteome.7b00391.

Mizuguchi, Yasuhiko, Taku Sakamoto, Taiki Hashimoto, Shunsuke Tsukamoto, Satoru Iwasa, Yutaka Saito, and Shigeki Sekine. 2019. “Identification of a Novel PRR15L-RSPO2 Fusion Transcript in a Sigmoid Colon Cancer Derived from Superficially Serrated Adenoma.” *Virchows Archiv* 475 (5): 659–63. https://doi.org/10.1007/s00428-019-02604-x.

Nielsen, Morten S., Sady J. Keat, Jida W. Hamati, Peder Madsen, Jakob J. Gutzmann, Arne Engelsberg, Karen M. Pedersen, et al. 2008. “Different Motifs Regulate Trafficking of SorCS1 Isoforms.” *Traffic* 9 (6): 980–94. https://doi.org/10.1111/j.1600-0854.2008.00731.x.

Noah, Taeko K., Bridgitte Donahue, and Noah F. Shroyer. 2011. “Intestinal Development and Differentiation.” *Experimental Cell Research* 317 (19): 2702–10. https://doi.org/10.1016/j.yexcr.2011.09.006.

Norreen-Thorsen, Marthe, Eike Christopher Struck, Sofia Öling, Martin Zwahlen, Kalle Von Feilitzen, Jacob Odeberg, Cecilia Lindskog, et al. 2022. “A Human Adipose Tissue Cell-

Type Transcriptome Atlas." *Cell Reports* 40 (2): 111046. https://doi.org/10.1016/j.celrep.2022.111046.

O'Flanagan, Ciara H, Kieran R Campbell, Allen W Zhang, Farhia Kabeer, Jamie LP Lim, Justina Biele, Peter Eirew, et al. 2019. "Dissociation of Solid Tumour Tissues with Cold Active Protease for Single-Cell RNA-Seq Minimizes Conserved Collagenase-Associated Stress Responses." *bioRxiv*, January, 683227. https://doi.org/10.1101/683227.

Okochi-Takada, Eriko, Kazuyuki Nakazawa, Mika Wakabayashi, Akiko Mori, Shizue Ichimura, Toshiharu Yasugi, and Toshikazu Ushijima. 2006. "Silencing of the UCHL1 Gene in Human Colorectal and Ovarian Cancers." *International Journal of Cancer* 119 (6): 1338–44. https://doi.org/10.1002/ijc.22025.

Ou, Minghui, Xia Li, Shibo Zhao, Shichao Cui, and Jie Tu. 2020. "Long Non-Coding RNA CDKN2B-AS1 Contributes to Atherosclerotic Plaque Formation by Forming RNA-DNA Triplex in the CDKN2B Promoter." *EBioMedicine* 55 (May): 102694. https://doi.org/10.1016/j.ebiom.2020.102694.

Pasquale, Louis R., Stephanie J. Loomis, Jae H. Kang, Brian L. Yaspan, Wael Abdrabou, Donald L. Budenz, Teresa C. Chen, et al. 2013. "CDKN2B-AS1 Genotype–Glaucoma Feature Correlations in Primary Open-Angle Glaucoma Patients From the United States." *American Journal of Ophthalmology* 155 (2): 342-353.e5. https://doi.org/10.1016/j.ajo.2012.07.023.

Pira, Giovanna, Paolo Uva, Antonio Mario Scanu, Paolo Cossu Rocca, Luciano Murgia, Elena Uleri, Claudia Piu, et al. 2020. "Landscape of Transcriptome Variations Uncovering Known and Novel Driver Events in Colorectal Carcinoma." *Scientific Reports* 10 (1): 1–12. https://doi.org/10.1038/s41598-019-57311-z.

Ponten, F., K. Jirstrom, and M. Uhlen. 2008. "The Human Protein Atlas - a Tool for Pathology." *Journal of Pathology* 216 (4): 387–93. https://doi.org/10.1002/path.2440.

Purcell, Scott H., Jeremy D. Cantlon, Casey D. Wright, Luiz E. Henkes, George E. Seidel Jr., and Russell V. Anthony. 2009. "The Involvement of Proline-Rich 15 in Early Conceptus Development in Sheep1." *Biology of Reproduction* 81 (6): 1112–21. https://doi.org/10.1095/biolreprod.109.076190.

Rademakers, Glenn, Maartje Massen, Alexander Koch, Muriel X. Draht, Nikkie Buekers, Kim A. D. Wouters, Nathalie Vaes, et al. 2021. "Identification of DNA Methylation Markers for Early Detection of CRC Indicates a Role for Nervous System-Related Genes in CRC." *Clinical Epigenetics* 13 (1): 80. https://doi.org/10.1186/s13148-021-01067-9.

Rankin, Carl Robert, Zulfiqar Ali Lokhandwala, Raymond Huang, Joel Pekow, Charalabos Pothoulakis, and David Padua. 2019. "Linear and Circular CDKN2B-AS1 Expression Is Associated with Inflammatory Bowel Disease and Participates in Intestinal Barrier Formation." *Life Sciences* 231 (August): 116571. https://doi.org/10.1016/j.lfs.2019.116571.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. "The Human Cell Atlas." Edited by Thomas R Gingeras. *eLife* 6 (December): e27041. https://doi.org/10.7554/eLife.27041.

Rheinbay, Esther, Morten Muhlig Nielsen, Federico Abascal, Grace Tiao, Henrik Hornshøj, Julian M. Hess, Randi Istrup Pedersen, et al. 2017. "Discovery and Characterization of Coding and Non-Coding Driver Mutations in More than 2,500 Whole Cancer Genomes." bioRxiv. https://doi.org/10.1101/237313.

Ribeiro, Luís F., Ben Verpoort, Julie Nys, Kristel M. Vennekens, Keimpe D. Wierda, and Joris de Wit. 2019. "SorCS1-Mediated Sorting in Dendrites Maintains Neurexin Axonal Surface Polarization Required for Synaptic Function." *PLOS Biology* 17 (10): e3000466. https://doi.org/10.1371/journal.pbio.3000466.

Rochet, Elise, Binoy Appukuttan, Yuefang Ma, Liam M. Ashander, and Justine R. Smith. 2019. "Expression of Long Non-Coding RNAs by Human Retinal Müller Glial Cells Infected with Clonal and Exotic Virulent Toxoplasma Gondii." *Non-Coding RNA* 5 (4): 48. https://doi.org/10.3390/ncrna5040048.

Rodriguez Sawicki, Luciana, Natalia María Bottasso Arias, Natalia Scaglia, Lisandro Jorge Falomir Lockhart, Gisela Raquel Franchini, Judith Storch, and Betina Córsico. 2017. “FABP1 Knockdown in Human Enterocytes Impairs Proliferation and Alters Lipid Metabolism.” *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* 1862 (12): 1587–94. https://doi.org/10.1016/j.bbalip.2017.09.006.

Santa Barbara, P. de, G. R. van den Brink, and D. J. Roberts. 2003. “Development and Differentiation of the Intestinal Epithelium.” *Cellular and Molecular Life Sciences (CMLS)* 60 (7): 1322–32. https://doi.org/10.1007/s00018-003-2289-3.

Savas, Jeffrey N., Luís F. Ribeiro, Keimpe D. Wierda, Rebecca Wright, Laura A. DeNardo-Wilke, Heather C. Rice, Ingrid Chamma, et al. 2015. “The Sorting Receptor SorCS1 Regulates Trafficking of Neurexin and AMPA Receptors.” *Neuron* 87 (4): 764–80. https://doi.org/10.1016/j.neuron.2015.08.007.

Schneider, Sabine, Christina M. Wright, and Robert O. Heuckeroth. 2019. “Unexpected Roles for the Second Brain: Enteric Nervous System as Master Regulator of Bowel Function.” *Annual Review of Physiology* 81 (1): 235–59. https://doi.org/10.1146/annurev-physiol-021317-121515.

Serigado, Joao M., Jennifer Foulke-Abel, William C. Hines, Joshua A Hanson, Julie In, and Olga Kovbasnjuk. 2022. “Ulcerative Colitis: Novel Epithelial Insights Provided by Single Cell RNA Sequencing.” *Frontiers in Medicine* 9 (April): 868508. https://doi.org/10.3389/fmed.2022.868508.

Sha, Lixiao, Lingxiao Huang, Xishao Luo, Jiaping Bao, Lijun Gao, Qionghui Pan, Min Guo, Feiyun Zheng, and Hanchu Wang. 2017. “Long Non-Coding RNA LINC00261 Inhibits Cell Growth and Migration in Endometriosis.” *Journal of Obstetrics and Gynaecology Research* 43 (10): 1563–69. https://doi.org/10.1111/jog.13427.

Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. “Single-Cell Sequencing-Based Technologies Will Revolutionize Whole-Organism Science.” *Nature Reviews Genetics* 14 (9): 618–30. https://doi.org/10.1038/nrg3542.

Sheng, Y. H., S. Triyana, R. Wang, I. Das, K. Gerloff, T. H. Florin, P. Sutton, and M. A. McGuckin. 2013. “MUC1 and MUC13 Differentially Regulate Epithelial Inflammation in Response to Inflammatory and Infectious Stimuli.” *Mucosal Immunology* 6 (3): 557–68. https://doi.org/10.1038/mi.2012.98.

Shi, Jingli, Huimin Ma, Huaixi Wang, Weiyan Zhu, Shuting Jiang, Rui Dou, and Beizhan Yan. 2019. “Overexpression of LINC00261 Inhibits Non–Small Cell Lung Cancer Cells Progression by Interacting with miR-522-3p and Suppressing Wnt Signaling.” *Journal of Cellular Biochemistry* 120 (10): 18378–87. https://doi.org/10.1002/jcb.29149.

Silberg, Debra G., Gary P. Swain, Eun Ran Suh, and Peter G. Traber. 2000. “Cdx1 and Cdx2 Expression during Intestinal Development.” *Gastroenterology* 119 (4): 961–71. https://doi.org/10.1053/gast.2000.18142.

Smillie, Christopher S., Moshe Biton, Jose Ordovas-Montanes, Keri M. Sullivan, Grace Burgin, Daniel B. Graham, Rebecca H. Herbst, et al. 2019. “Intra- and Inter-Cellular Rewiring of the Human Colon during Ulcerative Colitis.” *Cell* 178 (3): 714-730.e22. https://doi.org/10.1016/j.cell.2019.06.029.

Spencer, Nick J., and Hongzhen Hu. 2020. “Enteric Nervous System: Sensory Transduction, Neural Circuits and Gastrointestinal Motility.” *Nature Reviews. Gastroenterology & Hepatology* 17 (6): 338–51. https://doi.org/10.1038/s41575-020-0271-2.

Squair, Jordan W., Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, et al. 2021. “Confronting False Discoveries in Single-Cell Differential Expression.” *Nature Communications* 12 (1): 5692. https://doi.org/10.1038/s41467-021-25960-2.

Tabula Sapiens, Consortium, R. C. Jones, J. Karkanias, M. A. Krasnow, A. O. Pisco, S. R. Quake, J. Salzman, et al. 2022. “The Tabula Sapiens: A Multiple-Organ, Single-Cell Transcriptomic Atlas of Humans.” *Science* 376 (6594): eabl4896. https://doi.org/10.1126/science.abl4896.

Tanaka, Tomoyuki, Yaeko Nakajima-Takagi, Kazumasa Aoyama, Shiro Tara, Motohiko Oshima, Atsunori Saraya, Shuhei Koide, et al. 2017. “Internal Deletion of BCOR

Reveals a Tumor Suppressor Function for BCOR in T Lymphocyte Malignancies." *Journal of Experimental Medicine* 214 (10): 2901–13. https://doi.org/10.1084/jem.20170167.

Tang, Yanyan, Yi He, Ping Zhang, Jinpeng Wang, Chunmei Fan, Liting Yang, Fang Xiong, et al. 2018. "LncRNAs Regulate the Cytoskeleton and Related Rho/ROCK Signaling in Cancer Metastasis." *Molecular Cancer* 17 (1): 77. https://doi.org/10.1186/s12943-018-0825-x.

Teranishi, Nobuhisa, Zenya Naito, Toshiyuki Ishiwata, Noritake Tanaka, Kiyonori Furukawa, Tomoko Seya, Seiichi Shinji, and Takashi Tajiri. 2007. "Identification of Neovasculature Using Nestin in Colorectal Cancer." *International Journal of Oncology* 30 (3): 593–603. https://doi.org/10.3892/ijo.30.3.593.

Thompson, Cayla A., Ann DeLaForest, and Michele A. Battle. 2018. "Patterning the Gastrointestinal Epithelium to Confer Regional-Specific Functions." *Developmental Biology* 435 (2): 97–108. https://doi.org/10.1016/j.ydbio.2018.01.006.

Tian, Yuanyuan, Lujia Cui, Cheng Lin, Yuxuan Wang, Zhanju Liu, and Xinpu Miao. 2020. "LncRNA CDKN2B-AS1 Relieved Inflammation of Ulcerative Colitis via Sponging miR-16 and miR-195." *International Immunopharmacology* 88 (November): 106970. https://doi.org/10.1016/j.intimp.2020.106970.

Uhlen, M., L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, et al. 2015. "Proteomics. Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419. https://doi.org/10.1126/science.1260419.

Uhlen, M., C. Zhang, S. Lee, E. Sjostedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, et al. 2017. "A Pathology Atlas of the Human Cancer Transcriptome." *Science* 357 (6352). https://doi.org/10.1126/science.aan2507.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. "Tissue-Based Map of the Human Proteome." *Science* 347 (6220): 1260419. https://doi.org/10.1126/science.1260419.

Wang, Yalong, Wanlu Song, Jilian Wang, Ting Wang, Xiaochen Xiong, Zhen Qi, Wei Fu, Xuerui Yang, and Ye-Guang Chen. 2020. "Single-Cell Transcriptome Analysis Reveals Differential Nutrient Absorption Functions in Human Intestine." *Journal of Experimental Medicine* 217 (2). https://doi.org/10.1084/jem.20191130.

Wei, Gai-Gai, Xiao-Chuan Geng, Ming-Sheng Fu, Jun Wei, Hai-Bo Yu, Xiao-Hui Li, and Gang-Long Gao. 2021. "Molecular Signature of Gastric Cancer Progression in Clinical Using Whole Genome Sequencing and the Cancer Genome Atlas (TCGA) Analysis." *Translational Cancer Research* 0 (0): 0–0. https://doi.org/10.21037/tcr-21-1011.

Willnow, Thomas E., Claus M. Petersen, and Anders Nykjaer. 2008. "VPS10P-Domain Receptors — Regulators of Neuronal Viability and Function." *Nature Reviews Neuroscience* 9 (12): 899–909. https://doi.org/10.1038/nrn2516.

Wu, Qiong, Min Shi, Wenying Meng, Yugang Wang, Pingping Hui, and Jiali Ma. 2019. "Long Noncoding RNA FOXD3-AS1 Promotes Colon Adenocarcinoma Progression and Functions as a Competing Endogenous RNA to Regulate SIRT1 by Sponging miR-135a-5p." *Journal of Cellular Physiology* 234 (12): 21889–902. https://doi.org/10.1002/jcp.28752.

Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. "clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data." *The Innovation* 2 (3): 100141. https://doi.org/10.1016/j.xinn.2021.100141.

Xi, C., N.-Y. Ye, and Y.-B. Wang. 2020. "LncRNA LINC01278 Accelerates Colorectal Cancer Progression via miR-134-5p/KDM2A Axis." *European Review for Medical and Pharmacological Sciences* 24 (20): 10526–34. https://doi.org/10.26355/eurrev_202010_23405.

Yaeger, Rona, Walid K. Chatila, Marla D. Lipsyc, Jaclyn F. Hechtman, Andrea Cercek, Francisco Sanchez-Vega, Gowtham Jayakumaran, et al. 2018. "Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer." *Cancer Cell* 33 (1): 125-136.e3. https://doi.org/10.1016/j.ccell.2017.12.004.

Yan, Dongsheng, Weidong Liu, Yeliu Liu, and Man Luo. 2019. “LINC00261 Suppresses Human Colon Cancer Progression via Sponging miR-324-3p and Inactivating the Wnt/β-Catenin Pathway.” *Journal of Cellular Physiology* 234 (12): 22648–56. https://doi.org/10.1002/jcp.28831.

Yang, Xiufang, Huilan Du, Wenhui Bian, Qingxue Li, and Hairu Sun. 2021. “FOXD3-AS1/miR-128-3p/LIMK1 Axis Regulates Cervical Cancer Progression.” *Oncology Reports* 45 (5): 1–12. https://doi.org/10.3892/or.2021.8013.

Yates, Andrew D, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, et al. 2020. “Ensembl 2020.” *Nucleic Acids Research* 48 (D1): D682–88. https://doi.org/10.1093/nar/gkz966.

Yu, Da-Hai, Carol Ware, Robert A. Waterland, Jiexin Zhang, Miao-Hsueh Chen, Manasi Gadkari, Govindarajan Kunde-Ramamoorthy, Lagina M. Nosavanh, and Lanlan Shen. 2013. “Developmentally Programmed 3′ CpG Island Methylation Confers Tissue- and Cell-Type-Specific Transcriptional Activation.” *Molecular and Cellular Biology* 33 (9): 1845–58. https://doi.org/10.1128/MCB.01124-12.

Yu, Yingcong, Linjin Li, Zhiqiang Zheng, Senrui Chen, Ende Chen, and Yiren Hu. 2017. “Long Non-Coding RNA Linc00261 Suppresses Gastric Cancer Progression via Promoting Slug Degradation.” *Journal of Cellular and Molecular Medicine* 21 (5): 955–67. https://doi.org/10.1111/jcmm.13035.

Zhang, Baogang, Changfeng Li, and Zhixia Sun. 2018. “Long Non-Coding RNA LINC00346, LINC00578, LINC00673, LINC00671, LINC00261, and SNHG9 Are Novel Prognostic Markers for Pancreatic Cancer.” *American Journal of Translational Research* 10 (8): 2648–58.

Zhang, Hai-Feng, Wei Li, and Yi-Di Han. 2018. “LINC00261 Suppresses Cell Proliferation, Invasion and Notch Signaling Pathway in Hepatocellular Carcinoma.” *Cancer Biomarkers* 21 (3): 575–82. https://doi.org/10.3233/CBM-170471.

Zhang, Menggang, Fang Gao, Xiao Yu, Qiyao Zhang, Zongzong Sun, Yuting He, and Wenzhi Guo. 2021. “LINC00261: A Burgeoning Long Noncoding RNA Related to Cancer.” *Cancer Cell International* 21 (1): 274. https://doi.org/10.1186/s12935-021-01988-8.

Zhang, X., Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, et al. 2019. “CellMarker: A Manually Curated Resource of Cell Markers in Human and Mouse.” *Nucleic Acids Res* 47 (D1): D721–28. https://doi.org/10.1093/nar/gky900.

Zhang, Yiyun, Jiangyun Sun, Yuan Qi, Yimin Wang, Yu Ding, Kun Wang, Qingxin Zhou, et al. 2020. “Long Non-Coding RNA TPT1-AS1 Promotes Angiogenesis and Metastasis of Colorectal Cancer through TPT1-AS1/NF90/VEGFA Signaling Pathway.” *Aging (Albany NY)* 12 (7): 6191–6205. https://doi.org/10.18632/aging.103016.

Zhang, Yuanyuan, Lifeng Yang, and Xiong Jiao. 2022. “Analysis of Breast Cancer Differences between China and Western Countries Based on Radiogenomics.” *Genes* 13 (12): 2416. https://doi.org/10.3390/genes13122416.

Zhang, Yunbin, Jingjing Song, Zhongwei Zhao, Mengxuan Yang, Ming Chen, Chenglong Liu, Jiansong Ji, and Di Zhu. 2020. “Single-Cell Transcriptome Analysis Reveals Tumor Immune Microenvironment Heterogenicity and Granulocytes Enrichment in Colorectal Cancer Liver Metastases.” *Cancer Letters* 470 (February): 84–94. https://doi.org/10.1016/j.canlet.2019.10.016.

Zhou, Yang, Yang Guo, and Yuanhe Wang. 2022. “Identification and Validation of a Seven-Gene Prognostic Marker in Colon Cancer Based on Single-Cell Transcriptome Analysis.” *IET Systems Biology* 16 (2): 72–83. https://doi.org/10.1049/syb2.12041.

Zhu, Kongxi, Yunxia Wang, Lan Liu, Shuai Li, and Weihua Yu. 2020. “Long Non-Coding RNA MBNL1-AS1 Regulates Proliferation, Migration, and Invasion of Cancer Stem Cells in Colon Cancer by Interacting with MYL9 via Sponging microRNA-412-3p.” *Clinics and Research in Hepatology and Gastroenterology* 44 (1): 101–14. https://doi.org/10.1016/j.clinre.2019.05.001.

Zhu, Lihong, Quanhua Zhang, Shaoping Li, Shan Jiang, Jingjing Cui, and Ge Dang. 2019. “Interference of the Long Noncoding RNA CDKN2B-AS1 Upregulates miR-181a-5p/TGFβI Axis to Restrain the Metastasis and Promote Apoptosis and Senescence of

Cervical Cancer Cells.” *Cancer Medicine* 8 (4): 1721–30. https://doi.org/10.1002/cam4.2040.

A

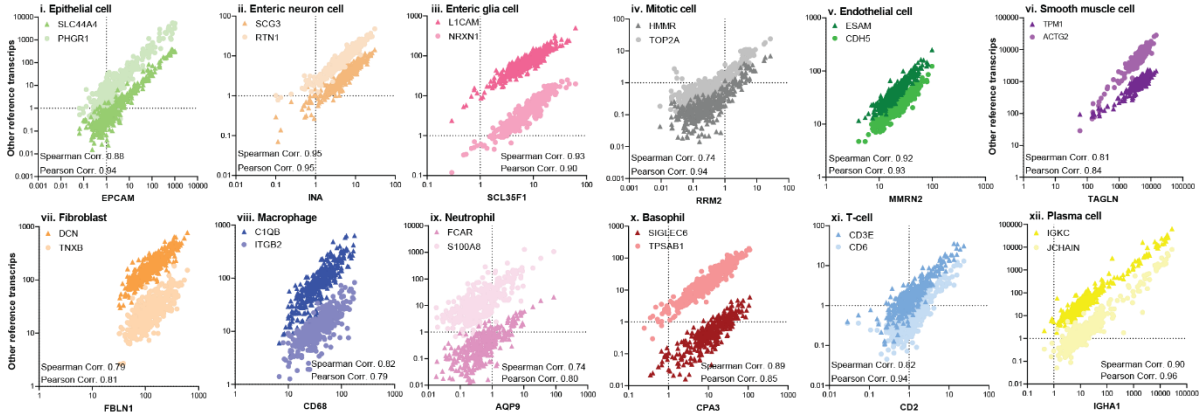


Figure S1. Expression distribution and correlation between human sigmoid colon cell type reference transcripts. Related to Figure 1. (A) Expression of Ref.T. selected to represent: (i) epithelial cell, (ii) intestinal endocrine cell, (iii) enteric glial cell, (iv) mitotic cell, (v) endothelial cell, (vi) smooth muscle cell, (vii) fibroblast, (viii) macrophage, (ix) neutrophil, (x) basophil, (xi) T-cell and (xii) plasma cell, across the sample set.

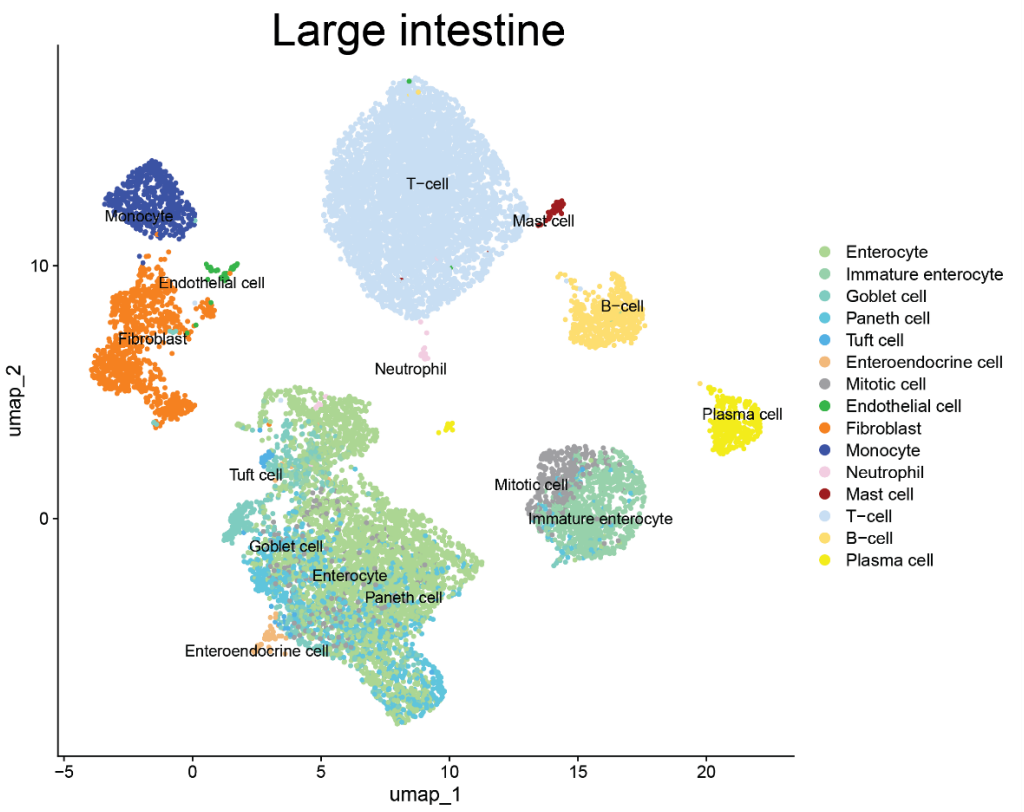


Figure S2. Single cell RNAseq (scRNAseq) annotations. Related to Figure 6. scRNAseq data was sourced from Tabula Sapiens. UMAP plots showing original annotations of cell clusters in large intestine.

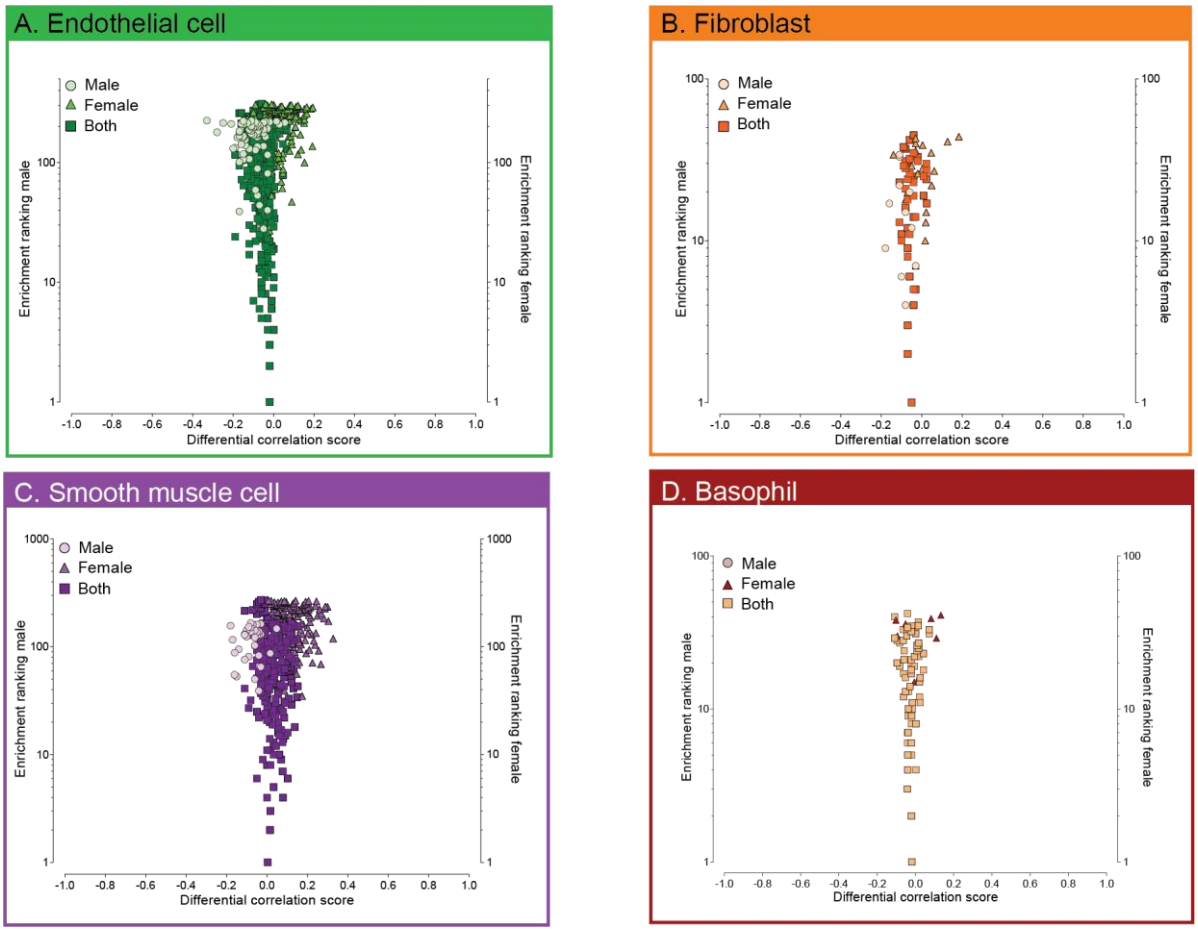


Figure S3. Identification of sex-specific cell type enriched genes in human sigmoid colon. Related to figure 7. Human colon RNAseq data was retrieved from GTEx and divided into female (n=133) and male (n=240) subgroups before analyzing for sex-specific cell type-enriched transcripts. For transcripts classified as **(A)** endothelial, **(B)** fibroblast, **(C)** smooth muscle cell or **(D)** basophil enriched the 'sex differential corr. score' was plotted vs. 'enrichment ranking' was plotted. Transcripts classified as enriched in both sexes are represented by square symbols and transcripts that are classified as enriched in only male or female are represented by circle or triangle symbols, respectively.

Paper III

A tissue centric atlas of cell type transcriptome enrichment signatures

P Dusart [1,2,3], S Öling[3], E Struck[3], M Norreen-Thorsen[3], M Zwahlen[2], K von Feilitzen[2], P Oksvold[2], M Bosic[4,5], MJ Iglesias[2], T Renné[6], J Odeberg[2,3,7,8], F Pontén[4], C Lindskog[4], M Uhlén[2], LM Butler [1,2,3,9*]

[1] Clinical Chemistry and Blood Coagulation Research, Department of Molecular Medicine and Surgery, Karolinska Institute, 171 76 Stockholm, Sweden

[2] Science for Life Laboratory, Department of Protein Science, Royal Institute of Technology (KTH), 171 21 Stockholm, Sweden

[3] Translational Vascular Research, Department of Clinical Medicine, The Arctic University of Norway, 9019 Tromsø, Norway

[4] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, 752 37 Uppsala, Sweden

[5] Institute of Pathology, Faculty of Medicine, University of Belgrade, 11000 Belgrade, Serbia

[6] Institute for Clinical Chemistry and Laboratory Medicine, University Medical Centre Hamburg-Eppendorf, D-20246 Hamburg, Germany

[7] The University Hospital of North Norway (UNN), 9019 Tromsø, Norway

[8] Coagulation Unit, Department of Hematology, Karolinska University Hospital, 171 76 Stockholm, Sweden

[9] Clinical Chemistry, Karolinska University Laboratory, Karolinska University Hospital, 171 76 Stockholm, Sweden

Correspondence to:

Dr. L.M Butler, PhD

Clinical Chemistry and Blood Coagulation Research,

Department of Molecular Medicine and Surgery,

Karolinska Institute,

SE-171 76 Stockholm, Sweden

Email: Lynn.butler@ki.se

Key words: Cell profiling, bulk RNAseq, gene enrichment

SUMMARY

Genes with cell type specific expression typically encode for proteins that have cell type specific functions. Single cell RNAseq (scRNAseq) has facilitated the identification of such genes, but various challenges limit the analysis of certain cell types and lowly expressed genes. Here, we performed an integrative network analysis of over 6000 bulk RNAseq datasets from 15 human organs, to generate a tissue-by-tissue cell type enrichment prediction atlas for all protein coding genes. We profile all the major constituent cell types, including several that are fragile or difficult to process and thus absent from existing scRNAseq-based atlases. The stability and read depth of bulk RNAseq data, and the high number of biological replicates analysed, allowed us to identify lowly expressed cell type enriched genes that are difficult to classify using existing methods. We identify co-enriched gene panels shared by pancreatic alpha and beta cells, chart temporal changes in cell enrichment signatures during spermatogenesis, and reveal that cells in the hair root are a major source of skin enriched genes. In a cross-tissue analysis, we identify shared gene enrichment signatures between highly metabolic and motile cell types, and core identity profiles of cell types found across tissue types. Our study provides the only cell type gene enrichment atlas generated independently of scRNAseq, representing a new addition to our existing toolbox of resources for the understanding of gene expression across human tissues.

INTRODUCTION

Cell type can be categorised by function, origin, location, morphology and, more recently, global transcriptome. Transcriptional profiles depend on both intrinsic cell characteristics and transient states, but selective expression of genes typically required for cell type specialised functions currently underlie our definition of cell type. Large-scale projects, such as the Human Cell Atlas (www.humancellatlas.org) [1] and the Human Protein Atlas (www.proteinatlas.org/) [2,3] contain single-cell RNA sequencing (scRNA-seq) data from thousands of cells, which can be used to further understand human health and disease, through, for example, targeted biomarker discovery [4], or elucidation of disease associated gene expression [5,6].

However, scRNA-seq has limitations; cell processing can cause artefactual modification of gene expression, through induction of the stress response [7,8] or as a consequence of removal from the microenvironment [9]. Some cell types are sensitive to extraction protocols, e.g., kidney podocytes [8], whilst others require extensive, damaging proteolytic digestion to isolate e.g., adipocytes [10,11]; such cell types are absent from the major databases [3,12,13]. Single nuclei sequencing is an alternative tool for analysing such cell types [14], but resultant expression profiles are incomplete [15]. Compared to bulk RNA-seq, where all cell types in a tissue are sequenced without prior separation, scRNAseq produces less stable and more variable data, with a high number of zero values, particularly for lowly expressed genes [16-19], requiring computational imputation for interpretation [20,21], with methods remaining controversial [22]. Typically, tissues from a limited number of donors are analysed, resulting in underestimation of biological variance of gene expression and potential false discoveries when analysing differential expression between cell types or conditions [23-25]. Differentially expressed genes identified using scRNAseq typically have higher expression and smaller fold changes than those identified with bulk RNAseq [24].

We previously developed and validated an integrative correlation analysis method to identify cell type-enriched transcriptome profiles from unfractionated tissue RNAseq [26-28]. Our method circumvents some limitations of scRNAseq; hundreds of samples are analysed concurrently to

reduce the influence of biological variation and batch effects, cell types that are technically challenging to process can be analysed, and lowly expressed cell enriched transcripts classified [28]. Here, we analysed over 6000 bulk RNAseq datasets from Genotype-Tissue Expression (GTEx) to generate a genome-wide, tissue-by-tissue cell type enrichment prediction atlas for all protein coding transcripts in 15 different human tissues. We provide gene enrichment signatures for all major constituent cell types, including those that are fragile or difficult to process, such as podocytes in the kidney and adipocytes in the breast, as well as for minority cell types, such as those in the hair follicles of the skin. We identify co-enriched genes shared by related cell types, such as pancreatic alpha and beta cells, and chart temporal changes in gene enrichment during spermatogenesis. In a cross-tissue analysis, we identify common gene enrichment signatures, e.g., between respiratory ciliated cells and spermatids, endocrine cells in the pancreas, colon, thyroid, and stomach, and between cell types found in all or most tissues, such as endothelial and immune cell types.

All data is available on the Human Protein Atlas (HPA) (www.proteinatlas.org/humanproteome/tissue+cell+type).

RESULTS

Cell type reference transcripts correlate across unfractionated tissue RNAseq data

Bulk RNAseq datasets for 15 human tissue types were retrieved from Genotype-Tissue Expression (GTEx) V8 (www.gtexportal.org) [29] (Figure 1A). To identify cell type-enriched transcript profiles, we performed an integrative correlation analysis on each dataset, using our previously published method [26-28].

As the tissue is unfractionated prior to sequencing, constituent cell types are present in different proportions in each sample (Figure 1 B.i [lung as an illustrative example]). Thus, each cell contributes mRNAs subsequently measured by RNAseq (Figure 1 B.ii), which can be: predominantly expressed in that cell type (cell type enriched), selectively expressed in two cell types (co-enriched), or expressed in several, or all, cell types within the tissue. For the main constituent cell types in each tissue (Figure 1 B.iii) marker ´reference transcripts´ [*Ref.T.*] were shortlisted (n=10-30), including: (**i**) those identified through in house tissue protein profiling [2] (**ii**) established markers identified in older ´none-omics´ studies, (**iii**) those identified by scRNAseq of mouse [13] or human [30] tissue, and (**iv**) markers from databases containing multiple studies e.g., Cell Marker [31], PanglaoDB [32] (Figure 1 B.iv). Spearman correlation coefficients were generated between all shortlisted candidate *Ref.T*. across each sample set, and three were selected to represent each cell type (for lung see Figure 1 B.v), based on the following criteria: (**i**) a high correlation between *Ref.T.* within each cell type panel (FDR <0.00001), consistent with cell type co-expression, (**ii**) a low correlation coefficient between *Ref.T.* in different cell type panels*,* consistent with high *specificity* of each panel (Figure 1 B.v) and (**iii**) a normal expression distribution of *Ref.T.* across samples. For all cell types, corresponding *Ref.T* and intra/inter *Ref.T* panel correlation coefficients in each tissue see Table S1, Tab 1, Table A-O.

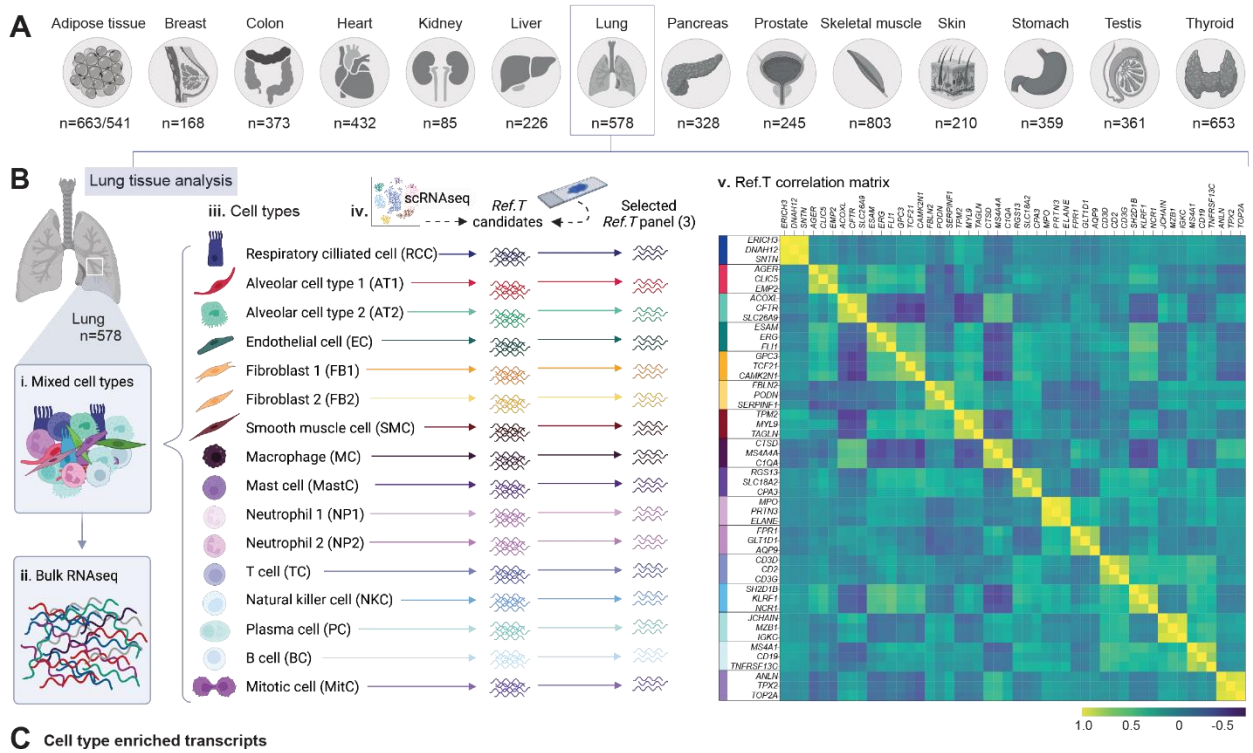

A
Adipose tissue
Breast
Colon
Heart
Kidney
Liver
Lung
Pancreas
Prostate
Skeletal muscle
Skin
Stomach
Testis
Thyroid
n=663/541
n=168
n=373
n=432
n=85
n=226
n=578
n=328
n=245
n=803
n=210
n=359
n=361
n=653
B
Lung tissue analysis
Lung
n=578
i. Mixed cell types
ii. Bulk RNAseq
iii. Cell types
iv.
scRNAseq
Ref.T candidates
Selected Ref.T panel (3)
Respiratory cilliated cell (RCC)
Alveolar cell type 1 (AT1)
Alveolar cell type 2 (AT2)
Endothelial cell (EC)
Fibroblast 1 (FB1)
Fibroblast 2 (FB2)
Smooth muscle cell (SMC)
Macrophage (MC)
Mast cell (MastC)
Neutrophil 1 (NP1)
Neutrophil 2 (NP2)
T cell (TC)
Natural killer cell (NKC)
Plasma cell (PC)
B cell (BC)
Mitotic cell (MitC)
v. Ref.T correlation matrix
1.0
0.5
0
-0.5
C
Cell type enriched transcripts
i. Respiratory ciliated
ii. Alveolar type 1
iii. Alveolar type 2
iv. Endothelial cell
v. Fibroblast 1 (alveolar)
vi. Fibroblast 2 (adventitial)
vii. Smooth muscle cell
viii. Macrophage
Mean correlation value
Ref.T panel
GO terms
Total = 681
Cilium organisation
4.4x10-63
Total = 132
Cell projection morphogenesis
3.5x10-6
Total = 363
Lipid metabolic process
4.6x10-6
Total = 97
Circulatory system development
4.3x10-5
Total = 115
Cell adhesion
6.5x10-7
Total = 136
Extracellular matrix organisation
1.1x10-8
Total = 63
Muscle structure development
1.1x10-10
Total = 160
Inflammatory response
4.7x10-9
ARMC3
LMO7
SFTPC
PECAM1
CDH11
MFAP5
CALD1
CD163
ix. Mast cell
x. Neutrophil 1
xi. Neutrophil 2
xii. T-cell
xiii. Natural killer cell
xiv. Plasma cell
xv. B-cell
xvi. Mitotic cell
Total = 17
No signicantly enriched terms
Total = 31
Defence response to bacterium
8.5x10-11
Total = 129
Immune system process
5.0x10-20
Total = 89
Adaptive immune response
1.7x10-65
Total = 55
Immune system process
9.1x10-13
Total = 185
Adaptive immune response
3.04x10-189
Total = 11
No signicantly enriched terms
Total = 153
Cell cycle
9.2x10-101
KIT
AZU1
CD177
CD8A
GZMB
MZB1
CD19
CENPF
50µm

Figure 1. Integrative co-expression analysis of unfractionated human lung tissue RNAseq can resolve constituent cell type enriched genes. (**A**) Bulk RNAseq datasets were retrieved from GTEx V8 and analysed by tissue type (n=sample number). (**B**) Analysis concept, using lung as an illustrative example: (i) each sample (n=578) contained mixed cell types, contributing (ii) differing proportions of mRNA to each sequenced dataset. To profile cell type-enriched transcriptomes (iii) constituent cell types for each tissue were identified and for each (iv) candidate reference transcripts (*Ref.T.*) for ´virtual tagging´ were shortlisted, primarily based on predicted cell specificity from existing literature and/or in house protein profiling. (v) Matrix of correlation coefficient values between selected *Ref.T.* across the sample set. (**C**) Mean correlation coefficients between genes above designated thresholds for classification as cell-type enriched in: (i) respiratory ciliated [RCC], (ii) alveolar type I [AT1], (iii) alveolar type II [AT2], (iv) endothelial [EC], (v) alveolar fibroblasts [FB1], (vi) adventitial fibroblasts [FB2], (vii) smooth muscle cell [SMC], (viii) macrophage [MC], (ix) mast cell [MastC], (x-xi) neutrophil [NP1 and NP2], (xii) T-cell [TC], (xiii) natural killer cell [NK], (xiv) plasma cell [PC], (xv) B-cell [BC], or (xvi) mitotic cell [MitC], and all *Ref.T.* panels. Total number, most significant gene ontology (GO) terms and illustrative protein profiling in human lung tissue are provided for each cell type. See also Table S1, Figure S1 and S2.

Reference transcripts analysis can identify cell-type enriched gene signatures

For each tissue type analysed, the proportion of constituent cell types between samples vary, due to sampling and genetic factors [33,34], but ratios between constitutively expressed cell-specific genes remain relatively constant. Thus, high correlation of a given transcript with all *Ref.T.* in any one panel is consistent with selective expression in the corresponding cell type [28]. For all tissues, we generated correlation coefficients between each *Ref.T.* and all other sequenced transcripts (´test-transcripts´) and produced a list of provisional cell type-enriched transcripts, based on the following criteria: (**i**) the test-transcript had a mean correlation with a given *Ref.T.* panel ≥0.50 (FDR <0.0001), which was (**ii**) higher than the mean correlation with *any other Ref.T.* panel. Resultant transcripts for each cell type were generally well separated from all others e.g., for lung: respiratory ciliated cells (RCC; Figure S1 A.i) and alveolar cell type 1 (AT1; Figure S1 Bi). However, in some cases, test-transcripts correlated well with more than one *Ref.T.* panel; panels typically representing closely related cell types, e.g., natural killer and T-cells (NK and TC; Figure S1 C.i), or those with functional commonalities, e.g., macrophages and alveolar type 2 (AT2) cells [35] (MC and AT2; Figure S1 D.i). To more carefully analyse the relationship between transcripts, the following was calculated for each to compare cell type lists: (**i**) the '*differential correlation score*', defined as the difference between the mean correlation of the test-transcript with the two sets of *Ref.T.*, e.g., respiratory ciliated cell (RCC) type panel [*ERICH3*, *DNAH12*, *SNTN*] and smooth muscle cell (SMC) panel [*TPM2*, *MYL9*, *TAGLN*] (Figure S1 A.ii) and (**ii**) the '*enrichment ranking*', based on the mean correlation value of the test-transcript with the *Ref.T.* panel (rank 1 = highest corr.). Transcripts that most highly correlated with the RCC *Ref.T.* panel separated well, from even the next closest cell type, SMC (Figure S1 A.ii), as did those most highly correlating with the alveolar cell type 1 (AT1) *Ref.T.* panel (Figure S1 B.ii). A panel of transcripts that most highly correlated with *Ref.T.* representing NK (Figure S1 C.ii, right side) or MC (Figure S1 D.ii, right side) had a low differential correlation score with *Ref.T.* for TC or AT2, respectively (Figure S1 C.ii and D.ii, left side), consistent with co-enrichment in both cell types, as we previously demonstrated [28].

scRNAseq data from human lung [36] was used to verify expression profiles of selected transcripts with predicted enrichment in one (Figure S1 A-D.iii and v) or both cell types (Figure S1 A-D.iv). For classification as single cell-type enriched, any transcript with a differential correlation score <0.15 vs. any *Ref.T.* panel representing a different cell type was excluded, on the basis of predicted co-enrichment (e.g., Figure S1 A-D.ii, grey shaded area). Application of these criteria across tissues generally resulted in intra-tissue cell-enriched gene panels that were well separated from each other (example for lung; Figure 1 C.i-xvi). For some cell types, these default thresholds were decreased when overlap with other *Ref.T*. panels was absent e.g., for erythroid cells in the liver (Figure S1 E.i and ii) or increased when overlap remained (details provided in Table S1, Tab 3). Gene ontology (GO) analysis [37], performed to identify over-represented classes and pathways among genes identified as cell type enriched produced resultant terms consistent with expected cell type functions, e.g. for lung respiratory ciliated cells, significant terms included ´*cilium organisation*´ (FDR 4.4 x10^{-63}) (Figure 1 C.i), and for plasma cells ´*adaptive immune response*´ (FDR 3.0 x10^{-189}) (Figure 1C.xiv). Tissue profiling for selected proteins encoded by predicted cell type enriched genes had expression consistent with our classifications (Figure 1 C.i-xvi).

Weighted network correlation analysis supports cell type enrichment predictions

As our analysis method is based on manually selected *Ref.T.*, cell type classification is subject to an input bias. However, we previously showed that unbiased weighted network correlation analysis (WGCNA) [38], where correlation coefficients between all transcripts are calculated and subsequently clustered into related groups (based on expression similarity), supports *Ref.T.* based analysis cell type enrichment predictions [27,28]. Here, we performed WGNCA of lung and liver samples (Figure S2). Both *Ref.T* (Figure S2 A-B.i) and predicted cell-type enriched gene panels (Figure S2 A-B.ii-ix) clustered into the same, or closely related WGCNA groups when the differential correlation for exclusion was set at >0.15 (as described above) (Figure S2 A-B.v). When the differential correlation was increased in increments of 0.05 (Figure S2 A-B.vi-ix) the number of predicted cell type enriched genes outside the predominant WGCNA clusters decreased (see red dashed box), consistent with higher enrichment specificity. Gene

enrichment could thus be categorised into *very high*, *high* or *moderate*, corresponding to a differential score vs. other profiled cell types within the tissue of >0.35, >0.25 or >0.15, respectively (see Table S1, Tab 3 for total number in each category for all cell types/tissues).

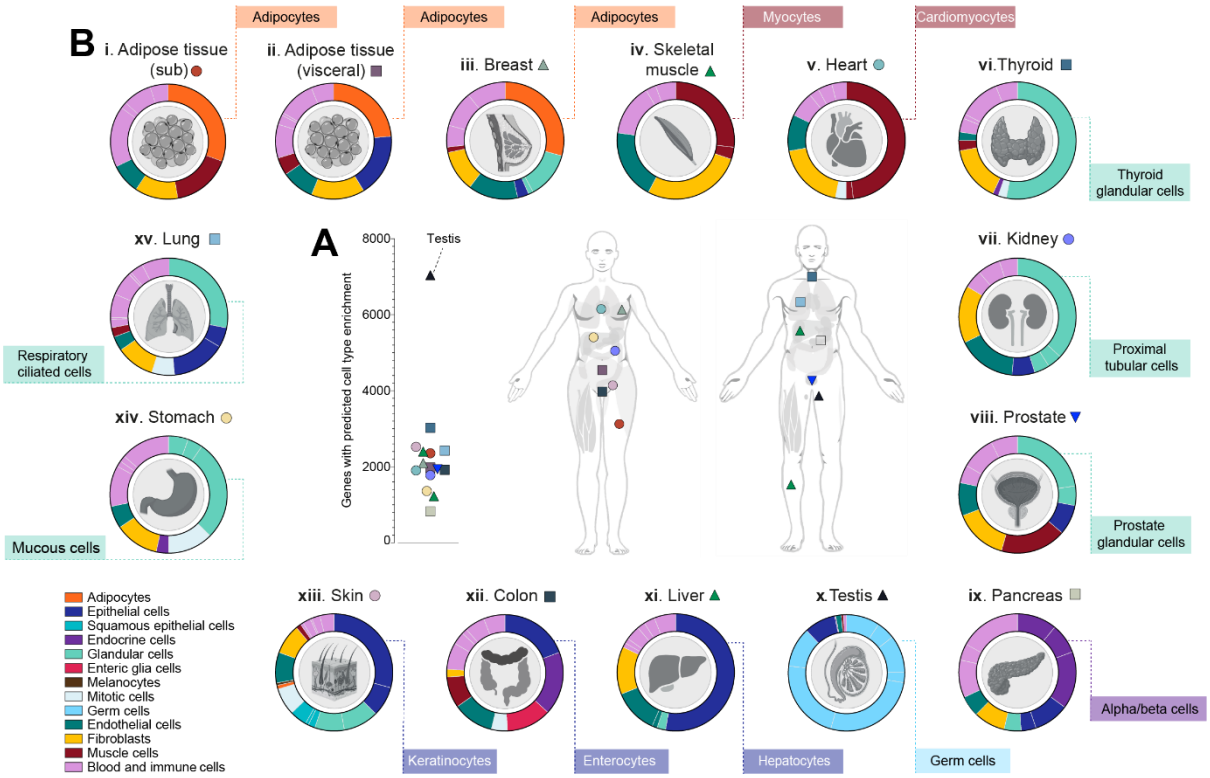
Specialised cell types have the highest number of enriched genes within tissues

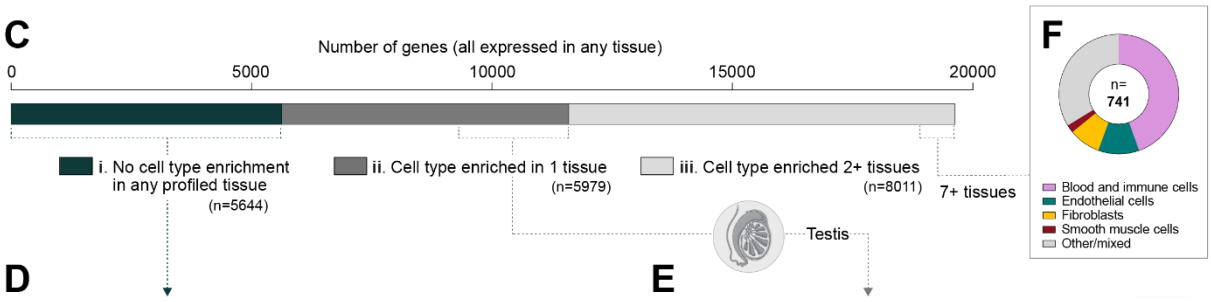
The total number of genes with predicted cell type enrichment (*very high*, *high* or *moderate*) within each tissue ranged from 7041 (testis) to 829 (pancreas) (Figure 2 A) (Table S1, Tab 3). The number of cell types analysed in each tissue type ranged from 7-18; with the lowest number profiled in skeletal muscle and subcutaneous adipose tissue (n=7 and 8, respectively) and the highest in skin and lung (n=18 and 14, respectively) (Table S1, Tab 1).Tissue specialised cell types had the highest number of enriched genes, such as cardiomyocytes in the heart (number/total enriched in all cell types in that tissue: 916/1902 [48%]) (Figure 2 B.v), proximal tubular cells in the kidney (657/1778 [37%]) (Figure 2 B.vii), hepatocytes in the liver (1264/2393 [53%]) (Figure 2 B.xi), keratinocytes in the skin (945/2460 [38.4%]) (Figure 2 B.xiii), gastric mucosal cells in the stomach (379/1361 [28%]) (Figure 2 B.xiv) and respiratory ciliated cells in the lung (681/2419 [28%]) (Figure 2 B.xv).

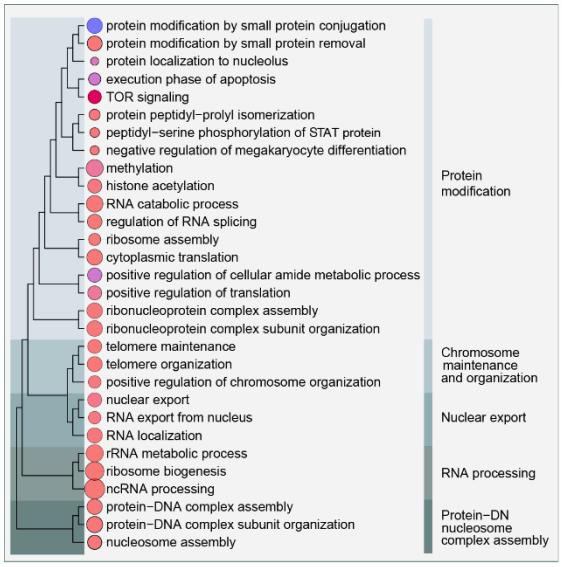
Of the 19,634 protein coding genes expressed in one or more tissues, 5644 (28.7%) were not predicted to be cell type enriched in any tissue (Figure 2 C.i). GO analysis identified the most significant over-represented pathways among these genes as ´*metabolism of RNA*´ (FDR 4.6 $x10^{-21}$), ´*gene expression (transcription)*´ (FDR 2.3 $x10^{-11}$) ´*RNA polymerase II transcription*´ (FDR 5.4 $x10^{-10}$) and ´*rRNA processing* ´ (FDR 5.8 $x10^{-10}$) (subgroups shown in Figure 2 D), consistent with housekeeping function. Indeed, 2893 of these 5644 genes (52.3%, $p<10^{-15}$) had been previously categorised as members of the housekeeping proteome [2].

5979 (30.4%) genes were classified as cell type enriched in only a single tissue (Figure 2 C.ii), the largest proportion of which were in testis (n=3141) (Table S1, Tab 4). GO term analysis of this gene group identified the most significant over-represented pathways as ´*sexual reproduction*´ (FDR 3.7 $x10^{-32}$) and ´*spermatogenesis*´ (FDR 2.9 $x10^{-30}$) (subgroups shown in Figure 2 E). Of the 8011 genes predicted to be cell type enriched in multiple tissues (Figure 2 C.iii), a small number (741, 9.2%) were enriched in seven or more; the majority of which were

predicted to be immune cell-, endothelial cell- or stromal cell- enriched (Figure 2 F), i.e., in cell types profiled in all, or most, tissues. Enrichment scores for all genes in cell types by tissue type can be found in Table S2 (summary of cell type gene enrichment across tissue in Table S1, Tab 4).


B
i. Adipose tissue (sub)
ii. Adipose tissue (visceral)
iii. Breast
iv. Skeletal muscle
v. Heart
vi.Thyroid
Adipocytes
Adipocytes
Adipocytes
Myocytes
Cardiomyocytes
Thyroid glandular cells
xv. Lung
Respiratory ciliated cells
A
Testis
Genes with predicted cell type enrichment
8000
6000
4000
2000
0
vii. Kidney
Proximal tubular cells
xiv. Stomach
Mucous cells
viii. Prostate
Prostate glandular cells
Adipocytes
Epithelial cells
Squamous epithelial cells
Endocrine cells
Glandular cells
Enteric glia cells
Melanocytes
Mitotic cells
Germ cells
Endothelial cells
Fibroblasts
Muscle cells
Blood and immune cells
xiii. Skin
xii. Colon
xi. Liver
x.Testis
ix. Pancreas
Alpha/beta cells
Keratinocytes
Enterocytes
Hepatocytes
Germ cells


C
Number of genes (all expressed in any tissue)
0
5000
10000
15000
20000
i. No cell type enrichment in any profiled tissue
(n=5644)
ii. Cell type enriched in 1 tissue
(n=5979)
iii. Cell type enriched 2+ tissues
(n=8011)
7+ tissues
Testis
F
n= 741
Blood and immune cells
Endothelial cells
Fibroblasts
Smooth muscle cells
Other/mixed


D
protein modification by small protein conjugation
protein modification by small protein removal
protein localization to nucleolus
execution phase of apoptosis
TOR signaling
protein peptidyl−prolyl isomerization
peptidyl−serine phosphorylation of STAT protein
negative regulation of megakaryocyte differentiation
methylation
histone acetylation
RNA catabolic process
regulation of RNA splicing
ribosome assembly
cytoplasmic translation
positive regulation of cellular amide metabolic process
positive regulation of translation
ribonucleoprotein complex assembly
ribonucleoprotein complex subunit organization
telomere maintenance
telomere organization
positive regulation of chromosome organization
nuclear export
RNA export from nucleus
RNA localization
rRNA metabolic process
ribosome biogenesis
ncRNA processing
protein−DNA complex assembly
protein−DNA complex subunit organization
nucleosome assembly
Protein modification
Chromosome maintenance and organization
Nuclear export
RNA processing
Protein−DN nucleosome complex assembly
p.adjust
0.0100
0.0075
0.0050
0.0025
Gene number
50
100
150
200

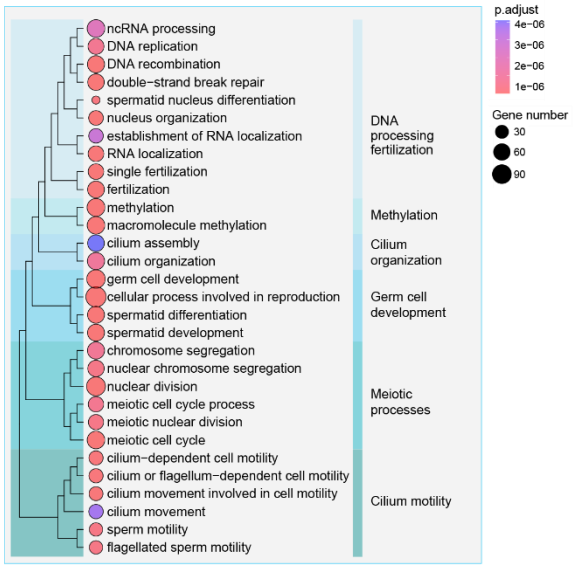

E
ncRNA processing
DNA replication
DNA recombination
double−strand break repair
spermatid nucleus differentiation
nucleus organization
establishment of RNA localization
RNA localization
single fertilization
fertilization
methylation
macromolecule methylation
cilium assembly
cilium organization
germ cell development
cellular process involved in reproduction
spermatid differentiation
spermatid development
chromosome segregation
nuclear chromosome segregation
nuclear division
meiotic cell cycle process
meiotic nuclear division
meiotic cell cycle
cilium−dependent cell motility
cilium or flagellum−dependent cell motility
cilium movement involved in cell motility
cilium movement
sperm motility
flagellated sperm motility
DNA processing fertilization
Methylation
Cilium organization
Germ cell development
Meiotic processes
Cilium motility
p.adjust
4e−06
3e−06
2e−06
1e−06
Gene number
30
60
90

Figure 2. Overview of cell type enriched gene profiles across tissue types. Bulk RNAseq datasets were retrieved from GTEx V8 and cell type enriched transcriptome predictions made using integrative correlation analysis. (**A**) Number of genes with predicted cell type enrichment in each analysed tissue type. (**B**) Circular plots showing broad classification of genes predicted to be cell type enriched in: (i) subcutaneous adipose tissue, (ii) visceral adipose tissue, (iii) breast, (iv) skeletal muscle, (v) heart, (vi) thyroid, (vii) kidney, (viii) prostate, (ix) pancreas, (x) testis, (xi) liver, (xii) colon, (xiii) skin, (xiv) stomach and (xv) lung, with majority cell types indicated in connected boxes. (**C**) Total number of expressed genes (in at least one tissue type) by respective status: (i) no cell type enrichment in any tissue, (ii) prediction as cell type enriched in one tissue, or (iii) predicted to be cell type enriched in two or more tissues. Gene ontology overrepresented terms for genes with: (**D**) no predicted cell type enrichment and (**E**) predicted enrichment only in testis. (**F**) Cell type enrichment predictions for genes classified as enriched in seven or more tissue types. See also Table S1 and S2 and Figure

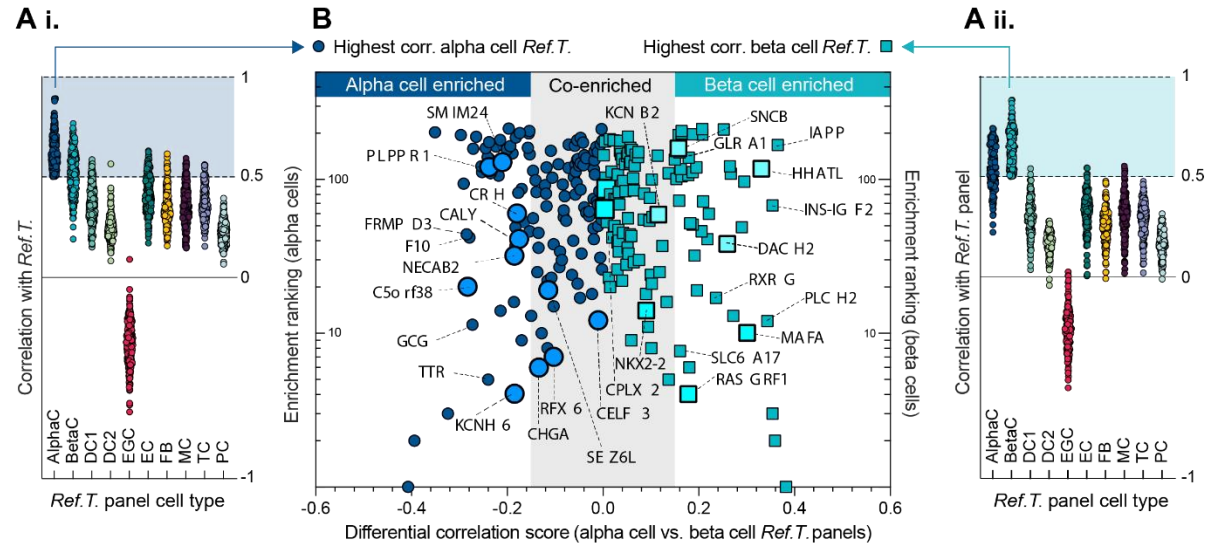
***Ref.T.* analysis can predict source of tissue enriched genes**

RNAseq data from unfractionated human tissues can be used to identify genes with higher expression in any given tissue, compared to others. For genes classified as *tissue enriched* in the Human Protein Atlas (HPA) [2], those we classified as cell type enriched were predominantly expressed in tissue specialised cell types, for example, heart enriched genes were predominantly cardiomyocyte enriched and liver-enriched genes predominantly hepatocyte enriched (Figure S3 A). A hypergeometric test was performed to determine similarity between predicted cell type enriched genes and the top 300 enriched genes in each tissue in the GTEx data [29] (as collated in the Harminozome database [39]); similar to the comparison with the HPA data, the highest statistical overlap between tissue enriched genes and cell enriched genes were predominantly with tissue specialised cell types (Figure S3 Bi-vi). This highlights the usefulness of our analysis of bulk RNAseq to disentangle cell type variance across the different tissues in the human body, independent of scRNAseq data.

Pancreatic alpha and beta cells have both specific and shared gene enrichment profiles

Alpha and beta cells, the most abundant endocrine cell types in the pancreatic islet of Langerhans [40], are defined by their expression of the blood glucose elevating or lowering hormones, glucagon (*GCG*) and insulin (*INS*), respectively. As a general rule, transcripts predicted to be cell type enriched generally separated well from others, but analysis of pancreas samples (n=328) revealed that many transcripts that correlated most highly with the alpha cell *Ref.T.* panel also correlated well with the beta cell *Ref.T.* panel (Figure 3 A.i), and *vice versa* (Figure 3 A.ii). Analysis of individual transcripts revealed 131 genes highly and selectively correlated with the *Ref.T*. panels for *both* alpha and beta-cells (Figure 3B, [grey central panel; mean differential corr. between Ref.T panels <0.15]). GO and reactome analysis [41] of these 131 co-enriched genes revealed over-represented classes and pathways included *´regulation of secretion by cell´* (FDR 7.5 $x10^{-11}$), *´neuronal system´* (FDR 9.9 $x10^{-7}$) and *´synapse´* (FDR 1.5 $x10^{-15}$) (Table S3, Tab 1, Tables A-C). Synapse related proteins (n=44) included members of the synaptotagmin (*SYT4*, *5*, *7*, *13, 14*), and glutamate receptor

(*GRIA2*,*3*) families (Table S3, Tab 2), many of which are reported to be important for pancreatic endocrine cell function e.g., *SYT4* [42] and *SYT13* [43,44], whilst the function of others in this context is not currently known e.g., *FRRS1L and NSG1.* Alpha and beta cell co-enriched genes included several encoding for transcription factors involved in islet cell specification, e.g., *NKX2-2*, [45], *NEUROD1* [46], *RFX6* [47], *INSM1* [48], *PAX6* [49] and *MYT1* [50], as well as those with no currently reported function in these cell types, e.g., *CELF3* and *MYT1L* one could speculate such genes likely have a role in neuroendocrine cell function. 91 genes had predicted alpha cell-enrichment, including *GCG*, *TTR* and *KCNH6* (Figure 3 B, left side); all of which are involved in glucose homeostasis [51-53], and other genes with, as yet, no described function in this cell type e.g., *SMIM24*, *CALY* and *C5orf38* (Figure 3 B, left side). 69 genes had predicted beta cell enrichment, including those encoding proteins with known beta cell-specific functions, e.g., *IAPP* and *MAFA* [54,55], as well as those with no reported function in this cell type, e.g., *HHATL, SNCB* and *SLC6A17* (Figure 3 B, right side). Tissue profiling for selected genes showed protein expression consistent with our classifications (Figure C-D top panel). We sourced data from scRNAseq of human pancreas [36], to compare the expression profiles of selected predicted alpha- (Figure 3 C.i-iv), beta- (Figure 3 E.i-iv) or co- (Figure 3 D.i-iv) enriched genes; categorisation was largely consistent between datasets. A small number of genes we predicted to be alpha-, beta or co-enriched had a mean expression <0.1 TPM in the analysed bulk RNAseq dataset (gene n=11, 6 and 4, respectively, Figure S4 A). Despite this low expression, our predicted expression of these genes was consistent with the scRNAseq analysis; with most (21/22 [95%]) detected predominantly in the correspondingly annotated cell types (Figure S4 C-E). However, for several of these genes, detectable expression by scRNAseq was low, or only evident in a small number of cells within the cluster, e.g., *GLB1L3* (Figure S4 C.ii). The interpretation of such scRNAseq data is challenging; thus, our classifications, based on a completely independent data collection and analysis method, can be used to verify that low or irregular detection of gene expression by scRNAseq in an annotated cell type supports a real biological phenomenon, as opposed to noise or imputation artefact.


A i.
B
A ii.
Highest corr. alpha cell Ref.T.
Highest corr. beta cell Ref.T.
Alpha cell enriched
Co-enriched
Beta cell enriched
Correlation with Ref.T.
Ref.T. panel cell type
AlphaC
BetaC
DC1
DC2
EGC
EC
FB
MC
TC
PC
Enrichment ranking (alpha cells)
Enrichment ranking (beta cells)
Differential correlation score (alpha cell vs. beta cell Ref.T. panels)
SM IM24
KCN B2
SNCB
GLR A1
IAPP
PLPP R 1
HHATL
CR H
INS-IG F2
CALY
FRMP D3
F10
DAC H2
NECAB2
C5o rf38
RXR G
PLC H2
GCG
MA FA
SLC6 A17
NKX2-2
TTR
RAS GRF1
CPLX 2
RFX 6
CELF 3
KCNH 6
CHGA
SE Z6L

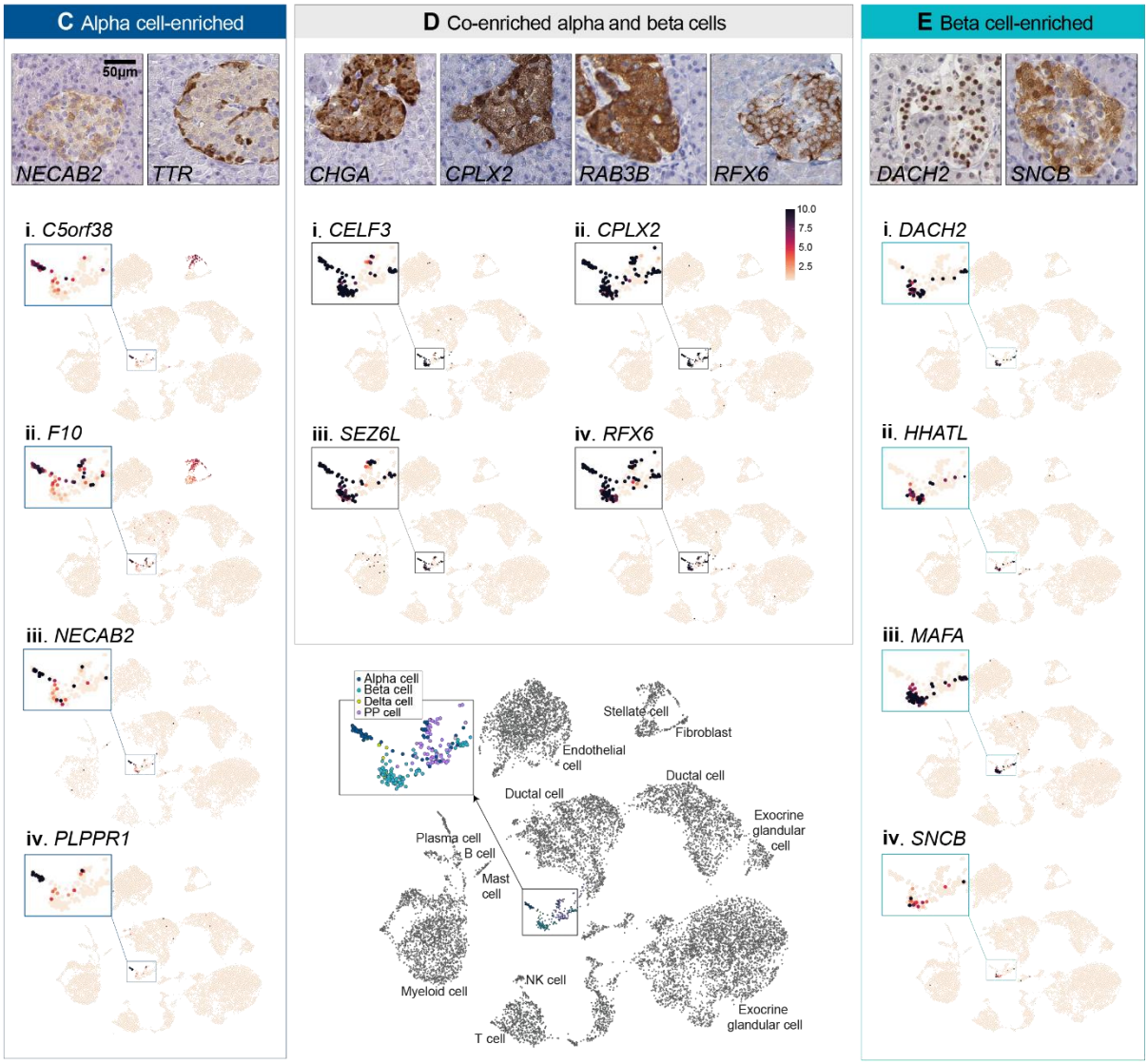

C Alpha cell-enriched
D Co-enriched alpha and beta cells
E Beta cell-enriched
50µm
NECAB2
TTR
CHGA
CPLX2
RAB3B
RFX6
DACH2
SNCB
i. C5orf38
ii. F10
iii. NECAB2
iv. PLPPR1
i. CELF3
ii. CPLX2
iii. SEZ6L
iv. RFX6
i. DACH2
ii. HHATL
iii. MAFA
iv. SNCB
Alpha cell
Beta cell
Delta cell
PP cell
Stellate cell
Fibroblast
Endothelial cell
Ductal cell
Exocrine glandular cell
Plasma cell
B cell
Mast cell
Myeloid cell
NK cell
T cell

Figure 3. Pancreatic alpha and beta cells express respective cell type enriched genes and a panel of shared co-enriched genes. RNAseq datasets for human pancreas (n=328) were retrieved from GTEx V8 and correlation coefficients between selected cell type *Ref T.* and all others were generated. Mean correlation values between protein coding genes that correlated most highly with (i) alpha or (ii) beta cell *Ref.T.* (above >0.50) and all *Ref.T.* panels. (**B**) For these transcripts, the '*differential correlation score'* (difference between mean correlation with alpha and beta cell *Ref.T.*) was plotted vs. 'enrichment ranking' (position in each respective list, highest correlation = rank 1). Shaded grey box highlights genes with differential correlation <0.15. Genes highlighted in bold correspond to those featured in the lower panels. Tissue protein profiling of selected genes predicted to be (**C**) alpha cell-enriched, (**D**) co-enriched in both alpha and beta cells, or (**E**) beta cell-enriched, in human pancreas samples. scRNAseq data from analysis of human pancreas was sourced from Tabula Sapiens [36], and used to generate UMAP plots, showing the expression profiles of example genes we predicted as being (**C**) alpha cell-enriched; (i) *C5orf58*, (ii) *F10* (iii) *NECAB2* and (iv) P*LPPR1*, (**D**) co-enriched in both alpha and beta cells; (i) *CELF3*, (ii) *CPLX2*, (iii) *SEZ6L* and (iv) *RFX6*, or (**E**) beta cell-enriched; (i) *DACH2*, (ii) *HHATL*, (iii) *MAFA* and (iv) *SNCB*. scRNAseq cell type annotations are displayed on lower central plot. AlphaC; alpha cell, BetaC; beta cell, DC1; ductal cell 1, DC2; ductal cell 2, EC; endothelial; FB1/2; fibroblast 1/2, SMC; smooth muscle cell, MC; macrophage, MastC; mast cell, NP1/2; neutrophil 1/2, TC; T-cell, PC; plasma cell. See also Table S3 and Figure S4.

Temporal changes in gene enrichment signature underlie the process of spermatogenesis

Precise definitions, markers and terminology used for the respective cell types in the different stages of spermatogenesis vary between studies. Our analysis was based on four *Ref.T.* panels (S1-S4) that were selected to represent the temporal order of development: S1, germ cell expressed [*MAGEB2*, *KDM1B, PIWIL4*] (spermatogonia), S2, meiotic cell cycle expressed [*ANKRD31*, *RBM44, TOP2A*] (spermatocytes), S3, spermatid structure-related [*CEP55, KPNA5*, *PBK*] (round/early elongating spermatids) and S4, nuclear condensation/protamine repackaging factors [*PRM1*, *PRM2*, *TNP1*] (late/elongated spermatids) (Figure 4 A and Table S1, Tab 1, Table N). When the sample set was analysed by WGNCA, *Ref.T.* within each respective panel were all in a common module (Figure S5 A). Principle component analysis of the corr. values of cell-enriched genes vs. all *Ref.T*. panels revealed the greatest proportion of variance in enrichment, and thus uniqueness vs. other cell types, was driven by cell types S1, S2, S3, S4 (Figure 4 B). Tissue profiling for proteins encoded by a panel of genes predicted to be enriched in cell types outside those in the spermatogenesis pathway revealed expression consistent with our classifications (Figure 4 C). 6179 genes were enriched in one or more of the germ cell types representing the different stages of spermatogenesis, vs. non-germ cell types (Figure S5 B.i and Table S4, Tab 1 [correlation with respective Ref.T panel >0.50, differential correlation vs. all non-germ cell types >0.15] columns H-K and Q). GO and reactome analysis of this gene list revealed that the most significantly over-represented terms included ´*sexual reproduction*´ (FDR 3.1 x10^{-27}), ´*microtubule-based processes*´ (FDR 2.2 x10^{-26}), ´*male gamete generation*´ (FDR 2.3 x10^{-26}) and ´*cell cycle*´ (FDR 4.6 x10^{-19}) (Table S4, Tab 2, Tables A and B) (Figure S5 B.ii [summary plot of GO terms, made with REVIGO [56]]). Genes that correlated with *Ref.T.* panels representing cells at different stages of spermatogenesis had two main profiles; they were enriched at a specific developmental stage, i.e., S1 (Figure 4 D.i), S2 (Figure 4 D.ii) S3 (Figure 4 D.iii) or S4 (Figure 4 D.iv) (for all see Figure S5 C .i and ii) or, they were co-enriched in adjacent cell types on the developmental trajectory: i.e., S1 and

S2 (Figure 4 D.v), S2 and S3 (Figure 4 D.vi), S3 and S4 (Figure 4 D.vii) or S2, S3 and S4 (Figure 4 D.viii) (for all see Figure S5 D.i and ii). Each plot shows five illustrative genes for each enrichment profile type (Figure 5 Ei-vii), including genes encoding for proteins with a previously reported function at the corresponding stage of spermatogenesis e.g., for S1: *FGFR3* [57], and those with no known role in this context e.g., *C19orf84*(Figure 4 E.i). Protein profiling revealed spatial distribution for those encoded by genes classified as S1, S2, S3 or S4-enriched or co-enriched, with positive signals observed progressively closer to the centre of the seminiferous tubule with each subsequent developmental stage (Figure 4 E.i-vi). GO analysis revealed over-represented classes in genes predicted to be S1, S2- or S1 & S2 enriched included developmental, cell cycle and meiotic processes (Figure 4 F.i, ii and v), whilst organelle assembly, microtubule processes and cilium and flagellum organisation and motility associated genes were over represented in S3-, S4- and S3 & S4-enriched genes (Figure 4 F.iii, iv and vii) (Table S4, Tab 3). No transcripts were predicted to be co-enriched in non-adjacent cell stages along the developmental trajectory (e.g., S1 and S3, or S2 and S4), consistent with a coordinated temporal modification in gene enrichment signatures between subsequent stages. A single gene, *MEIOC*, was predicted to be enriched in 3 stages - S2, S3 and S4. *MEIOC* is required for germ cells to properly transition to a meiotic cell cycle program, together with binding partners *YTHDC2* and *RBM46* [58]; both of which we also predicted as enriched in cells in S2 and, to a lesser extent S3 (Table S4, Tab 1). Data from scRNAseq of human adult testis [59] supported our predictions, showing *MEIOC* enrichment in cell clusters broadly corresponding to our classification of S2, S3 and S4 (Figure 4 G.i) (cell type annotation UMAP as in the original publication in Figure S5). In contrast, we predicted that the related transcript *MEIOB* had specific enrichment at stage S2 (Figure 4 D.ii and E.ii), which was also verified by scRNAseq (Figure 4 G.ii). scRNAseq for selected less well characterised genes that we predicted as enriched in either S1, S2, S3 or S4 cells (Figure S5 C.iii), or gene predicted to be co-enriched in two stages (Figure S5 D.iii) also showed agreement with our classifications. A number of genes that were predicted to be enriched in one or more of the germ cell stages were lowly expressed (n=240 with mean TPM<0.1), several of which did not

appear in the scRNAseq dataset [59], presumably due to a lack of detection. Of the 100 most lowly expressed genes for which scRNAseq data was available, most (>80%) had expression profiles consistent with our predictions in the scRNAseq data (examples in Figure S6), but in many cases transcripts were detected at low levels in only a few cells in the corresponding cluster, e.g., *FZD10* (Figure S6 D), *LEP* (Figure S6 F) and *SIGLEC15* (Figure S6 J), making interpretation of this scRNAseq data in isolation challenging. Thus, we show that analysis of bulk RNAseq can identify differentially enriched genes associated with one or multiple stages of the developmental trajectory during spermatogenesis, including genes that are likely too lowly expressed for detection or classification as cell type enriched by scRNAseq.

RNAseq data from unfractionated tissue can be used to identify genes with enriched expression in testis vs. other tissues, as we previously described [2]. The vast majority of genes with testis-enriched expression were predicted to be enriched in one or more germ cell type (845/871 [97%]), with a smaller number predicted to be enriched in sertoli (24/871 [2.8%]), Leydig (24/871 [0.1%]) or peritubular cells (24/871 [0.1%]) (Figure 4H). No testis enriched genes were classified as endothelial or macrophage-enriched in our analysis, reflecting their presence in other tissues.

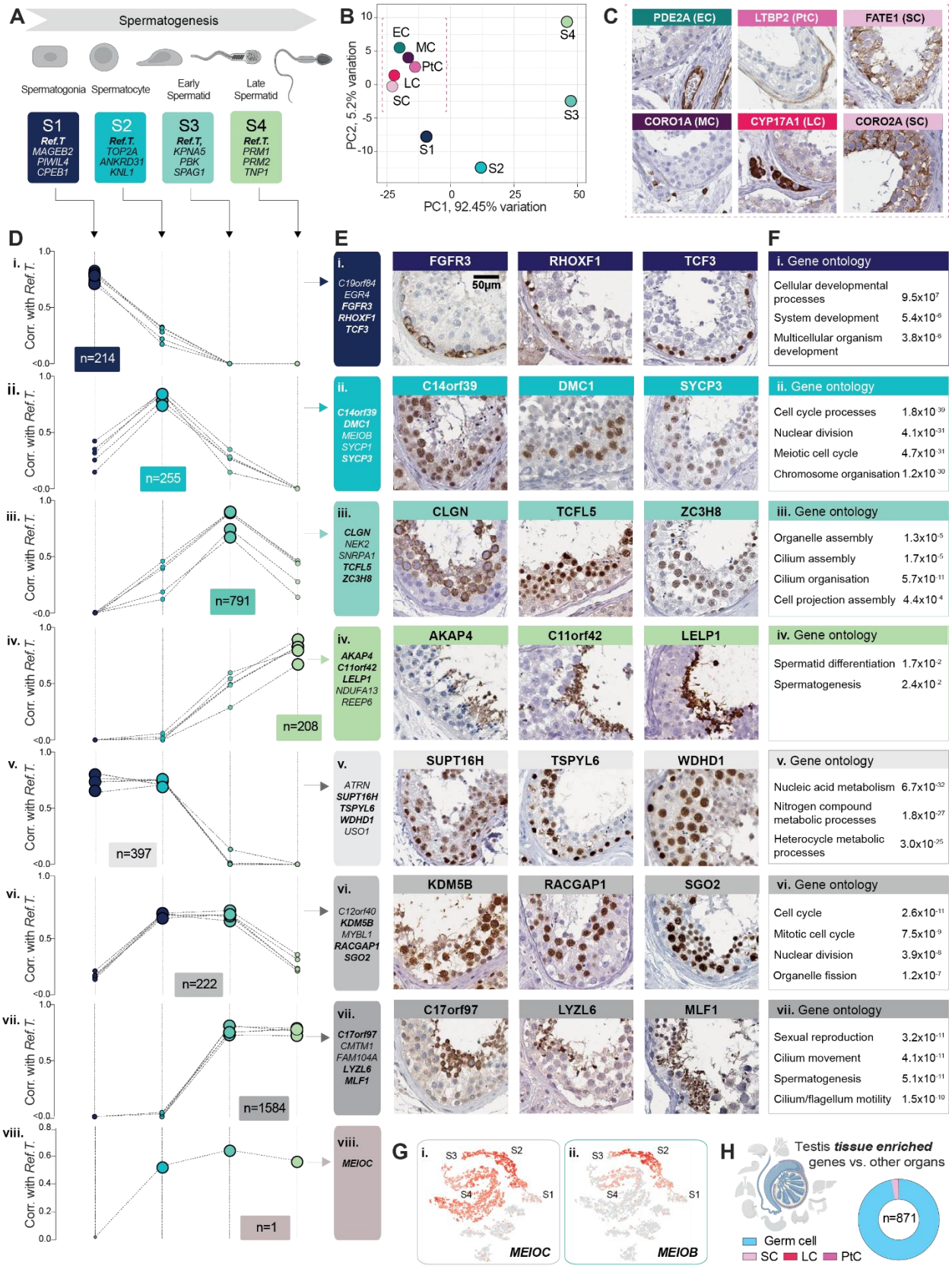

A
Spermatogenesis
Spermatogonia
Spermatocyte
Early Spermatid
Late Spermatid
S1
Ref.T
MAGEB2
PIWIL4
CPEB1
S2
Ref.T.
TOP2A
ANKRD31
KNL1
S3
Ref.T,
KPNA5
PBK
SPAG1
S4
Ref.T.
PRM1
PRM2
TNP1
B
EC
MC
PtC
LC
SC
S1
S2
S3
S4
PC2, 5.2% variation
PC1, 92.45% variation
C
PDE2A (EC)
LTBP2 (PtC)
FATE1 (SC)
CORO1A (MC)
CYP17A1 (LC)
CORO2A (SC)
D
Corr. with Ref. T.
n=214
n=255
n=791
n=208
n=397
n=222
n=1584
n=1
E
i. C19orf84 EGR4 FGFR3 RHOXF1 TCF3
ii. C14orf39 DMC1 MEIOB SYCP1 SYCP3
iii. CLGN NEK2 SNRPA1 TCFL5 ZC3H8
iv. AKAP4 C11orf42 LELP1 NDUFA13 REEP6
v. ATRN SUPT16H TSPYL6 WDHD1 USO1
vi. C12orf40 KDM5B MYBL1 RACGAP1 SGO2
vii. C17orf97 CMTM1 FAM104A LYZL6 MLF1
viii. MEIOC
FGFR3
RHOXF1
TCF3
50µm
C14orf39
DMC1
SYCP3
CLGN
TCFL5
ZC3H8
AKAP4
C11orf42
LELP1
SUPT16H
TSPYL6
WDHD1
KDM5B
RACGAP1
SGO2
C17orf97
LYZL6
MLF1
F
i. Gene ontology
Cellular developmental processes 9.5x10-7
System development 5.4x10-6
Multicellular organism development 3.8x10-6
ii. Gene ontology
Cell cycle processes 1.8x10-39
Nuclear division 4.1x10-31
Meiotic cell cycle 4.7x10-31
Chromosome organisation 1.2x10-30
iii. Gene ontology
Organelle assembly 1.3x10-5
Cilium assembly 1.7x10-5
Cilium organisation 5.7x10-11
Cell projection assembly 4.4x10-4
iv. Gene ontology
Spermatid differentiation 1.7x10-2
Spermatogenesis 2.4x10-2
v. Gene ontology
Nucleic acid metabolism 6.7x10-32
Nitrogen compound metabolic processes 1.8x10-27
Heterocycle metabolic processes 3.0x10-25
vi. Gene ontology
Cell cycle 2.6x10-11
Mitotic cell cycle 7.5x10-9
Nuclear division 3.9x10-8
Organelle fission 1.2x10-7
vii. Gene ontology
Sexual reproduction 3.2x10-11
Cilium movement 4.1x10-11
Spermatogenesis 5.1x10-11
Cilium/flagellum motility 1.5x10-10
G
i.
ii.
MEIOC
MEIOB
H
Testis tissue enriched genes vs. other organs
n=871
Germ cell
SC
LC
PtC

Figure 4. Analysis of pseudo temporal changes during spermatogenesis reveals stage-specific and common stage-shared gene enrichment signatures. (**A**) Cell types at the different stages of spermatogenesis were defined based on *Ref.T.* selected to broadly represent the developmental stage classifications spermatogonia [S1], spermatocytes [S2], early spermatids [S3] and late spermatids [S4]. (**B**) Principal component analysis of comparative correlation profiles of cell-enriched genes in S1, S2, S3, S4, sertoli cells (SC), Leydig cells (LC), peritubular cells (PtC), endothelial cells (EC) or macrophages (MC) vs. all *Ref.T*. panels. (**C**) Tissue profiling for proteins encoded by example genes we predicted to be enriched in non-germ cell types. (**D**) Pseudo trajectories of gene enrichment signatures over time, showing enrichment values for each developmental stage, using (**E**) illustrative genes predicted to be: (i) S1, (ii) S2, (iii) S3, (iv) S4, (v) S1 and S2, (vi) S2 and S3, (vii) S2, S3 and S4, enriched, with corresponding tissue protein profiling. (**F**) Over-represented gene ontology terms and significance corrected FDR values for all genes classified as: (i-iv) highly enriched at a specific stage, or (v-vii) co-enriched at one or more stages of development. (**G**) UMAP plots showing gene expression profile in the Human Testis Atlas scRNAseq data (Guo et al., 2018) of: (i) the S2, S3 and S4 predicted enriched gene *MEIOC* and (ii) the S2 predicted enriched gene *MEIOB*. (**H**) Classification of testis tissue enriched genes that we predicted to be cell type enriched. See also Table S4, Figure S5 and S6

Constituent cells of the skin hair root are the primary source of skin tissue enriched genes

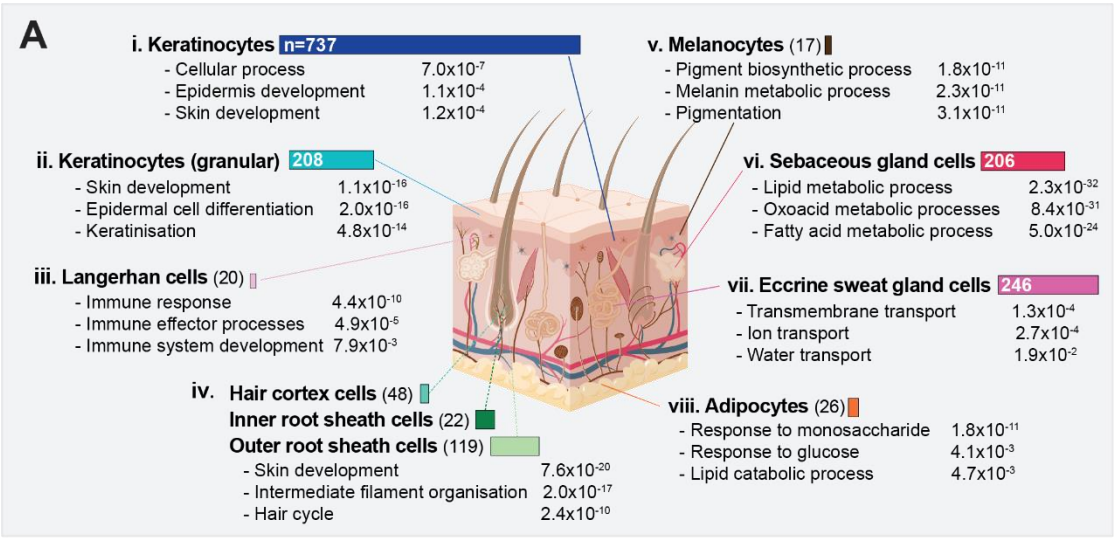
The skin is one of the most complex tissue types in the human, with multiple layers and diverse constituent cell types. We profiled 18 different cell types in the skin, many of which are not represented in scRNAseq data in Tabula Sapiens [36] or the Human Protein Atlas (HPA) (www.proteinatlas.org/) [2,3], e.g. sebaceous gland cells, eccrine sweat gland cells, adipocytes, hair cortex and inner/outer root cells. Keratinocytes expressed the largest proportion of predicted cell type-enriched genes; 737 in the sub-granular layers (Figure 5 A.i) and 208 in the granular layer (Figure 5 A.ii). Analysis of the sub-granular keratinocyte layers at a higher cell type resolution was not possible, as a *Ref.T* panel with high, consistent, specificity for sub-population(s) of basal and suprabasal keratinocytes could not be identified. Similarly, when the dataset was analysed by WGNCA, most genes we predicted to be sub-granular keratinocyte enriched clustered in multiple groups on common clades (552/737 [75%]), the constituent groups of which contained a combination of genes considered basal e.g., *COL17A1* or suprabasal e.g., *DSG1* keratinocyte markers (Figure S7 A.i). In contrast, *Ref.T.* representing granular keratinocytes and the majority of genes predicted to be enriched in this cell type (181/208 [87%]), clustered in two groups on a single clade (Figure S7 A.ii). These results are consistent with keratinocyte development being associated with a shift in absolute gene expression levels, as opposed to a defined transition between highly distinct cell states that express many specific markers (prior to terminal differentiation in the granular layer).

For genes identified as cell type enriched, GO analysis revealed over-represented classes consistent with cell type annotation, e.g. for granular keratinocytes significant terms included *´epidermal cell differentiation´* (FDR 2.0 $x10^{-16}$) (Figure 5 A.ii) and for sebaceous gland cells *´lipid metabolic processes´* (FDR 2.3 $x10^{-32}$) (Figure 5 A.vi). Of the skin-specific cell types profiled, melanocytes had the fewest enriched genes (n=17) (Figure 5 A.v), including highly expressed genes with known cell type-specific functions e.g., *PMEL*, *DCT* (mean TPM in skin RNAseq 58.2 and 29.6, respectively). More lowly expressed melanocyte-enriched genes

included *SLC24A5, CA14* and *SLC45A* (mean TPM in skin RNAseq 0.5, 1.9 and 5.7, respectively). In skin scRNAseq data from Tabula Sapiens [36] (Figure S7 B.i) *SLC24A5* was predominantly expressed in a sub-population of cells in melanocyte annotated cluster (Figure S7 B.ii), but *CA14* and *SLC45A2* were not as clearly enriched in this cell type (Figure S7 B.iii and iv). However, our classifications of these genes as melanocyte-enriched are supported by other studies showing that *SLC45A2* has a role in deacidification of maturing melanosomes to support melanin synthesis [60] and that *CA14* is downregulated in vitiligo skin samples, compared to normal, along with other genes we classified as melanocyte enriched [61]. Furthermore, all three of these genes were clustered together with the melanocyte *Ref.T* when the dataset was analysed by WGNCA (Figure S7 A.iii). Thus, as we demonstrated for alpha and beta cells in the pancreas and germ cells in the testis, our analysis can identify cell-type enriched genes that are not always detectable as such by scRNAseq.

RNAseq data from unfractionated tissue was used to identify 188 genes as skin enriched vs. other tissues in the HPA tissue section [2], of which 151 were also identified as such in a similar analysis of tissues in GTEx [29], collated in the Harminozome database [39]. Of these, 96/151 [63%] were predicted to be cell type enriched in our analysis (Figure 5 B.i); most frequently in cells of the hair root (hair cortex or inner root sheath cell), granular keratinocytes or other keratinocytes. Other skin enriched genes were predicted to be enriched in melanocytes, sweat gland or sebaceous gland cells (Figure 5 B.i). Tissue profiling of proteins encoded by selected genes supported our classifications (Figure 5 B.ii). No skin enriched genes were predicted to be cell type enriched in endothelial cells, smooth muscle cells, fibroblasts, macrophages, or other immune cell types - consistent with their presence in multiple tissue beds, and thus lack of specificality to skin. Of those cells that were skin enriched, but not classified as cell type enriched in our analysis (Figure S7 C.i [rows lacking an enlarged circle]) most had co-enrichment profiles in multiple cell types in the hair root (Figure S7 C.ii). These genes included *PSORS1C2*, a member of the psoriasis susceptibility locus [62]. Enrichment of this gene in cell types of the hair follicle is supported by studies showing that ´near naked hairless´ mice, which have a spontaneous mutation preventing the development of a normal coat, have significantly

reduced expression of *PSORS1C2* [63], together with others highlighted here e.g., *S100A3* and *KRTAP16-1* (Figure S7 C.ii) [63]. In depth skin tissue profiling showed expression of selected encoded proteins consistent with enrichment in the hair root (Figure S7 C.iii). Previously, keratinocytes, the majority cell type in the skin, have been annotated as the site of expression for the majority of skin enriched genes [3]. However, this is likely due to the lack of hair root cells in the scRNAseq data on which these classifications are based. Here, we show that a minority cell type represents the most common source of skin enriched genes.


A
i. Keratinocytes n=737
- Cellular process 7.0x10-7
- Epidermis development 1.1x10-4
- Skin development 1.2x10-4
ii. Keratinocytes (granular) 208
- Skin development 1.1x10-16
- Epidermal cell differentiation 2.0x10-16
- Keratinisation 4.8x10-14
iii. Langerhan cells (20)
- Immune response 4.4x10-10
- Immune effector processes 4.9x10-5
- Immune system development 7.9x10-3
iv. Hair cortex cells (48)
Inner root sheath cells (22)
Outer root sheath cells (119)
- Skin development 7.6x10-20
- Intermediate filament organisation 2.0x10-17
- Hair cycle 2.4x10-10
v. Melanocytes (17)
- Pigment biosynthetic process 1.8x10-11
- Melanin metabolic process 2.3x10-11
- Pigmentation 3.1x10-11
vi. Sebaceous gland cells 206
- Lipid metabolic process 2.3x10-32
- Oxoacid metabolic processes 8.4x10-31
- Fatty acid metabolic process 5.0x10-24
vii. Eccrine sweat gland cells 246
- Transmembrane transport 1.3x10-4
- Ion transport 2.7x10-4
- Water transport 1.9x10-2
viii. Adipocytes (26)
- Response to monosaccharide 1.8x10-11
- Response to glucose 4.1x10-3
- Lipid catabolic process 4.7x10-3

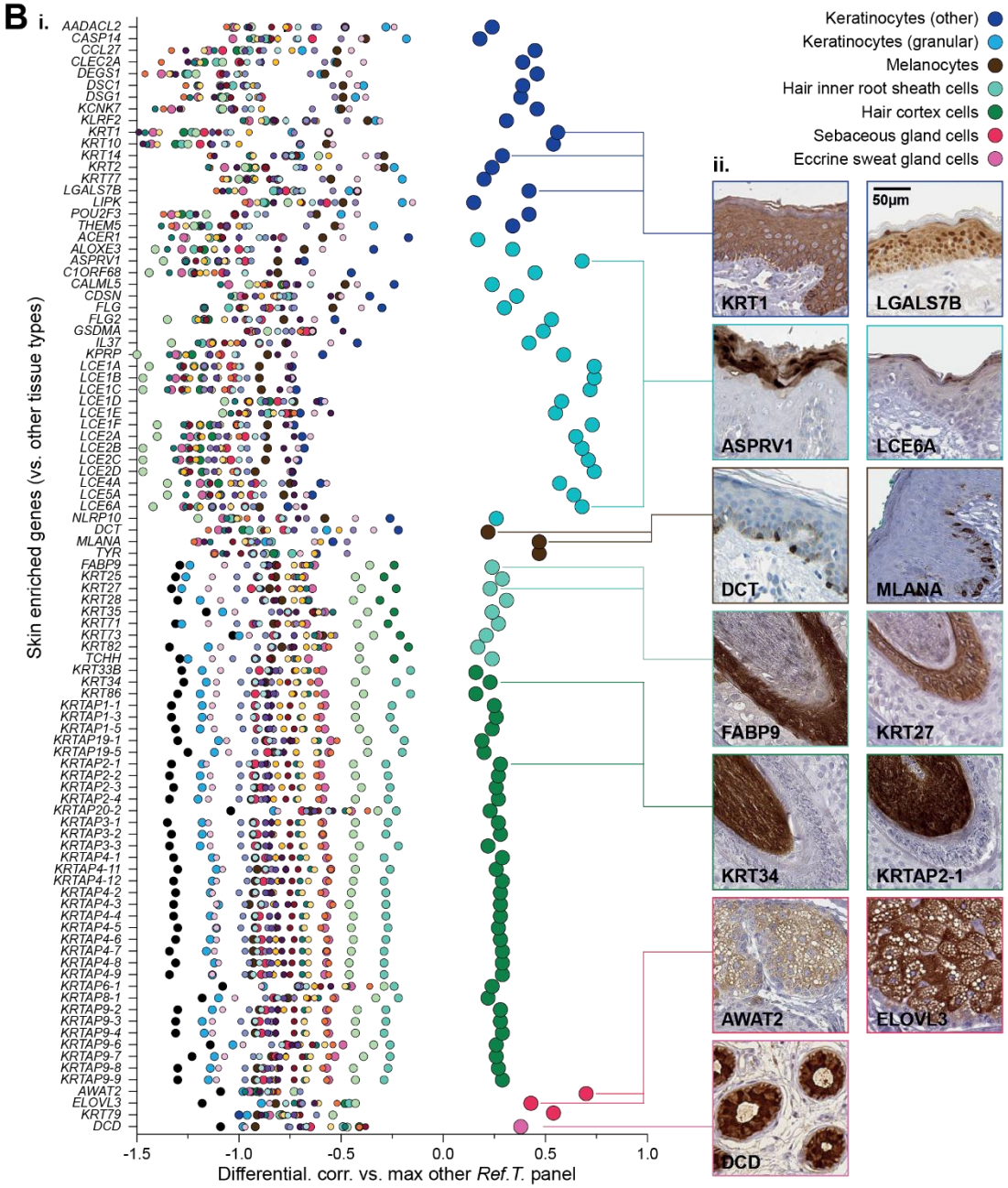

B i.
Keratinocytes (other)
Keratinocytes (granular)
Melanocytes
Hair inner root sheath cells
Hair cortex cells
Sebaceous gland cells
Eccrine sweat gland cells
Skin enriched genes (vs. other tissue types)
Differential. corr. vs. max other Ref.T. panel
-1.5
-1.0
-0.5
0.0
0.5
1.0
ii.
50µm
KRT1
LGALS7B
ASPRV1
LCE6A
DCT
MLANA
FABP9
KRT27
KRT34
KRTAP2-1
AWAT2
ELOVL3
DCD

Figure 5. Constituent cells of the skin hair root are the primary source of skin tissue enriched genes. (**A**) Number of genes predicted to be cell type enriched, and corresponding over-represented gene ontology terms and significance corrected FDR values, for skin specialised cell types profiled: (i) keratinocytes, (ii) granular keratinocytes, (iii) Langerhans cells, (iv) hair cortex cells, inner and outer root sheath cells, (v) melanocytes, (vi) sebaceous gland cells (vii) eccrine sweat gland cells and (viii) adipocytes. (**B**) (i) Genes enriched in skin vs. other organs, which were predicted to be cell type enriched in our analysis, were plotted to show the min differential values between the mean correlation coefficients with the *Ref.T.* panels for each cell type. Enlarged circles represent classification as predicted cell type enriched. (**ii**) Tissue profiling for proteins encoded by skin tissue enriched genes with predicted enrichment in the indicted cell type. See also Figure S7.

Cross-tissue analysis reveals similarities in cell type gene enrichment signatures

8011 genes were predicted to be cell type enriched in more than two tissue types. To explore the relationship between these cell type gene enrichment signatures, we performed a hypergeometric test to determine the degree of similarity between all cell types in all tissues. As cell type gene enrichment signatures are generated via a correlation-based analysis, independent of cell type absolute gene expression levels, such comparisons between tissue datasets can be made without correction for batch effects, the analysis platform used, or requirement for normalisation.

Organ-specific cell types can have common gene enrichment signature panels

Organ specific cell types (i.e. excluding those found in all or multiple organs, e.g., endothelial cells, fibroblasts and immune cell types) had gene enrichment signatures with: little or no similarity to other cell types e.g., hair inner root sheath cells and melanocytes (Figure 6 A.ii and iii), significant similarity to one other cell type, e.g., skeletal myocytes and cardiomyocytes (Figure 6 A.iv) or significant similarity with multiple other cell types, e.g., endocrine cells from several tissues; alpha and beta cells from the pancreas, enteroendocrine cells from the colon and stomach, and parafollicular cells from the thyroid (Figure 6 A.vi). We found a significant overlap between the enriched gene signatures of adipocytes (subcutaneous adipose, visceral adipose and breast), skin sebaceous gland cells, liver hepatocytes, and kidney proximal tubular cells (Figure 6 A.vii). GO analysis of the 41 genes predicted to be enriched in at least three of the aforementioned cell types (Figure 6 B.i [green box]) (Table S5, Tab 1, Table A) revealed significant terms all related to metabolic processes, including *´carboxylic acid metabolic processes´* (FDR 8.8 $x10^{-26}$) and *´organic acid metabolic processes´* (FDR 9.4 $x10^{-26}$) (Table S5, Tab 1, Table B) (Figure 6 B.ii). 22 of these 41 genes were also predicted to be enriched in cardiomyocytes, another highly metabolically active cell type with a significant overlap in gene enrichment signature with both adipocytes and hepatocytes (Figure 6 A and Table S5, Tab 1, Table A). Illustrative protein profiling showed selective expression of ACO1 and HADH in adipocytes in adipose tissue, sebaceous gland cells in skin, hepatocytes in liver and proximal tubular cells in kidney (Figure 6 B.iii). The enrichment of such genes in many

highly metabolically active cells is consistent with a common shared function across tissue types. In contrast, cell type enriched genes classified as such in only one tissue are likely key for highly specialised cell functions, e.g., complement and coagulation factor genes were predicted to be enriched only in hepatocytes (including *C4B*, *C8A*, *C9*, *CFHR1/2/4/5* and others) and specific solute transporters only in proximal tubular cells (e.g., *SLC13A1*, *SLC22A13*, *SLC22A6*, *SLC22A8*). Tissue profiling for proteins encoded by example genes predicted to be enriched in only one of these four cell types; adipocytes (*PRKAR2B*), sebaceous gland cells (*TMEM97*), hepatocytes (*OTC*) or proximal tubular cells (*TMEM174*) showed positive staining in only the respective cell types (Figure 6 B.iv). In contrast to ACO1 and HADH, which were expressed mean TMP >10 in the RNAseq datasets analysed (Figure 6 B.v), expression values of these example genes were highest in the corresponding tissue, with low or no expression in the others (Figure 6 B.vi).

Our analysis also revealed a significant overlap between the gene enrichment signatures of early and late spermatids in the testis and respiratory ciliated cells in the lung (Figure 6 A.v and C.i). GO term analysis of these 441 shared genes (Table S5, Tab 2, column A-B) revealed the most significant terms were related to cilia (Figure 6 C.ii), which are important for both clearance of fluid from the airways and movement of the sperm flagellum, including *´cilium organisation´* (FDR 3.6 x10^{-69}) and *´cilium movement´* (FDR 9.5 x10^{-64}) (Table S5, Tab 2, Table 1). The top 50 genes predicted to be most highly enriched in both early and late spermatids and RCC (Figure 6 C.iii) had variable absolute expression in the respective tissues. *LMNTD1* and *MROH9* had very low expression in the lung RNAseq (mean TMP 0.42 and 0.68, respectively) (Figure 6 C.iii) and scRNAseq data from human lung [36] revealed highly specific, but variable expression (or detection) of these genes in RCC (Figure S8 A.ii and iii). Predicted expression in S3 and S4 cells in testis was also supported by scRNAseq from the Human Testis Atlas [59] (Figure S8 B.ii and iii). Despite the highly specific enrichment profiles of *LMNTD1* and *MROH9*; neither were predicted to be enriched in any other cell type across all tissues analysed (Figure 6 C.iii), there are no existing studies of these genes in this context. Some other genes with highly predicted enrichment in early and late spermatids and RCC were also

predicted to be enriched in several other cell types e.g., *PACRG* (Figure 6 C.iii), which is well studied in the context of motile cilia, particularly in sperm [64], but has also been proposed to have other roles, such as in primary cilia [65] and even inflammatory pathway signalling [66]; perhaps explaining its more widespread enrichment profiles in our analysis. Tissue profiling for proteins encoded by example genes enriched in both RCC and early and late spermatids (Figure 6 C.iv), or RCC only (Figure 6 C.v) showed expression consistent with our predictions. GO analysis of genes predicted to be highly enriched in spermatids, but not RCC, revealed the most significant terms were unrelated to cilia formation, including ´*spermatogenesis*´ (FDR 1. $x10^{-25}$), ´*multicellular organism reproduction*´ (FDR 1.5 $x10^{-19}$), ´*spermatid development*´ (FDR 6.7 $x10^{-14}$) and ´*fertilisation*´ (FDR 1.4 $x10^{-10}$) (Table S5, Tab 3, Column A-B and Table 1); reflecting an enrichment for genes with highly specialised function within the testis only, e.g., *CALR3*, *LELP1* and *SMCP* (Figure 6 C.vi).

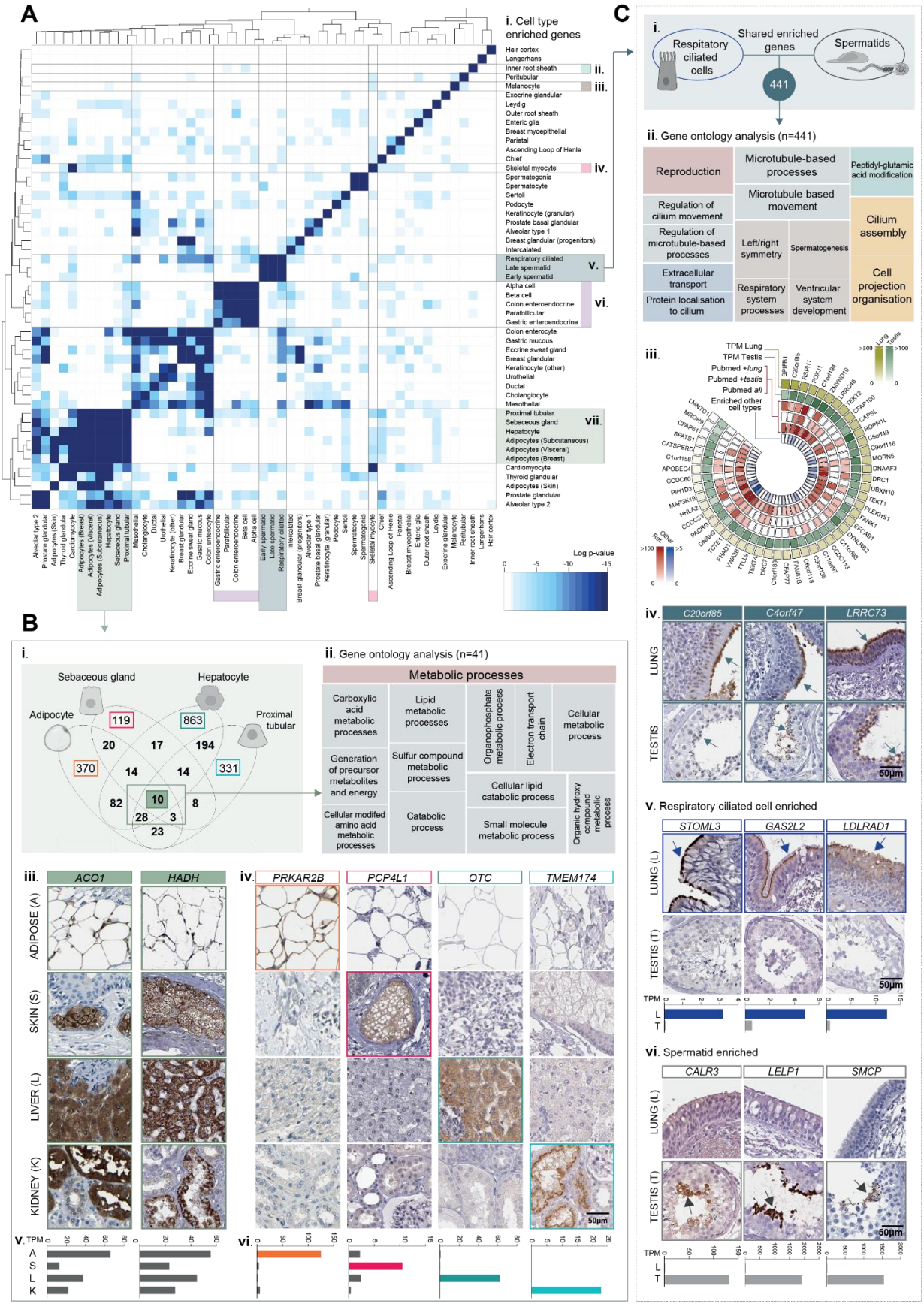

A
i. Cell type enriched genes
ii.
iii.
iv.
v.
vi.
vii.
Log p-value
B
i.
Sebaceous gland
Hepatocyte
Adipocyte
Proximal tubular
119
863
20
17
194
370
14
14
331
82
10
8
28
3
23
ii. Gene ontology analysis (n=41)
Metabolic processes
Carboxylic acid metabolic processes
Lipid metabolic processes
Organophosphate metabolic process
Electron transport chain
Cellular metabolic process
Generation of precursor metabolites and energy
Sulfur compound metabolic processes
Cellular lipid catabolic process
Organic hydroxy compound metabolic process
Cellular modifed amino acid metabolic processes
Catabolic process
Small molecule metabolic process
iii.
ACO1
HADH
ADIPOSE (A)
SKIN (S)
LIVER (L)
KIDNEY (K)
iv.
PRKAR2B
PCP4L1
OTC
TMEM174
50µm
v. TPM
vi.
C
i.
Respitatory ciliated cells
Shared enriched genes
Spermatids
441
ii. Gene ontology analysis (n=441)
Reproduction
Microtubule-based processes
Peptidyl-glutamic acid modification
Microtubule-based movement
Regulation of cilium movement
Cilium assembly
Regulation of microtubule-based processes
Left/right symmetry
Spermatogenesis
Extracellular transport
Cell projection organisation
Respiratory system processes
Ventricular system development
Protein localisation to cilium
iii.
TPM Lung
TPM Testis
Pubmed +lung
Pubmed +testis
Pubmed all
Enriched other cell types
iv.
C20orf85
C4orf47
LRRC73
LUNG
TESTIS
50µm
v. Respiratory ciliated cell enriched
STOML3
GAS2L2
LDLRAD1
LUNG (L)
TESTIS (T)
50µm
TPM
L
T
vi. Spermatid enriched
CALR3
LELP1
SMCP
LUNG (L)
TESTIS (T)
50µm
TPM
L
T

Figure 6. Organ specific cell types can have gene enrichment signature similarities. (**A**) Heatmap showing significance p-values for similarity scores between predicted cell type enriched genes, calculated using a hypergeometric test for: (i) all organ-specific cell type enriched gene signatures, (ii) skin inner root sheath hair cells, (iii) skin melanocytes, (iv) skeletal myocytes, (v) lung respiratory ciliated cells and testis spermatids, (iv) pancreatic alpha and beta cells, colon enteroendocrine cells, thyroid parafollicular and stomach gastric enteroendocrine cells and (vii) kidney proximal tubular cells, sebaceous gland cells, hepatocytes, and adipocytes in subcutaneous adipose tissue, visceral adipose tissue and breast. (**B**) (i) Number of individual or common enriched genes for adipocytes (in at least 2/3 tissues profiled) kidney proximal tubular cells, sebaceous gland cells and hepatocytes. (ii) Over-represented gene ontology terms among the 41 genes featuring in the gene enrichment signature of at least 3/4 cell types, displayed in tree map format, generated using REVIGO (areas proportional to Log10 significance values). Tissue profiling for proteins encoded by genes that were: (iii) part of the shared gene enrichment signature [ACO1, HADH] or (iv) classified as cell type enriched only in one cell type [PRKAR2B, TMEM97, OTC, TMEM174] and corresponding mean TMP expression in the corresponding tissue RNAseq datasets (v) and (vi), respectively. (**C**) (i) Number of genes that featured in gene enrichment signatures for respiratory ciliated cells, early and late spermatids and (ii) the over-represented gene ontology terms among these shared enriched genes. (iii) Circular plot showing up to the top 50 most enriched genes in respiratory ciliated cells and spermatids, displaying the mean TMP values in the lung and testis RNAseq datasets, the number of mentions in Pubmed of gene *and* corresponding tissue (´*Pubmed* + *lung/testis*´) or the gene alone (´*Pubmed all*´), and the number of other cell types in which the gene was also predicted to be enriched (´*enriched other cell type*´). Tissue profiling for proteins encoded by genes with predicted enrichment in: (iv) both respiratory ciliated cells and spermatids, (v) respiratory ciliated cells only and (vi) spermatids only and corresponding mean TMP values in the tissue RNAseq. See also Table S5 and Figure S8.

Core cell types have common gene enrichment signature panels across tissues

Eight cell types were profiled in all, or most, tissue types (termed “core cell types”); endothelial cells [n=15 tissues], smooth muscle cells [n=10], fibroblasts (including hepatic stellate cells (HSC) in the liver, and adipose progenitor cells (APC) in adipose tissue) [n=14], macrophages (including Kupffer cells in the liver) [n=15], neutrophils [n=8], mast cells [n=5], T-cells [n=13] and plasma cells [n=14] (Figure 7 A.i). Gene enrichment signatures of the same core cell type in different tissues had high similarity, with little, or no, crossover between different cell types (Figure 7A). Notable exceptions included hepatic stellate cells (HSC) and fibroblasts in liver and kidney, respectively, which had some commonality with smooth muscle cell gene enrichment signatures in other tissues (Figure 7 A.ii), in line with reports that liver HSC can have contractile properties [67] and potentially reflecting the presence of a kidney myofibroblast-like population, and lung neutrophils, which had some similarity to macrophages in several other tissues (Figure 7 A.iii). Enrichment signatures of core cell types had little or no cross over with those of organ specific cell types (Figure S8 C.i), except for lung macrophages, which had a significant similarity with the cell type group we previously identified as having shared gene enrichment signatures related to metabolic processes (Figure 6 B), including adipocytes, hepatocytes, proximal tubular cells (Figure S8 C.ii). One could speculate that this indicates macrophages in the lung have specific metabolic characteristics, in keeping with recent studies indicating that their metabolic responses to infectious pathogens or other insults may be distinct from other macrophage subtypes [68].

Endothelial cells had strong gene enrichment signature similarities across tissues (Figure 7 A), with the exception of liver sinusoidal endothelial cells (LSEC), where over half of the enriched genes (19/34 [56%]) were not enriched in endothelial cells in any other tissue, consistent with their unique structural and phenotypic features, and highly specialised function [69]. Despite this, overall, they did have greatest similarity with vascular endothelial cells vs. any other core cell type (Figure 7 A.iv). Tissue profiling for proteins encoded by *CLEC4G* (Figure 7 B.i) and *CD36*

(Figure 7 B.ii) showed expression consistent with our predictions of LSEC enrichment only, or vascular endothelial and LSEC enrichment, respectively.

To define key components of the gene enrichment signature for each core cell type, we identified genes predicted to enriched in at least half of the tissues profiled (Table S6), e.g., in at least 8/15 tissues for EC and MC (Figure 7 C-Gi and ii and Table S6, Tab 1). To assess existing reports for each gene in each given cell type, we used PubMed to search for the number of studies citing both gene name and cell type together (Figure 7C-G.iii), or gene name alone (Figure 7 C-G.iv). Many were well characterised genes on which a plethora of studies have been performed, e.g., *CD3E* and *CD2* in T-cells (Figure 7 C.iii and iv), others were poorly studied in the cell type context. For example, *SHANK3,* predicted to be endothelial cell enriched in 10 tissues (Figure 7 D), has been researched predominantly in the context of neurons and autism [70]. *TSPAN7* (Figure 7D), predicted to be endothelial cell enriched in 8 tissues, has only been identified in endothelial cells in the context of tumour associated vasculature and metastasis [71]. There is little information in the literature about the function of *LRRN4CL* (Figure 7 E), a gene we predicted to be fibroblast enriched in 7 tissues, except for its elevated expression in skin melanoma metastases and breast cancer samples [72,73]. In contrast, *MFAP4* (Figure 7 E) is a well-known gene in this cell context [74]. *TBXAS1* was identified as a macrophage core enriched gene, and its enzymatic product, thromboxane A2, is linked to vasoconstriction and platelet aggregation, with links to innate immunity [75], but little knowledge exists in the macrophage context. Tissue profiling for proteins encoded by predicted endothelial enriched genes *SHANK3* (Figure 7 D.v) and *TSPAN7* (Figure 7 D.vi), fibroblast enriched genes *LRRN4CL* (Figure 7 E.v) and *MFAP4* (Figure 7 E.vi) and the macrophage enriched gene *TBXAS1* (Figure 7 G.vi), revealed selective expression consistent with our predictions.

Whereas most core endothelial and fibroblast enriched genes were not predicted to be enriched in any other cell types in our analysis (Figure 7 D-E.v), several T-cell (Figure 7 C.v) smooth muscle cell (Figure 7 F.v), and macrophage (Figure G.v) enriched genes where predicted to be enriched in an additional cell type(s). *BCL11B*, a gene we predicted to be T-

cell enriched gene in 10 tissues (Figure 7 C) with a known role in T-cell development [76], was also predicted to be enriched in skin keratinocytes, consistent with its role in dermal development in mice [77], and unexpectedly, in the absence of any existing reports, in spermatogonia in the testis; expression profiles we verified by protein profiling (Figure 7 C.vii). *AIF1*, predicted to be macrophage enriched in 12 tissues (Figure 7 G.i and viii), consistent with its known expression in this cell type [78], was also classified as kidney podocyte enriched; previously reported in only a single study [79], a prediction we again verified with tissue protein profiling (Figure 7 G.vi). Of all the core cell types, smooth muscle cell enriched genes were most likely to have predicted enrichment in another cell type, most frequently cardiomyocytes, skeletal myocytes or breast myoepithelial cells, e.g., *SYNM* (Figure 7 F.v) and *TPM1* (Figure 7 F.vi).

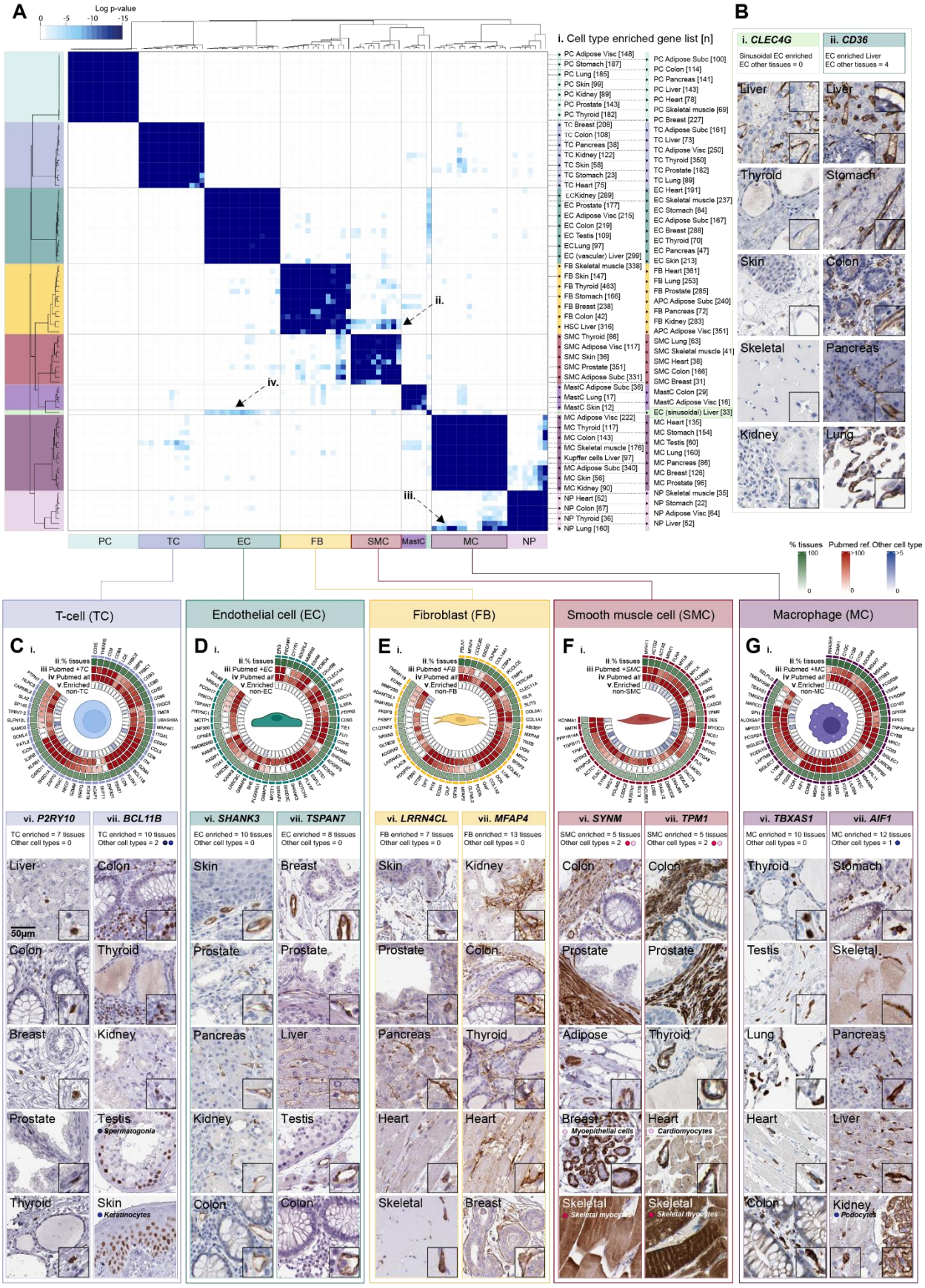

A
Log p-value
0 -5 -10 -15
i. Cell type enriched gene list [n]
PC Adipose Visc [148]
PC Stomach [187]
PC Lung [185]
PC Skin [99]
PC Kidney [89]
PC Prostate [143]
PC Thyroid [182]
TC Breast [208]
TC Colon [108]
TC Pancreas [38]
TC Kidney [122]
TC Skin [58]
TC Stomach [23]
TC Heart [75]
EC Kidney [289]
EC Prostate [177]
EC Adipose Visc [215]
EC Colon [219]
EC Testis [109]
EC Lung [97]
EC (vascular) Liver [299]
FB Skeletal muscle [338]
FB Skin [147]
FB Thyroid [463]
FB Stomach [166]
FB Breast [238]
FB Colon [42]
HSC Liver [316]
SMC Thyroid [88]
SMC Adipose Visc [117]
SMC Skin [36]
SMC Prostate [351]
SMC Adipose Subc [331]
MastC Adipose Subc [36]
MastC Lung [17]
MastC Skin [12]
MC Adipose Visc [222]
MC Thyroid [117]
MC Colon [143]
MC Skeletal muscle [176]
Kupffer cells Liver [97]
MC Adipose Subc [340]
MC Skin [56]
MC Kidney [90]
NP Heart [52]
NP Colon [87]
NP Thyroid [36]
NP Lung [160]
PC Adipose Subc [100]
PC Colon [114]
PC Pancreas [141]
PC Liver [143]
PC Heart [78]
PC Skeletal muscle [69]
PC Breast [227]
TC Adipose Subc [161]
TC Liver [73]
TC Adipose Visc [250]
TC Thyroid [350]
TC Prostate [182]
TC Lung [89]
EC Heart [191]
EC Skeletal muscle [237]
EC Stomach [84]
EC Adipose Subc [167]
EC Breast [288]
EC Thyroid [70]
EC Pancreas [47]
EC Skin [213]
FB Heart [361]
FB Lung [253]
FB Prostate [285]
APC Adipose Subc [240]
FB Pancreas [72]
FB Kidney [283]
APC Adipose Visc [351]
SMC Lung [63]
SMC Skeletal muscle [41]
SMC Heart [38]
SMC Colon [166]
SMC Breast [31]
MastC Colon [29]
MastC Adipose Visc [16]
EC (sinusoidal) Liver [33]
MC Heart [135]
MC Stomach [154]
MC Testis [60]
MC Lung [160]
MC Pancreas [86]
MC Breast [126]
MC Prostate [96]
NP Skeletal muscle [35]
NP Stomach [22]
NP Adipose Visc [54]
NP Liver [52]
i.
ii.
iii.
iv.
PC TC EC FB SMC MastC MC NP
B
i. CLEC4G
Sinusoidal EC enriched
EC other tissues = 0
ii. CD36
EC enriched Liver
EC other tissues = 4
Liver
Liver
Thyroid
Stomach
Skin
Colon
Skeletal
Pancreas
Kidney
Lung
% tissues
Pubmed ref.
Other cell type
T-cell (TC)
Endothelial cell (EC)
Fibroblast (FB)
Smooth muscle cell (SMC)
Macrophage (MC)
C i.
D i.
E i.
F i.
G i.
ii. % tissues
iii Pubmed +TC
iii Pubmed +EC
iii Pubmed +FB
iii Pubmed +SMC
iii Pubmed +MC
iv Pubmed all
v Enriched non-TC
v Enriched non-EC
v Enriched non-FB
v Enriched non-SMC
v Enriched non-MC
vi. P2RY10
TC enriched = 7 tissues
Other cell types = 0
vii. BCL11B
TC enriched = 10 tissues
Other cell types = 2
vi. SHANK3
EC enriched = 10 tissues
Other cell types = 0
vii. TSPAN7
EC enriched = 8 tissues
Other cell types = 0
vi. LRRN4CL
FB enriched = 7 tissues
Other cell types = 0
vii. MFAP4
FB enriched = 13 tissues
Other cell types = 0
vi. SYNM
SMC enriched = 5 tissues
Other cell types = 2
vii. TPM1
SMC enriched = 5 tissues
Other cell types = 2
vi. TBXAS1
MC enriched = 10 tissues
Other cell types = 0
vii. AIF1
MC enriched = 12 tissues
Other cell types = 1
Liver
Colon
50µm
Colon
Thyroid
Breast
Kidney
Prostate
Testis
Spermatogonia
Thyroid
Skin
Keratinocytes
Skin
Breast
Prostate
Prostate
Pancreas
Liver
Kidney
Testis
Colon
Colon
Skin
Kidney
Prostate
Colon
Pancreas
Thyroid
Heart
Heart
Skeletal
Breast
Colon
Colon
Prostate
Prostate
Adipose
Thyroid
Breast
Myoepithelial cells
Heart
Cardiomyocytes
Skeletal
Skeletal myocytes
Skeletal
Skeletal myocytes
Thyroid
Stomach
Testis
Skeletal
Lung
Pancreas
Heart
Liver
Colon
Kidney
Podocytes

Figure 7. Core cell types share gene enrichment signatures across organs. (**A**) Heatmap showing significance p-values for similarity scores in cell type gene enrichment signatures, calculated using a hypergeometric test, between (**i**) plasma cells (PC), T-cells (TC), endothelial cells (EC), fibroblasts (FB), smooth muscle cells (SMC), mast cells (MastC), macrophages (MC) and neutrophils (NP) in different tissues. (**B**) Tissue profiling for proteins encoded by (i) the sinusoidal EC enriched gene *CLEC4G* and (ii) the vascular and sinusoidal EC enriched gene *CD36*, in different tissue beds. Circular plots showing up to the top 50 genes most frequently predicted as enriched in (**C**) TC, (**D**) EC, (**E**) FB, (**F**) SMC and (**G)** MC in different organs, displaying (i) the percentage of tissues in which the gene was classified as enriched in the given cell type (´*% tissues*´), the number of mentions in Pubmed of (ii) gene *and* corresponding core cell type (´*Pubmed + cell type*´) together or (iii) the gene alone (´*Pubmed all*´), and (iv) the number of other cell types (including non-core cell types) in which the gene was also predicted as enriched (´*enriched non-cell type*´). (v-vi) Tissue profiling of proteins encoded by selected genes predicted to be core cell type enriched. See also Table S6 and Figure S8.

DISCUSSION

Here we present a tissue-centric, cell type gene enrichment atlas, generated from the analysis of hundreds of biological replicates. Although it is frequently stated that cell-type gene expression profiles cannot be extracted from bulk RNAseq, e.g., [80,81], here we have identified cell type enriched or co-enriched genes, and charted temporal transcriptome changes underlying cell type differentiation. We made comparisons between cell type enrichment signatures across tissues, without the requirement for normalisation or batch effect adjustments, a significant issue when handling scRNAseq datasets, for which currently no universal solution [82,83]. Our analysis included cell types that are difficult to extract from tissue, e.g., adipocytes, and those that are sensitive to processing, e.g., kidney podocytes; issues that can hinder analysis [8,10], but are circumvented here as cell removal from tissue was not required. We identify lowly expressed transcripts as cell type enriched, many of which can be detected only in a small minority of cells annotated as a given type by scRNAseq, possibly due to limited read depth and high number of drop-out events [18]. Transcript level alone is not sufficient to predict protein levels [84] and so potential function of proteins encoded by such genes may have been overlooked.

Our study is the only cell type gene enrichment atlas generated independently of scRNAseq. Comparison of scRNAseq datasets generated from analysis of the same tissue type can reveal surprisingly low agreement between studies [22,23], possibly due to the low number of samples typically analysed, and the associated lack of biological varience. For top cell type enriched genes in adipose tissue, agreement between data generated using our analysis method and several scRNAseq studies was equivalent or greater than between the scRNAseq studies themselves [28]. This could reflect the large sample set analysed and the associated biological variance represented. Our method also has scope for well-powered comparisons of cell type enrichment profiles between healthy and disease states, sexes [28], ages, developmental stages, or metabolic states, using existing RNAseq resources for which phenotypic data is available, such as GTEx [33] or TCGA (https://www.cancer.gov/tcga) .

Various deconvolution algorithms have been developed to determine proportions of constituent cell types in bulk RNAseq, e.g., CIBERSORT and others [85-87]. Such analyses typically depend on input expression matrices of cell type reference genes, generated from transcriptome analysis of isolated cells or cell types. The accuracy of input matrices can be affected by various factors, such as technical artifacts due to cell extraction and processing, the presence of contaminating cell types, and limited input data availability for some cell types. Cross checking input matrices against our dataset could optimise such analysis, by identifying the most likely highly enriched genes *in vivo*.

Limitations of the study

There are limitations to our study. In some tissues, we do not profile specific cell subtypes, e.g., basal and suprabasal keratinocytes in the skin, which were handled as one cell type in our analysis. In such cases, we failed to identify genes that fulfilled the criteria for use as input *Ref.T..* In keeping with our observations, scRNAseq analysis of skin showed that genes considered to be basal keratinocyte markers e.g., *COL17A1* and *KRT5*, were indeed most highly expressed in this cell type, but were also co-enriched within the tissue in suprabasal keratinocytes [3]. Thus, such cell subtype definitions are likely primarily governed by variation in absolute mRNA expression levels, rather that the presence or absence of a large number of uniquely enriched genes.

As our analysis end point is a gene enrichment score, we do not provide information on absolute mRNA expression profiles on a cell type basis, such as that generated by scRNAseq analysis.

As the prediction of cell type enriched genes is dependent on known input *Ref.T.*, we cannot identify novel cell (sub)types for which *Ref.T.* have not yet been described.

We analysed samples from a total of 933 individuals from the GTEx portal [33], with diverse health status, whose ages skewed older (ages 20-29: 8.5%, 30-39: 8.1%, 40-49: 15.6%, 50-59: 31.9%, 60-69: 32.4%, 70-79: 3.4%). Thus, the input dataset represents a limited age demographic, and a health status that may not represent the general population.

Expression of certain genes are strongly modified by environmental (e.g., eating, exercise, inflammation etc.), or genetic factors [88]. Such genes may therefore lack correlation with the constitutively expressed *Ref.T.* selected to represent the cell type in which they are predominantly transcribed, and thus could be considered a type of false negative in our analysis. One such example is *SELE*, an endothelial cell specific gene that is highly upregulated during inflammation and expressed at very low levels, if at all, in resting state [89]. *SELE*, despite its highly endothelial cell restricted expression profile, is not classified as EC enriched in our analysis, due to the variable nature of its expression.

We used relatively high thresholds for classification of cell type enriched genes. It is likely that some cell type enriched genes may be false negatives in our analysis, as they fall just below the thresholds required for classification as such. For example, the gene *KANK3* is classified as endothelial cell enriched in 9 tissue types, in the remaining 6 the highest enrichment score is also in endothelial cells, vs. all other types profiled, although it did not reach classification threshold. Thus, our classifications are intended only as a guide, and the reader should consider the data on a transcript-by-transcript basis.

All data generated in this study is available on the Human Protein Atlas in the ´Tissue Cell Type´ section (www.proteinatlas.org/humanproteome/tissue+cell+type), and can be used alongside data generated from scRNAseq in the ´Single Cell Type Section´ [3], and the antibody based tissue protein profiling in the ´Tissue Section´ [2].

METHODS AND RESOURCES

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact: Dr. Lynn Marie Butler. Email: Lynn.butler@ki.se

This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Bulk RNAseq data analysed in this study was obtained from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) V8 [29] on 2021/04/26 (dbGaP Accession phs000424.v8.p2). Protein coding genes were categorised according to Biotype definitions in ENSEMBL release 102 [90] inclusive of those defined as “protein_coding”, “IG_C_gene”, “IG_D_gene”, “IG_J_gene”, “IG_V_gene”, “TR_C_gene”,“TR_D_gene” ,“TR_J_gene” and “TR_V_gene”. All other categorisations were classified as “non-protein coding” and were excluded from the analysis. Human tissue protein profiling was performed in house as part of the Human Protein Atlas project [2,91,92] (www.proteinatlas.org). Normal tissue samples were obtained from the Department of Pathology, Uppsala University Hospital, Uppsala, Sweden, as part of the Uppsala Biobank. Samples were handled in accordance with Swedish laws and regulations, with approval from the Uppsala Ethical Review Board (Uhlen et al., 2015).

METHOD DETAILS

Sample inclusion

All samples in each GTEx tissue type dataset were included in the analyses, with the exception of: (i) ***Skin-not Sun Exposed (suprapubic)***: *Ref.T.* selected to represent hair root cells were absent or very lowly expressed in a large number of samples, presumably due to the lack of such structures in the selected region of tissue analysed. Thus, only samples with mean TPM >0.1 for hair follicle expressed transcripts trichohyalin (*TCHH*), keratins 25 (*KRT25*) and 71 (*KRT71*) were included for in the analysis (n=177). (ii) ***Breast – Mammary Tissue***: The GTEx breast dataset contains samples from both male and female donors, we analysed those from

only from females. In both cases, sample IDs included in the analysis can be found in Table S2, Tab ´Sample IDs´.

QUANTIFICATION AND STATISTICAL ANALYSIS

Reference transcript-based correlation analysis

This method was based on that we previously developed [26-28]. Pairwise Spearman correlation coefficients between reference transcripts (*Ref.T.*), selected as proxy markers for each cell type (see Table S1, Tab 1, Table A-O), and all other transcripts were calculated in R using the *corr.test* function from the *psych* package (v 1.8.4). False Discovery Rate (FDR) adjusted p-values (using Bonferroni correction) <0.0001 were considered significant. Genes were predicted to be cell type enriched if they fulfilled the criteria as described in the results section. In cases where a given cell type was represented by more than one *Ref.T* panel, or they could be considered related sub-cell types, the minimum differential score required vs. other *Ref.T.* panels was calculated excluding each the other (i.e., genes that correlated highly with both *Ref.T.* panels representing the same (sub)cell type were *not* excluded from classification as cell type enriched, but included in both – see Table S1, Tab 2).

Weighted correlation network (WGCNA) analysis

The R package WGCNA [38] was used to perform co-expression network analysis for gene clustering, on log2 expression TPM values. The analysis was performed according to recommended conditions in the WGCNA manual. Non-protein coding transcripts and transcripts with too many missing values were excluded using the goodSamplesGenes() function.

Gene Ontology

The Gene Ontology Consortium [41] and PANTHER classification resource [93,94] were used to identify over represented terms in gene lists from the GO ontology (release date 2022-07-01) or reactome (release date 2021-10-01) databases. Plots of GO terms were created using the Clusterprofiler package in R [95] or REVIGO [56], as specified.

Additional datasets and analysis

Single cell RNAseq data from Tabula Sapiens [36] was downloaded and UMAP plots created using the Seurat package in R [96]. Human testis scRNAseq data was sourced from the human testis atlas [59]. Tissue enriched genes were downloaded from the Human Protein Atlas (HPA) tissue atlas [2] or GTEx database [29], as collated in the Harminozome database [39].

Tissue Profiling: Human tissue sections

Immunohistochemistry (IHC) stained tissue sections were stained, as previously described [2,91]. Briefly, formalin fixed and paraffin embedded tissue samples were sectioned, de-paraffinised in xylene, hydrated in graded alcohols and blocked for endogenous peroxidase in 0.3% hydrogen peroxide diluted in 95% ethanol. For antigen retrieval, a Decloaking chamber® (Biocare Medical, CA) was used. Slides were boiled in Citrate buffer®, pH6 (Lab Vision, CA). Primary antibodies and a dextran polymer visualization system (UltraVision LP HRP polymer®, Lab Vision) were incubated for 30 min each at room temperature and slides were developed for 10 minutes using Diaminobenzidine (Lab Vision) as the chromogen. Slides were counterstained in Mayers hematoxylin (Histolab) and scanned using Scanscope XT (Aperio). Primary antibodies used for IHC staining are listed in Table S7.

Other visualisation and analysis tools

Graphs and plots were made using Graphpad prism or the ggplot2 package in R [97], unless otherwise specified. Circular plots were constructed using the R package *circlize* [98] and pubmed data was extracted using the easyPubMed package in R (https://CRAN.R-project.org/package=easyPubMed). Some figure illustrations were created using BioRender.com.

DATA AVAILABILITY

Data for all protein coding genes and antibody-based protein profiling is provided on the Human Protein Atlas (Tissue Cell Type section) (www.proteinatlas.org/humanproteome/tissue+cell+type). This article also includes all

individual tissue datasets generated (Table S2) and cell type enrichment categorisations (Table S1, Tab 4).

AUTHOR CONTRIBUTIONS

Conceptualisation: LMB. Methodology: PD, LMB. Formal analysis: PD, SO, ES, MNT, MJI, LMB. Investigation: PD, SO, ES, MNT, MJI, FP, CL, LMB. Resources: FP, CL, JO, MU, LMB. Writing – Original Draft: PD, LMB. Writing – Review & Editing: All, Visualisation: PD, LMB. Supervision: LMB, PD. Funding Acquisition: JO, FP, CL, MU, LMB.

ACKNOWLEDGEMENTS

This work was supported by funding granted to LMB from Hjärt Lungfonden (20170759, 20170537) and the Swedish Research Council (2019-01493). Main funding for the Human Protein Atlas was provided from the Knut and Alice Wallenberg Foundation (WCPR) and the Erling Persson Foundation (KCAP). We acknowledge the staff of the Human Protein Atlas program and the Science for Life Laboratory (SciLifeLab) for their valuable contributions. **Data usage:** We used data from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) [29], supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL TABLE LEGENDS

Table S1. Reference transcript selection and analysis summary

[Tab 1]: Correlation coefficient values between selected *Ref.T.* in each tissue. [Tab 2]: Cell subtypes represented by different *Ref.T*. panels within a single tissue and corresponding annotations in the Tabula Sapiens and HPA databases. [Tab 3]: Analysis criteria and totals for very high, high and moderately enriched genes within each cell type. [Tab 4]: Cell type enrichment predictions for all protein coding genes.

Table S2. Sample IDs and tissue-by-tissue data

[Tab: Sample IDs]: Analysed sample IDs (GTEx). [Other tabs]: Details for each tissue type (see key).

Table S3. Gene ontology (GO) terms in alpha and beta cells of pancreas

[Tab 1]: GO biological process, reactome, and cellular component analysis for genes predicted to be co-enriched in alpha and beta cells of the pancreas. [Tab 2]: Synapse-linked genes with predicted co-enrichment in alpha and beta cells of the pancreas.

Table S4. Values and gene ontology (GO) analysis of germ cell enriched genes

[Tab 1]: Genes predicted to be enriched in germ cells of the testis (see key). [Tab 2]: GO biological process and reactome analysis of germ cell enriched genes. [Tab 3]: GO biological process analysis for germ cell subtype predicted enriched genes.

Table S5. GO analysis of genes enriched in multiple cell types

[Tab 1]: Table A: Genes predicted to be enriched in 3 or more of: adipocytes, sebaceous gland cells, hepatocytes, and proximal tubular cells. Table B: GO Biological Process analysis for genes in Table A. [Tab 2]: Table A: Genes predicted to be enriched in respiratory ciliated cells of the lung and S3 and/or S4 cells (early or late spermatids) of the testis. Table B: Enriched GO biological process anaysis for genes listed in Table A. [Tab 3]: Table A: Genes predicted to be enriched in S3 and/or S4 cells (early/late Spermatids) of the testis, but not in respiratory ciliated cells of the lung. Table B: GO biological process analysis for genes listed in Table A.

Table S6. Core cell type predicted enriched genes

Genes predicted to be enriched in the same cell type in at least half of the tissues where profiled.

Table S7. Primary antibodies

IDs for all primary antibodies used to stain all immunohistochemistry images used in this study.

REFERENCES

1. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. Elife *6*. 10.7554/eLife.27041.

2. Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., et al. (2015). Proteomics. Tissue-based map of the human proteome. Science *347*, 1260419. 10.1126/science.1260419.

3. Karlsson, M., Zhang, C., Mear, L., Zhong, W., Digre, A., Katona, B., Sjostedt, E., Butler, L., Odeberg, J., Dusart, P., et al. (2021). A single-cell type transcriptomics map of human tissues. Sci Adv *7*. 10.1126/sciadv.abh2169.

4. Iglesias, M.J., Kruse, L.D., Sanchez-Rivera, L., Enge, L., Dusart, P., Hong, M.G., Uhlen, M., Renne, T., Schwenk, J.M., Bergstrom, G., et al. (2021). Identification of Endothelial Proteins in Plasma Associated With Cardiovascular Risk Factors. Arterioscler Thromb Vasc Biol *41*, 2990-3004. 10.1161/ATVBAHA.121.316779.

5. Cano-Gamez, E., and Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet *11*, 424. 10.3389/fgene.2020.00424.

6. Wang, R.D.-Y., L.; Jiang, Y. (2021). EPIC: inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. www.biorxiv.org. https://doi.org/10.1101/2021.06.09.447805.

7. van den Brink, S.C., Sage, F., Vertesy, A., Spanjaard, B., Peterson-Maduro, J., Baron, C.S., Robin, C., and van Oudenaarden, A. (2017). Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods *14*, 935-936. 10.1038/nmeth.4437.

8. Denisenko, E., Guo, B.B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clement, O., Simmons, R.K., Lister, R., and Forrest, A.R.R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol *21*, 130. 10.1186/s13059-020-02048-6.

9. Massoni-Badosa, R., Iacono, G., Moutinho, C., Kulis, M., Palau, N., Marchese, D., Rodriguez-Ubreva, J., Ballestar, E., Rodriguez-Esteban, G., Marsal, S., et al. (2020). Sampling time-dependent artifacts in single-cell genomics studies. Genome Biol *21*, 112. 10.1186/s13059-020-02032-0.

10. Rondini, E.A., and Granneman, J.G. (2020). Single cell approaches to address adipose tissue stromal cell heterogeneity. Biochem J *477*, 583-600. 10.1042/BCJ20190467.

11. Viswanadha, S., and Londos, C. (2006). Optimized conditions for measuring lipolysis in murine primary adipocytes. J Lipid Res *47*, 1859-1864. 10.1194/jlr.D600005-JLR200.

12. Quake, T.T.S.C.S.R. (2021). The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. www.biorxiv.org *https://doi.org/10.1101/2021.07.19.452956*.

13. Tabula Muris, C., Overall, c., Logistical, c., Organ, c., processing, Library, p., sequencing, Computational data, a., Cell type, a., Writing, g., et al. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. Nature *562*, 367-372. 10.1038/s41586-018-0590-4.

14. Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods *14*, 955-958. 10.1038/nmeth.4407.

15. Thrupp, N., Sala Frigerio, C., Wolfs, L., Skene, N.G., Fattorelli, N., Poovathingal, S., Fourne, Y., Matthews, P.M., Theys, T., Mancuso, R., et al. (2020). Single-Nucleus RNA-Seq Is Not Suitable for Detection of Microglial Activation Genes in Humans. Cell Rep *32*, 108189. 10.1016/j.celrep.2020.108189.

16. Haque, A., Engel, J., Teichmann, S.A., and Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med *9*, 75. 10.1186/s13073-017-0467-4.

17. Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Mol Cell *58*, 610-620. 10.1016/j.molcel.2015.04.005.

18. Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics *19*, 562-578. 10.1093/biostatistics/kxx053.

19. Zheng, Y., Zhong, Y., Hu, J., and Shang, X. (2021). SCC: an accurate imputation method for scRNA-seq dropouts based on a mixture model. BMC Bioinformatics *22*, 5. 10.1186/s12859-020-03878-8.

20. Chen, G., Ning, B., and Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. Front Genet *10*, 317. 10.3389/fgene.2019.00317.

21. Hou, W., Ji, Z., Ji, H., and Hicks, S.C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. Genome Biol *21*, 218. 10.1186/s13059-020-02132-x.

22. Jiang, R., Sun, T., Song, D., and Li, J.J. (2022). Statistics or biology: the zero-inflation controversy about scRNA-seq data. Genome Biol *23*, 31. 10.1186/s13059-022-02601-5.

23. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. Nat Commun *12*, 5692. 10.1038/s41467-021-25960-2.

24. Denninger, J.K., Walker, L.A., Chen, X., Turkoglu, A., Pan, A., Tapp, Z., Senthilvelan, S., Rindani, R., Kokiko-Cochran, O.N., Bundschuh, R., et al. (2022). Robust Transcriptional Profiling and Identification of Differentially Expressed Genes With Low Input RNA Sequencing of Adult Hippocampal Neural Stem and Progenitor Populations. Front Mol Neurosci *15*, 810722. 10.3389/fnmol.2022.810722.

25. Trostle, A.J., Wang, J., Li, L., Wan, Y., and Liu, Z. (2022). Most High Throughput Expression Data Sets Are Underpowered. biorxiv.org. 10.1101/2022.08.03.502688

26. Butler, L.M., Hallstrom, B.M., Fagerberg, L., Ponten, F., Uhlen, M., Renne, T., and Odeberg, J. (2016). Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome. Cell Syst *3*, 287-301 e283. 10.1016/j.cels.2016.08.001.

27. Dusart, P., Hallstrom, B.M., Renne, T., Odeberg, J., Uhlen, M., and Butler, L.M. (2019). A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes. Cell Rep *29*, 1690-1706 e1694. 10.1016/j.celrep.2019.09.088.

28. Norreen-Thorsen, M., Struck, E.C., Oling, S., Zwahlen, M., Von Feilitzen, K., Odeberg, J., Lindskog, C., Ponten, F., Uhlen, M., Dusart, P.J., and Butler, L.M. (2022). A human adipose tissue cell-type transcriptome atlas. Cell Rep *40*, 111046. 10.1016/j.celrep.2022.111046.

29. Consortium, G.T. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648-660. 10.1126/science.1262110.

30. Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W., et al. (2020). Construction of a human cell landscape at single-cell level. Nature *581*, 303-309. 10.1038/s41586-020-2157-4.

31. Zhang, X., Lan, Y., Xu, J., Quan, F., Zhao, E., Deng, C., Luo, T., Xu, L., Liao, G., Yan, M., et al. (2019). CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res *47*, D721-D728. 10.1093/nar/gky900.

32. Franzen, O., Gan, L.M., and Bjorkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) *2019*. 10.1093/database/baz046.

33. Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318-1330. 10.1126/science.aaz1776.

34. Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun *10*, 380. 10.1038/s41467-018-08023-x.

35. Debbabi, H., Ghosh, S., Kamath, A.B., Alt, J., Demello, D.E., Dunsmore, S., and Behar, S.M. (2005). Primary type II alveolar epithelial cells present microbial antigens to

antigen-specific CD4+ T cells. Am J Physiol Lung Cell Mol Physiol *289*, L274-279. 10.1152/ajplung.00004.2005.

36. Tabula Sapiens, C., Jones, R.C., Karkanias, J., Krasnow, M.A., Pisco, A.O., Quake, S.R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., et al. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science *376*, eabl4896. 10.1126/science.abl4896.

37. Gene Ontology, C. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res *49*, D325-D334. 10.1093/nar/gkaa1113.

38. Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559. 10.1186/1471-2105-9-559.

39. Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford) *2016*. 10.1093/database/baw100.

40. Moede, T., Leibiger, I.B., and Berggren, P.O. (2020). Alpha cell regulation of beta cell function. Diabetologia *63*, 2064-2075. 10.1007/s00125-020-05196-3.

41. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29. 10.1038/75556.

42. Huang, C., Walker, E.M., Dadi, P.K., Hu, R., Xu, Y., Zhang, W., Sanavia, T., Mun, J., Liu, J., Nair, G.G., et al. (2018). Synaptotagmin 4 Regulates Pancreatic beta Cell Maturation by Modulating the Ca(2+) Sensitivity of Insulin Secretion Vesicles. Dev Cell *45*, 347-361 e345. 10.1016/j.devcel.2018.03.013.

43. Tarquis-Medina, M., Scheibner, K., Gonzalez-Garcia, I., Bastidas-Ponce, A., Sterr, M., Jaki, J., Schirge, S., Garcia-Caceres, C., Lickert, H., and Bakhti, M. (2021). Synaptotagmin-13 Is a Neuroendocrine Marker in Brain, Intestine and Pancreas. Int J Mol Sci *22*. 10.3390/ijms222212526.

44. Bakhti, M., Bastidas-Ponce, A., Tritschler, S., Tarquis-Medina, M., Nedvedova, E., Scheibner, K., Jaki, J., Cota, P., Salinno, C., Boldt, K., et al. (2021). Synaptotagmin 13 orchestrates pancreatic endocrine cell egression and islet morphogenesis. www.biorxiv.org *https://doi.org/10.1101/2021.08.30.458251*.

45. Churchill, A.J., Gutierrez, G.D., Singer, R.A., Lorberbaum, D.S., Fischer, K.A., and Sussel, L. (2017). Genetic evidence that Nkx2.2 acts primarily downstream of Neurog3 in pancreatic endocrine lineage development. Elife *6*. 10.7554/eLife.20010.

46. Bohuslavova, R., Smolik, O., Malfatti, J., Berkova, Z., Novakova, Z., Saudek, F., and Pavlinkova, G. (2021). NEUROD1 Is Required for the Early alpha and beta Endocrine Differentiation in the Pancreas. Int J Mol Sci *22*. 10.3390/ijms22136713.

47. Soyer, J., Flasse, L., Raffelsberger, W., Beucher, A., Orvain, C., Peers, B., Ravassard, P., Vermot, J., Voz, M.L., Mellitzer, G., and Gradwohl, G. (2010). Rfx6 is an Ngn3-dependent winged helix transcription factor required for pancreatic islet cell development. Development *137*, 203-212. 10.1242/dev.041673.

48. Liang, X., Duan, H., Mao, Y., Koestner, U., Wei, Y., Deng, F., Zhuang, J., Li, H., Wang, C., Hernandez-Miranda, L.R., et al. (2021). The SNAG Domain of Insm1 Regulates Pancreatic Endocrine Cell Differentiation and Represses beta- to delta-Cell Transdifferentiation. Diabetes *70*, 1084-1097. 10.2337/db20-0883.

49. Hart, A.W., Mella, S., Mendrychowski, J., van Heyningen, V., and Kleinjan, D.A. (2013). The developmental regulator Pax6 is essential for maintenance of islet cell function in the adult mouse pancreas. PLoS One *8*, e54173. 10.1371/journal.pone.0054173.

50. Wang, S., Zhang, J., Zhao, A., Hipkens, S., Magnuson, M.A., and Gu, G. (2007). Loss of Myt1 function partially compromises endocrine islet cell differentiation and pancreatic physiological function in the mouse. Mech Dev *124*, 898-910. 10.1016/j.mod.2007.08.004.

51. Su, Y., Jono, H., Misumi, Y., Senokuchi, T., Guo, J., Ueda, M., Shinriki, S., Tasaki, M., Shono, M., Obayashi, K., et al. (2012). Novel function of transthyretin in pancreatic alpha cells. FEBS Lett *586*, 4215-4222. 10.1016/j.febslet.2012.10.025.

52. Noguchi, G.M., and Huising, M.O. (2019). Integrating the inputs that shape pancreatic islet hormone release. Nat Metab *1*, 1189-1201. 10.1038/s42255-019-0148-2.

53. Yang, J.K., Lu, J., Yuan, S.S., Asan, Cao, X., Qiu, H.Y., Shi, T.T., Yang, F.Y., Li, Q., Liu, C.P., et al. (2018). From Hyper- to Hypoinsulinemia and Diabetes: Effect of KCNH6 on Insulin Secretion. Cell Rep *25*, 3800-3810 e3806. 10.1016/j.celrep.2018.12.005.

54. Westermark, P., Andersson, A., and Westermark, G.T. (2011). Islet amyloid polypeptide, islet amyloid, and diabetes mellitus. Physiol Rev *91*, 795-826. 10.1152/physrev.00042.2009.

55. Nishimura, W., Takahashi, S., and Yasuda, K. (2015). MafA is critical for maintenance of the mature beta cell phenotype in mice. Diabetologia *58*, 566-574. 10.1007/s00125-014-3464-9.

56. Supek, F., Bosnjak, M., Skunca, N., and Smuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One *6*, e21800. 10.1371/journal.pone.0021800.

57. Ewen, K.A., Olesen, I.A., Winge, S.B., Nielsen, A.R., Nielsen, J.E., Graem, N., Juul, A., and Rajpert-De Meyts, E. (2013). Expression of FGFR3 during human testis development and in germ cell-derived tumours of young adults. Int J Dev Biol *57*, 141-151. 10.1387/ijdb.130022er.

58. Qian, B., Li, Y., Yan, R., Han, S., Bu, Z., Gong, J., Zheng, B., Yuan, Z., Ren, S., He, Q., et al. (2022). RNA binding protein RBM46 regulates mitotic-to-meiotic transition in spermatogenesis. Sci Adv *8*, eabq2945. 10.1126/sciadv.abq2945.

59. Guo, J., Grow, E.J., Mlcochova, H., Maher, G.J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., et al. (2018). The adult human testis transcriptional cell atlas. Cell Res *28*, 1141-1157. 10.1038/s41422-018-0099-2.

60. Le, L., Escobar, I.E., Ho, T., Lefkovith, A.J., Latteri, E., Haltaufderhyde, K.D., Dennis, M.K., Plowright, L., Sviderskaya, E.V., Bennett, D.C., et al. (2020). SLC45A2 protein stability and regulation of melanosome pH determine melanocyte pigmentation. Mol Biol Cell *31*, 2687-2702. 10.1091/mbc.E20-03-0200.

61. Yu, R., Broady, R., Huang, Y., Wang, Y., Yu, J., Gao, M., Levings, M., Wei, S., Zhang, S., Xu, A., et al. (2012). Transcriptome analysis reveals markers of aberrantly activated innate immunity in vitiligo lesional and non-lesional skin. PLoS One *7*, e51040. 10.1371/journal.pone.0051040.

62. Abbas Zadeh, S., Mlitz, V., Lachner, J., Golabi, B., Mildner, M., Pammer, J., Tschachler, E., and Eckhart, L. (2017). Phylogenetic profiling and gene expression studies implicate a primary role of PSORS1C2 in terminal differentiation of keratinocytes. Exp Dermatol *26*, 352-358. 10.1111/exd.13272.

63. Liu, Y., Das, S., Olszewski, R.E., Carpenter, D.A., Culiat, C.T., Sundberg, J.P., Soteropoulos, P., Liu, X., Doktycz, M.J., Michaud, E.J., and Voy, B.H. (2007). The near-naked hairless (Hr(N)) mutation disrupts hair formation but is not due to a mutation in the Hairless coding region. J Invest Dermatol *127*, 1605-1614. 10.1038/sj.jid.5700755.

64. Li, W., Tang, W., Teves, M.E., Zhang, Z., Zhang, L., Li, H., Archer, K.J., Peterson, D.L., Williams, D.C., Jr., Strauss, J.F., 3rd, and Zhang, Z. (2015). A MEIG1/PACRG complex in the manchette is essential for building the sperm flagella. Development *142*, 921-930. 10.1242/dev.119834.

65. Dawe, H.R., Farr, H., Portman, N., Shaw, M.K., and Gull, K. (2005). The Parkin co-regulated gene product, PACRG, is an evolutionarily conserved axonemal protein that functions in outer-doublet microtubule morphogenesis. J Cell Sci *118*, 5421-5430. 10.1242/jcs.02659.

66. Meschede, J., Sadic, M., Furthmann, N., Miedema, T., Sehr, D.A., Muller-Rischart, A.K., Bader, V., Berlemann, L.A., Pilsl, A., Schlierf, A., et al. (2020). The parkin-coregulated gene product PACRG promotes TNF signaling by stabilizing LUBAC. Sci Signal *13*. 10.1126/scisignal.aav1256.

67. Soon, R.K., Jr., and Yee, H.F., Jr. (2008). Stellate cell contraction: role, regulation, and potential therapeutic target. Clin Liver Dis *12*, 791-803, viii. 10.1016/j.cld.2008.07.004.

68. Khaing, P., and Summer, R. (2020). Maxed Out on Glycolysis: Alveolar Macrophages Rely on Oxidative Phosphorylation for Cytokine Production. Am J Respir Cell Mol Biol *62*, 139-140. 10.1165/rcmb.2019-0329ED.

69. Shetty, S., Lalor, P.F., and Adams, D.H. (2018). Liver sinusoidal endothelial cells - gatekeepers of hepatic immunity. Nat Rev Gastroenterol Hepatol *15*, 555-567. 10.1038/s41575-018-0020-y.

70. Delling, J.P., and Boeckers, T.M. (2021). Comparison of SHANK3 deficiency in animal models: phenotypes, treatment strategies, and translational implications. J Neurodev Disord *13*, 55. 10.1186/s11689-021-09397-8.

71. Sawada, J., Hiraoka, N., Qi, R., Jiang, L., Fournier-Goss, A.E., Yoshida, M., Kawashima, H., and Komatsu, M. (2022). Molecular Signature of Tumor-Associated High Endothelial Venules That Can Predict Breast Cancer Survival. Cancer Immunol Res *10*, 468-481. 10.1158/2326-6066.CIR-21-0369.

72. van der Weyden, L., Harle, V., Turner, G., Offord, V., Iyer, V., Droop, A., Swiatkowska, A., Rabbie, R., Campbell, A.D., Sansom, O.J., et al. (2021). CRISPR activation screen in mice identifies novel membrane proteins enhancing pulmonary metastatic colonisation. Commun Biol *4*, 395. 10.1038/s42003-021-01912-w.

73. Zhang, Y., Tong, G.H., Wei, X.X., Chen, H.Y., Liang, T., Tang, H.P., Wu, C.A., Wen, G.M., Yang, W.K., Liang, L., and Shen, H. (2021). Identification of Five Cytotoxicity-Related Genes Involved in the Progression of Triple-Negative Breast Cancer. Front Genet *12*, 723477. 10.3389/fgene.2021.723477.

74. Lin, Y.J., Chen, A.N., Yin, X.J., Li, C., and Lin, C.C. (2020). Human Microfibrillar-Associated Protein 4 (MFAP4) Gene Promoter: A TATA-Less Promoter That Is Regulated by Retinol and Coenzyme Q10 in Human Fibroblast Cells. Int J Mol Sci *21*. 10.3390/ijms21218392.

75. Sellers, M.M., and Stallone, J.N. (2008). Sympathy for the devil: the role of thromboxane in the regulation of vascular tone and blood pressure. Am J Physiol Heart Circ Physiol *294*, H1978-1986. 10.1152/ajpheart.01318.2007.

76. Hosokawa, H., Romero-Wolf, M., Yang, Q., Motomura, Y., Levanon, D., Groner, Y., Moro, K., Tanaka, T., and Rothenberg, E.V. (2020). Cell type-specific actions of Bcl11b in early T-lineage and group 2 innate lymphoid cells. J Exp Med *217*. 10.1084/jem.20190972.

77. Golonzhka, O., Liang, X., Messaddeq, N., Bornert, J.M., Campbell, A.L., Metzger, D., Chambon, P., Ganguli-Indra, G., Leid, M., and Indra, A.K. (2009). Dual role of COUP-TF-interacting protein 2 in epidermal homeostasis and permeability barrier formation. J Invest Dermatol *129*, 1459-1470. 10.1038/jid.2008.392.

78. Zhao, Y.Y., Yan, D.J., and Chen, Z.W. (2013). Role of AIF-1 in the regulation of inflammatory activation and diverse disease processes. Cell Immunol *284*, 75-83. 10.1016/j.cellimm.2013.07.008.

79. Tsubata, Y., Sakatsume, M., Ogawa, A., Alchi, B., Kaneko, Y., Kuroda, T., Kawachi, H., Narita, I., Yamamoto, T., and Gejyo, F. (2006). Expression of allograft inflammatory factor-1 in kidneys: A novel molecular component of podocyte. Kidney Int *70*, 1948-1954. 10.1038/sj.ki.5001941.

80. Li, W.V., and Li, J.J. (2018). An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun *9*, 997. 10.1038/s41467-018-03405-7.

81. Mou, T., Deng, W., Gu, F., Pawitan, Y., and Vu, T.N. (2020). Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. Front Genet *10*, 1331. 10.3389/fgene.2019.01331.

82. Chu, S.K., Zhao, S., Shyr, Y., and Liu, Q. (2022). Comprehensive evaluation of noise reduction methods for single-cell RNA sequencing data. Brief Bioinform *23*. 10.1093/bib/bbab565.

83. Mandelboum, S., Manber, Z., Elroy-Stein, O., and Elkon, R. (2019). Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. PLoS Biol *17*, e3000481. 10.1371/journal.pbio.3000481.

84. Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535-550. 10.1016/j.cell.2016.03.014.

85. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Methods *12*, 453-457. 10.1038/nmeth.3337.

86. Glastonbury, C.A., Couto Alves, A., El-Sayed Moustafa, J.S., and Small, K.S. (2019). Cell-Type Heterogeneity in Adipose Tissue Is Associated with Complex Traits and Reveals Disease-Relevant Cell-Specific eQTLs. Am J Hum Genet *104*, 1013-1024. 10.1016/j.ajhg.2019.03.025.

87. Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietilainen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat Commun *11*, 1971. 10.1038/s41467-020-15816-6.

88. Gibson, G. (2008). The environmental contribution to gene expression profiles. Nat Rev Genet *9*, 575-581. 10.1038/nrg2383.

89. Vestweber, D., and Blanks, J.E. (1999). Mechanisms that regulate the function of the selectins and their ligands. Physiol Rev *79*, 181-213. 10.1152/physrev.1999.79.1.181.

90. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res *48*, D682-D688. 10.1093/nar/gkz966.

91. Ponten, F., Jirstrom, K., and Uhlen, M. (2008). The Human Protein Atlas - a tool for pathology. J Pathol *216*, 387-393. 10.1002/path.2440.

92. Uhlen, M., Zhang, C., Lee, S., Sjostedt, E., Fagerberg, L., Bidkhori, G., Benfeitas, R., Arif, M., Liu, Z., Edfors, F., et al. (2017). A pathology atlas of the human cancer transcriptome. Science *357*. 10.1126/science.aan2507.

93. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic Acids Res *44*, D336-342. 10.1093/nar/gkv1194.

94. Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nature protocols *8*, 1551-1566. 10.1038/nprot.2013.092.

95. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) *2*, 100141. 10.1016/j.xinn.2021.100141.

96. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587 e3529. 10.1016/j.cell.2021.04.048.

97. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

98. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics *30*, 2811-2812. 10.1093/bioinformatics/btu393.

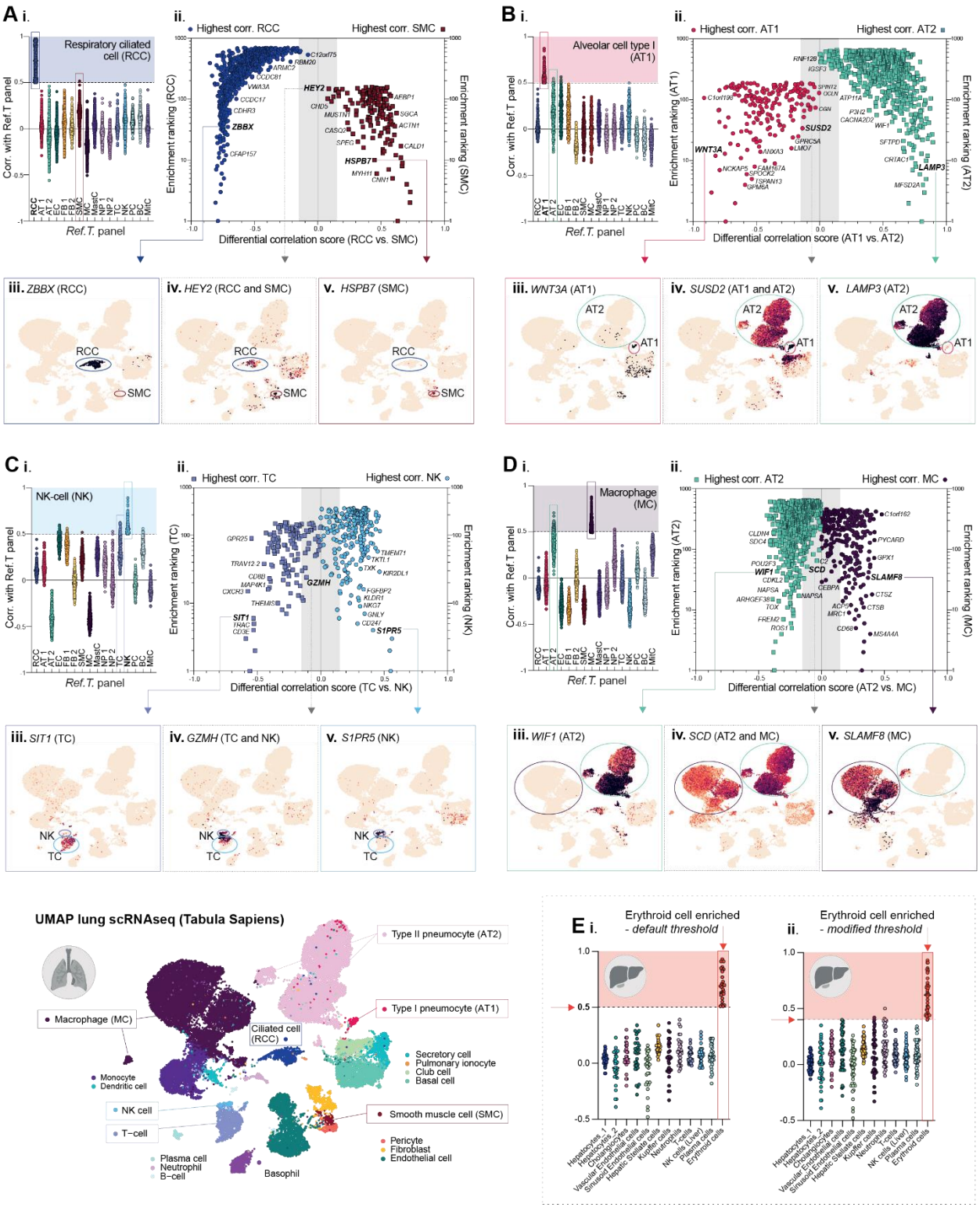

A i.
Respiratory ciliated cell (RCC)
Corr. with Ref.T panel
Ref.T. panel
ii.
Highest corr. RCC
Highest corr. SMC
Enrichment ranking (RCC)
Enrichment ranking (SMC)
Differential correlation score (RCC vs. SMC)
iii. ZBBX (RCC)
iv. HEY2 (RCC and SMC)
v. HSPB7 (SMC)
RCC
SMC
B i.
Alveolar cell type I (AT1)
ii.
Highest corr. AT1
Highest corr. AT2
Enrichment ranking (AT1)
Enrichment ranking (AT2)
Differential correlation score (AT1 vs. AT2)
iii. WNT3A (AT1)
iv. SUSD2 (AT1 and AT2)
v. LAMP3 (AT2)
AT1
AT2
C i.
NK-cell (NK)
ii.
Highest corr. TC
Highest corr. NK
Enrichment ranking (TC)
Enrichment ranking (NK)
Differential correlation score (TC vs. NK)
iii. SIT1 (TC)
iv. GZMH (TC and NK)
v. S1PR5 (NK)
NK
TC
D i.
Macrophage (MC)
ii.
Highest corr. AT2
Highest corr. MC
Enrichment ranking (MC)
Differential correlation score (AT2 vs. MC)
iii. WIF1 (AT2)
iv. SCD (AT2 and MC)
v. SLAMF8 (MC)
UMAP lung scRNAseq (Tabula Sapiens)
Type II pneumocyte (AT2)
Type I pneumocyte (AT1)
Macrophage (MC)
Ciliated cell (RCC)
Secretory cell
Pulmonary ionocyte
Club cell
Basal cell
Monocyte
Dendritic cell
NK cell
T−cell
Smooth muscle cell (SMC)
Pericyte
Fibroblast
Endothelial cell
Plasma cell
Neutrophil
B−cell
Basophil
E i.
Erythroid cell enriched
- default threshold
ii.
Erythroid cell enriched
- modified threshold

Figure S1. Integrative co-expression analysis of unfractionated human tissue RNAseq can resolve constituent cell type enriched genes. Related to Figure 1.

RNAseq datasets for human lung (n=578) were retrieved from GTEx V8 and correlation coefficients between selected cell type *Ref. T.* and all other sequenced transcripts generated. Correlation values vs. all other cell type *Ref.T.* panels for transcripts reaching the designated threshold with *Ref. T.* for (**A**) (i) respiratory ciliated cells (RCC) (**B**) (i) alveolar type I cells (AT1), (**C**) (i) natural killer cells (NK) or (**D**) (i) macrophages (MC). The '*differential correlation score'* and respective enrichment rankings for transcripts reaching the designated threshold with *Ref. T.* for (**A**) (ii) RCC or SMC, (**B**) (ii) AT1 or AT2, (**C**) (ii) NK or TC and (**D**) (ii) MC and AT2. scRNAseq data from analysis of human lung was sourced from Tabula Sapiens (Tabula Sapiens et al., 2022) and used to generate UMAP plots, showing the expression profiles of example genes we predicted as being enriched in (**A**) (iii) RCC only, (iv) RCC and SMC or (v) SMC only, (**B**) (iii) AT1 only, (iv) AT1 and AT2 or (v) AT2 only, (**C**) (iii) TC only, (iv) TC and NK or (v) NK only, or (**D**) (iii) AT2 only, (iv) AT2 and MC or (v) MC only. (**E**). RNAseq datasets for human liver (n=226) were retrieved from GTEx V8 and analysed as described for lung. Correlation values vs. all cell type *Ref.T.* panels for transcripts reaching the (i) designated or (ii) modified threshold for classification as erythroid cell enriched. EC; Endothelial cell, FB1/FB2; fibroblast, MC; macrophage, MastC; mast cell, NP1/NP2; neutrophil, TC; T-cell, NK; natural killer cell, PC; plasma cell, BC; B-cell.

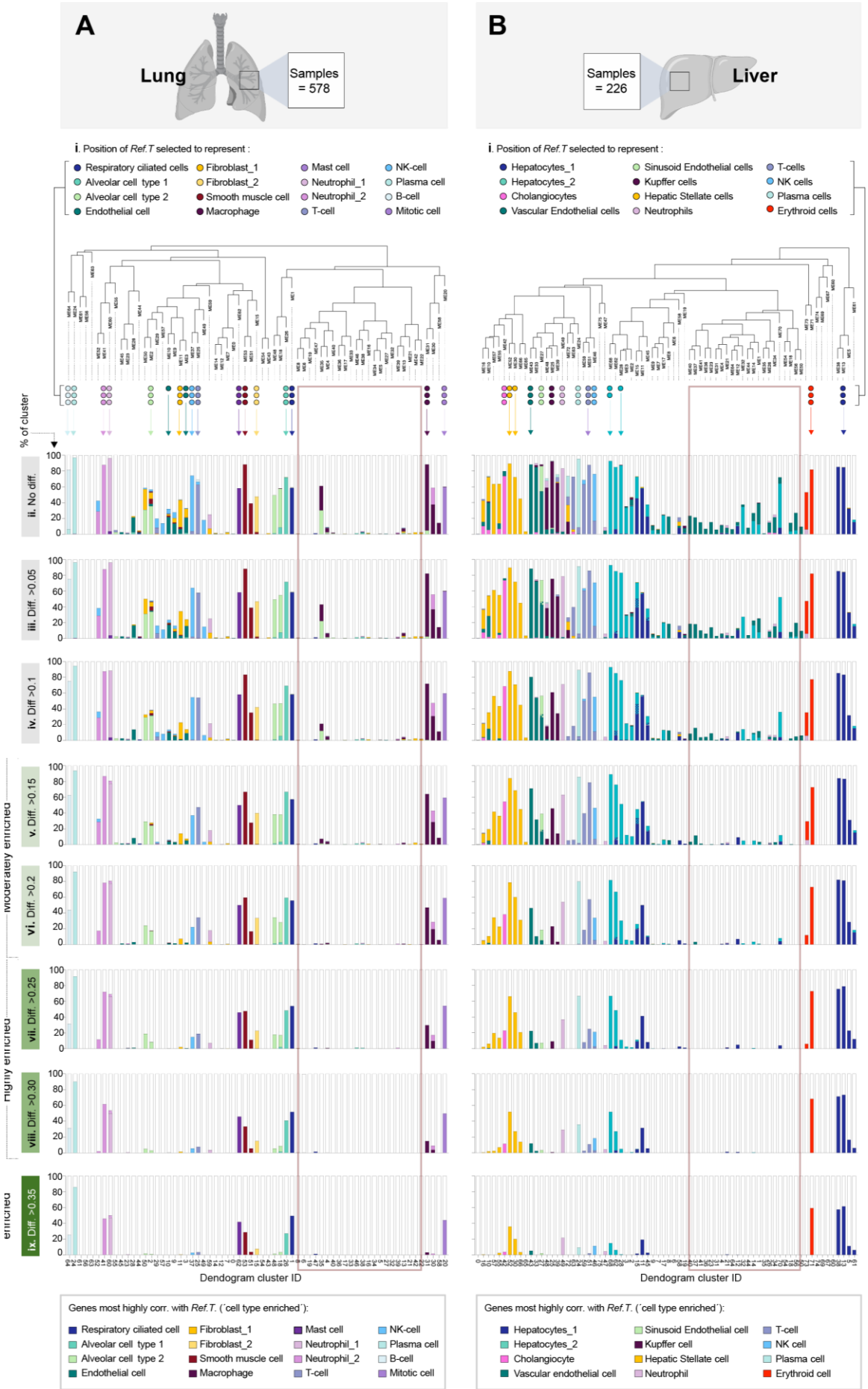

A
Lung
Samples = 578
B
Samples = 226
Liver
i. Position of Ref.T selected to represent :
Respiratory ciliated cells
Alveolar cell type 1
Alveolar cell type 2
Endothelial cell
Fibroblast_1
Fibroblast_2
Smooth muscle cell
Macrophage
Mast cell
Neutrophil_1
Neutrophil_2
T-cell
NK-cell
Plasma cell
B-cell
Mitotic cell
Hepatocytes_1
Hepatocytes_2
Cholangiocytes
Vascular Endothelial cells
Sinusoid Endothelial cells
Kupffer cells
Hepatic Stellate cells
Neutrophils
T-cells
NK cells
Plasma cells
Erythroid cells
% of cluster
ii. No diff.
iii. Diff. >0.05
iv. Diff >0.1
v. Diff. >0.15
vi. Diff. >0.2
vii. Diff. >0.25
viii. Diff. >0.30
ix. Diff. >0.35
Moderately enriched
Highly enriched
enriched
Dendogram cluster ID
Genes most highly corr. with Ref.T. (´cell type enriched´):
Respiratory ciliated cell
Alveolar cell type 1
Alveolar cell type 2
Endothelial cell
Fibroblast_1
Fibroblast_2
Smooth muscle cell
Macrophage
Mast cell
Neutrophil_1
Neutrophil_2
T-cell
NK-cell
Plasma cell
B-cell
Mitotic cell
Hepatocytes_1
Hepatocytes_2
Cholangiocyte
Vascular endothelial cell
Sinusoid Endothelial cell
Kupffer cell
Hepatic Stellate cell
Neutrophil
T-cell
NK cell
Plasma cell
Erythroid cell

Figure S2. Unsupervised weighted network correlation analysis (WGNCA) is consistent with *Ref.T.* analysis. Related to Figure 1. RNAseq data from human (A) lung (n=578 individuals) or (B) pancreas (n=328) was subject to weighted correlation network analysis (WGCNA). In the resultant dendrograms, the position of (i) *Ref.T*. selected to represent each cell type and (ii) the % of the cluster containing transcripts that had a correlation with any *Ref.T*. panel above the designated threshold, are indicated; colour representing the cell type classification (see bottom panel) (Table S1, Tab 5 for thresholds). Distribution of transcripts for each cell type classification when the highest correlation with any given *Ref.T*. panel was a minimum of (ii) 0, (iii) 0.05, (iv) 0.10, (v) 0.15 [moderately enriched], (vi) 0.20, (vii) 0.25 [highly enriched] or (viii) 0.30 or (ix) 0.35 [very highly enriched] greater that the next highest with a different *Ref.T.* panel (´differential correlation score´).

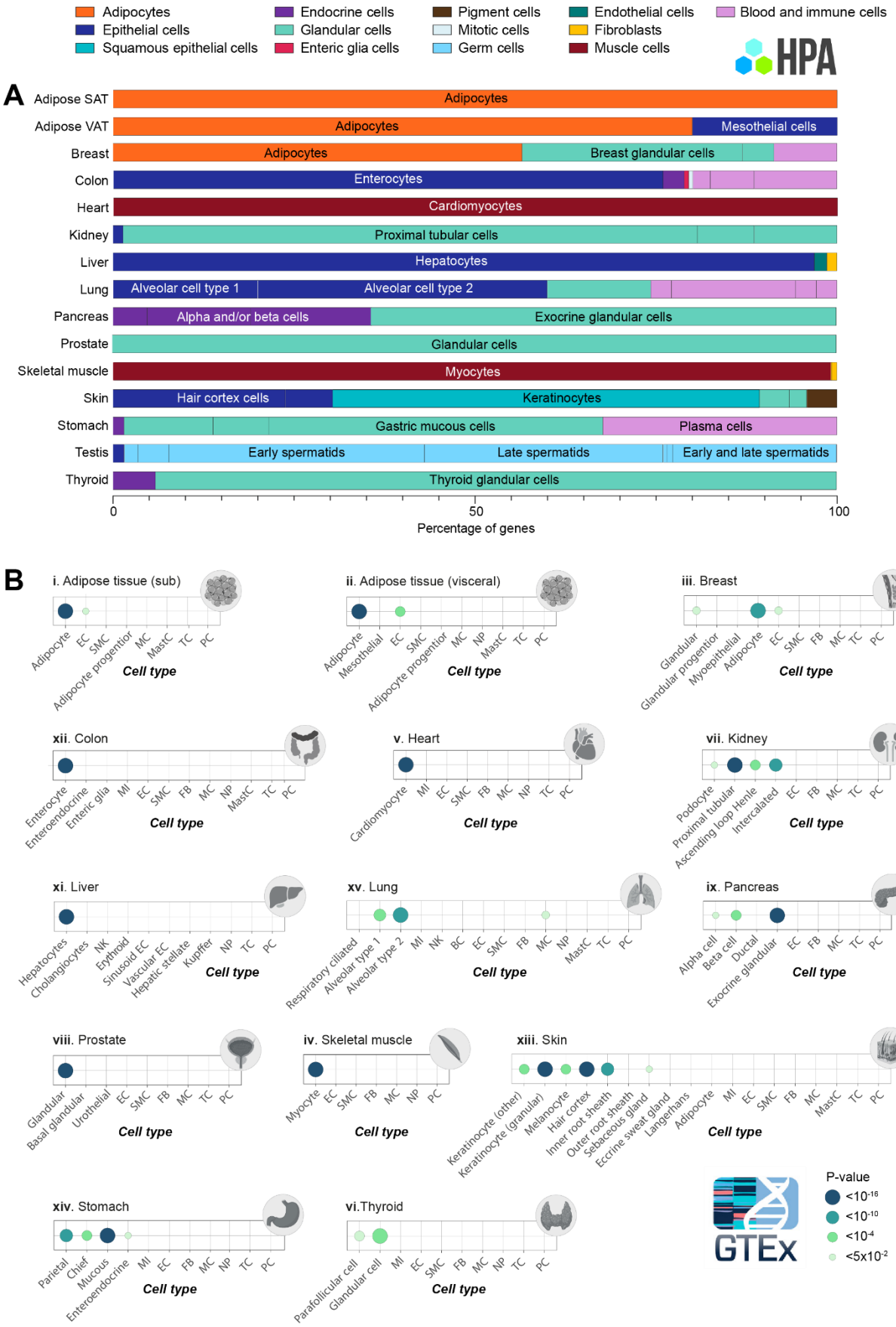
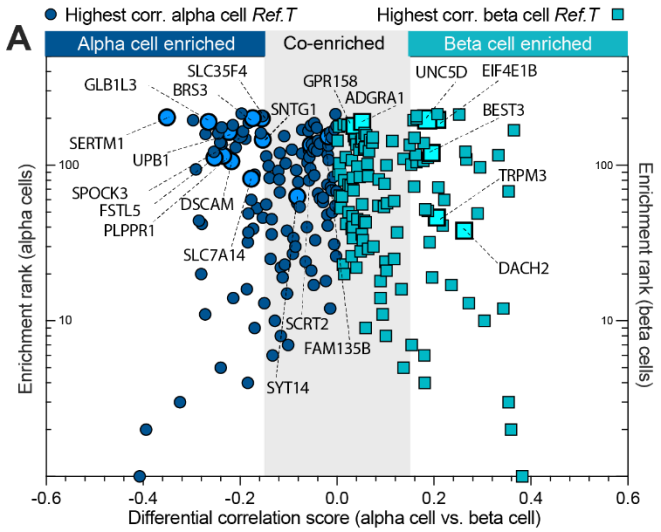
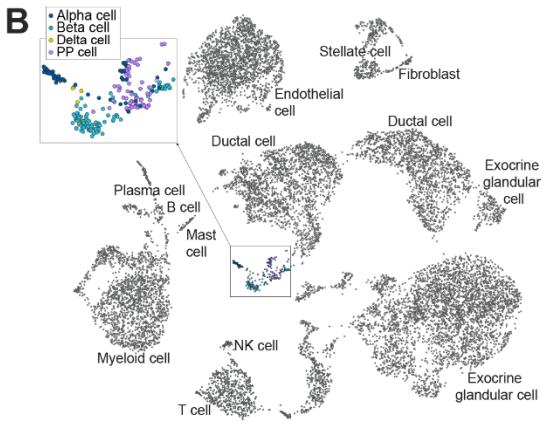

Adipocytes
Epithelial cells
Squamous epithelial cells
Endocrine cells
Glandular cells
Enteric glia cells
Pigment cells
Mitotic cells
Germ cells
Endothelial cells
Fibroblasts
Muscle cells
Blood and immune cells
HPA
A
Adipose SAT
Adipocytes
Adipose VAT
Adipocytes
Mesothelial cells
Breast
Adipocytes
Breast glandular cells
Colon
Enterocytes
Heart
Cardiomyocytes
Kidney
Proximal tubular cells
Liver
Hepatocytes
Lung
Alveolar cell type 1
Alveolar cell type 2
Pancreas
Alpha and/or beta cells
Exocrine glandular cells
Prostate
Glandular cells
Skeletal muscle
Myocytes
Skin
Hair cortex cells
Keratinocytes
Stomach
Gastric mucous cells
Plasma cells
Testis
Early spermatids
Late spermatids
Early and late spermatids
Thyroid
Thyroid glandular cells
0
50
100
Percentage of genes
B
i. Adipose tissue (sub)
ii. Adipose tissue (visceral)
iii. Breast
xii. Colon
v. Heart
vii. Kidney
xi. Liver
xv. Lung
ix. Pancreas
viii. Prostate
iv. Skeletal muscle
xiii. Skin
xiv. Stomach
vi.Thyroid
Cell type
GTEx
P-value
<10-16
<10-10
<10-4
<5x10-2

Figure S3. Integrative co-expression analysis of unfractionated human tissue RNAseq can resolve tissue enriched genes into single cell type expression source. Related to Figure 2. (**A**) Bar plot showing the fraction of predicted cell type enriched genes among the tissue, or tissue-group, enriched genes in Human Protein Atlas (HPA). Colour indicates cell type group. The cell type with the most shared enriched genes with tissues are labelled. (B) Bubble plots showing the significance (indicated by dot size and colour) of similarity between the top 300 tissue enriched genes in GTEx and the predicted cell type enrichment signatures Where overlap is not statistically significant (hypergeometric test, $P > 0.05$), the corresponding dot is removed. EC; endothelial cell, SMC; Smooth muscle cell, MC; macrophage, MastC; mast cell, TC; T-cell, PC; Plasma cell, NP; Neutrophil, MI; Mitotic cell, NK; Natural killer cell.


A
Highest corr. alpha cell Ref.T
Highest corr. beta cell Ref.T
Alpha cell enriched
Co-enriched
Beta cell enriched
GLB1L3
SLC35F4
BRS3
GPR158
ADGRA1
SNTG1
UNC5D
EIF4E1B
BEST3
SERTM1
UPB1
TRPM3
SPOCK3
FSTL5
PLPPR1
DSCAM
SLC7A14
DACH2
SCRT2
FAM135B
SYT14
Enrichment rank (alpha cells)
Enrichment rank (beta cells)
100
10
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Differential correlation score (alpha cell vs. beta cell)


B
Alpha cell
Beta cell
Delta cell
PP cell
Stellate cell
Fibroblast
Endothelial cell
Ductal cell
Ductal cell
Exocrine glandular cell
Plasma cell
B cell
Mast cell
Myeloid cell
NK cell
T cell
Exocrine glandular cell

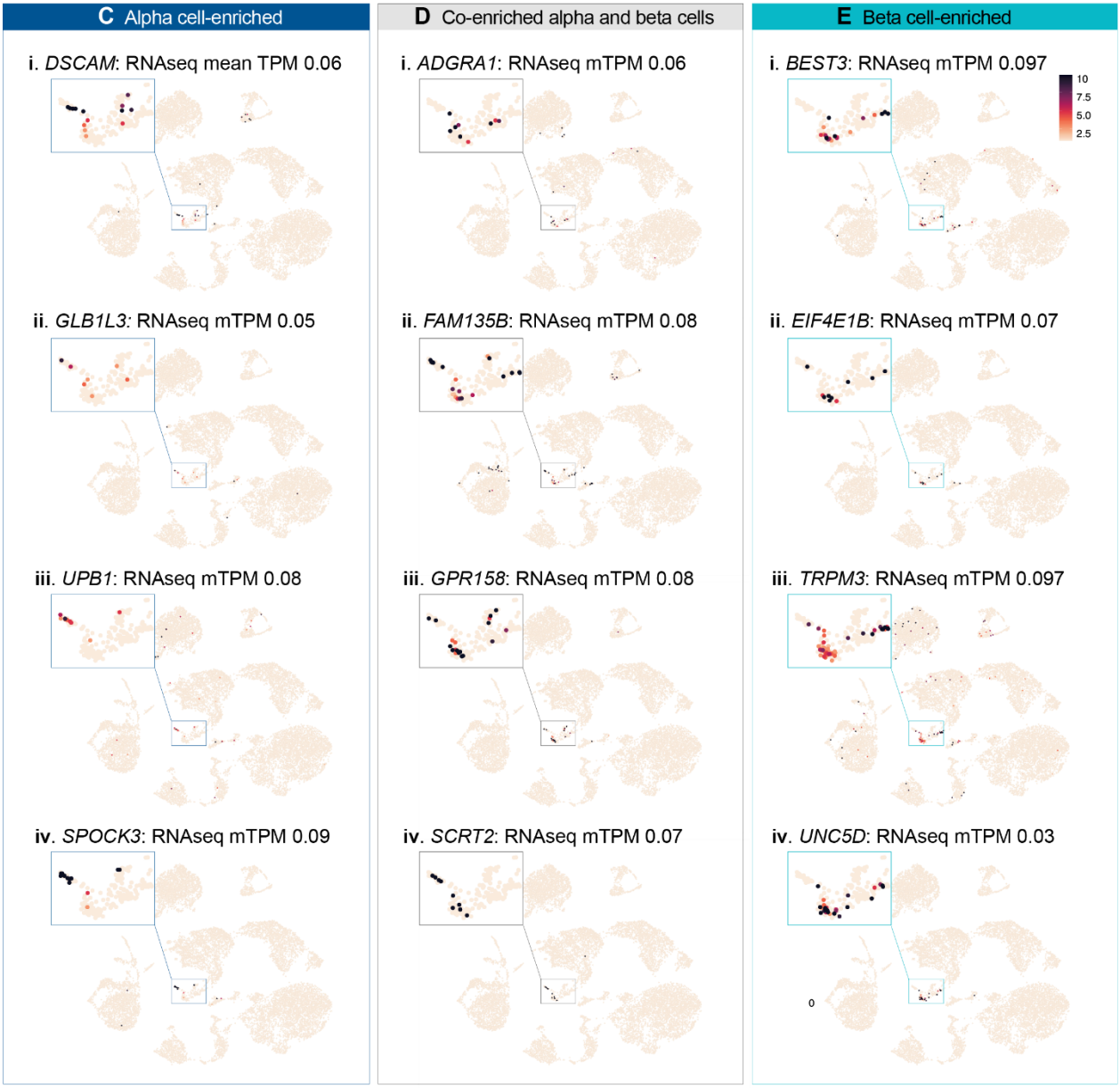

C Alpha cell-enriched
i. DSCAM: RNAseq mean TPM 0.06
ii. GLB1L3: RNAseq mTPM 0.05
iii. UPB1: RNAseq mTPM 0.08
iv. SPOCK3: RNAseq mTPM 0.09
D Co-enriched alpha and beta cells
i. ADGRA1: RNAseq mTPM 0.06
ii. FAM135B: RNAseq mTPM 0.08
iii. GPR158: RNAseq mTPM 0.08
iv. SCRT2: RNAseq mTPM 0.07
E Beta cell-enriched
i. BEST3: RNAseq mTPM 0.097
10
7.5
5.0
2.5
ii. EIF4E1B: RNAseq mTPM 0.07
iii. TRPM3: RNAseq mTPM 0.097
iv. UNC5D: RNAseq mTPM 0.03
0

Figure S4. Reference transcript-based identification of lowly expressed pancreatic alpha and beta cell type enriched and co-enriched genes. Related to Figure 3. RNAseq datasets for human pancreas (n=328) were analysed to generate correlation coefficient values between all protein coding genes and *Ref T*. (**A**) For genes that correlated most highly with alpha (dark blue) or beta cell (turquoise) *Ref.T* (above >0.50), the '*differential correlation score'* (difference between mean corr. with alpha and beta cell *Ref.T.*) was plotted vs. 'enrichment ranking' (position in each respective list, highest corr. = rank 1). Shaded grey box highlights genes enriched in both cell types (co-enriched). Genes highlighted in bold correspond to those featured in the lower panels. scRNAseq data from analysis of human pancreas was sourced from Tabula Sapiens [36], and used to generate UMAP plots showing (**B**) scRNAseq cell type annotations, and the expression profiles of genes we predicted as being (**C**) alpha cell-enriched; (i) *DSCAM*, (ii) *GLB1L3*, (iii) *UPB1* and (iv) *SPOCK3*, (**D**) co-enriched in both alpha and beta cells; (i) *ADGRA1*, (ii) *FAM135B*, (iii) *GPR158* and (iv) *SCRT2*, or (**E**) beta cell-enriched; (i) *BEST3*, (ii) *EIF4E1B*, (iii) *TRPM3* and (iv) *UNC5D.*

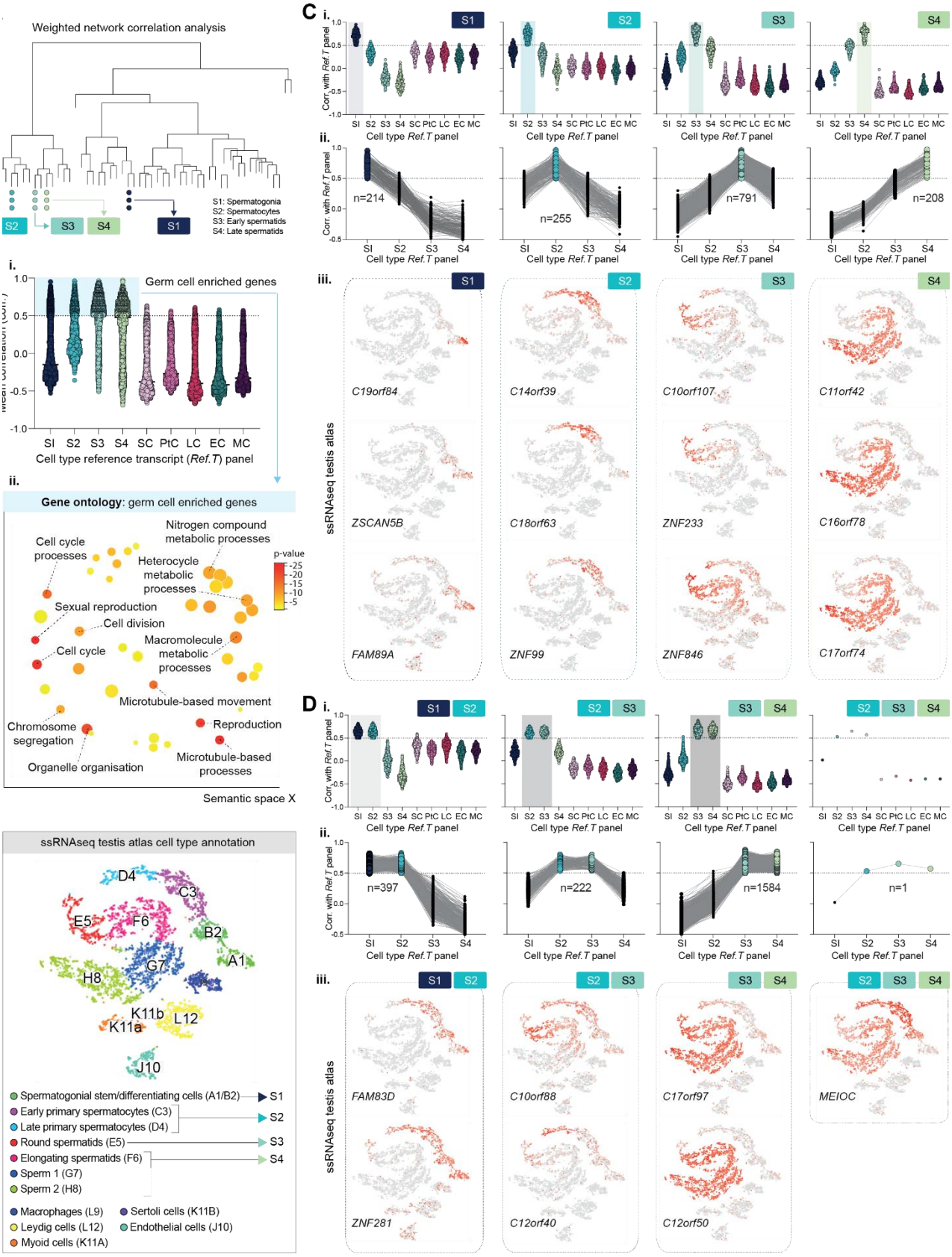

Weighted network correlation analysis
S1: Spermatogonia
S2: Spermatocytes
S3: Early spermatids
S4: Late spermatids
i.
Germ cell enriched genes
Cell type reference transcript (Ref.T) panel
ii.
Gene ontology: germ cell enriched genes
Nitrogen compound metabolic processes
Cell cycle processes
Heterocycle metabolic processes
p-value
Sexual reproduction
Cell division
Macromolecule metabolic processes
Cell cycle
Microtubule-based movement
Chromosome segregation
Reproduction
Microtubule-based processes
Organelle organisation
Semantic space X
ssRNAseq testis atlas cell type annotation
D4
C3
E5
F6
B2
A1
G7
H8
K11b
L12
K11a
J10
Spermatogonial stem/differentiating cells (A1/B2)
Early primary spermatocytes (C3)
Late primary spermatocytes (D4)
Round spermatids (E5)
Elongating spermatids (F6)
Sperm 1 (G7)
Sperm 2 (H8)
Macrophages (L9)
Sertoli cells (K11B)
Leydig cells (L12)
Endothelial cells (J10)
Myoid cells (K11A)
C
i.
ii.
iii.
Corr. with Ref.T panel
Cell type Ref.T panel
n=214
n=255
n=791
n=208
ssRNAseq testis atlas
C19orf84
C14orf39
C10orf107
C11orf42
ZSCAN5B
C18orf63
ZNF233
C16orf78
FAM89A
ZNF99
ZNF846
C17orf74
D
n=397
n=222
n=1584
n=1
FAM83D
C10orf88
C17orf97
MEIOC
ZNF281
C12orf40
C12orf50

Figure S5. Analysis of pseudo temporal changes during spermatogenesis reveals stage-specific and common stage-shared gene enrichment signatures. Related to Figure 4. (**A**) weighted network correlation analysis of human testis RNAseq data (n=361) annotated to show position of genes in *Ref.T*. panels (each indicated with single circle) selected to represent cell types at the different stages of spermatogenesis: S1 (spermatogonia), S2 (spermatocytes), S3 and S4 (early and late spermatids, respectively). (**B**) For genes with predicted cell-type enrichment in S1, S2, S3 or S4 (i) mean correlation coefficients with *Ref.T*. for S1, S2, S3, S4 and sertoli cells (SC), Leydig cells (LC), peritubular cells (PtC), endothelial cells (EC) or macrophages (MC) and (ii) over-represented gene ontology terms, summarised and visualised using REVIGO. For all genes predicted to be: (**C**) highly cell type enriched at one stage of spermatogenesis or (**D**) co-enriched at two or more stages of spermatogenesis (category indicated in top left of each plot): (i) mean correlation coefficients with *Ref.T*. for S1, S2, S3, S4, SC, LC, PtC, EC or MC, (ii) mean correlation coefficients with *Ref.T*. for S1, S2, S3, S4 with linkage lines connecting each individual gene (iii) expression profiles in Human Testis Atlas scRNAseq data (Guo et al., 2018) for selected lesser known genes appearing in each respective category. UMAP from the Human Testis Atlas shows original cell type annotations (bottom left), with arrows to indicating the broad equivalence classifications in our analysis.

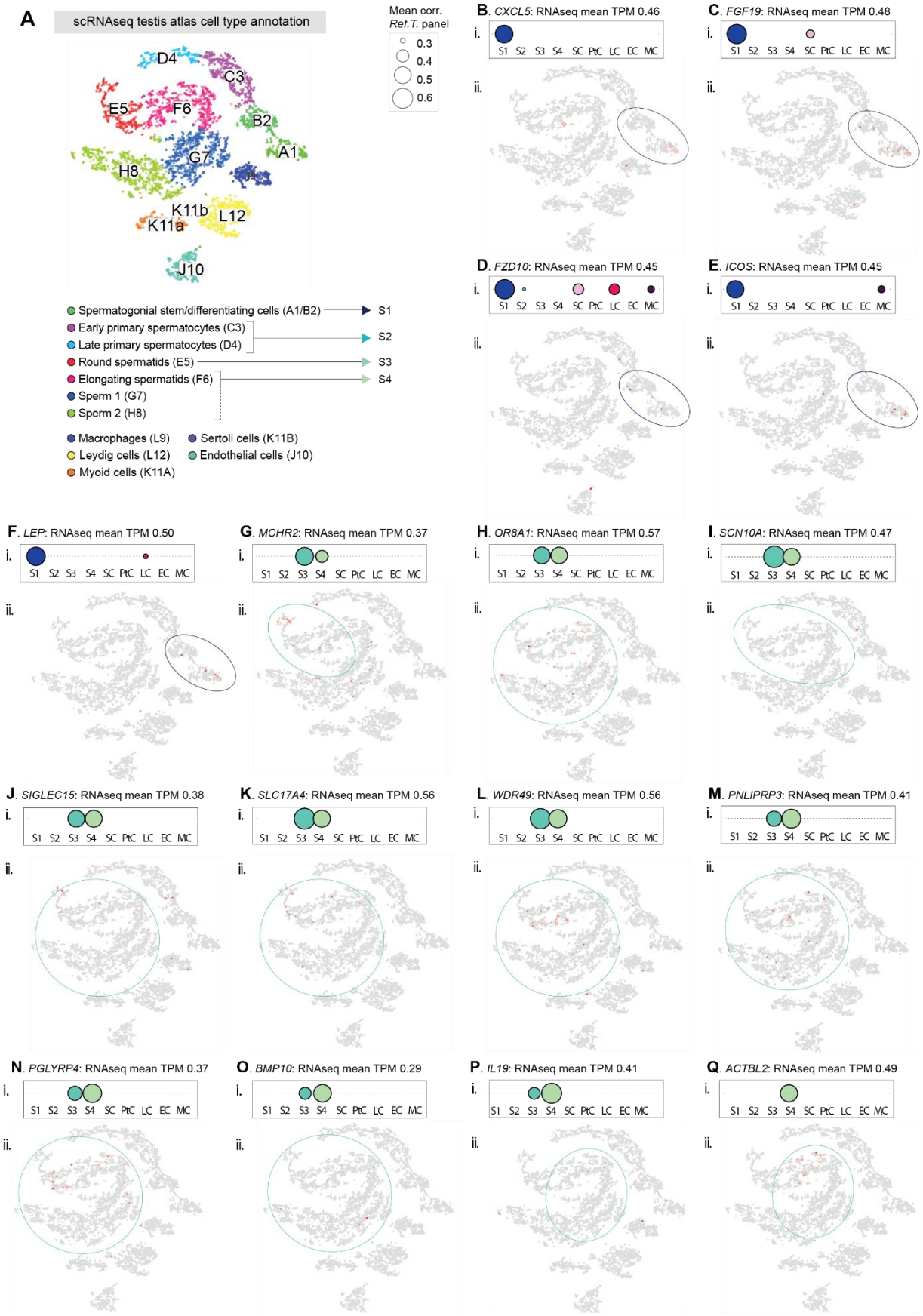

A
scRNAseq testis atlas cell type annotation
Mean corr. Ref.T. panel
0.3
0.4
0.5
0.6
D4
C3
E5
F6
B2
A1
G7
H8
K11b
L12
K11a
J10
Spermatogonial stem/differentiating cells (A1/B2)
Early primary spermatocytes (C3)
Late primary spermatocytes (D4)
Round spermatids (E5)
Elongating spermatids (F6)
Sperm 1 (G7)
Sperm 2 (H8)
S1
S2
S3
S4
Macrophages (L9)
Sertoli cells (K11B)
Leydig cells (L12)
Endothelial cells (J10)
Myoid cells (K11A)
B. CXCL5: RNAseq mean TPM 0.46
C. FGF19: RNAseq mean TPM 0.48
D. FZD10: RNAseq mean TPM 0.45
E. ICOS: RNAseq mean TPM 0.45
F. LEP: RNAseq mean TPM 0.50
G. MCHR2: RNAseq mean TPM 0.37
H. OR8A1: RNAseq mean TPM 0.57
I. SCN10A: RNAseq mean TPM 0.47
J. SIGLEC15: RNAseq mean TPM 0.38
K. SLC17A4: RNAseq mean TPM 0.56
L. WDR49: RNAseq mean TPM 0.56
M. PNLIPRP3: RNAseq mean TPM 0.41
N. PGLYRP4: RNAseq mean TPM 0.37
O. BMP10: RNAseq mean TPM 0.29
P. IL19: RNAseq mean TPM 0.41
Q. ACTBL2: RNAseq mean TPM 0.49
S1 S2 S3 S4 SC PtC LC EC MC

Figure S6. Reference transcript-based identification of lowly expressed germ cell enriched genes in the human testis. Related to Figure 4. (**A**) UMAP and cell type annotations as defined in the scRNAseq Human Testis Atlas (Guo et al., 2018), with arrows to indicate the broad equivalence classifications in our analysis. (i) Enrichment scores in all cell types profiled for genes predicted to be (**B**-**F**) S1 enriched, (**H**-**N**) S3 and S4 enriched or (**O**-**Q**) S4 enriched, with (ii) corresponding UMAP expression plots from the scRNAseq Human Testis Atlas (Guo et al., 2018).

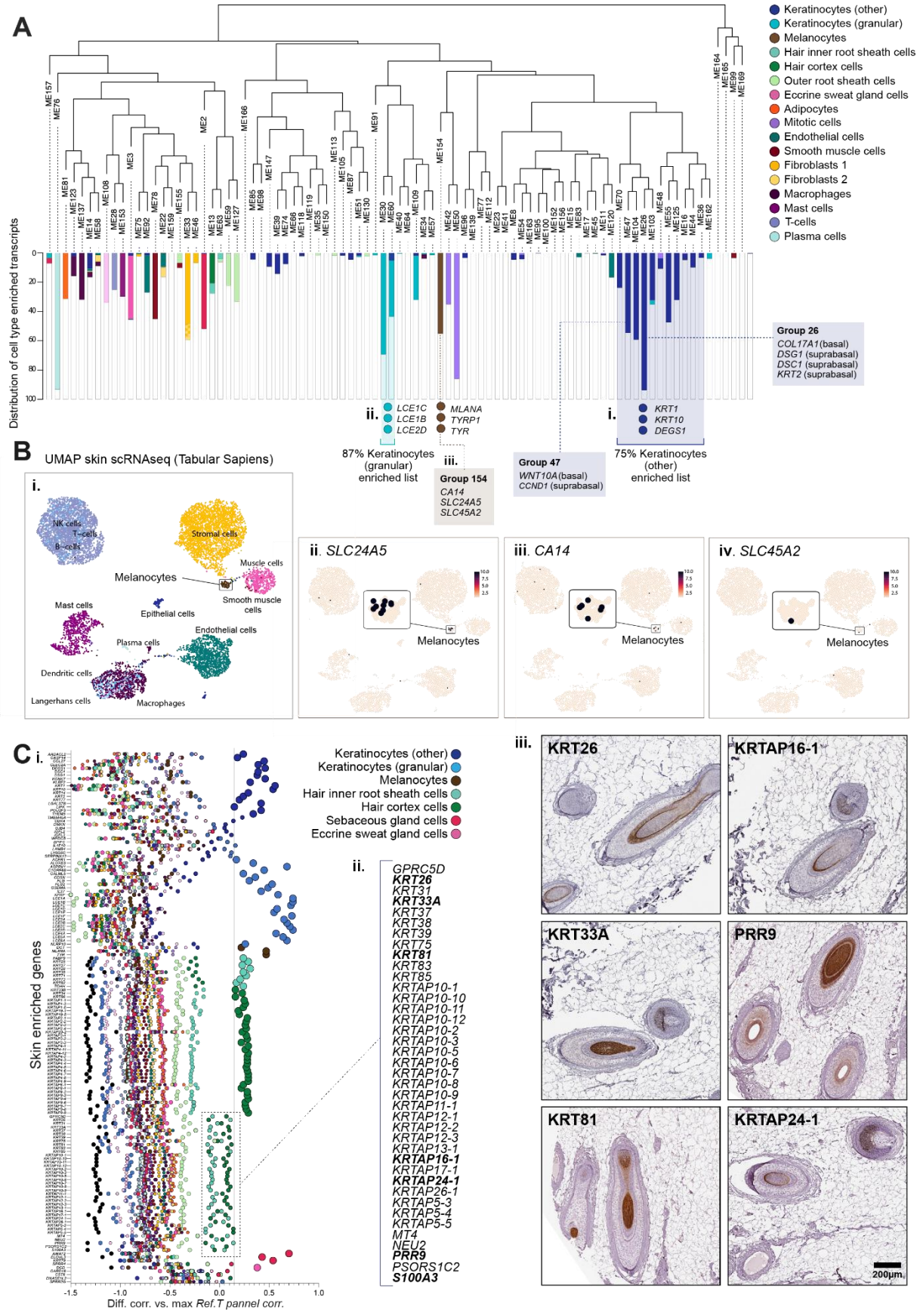

A
Keratinocytes (other)
Keratinocytes (granular)
Melanocytes
Hair inner root sheath cells
Hair cortex cells
Outer root sheath cells
Eccrine sweat gland cells
Adipocytes
Mitotic cells
Endothelial cells
Smooth muscle cells
Fibroblasts 1
Fibroblasts 2
Macrophages
Mast cells
T-cells
Plasma cells
Distribution of cell type enriched transcripts
Group 26
COL17A1 (basal)
DSG1 (suprabasal)
DSC1 (suprabasal)
KRT2 (suprabasal)
ii. LCE1C LCE1B LCE2D
MLANA TYRP1 TYR
i. KRT1 KRT10 DEGS1
87% Keratinocytes (granular) enriched list
iii.
Group 154
CA14
SLC24A5
SLC45A2
Group 47
WNT10A (basal)
CCND1 (suprabasal)
75% Keratinocytes (other) enriched list
B
UMAP skin scRNAseq (Tabular Sapiens)
i.
NK cells
T-cells
B-cells
Stromal cells
Muscle cells
Melanocytes
Smooth muscle cells
Epithelial cells
Mast cells
Endothelial cells
Plasma cells
Dendritic cells
Langerhans cells
Macrophages
ii. SLC24A5
iii. CA14
iv. SLC45A2
Melanocytes
C i.
Keratinocytes (other)
Keratinocytes (granular)
Melanocytes
Hair inner root sheath cells
Hair cortex cells
Sebaceous gland cells
Eccrine sweat gland cells
Skin enriched genes
Diff. corr. vs. max Ref.T pannel corr.
ii.
GPRC5D
KRT26
KRT31
KRT33A
KRT37
KRT38
KRT39
KRT75
KRT81
KRT83
KRT85
KRTAP10-1
KRTAP10-10
KRTAP10-11
KRTAP10-12
KRTAP10-2
KRTAP10-3
KRTAP10-5
KRTAP10-6
KRTAP10-7
KRTAP10-8
KRTAP10-9
KRTAP11-1
KRTAP12-1
KRTAP12-2
KRTAP12-3
KRTAP13-1
KRTAP16-1
KRTAP17-1
KRTAP24-1
KRTAP26-1
KRTAP5-3
KRTAP5-4
KRTAP5-5
MT4
NEU2
PRR9
PSORS1C2
S100A3
iii.
KRT26
KRTAP16-1
KRT33A
PRR9
KRT81
KRTAP24-1
200µm

Figure S7. Constituent cells of the skin hair root are the primary source of skin tissue enriched genes. Related to Figure 5. (**A**) Weighted network correlation analysis (WGNCA) of human skin samples (n=210) with coloured coded bars showing distribution of genes predicted to be cell type enriched. Position of *Ref.T*. and example cell-type enriched genes are highlighted for: (i) supra-basal keratinocytes, (ii) granular keratinocytes and (iii) melanocytes. (**B**) scRNAseq data and cell type definitions were sourced for human skin from Tabula Sapiens (Tabula Sapiens et al., 2022) and used to generate UMPA plots showing: (i) cell type annotations or expression profiles for genes we predicted to be melanocyte enriched (ii) *SLC24A5*, (iii) *CA14* and (iv) *SLC45A2*. (**C**) Skin enriched genes (vs. other tissue types) were identified and (i) corresponding cell type enrichment profiles in skin plotted, a panel of which (ii) did not reach the threshold for classification as enriched in a single cell type but had highest enrichment scores in one or more hair cell types. (iii) Expression of proteins encoded by selected examples were profiled in human skin tissue containing hair roots.

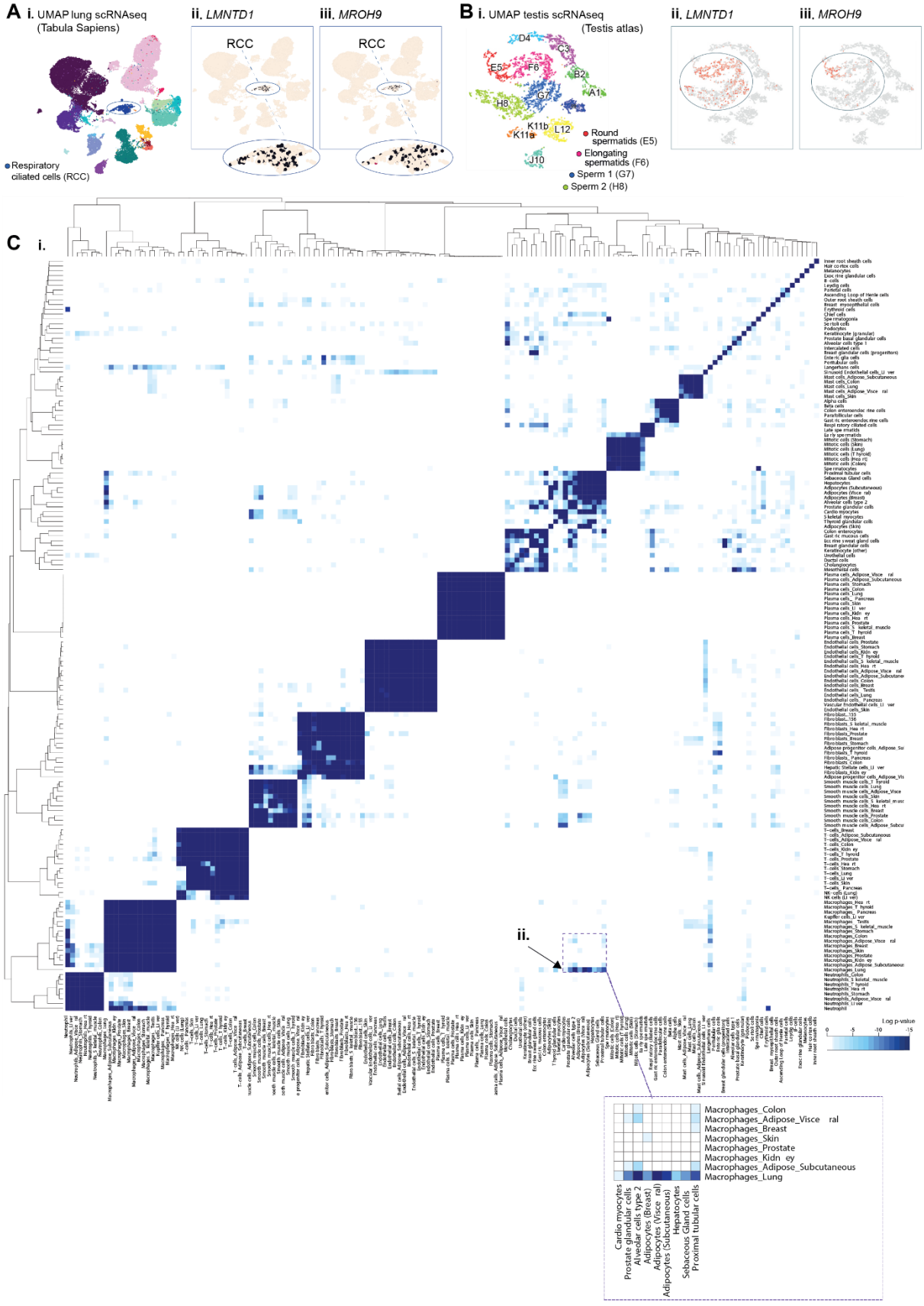

A i. UMAP lung scRNAseq
(Tabula Sapiens)
ii. LMNTD1
iii. MROH9
RCC
RCC
Respiratory ciliated cells (RCC)
B i. UMAP testis scRNAseq
(Testis atlas)
ii. LMNTD1
iii. MROH9
D4
C3
E5
F6
B2
A1
G7
H8
K11b
K11a
L12
J10
Round spermatids (E5)
Elongating spermatids (F6)
Sperm 1 (G7)
Sperm 2 (H8)
C i.
ii.
Log p-value
0
-5
-10
-15
Macrophages_Colon
Macrophages_Adipose_Visce ral
Macrophages_Breast
Macrophages_Skin
Macrophages_Prostate
Macrophages_Kidn ey
Macrophages_Adipose_Subcutaneous
Macrophages_Lung
Cardio myocytes
Prostate glandular cells
Alveolar cells type 2
Adipocytes (Breast)
Adipocytes (Visce ral)
Adipocytes (Subcutaneous)
Hepatocytes
Sebaceous Gland cells
Proximal tubular cells

Figure S8. Cell type enriched signature comparisons. Related to Figure 6 and 7. scRNAseq data was sourced for human (**A**) lung from Tabula Sapiens (Tabula Sapiens et al., 2022) or (**B**) testis from the Human Testis Atlas (Guo et al., 2018), and used to generate UMAP plots to show (i) cell type annotation as according to the original studies, or expression profiles of (ii) *LMNTD1* or (iii) *MROH9*. (**C**) Heatmap showing significance p-values for similarity scores, calculated using a hypergeometric test, between: (i) all predicted cell type enriched genes, and (ii) lung macrophages vs. other non-macrophage cell types.

Paper IV

KANK3 is a shear stress regulated endothelial protein with a role in cell migration and tissue factor regulation

Eike Christopher Struck[1], Sofia Maria Öling[1], Philip James Dusart[2], Marthe Norreen-Thorsen[1], Julian Connor Eckel, Larissa Dorothea Kruse[3], Casper Ullsten-Wahlund[1], Jacob Odeberg[1, 2, 4], Clément Naudin[1,2,] Lynn Marie Butler[1, 2, 5]

[1]Department of Clinical Medicine, The Arctic University of Norway, N-9037, Tromsø, Norway

[2]Science for Life Laboratory, Department of Protein Science, School of Engineering Sciences, Stockholm, Sweden

[3]Department of Medical Biology, The Arctic University of Norway, N-9037, Tromsø, Norway

[4]The University Hospital of North Norway (UNN), PB100, 9038 Tromsø, Norway

[5]Clinical Chemistry and Blood Coagulation Research, Department of Molecular Medicine and Surgery, Karolinska Institute, SE-171 76 Stockholm, Sweden

Correspondence information:

Dr. L.M Butler, PhD

Department of Clinical Medicine,

The Arctic University of Norway,

N-9037, Tromsø,

Norway

Email: Lynn.m.butler@uit.no

KEY WORDS: Endothelium, KANK3, focal adhesions, cytoskeleton, cell migration, coagulation

ABSTRACT

The endothelium is the innermost layer of all blood vessels. Endothelial cells (EC) play a central role in the regulation of vascular processes, such as coagulation, inflammation, and angiogenesis. Proteins with EC restricted expression tend to be critical for such cell type specific functions. In a previously published bioinformatic based analysis of RNAseq, we predicted that *KANK3,* which encodes an uncharacterised protein, had body wide enriched expression in human EC. Here, we verify that KANK3 is a body-wide endothelial-enriched protein at the transcript and protein level. We characterise its subcellular distribution in primary EC and uncover that its expression is strongly induced in response to shear stress exposure. When KANK3 protein was depleted using siRNA, the distribution of the EC intermediate filament vimentin was disrupted in both static and shear stress exposed cultures, indicating a direct or indirect interaction between these proteins. Correspondingly, in a wound healing model, depletion of KANK3 increased EC migratory capacity, but did not increase proliferative capacity. Furthermore, we observed an increase in the expression of the pro-coagulant protein tissue factor in KANK3 depleted EC, indicating that it could have further regulatory roles, beyond those associated with cytoskeletal modification and motility.

INTRODUCTION

The vascular endothelium lines the inside of all blood and lymphatic vessels and has numerous functions, including in the regulation of inflammation, haemostasis, and blood pressure [1, 2]. Proteins specifically expressed in endothelial cells (EC) tend to have central roles in cell specialised functions, e.g., cadherin-5 (CDH5), claudin-5 (CLDN5) and endothelial cell-selective adhesion molecule (ESAM), play established roles in EC integrity, polarity and shape, vessel permeability and signalling [3-6], and the vascular endothelial growth factor receptor 1 (FLT1) and 2 (KDR) are central to angiogenesis [7]. In earlier work, based on bioinformatic analysis of bulk RNAseq, we predicted that the gene encoding the uncharacterised protein KN Motif And Ankyrin Repeat Domains 3 (KANK3) had body wide enriched expression in human EC [8].

The KANK family consist of four members (KANK 1-4), which arose through gene duplication and diversification, with strong conservation across the evolutionary tree [9, 10]. They are defined by their unique structure, consisting of a variable number of coiled-coil motifs in the central N-terminal regions, five ankyrin repeats in the C-terminal region and a talin-binding KN-motif domain at the N-terminus [10, 11]. The interaction between KANK1 and talin regulates the recruitment of complexes that stabilize cortical microtubules to focal adhesions [11]. KANK2 promotes the creation of central adhesions by triggering talin activation and is responsible for the reduction of force transduction across integrins [12]. KANK1 and 2 are involved in cell migration and adhesion, via interactions with kinesin family member 21A (KIF21A), and the regulation of its activity through its coiled-coil domain [10, 13].

The cytoskeleton and focal adhesions are crucial for various EC specialised functions, such as the maintenance of the structural integrity required to withstand the mechanical forces exerted by the blood flow [14], to control movement and migration during processes such as angiogenesis [15], to stabilise junctional connections and control vascular permeability [16] and in processes such as coagulation [17] and inflammation [18].

While KANK1 and KANK2 are well relatively studied, KANK3 is comparably poorly described; currently, it has no reported function in a vascular context in vertebrates. A homologue of KANK3 has been described in vascular EC in of zebrafish embryos, where it was essential for embryonic development and survival, with a potential role in cell adhesion and tissue integrity [19, 20]. Over expression of KANK3 in NIH3T3 cells revealed a possible role in actin stress fibre formation [21], and other studies have indicated a role in the regulation of cell migration in hepatocellular carcinoma [22] and lung adenocarcinoma [23].

In this study, we verify that KANK3 is a body-wide endothelial-enriched protein. We characterise its subcellular distribution in primary EC, and report that it is shear stress-induced gene. We show that KANK3 depletion modifies the subcellular distribution of the EC intermediate filament vimentin and increases EC motility in a gap closing assay. We observed an increase in the expression of the pro-coagulant protein tissue factor in KANK3 depleted EC, particularly under inflammatory conditions, together with an increased capacity to induce thrombin generation in plasma. Thus, we demonstrate a role for the EC enriched protein KANK3 in EC specialised functions.

RESULTS

KANK3 IS AN ENDOTHELIAL ENRICHED PROTEIN IN HUMAN

KANK3 mRNA expression correlates with endothelial cell genes in human tissues

Proteins expressed specifically by EC tend to be critical for EC specialised functions. Previously, using mixed tissue bulk RNAseq, we found that *KANK3* expression was strongly correlated with EC marker genes, indicating EC specificity [8]. More recently, using a similar approach, we profiled gene enrichment signatures for cell types in 15 individual tissue datasets [24-27] (data is displayed on the Human Protein Atlas [www.proteinatlas.org/humanproteome/tissue+cell+type]). Here, *KANK3* was predicted to be EC enriched in multiple vascular beds (Figure S1 A.i), whilst other KANK family members, *KANK1*, *2* or *4* were not (Figure S1 B-D.i).

To further explore this potential relationship using a *KANK3*-centric approach in an expanded dataset, we retrieved bulk RNAseq datasets for 36 human tissue types from Genotype-Tissue Expression (GTEx) V8 (www.gtexportal.org)[28] (mean samples/tissue =377, range 85-803) (Table S1A). For each dataset, we calculated Pearson correlation coefficient values between *KANK3* and all other mapped genes (Table S1B). The top 100 most highly correlating genes with *KANK3* (all correlation coefficient [corr.] >0.5, p-value<0.0001) in each tissue type (Table S1C) were cross-compared, to identify 67 genes that were highly correlated with *KANK3* in 10 or more tissue types (Figure 1A) (Table S1D). These genes included *EGFL7* (32 tissues; mean corr. =0.76), *ROBO4* (29 tissues; mean corr. =0.74), *ESAM* (29 tissues; mean corr. =0.76) and *CDH5* (29 tissues; mean corr. =0.72); all of which have key roles in EC specific functions [29-32].

Gene ontology (GO) analysis [33] was performed to identify over represented groups within this list of 67 genes (Figure 1B and Table S1E). Over-represented terms were related to vascular or EC function and included ´*vasculature development*´ ($p=1.1 \times 10^{13}$), ´*angiogenesis*´ ($p=1.8 \times 10^{11}$) and ´*establishment of endothelial barrier*´ ($p=3.2 \times 10^{7}$) (Figure 1 B) (Table S1E). As high correlation values between genes within tissue can indicate co-expression in a common cell-type, these results are consistent with our prediction that *KANK3* is an EC enriched gene.

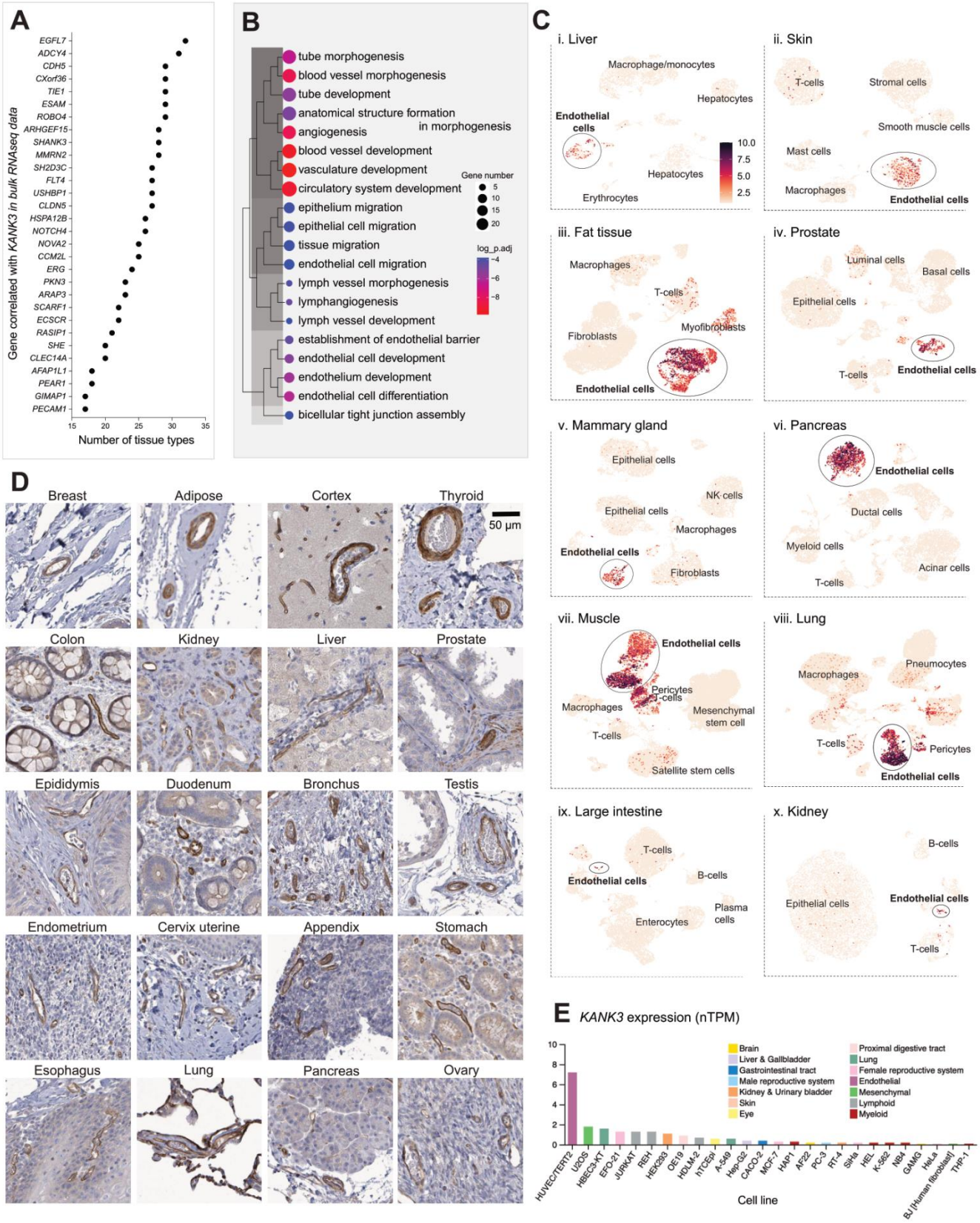

A
Gene correlated with KANK3 in bulk RNAseq data
EGFL7
ADCY4
CDH5
CXorf36
TIE1
ESAM
ROBO4
ARHGEF15
SHANK3
MMRN2
SH2D3C
FLT4
USHBP1
CLDN5
HSPA12B
NOTCH4
NOVA2
CCM2L
ERG
PKN3
ARAP3
SCARF1
ECSCR
RASIP1
SHE
CLEC14A
AFAP1L1
PEAR1
GIMAP1
PECAM1
15
20
25
30
35
Number of tissue types
B
tube morphogenesis
blood vessel morphogenesis
tube development
anatomical structure formation in morphogenesis
angiogenesis
blood vessel development
vasculature development
circulatory system development
epithelium migration
epithelial cell migration
tissue migration
endothelial cell migration
lymph vessel morphogenesis
lymphangiogenesis
lymph vessel development
establishment of endothelial barrier
endothelial cell development
endothelium development
endothelial cell differentiation
bicellular tight junction assembly
Gene number
5
10
15
20
log_p.adj
−4
−6
−8
C
i. Liver
Macrophage/monocytes
Hepatocytes
Endothelial cells
Hepatocytes
Erythrocytes
10.0
7.5
5.0
2.5
ii. Skin
T-cells
Stromal cells
Smooth muscle cells
Mast cells
Macrophages
Endothelial cells
iii. Fat tissue
Macrophages
T-cells
Myofibroblasts
Fibroblasts
Endothelial cells
iv. Prostate
Luminal cells
Basal cells
Epithelial cells
T-cells
Endothelial cells
v. Mammary gland
Epithelial cells
NK cells
Epithelial cells
Macrophages
Endothelial cells
Fibroblasts
vi. Pancreas
Endothelial cells
Ductal cells
Myeloid cells
Acinar cells
T-cells
vii. Muscle
Endothelial cells
Pericytes
T-cells
Macrophages
Mesenchymal stem cell
T-cells
Satellite stem cells
viii. Lung
Macrophages
Pneumocytes
T-cells
Pericytes
Endothelial cells
ix. Large intestine
T-cells
Endothelial cells
B-cells
Plasma cells
Enterocytes
x. Kidney
B-cells
Epithelial cells
Endothelial cells
T-cells
D
Breast
Adipose
Cortex
Thyroid
50 µm
Colon
Kidney
Liver
Prostate
Epididymis
Duodenum
Bronchus
Testis
Endometrium
Cervix uterine
Appendix
Stomach
Esophagus
Lung
Pancreas
Ovary
E
KANK3 expression (nTPM)
10
8
6
4
2
0
Brain
Liver & Gallbladder
Gastrointestinal tract
Male reproductive system
Kidney & Urinary bladder
Skin
Eye
Proximal digestive tract
Lung
Female reproductive system
Endothelial
Mesenchymal
Lymphoid
Myeloid
HUVEC/TERT2
U2OS
HBEC3-KT
EPO-21
JURKAT
REH
HEK293
OE19
HDLM-2
hTCtpl
A-549
Hep-G2
CACO-2
MCF-7
HAP1
AF22
PC-3
RT-4
SiHa
HEL
K-562
NB4
GAMG
HeLa
BJ [Human fibroblast]
THP-1
Cell line

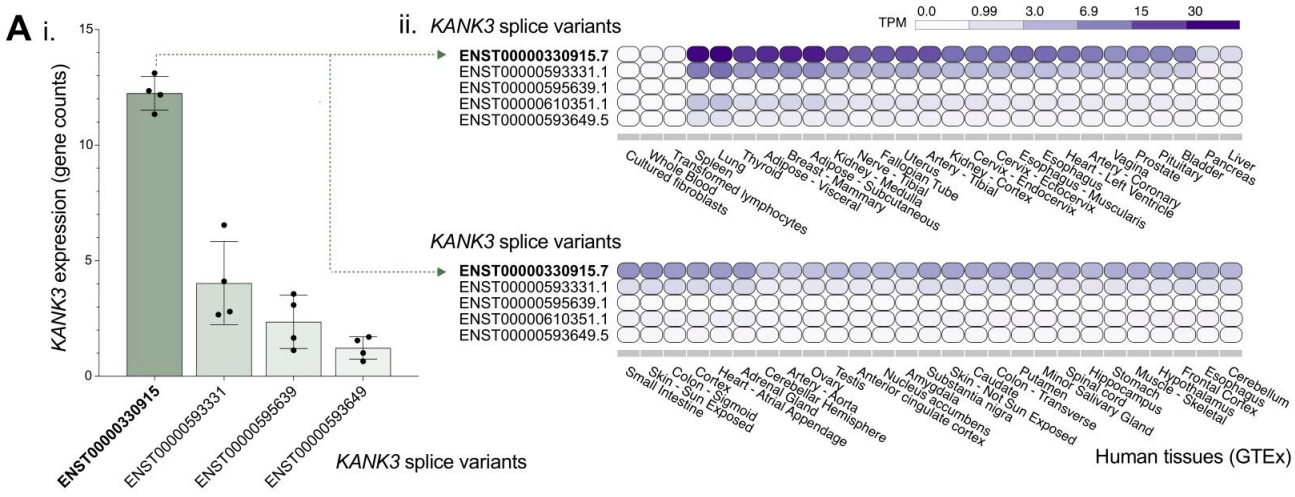
Figure 1. KANK3 is an endothelial cell enriched protein in the human. RNAseq data from 36 human tissue types was sourced from Genotype-Tissue Expression (www.gtexportal.org) [28]. Pearson correlation coefficient values between *KANK3* and all other mapped genes were calculated for each. (**A**) Genes most frequently among the top 100 most highly *KANK3* correlated genes (all >0.50, p<0.0001) across tissue types (**B**) Gene Ontology over-represented terms associated with genes among the top 100 most highly *KANK3* correlated genes in ≥10 tissues. (**C**) Data was downloaded from Tabula Sapiens [34] and used to generate Uniform manifold approximation and projection (UMAP) visualizations for *KANK3* expression in human: (i) liver, (ii) skin, (iii) fat, (iv) prostate, (v) mammary gland, (vi) pancreas, (vii) muscle, (viii) lung, (ix) large intestine and (x) kidney. (**D**) Protein profiling for KANK3 across human tissue types. (**D**) Immortalised human cell lines in the panel tested that had with the highest expression of *KANK3*, generated as part of the Human Protein Atlas project (ref).

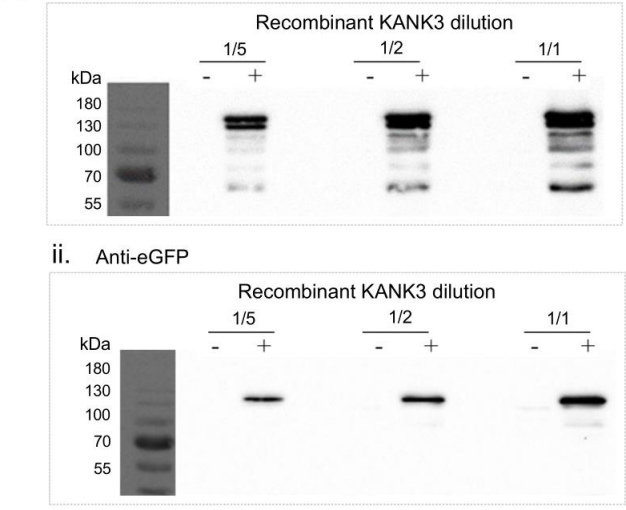
KANK3 endothelial enriched expression can be verified by scRNAseq and protein profiling

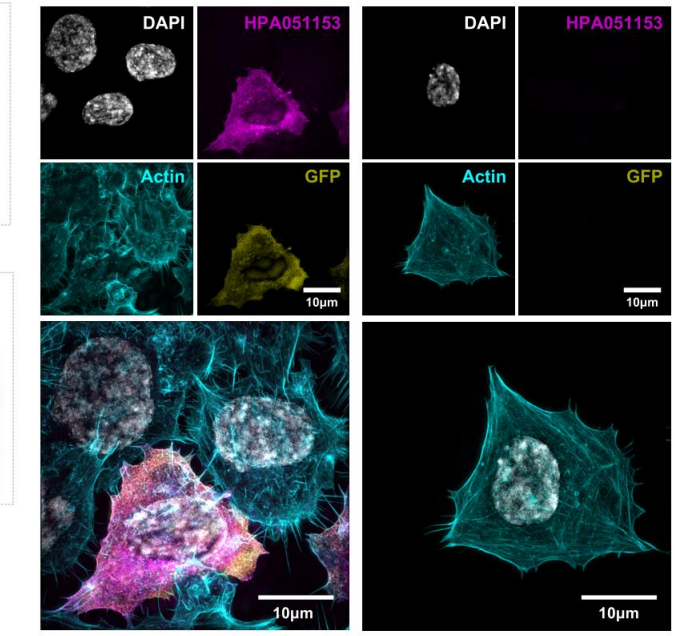
Single cell RNAseq data from the Tabula Sapiens [34] was used to explore expression profiles of KANK3 in skin, liver, fat, prostate, mammary gland, pancreas, muscle, lung, large intestine, and kidney (Figure 1 C.i-x). In all cases, KANK3 was predominantly expressed within clusters annotated as EC. Low levels of *KANK3* were detected in myofibroblasts in fat (Figure S1 C.iii), and pericytes in muscle and lung (Figure 1C.vii and viii). There was little or no *KANK3* expression in tissue specific cell types, e.g., hepatocytes in the liver (Figure 1 C.i) or pneumocytes in the lung (Figure 1 C.viii). Protein profiling confirmed EC expression of KANK3 in multiple tissues, including colon, kidney, liver, breast, adipose, cortex, prostate, skeletal muscle, thyroid (Figure 1D) and others (Figure S1B) [35].

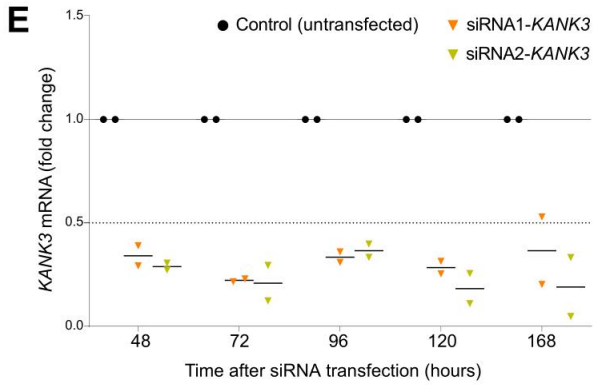
KANK3 expression is enriched in cell lines of endothelial origin

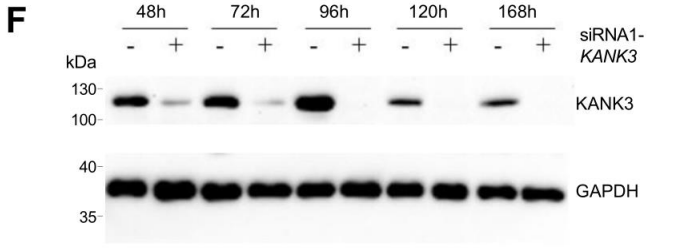
To determine if *KANK3* expression is maintained in cells of EC origin following immortalisation, we examined its expression in RNA-sequencing data from different 41 cell lines from the HPA [35]. *KANK3* was not detectable, or detectable only at very low levels (nTPM ≤ 0.5) in 30/41 (73%) of the cell lines tested. The highest expression was detected in human umbilical vein endothelial cells (HUVEC), HUVEC/TERT2 cells (7.2 nTPM), followed by the mesenchymal cell line U2OS cells (1.8 nTPM) (Figure 1E) (data for all shown in Figure S1). In comparison, other members of the KANK family (KANK1, 2 and 4), which we have previously predicted to lack EC specificity across human tissue types [36], show no EC specificity in immortalised cell lines (Figure S2) or in scRNAseq data from Tabula Sapiens (Figure S2). Together, these data support our prediction [8, 24, 27, 36] that *KANK3*, but not *KANK1*, *KANK2* or *KANK4*, is a human endothelial enriched gene.


A i.
KANK3 expression (gene counts)
ENST00000330915
ENST00000593331
ENST00000595639
ENST00000593649
KANK3 splice variants
ii. KANK3 splice variants
TPM
0.0 0.99 3.0 6.9 15 30
ENST00000330915.7
ENST00000593331.1
ENST00000595639.1
ENST00000610351.1
ENST00000593649.5
Human tissues (GTEx)


B i. HPA051153
Recombinant KANK3 dilution
1/5
1/2
1/1
kDa
180
130
100
70
55
ii. Anti-eGFP
Recombinant KANK3 dilution


C i. KANK3 transfected HEK293
ii. Untransfected HEK293
DAPI
HPA051153
Actin
GFP
10µm


E
Control (untransfected)
siRNA1-KANK3
siRNA2-KANK3
KANK3 mRNA (fold change)
Time after siRNA transfection (hours)


F
48h
72h
96h
120h
168h
siRNA1-KANK3
kDa
KANK3
GAPDH

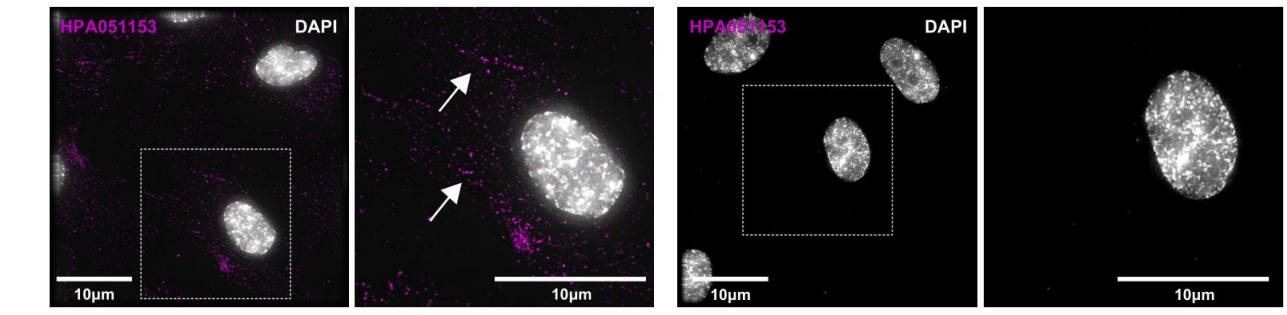

G i. Untransfected HUVEC
ii. siRNA1-KANK3 transfected HUVEC
HPA051153
DAPI
10µm

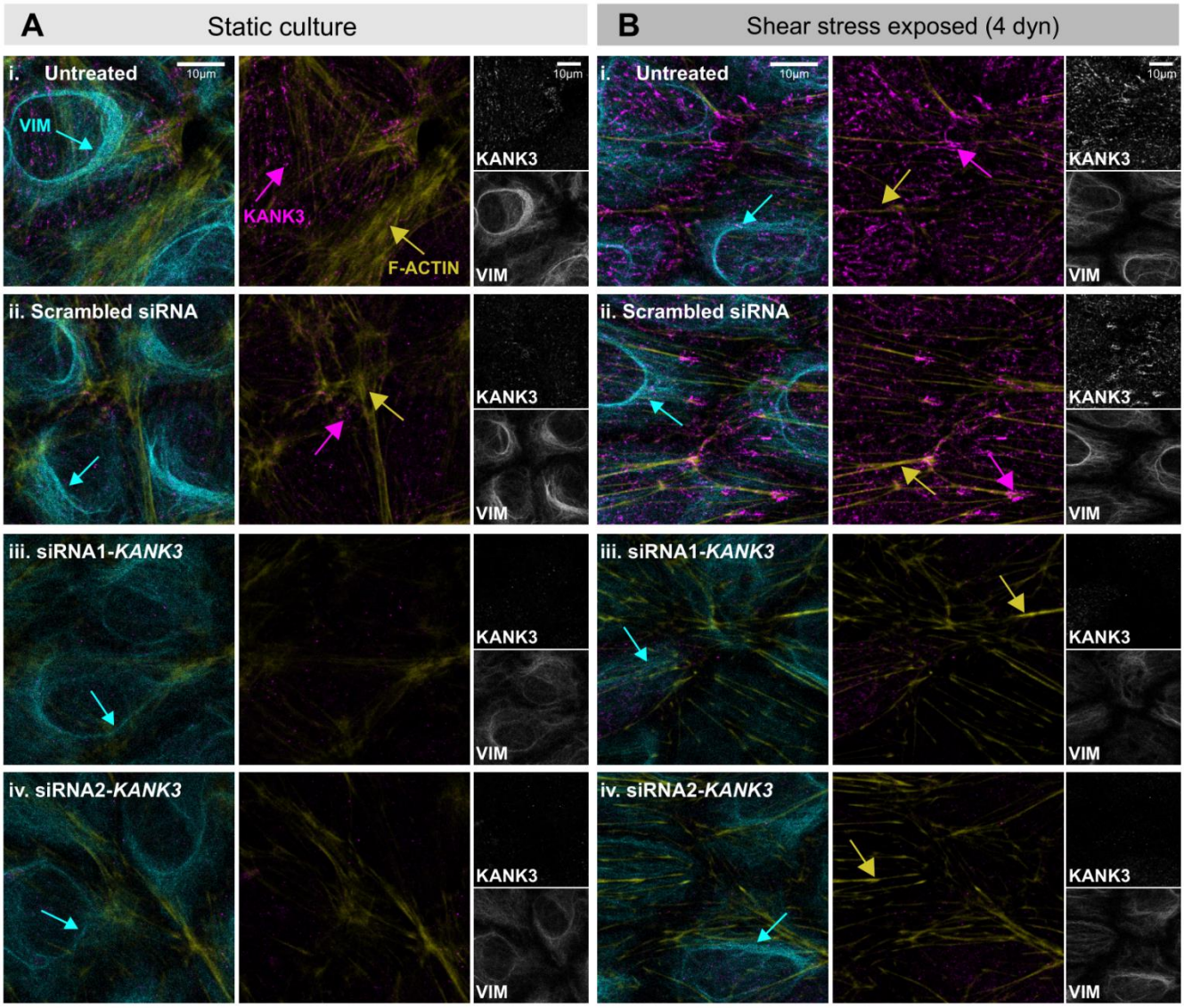
Figure 2. Validation of tools for the study of KANK3 function. (**A**) Detection of KANK3 splice variants by RNAseq in: (i) primary HUVEC and (ii) unfractionated human tissue from Genotype-Tissue Expression (www.gtexportal.org). HEK293 cells were untreated or transfected with KANK3-eGFP expressing plasmids and used to generate (**B**) cell lysates for Western Blot analysis using primary antibodies: (i) HPA051153, a rabbit polyclonal antibody raised to target KANK3, or (ii) an anti-eGFP antibody, or (**C**) immunocytochemistry staining to show signals from eGFP (yellow), HPA051153 (Magenta), Actin (Phalloidin-647; yellow) or DAPI (Nuclear staining, Gray), in (i) KANK3 transfected or (ii) untreated cells. HUVEC were transfected with siRNAs targeting KANK3 and cultured for between 48 and 168h before cell lysis and (**E**) measurement of KANK3 mRNA expression using real time qPCR and (**F**) Western blot analysis using primary antibody HPA051153 and an anti-GAPDH as loading control, or (G) immunocytochemistry using HPA051153 (Magenta), and DAPI (Nuclear staining, grey) (72 hours post transfection). All immunocytochemistry images were captured using captured using structured illumination microscopy (SIM).

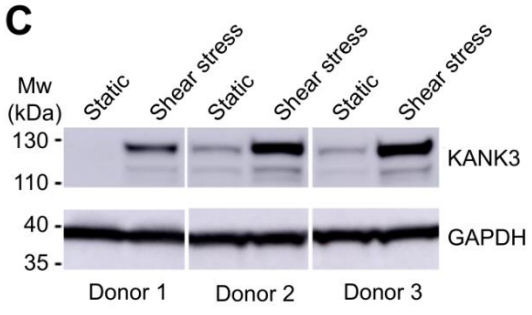
GENERATION AND VERIFICATION OF TOOLS FOR THE STUDY OF KANK3

The search term KANK3 returns 11 hits on PubMed (https://pubmed.ncbi.nlm.nih.gov/), confirming this protein is not well studied. One reason why under studied proteins are unattractive targets for functional analysis can be a lack, or unknown reliability, of research tools with which to investigate them [37, 38]. Thus, we tested our model system and generated or purchased reagents prior to study of KANK3 functional role in EC.

KANK3 splice variant expression profile in HUVEC reflects that found in vivo

Having confirmed that KANK3 is an EC enriched protein across tissue types, we went on to verify its expression profile in freshly isolated primary HUVEC, which we planned to use as an experimental model. RNA sequencing of *in vitro* cultured HUVEC (n=5) revealed that *KANK3* was reasonably highly expressed (mean 26.37 TPM ± std dev 2.25). Other KANK family members were expressed at similar levels: *KANK1* (39.5 TPM ± std dev 3.31), *KANK2* (27.3 TPM ± std dev 2.62), with the exception of *KANK4*, which was very lowly expressed (0.04 TPM ± std dev 0.03). *KANK3* splice variant ENST00000330915 was the most common isoform in HUVEC (61.6%; length: 821 aa), whilst the other variants were expressed at lower levels: ENST00000593331 (20.3%; non-protein coding), ENST00000595639 (11.9%; length: 146 aa) and ENST00000593649 (6.2%; length: 840 aa) (Figure 2A.i). Data from bulk sequencing of unfractionated human tissues in GTEx revealed that, similar to HUVEC, ENST00000330915 was the most highly expressed *KANK3* splice variant, followed by the non-protein coding variant ENST00000593331 (Figure 2 A.ii). The transcript ENST00000610351.1 in the GTEx data (Figure 1 A.ii) was retired after ENSEMBL version 104 and is not part of the current ENSEMBL gene set (V110), which our sequencing data was mapped against (hence its absence from Figure 1 A.i). It can be assumed, based on the verification of KANK3 as an endothelial enriched protein (Figure 1), that the expression site of the *KANK3* isoforms within GTEx tissues is largely EC restricted. Thus, the relative expression profile of *KANK3* variants in HUVEC reflects that found *in vivo*.


A
Static culture
B
Shear stress exposed (4 dyn)
i. Untreated
10μm
VIM
KANK3
F-ACTIN
ii. Scrambled siRNA
iii. siRNA1-KANK3
iv. siRNA2-KANK3


C
Mw (kDa)
Static
Shear stress
130
110
40
35
KANK3
GAPDH
Donor 1
Donor 2
Donor 3

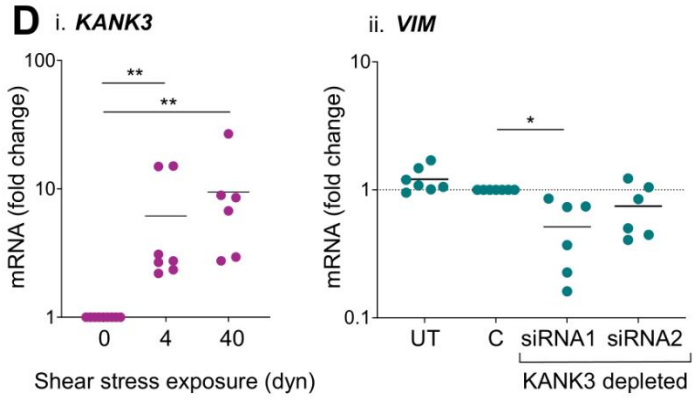

D
i. KANK3
ii. VIM
mRNA (fold change)
100
10
1
0.1
**
*
0
4
40
Shear stress exposure (dyn)
UT
C
siRNA1
siRNA2
KANK3 depleted

Figure 3: KANK3 is a shear stress regulated protein. HUVEC were cultured under (**A**) static or (**B**) shear stress exposed conditions (4 or 40 dyn/cm^2) for 48 hours after they were (i) untreated, or transfected with (ii) scrambled control siRNA, (iii) siRNA1-*KANK3* or (iv) siRNA2-*KANK3*. Immuno-cytochemistry was performed using primary antibodies targeting KANK3 (magenta), F-actin (yellow) or vimentin (VIM; cyan). (**C**) Untreated HUVEC from 3 donors were cultured under static or shear stress exposed conditions (4 dyn/cm^2) for 48 hours (´flow´) before analysis of KANK3 protein expression by Western blot, with GAPDH as a loading control. (**D**) Expression of (i) *KANK3* mRNA in HUVEC following static or shear stress exposure or (ii) *VIM* in HUVEC following transfection with or without siRNAs targeting *KANK3*. ** $p<0.001$, * $p<0.01$

Verification of KANK3 antibody specificity and siRNA knockdown efficiency

Antibody reliability can be a problematic issue and a major source of research waste; it has been suggested that up to half of all commercially available antibodies have significant issues with sensitivity and specificity [39, 40]. Furthermore, the availability and testing of such reagents targeting understudied proteins is limited [41] and thus, the validation of antibody reagents is important.

We obtained an in house generated rabbit polyclonal antibody targeting KANK3 (HPA051153) [35]. To verify its binding specificity, we took a twofold approach. Firstly, we designed a plasmid vector coding for *KANK3* with an eGFP tag, based on the sequence of the most common KANK3 isoform (ENST00000330915). This was used to express recombinant KANK3 in HEK293 cells, which do not express endogenous *KANK3*. Western Blot analysis of cell lysates with antibody HPA051153 detected bands corresponding to the size of the KANK3 protein (130kDA) (Figure 2 B.i), and staining with an anti-eGFP antibody gave similar results to HPA051153, over a range of dilutions (Figure 2B.ii).

Immunofluorescence staining of KANK3-eGFP transfected HEK293 cells with antibody HPA051153 (Figure 2 C.i) showed selective binding (pink) to cells expressing recombinant KANK3-eGFP (yellow) (Figure 1 C.i, large panel). HPA051153 did not bind KANK3-eGFP negative HEK293 cells within the transfected culture (Figure 1C.i, large panel), or to untreated HEK293 cells (Figure 2 C.ii). Secondly, to test antibody HPA051153 specificity for HUVEC KANK3, we used siRNA to deplete the protein. Two different siRNAs (siRNA1-*KANK3* and siRNA2-*KANK3*) effectively depleted HUVEC *KANK3* mRNA expression over an extended time course (fold change at 48h [mean ± std dev]: siRNA1 0.22 ± 0.006, siRNA2 0.21 ± 0.086) (Figure 2E).

Correspondingly, subsequent Western Blot analysis with HPA051153 showed that bands of a size corresponding to KANK3 protein (130kDA) were smaller, or absent, in cell lysates from siRNA transfected HUVEC (Figure 2F); an inhibition that was maintained over several days. Having verified efficient *KANK3* knockdown in HUVEC using siRNA, we performed immunofluorescence

staining using HPA051153 as a primary antibody (Figure 2G). HPA051153 showed clear punctate staining in untreated HUVEC (Figure 2 G.i), which was absent when cells were transfected with siRNA targeting *KANK3* (Figure 2 G.ii). Thus, we can have high confidence that HPA051153 selectively binds KANK3 protein in both Western Blot and immunofluorescence staining applications. Furthermore, HUVEC appear to be a suitable model system for the functional investigation of KANK3, and siRNA-mediated depletion induced a robust knockdown of KANK3 protein for several days after transfection.

KANK3 localizes within the cytoplasm and accumulates in cell-cell interaction sites

A recent study found KANK3 to be expressed at the plasma membrane of mouse EC in dermal and lymphatic vessels (S. S. Guo et al. 2021), with more diffuse staining in kidney, lung brain and oesophagus EC. Immunofluorescence staining of native KANK3 expression in endothelial cells such as HUVEC and mouse LSEC however, shows punctate distribution of KANK3 in the cytoplasm and accumulation in cell-cell interaction sites (FIGURE S2B) as well as partial colocalization to the cytoskeleton (Figure S2A).

Expression of KANK3 is enhanced under flow versus static conditions

Whilst KANK3 is poorly studied in a functional context, it shares structural homology with other members of the KANK family, which have been shown to have a role in cytoskeletal organisation, and focal adhesion formation [21, 42]. As such processes are key in the EC response to shear stress [14, 43], we investigated KANK3 expression and distribution in this context.

Untreated HUVEC, or those transfected with siRNA1-*KANK3*, siRNA2-*KANK3* or a scrambled siRNA control were cultured under static or shear stress exposed conditions (4 or 40 dyne/cm^2) for 48 hours. Cell were fixed and stained for KANK3, vimentin - the major endothelial intermediate filament (IF) that is a key regulator of focal contact size and cell-matrix adhesions in EC subjected to shear stress [44, 45] and the actin cytoskeleton which has a vital role in cell-cell adhesions [46]. Under static conditions, in both untreated and control HUVEC, KANK3 had a diffuse punctate distribution (Figure 3 A.i and ii, magenta arrows) and its expression was markedly up regulated

following shear stress exposure (Figure 3 B.i and ii, magenta arrows). Staining patterns revealed distinct areas of dense KANK3 expression. Western blotting confirmed that a significant up regulation of KANK3 protein was induced in HUVEC that had been exposed to shear stress, in 3 biological replicates (Figure 3C). Measurement of *KANK3* mRNA showed significantly elevated levels following exposure to shear stress of both 4 and 40 dyne/cm^2 (fold change vs. static ± std dev: 4 dyne/cm2 6.15 ± 5.6 p=0.004, 40 dyne/cm2 9.5 ± 8.1 p=0.002) (Figure 3D).

Under static conditions, in both untreated and control HUVEC, vimentin was located to the endogenous IF network (Figure 3 A.i and ii, turquoise arrows) with notable directional redistribution following shear stress exposure (Figure 3 B.i and ii, turquoise arrows), as previously described [47]. In KANK3 depleted HUVEC, cultured under static conditions, vimentin expression was markedly reduced compared to untreated or control HUVEC (Figure 3 A.iii and iv, turquoise arrows, Figure S2C & D). Shear stress exposure failed to induce a recovery of vimentin expression or a more typical redistribution pattern in KANK3 depleted cells (Figure 3 B.iii and iv, turquoise arrows).

In EC cultured under static conditions, *VIM* mRNA expression was lower in siRNA1-*KANK3* treated HUVEC, compared to untreated or control HUVEC, but not in siRNA2-*KANK3* treated HUVEC (fold change control EC vs. siRNA1-*KANK3* ± std dev 0.51 ± 0.27 p=0.03, siRNA2-*KANK3* 0.74 ± 0.25 p=0.10) (Figure 3 D.ii). Thus, the effects of EC *KANK3* depletion on vimentin expression/cellular distribution is unlikely to be driven by changes at the transcriptional level. Actin was diffusely expressed under static conditions, in both untreated and control HUVEC (Figure 3 A.i and ii, yellow arrows). Actin was redistributed to align with flow direction following shear stress exposure, as previously described [48] (Figure 3 B.i and ii, yellow arrows). Unlike vimentin, actin redistribution in response to shear stress was not markedly modified by KANK3 depletion (Figure 3 B.iii and iv, yellow arrows).

Thus, our results show that KANK3 levels are increased in response to EC and that KANK3 has a previously unreported direct or indirect link to vimentin distribution.

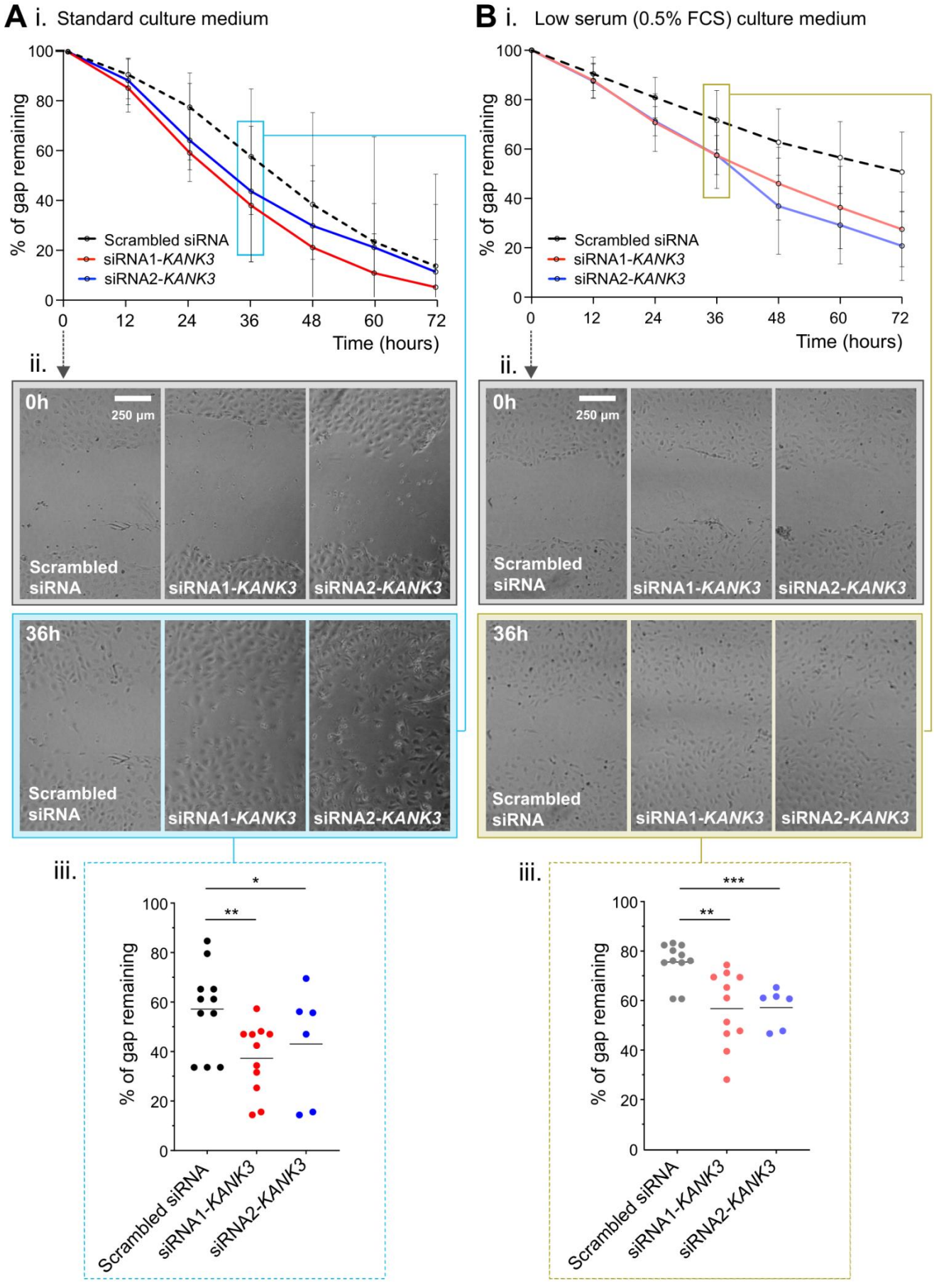

A i. Standard culture medium
B i. Low serum (0.5% FCS) culture medium
% of gap remaining
Time (hours)
Scrambled siRNA
siRNA1-KANK3
siRNA2-KANK3
ii.
0h
36h
250 µm
Scrambled siRNA
siRNA1-KANK3
siRNA2-KANK3
iii.
*
**
***

Figure 4. KANK3 depletion increases EC migration *in vitro*. HUVEC were cultured to confluence in (**A**) standard or (**B**) low serum (0.5%) culture medium, following transfection with scrambled control siRNA, siRNA1-*KANK3* or siRNA2-*KANK3*. A ´wound´ was created in the monolayer, using a pipette tip, and (i) gap closure was monitored over 72 hours. (ii) Representative phase contrast images and (iii) corresponding data points from individual experiments, from the 36-hour time point. *** $p<0.001$, ** $p<0.01$, *$p<0.05$

KANK3 has a role in endothelial cell migration

KANK3 has not been studied in the context of EC motility, but it has been reported to have a role in the regulation of cell motility in cancer cells [22, 23]. Furthermore, we observed that KANK3 depletion modified EC vimentin distribution, a protein with a key role in EC migration[49]. Therefore, we analysed the influence of KANK3 depletion in an EC gap closing assay. EC were transfected with one of 2 siRNAs targeting KANK3, or a scrambled siRNA control and cultured to confluence before a gap was created in the monolayer. Gap closure was monitored in real time, using phase contrast microscopy, and the wound area was measured at 24h, 36h, 48h and 72h. *KANK3* siRNA treated EC tended to close the gap faster than those treated with the scrambled control EC (p for trend C vs siRNA 1 p for trend 0.0142, C vs siRNA 2 p for trend: 0.0505) (Figure 4 A.i). Representative phase contrast images show the gap size at 0 and 36 hours (Figure 4 A.ii), with corresponding data points for replicate experiments (Figure 4 A.iii) (Change in gap closure normalised to control [%] ± std dev: siRNA1-*KANK3*: +30.9 ± 24.9, p=0.004, siRNA2-*KANK3*: +36.0 ± 24.6, p=0.028). Accelerated gap closure could be either due to increased migratory capacity of cells in which KANK3 has been depleted, or an increase in cell proliferation.

To assess the relative role of each, we performed the same experiment in low serum culture medium as we have previously showed that proliferation is inhibited in this condition [50]. As expected, closure of the gap was inhibited in low serum medium (Figure 2 A.ii) (% of gap remaining [control HUVEC] standard vs. low serum ± std dev: 36h 57.1 ± 16.8 vs. 75.9 ± 7.5 and 72h 10.3 ± 11.8 vs. 50.3 ± 13.7).

As observed for standard medium, *KANK3* siRNA treated EC tended to close the gap faster than control EC (p for trend C vs siRNA 1 p for trend 0.0307, C vs siRNA 2 p for trend: 0.0402) (Figure 4 B.i), but no gap closed completely within the 72-hour time frame. Representative phase contrast images of the gap size at 0 and 36 hours (Figure 4 B.ii) and corresponding data points for replicate experiments (Figure 4B A.iii) (Change in gap closure normalised to control [%] ± std dev: siRNA1-*KANK3*: 23.3 ± 21.0, p=0.018, siRNA2-*KANK3*: +23.5 ± 6.0, p=0.0004), showed similar results as

those obtained in standard culture medium. Furthermore, measurement of PCNA mRNA revealed no difference in expression between control of KANK3 depleted cells (fold change control EC vs. siRNA1-*KANK3* ± std dev 0.71 ± 0.11, siRNA2-*KANK3* 1.08 ± 0.07). Therefore, the increased rate of gap closing observed in KANK3 depleted HUVEC appears to be primarily driven by an increased migratory capacity, as opposed to an increased rate of proliferation.

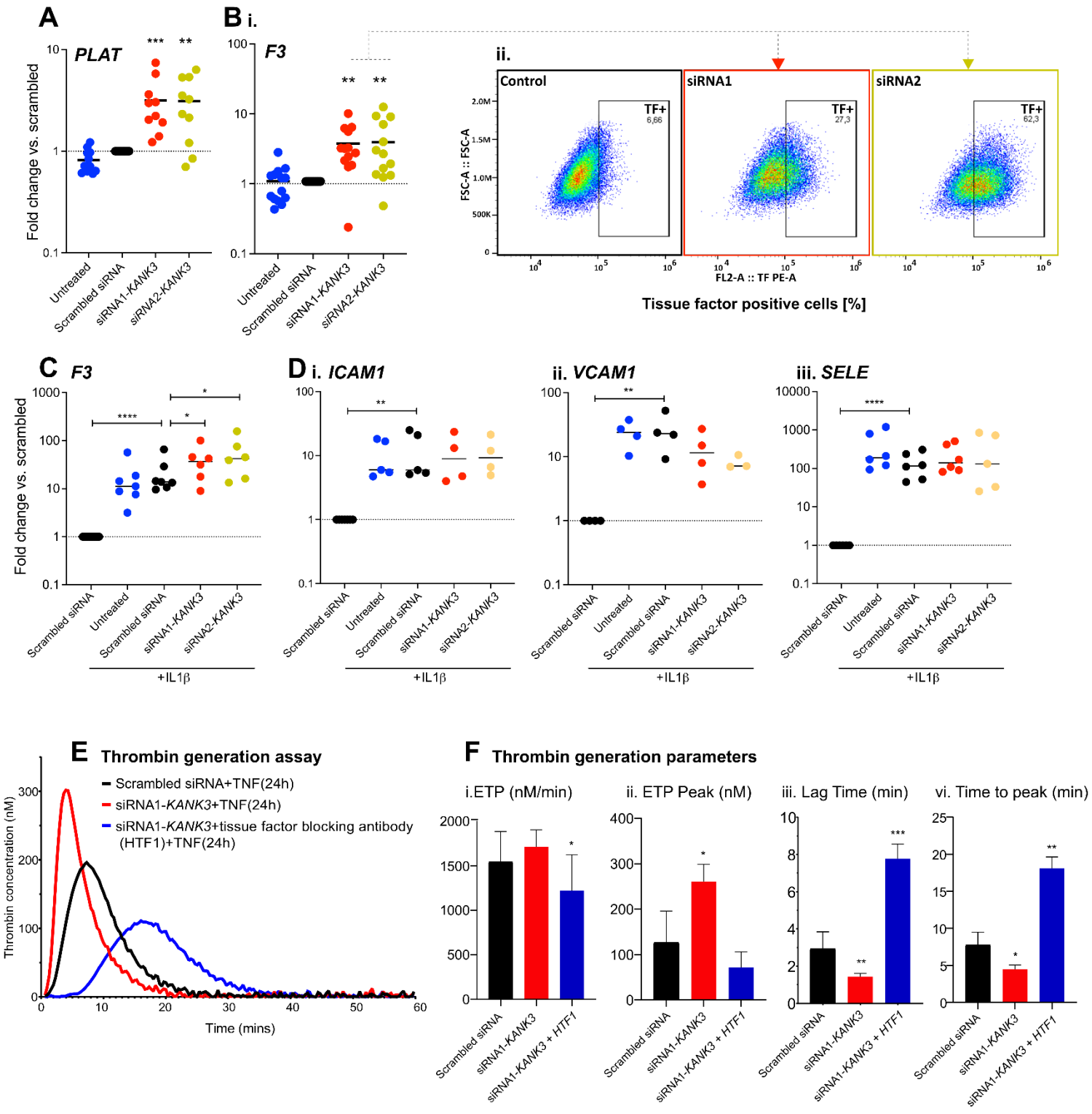

A
PLAT
Fold change vs. scrambled
Untreated
Scrambled siRNA
siRNA1-KANK3
siRNA2-KANK3
B i.
F3
ii.
Control
siRNA1
siRNA2
TF+
6,66
27,3
62,3
FSC-A :: FSC-A
FL2-A :: TF PE-A
Tissue factor positive cells [%]
C F3
D i. ICAM1
ii. VCAM1
iii. SELE
+IL1β
E Thrombin generation assay
Scrambled siRNA+TNF(24h)
siRNA1-KANK3+TNF(24h)
siRNA1-KANK3+tissue factor blocking antibody (HTF1)+TNF(24h)
Thrombin concentration (nM)
Time (mins)
F Thrombin generation parameters
i.ETP (nM/min)
ii. ETP Peak (nM)
iii. Lag Time (min)
vi. Time to peak (min)
siRNA1-KANK3 + HTF1

Figure 5: Effects of endothelial KANK3 knockdown on coagulation related proteins. HUVEC were untreated, or transfected with scrambled control siRNA, siRNA1-*KANK3* or siRNA2-*KANK3*. Measurement of: (**A**) *PLAT* or (**B**) (i) *F3* mRNA expression by qPCR, or (ii) cell surface tissue factor protein by flow cytometry. HUVEC were treated with or without IL1β (10 ng/ml) before measurement of mRNA encoding for (**C**) F3, or (**D**) the cytokine responsive adhesion molecules (i) *ICAM1*, (ii) *VCAM1*, (iii) *SELE*. (**E**) Calibrated automated thrombogram (CAT) assay was used to assess the thrombin generation potential of HUVEC treated with TNF (10 ng/ml) for 24 hours with or without pre-incubation with a function blocking anti-tissue factor antibody (HTF1). (F) Bar plots show the (i) total endogenous thrombin potential (ii) maximum endogenous thrombin potential (iii) lag time until the beginning of thrombin production (iv) time until peak thrombin production. *** $p<0.001$, ** $p<0.01$, *$p<0.05$ vs. scrambled control.

KANK3 effects tissue factor and tissue type plasminogen activator expression

As other members of the KANK family have potential functions beyond cytoskeletal regulation, e.g. KANK4 can regulate VEGFR2 signalling via its interaction with talin [51], we screened untreated, scrambled siRNA transfected and KANK3 depleted EC for expression profiles of the following gene panels, which are related to EC specialised functions: (i) *Coagulation related*: Factor 8 (*F8*), protein C receptor (*PROCR*), protein S (*PROS*), tissue factor (*F3*), tissue factor pathway inhibitor (*TFPI*), tissue plasminogen activator (*PLAT*) and von Willebrand Factor (*VWF*) (ii) *Inflammation related*: intracellular adhesion molecule 1 (*ICAM1*), vascular cell adhesion molecule 1 (*VCAM1*), E-selectin (*SELE*) and (iii) *angiogenesis related*: angiopoietin-1 receptor (TEK), angiopoietin-2 (ANGPT2), vascular endothelial growth factor A (VEGFA) and vascular endothelial growth factor receptor 1 (FLT1), and kinase insert domain receptor (KDR, also known as vascular endothelial growth factor receptor 2).

Two genes in the coagulation related panel, *PLAT* and *F3,* were expressed at higher levels in both siRNA1-*KANK3* and siRNA2-*KANK3* treated EC, compared to the scrambled control (mean fold change ± std dev *F3:* 4.1 ± 2.8, *PLAT*: 3.1 ± 1.9, both $p<0.05$) (Figure 5 A and B.i). No other genes tested were consistently elevated, or reduced, in KANK3 depleted EC (Figure S3). As the *F3* gene encodes for the protein tissue factor, which is the key initiator of the extrinsic coagulation cascade, we investigated its relationship with KANK3 further. In line with the changes observed at the transcript level, flow cytometry confirmed an increase in the cell surface expression of tissue factor (TF) protein on KANK3 depleted HUVEC, compared to the scrambled control (Figure 5 B.ii) (scrambled control: MFI 29489; 6.66% cells TF positive, siRNA1-*KANK3*: MFI 47548; 27.3% cells TF positive, siRNA2-*KANK3:* MFI 93347; 62.3% cells TF positive).

F3 expression is relatively low on resting EC, but strong induced by inflammatory cytokines [52], so we tested if KANK3 depletion would modify this response. HUVEC transfected with scrambled siRNA, siRNA1-*KANK3*, or siRNA2-*KANK3* were treated with the inflammatory cytokine IL-1β for

24h. IL-1β strongly induced *F3* mRNA expression in scrambled siRNA HUVEC (scrambled siRNA vs. scrambled siRNA +IL-1β [mean fold change ± std dev] 22.4 ± 18.7), an increase that was exacerbated further by KANK3 depletion (scrambled siRNA +IL-1β vs. siRNA1-*KANK3*+IL-1β [mean fold change ± std dev]: 2.1 ± 1.0, siRNA2-*KANK3*+IL-1β: 3.2 ± 1.6) (Figure 5C).

To test if this was a consequence of a general enhancement of IL-1B signalling, we measured the expression of EC adhesion molecules. As expected, all were induced by IL-1β treatment (Figure 5 D i-iii), but we did not observe any further increase in expression in KANK3 depleted EC (Figure 5 D i-iii). Thus, KANK3 depletion does not appear to enhance cytokine signalling *per se*.

F3 is also induced by the inflammatory cytokine TNF, which has many signalling pathways in common with IL-1β [53]. To test if there was a functional consequence of this enhanced expression of *F3* in KANK depleted EC, thrombin generation potential was measured using the calibrated automated thrombograph (CAT) (Figure 5E). Relative to scrambled siRNA control, the depletion of KANK3 resulted in enhanced thrombin generation, as shown by a representative curve (Figure 5E). Whilst no statistically significant difference in endogenous thrombin potential (ETP) was observed (Figure 5 D.i), response time (scrambled siRNA vs. siRNA1-*KANK3* [difference [min;%], ± std dev]: -1.5 min, -52.2 %), time to peak (scrambled siRNA vs. siRNA1-*KANK3* [difference [min;%], ± std dev]: -1.5 min, -52.2 % -3.3min; -42.2 %), and peak thrombin generation levels (scrambled siRNA vs. siRNA1-*KANK3* [difference [levels; %], ± std dev]: +134.04; +105.1%) were all enhanced in KANK3 depleted EC (Figure 5 F.i-iv). Similar results were observed with siRNA2-*KANK3* (Figure S3 B and C). When KANK3 depleted EC were pre-treated with tissue factor function blocking antibodies, these effects were largely abolished (Figure 5 E and F ii-vi), consistent with tissue factor being the driver behind the increased thrombin generation.

DISCUSSION

Previously, we generated the first prediction that *KANK3* was an EC enriched protein in the human [8]. Here we confirm that KANK3 has a body wide EC enrichment at both the transcript and protein level. Recent work in mice also showed KANK3 expression exclusively in the vasculature, whilst KANK1 was expressed at the basal side of epithelial cells in various tissues, and KANK2 was observed predominantly at the plasma membrane and/or in the cytoplasm of mesenchymal cells [54]. Thus, the strictly restricted cell type expression profile of KANK3 is unique among the KANK family, the other members of which are also expressed more broadly in the human [26]. The EC type specific profile of KANK3 could underlie the lack of studies on its function, as indeed a several studies regarding KANK3 primarily center on cancer and overlook EC types in experimental design [9, 22, 23, 55], presumably due to the fact that EC are a minority cell type within any given tissue [8]. In the absence of understanding the likely context (cell type) in which a protein functions, it can be challenging to functionally characterise it. Understudied proteins can be unattractive targets for functional analysis due to a lack, or unknown reliability, of research tools with which to investigate them [37, 38]. Here, we validated our in house generated anti-KANK3 antibody, using both over expression and knockdown systems, as antibody specificity and reliability can be a problematic issue [39, 40, 56]. A study from 2008 showed that this problem is systemic. Of 6000 tested antibodies, fewer than 3000 were able to bind their target correctly. Due to the increased use of commercial antibodies, without validation, this might mean that an entire project could be based on artifacts. Due to the widespread utilisation of research antibodies, this is potentially a billion-dollar problem, with approximately 1.7 billion USD that have been wasted to antibodies in 2019. Hence, validation of any antibody used in a project is an essential step in research [57, 58]. The KANK family have a unique shared structure, consisting of a small N-terminal motif (“KN-motif”), C-terminal coiled-coil domains and ankyrin repeats (Zhu et al. 2008; Kakinuma et al. 2009).

The protein structure of KANK3 would indicate a function within cytoskeletal organization and in focal adhesions, due to the presence of a talin binding domain. It consists of a liprin binding domain and a KIF21A binding domain, which links them to the cytoskeleton [11, 59]. Immunofluorescence staining of endogenous KANK3 expression in EC, including HUVEC and mouse LSEC, reveals a distinctive punctate pattern within the cytoplasm. Notably, there is an accumulation of KANK3 at sites of cell-cell interactions in murine LSEC. This supports the role of KANK3 being linked to cell adhesions and to the basement membrane, which are important for integrity of the endothelial layer under stressful conditions such as flow or inflammation [60, 61] Crosstalk between focal adhesion proteins and cell junction proteins has been previously described in the regulation of endothelial barrier function [62]. Considering the critical role of EC barrier function and the established involvement of the cytoskeleton and focal adhesions in the maintenance of cellular junctions [16], it could be speculated that KANK3 has a role in anchoring EC to the extracellular matrix and neighbouring cells. Furthermore, KANK3 might play a role in cell junction processes in some vascular beds.

EC shear forces have been shown to modulate various biochemical processes, in addition to cellular morphology and reorganisation of the cytoskeleton [63]. Shear stress has previously been shown to induce focal assembly, recruit signalling complexes to FA, and induce redistributions in stress fibres in EC [14]. We observe a significant upregulation of KANK3 gene and protein expression following EC exposure to shear stress. Similar mechanoregulation has been previously reported for different focal adhesion proteins through vinculin-vinculin regulation [64]. Laminar shear stress has been shown to induce integrin expression [65], α-actinin recruitment [66] and to mediate redistribution of intracellular stress fibers. Consequently, under the influence of this shear stress, there is a force-dependent alteration in the dynamics of FAs due to enhancement in actin fibers. The enhancement in actin fibers results in the sustenance or growth of the FAs connected to them [67]. FA have been shown to be involved in matrix-adhesion and mechanosensing in EC

[14], and vascular resistance in SMC [68]. Due to its vascular specificity, one could speculate, that KANK3 is involved in similar vascular functions.

We also observed that KANK3 depletion had a marked effect on the distribution of vimentin. This did not appear to be driven by a modification in VIM transcription, and could therefore be driven by a direct or indirect KANK3-vimentin protein-protein interaction, such as is observed with other cytoskeletal components, e.g., actin and myosin [69] or intermediate filaments and microtubules [70]. Vimentin is a major intermediate filament in EC [44], which has roles in cell migration and polarity, cell structure and integrity, response to mechanical stress and EC differentiation [71-74] and has been described as integral to cell adhesion and EC sprouting [44]. Although there is a lack of research regarding the interactions between KANK proteins and vimentin, it's worth noting that vimentin plays a role in cell migration. Previous research has indicated its involvement in determining cellular polarity, regulating the formation of cell contacts, and organizing and transporting signalling proteins that contribute to cell motility [75]. Vimentin increases cell stiffness and promotes cell migration when cells are densely cultured. However, its impact on the migration of cells plated sparsely is minimal or negligible [76].

Here we show that KANK3 depletion increases EC migratory capacity. Previously, talin has been identified as a regulator of cell motility [77] and the stability of talin rods has been shown to control cell migration [78]. It could be reasonably assumed that KANK3 migratory control is driven through KANK3-talin interactions. Although to our knowledge, KANK3 has not been studied in the context of EC motility, previous reports showed it had a role in the regulation of oxygen dependent suppression of cell motility in hepatocellular carcinoma cells [22], and the inhibition of invasion and migration of lung adenocarcinoma [23] and was therefore considered a valuable target in cancer research. The increase of cell motility in KANK3 depleted cells could suggest a protective effect against shear forces and increase in matrix adhesion, similar to the contribution of vimentin networks to the stiffening of cells, which allows them to withstand mechanical forces [79].

Focal adhesions play a pivotal role in relaying mechanical forces and external signals to intracellular pathways. Disturbed multidirectional shear stress of the vasculature is recognised for its role in triggering the activation of atherogenic and thrombogenic genes in EC and SMC [80]. Hence, it is plausible that KANK3 mediated modifications in the cytoskeleton might initiate signalling pathways linked to the regulation of genes associated with coagulation. FA activation in vascular smooth muscle cells regulates arterial stiffness and procoagulant properties of the vessel wall. Their findings revealed a decrease in thrombin generation potential of vascular smooth muscle cells (VSMCs) as the matrix stiffness increases On a rigid matrix, the presence of αvb3 integrin within the FA complex diminishes the accessibility of binding sites for prothrombin. As a consequence, this leads to a reduction in the generation of thrombin on VSMCs. Conversely, it could be hypothesized that this outcome is reversed when dealing with a less rigid matrix [81]. This connection underscores the intricate interplay between FA signalling, vascular mechanics, and thrombotic potential.

In summary, our study provides insight into the function of KANK3 in the vascular compartment. Our findings are consistent with it having EC specific functions, in line with its enriched expression profile in this cell type.

METHODS

Tissue profiling

Protein profiling was performed as part of the Human Protein Atlas (HPA) project. Tissue sections from breast, adipose tissue, cortex, thyroid gland, colon, kidney, liver, prostate, epididymis, duodenum, bronchus, testis, endometrium, cervix, appendix, stomach, oesophagus, lung, pancreas and ovary were generated and stained, as previously described (Pontén, Jirström, and Uhlen 2008; Uhlen et al. 2015). Briefly, formalin fixed, and paraffin embedded tissue samples were sectioned, de-paraffinized in xylene, hydrated in graded alcohols and blocked for endogenous peroxidase in 0.3% hydrogen peroxide diluted in 95% ethanol. For antigen retrieval, a Decloaking chamber® (Biocare Medical, CA) was used. Slides were boiled in Citrate buffer®, pH6 (Lab Vision, CA). Primary antibody against KANK3 (HPA051153) and a dextran polymer visualization system (UltraVision LP HRP polymer®, Lab Vision) were incubated for 30 min each at room temperature and slides were developed for 10 min using Diaminobenzidine (Lab Vision) as the chromogen. Slides were counterstained in Mayers haematoxylin (Histolab) and scanned using Scanscope XT (Aperio).

Isolation and culture of human umbilical vein endothelial cells

Ethical approval for endothelial cell isolation and subsequent experimentation was granted by *Regionala etikprövningsnämnden i Stockholm* (diarienummer 2015/1294-31/2). Human umbilical vein endothelial cells (HUVEC) were isolated from human umbilical cords, collected from Karolinska Hospital (Stockholm, Sweden) and from the University Hospital of Northern Norway (UNN; Tromsø, Norway), as previously described (Cooke et al. 1993). HUVEC were cultured in Medium M199, supplemented with 10% fetal bovine serum (FBS) (or 0.5% FBS in some experiments), 10ml/l Penicillin-Streptomycin, 2.5mg/l Amphotericin B (all ThermoFisher, Gibco), 1mg/l Hydrocortisone 1µg/l and human Epidermal Growth Factor (hEGF) (both Merck). In some

experiments, EC were cultured under laminar shear stress (4dyn or 40dyn) for 48 hours in flow chamber slides (μ-slide VI 0.4, Ibidi), integrated into an Ibidi flow pump system.

HEK293 cells were obtained in frozen vials from ATCC (HEK-293 CRL-1573) and cultured in DMEM Cell culture medium supplemented with 10ml/l Penicillin-Streptomycin and 10% fetal bovine serum (FBS).

Mouse liver sinusoidal endothelial cells (mLSEC) were gifted from *Vascular Biology Research Group (VBRG) at UiT The Arctic University of Norway*, isolated and cultured in RPMI 1640 supplemented with L-Glutamine (300mg/l), 10ml/l Penicillin-Streptomycin, as previously described [82].

siRNA transfection

HUVECs and HEK239 cells were transfected with siRNA sequences targeting KANK3 (silencer select siRNA s230059, s230061, ThermoFisher) or Silencer Select negative control siRNA (ThermoFisher: 4390843). Transfection was performed using Lipofectamine RNAiMAX transfection reagent (Invitrogen), according to manufacturer instructions, at 60-80% confluency in Opti-MEM reduced serum medium (ThermoFisher, Gibco) without additives for 4h. Medium was changed to standard cell culture medium. Knockdown efficiency was accessed after 48h by qPCR, Western blot, or immunofluorescence staining.

Recombinant KANK3-eGFP protein expression

Transient transfection of vector coding for KANK3_eGFP (GenScript) into HEK293 (ATCC: CRL-3216) cells was done using Lipofectamine 3000 (ThermofisherScientific), according to the manufacturer's instructions. Plasmid transfection was performed at cell confluency of 60-80% in Opti-MEM reduced serum medium without additives for 5h using Lipofectamine 3000 transfection reagent (Invitrogen), according to manufacturer instructions. 48 h after transfection, transfected cells were lysed with RIPA cell Lysis Buffer and sample was frozen, or cells were fixed for further analyses.

RNA sequencing

RNA isolation and purification was performed using the RNAeasy mini kit (Qiagen). RNA concentration was measured using Nanodrop 2000 spectrophotometer and RNA integrity number (RIN) determined using Agilent 2100 Bioanalyzer (RIN>9 required for inclusion). Library preparation and RNA sequencing was performed by the National Genomics Infrastructure Sweden (NGI) using Illumina stranded TruSeq poly-A selection kit and Illumina NovaSeq6000S (4 lanes, 2x 150bp reads, incl 2Xp kits). The data was processed using demultiplexing. Data storage and initial analyses were performed using server sided computation supplied by the Swedish National Infrastructure for Computing (SNIC). Genome assembly used for sequence alignment: Homo_sapiens.GRCh38.dna.primary_assembly.fa and annotation performed using: Homo_sapiens.GRCh38.96.gtf. Sequence alignment was carried out using STAR/2.5.3a. Gene mapping has been carried out using subread/1.5.2 and the module feature counts. Transcript mapping carried out using Salmon/0.9.1.

Gap closing ("scratch") assay

A ´gap´ was created in a confluence EC monolayer using a 100 µl pipette tip. Gap size was monitored with an Olympus IXplore Live microscope in phase/contrast mode in 10x magnification, with cells in a 37°C, 5% CO_2 on stage incubator chamber. Gaps were imaged every 30min for 96h. Gap size was measured every 6h in Fiji using ImageJ2 graphics procession software.

Shear stress exposure

Endothelial cells were cultured in flow chamber slides (µ-slide VI 0.4, Ibidi) until confluence. The slide was connected to an Ibidi flow pump system and cultured under laminar shear stress (4dyn or 40dyn) for 48 hours. The cells were then lysed for qPCR, Western blot, and fixed for confocal microscopy.

qPCR

Cell lysis and cDNA creation were performed using the 2-Step Fast-Cells-to-CT-Kit (Invitrogen, ThermoFisher) according to their protocols. qPCR was performed using TaqMan Fast Universal

PCR mix. Target primer conjugated to FAM-probe (4448892, ThermoFisher) was used to access KANK3 levels. 18s rRNA (4319413E conjugated to VIC probe, ThermoFisher) was used as endogenous control. qPCR was performed using a RealTime PCR LightCycler 96 ® system (Roche Life Sciences).

SDS-PAGE and Western blot

Lysate was analysed for KANK3 expression by western blotting with rabbit polyclonal anti-KANK3 antibody (1:250, HPA051153) and horseradish peroxidase (HRP)-coupled goat anti-rabbit antibody (1:2000, Dako). After chemiluminescence detection, membrane was washed, incubated in stripping buffer, and analysed for GAPDH housekeeping gene expression. Additionally, mouse monoclonal anti-eGFP antibody (1:1000) with secondary HRP-coupled goat anti-mouse antibody (1:2000, Dako), were used.

Flow cytometry

Endothelial cells were cultured in 6 well plates and transfected with 2 different siRNAs targeting KANK3 or a scrambled siRNA control 48h prior 2 harvesting cells. 4h prior to harvest cells were stimulated with 10ng/ml TNF. Cells were harvested by trypsin digest followed by centrifugation (x350g, 7min) and separation, decantation of supernatant and resuspension of EC in ice-cold PBS (Gibco, ThermoFisher). Cells were split into two tubes and treated with PE-conjugated anti-CD142 Clone NY2 (30 µl/ml) and isotype-matched control mouse-IgG1 (6 µl/ml) and incubated on ice for 30min, followed by centrifugation (x350g, 7min), decantation of supernatant and resuspension of cells in PBS. Flow cytometry was performed in Beckman Coulter CytoFLEX Flow Cytometer (acquisition settings FSC 20V, SSC 150V, PE 130V). Gating and data analysis was performed using CytExpert for CytoFLEX Acquisition and Analysis Software and FlowJo™ v10.7. Gating was performed for live vs dead cells and singlets vs doublets. Dead cells and doublets were excluded, followed by gating for TF positive and TF negative cells. Isotype control signal was subtracted from full stain for each sample and median fluorescence intensity (MFI) and TF positive cells (%) were identified for each condition.

Confocal microscopy

Cells were fixed in 4 % paraformaldehyde in PBS, permeabilised in 0.5% triton X-100 and blocked using 5% BSA. Primary antibodies against KANK3 (HPA051153, Atlas antibodies) and vimentin (OMA1-06001, Invitrogen) were incubated on cells for 20 minutes, followed by FITC-conjugated anti-rabbit antibody (F-9887, Sigma), Alexa 555-conjugated anti mouse IgG, and either TRITC-conjugated (P1951, Sigma) or Atto-647N-conjugated (65906, Sigma) phalloidin, depending on experiment, and were then coated in mounting medium (VectaShield) containing DAPI nuclear stain for storage and imaging. Images were taken using a Leica TP5 SP5 confocal microscope and image analysis was performed in Fiji ImageJ2 graphics procession software.

Structured Illumination Microscopy and Deconvolution imaging

48h after siRNA or plasmid transfection, cells were plated on fibronectin coated (1 µg/cm^2) #1.5 glass coverslips (Zeiss) subconfluently (30-50k cells per cm^2) and cultivated for 1h (HEK293 cells) or 4h (HUVECs). Afterwards, cells were fixed in 4% paraformaldehyde (Merck, Sigma) in PBS (Dulbecco, Sigma) for 20 min, washed with PBS (Dulbecco, Sigma) and left in PBS until further analysis. Samples were permeabilised in 0.05% Triton X-100 (Sigma) in PBS and blocked in 3% BSA in PBS. Primary antibody from Rabbit against KANK3 (HPA051153, Atlas antibodies) was prepared in blocking buffer, followed by incubation with phalloidin conjugated to atto-647, secondary anti rabbit IgG conjugated to Alexa 555 and anti-mouse IgG conjugated to Alexa 488 for 30 minutes at RT, nuclear stain was performed with DAPI for 20 minutes at RT in the dark. Samples were mounted using hardset antifade mounting medium (VectaShield). Images were taken in an OMX Blaze SIM microscope using a 60X 1.42NA oil-immersion objective (GE Healthcare; Olympus). 3D-SIM images stacks of up to 3 µm were acquired every 125 nm in five phases and three angles, resulting in 15 raw images per z-plane and total of 24 focal planes. Reconstruction used SoftWoRx software (GE Healthcare). Image analysis was performed in Fiji ImageJ2 graphics procession software.

Calibrated automated thrombinoscope (CAT) assay

HUVECs were cultivated until confluency in medium M199 and, if stated, knocked down as described above. Cells were transferred to flat-bottom 96 well plates (VWR) treated with tumour necrosis factor alpha (TNF; 10 ng/mL) (ThermoFisher) for 24h before thrombin generation assay. After washing the cells with PBS, thrombin formation was initiated in 120 µL reaction mixtures containing human citrated plasma, 4 µM phospholipids (Thrombinoscope BV), 16.6 mM Ca2 + and 2.5 mM fluorogenic substrate (Z-Gly-Gly-Arg-AMC, Thrombinoscope BV). As controls, Tissue factor (1 pM, Dade Innovin), mouse monoclonal anti-TF antibody (12.5 µg/ml, HTF-1, BD Pharmingen) or corn trypsin inhibitor were added 15 min before adding the substrate. All real time thrombin formation experiments were run in triplicates. Thrombin generation was quantified using the Thrombinoscope software package (Version 5.0.0.742) that reported means ± SD.

Gene ontology analysis

The Gene Ontology Consortium [33] and PANTHER classification resource (Mi, 2019) were used to identify overrepresented terms in gene lists using the GO databases (release date 2023-07-05). Plots of GO terms were created using the R package clusterProfiler [83].

Data usage and analysis

Human tissue RNAseq data was retrieved from Genotype-Tissue Expression (GTEx) portal V8 (www.gtexportal.org) [84]. Statistical analyses were done in RStudio (R V 4.0.3), using the corr.test function from the additional package psych (V 2.0.12) (Pearson correlation coefficient). Single Cell sequencing data was sourced from data collected into the Tabula Sapiens [85].

Software: Image analyses

Image analysis was performed using ImageJ2 Fiji using ImageJ2 graphics procession software [86].

Software: Graphs, Figures and Tables

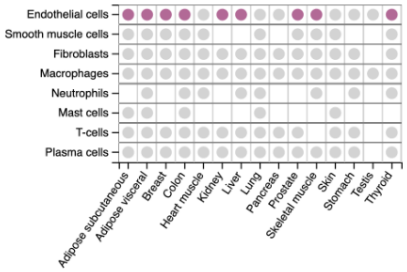
Graphs and calculation tables were created using GraphPad Prism (V. 8.4.3) and Microsoft Excel 2019 (Office 365). Figures were created in Photoshop.

Software: Statistical analyses

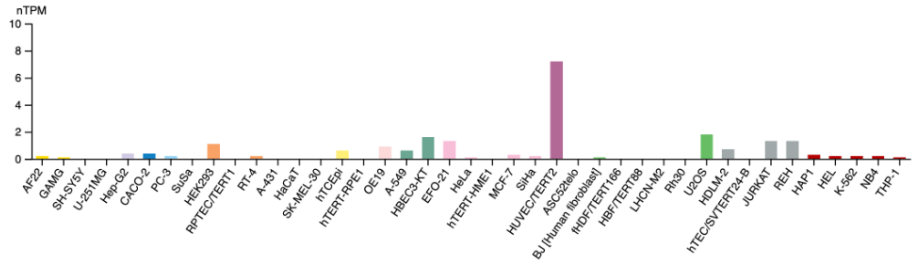
Statistical analyses were performed in RStudio (R version 4.0.3) using the following additional packages: psych, readr dplyr, data.table and tools.

A *KANK3*

i. Cell type enrichment

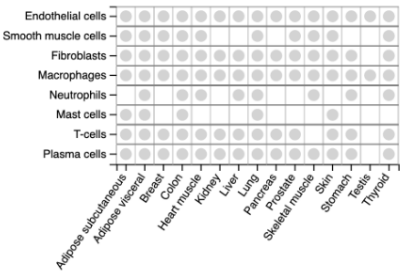


ii. Cell line expression

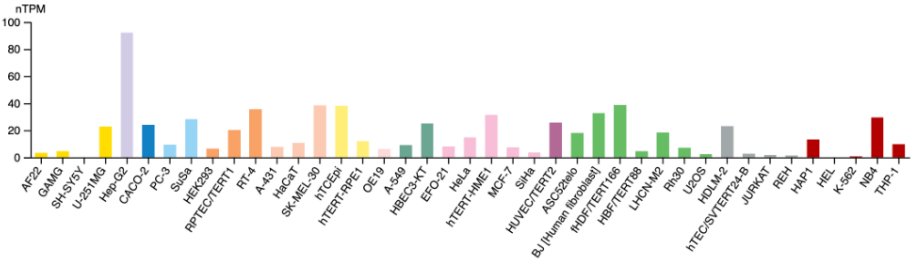


B *KANK1*

i. Cell type enrichment

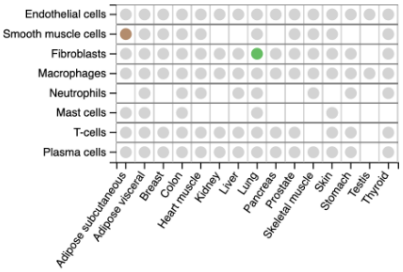


ii. Cell line expression

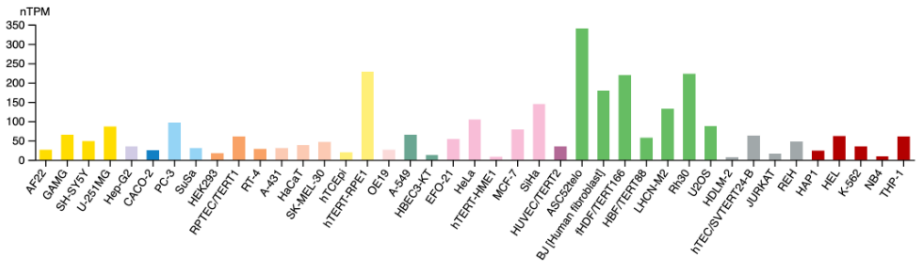


C *KANK2*

i. Cell type enrichment

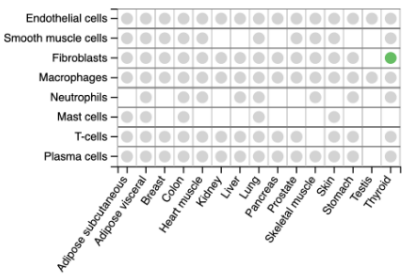


ii. Cell line expression



D *KANK4*

i. Cell type enrichment



ii. Cell line expression

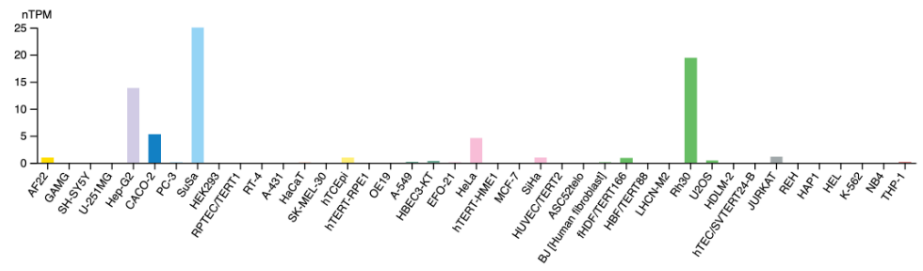


Figure S1. Enrichment of KANK family members in human tissues and immortalized cell lines. Expression profiles for (**A**) *KANK3*, (**B**) *KANK1*, (**C**) *KANK2* and (**D**) *KANK4* in: (i) human cell types profiled using bioinformatic based analysis of bulk RNAseq data [26], or in (ii) immortalised cell lines [35].

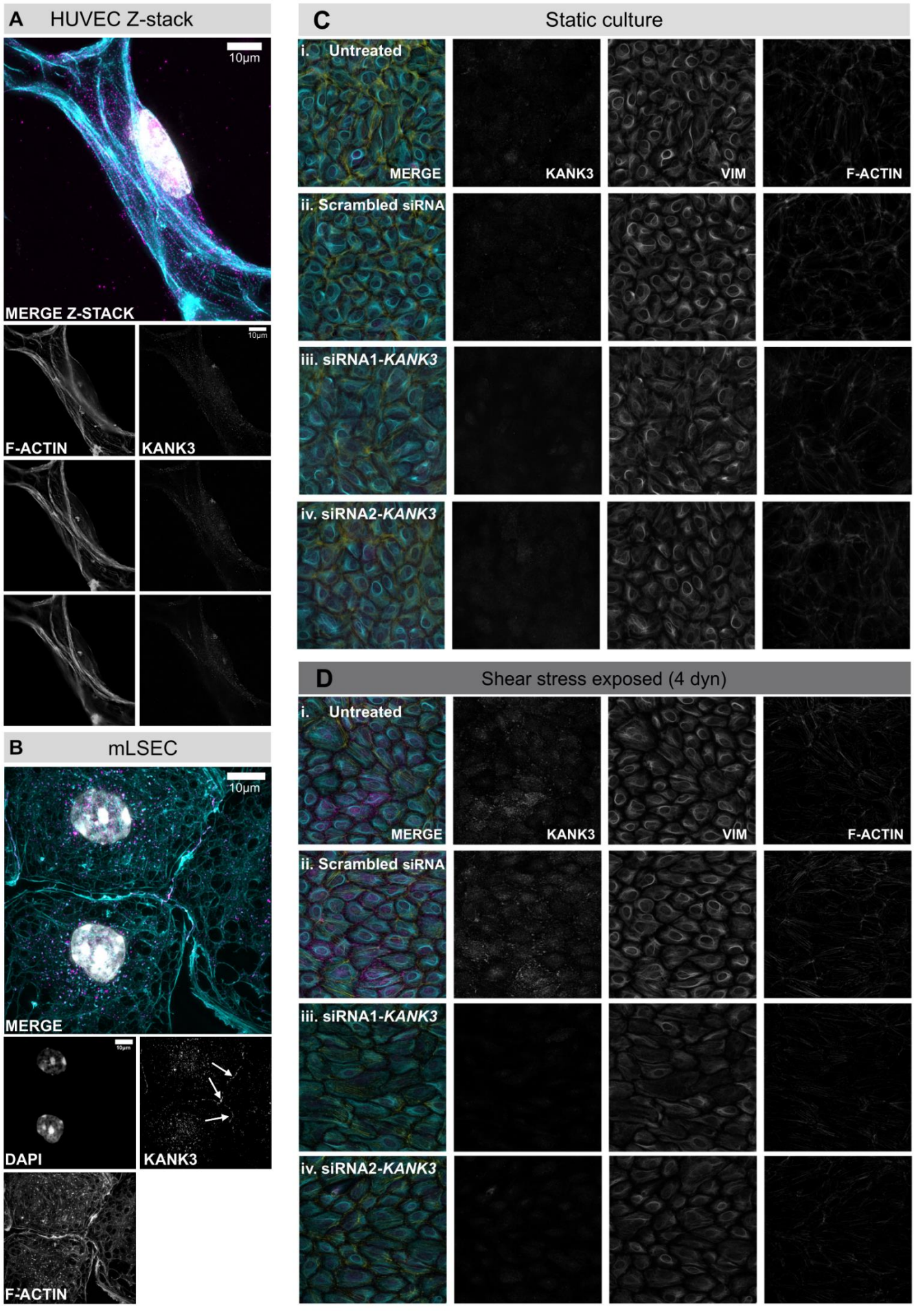

A HUVEC Z-stack
10μm
MERGE Z-STACK
F-ACTIN
KANK3
B mLSEC
10μm
MERGE
DAPI
KANK3
F-ACTIN
C Static culture
i. Untreated
MERGE
KANK3
VIM
F-ACTIN
ii. Scrambled siRNA
iii. siRNA1-KANK3
iv. siRNA2-KANK3
D Shear stress exposed (4 dyn)
i. Untreated
MERGE
KANK3
VIM
F-ACTIN
ii. Scrambled siRNA
iii. siRNA1-KANK3
iv. siRNA2-KANK3

Figure S2. Subcellular location of KANK3 in static and flow culture: Immunocytochemistry staining using anti-KANK3 antibody (magenta), DAPI nuclear stain (grey), and phalloidin F-actin stain (cyan) for (A) HUVEC Z stack average (B) mLSEC. (C,D) Immunocytochemistry staining using anti-KANK3 antibody (magenta), anti-vimentin antibody (yellow), DAPI nuclear stain (grey), and phalloidin F-actin stain (cyan) in (C) static cultured and (D) shear stress exposed HUVEC.

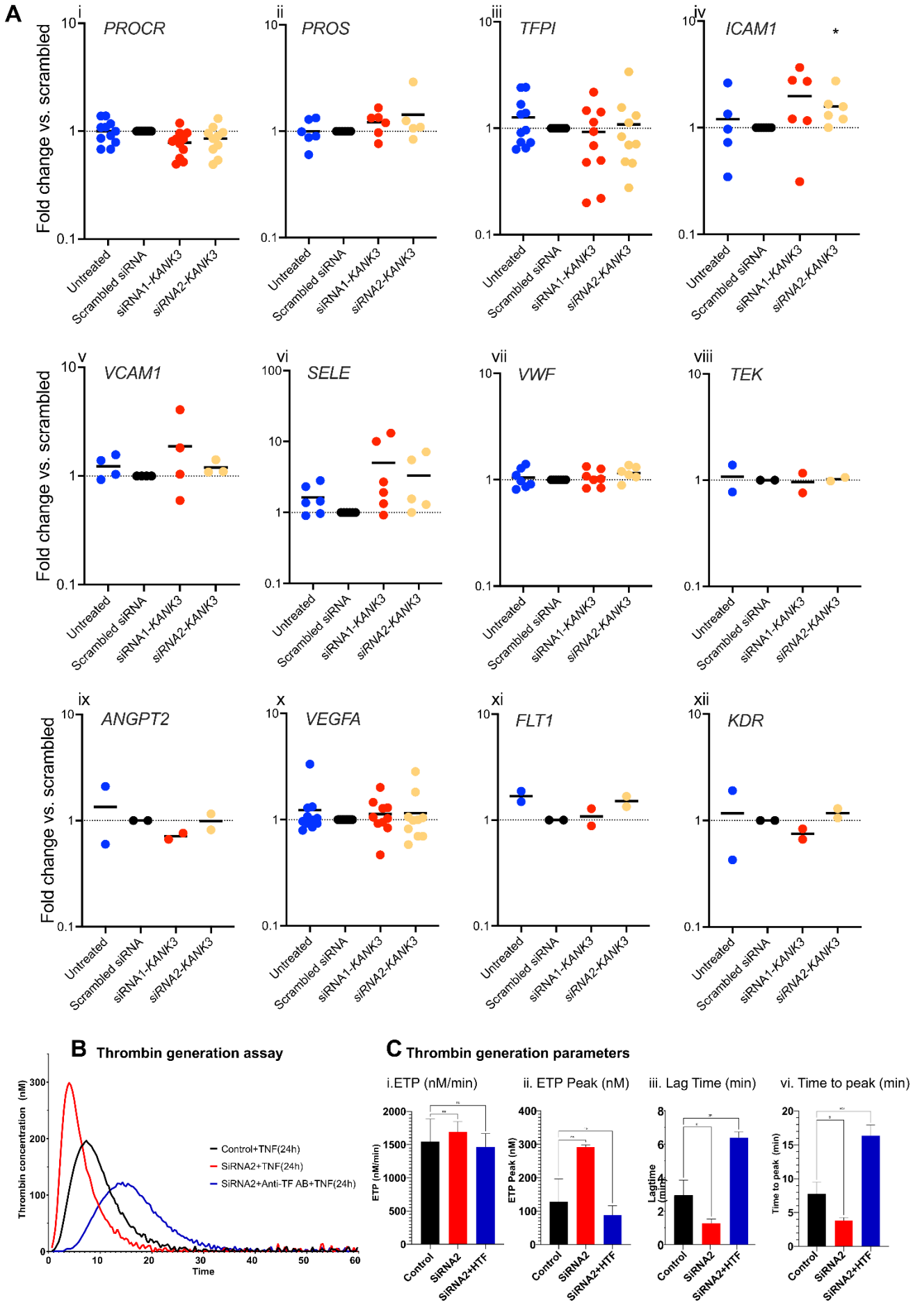

A
i
PROCR
ii
PROS
iii
TFPI
iv
ICAM1
*
v
VCAM1
vi
SELE
vii
VWF
viii
TEK
ix
ANGPT2
x
VEGFA
xi
FLT1
xii
KDR
Fold change vs. scrambled
Untreated
Scrambled siRNA
siRNA1-KANK3
siRNA2-KANK3
B Thrombin generation assay
Thrombin concentration (nM)
Time
Control+TNF(24h)
SiRNA2+TNF(24h)
SiRNA2+Anti-TF AB+TNF(24h)
C Thrombin generation parameters
i.ETP (nM/min)
ii. ETP Peak (nM)
iii. Lag Time (min)
vi. Time to peak (min)
Control
SiRNA2
SiRNA2+HTF

Figure S3: Effect of KANK3 depletion on gene expression. (A) HUVEC were untreated, or transfected with scrambled control siRNA, siRNA1-*KANK3* or siRNA2-*KANK3* before measurement of mRNA level of genes indicated by qPCR. *p<0.05 vs. scrambled control. (**B**) Calibrated automated thrombogram (CAT) assay was used to assess the thrombin generation potential of HUVEC treated with TNF (10 ng/ml) for 24 hours with or without pre-incubation with a function blocking anti-tissue factor antibody (HTF1). (C) Bar plots show the (i) total endogenous thrombin potential (ii) maximum endogenous thrombin potential (iii) lag time until the beginning of thrombin production (iv) time until peak thrombin production. *** p<0.001, ** p<0.01, *p<0.05 vs. scrambled control.

REFERENCES

1. Galley, H.F. and N.R. Webster, *Physiology of the endothelium.* Br J Anaesth, 2004. **93**(1): p. 105-13.
2. Pober, J.S. and W.C. Sessa, *Evolving functions of endothelial cells in inflammation.* Nat Rev Immunol, 2007. **7**(10): p. 803-15.
3. Dejana, E., *Endothelial cell–cell junctions: happy together.* Nature Reviews Molecular Cell Biology, 2004. **5**(4): p. 261-270.
4. Gavard, J., *Endothelial permeability and VE-cadherin: a wacky comradeship.* Cell Adh Migr, 2014. **8**(2): p. 158-64.
5. Scalise, A.A., et al., *The blood-brain and gut-vascular barriers: from the perspective of claudins.* Tissue Barriers, 2021. **9**(3): p. 1926190.
6. Hirata, K., et al., *Cloning of an immunoglobulin family adhesion molecule selectively expressed by endothelial cells.* J Biol Chem, 2001. **276**(19): p. 16223-31.
7. Melincovici, C.S., et al., *Vascular endothelial growth factor (VEGF) - key factor in normal and pathological angiogenesis.* Rom J Morphol Embryol, 2018. **59**(2): p. 455-467.
8. Butler, L.M., et al., *Analysis of Body-wide Unfractionated Tissue Data to Identify a Core Human Endothelial Transcriptome.* Cell Syst, 2016. **3**(3): p. 287-301 e3.
9. Tadijan, A., et al., *KANK family proteins in cancer.* Int J Biochem Cell Biol, 2021. **131**: p. 105903.
10. Kakinuma, N., et al., *Kank proteins: structure, functions and diseases.* Cell Mol Life Sci, 2009. **66**(16): p. 2651-9.
11. Bouchet, B.P., et al., *Talin-KANK1 interaction controls the recruitment of cortical microtubule stabilizing complexes to focal adhesions.* Elife, 2016. **5**.
12. Sun, Z., et al., *Kank2 activates talin, reduces force transduction across integrins and induces central adhesion formation.* Nature Cell Biology, 2016. **18**(9): p. 941-953.
13. Guo, Q., et al., *Structural basis for the recognition of kinesin family member 21A (KIF21A) by the ankyrin domains of KANK1 and KANK2 proteins.* J Biol Chem, 2018. **293**(2): p. 557-566.
14. Katoh, K., Y. Kano, and S. Ookawara, *Role of stress fibers and focal adhesions as a mediator for mechano-signal transduction in endothelial cells in situ.* Vasc Health Risk Manag, 2008. **4**(6): p. 1273-82.
15. Ramjaun, A.R. and K. Hodivala-Dilke, *The role of cell adhesion pathways in angiogenesis.* Int J Biochem Cell Biol, 2009. **41**(3): p. 521-30.
16. Wu, M.H., *Endothelial focal adhesions and barrier function.* J Physiol, 2005. **569**(Pt 2): p. 359-66.
17. Vouret-Craviari, V., et al., *Regulation of the actin cytoskeleton by thrombin in human endothelial cells: role of Rho proteins in endothelial barrier function.* Mol Biol Cell, 1998. **9**(9): p. 2639-53.
18. Lee, J., et al., *Endothelial Cell Focal Adhesion Regulates Transendothelial Migration and Subendothelial Crawling of T Cells.* Front Immunol, 2018. **9**: p. 48.
19. Boggetti, B., et al., *NBP, a zebrafish homolog of human Kank3, is a novel Numb interactor essential for epidermal integrity and neurulation.* Dev Biol, 2012. **365**(1): p. 164-74.
20. Hensley, M.R., et al., *Evolutionary and developmental analysis reveals KANK genes were co-opted for vertebrate vascular development.* Sci Rep, 2016. **6**: p. 27816.
21. Zhu, Y., et al., *Kank proteins: a new family of ankyrin-repeat domain-containing proteins.* Biochim Biophys Acta, 2008. **1780**(2): p. 128-33.
22. Kim, I., et al., *A novel HIF1AN substrate KANK3 plays a tumor-suppressive role in hepatocellular carcinoma.* Cell Biol Int, 2018. **42**(3): p. 303-312.

23. Dai, Z., et al., *KANK3 mediates the p38 MAPK pathway to regulate the proliferation and invasion of lung adenocarcinoma cells.* Tissue Cell, 2023. **80**: p. 101974.
24. Norreen-Thorsen, M., et al., *A human adipose tissue cell-type transcriptome atlas.* Cell Rep, 2022. **40**(2): p. 111046.
25. Dusart, P., et al., *A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes.* Cell Rep, 2019. **29**(6): p. 1690-1706.e4.
26. Dusart, P., et al., *A tissue centric atlas of cell type transcriptome enrichment signatures.* bioRxiv, 2023: p. 2023.01.10.520698.
27. Öling, S., et al., *A human stomach cell type transcriptome atlas.* bioRxiv, 2023: p. 2023.01.10.520700.
28. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
29. Usuba, R., et al., *EGFL7 regulates sprouting angiogenesis and endothelial integrity in a human blood vessel model.* Biomaterials, 2019. **197**: p. 305-316.
30. Dai, C., et al., *Regulatory mechanisms of Robo4 and their effects on angiogenesis.* Biosci Rep, 2019. **39**(7).
31. Ishida, T., et al., *Targeted disruption of endothelial cell-selective adhesion molecule inhibits angiogenic processes in vitro and in vivo.* J Biol Chem, 2003. **278**(36): p. 34598-604.
32. Sauteur, L., et al., *Cdh5/VE-cadherin promotes endothelial cell interface elongation via cortical actin polymerization during angiogenic sprouting.* Cell Rep, 2014. **9**(2): p. 504-13.
33. Ashburner, M., et al., *Gene Ontology: tool for the unification of biology.* Nature Genetics, 2000. **25**(1): p. 25-29.
34. Tabula Sapiens, C., et al., *The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.* Science, 2022. **376**(6594): p. eabl4896.
35. Uhlén, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
36. Dusart, P., et al., *A Systems-Based Map of Human Brain Cell-Type Enriched Genes and Malignancy-Associated Endothelial Changes.* Cell Rep, 2019. **29**(6): p. 1690-1706 e4.
37. Kustatscher, G., et al., *An open invitation to the Understudied Proteins Initiative.* Nat Biotechnol, 2022. **40**(6): p. 815-817.
38. Kustatscher, G., et al., *Understudied proteins: opportunities and challenges for functional proteomics.* Nat Methods, 2022. **19**(7): p. 774-779.
39. Baker, M., *Reproducibility crisis: Blame it on the antibodies.* Nature, 2015. **521**(7552): p. 274-6.
40. Weller, M.G., *Quality Issues of Research Antibodies.* Anal Chem Insights, 2016. **11**: p. 21-7.
41. Stoeger, T., et al., *Large-scale investigation of the reasons why potentially important genes are ignored.* PLoS Biol, 2018. **16**(9): p. e2006643.
42. Rafiq, N.B.M., et al., *A mechano-signalling network linking microtubules, myosin IIA filaments and integrin-based adhesions.* Nat Mater, 2019. **18**(6): p. 638-649.
43. Tzima, E., et al., *Activation of Rac1 by shear stress in endothelial cells mediates both cytoskeletal reorganization and effects on gene expression.* EMBO J, 2002. **21**(24): p. 6791-800.
44. Dave, J.M. and K.J. Bayless, *Vimentin as an integral regulator of cell adhesion and endothelial sprouting.* Microcirculation, 2014. **21**(4): p. 333-44.
45. Tsuruta, D. and J.C. Jones, *The vimentin cytoskeleton regulates focal contact size and adhesion of endothelial cells subjected to shear stress.* J Cell Sci, 2003. **116**(Pt 24): p. 4977-84.
46. Hoelzle, M.K. and T. Svitkina, *The cytoskeletal mechanisms of cell-cell junction formation in endothelial cells.* Mol Biol Cell, 2012. **23**(2): p. 310-23.

47. Helmke, B.P., R.D. Goldman, and P.F. Davies, *Rapid displacement of vimentin intermediate filaments in living endothelial cells exposed to flow.* Circ Res, 2000. **86**(7): p. 745-52.
48. Wechezak, A.R., R.F. Viggers, and L.R. Sauvage, *Fibronectin and F-actin redistribution in cultured endothelial cells exposed to shear stress.* Lab Invest, 1985. **53**(6): p. 639-47.
49. Battaglia, R.A., et al., *Vimentin on the move: new developments in cell migration.* F1000Res, 2018. **7**.
50. Dusart, P., et al., *A systems-approach reveals human nestin is an endothelial-enriched, angiogenesis-independent intermediate filament protein.* Sci Rep, 2018. **8**(1): p. 14668.
51. Zhang, C., et al., *KANK4 Promotes Arteriogenesis by Potentiating VEGFR2 Signaling in a TALIN-1-Dependent Manner.* Arterioscler Thromb Vasc Biol, 2022. **42**(6): p. 772-788.
52. Witkowski, M., U. Landmesser, and U. Rauch, *Tissue factor as a link between inflammation and coagulation.* Trends in Cardiovascular Medicine, 2016. **26**(4): p. 297-303.
53. Ott, L.W., et al., *Tumor Necrosis Factor-alpha- and interleukin-1-induced cellular responses: coupling proteomic and genomic information.* J Proteome Res, 2007. **6**(6): p. 2176-85.
54. Guo, S.S., et al., *Tissue distribution and subcellular localization of the family of Kidney Ankyrin Repeat Domain (KANK) proteins.* Exp Cell Res, 2021. **398**(1): p. 112391.
55. Zhu, Y., et al., *Identification of three immune subtypes characterized by distinct tumor immune microenvironment and therapeutic response in stomach adenocarcinoma.* Gene, 2022. **818**: p. 146177.
56. Fredolini, C., et al., *Systematic assessment of antibody selectivity in plasma based on a resource of enrichment profiles.* Sci Rep, 2019. **9**(1): p. 8324.
57. Bradbury, A. and A. Plückthun, *Reproducibility: Standardize antibodies used in research.* Nature, 2015. **518**(7537): p. 27-29.
58. Schonbrunn, A., *Editorial: Antibody can get it right: confronting problems of antibody specificity and irreproducibility.* Mol Endocrinol, 2014. **28**(9): p. 1403-7.
59. Sun, Z., et al., *Kank2 activates talin, reduces force transduction across integrins and induces central adhesion formation.* Nat Cell Biol, 2016. **18**(9): p. 941-53.
60. Davis, G.E. and D.R. Senger, *Endothelial Extracellular Matrix.* Circulation Research, 2005. **97**(11): p. 1093-1107.
61. Thomsen, M.S., L.J. Routhe, and T. Moos, *The vascular basement membrane in the healthy and pathological brain.* J Cereb Blood Flow Metab, 2017. **37**(10): p. 3300-3317.
62. Quadri, S.K., *Cross talk between focal adhesion kinase and cadherins: role in regulating endothelial barrier function.* Microvasc Res, 2012. **83**(1): p. 3-11.
63. Mott, R.E. and B.P. Helmke, *Mapping the dynamics of shear stress-induced structural changes in endothelial cells.* Am J Physiol Cell Physiol, 2007. **293**(5): p. C1616-26.
64. Carisey, A., et al., *Vinculin regulates the recruitment and release of core focal adhesion proteins in a force-dependent manner.* Curr Biol, 2013. **23**(4): p. 271-81.
65. Urbich, C., et al., *Laminar Shear Stress Upregulates Integrin Expression.* Circulation Research, 2000. **87**(8): p. 683-689.
66. Ye, N., et al., *Direct observation of α-actinin tension and recruitment at focal adhesions during contact growth.* Exp Cell Res, 2014. **327**(1): p. 57-67.
67. Verma, D., et al., *Flow-induced focal adhesion remodeling mediated by local cytoskeletal stresses and reorganization.* Cell Adh Migr, 2015. **9**(6): p. 432-40.
68. Sun, Z., et al., *Extracellular matrix-specific focal adhesions in vascular smooth muscle produce mechanically active adhesion sites.* Am J Physiol Cell Physiol, 2008. **295**(1): p. C268-78.
69. Fletcher, D.A. and R.D. Mullins, *Cell mechanics and the cytoskeleton.* Nature, 2010. **463**(7280): p. 485-92.

70. Powell , K., *Actin and microtubules interact via MAPs.* Journal of Cell Biology, 2005. **170**(5): p. 701-701.
71. Pogoda, K., et al., *Unique Role of Vimentin Networks in Compression Stiffening of Cells and Protection of Nuclei from Compressive Stress.* Nano Letters, 2022. **22**(12): p. 4725-4732.
72. Satelli, A. and S. Li, *Vimentin in cancer and its potential as a molecular target for cancer therapy.* Cell Mol Life Sci, 2011. **68**(18): p. 3033-46.
73. Ridge, K.M., et al., *Roles of vimentin in health and disease.* Genes Dev, 2022. **36**(7-8): p. 391-407.
74. Boraas, L.C. and T. Ahsan, *Lack of vimentin impairs endothelial differentiation of embryonic stem cells.* Scientific Reports, 2016. **6**(1): p. 30814.
75. Chernoivanenko, I.S., A.A. Minin, and A.A. Minin, *[Role of vimentin in cell migration].* Ontogenez, 2013. **44**(3): p. 186-202.
76. Messica, Y., et al., *The role of Vimentin in Regulating Cell Invasive Migration in Dense Cultures of Breast Carcinoma Cells.* Nano Letters, 2017. **17**(11): p. 6941-6948.
77. Lawson, C., et al., *FAK promotes recruitment of talin to nascent adhesions to control cell motility.* J Cell Biol, 2012. **196**(2): p. 223-32.
78. Rahikainen, R., et al., *Mechanical stability of talin rod controls cell migration and substrate sensing.* Sci Rep, 2017. **7**(1): p. 3571.
79. Pogoda, K., et al., *Unique Role of Vimentin Networks in Compression Stiffening of Cells and Protection of Nuclei from Compressive Stress.* Nano Lett, 2022. **22**(12): p. 4725-4732.
80. Peng, Z., et al., *Endothelial Response to Pathophysiological Stress.* Arteriosclerosis, Thrombosis, and Vascular Biology, 2019. **39**(11): p. e233-e243.
81. Raoul, A., et al., *Role of focal adhesions of vascular smooth muscle cells in thrombin generation.* Archives of Cardiovascular Diseases Supplements, 2020. **12**(1): p. 145.
82. Elvevold, K., I. Kyrrestad, and B. Smedsrød, *Protocol for Isolation and Culture of Mouse Hepatocytes (HCs), Kupffer Cells (KCs), and Liver Sinusoidal Endothelial Cells (LSECs) in Analyses of Hepatic Drug Distribution*, in *Antisense RNA Design, Delivery, and Analysis*, V. Arechavala-Gomeza and A. Garanto, Editors. 2022, Springer US: New York, NY. p. 385-402.
83. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics data.* The Innovation, 2021. **2**(3): p. 100141.
84. Lonsdale, J., et al., *The Genotype-Tissue Expression (GTEx) project.* Nature Genetics, 2013. **45**(6): p. 580-585.
85. Jones, R.C., et al., *The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans.* Science, 2022. **376**(6594): p. eabl4896.
86. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nat Methods, 2012. **9**(7): p. 676-82.

ACKNOWLEDGEMENTS

Funding was granted to L.M.B. from Hjärt Lungfonden (20170759, 20170537, 20200705) and Vetenskapsrådet (2019-01493). The Human Protein Atlas (HPA) is funded by The Knut and Alice Wallenberg Foundation. **Data usage:** We used data from the Genotype-Tissue Expression (GTEx) Project (gtexportal.org) [84] which was supported by the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.