From Department of Oncology-Pathology
Karolinska Institutet, Stockholm, Sweden

Long Non-coding RNAs and Cellular Interactions: Investigating Underlying Mechanisms of Oncogenesis

Jason T. Serviss

Karolinska Institutet

Stockholm 2019

Published by Karolinska Institutet.
Printed by Eprint AB 2019

ISBN 978-91-7831-586-4

Long non-coding RNAs and Cellular Interactions: Investigating Underlying Mechanisms of Oncogenesis
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Jason T. Serviss, MSc.

Principal Supervisor:
Assistant Professor Martin Enge
Karolinska Institutet
Department of Oncology-Pathology

Co-supervisor(s):
Professor Joakim Lundeberg
Science for Life Laboratory
KTH Royal Institute of Technology
Department of Gene Technology

Opponent:
Professor Peter Kharchenko
Harvard University
Department of Biomedical Informatics

Examination Board:
Professor Chandrasekhar Kanduri
University of Gothenburg
Department of Medical Biochemistry and Cell Biology

Associate Professor Gonçalo Castelo-Branco
Karolinska Instutitet
Department of Medical Biochemistry and Biophysics

Associate professor Ola Larsson
Karolinska Institutet
Department of Oncology-Pathology

"Our notions of law and harmony are commonly confined to those instances which we detect; but the harmony which results from a far greater number of seemingly conflicting, but really concurring, laws, which we have not detected, is still more wonderful." -Henry David Thoreau

ABSTRACT

Cancer is a leading cause of death worldwide with one in 8 men and one in 11 women dying from the disease (World Health Organization, 2018). Despite vast improvements in cancer diagnosis and therapy, the global cancer burden continues to rise in unison with population growth and longevity. Although cancer presents itself as a heterogeneous group of diseases, often divided by tissue of origin, tumor characterization increasingly identifies molecular level commonalities and patterns that are similar across all cancers. Expanding our knowledge of these molecular characteristics, together with the development of new tools and technologies, has historically been one of the most efficient ways to increase the effectivity of cancer therapies and thus, decrease the cancer burden of the population. This thesis investigates two newly identified molecular mechanisms, long non-coding RNAs and cell-cell interactions, whose role are increasingly appreciated in tumor progression and development. In addition, the thesis reports the development of methods and tools that have been established to facilitate further investigation of cancers molecular attributes by the scientific community.

LIST OF SCIENTIFIC PAPERS

I. **Serviss, J. T.***, Gådin, J. R.*, Eriksson, P., Folkersen, L., Grandér, D. (2017). ClusterSignificance: A bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data. *Bioinformatics*, *33*(19), 3126-3128.

II. **Serviss, J. T.**, Andrews, N., Van den Eynden, J., Richter, F. C., Houtman, M., Vesterlund, M., Schwarzmueller, L., Johnsson, P., Larsson, E., Grandér, D., Pokrovskaja Tamm, K. (2018). An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a. *Cell Death and Disease*, *9*(736).

III. **Serviss, J.T.***, Andrews, N.*, Andersson, A.B., Dzwonkowska, E., Heijboer, R., Gerling, M. & Enge, M. Unsupervised cell interaction profiling based on multiplet RNA sequencing reveals major architectural differences between small intestinal and colonic epithelium. *Manuscript.*

*Equal contribution

Publications not included in the thesis:

Serviss, J. T., Johnsson, P., Grandér, D. (2014). An emerging role for long non-coding RNAs in cancer metastasis. *Front. Genet.*, 5(234).

Pellegrini, P., **Serviss, J. T.**, Lundbäck, T., Bancaro, N., Mazurkiewicz, M., Kolosenko, I., Yu, D., Haraldsson, M., D'Arcy, P., Linder, S., De Milito, A. (2018). A drug screening assay on cancer cells chronically adapted to acidosis. *Cancer Cell International*, 18(147).

Edsbäcker, E., **Serviss, J. T.**, Kolosenko, I., Palm-Apergi, C., De Milito, A., Pokrovskaja Tamm, K. (2019). STAT3 is activated in multicellular spheroids of colon carcinoma cells and mediates expression of IRF9 and interferon stimulated genes. *Scientific Reports*, 9(536).

CONTENTS

1 Introduction 5
1.1 Long non-coding RNA 5
1.2 Cell-Cell interactions 7
2 Methodological Considerations 9
2.1 RNA sequencing 9
2.2 Dimensionality reduction 9
2.3 Classification 10
2.4 Particle swarm optimization 10
2.5 Principal curve 10
3 Aims of the Thesis 12
4 Results 13
4.1 Paper I: *ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data* 13
4.2 Paper II: *An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a* 14
4.3 Paper III: *Unsupervised cell interaction profiling based on multiplet RNA sequencing reveals major architectural differences between small intestinal and colonic epithelium* 15
5 Discussion 18
5.1 Paper I: *ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data* 18
5.2 Paper II: *An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a* 19
5.3 Paper III: *Unsupervised cell interaction profiling based on multiplet RNA sequencing reveals major architectural differences between small intestinal and colonic epithelium* 20
6 Concluding remarks 23
7 Acknowledgements 24
8 References 25

LIST OF ABBREVIATIONS

RNA	Ribonucleic acid
lncRNA	Long non-coding RNA
ncRNA	Non-coding RNA
miRNA	Micro-RNA
HOTAIR	Hox antisense intergenic RNA
TME	Tumor microenvironment
FISH	Fluorescence *in situ* hybridization
RNAseq	RNA sequencing
scRNAseq	Single cell RNA sequencing
lncTAM34a	Long non-coding transcriptional activator of miR34a
ChIP	Chromatin immunoprecipitation
CIM-seq	Cell interactions by multiplet sequencing
SI	Small intestine
FACS	Fluorescence-activated cell sorting
PCA	Principal component analysis
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection

1 INTRODUCTION

In recent years, considerable progress in the treatment of cancer has been witnessed by the medical and research communities. From major advances in chemotherapy treatment, to various forms of targeted therapies, and vaccines aiding in the prevention of cervical cancer, the last four decades have seen rapid improvements in cancer therapy. Many of these discoveries can be directly related to improving technologies that facilitate a better understanding of the underlying biology of cancer. Armed with this understanding, researchers have been able to develop methods to target cancers vulnerabilities via both general and specific mechanisms. For example, in the 1960's-1970's the discovery of the proto-oncogene Src (Oppermann et al., 1979) and epidermal growth factor receptor (Cohen and Elliott, 1963), as well as the role these play in growth factor signaling, opened the door for the development of drugs targeting tyrosine kinases and also initiated the era of "targeted therapies". Among the more recently discovered players in oncogenesis, long non-coding RNAs (lncRNA) and cell-cell interactions have both been in the limelight of the scientific community due to their evolving role in cancer biology and their promising potential as therapeutic targets (Balas and Johnson, 2018; Kamińska et al., 2015). This thesis presents a multi-level investigation of these newly appreciated molecular elements as well as several methodologies and high-throughput bioinformatic-based tools developed for their interrogation in cancer development and progression.

1.1 LONG NON-CODING RNA

Characteristics of lncRNAs. Advances in sequencing technology have revealed that a large portion of the genome is transcribed while only a fraction of transcripts gives rise to proteins (Djebali et al., 2012; International Human Genome Sequencing Consortium, 2004). These non-translated RNAs are collectively termed non-coding RNAs (ncRNAs) and are typically further sub-classified by their size and/or their location in relation to other genomic features. One of these classes, microRNAs (miRNA), have been intensively studied and are involved in a wide range of disease types such as inflammatory, neurodevelopmental, and autoimmune diseases, as well as cancer (Ardekani and Naeini, 2010). The roles of another class of ncRNAs, lncRNA, are currently being unraveled and have been shown to be involved in the regulation of key cellular processes, such as chromosome inactivation, differentiation, cell cycle, and apoptosis (Johnsson et al., 2014). Despite this, to date only a handful lncRNAs have been functionally characterized. lncRNAs are, per definition, non-coding and longer than 200 nucleotides and also tend to show a lower rate of conservation and expression but higher rate of cell-type specificity than protein coding genes (Uszczynska-Ratajczak et al., 2018). Various mechanisms, mediated by RNA:RNA, RNA:protein, and RNA:DNA interactions, have been shown to be implemented by lncRNAs to control transcriptional and post-transcriptional regulation (**Figure 1**). These include transcriptional regulation via epigenetic mechanisms in both cis and trans, altering the stability of protein-coding RNAs, and inhibition of post-translational modifications (Johnsson et al., 2013; Mahmoudi et al., 2016; Wang et al., 2014).

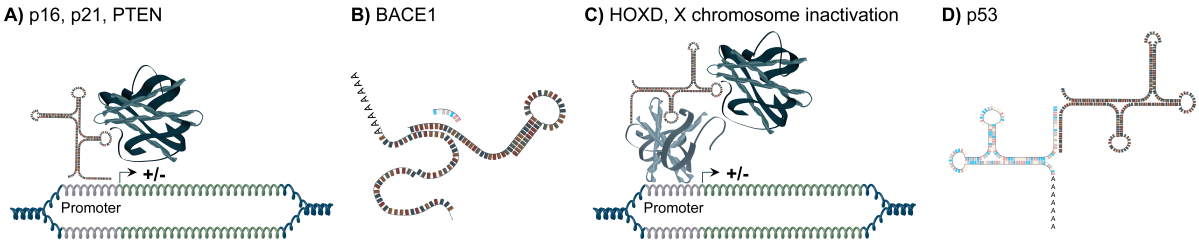

Figure 1. Mechanisms of lncRNA-mediated gene regulation. Genes reported to be regulated by the illustrated mechanisms are listed above each respective panel. **A)** lncRNAs can bind directly to DNA at gene promoters and subsequently recruit additional factors that can both positively or negatively regulate gene expression. **B)** lncRNAs can function as competing endogenous RNAs blocking miRNA-binding sites resulting in de-repression of miRNA target genes. **C)** Anchor-like functions have been reported for lncRNAs where they serve as scaffolds for multiple factors e.g. at gene promoters. **D)** Direct binding of lncRNAs to mRNA can post-transcriptionally increase or decrease mRNA degradation.

lncRNAs in cancer. Early indications arose that lncRNAs may be involved in cancer development and progression when many were found to be dysregulated in various cancer types (Niknafs et al., 2016; Prensner et al., 2011). Individual functional studies revealed the specific interactions of several of these lncRNAs and, as well, demonstrated their utility as diagnostic and prognostic markers (Gupta et al., 2010; Redis et al., 2013). The lncRNA antisense noncoding RNA in the INK4 locus (ANRIL), for example, has been shown to epigenetically silence the tumor suppressor gene INK4/p15 by recruitment of the polycomb repressor complex 1 component, chromobox 7 (Yap et al., 2010; Yu et al., 2008). High levels of ANRIL are associated with poor prognosis in numerous cancer types, such as hepatocellular carcinoma, lung cancer, and cervical cancer (Hua et al., 2015; Lin et al., 2015; Zhang et al., 2017). Another well-studied lncRNA, Hox antisense intergenic RNA (HOTAIR), serves as a scaffold for several protein complexes including members of the polycomb repressive complex 2 and Lsd1. Recruitment of these chromatin modifying factors to specific genomic loci deposits the repressive H3K27me3 histone modification and removes the activating H3K4me2 histone modification resulting in a net repression of the loci (Tsai et al., 2010). HOTAIR was initially implicated in the regulation of the HOXD gene in trans but was subsequently shown to regulate multiple genes across the genome including the metastasis suppressors PCDH10, PCDHB5, and JAM2 (Gupta et al., 2010; Rinn et al., 2007). Regulation of these metastasis suppressor genes is thought to give rise to the poor survival and increased occurrence of metastasis associated with high HOTAIR levels in breast and colorectal cancers, among others (Gupta et al., 2010; Kogo et al., 2011).

The future of lncRNAs in the clinic. Although lncRNAs are already proving to be important diagnostic and prognostic biomarkers, their use in treatment regimes faces a long road ahead with many obstacles to be overcome (Mouraviev et al., 2016; Sánchez and Huarte, 2013; Saus et al., 2016). Due to lncRNAs cell-type specificity and lack of post-translational modifications, they may be especially suited as biomarkers. Although utilization of lncRNAs as biomarkers in the clinic is currently limited, initial studies show promising results (Lee et al., 2011; Qi et al., 2016). Despite the fact that lncRNA-based therapies have not yet advanced to clinical studies, other non-coding RNAs have seen some clinical progress (Beg et al., 2017; Kao et al., 2015). Several strategies can be envisioned for modulating lncRNAs clinically such as targeting a) over-expressed lncRNAs, b) lncRNA interactions, or c) lncRNA structure. Repression of lncRNAs directly may potentially be mediated via siRNAs or other modified short nucleic acids, the utility of which has shown to be promising in multiple clinical studies (Kanasty et al., 2013; Zuckerman and Davis, 2015). Regulating lncRNA interactions offers an alternative approach where either the interacting region of the lncRNA may be targeted or, in the case of RNA:protein interactions, the interacting partner may be targeted, potentially via small molecule inhibitors. Finally, due the proposed importance of lncRNA secondary structure

for lncRNA functionality, structural disturbance of these could also be a valid route towards mediating the effect of specific lncRNAs (Li et al., 2016). Currently, fast and reliable prediction of lncRNA secondary structure and its associated function is limited and advances in this field would need to be accomplished before this approach would be feasible.

1.2 CELL-CELL INTERACTIONS

Cell-cell interactions during tissue homeostasis. Strict control of tissue architecture is necessary to maintain continued functionality of the tissue throughout the life of multicellular organisms (Morrison and Scadden, 2014; Sato et al., 2011). For example, the small intestine (SI) is organized into discrete structures, known as villi and crypts, which increase the tissues surface area and are essential for its ability to effectively perform its absorptive functions (**Figure 2**). It is estimated that some 10^{11} cells per day are shed from the human intestine (mice ≈ 2 x 10^{8}) underlining the continuous need for organizational control over the massive amount of turnover and regeneration that is continuously ongoing in this tissue (Williams et al., 2015). Even in cases such as the SI, where the tissue architecture is well characterized, mechanisms that regulate this organization are not well understood, although cell-cell junctions are thought to play a leading role.

Cell-cell interactions are also important for controlling various cellular functions and the maintenance of cell identity (Morrison and Scadden, 2014; Sato et al., 2011). For example, a stem cell population residing at the bottom of the SI crypts produces the SI cell types. These stem cells undergo symmetric cell division where the daughter cells are initially equivalent but either maintain their stem cell identity or become differentiated cells dependent on cellular context (**Figure 2,** Lopez-Garcia et al., 2010; Snippert et al., 2010). The cells are positioned as such that when cell divisions take place they may lose

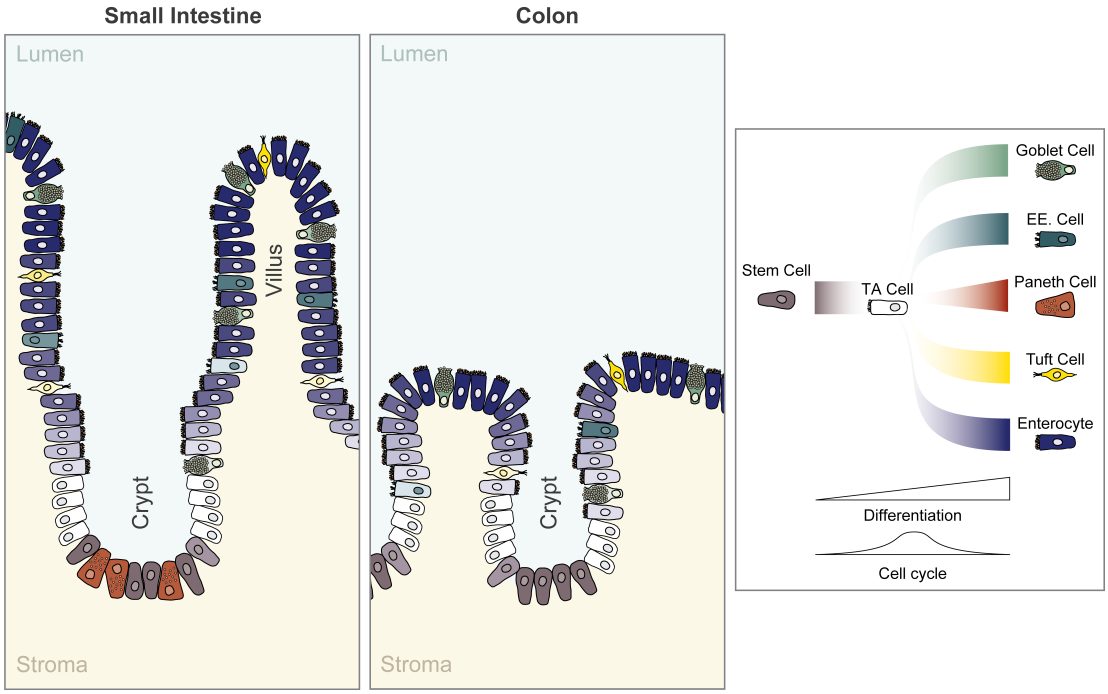

Figure 2. Gut architecture. The architecture of the small intestine (left panel) and colon (middle panel) showing major cell types. The legend (right panel) illustrates the differentiation scheme for gut cell types. Transit amplifying (TA), Enteroendocrine (EE).

contact with a specific neighboring cell type, known as a paneth cell. Interaction with paneth cells promotes the preservation of the stem cell phenotype and thus, it is the loss or maintenance of this contact that determines if the daughter cells remains a stem cell or differentiates. Other well studied changes in cell identity regulated by cell-cell interactions include epithelial-mesenchymal transition (Arias, 2001), retinal vascular and endothelial cell survival (Giannotta et al., 2013; Roy et al., 2017), and epidermal vs. neuronal fate specification via lateral inhibition (Alberts et al., 2002).

The role of cell-cell interactions in the tumor microenvironment. The previous view that a tumor is wholly composed of a relatively homogeneous group of uncontrollably proliferating cells, has been gradually replaced with the insight that a tumor's complexity actually reflects or exceeds that of a normal organ, as well as the tumor microenvironment's (TME) instrumental role in shaping that complexity (Egeblad et al., 2010). The TME constitutes the non-cancerous cell types surrounding and infiltrating the tumor, the extracellular matrix, as well as conditions that are shaped collectively by these and the tumor cells such as hypoxia, pH, and interstitial fluid pressure. In recent years, advances in high-throughput single cell applications have begun to uncover the heterogeneity of tumor cell populations, on both a genetic and transcriptomic level, although the functional characterization of these sub-populations still remains challenging (Gawad et al., 2014; Patel et al., 2014). During tumor development, the TME is gradually transformed from a structurally organized normal tissue to a heterogeneous group of cells and conditions reflecting the environment formed by the various cell populations residing within the tumor (Marusyk et al., 2012; Wells et al., 2015). In addition, the TME provides the selective pressure for the ongoing Darwinian selection process of the tumor cell population, resulting in clones with improved fitness characteristics. Efforts are ongoing to characterize the interactions of the tumor cell populations with their TME "niche" as this interaction has been reported to support multiple features of cancer progression, such as increased cell survival, metastasis, therapy resistance, inflammation, and angiogenesis, among others (Hanahan and Weinberg, 2011; Junttila and de Sauvage, 2013).

Detecting cell-cell interactions. Current methods capable of interrogating cell-cell interactions in a high throughput and quantitative manner can be roughly divided into four main groups: array-based, fluorescence *in situ* hybridization (FISH)-based, microdissection-based, and *in situ* sequencing-based (Boisset et al., 2018; Codeluppi et al., 2018; Crosetto et al., 2015; Ståhl et al., 2016; Wang et al., 2018). In addition to detecting cell-cell adjacency, these methods all quantify gene expression to determine individual cell types and, thus, the gene expression quantification values are included as part of the methods output. In several cases, these methods are also able to relate changes in gene expression to a specific interaction type. Despite this, all of these methods suffer from a low spatial resolution, reliance on pre-defined cell type biomarkers, a limited number of genes that can be quantified, RNA diffusion, or a need for non-standard specialized equipment.

2 METHODOLOGICAL CONSIDERATIONS

The projects contained within the thesis make use of previously developed statistics, algorithms, and scientific methods. A brief overview of some of these methods is provided below to give sufficient background to the reader.

2.1 RNA SEQUENCING

Quantification of gene expression via measurement of RNA abundance can be accomplished using multiple methods on both a targeted and global analysis level. Of these, RNA sequencing (RNAseq) uses next generation sequencing technology to identify and quantify the RNAs present in cells and can be performed at both bulk and single cell resolution (Stark et al., 2019). Briefly, the *in vitro* elements of RNAseq are carried out by isolating RNA, performing reverse transcription, cDNA fragmentation, size selection, and addition of sequencing linkers. The resulting libraries are sequenced and result in generation of data in the form of RNA sequences or "reads". In order to identify which genes the reads correspond to and quantify the gene expression, the *in silico* analysis begins by aligning the reads to an annotated reference genome with known gene coordinates. Quantification can subsequently be performed in several ways, such as counting of reads that are aligned to a specific area of the genome that corresponds to a known gene.

Single cell (sc) RNAseq protocols, generally follows the same steps as bulk RNAseq with several modifications to overcome the reduced amounts of starting material. Individual cells are isolated into wells, typically using fluorescence-activated cell sorting (FACS), or individual droplets using microfluidics-based techniques (Picelli et al., 2014; Zheng et al., 2017). Technologies exist enabling the sequencing of 5-prime, 3-prime, or full-length sequences and may or may not facilitate the use of unique molecular identifiers (Islam et al., 2014; Kivioja et al., 2012).

2.2 DIMENSIONALITY REDUCTION

Dimensionality reduction is a technique to reduce high dimensional data to a lower dimensional representation and, thus, limit the number of features needed to represent the data. Oftentimes, high dimensional data has multiple features that strongly correlate with each other and can be represented in a simplified but sufficient fashion by merging these features. This is especially relevant in the case of gene expression where e.g. an environmental signal initiating a change in gene expression causes up- and down-regulation of multiple genes. Dimensionality reduction facilitates visualization of the data and is often used as a feature selection and noise reduction pre-processing step for further downstream analysis.

Multiple dimensionality reduction methods exist, (e.g. Principal Component Analysis [PCA], t-Distributed Stochastic Neighbor Embedding [t-SNE], and Uniform Manifold Approximation and Projection [UMAP]), and vary in their suitability in a specific case depending on the type of input data and the goals of the analysis (Hotelling, 1933; Maaten and Hinton, 2008; McInnes et al., 2018; Pearson, 1901). Some dimensionality reduction methods, e.g. PCA, seek a linear combination of the features whereas others, e.g. t-SNE or UMAP, are non-linear. In addition, some of dimensionality reduction methods aim to maintain global data structures whereas others do not. t-SNE, for example, was primarily designed as a visualization aid and, hence, does not maintain global data structures well whereas PCA and UMAP have a stronger capability to do so and are therefore more suitable for pre-processing applications. Together, the attributes of the dimensionality

reduction algorithm and the goals of the analysis help to aid in deciding which method is most appropriate for the application at hand.

2.3 CLASSIFICATION

Classification aims to identify which group within a population a new sample belongs to depending on the similarity of the sample to already defined samples. Classification can be performed in either a supervised or unsupervised fashion. Supervised classification uses a subset of the total data where the classifications are known to "learn" how to identify the classes before being utilized on another dataset where the classes are unknown. Unsupervised classification, on the other hand, is performed without the algorithm obtaining prior information regarding the classes and performs the classification solely on the basis of the data provided. This results in the definition of groups of samples within the data that have similar feature patterns to each other. Some examples of unsupervised classification include k-means clustering, autoencoders, and graph-based methods.

Graph-based classification methods utilize data representations if the form of a network structure that can be created by computing a distance/similarity metric to portray the relationship of the samples to each other. One such graph-based classification algorithm, Louvain community detection, functions by optimizing graph modularity and is quick and effective at resolving classifications in large datasets (Andrea Lancichinetti and Santo Fortunato, 2010; Blondel et al., 2008). Despite this, graph-based classification methods often require input from the user based on assumptions reflecting the number of classes that are expected to be present in the data. In addition, they are known to suffer from a resolution limit under which their capability of detecting clusters is diminished (Fortunato and Barthélemy, 2007). This is especially important to take into account when utilizing graph-based classification algorithms to analyze scRNAseq data when rare cell populations are present.

2.4 PARTICLE SWARM OPTIMIZATION

Particle swarm optimization is one of many types of mathematical optimization whose goal is to select an optimal value from many choices based on a set of criteria. Typically, optimization is used to minimize or maximize a function by iteratively and systematically altering candidate solutions. With particle swarm optimization these candidate solutions, known as particles, move in the search space under the influence of the currently best-identified solution, as well as mathematical constraints such as particle position, velocity, and user defined constraints. The algorithm terminates upon reaching various stopping criteria providing both the best identified candidate solution and the value produced by the function using said solution. Particle swarm optimization is a robust stochastic optimization method and, due to its emulation of biological behavior, is especially suited to situations where it is difficult to assess the validity of assumptions concerning the scale, differentiability, or shape of the problem (Ab Wahab et al., 2015).

2.5 PRINCIPAL CURVE

Principal curve analysis is a method that aims to produce a smooth and nonlinear summary of multidimensional data. It is approximated in such a way that each point on the curve is an average of the surrounding data points and passes through the middle of the data in an orthogonal sense (**Figure 3**). The principal curve algorithm typically begins by using the first principal component as a line through the data after which the algorithm optimizes the average distance in arc length of the p-dimensional points from the previous iteration until self-consistency is reached. Self-consistency, in this case, means that the position of the curve at any individual point can be identified by calculating the average of all data points

projecting to that point. Hastie and Stuetzle state “The human eye is skilled at making trade-offs between smoothness and fidelity to the data; we would like a procedure that makes this judgment automatically” (Hastie and Stuetzle, 1989). Due to this, principal curves are especially useful in cases where it is necessary to easily understand and judge the relationship between the data and the curve.

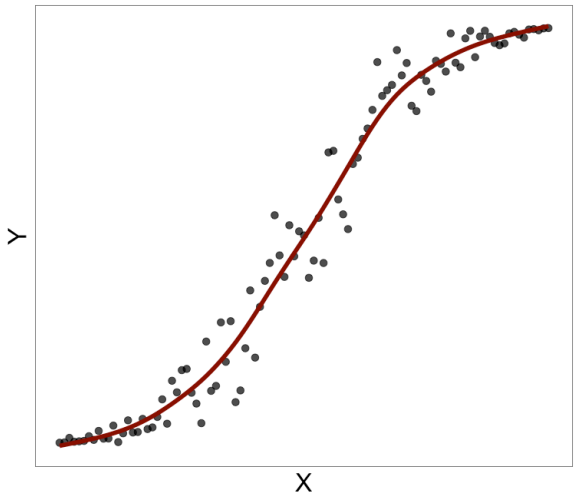

Figure 3. Principal curve. The figure illustrates two-dimensional *in silico*-generated data (black points) with a principal curve (red line) indicating how the curve produces a smooth summary and passes through the middle of the data.

3 AIMS OF THE THESIS

The overall aim of the thesis was to investigate mechanisms that play a pivotal role in oncogenesis. This was accomplished by specifically focusing on two relatively novel players in tumor development: lncRNAs and cell-cell interactions. The thesis also aims to describe several bioinformatics-based methods and software that were designed and utilized to help achieve the overall goals of the thesis.

The specific aim of each paper was:

Paper I: Here we present ClusterSignificance; a software and statistical methodology allowing for the determination of class separations in dimensionality reduced data. In addition, the paper aims to use ClusterSignificance to evaluate a potential role of lncRNAs in the identity of hematological malignancies.

Paper II: In this paper we aim to characterize lncTAM34a, a long non-coding RNA in antisense orientation to the tumor suppressor micro-RNA 34a, and its role in oncogenesis.

Paper III: This paper aims to present Cell Interaction by Multiplet Sequencing (CIM-seq), a high-throughput and unsupervised method for investigating cell-cell interactions. We subsequently utilize CIM-seq to explore and gain a better understanding of the architecture of murine colonic crypts.

4 RESULTS

4.1 PAPER I: *ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data*

In this paper we describe the development ClusterSignificance, a software and method that facilitates the evaluation of a separation between known classes in dimensionality-reduced data. Although routinely used in research, the output of dimensionality reduction algorithms is typically evaluated in a subjective manner through visualization of their output. ClusterSignificance provides statistical rigor to this process and formally evaluates the independence of the known class labels and the features (genes). ClusterSignificance is useful when e.g. a set of genes is thought to characterize two patient groups but it is unclear if these two groups are separated when evaluating the output from dimensionality reduction algorithms. The method works by further decreasing the data dimensionality into one dimension before scoring the separation based on the user defined classes and comparing this separation score to separation scores from permuted data. ClusterSignificance is also an intuitive and strongly visual method and has limited assumptions, making it useful for a broad range of researchers and research questions.

ClusterSignificance requires the known class labels and the data representation post-reduction as input after which the workflow proceeds in three distinct stages, 1) Principal Curve Projection, 2) Separation Classification, and 3) Score Permutation (**Paper I**, Fig 1a-1c). A brief description of each stage is outlined below:

1) Principal Curve Projection: The goal of this stage is to further reduce the data to one dimension. This is accomplished by projecting the data points onto a principal curve and subsequently calculating the Euclidean distance between the points, thus completing the one dimensional projection.
2) Separation Classification: In this stage, the previously projected data is used to determine the optimal separation between classes in the projected space using the class definitions provided by the user. Each possible separation is scored by calculating the complement of the Euclidean distance along the ROC curve to the operating point. The highest separation score is retained for later use.
3) Score Permutation: In the final stage, the user defined classes are randomly assigned to the data and steps 1-2 are re-run *n* times with the best separation score being recored each time. Once *n* runs are completed, a p-value for the separation is calculated as the fraction of separation scores for the permuted data that were higher than the original data.

ClusterSignificance was tested using *in silico* data generated to have either 0 or 100% overlap of the classes and was shown to correctly characterize true and false separations in contexts where they are known to exist (**Paper I**, Supplementary Fig. 1a-1c).

Finally, we used ClusterSignificance to determine if patients with multiple types of hematological malignancies could be defined by their lncRNA expression alone. To this end, we utilized 2096 microarray samples representing 6 different hematological

malignancies as well as non-leukemic and healthy patients. The results from ClusterSignificance indicated that the vast majority of malignancies were significantly separated, suggesting the importance of lncRNAs in the identity of these cancers and prompting further investigation to better understand their role in these diseases (**Paper I**, Fig. 2a-2b).

4.2 PAPER II: *An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a*

Here we identify and characterize *long non-coding transcriptional activator of micro-RNA34a* (lncTAM34a), a lncRNA transcribed in a head-to-head antisense orientation to the micro-RNA34a (miR34a) tumor suppressor gene (**Paper II**, Fig. 1a). Due to miR34a's known role as a master regulator of tumor suppression it is not surprising that its expression is dysregulated in a broad range of hematological and solid tumors. Despite this, the mechanisms underlying this dysregulation are largely unknown. We hypothesized that lncTAM34a may be important in the regulation of miR34a expression and, therefore, began by examining its expression in a panel of cell lines and a large dataset of primary human cancers. Due to the fact that miR34a is a direct downstream target of TP53, we chose to include cell lines and samples with varying but known TP53 status. Collectively these results indicated that miR34a and lncTAM34a are positively correlated in both TP53 wild-type and mutated settings (**Paper II**, Fig. 1b-1c, Supplementary Fig. 1a). In addition, we found that both lncTAM34a and miR34a expression levels are reduced in $TP53^{-/-}$ cell lines and patients with nonsynonymous TP53 mutations (**Paper II**, Fig. 1b-1c, Supplementary Fig. 1b).

We next performed a thorough molecular characterization of lncTAM34a by first determining its transcription start and stop sites (**Paper II**, Fig. 1d-1e, Supplementary Fig. 2a), polyadenylation status (**Paper II**, Supplementary Fig. 2b), alternative splicing isoforms (**Paper II**, Supplementary Fig. 2c), and nuclear localization within the cell (**Paper II**, Supplementary Fig. 2d). We also confirmed lncTAM34a's non-coding status using two separate bioinformatics-based methods and, in addition, searching a large dataset of mass spectrometry peptides from eleven cancer cell lines for peptides corresponding to lncTAM34a (**Paper II**, Fig. 1f, Supplementary Fig. 2e).

miR34a and lncTAM34a expression have been previously reported to be increased upon cellular stress that results in the activation of TP53. As such, we investigated this in multiple experimental systems and showed miR34a and lncTAM34a expression to be induced upon DNA damage-mediated TP53 activation (**Paper II**, 2a-2b). We furthermore demonstrated that TP53-mediated miR34a and lncTAM34a expression can be regulated from a single promoter and that the expression takes place in a bidirectional manner (**Paper II**, 2c).

We next examined the function of lncTAM34a hypothesizing that it may serve to regulate the expression of miR34a host gene (HG). Using si- and sh-RNA in several cell lines and in the presence or absence of activated TP53, we demonstrated the ability of lncTAM34a to positively regulate miR34a expression (**Paper II**, Fig. 2d-2e, Supplementary Fig. 3c). Despite the confirmed ability of TP53 to regulate the miR34a/lncTAM34a locus, previous results had indicated that other factors were also able to regulate this locus (**Paper II**, Fig. 2b). In order to better understand the regulation dynamics of lncTAM34a and miR34a in the absence of TP53, we stably overexpressed lncTAM34a in three different TP53-null cell lines. Our results indicated that lncTAM34a overexpression was sufficient to rescue miR34a expression even in the absence of TP53 (**Paper II**, Fig. 3a). miR34a is known for

its role in the regulation of multiple oncogenesis-related phenotypes, such as cell cycle and cell growth, among others. To further confirm the functionality of lncTAM34a-mediated miR34a expression in a TP53-null background, we examined cell cycle (**Paper II**, Fig. 3b, Supplementary Fig. 4b), miR34a-mediated decrease of cell cycle regulators (**Paper II**, Supplementary Fig. 4b-4c), and cell growth in this system (**Paper II**, Fig. 3c, Supplementary Fig. 5a-5b). In summary, these results indicate that increased expression of lncTAM34a in a TP53-null background is sufficient to upregulate miR34a and lead to known miR34a induction phenotypes. Finally, these results (**Paper II**, Fig. 3c), together with others (**Paper II**, Fig. 2e), indicate that lncTAM34a-mediated miR34a regulation is especially crucial to drive the appropriate cellular responses when cells encounter stress conditions.

We next asked at what level lncTAM34a regulates miR34a. Due to the fact that several of the cell lines that were engineered to stably overexpress lncTAM34a had an undetectable level of miR34a previous to lncTAM34a overexpression, we hypothesized that the regulation level was transcriptional. Therefore, we began by performing phosphorylated polymerase II chromatin immunoprecipitation (ChIP) at the miR34a promoter. Importantly, primers detecting phosphorylated polymerase II enrichment were located outside of the cloned and overexpressed lncTAM34a region. The results showed that overexpression of lncTAM34a increased phosphorylated polymerase II binding at the miR34a promoter and therefore indicate that lncTAM34a regulates miR34a at the transcriptional level (**Paper II**, Fig. 3d).

Finally, we utilized RNAseq data from The Cancer Genome Atlas, comprised of 17 different cancer types, to understand the association between lncTAM34a expression and survival. Our results indicate that there is indeed an association between decreased lncTAM34a expression and decreased survival in multiple cancer types (**Paper II**, Fig. 4a-4b). Despite the fact that this result does not implicate any causal relationship, we believe that it provides a basis for further investigation using controlled trials.

In summary, we identify and characterize lncTAM34a finding it to positively regulate miR34a transcription in both TP53 wild type and deficient cells. Although previous studies have used various molecular biology methods to upregulate miR34a expression, this is, to our knowledge, the first time an endogenous method has been shown to be able to achieve this in the absence of TP53.

4.3 PAPER III: *Unsupervised cell interaction profiling based on multiplet RNA sequencing reveals major architectural differences between small intestinal and colonic epithelium*

In this work we describe Cell Interactions by Multiplet Sequencing (CIM-seq), a high throughput, hypothesis free, intuitive, and easily implemented method to interrogate global cell-cell interactions within a tissue. CIM-seq relies on RNA sequencing of incompletely disassociated cells (multiplets) that are a common by-product in single cell RNAseq experiments. With CIM-seq we repurpose these multiplets to give us information concerning the cell-cell interactions in the intact tissue and simultaneously use fully dissociated single cells (singlets) to procure single cell resolution RNAseq data. This allows us not only to deduce the global map of cell-cell interactions in the tissue but also relate gene expression changes to specific cell-cell interactions without a predefined hypothesis. We subsequently utilize CIM-seq to explore cell-cell interactions within the architecture of mouse colonic crypts.

CIM-seq works by first performing an incomplete dissociation of the target tissue and subsequent FACS sorting of single cells and multiplets separately (**Paper III**, Fig. 1a). Both singlets and multiplets are then RNA sequenced and the data is analysed individually. The singlets are used to form a blueprint of the cell types in the tissue via unsupervised graph-based classification methods. Given the set of transcriptional profiles corresponding to the different cell types and an estimate of cell numbers in each multiplet, the multiplets are deconvoluted, resulting in a fractional contribution of each cell type for all of the multiplets under consideration. Finally, in order to gain an understanding of which interactions are overrepresented in the data, we calculate an enrichment score (observed / expected) and probability for each of the observed cell-cell interaction types.

In order to verify the assumptions necessary for the CIM-seq method to work, we began by examining the propensity of cells to re-associate after incomplete dissociation, which would cause the detection of cell-cell interactions that are not representative of interactions present in the tissue. Our results indicate that singlet reassociation was less than 0.5% after 2 hours (**Paper III**, Extended Data Fig. 1a). By visually examining cells after incomplete dissociation we found that the majority of multiplets were comprised of two cells bound together (**Paper III**, Fig. 1b). To test CIM-seq in a controlled setting, we next sequenced singlets and multiplets of a known composition from three cell lines. By examining the fraction of ERCC reads at different known cell counts we could show that they provide a reasonable proxy for cell number and correspond well to the known number of cells in the multiplets (**Paper III**, Fig. 1c-1d). The singlets were then subjected to graph-based classification to distinguished the different cell types and thus provided a blueprint for the deconvolution algorithm (**Paper III**, Extended Data Fig. 1c). Finally, the deconvolution revealed that CIM-seq was capable of accurately recovering the expected connections with an average misclassification rate of < 5% in all of the examined cell compositions (**Paper III**, Fig. 1e, Extended Data Fig. 1d-1e).

Next, we utilized mouse SI to evaluate the performance of CIM-seq in a complex tissue. SI stem cells reside at the base of the intestinal crypts and maintain contact with paneth cells that provide Wnt signaling and, thus, facilitate the stem cells ability to maintain their stemness characteristics. We tested the ability of CIM-seq to detect this previously known cell-cell interaction in an unsupervised manner. As such we sorted, sequenced, and classified 1214 single cells from the SI epithelium. Classification of these singlets revealed previously reported cell types and states known to exist in the tissue (**Paper III**, Fig. 2a). Due to the fact that the validity of the classifications is essential to the interpretation of the results from the CIM-seq algorithm, multiple additional steps were taken to verify the soundness of these classifications. In summary, the results indicated that the classification procedure successfully identified *bona fide* cell types and states with differential gene expression in all cases (**Paper III**, Extended Data Fig. 2a-2b). Subsequent deconvolution of 451 multiplets isolated from the same suspensions as the singlets, revealed the frequencies of cell types detected in the multiplets to strongly correspond to the cell type frequencies in singlets (**Paper III**, Fig. 2b). Enrichment analysis showed that cell types that are known to be equally distributed throughout the crypt were rich in connections but none of these were significantly enriched (**Paper III**, Fig. 2c). On the other hand, paneth and stem cells showed a highly enriched connection reflecting the previously known SI crypt architecture. RNA *in situ* hybridization (ISH) verified the adjacency of Lgr5+ stem cells and Lyz1+ paneth cells *in vivo* (**Paper III**, Fig. 2d). In summary, these results indicate that CIM-seq is capable of accurately detecting known cell-cell interactions and functions as expected in a complex tissue.

Although the colon has a similar crypt structure as the SI, it lacks villi and its crypt architecture is not as well defined. To gain a better understanding of the similarities and differences between the two tissues, we sequenced 2462 single cells isolated from the mouse colon. Our results showed a larger and more diverse goblet cell population in the colon than had been observed in the SI (**Paper III**, Fig. 3a) with two of these classes expressing the wound-healing marker Plet1 (**Paper III**, Extended Data Fig. 3a). Deconvolution of 1703 multiplets showed a distinct interaction pattern when comparing Plet1+ and Plet1- goblet cells (**Paper III**, Fig. 3c). Whereas Plet1+ goblet cells had a preferential interaction with the most highly Lgr5 expressing stem cells, Plet1- preferred to interact with stem cells located further along the differentiation trajectory. ISH for Lgr5 and Plet1 showed that Plet1+ goblet cells adjacent to Lgr5+ stem cells (**Paper III**, Fig. 3d). Finally, quantification of Plet1 along the longitudinal crypt axis revealed Plet1+ goblet cells to be localized at the base of the crypt and their distribution to mirror that of Lyz1+ paneth cells in the SI (**Paper III**, Fig. 3e). Collectively, these results indicate that CIM-seq identified a novel cell-cell interaction between Plet1+ goblet cells and Lgr5+ stem cells in the colon stem cell niche.

Although the dogma regarding paneth cells as the main source of stemness signaling in the SI is generally established, there are multiple competing theories regarding how this takes place in the colon. One theory postulates that there is a paneth cell equivalent that resides in the colon whereas another claims that the stroma is largely or entirely responsible for providing the necessary environment to facilitate stemness. We wanted to investigate the possibility that Plet1+ cells are a source of stemness signaling in the colon and began by searching our singlet expression data for expression of classical stemness ligands. Our results showed that both paneth cells as well as the stroma are responsible for stemness signaling in the SI, in agreement with previous results (**Paper III**, Fig. 4a-4b, Degirmenci et al., 2018; Shoshkes-Carmel et al., 2018; Valenta et al., 2016). In the colon, we also identified stemness-signaling originating in the stroma, although the role of the epithelial compartment is not as clear. Stemness factors can be seen to originate from several different colonic epithelial cells, including Plet1+ goblet cells, but additional experiments would be needed to determine if the stemness factors originating from these cell types are essential for the maintenance of colonic stem cells (**Paper III**, Fig. 4a-4b). Thus, our results indicate that Plet1+ goblet cells are unlikely to be a specific source of stemness signaling in the colon.

Interestingly, Plet1 has been previously shown to be important both in wound healing in the intestinal epithelium and in cell migration. We detected high expression levels of genes involved in cell-cell adhesion, tissue organization, inflammation, and cellular signaling in Plet1+ goblet cells (**Paper III**, Extended Data Fig. 3a). To further investigate Plet1+ goblet cells role in wound healing and maintenance of tissue architecture, we induced epithelial injury via dextran sodium sulfate (DSS) treatment and subsequently analysed the expression of Plet1. Our results indicated an increased expression of Plet1 at areas of DSS induced erosion (**Paper III**, Fig. 4c). In summary, these results indicate that Plet1+ goblet cells are an important component in wound healing and maintenance of intestinal tissue integrity.

In conclusion, we developed CIM-seq, a high throughput method for interrogating cell-cell interactions. Analysis of the mouse colon identified a cell-cell interaction between a subset of Plet1+ goblet cells and stem cells with additional results indicating that the interaction may be crucial for the maintenance of tissue architecture. We believe that CIM-seq will be capable of addressing a wide array of scientific questions and shed light on currently unknown cell-cell interactions giving rise to specific changes in cell identity.

5 DISCUSSION

This section includes some project-specific discussions, "behind the scenes" details, and successes and failures encountered during the realization of the projects described in the thesis. It is not intended to be an in-depth discussion of each project as a whole but, instead, select things that I have found to be interesting or challenging, points that I would like to highlight, or reflections since the project was completed.

5.1 PAPER I: *ClusterSignificance: a bioconductor package facilitating statistical analysis of class cluster separations in dimensionality reduced data*

The separation score function. For much of the development phase of ClusterSignificance the function to calculate the separation score was different than that which was used in the final publication. The original function basically calculated the number of correct classifications minus the incorrect classifications at each possible projected separation. In the course of the algorithms development, some testing revealed strange results that we could not explain immediately. After some further investigation, these results seemed to stem from the fact that, in the specific test examples we were using, the class sizes were very different and this was causing false positive results. While searching for a solution I came across the Song et al. paper that described a calculation that seemed as though it would resolve this issue. In addition, the new calculation had the added advantage of being based on a receiver operating characteristic (ROC) curve which many people are already familiar with. We implemented the complement of the Song et al. equation (**Equation 1**) as the new score calculation and it proved to work as well as we had hoped.

Equation 1. $$1 - \sqrt{(1 - specificity)^2 + 1 - sensitivity^2}$$

Quantifying a separation. With the new score calculation it was also possible to describe an effect size, in the form of area under the curve (AUC), for a specific separation. Although we became aware of this almost immediately, we were hesitant to implement it in the software. One of the criticisms typically leveled at ClusterSignificance is, just because a separation is significant does not mean that it has any biological meaning. Typically, in other types of quantification-based analysis, people use an effect size to estimate if something that is statistically significant actually has any biological effect. Despite the fact that, in my humble opinion, setting arbitrary cutoffs is less than optimal, it is typical operating procedure due to the lack of a better alternative and, in most cases serves, to limit the false positives discovered in the analysis. Specifically, we were aware that there were very few instances that the AUC could actually tell the users anything realistic about whether or not a result was biologically meaningful. Therefore, we believed that providing a metric that quantified separation would most often lead to its misinterpretation and decided it was best to not include it. Nevertheless, during the review process the issue of effect size and biologically meaningful separations was brought up and we were persuaded to include the AUC in the output of the separation scoring procedure.

Confirmatory not exploratory. There are several elements concerning the use of ClusterSignificance that, in hindsight, I believe could have been highlighted better in the

final publication. The first is that ClusterSignificance is intended to be a confirmatory method not an exploratory method and is designed, and tested to aid in the confirmation of a preconceived hypothesis. Due to this, using it as an exploratory method may give inaccurate results. Somewhat related is that, it is not valid to use the gene expression data itself to define the classes. The best way that I have found to explain why this is the case is because we are formally testing the independence of the labels and the features, if the features define the labels we basically already know the result without running the test.

5.2 PAPER II: *An antisense RNA capable of modulating the expression of the tumor suppressor microRNA-34a*

A lesson in determination. The lncTAM34a project was the longest running project included in this thesis. I actually started working on it as a master's thesis project and didn't publish it until the later stages of the PhD. Most of the initial identification and characterization of the transcript itself went very quickly and even the analysis concerning the TP53-mediated control over transcription was straightforward. The functional characterization, on the other hand, was much more time consuming. I felt that there were two things that caused a lot of difficulty. One was that we were often plagued by was the lack of an experimental method that could efficiently measure the expression of lncTAM34a in some of the experimental systems we used. Although lncTAM34a is expressed at a reasonable level in many TP53 wild-type systems, the levels are almost non-existent in TP53 null cell lines. This meant that, when working in the TP53 null systems where lncTAM34a was overexpressed, we were continuously trying to measure extremely minute levels of lncTAM34a in our controls. This led to us working very close to the Q-PCR detection limit and made each experiment extremely sensitive to small deviations of the protocol. At one point we tried to utilize digital PCR to address this problem and, although it did not work well for this issue specifically, it did confirm that our control cells were averaging much less than one lncTAM34a transcript per cell. It was necessary for us to set strict quality control guidelines for the Q-PCR experiments and they were often repeated additional times to gain more confidence in our results.

This missing link. The second aspect of the project that made deducing lncTAM34a function difficult was the "missing link", i.e. the details concerning the mechanism by which lncTAM34a functions. I would estimate that some 50% of the time that the project was ongoing we were focused on identifying the mechanism that lncTAM34a used to regulate miR34a expression. We already knew that lncTAM34a positively regulated miR34a expression and, as well, that it was mediated through polymerase II recruitment. Despite this, we didn't know the details of how this was occurring. Was lncTAM34a binding directly to the DNA and recruiting polymerase II? Was it acting as a scaffold that enabled the binding of other transcription factors that in turn recruited polymerase II? Were we dealing with some new and yet unreported mechanism by which lncRNAs can regulate gene transcription? This felt like the golden key to unlock the rest of the mystery and complete the story and I remained very determined to find it for a very long time. Nevertheless, despite lots of time, money, and bloodshed we never did manage to discover these details. Mostly this was due to both various technical difficulties and a lack of a more high throughput analysis that we had faith in to help resolve the question. I firmly believe that discovering this missing link would have shed light on a lot of the "head scratchers" we had experienced during the project. One criticism of some of the results might be the relatively small effect size that is seen in some of the experiments. This can be explained in several ways. One would be that lncTAM34a is primarily important when the cell is encountering stress and, therefore, in all cases where we did not induce stress, the effects of

lncTAM34a modulation on miR34a are minimal. Another possibility is that lncTAM34a is a fine-tuning regulator and that its role is primarily to serve as one of several factors that regulate the strength of the miR34a response. Another explanation would be that the factors that allow lncTAM34a to regulate miR34a expression are not present, or not induced, in those systems where we overexpress lncTAM34a and thus create a bottleneck. Having discovered the mechanism by which lncTAM34a functions would not only have allowed us to provide a more complete account for this lncRNA but also helped us to understand more about our other observations and potentially, establish better systems to examine the function of lncTAM34a in future experiments.

5.3 PAPER III: *Unsupervised cell interaction profiling based on multiplet RNA sequencing reveals major architectural differences between small intestinal and colonic epithelium*

Need for speed. Since this project is still, to some extent, underway, I don't have as much benefit of hindsight as I do with the other projects and it is somewhat harder to reflect on it in the same way. Despite this, I do think that there are a few interesting things that haven't really found their place in any of the text that would be worthwhile to bring up. One thing about CIM-seq that will not be apparent to anyone reading the paper is that it takes some time to run. Despite having worked extensively to improve the algorithms efficiency, to the best of our abilities, it still borderlines on reasonably slow. How slow? Well, run times are strongly dependent on the input parameters. The number of classes and features, the number of synthetic multiplets and permitted optimization iterations, as well as, the number of multiplets to deconvolute all affect runtimes. I routinely ran all of our deconvolutions on the Uppmax supercomputer at Uppsala University and would utilize around 100 cores per run. Under these circumstances, all analyses completed in less than 24 hours but that doesn't really help the guy who wants to run CIM-seq on his Macbook Pro. One of the post publication project goals is to work further, perhaps in collaboration, to find additional ways to reduce the algorithms speed and make it more accessible to a wider variety of users.

Why empirical? One of the reasons that the algorithm takes so much time is that the cost function that is optimized during the deconvolution is fairly computationally complex. This is due to the fact that we empirically model the multiplet values using the blueprint, formed by the singlets, in order to estimate the fraction of contribution for each possible cell type. There are theoretically much easier, and faster, ways to do this and, in fact, we have experimented with multiple cost functions over the course of CIM-seqs development. One of the early cost functions used the mean gene expression of each cell type defined in the blueprint. After adjusting these mean gene expression values for each cell type using the candidate solution and calculating the mean for each gene, we calculated the sum of differences between these values and the real multiplet gene expression values. This sum of differences was the cost that was being minimized by the particle swarm optimization algorithm. This was blazing fast and performed fairly well with the sorted cell line multiplets but we ultimately found that it was not sufficient when utilized in a complex tissue. This led us to the conclusion that, instead of trying to find a better metric that would apply well to all genes, it would be better to empirically model the multiplets using the singlet gene expression values. This led to the current cost calculation that was found to perform much better in real world applications but had a larger time penalty.

Cell types, states, and trajectories. During the development of CIM-seq there was one consideration that arose repeatedly and it took us some time to settle how we wanted to deal

with it. CIM-seq is developed in such a way that it requires discrete cell types to be classified before the deconvolution can take place and this has both pros and cons associated with it. Humans tend to like to classify things into individual categories in order to simplify problems but this is often at the expense of appreciating the complex reality of the situation. Cell types have traditionally been viewed as discrete entities although; in part by the advent of scRNAseq, this viewpoint is being questioned more and more often. It has in fact been argued that cell identity exists on a manifold where some paths are more easily traveled than others but all space on the manifold can theoretically be occupied. In this case, different end-states could be viewed as discrete entities but cell types as a whole are connected to each other via continuous trajectory gradients. In many tissues where CIM-seq could be utilized, we would expect to find a mix of both discrete end-points (i.e. cell types) and continuous trajectories (i.e. differentiating cells or cell states) and therefore, classifying all of these as discrete cell types is sub-optimal. During the deconvolution stage, the fractional contribution vector is the length of the number of cell types in the blueprint and, hence, we constrain the problems solution into a discrete space when, in fact, some of the discrete variables may be better represented simultaneously as a continuous process. One possible solution to this would be, to not use the classification in the deconvolution but instead consider each singlet as an individual cell type and then, post-deconvolution, determine which class, or where on the cell type manifold, that cell is. Although this may better match the present view concerning how cell types exist on a manifold, it is currently infeasible due to the computational complexity of the problem and the difficulty of interpretation of the output by the user. In practice we believe that these considerations will not have a drastic effect on the interpretation of the results and, in addition, feel as though the current algorithm makes reasonable compromises between theoretical accuracy and enabling user interpretation.

6 CONCLUDING REMARKS

Despite the widespread use of dimensionality reduction algorithms, their output is typically solely judged on a subjective basis via visualization. In **Paper I** we develop ClusterSignificance, a method and software that facilitates the analysis of statistical significance of class separations in dimensionality reduced data. We subsequently utilize ClusterSignificance to identify a role for lncRNAs in the identity of multiple hematological malignancies.

miR34a is a well characterized tumor suppressor that is a direct downstream target of TP53 and a master regulator of tumor suppression. The anti-tumor effects of increased miR34a expression are so widespread and ubiquitous that it is one of the very few miRNA that have been involved in clinical trials via miRNA replacement therapy. miR34a has been seen to be dysregulated in a multitude of tumor types both hematological and solid in nature, although the underlying mechanisms causing this are largely uninvestigated. In order to gain a deeper understanding of mechanisms that regulate miR34a, **Paper II** characterizes lncTAM34a, a lncRNA located in the antisense orientation to miR34a. Our results indicate that lncTAM34a positively regulates miR34a transcriptionally via recruitment of polymerase II. Our results further indicate that this regulation is especially crucial in contexts of cellular stress such as those routinely experienced during tumor development.

scRNA-seq has provided an unprecedented insight into the myriad of cell types present in various multicellular organisms. Efforts such as The Human Cell Atlas (Regev et al., 2018) will surely uncover additional levels and complexity while providing a broad overview of cell identity. Despite this high-resolution glimpse at the manifold on which cell types can exist, it remains challenging to attribute functions to these as they are discovered. In **Paper III** we develop CIM-seq, a high throughput hypothesis-free method to profile cell-cell interactions. Analysis of the scRNAseq data from mouse colonic crypts identified two novel goblet cell subsets with specific expression of the Plet1 gene. Subsequent utilization of CIM-seq to profile cell-cell interactions in the colonic crypts identified a previously unknown interaction in the stem cell niche between stem cells and Plet1+ goblet cells with further investigation indicating the potential role of these goblet cells in the maintenance of tissue architecture. In summary, CIM-seq is a widely applicable method facilitating high throughput cell-cell interaction profiling and is currently the only method that can relate cell-cell interactions to specific changes in gene expression without a previous hypothesis.

Despite the fact that, in the scope of this work, CIM-seq is not utilized to examine cell-cell interactions in tumors or the tumor microenvironment, hopefully the development and characterization of the algorithm completed thus far will aid in such analyses at a future time.

7 ACKNOWLEDGEMENTS

To my friends, family, and loved ones; to those that have collaborated with me, mentored, helped, or inspired me; to those who have taken the time to show an interest in, thoroughly examine, or offer advice concerning my scientific work, thank you all.

8 REFERENCES

Ab Wahab, M.N., Nefti-Meziani, S., Atyabi, A., 2015. A Comprehensive Review of Swarm Optimization Algorithms. PLoS ONE 10. https://doi.org/10.1371/journal.pone.0122827

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P., 2002. Organogenesis and the Patterning of Appendages. Mol. Biol. Cell 4th Ed.

Andrea Lancichinetti, Santo Fortunato, 2010. Community detection algorithms: a comparative analysis. arXiv.

Ardekani, A.M., Naeini, M.M., 2010. The Role of MicroRNAs in Human Diseases. Avicenna J. Med. Biotechnol. 2, 161–179.

Arias, A.M., 2001. Epithelial Mesenchymal Interactions in Cancer and Development. Cell 105, 425–431. https://doi.org/10.1016/S0092-8674(01)00365-8

Balas, M.M., Johnson, A.M., 2018. Exploring the mechanisms behind long noncoding RNAs and cancer. Non-Coding RNA Res. 3, 108–117. https://doi.org/10.1016/j.ncrna.2018.03.001

Beg, M.S., Brenner, A.J., Sachdev, J., Borad, M., Kang, Y.-K., Stoudemire, J., Smith, S., Bader, A.G., Kim, S., Hong, D.S., 2017. Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors. Invest. New Drugs 35, 180–188. https://doi.org/10.1007/s10637-016-0407-y

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 2008, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boisset, J.-C., Vivié, J., Grün, D., Muraro, M.J., Lyubimova, A., Oudenaarden, A. van, 2018. Mapping the physical network of cellular interactions. Nat. Methods 15, 547–553. https://doi.org/10.1038/s41592-018-0009-z

Codeluppi, S., Borm, L.E., Zeisel, A., Manno, G.L., Lunteren, J.A. van, Svensson, C.I., Linnarsson, S., 2018. Spatial organization of the somatosensory cortex revealed by osmFISH. Nat. Methods 15, 932–935. https://doi.org/10.1038/s41592-018-0175-z

Cohen, S., Elliott, G.A., 1963. The Stimulation of Epidermal Keratinization by a Protein Isolated from the Submaxillary Gland of the Mouse**From the Department of Biochemistry and the Department of Pathology, Vanderbilt University School of Medicine, Nashville, Tennessee. J. Invest. Dermatol. 40, 1–5. https://doi.org/10.1038/jid.1963.1

Crosetto, N., Bienko, M., van Oudenaarden, A., 2015. Spatially resolved transcriptomics and beyond. Nat. Rev. Genet. 16, 57–66. https://doi.org/10.1038/nrg3832

Degirmenci, B., Valenta, T., Dimitrieva, S., Hausmann, G., Basler, K., 2018. GLI1-expressing mesenchymal cells form the essential Wnt-secreting niche for colon stem cells. Nature 558, 449–453. https://doi.org/10.1038/s41586-018-0190-3

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R., 2012. Landscape of transcription in human cells. Nature 489, 101–108. https://doi.org/10.1038/nature11233

Egeblad, M., Nakasone, E.S., Werb, Z., 2010. Tumors as Organs: Complex Tissues that Interface with the Entire Organism. Dev. Cell 18, 884–901. https://doi.org/10.1016/j.devcel.2010.05.012

Fortunato, S., Barthélemy, M., 2007. Resolution limit in community detection. Proc. Natl. Acad. Sci. 104, 36–41. https://doi.org/10.1073/pnas.0605965104

Gawad, C., Koh, W., Quake, S.R., 2014. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proc. Natl. Acad. Sci. 111, 17947–17952. https://doi.org/10.1073/pnas.1420822111

Giannotta, M., Trani, M., Dejana, E., 2013. VE-Cadherin and Endothelial Adherens Junctions: Active Guardians of Vascular Integrity. Dev. Cell 26, 441–454. https://doi.org/10.1016/j.devcel.2013.08.020

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R.B., Vijver, M.J. van de, Sukumar, S., Chang, H.Y., 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076. https://doi.org/10.1038/nature08975

Hanahan, D., Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. Cell 144, 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hastie, T., Stuetzle, W., 1989. Principal Curves. J. Am. Stat. Assoc. 84, 502–516. https://doi.org/10.1080/01621459.1989.10478797

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441. https://doi.org/10.1037/h0071325

Hua, L., Wang, C.-Y., Yao, K.-H., Chen, J.-T., Zhang, J.-J., Ma, W.-L., 2015. High expression of long non-coding RNA ANRIL is associated with poor prognosis in hepatocellular carcinoma. Int. J. Clin. Exp. Pathol. 8, 3076–3082.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945. https://doi.org/10.1038/nature03001

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods 11, 163–166. https://doi.org/10.1038/nmeth.2772

Johnsson, P., Ackley, A., Vidarsdottir, L., Lui, W.-O., Corcoran, M., Grandér, D., Morris, K.V., 2013. A pseudogene long-noncoding-RNA network regulates *PTEN* transcription and translation in human cells. Nat. Struct. Mol. Biol. 20, 440–446. https://doi.org/10.1038/nsmb.2516

Johnsson, P., Lipovich, L., Grandér, D., Morris, K.V., 2014. Evolutionary conservation of long noncoding RNAs; sequence, structure, function. Biochim. Biophys. Acta 1840, 1063–1071. https://doi.org/10.1016/j.bbagen.2013.10.035

Junttila, M.R., de Sauvage, F.J., 2013. Influence of tumour micro-environment heterogeneity on therapeutic response. Nature 501, 346–354. https://doi.org/10.1038/nature12626

Kamińska, K., Szczylik, C., Bielecka, Z.F., Bartnik, E., Porta, C., Lian, F., Czarnecka, A.M., 2015. The role of the cell–cell interactions in cancer progression. J. Cell. Mol. Med. 19, 283–296. https://doi.org/10.1111/jcmm.12408

Kanasty, R., Dorkin, J.R., Vegas, A., Anderson, D., 2013. Delivery materials for siRNA therapeutics. Nat. Mater. 12, 967–977. https://doi.org/10.1038/nmat3765

Kao, S.C., Fulham, M., Wong, K., Cooper, W., Brahmbhatt, H., MacDiarmid, J., Pattison, S., Sagong, J.O., Huynh, Y., Leslie, F., Pavlakis, N., Clarke, S., Boyer, M., Reid, G., van Zandwijk, N., 2015. A Significant Metabolic and Radiological Response after a Novel Targeted MicroRNA-based Treatment Approach in Malignant Pleural Mesothelioma. Am. J. Respir. Crit. Care Med. 191, 1467–1469. https://doi.org/10.1164/rccm.201503-0461LE

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., Taipale, J., 2012. Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods 9, 72–74. https://doi.org/10.1038/nmeth.1778

Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., Miyano, S., Mori, M., 2011. Long Noncoding RNA HOTAIR Regulates Polycomb-Dependent Chromatin Modification and Is Associated with Poor Prognosis in Colorectal Cancers. Cancer Res. 71, 6320–6326. https://doi.org/10.1158/0008-5472.CAN-11-1021

Lee, G.L., Dobi, A., Srivastava, S., 2011. Diagnostic performance of the PCA3 urine test. Nat. Rev. Urol. 8, 123–124. https://doi.org/10.1038/nrurol.2011.10

Li, R., Zhu, H., Luo, Y., 2016. Understanding the Functions of Long Non-Coding RNAs through Their Higher-Order Structures. Int. J. Mol. Sci. 17. https://doi.org/10.3390/ijms17050702

Lin, L., Gu, Z.-T., Chen, W.-H., Cao, K.-J., 2015. Increased expression of the long non-coding RNA ANRIL promotes lung cancer cell metastasis and correlates with poor prognosis. Diagn. Pathol. 10, 14. https://doi.org/10.1186/s13000-015-0247-7

Lopez-Garcia, C., Klein, A.M., Simons, B.D., Winton, D.J., 2010. Intestinal Stem Cell Replacement Follows a Pattern of Neutral Drift. Science 330, 822–825. https://doi.org/10.1126/science.1196236

Maaten, L. van der, Hinton, G., 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Mahmoudi, S., Henriksson, S., Corcoran, M., Méndez-Vidal, C., Wiman, K.G., Farnebo, M., 2016. Wrap53, a Natural p53 Antisense Transcript Required for p53 Induction upon DNA Damage. Mol. Cell 64, 1009. https://doi.org/10.1016/j.molcel.2016.11.027

Marusyk, A., Almendro, V., Polyak, K., 2012. Intra-tumour heterogeneity: a looking glass for cancer? Nat. Rev. Cancer 12, 323–334. https://doi.org/10.1038/nrc3261

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat.

Morrison, S.J., Scadden, D.T., 2014. The bone marrow niche for haematopoietic stem cells. Nature 505, 327–334. https://doi.org/10.1038/nature12984

Mouraviev, V., Lee, B., Patel, V., Albala, D., Johansen, T.E.B., Partin, A., Ross, A., Perera, R.J., 2016. Clinical prospects of long noncoding RNAs as novel biomarkers and therapeutic targets in prostate cancer. Prostate Cancer Prostatic Dis. 19, 14–20. https://doi.org/10.1038/pcan.2015.48

Niknafs, Y.S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., Iyer, M.K., Pitchiaya, S., Malik, R., Hosono, Y., Prensner, J.R., Poliakov, A., Singhal, U., Xiao, L., Kregel, S., Siebenaler, R.F., Zhao, S.G., Uhl, M., Gawronski, A., Hayes, D.F., Pierce, L.J., Cao, X., Collins, C., Backofen, R., Sahinalp, C.S., Rae, J.M., Chinnaiyan, A.M., Feng, F.Y., 2016. The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. Nat. Commun. 7, 1–13. https://doi.org/10.1038/ncomms12791

Oppermann, H., Levinson, A.D., Varmus, H.E., Levintow, L., Bishop, J.M., 1979. Uninfected vertebrate cells contain a protein that is closely related to the product of the avian sarcoma virus transforming gene (src). Proc. Natl. Acad. Sci. U. S. A. 76, 1804–1808.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., Louis, D.N., Rozenblatt-Rosen, O., Suvà, M.L., Regev, A., Bernstein, B.E., 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science 344, 1396–1401. https://doi.org/10.1126/science.1254257

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2, 559–572. https://doi.org/10.1080/14786440109462720

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R., 2014. Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. 9, 171–181. https://doi.org/10.1038/nprot.2014.006

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., Cao, X., Jing, X., Wang, X., Siddiqui, J., Wei, J.T., Robinson, D., Iyer, H.K., Palanisamy, N., Maher, C.A., Chinnaiyan, A.M., 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1 , an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. 29, 742–749. https://doi.org/10.1038/nbt.1914

Qi, P., Zhou, X., Du, X., 2016. Circulating long non-coding RNAs in cancer: current status and future perspectives. Mol. Cancer 15. https://doi.org/10.1186/s12943-016-0524-4

Redis, R.S., Sieuwerts, A.M., Look, M.P., Tudoran, O., Ivan, C., Spizzo, R., Zhang, X., de Weerd, V., Shimizu, M., Ling, H., Buiga, R., Pop, V., Irimie, A., Fodde, R., Bedrosian, I., Martens, J.W., Foekens, J.A., Berindan-Neagoe, I., Calin, G.A., 2013. CCAT2, a novel long non-coding RNA in breast cancer: expression study and clinical correlations. Oncotarget 4, 1748–1762.

Regev, A., Teichmann, S., Rozenblatt-Rosen, O., Stubbington, M., Ardlie, K., Amit, I., Arlotta, P., Bader, G., Benoist, C., Biton, M., Bodenmiller, B., Bruneau, B., Campbell, P., Carmichael, M., Carninci, P., Castelo-Soccio, L., Clatworthy, M., Clevers, H., Conrad, C., Eils, R., Freeman, J., Fugger, L., Goettgens, B., Graham, D., Greka, A., Hacohen, N., Haniffa, M., Helbig, I., Heuckeroth, R., Kathiresan, S., Kim, S., Klein, A., Knoppers, B., Kriegstein, A., Lander, E., Lee, J., Lein, E., Linnarsson, S., Macosko, E., MacParland, S., Majovski, R., Majumder, P., Marioni, J., McGilvray, I., Merad, M., Mhlanga, M., Naik, S., Nawijn, M., Nolan, G., Paten, B., Pe'er, D., Philippakis, A., Ponting, C., Quake, S., Rajagopal, J., Rajewsky, N., Reik, W., Rood, J., Saeb-Parsy, K., Schiller, H., Scott, S., Shalek, A., Shapiro, E., Shin, J., Skeldon, K., Stratton, M., Streicher, J., Stunnenberg, H., Tan, K., Taylor, D., Thorogood, A., Vallier, L., van Oudenaarden, A., Watt, F., Weicher, W., Weissman, J., Wells, A., Wold, B., Xavier, R., Zhuang, X., Committee, H.C.A.O., 2018. The Human Cell Atlas White Paper. ArXiv181005192 Q-Bio.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., Chang, H.Y., 2007. Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell 129, 1311–1323. https://doi.org/10.1016/j.cell.2007.05.022

Roy, S., Kim, D., Lim, R., 2017. Cell-cell communication in diabetic retinopathy. Vision Res., Diabetic Retinopathy - an Overview 139, 115–122. https://doi.org/10.1016/j.visres.2017.04.014

Sánchez, Y., Huarte, M., 2013. Long Non-Coding RNAs: Challenges for Diagnosis and Therapies. Nucleic Acid Ther. 23, 15–20. https://doi.org/10.1089/nat.2012.0414

Sato, T., van Es, J.H., Snippert, H.J., Stange, D.E., Vries, R.G., van den Born, M., Barker, N., Shroyer, N.F., van de Wetering, M., Clevers, H., 2011. Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. Nature 469, 415–418. https://doi.org/10.1038/nature09637

Saus, E., Brunet-Vega, A., Iraola-Guzmán, S., Pegueroles, C., Gabaldón, T., Pericay, C., 2016. Long Non-Coding RNAs As Potential Novel Prognostic Biomarkers in Colorectal Cancer. Front. Genet. 7. https://doi.org/10.3389/fgene.2016.00054

Shoshkes-Carmel, M., Wang, Y.J., Wangensteen, K.J., Tóth, B., Kondo, A., Massasa, E.E., Itzkovitz, S., Kaestner, K.H., 2018. Subepithelial telocytes are an important source of Wnts that supports intestinal crypts. Nature 557, 242–246. https://doi.org/10.1038/s41586-018-0084-4

Snippert, H.J., Flier, L.G. van der, Sato, T., Es, J.H. van, Born, M. van den, Kroon-Veenboer, C., Barker, N., Klein, A.M., Rheenen, J. van, Simons, B.D., Clevers, H., 2010. Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. Cell 143, 134–144. https://doi.org/10.1016/j.cell.2010.09.016

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P.I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., Frisén, J., 2016. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science 353, 78–82. https://doi.org/10.1126/science.aaf2403

Stark, R., Grzelak, M., Hadfield, J., 2019. RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656. https://doi.org/10.1038/s41576-019-0150-2

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., Chang, H.Y., 2010. Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. Science 329, 689–693. https://doi.org/10.1126/science.1192002

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R., Johnson, R., 2018. Towards a complete map of the human long non-coding RNA transcriptome. Nat. Rev. Genet. 19, 535–548. https://doi.org/10.1038/s41576-018-0017-y

Valenta, T., Degirmenci, B., Moor, A.E., Herr, P., Zimmerli, D., Moor, M.B., Hausmann, G., Cantù, C., Aguet, M., Basler, K., 2016. Wnt Ligands Secreted by Subepithelial Mesenchymal Cells Are Essential for the Survival of Intestinal Stem Cells and Gut Homeostasis. Cell Rep. 15, 911–918. https://doi.org/10.1016/j.celrep.2016.03.088

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., Jiang, Z., Xu, J., Liu, Q., Cao, X., 2014. The STAT3-Binding Long Noncoding RNA lnc-DC Controls Human Dendritic Cell Differentiation. Science 344, 310–313. https://doi.org/10.1126/science.1251456

Wang, X., Allen, W.E., Wright, M.A., Sylwestrak, E.L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G.P., Bava, F.-A., Deisseroth, K., 2018. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. Science 361, eaat5691. https://doi.org/10.1126/science.aat5691

Wells, D.K., Chuang, Y., Knapp, L.M., Brockmann, D., Kath, W.L., Leonard, J.N., 2015. Spatial and Functional Heterogeneities Shape Collective Behavior of Tumor-Immune Networks. PLOS Comput. Biol. 11, e1004181. https://doi.org/10.1371/journal.pcbi.1004181

Williams, J.M., Duckworth, C.A., Burkitt, M.D., Watson, A.J.M., Campbell, B.J., Pritchard, D.M., 2015. Epithelial Cell Shedding and Barrier Function. Vet. Pathol. 52, 445–455. https://doi.org/10.1177/0300985814559404

World Health Organization, 2018. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018.

Yap, K.L., Li, S., Muñoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., Zhou, M.-M., 2010. Molecular Interplay of the Noncoding RNA ANRIL and Methylated Histone H3 Lysine 27 by Polycomb CBX7 in Transcriptional Silencing of INK4a. Mol. Cell 38, 662–674. https://doi.org/10.1016/j.molcel.2010.03.021

Yu, W., Gius, D., Onyango, P., Muldoon-Jacobs, K., Karp, J., Feinberg, A.P., Cui, H., 2008. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature 451, 202–206. https://doi.org/10.1038/nature06468

Zhang, D., Sun, G., Zhang, H., Tian, J., Li, Y., 2017. Long non-coding RNA ANRIL indicates a poor prognosis of cervical cancer and promotes carcinogenesis via PI3K/Akt pathways. Biomed. Pharmacother. 85, 511–516. https://doi.org/10.1016/j.biopha.2016.11.058

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H., 2017. Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8, 1–12. https://doi.org/10.1038/ncomms14049

Zuckerman, J.E., Davis, M.E., 2015. Clinical experiences with systemically administered siRNA-based therapeutics in cancer. Nat. Rev. Drug Discov. 14, 843–856. https://doi.org/10.1038/nrd4685