Functional genomics studies of cancer and cells of origin:

From pan-cancer to single-cell

Zhiyuan Hu

St Cross College

Nuffield Department of Medicine

University of Oxford

United Kingdom

A thesis presented for the degree of

Doctor of Philosophy in Clinical Medicine

February 19, 2020

Functional genomics studies of cancer and cells of origin:
From pan-cancer to single-cell

Zhiyuan Hu
St Cross College

A thesis presented for the degree of
Doctor of Philosophy in Clinical Medicine
Trinity Term 2019

Abstract

The development of stratification systems and therapeutical targets is critical for cancer research. Recent advances in high-throughput sequencing enable the generation of a large amount of cancer genomic and transcriptomic data, which can be used to answer the questions on cancer diagnosis and treatment. However, it is still a great challenge to interpret a mass of data and get reliable conclusions.

In this study, I used the functional genomics methodologies to study cancer and cells of origin. In the pan-cancer study, I developed an algorithm, Masonmd, to predict the targets of nonsense-mediated decay (NMD). This algorithm identified over 140 thousand NMD-elicit mutations from the genomic data of The Cancer Genome Atlas (TCGA) pan-cancer cohort. Further analysis showed that NMD was prevalent across cancer types and, more notably, it could play a permissive role in cancer progression. The results support the notion that the inhibition of NMD is a promising therapeutic strategy.

In the work on single-cell RNA sequencing (scRNA-seq), I first developed a differential-expression-based clustering algorithm, ClinCluster, tailored for clinical scRNA-seq data. The benchmarking analysis demonstrated that ClinCluster outperformed other commonly used pipelines in both simulated and real datasets. I next generated and analysed the scRNA-seq data from human fallopian tubes, which certain high-grade serous ovarian cancer (HGSOC) arose from. The analysis constructed a cell atlas of human fallopian tubes and identified four novel secretory subtypes. Based on the discovery of cell phenotypic repertoire, the deconvolution analysis of the TCGA data refined the molecular subtyping of HGSOC. This gave rise to a robust prognostic predictor, which was validated in eight independent expression datasets. Together, the evidence suggests that the phenotypic heterogeneity in cells of origin can be inherited by tumour cells, which affects patients' overall survival. It shed light on a refined molecular diagnosis for HGSOC.

Acknowledgements

First and foremost, I offer my sincere thanks to my supervisors, Professor Ahmed Ashour Ahmed and Dr Christopher Yau, who have been offering me generous support and invaluable suggestions throughout my DPhil. My vocabulary could not be enough to express even one percent of my gratitude. I thank the China Scholarship Council and Nuffield Department of Medicine for funding my studentship. I would like to thank St Cross College for supporting me over my course.

I am grateful to Professor Cecilia Lindgren and Professor Johann de Bono, who acted as the examiners of my viva, Professor Krina Zondervan, who acted as the assessor for my transfer of status, Dr Benjamin Fairfax, who acted as the assessor for my confirmation of status, and Dr Gareth Bond, who acted as the assessor for both, for giving up their time to familiarise themselves with my research. Their feedback was not only useful but highly encouraging.

I am also grateful for the support and help from all the present and past members of our lab. I would like to thank Mara Artibani and Donatien Chedom Fosto for their guidance upon my joining the lab. I would like to thank Kay Chong specifically for help polishing my thesis. I would like to offer my gratitude to my fellow students for mutual support over the past four years.

I also thank all my colleagues who have helped me. I would like to thank the WIMM Flow Cytometry Facility and the WIMM Single Cell Facility for their technical support.

Finally, I would like to thank my husband and my dearest friend, Yibin Liu, for his continued support and enjoyable company throughout my studies and my life.

Declaration

I declare that no part of this thesis has been accepted or submitted for any degree, diploma, certificate or other qualification in this University or elsewhere. The results presented herein are my own.

The work in Chapter 2 was published on *Nature Communications* in 2017 (listed below). The results presented in the thesis were after reanalysis. All figures presented in this thesis differ from the published ones. The work presented in Chapter 3 and Chapter 4 is related to the publication in *Cancer Cell* and a patent (listed below).

Publications

Z. Hu, M. Artibani, A. Alsaadi, N. Wietek, ..., T. Sauka-Spengler, C. Yau*, A. A. Ahmed*. (2020). The repertoire of serous ovarian cancer non-genetic heterogeneity revealed by single-cell sequencing of normal fallopian tube epithelial cells. *Cancer Cell* **37** (2), 226-242. e7. `https://doi.org/10.1016/j.ccell.2020.01.003`

Z. Hu, C. Yau* and A. A. Ahmed*. (2017). A pan-cancer genome-wide analysis reveals tumour dependencies by induction of non-sense mediated decay. *Nature Communications* **8**, 15943 (2017). doi: 10.1038/ncomms15943.

Patent

UK Patent Application No. 1902653.3 Ovarian Cancer Biomarkers. Filing date: 27 February 2019

Contents

List of Acronyms

AP affinity propagation.

ARI adjusted Rand index.

CDS coding sequence.

CNV copy number variation.

DE differential expression.

DEG differentially expressed gene.

ECM extracellular matrix.

EMT epithelial-mesenchymal transition.

FACS fluorescence-activated cell sorting.

FDR false discovery rate.

FTE fallopian tube epithelium.

FTESC fallopian tube epithelial secretory cell.

GEO Gene Expression Omnibus.

HGSOC high-grade serous ovarian cancer.

HVG high variance gene.

indel insertion and deletion.

IPV impact of inter-patient variability.

LT long-term.

MNN mutual nearest neighbours.

NGS next-generation sequencing.

NMD nonsense-mediated decay.

PCA principal component analysis.

PTC premature stop codon.

RNA-seq RNA sequencing.

RT reverse transcription.

scRNA-seq single-cell RNA sequencing.

SNVs single-nucleotide variants.

t-SNE t-distributed stochastic neighbour embedding.

TCGA The Cancer Genome Atlas.

TSGs tumour suppressor genes.

UTR untranslated region.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Cancer genomics

Cancer is the second leading cause of death, which claims millions of lives per year. It refers to a collection of heterogeneous and complex diseases characterised by runaway cell division and metastasis, in which cells disseminate from one site to another in the body. The current theory states that genetic aberrations contribute to tumorigenesis. This means that certain normal cells, termed cells of origin, acquire driver mutations and gradually evolve into tumours. The hallmarks of tumorigenesis and cancer progression have been comprehensively reviewed by Hanahan and Weinberg [62].

Decades ago, researchers described a connection between tumours and genes. One of the milestone studies was the discovery of the oncogene from avian sarcoma viruses [175]. Subsequently, the discovery of a pivot tumour suppressor was reported by multiple research groups in 1979 [44, 99, 112, 126, 171]. The name of this tumour suppressor was later unified as p53 [109], which is encoded by the gene *TP53*. *BRCA1* and *BRCA2*, two cancer-related genes closely related to breast cancer and ovarian cancer, were cloned for the first time in 1994 and 1995 respectively [127, 199]. These seminal studies, amongst many others, have laid the groundwork for subsequent cancer genomics studies.

Advances in sequencing techniques have enabled the profiling of numerous cancer genomes. This has solidified the notion that substantial heterogeneity is one of the most significant and intriguing features of cancer genomes. Genetic heterogeneity can be principally divided into two categories, intra-tumour genetic heterogeneity and inter-tumour genetic heterogeneity. Intra-tumour genetic heterogeneity refers to the coexistence of multiple genetic clones within one tumour, while inter-tumour genetic heterogeneity describes the genetic diversity across tumours from different patients. A common methodology used to study inter-tumour heterogeneity is collecting tumour samples from a cohort of patients and conducting whole-genome sequencing on these samples. A range of computational tools has empowered the detection of events in cancer genomes, such as single-nucleotide variants (SNVs), insertions/deletions, copy number variations (CNVs) and other chromosomal abnormalities [46].

A clinical application of studying inter-tumour heterogeneity is personalised medicine, whereby the genetic information from a patient is used to inform the choice of optimal therapeutic strategies for this individual. For instance, many ovarian cancer patients carrying *BRCA1* and *BRCA2* mutations respond well to the poly(ADP-ribose) polymerase (PARP) inhibitors [105]. Inter-tumour heterogeneity also manifests in mutation signatures, which are the decomposed mutation patterns of the nucleotide switching based on a nonnegative matrix factorization method [2, 67, 135]. Such patterns can reflect the underpinning mutation process of cancer development.

Although genomics studies have enhanced our understanding of cancer, it has evident limitations. Genomics studies are restricted to genomic sequence information, which provides limited insight into their biological functions. This hinders the interpretation of their functional significance. In other words, there is a gap between the genotype and the phenotype. Consequently, a complementary research field has been built to investigate the functions of genomic elements.

1.2 Functional genomics and transcriptomics

Understanding gene functions is the foundation for cancer research and many other fields. In contrast to the traditional method that studies the functions of genes one by one, functional genomic research intends to study the functions and interactions of genes in a genome-wide way fueled by a large amount of data [75].

Transcriptomics is a vital modality of functional genomics studies. In contrast to the static genome, the transcriptome is dynamic. Transcriptomic profiling can be likened to taking a snapshot of the cell phenotype or the cell state under a certain condition. Thus, transcriptomics approaches can provide insight into the regulation of gene expression and the relationship between cell phenotypes and expression profiles.

Functional genomics studies consist of two core elements, data and data mining. Accordingly, there are two strategic points of performing functional genomics studies for cancer research. The first is how to generate data or which data to be used (experimental techniques), and the second is how to extract useful information from a massive volume of data (informatics tools). These two points will be expounded in the next part.

1.2.1 Experimental techniques

The depiction of the cellular transcriptome means quantifying the level of transcripts that have been transcribed from a genome. Many methods have been developed to achieve this purpose. The invention of reverse transcription-quantitative polymerase chain reaction (RT-qPCR) enabled the relative quantification of RNA molecules encoded by certain genes [34, 65, 69]. The subsequently developed high-throughput techniques, such as the microarray and RNA sequencing (RNA-seq), have revolutionised the field of cancer research.

1.2.1.1 Microarrays and RNA-seq

Prior to the advent of RNA-seq, microarrays were widely used to profile transcriptomes [165]. The principle of microarrays is complementary base pairing. In a microarray experiment, mRNA is first reverse-transcribed into cDNA and then the cDNA target is marked with fluorescent probes. The cDNA sample is loaded onto a probe-attaching chip, where the cDNAs complementary to probes stay on the chip and the others are washed away. The fluorescent strength is used to calculate the expression of genes. However, this method is limited to measuring the expression levels of genes with at least a partially known sequence. Moreover, it is also affected by cross-hybridisation artefacts [101].

The invention of next-generation sequencing (NGS) has changed the game. It enables simultaneously profiling the expression levels of more than 20,000 protein-encoding genes as well as non-coding RNAs, depending on the protocol used. The main steps of RNA-seq include RNA isolation (and selection), cDNA synthesis, library preparation and NGS. It expands both the breadth and the depth of transcriptomic profiling. As the sequencing cost has decreased substantially, researchers have been frequently using RNA-seq to profile cancer transcriptomes. This technique also became a key component in large-scale projects of cancer research, such as Cancer Cell Line Encyclopedia [9] and The Cancer Genome Atlas [29].

1.2.1.2 Single-cell RNA sequencing

With the advent of single-cell RNA sequencing (scRNA-seq), researchers can obtain knowledge of the transcriptome from single cells. The blueprint of a cancer single-cell atlas is being achieved. scRNA-seq has been conducted on many cancer types to delineate the intra-tumour heterogeneity and the tumour microenvironment [35, 110, 143, 154, 157, 183].

Two commonly used techniques are the plate-based Smart-Seq2 method [147, 148] and the droplet-based 10x Genomics method.

Smart-Seq2 generates the single-cell sequencing library individually and pools all

libraries together before sequencing. To achieve this, the bulk sample is dissociated into single-cell suspension and the single cell is isolated by the fluorescence-activated cell sorting (FACS). One cell is sorted into one well, in which it undergoes a series of steps, including lysis, reverse transcription, first-round PCR, tagmentation reaction and second-round PCR.

10x Genomics is a commercial technique. It involves the special chip with capillary channels. The single-cell suspension combined with reverse transcription master mix, barcoded beads and oil are loaded into different wells on a chip, which is then put inside a box-like controller. In the controller, all the reagents are pushed through the capillaries to form aqueous droplets diffused in the oil. Each droplet contains ideally no more than one bead and one cell. The mixture will undergo the reverse transcription reaction, in which the cDNAs from each cell are labelled with a cell-specific barcode. Then all the cDNAs are pool together and undergo the pre-amplification and the steps of library preparation.

	Smart-Seq2	10x Genomics
Purchase	In-house	Commercial
Type	Plate-based	Droplet-based
Throughput	Low	High
Experimental length	Long	Short
Cost per cell	High	Low
Gene coverage	High	Low
Controllability	High	Low

Table 1.1. The pros and cons of two scRNA-seq techniques.

Although both techniques have been widely used for scRNA-seq, there are advantages and disadvantages (Table 1.1) [47]. The advantages of 10x Genomics are high throughput and short experimental length. The sequencing library of 2000 cells can be prepared in two or three days if 10x Genomics is used. The average cost per cell is also much lower than Smart-Seq2. However, 10x Genomics usually cov-

ers fewer genes and, unlike Smart-Seq2, the quality of each cell cannot be traced in the procedure since all the cells are pooled together from the beginning of the 10x Genomics protocol. Although 10x Genomics cannot sequence single cells as deeply as Smart-Seq2, it can robustly capture differentiated populations. There is a trade-off between cell number and read depth. Two recent preprints have quantified this trade-off in certain scenarios [169, 178], but the model they proposed cannot be generalised to all systems. Hence, the rule of thumb is to select the scRNA-seq technique based on the questions of interest.

1.2.2 Informatics resources and methodologies

1.2.2.1 Public databases

With the popularisation of sequencing techniques, many large-scale databases have sprung up. There are two major types of databases. One is the project-centred, such as The Cancer Genome Atlas (TCGA), and the other is the collection of data repositories, such as Gene Expression Omnibus (GEO) (`https://www.ncbi.nlm.nih.gov/geo/`).

TCGA is a cancer genomics program that has generated a comprehensive database containing genomic, transcriptomic, epigenomic, proteomic and clinical data from thousands of tumour samples [27, 28, 76]. Apart from the original studies, researchers have been reanalysing the data with novel techniques and refreshing our understanding of cancer [29, 70, 118].

On the other hand, GEO is a database repository that stores the gene expression data generated by researchers from all over the world. Anyone can submit their data, usually along with the submission of a manuscript, to GEO, which will be censored and curated before being incorporated into this platform. GEO has catalogued the expression data of millions of samples, not only limited to tumour samples. There are more repositories similar to GEO. The accumulation of previously published datasets serves as a powerful tool for comprehensive studies.

In this informatics era, the accessibility of databases provides researchers with an

opportunity of standing on the shoulders of giants. Many published datasets contain a huge amount of information that cannot be fully exploited in the original analysis. Reusing or reanalysing published datasets with new methods is, thus, a cost-efficient strategy that sometimes generates refreshing findings [45]. For instance, Berger et al. reanalysed the TCGA data of four gynaecological cancer types and breast cancer and found novel features that characterised Pan-Gyn tumours [108]. Another example is the combination of scRNA-seq data and TCGA data that was performed recently by Neftel et al. [131]. In this study, researchers found that a collection of tumour cell states coexisted in glioblastoma and revealed the connection between the heterogeneity of single-cell states and the tumour subtypes previously defined by TCGA. Another advantage of reanalysing published datasets is that it can push through the limitation of a single study. The sample size of individual studies is usually limited, resulting in low reproducibility of findings. Reanalysing the integrative data of multiple published small studies enables the development of more powerful conclusions [33].

Apart from those data repositories, there are many functional genomics databases documenting the annotations of gene or protein functions, such as NCBI RefSeq (`https://www.ncbi.nlm.nih.gov/`) and UniProt (`https://www.uniprot.org`), as well as pathway annotations, such as gene ontology, KEGG (`https://www.genome.jp/kegg/`) and MSigBD (`http://software.broadinstitute.org/gsea/index.jsp`) [176]. A handy integrative database is GeneCard (`https://www.genecards.org`). The curated annotations in these databases are mostly derived from previous publications and computational predictions. Those knowledge-based databases empower the interpretation of gene expression data. For example, the knowledge on the functions of marker genes in the scRNA-seq data can be used to infer the functions of previously unknown cell populations.

1.2.2.2 Computational methods

Miscellaneous computational tools have been developed to study cancer transcriptomics or functional genomics. These tools can be classified into five major groups [37].

The first type is *global analysis*, also termed as pan-cancer analysis. Motivated by the massive data generated from multiple cancer types, researchers use the global analysis to find the generality shared across cancer types and the peculiarity of each cancer type [29, 108]. In general, pan-cancer analyses are focused on inter-cancer-type heterogeneity or homogeneity.

Relative analyses intend to dig up the diversity within one single tumour type by classifying or scoring tumour samples of the same cancer type. The clinical utility of relative analyses includes identifying molecular subtypes and building prognostic predictors. Tothill et al. used microarrays to profile more than 200 ovarian tumours and identified four molecular subtypes of high-grade serous ovarian cancer by k-means clustering, in which the mesenchymal subtype was associated with poor prognosis [184]. Another example is the prognostic predictor (Prediction Analysis of Microarray 50, PAM50) of breast cancer. This 50-gene panel can be used to identify five breast cancer molecular subtypes, namely luminal A, luminal B, HER2-enriched, basal-like and normal-like, based on the expression levels of 50 genes and can predict the probability of tumour metastasis. PAM50 is also found to be able to provide insight into treatment efficacy [140]. These cases illustrate the clinical value of cancer transcriptomics studies.

While the inter-sample variability is studied by global or relative analyses, *differential analyses* are aimed to compare samples from the same patients, e.g. normal versus tumour and primary versus metastatic. This computational technique can be used to study the genomic processes involved in cancer progression. It is more about intra-patient not inter-sample heterogeneity.

Conversely, *compositional analyses* (or called *deconvolution analyses*) are tailored to study the intra-sample heterogeneity, i.e. the composition of a single tumour [202]. In the field of cancer research, deconvolution analysis aims to break down the complexity of cancer. Rather than a single population, tumours usually consist of multiple cell types, including tumour cells, tumour-infiltrated lymphocytes and cancer-associated fibroblasts [18]. Newman et al. developed a computational tool, named CIBERSORT, to impute the proportions of individual cell types in a bulk tissue based on the linear support vector regression [132]. Later, BSEQ-sc and other

methods have been developed to use the single-cell RNA-seq data as the reference for deconvolution [8, 198]. The increasing amount of scRNA-seq data has boosted such analyses by providing the transcriptomic signature of individual cell type within a tumour [8, 154, 198].

Last but not least, *integrative analyses* usually overlap with the above four methods. Integrative analyses involve the combined analysis of multiple data types, such as genomics, transcriptomes, epigenomics and clinical data [11, 76, 90], since the information from one single data type can be *ex parte*.

As an example of comprehensively using the above methodologies, a recent study applied single-cell DNA and RNA sequencing on both pre- and post-chemotherapy breast tumour samples [90]. By comparing the pre- and post-treatment samples, they found that the transcriptomes of tumour cells underwent reprogramming during treatment. Several pathways, such as epithelial-mesenchymal transition (EMT) and hypoxia, were upregulated in chemoresistant tumours after neoadjuvant chemotherapy. This work used both differential and integrative approaches to elucidate the molecular mechanism of chemoresistance at the single-cell level.

1.3 Aims and objectives

The focus of this thesis is to use the experimental and computational methods of functional genomics to address questions of cancer biology. The questions are related to cancer diagnosis and treatments.

Chapter 3 describes the usage of global technique to study the connection between cancer genomes and transcriptomes by studying a common type of loss-of-expression mutation. This project is aimed to assess the probability of using nonsense-mediated decay (NMD) as a therapeutical target.

Chapters 4 and 5 are related to the scRNA-seq technique. Chapter 4 aims to develop a clustering method tailored for the clinical scRNA-seq data with severe confounding factors. Chapter 5 will present the scRNA-seq data that dissects the human fallopian tubes, which are tissues of origin of ovarian cancer. This work will seek for

new subpopulations in this tissue. The last section of Chapter 5 will use the compositional, relative and integrative approaches to examine the composition of bulk ovarian tumours, refine the molecular classifier and test its prognostic value. This work is aimed to provide information to improve the diagnosis of ovarian cancer.

Chapter 2

Methods

2.1 The pan-cancer analysis of nonsense-mediated decay (NMD)

2.1.1 Data and data preprocessing

I downloaded the data (version 2016-12-29) of The Cancer Genome Atlas (TCGA) Pan-Cancer (PANCAN) cohort from the Xena (`https://xenabrowser.net`) [195], including the batch-effect normalised RNA-seq data (hg19), the somatic mutation (SNP and INDEL) MC3 data (hg19), the gene-level CNV data (hg19) and the phenotype data.

Before filtering, there were 2,845,316 somatic mutations in 8,784 samples from 32 cancer types. I firstly filtered out mutations in intronic regions (85809, 3%) or RNA regions (40381, 1.4%) that did not affect the coding sequences (CDSs). In total, 11455 (0.4%) mutations in 579 (2.9%) genes could not be mapped to Entrez IDs, while 2707671 (99.6%) mutations in 19548 (97.1%) genes were mapped to Entrez IDs. Those mutations without Entrez ID-annotated genes were filtered out. Then I filtered out 25412 mutations of which the genes had no CDS record, 7768 with the CDSs not ending with typical stop codons, 235895 located at 3'untranslated region (UTR), 63756 located at 5'UTR, 14638 located at 3' Flank and 8496 located at 5' Flank. The after-filtering set (including silent mutations) had 2,351,706 somatic

mutations. After excluding silent mutations, there were 1,706,218 somatic SNVs and indels in the coding region.

2.1.2 Prediction of NMD-elicit mutations

I developed an algorithm to predict the NMD-elicit mutations, termed as Masonmd (MAke Sense Of NMD). This algorithm can be applied on somatic mutations to predict if the mutation can introduce a PTC and activate the NMD pathway to cause the loss-of-expression of the mutation-harbouring gene. The algorithm has three components.

First, it infers the CDSs of mutated genes based on the mutation data, CDSs and positional annotations of transcripts provided by UCSC (hg19/NCBI Build 37, Feb. 2009), i.e. the TxDb R package and the BSgenome R package. The canonical isoform was defined as the isoform that had the longest coding sequence (CDS).

The prediction algorithm then detected the existence of any inframe premature stop codon (PTC) (TGA, TAA and TAG) in the mutated CDS. The potential stop positions were selected by two criteria: (1) the stop codon was in the same frame as the start codon, because the CDS must be the multiples of three; (2) the stop codon was the closest one to the start codon among the candidates.

After calculating the relative positions of the stop codon and the last exon-exon junction in the variant sequence, the algorithm performed the classification of NMD by following three rules. The first rule was that the gene has >1 exons in its longest CDS. The second one was that the stop codon in the variant was >50 bp upstream from the last exon-exon junction, and the last one was that the PTC was > 200 bp downstream from the last exon-exon junction. If the mutation satisfied all the above rules, it would be classified as an NMD-elicit mutation. If it led to a PTC but was not labelled as NMD-elicit, it was catalogued as an NMD-escape mutation. If the mutation did not lead to a PTC, the mutations would be termed as a non-PTC mutation. The code of Masonmd is available on `https://github.com/zhiyhu/masonmd`.

2.1.3 Statistical analyses

Novel annotations compared to TCGA TCGA database catalogued all somatic mutations into 14 classes, namely missense mutation, nonsense mutation, silent mutation, nonstop mutation, mutation on the translation start site, mutation in intron regions, mutation in 3'UTR, mutation in 5'UTR, inframe deletion, inframe insertion, frameshift deletion, frameshift insertion and mutation on the splice site. Among them, Masonmd catalogued missense mutation, nonsense mutation, silent mutation, nonstop mutation, mutation on the translation start site, inframe indels and frameshift indels into three categories (NMD-elicit, NMD-escape and non-PTC). I compared the prediction results with the original TCGA annotations.

Comparison of expression levels To validate that NMD-elicit mutations can influence the expression levels, I compared the expression levels of genes harbouring NMD-elicit mutations to the expression levels of genes harbouring NMD-escape, non-PTC and silent mutations. The expression levels were normalised with the background expression level in a given cancer type. For a certain gene, the background expression level was defined as the collection of expression levels from the samples that had no somatic mutation or CNV and belonged to the same cancer type. The normalised expression ranged from 0 to 1. The comparison was conducted with the non-parametric statistical test, Wilcoxon rank-sum test, because the data did not follow the normal distribution.

Annotations of cancer-related genes The list of cancer-related genes was downloaded from the COSMIC database (`http://cancer.sanger.ac.uk/census`). [179]

Measurements of NMD prevalence To depict the gene-wise prevalence of NMD-elicit mutations, the enrichment of NMD-elicit mutation (termed NMD enrichment) refers to as the ratio between the frequency of NMD-elicit mutations and the frequency of silent mutations in a certain gene. The cancer-specific NMD targets were defined by the genes that have NMD enrichment equal to or more than 1 in a

certain cancer type.

2.2 ClinCluster: a clustering method tailored for clinical scRNA-seq data

2.2.1 Adjusted Rand index: evaluation of cluster analysis

This study used the adjusted Rand index to empirically assess various clustering methods. The ARI was derived from of the Rand index R [74, 155, 194]. The Rand index measured the similarity between the predicted clustering result and the ground truth.

Assume that there is a set of data points, where n represents the number of data points in this set. The variable a is the number of pairwise connections of data points that belong to the same cluster in both the predicted result **B** and the ground truth **A**, while b is the number of pairwise connections of two data points that belong to two different clusters for both the predicted result **B** and the ground truth **A** (Table 2.1).

	Pairs assigned to the same group (A)	Pairs assigned to different groups (A)
Pairs assigned to the same group (B)	a	d
Pairs assigned to different groups (B)	c	b

Table 2.1. Four options for a pairs of data points from a data set that has been pre-classified.
The ground truth is **A**, *while the predicted clustering pattern is* **B**.

The Rand index can be described by Equation 2.1.

$$R = \frac{a+b}{n(n-1)/2} \tag{2.1}$$

In other words, the numerator $a+b$ represents the consistency between the prediction

and the ground truth. The denominator $n(n-1)/2$ represents the total number of pairwise connections $\binom{n}{2}$, which is equal to the sum of a, b, c, d shown in Table 2.1. The Rand index is a value between 0 to 1. When $R = 1$, the prediction matches the ground truth perfectly.

However, the Rand index is affected by random association. To remove these effects, the permutation model is used to calculate the expected index and deduct it from the Rand index (Equation 2.2), which gives the adjusted Rand index (ARI). Thus, $ARI = 0$ means that the prediction is equivalent to random guessing.

$$ARI = \frac{R - R_{expected}}{R_{max} - R_{expected}} \tag{2.2}$$

Given the contingency matrix (Table 2.2),

	Y_1	Y_2	...	Y_r	Sums
X_1	n_{11}	n_{12}	...	n_{1r}	a_1
X_2	n_{21}	n_{22}	...	n_{2r}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_s	n_{s1}	n_{s2}	...	n_{sr}	a_s
Sums	b_1	b_2	...	b_r	

Table 2.2. Contingency matrix

The data has two sets of clustering patterns, $\mathbf{X}$ *and* $\mathbf{Y}$*. In pattern* $\mathbf{X}$*, cells are grouped into clusters,* $\mathbf{X_1}, \mathbf{X_2} \ldots \mathbf{X_s}$*; in pattern* $\mathbf{Y}$*, cells are grouped into clusters,* $\mathbf{Y_1}, \mathbf{Y_2} \ldots \mathbf{Y_r}$*. The entry* n_{ij} *represents the number of cells that belong to both* $\mathbf{X_i}$ *and* $\mathbf{Y_j}$*.*

The ARI can be described by Equation 2.3.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{2.3}$$

In this study, I used the `adjustedRandIndex` function from R package mclust [166].

2.2.2 Benchmarking on simulated data

The simulated datasets were generated by using R package Splatter [205]. I generated the datasets with randomised numbers of cells per batch in normal distribution and the ratios of each cell population/group in uniform distribution. The featured named group defined the ground truth of cell populations/clusters, and the feature named batch referred to the simulated batches.

In this comparison, I considered scenarios with various numbers of cell populations ($g = 2, 4, 8$) and batches ($b = 2, 4, 8$). Each condition had 7 replicates. The clustering algorithms were evaluated by the ARI between the true groups and the predicted ones. I also quantified the impact of batch effects, which was defined by the ARI between batches and predicted groups to measure the influence of batch effects on clustering results. Because it was no possible to assign a cluster number to Seurat, I used the recommended resolution of 0.6.

Given that the spectral clustering used iterative eigenvalue solvers [133, 196], I tested its stability by repeating only the initial spectral clustering on the simulated data and measuring the ARI [74] between the prediction and the group truth.

ClinCluster was benchmarked with other three frequently used clustering algorithms: Seurat [24], RaceID3 [68] and SC3 [95]. The performance of these methods was measured by the ARI between the clustering results and the group truth.

ClinCluster was also compared with a batch effect removal algorithm, mutual nearest neighbours (MNN). As the output of the MNN function `bachelor::fastMNN()` is a matrix of corrected principal components, I appllied spectral clustering on the output matrix to cluster the cells.

2.2.3 Benchmarking on the melanoma dataset

This dataset was downloaded from the accession GSE72056 [183]. It contained over 4000 single cells from 19 clinical melanoma samples. This study used CNV inferred from the scRNA-seq data to separate the malignant cells from the non-malignant ones, based on the knowledge that the melanoma cells harbour frequent CNVs.

In the benchmarking analysis, I combined the nonmalignant cells from seven patients that each had more than 200 cells sequenced. In the ClinCluster pipeline, I set the number of initial clusters per patient as 7 and the number of final clusters as 6.

To measure to which extent the clustering was affected by the inter-patient variability, I used the ARI between patients and clustering results to denote impact of inter-patient variability (IPV). This measurement was based on a hypothesis that the cell populations were shared across all patients. Thus, if a clustering algorithm was only affected by batch effects but not cell subtypes, the IPV would equal to 1. In the real situation, the lowest IPV would not reach 0, because the cell populations was often not evenly distributed across patients.

2.3 Single-cell transcriptomic profiling of human fallopian tubes and compositional analysis of ovarian tumours

2.3.1 Patient samples

Patients in this project were recruited under the Gynecological Oncology Targeted Therapy Study 01 (GO-Target-01, research ethics approval #11-SC-0014). All patients were appropriately informed and consented before the surgeries.

2.3.2 Experimental procedures

2.3.2.1 Sample processing

Sample dissociation The fresh fallopian tube tissue was washed with Phosphate-Buffered Saline (PBS), pH 7.4 (Gibco) and cut into 2 mm pieces. The minced tissue was digested with the Tumor Dissociation Kit, human (Miltenyi Biotec) at 37°C for 1 hour. The cell suspension was filtered through a 100 μm cell stainer to exclude clumps.

Fluorescence-activated cell sorting The prepared single-cell suspension was stained with the EpCAM-APC antibody (Clone REA764, Miltenyi Biotec) and the CD45-FITC antibody (Clone REA747, Miltenyi Biotec) with the dilution of 1:40 at 4°C in dark for 30 min. After staining, the cells were washed with 1-2 mL PBS and resuspended in 400 μL to 2 mL fluorescence-activated cell sorting (FACS) buffer, which contained 1× PBS, 1% bovine serum albumin (BSA, Sigma-Aldrich) and 2 mM RNase-free EDTA pH 8.0 (Invitrogen). DAPI was added at 2-10 μg/mL into the cell suspension 10 min before sorting. I performed cell sorting on SH800 sorter (Sony). The EpCAM+CD45- epithelial population and two control groups (EpCAM-CD45+ and EpCAM-CD45-) were sorted into 96-well PCR plates (Figure 2.1). Each 96-well plate has 1 or 2 empty wells as negative controls to exclude the possibility of contamination.

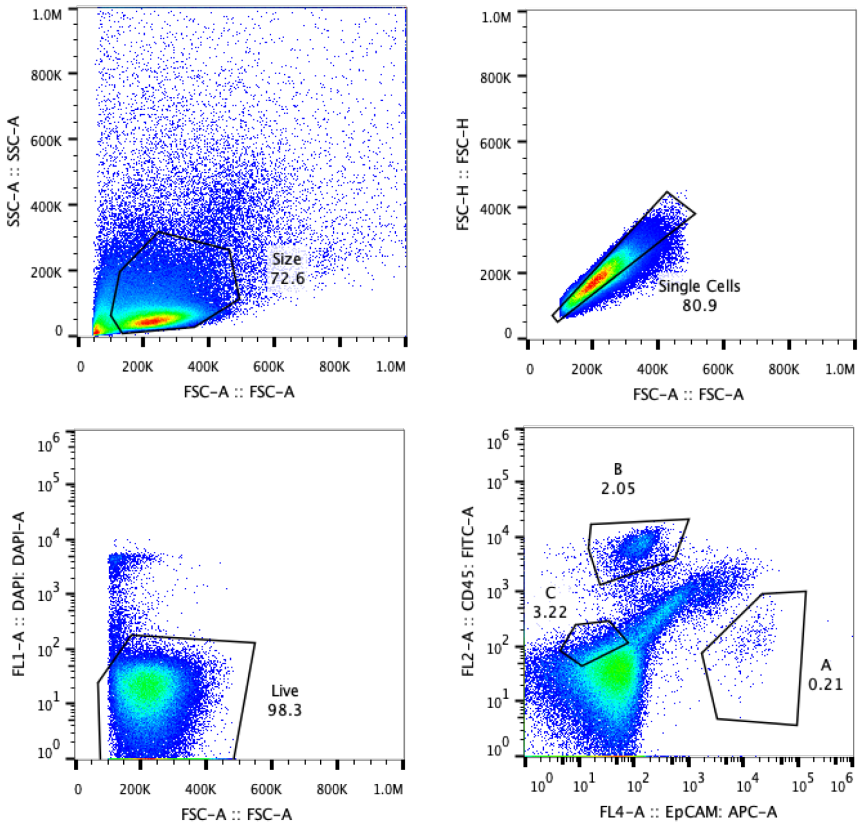

Figure 2.1. Diagram of gating strateies in FACS.

The example is plotted with the data of samples from Patient 15077. (Top left) the gating selects the cells with right sizes. (Top right) the gating selects the singlets, with approximately equal forward scatter height (FSC-H) and area (FSC-A). (Bottom left) the gating selects the DAPI negative live cells. (Right bottom) Gate A, EpCAM+CD45-, i.e. epithelial cells; Gate B, EpCAM-CD45+, i.e. blood cells; Gate C, EpCAM-CD45-.

2.3.2.2 Primary cell culture

To investigate the effect of cell culture on transcriptomes, some fallopian tube epithelial cells were cultured to be compared with fresh ones. The dissociated fallopian tube cells were cultured with the BM2 medium containing 41.7 mL Advanced DMEM/F12 (Gibco), 540 μL HEPES (Gibco), 450 μL Penicillin-Streptomycin (Gibco), 4.5 μL (10 ng/mL) epidermal growth factor (EGF) (Sigma-Aldrich), 2.25 mL heati-nactivated Fetal Bovine Serum (FBS, Gibco) (5%), 45 μL (9 μM) Rock inhibitor (Y-27632 dihydrochloride, Sigma-Aldrich) in the 6-well plate at 37°C, 20% O_2 and 5% CO_2.

2.3.2.3 Smart-Seq2

I followed the Smart-Seq2 protocol [147, 148] with modifications and automation. I used the 5′-biotinylated oligo-dT (Table 2.5) and the 5′-biotinylated ISPCR primer (Table 2.7) to minimise the generation of reaction byproducts.

Before sorting, prepare plates of lysis buffer in the PCR Clean Room that belongs to WIMM Single Cell Facility. First, thaw oligo-dT30 and dNTP mix in fridge 30 min before preparing plates and put Triton-X on a heat block to make it less viscous. Prepare 0.4% Trion by the dilution steps shown in Tables 2.3 and 2.4.

Triton-X 100 (pure and pre-heated)	100 μL
H_2O	900 μL
10% Triton	1 mL

Table 2.3. The recipe for the preparation of 10% Triton.

10% Triton	65 μL
H_2O	1560 μL
0.4% triton	1625 μL

Table 2.4. The recipe for the preparation of 0.4% Triton.

Then prepare the lysis buffer by following Table 2.5.

	1x	1 96-well plate
RNase inhibitor	0.1 μL	10.5 μL
0.4% Triton X	1.9 μL	199.5 μL
5′-biotinylated oligo-dT30 (10 μM, IDT)	1 μL	105 μL
dNTP 10 mM (Thermo Scientific, cat R0192)	1 μL	105 μL
Total	4 μL	420 μL

Table 2.5. The recipe for the preparation of lysis buffer.
Oligo-dT sequence is 5′-biotinylated-AAGCAGTGGTATCAACGCAGAGTACT30VN-3′.

Vortex and spin down the mix, and then add 4 μL lysis buffer per well. The dispensing of lysis buffer can be achieved with the MANTIS liquid handler (Formulatrix). The plates with lysis buffer were sealed with films and brought to the sorting facility. One cell was sorted to a single well. Each batch of sorting had at least a bulk control (10 or 100 cells). Each plate had at least one negative control (no cell). After sorting, the plates were spun down, snap-frozen on dry ice, sealed with foil films and stored at -80°C.

The reverse transcription (RT) reaction and the setup of PCR reaction were also performed in a PCR Clean Room.

Before the set up of RT master mix, thaw SuperScript II buffer and DTT at room temperature and the template-switching oligo (TSO) at 4°C. Once SuperScript II buffer and DTT were fully defrosted, move them onto an ice block. Take the plate containing samples from -80°C, store it on dry ice and prepare the RT master mix by following Table 2.6.

Before adding the SuperScript RT enzyme, heat the sample at 72°C for 3 min in a preheated, hot-lid thermal cycler. During the 3 min, add RT enzyme into the master mix, vortex it and centrifuge. After the sample cooled down, add 6 μL master mix per well including the negative control. Do not touch the lysis buffer with the tip during adding the RT master mix to avoid taking out any material. Then spin down the plate and incubate it on a thermal cycler by the program: 42°C for 90 min, followed by 10 cycles of 2 min 50°C and 2 min 42°C, and 70°C for 15 min as

Reagent	Supplier	1× μL	1× plate	Final
SuperScript II buffer (5x)	Invitrogen, cat 18064014	2	211.2	1x
DTT (100 mM)	Invitrogen, cat 18064014	0.5	52.8	5 mM
Betaine (5 M)	Sigma-Aldrich, cat B0300-1VL	2	211.2	1 M
$MgCl_2$ (1 M)	Invitrogen, cat AM9530G	0.06	6.336	0.01 M
Rnase Inhibitor (40 U/μL)	TaKaRa, cat 2313A	0.25	26.4	1 U/μL
TSO (100 μM)	Qiagen	0.1	10.56	1 μM
RT-PCR grade water	Invitrogen, cat AM9935	0.84	88.704	-
RT enzyme (200 U/μL)	Invitrogen, cat 18064014	0.25	26.4	50 U
Total	-	6	633.6	-

Table 2.6. The recipe for the preparation of reverse transcription master mix. *TSO sequence is 5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′. Add reagents strictly by the order listed in the table. Cat, Catalogue number. The protocol is adapted from Picelli et al. [147]*

previously described [147]. This RT step took around 3 hours.

After the reaction finished, go back to the clean room to set up the PCR reaction. First prepare the PCR master mix by following Table 2.7 and add 15 μL to each well. Vortex and spin down the plate.

Reagent	Supplier	1× μL	1× plate
KAPA Hifi HS Ready Mix (2x)	Roche, cat KK2602	12.5	1320
ISPCR primrs (10uM)	IDT	0.125	13.2
Water	Invitrogen, cat AM9935	2.375	250.8
Total	-	15	1584

Table 2.7. The recipe for the preparation of PCR master mix. *ISPCR primer sequence is 5′-biotinylated-AAGCAGTGGTATCAACGCAGAGT-3′. Add reagents strictly by the order listed in the table. The protocol is adapted from Picelli et al. [147]*

After adding the PCR master mix, bring the plates out of the clean room and run the PCR program (Table 2.8) on a thermal cycler.

The last step of PCR was usually held at 10°C overnight. The PCR product was

Cycle	Denature	Anneal	Extend	Hold
1	98°C, 3 min	-	-	-
24	98°C, 20 s	67°C, 15 s	72°C, 6 min	-
1	-	-	72°C, 5 min	-
1	-	-	-	4°C or 10°C

Table 2.8. The PCR program to pre-amplify single-cell cDNA.

cleaned up with Ampure XP beads (Beckman Coulter) on the second day by using the Biomek FxP Workstation (Beckman Coulter). For each well, I used 20 μL beads to bind the cDNA and 20 μL Buffer EB (Qiagen) to elute. The after-clean PCR product was stored in 384-well plates at -20°C.

To quantify the cDNA concentration, I used the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen, cat No. P11496). The dye was diluted in 1× TE (diluted from 20× TE provided by the kit) at 1:200. I added 9.5 μL diluted dye and 0.5 μL cDNA to each well of a black flat-bottom 384-well plate (Grenier, car No. 781076). The cDNA was dispensed by the Mosquito HTS robot (TTP Labtech). I diluted the DNA ladder into a concentration gradient (ng/μL) of 2, 1, 0.5, 0.25, 0.125, 0.0625, 0.03125 and 0. I added 0.5 μL ladder into 9.5 μL diluted dye, and then used a microplate reader CLARIOstar (BMG LABTECH) to read the fluorescent signal in each well. The regression line was computed between the fluorescent signal and the cDNA concentration. This relationship was used to calculate the concentrations of single-cell cDNA.

Because the concentration could be affected by primer dimers, I used qPCR of the housekeeping gene, *GAPDH*, which had higher specificity than the PicoGreen assay. After the first round of PCR and beads cleanup, I set up and ran 4 μL qPCR reaction by using the QuantiNova SYBR Green PCR Kit (Qiagen). For a new tube of the SYBR Green PCR Master Mix, add 17 μL ROX reference dye to 1.7 mL master mix. To set up the master mix, first thaw 2x QuantiNova SYBR Green PCR Master Mix, template gDNA or cDNA, primers, QN ROX Reference Dye and RNase-Free water. Vortex to mix the individual solutions. The sequences of GAPDH primers were 5′-CCCCCACCACACTGAATCTC-3′ (forward) and 5′-

GGTACATGACAAGGTGCGGC-3′ (reverse). Set up the qPCR master mix by following Table 2.9.

Component	1× (μL)	1× 384 plate	Final
2x QuantiNova SYBR Master Mix	2.02	852	1x
Primers (10 μM)	0.28	18	0.7 μM
RNase-Free water	1.3	549	-
cDNA (add later)	0.4	-	$\leq$ 100 ng/reaction

Table 2.9. The recipe for the preparation of QuantiNova qPCR master mix.

Mix the master mix thoroughly by vortexing and dispense 3.6 μL per well into a PCR plate (4titute FrameStar frosted well 384 plate, cat No. 4ti-0387). Then add 0.4 μL template cDNA to each well by using the Mosquito HTS robot. The qPCR plate was sealed and spun down briefly. The qPCR reaction was performed in the QuantStudio 7 Flex Real-Time PCR System (Thermo Fisher) with the program shown in Table 2.10.

Step	Time	Temperature	Ramp rate
PCR initial activation step	2 min	95°C	Fast mode
Denaturation	5 s	95°C	40 cycles
Annealing/extension	20s	60°C	
Melting curve analysis			

Table 2.10. The program of the QuantiNova qPCR reaction to measure the cDNA quality.

I also picked some wells and checked them with High Sensitivity D5000 reagents and screentapes on the 2200 TapeStation system (Agilent). The good-quality cDNA had a peak of around 1 kb (Figure 2.2). It is common to see a small peak at around 100 bp and a wide range of peaks from 300 bp to 5000 bp in clinical samples. The former refers to the primer dimer, and the latter corresponds to the slight degradation of cDNA. Yet these factors usually would not affect the library preparation and sequencing.

The wells with good-quality cDNA (positive expression of *GAPDH*) were cherry-

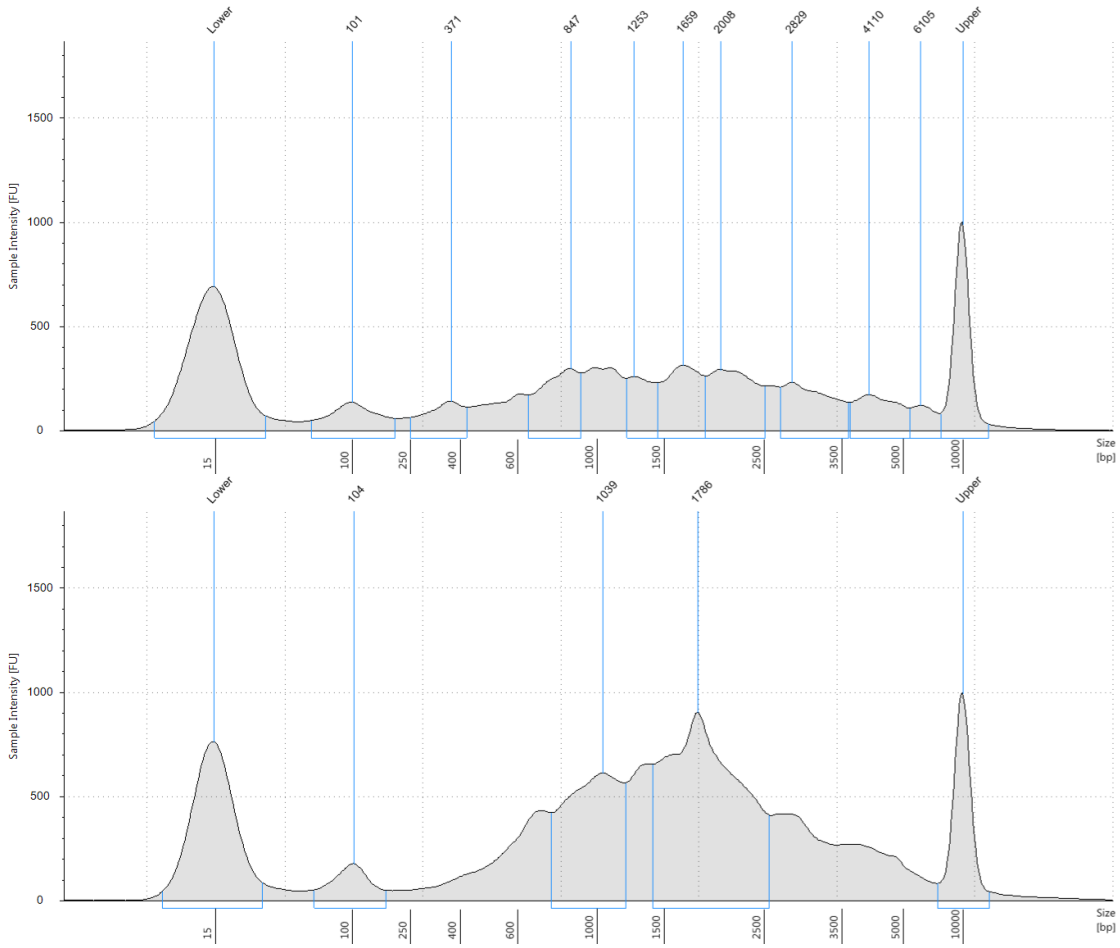

Figure 2.2. TapeStation results of a pre-amplified cDNA library

picked to a new 384-well plate and normalised to 2 ng/μL with EB by using the Biomek FxP Workstation (Beckman Coulter).

Then I used a miniaturised 4 μL Nextera reaction (Illumina, cat No. FC-131-1096) with 400 nL cDNA input to generate scRNA-seq libraries. The miniaturised reaction was performed with the Mosquito HTS robot. This library preparation step followed the protocol provided by TTP Labtech, titled Nextera XT Miniaturization down to 4 μL (version 1.0; 03/15). First, it dispensed 800 nL TD buffer (tagmentation buffer), 400 nL cDNA and 400 nL ATM (containing Tn5 transposase) into each well of a 384-well PCR plate. The plate was sealed, vortexed, spun down and incubated in the thermal cycler at 55°C for 10 min. Then immediately add 400 nL NT buffer per well to terminate the tagmentation reaction. Incubate the plate at room temperature for 5 min and then dispense 1200 nL NPM (PCR master mix) per well. Afterwards, add 400 nL index i5 and 400 nL index i7 per well.

For the indexes (Illumina, cat No. FC-131-2001 and FC-131-2004), I used the layout shown in Tables 2.11 and 2.12. The combination of 16 indexes i5 and 24 indexes i7 allowed the pooling of 384 single-cell libraries.

Row	Index i5	Row	Index i5
A	502	B	503
C	505	D	506
E	507	F	508
G	510	H	511
I	513	J	515
K	516	L	517
M	518	N	520
O	521	P	522

Table 2.11. The layout of Illumina Nextera index i5 in a 384-well plate.

Column	1	3	5	7	9	11	13	15	17	19	21	23
Index i7	701	703	705	707	711	714	716	719	721	723	726	728
Column	2	4	6	8	10	12	14	16	18	20	22	24
Index i7	702	704	706	710	712	715	718	720	722	724	727	729

Table 2.12. The layout of Illumina Nextera index i7 in a 384-well plate.

After adding the indexes, the plate was spun down and put on a thermal cycler to run the PCR programme shown in Table 2.13.

Following the PCR amplification, 800 nL PCR product from each well was pooled into a 1.5 mL Eppendorf tube and cleaned with 0.6:1 Ampure beads. The library was eluted into 25 μL EB and measured by the Qubit dsDNA HS (High Sensitivity) Assay (ThermoFisher, cat No. Q32854) and TapeStation HS D5000 reagents. The fragment size of the pooled library was a bell shape with a peak between 600 bp and 700 bp (Figure 2.3).

Each pooled library contained 384 combinations of indexes. At least one well was negative control. The library was sequenced on the NextSeq 500 sequencer (Illu-

Cycle	Denature	Anneal	Extend	Hold
1	-	-	72°C, 3 min	-
1	95°C, 30 s	-	-	-
12	95°C, 10 s	55°C, 30 s	72°C, 30 s	-
1	-	-	72°C, 5 min	-
1	-	-	-	10°C

Table 2.13. The PCR program for the library preparation.

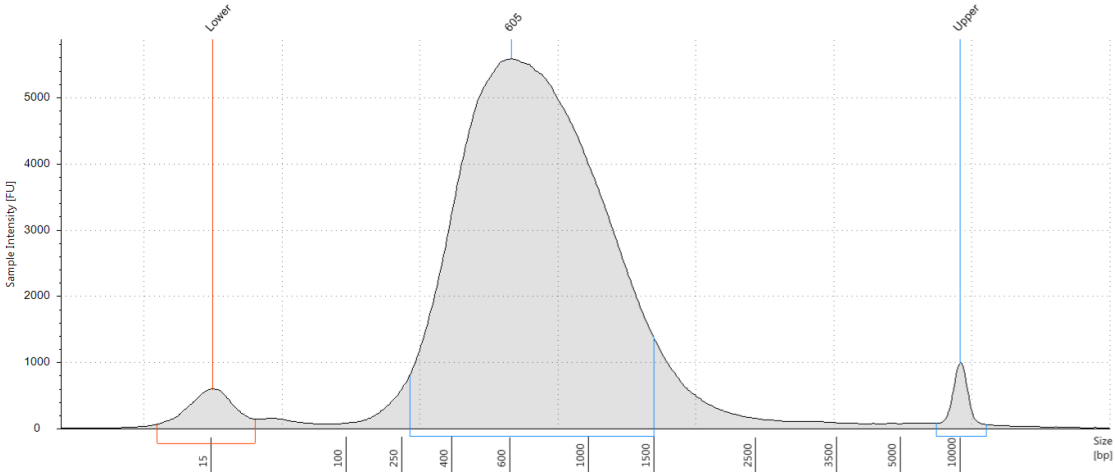

Figure 2.3. The size of a pooled sequencing library.

mina) with 75 cycles high output NextSeq kits v2 (Illumina) at the WIMM Sequencing Facility.

2.3.2.4 10x Genomics

I used the 10x Genomics Chromium Gene Expression Kits and Chips (v2). For each sample, I sorted 10 thousand DAPI-negative alive single cells and followed the protocol provided by 10x. In principle, the cells and barcoded beads were firstly mixed on the chip to form the one-cell-one-bead droplet. The reverse transcription reaction was achieved in the droplet, in which the cDNA from each cell was labelled with a unique barcode. Afterwards, cDNA from all cells within a sample was pooled together to be amplified and transformed into sequencing libraries by shortening cDNA and adding sequencing adapters. The prepared libraries were sent to the Novogene Company and sequenced on NovoSeq or HiSeq sequencers (Illumina)

following the instruction provided with 10x kits.

2.3.3 Bioinformatic and quantitative analysis

2.3.3.1 Impact of culture on transcriptomes

Preprocessing of Smart-Seq2 data The fastq files were merged by `cat`, trimmed by Trim Galore v0.4.1 (Cutadapt 1.2.1 incorporated, `trim_galore -- clip_R1 15 -- three_prime_clip_R1 3`), mapped by STAR v2.4.2a (`STAR --genomeDir ref.genome --readFilesIn input.fq.gz --readFilesCommand zcat --outFileNamePrefix output --runThreadN 25 --outSAMtype BAM SortedByCoordinate`) and counted by FeatureCount v1.4.5 (`featureCounts -T 20 -a genes.gtf -t exon -g genes.gtf id -o outputpath -s 0 *.bam`). The reference genome used was UCSC hg19 located at `/databank/igenomes/Homo_sapiens/UCSC/hg19/` (version `2013-03-06-11-23-03`) on the CBRG Deva server. The low-quality cells were excluded with the cutoffs of 1200 genes and 10,000 reads per cells.

Pseudotime analysis This analysis used the scRNA-seq data from two patients, 11553 and 15072. The culture time of three groups, fresh, overnight-cultured and long-term cultured, was set as 0, 0.5 and 2. I applied PhenoPath [26] on the normalised count matrix to calculate the pseudotime values for each cell. PhenoPath exploited a Bayesian model to align cells onto a one-dimensional trajectory, which is the pseudotemporal process. I used the culture time as a covariate in this model. The closeness of the pseudotime between two cells represented that similarity of their molecular data in terms of a one-dimensional progression.

Differential expression analysis and pathway analysis I used differential expression analysis to study the differences between fresh and cultured cells, i.e. the effect of cell culture on single-cell transcriptomes. I used limma voom [103], which was designed for bulk scRNA-seq data but outperformed other single-cell differential expression (DE) analysis tools in a recent benchmarking study [174]. I did three comparisons, namely fresh versus overnight, long-term versus fresh, and

long-term versus overnight. The genes with false discovery rate (FDR) less than 0.05 and $\log_2$ fold-change no less than 0.4 were labelled as differentially expressed genes (DEGs). Pathway analysis [190] was performed on the DEGs to identify the active subnetworks of gene sets.

2.3.3.2 Constructing the cell atlas of human fallopian tube

Preprocessing 10x data The fastq files were put through the CellRanger v3.0.1 with the command `cellranger count`. I used the cellranger reference `hg19-3.0.0`. The low-quality cells were excluded by the following thresholds: the number of genes detected per cell > 200 and < 4000, and the percentage of mitochondrial RNA < 10.

Clustering by Seurat I used the recommended pipeline of Seurat v3 to analyse the 10x data. The pipeline included the following steps: normalising data, selecting high variance genes, scaling data, calculating principal components (PCs), picking up significant PCs, computing the nearest neighbour graph and finding clusters.

Excluding putative doublets I filtered out the presumable doublets by using DoubletFinder [124]. DoubletFinder used the idea of comparing the real data points and the simulated data points of doublets from two heterogeneous populations. In the process, the top 10 PCs were used. DoubletFinder found 102 cells with a high probability of being doublets and 15 cells with a low probability. After filtering, 1446 singlets were kept for the following analysis.

Finding marker genes The marker genes of each cluster were selected by the thresholds of $\log_2$ fold-change more than 0.4 and FDR less than 0.05.

Analysis of cell-cell interactions I used CellPhoneDB to study the intercellular communications [193]. CellPhoneDB was based on the ligand-receptor interaction database. It calculated an empirical p-value of a certain interaction by using the

random shuffling combinations as background. The strength of interaction was calculated from the average of expression levels in the two cell populations.

2.3.3.3 Extensively profiling of secretory and ciliated populations

Clustering For the epithelial cells, I first applied Seurat v3 on the data to separate the fully differentiated populations, secretory and ciliated cells. The subpopulations of secretory cells were masked by patient-specific variability. Thus, I used the in-house algorithm, ClinCluster. As aforementioned, the principle of this algorithm was using the differential expression analysis to sift out differentially expressed functional modules from confounding effects. The cells from each patient were firstly clustered by KNN-based spectral clustering. Then the inter-group distances were calculated by using limma voom [103]. The final clusters were merged by the hierarchical clustering based on the inter-group distance matrix.

The identification of marker genes for each cluster was also based on limma voom between the target cluster and other clusters. The cutoffs of differentially expressed genes (or marker genes) were $\log_2$ fold-change $\geq$ and FDR < 0.05.

Gene Ontology analysis The Gene Ontology analysis was performed by using the website interface of GORilla [48, 49]. I used the cutoff of FDR < 0.05 to pick up the enriched biological pathways.

2.3.4 Deconvolution analysis of bulk ovarian tumours

2.3.4.1 Deconvolution analysis

I downloaded the TCGA RNA-seq data (IlluminaHiSeq UNC RNA-seq dataset, version 2017-10-13), clinical data and molecular subtyping (version 2016-05-27) data from the UCSC Xena portal. The dataset had 308 clinical samples, in which one patient's survival data was missing. Thus, the number of patients with valid survival data was 307.

To perform the deconvolution analysis, the top marker genes were selected for each of the five fallopian tube epithelial populations, namely KRT-MHC, differentiated, extracellular matrix (ECM), cell cycle and ciliated. The selection of marker genes followed the criteria: (1) $\log_2$ fold-change was no less than 1.5; (2) FDR was no more than 0.01; (3) variable loadings in the principal components of TCGA expression data was equal or more than 0.2; (4) correlation of expression levels with other marker genes was less than 0.9. This selection led to 52 top marker genes. Each cluster had 7-15 genes to represent its transcriptomic signature.

I exploited BSEQ-sc [8] to scale the expression levels of each marker genes across all five cell subtypes, which generated a reference matrix as the parameter for the deconvolution model. The deconvolution tool, Cibersort, was based on a linear support vector regression [132], which was more robust to the outliers than the regular linear models. Cibersort took the input of a bulk expression vector and a reference matrix to compute the proportion of each cell type in a bulk sample, based on the assumption that the expression profile of the bulk sample is the linear combination of each component. In other words, the deconvolution analysis estimated the composition of bulk tumour samples. I applied the deconvolution analysis on the expression data of the TCGA ovarian cancer cohort. The output matrix had 308 rows and 5 columns. Each column corresponded to a cell subtype and each row a bulk tumour. The sum of each row was equal to 1. The entries of the matrix represented the proportion of a given cell subtype in a bulk tumour.

2.3.4.2 Correlation analysis

I performed a correlation analysis to study the relationship between the ECM proportion and clinical factors. Continuous factors were dichotomised. The stage was split into two groups: early (I & II) and late (III & IV). The primary therapy outcome was grouped into complete respond and incomplete respond. The residual disease was dichotomised into two groups: residual (> 20 mm and 11-20 mm) and non-residual (no macroscopic disease and 1-10 mm). The correlation tested used the Pearson Correlation test in R.

2.3.4.3 Survival analysis

I studied the relationship between the ECM proportion and overall survival by using the Cox hazard regression model. To be specific, the function `coxph` in R was used. When I compared the new grouping criteria with the TCGA molecular subtyping system, all samples were divided into four groups, namely ECM-high and mesenchymal (Mes, n = 66), ECM-high and non-Mes (n = 88), ECM-low and Mes (n = 3), as well as ECM-low and non-Mes (n = 151). Since the ECM-low and Mes group only had three samples, it was excluded from the multiple-group survival analysis.

Chapter 3

A pan-cancer analysis of nonsense-mediated decay

3.1 Abstract

The nonsense-mediated decay (NMD) pathway can spot and eliminate mutation-harbouring mRNAs to prevent the expression of truncated proteins. Recent studies found that NMD targets several tumour suppressor genes (TSGs) in cancer samples. To systematically assess the role of NMD in cancer, I developed a prediction algorithm and applied it on near 8 thousand genomes of 32 cancer types from The Cancer Genome Atlas (TCGA) and identified a large number of NMD-elicit mutations. Further analysis demonstrates that these NMD-elicit mutations were concentrated in the TSGs and other cancer-related genes. The results imply NMD as a potential drug target for cancer therapy.

3.2 Introduction

One ubiquitous feature of cancer genomes is the mutation [2], including single-nucleotide variants (SNVs) and small insertions and deletions (indels). A subset of the mutations can disturb the expression level or the protein function of the mutated genes. In this chapter, I looked into the phenomenon that the mutations can cause the loss-of-expression of the mutated genes by activating a pathway termed

nonsense-mediated decay (NMD).

3.2.1 NMD is an mRNA surveillance pathway

In a normal condition, mRNAs are translated into functional proteins by the ribosomes. Yet mutations can sometimes introduce an aberrant stop codon upstream from the normal one, termed as the *premature stop codon (PTC)*. A PTC in the mRNA can putatively result in the synthesis of truncated proteins (Figure 3.1), which may possess the deleterious *dominant-negative effect* [56, 61, 120]. The NMD pathway eliminates such PTC-harbouring mRNAs [113] and, thus, prevents PTCs from misleading the translation into stopping too early. Hence, the NMD pathway plays an indispensable role in the cells of all eukaryotes by protecting cells from potentially hazardous truncated proteins.

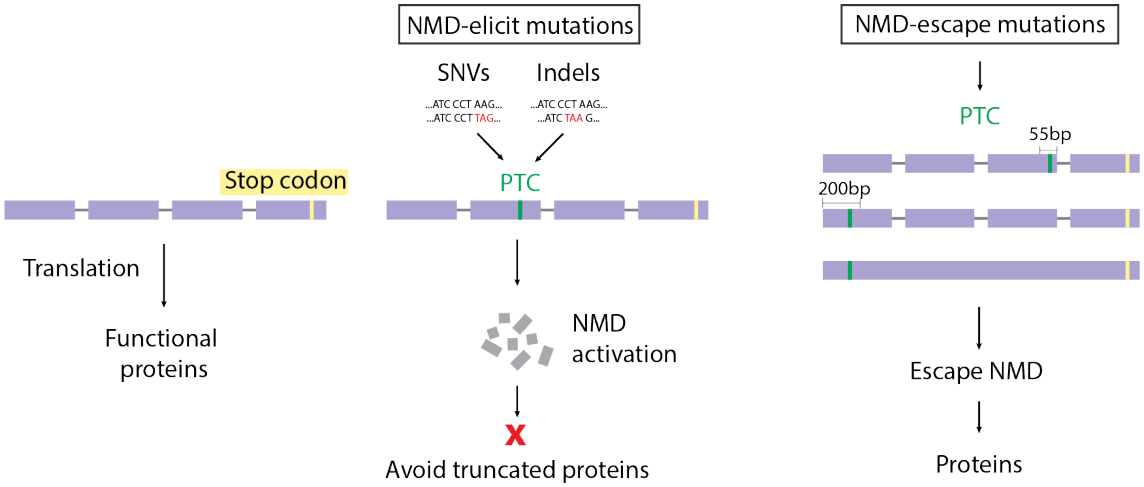

Figure 3.1. Diagram of the nonsense-mediated decay.
SNVs, single-nucleotide variants; indels, small insertions or deletions; PTC, premature termination codon; NMD, nonsense-mediated decay.

Notably, not all PTC-harbouring mRNA variants can activate NMD. Previous studies have discovered and validated three rules on the selectivity of NMD targeting (Figure 3.1) [111, 180, 209]. The first one is termed as the 50 bp rule that the NMD pathway only targets the PTC that is >50-54 bp upstream from the last exon-exon junction [72, 153]. Secondly, the position of PTC is more than 200 bp downstream from the start codon [206]. Thirdly, the number of exons is more than 1. When all of the rules are satisfied, the PTC-harbouring mRNA tends to be decayed. Ac-

cordingly, the mutation leading to this PTC is termed as the NMD-elicit mutation. When the rules are not satisfied at the same time, the PTC is unlikely to be able to activate the NMD pathway and the mRNA can escape the degradation. The causative mutation is, therefore, termed as the NMD-escape mutation.

The 50 bp rule is based on the canonical exon junction complex model, which is well supported [180, 208]. However, studies have reported exceptions [23], such as long 3′ untranslated region (UTR) triggering NMD [119]. A recent study systematically evaluated the rules affecting NMD efficiency and found that only 75% of the variance in NMD efficiency can be explained by the identified rules [111], indicating that although the model can efficiently predict most of the NMD targets, we should not neglect the uncertainty embedded in it.

3.2.2 NMD and cancer

NMD and PTCs are related to more than one-third of the genetic diseases [130]. Previous studies have reported that the expression of tumour suppressor genes (TSGs) harbouring nonsense SNVs is suppressed by NMD in cancer, while these mutated TSGs still keep partial functions [41, 88, 191]. In this scenario, NMD facilitates cancer progression. Another study showed that the semi-functional tumour suppressor can be re-expressed by inhibiting NMD, which represses the survival of cancer cells [121].

Therefore, inhibition of NMD is a potential strategy to treat cancer [58, 107]. A chemotherapeutic drug, doxorubicin, has been proven to promote apoptosis by restraining the NMD pathway [152]. Many small molecules have been identified as NMD inhibitors to restore the full-length tumour suppressor proteins when combined with the “read-through” drug at a nontoxic level [121].

3.2.3 Motivations

Although previous studies have shed light on the possibility of using NMD inhibitors to treat cancer, the global effect of NMD on the pan-cancer transcriptomes still

remains elusive. Before using PTC-harbouring variants as drug targets, it is critical to predict the NMD targets in cancer to prevent the runaway of seriously deleterious truncated proteins. Previous meta-analyses studied nonsense mutations in various diseases and cancers, but the authors only focused on the nonsense SNVs [80, 130]. The frameshift mutations have not been annotated on the capability to active NMD, whilst some of them can produce PTCs and elicit NMD. A recent study illustrated the effect of protein-truncating genetic variants on transcriptomes [159], but it was based on the healthy cohort and neglected the classification of frameshift indels by generating PTCs or not. Another study validated the NMD rules with the TCGA data, but the effect of NMD on cancer was unclear due to the strict filtering of mutations [111].

There was no well-developed algorithm to identify the putative NMD targets caused by frameshift indels. The annotations of NMD targets in cancer are also absent in the public databases, especially for the frameshift mutations. For example, the Catalogue of Somatic Mutations in Cancer (COSMIC) documented a large number of cancer-related somatic mutations [179], but the annotations of variants lack the insight into NMD. The TCGA database annotated the PTC-inducing SNVs as nonsense mutations, but it dismissed the frameshift mutations that can elicit NMD. Moreover, the NMD-elicit mutations have not been evaluated by the expression-based variant-impact analysis [14]. Hence, the efficiency and the prevalence of NMD in cancers have been unclear so far.

This study is aimed to estimate the extent to which tumours are affected by NMD. This study exploited a pipeline to identify the potential NMD-targeted genes, especially the ones caused by frameshift indels, in pan-cancer genomes.

3.3 Results and Discussion

3.3.1 Design of the pan-cancer analysis

This study analysed the somatic mutation data and the RNA-seq data of 32 cancer types from the TCGA database (Figure 3.2). I developed a pipeline, Masonmd, and

exploited it to label the somatic mutations as NMD-elicit, NMD-escape and non-PTC. This work validated the effect of NMD on mRNA levels by comparing the expression levels of genes that harbour NMD-escape mutations or other mutations. It also looked into the prevalence of NMD-elicit mutations in cancer-related genes and pathways across cancer types.

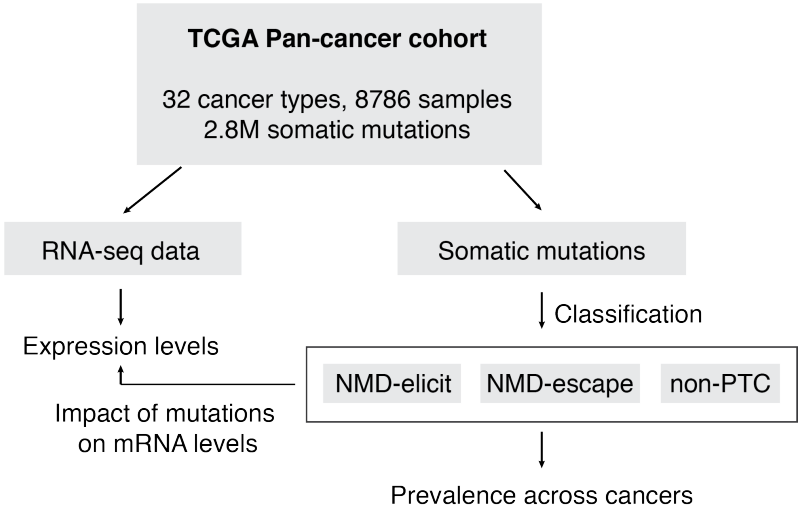

Figure 3.2. Diagram of the workflow design.

3.3.2 Novel annotations revealed prevalent NMD in cancers

In total, this pipeline detected 140,020 (8%) NMD-elicit mutations and 61,550 (3.6%) NMD-escape mutations out of 1,706,218 somatic non-silent SNVs and indels in the coding region (mutations on translation start sites and splice sites excluded). To check if my work has any additional contribution compared to the previous knowledge, I compared the prediction results to the original TCGA annotations of mutations. The analysis revealed that 31% nonsense mutations annotated by TCGA were predicted to escape NMD (Table 3.1). Frameshift mutations were thought to only be able to mess up the sequence of translated proteins, while the analysis reveals that two-thirds of the frameshift indels can cause loss-of-expression leading to further disturbance.

I next summarised the prevalence of predicted NMD-elicit mutations across cancer types (Figure 3.3). In general, the average frequency of NMD-elicit mutations per sample was positively associated with the frequency of total mutations (correlation

TCGA	NMD-elicit	NMD-escape	nonPTC
Frameshift deletion	46881	19863	0
Frameshift insertion	13706	5838	0
Inframe deletion	0	0	7985
Inframe insertion	0	0	698
Missense	14	26	1463835
Nonsense	79419	35823	53

Table 3.1. Comparison between the original TCGA annotations and the new annotations predicted by Masonmd.

= 0.94, p-value = 1.204e−15, by Pearson correlation test). The five cancer types with the highest proportion of NMD-elicit mutations were breast invasive carcinoma, stomach adenocarcinoma, kidney clear cell carcinoma, kidney papillary cell carcinoma and colon adenocarcinoma. It suggests that NMD-elicit mutations commonly occur in cancer genomes, especially for certain cancer types.

3.3.3 Predicted NMD-elicit mutations are associated with decreased mRNA levels

I assumed that the NMD-elicit mutation can trigger the NMD machinery and degrade mutation-harbouring mRNAs. To validate this assumption, I compared the gene-wise mRNA levels between the samples that harboured NMD-elicit mutations and the ones that harboured other mutations. To mask the gene- and tissue-specific effects on expression levels, I normalised the expression levels of variant-harbouring genes with the expression levels of the genes free from somatic mutations or copy number variations (CNVs) within the same cancer type.

Compared to NMD-escape, non-PTC and silent mutations, the normalised mRNA levels were significantly lower for the genes with NMD-elicit mutations (ratio of mean = 0.19/0.53, p-value < 2.2e−16, by Wilcoxon rank-sum test, Figure 3.4). The expression-based variant-impact analysis demonstrates that the predicted NMD-elicit mutations are associated with the decreased mRNA levels putatively owing to

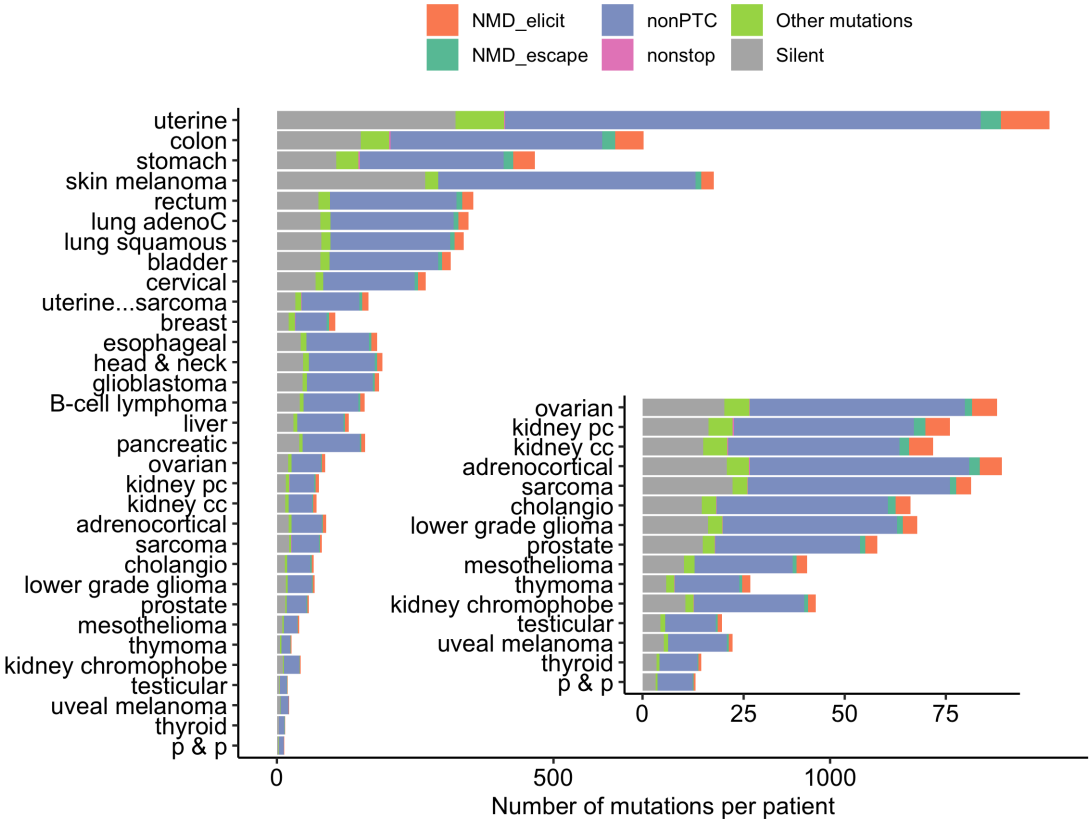

Figure 3.3. Overview of the frequency of NMD-elicit mutations across cancer. *The x-axis represents the average number of mutations per patient for each cancer type. The cancer types are ordered based on the average number of NMD-elicit mutation per sample. The bar is coloured by the annotated type of mutations. The subfigure zooms into the last 15 cancer types for better visualisation. Lung adenoC, lung adenocarcinoma; uterine...sarcoma, uterine carcinosarcoma; kidney pc, kidney papillary cell carcinoma; kidney cc, kidney clear cell carcinoma; p & p, pheochromocytoma & paraganglioma.*

the activation of NMD.

3.3.4 NMD prefers targeting cancer-related genes

This study scored the prevalence of NMD-elicit mutations in TSGs (COSMIC Tier 1) across cancer types. Overall, the NMD-elicit mutations targeting TSGs occurred in 48% (4222 out of 8786) samples across all cancer types (Figure 3.5), suggesting that TSGs are commonly targeted by NMD in cancer. There was substantial dispersion in terms of such prevalence among cancer types. For bladder, uterine corpus, lung squamous cell, colon and head & neck cancers, the TSGs were putatively targeted by NMD in more than 65% samples. However, this phenomenon was relatively rare in thymoma and thyroid cancers. It implies that the therapy targeting NMD is more

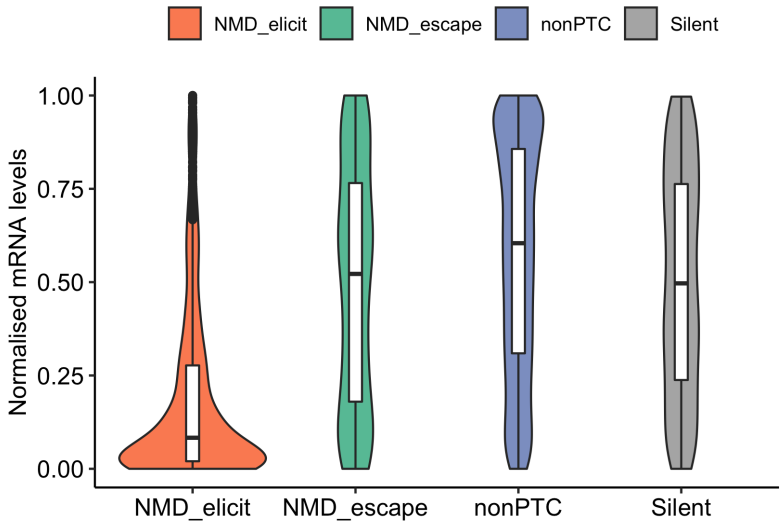

Figure 3.4. Effect of NMD-elicit mutations on mRNA levels.

The violin plot visualises the distribution of the normalised mRNA levels in each group coloured by the type of mutations. The embedded boxplot shows the upper quantile, the median and the lower quantile of each group.

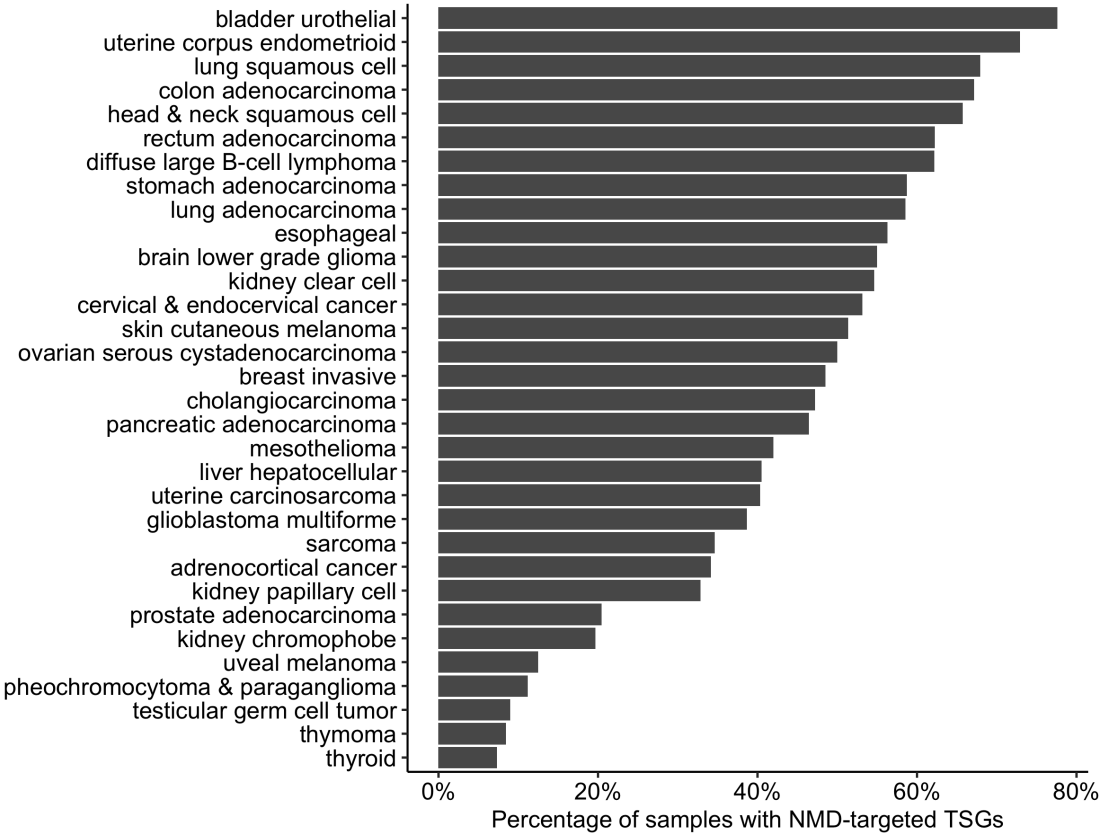

Figure 3.5. Prevalence of TSGs that harbour NMD-elicit mutations.

promising for certain cancer types that are more often affected by NMD.

To eliminate the random effect of mutations, I defined the enrichment of NMD-

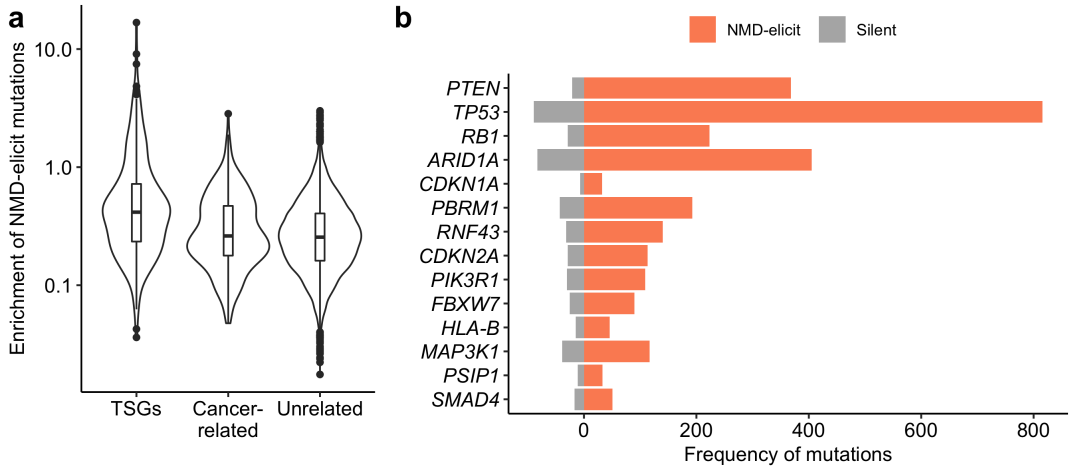

Figure 3.6. The enrichment of NMD-elicit mutations in TSGs.

(a) NMD-elicit mutations are more enriched in TSGs compared to cancer-unrelated genes. The violin plot visualises the distribution of the gene-wise enrichment of NMD-elicit mutations. (b) The gene-wise enrichment of NMD-elicit mutations. The barplot is coloured by the type of mutations, silent or NMD-elicit. The genes are ordered descendingly by the enrichment, calculated as the ratio between NMD-elicit and silent mutations.

elicit mutation as the ratio between the frequency of NMD-elicit mutations and the frequency of background silent mutations. Compared to silent mutations, NMD-elicit mutations were significantly enriched in the annotated TSGs compared to the cancer-unrelated genes (ratio of mean = 0.78/0.32, p-value = 6.711e−08, by Welch Two Sample *t*-test; Figure 3.6a). The following analysis discovered that NMD-elicit mutations were mostly enriched in *PTEN*, *TP53*, *RB1*, *ARID1A* and *CDNK1A* (Figure 3.6b), all of which are annotated as TSGs in COSMIC (release v89, 15th May 2019) [179]. All the evidence suggests that NMD tends to target TSGs in tumours and, consequently, promotes the cancer cell survival.

So far, this study has answered how prevalent the NMD-elicit mutation is and which gene such mutations tend to target. The next step is to study the inter-cancer-type heterogeneity of NMD targets.

I calculated the gene-wise enrichment of NMD-elicit mutations in individual cancer types, showing that the NMD targets were heterogeneous across various cancer types (Figure 3.7). The top 37 NMD targets with enrichment ≥ 1 were selected. Among them, 26 (70%) were annotated as TSGs by COSMIC, indicating that the majority of cancer-specific NMD targets function as tumour suppressors. For instance, two well-known TSGs, *TP53* and *PTEN*, were the pan-cancer NMD targets. Additionally,

ATRX was frequently targeted by NMD in lower grade glioma and sarcoma, and *APC* in rectum and colon cancer. However, seven of the NMD targets were not annotated as cancer-related genes in COSMIC. For example, *BRD7* was the NMD target in liver cancer and melanoma, and this gene was reported to regulate cell cycle [210]. It implies that this gene is likely related to the development of liver cancer or melanoma, although this relationship has not been catalogued by the public database.

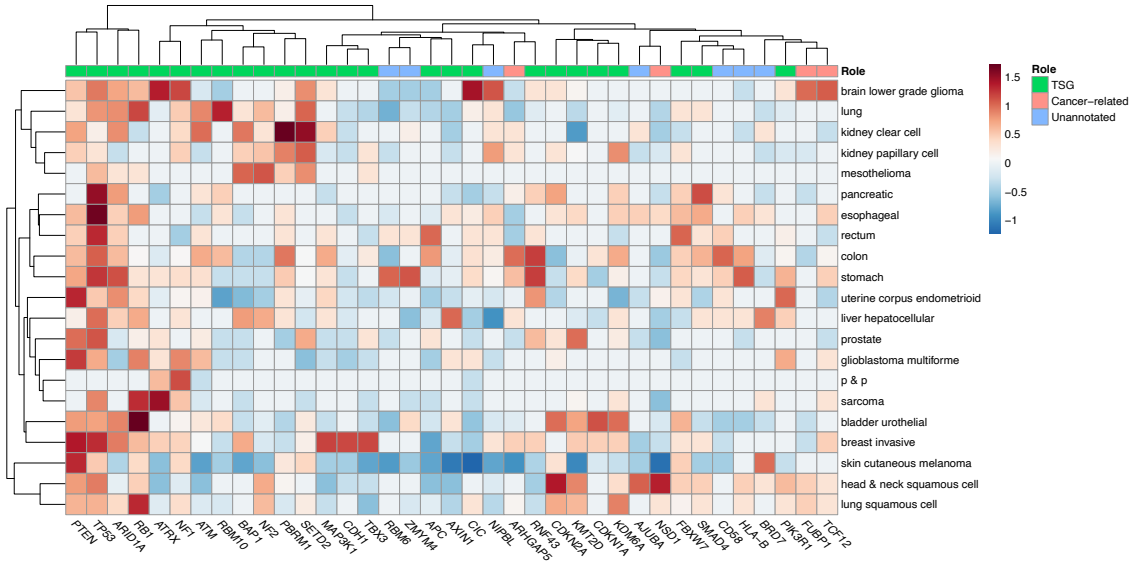

Figure 3.7. Overview of the cancer-specific NMD targets.

The rows are cancer types and the columns are genes. The fill gradient represents the enrichment of NMD-elicit mutations, aka the ratio between the NMD-elicit mutations and the silent mutations. Only genes and cancers with enrichment ≥ 1 are shown here. The columns and rows are hierarchically clustered by the complete linkage method.

3.3.5 NMD-elicit mutations are associated with hypermutation

In Section 3.3.2, this study showed that the NMD-elicit mutations were enriched in certain cancers, such as uterine, colon and stomach cancers. Intriguingly, these cancer types are also prone to develop the hypermutation phenotype, in which one tumour harbours thousands of somatic mutations. This study next focused on the five cancer types with more than 20 hypermutated samples (mutation load over one thousand): colon adenocarcinoma, lung adenocarcinoma, skin cutaneous melanoma, stomach adenocarcinoma and uterine corpus endometrioid carcinoma.

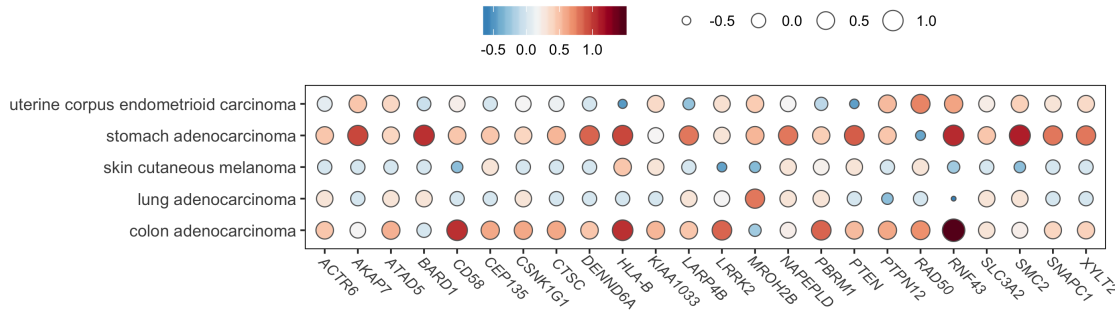

Figure 3.8. Difference in the gene-wise enrichment between hypermutated and hypomutated samples.

The size and the colour of each dots represent the difference in the gene-wise enrichment. The larger the value is, the more specifically the NMD-elicit mutations are enriched in the gene (column) in a hypermutated-specific manner. That is, the red colour means that the certain gene (column) is more frequently targeted by NMD in hypermutated samples than the hypomutated ones.

To study the question which gene is more likely to be targeted by NMD in the hypermutated samples, I compared the gene-wise enrichment of NMD-elicit mutations between the hypomutated (number of somatic mutations $[N] < 100$) and the hypermutated ($N \geq 1000$) samples. The analysis showed that the NMD-elicit mutations were more abundant in 1560 genes within the hypermutated samples than the hypomutated ones. These hypermutation-related NMD targets were enriched in the pathways of cell cycle process, chromosome organisation and DNA repair (FDR < 0.003). I infer that the NMD-elicit mutations lead to the dysfunction of these pathways, which may consequently induce or exacerbate the mutator phenotype.

Furthermore, out of the top 24 hypermutation-related NMD targets (difference $\geq$ 1.2), only five genes have been annotated as TSGs, namely *BARD1*, *LARP4B*, *PBRM1*, *PTEN* and *RNF43* (3.8). As for two examples in the unannotated genes, *RAD50* and *PBRM1* are involved in double-strand break repair and chromatin remodelling respectively [51, 211]. It implies that NMD in the hypermutated tumours is inclined to silence the genes that play a role in the DNA repair or chromatin organisation.

3.4 Summary

The algorithm presented in this study, Masonmd, efficiently identified over 140 thousand putative NMD-elicit somatic mutations in the TCGA cancer genomes, which was validated by the significant decrease in the mRNA levels. It discovered that 60K (two-thirds) frameshift indels were NMD-elicit and that 26K (one-third) nonsense SNVs were NMD-escape.

In addition to the technical annotations, my analysis reveals that NMD has the tendency to repress the mRNA levels of TSGs. Moreover, this study found that the expression of genes related to DNA repair was inhibited by NMD in the hypermutated tumours. Although more experimental work is needed to validate this causality, my results demonstrate that NMD is associated with the repression of TSGs in tumour samples, suggesting that the NMD pathway may play a key role in cancer development.

The systematic findings in this study support the previous assumption that the NMD inhibitors or the read-through chemicals can potentially be used to treat cancer [4, 107, 153]. Those molecules can repress the NMD pathway and re-express the truncated or even full-length tumour suppressors in tumour cells, which can kill the cancer cells. Alternatively, inhibiting NMD can result in the expression of neo-antigens, which can improve the efficacy of immunotherapy [142]. The major limitation of using NMD inhibition as a therapeutic strategy is the possibility of re-expressing dominant-negative mutated proteins, which can be deleterious to other normal cells in the body. This limitation can potentially be overcome by limiting NMD inhibition to cancer cells (target therapy) or only applying this to patients without potential hazardous truncated proteins in normal cells (personalised medicine).

This study has two major limitations though. First my analysis was based on the level 3 public datasets that did not contain specific sequence information. Thus, this work could not classify the splice-site mutations or translation start site mutations, even though they may induce alternative splicing to cause intron retention or NMD. Secondly, the study did not include other mRNA surveillance systems, such as the

non-stop decay [197], of which the rules have not been well studied. However, as shown in Figure 3.3, the proportion of non-stop mutations was much lower than the NMD-elicit mutations. Thus, the omission of those non-stop mutations is unlikely to compromise the conclusions presented in this study.

Overall, this study illustrated that NMD may assist in the cancer progression by frequently interfering with the expression of tumour suppressors and other cancer-related genes. It clarifies that the NMD pathway could be a promising target for cancer therapy.

Chapter 4

ClinCluster: a novel clustering method for clinical single-cell RNA-seq data

4.1 Abstract

Single-cell RNA sequencing (scRNA-seq) datasets that are produced from multiple human samples are usually confounded by batch effects and inter-patient variability. Existing batch effect removal methods require strong hypotheses of the identical composition of cell populations across all patients. Here I present a clustering strategy, ClinCluster, to overcome inter-patient variability based on the differential expression analysis. ClinCluster was benchmarked with other scRNA-seq clustering methods and a bench correction method by using both simulated and real datasets. The results demonstrate that ClinCluster outperformed other methods in datasets that are confounded by batch effects and inter-patient variability.

4.2 Introduction

Clustering is a primitive intuition for human beings. It is a process of grouping items into catalogues. An interesting example is the Chinese literature titled Er-ya that can be traced back to 1100 - 200 BC, in which the authors not only documented

hundreds of plants and animals but also classified them based on their characteristics. For instance, the plants were classified to herbaceous and ligneous plants, and the ligneous plants were further grouped into arbour, shrub and palm, while the animals were divided into insects, fish, birds and mammal. Interestingly, the birds were defined by animals with two feet and feathers; the mammals were defined by ones with four feet and fur.

Twenty-two centuries have passed, we are still following this notion, but with more refined systems. For example, the living beings are catalogued by the modern system of Kingdom, Phylum, Class, Order, Family, Genus and Species. Within one living being such as a human being, the cells can be grouped into diverse cell types, including epithelial cells, blood cells and adipocytes, based on their morphology and biological functions. Unprecedentedly, we have entered a new era to more accurately classify the cell types owing to an emerging technique.

4.2.1 General steps of clustering scRNA-seq data

Single-cell RNA sequencing (scRNA-seq) has been used in a rapidly increasing number of biological studies from a variety of fields, including cancer research [110, 143, 154, 183], cell communication [193], development [71, 188] and neuroscience [43]. Given such widespread application of this technique, the analysis and interpretation of scRNA-seq data are important.

Different cell types serve diverse functions by differentially expressing certain genes at a measurable level, while these differentially expressed genes (DEGs) are related to specific biological processes. Based on this notion, scRNA-seq is used to profile cell populations and discover new cellular subtypes in the heterogeneous systems [30]. The process, in which single cells are classified into populations based on their transcriptomes, is termed as clustering.

Clustering is an unsupervised statistical learning technique. It can effectively partition cells into populations by exploiting the concept of similarity. In principle, cells with highly similar expression profiles are assigned to the same group, while cells with dissimilar profiles go to separate groups.

To prepare for the clustering analysis, it is necessary to preprocess the sequencing data. Raw sequencing data, such as the FASTQ file, is processed to generate a count matrix with rows as genes (or transcripts) and columns as cells. (Figure 4.1). The first thing after generating the count matrix is filtering out the low-quality cells. The filtering step is based on certain criteria, such as sequencing depth and gene coverage. Afterwards, the data matrix is normalised to eliminate the unwanted variability caused by inconsistent sequencing depth. To improve the signal to noise ratio, we will identify the high variance genes (HVGs) from the data matrix. Only the HVGs are kept for downstream analysis. The data then undergoes further dimensionality reduction, which can improve the computational efficiency of the clustering analysis.

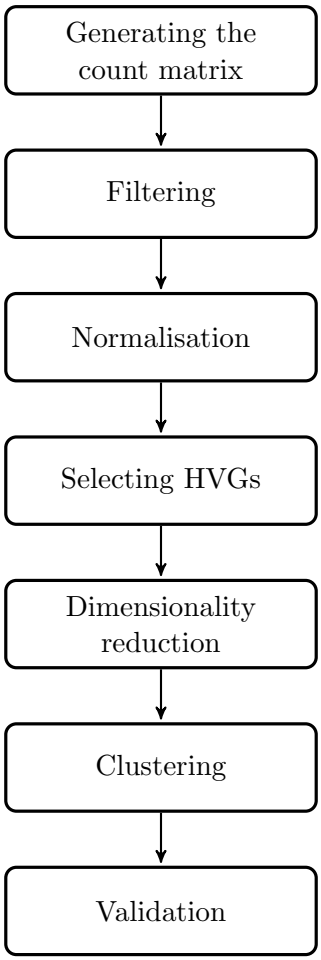

Figure 4.1. Flowchart shows the pipeline of clustering scRNA-seq data. *HVGs, high variance genes. See text for more details.*

After these preprocessing steps, the selected clustering algorithm is applied on the data matrix to catalogue cells into groups (or populations, clusters). A common

way of describing cell populations is using marker genes. The term marker gene refers to a gene that is specifically expressed or significantly upregulated in the given population compared to other populations. The identified marker genes can be put into the pathway analysis, which can reveal the biological functions of this cellular population.

Last but not least, it is essential to verify if the analysis result is reproducible and if the *in-silico* clusters can genuinely represent the *in-situ* cell populations. The scRNA-seq data and the computational analysis are not free from artefacts. For example, the clustering algorithm may force one cell population into multiple clusters (over-clustered) due to technical variance. Moreover, the dissociation of bulk samples can trigger the expression of early response genes, such as *FOS* and *JUN* [136].

To test the reproducibility, there are at least three approaches. First, researchers can use alternative tools to reanalyse the data. Second, performing scRNA-seq on multiple samples can tell if the discovery is robust from sample to sample. Another way is to use another platform. For instance, if a discovery is based on the data generated by the 10x pipeline, researchers can use Smart-seq2 or single-cell qPCR to confirm it.

The validation can be performed in two ways. The first one is to verify the existence of cell populations by using RNA *in situ* hybridisation (e.g. RNAscope) and protein-based assays (e.g. immunofluorescent staining, immunohistochemistry and fluorescence-activated cell sorting), even though they are not on the same scale of RNA-seq in terms of gene coverage. The other type of validation tests the function of a newly discovered cell type or a new marker gene by using functional assays.

4.2.2 Miscellaneous clustering algorithms

Here I will introduce several clustering algorithms tested in this study.

k-means clustering is one of the most classic and common approaches. It was first brought up by MacQueen in 1967 [116]. An important concept used by k-means

clustering is the centroid, which refers to the average (or the centre) of a group of observations. After the number of clusters (k) is designated, the algorithm randomly picks up k data points as the initial centroids and then optimises the centroids by regrouping the points and rechoosing the centroids in a stepwise manner. When the centroids stop changing, the grouping pattern is the final clustering result.

Hierarchical clustering is also used frequently. It consists of two types of algorithms, namely agglomerative and divisive. They differ in the way of constructing the hierarchical tree. The agglomerative hierarchical clustering starts with individual points, merges them in a series of steps based on the similarity matrix, and eventually reaches a single large cluster containing all data points. Oppositely, the divisive method starts with one group containing all data points and divides the existing groups gradually until each group only has one point. To partition data points into groups, a cutoff will be assigned to cut the trees at a certain height (distance).

Another two clustering algorithms were proposed more recently. *Spectral clustering* uses the graph theory to extract information on how to segment the data points [83, 133, 196]. It contains three steps. The first step is to construct a similarity graph. There are many options, such as the k-nearest neighbour graph or the fully connected graph based on various similarity functions (e.g. the Gaussian kernel). In the second step, the data is embedded into a low-dimensional space by using the top eigenvectors of the Graph Laplacian, in which the structure of clusters becomes more evident. In the last step, the embedding was partitioned into groups by k-means clustering. Compared to the k-means clustering, spectral clustering does not require assumptions on the structure of data or the shape of clusters.

Affinity propagation (AP) is based on the idea of exchanging messages between data points [54]. The messages passed between two data points contain the information about whether one point is suitable and available to be the exemplar of the other point. Unlike the k-means clustering, AP does not need random initial centres. It updates the selection of exemplars iteratively until convergence. The clusters are decided based on the final exemplars. This clustering method is suitable for the dataset containing a large number of clusters.

These clustering methods have their advantages and disadvantages [145]. First, k-

means and hierarchical clustering have higher scalability than spectral clustering, while AP is not scalable with the number of samples. These methods are suitable for diverse scenarios. For example, hierarchical clustering and AP perform well with large numbers of clusters, while k-means and spectral clustering are better for fewer numbers of clusters. Furthermore, whilst AP can handle non-even cluster sizes well, k-means and spectral clustering cannot. In the Results section, I will benchmark these clustering methods by using simulated scRNA-seq datasets.

A wide range of clustering pipelines has been developed for scRNA-seq data [94]. Basic techniques, such as k-means, hierarchical and graph-based clustering, have been frequently incorporated into these pipelines. For example, Seurat measures the similarity between data points by using the shared nearest neighbour [201] and assigns cells into groups by using the graph-based clustering. In RaceID3, large populations are detected by using k-medoids clustering, in which the gap statistic is used to determine the number of clusters; rare populations are identified by examining the outliers [57, 68]. SC3 uses consensus clustering to summarise the results of multiple k-means clustering [95]. However, these existing clustering pipelines were designed for general scRNA-seq data rather than specifically for clinical scRNA-seq data [94].

4.2.3 Challenges in analysing clinical scRNA-seq data

Clustering enables the identification of cellular heterogeneity, but it has been confronted with challenges. First, scRNA-seq data can be represented by a high-dimensional count matrix, which usually has tens of thousands of rows (genes or transcripts) and thousands to millions of columns (cells). This can cause a problem known as the curse of dimensionality [12]. Second, the count matrix is highly sparse, in which over 80% of the entries are zeros [150]. The sparsity is due to either the absence of the transcript or the technical dropout. The term, dropout, means that this transcript is expressed in the cell but not captured in the sequencing data owing to technical limitations. Finally, unlike cell lines and mouse samples, scRNA-seq data from clinical samples is usually confounded with batch effects and inter-patient variability. If clinical samples are processed in separate batches, batch effects will

be introduced into the data. On the other hand, the inter-patient variability usually arises from the diverse genetic background across patients.

Normalisation and batch effect correction methods are often used to eliminate the batch effects. Normalisation approaches, such as TMM (the weighted trimmed mean of M-values) [161] and upper quartile, can correct the batch effects caused by the inconsistency in sequencing depth.

Another approach is the batch effect correction, such as mutual nearest neighbours (MNN) [60]. This method is based on two assumptions. One is that at least one population is shared between two batches. Secondly, batch effects and inter-population variability are orthogonal to each other. By computing the anchor vector that connects the subspaces of two batches, MNN can remove the batch effects and merge the data.

Unlike batch effects, inter-patient variability only affects a group of genes, not the overall sequencing depth. Thus, normalisation is not suitable to deal with clinical scRNA-seq data.

Moreover, the inter-patient variability is not necessarily orthogonal to the patient-specific subspaces, which will lead to the incorrect removal of biological variability of interest. No existing algorithm has been developed to specifically address the inter-patient variability.

4.2.4 Differential expression analysis

The purpose of differential expression (DE) analysis is to identify genes or transcripts that are expressed at significantly different levels between two conditions. Computational approaches for DE analysis have been developed for bulk RNA-seq data [173] (e.g. edgeR [162], DESeq2 [114] and limma [158] and scRNA-seq data (e.g. MAST [53]).

Recently Soneson and Robinson [174] carried out an empirical study, in which they compared 36 types of DE analysis tools for scRNA-seq data. Their criteria consisted of the ability to control false discovery rate, computational efficiency, scalability and

sensitivity. In the comparative study, they introduced a method, termed as detection rate (DetRate), in which they used the proportion of genes detected per cell as a covariate in the DE analysis model. According to their results, the top-ranking methods are edgeR/quasi-likelihood approach (QLF)/DetRate [115], MAST/CPM (counts per million)/DetRate [53], limma-trend [103], MAST/TPM (transcripts per million)/DetRate, edgeR/QLF and limma-voom [103]. Surprisingly the methods for bulk data outperformed the ones that were designed specifically for single-cell data.

As a mature technique for analysing RNA-seq data, DE analysis is intrinsically more tolerant of unwanted variability than clustering, owing to the embedded linear model. By including the batch as a covariate in the model, DE analysis is capable of excluding batch effects. The unique advantage of DE analysis provides insight into an alternative way of addressing the inter-patient variability in scRNA-seq data.

4.2.5 Motivations

In the scRNA-seq clustering pipelines, the commonly used dissimilarity measures, such as Pearson correlation and Euclidean distance, suffer from the interference of batch effects. Such limitation tampers the correct clustering of single-cell transcriptomes. Given the aforementioned advantages of the DE analysis, I proposed a scheme, in which DE analysis is used to compute dissimilarity and hereby cope with the batch effects and inter-patient variability in clinical scRNA-seq data.

4.3 Results

4.3.1 Design of workflow

The DE-based clustering has three steps (Figure 4.2). The first step was termed as *initial clustering*, in which cells were grouped into patient-specific initial clusters. It was followed by the calculation of an inter-group distance matrix based on the similarity of functional behaviours. The third step, termed as *final clustering*, merged the initial clusters based on the inter-group similarity to generate cross-patient final

clusters.

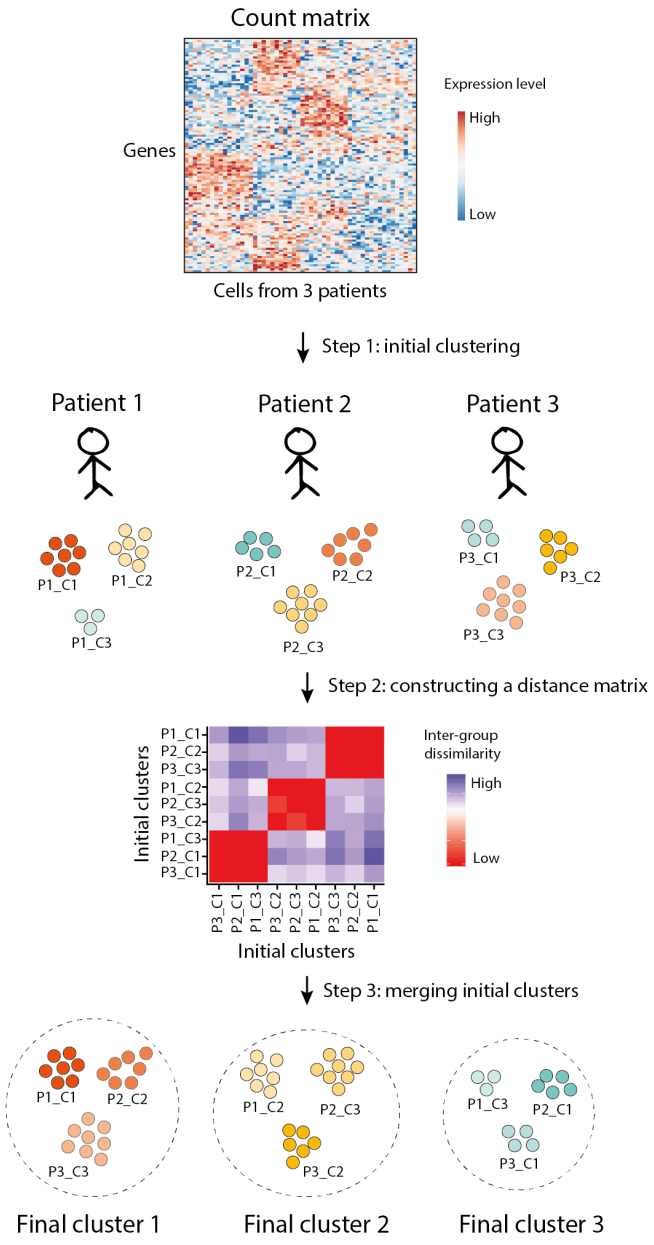

Figure 4.2. Diagram shows the workflow of ClinCluster.

ClinCluster is a three-step clustering algorithm tailored for clinical scRNA-seq data. The input is the count matrix with rows as genes and columns as cells from multiple patients. The first step is called initial clustering, in which the cells are grouped into patient-specific clusters, named initial clusters. In the diagram, the cells (dots) from each patient are catalogued into three initial clusters, e.g. P1_C1, Cluster 1 of Patient 1. The next step is constructing an inter-group dissimilarity matrix (the heatmap), in which rows and columns are the initial clusters and entries represent the inter-group distance as shown in the colour bar. The last step is merging the initial clusters into final clusters (dashed circles) based on the inter-group similarity.

4.3.1.1 Selecting the algorithm for the initial clustering

To generate the patient-specific initial clusters, there were several candidate algorithms that have been introduced in Section 4.2.2, including k-means clustering (R function `kmeans`), Spectral clustering (R package `kernlab` [83]), Spectral clustering based on the k-nearest-neighbour graph (R package `kknn` [64]) and AP (R package `apcluster` [20]).

To select a suitable algorithm, I compared these approaches with the simulated data consisting of only one batch. First a dataset with n genes and m cells was generated by R package Splatter [205]. Before the clustering step, I preprocessed the datasets by performing log-normalisation and HVG selection. Denote the preprocessed expression matrix as M $(k \times m)$, in which the k rows correspond to the number of HVGs and the m columns refer to the number of cells.

Here the adjusted Rand index (ARI) [74, 155, 194] was used to evaluate the performance of clustering algorithms. The ARI is one of the external evaluation measures, which refer to benchmarking predicted clusters with ground truth for the preclassified datasets. In brief, the ARI represents the frequency of correct decisions that a clustering algorithm makes compared with the ground truth (see Methods, Equations 2.2 and 2.3).

The simulation was repeated with different sets of parameters: the number of cells $m = 50, 150, 500$ and the number of true populations $g = 2, 4, 8$. After 20 times of simulation for each set of parameters, the spectral clustering methods and k-means behaved better than AP (Figure 4.3).

To further investigate the advantages and disadvantages of these methods, I plotted out their performance under disparate conditions (Figure 4.4). Although AP underperformed the other three methods, its accuracy increased when there were more populations ($m = 500$, $g = 8$). However, for the small group size and relatively simple composition, AP behaved the worst.

The hypothesis used to build simulated datasets was strong, and it simplified the structure. In clinical data, due to sample degradation or prolonged processing time,

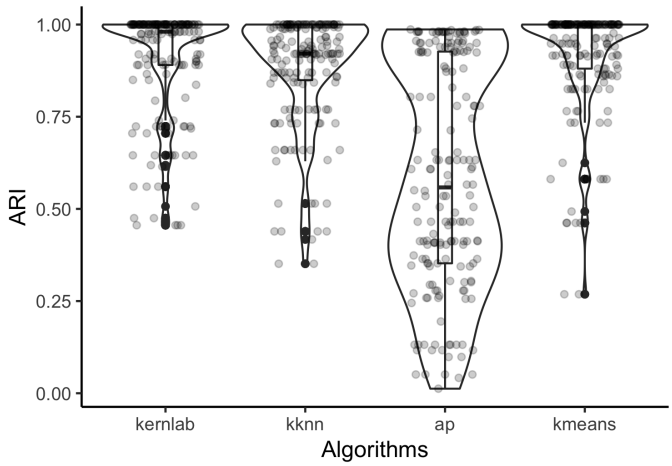

Figure 4.3. Comparison of four methods for the initial clustering.
The violin plots represent the distribution of adjusted Rand index. The upper and lower borders of the boxplots correspond to the upper and lower quantiles respectively, and the bold vertical lines in the middle the medians. Each dot represents the result of one simulation. The x-axis refers to the tested clustering algorithms, and the y-axis the adjusted Rand index (ARI). the ARI is an external evaluation measure of clustering methods. Higher ARIs correspond to better performance.

the data may suffer from a higher rate of outliers. Therefore, I decided to include this feature when simulating clinical-like datasets. In this scenario (ratio of outliers = 0.2 compared to the default 0.05), all the four methods performed worse (Figure 4.5); The ARIs dropped for all methods compared to Figure 4.4. AP still had the lowest ARI, suggesting that it is not a suitable method. From another perspective, all methods had the lowest accuracy when there were 8 population in 50 cells. This extreme case could be caused by the the insufficient sample size. After excluding this situation, spectral clustering outperformed the other methods (Figure 4.6). The k-means clustering method had higher variance, and it performed worse for the datasets with small sample sizes ($n = 50$). The spectral clustering function from R package `kernlab` slightly outperformed the `kknn` function.

Therefore, the spectral clustering method `specc` from R package `kernlab` was used for further analysis, because it was stable for simulated datasets with complications, i.e. outliers and limited sample sizes. Yet one thing to keep in mind is that the choice of the initial clustering method will largely depend on the structure of the dataset.

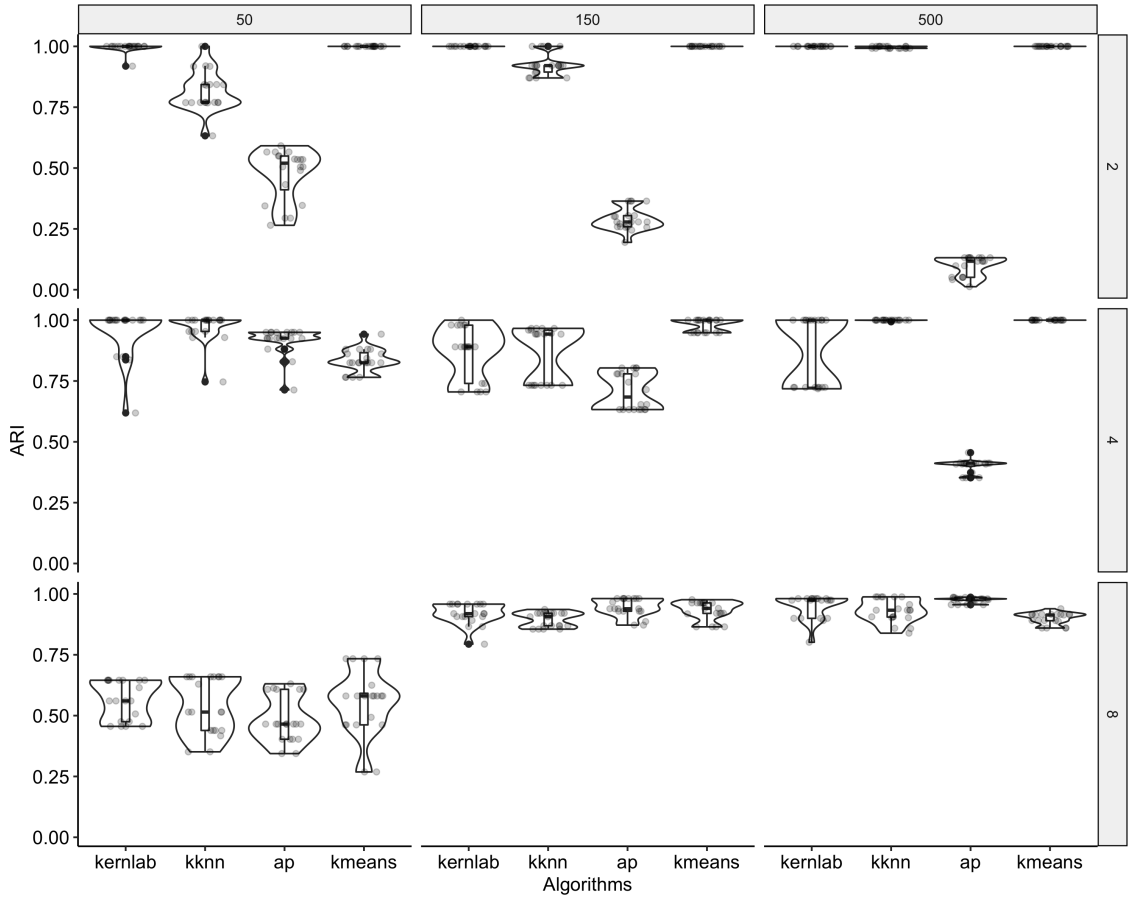

Figure 4.4. Comparison of four methods for the initial clustering under diverse conditions.

This study tested the four algorithms with combinations of two parameters: the number of cells $m = 50, 150, 500$ and the number of true populations $g = 2, 4, 8$. The subfigure panels are arranged by the number of populations (rows) and the number of cells (columns) as marked. The x-axis represents the tested algorithms, and the y-axis refers to the performance of each method in multiple tests (dots). The upper and lower borders of the boxplots correspond to the upper and lower quantiles respectively, and the bold vertical lines in the middle the medians. Each dot represents the result of one simulation. See also Figure 4.3.

4.3.1.2 Constructing the distance matrix

After the initial spectral clustering, each cell was assigned to a patient-specific initial cluster. Following this clustering step, the DE analysis was carried out between each pair of initial clusters across patients. The DE analysis was based on the generalised linear models to evaluate whether the expression levels of a gene were distinguishable between two or more conditions [115]. Based on the strength and the number of DEGs between two initial clusters, the following equation (Equation 4.1) was used to generate the entry, d_{ij}, that represented the dissimilarity between the ith and jth

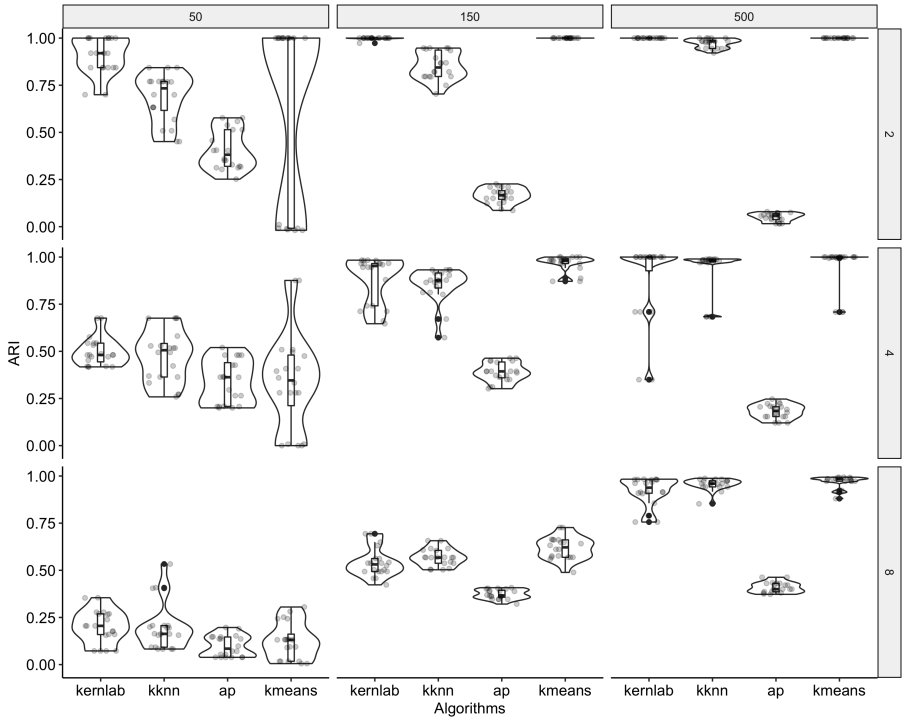

Figure 4.5. Comparison of four methods for the initial clustering in more severe conditions.

The subfigure panels are organised by the number of cells (columns) and the number of groups (rows). The candidates are tested with simulated data with the outlier rate of 0.2. Overall, the spectral clustering still outperforms the other methods by examining the ARI. As for the the outlier rate of 0.05 (default), see Figure 4.4.

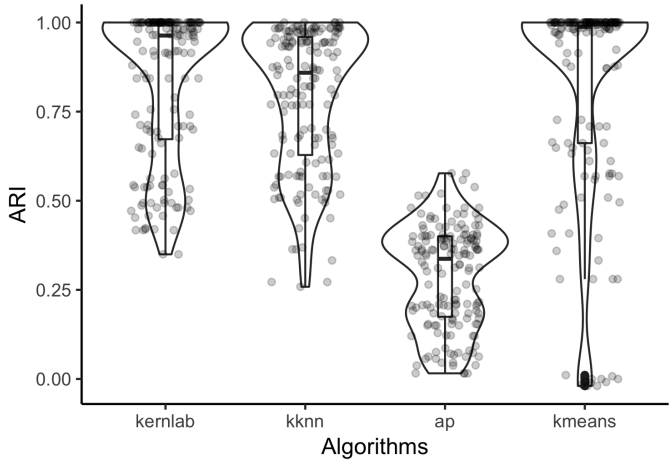

Figure 4.6. Comparison of four methods for the initial clustering in more severe conditions.

This figure summarised all conditions from Figure 4.5, except one ($n = 50$ and $g = 8$). Two spectral clustering algorithms from kernlab and kknn outperform AP and k-means, while kernlab is slightly better than kknn.

initial clusters in a distance matrix.

$$d_{ij} = \sum_{n=1}^{k} \frac{|f_n|(e_{ni} + l)}{(e_{nj} + l)} \tag{4.1}$$

Herein, k refers to the number of DEGs between the ith and jth initial clusters. In the numerator, f_n denotes the $\log_2$ fold-change of the nth DGE, while e_{ni} and e_{nj} denote the ratio of cells respectively in clusters i and j that express the nth DEG. l is a constant (usually $l = 0.01$) to avoid the denominator being zero.

In other words, the numbers of DEGs and the strength of DE represent the functional dissimilarity between two initial clusters. The distance matrix of functional dissimilarity can be described as

$$\mathbf{D} = \{d_{ij}\}, \tag{4.2}$$

in which the dissimilarity d_{ij} is defined in Equation 4.1 between the ith and j th initial clusters.

The values of cut-offs for DEGs should be optimised to maximise the functional dissimilarity across different cell populations and to minimise the dissimilarity caused by structured batch effects. There were three types of parameters to optimise. The first one was the fold-change of expression levels between two clusters. The second was the detection ratio and the third was the false discovery rate (FDR) calculated by the DE analysis. I carried out the DE analyses with different combinations of parameters (Figure 4.7). For the $\log_2$ fold-change between 0.6 and 0.8 and the ratio of expression between 0.25 and 0.75, the dissimilarity between different populations exceeded the one caused by structured batch effects.

To test whether there was any benefit from this DE step, I compared ClinCluster to spectral clustering on the simulated data with three batches (300 cells per batch) and four populations (fractions = 0.23, 0.34, 0.12 and 0.31). When the spectral clustering was directly conducted on this dataset, the clustering results were misled by the batch effects (Figure 4.8). Most obviously, the original Group 3 and Group 4 were shuffled and divided into two new groups. It demonstrated that a method that can cope with batch effects is necessary for this type of single-cell data.

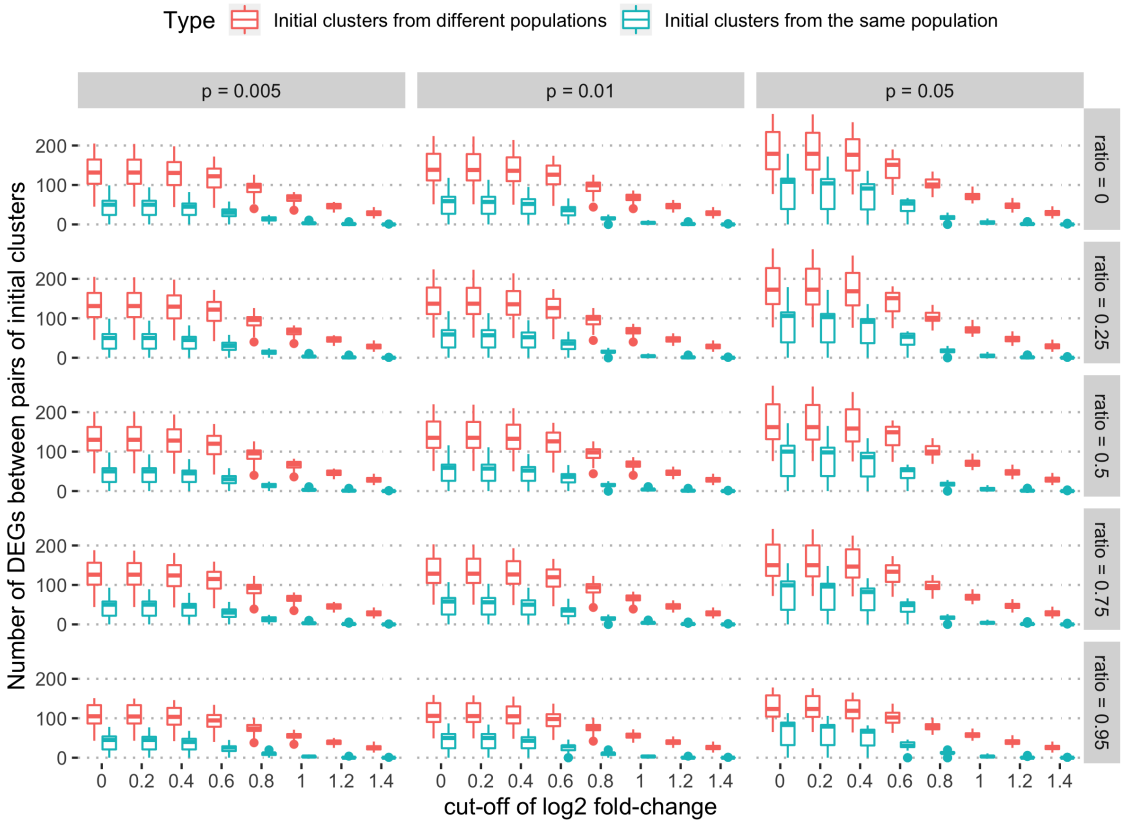

Figure 4.7. Optimising the parameters to maximise biological dissimilarity and minimise batch effects.

The y-axis denoted the numbers of DEGs between pairs of initial clusters. For the red boxes, two initial clusters in a given pair are from the different populations, whereas for the green boxes, both initial clusters are from the same population. In other words, the values of the red boxes represent the biological variability, whilst the ones of the green boxes represent the inter-batch variability. Ideally, the selection of parameters is needed to maximise the biological variability and minimise the inter-batch variability. I tested three types of parameters, p (adjusted p-values, FDR), ratio (the proportion of cells in a given initial cluster that express the DGE) and $\log_2$ *fold-change. Each subfigure panel has a combination of p and ratio as shown in the grey strips. The x-axis denotes the* $\log_2$ *fold-change. For the* $\log_2$ *fold-change between 0.6 and 0.8 and the ratio of expression between 0.25 and 0.75, the dissimilarity between different populations (red) exceeded the one caused by structured batch effects (green).*

As for the step of the DE analysis, there were several options. I compared four top methods from a previous benchmarking study [174], namely edgeR/QLF [115], MAST/CPM [53], limma-trend [103] and limma-voom [103], with a simulated dataset.

First, I applied the initial clustering on this dataset as described above (Figure 4.9). Then a distance matrix (Equations 4.1 and 4.2) among those initial clusters was computed with edgeR, limma-trend and limma-voom. All of these three methods accurately recognised the structure of the initial groups (Figure 4.10). The pairwise

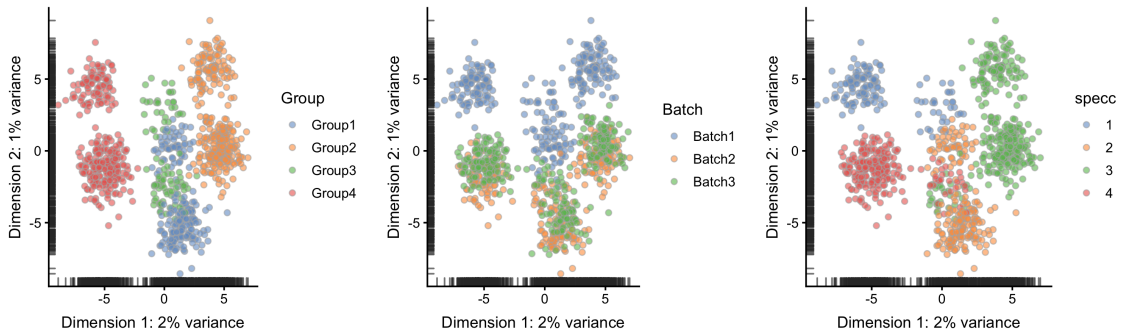

Figure 4.8. Directly using spectral clustering is influenced by batch effects.
Left panel, cells (dots) are coloured by groups. Middle panel, cells are coloured by batches. Right panel, cells are coloured by the clustering results from direct spectral clustering (specc) without the DE step. When the spectral clustering was directly conducted on this dataset, the clustering results were misled by the batch effects (ARI = 0.75). As shown in the right panel, the original Group 3 and Group 4 were shuffled and divided into two new groups. It demonstrates that an additional step, such as DE analysis, is necessary to deal with the batch effects.

distances were close to zero between the initial clusters that belonged to the same population, but over 100 for the ones belonging to different populations. However, one obvious disadvantage of the edgeR method was its calculation speed. To complete the same mission on a MacBookPro (2.8 GHz processor), limma-trend and limma-voom only took 35 s and 45 s respectively, while edgeR took 45 min. Hence, I decided to move forward with limma-trend and limma-voom.

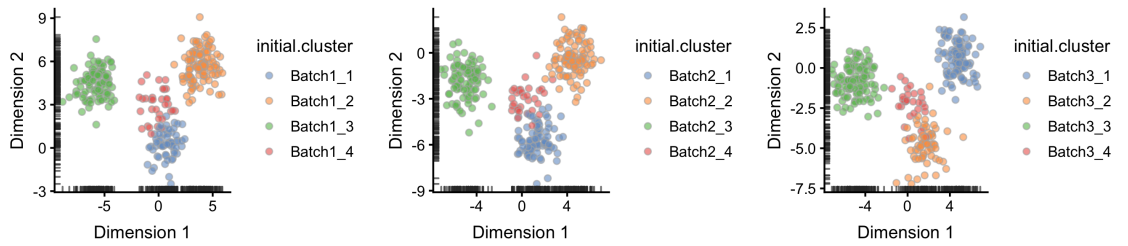

Figure 4.9. PCA plots visualise the initial clusters of a simulated dataset.

After applying the initial clustering on the simulated dataset, cells from three different batches (subfigures) are grouped into initial clusters (colours). The initial clustering result will be the input of the next step (see Figure 4.10).

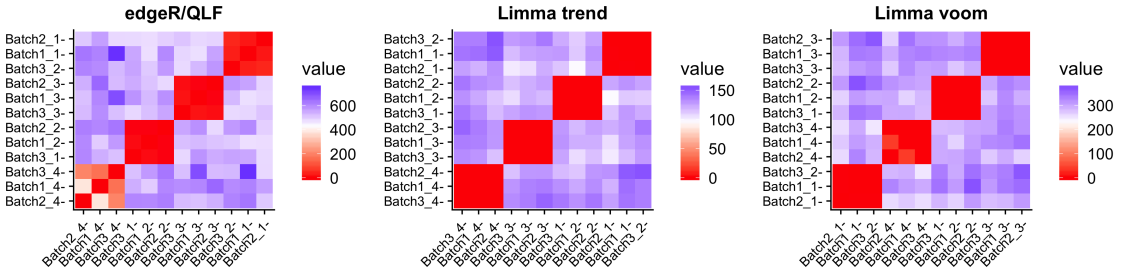

Figure 4.10. All three DE methods recognised the structure underlying initial clusters.

The distance matrix (heatmap) represents the inter-group dissimilarity between the initial clusters shown in Figure 4.9. Each subfigure panel is the result of one DE analysis algorithm as shown in the title. The x- and y-axes of heatmaps denote the initial clusters, while the colour represents the inter-group distance. The pairwise distances were close to zero between the initial clusters from the same population, but over 100 for the ones from different populations.

4.3.1.3 Hierarchical clustering to merge initial clusters

In this step, hierarchical clustering assigned all initial clusters to final clusters based on the distance matrix. Accordingly, each cell was assigned to the corresponding final cluster (Figure 4.2). The hierarchical dendrogram can be plotted to help determine the height to cut the tree.

Following the above example (Figures 4.9 and 4.10), the distance matrices are visualised in the format of hierarchical dendrograms (Figure 4.11). The branching structure can facilitate the decision of tree cutting. For example, in the tree titled limma-trend, the three initial clusters, namely Batch3_4, Batch1_4 and Batch2_4, were assigned into one final cluster, along with the single cells that belonged to those initial clusters. The ARI of the final clustering result was equal to 1. It was better than the result of directly using spectral clustering (ARI = 0.75, Figure 4.8), confirming the essentiality of using ClinCluster for batch-confounding data.

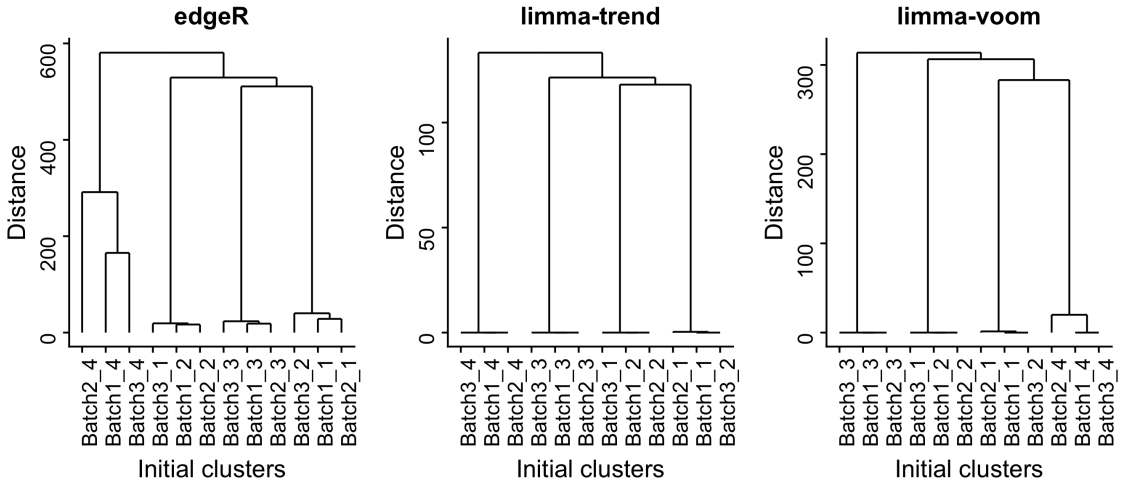

Figure 4.11. Hierarchical dendrograms are constructed based on distance matrices.

The dendrogram visualises the distance matrix among initial clusters (nodes on the x-axis). The y-axis represents the length of edges (distances). I used three DE analysis methods to generate the distance matrices as shown in the three subfigures. Based on the branching structure, we can decide where to cut the trees. For example, in the middle panel (titled limma-trend), the cutoff can be set at around 50 whereby the three initial clusters (Batch3_4, Batch1_4 and Batch2_4) are assigned into one final cluster, along with the single cells that belong to those initial clusters.

4.3.2 Benchmarking ClinCluster on simulated datasets

Next, I benchmarked ClinCluster with other scRNA-seq clustering pipelines. The comparison used datasets generated by R package Splatter [205]. ClinCluster was compared with three commonly used techniques [182]: Seurat [24], RaceID3 [57, 68] and SC3 [95]. Regarding the ability to cope with batch effects, ClinCluster was also compared with a batch correction method, mutual nearest neighbours (MNN), which is reported to outperform other batch correction methods [60].

The results reveal that ClinCluster outperformed other clustering pipelines in the scenario where the data is confounded by batch effects (Figure 4.12). ClinCluster reached the ARIs of over 0.8 in 89% cases (56 out of 63). The accuracy (measured by the average ARI) of ClinCluster was higher than the other four methods (Table 4.1). MNN was second to ClinCluster. It did well in certain situations (e.g. the datasets with 2 groups), but poorly in others (e.g. the datasets with 4 batches and 8 groups).

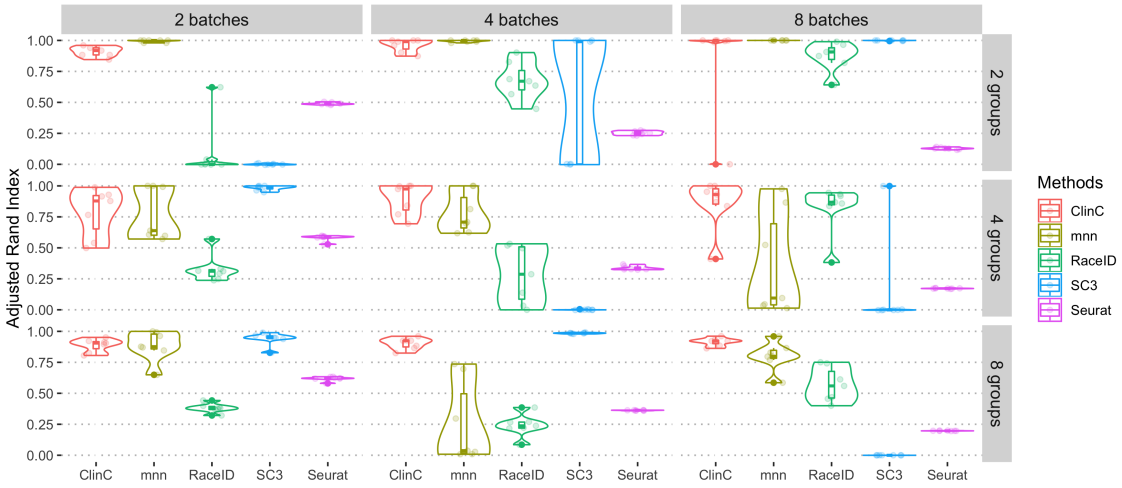

Figure 4.12. ClinCluster outperforms the other methods on multiple-batch simulated data.

This study compared several clustering pipelines (x-axis) with simulated scRNA-seq data under different conditions (subfigures). ClinCluster (ClinC) reached the ARIs of over 0.8 in 89% cases (56 out of 63). Evaluated by ARIs (the y-axis), ClinCluster achieves accuracy prediction in all situations. MNN was second to ClinCluster, and its accuracy was limited with the 4-batch and 8-group datasets.

The stability (measured by the standard deviation of ARIs) of ClinCluster also exceeded the other methods (Table 4.1). Opposite to ClinCluster, the ARIs of

Method	Average ARI	Dispersion ARI
ClinCluster	0.88	0.16
MNN	0.76	0.32
RaceID	0.47	0.30
SC3	0.51	0.49
Seurat	0.35	0.17

Table 4.1. ClinCluster outperforms the other clustering and batch correction methods in simulated datasets.

The average ARI is used to evaluate the accuracy of clustering. Higher average ARIs correspond to higher accuracy. The dispersion (the standard deviation of ARIs) is used as a measurement of the stability. Lower dispersion represents higher stability.

SC3 ranged from 0 to 1 (e.g. in the 4-batch and 2-group scenario) (Figure 4.12), suggesting a lack of stability. There was one outlier (ARI = 0) in the 8-batch and 2-group case for ClinCluster in Figure 4.12, in which ARI was equal to 0. I looked into this case and found that spectral clustering picked up the wrong initial grouping. It suggested that there was room for improvement of the initial clustering algorithm to reach higher robustness.

I assumed that the poor performance of other clustering pipelines was caused by the batch effects. To test this assumption, I calculated the ARI between the clustering results and the batches (rather than the populations), which was used to measure the extent to which the clustering results were impacted by inter-batch variability. The result showed that the other three clustering methods were influenced by batch effects (Figure 4.13). For instance, the result of Seurat was influenced by the batch effects under all situations (Figure 4.13). It verified the hypothesis that batch effects interfere with the accuracy of other clustering pipelines.

I also compared the time efficiency of these methods. ClinCluster was comparable with MNN and RaceID. All of them were better than SC3 (Figure 4.14). Because ClinCluster involved pairwise comparisons between initial clusters, its running time increased with the number of initial clusters ($O(n^2)$, where n is the number of initial clusters). The running time of MNN, RaceID and SC3 was associated with the

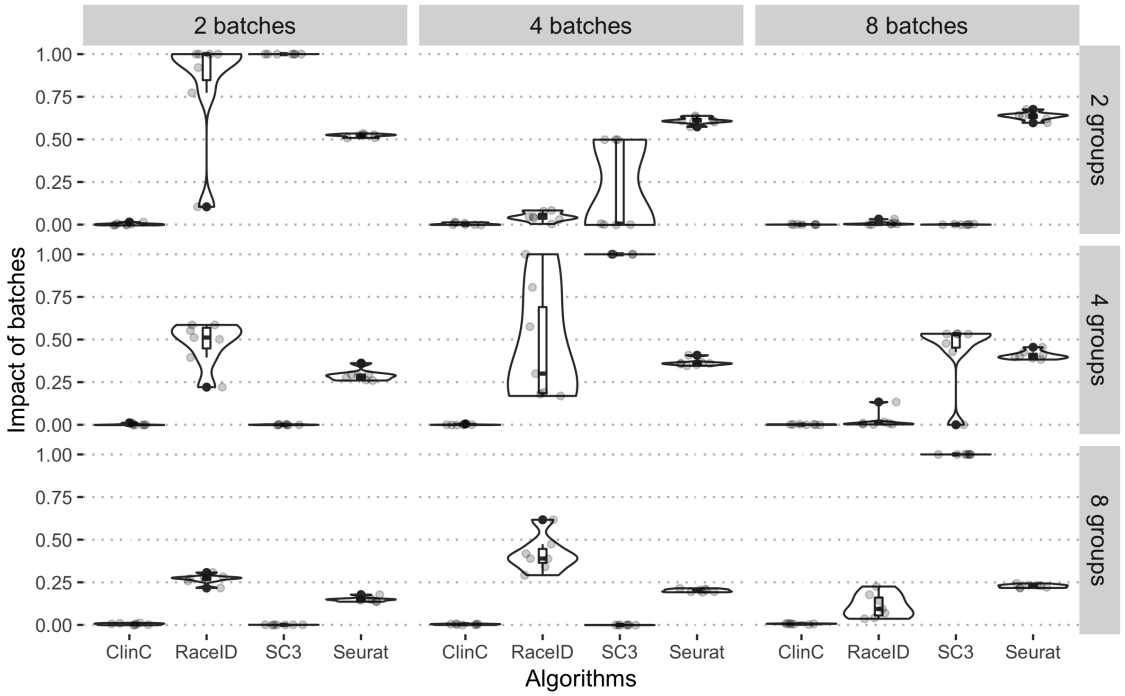

Figure 4.13. Other clustering methods are swayed by batch effects.

To measure the impact of batch effects, the ARI was computed between the clustering results and the batches (rather than the populations). The y-axis represents the extent to which clustering results are impacted by inter-batch variability. Unlike other clustering pipelines, ClinCluster (ClinC) is free from the impact of batch effects under miscellaneous scenarios (subfigures).

number of total cells, while Seurat was not affected by the data size.

Overall, the results show that ClinCluster can overcome the batch effects in a range of simulated datasets.

4.3.3 Benchmarking on the clinical dataset

I next evaluated ClinCluster with a clinical dataset of patients' melanoma samples [183]. For clinical samples, an outlier mode was introduced. It was assumed that the small sample size of scRNA-seq could cause a higher false positive rate for the DE analysis, which could jeopardise the structure of the hierarchical tree. Therefore, I set up an optional mode, in which the small initial clusters were excluded from the construction of hierarchical trees.

I tested the ClinCluster pipeline with or without the outlier mode on the melanoma dataset (Figure 4.15). When the small-size initial clusters ($n \leq 10$, red labels in

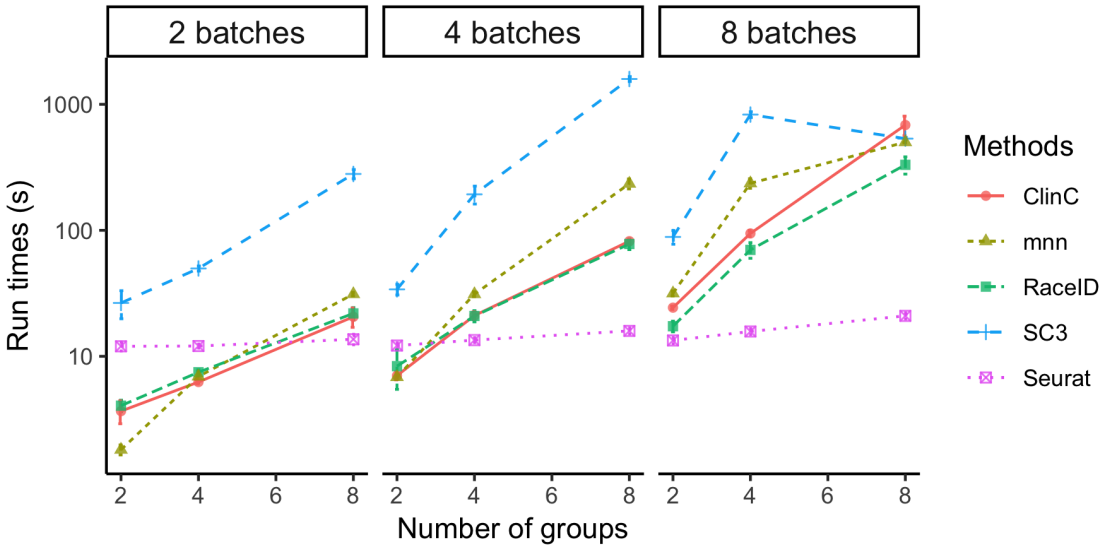

Figure 4.14. Time efficiency of clustering methods.

The time efficiency of various clustering pipelines (shown in the legend) was tested. The x-axis denotes the number of groups in a dataset, and the y-axis shows the run time of each method. The subfigures represent different numbers of batches. Seurat has the highest efficiency, while SC3 has the lowest. The performance of ClinCluster, MNN and RaceID is between those two extreme cases.

Figure 4.15a) were included, these clusters formed a structure, in which the outlier group did not have any mutual nearest neighbour. These outlier groups can make the final clustering less accurate. For example, the existence of initial cluster Mel60_2 hindered the correct partition between two clusters (coloured by purple and pink in Figure 4.15a). Under the outlier mode, the small-size initial clusters were excluded from the hierarchical tree at first (Figure 4.15b) and then merged into the major clusters. It ensured that the final clustering step would not be influenced by the outliers.

I compared the results of ClinCluster (outlier mode) with other four clustering pipelines as well as the spectral clustering. In Tirosh et al.'s study, 2223 cells from seven donors were classified into 7 groups (Figure 4.16a and b).

- Group 0, unclassified cells
- Group 1, T cells
- Group 2, B cells
- Group 3, macrophages

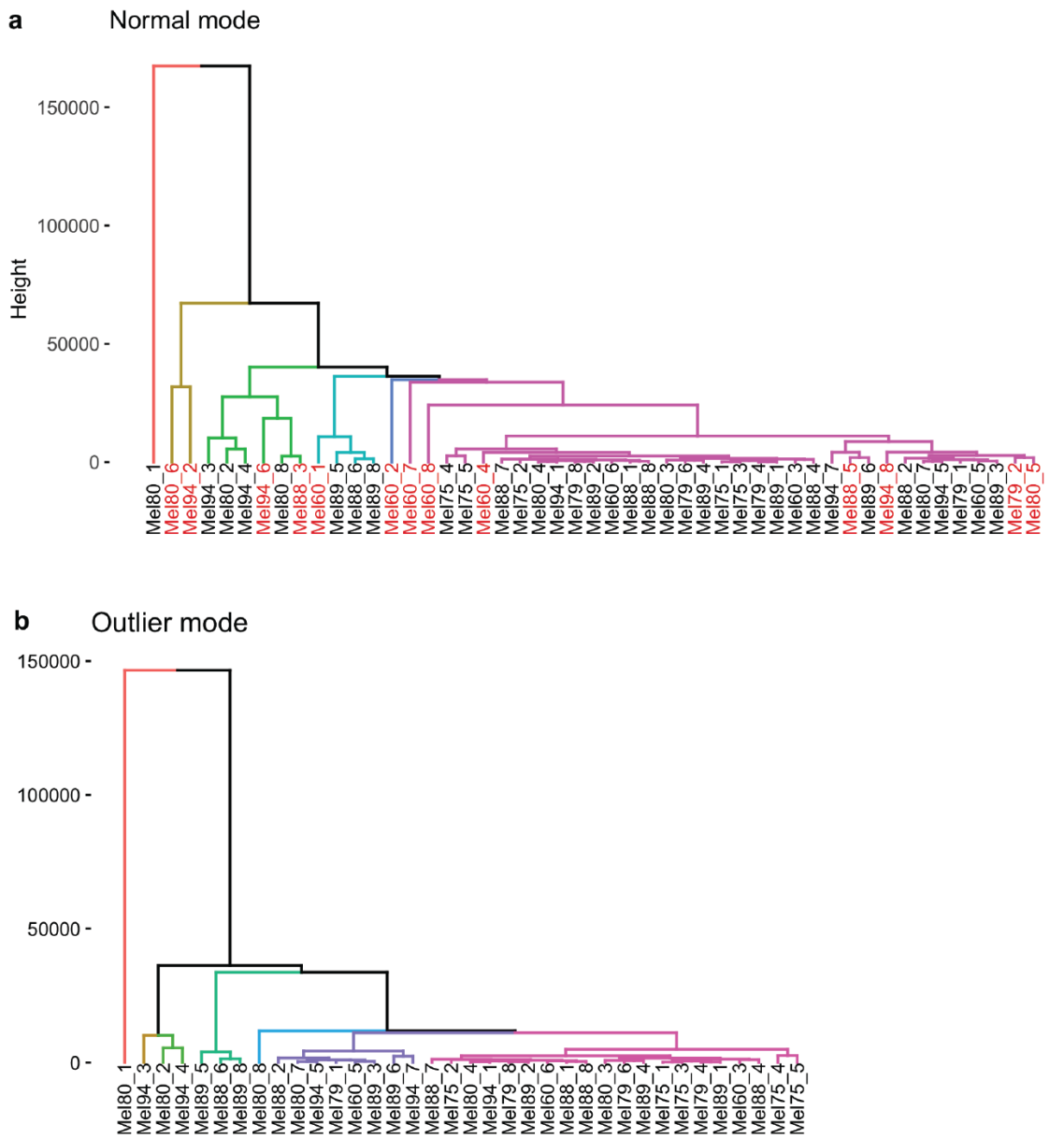

Figure 4.15. The outlier mode restores the correct tree structure.

For the real dataset, the outlier mode was introduced, in which the small initial clusters (red labels in Subfigure a) were treated as outliers in the last step. The colours of branches represent the final clustering results. (a) The dendrogram shows the hierarchical tree built under the normal mode. The red labels denote the small-size initial clusters. Some of them, such as Mel60_2 and Mel60_7, form a singlet node with no mutual closest neighbours. (b) The dendrogram shows the hierarchical tree built under the outlier mode, in which the small-size initial clusters were excluded from it. It facilitates the selection of a cutoff. In the subfigure (a), the existence of initial cluster Mel60_2 interferes with the correct partition between two clusters (coloured by purple and pink).

- Group 4, endothelial cells
- Group 5, cancer-associated fibroblasts (CAFs)

- Group 6, natural killers (NKs)

By using the original annotations as the ground truth, the ARIs were computed to evaluate the accuracy of clustering approaches. The comparison revealed that ClinCluster outperformed other clustering approaches. It correctly identified T cells (C1), B cells (C2), endothelial cells (C4), CAFs (C5) and macrophages (C6). The NK cells were assigned to the T cell population, due to the small difference between these two cell types (Figure 4.16b). Compared to the original annotation, ClinCluster successfully assigned the unclassified cells to the individual groups (Figure 4.16c).

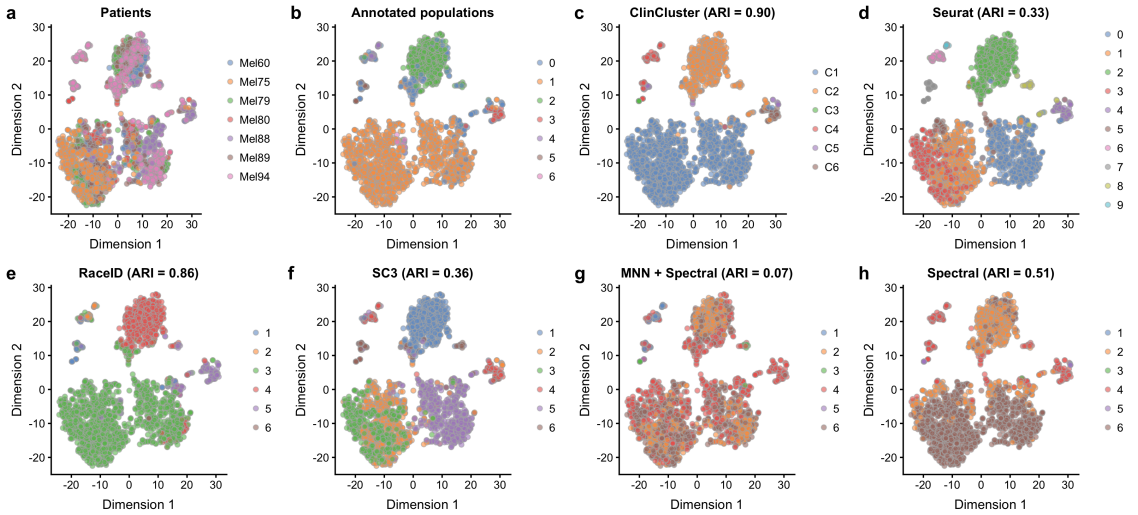

Figure 4.16. ClinCluster outperforms other methods in the melanoma dataset. *The x- and y-axes denote the first two components of t-distributed stochastic neighbour embedding (t-SNE), which is a dimensionality reduction technique. Each dot represents a single cell. (a) Cells are coloured by patients. (b) Cells are coloured by the original annotations (ground truth). (c-h) Cells are coloured by the results of miscellaneous clustering pipelines: ClinCluster (c), Seurat (d), RaceID (e), SC3 (f), MNN + spectral clustering (g) and spectral clustering (f). Evaluated by the ARI, ClinCluster has the highest accuracy by correctly identifying T cells (C1), B cells (C2), endothelial cells (C4), CAFs (C5) and macrophages (C6).*

The second best one was RaceID. It also identified the main populations but misclassified some B cells (RaceID/Cluster 4) and T cells (RaceID/Cluster 3) (Figure 4.16e). SC3 failed in separating macrophages and endothelial cells. Besides, it divided the T cells into three sub-groups, one of which (SC3/Cluster 3) is a patient-specific cluster (Figure 4.16f). This cluster contained 311 cells from only one patient, Mel75. It suggests that SC3 over-clustered the T cells due to inter-patient variability. The results of Seurat also contained a patient-specific cluster Seurat/Cluster 3, in which

243 out of 250 cells were from Mel75. The impact of inter-patient variability (IPV), which referred to the ARI computed between patients and predicted clusters, also indicated that Seurat and SC3 were affected by inter-sample variability (IPV = 0.19 for SC3 and 0.13 for Seurat, compared to IPV = 0.05 for ClinCluster). When the spectral clustering was directly applied on the data, it could not identify the main populations owing to the interference of outliers. For instance, spectral clustering picked up two small-size clusters (Spectral/Cluster 3 and Spectral/Cluster 5) that consisted of outliers (Figure 4.16h).

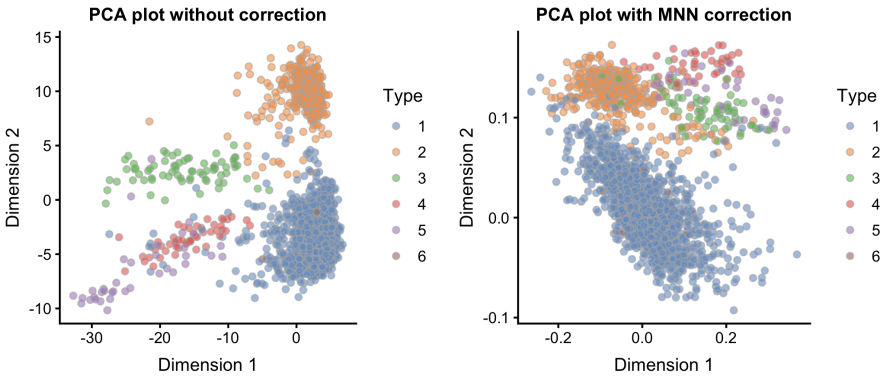

Figure 4.17. PCA plots before and after the MNN corrections.

The x- and y-axes represent the first two principal components. The comparison between the before- and after-MNN data reveals that the discrepancies between cell populations are shrunk by the MNN correction. For instance, the difference between Cluster 3 and 5 was evident in the left panel (uncorrected data), but vague in the right panel (correct data). The lose of key information hinders the correct partition of data points.

The combination of MNN and spectral clustering also did not work well (Figure 4.16g). By comparing the PCA results before and after the MNN correction, the analysis revealed that the MNN correction shrank the discrepancies between cell populations (Figure 4.17). I inferred that the nonorthogonality between the inter-patient variability and the patient-specific biological subspace did not meet the orthogonality prerequisites of the MNN approach [60]. In other words, MNN was tailored to cope with batch effects and technical variability rather than inter-patient variability. Among these methods, ClinCluster can most effectively overcome inter-donor diversity.

4.4 Discussion

Clinical scRNA-seq datasets are often confounded by both batch effects and inter-donor variability, which interfere with the accurate interpretation of data. The batch effects are usually caused by technical factors; for example, different lots of reagents are used. However, the inter-donor variability has not been fully studied. The inter-donor variability can be caused by diverse genotypes or other factors, such as microenvironment or epigenetic changes.

To address the confounding factors in clinical scRNA-seq, this study presented a clustering method, named ClinCluster, which was tailored for clinical scRNA-seq data. ClinCluster was based on the hypothesis that the inter-population discrepancy is measurably larger than the inter-batch and inter-patient variability. By benchmarking ClinCluster with other existing clustering and batch correction techniques, this study demonstrates that ClinCluster can surmount the confounding factors and recognise the cross-donor populations.

ClinCluster is designed for clinical samples. On the other hand, it is unnecessary to use ClinCluster for the datasets without inter-batch confounding factors. As the rule of thumb, the selected clustering algorithm should be suitable for the given data type. To decide if ClinCluster fits a given dataset, multiple clustering methods can be applied on it first. If the result of other clustering methods is disturbed by batch effects (e.g. IPV > 0.1), ClinCluster will provide a more accurate result.

The advantage of ClinCluster is that it does not need harsh simplifying hypotheses compared to other batch correction methods. Although MNN can successfully remove technical confounding factors [60], it requires orthogonality between batch effects and inter-population variability. However, inter-patient diversity is not always orthogonal to patient-specific subspaces (Figure 4.17).

The first limitation of ClinCluster is the lack of scalability. Scalability is an important feature for the scRNA-seq analysis tools because the throughput of scRNA-seq experiments is exponentially increasing. More optimisation work will be done to make ClinCluster suitable for large datasets. Potential solutions include using the notion of metacells [7], i.e. merging cells that statistically represent one closely

related group, and then applying clustering on the metacells.

Secondly, ClinCluster cannot predict the number of clusters, which currently needs to be assigned empirically. Incorporating this feature into ClinCluster needs to be done with more work. A potential solution is using the notion of gap statistics to find the suitable number of clusters [196].

Besides, the results of ClinCluster will rely on the selection of HVGs. More work is needed to evaluate the effect of using different sets of HVGs to calculate the between-cell dissimilarity. The ideal set of HVGs will mostly represent the true biology variance.

Furthermore, ClinCluster can identify small populations to a certain extent, such as the cycle cell cluster that I will show in the next Chapter. However, when a population is too rare (such as one in ten thousand cells) to be identified by any existing clustering method, we should consider deal with it experimentally. For example, samples can be treated with a condition that enriches the rare population, and then we apply scRNA-seq on the treated samples. In this way, the rare population (e.g. the stem cell population) can be more easily identified and profiled.

Although many computational methods are available to correct batch effects [186], it is vital to minimise batch effects at the wet lab level. One example of the best practice is to process multiple samples in one batch. For instance, if an experiment requires cells from mice treated by different conditions, try to process all groups in the same batch. However, this is often not the case for clinical samples. As I will describe in the next chapter, although I performed reverse transcription and other reactions on the cells from multiple donors in the same batch, the sample from a donor were dissociated on the same day as her surgery. It means that samples from different donors have to be dissociated and sorted separately. Moreover, researchers usually cannot control how long it takes for the sample to be shipped to the laboratory after it comes out from the donor's body. This variability is also manifested as the batch effect. Thus, in the scenario of clinical samples, batch correction at the computational level is essential to account for those uncertainties.

In this chapter, I described a clustering algorithm, ClinCluster, which was tailored

for clinical scRNA-seq data. In the next chapter, this algorithm will be applied on the scRNA-seq dataset that I generated from human fallopian tubes.

Chapter 5

Single-cell transcriptomic profiling of human fallopian tubes and decomposition of bulk ovarian tumours

5.1 Abstract

It has been challenging to develop a robust classifier of high-grade serous ovarian cancer (HGSOC). Given the resemblance between HGSOC and its cell of origin, this study hypothesised that the presumable phenotypic heterogeneity within cells of origin might be inherited by cancerous cells. Fallopian tubes are an essential part of the female reproductive system and the origin of certain serous ovarian cancer. However, the knowledge of fallopian tubes at the molecular level is limited. In this work, I profiled the single-cell transcriptomes of epithelial cells, lymphocytes, fibroblasts, *etc.* from human fallopian tubes by using scRNA-seq. My analysis revealed not only the communication network among those populations but also four novel secretory subpopulations. This work provides the first benchmarking dataset that delineates the transcriptomic landscape of human fallopian tubes at the single-cell level. By using the transcriptomic signatures derived from the scRNA-seq data of the fallopian tube epithelium, I decomposed TCGA HGSOC bulk samples and found that the enrichment of the ECM signature was associated with poor prognosis

and treatment response.

5.2 Introduction

5.2.1 Ovarian cancer

Ovarian cancer is the most lethal gynaecological malignancy, which causes over 4 thousand deaths annually in the UK [77]. Based on aetiology and morphology, ovarian cancers are classified into four major histopathological types, namely serous, mucinous, endometrioid and clear cell [79]. Based on cell morphology and differentiation, ovarian cancers are also described as low-grade and high-grade [117, 192]. Among all these subtypes, high-grade serous ovarian cancer (HGSOC) is the most common (70-80%) subtype of ovarian cancer with the lowest survival rate.

The most significant genetic feature of HGSOC is the ubiquitous *TP53* mutation [1]. The *TP53* mutation rate was reported to be around 95% in HGSOC. The other feature is the frequent occurrence of copy number variants [11], which is associated with the defective homologous recombination pathway [177] in around half of the HGSOC cases.

5.2.2 Molecular subtyping of HGSOC

Molecular subtyping is important for stratifying the patients and delivering the most effective treatment. Although the stratification still mostly relies on DNA-based assays, RNA-seq and microarrays have provided unique insights into precision medicine. For example, Parker et al. developed a 50-gene signature (termed as the PAM50 [PAM is short for Prediction Analysis of Microarray]) to classify breast cancer into four subtypes: basal, HER2-enriched, luminal A and luminal B based on expression levels [140]. A test based on PAM50 can predict the risk of recurrence and enable rational choice of treatment.

Previous studies have made progress on subtyping serous ovarian cancer. Tothill

et al. exploited the microarray to profile the transcriptomes of 285 serous and endometrioid tumour samples and for the first time classified serous ovarian cancer into four molecular subtypes [184]. Moreover, they found that a mesenchymal subtype was associated with patient's survival. Such classification was later validated by the TCGA study in a larger cohort [11]. However, the correlation between this mesenchymal subtype and poor prognosis was not significant in the TCGA data, implying that this clustering-based classifier is unstable. Researchers have made some progress to refine this subtype classifier [33, 98], but their methods are limited to the use of clustering to identify four discrete groups. Given this situation, I speculated whether a methodology other than clustering could improve the molecular subtyping system of HGSOC.

5.2.3 Tubal origin hypothesis of ovarian cancer

Ovarian cancer was previously believed to arise from the ovarian surface epithelium. Recently, an increasing number of studies support the tubal origin hypothesis that the fallopian tube epithelial secretory cell (FTESC) population is an alternative cell of origin of serous ovarian cancer [134]. This hypothesis proposes that these cells of origin acquire mutations and evolve into early lesions, which then exfoliate and fall into ovaries, where the mutated cells obtain further genetic alterations and eventually develop into HGSOC.

This hypothesis was brought up because a type of presumed early lesions, known as serous tubal intraepithelial carcinomas (STICs), were frequently founded in genetically high-risk women such as *BRCA*-mutated carriers [92, 106, 149]. The fact that STICs share the same morphology and markers (such as Ki67 [100, 168]) with serous ovarian cancer gives more power to the tubal origin hypothesis [32]. Another putative early lesion is the p53 signature, which is defined by a serial of p53-expressing secretory cells with normal cell morphology in the fallopian tube epithelium.

The evidence supporting the tubal origin hypothesis has been accumulating. A previous study reported that the immortalised and transformed FTESCs derived from healthy fallopian tubes can mimic HGSOC [85]. In the mouse model, Perets et

al. demonstrated that HGSOC can originate from FTESCs by specifically knocking out Brca, Tp53 and Pten in the *PAX8*-expressing secretory cells [146]. At the genetic level, Labidi-Galy et al. analysed the whole exome of microdissected early tubal lesions and ovarian cancers [102]. Their evolutionary analysis revealed that the tumour-specific mutations and copy number variants present in tumours also existed in the early lesions, including the mutations affecting tumour suppressor genes such as *TP53* and *PTEN*. It suggests that the tubal early lesions may be the precursors of ovarian cancer based on the evolutionary paradigm.

Moreover, several population-based studies reveal that the fallopian tube resection can effectively reduce the risk of serous ovarian cancer [36, 86, 156], supporting that fallopian tube is at least one of the origins of HGSOC.

5.2.4 Previous knowledge of the cellular landscape in fallopian tubes

The fallopian tube is a ductus connecting the ovary and the uterus in the female reproductive system (Figure 5.1). It plays an important role in transporting the ovulated eggs to the endometrium, and its cavity is the usual place where the fertilisation happens. A layer of epithelial cells, termed as the fallopian tube epithelium (FTE), spreads above the inner cavity of fallopian tubes. Dysfunction of the FTE is associated with a variety of diseases ranging from infertility to cancer [164].

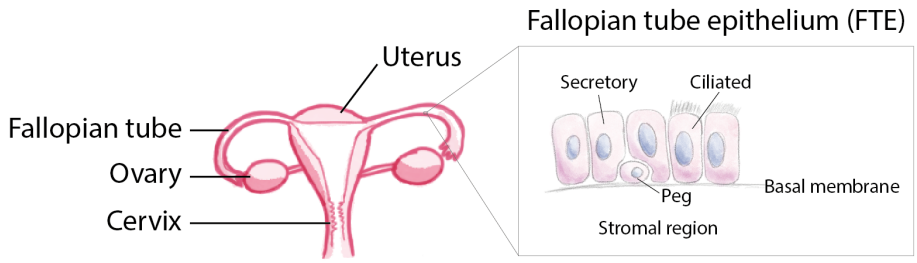

Figure 5.1. Diagram of the human female reproductive system and the fallopian tube epithelium.

The previous literature has documented three major cell types in the FTE [39] (Figure 5.1). The Pax8 positive secretory cell population secretes the fluid nourishing

eggs, while the Tubb4 positive ciliated population flaps the cilia to drift eggs towards the endometrial cavity. The third cell type is termed as the peg cell or the basal cell. Unlike the columnar secretory and ciliated cells, the basal-located peg cells are round-shaped.

5.2.5 Motivations

As the origin of ovarian cancer, a comprehensive understanding of the FTE can provide clues to develop early screening approaches and to improve treatments for ovarian cancer. However, our knowledge of FTE cellular types is limited to cell morphology and few markers, such as the aforementioned Pax8 and Tubb4. Another question that remains elusive is whether there is any unrecognised cell subtype in the FTE. Thus, this work is aimed to reveal the expression profile of individual cell types and identify novel cell populations in the FTE by using scRNA-seq.

Given the reported similarity between FTESCs and HGSOC [85, 146], I assumed that the repertoire of cell states in the cell of origin was inherited by cancer cells. Hence, I speculated that I could compute the composition of cell states in bulk tumours by using the transcriptomic signatures of nonmalignant FTE cell subtypes. In this study, I aimed to exploit the new knowledge obtained from FTE to seek a deeper understanding of the molecular heterogeneity of ovarian cancer.

5.3 Results and Discussion

I designed an experimental workflow to census cell populations in the FTE by exploiting scRNA-seq. The overall design included three parts. The first part (Section 5.3.1) was a technical study, in which I tried to optimise the sequencing conditions. In Section 5.3.2, I used the 10x Genomics technique to portray the global picture of fallopian tubes, including epithelial cells, stromal cells and lymphocytes. The rest part was focused on the epithelial cells, especially secretory cells, to thoroughly scrutinise the subpopulations and generate a new model of the cellular landscape in the FTE.

5.3.1 Impact of culture on transcriptomes

To maintain primary cells from fresh clinical samples, cell culture techniques have been frequently used when fresh samples cannot be easily obtained [104, 167]. The primary cell culture is also a typical model to study the underlying mechanism of diseases or other biological questions. However, the effects of cell culture or other cell maintenance techniques on single-cell transcriptomes have not been well studied.

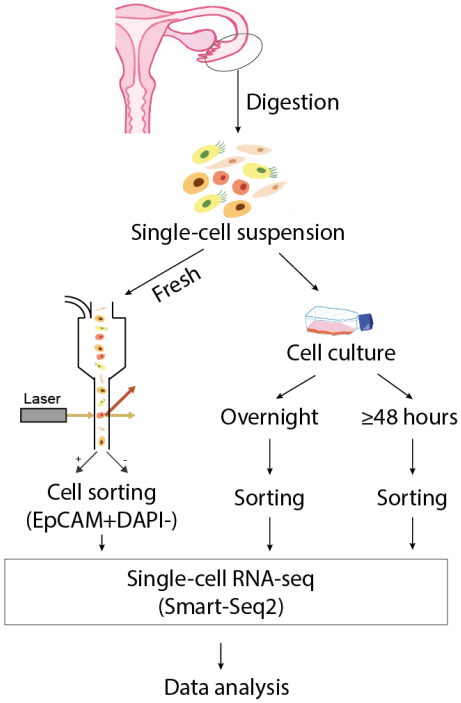

Figure 5.2. Diagram shows the workflow to study the effect of cell culture.

To compare the cells under the fresh condition with the ones under the culture condition, I applied Smart-Seq2 [147, 148] on fresh and cultured fallopian tube epithelial cells (Figure 5.2, Methods see Section 2.3.2.2). The cells from different conditions were gathered into different groups based on the uniform manifold approximation and projection (UMAP) [125] (Figure 5.3a), while the cells from different patients coexisted in all the clusters (Figure 5.3b). UMAP is a dimensionality reduction approach, which is aimed to select representative features or variates and, thus, achieve interpretable visualisation of high dimensional data. Apart from UMAP, another two methods are frequently used for scRNA-seq data, namely t-distributed stochastic neighbour embedding (t-SNE) and principal component analysis (PCA).

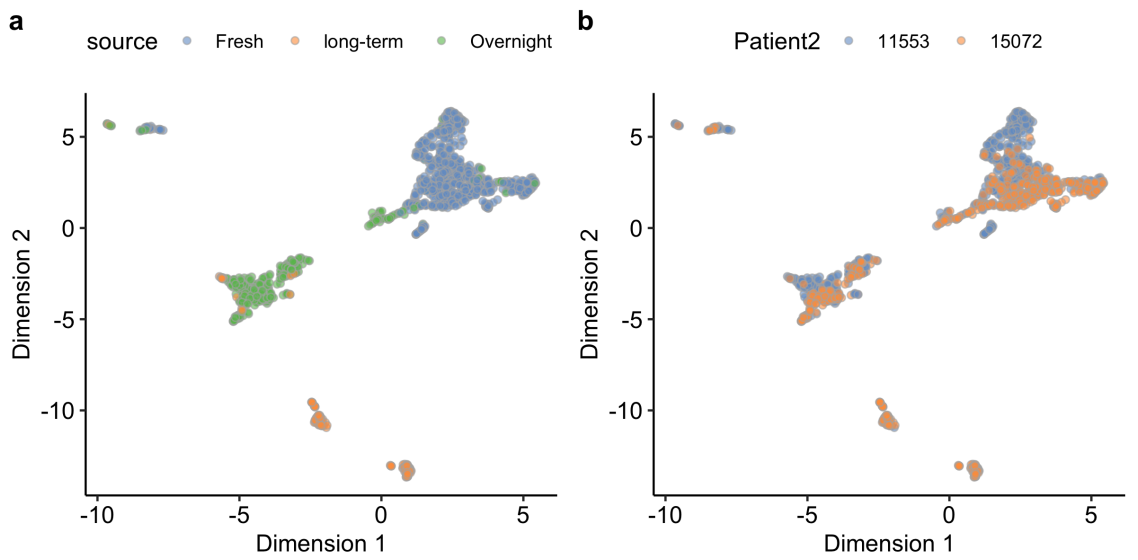

Figure 5.3. UMAPs visualise single cells from two patients across three conditions.

Each dot denotes a cell. Cells are coloured by conditions in ***(a)*** *and by donors in* ***(b)****.*

5.3.1.1 Pseudotime analysis

Firstly, I ran a pseudotime analysis to study the latent variables of culturing effect [26]. Unlike clustering, pseudotime analysis projects cells onto one dimension (also known as the trajectory), which can imply the underlying biological process. The analysis assigns a pseudotemporal value to each cell. The pseudotemporal position of each cell on the trajectory reflects the stage in a certain process, such as the organ development [160, 163, 187]. The proximity of two pseudotemporal values implies the similarity between the two cells.

Our analysis showed that both the median and the dispersion of pseudotemporal values of cultured cells were higher than the ones of fresh cells (Figure 5.4a). The result also showed that the cells reached the highest pseudotime in the overnight-cultured group and the pseudotime of long-term-cultured (LT-cultured) cells followed a bimodal distribution.

I then used PCA to project the single-cell data into two dimensions for visualisation. Consistent with the violin plot, the PCA plot also showed a split of the LT-cultured group into two sub-groups possibly owing to the effect of culture condition (Figure 5.4b). This result implied that tremendous changes were caused by the procedure

of cell culture.

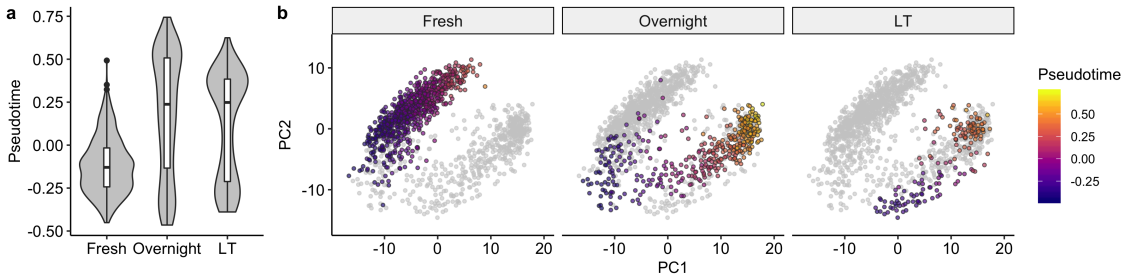

Figure 5.4. Pseudotime analysis of cells under different conditions.

5.3.1.2 Differential expression analysis and pathway analysis

Following the pseudotime analysis, I performed the differential expression analysis among the three conditions. In total, 10,640 genes were differentially expressed across these conditions ($\log_2$ fold-change ≥ 0.4, FDR < 0.05). I summarised the number of differentially expressed genes (DEGs) in each comparison (Table 5.1). It is not surprising that a substantial number of DEGs existed between the fresh and the overnight-cultured, but the discrepancy between the overnight-cultured and the LT-cultured was also significant. It indicated that the overnight-cultured cells did not reach a stable state. Thus, the short-term culture can be regarded as a disturbance in the gene expression.

	Upregulated	Downregulated
Fresh to Overnight	2438	4074
Overnight to LT	5474	2319
Fresh to LT	4353	2828

Table 5.1. The differential expression of genes after cell culture.

I next analysed which pathways were significantly affected by cell culture [81, 190]. The analysis revealed that the pathways affected by the overnight culture condition were involved in DNA, RNA, proteins and many cancer-related pathways (Figure 5.5a). The change in the cancer-related pathways may result from the similarity between the cell culture and the cancerization, in which the negative regulation of cell division was relieved.

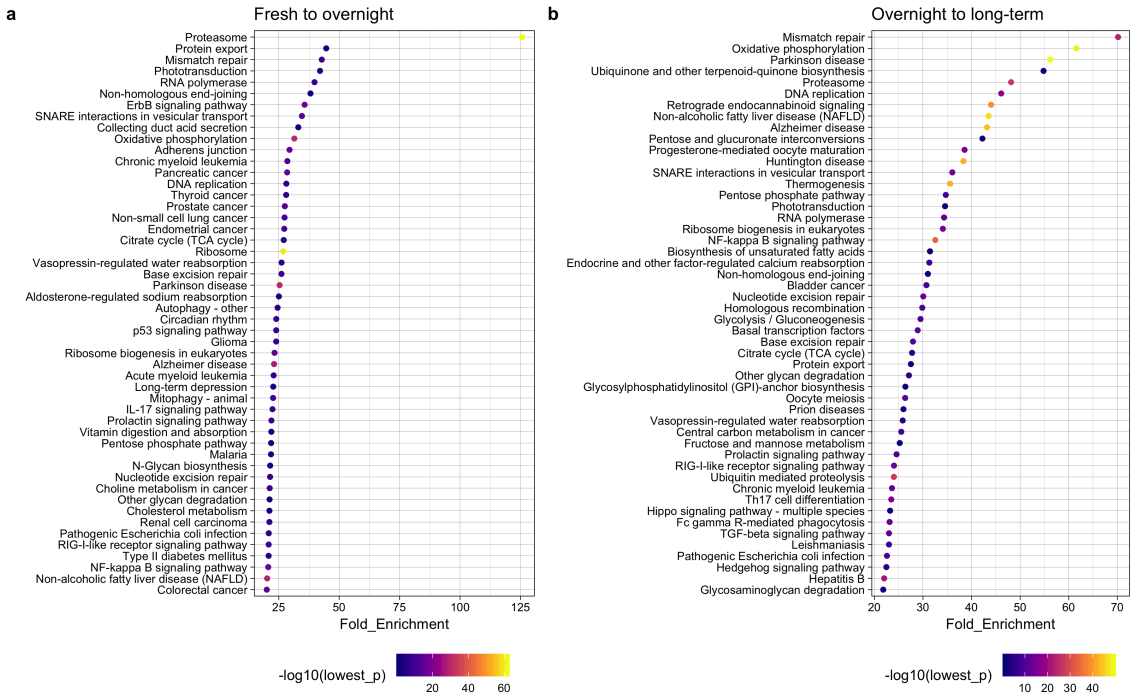

Figure 5.5. Pathway analysis of differentially expressed genes.
The top 50 pathways ordered by the fold enrichment are shown.

Cell cycle The activation of DNA repair and DNA replication pathways was probably due to the preparation for cell division. For example, the genes that restrict division, such as *CDKN1A* (as known as p21) [63] and *TP53* [31], were downregulated, while the cell-cycle-promoting genes, such as *CDC27* (Cell Division Cycle 27) [78] and *CCND1* (Cyclin D1) [123], were upregulated. After the prolonged culture, the DNA replication pathway was further activated. For instance, the genes composing the DNA polymerase, such as the MCM complex (also known as the helicase) [19, 189], were upregulated in the LT-cultured group compared to the overnight-cultured one. This implies that the epithelial cells may be suppressed by the extracellular signal *in situ*, which no longer exists in the culture condition so that the previously suppressed progenitor cells start entering the cell cycle.

Signalling pathways Many signalling pathways are also affected by cell culture (Figure 5.6), including the ErbB, TGF-beta, Notch, Wnt and PI3K-Akt signalling pathways, which are known to modulate cell proliferation and homeostasis [38, 89].

The important components of the ErbB signalling pathway include the membrane receptors expressed by genes *EGFR*, *ERBB2*, *ERBB3*, *ERBB4* [6, 137]. Among

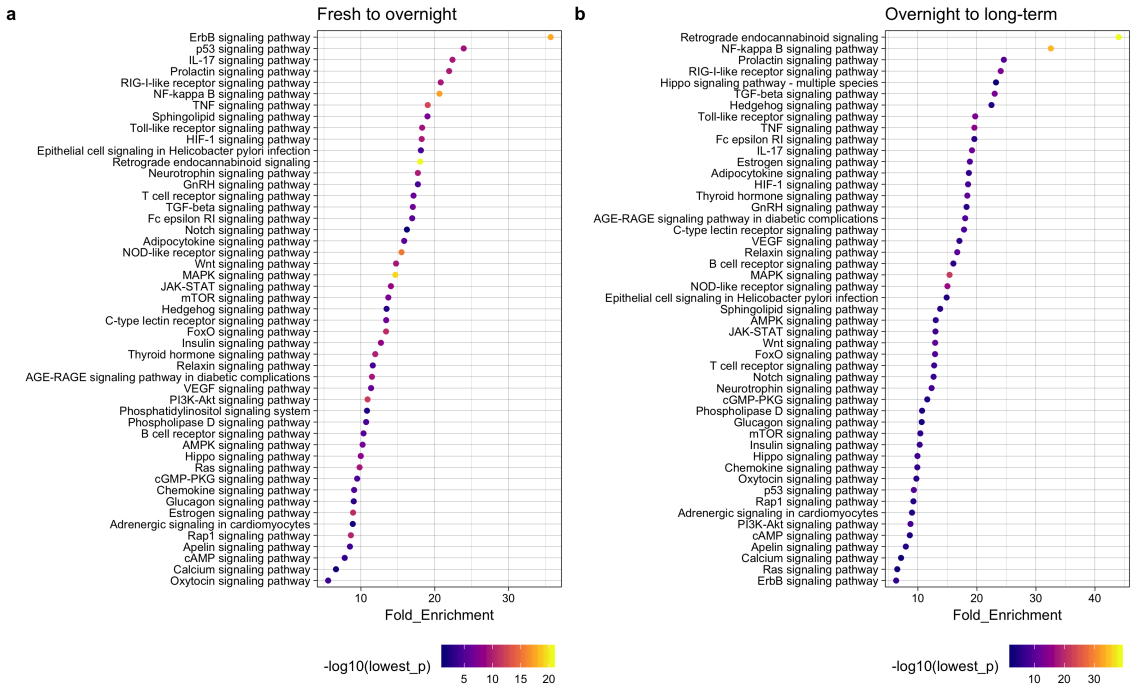

Figure 5.6. Pathway analysis of differentially expressed genes.
The top 50 signalling pathways ordered by the fold enrichment are shown.

them, the *EGFR* was upregulated after overnight culture, while the other three were downregulated. It may be caused by the addition of the EGF factor in the culture medium.

The estrogen signalling pathway was also affected. Compared to the fresh cells, the estrogen receptors, *ESR1* and *ESR2*, and the progestogen receptor, *PGR*, were all downregulated in the culture condition, possibly owing to the depletion of estrogen and progestogen from the environment.

Moreover, the pathways related to the immunomodulation were also affected, such as IL-17 signalling pathway, NK-kappa B signalling pathway and TNF signalling pathway, possibly due to the lack of inflammatory factors in the culture medium.

Next, I focused on two key pathways for cell stemness, Notch and Wnt signalling pathways [5].

Notch signalling pathway In the Notch signalling pathway, 23 genes were differentially expressed. The function of this signalling pathway is to modulate the stem cell fate and is related to the development of organs and cancers [5]. As for the

extracellular components, *ADAM17*, an activator of Notch signalling pathway [22], was downregulated by the culture condition, while the inhibitors, *JAG1* and *JAG2*, were upregulated. For the cytoplasmic components, the repressors, such as *DVL1* and *DVL2*, were upregulated, while the activators, such as *CREBBP* and *EP300*, were downregulated. A previous study of fallopian tube organoids suggested that the differentiation direction relied on the strength of Notch signalling [89]. As this study also observed the decreasing proportion of ciliated cells in the culture condition, I speculated that the Notch pathway was moderately repressed and, thus, the differentiation from progenitor cells to ciliated cells was blocked under the culture condition. As ciliated cells are unable to renew themselves, the repressed Notch signalling lead to a reduction in the proportion of ciliated cells.

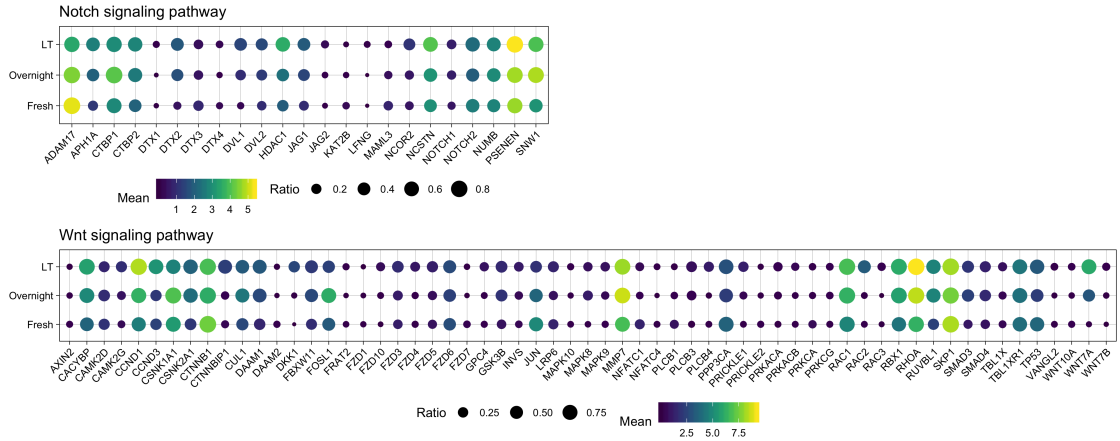

Figure 5.7. DE genes related to the Wnt or Notch signalling pathway.

Wnt signalling pathway In the Wnt signalling pathway, 61 genes were differentially expressed (Figure 5.7). Among these DEGs, Wnt7a is essential in the embryonic development of fallopian tubes [141], while its function in the FTE cells from adults has not been well studied. The results demonstrate that *WNT7A* was significantly upregulated in the culture condition, implying that Wnt7a may play an important role in the mature fallopian tube epithelial cells. On the other hand, some genes related to the Wnt pathway were transiently affected in the culture condition. For example, *FOSL1* was firstly upregulated overnight and then downregulated after 48 hours; *FZD3*, *FZD5*, *FZD6* and *LRP6* were firstly downregulated and then upregulated. These alterations in key pathways elucidate the bias that can be introduced by cell culture.

The overall result illustrates that the culture condition with the addition of EGF, FBS and Rock inhibitor (see Section 2.3.2.2) has a global effect on the transcriptome of primary FTE cells, including substantial changes in the expression of genes related to immune response, metabolism and cell stemness.

5.3.2 The cell atlas of human fallopian tubes

To view the global picture of the cellular ecosystem in human fallopian tubes, I applied the 10x Genomics scRNA-seq method to the fresh fallopian tube samples from two premenopausal donors. The biopsies were dissociated into single-cell suspension, stained with antibodies against EpCAM and CD45 and sorted into three groups (EpCAM+, EpCAM- CD45+ and EpCAM- CD45-). The three groups were evenly mixed and processed following the 10x Genomics protocol (Figure 5.8). In total, 1,563 single cells were captured after filtering by the cell quality. I further excluded 117 putative doublets by DoubletFinder [124], leaving 1,446 cells for further analysis.

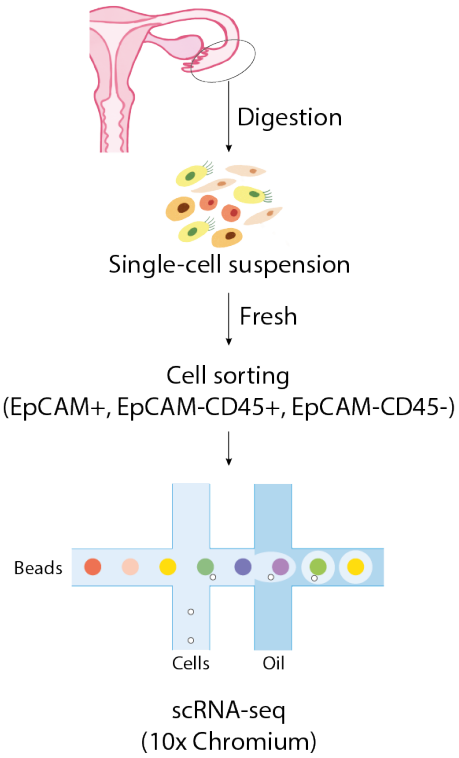

Figure 5.8. Diagram shows the workflow of 10x Genomics scRNA-seq.
The diagram of the 10x Genomics technique is adapted from the Chromium official brochure.

5.3.2.1 Clustering results show cell populations

The remaining cells were assigned into 8 groups by using the Seurat clustering [24] (Figure 5.9a). The clustering identified two epithelial populations (ciliated and secretory), three blood cell populations (lymphocytes, monocytes and dendritic cells) and three stromal populations (fibroblast 1, fibroblast 2 and endothelial cells). All of the populations existed in both patients, Pre2 and Pre3 (Figure 5.9b), suggesting that these populations were robust. The marker genes are listed in Table 7.1.

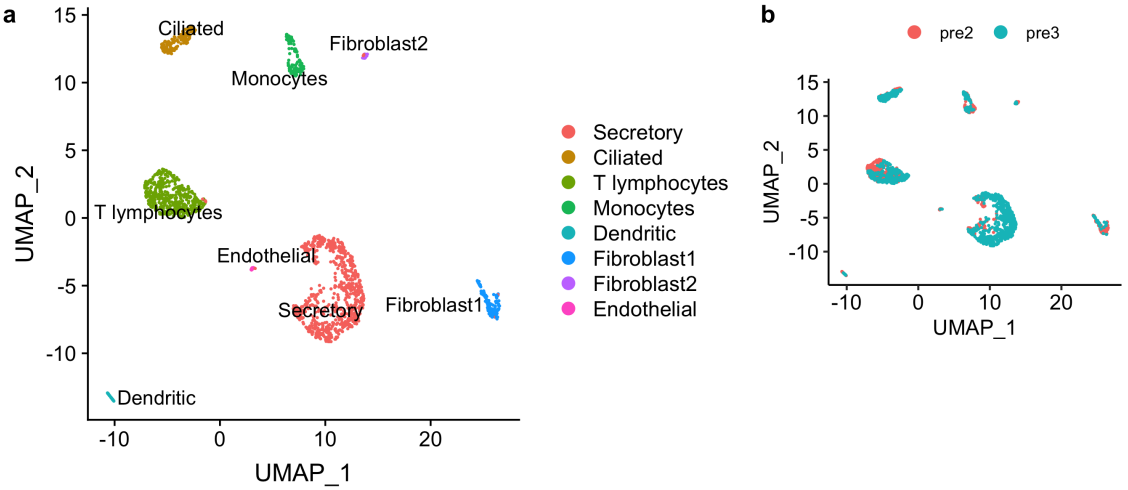

Figure 5.9. UMAPs visualise cell populations in human fallopian tubes.
*(**a**) A overview of the clustering results. Each dot represents one cell coloured by the assigned population.*
*(**b**) The cells are coloured by the source of donors.*

5.3.2.2 Cell-cell communications between cell types

To explore the network of cell-cell communications, I applied CellPhoneDB [193] on the count matrix. CellPhoneDB was built based on the curated receptor-ligand interaction database, Protein Data Bank (PDB) [66]. CellPhoneDB first identified which receptor or ligand was expressed in a certain cell population, and then searched for matched receptor-ligand pairs. The permutation that assigned cells to random clusters enabled the calculation of an empirical p-value for each potential interaction.

Our analysis reveals plenty of crosstalk between pairwise populations in fallopian tubes (Figure 5.10). The most interactive populations were endothelial cells (Endo), fibroblast (F2) and monocytes (Mono). In the epithelial cells, secretory cells (SE)

were more talkative than the ciliated cells (CI).

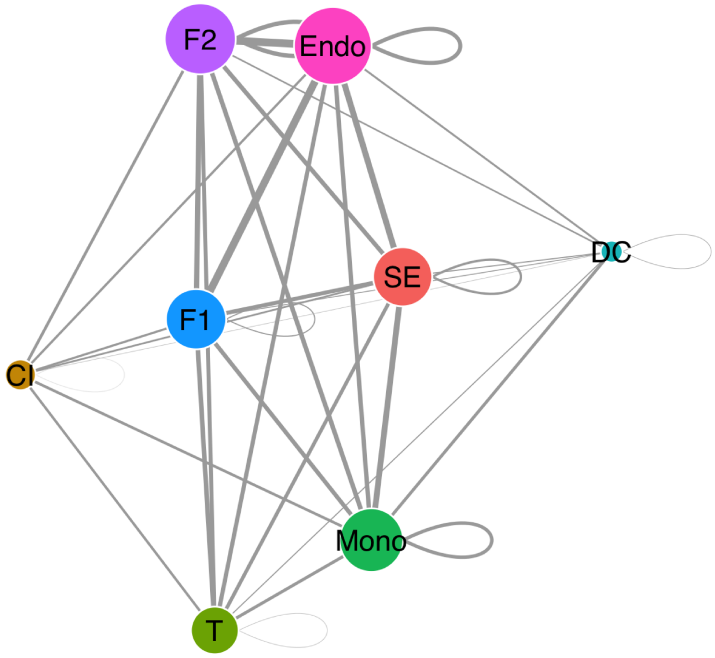

Figure 5.10. Network graph shows the cell-cell communications among fallopian tube cell populations.

The nodes are cell populations and the node size denotes the degree of each node, that is the sum of weights of all edges connecting to a given node. The width of each edge represents the number of significant interactions between a pair of cell populations. F1, fibroblast 1; F2, fibroblast 2; Endo, endothelial; SE, secretory; Mono, monocyte; T, lymphocyte; CI, ciliated; DC, dendritic cells.

I found that the ligand-receptor interactions in fallopian tubes were related to the immunomodulation, angiogenesis (angio.), growth factors, Wnt and Notch signalling. (Figure 5.11). The *CXCL12-CXCR4* and *CXCL16-CXCR6* were used frequently to attract T lymphocytes. The angiogenesis interactions, such as *NRP2-VEGFA*, existed between endothelial, dendritic and monocytes. The *FGF7-FGFR2* between fibroblasts and secretory cells potentially supported the proliferation of normal epithelial cells. This is consistent with a previous finding that the FGFR2 signalling can regulate cell proliferation [91]. *CD74* and *CD44* were involved in the communication between the epithelial population and other populations, implying that they play an important role to facilitate the communications between FTE cells and other non-epithelial cells.

Notably, the *WNT4-FZD6* interaction existed between the fibroblasts and epithelial

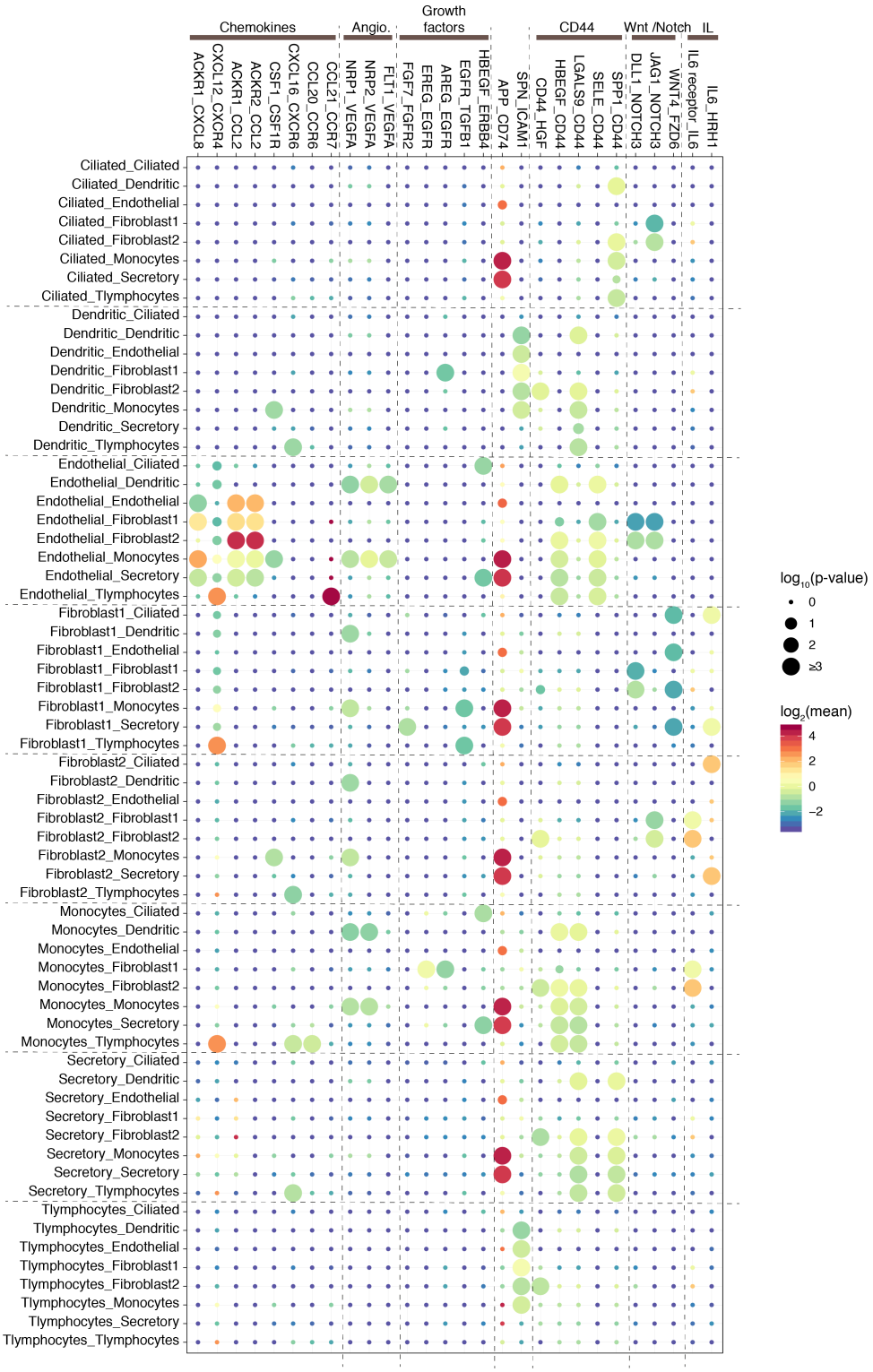

Figure 5.11. An overview of the ligand-receptor interactions in fallopian tubes.

The columns represent the pairs of ligands and receptors, and the rows represent the pairs of the ligand-harbouring population and the receptor-harbouring population. The dots are coloured by the expression levels of two genes, and the dot sizes correspond to the p-values as shown in the legends.

cells. Wnt4 is a versatile factor involved in many biological functions [15], including the sex determination [138] and the formation of the female reproductive system. Moreover, an interaction involving the Notch signalling pathway, *JAG1-NOTCH3*, existed between ciliated cells and fibroblasts (Fibroblast 2), while *JAG1* has been reported to regulate cell differentiation [129]. On the other hand, Zimrim et al. reported that *JAG1* could boost the fibroblast growth factor (FGF)-dependent angiogenesis in the *in-vitro* assay [212]. Therefore, the *WNT4-FZD6* and *JAG1-NOTCH3* existing between epithelial cells and other cell populations may play an important role in regulating cell homeostasis, while the underlying mechanism needs further experimental validation.

A previous study used the culture medium with supplements of EGF, FGF10 and TGF-beta, whilst it remains unclear whether those factors are appropriate or essential for the growth of FTE cells [89]. This study discovered several factors that could be critical for the microenvironment around the FTE cells. The results also confirmed the importance of the intercellular interactions, *TGFB-EGFR* and *EGF-EGFR* in the human fallopian tubes, but they are not expressed in secretory cells. On the other hand, the analysis found that the *HBEGF* (Heparin Binding EGF Like Growth Factor)-*ERBB4* and the *FGF7-FGFR2* interactions existed between secretory cells and other cell populations, which may be more essential for secretory cells. Therefore, these two factors should be considered as potential supplements added to the culture of FTE cells.

5.3.3 Extensively profiling of secretory and ciliated types

Following the tissue-level profiling of all cell populations, I next focused on the fallopian tube epithelial cells, which are the cells of origin of HGSOC.

I isolated 2275 EPCAM+ epithelial cells, 108 CD45+EPCAM- cells and 72 CD45-EPCAM- cells by using fluorescence-activated cell sorting (FACS) and sequenced them by using the Smart-Seq2 protocol [147]. The cells came from the fresh human fallopian tube biopsies of six patients (Table 5.2). Three patients were diagnosed with ovarian or peritoneal cancer, and the other three endometrial cancer.

Patient IDs	Age	Diagnosis	Tubal status	No. cells
11543	73	OC	R: STIC	224
11545	66	PPC	Normal	516
11553	77	HGSOC	R: MC	463
15066	52	EC	Normal	599
15072	62	EC	Normal	319
15091	82	EC	Normal	322

Table 5.2. Overview of patients' information.

OC, ovarian cancer; STIC, serous tubal intraepithelial carcinomas; PPC, Primary peritoneal cancer; HGSOC, high-grade serous ovarian cancer; MC, mucosal carcinoma; EC, endometrial cancer. The number of cells is after the QC and doublet filtering.

I applied Seurat (v3) on the filtered cells to identify subpopulations of secretory cells (Figure 5.12) and annotated the Seurat clusters based on the known marker genes (Figure 5.13 and Table 5.3). *EPCAM* is the marker of epithelial cells, *PTPRC* (also known as CD45) immune cells, *COL1A1* fibroblast, *PAX8* secretory cells, and *FOXJ1* ciliated cells.

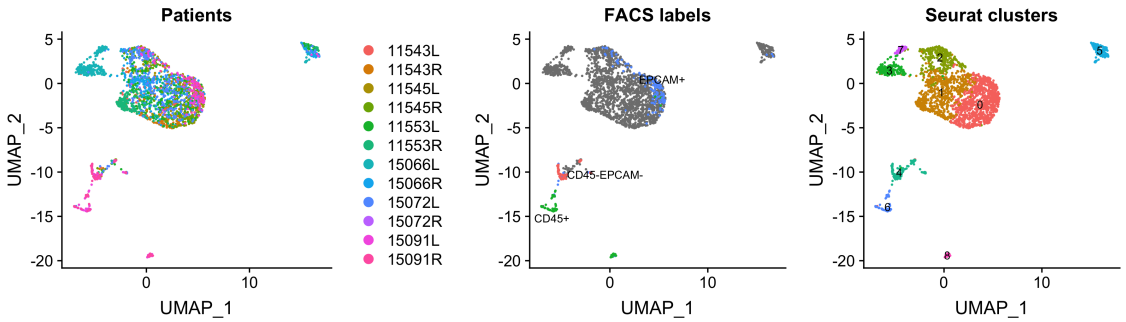

Figure 5.12. UMAPs visualise the cells profiled by Smart-Seq2.

The cells are coloured by the donors, FACS labels and Seurat labels in the sub-panels from left to right. In the colour legend of the left panel, L denotes the left fallopian tube, and R the right tube. In the middle panel, grey denotes unlabeled cells, blue EPCAM+ cells, green CD45+ cells and red CD45-EPCAM- cells.

Non-epithelial cells Clusters 4, 6 and 8 were negative for the epithelial marker *EPCAM* (Figure 5.13), suggesting that they are non-epithelial cells. Besides, Clusters 6 and 8 were positive for CD45, suggesting that these two clusters are leukocytes.

Cluster 4 was negative for the lymphocyte common antigen, *PTPRC* (also known as

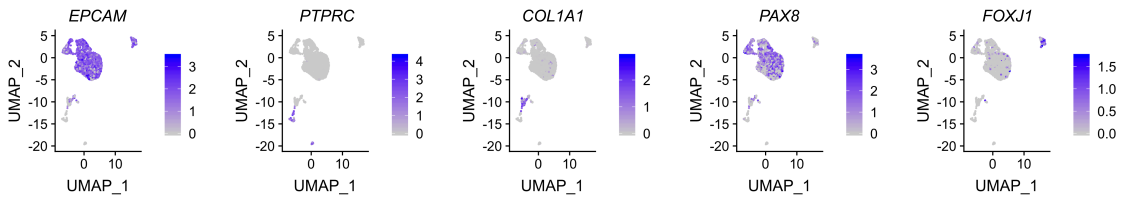

Figure 5.13. Known markers label the clusters.

Cells are coloured by the expression levels of the marker genes that shown in the titles.

Clusters	Annotations	Marker genes
0	Secretory	
1	Secretory	*CXCL1, CCL2, IL8, CXCL3, MMP7*
2	Secretory	*SCGB1A1, MUC6, RNASE1, SCGB3A1, SNCG*
3	Secretory	*LTF, HP, PAEP, SERPINA3, CHI3L1*
4	Fibroblast	*DCN, ACTA2, SFRP4, FBLN1, MFAP4*
5	Ciliated	*TPPP3, C20orf85, ZMYND10, PIFO, CAPS*
6	Leukocytes	*SRGN, SLC2A3, IL1B, ITGAX, RGS2*
7	Secretory	*REG1A, FGG, SPP1, SERPINA1, A2M*
8	Leukocytes	*IGJ, GPR183, CCR7, CD79A, CXCR4*

Table 5.3. The marker genes and annotations of Seurat clusters.

The cutoffs of marker genes: logFC $\geq$ 1, FDR $<$ 0.05.

CD45), and positive for the fibroblast marker, Collagen (*COL1A1*). It suggests that Cluster 4 is the fibroblast population with the expression of certain representative markers, such as Actin Alpha 2, Smooth Muscle (*ACTA2*), Decorin (*DCN*) and Secreted frizzled-related protein 4 (*SFRP4*). *SFRP4* was reported to interact with the Wnt ligand or the Wnt signalling as an antagonist [16]. As I mentioned that the interaction between Wnt ligands and their receptors exists in fallopian tubes (see Section 5.3.2.2), it will be interesting to study whether *SFRP4* also plays a role in the Wnt signalling network in human fallopian tubes.

Epithelial cells Clusters 0, 1, 2, 3, 5 and 7 were *EPCAM* positive epithelial cells based on the mRNA level and the FACS label. Except Cluster 5, the other five clusters of epithelial cells were positive for *PAX8*, the marker of secretory cells, indicating that these clusters are secretory cells. I will continue to discuss the

secretory cells in Section 5.3.4.

Ciliated cells Cluster 5 was positive for *FOXJ1* and *PIFO*, which are involved in ciliogenesis [204] and cilia disassembly [93] respectively. The expression of the two markers indicates that Cluster 5 is the ciliated cell population.

In total, I identified 198 marker genes of ciliated cells (logFC [$\log_2$ fold-change] ≥ 1, FDR < 0.05, pct.1 [the percentage of cells expressing this marker genes in the given cluster] > 0.8). These markers were enriched in the pathways related to cilium movement and nucleoside diphosphate phosphorylation (Benjamini-corrected p-values < 0.05, by DAVID). I speculate that the cilia-related genes are expressed to assemble cilia, while the nucleoside diphosphate phosphorylation pathway modulates such movement [172]. Out of the 198 markers, 15 genes belonged to the coiled-coil domain containing (CCDC) family, such as *CCDC78*, *CCDC17* and *CCDC164*. It is consistent with the knowledge that the CCDC family is important for ciliary assembly [10, 96, 144].

5.3.4 Identifying novel secretory cell subtypes

Given the high capability of metastasis of ovarian tumours, fallopian tube samples may be contaminated by the metastatic tumour cells. To avoid this contamination in secretory cells, I first checked if the samples used for scRNA-seq were free from visible tumours before the dissociation step. I only proceeded with samples that had normal morphology. Yet this cannot exclude the possiblity of microscopic metastasis. The TCGA study demonstrated that ovarian cancer cells tend to have abundant large-scale copy number variations (CNVs) [11]. HoneyBADGER is based on an assumption that the presence of chromosome-scale CNVs will substantially affect the expression levels of genes in the affected region [52]. It uses the average scaled expression data to identify the chromosomal regions with significant upregulation or downregulation, which correspond to potentially amplified or deleted copy numbers. To infer CNVs of single cells, I applied HoneyBADGER on the scRNA-seq data of secretory cells from cancer patients. In one patient (Patient 11528), I observed some cells with abnormal expression profiles (Figure 5.14). Compared to other secretory

cells from the same patient, these tumour-like cells have putative amplification on chromosome 6 and putative deletion on chromosomes 13, 15 and 22, and many of they also shared the same pathological *TP53* mutation with the tumour sample, which was profiled by the whole-exome sequencing. It indicates that HoneyBADGER can tell the presumable tumour cell contamination from the normal FTE cells based on scRNA-seq data. I applied HoneyBADGER on the single cells from other patients, and it did not detect any chromosome-level CNV event. It suggests that the samples from other patients are highly likely free from tumour contamination. The downstream analysis did not include any putative tumour cell.

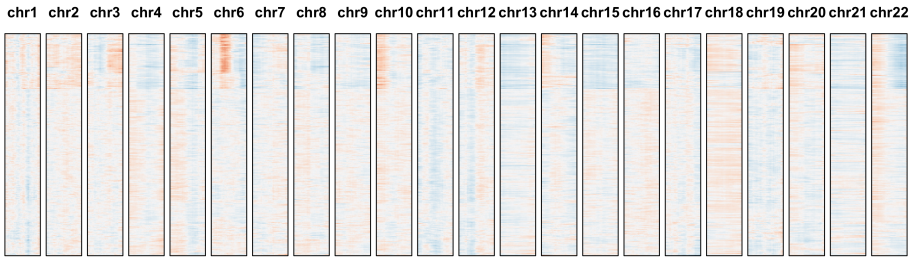

Figure 5.14. Heapmap shows putative CNVs called by HoneyBADGER in single cells from HoneyBADGER.

Each row is a cell and the columns are chromosomal positions. Red represents putative regions with gain of copy numbers, while blue represents putative ones with loss of copy numbers. The darker the colour, the higher probability of CNVs. Some cells at the top of the plot have high likelihood of duplication in Chromosome 6 and deletion in Chromosomes 13, 15 and 22. This is the consistent with the whole-exome sequencing result of the tumour sample from the same patient. Most of the single cells that harboured Chromosome 6 amplification and had coverage on gene TP53 also harboured the same pathological TP53 mutation as the tumour cells, while the other normal cells did not have this TP53 mutation.

To address the aforementioned question on whether there is any additional secretory subtype, the next step is to cluster secretory cells and look for new subpopulations. I applied the clustering method, ClinCluster, on the fresh secretory dataset. Based on the clustering results, the secretory cells were assigned into seven clusters (Figure 5.15 and Table 5.4).

I looked into the marker genes of each cluster and annotated the clusters based on the functions of the marker genes. The most featured population was the KRT-MHC group (C1) with upregulated expression levels of keratins (such as *KRT17*, *KRT14* and *KRT5*) and major histocompatibility complex (MHC) II (such as *HLA-DPA1*, *HLA-DPB1* and *HLA-DQA1*). One of the top markers is Polymeric Immunoglobulin

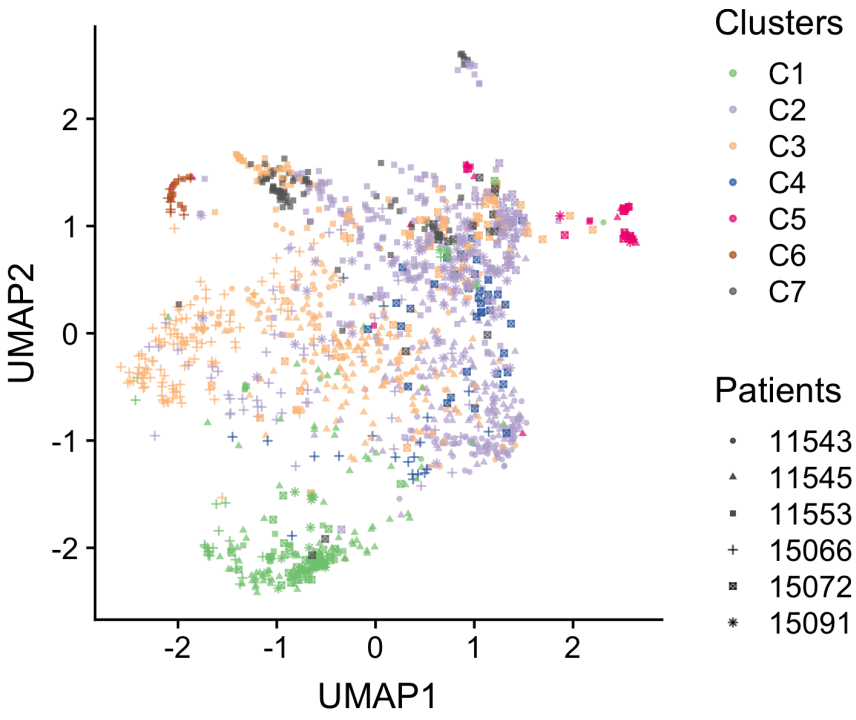

Figure 5.15. The UMAP plot visualises the subpopulations of secretory cells. *The cells are coloured by the clustering output of ClinCluster and shaped by the source of donors as shown in the legends.*

Receptor, *PIGR* (Figure 5.16). According to the previous studies, *PIGR* is related to the immunoglobulin transcytosis in epithelial cells [50] and it is also slightly associated with longer overall survival of ovarian cancer patients [17].

Annotations	Cluster IDs	Representative markers
KRT-MHC	C1	*PIGR, KRT17, HLA-DQA1*
Stressed or Quiescent	C2, C4	*JUN, EGR1*
Differentiated	C3	*PTGS1, FMOD*
ECM	C5	*TIMP3, SPARC, TPM2*
Cell cycle	C6	*ZWINT, STMN1, TK1*
Inflammatory	C7	*IL8, CXCL11*

Table 5.4. The marker genes and annotations of secretory clusters. *The cutoffs of marker genes: logFC $\geq$ 1, FDR $<$ 0.05.*

I studied the enriched Gene Ontology biological pathways of each cluster by using GORilla [48, 49] (FDR $<$ 0.05, Table 5.5). The markers of the KRT-MHC group were enriched in the pathways of the immune system process, cell surface receptor

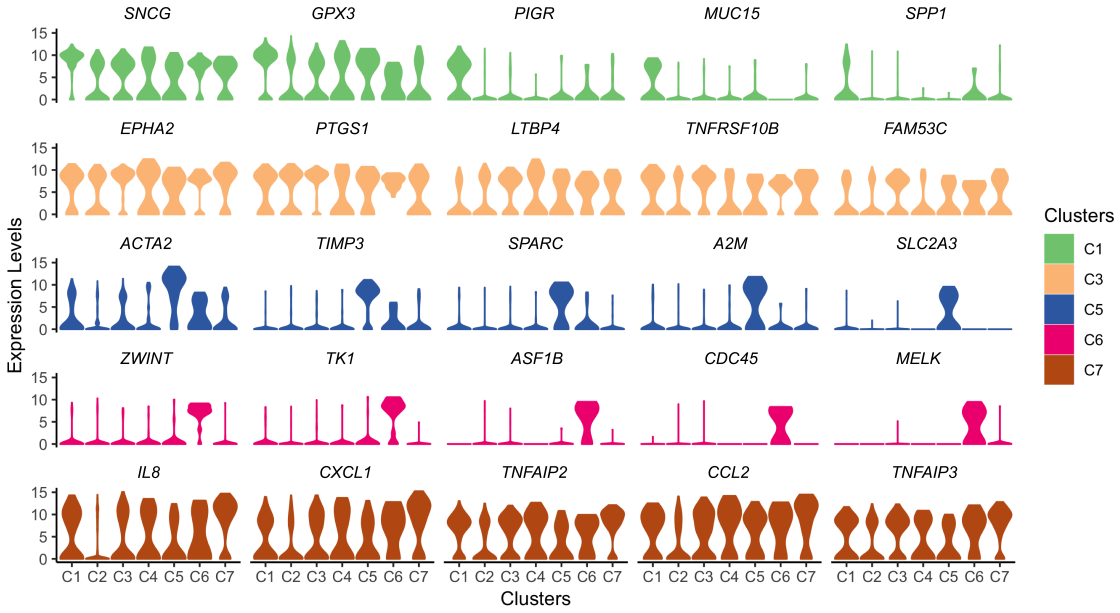

Figure 5.16. Expression levels of marker genes across cell subpopulations.
The violin plots show the upregulation of representative marker genes in each cluster. The subfigures are coloured by the cluster represented by the certain marker. The x axis denotes clusters and the y axis expression levels.

signalling pathway and primary alcohol catabolic process. Although the functions of this subpopulation have not been fully understood, the expression profile indicates that this population plays an important role in interacting with T lymphocytes and other immunocytes, due to the expression of MHC II genes [200]. Functional assays are needed to investigate the functions of this subpopulation.

I repeated the annotating workflow on other clusters. Based on the known functions of marker genes and the enriched pathways, the rest of those clusters were annotated as Differentiated (C3), Extracellular matrix (ECM) (C5) and Cell cycle (C6). The ECM subtype may be related to the epithelial-mesenchymal transition (EMT) process, because a well known EMT marker *SNAI2* was upregulated in this group. Moreover, the Quiescent/Stressed Group (C2 and C7) did not have significantly enriched pathways or specific marker genes, probably due to the fact that the donors were post-menopausal (Table 5.2). Thus, the largest population in the fallopian tubes was in the senescent state.

Annotations	Clusters	Representative pathways
KRT-MHC	C1	Immune system process, cell surface receptor signalling pathway and primary alcohol catabolic process
Differentiated	C3	mRNA metabolic process and regulation of stem cell differentiation
ECM	C5	Extracellular matrix and epithelial cell migration
Cell cycle	C6	Cell cycle, DNA repair and chromosome organization
Inflammatory	C7	Inflammatory response pathway

Table 5.5. The representative enriched pathways of secretory subtypes. The cutoffs of the GO enrichment analysis: FDR < 0.05.

5.3.5 Deconvolution reveals inter- and intra-tumour transcriptomic heterogeneity

To represent the expression signatures of four secretory cell subtypes and a ciliated cell type, I picked up 52 marker genes. The compositional analysis has two steps. The first one is calculating the scaled expression levels of these markers in the five FTE cellular subtypes. This step generates a reference matrix, in which rows denote the cells and columns denotes marker genes. The second step is using a linear support regression model to estimate the proportion of each cell subtype in a given bulk tumour [8, 132].

I applied the deconvolution method on 308 HGSOC samples of TCGA. Based on the composition of each tumour, the tumours were arbitrarily classified into five molecular subtypes: differentiated, KRT-MHC, ECM, cell cycle and mixed (Figure 5.17). The tumour subtypes were defined by the major cell state in a certain tumour (proportion > 0.5). If there was no major cell state, the tumour would be classified as mixed. This result reveals the inter-patient and intra-tumour heterogeneity and demonstrates that multiple states of cancer cells can coexist in a single tumour.

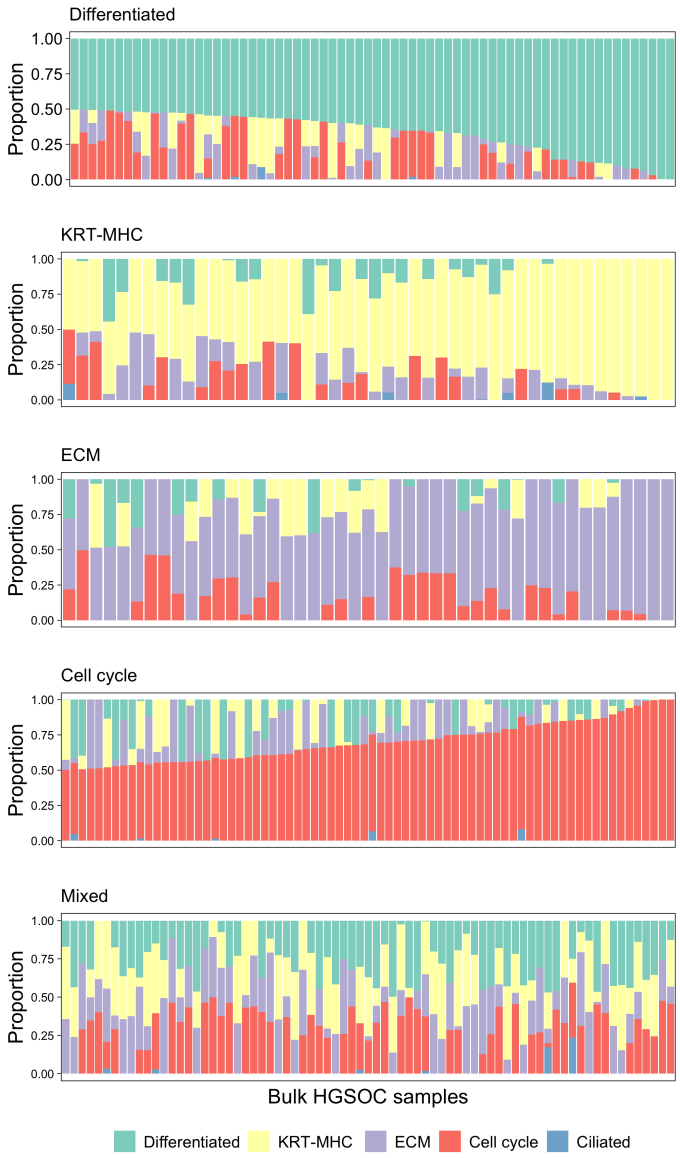

Figure 5.17. Regrouping of TCGA ovarian cancer based on the proportions of five cells states.

Each column in this barplot denotes the composition of a bulk ovarian tumour. The tumours are grouped into five sub-figures based on the dominant cell state, the proportion of which is more than fifty percent in a given tumour. The colour of each component in a bar represents the cell state as shown in the figure legend at the bottom. The height of each bar component denotes the proportion of the corresponding cell state in the given tumour.

5.3.5.1 ECM signature is associated with epithelial-mesenchymal transition

To study the biological pathways underlying the enrichment of the ECM cell state in ovarian tumours, I performed differential expression analysis between the ECM-high and ECM-low tumours. My analysis identified 1394 differentially expressed

genes (FDR < 0.05, $\log_2$FC > 1), including the known EMT markers [82], *SNAI2*, *TWIST1*, *TWIST2*, *ZEB1* and *ZEB2* (also known as SIP1) (Figure 5.18). It suggests that the EMT process is associated with the accumulation of the ECM cell state in HGSOC.

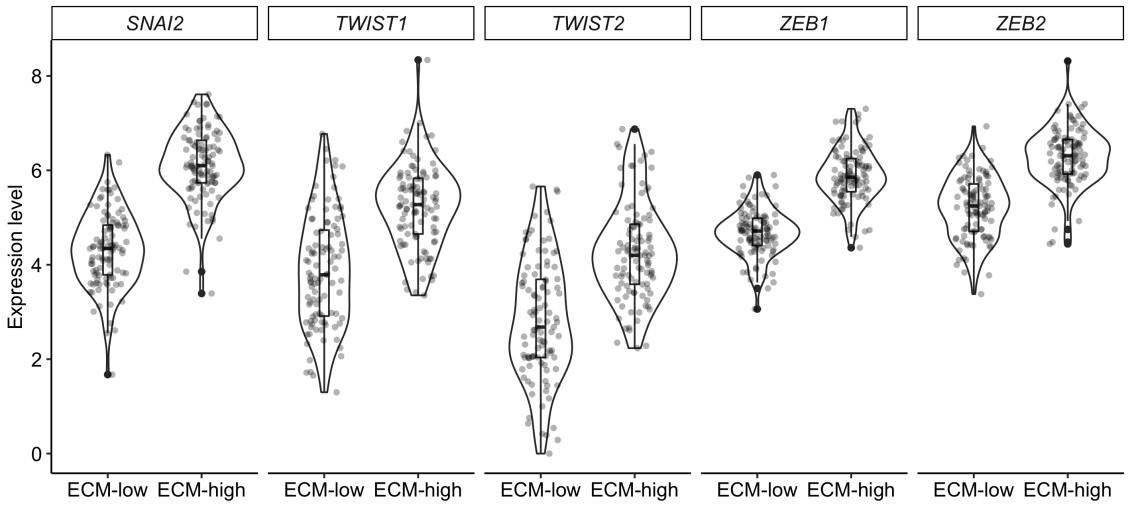

Figure 5.18. The EMT markers are upregulated in the ECM-high tumours.
Each panel corresponds to an EMT marker gene as shown in the sub-figure title. The x-axis denotes two groups of tumours, ECM-low and ECM-high, which are dichotomised by the median of ECM proportions. The y-axis denotes the expression levels of the marker genes. Each dot denotes one tumour samples. The upper border, middle line and the lower border of the boxplot represent the upper quantile, median and lower quantile of each group respectively.

5.3.5.2 EMT enrichment is negatively associated with poor prognosis

The association analysis was conducted between clinical factors and the EMT enrichment, which was defined by the proportion of the ECM cell state in a bulk tumour. This analysis revealed the significant association between the EMT enrichment and the factors (stage, age and incomplete response to treatment) was significant ($p < 0.05$, Figure 5.19 and Table 5.6). The relationship between the residual disease and the EMT enrichment was positive but not significant, possibly due to the small sample size. I also tested the correlation between the relapse-free survival and the EMT enrichment. Although the correlation was negative, it was not significant.

Moreover, the EMT enrichment was associated with poor survival of HGSOC patients (Figure 5.20 and Table 5.6). Given the potential connection between the ECM

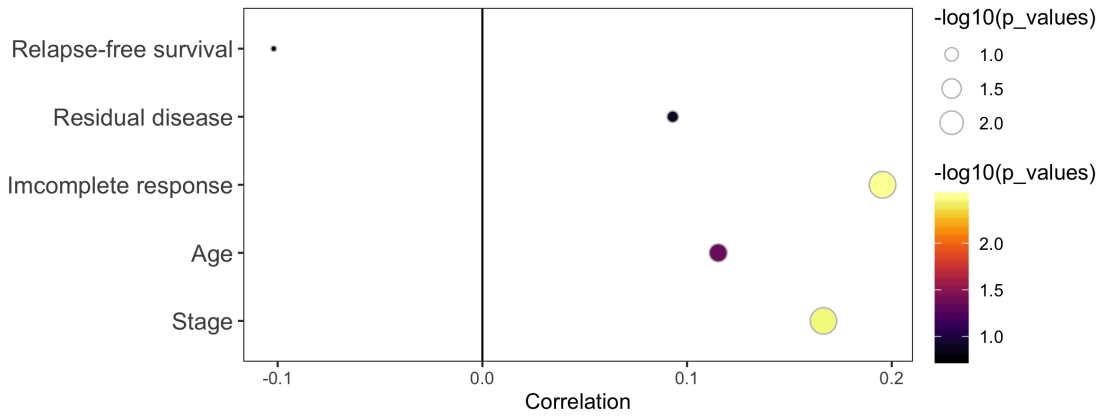

Figure 5.19. The association between the EMT enrichment and other clinical features.

textslThe x-axis denotes the Pearson correlation coefficient, and the y-axis denotes the clinical factors. The size and colour of dots represent the $\log_{10}$(p-value) as shown in the legend. The vertical black line shows where the correlation of 0 is.

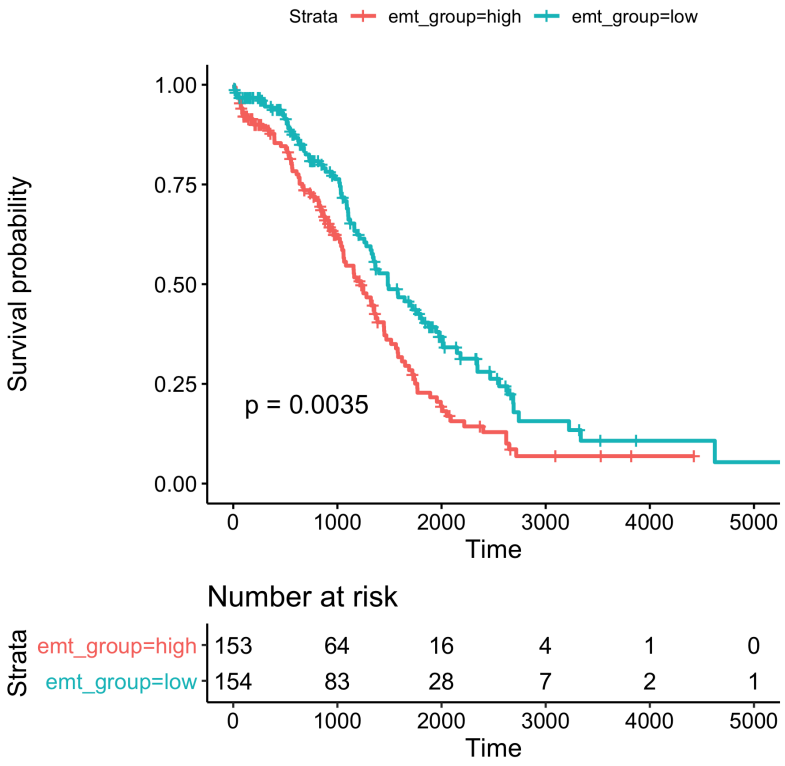

Figure 5.20. Kaplan–Meier curve of the ECM-high and ECM-low tumours.

The tumours are dichotomised by the median of the EMT enrichment (0.146). Harzard ratio = 1.54 by the Cox proportional hazards regression model.

cell state and the EMT process, my finding is consistent with the notion that EMT can promote cancer progression [181].

Previous studies also reported that a mesenchymal subtype is associated with poor

Features (Number of patients)	p-values	Correlation
Stage (n = 306)	0.003	0.17
Age (n = 308)	0.043	0.12
Imcomplete response (n = 226)	0.003	0.20
Residual disease (n = 270)	0.127	0.09
Relapse-free survival (n = 179)	0.174	-0.10
Poor overall survival (n = 307)	0.006	2.29*

Table 5.6. The association between the continuous EMT enrichment and other clinical features.

The correlation of stage, age, therapy response, residual disease and relapse-free survival was calculated by Pearson's product-moment correlation. The p-values have not been corrected for multiple comparisons. The Stages were dichotomised by Stages 1 & 2 or Stages 3 & 4, and residual disease $\leq$ 10 mm or $\geq$ 11 mm. The correlation between overall survival and the EMT enrichment were calculated by the log-rank test.
* *Harzard ratio = 2.29.*

prognosis [33, 184]. This mesenchymal subtype was defined by applying the non-negative matrix factorization consensus clustering on the expression profile of 1,500 high variance genes. I compared the classification results with the mesenchymal subtypes defined by the TCGA study [11]. All the samples were catalogued into four groups, namely ECM-high mesenchymal, ECM-high non-mesenchymal, ECM-low mesenchymal ($N = 3$, not included in the following analysis due to the small size) and ECM-low non-mesenchymal. The ECM-high and ECM-low are defined based on the median of the EMT enrichment. This study showed that both the ECM-high non-mesenchymal group had worse survival than the ECM-low non-mesenchymal group (Figure 5.21), while the former had similar prognosis compared to the ECM-high mesenchymal group. It indicates that the method presented here can detect poor-prognostic patients more sensitively.

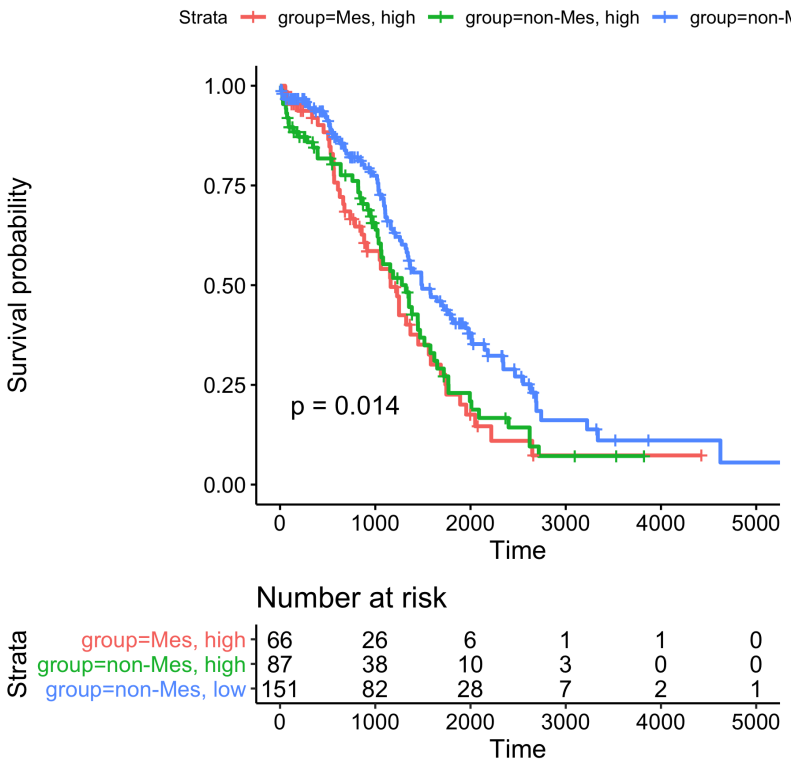

Figure 5.21. Comparison between the Tothill et al.'s four-subtype classifier and the 52-gene predictor in the TCGA data.

I dichotomised the TCGA HGSOCs into four groups based on the previous classifier and the predictor. The previous classifier defined the mesenchymal (Mes) subtype as high risk and other subtypes as low risk. The predictor catalogues samples into ECM-high and ECM-low, dichotomised by the median. The ECM-high cases had a higher risk than the ECM-low ones. The comparison found that the non-mesenchymal ECM-high tumours (green curve) had a similar survival compared with the mesenchymal ECM-high tumours (red curve). Quoted from Fig S6A, Hu et al. 2020 [73]

To test the reproducibility of the prognostic prediction, I repeated the analysis on another seven independent datasets that profiled the bulk transcriptomes of serous ovarian cancer [55]. These datasets were generated by various microarray platforms. The association between deconvolution results and poor survival outcome ($p < 0.05$, hazard ratio between 1.88 to 3.31) was reproducible in totally eight independent datasets, including TCGA dataset, AOCS dataset and six additional microarray datasets from the CuratedOvarianData database [55] (Table 5.7). This result confirmed that the deconvolution analysis using the identified panel of genes was strongly predictive of prognosis in serous ovarian cancer.

GEO entry	n	Hazard ratio	p-value	95% CI
E.MTAB.386 [13]	129	3.31	0.0120	1.30, 8.41
GSE49997 [151]	194	3.08	0.0038	1.44, 6.60
GSE13876 [42]	144	2.58	0.0272	1.06, 5.11
GSE26712 [21]	185	2.03	0.0315	1.07, 3.89
GSE26193 [122]	107	2.33	0.0348	1.06, 5.11
GSE51088 [84]	139	1.89	0.0285	1.07, 3.33
GSE32062.GPL6480 [203]	260	1.88	0.0556	0.98, 3.58
AOCS (Grade $\geq$ 2) [184]	253	2.69	0.0004	1.56, 4.66
TCGA RNA-seq [11]	307	2.30	0.0047	1.29, 4.09

Table 5.7. The high EMT enrichment is robustly correlated with poor prognosis across multiple datasets.

This table shows the survival analysis performed on 8 independent datasets. These datasets were generated by various microarray platforms. The association between deconvolution results and poor survival outcome was reproducible in totally eight independent datasets, including TCGA dataset, AOCS dataset and six additional microarray datasets from the CuratedOvarianData database [55]. The n refers to the number of samples in a dataset. CI, confidence interval.

5.4 Discussion

In this chapter, I used the droplet- and plate-based scRNA-seq methods to profile the cellular transcriptomes of human fallopian tubes and identified the complex interactive network between cell populations as well as novel secretory populations. This study is the first comprehensive transcriptomic profiling of human fallopian tubes at the single-cell level. The catalogues of all cell subtypes will lay a foundation for future studies in both ovarian cancer and the female reproductive system.

As I mentioned in the Introduction section, the third known epithelial population is the basal cells. A previous study reported that the basal cells were positive for CD44 and EpCAM but negative for Pax8 or ciliated markers [139]. However, another study suggested that the basal population are infiltrated T lymphocytes [3]. If so, these basal T lymphocytes are positive for the epithelial marker, EpCAM. The underlying mechanism has not been studied. However, the obstacle to studying the transcriptome of this population is its scarcity. This population is, consequentially, difficult to be obtained with the cell sorting technique. Moreover, because this population is double-positive for EpCAM and CD3, cell sorting isolation can be easily interfered by the auto-fluorescent cells or doublets, which are also double positive. One plausible solution is to use infection to boost the number of infiltrated T lymphocytes. Chlamydia is the most common infection in fallopian tubes, which is reported to be potentially associated with an increased risk of ovarian cancer [185]. Therefore, as for the future work, we could use Chlamydia to infect mouse oviductal cells and study the oviductal basal cells in this model.

With the advent of scRNA-seq techniques, a more comprehensive understanding of tumours is obtained. Rather than a bunch of monotonous cells under uncontrollable cell division, cancer consists of a complicated ecosystem [131]. In this project, I used the deconvolution method to score HGSOC bulk samples and found that the proportion of the ECM cell state is associated with patient survival and other clinical factors. The definition of the ECM-high subtype in this study overlaps with the previously defined mesenchymal subtype [184], but the scoring system proposed in my study outperformed the previous clustering-based subtyping system.

The association between the EMT enrichment and clinical factors, though not having correlations of over 0.5, should not be overlooked, as causality may exist. In my finding, the patient's age and the EMT enrichment have a positively correlated relationship. As a support for my finding, a recent study showed that ageing is associated with fibroblast senesce that secrets sFRP2, which enhances tumour invasion and drug resistance in melanoma [87]. sFRP4, as one of the ECM markers, is another soluble frizzled-related protein, I speculate that the association between ageing and EMT enrichment may also have a similar mechanism involving sFRP4.

This study elucidates the association between the cell subtypes of origin and cell states in bulk cancer, which can be explained by two existing models. In the first model, multiple cancer molecular subtypes arise from one specific cell type in the tissue-of-origin. The tumour cells follow the differentiation trajectory inherited from the cell-of-origin, and differentiate into diverse subtypes, which is regulated by genetic, epigenetic and other mechanisms. In the alternative model, each tumour subtype originates from its corresponding cell subtype in the tissue of origin, while the plasticity allows cancer cells to transfer from one state to another [59]. Both hypotheses support the association found in this study. To further investigate the underpinning, future studies can test the capability of tumorigenesis of each secretory subtype in vitro and in vivo models [146].

There is a large amount of bulk gene expression data available (see Section 1.2.2.1). CIBERSORT and other deconvolution algorithms act as the connection between scRNA-seq data and bulk expression data. By using the transcriptomic signatures derived from scRNA-seq data, the deconvolution analysis can reveal the complexity in the bulk tissues. Newman et al. showed that CIBERSORT predicted the proportions of cellular components in lung and other tissues with relatively high accuracy by comparing with the flow cytometry fractions [132]. Moreover, this computational method is cost efficient and allows researchers to take full advantage of previously published datasets. CIBERSORT, in essence, is a linear regression model, to be more specific, a support vector regression; thus, the alternative options for it include other linear regression models, such as the classic (e.g. ordinary least squares) and feature selection models (e.g. ridge and lasso). Many other avaiable tools are listed

in a recent benchmarking preprint [40]. Although CIBERSORT can often generate useful estimation, it is difficult to anticipate its accuracy in a tissue where it has not been fully validated by experiments. CIBERSORT relies on the assumption that the the relationship between bulk expression data and transcriptomic signatures of cell subtypes is linear. Thus, if this relationship is nonlinear due to technical issues or some key variables (cell subtypes) are not included in the model, the acurracy of prediction will not be desirable. When a high accuracy is required, it is necessary to validate the predicted result by experiments, such as using flow cytometry to estimate the fraction of each cell type in a bulk tissue.

My work only proves the association between the tumour subtypes and the FTE cell subtypes, but the underlying mechanism remains elusive. Future work is needed to study whether the diverse tumour subtypes have the same or different cell subtypes of origin.

According to these results, the EMT process may play an important role in the ECM-high tumours. It will be an interesting direction for future studies. Moreover, previous studies also revealed that the mesenchymal (ECM-high) subtype is related to the infiltration of stromal cells [207] and certain drugs may outperform the traditional treatment in this subtype [97, 128]. However, more work will be needed to explore the therapeutical opportunities for the ECM-high subtype.

Overall, this work has comprehensively elucidated the cellular subtypes in the human fallopian tubes and shed light on how the finding of novel secretory populations facilitates the understanding of ovarian cancer.

Chapter 6

Discussion and conclusions

6.1 Overview

In this thesis, I exploited a variety of functional genomics approaches to study cancer and cells of origin. These studies were aimed at improving our understanding of cancer, which would facilitate the advancement of molecular diagnosis and cancer treatment.

In the pan-cancer analysis of NMD, I developed an algorithm to predict the NMD-elicit mutations that can cause loss of expression, and then used it to identify the gene targets of NMD in the TCGA pan-cancer cohort (see Chapter 3.1). The systematical analysis revealed not only the prevalence of NMD in the genomes across cancer types but also the permissive role that NMD plays in cancer progression.

Next, I moved from the pan-cancer study to the single-cell field, whilst still using functional genomics methodologies. The following two chapters were focused on an emerging technique, scRNA-seq. I used scRNA-seq to depict the cell populations and subpopulations in the human fallopian tube epithelium, which is the putative origin of serous ovarian cancer. First I built a clustering algorithm (ClinCluster) that was tailored for clinical scRNA-seq data (see Chapter 4.1). By benchmarking this algorithm with other frequently used computational tools, the analysis demonstrated that it outperformed the other methods in both simulated and real datasets with

confounding inter-patient variability.

The last part was aimed at improving the molecular classifier of ovarian cancer, because the existing ones were not reproducible across independent datasets. The scRNA-seq data of cells was acquired from human fallopian tubes. By applying ClinCluster on this dataset, this work profiled previously reported cell types (see Section 5.3.2) and discovered new secretory subtypes (see Section 5.3.4). Based on this knowledge, the analysis decomposed the expression data of bulk ovarian cancer. This led to the establishment of a molecular classifier that can predict prognosis with higher sensitivity (see Section 5.3.5).

The contribution of the work presented in this thesis is manifested in three aspects. The first aspect is the development and validation of new computational methods. This thesis demonstrated two algorithms, masonmd and ClinCluster. Masonmd has been released as an R package (`https://github.com/zhiyhu/masonmd`). Previous studies have suggested the importance of NMD in cancer development [111] and the possibility of treating cancer with NMD inhibitors [107]. However, no genome-wide NMD predictor or annotation was accessible for cancer genomes. This work provides an accessible and efficient tool that can be used to predict NMD targets. Moreover, the integrative analysis of TCGA expression data and genomic data shows that the NMD-elicit mutation was associated with the significant downregulation of expression levels in the mutation-harbouring genes. It validated the effectiveness of masonmd.

The other algorithm, ClinCluster, was designed for the clinical scRNA-seq data. Although there are plenty of clustering algorithms for scRNA-seq data [24, 68, 95, 182], no method has been tailored specifically for the data generated from clinical samples. Furthermore, the performance of clustering algorithms largely depends on the structure of data. Arguably, ClinCluster narrows the gap between data acquiring and data interpretation of clinical samples. Unlike other batch correction approaches, ClinCluster does not require strong hypotheses that batch effects are orthogonal to biological variability and that the composition of cellular types is consistent across samples. It is based on a functional similarity hypothesis, in which the inter-population discrepancy exceeds inter-sample diversity in terms of differential expres-

sion. With the rapid growth of clinical scRNA-seq data [110, 131, 143, 154, 183], ClinCluster can facilitate rational interpretation of this type of datasets.

The second aspect of significance is the construction of the first cell atlas of the human fallopian tubes. Further work in our group has validated the novel secretory subtypes using protein assays [73]. Whether a key component of the female reproductive system or the origin of ovarian cancer, the fallopian tube, though small in size, is a noteworthy tissue in research [164]. The knowledge of this tissue is also valuable for studies on ovarian cancer. For example, Perets et al. used Pax8 to specifically knockout tumour suppressor genes in the FTESCs, proving that HGSOC can originate from secretory cells [146]. However, previous knowledge was limited to a monotonous Pax8-expressing secretory population, which hindered a more accurate definition of cell subtypes of origin. Nevertheless, the comprehensive profiling of secretory subtypes provides an opportunity to identify a more narrow window of tumorigenesis for HGSOC.

Thirdly, this study found new strategies for cancer diagnosis and treatment. The pan-cancer study confirmed that NMD is a potential therapeutic target by analysing over eight thousand cancer genomes. Apart from the NMD inhibitors that can rescue mutated tumour suppressors [25, 107, 121], inhibition of NMD can also be used to induce the expression of neo-antigens in cancer cells, which can be potentially used in chemotherapy [142]. The clinical implications of this ovarian cancer study are reflected in the refinement of the HGSOC molecular classifier, which is significantly associated with patients' overall survival. More work will be done to verify its robustness and explore the potential clinical utility for precise medicine.

6.2 Future directions

The prognostic predictor can potentially improve the stratification of patients and the selection of therapies. However, this work has only validated the association between EMT enrichment and patients' overall survival in the published dataset. Future work will further validate this relationship in independent cohorts. Moreover, since the treatment outcome information is incomplete in the TCGA dataset, the

statistical testing between the chemotherapy response and the proportion of EMT cells did not reach a significant p-value. This association will be tested in the new cohorts that have complete data of treatment outcome.

Although the deconvolution of bulk tumours reflected the intra-tumour heterogeneity at the transcriptomic level, the biological underpinning remains unclear. Further work of our group found that the EMT enrichment was significantly increased after chemotherapy in patients by comparing the expression data of pre- and post-chemotherapy tumour biopsies from matching patients (M Artibani et al., unpublished data). This is consistent with a recent finding that an EMT signature was upregulated in the resistant triple-negative breast tumours after neoadjuvant chemotherapy [90]. I speculate that the poorer survival of the EMT-high HGSOC subtype is caused by drug resistance. This notion has been introduced previously [170]. More work is needed to verify this notion in HGSOC and explore the way of repressing EMT, which will sensitise tumour cells for chemotherapy and improve patients' survival.

6.3 Conclusions

Overall, this work has identified the NMD pathway as a promising drug target across multiple cancer types, constructed a cell atlas of human fallopian tubes, and developed a classifier of molecular subtypes for HGSOC. It also reflects the genomic and phenotypic heterogeneity of tumours at multiple levels.

Chapter 7

Appendix

Table 7.1. Marker genes of cell populations in human fallopian tubes.

Avg logFC	Pct.1	Pct.2	Adjusted p	Cluster	Gene
1.68	0.936	0.213	2.81e−109	Secretory	*CLDN4*
1.59	0.906	0.223	9.46e−98	Secretory	*ATF3*
1.48	0.7	0.178	6.02e−55	Secretory	*CXCL2*
1.45	0.995	0.436	1.13e−76	Secretory	*JUN*
1.41	0.872	0.19	2.34e−87	Secretory	*ELF3*
1.41	0.64	0.109	1.50e−69	Secretory	*HES1*
1.39	0.921	0.258	1.52e−84	Secretory	*TACSTD2*
1.37	0.877	0.171	1.35e−95	Secretory	*CYR61*
1.18	0.961	0.416	8.86e−66	Secretory	*KRT18*
1.18	0.995	0.544	2.45e−54	Secretory	*FOS*
1.17	0.961	0.359	1.21e−69	Secretory	*KRT8*
1.17	0.962	0.364	1.04e−73	Secretory	*MMP7*
1.14	0.498	0.15	1.86e−26	Secretory	*OVGP1*
1.13	1	0.586	8.58e−78	Secretory	*CLU*

1.13	0.621	0.136	1.79e−51	Secretory	*CTGF*
3.85	0.986	0.16	1.42e−83	Ciliated	*TPPP3*
3.49	1	0.225	8.31e−70	Ciliated	*CAPS*
3.26	0.971	0.066	1.10e−135	Ciliated	*C9orf24*
3.19	1	0.044	1.03e−171	Ciliated	*C20orf85*
3.18	0.929	0.068	6.67e−123	Ciliated	*C1orf194*
3.16	0.871	0.052	4.59e−127	Ciliated	*TMEM190*
3.13	0.986	0.05	3.36e−159	Ciliated	*FAM183A*
2.82	0.929	0.252	7.88e−51	Ciliated	*AGR2*
2.80	0.943	0.206	5.25e−63	Ciliated	*AGR3*
2.72	0.986	0.06	1.59e−146	Ciliated	*RSPH1*
2.66	0.929	0.029	2.28e−180	Ciliated	*SNTN*
2.65	0.957	0.044	2.84e−160	Ciliated	*C5orf49*
2.60	0.914	0.054	3.70e−135	Ciliated	*PIFO*
2.58	0.957	0.266	1.25e−55	Ciliated	*CETN2*
2.53	0.9	0.028	2.84e−174	Ciliated	*C11orf88*
1.93	0.926	0.158	5.11e−130	Lymphocyte	*CCL5*
1.79	0.477	0.074	5.35e−58	Lymphocyte	*IL7R*
1.78	0.803	0.101	3.90e−126	Lymphocyte	*CD7*
1.78	0.677	0.194	4.28e−56	Lymphocyte	*CCL4*
1.75	0.432	0.103	2.99e−35	Lymphocyte	*KLRB1*
1.68	0.932	0.392	5.54e−74	Lymphocyte	*CXCR4*
1.67	0.773	0.265	6.07e−67	Lymphocyte	*RGCC*
1.58	0.978	0.612	2.15e−86	Lymphocyte	*ZFP36L2*
1.57	0.865	0.287	3.77e−80	Lymphocyte	*CREM*

1.55	0.895	0.225	1.26e−95	Lymphocyte	*IL32*
1.55	0.301	0.058	7.54e−28	Lymphocyte	*GNLY*
1.54	0.3	0.061	7.56e−25	Lymphocyte	*GZMK*
1.54	0.699	0.099	4.33e−93	Lymphocyte	*NKG7*
1.50	0.345	0.027	3.71e−56	Lymphocyte	*XCL1*
1.50	0.572	0.05	2.12e−96	Lymphocyte	*CD8A*
3.69	0.759	0.024	7.92e−152	Monocyte	*LYZ*
3.16	0.566	0.043	2.23e−72	Monocyte	*IL1B*
3.00	0.518	0.021	5.02e−91	Monocyte	*C1QA*
2.85	0.892	0.028	1.60e−183	Monocyte	*AIF1*
2.81	0.434	0.073	5.83e−27	Monocyte	*CCL3*
2.69	0.325	0.007	2.42e−69	Monocyte	*S100A8*
2.68	0.325	0.016	9.43e−48	Monocyte	*CCL3L1*
2.67	0.47	0.021	5.51e−78	Monocyte	*C1QB*
2.66	0.904	0.084	2.75e−110	Monocyte	*TYROBP*
2.59	0.53	0.091	2.13e−34	Monocyte	*IL8*
2.58	0.361	0.034	1.32e−35	Monocyte	*S100A9*
2.58	0.422	0.067	2.45e−27	Monocyte	*APOE*
2.57	0.819	0.14	6.31e−66	Monocyte	*PLAUR*
2.56	0.482	0.017	2.00e−88	Monocyte	*C1QC*
2.51	0.88	0.073	2.65e−113	Monocyte	*FCER1G*
5.69	1	0.011	3.54e−194	Dendritic	*TPSAB1*
3.03	0.923	0.011	8.06e−177	Dendritic	*HPGDS*
2.39	0.731	0.002	6.31e−195	Dendritic	*CPA3*
2.37	0.731	0.066	3.47e−36	Dendritic	*VWA5A*

2.26	0.769	0.002	1.19e−206	Dendritic	*RGS13*
2.21	0.654	0.028	1.63e−57	Dendritic	*GATA2*
2.19	0.731	0.356	1.54e−05	Dendritic	*GLUL*
2.13	0.654	0	1.22e−201	Dendritic	*HDC*
2.11	0.731	0	5.78e−226	Dendritic	*SLC18A2*
2.04	0.885	0.501	9.48e−09	Dendritic	*LMNA*
1.96	0.846	0.322	5.90e−10	Dendritic	*BIRC3*
1.96	0.731	0.124	7.88e−18	Dendritic	*PLIN2*
1.95	0.692	0.145	1.17e−12	Dendritic	*GPR65*
1.88	0.192	0	5.64e−57	Dendritic	*CTSG*
1.87	0.462	0.032	6.26e−26	Dendritic	*HPGD*
3.85	0.92	0.039	2.06e−184	Fibroblast 1	*SFRP4*
3.77	0.95	0.074	1.66e−150	Fibroblast 1	*DCN*
3.10	0.89	0.045	6.73e−163	Fibroblast 1	*IGFBP5*
3.05	0.93	0.078	1.96e−141	Fibroblast 1	*FBLN1*
3.00	0.81	0.025	9.17e−177	Fibroblast 1	*PTGDS*
2.73	0.67	0.022	4.34e−140	Fibroblast 1	*LUM*
2.72	0.7	0.004	1.81e−195	Fibroblast 1	*MEG3*
2.61	0.91	0.317	1.86e−59	Fibroblast 1	*SPARCL1*
2.60	0.77	0.079	9.94e−98	Fibroblast 1	*IGFBP6*
2.59	0.46	0.021	1.55e−80	Fibroblast 1	*APOD*
2.55	0.87	0.026	2.96e−192	Fibroblast 1	*SERPINF1*
2.23	0.68	0.109	6.90e−59	Fibroblast 1	*ACTA2*
2.15	0.94	0.319	3.19e−61	Fibroblast 1	*LGALS1*
2.12	0.64	0.009	1.54e−160	Fibroblast 1	*IGF2*

2.07	0.57	0.092	8.38e−45	Fibroblast 1	*MGP*
3.93	0.957	0.042	1.48e−79	Fibroblast 2	*RGS5*
3.33	0.826	0.149	6.40e−18	Fibroblast 2	*CCL2*
2.92	0.913	0.086	3.96e−40	Fibroblast 2	*TAGLN*
2.89	0.957	0.442	4.73e−10	Fibroblast 2	*MT2A*
2.84	1	0.298	7.89e−19	Fibroblast 2	*ADIRF*
2.79	0.913	0.136	1.93e−25	Fibroblast 2	*ACTA2*
2.75	0.87	0.07	9.54e−43	Fibroblast 2	*C11orf96*
2.59	0.87	0.009	5.12e−163	Fibroblast 2	*NDUFA4L2*
2.47	1	0.327	1.54e−17	Fibroblast 2	*MYL9*
2.46	0.739	0.067	4.82e−31	Fibroblast 2	*GEM*
2.45	0.348	0.041	2.89e−08	Fibroblast 2	*IL6*
2.44	0.826	0.003	2.06e−210	Fibroblast 2	*HIGD1B*
2.39	0.87	0.058	4.34e−51	Fibroblast 2	*GJA4*
2.39	0.217	0.001	2.17e−53	Fibroblast 2	*FABP4*
2.35	0.609	0.006	7.95e−113	Fibroblast 2	*MT1A*
4.89	0.389	0.015	1.27e−25	Endothelial	*CCL21*
2.88	0.944	0.128	4.61e−25	Endothelial	*GNG11*
2.83	0.778	0.025	3.68e−66	Endothelial	*SDPR*
2.74	0.889	0.334	2.58e−08	Endothelial	*TM4SF1*
2.71	0.889	0.153	1.39e−17	Endothelial	*EMP1*
2.65	0.833	0.015	1.29e−103	Endothelial	*RAMP2*
2.26	0.889	0.276	3.08e−09	Endothelial	*SOCS3*
2.26	0.5	0.002	5.22e−114	Endothelial	*CLDN5*
2.22	0.389	0.06	7.25e−05	Endothelial	*TFF3*

2.06	0.667	0.003	7.16e−154	Endothelial	*ECSCR*
2.02	0.667	0.028	1.53e−43	Endothelial	*THBD*
1.96	0.889	0.277	8.87e−09	Endothelial	*CDKN1A*
1.92	0.778	0.003	2.90e−187	Endothelial	*EMCN*
1.88	0.5	0.076	6.96e−08	Endothelial	*A2M*
1.82	0.667	0.013	5.57e−78	Endothelial	*AQP1*

References

[1] Ashour Ahmed Ahmed, Dariush Etemadmoghadam, Jillian Temple, Andy G Lynch, Mohamed Riad, Raghwa Sharma, Colin Stewart, Sian Fereday, Carlos Caldas, Anna DeFazio, David Bowtell, James D Brenton, and Australian Ovarian Canc Study Grp. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *Journal of Pathology*, 221(1):49–56, May 2010.

[2] Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel A J R Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Åke Borg, Anne-Lise Børresen-Dale, Sandrine Boyault, Birgit Burkhardt, Adam P Butler, Carlos Caldas, Helen R Davies, Christine Desmedt, Roland Eils, Jórunn Erla Eyfjörd, John A Foekens, Mel Greaves, Fumie Hosoda, Barbara Hutter, Tomislav Ilicic, Sandrine Imbeaud, Marcin Imielinski, Natalie Jäger, David T W Jones, David Jones, Stian Knappskog, Marcel Kool, Sunil R Lakhani, Carlos López-Otín, Sancha Martin, Nikhil C Munshi, Hiromi Nakamura, Paul A Northcott, Marina Pajic, Elli Papaemmanuil, Angelo Paradiso, John V Pearson, Xose S Puente, Keiran Raine, Manasa Ramakrishna, Andrea L Richardson, Julia Richter, Philip Rosenstiel, Matthias Schlesner, Ton N Schumacher, Paul N Span, Jon W Teague, Yasushi Totoki, Andrew N J Tutt, Rafael Valdés-Mas, Marit M van Buuren, Laura van t Veer, Anne Vincent-Salomon, Nicola Waddell, Lucy R Yates, Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Jessica Zucman-Rossi, P Andrew Futreal, Ultan McDermott, Peter Lichter, Matthew Meyerson, Sean M Grimmond, Reiner Siebert, Elias Campo, Tatsuhiro Shibata, Stefan M Pfister, Peter J Campbell,

and Michael R Stratton. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.

[3] Ardighieri, Laura, Lonardi, Silvia, Moratto, Daniele, Facchetti, Fabio, Shih, Ie-Ming, Vermi, William, and Kurman, Robert J. Characterization of the immune cell repertoire in the normal fallopian tube. *International journal of gynecological pathology : official journal of the International Society of Gynecological Pathologists*, 33(6):581–591, November 2014.

[4] Arribere, Joshua A and Gilbert, Wendy V. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome research*, 23(6):977–987, June 2013.

[5] S Artavanis-Tsakonas, M D Rand, and R J Lake. Notch signalling: cell fate control and signal integration in development. *Science (New York, N.Y.)*, 284(5415):770–776, April 1999.

[6] Carlos L Arteaga and Jeffrey A Engelman. ERBB Receptors: From Oncogene Discovery to Basic Science to Mechanism-Based Cancer Therapeutics. *Cancer Cell*, 25(3):282–303, March 2014.

[7] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome biology*, 20(1):206, October 2019.

[8] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, 3(4):346–+, 2016.

[9] Barretina, Jordi, Caponigro, Giordano, Stransky, Nicolas, Venkatesan, Kavitha, Margolin, Adam A, Kim, Sungjoon, Wilson, Christopher J, Lehár, Joseph, Kryukov, Gregory V, Sonkin, Dmitriy, Reddy, Anupama, Liu, Manway, Murray, Lauren, Berger, Michael F, Monahan, John E, Morais, Paula,

Meltzer, Jodi, Korejwa, Adam, Jané-Valbuena, Judit, Mapa, Felipa A, Thibault, Joseph, Bric-Furlong, Eva, Raman, Pichai, Shipway, Aaron, Engels, Ingo H, Cheng, Jill, Yu, Guoying K, Yu, Jianjun, Aspesi, Peter, de Silva, Melanie, Jagtap, Kalpana, Jones, Michael D, Wang, Li, Hatton, Charles, Palescandolo, Emanuele, Gupta, Supriya, Mahan, Scott, Sougnez, Carrie, Onofrio, Robert C, Liefeld, Ted, MacConaill, Laura, Winckler, Wendy, Reich, Michael, Li, Nanxin, Mesirov, Jill P, Gabriel, Stacey B, Getz, Gad, Ardlie, Kristin, Chan, Vivien, Myer, Vic E, Weber, Barbara L, Porter, Jeff, Warmuth, Markus, Finan, Peter, Harris, Jennifer L, Meyerson, Matthew, Golub, Todd R, Morrissey, Michael P, Sellers, William R, Schlegel, Robert, and Garraway, Levi A. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, March 2012.

[10] Anita Becker-Heck, Irene E Zohn, Noriko Okabe, Andrew Pollock, Kari Baker Lenhart, Jessica Sullivan-Brown, Jason McSheene, Niki T Loges, Heike Olbrich, Karsten Haeffner, Manfred Fliegauf, Judith Horvath, Richard Reinhardt, Kim G Nielsen, June K Marthin, Gyorgy Baktai, Kathryn V Anderson, Robert Geisler, Lee Niswander, Heymut Omran, and Rebecca D Burdine. The coiled-coil domain containing protein CCDC40 is essential for motile cilia function and left-right axis formation. *Nature genetics*, 43(1):79–84, January 2011.

[11] D Bell, A Berchuck, M Birrer, J Chien, D W Cramer, F Dao, R Dhir, P DiSaia, H Gabra, P Glenn, A K Godwin, J Gross, L Hartmann, M Huang, D G Huntsman, M Iacocca, M Imielinski, S Kalloger, B Y Karlan, G B Mills, C Morrison, D Mutch, N Olvera, S Orsulic, K Park, N Petrelli, B Rabeno, J S Rader, B I Sikic, K Smith-McCune, A K Sood, D Bowtell, J R Testa, K Chang, H H Dinh, J A Drummond, G Fowler, P Gunaratne, A C Hawes, C L Kovar, L R Lewis, M B Morgan, I F Newsham, J Santibanez, J G Reid, L R Trevino, Y Q Wu, M Wang, D M Muzny, D A Wheeler, R A Gibbs, K Cibulskis, A Y Sivachenko, C Sougnez, J Wilkinson, T Bloom, T Fennell, J Baldwin, S Gabriel, E S Lander, L Ding, R S Fulton, D C Koboldt, M D McLellan, T Wylie, J Walker, M O'Laughlin, D J Dooling, L Fulton, R Abbott, N D Dees, Q Zhang, C Kandoth, M Wendl, W Schierding, D Shen, C C Harris,

H Schmidt, J Kalicki, K D Delehaunty, C C Fronick, R Demeter, L Cook, J W Wallis, L Lin, V J Magrini, J S Hodges, J M Eldred, S M Smith, C S Pohl, F Vandin, B J Raphael, G M Weinstock, R Mardis, R K Wilson, M Meyerson, W Winckler, G Getz, R G W Verhaak, S L Carter, C H Mermel, H Nguyen, R C Onofrio, M S Lawrence, D Hubbard, S Gupta, A Crenshaw, A H Ramos, K Ardlie, L Chin, A Protopopov, Juinhua Zhang, T M Kim, I Perna, Y Xiao, G Ren, N Sathiamoorthy, R W Park, E Lee, R Kucherlapati, D M Absher, L Waite, G Sherlock, J D Brooks, J Z Li, J Xu, R M Myers, P W Laird, L Cope, J G Herman, H Shen, D J Weisenberger, H Noushmehr, F Pan, T Jr Triche, B P Berman, D J Van den Berg, J Buckley, S B Baylin, P T Spellman, P Neuvial, H Bengtsson, L R Jakkula, S Dorton, H Marr, Y G Choi, V Wang, N J Wang, J Ngai, J G Conboy, H S Feiler, N D Socci, Y Liang, B S Taylor, A E Lash, C Brennan, A Viale, K A Hoadley, S Meng, Y Du, Y Shi, L Li, Y J Turman, D Zang, E B Helms, S Balu, X Zhou, J Wu, M D Topal, D N Hayes, C M Perou, D Voet, G Saksena, Junihua Zhang, H Zhang, C J Wu, S Shukla, A Sivachenko, R Jing, Y Liu, P J Park, M Noble, H Carter, D Kim, R Karchin, E Purdom, S Durinck, J Han, J E Korkola, L M Heiser, R J Cho, Z Hu, B Parvin, T P Speed, J W Gray, N Schultz, E Cerami, A Olshen, B Reva, Y Antipin, R Shen, P Mankoo, R Sheridan, G Ciriello, W K Chang, J A Bernanke, L Borsu, D A Levine, M Ladanyi, C Sander, D Haussler, C C Benz, J M Stuart, S C Benz, J Z Sanborn, C J Vaske, J Zhu, C Szeto, G K Scott, C Yau, M D Wilkerson, N Zhang, R Akbani, K A Baggerly, W K Yung, J N Weinstein, R Penny, T Shelton, D Grimm, M Hatfield, S Morris, P Yena, P Rhodes, M Sherman, J Paulauskis, S Millis, A Kahn, J M Greene, R Sfeir, M A Jensen, J Chen, J Whitmore, S Alonso, J Jordan, A Chu, Jinghui Zhang, A Barker, C Compton, G Eley, M Ferguson, P Fielding, D S Gerhard, R Myles, C Schaefer, K R Mills Shaw, J Vaught, J B Vockley, P J Good, M S Guyer, B Ozenberger, J Peterson, E Thomson, and Canc Genome Atlas Res Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 2011.

[12] Richard Bellman. Dynamic Programming. Princeton University Press, 1957.

[13] Bentink, Stefan, Haibe-Kains, Benjamin, Risch, Thomas, Fan, Jian Bing,

Hirsch, Michelle S, Holton, Kristina, Rubio, Renee, April, Craig, Chen, Jing, Wickham-Garcia, Eliza, Liu, Joyce, Culhane, Aedin, Drapkin, Ronny, Quackenbush, John, and Matulonis, Ursula A. Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS one*, 7(2):e30269, February 2012.

[14] Berger, Alice H, Brooks, Angela N, Wu, Xiaoyun, Shrestha, Yashaswi, Chouinard, Candace, Piccioni, Federica, Bagul, Mukta, Kamburov, Atanas, Imielinski, Marcin, Hogstrom, Larson, Zhu, Cong, Yang, Xiaoping, Pantel, Sasha, Sakai, Ryo, Watson, Jacqueline, Kaplan, Nathan, Campbell, Joshua D, Singh, Shantanu, Root, David E, Narayan, Rajiv, Natoli, Ted, Lahr, David L, Tirosh, Itay, Tamayo, Pablo, Getz, Gad, Wong, Bang, Doench, John, Subramanian, Aravind, Golub, Todd R, Meyerson, Matthew, and Boehm, Jesse S. High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell*, 30(2):214–228, August 2016.

[15] Pascal Bernard and Vincent R Harley. Wnt4 action in gonadal development and sex determination. *The international journal of biochemistry & cell biology*, 39(1):31–43, 2007.

[16] Theresa Berndt, Theodore A Craig, Ann E Bowe, John Vassiliadis, David Reczek, Richard Finnegan, Suzanne M Jan De Beur, Susan C Schiavi, and Rajiv Kumar. Secreted frizzled-related protein 4 is a potent tumor-derived phosphaturic agent. *The Journal of clinical investigation*, 112(5):785–794, September 2003.

[17] Jonna Berntsson, Sebastian Lundgren, Björn Nodin, Mathias Uhlén, Alexander Gaber, and Karin Jirström. Expression and prognostic significance of the polymeric immunoglobulin receptor in epithelial ovarian cancer. *Journal of ovarian research*, 7(1):26, February 2014.

[18] Mikhail Binnewies, Edward W Roberts, Kelly Kersten, Vincent Chan, Douglas F Fearon, Miriam Merad, Lisa M Coussens, Dmitry I Gabrilovich, Suzanne Ostrand-Rosenberg, Catherine C Hedrick, Robert H Vonderheide, Mikael J Pittet, Rakesh K Jain, Weiping Zou, T Kevin Howcroft, Elisa C

Woodhouse, Robert A Weinberg, and Matthew F Krummel. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nature medicine*, 24(5):541–550, May 2018.

[19] Matthew L Bochman and Anthony Schwacha. The Mcm complex: unwinding the mechanism of a replicative helicase. *Microbiology and molecular biology reviews : MMBR*, 73(4):652–683, December 2009.

[20] Ulrich Bodenhofer, Andreas Kothmeier, and Sepp Hochreiter. APCluster: an R package for affinity propagation clustering. *Bioinformatics (Oxford, England)*, 27(17):2463–2464, September 2011.

[21] Bonome, Tomas, Levine, Douglas A, Shih, Joanna, Randonovich, Mike, Pise-Masison, Cindy A, Bogomolniy, Faina, Ozbun, Laurent, Brady, John, Barrett, J Carl, Boyd, Jeff, and Birrer, Michael J. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer research*, 68(13):5478–5486, July 2008.

[22] Marko T Boskovski, Shiaulou Yuan, Nis Borbye Pedersen, Christoffer Knak Goth, Svetlana Makova, Henrik Clausen, Martina Brueckner, and Mustafa K Khokha. The heterotaxy gene GALNT11 glycosylates Notch to orchestrate cilia type and laterality. *Nature*, 504(7480):456–459, December 2013.

[23] Saverio Brogna and Jikai Wen. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature structural & molecular biology*, 16(2):107–113, February 2009.

[24] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–+, May 2018.

[25] Bykov, Vladimir J N, Eriksson, Sofi E, Bianchi, Julie, and Wiman, Klas G. Targeting mutant p53 for efficient cancer therapy. *Nature Reviews Cancer*, 18(2):89–102, February 2018.

[26] Kieran R Campbell and Christopher Yau. Uncovering pseudotemporal tra-

jectories with covariates from single cell and bulk expression data. *Nature communications*, 9(1), 2018.

[27] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, October 2008.

[28] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[29] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, October 2013.

[30] Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science (New York, N.Y.)*, 357(6352):661–667, August 2017.

[31] D A Carson and A Lois. Cancer progression and p53. *Lancet (London, England)*, 346(8981):1009–1011, October 1995.

[32] Chen, F, Gaitskell, K, Garcia, M J, Albukhari, A, Tsaltas, J, and Ahmed, A A. Serous tubal intraepithelial carcinomas associated with high-grade serous ovarian carcinomas: a systematic review. *BJOG : an international journal of obstetrics and gynaecology*, 124(6):872–878, May 2017.

[33] Chen, Gregory M, Kannan, Lavanya, Geistlinger, Ludwig, Kofia, Victor, Safikhani, Zhaleh, Gendoo, Deena M A, Parmigiani, Giovanni, Birrer, Michael, Haibe-Kains, Benjamin, and Waldron, Levi. Consensus on Molecular Subtypes of High-Grade Serous Ovarian Carcinoma. *Clinical Cancer Research*, 24(20):5037–5047, October 2018.

[34] P W Chiang, W J Song, K Y Wu, J R Korenberg, E J Fogel, M L Van Keuren, D Lashkari, and D M Kurnit. Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome research*, 6(10):1013–1026, October 1996.

[35] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han, and Woong-Yang Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8, 2017.

[36] D Cibula, M Widschwendter, O Májek, and L Dusek. Tubal ligation and the risk of ovarian cancer: review and meta-analysis. *Human reproduction update*, 17(1):55–67, January 2011.

[37] Cieślik, Marcin and Chinnaiyan, Arul M. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93–109, February 2018.

[38] H Clevers. Identification of stem cells in small intestine and colon by a single marker gene Lgr5. *European Journal of Cancer Supplements*, 6(9):1, July 2008.

[39] M J Clyman. Electron microscopy of the human fallopian tube. *Fertility and sterility*, 17(3):281–301, May 1966.

[40] Francisco Avila Cobos, José Alquicira-Hernandez, Joseph Powell, Pieter Mestdagh, and Katleen De Preter. Comprehensive benchmarking of computational deconvolution of transcriptomics data. *bioRxiv*, 24:2020.01.10.897116, January 2020.

[41] G E Crawford, J A Faulkner, R H Crosbie, K P Campbell, S C Froehner, and J S Chamberlain. Assembly of the dystrophin-associated protein complex does not require the dystrophin COOH-terminal domain. *Journal of Cell Biology*, 150(6):1399–1410, September 2000.

[42] Crijns, Anne P G, Fehrmann, Rudolf S N, de Jong, Steven, Gerbens, Frans, Meersma, Gert Jan, Klip, Harry G, Hollema, Harry, Hofstra, Robert M W, te Meerman, Gerard J, de Vries, Elisabeth G E, and van der Zee, Ate G J. Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS medicine*, 6(2):e24, February 2009.

[43] Davie, Kristofer, Janssens, Jasper, Koldere, Duygu, De Waegeneer, Maxime, Pech, Uli, Kreft, Łukasz, Aibar, Sara, Makhzami, Samira, Christiaens, Valerie, Bravo González-Blas, Carmen, Poovathingal, Suresh, Hulselmans, Gert, Spanier, Katina I, Moerman, Thomas, Vanspauwen, Bram, Geurs, Sarah, Voet, Thierry, Lammertyn, Jeroen, Thienpont, Bernard, Liu, Sha, Konstantinides, Nikos, Fiers, Mark, Verstreken, Patrik, and Aerts, Stein. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell*, 174(4):982–998.e20, August 2018.

[44] A B DeLeo, G Jay, E Appella, G C Dubois, L W Law, and L J Old. Detection of a transformation-related antigen in chemically induced sarcomas and other transformed cells of the mouse. *Proceedings of the National Academy of Sciences*, 76(5):2420–2424, May 1979.

[45] Deniz Demircioğlu, Engin Cukuroglu, Martin Kindermans, Tannistha Nandi, Claudia Calabrese, Nuno A Fonseca, André Kahles, Kjong-Van Lehmann, Oliver Stegle, Alvis Brazma, Angela N Brooks, Gunnar Rätsch, Patrick Tan, and Jonathan Göke. A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell*, 178(6):1465–1477.e17, September 2019.

[46] Li Ding, Michael C Wendl, Joshua F McMichael, and Benjamin J Raphael. Expanding the computational toolbox for mining cancer genomes. *Nature Reviews Genetics*, 15(8):556–570, July 2014.

[47] Ding, Jiarui, Adiconis, Xian, Simmons, Sean K, Kowalczyk, Monika S, Hession, Cynthia C, Marjanovic, Nemanja D, Hughes, Travis K, Wadsworth, Marc H, Burks, Tyler, Nguyen, Lan T, Kwon, John Y H, Barak, Boaz, Ge, William, Kedaigle, Amanda J, Carroll, Shaina, Li, Shuqiang, Hacohen, Nir,

Rozenblatt-Rosen, Orit, Shalek, Alex K, Villani, Alexandra-Chloé, Regev, Aviv, and Levin, Joshua Z. Systematic comparative analysis of single cell RNA-sequencing methods. *bioRxiv*, 10:632216, May 2019.

[48] Eran Eden, Roy Navon, Israel Steinfeld, Doron Lipson, and Zohar Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *Bmc Bioinformatics*, 10(1), 2009.

[49] Eden, Eran, Lipson, Doron, Yogev, Sivan, and Yakhini, Zohar. Discovering motifs in ranked lists of DNA sequences. *PLoS computational biology*, 3(3):508–522, March 2007.

[50] Chris D Emmerson, Els J van der Vlist, Myrthe R Braam, Peter Vanlandschoot, Pascal Merchiers, Hans J W de Haard, C Theo Verrips, Paul M P van Bergen en Henegouwen, and Edward Dolk. Enhancement of polymeric immunoglobulin receptor transcytosis by biparatopic VHH. *PloS one*, 6(10):e26299, 2011.

[51] Ghia Euskirchen, Raymond K Auerbach, and Michael Snyder. SWI/SNF chromatin-remodeling factors: multiscale analyses and diverse functions. *The Journal of biological chemistry*, 287(37):30897–30905, September 2012.

[52] Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-Eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolas Barkas, Peter J Park, Woong-Yang Park, and Peter V Kharchenko. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome research*, page gr.228080.117, June 2018.

[53] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, Peter S Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology*, 16:278, December 2015.

[54] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814):972–976, February 2007.

[55] Ganzfried, Benjamin Frederick, Riester, Markus, Haibe-Kains, Benjamin, Risch, Thomas, Tyekucheva, Svitlana, Jazic, Ina, Wang, Xin Victoria, Ahmadifar, Mahnaz, Birrer, Michael J, Parmigiani, Giovanni, Huttenhower, Curtis, and Waldron, Levi. curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database : the journal of biological databases and curation*, 2013:bat013, 2013.

[56] Greenman, Christopher, Stephens, Philip, Smith, Raffaella, Dalgliesh, Gillian L, Hunter, Christopher, Bignell, Graham, Davies, Helen, Teague, Jon, Butler, Adam, Stevens, Claire, Edkins, Sarah, O'Meara, Sarah, Vastrik, Imre, Schmidt, Esther E, Avis, Tim, Barthorpe, Syd, Bhamra, Gurpreet, Buck, Gemma, Choudhury, Bhudipa, Clements, Jody, Cole, Jennifer, Dicks, Ed, Forbes, Simon, Gray, Kris, Halliday, Kelly, Harrison, Rachel, Hills, Katy, Hinton, Jon, Jenkinson, Andy, Jones, David, Menzies, Andy, Mironenko, Tatiana, Perry, Janet, Raine, Keiran, Richardson, Dave, Shepherd, Rebecca, Small, Alexandra, Tofts, Calli, Varian, Jennifer, Webb, Tony, West, Sofie, Widaa, Sara, Yates, Andy, Cahill, Daniel P, Louis, David N, Goldstraw, Peter, Nicholson, Andrew G, Brasseur, Francis, Looijenga, Leendert, Weber, Barbara L, Chiew, Yoke-Eng, DeFazio, Anna, Greaves, Mel F, Green, Anthony R, Campbell, Peter, Birney, Ewan, Easton, Douglas F, Chenevix-Trench, Georgia, Tan, Min-Han, Khoo, Sok Kean, Teh, Bin Tean, Yuen, Siu Tsan, Leung, Suet Yi, Wooster, Richard, Futreal, P Andrew, and Stratton, Michael R. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, March 2007.

[57] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, September 2015.

[58] Gudikote, Jayanthi P, Zhao, Hao, Wang, Jing, Fan, Youhong, Diao, Lixia, Byers, Lauren A, Giri, Uma, Nilsson, Monique, and Heymach, John V. Abstract 2118: Nonsense-mediated decay regulates the expression of tumor suppressor

transcripts and cancer pathways in non-small cell lung cancer cell lines. *Cancer research*, 75(15 Supplement):2118–2118, August 2015.

[59] Piyush B Gupta, Christine M Fillmore, Guozhi Jiang, Sagi D Shapira, Kai Tao, Charlotte Kuperwasser, and Eric S Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, August 2011.

[60] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–+, May 2018.

[61] Hall, G W and Thein, S. Nonsense codon mutations in the terminal exon of the beta-globin gene are not associated with a reduction in beta-mRNA accumulation: a mechanism for the phenotype of dominant beta-thalassemia. *Blood*, 83(8):2031–2037, April 1994.

[62] Douglas Hanahan and Robert A Weinberg. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5):646–674, 2011.

[63] J W Harper, G R Adami, N Wei, K Keyomarsi, and S J Elledge. The p21 Cdk-interacting protein Cip1 is a potent inhibitor of G1 cyclin-dependent kinases. *Cell*, 75(4):805–816, November 1993.

[64] K Hechenbichler and K Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.

[65] C A Heid, J Stevens, K J Livak, and P M Williams. Real time quantitative PCR. *Genome research*, 6(10):986–994, October 1996.

[66] D W Heinz, W A Baase, F W Dahlquist, and B W Matthews. How amino-acid insertions are allowed in an alpha-helix of T4 lysozyme. *Nature*, 361(6412):561–564, February 1993.

[67] Thomas Helleday, Saeed Eshtad, and Serena Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nature reviews. Genetics*, 15(9):585–598, September 2014.

[68] Josip S Herman, Sagar, and Dominic Grün. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386, May 2018.

[69] R Higuchi, C Fockler, G Dollinger, and R Watson. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology (Nature Publishing Company)*, 11(9):1026–1030, September 1993.

[70] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max D M Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A Margolin, Laura J Van't Veer, Nuria Lopez-Bigas, Peter W Laird, Benjamin J Raphael, Li Ding, A Gordon Robertson, Lauren A Byers, Gordon B Mills, John N Weinstein, Carter Van Waes, Zhong Chen, Eric A Collisson, Cancer Genome Atlas Research Network, Christopher C Benz, Charles M Perou, and Joshua M Stuart. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, August 2014.

[71] Hochgerner, Hannah, Zeisel, Amit, Lobnnerberg, Peter, and Linnarsson, Sten. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat Neurosci*, 21(2):290–+, February 2018.

[72] Holbrook, Jill A, Neu-Yilik, Gabriele, Hentze, Matthias W, and Kulozik, Andreas E. Nonsense-mediated decay approaches the clinic. *Nature Genetics*, 36(8):801–808, August 2004.

[73] Zhiyuan Hu, Mara Artibani, Abdulkhaliq Alsaadi, Nina Wietek, Matteo Morotti, Tingyan Shi, Zhe Zhong, Laura Santana Gonzalez, Salma El-Sahhar, Mohammad KaramiNejadRanjbar, Garry Mallett, Yun Feng, Kenta Masuda, Yiyan Zheng, Kay Chong, Stephen Damato, Sunanda Dhar, Leticia Campo, Riccardo Garruto Campanile, Hooman Soleymani Majd, Vikram Rai, David Maldonado-Perez, Stephanie Jones, Vincenzo Cerundolo, Tatjana Sauka-Spengler, Christopher Yau, and Ahmed Ashour Ahmed. The Repertoire

of Serous Ovarian Cancer Non-genetic Heterogeneity Revealed by Single-Cell Sequencing of Normal Fallopian Tube Epithelial Cells. *Cancer cell*, 37(2):226–242.e7, February 2020.

[74] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, December 1985.

[75] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade, P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraburtty, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, July 2000.

[76] Carolyn Hutter and Jean Claude Zenklusen. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*, 173(2):283–285, April 2018.

[77] Ian J Jacobs, Usha Menon, Andy Ryan, Aleksandra Gentry-Maharaj, Matthew Burnell, Jatinderpal K Kalsi, Nazar N Amso, Sophia Apostolidou, Elizabeth Benjamin, Derek Cruickshank, Danielle N Crump, Susan K Davies, Anne Dawnay, Stephen Dobbs, Gwendolen Fletcher, Jeremy Ford, Keith Godfrey, Richard Gunu, Mariam Habib, Rachel Hallett, Jonathan Herod, Howard Jenkins, Chloe Karpinskyj, Simon Leeson, Sara J Lewis, William R Liston, Alberto Lopes, Tim Mould, John Murdoch, David Oram, Dustin J Rabideau, Karina Reynolds, Ian Scott, Mourad W Seif, Aarti Sharma, Naveena Singh, Julie Taylor, Fiona Warburton, Martin Widschwendter, Karin Williamson, Robert Woolas, Lesley Fallowfield, Alistair J McGuire, Stuart Campbell, Mahesh Parmar, and Steven J Skates. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *Lancet*, 387(10022):945–956, 2016.

[78] Lingyan Jin, Adam Williamson, Sudeep Banerjee, Isabelle Philipp, and Michael Rape. Mechanism of ubiquitin-chain formation by the human anaphase-promoting complex. *Cell*, 133(4):653–665, May 2008.

[79] Tsunehisa Kaku, Shinji Ogawa, Yoshiaki Kawano, Yoshihiro Ohishi, Hiroaki Kobayashi, Toshio Hirakawa, and Hitoo Nakano. Histological classification of

ovarian cancer. *Medical electron microscopy : official journal of the Clinical Electron Microscopy Society of Japan*, 36(1):9–17, March 2003.

[80] Kandoth, Cyriac, McLellan, Michael D, Vandin, Fabio, Ye, Kai, Niu, Beifang, Lu, Charles, Xie, Mingchao, Zhang, Qunyuan, McMichael, Joshua F, Wyczalkowski, Matthew A, Leiserson, Mark D M, Miller, Christopher A, Welch, John S, Walter, Matthew J, Wendl, Michael C, Ley, Timothy J, Wilson, Richard K, Raphael, Benjamin J, and Ding, Li. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339, October 2013.

[81] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, January 2000.

[82] Yibin Kang and Joan Massagué. Epithelial-mesenchymal transitions: twist in development and metastasis. *Cell*, 118(3):277–279, August 2004.

[83] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab- An S4Package for Kernel Methods in R. *Journal of statistical software*, 11(9):1–20, 2004.

[84] Karlan, Beth Y, Dering, Judy, Walsh, Christine, Orsulic, Sandra, Lester, Jenny, Anderson, Lee A, Ginther, Charles L, Fejzo, Marlena, and Slamon, Dennis. POSTN/TGFBI-associated stromal signature predicts poor prognosis in serous epithelial ovarian cancer. *Gynecologic oncology*, 132(2):334–342, February 2014.

[85] Karst, Alison M, Levanon, Keren, and Drapkin, Ronny. Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7547–7552, May 2011.

[86] Noah D Kauff, Jaya M Satagopan, Mark E Robson, Lauren Scheuer, Martee Hensley, Clifford A Hudis, Nathan A Ellis, Jeff Boyd, Patrick I Borgen, Richard R Barakat, Larry Norton, Mercedes Castiel, Khedoudja Nafa, and Kenneth Offit. Risk-reducing salpingo-oophorectomy in women with a BRCA1 or BRCA2 mutation. *The New England journal of medicine*, 346(21):1609–1615, May 2002.

[87] Amanpreet Kaur, Marie R Webster, Katie Marchbank, Reeti Behera, Abibatou Ndoye, Curtis H III Kugel, Vanessa M Dang, Jessica Appleton, Michael P O'Connell, Phil Cheng, Alexander A Valiga, Rachel Morissette, Nazli B McDonnell, Luigi Ferrucci, Andrew V Kossenkov, Katrina Meeth, Hsin-Yao Tang, Xiangfan Yin, William H III Wood, Elin Lehrmann, Kevin G Becker, Keith T Flaherty, Dennie T Frederick, Jennifer A Wargo, Zachary A Cooper, Michael T Tetzlaff, Courtney Hudgens, Katherine M Aird, Rugang Zhang, Xiaowei Xu, Qin Liu, Edmund Bartlett, Giorgos Karakousis, Zeynep Eroglu, Roger S Lo, Matthew Chan, Alexander M Menzies, Georgina V Long, Douglas B Johnson, Jeffrey Sosman, Bastian Schilling, Dirk Schadendorf, David W Speicher, Marcus Bosenberg, Antoni Ribas, and Ashani T Weeraratna. sFRP2 in the aged microenvironment drives melanoma metastasis and therapy resistance. *Nature*, 532(7598):250–+, 2016.

[88] Tim P Kerr, Caroline A Sewry, Stephanie A Robb, and Roland G Roberts. Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? *Human Genetics*, 109(4):402–407, October 2001.

[89] Mirjana Kessler, Karen Hoffmann, Volker Brinkmann, Oliver Thieck, Susan Jackisch, Benjamin Toelle, Hilmar Berger, Hans-Joachim Mollenkopf, Mandy Mangler, Jalid Sehouli, Christina Fotopoulou, and Thomas F Meyer. The Notch and Wnt pathways regulate stemness and differentiation in human fallopian tube organoids. *Nature communications*, 6(1):8989, December 2015.

[90] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E Navin. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, 173(4):879–893.e13, May 2018.

[91] Yuna Kim, Nathan Bingham, Ryohei Sekido, Keith L Parker, Robin Lovell-Badge, and Blanche Capel. Fibroblast growth factor receptor 2 regulates proliferation and Sertoli differentiation during male sex determination. *Proceedings of the National Academy of Sciences of the United States of America*, 104(42):16558–16563, October 2007.

[92] David W Kindelberger, Yonghee Lee, Alexander Miron, Michelle S Hirsch, Colleen Feltmate, Fabiola Medeiros, Michael J Callahan, Elizabeth O Garner, Robert W Gordon, Chandler Birch, Ross S Berkowitz, Michael G Muto, and Christopher P Crum. Intraepithelial carcinoma of the fimbria and pelvic serous carcinoma: Evidence for a causal relationship. *The American journal of surgical pathology*, 31(2):161–169, February 2007.

[93] Doris Kinzel, Karsten Boldt, Erica E Davis, Ingo Burtscher, Dietrich Trumbach, Bill Diplas, Tania Attie-Bitach, Wolfgang Wurst, Nicholas Katsanis, Marius Ueffing, and Heiko Lickert. Pitchfork Regulates Primary Cilia Disassembly and Left-Right Asymmetry. *Developmental Cell*, 19(1):66–77, 2010.

[94] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5):273–282, May 2019.

[95] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, May 2017.

[96] Deborah A Klos Dehring, Eszter K Vladar, Michael E Werner, Jennifer W Mitchell, Peter Hwang, and Brian J Mitchell. Deuterosome-mediated centriole biogenesis. *Developmental cell*, 27(1):103–112, October 2013.

[97] Stefan Kommoss, Boris Winterhoff, Ann L Oberg, Gottfried E Konecny, Chen Wang, Shaun M Riska, Jian-Bing Fan, Matthew J Maurer, Craig April, Viji Shridhar, Friedrich Kommoss, Andreas du Bois, Felix Hilpert, Sven Mahner, Klaus Baumann, Willibald Schroeder, Alexander Burges, Ulrich Canzler, Jeremy Chien, Andrew C Embleton, Mahesh Parmar, Richard Kaplan, Timothy Perren, Lynn C Hartmann, Ellen L Goode, Sean C Dowdy, and Jacobus Pfisterer. Bevacizumab May Differentially Improve Ovarian Cancer Outcome in Patients with Proliferative and Mesenchymal Molecular Subtypes. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 23(14):3794–3801, July 2017.

[98] Gottfried E Konecny, Chen Wang, Habib Hamidi, Boris Winterhoff, Kimberly R Kalli, Judy Dering, Charles Ginther, Hsiao-Wang Chen, Sean Dowdy, William Cliby, Bobbie Gostout, Karl C Podratz, Gary Keeney, He-Jing Wang, Lynn C Hartmann, Dennis J Slamon, and Ellen L Goode. Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. *Journal of the National Cancer Institute*, 106(10):11, October 2014.

[99] M Kress, E May, R Cassingena, and P May. Simian virus 40-transformed cells express new species of proteins precipitable by anti-simian virus 40 tumor serum. *Journal of virology*, 31(2):472–483, August 1979.

[100] Elisabetta Kuhn, Robert J Kurman, Ann Smith Sehdev, and Ie-Ming Shih. Ki-67 labeling index as an adjunct in the diagnosis of serous tubal intraepithelial carcinoma. *International journal of gynecological pathology : official journal of the International Society of Gynecological Pathologists*, 31(5):416–422, September 2012.

[101] Kimberly R Kukurba and Stephen B Montgomery. RNA Sequencing and Analysis. *Cold Spring Harbor protocols*, 2015(11):951–969, April 2015.

[102] S Intidhar Labidi-Galy, Eniko Papp, Dorothy Hallberg, Noushin Niknafs, Vilmos Adleff, Michael Noe, Rohit Bhattacharya, Marián Novak, Siân Jones, Jillian Phallen, Carolyn A Hruban, Michelle S Hirsch, Douglas I Lin, Lauren Schwartz, Cecile L Maire, Jean-Christophe Tille, Michaela Bowden, Ayse Ayhan, Laura D Wood, Robert B Scharpf, Robert Kurman, Tian-Li Wang, Ie-Ming Shih, Rachel Karchin, Ronny Drapkin, and Victor E Velculescu. High grade serous ovarian carcinomas originate in the fallopian tube. *Nature communications*, 8(1):1093, October 2017.

[103] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology*, 15(2):R29, 2014.

[104] Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V Sivakamasundari, Ina Kycia, Paul Robson, and Michael L Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type

specific expression changes in type 2 diabetes. *Genome Research*, 27(2):208–222, February 2017.

[105] Jonathan Ledermann, Philipp Harter, Charlie Gourley, Michael Friedlander, Ignace Vergote, Gordon Rustin, Clare Scott, Werner Meier, Ronnie Shapira-Frommer, Tamar Safra, Daniela Matei, Euan Macpherson, Claire Watkins, James Carmichael, and Ursula Matulonis. Olaparib maintenance therapy in platinum-sensitive relapsed ovarian cancer. *The New England journal of medicine*, 366(15):1382–1392, April 2012.

[106] Katherine Leeper, Rochelle Garcia, Elizabeth Swisher, Barbara Goff, Benjamin Greer, and Pamela Paley. Pathologic findings in prophylactic oophorectomy specimens in high-risk women. *Gynecologic oncology*, 87(1):52–56, October 2002.

[107] Lejeune, Fabrice. Triple Effect of Nonsense-Mediated mRNA Decay Inhibition as a Therapeutic Approach for Cancer. *Single Cell Biology*, 5(2), 2016.

[108] Walter Lenoir, Wenbin Liu, Huihui Fan, Hui Shen, Visweswaran Ravikumar, Arvind Rao, Andre Schultz, Xubin Li, Pavel Sumazin, Cecilia Williams, Pieter Mestdagh, Preethi H Gunaratne, Daniel G Tiezzi, Chen Wang, Nicole M Kuderer, Janet S Rader, Rosemary E Zuna, Anil K Sood, Akinyemi I Ojesina, Keith A Baggerly, Ting-Wen Chen, Hua-Sheng Chiu, Steve Lefever, Liang Liu, Karen MacKenzie, Sandra Orsulic, Jason Roszik, Qianqian Song, Christopher P Vellano, Gordon B Mills, Douglas A Levine, Rehan Akbani, Samantha J Caesar-Johnson, John A Demchok, Ina Felau, Melpomeni Kasapi, Martin L Ferguson, Carolyn M Hutter, Heidi J Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C Zenklusen, Jiashan Julia Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I Heiman, Jaegil Kim, Michael S Lawrence, Pei Lin, Sam Meier, Michael S Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Bradley M Broom, Apurva M

Hegde, Zhenlin Ju, Rupa S Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Yiling Lu, Kwok-Shing Ng, Michael Ryan, Jing Wang, John N Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E Gross, Zachary J Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G Nissan, Angelica Ochoa, Sarah M Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz, Robert Sheridan, S Onur Sumer, Yichao Sun, Barry S Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M Stuart, Christopher K Wong, Christina Yau, D Neil Hayes, Joel S Parker, Matthew D Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J M Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A Marra, Michael Mayo, Richard A Moore, Andrew J Mungall, Karen Mungall, A Gordon Robertson, Sara Sadeghi, Jacqueline E Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C Berger, Rameen Beroukhim, Andrew D Cherniack, Carrie Cibulskis, Stacey B Gabriel, Galen F Gao, Gavin Ha, Matthew Meyerson, Steven E Schumacher, Juliann Shih, Melanie H Kucherlapati, Raju S Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S Bootwalla, Phillip H Lai, Dennis T Maglinte, David J Van Den Berg, Daniel J Weisenberger, J Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A Hoadley, Alan P Hoyle, Stuart R Jefferys, Corbin D Jones, Shaowu Meng, Piotr A Mieczkowski, Lisle E Mose, Amy H Perou, Charles M Perou, Jeffrey Roach, Yan Shi, Janae V Simons, Tara Skelly, Matthew G Soloway, Donghui Tan, Umadevi Veluvolu, Toshinori Hinoue, Peter W Laird, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A Donehower, Jennifer Drummond, Richard A Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L

Appelbaum, Matthew Bailey, Matthew G Cordes, Li Ding, Catrina C Fronick, Lucinda A Fulton, Robert S Fulton, Cyriac Kandoth, Elaine R Mardis, Michael D McLellan, Christopher A Miller, Heather K Schmidt, Richard K Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, and Scott a... Morris. A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell*, April 2018.

[109] Arnold J Levine and Moshe Oren. The first 30 years of p53: growing ever more complex. *Nature Reviews Cancer*, 9(10):749–758, October 2009.

[110] Li, Huipeng, Courtois, Elise T, Sengupta, Debarka, Tan, Yuliana, Chen, Kok Hao, Goh, Jolene Jie Lin, Kong, Say Li, Chua, Clarinda, Hon, Lim Kiat, Tan, Wah Siew, Wong, Mark, Choi, Paul Jongjoon, Wee, Lawrence J K, Hillmer, Axel M, Tan, Iain Beehuat, Robson, Paul, and Prabhakar, Shyam. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5):708–718, May 2017.

[111] Rik G H Lindeboom, Fran Supek, and Ben Lehner. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature Genetics*, 48(10):1112–1118, October 2016.

[112] D I Linzer and A J Levine. Characterization of a 54K dalton cellular SV40 tumor antigen present in SV40-transformed cells and uninfected embryonal carcinoma cells. *Cell*, 17(1):43–52, May 1979.

[113] R LOSSON and F LACROUTE. Interference of Nonsense Mutations with Eukaryotic Messenger-Rna Stability. *Proceedings of the National Academy of Sciences*, 76(10):5134–5137, 1979.

[114] Love, Michael I, Huber, Wolfgang, and Anders, Simon. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):31, December 2014.

[115] Aaron T L Lun, Yunshun Chen, and Gordon K Smyth. It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using

Quasi-Likelihood Methods in edgeR. *Methods in molecular biology (Clifton, N.J.)*, 1418:391–416, 2016.

[116] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[117] Anais Malpica, Michael T Deavers, Karen Lu, Diane C Bodurka, Edward N Atkinson, David M Gershenson, and Elvio G Silva. Grading ovarian serous carcinoma using a two-tier system. *The American journal of surgical pathology*, 28(4):496–504, April 2004.

[118] Tathiane M Malta, Artem Sokolov, Andrew J Gentles, Tomasz Burzykowski, Laila Poisson, John N Weinstein, Bożena Kamińska, Joerg Huelsken, Larsson Omberg, Olivier Gevaert, Antonio Colaprico, Patrycja Czerwińska, Sylwia Mazurek, Lopa Mishra, Holger Heyn, Alex Krasnitz, Andrew K Godwin, Alexander J Lazar, Cancer Genome Atlas Research Network, Joshua M Stuart, Katherine A Hoadley, Peter W Laird, Houtan Noushmehr, and Maciej Wiznerowicz. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*, 173(2):338–354.e15, April 2018.

[119] David A Mangus, Matthew C Evans, and Allan Jacobson. Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome biology*, 4(7):223, 2003.

[120] Maquat, L E. When cells stop making sense: effects of nonsense codons on RNA metabolism in vertebrate cells. *Rna*, 1(5):453–465, July 1995.

[121] Martin, Leenus, Grigoryan, Arsen, Wang, Ding, Wang, Jinhua, Breda, Laura, Rivella, Stefano, Cardozo, Timothy, and Gardner, Lawrence B. Identification and Characterization of Small Molecules That Inhibit Nonsense-Mediated RNA Decay and Suppress Nonsense p53 Mutations. *Cancer research*, 74(11):3104–3113, 2014.

[122] Mateescu, Bogdan, Batista, Luciana, Cardon, Melissa, Gruosso, Tina, de Feraudy, Yvan, Mariani, Odette, Nicolas, André, Meyniel, Jean-Philippe, Cottu,

Paul, Sastre-Garau, Xavier, and Mechta-Grigoriou, Fatima. miR-141 and miR-200a act on ovarian tumorigenesis by controlling oxidative stress response. *Nature medicine*, 17(12):1627–1635, November 2011.

[123] Isao Matsuura, Natalia G Denissova, Guannan Wang, Dongming He, Jianyin Long, and Fang Liu. Cyclin-dependent kinases regulate the antiproliferative function of Smads. *Nature*, 430(6996):226–231, July 2004.

[124] Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4):329–337.e4, April 2019.

[125] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[126] J A Melero, D T Stitt, W F Mangel, and R B Carroll. Identification of new polypeptide species (48-55K) immunoprecipitable by antiserum to purified large T antigen and present in SV40-infected and -transformed cells. *Virology*, 93(2):466–480, March 1979.

[127] Y Miki, J Swensen, D Shattuck-Eidens, P A Futreal, K Harshman, S Tavtigian, Q Liu, C Cochran, L M Bennett, and W Ding. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science (New York, N.Y.)*, 266(5182):66–71, October 1994.

[128] Q H Miow, T Z Tan, J Ye, J A Lau, T Yokomizo, J-P Thiery, and S Mori. Epithelial-mesenchymal status renders differential responses to cisplatin in ovarian cancer. *Oncogene*, 34(15):1899–1907, April 2015.

[129] Munemasa Mori, John E Mahoney, Maria R Stupnikov, Jesus R Paez-Cortez, Aleksander D Szymaniak, Xaralabos Varelas, Dan B Herrick, James Schwob, Hong Zhang, and Wellington V Cardoso. Notch3-Jagged signalling controls the pool of undifferentiated airway progenitors. *Development (Cambridge, England)*, 142(2):258–267, January 2015.

[130] Mort, Matthew, Ivanov, Dobril, Cooper, David N, and Chuzhanova, Nadia A. A meta-analysis of nonsense mutations causing human genetic disease. *Human Mutation*, 29(8):1037–1047, August 2008.

[131] Cyril Neftel, Julie Laffy, Mariella G Filbin, Toshiro Hara, Marni E Shore, Gilbert J Rahme, Alyssa R Richman, Dana Silverbush, McKenzie L Shaw, Christine M Hebert, John Dewitt, Simon Gritsch, Elizabeth M Perez, L Nicolas Gonzalez Castro, Xiaoyang Lan, Nicholas Druck, Christopher Rodman, Danielle Dionne, Alexander Kaplan, Mia S Bertalan, Julia Small, Kristine Pelton, Sarah Becker, Dennis Bonal, Quang-De Nguyen, Rachel L Servis, Jeremy M Fung, Ravindra Mylvaganam, Lisa Mayr, Johannes Gojo, Christine Haberler, Rene Geyeregger, Thomas Czech, Irene Slavc, Brian V Nahed, William T Curry, Bob S Carter, Hiroaki Wakimoto, Priscilla K Brastianos, Tracy T Batchelor, Anat Stemmer-Rachamimov, Maria Martinez-Lage, Matthew P Frosch, Ivan Stamenkovic, Nicolo Riggi, Esther Rheinbay, Michelle Monje, Orit Rozenblatt-Rosen, Daniel P Cahill, Anoop P Patel, Tony Hunter, Inder M Verma, Keith L Ligon, David N Louis, Aviv Regev, Bradley E Bernstein, Itay Tirosh, and Mario L Suvà. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*, 178(4):835–849.e21, August 2019.

[132] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–+, May 2015.

[133] Andrew Ng, Michael Jordan, and Yair Weiss. On Spectral Clustering: Analysis and an algorithm. *NIPS Proceedings*, pages 1–8, April 2002.

[134] Annie Ng and Nick Barker. Ovary and fimbrial stem cells: biology, niche and cancer origins. *Nature reviews. Molecular cell biology*, 16(10):625–638, October 2015.

[135] Serena Nik-Zainal, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, Jonathan Hinton, John

Marshall, Lucy A Stebbings, Andrew Menzies, Sancha Martin, Kenric Leung, Lina Chen, Catherine Leroy, Manasa Ramakrishna, Richard Rance, King Wai Lau, Laura J Mudie, Ignacio Varela, David J McBride, Graham R Bignell, Susanna L Cooke, Adam Shlien, John Gamble, Ian Whitmore, Mark Maddison, Patrick S Tarpey, Helen R Davies, Elli Papaemmanuil, Philip J Stephens, Stuart McLaren, Adam P Butler, Jon W Teague, Göran Jönsson, Judy E Garber, Daniel Silver, Penelope Miron, Aquila Fatima, Sandrine Boyault, Anita Langerød, Andrew Tutt, John W M Martens, Samuel A J R Aparicio, Åke Borg, Anne Vincent Salomon, Gilles Thomas, Anne-Lise Børresen-Dale, Andrea L Richardson, Michael S Neuberger, P Andrew Futreal, Peter J Campbell, Michael R Stratton, and Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, May 2012.

[136] Ciara H O'Flanagan, Kieran R Campbell, Allen W Zhang, Farhia Kabeer, Jamie L P Lim, Justina Biele, Peter Eirew, Daniel Lai, Andrew McPherson, Esther Kong, Cherie Bates, Kelly Borkowski, Matt Wiens, Brittany Hewitson, James Hopkins, Jenifer Pham, Nicholas Ceglia, Richard Moore, Andrew J Mungall, Jessica N McAlpine, CRUK IMAXT Grand Challenge Team, Sohrab P Shah, and Samuel Aparicio. Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses. *Genome biology*, 20(1):210, October 2019.

[137] M A Olayioye, I Beuvink, K Horsch, J M Daly, and N E Hynes. ErbB receptor-induced activation of stat transcription factors is mediated by Src tyrosine kinases. *The Journal of biological chemistry*, 274(24):17209–17218, June 1999.

[138] Chris Ottolenghi, Emanuele Pelosi, Joseph Tran, Maria Colombino, Eric Douglass, Timur Nedorezov, Antonio Cao, Antonino Forabosco, and David Schlessinger. Loss of Wnt4 and Foxl2 leads to female-to-male sex reversal extending to germ cells. *Human molecular genetics*, 16(23):2795–2804, December 2007.

[139] Daniel Y Paik, Deanna M Janzen, Amanda M Schafenacker, Victor S Velasco, May S Shung, Donghui Cheng, Jiaoti Huang, Owen N Witte, and Sanaz

Memarzadeh. Stem-like epithelial cells are concentrated in the distal end of the fallopian tube: a site for injury and serous cancer initiation. *Stem cells (Dayton, Ohio)*, 30(11):2487–2497, November 2012.

[140] Parker, Joel S, Mullins, Michael, Cheang, Maggie C U, Leung, Samuel, Voduc, David, Vickery, Tammi, Davies, Sherri, Fauron, Christiane, He, Xiaping, Hu, Zhiyuan, Quackenbush, John F, Stijleman, Inge J, Palazzo, Juan, Marron, J S, Nobel, Andrew B, Mardis, Elaine, Nielsen, Torsten O, Ellis, Matthew J, Perou, Charles M, and Bernard, Philip S. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(8):1160–1167, March 2009.

[141] B A Parr and A P McMahon. Sexually dimorphic development of the mammalian reproductive tract requires Wnt-7a. *Nature*, 395(6703):707–710, October 1998.

[142] Fernando Pastor, Despina Kolonias, Paloma H Giangrande, and Eli Gilboa. Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. *Nature*, 465(7295):227–230, May 2010.

[143] Patel, Anoop P, Tirosh, Itay, Trombetta, John J, Shalek, Alex K, Gillespie, Shawn M, Wakimoto, Hiroaki, Cahill, Daniel P, Nahed, Brian V, Curry, William T, Martuza, Robert L, Louis, David N, Rozenblatt-Rosen, Orit, Suvà, Mario L, Regev, Aviv, and Bernstein, Bradley E. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.)*, 344(6190):1396–1401, 2014.

[144] Gregory J Pazour, Nathan Agrin, John Leszyk, and George B Witman. Proteomic analysis of a eukaryotic cilium. *The Journal of cell biology*, 170(1):103–113, July 2005.

[145] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[146] Ruth Perets, Gregory A Wyant, Katherine W Muto, Jonathan G Bijron, Barish B Poole, Kenneth T Chin, Jin Yun H Chen, Anders W Ohman, Corey D Stepule, Soongu Kwak, Alison M Karst, Michelle S Hirsch, Sunita R Setlur, Christopher P Crum, Daniela M Dinulescu, and Ronny Drapkin. Transformation of the fallopian tube secretory epithelium leads to high-grade serous ovarian cancer in Brca;Tp53;Pten models. *Cancer Cell*, 24(6):751–765, December 2013.

[147] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-Seq2. *Nature Protocols*, 9(1):171–181, January 2014.

[148] Picelli, Simone, Björklund, Åsa K, Faridani, Omid R, Sagasser, Sven, Winberg, Gösta, and Sandberg, Rickard. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10(11):1096–1098, November 2013.

[149] J M Piek, P J van Diest, R P Zweemer, J W Jansen, R J Poort-Keesom, F H Menko, J J Gille, A P Jongsma, G Pals, P Kenemans, and R H Verheijen. Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. *Journal of Pathology*, 195(4):451–456, November 2001.

[150] Emma Pierson and Christopher Yau. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology*, 16:241, November 2015.

[151] Pils, Dietmar, Hager, Gudrun, Tong, Dan, Aust, Stefanie, Heinze, Georg, Kohl, Maria, Schuster, Eva, Wolf, Andrea, Sehouli, Jalid, Braicu, Ioana, Vergote, Ignace, Cadron, Isabelle, Mahner, Sven, Hofstetter, Gerda, Speiser, Paul, and Zeillinger, Robert. Validating the impact of a molecular subtype in ovarian cancer on outcomes: a study of the OVCAD Consortium. *Cancer science*, 103(7):1334–1341, July 2012.

[152] Popp, Maximilian W and Maquat, Lynne E. Attenuation of nonsense-

mediated mRNA decay facilitates the response to chemotherapeutics. *Nature communications*, 6(1):6632, 2015.

[153] Popp, Maximilian W and Maquat, Lynne E. Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. *Cell*, 165(6):1319–1322, June 2016.

[154] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, Daniel G Deschler, Mark A Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W Rocco, William C Faquin, Derrick T Lin, Aviv Regev, and Bradley E Bernstein. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*, 171(7):1611.e1–1611.e24, December 2017.

[155] William M Rand. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, December 1971.

[156] Timothy R Rebbeck, Henry T Lynch, Susan L Neuhausen, Steven A Narod, Laura Van't Veer, Judy E Garber, Gareth Evans, Claudine Isaacs, Mary B Daly, Ellen Matloff, Olufunmilayo I Olopade, Barbara L Weber, and Prevention and Observation of Surgical End Points Study Group. Prophylactic oophorectomy in carriers of BRCA1 or BRCA2 mutations. *The New England journal of medicine*, 346(21):1616–1622, May 2002.

[157] American Association for Cancer Research. Single-Cell RNA-Seq Reveals Melanoma Transcriptional Heterogeneity. *Cancer discovery*, 6(6):570–570, June 2016.

[158] Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):–e47, 2015.

[159] Manuel A Rivas, Matti Pirinen, Donald F Conrad, Monkol Lek, Emily K Tsang, Konrad J Karczewski, Julian B Maller, Kimberly R Kukurba, David S

DeLuca, Menachem Fromer, Pedro G Ferreira, Kevin S Smith, Rui Zhang, Fengmei Zhao, Eric Banks, Ryan Poplin, Douglas M Ruderfer, Shaun M Purcell, Taru Tukiainen, Eric V Minikel, Peter D Stenson, David N Cooper, Katharine H Huang, Timothy J Sullivan, Jared Nedzel, GTEx Consortium, Geuvadis Consortium, Carlos D Bustamante, Jin Billy Li, Mark J Daly, Roderic Guigo, Peter Donnelly, Kristin Ardlie, Michael Sammeth, Emmanouil T Dermitzakis, Mark I McCarthy, Stephen B Montgomery, Tuuli Lappalainen, and Daniel G MacArthur. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348(6235):666–669, May 2015.

[160] Rizvi, Abbas H, Camara, Pablo G, Kandror, Elena K, Roberts, Thomas J, Schieren, Ira, Maniatis, Tom, and Rabadan, Raul. Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35(6):551–560, June 2017.

[161] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):R25, 2010.

[162] Robinson, M D, McCarthy, D J, and Smyth, G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, December 2009.

[163] Rosenberg, Alexander B, Roco, Charles M, Muscat, Richard A, Kuchina, Anna, Sample, Paul, Yao, Zizhen, Graybuck, Lucas T, Peeler, David J, Mukherjee, Sumit, Chen, Wei, Pun, Suzie H, Sellers, Drew L, Tasic, Bosiljka, and Seelig, Georg. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (New York, N.Y.)*, 360(6385):176–182, April 2018.

[164] Angshumoy Roy and Martin M Matzuk. Reproductive tract function and dysfunction in women. *Nature reviews. Endocrinology*, 7(9):517–525, May 2011.

[165] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring

of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–470, October 1995.

[166] Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R journal*, 8(1):289–317, August 2016.

[167] Åsa Segerstolpe, Athanasia Palasantza, Pernilla Eliasson, Eva-Marie Andersson, Anne-Christine Andreasson, Xiaoyan Sun, Simone Picelli, Alan Sabirsh, Maryam Clausen, Magnus K Bjursell, David M Smith, Maria Kasper, Carina Ammala, and Rickard Sandberg. Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, 24(4):593–607, 2016.

[168] Ann Smith Sehdev, Robert J Kurman, Elisabetta Kuhn, and Ie-Ming Shih. Serous tubal intraepithelial carcinoma upregulates markers associated with high-grade serous carcinomas including Rsf-1 (HBXAP), cyclin E and fatty acid synthase. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 23(6):844–855, June 2010.

[169] Morten Seirup, Li-Fang Chu, Srikumar Sengupta, Ning Leng, Christina M Shafer, Bret Duffin, Angela L Elwell, Jennifer M Bolin, Scott Swanson, Ron Stewart, Christina Kendziorski, James A Thomson, and Rhonda Bacher. Tradeoff between more cells and higher read depth for single-cell RNA-seq spatial ordering analysis of the liver lobule. *bioRxiv*, pages 1–32, September 2019.

[170] Tsukasa Shibue and Robert A Weinberg. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature reviews. Clinical oncology*, 14(10):611–629, October 2017.

[171] A E Smith, R Smith, and E Paucha. Characterization of different tumor antigens present in cells transformed by simian virus 40. *Cell*, 18(2):335–346, October 1979.

[172] Natasha T Snider, Peter J Altshuler, and M Bishr Omary. Modulation of cytoskeletal dynamics by mammalian nucleoside diphosphate kinase (NDPK)

proteins. *Naunyn-Schmiedeberg's archives of pharmacology*, 388(2):189–197, February 2015.

[173] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, 14(1):91, March 2013.

[174] Charlotte Soneson and Mark D Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 18(4):735–261, February 2018.

[175] D Stehelin, H E Varmus, J M Bishop, and P K Vogt. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173, March 1976.

[176] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[177] Patrick Sung and Hannah Klein. Mechanism of homologous recombination: mediators and helicases take on regulatory functions. *Nature reviews. Molecular cell biology*, 7(10):739–750, October 2006.

[178] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. *bioRxiv*, pages 3486–9, September 2019.

[179] John G Tate, Sally Bamford, Harry C Jubb, Zbyslaw Sondka, David M Beare, Nidhi Bindal, Harry Boutselakis, Charlotte G Cole, Celestino Creatore, Elisabeth Dawson, Peter Fish, Bhavana Harsha, Charlie Hathaway, Steve C Jupe, Chai Yin Kok, Kate Noble, Laura Ponting, Christopher C Ramshaw, Claire E Rye, Helen E Speedy, Ray Stefancsik, Sam L Thompson, Shicai Wang, Sari Ward, Peter J Campbell, and Simon A Forbes. COSMIC: the Catalogue Of

Somatic Mutations In Cancer. *Nucleic Acids Research*, 47(D1):D941–D947, 2019.

[180] Thermann, Rolf, Yilik, Gabriele Neu, Deters, Andrea, Frede, Ute, Wehr, Kristina, Hagemeier, Christian, Hentze, Matthias W, and Kulozik, Andreas E. Binary specification of nonsense codons by splicing and cytoplasmic translation. *The EMBO Journal*, 17(12):3484–3494, June 1998.

[181] Jean Paul Thiery. Epithelial-mesenchymal transitions in tumour progression. *Nature Reviews Cancer*, 2(6):442–454, June 2002.

[182] Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, Shalin H Naik, and Matthew E Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature methods*, 16(6):479–487, June 2019.

[183] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen, Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (New York, N.Y.)*, 352(6282):189–196, April 2016.

[184] Tothill, Richard W, Tinker, Anna V, George, Joshy, Brown, Robert, Fox, Stephen B, Lade, Stephen, Johnson, Daryl S, Trivett, Melanie K, Etemadmoghadam, Dariush, Locandro, Bianca, Traficante, Nadia, Fereday, Sian, Hung, Jillian A, Chiew, Yoke-Eng, Haviv, Izhak, Australian Ovarian Cancer Study Group, Gertig, Dorota, DeFazio, Anna, and Bowtell, David D L. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to

clinical outcome. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 14(16):5198–5208, August 2008.

[185] Britton Trabert, Tim Waterboer, Annika Idahl, Nicole Brenner, Louise A Brinton, Julia Butt, Sally B Coburn, Patricia Hartge, Katrin Hufnagel, Federica Inturrisi, Jolanta Lissowska, Alexander Mentzer, Beata Peplonska, Mark E Sherman, Gillian S Wills, Sarah C Woodhall, Michael Pawlita, and Nicolas Wentzensen. Antibodies Against Chlamydia trachomatis and Ovarian Cancer Risk in Two Independent Populations. *Journal of the National Cancer Institute*, 111(2):129–136, February 2019.

[186] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology*, 21(1):1–32, January 2020.

[187] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, April 2014.

[188] Tusi, Betsabeh Khoramian, Wolock, Samuel L, Weinreb, Caleb, Hwang, Yung, Hidalgo, Daniel, Zilionis, Rapolas, Waisman, Ari, Huh, Jun R, Klein, Allon M, and Socolovsky, Merav. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*, 555(7694):54–60, March 2018.

[189] B K Tye. MCM proteins in DNA replication. *Annual review of biochemistry*, 68(1):649–686, 1999.

[190] Ege Ulgen, Ozan Ozisik, and Osman Ugur Sezerman. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. *bioRxiv*, page 272450, January 2018.

[191] USUKI, F, YAMASHITA, A, KASHIMA, I, HIGUCHI, I, OSAME, M, and OHNO, S. Specific inhibition of nonsense-mediated mRNA decay components,

SMG-1 or Upf1, rescues the phenotype of ullrich disease fibroblasts. *Molecular Therapy*, 14(3):351–360, September 2006.

[192] Russell Vang, Ie-Ming Shih, and Robert J Kurman. Ovarian low-grade and high-grade serous carcinoma: pathogenesis, clinicopathologic and molecular biologic features, and diagnostic problems. *Advances in anatomic pathology*, 16(5):267–282, September 2009.

[193] Vento-Tormo, Roser, Efremova, Mirjana, Botting, Rachel A, Turco, Margherita Y, Vento-Termo, Miquel, Meyer, Kerstin B, Park, Jong-Eun, Stephenson, Emily, Polanski, Krzysztof, Goncalves, Angela, Gardner, Lucy, Holmqvist, Staffan, Henriksson, Johan, Zou, Angela, Sharkey, Andrew M, Millar, Ben, Innes, Barbara, Wood, Laura, Wilbrey-Clark, Anna, Payne, Rebecca P, Ivarsson, Martin A, Lisgo, Steve, Filby, Andrew, Rowitch, David H, Bulmer, Judith N, Wright, Gavin J, Stubbington, Michael J T, Haniffa, Muzlifah, Moffett, Ashley, and Teichmann, Sarah A. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature*, 563(7731):347–+, 2018.

[194] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *the 26th Annual International Conference*, pages 1073–1080, New York, New York, USA, June 2009. ACM.

[195] John Vivian, Arjun Arkal Rao, Frank Austin Nothaft, Christopher Ketchum, Joel Armstrong, Adam Novak, Jacob Pfeil, Jake Narkizian, Alden D Deran, Audrey Musselman-Brown, Hannes Schmidt, Peter Amstutz, Brian Craft, Mary Goldman, Kate Rosenbloom, Melissa Cline, Brian O'Connor, Megan Hanna, Chet Birger, W James Kent, David A Patterson, Anthony D Joseph, Jingchun Zhu, Sasha Zaranek, Gad Getz, David Haussler, and Benedict Paten. Toil enables reproducible, open source, big biomedical data analyses. *Nature biotechnology*, 35(4):314–316, April 2017.

[196] Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, August 2007.

[197] Wagner, Eileen and Lykke-Andersen, Jens. mRNA surveillance: the perfect persist. *Journal of cell science*, 115(Pt 15):3033–3038, August 2002.

[198] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications*, 10(1):380, January 2019.

[199] Wooster, R, Bignell, G, Lancaster, J, Swift, S, Seal, S, Mangion, J, Collins, N, Gregory, S, Gumbs, C, and Micklem, G. Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559):789–792, December 1995.

[200] Jonathan E Wosen, Dhriti Mukhopadhyay, Claudia Macaubas, and Elizabeth D Mellins. Epithelial MHC Class II Expression and Its Role in Antigen Presentation in the Gastrointestinal and Respiratory Tracts. *Frontiers in immunology*, 9:2144, 2018.

[201] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015.

[202] Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez, Rahulsimham Vegesna, Hoon Kim, Wandaliz Torres-Garcia, Victor Treviño, Hui Shen, Peter W Laird, Douglas A Levine, Scott L Carter, Gad Getz, Katherine Stemke-Hale, Gordon B Mills, and Roel G W Verhaak. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications*, 4:2612, 2013.

[203] Yoshihara, Kosuke, Tsunoda, Tatsuhiko, Shigemizu, Daichi, Fujiwara, Hiroyuki, Hatae, Masayuki, Fujiwara, Hisaya, Masuzaki, Hideaki, Katabuchi, Hidetaka, Kawakami, Yosuke, Okamoto, Aikou, Nogawa, Takayoshi, Matsumura, Noriomi, Udagawa, Yasuhiro, Saito, Tsuyoshi, Itamochi, Hiroaki, Takano, Masashi, Miyagi, Etsuko, Sudo, Tamotsu, Ushijima, Kimio, Iwase, Haruko, Seki, Hiroyuki, Terao, Yasuhisa, Enomoto, Takayuki, Mikami, Mikio, Akazawa, Kohei, Tsuda, Hitoshi, Moriya, Takuya, Tajima, Atsushi, Inoue, Ituro, Tanaka, Kenichi, and Japanese Serous Ovarian Cancer Study Group. High-risk ovarian cancer based on 126-gene expression signature is uniquely

characterized by downregulation of antigen presentation pathway. *Clinical Cancer Research*, 18(5):1374–1385, March 2012.

[204] Yingjian You, Tao Huang, Edward J Richer, Jens-Erik Harboe Schmidt, Joseph Zabner, Zea Borok, and Steven L Brody. Role of f-box factor foxj1 in differentiation of ciliated airway epithelial cells. *American journal of physiology. Lung cellular and molecular physiology*, 286(4):L650–7, April 2004.

[205] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1), 2017.

[206] J Zhang and L E Maquat. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. *The EMBO Journal*, 16(4):826–833, February 1997.

[207] Qing Zhang, Chen Wang, and William A Cliby. Cancer-associated stroma significantly contributes to the mesenchymal subtype signature of serous ovarian cancer. *Gynecologic oncology*, 152(2):368–374, February 2019.

[208] Zhang, J, Sun, X, Qian, Y, LaDuca, J P, and Maquat, L E. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Molecular and cellular biology*, 18(9):5272–5283, September 1998.

[209] Zhang, J, Sun, X, Qian, Y, and Maquat, L E. Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *Rna*, 4(7):801–815, July 1998.

[210] Ming Zhou, Huaying Liu, Xiaojie Xu, Houde Zhou, Xiaoling Li, Cong Peng, Shourong Shen, Wei Xiong, Jian Ma, Zhaoyang Zeng, Songqing Fang, Xinmin Nie, Yixin Yang, Jie Zhou, Juanjuan Xiang, Li Cao, Shuping Peng, Shufang Li, and Guiyuan Li. Identification of nuclear localization signal that governs nuclear import of BRD7 and its essential roles in inhibiting cell cycle progression. *Journal of cellular biochemistry*, 98(4):920–930, July 2006.

[211] X D Zhu, B Küster, M Mann, J H Petrini, and T de Lange. Cell-cycle-regulated association of RAD50/MRE11/NBS1 with TRF2 and human telomeres. *Nature genetics*, 25(3):347–352, July 2000.

[212] A B Zimrin, M S Pepper, G A McMahon, F Nguyen, R Montesano, and T Maciag. An antisense oligonucleotide to the notch ligand jagged enhances fibroblast growth factor-induced angiogenesis in vitro. *The Journal of biological chemistry*, 271(51):32499–32502, December 1996.