This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
The use of scRNA-seq to characterise the tumour microenvironment of High Grade Serous Ovarian Carcinomas (HGSOC)

Thomas William Parry

Submitted for the degree of Doctor of Philosophy

The University of Edinburgh

2022
Declaration:

I declare the this thesis has been composed entirely by myself. No section of this thesis has been used in any previous application for degree or qualification. Except where contributions and references are clearly indicated, the work herein is my own.

Thomas W. Parry (2022)
Final word count: 47,659 words excluding abstract, lay summary, acknowledgements, list of abbreviations, table of contents, list of figures, list of tables, figure titles, figure captions and bibliography.
Abstract:

High Grade Serous Ovarian Carcinoma (HGSOC) is the most common type of ovarian cancer. Patients with this disease typically experience relapse in their disease following surgical debulking and initially effective chemotherapy. HGSOC has been intensely studied at the genomic and transcriptomic levels in efforts to advance knowledge of the biological mechanisms that drive the behaviour of this malignancy, and so that new treatment strategies may curb the disease progression relapse.

This body of work contributes an optimised protocol for generating robust 10X scRNA-seq libraries from fresh and preserved HGSOC tissue, aiming to dissect the cellular heterogeneity of HGSOC’s Tumour microenvironment (TME). Through unsupervised clustering analysis, it uncovers distinct cellular communities, elucidates transcriptomic signatures across HGSOC tumours, and augments bulk RNA-seq datasets via computational deconvolution, enhancing understanding of HGSOC’s cellular complexity across an expanded clinical cohort.

The sequencing and analysis of these HGSOC patient tumours revealed 11 distinct cell types, including 2 that are novel in this tumour type; namely ciliated epithelial cells and metallothionein expressing T-cells. These 11 distinct cell types can be broadly categorised into 3 TME components (Tumour, Stroma and Immune) as in other previous tumour scRNA-seq studies. An additional analysis of these components examined the copy number variation (CNV) in the profiled cells and revealed HGSOC tumour cells to be mostly aneuploid while ciliated epithelial cells were diploid. A novel integrative subcluster analysis of HGSOC aneuploid tumour cells identified several apparently tumourigenic gene expression signatures. These include a KRT17+, protease inhibitory signature, an increased cellular metabolism signature, and an immune-reactive signature. Additionally, a ciliated cluster re-emerged within the HGSOC tumour cells, even though the diploid ciliated epithelial cells were not included in the integrative analysis.
Finally, the high granularity of HGSOC cellular composition revealed by scRNA-seq is utilised to perform deconvolution analyses to estimate cellular proportions and infer the TME of earlier bulk RNA-seq profiled HGSOC tumour samples. This investigation of earlier sequenced HGSOC samples revealed heterogeneity in the proportions of the TME compartments across the patient cohorts. Survival analysis using these inferred cellular proportions suggest that immune cell presence alone is not associated with survival, but metastatic fibroblast burden in tumour samples is significantly associated with worsen overall survival in HGSOC patients.

In conclusion, the laboratory protocol, the scRNA-seq datasets produced, and their analysis and application presented in this work expands the collective knowledge base of HGSOC. Specifically by characterising the cells of the HGSOC tumour microenvironment, and nuances of expression signatures of the malignant cells. The deconvolution approach showcases how scRNA-seq data can expand the clinical utility of earlier RNA-seq HGSOC datasets in a way that is scalable.
Lay Summary:

Ovarian cancer (OC) is the most lethal cancer that affects the ovaries. The most common form of ovarian cancer is called High Grade Serous Ovarian Carcinoma (HGSOC). Over 70% of ovarian cancers are of this variety. There is a severe unmet need to understand how patients with HGSOC often initially respond well to chemotherapy, but, after receiving treatment, frequently experience relapse of their condition with an often lethal chemotherapy-resistant variant.

The nucleus of the cell harbours the DNA, which in principle is the same in all nucleated cells. RNA are the chemical messengers send from the cell’s nucleus which tells the cell which genes to make into proteins. What differentiates cells are the proteins they produce. Therefore, the RNA, by reflecting the quantities and types of proteins provide a readout of the characteristics of that specific cell.

In order to understand HGSOC, a new method of sequencing RNA in individual cells, called single-cell RNA-sequencing (sc-RNAseq), was deployed to test the tumours of patients with HGSOC.

HGSOC tumours were collected from patients undergoing surgery. A method was created to harvest healthy cells from the solid tumours, enabling the sequencing of the RNA from these cells. After sequencing, the cells of the tumours were grouped and labelled based on the RNA molecules present inside them at the time of the patient’s surgery.

This work uses this single-cell RNA sequencing data of HGSOC tumours in two main ways to expand our understanding of the disease.

First, we use the knowledge about which RNA molecules are present in each cell in order to identify novel sub-types of cells that exist in the HGSOC tumour. We then determine which cells are cancerous, and attribute biological activities to them.
Understanding their biological activities helps us understand how this disease responds to the host environment, develops and acquires resistance to chemotherapy.

Second, we apply the information from the first work to mathematically determine the proportional make up of various cell types within tumours that have already had the RNA molecules assessed (although not on a cell by cell basis). This allows for the re-evaluation of HGSOC clinical datasets and leads to a novel finding which suggests patients whose HGSOC tumours have a high proportion of metastatic fibroblasts experience a lower survival rate.
Acknowledgments:

I would firstly like to thank my supervisors, Charlie Gourley and Colin Semple. Their support of me throughout the journey of my Ph.D., COVID19 and up to the submission of this work means more than words can convey. They believed in me in the moments I lost faith in myself, never did their patience with me wain. I feel truly blessed to be under their supervision for my Ph.D..

I wish to thank the members of Charlie Gourley’s lab, Robb Hollis, John Thompson and Mike Churchman for being exemplary role models and sincere friends throughout my time in Edinburgh. To this day, it feels surreal to be in the company of such impossibly intelligent, caring and funny people.

To my lab and project partner, Jess – you made the time working in this Ph.D. feel like no time at all. When things were difficult, you responded with strength and resilience. I feel privileged to have learned so much beyond what I could’ve learned in a Ph.D. from you. Thank you.

Dan, you’ve been an immensely good friend, funny and understanding. I fear to think how my time during my Ph.D. would’ve been had it not been for you. Thanks so much.

Gee and Ath, I have much to thank you both for. Your friendships made the times during my Ph.D. a time of laughter, joy and adventure. I am blessed for having met both of you.

Edwin, you are like an older brother to me. You helped me make the most of my time in Edinburgh and bringing me out my shell during this Ph.D.. Thank you.

James Doobs, Yalla Habibi, If I had to write every reason I have to thank you, the text would surpass the word count of this thesis. You’ve been with me long since this journey began, and I can rest assured you’ll be with me in the future. To you, I am forever grateful.
The Bean, your unshaking belief in me, encouragement, support and patience mean so much to me. I can never truly repay you. This work would never have been possible without the copious amount of guidance you’ve lent me over the years we’ve known each other. I hope to forever remain in your graces.

I would also like to thank the courageous and selfless ovarian cancer patients of this study, and other cancer patients who participate in studies like this. Their contribution is at the core of what makes this research possible, and imbue it with sense of what makes this work worth doing.
### List of Abbreviations:

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATAC-seq:</td>
<td>Assay for Transposase-Accessible Chromatin Sequencing</td>
</tr>
<tr>
<td>AUROC:</td>
<td>Area Under Receiver Operator Curve</td>
</tr>
<tr>
<td>CAF:</td>
<td>Cancer Associated Fibroblast</td>
</tr>
<tr>
<td>cDNA:</td>
<td>complementary Deoxyribose Nucleic Acid</td>
</tr>
<tr>
<td>CFTR:</td>
<td>Cystic Fibrosis Transmembrane conductance Regulator</td>
</tr>
<tr>
<td>CI:</td>
<td>Confidence Interval</td>
</tr>
<tr>
<td>CITE-seq:</td>
<td>Cellular Indexing of Transcriptomes and Epitopes Sequencing</td>
</tr>
<tr>
<td>CNV:</td>
<td>Copy Number Variation</td>
</tr>
<tr>
<td>CPM:</td>
<td>Counts per Million</td>
</tr>
<tr>
<td>DAPI:</td>
<td>4’,6-Diamidino-2-Phenylindole</td>
</tr>
<tr>
<td>DCR:</td>
<td>Dead Cell Removal</td>
</tr>
<tr>
<td>DMEM:</td>
<td>Dulbecco's Modified Eagle Medium</td>
</tr>
<tr>
<td>DMSO:</td>
<td>Dimethyl Sulfoxide</td>
</tr>
<tr>
<td>DNA:</td>
<td>Deoxyribose Nucleic Acid</td>
</tr>
<tr>
<td>ECFC:</td>
<td>Endothelial Colony Forming Cells</td>
</tr>
<tr>
<td>EMT:</td>
<td>Epithelial Mesenchymal Transition</td>
</tr>
<tr>
<td>FACS:</td>
<td>Fluorescence-activated Cell Sorting</td>
</tr>
</tbody>
</table>
FCS: Fetal Calf Serum
FDR: False Discovery Rate
FIGO: International Federation of Gynecology and Obstetrics
GEM: Gel Bead-In Emulsion
GEP: Gene Expression Profile
GLM: General Linear Module
GO: Gene Ontology
GSEA: Gene Set Enrichment Analysis
HGSOC: High Grade Serous Ovarian Carcinoma
HR: Hazard Ratio
HRD: Homologous Repair Deficient
HVG: Highly Variable Gene
IHC: Immuno-Histochemistry
KEGG: Kyoto Encyclopaedia of Genes and Genomes
KNN: K-Nearest Neighbours
LoF: Loss of Function
LGSOC: Low Grade Serous Ovarian Cancer
MAD: Median Absolute Deviation
MHTC: Multiple Hypothesis Test Correction
MRNA: Messenger Ribonucleic Acid

MSigDb: Molecular Signatures Database

MVW: Maximal Variance Weight

NEJM: New England Journal of Medicine

NNLS: Nonnegative Least Squares

OC: Ovarian Cancer

OS: Overall Survival

PARP: Poly ADP-Ribose Polymerase

PBMCS: Peripheral Blood Mononuclear Cells

PBS: Phosphate-Buffered Saline

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

PFS: Progression Free Survival

RBC: Red Blood Cell

RNA-seq: Ribonucleic Acid Sequencing

RNA: Ribonucleic Acid

RPM: Rotations per Minute

RSMD: Root Square Mean Difference

RT-PCR: Real Time - Polymerase Chain Reaction
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCE</td>
<td>Single Cell Experiment</td>
</tr>
<tr>
<td>scRNA-seq</td>
<td>Single Cell Ribonucleic Acid Sequencing</td>
</tr>
<tr>
<td>SNN</td>
<td>Shared Nearest Neighbours</td>
</tr>
<tr>
<td>SVR</td>
<td>Support Vector Regression</td>
</tr>
<tr>
<td>TAM</td>
<td>Tumour Associated Macrophage</td>
</tr>
<tr>
<td>TCI</td>
<td>Tumour Cell Isolation</td>
</tr>
<tr>
<td>TIL</td>
<td>Tumour Infiltrating Leukocyte</td>
</tr>
<tr>
<td>TME</td>
<td>Tumour Microenvironment</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer Ribonucleic Acid</td>
</tr>
<tr>
<td>tSNE</td>
<td>T-distributed Stochastic Neighbour Embedding</td>
</tr>
<tr>
<td>UMAP</td>
<td>Uniform Manifold Approximation and Projection</td>
</tr>
<tr>
<td>UMI</td>
<td>Unique Molecular Identifier</td>
</tr>
<tr>
<td>WT</td>
<td>Wild Type</td>
</tr>
</tbody>
</table>
1 Table of Contents

Declaration: ............................................................................................................. ii
Abstract: ................................................................................................................ iv
Lay Summary: .......................................................................................................... vi
Acknowledgments: ............................................................................................... viii
List of Abbreviations: ........................................................................................... x
List of Figures: ......................................................................................................... 20
List of Tables: ......................................................................................................... 1

1 Chapter 1: Introduction .................................................................................... 3

1.1 Fundamentals of Genomics ............................................................................. 3
  1.1.1 What genes are made of ............................................................................. 3
  1.1.2 The flow of genetic information, the central dogma ............................... 4
  1.1.3 RNA molecules .......................................................................................... 4
  1.1.4 Gene expression and protein synthesis ..................................................... 6

1.2 Genomics Variation and its implications ...................................................... 12
  1.2.1 Exceptions to Genomic uniformity in Somatic Cells .............................. 12
  1.2.2 Types and causes of DNA mutations ....................................................... 13
  1.2.3 Chromosomal instability ......................................................................... 14
  1.2.4 Epigenetic regulation of gene expression ................................................. 15

1.3 Measuring gene expression using RNA sequencing technology 16
  1.3.1 Microarrays: ............................................................................................. 17
  1.3.2 Quantitative Real-Time PCR (qRT-PCR) ............................................... 18
  1.3.3 Next Generation Sequencing: ................................................................. 21
  1.3.4 RNA-seq .................................................................................................. 22

1.4 Single cell RNA-seq ...................................................................................... 24
  1.4.1 ScRNA-seq approaches: ......................................................................... 25
1.5  Cell type identification: Microscopy to scRNA-seq ..........................28
  1.5.1  Microscopy .................................................................................28
  1.5.2  Immunohistochemistry .................................................................29
  1.5.3  Flow cytometry ............................................................................29
  1.5.4  Bulk RNA-seq ..............................................................................30
  1.5.5  scRNA-seq ..................................................................................30
  1.5.6  Practical limitations of scRNA-seq on cell-type classification .........31

1.6  The Tumour Microenvironment .........................................................32
  1.6.1  Introduction to cancer ....................................................................32
  1.6.2  Composition and variability of the tumour microenvironment .......32
  1.6.3  Immune surveillance, hot and cold TMEs ...................................33
  1.6.4  scRNA-seq view of the TME .........................................................34

1.7  High Grade Serous Ovarian Carcinoma (HGSOC) .........................35
  1.7.1  Pathogenesis of HGSOC ...............................................................35
  1.7.2  The genomic and mutational landscape of HGSOC ...................37
  1.7.3  Transcriptomic subtypes of HGSOC .........................................40

1.8  Thesis objective and aims .................................................................41

2  Chapter 2: The dissociation of physical HGSOC tumour specimens
   for use in 10x single cell RNA-seq studies .........................................43

2.1  Introduction .....................................................................................43
  2.1.1  Single cell RNA-sequencing .........................................................43
  2.1.2  Previous studies with dissociation of ovarian / endometrial tissue
         methodologies from Worthington’s Tissue Dissociation Manual ........43
  2.1.3  Dissociation workflows from previous HGSOC scRNA-seq studies 45

2.2  Aims ...............................................................................................46

2.3  Material and methods .....................................................................47
  2.3.1  Procurement of clinical HGSOC tumour specimens ....................47
  2.3.2  Cryopreservation of HGSOC tumour specimens .........................48
  2.3.3  Rescuing cryopreserved HGSOC tumour specimens ....................48
2.3.4 Manual dissociation of HGSOC tumour specimens ...........................................48
2.3.5 Automated dissociation of HGSOC tumours using a Miltenyi GentleMACS Octodissociator .................................................................................................49
2.3.6 Cell viability assessment ...................................................................................50
2.3.7 Dead cell and red blood cell removal ..............................................................50
2.3.8 Isolating PBMCS from patient whole blood samples ......................................50
2.3.9 Fluorescent activated cell sorting ....................................................................51
2.3.10 10X Chromium 3' Single cell RNA-sequencing .............................................51
2.3.11 DNA quality control and quantification .........................................................52
2.3.12 Library Sequencing .......................................................................................52
2.3.13 Alignment from raw base calls to filtered barcode matrices .......................52
2.3.14 Statistical Analysis .......................................................................................53

2.4 Results ..............................................................................................................53
2.4.1 Optimisation of the manual method of dissociation ........................................53
2.4.2 Determining the preferred enzymes in terms of recovering viable cell fractions from dissociated HGSOC tumour specimens ........................................54
2.4.3 Evaluating the consistency of dissociation performance of Liberase and Neutral Protease ........................................................................................................57
2.4.4 Evaluating post-dissociation viability enrichment techniques on manually dissociated HGSOC tumours ...........................................................................60
2.4.5 Summary results comparing FACS, Miltenyi tumour cell isolation kit and Miltenyi dead cell removal kit. ........................................................................66
2.4.6 Automated dissociation yields cell fractions of higher viability than manual dissociation using Liberase. ........................................................................66
2.4.7 Assessment of 10X single cell RNA-sequenced libraries of HGSOC tumours ..........................................................................................................................69
2.4.8 Comparing the quality of blood and tumour-dissociated derived 10X sc-RNAseq libraries .................................................................................................81
2.4.9 Reproducibility of the 10X single cell RNA-sequencing method using LGSOC samples ........................................................................................................83
3 Chapter 3: Characterising the single cell landscape of HGSOC

3.1 Introduction

3.1.1 The Tumour-Immune microenvironment of HGSOC

3.1.2 Current single cell RNA-seq research on HGSOC

3.2 Material and methods

3.2.1 Sequencing and alignment of HGSOC single cell RNA-seq libraries

3.2.2 Filtering scRNA-seq cells

3.2.3 Cell cycle and percent mitochondria scoring of single cells

3.2.4 SCTransformation, a novel single-cell count normalisation approach

3.2.5 Dimension reduction using PCA

3.2.6 Finding neighbours using shared nearest neighbours

3.2.7 Community detection and modularity optimisation

3.2.8 Visualising higher dimension scRNA-seq data using UMAP

3.2.9 Canonical marker identification to declare cell phenotype

3.2.10 Gene set enrichment analysis

3.2.11 Cell type correlation analysis

3.2.12 Copy number variation prediction in single cell RNA-sequencing data

3.3 Results

3.3.1 Overview of the single cell datasets from pilot sequencing batch 1

3.3.2 Single cell landscape across HGSOC

3.3.3 Differential gene expression across similar cell phenotypes in HGSOC

3.3.4 Cell types transcriptomes across HGSOC tumour sites, condition and patients are highly correlated
3.3.5 Using CNV estimation to distinguish between malignant and non-malignant epithelial cells in the HGSOC datasets ........................................119
3.3.6 Feature differences between aneuploid and diploid epithelial cells 120
3.3.7 Chromosomal CNV heterogeneity across HGSOC patients .......122
3.3.8 CNV subclone differential expression analysis .........................124
3.3.9 Integrative analysis of individual cell types pooled across HGSOC tumour samples ...........................................................................................................126

3.4 Discussion .................................................................................................131

4 Chapter 4: Application of HGSOC single-cell RNA-sequencing signatures onto HGSOC bulk-RNA sequenced data ..................................................136

4.1 Introduction .................................................................................................136
  4.1.1 Traditional techniques to assay complex tissues. .........................136
  4.1.2 RNA sequencing of complex tissues ...............................................137
  4.1.3 Leveraging single-cell RNAseq resolution to decompose RNA-seq HGSOC solid tissue specimens. .................................................................138

4.2 Aims ...........................................................................................................140

4.3 Material and methods ................................................................................140
  4.3.1 Acquisition and processing of C.Gourley HGSOC tumour samples 140
  4.3.2 Single-cell RNA sequencing ..............................................................141
  4.3.3 Single-cell RNA sequencing data processing ...............................141
  4.3.4 Shih 2018 HGSOC single-cell dataset ..........................................144
  4.3.5 Zhang 2019 HGSOC single-cell dataset ........................................146
  4.3.6 Pseudo-bulk RNA sequence data from Single cell RNA-sequencing data. 148
  4.3.7 Bulk RNA-sequencing data processing .......................................148
List of Figures:

Figure 1. 1 Steps involved in gene expression from pre-mature RNA to mature RNA. ....8
Figure 1. 2 Steps involved in translation .................................................................9
Figure 1. 3 Steps involved in post-translation gene expression .................................11
Figure 1. 4 Point mutations and structural chromosomal alterations .......................15
Figure 1. 5 Outline of bulk-RNAseq and scRNA-seq workflow ................................31
Figure 1. 6 Characteristics of cold and hot TMEs .......................................................34
Figure 2.1: Outcomes from 3 consecutive head-to-head trials of 3 different enzymes on a total of 9 cryopreserved primary ovarian specimens ...........................................56
Figure 2.2: Dot plots of cell concentrations and viability recovery of each cryopreserved ovarian specimen used in the Liberase vs Neutral Protease triplicate experiments. .....58
Figure 2.3 Scatter plots of cell counts and viability estimates of 3 cryopreserved HGSOC specimens from 1 patient sample (E05) assayed before and after FACS enrichment of viable cells ...........................................................................................................61
Figure 2.4 Trypan blue viability assay profiles from a Countess II. ...............................63
Figure 2.5 Plots of cells counts and viability estimates of 6 cryopreserved primary HGSOC specimens .................................................................................................................................65
Figure 2.6 Cell counts of 17 HGSOC samples across 3 sequencing batches ............72
Figure 2.7 Viability of input cells boxplot across 3 sequencing batches .....................73
Figure 2.8 Sequencing statistics for 17 sequenced 10X sc-RNAseq libraries across 3 independent sequencing batches, faceted across 7 measurements ......................77
Figure 2.9 Spearman correlation plot of aligned library features to cell viability .............80
Figure 3.1 Schematic of the key computational steps to identify cell-types in scRNA-seq data Beginning from step 1 assumes a normalised cells by gene matrix as outlined starting from method section 3.2.4. The dimension reduction from step 3.2.5 yields the cells by eigenvector matrix, which in these cases are PCAs. Steps 2-3 are outlined in method sections 3.2.6-3.2.7. Step 4 is the section 3.2.8 and lastly step 5 can involve labelling cells using known marker genes for expected cell-types as per section 3.2.9 or finding upregulated genes in novel cell communities .............................................99
Figure 3.2 Canonical marker gene expression of each cell type across patient 3 primary site, sample ID 9..................................................................................................................101
Figure 3.3 Colour labelled UMAPs of the 5 HGSOC scRNA-seq libraries in sequencing batch 1.................................................................................................................................106
Figure 3.4 UMAPs of 4 HGSOC scRNA-seq libraries from sequencing batch 2........109
Figure 3.5 UMAPs of 4 HGSOC scRNA-seq libraries from sequencing batch 3........110
Figure 3.6 Cell type composition (%) across the 8 HGSOC scRNA-seq libraries (batches 2 & 3)........................................................................................................................................111
Figure 3.7 Plots of differentially and non-differentially expressed genes between epithelial cells and ciliated (epithelial) cells ...........................................................................................................113
Figure 3.8 Volcano plots of differentially expressed genes between fibroblasts, mesenchymal cells, T cells and cytotoxic T cells ..................................................................................................114
Figure 3.9 MetaNeighbor AUROC correlation of 1,750 random labelled cells across the HGSOC scRNA-seq datasets. ..................................................................................................................117
Figure 3.10 Z-score Pearson correlation of 411 shared HGVs across 10 cell types uncovered across 8 HGSOC scRNA-seq tumour microenvironments ........................................................................118
Figure 3.11 UMAPs, volcano and violin plots of distinguished aneuploid and diploid epithelial cells ..............................................................................................................................................121
Figure 4.1 Introductory illustration of the concepts for leveraging scRNA-seq information to deconvolve bulk RNA-seq samples ...........................................................................................................139
Figure 4.2 Composition of the C.Gourley HGSOC scRNA-seq dataset ..................143
Figure 4.3 Composition of Shih et al. scRNA-seq dataset ........................................146
Figure 4.4 Composition of Zhang et al. scRNA-seq dataset ........................................147
Figure 4.5 Deconvolution evaluation framework ..............................................................156
Figure 4.6 Boxplots of the simulated proportions of 100 C.Gourley pseudo-bulk RNAseq samples .................................................................................................................................159
Figure 4.7 Boxplots of the simulated proportions of 100 Shih et al pseudo-bulk RNAseq samples ........................................................................................................................................160
Figure 4.8 Boxplots of the simulated proportions of 100 Zhang et al pseudo-bulk RNAseq samples .................................................................................................................................161
Figure 4.9 Heatmap of deconvolution estimates of C.Gourley scRNA-seq pseudo-bulk data by 5 scRNA-seq to bulk RNA-seq deconvolution algorithms. ..........................................................164
Figure 4.10 Line plots of the Pearson correlation (+ 95% CI) recorded in 4 deconvolution algorithms for each cell-type in C.Gourley scRNA-seq bulk RNA-seq data. ..................166
Figure 4.11 Heatmap of deconvolution estimates of Shih et al. scRNA-seq pseudo-bulk data by 5 scRNA-seq to bulk RNA-seq deconvolution algorithms .........................168
Figure 4.12 Line plots of the Pearson correlation (+ 95% CI) recorded in 4 deconvolution algorithms for each cell-type in Shih et al. scRNA-seq pseudo-bulk RNA-seq data. ...170
Figure 4.13 Heatmaps and row-wise summaries of the estimated proportions from 3 deconvolution algorithms using pseudo-bulk RNA-seq samples produced from Zhang et al. scRNA-seq data. ........................................................................................................172
Figure 4.14 Line plots and summary tables of the column-wise Pearson correlation from each deconvolution algorithm with 95% confidence intervals for each cell type on the Zhang et al. pseudo-bulk samples and training data..........................................................174
Figure 4.15 Basis matrix (𝔹) of Shih et al. scRNA-seq data constructed by CIBERSORTx ..........................................................................................................................179
Figure 4.16 Simplified\(\dagger\) deconvolution decomposition of 206 deconvolved HGSOC RNA-seq samples and their cellularity* estimates ..................................................180
Figure 4.17 Boxplot of simplified deconvolution estimates of 206 deconvolved HGSOC RNA-seq samples from 4 clinical cohorts........................................................................181
Figure 4.18 Boxplot of deconvolution estimates of 9 cell types found in Shih et al. scRNA-seq data in 206 deconvolved HGSOC RNA-seq samples from 4 clinical cohorts........182
Figure 4.19 Scatter plot of tumour proportion against estimated cellularity of 206 deconvolved HGSOC RNA-seq samples ..........................................................................183
Figure 4.20 Scatter plot of immune proportion against estimated cellularity of 206 deconvolved HGSOC RNA-seq samples ........................................................................184
Figure 4.21 Scatter plot of stromal proportion against estimated cellularity of 206 deconvolved HGSOC RNA-seq samples ..........................................................................185
Figure 4.22 Kaplan Meier curves of HGSOC patient overall survival and progression free survival stratified across HGSOC RNA-seq cohorts. ..................................................188
Figure 4.23 Diagnosis plots of the final OS Cox-proportional hazard models for the HGSOC RNA-seq cohorts.................................................................192
Figure 4.24 Diagnosis plots of the final PFS Cox-proportional hazard models for the HGSOC RNA-seq cohorts.................................................................193
Figure 4.25 Histogram of the distribution of the scaled metastatic fibroblast proportion across 180 HGSOC RNA-seq samples.................................................................194
Figure 4.26 Kaplan Meier curve of overall survival of HGSOC patients stratified by their metastatic fibroblast proportion.................................................................195
Figure 4.27 Volcano plot of Primary and Metastatic Fibroblasts DEGs .....................196
Figure 4.28 GO-plot of Metastatic Fibroblast DEGs....................................................198
Figure S1: Figure S1: Bioanalyzer traces of the cDNA and Indexed libraries of sequencing batch 1. The traces were produced following the reverse transcription step following GEM formation during the 10X step (cDNA) and after the library had been constructed (Indexed Library) ........................................................................................241
Figure S2 Bioanalyzer traces of the cDNA and Indexed libraries of sequencing batch 2. The traces were produced following the reverse transcription step following GEM formation during the 10X step (cDNA) and after the library had been constructed (Indexed Library) ........................................................................................242
Figure S3 Bioanalyzer traces of pooled library batches #2 and #3 prior to submission for sequencing ....................................................................................................243
List of Tables:

Table 2.1 Assessment of average and standard deviation key measurements from the triplicate experiments between enzymes Liberase and Neutral Protease from 2 separate patient samples E04 & E09 .......................................................... 59

Table 2.2 Summarisation of all the cryopreserved dissociation trials involving the use of the Dead Cell Removal (DCR) kit ................................................................. 68

Table 2.3 Summary information of each specimen that went on to be a sequenced 10X sc-RNAseq library ............................................................................................................................ 70

Table 2.4 Batch-wise summary of dissociation statistics of 17 sequenced 10X scRNA-seq libraries ........................................................................................................................ 74

Table 2.5 Sequencing batch-wise summary table of key 10X library preparation statistics of sequenced HGSOC samples* .................................................................................. 75

Table 2.6 Batch-wise average sequencing summary statistics of 17 sequenced 10X scRNA-seq libraries .................................................................................................................. 78

Table 2.8 Batch-wise average summary table of cell statistics of 17 sequenced 10X sc-RNAseq libraries across 3 batches ..................................................................................... 80

Table 2.9 Summary of each sequenced scRNA-seq library fraction of cells, viability and CellRanger number of cells ........................................................................................................ 82

Table 3.1 Canonical marker genes used for identifying cluster cell type and their reference source ................................................................................................................................. 100

Table 4.1 Sample accession numbers and name of the samples under the GSE118828 included in the Shih 2018 dataset for this report ............................................................... 144

Table 4.2 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on C.Gourley single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples) .................................................................................................................. 165

Table 4.3 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on Shih et al. single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples) ........................................................................................................ 169

Table 4.4 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on Zhang et al. single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples) ........................................................................................................ 173

Table 4.5 Clinical characteristics of the deconvolved HGSOC RNA-seq ovarian cancer samples by cohort .................................................................................................................. 187
Table 4.6 Base model of Overall Survival of HGSOC patients in the SHGSOC, AOCS and TCGA cohorts .......................................................... 189
Table 4.7 Base model for Progression Free Survival of HGSOC patients from the AOCS and SHGSOC cohorts ............................................................................................................. 189
Table 4.8 Top 15 genes expressed in the metastatic fibroblast cluster .................. 197

Table S1: Batch 1 Sample preparation [cells to indexed 10X library] quality control metrics. ........................................................................................................................................... 244
Table S2: Batch 2 Sample preparation quality control metrics ......................... 245
Table S3 Batch 3 sample preparation quality control metrics .......................... 246
1 Chapter 1: Introduction

Most of the genetic information of an organism is contained within its genome. The ability to decipher and utilise this information has ushered in biomedical and biotechnological revolutions from the advances made in the 20th century’s biological sciences. This work capitalises on the recent advances in cancer genomics, next generation RNA-sequencing technology and bio-informatic computation to delve into and characterise the tumour microenvironment of High Grade Serous Ovarian Carcinoma (HGSOC) using single-cell RNA sequencing.

This introductory chapter will briefly detail how genetic information is deciphered and how these principles were developed into the present-day field of genomics with an emphasis on the next-generation RNA sequencing technologies, bulk RNA-sequencing (RNA-seq) and single-cell RNA-sequencing (scRNA-seq). This introduction then touches the complexities of cell types and the tumour microenvironment, and how RNA sequencing technology is used to quantify the gene expression of the cells of the tumour and how research into this frontier can benefit cancer patients.

1.1 Fundamentals of Genomics

1.1.1 What genes are made of

Genomics is the study of the genes and genome of an organism. The entire collection of an organism’s genes is referred to as a genome. Genes are made out of deoxyribonucleic acids (DNA), and DNA consists of 4 distinct bases – A (adenine), C (cytosine), G (guanine) and T (thymine). These nucleotides concatenate in a specific sequence which can give rise to another unit, the gene. RNA is made out of 4 distinct bases, 3 of which are the same as in DNA, those being A, C and G. In sequences of RNA, there is no T base, instead it is replaced by U (Uracil).
1.1.2 The flow of genetic information, the central dogma

The central dogma of molecular biology is the explanation of how genetic information flows in biological systems. This information is encoded in genomic DNA, transcribed into closely related molecules called messenger Ribonucleic Acid (mRNA) and finally translated into proteins, the functional units of life\(^1\). Genes are the units of genomic DNA that are vital to this flow, they are the specific sequences of genomic DNA that are ‘read’ by cellular machinery and are made into proteins. In the tapestry of genomic DNA, genes are arranged in a specific order with interspersed sections of non-coding DNA (sections of DNA that are not made into proteins). The composite structure of coding and noncoding DNA is called a chromosome. Collectively, an organism’s complete set of chromosomes constitutes its genome.

1.1.3 RNA molecules

Elucidating the mechanism of information transference from the DNA confined to the nucleus is a cornerstone in molecular biology. Central to this process are the various types of RNA molecules, each serving unique functions and roles in the regulation of gene expression. RNA is an umbrella term that encompasses a diverse set of molecules, recent research has expanded our understanding of RNA’s functionality far beyond the canonical role of delivering genomic information to the protein producing machinery of the cells\(^2\).

1.1.3.1 Messenger RNAs (mRNA)

Messenger RNAs (mRNAs) are transcribed from protein-coding genes and serve as templates for protein synthesis. Once transcribed, mRNAs undergo various post-transcriptional modifications to become fully functional. One of these modifications is 5’ capping, where a modified guanine nucleotide is added to the 5’ end of the mRNA. This cap serves as a unique molecular module that recruits cellular proteins and mediate mRNA translation\(^3\). Another modification is 3’ polyadenylation, which involves the addition of a poly-A tail to the 3’ end of the mRNA. This tail plays a crucial role in mRNA stability.
once it is exported from the nucleus to the cytoplasm. In the study of gene expression, mRNA are used as a proxy for gene expression, primarily because the 5’ cap and 3’ poly-A tail modifications make mRNA more stable and easier to isolate, thus serving as a reliable indicators for gene activity in the cell. it is the level of mRNAs that are used as a proxy for gene expression\textsuperscript{4,5}.

1.1.3.2 \textit{Small nuclear RNA (snRNA)}

Small nuclear RNAs (snRNAs) are primarily involved in the splicing of pre-mature mRNA. They are usually transcribed from genes located in the introns of other genes or in specific clusters. snRNAs form complexes with other proteins to create small nuclear ribonucleoproteins (snRNPs), which are essential components of the spliceosome\textsuperscript{6,7}. The process of splicing involves the removal of intronic regions (introns) from pre-mature RNA to form the mature mRNA molecule. There are permutations in which different introns are excised or remain in the mature mRNA, in a process called alternative splicing\textsuperscript{6}. The phenomena of alternative splicing partially explains how a set number of catalogued genes in a genome can give rise to greater number of functional protein products.

1.1.3.3 \textit{Ribosomal RNA (rRNA)}

Ribosomal RNAs (rRNAs) are transcribed from ribosomal DNA (rDNA). These regions of the genome are organised into tandem repeats, allowing for the simultaneous transcription of multiple rRNA molecules\textsuperscript{8}. The ribosome is composed of both rRNA and ribosomal proteins. rRNA provides the structural and catalytic properties of the ribosome, specifically, rRNA is responsible for the formation of peptide bonds between amino acids, a key step in the translation process in the synthesis of proteins from mRNA templates\textsuperscript{8}.

1.1.3.4 \textit{Transfer RNA (tRNA)}

Transfer RNAs (tRNA) are transcribed from tRNA genes scattered throughout the genome, and are synthesised in the nucleus like other types of RNA. Individual tRNA molecules are uniquely structured to have an ‘anticodon’ region, a sequence of three nucleotides that is complementary for the codons on the mRNA. tRNAs are charged with an affinity for a specific amino acid, during translation the charged tRNAs chauffer amino
acids to the ribosome complex, where the tRNA transfer the amino acid to the growing amino acid chain during protein synthesis.

1.1.3.5 Micro RNA (miRNA)

MicroRNAs (miRNA) are small, non-coding RNA molecules that play a pivotal role in the regulation of gene expression. They are usually transcribed from non-coding regions of the genome and from the introns of protein-coding genes. The primary transcripts, known as primary miRNA (pri-miRNAs), as processed in the nucleus by the Drosha-DGCR8 complex to produce precursor miRNAs (pre-miRNAs). These pre-miRNAs are exported to the cytoplasm, where they undergo further processing by the Dicer enzyme to yield mature miRNAs. These mature miRNAs incorporate into a structure known as the RNA-induced Silencing Complex (RISC), a multi-protein complex that includes Argo2 and TRBP. Within RISC, the miRNA serves as a guide molecule, directing the complex to target mRNA based on sequence complementarity. The degree of complementarity between the miRNA and the target mRNA determines the fate of the mRNA: a perfect or near-perfect match generally leads to mRNA cleavage and degradation, while partial complementarity often results in translational repression. The role of miRNAs in post-transcriptional regulation is crucial for controlling gene expression and are implicated in various biological processes and diseases, including cancer and autoimmune disorders.

1.1.4 Gene expression and protein synthesis

The process by which biological information is transferred from DNA to RNA is called gene expression, while the subsequent conversion of RNA to a functional protein is called protein synthesis. These overarching processes can be further delineated into two key stages: transcription and translation, each of which is subject to intricate regulatory mechanisms at multiple levels.
1.1.4.1 Transcription

The initial step, transcription, involves using the DNA sequence of a gene as a template to synthesis a complementary RNA molecule\textsuperscript{13,14}. These RNA transcripts are single-stranded molecules that undergo several post-transcriptional modifications prior to the exiting the nucleus. First, a methylated guanosine ‘cap’ is added to the 5’ end of the RNA molecule\textsuperscript{3}. Next, a series of stepwise splicing events occur along the RNA molecule, excising introns and ligating exons together to form the matured RNA transcript. Importantly, a given transcribed RNA molecule can be spliced in different ways in a process called ‘alternative splicing’ to produce a variety of mature RNA transcripts from the same gene that are used to manufacture a variety of protein products\textsuperscript{6}. Lastly, a poly-A (a string of adenosine residues) ‘tail’ is added prior to exporting the matured RNA transcript from the nucleus to the cell’s cytoplasm. The length of this poly-A tail helps determine the longevity of the mature RNA transcript in the cytoplasm, which in turn influences the number of proteins produced by the single RNA transcript\textsuperscript{4,14}. 
Figure 1. 1 Steps involved in gene expression from pre-mature RNA to mature RNA, featuring the role of snRNA.

1.1.4.2 Translation

Translation is the next step in gene expression and occurs in 3 distinct steps. These steps are called initiation, elongation, and termination. During initiation, the matured RNA transcript engages with ribosomal subunits outside the nucleus. The ribosomal subunits coalesce around the mature mRNA and form a ribosomal complex. This complex reads the mRNA sequence in triplets of nucleotides called codons, these codons are recognised by specific transfer RNA (tRNA) molecules anticodons. tRNAs are RNA molecules with a specific amino acid bound. During elongation, several tRNA sequentially enter the ribosomal complex, bind to their corresponding codon along the mRNA, transferring their
amino acids to a growing chain of amino acids, then exit the ribosomal complex. Termination occurs when a ‘termination’ codon (UAA, UGA, or UAG) is read by the ribosomal complex, which signals the completion of the amino acid sequence. The ribosomal complex dissociates into its subunits, and the RNA transcript is disengaged, where it may again be read by another ribosome. The resultant amino acid sequence may fold to form the final protein product or undergo several other post-translational modifications prior to achieving its function form.

Figure 1. 2 Steps involved in translation, featuring the role of tRNA and rRNA

1.1.4.3 Post-translation

The fate of mRNA molecules after translation is subject to regulation. Several factors influence the longevity of mature mRNA transcripts outside the nucleus. The length of the
polyA tail, added during post-transcriptional modification is shortened overtime in a process called deadenylation. As the polyA tail is shortened, the mRNA transcript degrades and it no longer useful for producing proteins. Additionally, specific RNA-binding proteins called microRNAs can either stabilise mRNA transcripts, or accelerate their degradation\textsuperscript{10–12}. Some mRNA contain destabilising elements within their sequence which target them for rapid degradation. These regulatory mechanisms ensure that mRNA levels for specific gene products are finely tuned according the cell’s needs, allowing for dynamic changes in protein expression.
Advances in genomic technology have given us the capabilities of extracting and purifying the genetic material from biological material and deciphering the base codes of the genetic material at industrial scale. In turn, this has enabled the discovery, recognition, and cataloguing of genes throughout the genome of an organism. This is important for the...
scientific pursuit to be capable of systematically studying the role of a given gene or a set of genes in biological systems, such as in human health, disease, and development. The gene becomes the fundamental unit that underpins the insights gained from RNA-sequencing, which is the main technology utilised throughout the work presented in this thesis.

1.2 Genomics Variation and its implications

1.2.1 Exceptions to Genomic uniformity in Somatic Cells

1.2.1.1 Functional Deviations in Genomic DNA

Generally, the nucleotide sequence of DNA located in the nucleus of every somatic cell in an organism should be identical. However, there are important deviations from this rule that have functional significance for the expression of genes, and moreover- for human health. Gametes, for instance, contain unique genetic material due to meiotic recombination, a process during which homologous chromosomes exchange genetic material\textsuperscript{18}. This exchange produces a new combination of alleles, thus diversifying the genetic makeup of offspring. In the immune system, T-cells undergo a unique shuffling of gene segments, known as V(D)J recombination, allowing T-cells to produce a highly diverse repertoire of T-cell receptors\textsuperscript{19}. Similarly, B-cells undergo somatic hypermutation- a process that deliberately introduces mutations in variable regions of the antibodies they produce to increase their affinity for antigens\textsuperscript{20}. 

12
1.2.2 Types and causes of DNA mutations

1.2.2.1 Spontaneous Mutations

The internal cellular process that replicates DNA is integral for every cell as it divides into daughter cells. This replication of DNA ensures that sufficient genetic material is present in each subsequent generation of new cells. However, DNA replication is prone to error at a very low, but nonzero rate, leading to what are formally referred to as mutations in the DNA sequence\(^2\). Such mutations give rise to genetic variation. These spontaneous mutations can be neutral, beneficial, or deleterious in their effect on an organism. While some mutations may be repaired by cellular mechanism, others may persist and be passed onto the subsequent generations, contributing to evolutionary changes over time.

1.2.2.2 Induced Mutations

Additionally, DNA is vulnerable to insult from exogenous agents such as radioactive material, viral elements and carcinogenic substances\(^2\), causing mutations to individual nucleotides or entire sections of a chromosome. Prolonged exposure to carcinogens can result in mutations that have profound consequences for human health. Not only can mutations alter the function or expression of individual genes, some genes mutated in particular ways can have far-reaching impacts in highly interdependent cellular processes and pathways, culminating in alterations to an the organism’s phenotype – the set of observable characteristics, and in severe cases, result in various diseases, including cancer\(^2\).

1.2.2.3 Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) represent a prevalent type of genetic variation defined as when >1 of a population does not carry the same nucleotide at a specific position of a DNA sequence\(^3\). The 1000 Genomes Project revealed ~84.7 million SNPs in a population of 2,504 sequenced human genomes\(^4\), and this number is expected to increase as more genomes are sequenced, and rarer SNPS are revealed and verified.
1.2.2.4 Point mutations

While SNPs are relatively frequent, other types of genetic variations contribute towards genetic diversity and can have profound effect on human health and development\textsuperscript{25,26}. Point mutations, for instance, are a class of mutation which can include insertion or deletion, (referred to as Indels), and nucleotide substitution. Either type of point mutation may occur in critical regions of a gene, which may confer substantial functional changes or even loss of gene function. For instance, sickle cell anaemia arises from a single nucleotide substitution in the beta-haemoglobin gene\textsuperscript{27}, while one of the common mutations observed in cystic fibrosis is a deletion of a single nucleotide in the CFTR (Cystic Fibrosis Transmembrane Conductance Regulator) gene\textsuperscript{25}. Both diseases exemplify how specific point mutations can lead to distinct, and often severe health outcomes\textsuperscript{27}.

1.2.3 Chromosomal instability

Chromosomal instability refers to a condition characterised by a high frequency of chromosomal alterations, manifesting in the form of deletions, duplications, inversions, translocations and copy number variations (CNVs). This instability is often a hallmark of various diseases, including cancers like High-Grade Serous Ovarian Carcinoma (HGSOC) where chromosomal instability contributes to the disease’s aggressive nature and complex pathology\textsuperscript{28}. Chromosomal instability is also prevalent in colorectal cancer, where it is influenced by factors such as overexpression of the oncogene AURKA or the loss of the tumour suppressor gene CHK2\textsuperscript{29}.

1.2.3.1 Structural chromosomal mutations

Structural chromosomal mutations include inversions, deletions, duplications and translocations\textsuperscript{30}. Chromosomal inversions are when a specific region of a chromosome is excised, flipped, then reinserted into its original location in the DNA sequence. This can disrupt gene function and lead to abnormal gene regulation\textsuperscript{31}. Chromosomal deletions are
regions of a chromosome that is lost, resulting in the absence of all genes of that particular region, unless compensated by the other allele. Deletions can lead to haploinsufficiency, where the remaining copy of the gene is insufficient for normal function\(^{30}\). Chromosomal duplication is when a region of a chromosome is erroneously repeated and inserted back into the DNA sequence. This can result in gene dosage effects, where the increased number of gene copies can lead to harmful overexpression of that gene product. Chromosomal translocations involve the exchange of genetic material between two different chromosomes. A region from one chromosome is aberrantly incorporated into the DNA sequence of a different chromosome. Translocations can result in the formation of fusion genes, which may produce novel proteins with altered functions\(^{32}\).

![Figure 1. Schematic of point mutations and structural chromosomal alterations](image)

1. **Single nucleotide variants**
   - No mutation
   - Point mutations
     - Substitution
     - Deletion
     - Insertion
   - Functional gene product
   - Silent mutation
   - Dysfunctional gene product

2. **Structural chromosomal alterations**
   - No alteration / reference genome
   - Structural chromosomal instability
     - Inversion
     - Deletion
     - Insertion
   - Duplication
   - Translocation
   - Genes from a different chromosome

**Legends**
- Base pairs (bps)
- Regions of chromosome (Genes)

1.2.4 **Epigenetic regulation of gene expression**

Beyond the explicit nucleotide sequence of DNA and RNA lies an epigenetic layer of regulation that adds another dimension to gene expression. As initially emphasised, cells within an organism generally contain the same genetic information. It is important to recognise this level of regulation mediates the heterogeneous gene expression profiles that exist in the wide variety of cells and tissues in multicellular organisms and establishes tissue-specific gene transcription patterns\(^{33}\). Epigenetic control is exerted through several
mechanisms, each of which alters the physical structure of the DNA and has a subsequent influence on how genes are expressed. One such mechanism is DNA methylation, which involves the addition of a methyl group to DNA, the effect of which silences gene expression by recruiting proteins involved in gene repression or by inhibiting DNA transcription from commencing near the methylated DNA\(^3\). Likewise, histone modification can either promote or inhibit the unwinding of DNA – thereby controlling access of particular genes\(^3\). Lastly, the regulatory functions of noncoding RNAs such as miRNAs are a form of epigenetic control of gene expression\(^1\). These epigenetic mechanisms are not only reversible but also do not alter the underlying genetic sequence, contrasting them with mutations with genetic changes that are typically permanent unless remediated by cellular DNA repair machinery. Understanding epigenetic mechanisms is vital as they add further nuance to the already convoluted and intricate process of gene expression.

There is an extremely rich diversity of phenotypes that can arise from a given set genotypes of an individual organism. The tools in this era of next generation sequencing provide an increasingly comprehensive, scalable genomic lens at lower monetary costs to delve deeper than ever before into the many facets of biological systems and life itself.

1.3 Measuring gene expression using RNA sequencing technology

Measuring gene expression is important across various fields of biological and biomedical research, as physiological changes in organisms or cells manifest in altered gene expression patterns. Quantifying these profiles not only enriches our understanding of fundamental biology but also bridges the research gap between healthy and diseased states in biological systems. This insight sheds light on genetic factors contributing to disease pathology, susceptibility, and the potential for treatment efficacy. Building upon this, modern RNA-seq technologies evolved from predecessor technologies. This section briefly outlines three methods for transcriptome quantification, namely, microarrays, quantitative real-time PCR (qRT-PCR) and RNA-seq which have paved the way for and continue to contribute to seminal discoveries in molecular biology.
1.3.1 Microarrays:

Microarrays were first developed in the mid-late 1990s and emerged as a pioneering technology for high-throughput transcriptome analysis and remained predominant throughout the first decade of the 21st century. This technology revolutionized global gene expression studies by facilitating the simultaneous assay of thousands of genes. Over time, the manufacturing processes of microarrays have evolved, with numerous commercial providers contributing to expanding the scope of scientific inquiry they can support. Applications of microarrays extend beyond gene expression profiling; they are instrumental in tracking differential expression patterns across diverse tissues, evaluating individual variability, and are employed in Single Nucleotide Polymorphism (SNP) genotyping. Moreover, microarrays have been adapted for use in resequencing and comparative genomic hybridization, offering valuable insights into genomic copy number variations.

1.3.1.1 Principles and workflow

The foundational principle of microarrays is predicated on the hybridization of fluorescently-labelled complementary DNA (cDNA) molecules to specifically designed DNA probes immobilized on a slide. In a standard microarray experiment, the workflow encompasses several key steps: sample collection, reverse transcription of mRNA to cDNA, hybridization, washing, scanning, and data analysis. Starting with sample collection, RNA is meticulously extracted and isolated from a biological source, such as a tissue specimen, blood sample, or cultured cells. This extracted RNA is then subject to reverse transcription into cDNA. Central to this process is the use of an oligo(dT) primer—a short synthetic DNA sequence comprised of deoxythymidine (T) nucleotides—designed to anneal to the polyadenylated tail (poly-A tail) of mRNA molecules. This primer serves as the initiation point for the reverse transcriptase enzyme, thereby catalysing cDNA synthesis. As the enzyme traverses the mRNA template, it incorporates fluorescently-labelled nucleotides into the emergent cDNA strand. For hybridization to occur, the cDNA is denatured into single-stranded DNA prior to loading onto the microarray slide, permitting the binding of the complementary sequences present on the microarray slide. The microarray slide is populated with thousands to millions of DNA probes that are strategically fixed to its surface. These probes are short sequences, synthesised to be
complementary to the target genes under investigation. Post-hybridization, any unbound cDNA is rigorously washed away, setting the stage for the laser scanning of the microarray slide. The laser scanner detects the fluorescence emitted from the bound cDNA, quantifying it to ascertain gene expression levels. The fluorescence intensity serves as a proportional indicator of the specific RNA concentration within the sample, thereby enabling researchers to achieve either relative or absolute quantification of gene expression.

1.3.1.2 Strengths and limitations

The strengths of microarrays are that they are high-throughput, particularly for the early 2000s, the expression levels of thousands of genes can be assessed simultaneously. This feature also made microarrays cost-effective, especially on a per gene and per sample basis. This makes microarrays popular for large-scale studies\(^\text{37}\). Microarrays were one of the first high-throughput transcriptomic technologies, and continues to see wide-spread use today as it has matured and optimised in step with emergent next generation sequencing (NGS) technology\(^\text{38}\).

Limitations of microarray technology are that they are generally less sensitive to low abundant transcripts. Cross-hybridisation, where a single probe may bind to multiple cDNA molecules (meaning, multiple transcripts) can muddy the interpretation of the results, as it may result in the over-interpretation of the amount of a given transcript in a sample if multiple transcripts bind to the probe\(^\text{39}\). Lastly, microarrays require prior knowledge of the sequences to be studied- as this information is required for the probes present in the microarray slide. Therefore, microarrays are not suited for discovering novel genes or splice variants\(^\text{38}\).

1.3.2 Quantitative Real-Time PCR (qRT-PCR)

Quantitative Real-Time PCR (qRT-PCR) is another seminal technology in the field of molecular biology since its inception in the late 1990s\(^\text{40}\). For gene expression analysis,
qRT-PRC serves as a more accurate but less high-throughput alternative to microarrays. It utilises fluorescence-tagged probes or dyes into the Polymerase Chain Reaction (PCR) to quantify the amplification of specific RNA transcripts in real-time, thereby allowing researchers and clinicians to accurately measure the amount of particular RNA molecules within a biological sample. The advent of qRT-PCR marked a significant advancement in the realms of medical and biomedical sciences, from diagnostics, biomedical research and pharmacogenomics\textsuperscript{40}.

In a typical qRT-PRC experiment, much like for microarray experiments, the starting material is RNA isolated from a biological sample, be it tissue, cells or specific subcellular components. This isolated RNA serves as a template for the reverse transcription step, converting it into cDNA. It is from here the methodological sequences between Microarrays and qRT-PRC diverge. This cDNA then undergoes amplification through PCR, using primers specific to the target gene or genes of interest.

### 1.3.2.1 Polymerase Chain Reaction (PCR)

PCR is a molecular biology technique pivotal for amplifying specific DNA sequences from a complex mixture of DNA molecules. Developed by Kary B. Mullis in 1985\textsuperscript{41}, PCR has served as a cornerstone in various scientific applications, from diagnostics to genomics research. The core principle of PCR involves a cyclical series of temperature-controlled steps that are designed to result in the exponential amplification of a target DNA sequence. The main components of the PRC reaction mixture include

1. **DNA Template**: The DNA containing the sequence to be amplified, in the case of gene-expression analysis is the reverse-transcribed RNA from a gene recovered from a biological sample.
2. **Primers**: Short, single-stranded DNA fragments that are complementary to the flanking regions of the target sequence.
3. **DNA Polymerase**: A heat-stable enzyme that synthesizes the new DNA strands.
4. **Nucleotides (dNTPs)**: The building blocks for new DNA synthesis.
5. **Buffer**: Provides the optimal ionic and pH environment for the enzyme.
The PCR cycle consists of a series of temperature-controlled cycles, consisting of 3 key phases.

1. **Denaturation (94–98 °C):** The reaction mixture is heated to near boiling temperatures, for the purpose of separating the double-stranded DNA into single strands. This causes the separation of the two complementary strands.

2. **Annealing (50–65 °C):** This step cools the reaction mixture of denatured DNA strands and allows them to bind to the complementary sequences on the single-stranded DNA primer.

3. **Extension (72 °C):** The temperature is raised to the optimal activity temperature of the DNA polymerase, which incorporates available nucleotide (dNTPs) to the 3’ end of the annealed primer- forming new double-stranded DNA molecules.

As the extension step results in newly formed double-stranded DNA molecules, providing there are sufficient reagents available in the reaction mixture, the conditions are set to repeat the PCR cycle. A typical PCR amplification cycle undergoes 16-30 cycles, exponentially amplifying the target DNA sequence.

### 1.3.2.2 Real-time monitoring

What sets qRT-PCR apart from conventional PCR methods described above is its real-time monitoring capability. During the amplification cycles, fluorescent markers- either intercalating dyes like SYBR Green or sequence-specific probes like TaqMan are utilised to tag the DNA molecules. These markers emit fluorescence when exposed to a specific wavelength of light, and this fluorescent signal is captured at the end of each cycle. The intensity of the fluorescence is directly proportional to the amount of the target DNA of the probes, allowing for quantitative measurements of the amplified DNA.
1.3.2.3 Advantages and limitations

qRT-PCR offers a multitude of advantages that make it a go-to method for RNA / gene-expression studies. Its high sensitivity enables the detection of even minute quantities of target RNA, making it invaluable in studies requiring precise quantification. Additionally, its high specificity stems from the use of gene-specific primers, reducing the risk of non-specific amplification\textsuperscript{42}. The speed of qRT-PCR is noteworthy; a typical qRT-PCR experiment can be completed within a matter of hours, including sample preparation, amplification and data analysis.

A primary drawback of qRT-PCR is the need of prior knowledge of the target genes, which can be restrictive when investigating unknown or novel genetic landscapes. This is also true for microarrays, but the per-sample run gene-coverage afforded by microarrays make microarrays a preferred choice for exploring uncertain genetic samples. qRT-PCR are also less amenable to high-throughput, multiplex analyses compared to microarrays and next-generation sequencing methods like RNA-seq\textsuperscript{43}. The cost factor is another consideration; the reagents and specialised equipment required for qRT-PCR can make it a more expensive option for large-scale studies\textsuperscript{43}.

1.3.3 Next Generation Sequencing:

Understanding the transcriptome—the complete set of RNA transcripts produced by the genome—has emerged as a new cornerstone in unravelling both the functional elements of genomes and the biochemical capabilities of cells. While foundational methods like Microarrays and qRT-PCR have provided valuable insights, RNA-seq, enabled by Next-Generation Sequencing (NGS), represents a significant leap in gene expression analysis. NGS is a high-throughput technology made possible by a confluence of advancements, including innovations in DNA chemistry, increased computational resources, and economies of scale in sequencing\textsuperscript{44}. These collective advancements have democratized deep sequencing, making it accessible across various scales—from individual consumers seeking personalized genomic information to institutional and industrial research settings. This has contributed to a new and expanding frontiers in research, namely the ‘-omic’ era of many fields of biological, genetic and medical sciences\textsuperscript{45}. 
1.3.3.1 Deep sequencing

Deep sequencing, a subset of NGS, involves sequencing DNA or RNA molecules in a sample at high coverage, where 'coverage' refers to the number of times a particular nucleotide is read\(^4\). Typically, each nucleotide is sequenced multiple times—often ranging from 30x to 100x or more—to enhance quantitative accuracy. This depth of coverage is crucial for applications requiring precise quantification, such as translational medicine studies focused on disease biomarkers or diagnosis. Unlike qRT-PCR and Microarrays, which rely on prior knowledge of gene sequences for primer or probe design, deep sequencing provides an unbiased view of the transcriptome. This enables the discovery of novel transcripts, alternative splicing events, and other post-transcriptional modifications, thereby revolutionizing our understanding of cellular function\(^4\).

1.3.4 RNA-seq

The culmination of the advances in DNA chemistry and the increased availability of computational power are inherent to the RNA-seq methodology. There are 3 key steps involved in RNA-seq, consisting of library preparation, sequencing and data analysis\(^4\). Though before RNA-seq begins, careful consideration is given to the experimental design so that the appropriate methodology for library preparation, sequencing and data analysis are chosen for the research objective.

1.3.4.1 Library preparation

The library preparation step is a multi-step process that begins with the isolation of RNA from a biological sample, this RNA is reverse transcribed to create complementary cDNA as the template for sequencing (like for Microarrays and qRT-PCR methodologies)\(^4\). From here, RNA-seq library preparation diverges from the earlier methods. The cDNA is fragmented into smaller pieces then specialised adapters are ligated (joined) at either one or both ends of the fragmented cDNA. These adapters, composed of unique, predetermined DNA sequences, serve dual functions. First, they facilitate the anchoring
of cDNA fragments onto the sequencing platform, ensuring precise alignment for sequencing. Second, they can accommodate optional sample-specific barcodes, enabling the simultaneous sequencing of multiple samples in a single run. Following adapter ligation, a Polymerase Chain Reaction (PCR) is performed to amplify the adapter-ligated cDNA fragments, thereby increasing their quantity for sequencing. Prior to sequencing, it is crucial to address PCR artifacts, which are unintended by-products of the amplification process that can compromise the quality of the sequencing data. These artifacts include primer dimers, which are short double-stranded DNA molecules formed when primers anneal to each other rather than the target DNA, and nonspecific amplification, where the primers bind to unintended regions of the DNA template, leading to the amplification of undesired sequences that do not reflect the RNA content of the original biological sample, potentially introducing PCR bias.

To mitigate these issues, the final stage in library preparation is the clean-up process, an essential quality control measure. Various methods can be employed for this clean-up, each designed to selectively isolate the desired cDNA fragments while removing undesired elements like PCR artifacts, excess primers, and other impurities. One common approach uses magnetic beads coated with molecules that selectively bind to DNA; however, other techniques such as column-based purification or enzymatic digestion are also employed depending on the specific requirements of the experiment. Regardless of the method used, the goal is the same: to purify the library so that only the fragmented cDNA from the reverse transcribed input RNA is present to be sequenced and then represented in the data for analysis.

1.3.4.2 RNA-seq data analysis

The final aspect of RNA-seq lies in data interpretation, where invaluable insights into gene expression are gleaned. The democratization of computational power, facilitated by the advent of cloud-based services and increasingly affordable high-performance hardware, has made the analysis of voluminous sequencing data accessible to a broad range of researchers. Typically, raw sequencing data is exported from the sequencer in specialised files. These files are then computationally aligned to a reference genome or transcriptome using alignment algorithms, preparing them for downstream analysis.
Once aligned, various software tools and programming environments come into play for data interpretation. Genome viewers like the Integrative Genome Viewer (IGV)\(^{48}\) allow for the visualization of how reads align to the reference genome, providing an intuitive interface for data exploration. For more in-depth analysis, specialized bioinformatic frameworks such as DESeq2\(^{49}\) and edgeR\(^{50}\) are commonly used. These frameworks, often implemented in programming languages like R and Python, enable differential gene expression analysis, allowing researchers to identify genes that are upregulated or downregulated under specific conditions\(^{51}\).

### 1.3.4.3 Shortcomings of RNA-seq

A limitation of many RNA-seq approaches is that the RNA sequenced in these studies reflects a pooled average of RNA molecules from a large number of cells derived from tissues or cultivated cell lines\(^{52,53}\). The RNA content cannot be assayed to the resolution of the individual cell using the conventional RNA-seq approach\(^{54}\). This lack of resolution can lead to distortions in RNA-seq data where the gene expression profile of cells is a function of averaging or other forms of normalising that tends to emphasise the signal of the most abundant type of cell or expressed genes in the dataset. This can lead to the under-appreciation or total omission of the influence of rare but important types of cell in heterogeneous cellular communities such as progenitor stem cells.

The following sections introduces several key aspects directly involved and used in the work of this thesis. The most recent RNA-seq technology, called single cell RNA sequencing (scRNA-seq), how it is used to ameliorate the previous limitations of RNA-seq by resolving the gene-expression of single cells\(^{55,56}\).

### 1.4 Single cell RNA-seq

In the preceding sections, the ground-breaking capabilities of RNA-sequencing technologies in gene expression analysis were introduced. The focus for this section introduces the rapidly developing field of single cell RNA sequencing (scRNA-seq) and
how it ties into the study of the tumour microenvironment. It is important to underscore the symbiotic relationship between scRNA-seq and our evolving understanding of this complex biological landscape. scRNA-seq has endowed researchers with the capability to decipher the transcriptional activity within individual cancer cells, and also extends its utility to the comprehensive characterisation potentially all the cells of the tumour microenvironment for a given malignancy. The tumour microenvironment is comprised of a myriad of cell types and extracellular components and plays a pivotal role in cancer initiation, progression and response to treatment.

1.4.1 ScRNA-seq approaches:

Single cell RNA sequencing (scRNA-seq) is a recently developed, specialised adaptation of RNA-seq methodologies, designed to explore the transcriptome at the level of individual cells. The first scRNA-seq experiment was published in 2009\(^5\). The initial scRNA-seq were highly specialised endeavours requiring unique sample preparation, dedicated laboratory hardware and bespoke bioinformatic capabilities\(^6\). However, much like its RNA-seq predecessor, scRNA-seq quickly captured the attention of both academic researchers and biotechnology companies, accelerating its development and adoption. It is worth noting that the term 'bulk RNA-seq' gained prominence as a direct consequence of the rise of scRNA-seq. The prefix 'bulk' serves to distinguish traditional RNA-seq methodologies, which analyse RNA from a pooled mass of cells, from single-cell approaches. The term 'bulk' underscores the composite nature of the RNA source, derived from a heterogeneous mixture of cells. To avoid any ambiguity, the term 'bulk RNA-seq' will be used henceforth to refer to non-single-cell RNA-seq techniques.

1.4.1.1 Microfluidic approaches

There are two broadly categorised as scRNA-seq technology, based on the method of cell isolation and library preparation: microfluidic-based and plate-based methodologies. Microfluidic-based methods, such as the 10X Genomics Chromium system and DropSeq\(^5\), utilize specialized chips or devices to capture, sort, and process individual cells. These systems often employ droplet-based techniques where each cell is encapsulated in a droplet along with reagents for cell lysis and reverse transcription. The droplets serve as isolated microreactors, enabling parallel processing of thousands of
cells in a high-throughput manner. This approach is particularly advantageous for capturing a broad snapshot of cellular heterogeneity within a sample.

1.4.1.2 Plate-based approaches

In contrast, plate-based methods like Smart-seq\textsuperscript{259} rely on traditional well plates, often leveraging fluorescence-activated cell sorting (FACS) to deposit individual cells into separate wells. Each well then serves as a miniaturized reaction chamber for cell lysis, reverse transcription, and library preparation. While this method is generally lower in throughput compared to microfluidic systems, it often yields higher coverage and sensitivity, making it suitable for in-depth analysis of individual cells.

Microfluidic-based methods are often favoured for their higher throughput and lower reagent costs compared to plate-based methods, making them suitable for analysing a large number of cells\textsuperscript{60}. Plate-based methods, while more labour-intensive and costly, offer higher sensitivity and are better suited for specialised applications which require comprehensive coverage of the transcriptome\textsuperscript{60}. The choice between these methodologies often hinges on the specific research question, the complexity of the sample, and the level of detail required for downstream analyses\textsuperscript{61}.

1.4.1.3 scRNA-seq library preparation

A scRNA-seq experiment follows the same core steps as bulk RNA-seq: library preparation, sequencing and data analysis. However, this section will focus on the specialised adaptations within these procedural steps that were develop to address the unique challenges of single-cell analysis. For library preparation, one of the most notable innovations is the use of Unique Molecular Identifiers (UMIs)\textsuperscript{62}. While bulk RNA-seq often relies on read counts for quantification, this approach can risks introduction of amplification bias in the data. UMIs are short, random nucleotide sequences added during the reverse transcription step, tagging each original mRNA molecule with a unique identifier. This allows for the collapse of PCR duplicates during the data analysis step,
providing a more accurate measure of the original mRNA abundance in each cell. UMIs are particularly crucial in scRNA-seq, where the starting material is often extremely limited and thereby more confounded by amplification artifacts. Another distinguishing feature is the cell barcoding step, where each cell's cDNA is tagged with a unique barcode during library preparation. This enables the pooling of multiple cells in a single sequencing run while retaining the ability to deconvolute the data back to individual cells during analysis.

Developed in tandem with the hardware and methodology innovations were user-friendly computational pipelines designed for parsing and analysing single-cell data. User-friendly software packages freely available on Python and R include CellRanger, Seurat, the Bioconductor suite of packages such as Scater, and many more. The result of innovation in this area is a lowering of the computational barriers to entry, thereby making scRNA-seq accessible to a broader range of scientists. The analysis of single-cell data presents unique challenges that distinguish it from bulk RNA-seq analysis. For instance, scRNA-seq data is often sparser, with a higher frequency of zero counts for genes compared to what is observed from bulk RNA-seq data. Some strategies used involve specialisation imputation methods like MAGIC which are statistical techniques used to estimate and reduce the number of zero-count data points, reducing the sparsity of data. Alternatively, scRNA-seq specific normalisation approaches such as SCnorm—a method that estimates size factors through quantile regression which is specifically designed for scRNA-seq data.

Furthermore, single-cell analysis often involves the identification and interpretation of cell clusters, which represent different cell types or states. This is typically done through dimensionality reduction techniques like t-SNE or UMAP, the ability to analyse individual cells also allows for the tracking of cellular development over time, known as pseudo-time analysis, a feature unique to single-cell methodologies.

Furthermore, current innovation in scRNA-seq technology expanded the amount and variety of data captured from the assaying of single cells. These are known as multimodal single cell approaches such as CITE-seq (transcriptomics with cell surface proteomics) and 10X single cell transposase-accessible chromatin with sequencing (ATAC-seq) to elucidate epigenetic regulation of transcription factors and other regulatory elements can bind to the cell’s DNA. As a consequence, scRNA-seq has flourished into a well-recognised mainstream application used in a wide variety of research
endeavours, such as addressing many central questions in human genomics, such as the study of tumour heterogeneity\textsuperscript{75}, cancer evolution\textsuperscript{76} and bringing new resolution to the study of developmental biology\textsuperscript{77,78} and further characterising rare cell types\textsuperscript{79} and discovering entirely new ones\textsuperscript{80}.

1.5 Cell type identification: Microscopy to scRNA-seq

Multicellular organisms are comprised of a diverse set of cell types, each contributing to the overall function and complexity of the organism. These myriad cell communities enable the wide range of biological functions that characterise life, from the neural circuits in the brain, the hepatocytes of the liver cleansing the blood of the body, to the immune cells that defend against pathogens.

1.5.1 Microscopy

Traditionally, cell types have been defined based on a combination of morphological characteristics, functional roles, and marker gene expression. Microscopy is one of the earliest methods used for cell type identification, offering the unique advantage of visualising cells within their spatial context. It primarily relies on morphological characteristics such as cell size, shape, and the presence of specific organelles for categorisation. However, traditional microscopy is manually intensive and low-throughput, requiring skilled interpretation. This manual approach also posed challenges for standardised cross-comparison between observers due to its qualitative nature\textsuperscript{81}. Recent advancements in machine learning offer a remedy to some of these limitations in the form of digital pathology tools\textsuperscript{82}. A study from Pillar and Ozcan 2022 presented a workflow for the virtual tissue staining of unstained tissue sections using a deep convolutional neural network\textsuperscript{83}. QuPath can rapidly process numerous microscopy slides digitally, converting qualitative features into quantitative metrics and detecting subtle characteristics, such as variations in stain intensity or precise cellular dimensions, that can be overlooked by even skilled human observers\textsuperscript{84}.
1.5.2 Immunohistochemistry

Building on the foundational insights provided by microscopy, immunohistochemistry emerged as a technique that added both specificity and objectivity to cell type identification. This method utilises antibodies, which are immunological proteins with affinity to specific proteins targets, typically pathogen specific antigen. Used in the context of immunohistochemistry, antibodies are tagged with a fluorescent dye or an enzyme that produces a colour reaction, allowing for the visualisation of cells expressing a target protein under the microscope. For proteins differentially expressed in only certain cell types, these proteins are referred to as 'marker proteins' - which have corresponding 'marker genes'. Unlike traditional microscopy, which is largely based on morphological characteristics, immunohistochemistry provides a molecular-level overlay to aid identification by targeting specific proteins that are indicative of a particular cell type\textsuperscript{85,86}. This approach not only enhances the objectivity of cell type classification but also introduces the concept of 'marker genes,' which are genes that are characteristically expressed in specific cell types. These marker genes serve as molecular signatures that can be used for more precise and automated identification of cell types and are leveraged in methods such as flow cytometry and scRNA-seq.

1.5.3 Flow cytometry

Flow cytometry advances the field of cell type identification and characterisation by providing a high-throughput, quantitative method for analysing cells based on their physical and biochemical characteristics. Flow cytometry sorts fluorescently labelled cells suspended in fluid by passing them through a laser beam. The scattering of light and fluorescence emitted by the cells are measured, providing information on cell size, viability and expression of specific surface marker proteins. Flow cytometry is a high throughput technique and can simultaneously measure multiple parameters of cells, such as the expression levels of several marker proteins, so long as the fluorescent tags used to mark the proteins can be distinguished from one another. The multi-parametric analysis offered by flow cytometry is particularly useful for identifying rare cell populations from complex, highly heterogenous cell populations, making flow cytometry a powerful tool in both research and clinical settings.
1.5.4 Bulk RNA-seq

Bulk RNA-seq revolutionised the study of cell types by providing a genome-wide snapshot of gene expression within a sample. Unlike microscopy and flow cytometry, which focus on morphological and surface markers, bulk RNA-seq captures the entire transcriptomic landscape. This allows for a more nuanced understanding of cellular function and identity, as it can reveal the expression of thousands of genes simultaneously. However, the term 'bulk' is indicative of its primary limitation: it averages the gene expression profiles of all cells in the sample. While this approach is highly informative for identifying general trends in gene expression, it lacks the resolution to distinguish between individual cells within a heterogeneous population. Consequently, the unique transcriptomic signatures of rare or transient cell types can be masked, limiting our understanding of cellular diversity and function.

1.5.5 scRNA-seq

The advent of scRNA-seq has expanded our ability to explore and define cell types by building on earlier foundation set by fluorescent, marker based approaches and Bulk RNA-seq and directly addressing some of their limitations. Unlike traditional methods, scRNA-seq allows for the transcriptomic profiling of individual cells, thereby capturing the inherent heterogeneity within a given cell population. One of the transformative aspects of scRNA-seq in cell type research is its capacity for unsupervised clustering based on individual cellular RNA expression profiles. A key point raised earlier in this chapter is that while most cells within an organism share the same genomic DNA, what sets cells apart is the unique subset of genes they actively express from that shared library of genomic DNA, this information is captured in cellular RNA expression. Unsupervised clustering is a computational technique that categorises cells into groups based on the similarity in their gene expression profiles without need of predefined labels or categories of the cells, such as marker genes. However, specific marker genes in scRNA-seq data can be searched for to confirm the identity of certain clustered cell populations, and find new marker genes in previously unidentified cell-types. This mixture of cell-type foreknowledge from earlier
studies and the agnostic, computational approach to profiling cells is powerful for generating insights from scRNA-seq data, as the landscape of cellular communities can be complex, where the true number of cell types or cell states is unknown. Using this approach has led to the discovery of entirely new cell types, refined our understanding of known ones, discovered transitional states during cell differentiation and has provided a deeper characterisation of rare cell types that are often overlooked.

Figure 1.5 Outline of bulk-RNAseq and scRNA-seq workflow

1.5.6 Practical limitations of scRNA-seq on cell-type classification

The use of scRNA-seq in defining and understanding cell types is not without challenges and limitations. A large part of chapter 2 in this thesis evaluates the impact of cell stress and viability as a challenge on downstream scRNA-seq data quality. The process of dissociating tissues into single cells is not trivial, the structure of most cells that comprise tissue do not naturally exist in a free floating suspension. Dissociating cells from tissues involves exposing the tissue and therefore the cells to a combination of ischemic, mechanical and enzymatic stress- cells respond to these stresses and can be reflected in their gene expression profiles. There is a survivorship bias as well, as the cells that retrain their integrity are preferentially selected to make it through to library preparation, so more delicate or compromised cells are less likely to be represented in the final data, while
hardier, more durable cells are potentially overrepresented, leading to distortions in the actual cell type composition of tissues.

1.6 The Tumour Microenvironment

1.6.1 Introduction to cancer
Cancer is a complex disease characterised by the unconstrained growth and spread of abnormal cells. While decades of sustained cancer research have aimed to understand the characteristics that make these cells abnormally pervasive, [the precise biological mechanisms and timelines remain undetermined]. At its core, cancer originates from cells that have undergone genetic mutations that disrupt normal regulatory mechanisms governing cell growth, division and apoptosis (programmed, regulated cell death). The biological intricacies that underlie cancer genesis and its progression are elegantly captured in the seminal ‘Hallmark of Cancer’ papers.

1.6.2 Composition and variability of the tumour microenvironment
This complex interplay occurs within the tumour microenvironment (TME), a highly heterogeneous and dynamic cellular ecosystem. The characteristics of the TME can vary widely across different types of cancers and are shaped by the interactions between tumour cells, the cells that surround tumour cells and the host cells that infiltrate the tumour. Understanding the TME is not just about resolving the characteristics of malignant cells, but to understand their characteristics as they exist in a wider context of a tumour, how these cells influence, and are influenced by, their immediate cellular milieu, in the aspiring efforts to improve cancer patients outcomes.

The composition of the tumour microenvironment varies between cancer types, site of the lesion, and patient. For example, Lee et al. produced evidence using deconvolution that suggests that myeloid lineage cells are significantly enriched in metastatic bone lesions.
and lymph node lesions compared to metastatic breast lesions\(^{87}\). In general, the key constituents of a tumour microenvironment are the cancerous cells of the tumour, cancer-associated cells such as CAFS (cancer-associated fibroblasts) and TAMS (tumour-associated macrophages), the immune cells responding to the cancer; called TILs, (tumour infiltrating leukocytes) and various residential, non-cancerous cells of the tissue\(^{87-92}\).

### 1.6.3 Immune surveillance, hot and cold TMEs

Well-documented phenomena that occur in the tumour microenvironment include immune surveillance, whereby immune cells infiltrate the tumour in an attempt to redress the inflammation at these sites and mediate anti-tumour immunity\(^{93}\). However, research using high-dimensional genomic data seems to indicate that in certain instances the TILs in the TME exhibit diminished capacity for the cytotoxic elimination of tumour cells, with inhibited proliferation in the presence of particular tumour antigens, and are ultimately unable to mount a sufficient immune response to curtail the disease\(^{91,94}\). A facet in the study of the TME is the designation of ‘hot’ and ‘cold’ tumours, a term used to describe the density, diversity and activity of the TILs in the TME\(^{92}\). ‘Hot’ TMEs are those where TILs remain effective, showing signs of T-cell-mediated inflammation, recruitment and activation. ‘Hot’ TMEs usually predict a higher response rate to immunotherapies, such as anti-programmed death ligand (PD-L1) therapy, and typically confer a better prognosis for the patient\(^{95}\). In contrast, ‘cold’ TMEs exhibit a marked absence or exclusion of TILs and few/no signs of T-cell inflammation and typically confer a poor prognosis for the patient. As such, researchers have investigated how the TME can be manipulated to convert ‘cold’ TMEs to ‘hot’ so that the patients can benefit from immunotherapy treatments\(^{92,95}\).
1.6.4 scRNA-seq view of the TME

scRNA-seq offers a powerful point of view to study the inner workings of the individual patient’s tumour microenvironment. scRNA-seq possesses the capability of compartmentalising and recapitulating the cells that comprise the tumour and its microenvironment, while retaining transcriptomic level detail of each cell participating in that specific ecosystem\textsuperscript{55}. This granularity allows the decoding of the intercellular signalling networks that occur within the tumour microenvironment and allows researchers...
to infer important facets of the tumour microenvironments previously difficult to resolve, such as cell-cell interactions through ligand-receptor signaling\textsuperscript{96,97} A tool and database repository; CellPhoneDB, has been developed to resolve the cell-cell ligand-receptor communication present in scRNA-seq datasets\textsuperscript{97}.

1.7 High Grade Serous Ovarian Carcinoma (HGSOC)

Ovarian cancer (OC) is the most lethal of all gynaecological cancers, and ranks as the 7\textsuperscript{th} most frequent cancer among in women globally\textsuperscript{98}. Annually, around 295,000 women are diagnosed, and approximately 184,000 will die\textsuperscript{99}. Of ovarian cancers, high grade serous ovarian carcinoma (HGSOC) is the most common histological subtype of OC, encompassing approximately 70\% of Epithelial Ovarian Cancer (EOC) and 90-96\% of serous OC\textsuperscript{100}. One of the main feature of HGSOC that contributes to its high mortality is the inability to detect and diagnose the disease early on. Patients who present with HGSOC typically present at a late state of the disease. Only about 20\% of HGSOC cases are diagnosed while the cancer is limited to the ovaries\textsuperscript{101}, while the majority of patients are diagnosed at a late stage of the disease already exhibiting metastasis\textsuperscript{102}. These patients present with advanced, metastatic disease that has already spread to other abdominal structures such as the omentum and peritoneum. The late state of presentation poses significant challenges for effective treatment and contributes to the poor prognosis associated with HGSOC.

1.7.1 Pathogenesis of HGSOC

The origins of the pathogenesis of HGSOC have been the a focal point of extensive debate and research. The prevailing notion and evidence strongly indicates HGSOC initially arises from serous tubal intraepithelial carcinoma (STIC) lesions in the fallopian tubes. These lesions then migrate to the ovarian epithelium, where they proliferate\textsuperscript{98,103}. A study by Kindelberger et al, found that STIC lesions and concurrent HGSOC shared identical mutations in 27 out of 29 cases, reinforcing a clonal relationship between STIC and HSGSOC\textsuperscript{104}. The genetic link is further supported by another study by Kuhn et al which found that STICs were present in approximately 10-15\% of Fallopian tubes removed prophylactically from women at high risk of developing ovarian carcinoma due to a germline BRCA mutation. Moreover, STICs are detected in 50-60\% of cases of sporadic
ovarian, tubal and so-called primary peritoneal HGSCs\textsuperscript{105}. The cells of these STIC lesions are noteworthy for their near-universal presence of TP53 (Tumour-Protein 53) mutations, a feature absent in the surrounding ovarian epithelium which underscores the critical role of these mutations in the pathogenesis of HGSOC\textsuperscript{104}.

Over the past three decades, the prognosis of HGSOC has shown modest improvement, with 5-year survival rates rising from 25\% to 47\%\textsuperscript{103,106,107}. This improvement is partly attributed to advancements in diagnostic techniques and standardisation of treatment protocols. The standard treatment of HGSOC consists of primary surgical cytoreduction followed by platinum-containing chemotherapy\textsuperscript{108,109}. The rationale for aggressive cytoreductive surgery is supported by evidence that prognosis and survival are strongly correlated with the amount of postoperative residual disease\textsuperscript{108}. Platinum-based chemotherapy is administered owing to its significant association with prolonged survival in HGSOC patients\textsuperscript{108,110}. Platinum-based chemotherapy is effective in HGSOC due to the presence of Homologous Recombination Deficiency (HRD) in many cases. HRD leads to irreparable DNA damage when treated to platinum agents, resulting in preferential cell death of cells with HRD\textsuperscript{110}. More recently, the incorporation of anti-angiogenics like Bevacizumab has shown promise in HGSOC\textsuperscript{111}. These drugs work by inhibiting angiogenesis, preventing the mass of tumour cells from forming new blood vessels, starving the tumour of the blood supply required to sustain proliferation and survive. Additionally, poly(ADP-Ribose) polymerase (PARP) inhibitors such as Olaparib have been used as maintenance therapy for platinum-sensitive, relapsed cases. They are especially effective in cells with HRD, including those with BRCA mutations, as they prevent the repair of single-strand DNA breaks- leading to cell death\textsuperscript{109,112,113}.

Despite the increase in the patient 5-year survival rate and the advancements in both clinical and molecular understanding of HGSOC, which have led to more sophisticated invention options, the overall trajectory for patients remains troubling. While 60-80\% of patients initially respond well to standard treatments such as platinum-based chemotherapy, the rate of relapse is high\textsuperscript{101}. A typical feature of HGSOC is that a small number of drug-resistant cancerous cells persist even after effective treatment, this leads
to the current reality where >80% of patients eventually relapse with chemoresistant disease\textsuperscript{109}. This acquired resistance is not a coincidence, but an key clinical aspect of the clinical biology of HGSOC which stems from the genomic and mutational heterogeneity of the disease, which includes the undesirable capacity for the cancer to amount an adaptive response to treatment\textsuperscript{101,114,115}. The following sections will delve the genomic and mutational landscape of HGSOC to better understand the mechanism behind this chemoresistance.

1.7.2 The genomic and mutational landscape of HGSOC

The genomic and mutational landscape of HGSOC is characterised by its high heterogeneity\textsuperscript{28,106,116}. One of the most ubiquitous features across HGSOC tumours is the presence of pathogenic mutations in the TP53 gene\textsuperscript{28,98,116–118}. TP53 encodes the p53 protein, a critical tumour suppressor that plays a pivotal role in maintaining genomic stability by regulating gene transcription. The genes P53 regulates are related to cellular homeostasis and genome integrity. For instance, TP53 activates proteins in response to DNA damage, causing cells to arrest growth by disabling cell cycle progression beyond the G1/S stage\textsuperscript{119,120}. The majority of TP53 mutations in HGSOC are missense mutations, according for approximately 70% of mutations that result in dysfunction of the protein\textsuperscript{118}. The dysfunction of the p53 protein can manifest in two primary ways: a loss of WT (wild-type) function (LoF) mutation, where the protein loses its ordinary tumour-suppressing properties, and a gain new function (GoF) as a result of the mutation. The latter can lead to activities that promote malignant cell proliferation, metastasis and other hallmarks of cancer, including acquired chemoresistance\textsuperscript{120}.

In addition to TP53 mutations, HGSOC tumours often exhibit homologous recombination deficiency due to mutations- either somatic or germline- in \textit{BRCA1} and \textit{BRCA2} genes. These mutations are observed in up to 15-20% of HGSOC patients\textsuperscript{98,116}. The homologous recombination (HR) pathway is a high-fidelity DNA repair mechanism that is critical for repairing double-strands breaks in DNA\textsuperscript{114,116,121}. \textit{BRCA1} is involved in the initial recognition of DNA damage and acts as a scaffold for the assembly of other proteins involved in HR. \textit{BRCA2} is instrumental in the recruitment of \textit{RAD51}, a DNA recombinase, to the site of DNA damage. When either \textit{BRCA1} or \textit{BRCA2} is mutated and rendered dysfunctional, the HR pathway is compromised, leading to the genomic instability and
making cells more susceptible to DNA-damaging agents like platinum-based chemotherapy. PARP inhibitors exercise their effectiveness by inhibiting the poly(ADP-Ribose) polymerase (PARP) enzyme involved in the repair of single-strand DNA breaks. In the absence of functional HR, cells rely more heavily on PARP for DNA repair. By inhibiting PARP, these drugs exacerbate the DNA damage, leading to cell death and providing a synthetic lethal interaction\textsuperscript{112}.

HGSOC also presents with other genetic abnormalities that contribute to its complex genomic landscape. A combination of point mutations and gene breakage events in HGSOC commonly inactivate other tumour suppressor genes\textsuperscript{28}. For instance, the NF1 gene is a gene that encodes a protein called neurofibromin, which plays a critical role in regulating the Ras signalling pathway. The Ras signalling pathway is a key signalling pathway that promotes cell growth and survival by protecting cells from apoptosis. LoF in NF1 leads to uncontrolled Ras activity, which contributes to increased cell proliferation and tumourigenesis\textsuperscript{122}. Another tumour suppressor gene that is frequently observed in HGSOC are mutations in the RB1 gene. The RB1 gene encodes the retinoblastoma protein. This protein is pivotal in regulating the cell cycle, specifically overseeing the transition from the G1 to S phase. In cases where is RB1 mutated or deleted, this critical regulatory checkpoint is compromised, leading to uncontrolled cell division and, consequently, tumour formation. PTEN mutations are also prevalent in HGSOC and contribute to tumourigenesis. PTEN is a tumour suppressor gene that encodes a phosphatase protein, which acts as a negative regulator of the PI3K/AKT signalling pathway. The PI3K/AKT pathway is a cellular signalling cascade that plays critical roles in regulating various cellular functions such as metabolism, growth, survival and apoptosis. A cell with PTEN LoF loses the regulatory function of the phosphatase protein and therefore leads to unchecked activation of the PI3K/AKT pathway. This results in promoted cell survival and proliferation, contributing to the development and progression of tumours. Events that lead to the loss of function of these tumour suppressor genes are less frequent than TP53 and BRCA mutations in HGSOC, but each play a role in the heterogeneity of HGSOC and its response to treatment.
While mutations in key genes like TP53, BRCA1/2, NF1, RB1 and PTEN contribute to the pathogenesis of HGSOC, the disease is also characterised by a complex array of structural genomic changes. These include copy number variations (CNVs), chromosomal translocations, amplification and deletions, which add another layer of complexity to the genomic landscape of HGSOC. Unlike other solid tumours, HGSOC is characterised more by extensive somatic copy number changes rooted in chromosomal instability and defective DNA repair\textsuperscript{116}. A comprehensive whole genome characterisation of chemoresistant ovarian cancer published by Patch et al revealed that HGSOC is driven more by genomic copy number changes than by recurrent point mutations\textsuperscript{28}. The study reports a staggering 36,561 somatic structural variants in both primary and recurrent samples, with the number of variants per tumour ranging from 48 to 1,064. Interestingly, tumours with BRCA1 mutations had a higher number of structural variants, but also where associated with a pronounced intra-tumoural host immune response.

A reoccurring structural event witnessed in HGSOC is the amplification of localised breakpoint regions of chromosome 19 around the CCNE1 locus, this amplification results in the overexpression of the CCNE1 gene\textsuperscript{28}. The CCNE1 gene encodes cyclin E1, a protein involved in cell cycle regulation. Amplification of CCNE1 leads to overexpression of cyclin E1, tilting the cell towards uncontrolled cell cycle progression and contributing to tumourigenesis.

In addition to CCNE1 amplification, HGSOC tumour often exhibit chromosomal instability, which manifests as a high frequency of CNVs and translocations. These structural changes can disrupt the normal function of various genes and pathways, further driving the heterogeneity and adaptability of HGSOC tumours. For instance, Breakage-fusion-bridge amplification observed on chromosome 1 are associated to the chemoresistance phenotype of HGSOC tumours\textsuperscript{28}. The chromosomal instability is not just a by-product of the disease but a functional advantage that facilitates its survival and adaptability. This genomic disarray serves as fertile ground for the disease to evolve during treatment. The late-stage presentation of the majority of HGSOC cases exacerbates this issue, providing the disease with a larger initial pool of cells and distant metastatic sites and, consequently, more opportunity to evolve. This increases the likelihood that a residue number of initially susceptible cancer cells stumble across a permutation of the genomic landscape that
results in chemoresistance, allowing the disease to re-emerge with a more aggressive and lethal phenotype upon relapse.

1.7.3 Transcriptomic subtypes of HGSOC

Transcriptomics, briefly touched upon earlier, is the comprehensive study of RNA transcripts in a cell. In cancer research, transcriptomic studies often aim to provide a dynamic view of the different states of the tumour and elucidate biological pathways critical for disease progression. This approach has been applied to the context of HGSOC, a disease marked by its genomic heterogeneity and complexity.

Transcriptomic subtypes are essentially clusters of samples that share similar gene expression patterns. Experimentally, these are discovered by comparing the RNA transcripts in tumour tissues against those in healthy tissues. This contrast allows researchers to identify gene expression that are specifically related to the cancerous state. Two main frameworks have emerged for categorising transcriptomic subtypes in HGSOC: The Cancer Genome Atlas (TCGA)\(^\text{123}\) and the work by Tothill et al\(^\text{124}\). The TCGA framework identifies four subtypes- MES (mesenchymal), PRO (proliferative), IMR (immunoreactive), and DIF (differentiated). Similarly, Tothill et al. propose subtypes C1 (stromal), C2 (immune), C4 (differentiated) and C5 (mesenchymal/proliferative), each with distinct gene expression profiles and associated with varying levels of disease aggressiveness and response to treatment. While the frameworks were developed in independent settings, there are studies that have explored the correlation between these transcriptomic frameworks in HGSOC. A study by Chen et al. aimed to consolidate the transcriptomic subtypes identified by TCGA, Tothill and HGSOC subtypes found in another study by Konecny into a more generalisable consensus classifier called consensusOV\(^\text{125}\). While the paper finds the four-subtype classifications show significant concordance and association with patient survival, the conclusion of this paper suggested the subtypes are not robust when retrained on new datasets. However, a recent multiomic characterisation of HGSOC study from Hollis et al. using the tools provided in ConsensusOV to classify 362 HGSOC samples also found significant overlap between the
TCGA and Tothill frameworks. Interestingly, the authors link genomic and transcriptomic features, such as HRR related genes BRCA1/2 being mutated most frequently in the IMR/C2 subtypes, and least of all in the PRO/C5 subtypes, while CCNE1 amplification was lowest in the IMR/C2 subtypes and highest in PRO/C5. The presence of these subtypes has several implications for both patients outcomes and broader scientific research. Cases with aberrations in HRR genes showed a survival benefit across all transcriptomic subtypes. Additionally, linking the genomic and transcriptomic views revealed that cases with BRCA2 mutations experienced prolonged survival and significantly higher first- and second-line chemotherapy response rates.

While these transcriptomic subtypes are not yet part of standard clinical practice, they offer significant promise for the future of patient care in HGSOC. For instance, studies suggest that these subtypes may be crucial for determining the efficacy of specific treatments like antiangiogenic agents and PARP inhibitors. The utility of transcriptomic subtypes in HGSOC is an active area of research, and further studies are needed to validate these subtypes and explore their potential translational applications in the clinical setting, including their role in predicting treatment response and guiding therapeutic decisions.

1.8 Thesis objective and aims

The principal objective of this thesis is to harness the power of single-cell RNA sequencing (scRNA-seq) to dissect the cellular heterogeneity inherent in High-Grade Serous Ovarian Cancer (HGSOC). Specifically, the aim is to generate scRNA-seq datasets from HGSOC patient samples and utilise the high-resolution data to unravel the complex, heterogenous cellular landscape of HGSOC.

In the broadest terms, the aims of this thesis are to.

1. To conduct unsupervised clustering analysis on scRNA-seq data from HGSOC tumours, thereby identifying the distinct cellular communities that exist within the tumour microenvironment.
2. To discern transcriptomic signatures across HGSOC patient tumours microenvironments, with the goal of elucidated key pathways in the disease’s malignant behaviour.
3. To augment the interpretive power of existing bulk RNA-seq datasets by integrating them with scRNA-seq data. To be achieved through computational deconvolution methods aiming to retrospectively reveal the original cellular composition of bulk samples.

In an attempt to achieve the outlined aims, the work in the proceeding chapters of this thesis will delve into the methodologies, analyses, and interpretations that collectively contribute to a nuanced understanding of HGSOC at the single-cell level. By bridging the gap between single-cell and bulk transcriptomic data, this research aims to not only advance our understanding of HGSOC but also to pave the way for more targeted therapeutic approaches in the future.
2 Chapter 2: The dissociation of physical HGSOC tumour specimens for use in 10x single cell RNA-seq studies

2.1 Introduction

2.1.1 Single cell RNA-sequencing

scRNA-seq (single cell RNA-seq) is a new and rapidly developing field that enables the interrogation of complex biological systems. scRNA-seq involves the assaying of the RNA content of individual cells, in contrast to the preceding standard bulk RNA-seq technique, which assays RNA either extracted directly from homogenised tissue or an aggregate of cells\(^{127}\). Although both RNA-seq methods have already contributed enormously to the biomedical sciences; the source of the assayed RNA has important implications for the fundamental features of the information obtained. On a per-sample basis, the information derived from bulk RNA-seq represents a transcriptomic average; skewed by the most abundant cell populations\(^{52,54}\). As a consequence of this averaging, there is a risk that the impact of subtle but biologically important cell populations becomes obfuscated\(^{128,129}\). scRNA-seq addresses this limitation because it samples the RNA content of individual cells within a sample on a per-sample basis, producing an unprecedented high-resolution view of the heterogeneity of complex biological systems\(^{130–132}\).

2.1.2 Previous studies with dissociation of ovarian / endometrial tissue methodologies from Worthington’s Tissue Dissociation Manual

In spite of modern commercial platforms that enable the initial library generation and its preparation for sequencing, there lies a nontrivial challenge of liberating the cells from their composite 3D tissue of origin in such a way that makes them amenable to library generation that is an accurate reflection of the composition of the tissue of origin. Worthington’s Tissue Dissociation guide was consulted, and 3 publications were selected to develop a manual dissociation protocol to begin producing 10X HGSOC scRNA-seq libraries from.
Chen et al. dissociated endometrial biopsies into single cell suspensions by first transferring the material onto petri dishes soaked in media and used forceps and scalpels to mechanically dissect their biopsies into pieces. Then, they dissociated the biopsies enzymatically by incubating the dissected pieces in a digestion media of 1 gm/(249 U/mg) Collagenase I and 856U/mg Hyaluronidase in HBSS and DBPS while on a rotator at 37°C for between 1-2 hours. Finally, they pipetted the suspension and passed it through a 40μM cell strainer to be cultured for their research.

Evans, J. and Salamonsen, L liberated single cell suspensions from endometrial curettage samples. They described cutting the samples and then incubating them in PBS containing 7.5 IU (International Units)/mL collagenase III and 100mg/mL DNAse I at 37°C for 40 minutes while shaking at 130 rpm (rotations per minute). The dissociated suspension is vacuum filtered through 45 μM then 11 μM filters prior to pelleting by centrifugation.

The dissociation method described in Patel et al was to isolate ECFCs (Endothelial Colony-Forming Cells) from placenta tissue. The tissue was dissected prior to incubation in a digestion media consisting of 1mg/mL collagenase I, 1mg/mL DNAse I for 2 hours at 37°C. The suspension was filtered through a 100μM sieve and pelleted at 750g for 5 minutes. The pellet was resuspended in RBC (red blood cell) lysis buffer at room temperature for 10 minutes and pelleted again.

A common features found in the papers outlining their dissociation methods of endometrial specimens are the use of mechanical dissection/chopping prior to incubation in a digestion medium (enzymes featured: Collagenase(es) and DNAse) that is kept rotating and kept at body temperature 37°C for between 40 minutes and 2 hours. It is important to note that these papers were not estimating the viability of the single cells liberated by the dissociation process. Their aims were seeding culture plates or flow cytometry on the cell suspensions, which unlike scRNA-seq; are not as dependent on the integrity of the cells post-dissociation and washing for downstream application.
2.1.3 Dissociation workflows from previous HGSOC scRNA-seq studies

Winterhoff et al.\textsuperscript{136} was the first publication to feature scRNA-seq data on HGSOC specimens. Their publication featured a scRNA-seq analysis of 66 single cells from a single HGSOC patient. In their supplementary material (1), they describe collecting a fresh HGSOC tumour specimen from a patient undergoing primary debulking surgery. This specimen was left overnight in MACS tissue storage solution. The researchers used a Tumour Dissociation Kit and gentleMACs dissociation, both from Miltenyi Biotec, Inc. After dissociation, the cells were incubated with RBC lysis buffer followed by staining with calcein-AM and EthD-1 (Ethidium HomoDimer-1) to discern live-dead cells via flowcytometry.

The Shih et al.\textsuperscript{137} study obtained primary and metastatic pathological tissue, minced them prior to freezing them (in 40% FBS, 40% RPMI and 20% DMSO) for later processing. The authors describe using the Miltenyi dissociation kit, gentleMACS Octo-dissociator, and 70µM MACS Smart strainer. Then they used a Ficoll gradient to cleanse the suspension of RBCs and debris and a Miltenyi DCR (Dead Cell Removal) kit to enrich live cells prior to loading onto the BioRad droplet digital single-cell isolator.

The dissociation process of HGSOC tumour specimens described in Zhang et al.\textsuperscript{138} involved placing the fresh specimen in cold media. The authors described chopping the specimens and placing them in on the gentleMACS at 6°C for 1 hour, incubating with 1mL of \textit{Bacillus licheniformis} protease per 25mg of tissue. Afterwards, they dilute the cell suspension in cold Hank’s buffered saline, washing with trypsin, dispase, and DNAse while pipetting up and down. Finally, the cells were incubated with DCR microbeads and passed through MS columns on a MACS separator before a final viability assay and then sequencing.

The dissociation strategies in these three HGSOC scRNA-seq publications arrive at a broad consensus using various Miltenyi Biotec commercial solutions for the primary mechanical and enzymatic dissociation of fresh/preserved HGSOC tumours. In contrast to earlier studies dissociating ovarian/endometrial material, each of the scRNA-seq workflows included post-dissociation strategies to enrich the live cell fraction. The Miltenyi DCR kit (Dead Cell Removal) kit is used for post-dissociation enrichment in two of the publications while FACS is used in the other. The strategies used to dissociate tissue in
both the non-scRNA-seq, and scRNA-seq investigations, and the post-dissociation enrichment strategies are evaluated herein as part of the course for the development of an optimised protocol for the dissociation of HGSOC tumours to produce 10X scRNA-seq libraries.

The dissociation work in this chapter is presents important measurements of how 17 10X HGSOC scRNA-seq libraries were produced. There was a lot of trial, error, and analysis of data to arrive at the protocol detailed below. The rest of the work outlines in what areas the 10X scRNA-seq libraries produced from HGSOC tumours following the series of optimisations are of high quality compared to ideal 10X scRNA-seq libraries produced from blood. This work concludes with a novel pilot study evaluating the concordance of cell populations recovered in duplicate 10X scRNA-seq libraries, and discusses how it may impact the inference one can elucidate from a one-time sampling of single-cell mixture in similar studies of the tumour microenvironment.

## 2.2 Aims

The aims of optimising the dissociation of solid tumour tissue for scRNA-sequencing are to:

1. Understand the variables impacting the success and reproducibility of dissociating solid tissue with the goal of recovering a highly viable single cell suspension.
2. Develop a workflow that can reliably recover a highly viable single cell suspension amenable to scRNA-seq from fresh and cryopreserved HGSOC tumour specimens.
3. Demonstrate the reproducibility of the developed scRNA-seq workflow, from tumour to data.
2.3 Material and methods

This chapter outlines the methodological workflow from the initial acquisition of clinical HGSOC tumour specimens to the alignment of scRNA-seq datasets using CellRanger software. The results presented in this chapter encompass data generated within this specific range of steps. Chapter 3 details the analysis and interpretation of the scRNA-seq data generated from this chapter.

2.3.1 Procurement of clinical HGSOC tumour specimens.

Participants eligible for this study met the following inclusion criteria: 1) diagnosed with advanced-stage High-Grade Serous Ovarian Cancer (HGSOC) at FIGO stages III-IV; or 2) diagnosed with Low-Grade Serous Ovarian Cancer (LGSOC); and 3) scheduled for primary debulking surgery prior to the initiation of any chemotherapy regimen. Ethics approval for this study was granted by the Lothian Local Research Ethics Committee, approval number SR865/15/ES/0094.

During surgery, tumour samples were collected from both the primary site and, when available, multiple metastatic sites. These samples were immediately placed in MACS tissue storage solution (Miltenyi Biotec, Koln, Germany) and chilled on dry ice to preserve cellular integrity. The samples were then promptly transported from the operating theatre to the laboratory for immediate processing.

Two primary routes of sample processing were followed:

1. Cryopreservation: Samples were transferred from MACS tissue storage solution to cryopreservation media and stored at -80°C for later use.
2. Direct dissociation: Samples were processed immediately for scRNA-seq library preparation.

Each patient’s conditions underwent histopathological evaluation to confirm the diagnosis.
2.3.2 Cryopreservation of HGSOC tumour specimens.

Cryopreservation of HGSOC tumours was performed by placing the fresh tumour in a petri dish bathed in (chilled) 4.9mL of Gibco™ Dulbecco’s Phosphate-Buffered Salient (DPBS) with calcium and magnesium (Thermo Fisher Scientific, Waltham, MA, USA) supplemented with 0.1mL fetal calf serum (DPBS-FCS). Using scalpels, the specimen was cut into as many 1g pieces as desired for cryopreservation. 1g tumour pieces were placed in separate 5mL Nalgene (ThermoFisher) cryovials and filled with 3mL of cryopreservation media (80% Dulbecco’s Modified Eagle Medium, 10% dimethyl sulfoxide and 10% fetal calf serum.) The cryovials were placed in a CoolCell™ cryobox (Corning™, ThermoFisher, Waltham, MA, USA) and placed in a freezer at -80°C.

2.3.3 Rescuing cryopreserved HGSOC tumour specimens

Cryovials with tumour samples were removed from the cryobox stored at -80°C. The cryovials were placed on a tray chilled by dry ice and transferred to a pre-warmed a water bath to 37°C. The cryovials were visually monitored and removed once the frozen cryopreservation liquid was liquified. The tumour sample was removed from the cryopreservation media and replaced with chilled Gibco™ DPBS-FCS.

2.3.4 Manual dissociation of HGSOC tumour specimens

A 1g solid tumour specimen was initially placed in a petri dish and submerged in chilled Gibco™ DPBS-FCS. The tumour was mechanically dissociated by mincing it finely with scalpels over a 10-minute period. The minced tissue was then transferred to a 50mL Falcon tube, to which 9.94mL of Gibco™ DPBS, 400µL of fetal calf serum, 100µL of DNase (5U/mL) (Roche, Basel, Switzerland), and one of the following digestive enzyme combinations were added:

- 100µL of Liberase DL (0.26U/mL) (Roche) and 400µL of Liberase TL (0.5 U/mL) (Merck, Sigma-Aldrich, Burlington, MA, US),
- or 500µL of Neutral Protease (1 U/g tissue) (Roche),
• or 500µL of Collagenase (1 mg/mL) (Roche).

The tube was incubated at 37°C on a rotating Eppendorf ThermoMixer thermoblock set at 400rpm for 10 minutes. Subsequently, the contents were gently pipetted up and down using glass pipettes of decreasing sizes (25mL, 10mL, and then 5mL) for 2 minutes each. The suspension was diluted with 30mL of chilled staining buffer (Gibco™ DPBS without calcium and magnesium, and 2% (w/v) Bovine serum albumin (BSA)) and filtered through a 70µM strainer. The filtrate was centrifuged at 500rpm for 5 minutes at 4°C, after which the supernatant was aspirated and the pellet resuspended in a mixture of 1mL chilled staining buffer and 10mL 1x RBC lysis buffer (BioLegend, SD, California, US).

After vortexing for 5 seconds, the suspension was incubated on ice for 20 minutes. A final centrifugation step at 500rpm for 5 minutes at 4°C was performed, the supernatant was aspirated, and the pellet was resuspended in 1mL of staining buffer. At this point, the cell count and viability of the suspension were assessed as described in section 2.3.6.

2.3.5 Automated dissociation of HGSOC tumours using a Miltenyi GenlteMACS Octodissociator

Prior to dissociation, a cocktail of three proprietary enzymes- H, R and A- was prepared from the Miltenyi Tumour Dissociation Kit (Miltenyi Biotec) as per the manufacturer instructions. The enzyme cocktail HRA was prepared in a gentleMACS C-tube (Miltenyi Biotec) by mixing 200µL of enzyme H, 100µl of enzyme R, and 25µL of enzyme A. This mixture was reconstituted in 4.7mL of RPMI 1640 without L-glutamine and kept on ice while 1g of solid tumour was also transferred to the C-tube. The C-tube was then placed in the GentleMACS Octodissociator (Miltenyi Biotec) and processed using the pre-set program ‘37_h_TKD_2’ for primary and peritoneal specimens or ‘37_h_TKD_1’ for omental specimens.

Upon completion of the GentleMACS program, the C-tube was removed from the GentleMACS Octodissociator and was then passed through a pre-wetted 70µM MACS SmartStrainer (Miltenyi Biotec) into a 50mL Falcon tube. The suspension which passed the strainer was centrifuged at 500rpm for 7 minutes at 4°C, the supernatant was aspirated, and the cell pellet was resuspended in 1mL of RPMI 1640 without L-glutamine.
This 1mL suspension was then vortexed for 5 seconds. At this point a cell count and viability assay was performed as described in section 2.3.6. If the proportion of live cells to dead cells was too low (<70% live), the methods outlined in section 2.3.7 were carried out in the aim to increase the fraction of live cells in suspension by eliminating dead cell and debris.

2.3.6 Cell viability assessment

Following the preparation of a 1mL cell suspension, a 10µL aliquot of cell suspension was mixed with an equal volume of trypan blue. This 20µL mixture was then loaded into each chamber of a Countess II slide and read by a Countess II automated cell counter (Invitogen, Thermo Fisher Scientific) using its default settings. A cell count was performed and yielded the total number of cells, the trypan blue stain is used to determine the number of live and dead cells and their respective percentages.

2.3.7 Dead cell and red blood cell removal

Dead cells and cellular debris was removed by using the Dead Cell Removal Kit (Miltenyi Biotec) as per manufacturer instructions. Cell effluent from the Dead Cell Removal Kit were incubated in 10 volumes of RBC lysis buffer for 20 minutes.

After the 20 minutes incubation, the suspension was centrifuged at 500rpm for 7 minutes at 4°C, the supernatant was aspirated, and the cell pellet was resuspended in 1mL of RPMI 1640 without L-glutamine. This suspension would be reassessed for cellular viability as outlined in section 2.3.6.

2.3.8 Isolating PBMCs from patient whole blood samples

For some surgeries, along with tumour specimens, patient blood samples were delivered to the laboratory in 3mL purple blood vials for PBMC isolation and sequencing. PBMCs were isolated from patient blood using Ficoll-paque density gradient medium (Cytiva,
Merck, Sigma-Aldrich) following instructions from the Stemcell technologies website. Briefly, in a 10mL Falcon tube, a 1:1 volume ratio of DPBS-FCS was mixed with patient blood. Depending on the volume of blood, either 3mL or 10mL of Ficoll-Paque was added in a 50mL Falcon tube. Using a 10mL glass pipette, the blood-PBS mixture was gently layered atop the Ficoll-Paque in the 50mL Falcon tube. The 50mL Falcon tube was centrifuged at 800g for 25 minutes at room temperature (20-22°C), with centrifuge breaks turned off. After centrifugation, the top plasma layer was aspirated and discarded. Using a new pipette, the PBMC layer was transferred to a new 10mL Falcon tube.

The PBMCs were washed with an equal volume of DPBS-FCS, followed by centrifugation at 500 rpm for 2 minutes at 4°C, then resuspended in DPBS-FCS. From here a cell count and viability check were performed using the Countess II automated cell counter (Invitogen, Thermo Fisher Scientific) as described earlier in section 2.3.6. The routine cell viability measured following this method was >90% live cells.

2.3.9 Fluorescent activated cell sorting

Cell were suspended in 10mL PBS with 2% FCS and incubated with DAPI staining solution for 3 minutes, while protected from light. The cell sorter used for FACS in these experiments was a BD FACSArria II Cell sorter (BD Biosciences, New Jersey, US), operated by the flow cytometry facility manager Elizabeth Freyer of the Institute of Genetics and Cancer, University of Edinburgh, Scotland, UK. The cells were loaded into the FACSArria cell sorter, and FSC-A and FSC-H plots were used to gate singlets. Following singlet gating, additional gating filtering only cells that were DAPI negative was applied. These sorted cells were collected and suspended in 10mL DPBS-FCS.

2.3.10 10X Chromium 3' Single cell RNA-sequencing

Cell suspensions with the requisite viability >70% as measured using a trypan blue assay on the Countess II automated cell counter were put forward for 10X single cell sequencing using the 10X Genomics Chromium system. The cell suspension was diluted a final cell concentration of 1000 cells/µL, then the cell suspension was loaded targeting ~10,000 cells recovered as per the 10X Genomics Single Cell 3’ Gene Expression user manual.
2.3.11 DNA quality control and quantification.

The quality of the DNA libraries generated following the Single Cell 3' Gene Expression user manual were quantified using a Agilent Bioanalyser 2100 DNA 1000 kit (Agilent Technologies, Santa Clara, CA, US) performed by Sequencing Technologist Jeffrey Joseph of the Institute of Genetics and Cancer, University of Edinburgh, Scotland, UK. Further DNA QC was carried out using a Qubit Fluorometer High Sensitivity dsDNA assay (Invitrogen, Thermo Fisher Scientific) according to manufacturer’s instructions.

2.3.12 Library Sequencing

Constructed sc-RNAseq libraries were sent to Edinburgh Genomics (Scotland, UK) for single-cell RNA sequencing on a NovaSeq 6000 system (Illumina, San Diego, CA, US). Files with the read data were uploaded to and downloaded securely from the University of Edinburgh’s datastore via the Eddie compute cluster.

2.3.13 Alignment from raw base calls to filtered barcode matrices

Files with the raw base call (BCL) read data uploaded by Edinburgh Genomics were downloaded securely from the University of Edinburgh’s datastore via the Eddie compute cluster. The CellRanger v3.1.0 software command ‘cellranger mkfastq’ were used to generate FASTQ files from the BCL files. This mkfastq function is a wrapper of bcl2fastq made more convenient for handling 10X single-cell RNAseq data. The only optional flag used was –samplesheet which was set to a path to the sample sheet which contains the 10X genomic sample index names. After FASTQ files were generated, the next CellRanger function ‘cellranger count’ was used to align the FASTQ files to the GRCh38 reference transcriptome, no optional arguments were used, the –expect-cells flag if not specified lets the pipeline auto-estimate the number of expected cells.
2.3.14 Statistical Analysis

All statistical analysis was carried out using R version 4.0.3 ‘Bunny-Wunnies Freak Out’ (R Foundation for Statistical Computing, Vienna, Austria). Due to initial small sample availability, non-parametric tests were employed to compare mean cell viability across different experimental conditions. Specifically, R’s inbuilt Wilcoxon rank-sum test was used to assess differences in cell viability percentages across sequencing batches. This was executed by using the ‘wilcox.test()’ function from the base R ‘stats’ (R Core team (2022) R: A language and environment for statistical computing. R Foundation for Statistical Computer, Vienna, Austria) package.

For the comparison of Total Cell Count, Viable Cells, and Dead Cells across different enzymes, as well as for evaluating post-dissociation viability enrichment strategies, The Kruskal-Wallis rank sum test was applied using the ‘kruskal.test()’ function from the base R ‘stats’ package. Batch effect analysis is carried out extensively in this work, across the variables measured during tumour dissociation, library preparation, sequencing and mapping. The Kruskal-Wallis rank sum test was applied. This test was conducted using the ‘kruskal.test()’ function from the base R ‘stats’ package. For metrics that showed significant differences across batches, Dunn’s post-hoc test were performed for pairwise comparisons, using the ‘dunn.test’ package in R (Dinno A (2017): Dunn’s Test of Multiple Comparisons Using Rank Sums R version 1.3.5). The results were adjusted for multiple comparisons using the Bonferrroni method.

The relationship between cell viability and sequencing and mapping metrics was evaluated using Spearman’s rank correlation. This was executed using the ‘cor.test()’ function with the method set to ‘spearman’ from the base R ‘stats’ package. The corrplot (figure 2.9) was made using the ‘corrplot()’ function from the R package ‘corrplot’ (Taiyun Wei and Viliam Simko (2021), version 0.92).

2.4 Results

2.4.1 Optimisation of the manual method of dissociation

The foundation of the manual method described in the Method’s section 2.3.4 ‘Manual dissociation of HGSOC tumour specimens’ was adapted from the information and
citations presented in the Worthington tissue dissociation guide. The enzymes used by other researchers on human reproductive tissue indicated in the guide included Collagenase type 1 & Hyaluronidase\textsuperscript{133}, Collagenase type 3 & Deoxyribonuclease I\textsuperscript{134}, and various ratios of Collagenases and Neutral Protease\textsuperscript{135}. These investigations sought to determine which enzyme would result in the highest recovery of viable cells from solid-tissue dissociation. The optimisation process relied heavily on the use of cryopreserved, rather than fresh, tumour specimens. This is because the number of opportunities (i.e., patient cases) and number of samples to carry out optimisation experiments is greater in the cryopreserved setting. A single patient sample could be cut into several 1g tumour specimens, cryopreserved, and retrieved for dissociation at a later date. A key assumption in these optimisation experiments is that the adjustments that lead to improvement in these cryopreserved specimens will translate into dissociating fresh tumour specimens.

2.4.2 Determining the preferred enzymes in terms of recovering viable cell fractions from dissociated HGSOC tumour specimens.

The initial trials to evaluate the suitability of the candidate digestive enzymes consisted of processing the x3 1g tumour fractions from the same tumour specimen with each of the enzymes on three separate occasions. These early experiments indicated Liberase (average viability 42%, N = 3) and Neutral Protease (average viability of 47%, N = 3) were preferrable to collagenase (average viability 20% , N = 3) in terms of total cells recovered and the fraction of those cells being viable as determined using a trypan blue assay on a Countess II automated cell counter (Figure 2.1).

Together with the relatively low recovery of viable cells and comparable size of the dead cell fraction achieved with each enzyme; Collagenase, Liberase and Neutral Protease. The results of the Krustal-Wallis test for each enzyme across the total cell counts (p = 0.753), viable cells (p=0.255) and dead cells (p = 0.753) indicate no statistical difference in performance of these enzymes (Figure 2.1A). Plotting only the viable fraction as a percentage the performance of the neutral protease and liberase enzyme appear similar while collagenase yields on lower cell viability on average in 3 independent tests (Figure

54
2.1.B). A Kruskal-Wallis looking at the viability as found borderline statistical significance (P = 0.0343), followed by a Dunn’s test to see where average viability was detected showed an 18% average increase viability using Neutral protease compared to Collagenase (p=0.0345). Liberase showed yielded a modest 12% increase viability over collagenase, but this was not statistically significant (p=0.2). Neutral protease compared to liberase yielded average +7% viability across samples but the difference was not statistically significant (p=0.627). Taken together these results indicate collagenase would be an unsuitable enzyme going forward. Referring to Figure 2.1, it appears Neutral protease performed marginally better in terms of fraction viability than Liberase. However, given their performance similarity in fraction viability (Fig 2.1.B), before basing a decision on average performance across 3 samples with considerable variation (Fig 2.1.A) it was prudent to follow up this experiment by conducting triplicate experiments with the remaining cryopreserved specimens using only Liberase and Neutral Protease (Figure 2.2).
Figure 2.1: Outcomes from 3 consecutive head-to-head trials of 3 different enzymes on a total of 9 cryopreserved primary ovarian specimens.

A: Boxplot summaries of the total cell counts, viable cell and dead cell counts of the 3 head-to-head enzyme trials from each of the different enzymes.
2.4.3 Evaluating the consistency of dissociation performance of Liberase and Neutral Protease

In total 12 cryopreserved specimens were used in the triplicate experiment assessing performance and consistency of both enzymes Liberase and Neutral Protease. 6 specimens from patient samples E04 and E09. In both cases, there were 3 specimens for each enzyme. Further to optimising the dissociation protocol to attain sufficiently viable fractions (70%) from the specimens, an important objective is reliability of the protocol. To this aim, comparing the dissociation performances between Liberase and Neutral Protease in this series of experiments failed to discern meaningful difference. Referencing figure 2.2A: across the 12 dissociated samples, 6 each across the 2 patients, a Kruskal-Wallis test across each the measured variables total cell count resulted in ($p=0.963$), Viable cell count ($p=0.854$) and across dead cells ($p=0.788$).

In spite of there a lack of statistical difference, examining Table 2.1 which has the raw values across underscoring figures 2.2A and 2.2B could be read tentatively that the enzyme Liberase exhibited less variability as quantified by lower standard deviations compared to Neutral Protease in most measurements, most importantly in viability (%) recovered (E04 Liberase viability 42% (4% Sd) verses Neutral protease 32% (7% Sd) and E09 Liberase viability 40% (20% Sd) versus Neutral protease viability of 33% (27% Sd).
Figure 2.2: Dot plots of cell concentrations and viability recovery of each cryopreserved ovarian specimen used in the Liberase vs Neutral Protease triplicate experiments.

A: Dot plot of the cell concentration of the total cell count, viable cell count and dead cell count of each specimen used in the triplicate experiments. Dots are coloured by enzyme and shaped by patient of origin of the cryopreserved specimen used.

B: Viable fraction (%) recovered from each specimen. Separated by patient sample on the x-axis and dots coloured by enzyme used.
Table 2.1 Assessment of average and standard deviation key measurements from the triplicate experiments between enzymes Liberase and Neutral Protease from 2 separate patient samples E04 & E09.

<table>
<thead>
<tr>
<th>Enzymes Measurements</th>
<th>Patient sample E04 N = 6</th>
<th>Patient sample E09 N = 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Liberase N = 3</td>
<td>Neutral Protease N = 3</td>
</tr>
<tr>
<td>Total cells count</td>
<td>6,736,667 (808,723)</td>
<td>2,953,333 (583,980)</td>
</tr>
<tr>
<td>Viable cell count</td>
<td>2,786,533 (98,731)</td>
<td>943,333 (298,762)</td>
</tr>
<tr>
<td>Dead cell count</td>
<td>3,950,133 (710,432)</td>
<td>2,010,000 (459,627)</td>
</tr>
<tr>
<td>Viability (%)</td>
<td>42 (4)</td>
<td>32 (7)</td>
</tr>
</tbody>
</table>

Values: average (standard deviation)
2.4.4 Evaluating post-dissociation viability enrichment techniques on manually dissociated HGSOC tumours.

At this point during the optimisation, a total of 34 cryopreserved HGSOC tumour specimens had been manually dissociated, and not a single sample yielded viability of >70% viability, the minimum threshold for suitability for 10X sequencing. Other strategies beyond choice of digestive enzyme were investigated in the subsequent studies. An observation made from the earlier studies is that the current dissociation protocol yields adequate amounts of viable cells for sequencing, only the suspension is too contaminated with dead cells and debris. From here 3 strategies were considered for post-dissociation viability enrichment, FACS to enrich for DAPI- singlets and 2 Miltenyi antibody-bead conjugated kits: the Miltenyi Tumour cell isolation kit and, the Miltenyi Dead cell removal kit.

2.4.4.1 FACS to enrich for DAPI- singlets

Using FACS as a strategy to enrich for DAPI negative singlets 2 was evaluated for its suitability as post-dissociation viability enrichment methods following manual dissociation of the HGSOC tumour specimens (see methods 2.3.9). 3 dissociated cryopreserved HGSOC tumour specimens at viabilities of 40%, 39% and 35% respectively were processed on a BDF FACS Aria and these viability values were estimated again using a trypan blue assay on a Countess II following singlet gating and similar values were recorded for each of the samples 42%, 44% and 39% (an average of 4% increased viability, see Figure 2.3 B). Following singlet gating the cells were sorted based on their DAPI staining, collecting only cells that were DAPI negative (did not take the stain). These cells were re-sorted using the Aria, the percentage of cells without DAPI staining (DAPI-) in the each of the fractions after initial DAPI negative gating were 85%, 81% and 79% respectively. Taking the DAPI assay viability estimations, the FACS enriched the viability of the cells in these fractions by an average of 37%. However, assessing these same fractions using a trypan blue assay on the Countess II yielded viability estimates of 61% (+21), 61% (+22) and 62% (+27) respectively, demonstrating an average increase viability...
of 23% when taking the trypan blue assay from the Countess II values (Figure 2.3 A and B). This highlights a modest discordance between the recovered cell viability as measured using FACS to isolate DAPI negative singlets (the average viable fraction across the 3 cryopreserved tumours measured at ‘FACS DAPI negative’ was 82%) and measuring the same cell suspension via a trypan blue assay as measured on a Countess II automated cell counter (average viable fraction was 61%), resulting in a 19% discrepancy between measures of viability. The viability estimates from the Countess II were used as the principal indicator determining suitability for sequencing as this method for measuring viability is directly advised use that approach in the 10X single cell 3’ sequencing user manual.

Figure 2.3 Scatter plots of cell counts and viability estimates of 3 cryopreserved HGSOC specimens from 1 patient sample (E05) assayed before and after FACS enrichment of viable cells.

A: Total, dead and viable cell counts of 3 cryopreserved specimens before and after FACS enrichment.

B: Viability percentages of the dissociated cell fractions at 4 stages of the pilot experiment. `Pre-FACS Countess II` is the viability of the cell fractions immediately after washing the cells following the manual dissociation process, prior to incubating the cells with DAPI.

`FACS Singlet gate` is the estimated viability of the cells gated as singlets, but not sorted by DAPI positive/negative status. The viability values were determined using a Countess II with trypan blue staining.

`FACS DAPI negative` is the viability as defined by the fraction of cells that were classified as DAPI negative after gating.


`Post-FACS Countess II` viability percentages are from a Countess II using Trypan staining on the sorted, DAPI negative cell suspension retrieved from FACS.

### 2.4.4.2 Miltenyi Tumour Cell Isolation Kit

The Tumour Cell Isolation (TCI) kit developed by Miltenyi positively selects targets cells from a cellular suspension (i.e. tumour cells). 4 manually dissociated cryopreserved HGSOC specimens with viable fractions of 17%, 32%, 34% and 21% (TCI_1A – TCI_4A) respectively as determined by a trypan blue assay using a Countess II. These cell suspensions were incubated with the antibody-conjugate magnetic beads prior to being passed through a magnetic column. The cells in the solution that pass the column initially were discarded, the columns were then eluted and the following cell suspension was retained for another cell viability estimate. These samples exhibited viabilities of 65% (+48), 74% (+42), 68% (+34) and 64% (43+) (TCI_1B – TCI_4B) respectively (Figure 2.4), constituting an average viability improvement of 41% using a trypan blue stain on a Countess II.
Figure 2.4 Trypan blue viability assay profiles from a Countess II.

A: 4 cryopreserved HGSOC specimens (TCI_1 to TCI_4) viability profiles following manual dissociation.

B: The same 4 dissociated HGSOC specimens (TCI_1 to TCI_4) eluted suspension following Tumour Cell Isolation
2.4.4.3 Miltenyi Dead Cell Removal Kit

An alternative option to the Tumour Cell Isolation kit involved using a more generic purpose Dead Cell Removal (DCR) kit from Miltenyi Biotec (methods 2.3.7). The kit works in a similar way to the TCI but filters on negative selection. This means that the cells that pass through the magnetic column initially are kept, i.e., the ‘dead cells’ are retained within the magnetic column and the ‘not dead’ (i.e., viable) cells pass through. 6 cryopreserved HGSOC specimens were manually dissociated. The estimated viability percentages as determined using a trypan blue assay on a Countess II for each of these samples were 41%, 31%, 54%, 48%, 26% and 50% respectively (E1-E6, Figure 2.5 A and C). Following the application of the DCR kit on each of the samples, their respective viabilities were 45% (+4), 55% (+24), 71% (+17%), 45% (-3%), 59% (+33%) and 76% (+26%). The average increase of viability across these 6 samples was 17% (Figure 2.5 B).
Figure 2.5 Plots of cell counts and viability estimates of 6 cryopreserved primary HGSOC specimens

A: Boxplots featuring the cellular concentration (cells/mL) of 6 manually dissociated specimens. The X-axis denotes the concentrations before DCR and after DCR. The facets feature the total cell counts, dead cell count and viable cell count of each dissociated specimen identified by the colour of the dots.

B: Boxplot of viable fraction (%) of each dissociated cellular suspension.

C: Bar plot of net change (Δ) of each dissociated specimen after depleting dead cells using the DCR kit.

All viability estimates are produced from an automated cell counter (Countess II) utilising trypan blue as the indicator of cellular viability.
2.4.5 **Summary results comparing FACS, Miltenyi tumour cell isolation kit and Miltenyi dead cell removal kit.**

ANOVA tests were conducted to assess the efficacy of three post-dissociation methods in enhancing cell viability in HGSOC tumour samples. The analysis of pre-enrichment viability was performed to establish a baseline and to adjust for potential confounding factors in the post-enrichment viability tests. Specifically, this was done to account for the initial state of the tumour specimens, which could influence the effectiveness of any of the post-dissociation enrichment methods. No significant differences were found in pre-enrichment viability across the methods \( (p=0.428) \), allowing for a more straightforward comparison of post-enrichment viabilities. Similarly, no significant differences were observed in post-dissociation viability across the methods \( (p=0.385) \). However, a paired t-test across all 13 cryopreserved samples in these tests revealed a statistically significant improvement in cell viability across all post-dissociation enrichment methods (mean difference +29\%, \( p=1.854e-05 \)) compared to their viability pre-dissociation.

2.4.6 **Automated dissociation yields cell fractions of higher viability than manual dissociation using Liberase.**

14 cryopreserved HGSOC specimens were used to evaluate the manual dissociation with the preferred enzyme (Liberase) and an automated platform by Miltenyi Biotec, the GentleMACS Octo-dissociator (Table 2.2). 6 cryopreserved HGSOC specimens were dissociated manually and 2 sets of 4 cryopreserved HGSOC specimens were dissociated on the GentleMACS Octo-dissociator using different device settings (soft tissue and medium tissue settings). Each dissociation method liberated a similar number of total cells across the 10 samples on average (Manual: \( N=6, \bar{\mu} = 3,228,333 \) cells/mL, Automated: \( N = 8, \bar{\mu} = 3,523,750 \) cells/mL). The cell fractions immediately after the soft tissue setting were the least viable on average (12\%), while cell fractions after manual dissociation were more viable on average (42\%) and the medium setting cell fractions performing the best prior to any post dissociation enrichment (72\%).
Each cell fraction was ran through a DCR column and viability reassessed. Cell fractions from the soft tissue automated dissociation attained an average viability of 40%, an improvement of 28% on average. The manual cell fractions improved following DCR, going from an viability average of 42% to 59%, an increase of 17% on average. Dissociated samples from the medium tissue automated dissociation went from an average of 72% viability to 83% post DCR, an increase of 11%.
Table 2.2 Summarisation of all the cryopreserved dissociation trials involving the use of the Dead Cell Removal (DCR) kit.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-DCR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3,228,333</td>
<td>3,697,500</td>
<td>3,350,000</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(1,903,464)</td>
<td>(1,561,802)</td>
<td>(2,587,470)</td>
</tr>
<tr>
<td><strong>Dead</strong></td>
<td>1,936,900</td>
<td>3,284,875</td>
<td>732,125</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(1,267,442)</td>
<td>(1,418,292)</td>
<td>(304,673)</td>
</tr>
<tr>
<td><strong>Live</strong></td>
<td>1,291,433</td>
<td>412,625</td>
<td>2,617,875</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(846,058)</td>
<td>(188,805)</td>
<td>(2,438,760)</td>
</tr>
<tr>
<td><strong>Viability (%)</strong></td>
<td>42 (11)</td>
<td>12 (4.5)</td>
<td>72 (19)</td>
</tr>
<tr>
<td><strong>Post-DCR</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1,557,667</td>
<td>1,610,000</td>
<td>2,164,500</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(1,294,127)</td>
<td>(1,556,941)</td>
<td>(2,119,378)</td>
</tr>
<tr>
<td><strong>Dead</strong></td>
<td>736,173</td>
<td>1,276,475</td>
<td>246,635</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(743,841)</td>
<td>(1,624,395)</td>
<td>(101,159)</td>
</tr>
<tr>
<td><strong>Live</strong></td>
<td>821,493</td>
<td>333,525</td>
<td>1,917,865</td>
</tr>
<tr>
<td>(cells/mL)</td>
<td>(585,517)</td>
<td>(285,965)</td>
<td>(2,030,768)</td>
</tr>
<tr>
<td><strong>Viability (%)</strong></td>
<td>59 (13)</td>
<td>40 (26)</td>
<td>83 (9)</td>
</tr>
</tbody>
</table>

Values: Average (Standard deviation)
2.4.7 Assessment of 10X single cell RNA-sequenced libraries of HGSOC tumours

In total 17 libraries collected from 9 patients were sequenced using the 10X microfluidic platform. There were 4 main sequencing runs, featuring S1 (1 library), S2-S5 (4 libraries), then S6-S13 (8 libraries) followed by S14-S17 (4 libraries).

The first HGSOC library (S1) was sequenced with a batch of Glioblastoma (GBM) 10X libraries. The following batch of 4 libraries (S2-S5) were sent for sequencing shortly after. These libraries were prepared using the manual dissociation protocol outlined in methods section 2.3.4, prior to establishing the consistent dissociation protocol featuring the DCR kit and the automated GentleMACS Octo-dissociator, outlined in methods section 2.3.5. Further batch differences come from these 5 libraries being library prepped using 10X version 2 chemistry; while subsequent 10X libraries were prepared using updated version 3 chemistry. Table 2.2 summaries the information concerning the sequencing batch, specimen (HGSOC/LGSOC/PBMC), state (fresh/preserved) and dissociation methods of each of the 17 10X libraries. The earliest 5 libraries will be summarised together as one sequencing batch (Sequencing batch 1). The subsequent 12 libraries feature 8 HGSOC samples, 2 PBMC and 2 LGSOC samples. The 4 non-HGSOC samples were used as control libraries, and will be discussed in later sections of this chapter. This section will detail key indicators of the dissociation, library preparation and post-alignment QC for libraries in a batch-wise manner.
### Table 2.3 Summary information of each specimen that went on to be a sequenced 10X sc-RNAseq library

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Seq. Batch</th>
<th>Specimen</th>
<th>State</th>
<th>Site</th>
<th>Diss. Method</th>
<th>Post-Diss. Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Metastatic</td>
<td>Manual†</td>
<td>FACS</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>None</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>HGSOC</td>
<td>Cryo</td>
<td>Primary</td>
<td>Manual†</td>
<td>TCI</td>
</tr>
<tr>
<td>S4</td>
<td>1</td>
<td>HGSOC</td>
<td>Cryo</td>
<td>Primary</td>
<td>Manual†</td>
<td>TCI</td>
</tr>
<tr>
<td>S5</td>
<td>1</td>
<td>HGSOC</td>
<td>Cryo</td>
<td>Primary</td>
<td>Automated‡</td>
<td>TCI</td>
</tr>
<tr>
<td>S6</td>
<td>2</td>
<td>HGSOC</td>
<td>Cryo</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S7</td>
<td>2</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S8</td>
<td>2</td>
<td>PBMC</td>
<td>Fresh</td>
<td>Blood</td>
<td>Ficoll◊</td>
<td>None</td>
</tr>
<tr>
<td>S9</td>
<td>2</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S10</td>
<td>2</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Metastatic</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S11</td>
<td>2</td>
<td>PBMC</td>
<td>Fresh</td>
<td>Blood</td>
<td>Ficoll◊</td>
<td>None</td>
</tr>
<tr>
<td>S12</td>
<td>2</td>
<td>LGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S13</td>
<td>2</td>
<td>LGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S14</td>
<td>3</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Primary</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S15</td>
<td>3</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Metastatic</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S16</td>
<td>3</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Metastatic</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
<tr>
<td>S17</td>
<td>3</td>
<td>HGSOC</td>
<td>Fresh</td>
<td>Metastatic</td>
<td>Automated‡</td>
<td>DCR</td>
</tr>
</tbody>
</table>

†Manual protocol in section 2.3.5, ‡Automated protocol in section 2.3.6, ◊Centrifuge protocol in section 2.3.7

#### 2.4.7.1 Dissociation of 17 sequenced 10X single-cell RNAseq libraries

Samples S1, S3 and S4 were dissociated using the manual method, while S2 and S5 were dissociated on a trial run of the Miltenyi GentleMACS Octo-dissociator. None of these samples used the post-dissociation DCR kit. Samples S3-S5 benefitted from the TCI kit viability enrichment strategy. All samples following S5, i.e. from sequencing batch 70
2 onwards; followed the protocol outlined in section 2.3.5: Automated dissociation of HGSOC tumour specimens using Miltenyi GentleMACS Octo-dissociator. The only 2 exceptions are samples S8 and S11, which are the PBMC samples. These two were processed using the procedures outlined in section 2.3.8: Isolating PBMCs from patient whole blood samples. Samples S3, S4, S5 and S6 were the only samples prepared from cryopreserved specimens (see methods 2.3.2 and 2.3.3), all other samples were produced from freshly acquired tumours / blood.

Observable batch-effects are evident and are most pronounced from sequencing batch 1 to batches 2 and 3. Figure 2.6 presents analysis of cell counts across the three batches. The Kruskal-Wallis test revealed differences in both Total Cell Count (p=0.031) and Viable Cell Count (p=0.03) across the batches, with a marginal result for the Dead Cell Count (p=0.058). Dunn’s post-hoc tests further hones on these differences. Specifically, a significant difference in Total Cell Count was observed between batch 1 and batch 2 (Z = -2.624, p= 0.013) and for Viable Cell Count between batch 1 and batch 2 (Z=2.563, p= 0.015), showing batch 1 generally had lower counts. No significant differences were found between batch 3 and the other two at this level of analysis.

Figure 2.7 focuses on the variability in cell viability percentages across the three sequencing batches. Batch 1 displayed the lowest average viability, registering at 71%, while Batches 2 and 3 exhibited notably higher average viabilities of 88% and 93%, respectively (Table 2.4). Wilcoxon tests substantiate these observations, revealing significant differences in viability between batch 1 and both batch 2 (p=0.037) and batch 3 (p=0.023). Importantly, the comparable viabilities observed in Batches 2 and 3—which followed the same dissociation protocols—were not statistically different (p=0.11). These findings suggest that the lower viability in batch 1 is attributable to differences in the tumour dissociation process.
Figure 2.6 Cell counts of 17 HGSOC samples across 3 sequencing batches

The boxplots represent the distribution of total cell count, viable cell count and dead cell counts across 3 sequencing batches. Cell counts were generated using a Countess II automated cell counter with cells stained with trypan blue. Statistical comparisons were made using Kruskal-Wallis test.
Figure 2.7 Viability of input cells boxplot across 3 sequencing batches
Table 2.4 Batch-wise summary of dissociation statistics of 17 sequenced 10X scRNA-seq libraries

<table>
<thead>
<tr>
<th>Batch</th>
<th>Total cell count (cells/mL)</th>
<th>Viable cell count (cells/mL)</th>
<th>Dead cell count (cells/mL)</th>
<th>Viability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,205,750</td>
<td>1,006,978</td>
<td>405,722</td>
<td>71</td>
</tr>
<tr>
<td>2</td>
<td>3,642,250</td>
<td>3,034,048</td>
<td>608,203</td>
<td>88</td>
</tr>
<tr>
<td>3</td>
<td>3,165,000</td>
<td>3,023,350</td>
<td>141,650</td>
<td>93</td>
</tr>
</tbody>
</table>

Values are averages.

2.4.7.2 Library preparation quality control metrics of 17 sequenced 10X single-cell RNAseq libraries.

Qubit dsDNA high-sensitivity assays were done primarily as a quick way to verify the presence of genetic material following key PCR amplification steps during the 10X protocol. The Agilent 2000 Bioanalyzer is used primarily to produce traces featured in the appendix (Figures S1-S3) to assess the content of cDNA and to verify the success of the bead clean-up steps during 10X library preparation in removing undesired genetic material.

The most substantial differences regarding the library preparation quality control (QC) metrics were between libraries prepared in sequencing batch 1 and those prepared in sequencing batches 2 and 3. In general, the initial cDNA yields from dissociated tumour samples in batch 1 were lower than in subsequent sequencing batches. The mean cDNA yields in batch 1 was 94ng compared to 942ng and 471ng from batches 2 and 3 (Table 2.8) respectively. The 942ng average excludes control (i.e. the non-HGSOC) libraries. Table 2.5 features the average library summary QCs for the 3 sequencing batches. The bioanalyzer traces for sequencing 1 batch cDNA and library QC were mostly acceptable,
with an error of the bioanalyzer for sample S1 cDNA that appeared to not affect the library trace (appendix Figure S1).

Sequencing batch 2 contained the most 10X scRNA-seq libraries. The libraries were also the most varied compared to the other sequencing batches. Sequencing batch 2 featured 4 HGSOC solid tumour libraries (3 fresh, 1 cryopreserved), 2 HGSOC PBMC samples and a duplicated LGSOC library. The LGSOC sample had the lowest viability percentage (74%, Table 2.6). The bioanalyzer QC for the cDNA yields were good, though sample IDs 6, 11 and 13 exhibited low library peaks (Figure 2.8), follow-up Qubit high-sensitivity assays suggested samples 6 and 13 had acceptable DNA in their libraries, while sample 11 was borderline acceptable for sequencing. The bioanalyzer trace is of the pooled library samples before sequencing featured peaks of acceptable size, greenlighting submission to the sequencing facility (Figure 2.9 A).

The 10X library QCs for sequencing batch 3 were the best out of the other sequencing batches. These libraries featured the highest cell viabilities (82-99%) and none of the samples appeared to have issues with cDNA capture or library amplification (Table 2.7). The bioanalyzer trace available for this sequencing batch was the pooled library before sending them to the sequencing facility, all the detected peaks were in acceptable ranges for submission (Figure 2.9 B).

Table 2.5 Sequencing batch-wise summary table of key 10X library preparation statistics of sequenced HGSOC samples*

<table>
<thead>
<tr>
<th>Batch</th>
<th>Total Cell Count</th>
<th>Cell Viability</th>
<th>cDNA yield BA: (ng)</th>
<th>Library Yield BA: (ng/µL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,205,750</td>
<td>71</td>
<td>94</td>
<td>3.3</td>
</tr>
<tr>
<td>2</td>
<td>2,069,750</td>
<td>88</td>
<td>942</td>
<td>36.7</td>
</tr>
<tr>
<td>3</td>
<td>3,165,000</td>
<td>93</td>
<td>471</td>
<td>33.8</td>
</tr>
</tbody>
</table>

* excludes control samples from PBMC & LGSOC
2.4.7.3 Sequencing outcomes of 17 sequenced 10X scRNA-seq libraries indicates presence of batch effect

Analysis of sequencing metrics across the three batches revealed observable batch-effects in several quality control parameters. Specifically, the Kruskal-Wallis test identified significant differences in sequencing saturation (p=0.041), RNA quality (Q30_RNA, p=0.013), barcode quality (Q30_Bcode, p=0.003), sample index quality (Q30_SampleIndex, p=0.033), and UMI quality (Q30_UMI, p=0.001). Besides number of reads, which is not statistically evaluated because by experimental design led to batch 1 having far fewer reads / library (Table 2.6). The only sequencing feature not found different was the number of valid barcodes (p=0.273).

Furthermore, Spearman's rank correlation analysis was conducted to investigate the relationship between cell viability and these sequencing metrics. No significant correlation was found between cell viability and any of the evaluated metrics (rho = 0.23, p= 0.35), suggesting that the observed batch-effects are independent of the initial cell viability. This lack of correlation also indicates that the cell viability does not significantly influence the quality of the sequencing data, thereby validating the robustness of the dissociation process optimized during the initial phase of this research.
Figure 2.8 Sequencing statistics for 17 sequenced 10X sc-RNAseq libraries across 3 independent sequencing batches, faceted across 7 measurements
Table 2.6 Batch-wise average sequencing summary statistics of 17 sequenced 10X scRNA-seq libraries

<table>
<thead>
<tr>
<th>Batch</th>
<th>N reads</th>
<th>Valid Barcodes (%)</th>
<th>Sequencing Saturation (%)</th>
<th>Q30 RNA (%)</th>
<th>Q30 Barcodes (%)</th>
<th>Q30 Index (%)</th>
<th>Q30 UMI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6,817,890*</td>
<td>97</td>
<td>15*</td>
<td>88</td>
<td>96</td>
<td>91</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>265,026,714</td>
<td>96</td>
<td>49</td>
<td>94</td>
<td>97</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>259,934,277</td>
<td>97</td>
<td>31</td>
<td>94</td>
<td>97</td>
<td>94</td>
<td>96</td>
</tr>
</tbody>
</table>

*Value from sequenced sample S1 censored from these summaries due to being an outlier observation, causing a skew of the average value.
Post alignment statistics for each library were generated in a summary .html document following alignment using Cell Ranger v3.0.2. One of the key hypotheses of the work in this chapter is the influence of input cellular viability on downstream scRNA-seq library quality. Figure 2.9 outlines the Spearman correlation between the post-alignment variables and input cellular viability. The most singular important variable representing library quality was taken to be ‘Fraction’ – which is a percentage value of ‘Fraction of reads belonging to a cell’. The higher the fraction of reads retained to cells the better, and is indicative of low ambient RNA contamination from dying cells during the 10X microfluidic GEM formation (methods section 2.3.10).

While there was a positive correlation between input cellular viability and Fraction reads, it was not a statistically robust observation ($\rho = 0.4, p = 0.1$, $95\% \text{ CI} = 0.32-0.51$). The relationship between median UMI count per cell and the median number of genes was the highest exhibited correlation ($\rho = 0.85, p = < 2.2e^{-16}$), followed by the estimated number of cells from Cell Ranger (Est_cells_CellRange) and total number of genes ($\rho = 0.67, p = 0.002$), and MRPC (Mean Reads per Cell) and median UMI count per cell ($\rho = 0.67, p = 0.003$).
Figure 2.9 Spearman correlation plot of aligned library features to cell viability


Table 2.7 Batch-wise average summary table of cell statistics of 17 sequenced 10X sc-RNAseq libraries across 3 batches.

<table>
<thead>
<tr>
<th>Batch</th>
<th>N cells</th>
<th>% Reads in Cells</th>
<th>Transcriptome size (N genes)</th>
<th>Mean Reads/Cell</th>
<th>Median N Genes/Cell</th>
<th>Median N UMI/Cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,422</td>
<td>63</td>
<td>16,165</td>
<td>6,441*</td>
<td>545</td>
<td>1,167</td>
</tr>
<tr>
<td>2</td>
<td>11,574</td>
<td>78</td>
<td>23,645</td>
<td>27,244</td>
<td>1,095</td>
<td>2,556</td>
</tr>
<tr>
<td>3</td>
<td>18,135</td>
<td>69</td>
<td>24,887</td>
<td>14,804</td>
<td>912</td>
<td>1,649</td>
</tr>
</tbody>
</table>
S1 from batch 1 is omitted from the mean calculation because it was sequenced with libraries from an entirely separate project, resulting in an inflated read-count for the size of the library and a sequencing saturation of over 90%.

2.4.8 Comparing the quality of blood and tumour-dissociated derived 10X sc-RNAseq libraries

Using the PBMC 10X scRNA-seq libraries as ‘gold standard’ libraries, it is possible to evaluate to what extent the dissociation protocol impacts the success of producing 10X sc-RNAseq libraries directly from solid tumour material.

As shown earlier, the viability of the cells from the 2 PBMC libraries were exceptionally high (N = 2, viability = 92%) contrast to sequencing batch 1 (N = 5, viability = 71%), batch 2 (N = 6, viability = 84%). Sequencing batch 3 had comparable viability of input cellular material (N = 4, viability = 93%). The QC metric `fraction of reads/cells` is an indicator of ambient RNA contamination during GEM formation, and is the key metric to evaluate how different the cells of libraries from solid tumour-derived, dissociated specimens are from the cells of libraries produced from blood-derived, non-dissociated specimens.

The 2 PBMC libraries fraction reads/cell were the highest at S11 = 87% and S8 = 86% respectively. These values set the benchmark for the tumour-derived libraries. Table 2.12 shows fraction reads/cells of each sequenced 10X scRNA-seq library in this work. The average fraction of reads/cells of batch 1 was 63%, batch 2 it was 75% and for batch 3 it was 69%. Testing for correlation between fraction of reads/cells and viability indicate across the blood samples and the dissociated tumour samples revealed no association between them (df = 15, p = 0.2, rho = 0.32, 95% CI = -0.18 - 0.69).
Table 2.8 Summary of each sequenced scRNA-seq library fraction of cells, viability and CellRanger number of cells

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Tissue</th>
<th>Batch</th>
<th>Fraction reads / cells (%)</th>
<th>Viability (%)</th>
<th>CellRanger cell number</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Blood</td>
<td>2</td>
<td>87</td>
<td>93</td>
<td>18,575</td>
</tr>
<tr>
<td>8</td>
<td>Blood</td>
<td>2</td>
<td>86</td>
<td>91</td>
<td>5,957</td>
</tr>
<tr>
<td>9</td>
<td>HGSOC</td>
<td>2</td>
<td>83</td>
<td>95</td>
<td>15,862</td>
</tr>
<tr>
<td>7</td>
<td>HGSOC</td>
<td>2</td>
<td>81</td>
<td>81</td>
<td>12,918</td>
</tr>
<tr>
<td>6</td>
<td>HGSOC</td>
<td>2</td>
<td>77</td>
<td>83</td>
<td>5,094</td>
</tr>
<tr>
<td>10</td>
<td>HGSOC</td>
<td>2</td>
<td>76</td>
<td>96</td>
<td>13,539</td>
</tr>
<tr>
<td>5</td>
<td>HGSOC</td>
<td>1</td>
<td>75</td>
<td>84</td>
<td>2,480</td>
</tr>
<tr>
<td>14</td>
<td>HGSOC</td>
<td>3</td>
<td>74</td>
<td>94</td>
<td>21,819</td>
</tr>
<tr>
<td>2</td>
<td>HGSOC</td>
<td>1</td>
<td>71</td>
<td>54</td>
<td>2,042</td>
</tr>
<tr>
<td>15</td>
<td>HGSOC</td>
<td>3</td>
<td>70</td>
<td>98</td>
<td>13,879</td>
</tr>
<tr>
<td>13</td>
<td>LGSOC</td>
<td>2</td>
<td>70</td>
<td>72</td>
<td>13,398</td>
</tr>
<tr>
<td>1</td>
<td>HGSOC</td>
<td>1</td>
<td>70</td>
<td>63</td>
<td>786</td>
</tr>
<tr>
<td>17</td>
<td>HGSOC</td>
<td>3</td>
<td>66</td>
<td>99</td>
<td>19,942</td>
</tr>
<tr>
<td>16</td>
<td>HGSOC</td>
<td>3</td>
<td>64</td>
<td>82</td>
<td>16,828</td>
</tr>
<tr>
<td>12</td>
<td>LGSOC</td>
<td>2</td>
<td>62</td>
<td>72</td>
<td>7,245</td>
</tr>
<tr>
<td>3</td>
<td>HGSOC</td>
<td>1</td>
<td>58</td>
<td>73</td>
<td>1,300</td>
</tr>
<tr>
<td>4</td>
<td>HGSOC</td>
<td>1</td>
<td>40</td>
<td>80</td>
<td>505</td>
</tr>
</tbody>
</table>

The table is colour banned based the fraction reads/cells (%) metric. Gold = PBMC samples, green ≥ 70%, orange < 70%, ≥ 50% and grey < 50%.
2.4.9 Reproducibility of the 10X single cell RNA-sequencing method using LGSOC samples

The scRNAseq Libraries produced from sample IDs 12 & 13 were derived from the same LGSOC tumour specimen post dissociation. Despite a 53% difference in total cell numbers between the duplicate libraries (S12 = 6,268 cells, S13 = 11,911 cells, post-filtering values) and differences in the library preparation outcomes (table 2.13) the proportions of cells represented in their libraries were highly concordant, differing at most by 3% in the proportion of mesenchymal cells (figure 2.14). This provides confidence that differences in cell recovery was not a result of bias in any downstream analysis of cellular subtypes.

Table 2.13 Comparison of dissociation, library preparation and post-sequencing measurements between duplicate scRNA-seq samples 12 & 13

<table>
<thead>
<tr>
<th>Measurement</th>
<th>Sample ID</th>
<th>Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total cell count (cells/mL)</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>Viability (%)</td>
<td>5,405,000</td>
<td>Dissociation</td>
</tr>
<tr>
<td>Input volume (µL)</td>
<td>75</td>
<td>Library prep</td>
</tr>
<tr>
<td>cDNA yield (ng/µL)</td>
<td>53.9</td>
<td></td>
</tr>
<tr>
<td>Library yield (ng/µL)</td>
<td>13.3</td>
<td></td>
</tr>
<tr>
<td>Estimated N of cells</td>
<td>7,245</td>
<td>Post-sequencing</td>
</tr>
<tr>
<td>Read/cell</td>
<td>21,275</td>
<td>/Alignment</td>
</tr>
<tr>
<td>Fraction reads/cell (%)</td>
<td>62</td>
<td></td>
</tr>
</tbody>
</table>
2.5 Discussion

The goal of optimising a procedure to liberate the cells from clinical specimens for 10X scRNA-seq presents many challenges. The approach taken in this work separated the procedure into two key steps, both defined by measurable dependent variables such as total, live and dead cell counts; with the most important being dissociated cellular viability. The two key steps in which the procedure was decomposed to were i) the mechanical and enzymatic dissociation of the specimen and ii) the enrichment of the cell suspension for viable cells. The starting procedures were taken from earlier work published in Worthington’s Tissue Dissociation guide. The optimisation approach improved the procedure incrementally, eventually resulting in a protocol outlined in Section 2.3. This work experimentally shows how the protocol was developed, how each step of the dissociation experienced incremental improvements through experimentation and measurement taking. This effort is rewarded with evidence of its ability to reliably produce good quality 10X scRNA-seq libraries from freshly acquired HGSOC tumour specimens, demonstrating improvement between scRNA-seq sequencing batches 1 (pilot 10X scRNA-seq libraries) to batches 2 and 3 (analysed libraries in the Chapter 3). This work provides further evidence that 10X scRNA-seq libraries produced from dissociated tissue specimens using this protocol can be equitable to 10X sc-RNAseq libraries produced from non-dissociated, blood samples in terms of their library preparation, sequencing, and post-alignment quality indicators. Lastly, novel evidence is provided that suggests the composition of scRNA-seq libraries produced from tumour specimens using 10X are resistant to potential sampling variations from the initial input cellular suspension into the microfluidic capturing system (Figure 2.14).

The first step of the procedure is involves cutting the tumour specimens into 1g tumour samples. These tumour samples are then dissociated using a combination of mechanical and enzymatic forces. The initial strategy was to manually slice and shear the tumour with scalpels followed by incubation with various digestive enzymes on a rotatory thermoblock (Figures 2.1, 2.2). This approach was evaluated to and then superseded by an automated commercial solution, the Miltenyi Biotec GentleMACS Octo-dissociator (Table 2.2). This
was due to the higher viable fraction yield immediately following the dissociation and the
ability to dissociate up to 8 samples simultaneously, providing a more uniform and
consistent processing of the specimens throughout the procedure and across dissociated
prospective HGSOC specimens.

Following the dissociation step, there was the matter of enriching the dissociated cell
suspension for viable cells prior to loading them onto the 10X chip. Here, the use of FACS
(Figure 2.3) and two Miltenyi commercial kits, the tumour cell isolation (TCI) kit (see Figure
2.4) and the dead cell removal (DCR) (see figure 2.5) were evaluated. The dead cell
removal kit was preferred because of its ease of use and efficacy in enriching the viable
population of dissociated cell suspensions. Further benefits of the dead cell removal kit
were that it functions based on negative selection, meaning it is less likely to be
selectively depleting potential cells of interest in the HGSOC TME. Although the TCI kit
raised the viability of the dissociated cell suspension by an average of 41% in a trial of 4
dissociated cryopreserved samples, more than the DCR kit average of 17% increase from
6 samples. This enrichment of viability was achieved at the cost of depleting (via negative
selection) non EPCAM+ cells from the dissociated cell suspension. This results in the
depletion of PTPRC+ (immune) and ACTA2+ (fibroblast) cell populations from the cell
suspension. As one of the main aims of the investigation was to study the heterogeneity
of the tumour microenvironment of HGSOC, forfeiting major cell populations was not an
ideal solution to the problem of needing to recover sufficient viable cells. It is possible the
TCI kits enriched viability so effectively because it works on positive selection. Only cells
expressing the surface antigen of interest are retained, damaged or ruptured cells may
not express the typical surface antigen required for selection or are otherwise unable to
bind to the magnetic beads in the selection column. Additionally, tumour cells may be
more resilient to the stresses of dissociation than otherwise healthy cells. This may be
due to the need of these cells adapt to survive in adverse conditions.

The use of FACS and DAPI staining to enrich viable cells in suspension was beneficial to
the overall viability of the dissociated samples, increasing the viability from an average of
41% to 61% in 3 cryopreserved HGSOC tumour samples (Figure 2.3). The main issue of
using FACS was accessibility to and flexibility of the cytometry facility to accommodate
processing of freshly acquired clinical specimens. While it is not evident in the results
section of this work, the nontriviality of contending with coordinating the acquisition (i.e.
the debulking surgery), transportation, storage and dissociation of clinical specimens for these experiments should not be understated, particularly if one wishes to process freshly acquired specimens on the day of surgery in multi-process procedures using specialist flow cytometry equipment.

The optimised procedure concatenates the two experimentally optimised steps (dissociation and viability enrichment) to yield a cell suspension destined to producing 10X scRNA-seq libraries from HGSOC tumour specimens. A total of 17 10X scRNA-seq libraries were fully sequenced, aligned to the human genome, and subsequently analysed (see Chapter 3). Effort was taken to record quality indicators (QC) throughout the library preparation, sequencing, and post-alignment processing steps when producing each of the 10X scRNA-seq libraries for this work. As outlined in the results section, 18 QC metrics were recorded across the sequencing (7 QC metrics, Figure 2.10), mapping (5 QC metrics, Figure 2.11) and alignment (6 QC metrics, Figure 2.12) steps for each sequenced library and as sequencing batches (1,2 and 3). The sequencing metrics were influenced by the number of reads devoted to each scRNA-seq library on the sequencer. Each of the Q30 metrics (Q30 RNA, UMI, barcode and index) improved to a large extent as a function increased reads/cell, with only limited variation attributable to sequencing batch. However, there was some evidence of sequencing batch effect at the sequencing level. Sample 1 from batch 1 is an outlier from the other samples in this batch due to being sequenced separately with GBM libraries. As a result, the cells in this sample have a number of total reads comparable to the libraries of sequencing batches 2 and 3. Despite having a comparable number of reads, the Q30 values of S1 are on the lower end of what is seen in the other sequencing batches (Figure 2.10). This suggests dissociation and library preparation have downstream impacts on the quality of the DNA sent for sequencing. The mapping statistics were surprising because they indicated fewer reads (as a percentage) in sequencing batch 2 and 3 were being mapped to the transcriptome of the captured cells. However, this may be an expected trend when accounting for the difference in the number of reads devoted to these libraries compared to sequencing batch 1. An additional explanation for this observation comes from the alignment statistics in Figure 2.12,
bottom-left boxplot that shows the cells in the latest libraries (batch 2 and 3) have larger transcriptomes than the cells in the libraries of batch 1. The difference between the 10X sequencing chemistry version 2 (used in batch 1) and 3 (used in batches 2 and 3) may in part be responsible for this observation.

Sequencing batch 2, consisting of 8 10X scRNA-seq libraries, featured 4 libraries whose primary purpose was to act as experimental controls/validation for the other 4 HGSOC samples and any subsequent HGSOC samples (i.e., the 4 libraries sequenced in batch 3). Sample IDs 8 and 11 were libraries produced from blood samples from two different patients with HGSOC. The idea behind running two blood specimens on the 10X Chromium device was for them to serve as internal quality control for the dissociated tumour-derived libraries. The flagship cell atlases and datasets produced from 10X originated from blood samples63. Cells that exist in the blood are heterogeneous and are adapted for living in suspensions as single cells and navigating microfluidic systems (i.e., the human circulatory system), thereby representing the ideal input material for producing 10X single-cell RNAseq libraries. The process of obtaining these PBMCs from the blood was faster and less intensive (requiring no mechanical or enzymatic treatment) than obtaining cells from solid tumour material, which means that the blood specimen libraries act as ‘non-dissociation’ controls. Indeed, the cells derived from the blood specimens had exceptionally high viability (91% and 93%, respectively) without need any processing beyond the density gradient separation prior to loading on the 10X Chromium device. The contrast of QC metrics from these two PBMC libraries and the tumour specimen libraries was to reveal in what ways, if any; the dissociation procedure may technically bias the HGSOC 10X sc-RNAseq libraries.

The average viability of the input cells for the tumour-dissociated libraries in sequencing batches 1, 2 and 3 were 71%, 84%, and 93%, respectively. This suggests that sequenced cells in tumour-dissociated libraries, particularly batches 2 and 3 libraries; are comparable to cells derived from whole blood in terms of their general condition as determined from a trypan blue viability assay. In terms of library preparation after GEM formation and reverse transcription, the quantity of cDNA captured varied significantly between the PBMC libraries and the tumour-derived libraries. Sample S8 recorded 30ng of total cDNA and sample S11 yielded 280ng of total cDNA on the Bioanalyzer. The average amount of cDNA from tumour-derived libraries was 914ng of cDNA at this step. The total captured
RNA content should, in theory, be scaled to the predetermined input number of cells for each library, between 15,000 and 20,000 cells. With only 2 PBMC samples, it is difficult to determine what, if anything, lead to the observed difference in the RNA content of cells from the blood and the HGSOC tumour. After indexing, amplification, and size filtering of the fragments, the final library concentrations before sequencing were more similar between the PBMC and tumour-derived libraries, with sample S11 being the only outlier with only 3.3ng/µL cDNA. Sample S8 had 26.8ng/µL cDNA, which is still lower, but more in line with the average for the tumour-derived libraries at 38ng/µL cDNA.

The differing library cDNA concentrations in principle is scaled by the process of pooling the libraries, as the pooling ratios volumes are inverse to the recorded cDNA concentration. i.e., sample S11 with the lowest cDNA concentration had the greatest volume in the pooled library sent to sequencing.

A post-alignment metric anticipated to be important in distinguishing well produced 10X scRNA-seq libraries is the fraction of reads retained in cells upon GEM formation. The 10X CellRanger software flags any sequenced library with a fraction <70% with an amber warning indicating that cells in this library may be contaminated with ambient RNA, i.e., RNA molecules whose cell’s origin is unknown, not from the single cell formed in the GEM. Libraries with a retained read fraction <50% return a red warning, suggesting considerable ambient RNA contamination of the cell; see Table 2.12.

One of the main concerns with the tumour-dissociated 10X scRNA-seq libraries was the potential for RNA molecules leaking from any stressed/dying cells as a consequence of dissociation contaminating the GEMs. Recovering as high a fraction of viable cells in the suspension with frequent washes was assumed to be the best way to mitigate the risk of contaminating RNA that confounds the 10X scRNA-seq libraries. The RNA concentration in blood is reportedly an average of 14.59µg/mL (ranging between 6-23µg/mL in healthy donors). This is not adjusted for the Ficoll-plaque PBMC extraction and rounds of washing prior to loading on the 10X. Furthermore, the high viability of the PBMCs extracted from the blood samples suggests minimal leakage of intracellular RNA from the
PBMCs into the cell suspension and therefore the 10X Chromium microfluidic system during GEM formation. As shown in Table 2.12, the PBMC libraries retained the highest fraction of reads/cells as a percentage, 87% in sample S11 and 86% in sample S8. These values were marginally higher than the fraction of reads/cells recorded in the dissociated HGSOC tumour samples on average. Reaffirming the notion of PBMCs being an ideal tissue to derive cells from for 10X scRNA-seq libraries. Yet, while more difficult, we show it is feasible to attain comparable quality libraries from dissociated tumour specimens.

Sample IDs 12 & 13 were produced from a single LGSOC specimen. The idea behind these samples was to produce a technical replicate from identical input material. The inference intended to be made from this investigation is to evaluate how representative of the heterogenous TME of a tumour might be from what is usually the one-time assaying of a dissociated tumour (i.e., for the HGSOC clinical specimens). If the LGSOC duplicate samples differed in their cell type proportions or populations, this may call into question how accurate inferences can be made from any one-off sampling of a tumour microenvironment in scRNA-seq studies. These types of studies are not usually conducted because of the high cost (both in monetary and opportunity cost) of producing scRNA-seq libraries, especially from clinically valuable specimens. An observation based from Table 2.4 is that the journeys of these samples appear to differ immediately following the GEM formation out of the 10X Chromium device. The measured cDNA concentration differ between the replicates. This difference could be due to input cell sample variation within the 6.4 µL input volume, or from the mechanical and chemistry of the microfluidics and the molecular reactions that take place in the 10X Chromium device during GEM formation. Upon completing of the library preparation steps, LGSOC duplicate sample S12 and S13 had increasingly differing DNA concentrations as measured using the Qubit dsDNA high-sensitivity assay (Table 2.13). Although the magnitude of this difference looked smaller when measured using the Agilent 2000 Bioanalyzer dsDNA high-sensitivity assay. These LGSOC samples libraries were prepared in tandem, which is to say that there was minimal heterogeneity in the way these samples were prepared in the laboratory.

Post-sequencing metrics of the 2 LGSOC scRNA-seq libraries were broadly similar, with the biggest difference being the number of cells in the libraries, S12 = 7,245 cells and S13 = 13,398 cells, a difference of 53% between S13 and S12. Despite the difference in
cell number (which may in part explain the differing cDNA yields); the same cell types were found through clustering in these samples at highly concordant proportions (Figure 2.13 and 2.14). This suggests that assaying at least ~6-7K cells from the input cell suspension yields a functionally similar resolution of the heterogeneity of the tumour as ~11-13K cells from the same suspension. It also provides some evidence to suggest that little to no information or detail of the tumour is missed from the one-off time sampling of a dissociated tumour suspension in the curation of a 10X scRNA-seq library if a sufficient number of cells are sequenced. This evidence was obtained from duplicate sampling of a LGSOC tumour, it is assumed that this finding may be applicable when interpreting the HGSOC tumour 10X scRNA-seq libraries.
3 Chapter 3: Characterising the single cell landscape of HGSOC

3.1 Introduction

3.1.1 The Tumour-Immune microenvironment of HGSOC.

The tumour-immune microenvironment was known to have important implications for cancer patients before the introduction of single cell RNA-seq. A landmark study published in the New England Journal of Medicine (NEJM) by Zhang et al.\textsuperscript{140} used IHC and RT-PCR to quantify the presence of tumour infiltrating CD3+ T-lymphocytes in frozen OC samples. The authors demonstrated a 38% 5-year survival for patients whose tumours included the presence of CD3+ T-lymphocytes and only a 4.5% survival for patients with an absence of CD3+ T-lymphocytes. Following from the publication there was a study in which they performed IHC on 117 epithelial ovarian carcinoma cases and report the presence of specifically CD8+ (cytotoxic) T-lymphocytes that contributed to improved patient survival\textsuperscript{141}. These authors did not find evidence of CD3+ (naïve) T-lymphocytes independently contributing to increased patient survival, but they did note that the results seemed to indicate that their presence was beneficial to the CD8+ T-lymphocytes. This study established that a high ratio of high regulatory T-lymphocyte to CD8+ T-lymphocyte conferred an unfavourable prognosis for patients with HGSOC. A study by Clarke et al.\textsuperscript{142} investigated the impact of infiltrating CD3+ (naïve) and CD8+ (cytotoxic) T-lymphocytes separately and reports that the presence of CD8+ T-lymphocytes confers increased PFS and OS in serous and mucinous OC, but not in endometrioid or clear cell. The presence of CD3+ T cells was correlated with improved PFS in stage III serous OC cases. The discrepancy between the impact of the CD3+ T-lymphocytes between Sato et al. and Clarke et al. could be explained by the differences in IHC methodology as some investigators stained for CD4+ and FOXP3 and used this information to further subclassify the T lymphocyte population where other investigators did not. Just considering the T-lymphocyte subset suggests that the presence and absence of certain cells within a tumour environment have clinical implications for the patient.
3.1.2 Current single cell RNA-seq research on HGSOC

The first study of HGSOC using single cell RNA-seq was by Winterhoff et al.\textsuperscript{136} in 2017, which featured 66 cells from a single fresh HGSOC tumour sample, sequenced on the 10X platform. From these 66 cells they could analyse more than 24,000 genes and were able to classify the cells into cancer epithelial and cancer associated stromal cells using unsupervised hierarchical clustering, K-means clustering, and principal component analysis (PCA). Taking into account the 24,000 genes that annotate these cells, the authors were able to determine interesting characteristics of these cells, including conducting a gene set enrichment analysis (GSEA) to further characterise the biological function of the cells to determine the extent of malignancy within the epithelial and stromal cells. An interesting problem the authors found that is still relevant to contemporary single cell RNA-seq analysis is determining a consistent definition of rare cell populations (cancer stem cells) and identifying cells with novel biological features, in this case emerging or prior chemo-resistance.

Shortly after, a study was published by Shih et al.\textsuperscript{137} featuring single cell RNA-seq from 5 HGSOC patients, comprising 8 HGSOC tumour specimens with additional benign and low-grade OC tumours. The investigators used a BioRad droplet digital SEQ single-cell isolator followed by an Illumina SureCell whole transcriptome analysis 3’ library prep kit, making this study the first plate-based HGSOC scRNA-seq study published. This study featured a t-distributed stochastic neighbour embedding (tSNE) plot of 2,911 cells clustered into distinguishable cell-types and other subtypes including CD4+/CD8+ T-lymphocytes. The broad categorisation of cells into fibroblast/stroma, epithelial/mesothelial and leukocytes is a way many contemporary single cell RNA-seq experiments across various solid tissue tumours are presented. The epithelial cells in the study exhibited interpatient heterogeneity, with HGSOC epithelial, HGSOC-fallopian epithelial, LGSOC epithelial and specifically HGSOC patient 3 epithelial forming separate clusters, while the leukocytes and fibroblasts cell populations appeared to have mixed across the patients and serous OCs, with the exception of the cells from the benign sample.
3.2 Material and methods

3.2.1 Sequencing and alignment of HGSOC single cell RNA-seq libraries

All raw sequencing data were downloaded from the Edinburgh Genomics delivery server with their md5sums checked. The raw sequencing data was formatted into the .fastq file type using Cellranger 10x ‘mkfastq’ command. These .fastq files were subsequently aligned to the Homo sapiens reference genome build GRCh38 using the Cellranger 10x ‘count’ command with the default settings. The version of Cellranger used for was version 3.1.0. The resulting files generated include a .html report which contains summary information regarding the alignment, and both raw and filtered barcode matrix folders containing the cells (barcodes), genes (features) and count matrix associated with the alignment. The filtered output was used for subsequent data processing and analysis. The difference between the raw and filtered output from this pipeline is the filtered output has been through an additional process approximating the ‘EmptyDrops’ function to discern and retain barcodes that contain only a single cell rather than barcodes corresponding to either doublets (more than 1 single cell) or empty droplets; barcodes that may have captured ambient RNA without a single cell.

3.2.2 Filtering scRNA-seq cells

After the alignment process, cells underwent two sequential rounds of filtering. Initially, only cells present in the ‘filtered feature barcode’ matrices generated by the automate CellRanger count function were considered. These matrices exclusively contain barcodes associated with detected cells. These filtered matrices were then imported into R/RStudio using Seurat version 3\textsuperscript{143} via the ‘Read10X’ function. A second layer of filtering was subsequently applied, where cells were excluded if they deviated by more than three Median Absolute Deviations (MADS) in any of the following metrics:

- Percentage of reads that are mapped to the mitochondrial genome
- Total read count
- Total gene (feature) count
3.2.3 Cell cycle and percent mitochondria scoring of single cells

Seurat version 3 allows for the regression of unwanted sources of variation. In these cases, two unwanted sources of variation were in the RNA expression were a function of the percentage of mitochondrial genes expressed, and the cell’s current cell cycle state. Seurat has a function for calculating the percentage of counts from the mitochondrial genome, ‘PercentageFeatureSet’ and supplying the pattern ‘^MT-‘ which is a pattern all mitochondrial genes start with. As for cell cycle state, Seurat has a separate function ‘CellCycleScoring’ which compares the expression of a random subset of genes to canonical marker genes for a cell in both S phase and G2M phase. This approach along with the canonical set of genes Seurat uses to represent cell-cycle states G2M and S were presented in Nestorowa et al\textsuperscript{144}. A cell is received two scores, one for each G2M and S phase. The score range between 1 and -1, a positive value suggests the cell are in either G2M or S phase (whichever phase has the higher positive value), and a negative value suggests the cells are not cycling and are in G1 phase.

The regression is carried out secondarily to normalisation as detailed in section 2.3.16. Briefly, the unwanted variation is regressed out by constructing a linear regression model where the expression level of each gene serves as the dependent variable. The mitochondrial percentages & cell cycle scores are independent variables. The residuals from this model represent the portion of the gene expression not explained by the unwanted variation, and are used as the new expression values for each gene.

3.2.4 SCTransformation, a novel single-cell count normalisation approach

SCTransform is a novel statistical approach that normalises scRNA-seq count data by constructing a general linear model (GLM) to model the relationship between the observed UMI counts (response variable) and sequencing depth (explanatory variable). The authors of this approach argue that different group of genes cannot be effectively normalised by a single, constant factor- which is the approach taken by scaling-factor-based normalisation methods\textsuperscript{64}.
The GLM used in SCTransform is formulated as follows:

\[ Y_{ij} \sim \text{NB}(\mu_{ij}, \phi_i) \]  
\[ \log(\mu_{ij}) = \beta_{0i} + \beta_{1i} \log(X_j) \] 

Equation 1 Given in\textsuperscript{64}

Here, the term is \( Y_{ij} \) the UMI count for gene \( i \) in cell \( j \), \( \mu_{ij} \) is the expected UMI count, \( \phi_i \) is the dispersion parameter for gene \( i \), \( X_j \) is the sequencing depth for cell \( j \), and \( \beta_{0i} \) and \( \beta_{1i} \) are the intercept and slope parameters for gene \( i \), respectively.

The model uses a Negative Binomial (NB) distribution to account for both the mean and the variance of the UMI counts. However, directly fitting this model to each gene can lead to overfitting, particularly for genes with low expression levels. To mitigate this, SCTransform employs “regularised negative binomial regression,” which pools information across genes with similar expression levels to regularise the estimates of \( \beta_{0i} \), \( \beta_{1i} \) and \( \phi_i \).

Once the model is fitted, Pearson residuals are calculated or each gene-cell pair. The Pearson residual is defined as:

\[ \text{Pearson Residual}_{ij} = -\frac{Y_{ij} - \mu_{ij}}{\sqrt{\mu_{ij} + \phi_i\mu_{ij}^2}} \]

These residuals serve as normalised and variance-stabilised expression values. They aim to be devoid of technical artifacts related to sequencing depth across the cells, while retaining biological variability. These expression values are used for downstream analysis for dimension reduction and unsupervised clustering.

This method available in the Seurat version 3 package using the SCTransform function, there was an additional arguments supplied to this function. As mentioned in section 2.3.15, unwanted sources of variation are regressed following the normalisation step, the function was evoked as follows ‘SCTransform(data=data, vars.to.regress = c(‘percent.mt’,’G2M score’, S.score’).’
3.2.5 Dimension reduction using PCA

For each dataset 30 principal components were computed using Seurat’s ‘RunPCA’ function. In the PCA workflow, Single Value Decomposition (SVD) is applied to gene expression matrix $X$, yielding three matrices $U$, $D$, and $V^T$. The columns of $U$ represent the principal components (PCs), which are orthogonal vectors capturing the directions of maximum variance in the data. $N$ PCs are selected, in this case $N = 30$, the first 30 columns of $U$ are retained, corresponding to the $N$ largest singular values in $D$. These singular values, when squared and normalised by the sum of all squared singular values, indicate the proportion of variance explained by each PC. This collapses the dimensionality from $p$ (number of genes in transcriptome) to $N$.

3.2.6 Finding neighbours using shared nearest neighbours

In the construction of the shared nearest neighbour graph SNN graph for each dataset, the first step involved is identifying the k-nearest neighbours for each cell. This was based on the Euclidean distance in the PCA-reduced space, which was computed in the last step. The Seurat function `FindNeighbors` first constructs a k-nearest neighbour (KNN) graph featuring each cell, with the argument `compute.SNN=TRUE` the algorithm then calculates the shared neighbours between each pair of cells (A,B) that were neighbours in the KNN graph. The Jaccard index is employed to quantify between each pair of cells based on their shared neighbours. Mathematically, the Jaccard index for a pair of cells is computed as follows.

$$ Jaccard\ Index\ J(A,B) = \frac{\text{Number of shared neighbours of } A \text{ and } B}{\text{Total unique neighbours of } A \text{ and } B} = \frac{A \cap B}{A \cup B} $$

This index serves as a robust measure to determine the strength of the relationship between two cells.

Subsequently, an edge is created between two cells in the SNN graph if their Jaccard index exceeds its predefined threshold (set to default of 0.066) and the graph is pruned.
to remove edges of low Jaccard index values. The SNN graph is often more robust to noise and can reveal finer-grained structures in the data compared to KNN graphs.

### 3.2.7 Community detection and modularity optimisation

Using the Seurat function `FindClusters` there are various modularity optimisation algorithms used to identify clusters. In all cases for these datasets, the algorithm argument was set to 1 `FindClusters(algorithm=1)` which uses the Louvain modularity optimisation algorithm. Modularity is a measure of the density of connections of nodes (i.e. cells) within their own module (i.e. clusters) against their connection with nodes outside their module. The purpose of these modularity optimisation algorithms is to iteratively group cells to find the configuration of clusters which yields the highest modularity score. An important user input to these algorithms is the resolution parameter which specifies both a minimum threshold of density within a cluster, and a maximum density between clusters. Therefore, a higher resolution parameter will lead to numerous, highly dense clusters and a lower resolution results in fewer, larger clusters. The approach used to determine the resolution parameter for these HGSOC single cell RNA-seq involved iteratively running the Louvain algorithm at a range of resolution parameters (resolution set from 0.1 to 1.5 in 0.1 increments) and calculating some objective measures of the quality of the clustering. For each iteration the measurements made were a) maximum modularity, b) number of clusters and c) silhouette score. The preferred resolution parameter determined would be the value which a) maximized modularity while; b) preferring a lower number of clusters and c) maximizing the silhouette score. The resolution parameter that fit the majority of these scRNA-seq datasets were between the 0.1-0.3 boundary. These clusters are visualised on the UMAP dimensions.

### 3.2.8 Visualising higher dimension scRNA-seq data using UMAP

Cells are visualised using UMAP (Uniform Manifold Approximation and Projection) and was done by invoking Seurat’s `RunUMAP` function. The RunUMAP function utilises the UMAP implementation provided by the ‘uwot’ R package, which is based on the original UMAP algorithm developed by Leland McInnes, John Healy and James Melville.
The ‘RunUMAP’ function was applied to the SSN graph by running the command `RunUMAP(data=data, graph='SCT_snn')`, this means the UMAP algorithm uses the SNN graph to approximate the manifold structure of the data, preserving both local and global relationships between cells. The UMAP coordinates serve as a low-dimensional representation of the cells, which make for particularly useful visualisation of scRNA-seq data.

### 3.2.9 Canonical marker identification to declare cell phenotype

With the cells clustered in a lower dimensional space it is appropriate to return to the SCT-normalised count data to uncover which genes drive the separation between the clusters. The first step for each cluster in a dataset is the detection of marker genes using the Seurat function ‘FindAllMarkers’ which uses a Wilcoxon Rank Sum test to determine statistically significant differentially expressed genes within a given cluster relative to the rest of the dataset. This test is carried out on each cluster. The main element to consider when detecting differentially expressed genes is what constitutes ‘differentially expressed’. For this investigation a gene is differentially expressed if its log fold-change is at least 0.5 greater within the cluster compared to all other clusters and the gene must be expressed in a minimum of 25% of all cells within that cluster. This is a strict approach compared to how differentially expressed genes are usually defined in 10X single cell RNA-seq data which set the thresholds to 0.25 log fold-change and a 10% inside cluster expression, the Seurat default parameters\textsuperscript{146–148}. This stricter approach was used for these data so limit the opportunity of having the same canonical markers occur in multiple clusters to reduce ambiguity when designating clusters a phenotype. When cells are clustered successfully due to biological rather than technical variation the differentially expressed genes can reveal the phenotype of the cluster, as the genes will typically relate to known biological processes and can feature the same genes one would use to differentiate cell phenotypes in flow cytometry or histological cell staining. The canonical marker genes used in this analysis were taken from the CellMarker database [http://biocc.hrbmu.edu.cn/CellMarker]\textsuperscript{149} ‘Human-cell_markers’ section and the Human
Protein Atlas [http://www.proteinatlas.org](http://www.proteinatlas.org). If a given cluster highlights canonical marker genes within its set of differentially expressed genes it can be positively identified and declared to be of that cell phenotype. The R package ‘clusterProfiler’ version 4.0 was used to conduct the enrichment analysis for the differentially expressed genes of each cluster to the CellMarker database.

![Diagram of the key computational steps to identify cell-types in scRNA-seq data](image)

**Figure 3.1** Schematic of the key computational steps to identify cell-types in scRNA-seq data

Beginning from step 1 assumes a normalised cells by gene matrix as outlined starting from method section 3.2.4. The dimension reduction from step 3.2.5 yields the cells by eigenvector matrix, which in these cases are PCAs. Steps 2-3 are outlined in method sections 3.2.6-3.2.7. Step 4 is the section 3.2.8 and lastly step 5 can involve labelling cells using known marker genes for expected cell-types as per section 3.2.9 or finding upregulated genes in novel cell communities.
<table>
<thead>
<tr>
<th>Cell type</th>
<th>Marker genes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epithelial (Ovarian cancer)</td>
<td>EPCAM, MUC16, WFDC2, CD24</td>
<td>149,152</td>
</tr>
<tr>
<td>Ciliated epithelial</td>
<td>TPPP3, PIFO, FOXJ1</td>
<td>150</td>
</tr>
<tr>
<td>Endothelial cell</td>
<td>PECAM1, MCAM, VWF</td>
<td>149,152</td>
</tr>
<tr>
<td>Fibroblast</td>
<td>VIM, CD81, ENG, ACTA2, MMP2</td>
<td>149,150</td>
</tr>
<tr>
<td>T cell</td>
<td>PTPRC, CD3E, CD3D, IL7R, CD2</td>
<td>149,150</td>
</tr>
<tr>
<td>Macrophage / Myeloid</td>
<td>CD14, C1QC, LYZ, C1QB, APOE</td>
<td>149,152</td>
</tr>
<tr>
<td>B cell</td>
<td>CD79A, MS4A1, C79B, BANK1</td>
<td>149,152</td>
</tr>
<tr>
<td>Red blood cell</td>
<td>HBB, HBA2, HBA1</td>
<td>149,153</td>
</tr>
</tbody>
</table>
Figure 3.2 Canonical marker gene expression of each cell type across patient 3 primary site, sample ID 9.

This dataset contains all the non-novel cell-types, excluding ‘Immune’, ‘MT-Tcell’, Ciliated, Mesenchymal and Erythrocytes. Beginning in the top-right UMAP are 7 unlabelled clusters (clusters 0-6). The expression of canonical marker genes is scored in each cell and the average expression of these candidate marker genes across all the cells in the cluster is returned to identify which cluster is which cell type, based on canonical marker expression. The average SCT-gene expression value illustrates which cluster on the UMAPs the canonical marker genes are expressed in the most. Each cluster is labelled as per the last UMAP.
3.2.10 Gene set enrichment analysis

A secondary method to positively identify a cluster of cells in terms of a cell phenotype is to consider the entirety of the differentially expressed genes in a GSEA. The advantages of this approach are that it can consider the presences of all the differentially expressed genes of a cluster, not only the recognisable canonical markers. This approach can offer a more descriptive definition of a cluster in well recognised terms from several databases including GO, KEGG and MSigDb which can further annotate the biological features of a cluster, aiding in identifying the cells of a cluster when the use of canonical markers is impossible (the canonical markers could not be present, or feature in many clusters, causing ambiguity while designating a cell phenotype the given clusters). The use of GSEA and the descriptive definitions it provides are especially useful when conducting the integrative analysis, where canonical markers for rarer cell types are less prevalent, and the canonical markers that are present are shared across many of the sub-clusters. ClusterProfiler was used to carry out GSEA on the differentially expressed genes testing for significantly enriched GO ‘BP’ – biological processes terms to annotate integrated sub-clusters.

3.2.11 Cell type correlation analysis

Two strategies were used to assess the extent to which positively defined clusters are similar to equivalently defined clusters across datasets (e.g. T-cells are identified in 9 of the HGSOC scRNA-seq libraries), and are therefore comparable for integrative analysis. The first is a published method called MetaNeighbour\(^{154}\) which takes a merged log count gene by cell matrix annotated with cell-type label and dataset of origin and outputs a cell-type by cell-type mean AUROC matrix. First, the correlations across cell-types of comparison is calculated based on a learnt set of variable genes. Then a cross-dataset validation step is taken by masking cell type labels and running a neighbour voting algorithm to predict the masked cell type based on its similarity to the unmasked data. The second strategy is a homebrew inspired by the former which uses a matrix of averaged expression of a set of shared variable genes across each cell-type. This matrix
is annotated with cell-type, dataset and patient. Next, the Z-scores of gene expression across each cell-type is calculated and a Pearson’s correlation is computed. Lastly these correlation scores are clustered in an unsupervised manner with the aim to show this clustering will be led predominantly by cell-type rather than dataset or patient.

3.2.12 Copy number variation prediction in single cell RNA-sequencing data

As HGSOC is an malignancy arising in the epithelium, it is considered likely that the cells driving the malignancy are of epithelial origin. To distinguish between malignant and non-malignant epithelial cells in these HGSOC scRNA-seq datasets, a published R-packaged CopyKAT was used to predict the whether the epithelial cells in HGSOC datasets were diploid (normal) or aneuploid (malignant). Briefly, CopyKAT uses Bayesian segmentation analysis to attain a genome-wide copy number profile of each cell at a 5Mb resolution and compare these profiles for different cell-types. For each dataset with epithelial cells, a genome-wide copy number profile of select ‘normal’ cells (i.e. myeloid cells) was constructed to produce a ‘normal background’ then the same process was repeated for the epithelial cells. The copy number profiles between these cell-types are then overlaid to estimate where in the epithelial cells there is evidence of chromosome amplification or deletion, and finally CopyKAT predicts whether a given epithelial cell is diploid or aneuploid.

3.3 Results

As outlined in the previous chapter, a total of 17 scRNA-seq libraries across 3 separate sequencing batches were sequenced, and are available for single-cell analysis. Of these 17 libraries, 13 were produced from HGSOC tumour material for analysis. The other 4 libraries were non-HGSOC for control purposes. These consist of 2 libraries from HGSOC patient blood and duplicate (x2) libraries from a single LGSOC tumour specimen.
### Table 3.2 Summary of each sequenced HGSOC sc-RNAseq library.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Seq. Batch</th>
<th>Library/UMAP name</th>
<th>Cells / genes in the dataset</th>
<th>N cell-types</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>S1 Fresh Primary</td>
<td>634 / 9,465</td>
<td>3</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>S2 Fresh Primary</td>
<td>1,595 / 9,967</td>
<td>2</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>S3 Cryopreserved Primary</td>
<td>741 / 6,259</td>
<td>1</td>
</tr>
<tr>
<td>S4</td>
<td>1</td>
<td>S4 Cryopreserved Primary</td>
<td>424 / 8,320</td>
<td>0</td>
</tr>
<tr>
<td>S5</td>
<td>1</td>
<td>S5 Cryopreserved Primary</td>
<td>2,173 / 10,954</td>
<td>0</td>
</tr>
<tr>
<td>S6</td>
<td>2</td>
<td>Patient 1 Cryopreserved Primary</td>
<td>4,296 / 19,148</td>
<td>6</td>
</tr>
<tr>
<td>S7</td>
<td>2</td>
<td>Patient 2 CTx Peritoneal</td>
<td>8,371 / 17,571</td>
<td>6</td>
</tr>
<tr>
<td>S9</td>
<td>2</td>
<td>Patient 3 Primary</td>
<td>13,286 / 18,566</td>
<td>7</td>
</tr>
<tr>
<td>S10</td>
<td>2</td>
<td>Patient 3 Peritoneal</td>
<td>12,303 / 17,530</td>
<td>5</td>
</tr>
<tr>
<td>S14</td>
<td>3</td>
<td>Patient 4 Primary</td>
<td>19,579 / 18,652</td>
<td>6</td>
</tr>
<tr>
<td>S15</td>
<td>3</td>
<td>Patient 4 Peritoneal</td>
<td>11,150 / 18,893</td>
<td>6</td>
</tr>
<tr>
<td>S16</td>
<td>3</td>
<td>Patient 4 Small Bowel Nodule</td>
<td>16,337 / 17,791</td>
<td>5</td>
</tr>
<tr>
<td>S17</td>
<td>3</td>
<td>Patient 4 Omentum</td>
<td>15,458 / 18,714</td>
<td>6</td>
</tr>
</tbody>
</table>
3.3.1 Overview of the single cell datasets from pilot sequencing batch 1

The HGSOC libraries featured in sequencing batch 1 are the first HGSOC libraries produced for this work. These 5 libraries feature a total of 7,113 cells (barcodes from the sequencer), resulting in 5,567 cells after filtering and SC-Transformation. On a technical level the libraries of this batch differ systematically from subsequent HGSOC libraries from sequencing batches 2 and 3 in two key ways. Firstly, the libraries in batch 1 were produced using experimental and highly heterogenous dissociation protocols (i.e. both manual and automatic mechanical & enzymatic dissociation, differing post-dissociation cleaning). Whereas the HGSOC libraries in sequencing batches 2 and 3 used a single, optimised dissociation protocol. Secondly, the library preparation protocol for the libraries in batch 1 use an earlier version of the 10X chemistry (version 2). The subsequent libraries benefitted from using a revised protocol with newer chemistry (version 3).

The libraries in sequencing batch 1 on the single-cell analysis level differ from the later HGSOC scRNA-seq libraries. The 5 libraries of this batch feature considerably fewer cells on average than batches 2 and 3 (1,113 cells vs 9,564 cells and 15,631 cells respectively). Additionally the cells in batch 1 possess lower gene coverage of the transcriptome than the later batches (552 median genes/cell vs 1,117 and 840 median genes/cells respectively). As a result, the libraries in batch 1 consist of fewer clusters on a per library-level, with these clusters featuring several ambiguities in their cluster identities (Ribosomal and Metallothionein gene clusters, clusters featuring no differentially expressed genes in S5, the S4 library yields a solitary cluster) and only few recognisable cell phenotypes compare to what would be expected in the HGSOC tumour-microenvironment (Immune cells and Epithelial Cells). The cell phenotype clusters that do emerge in these libraries do not exhibit the heterogeneity observed in the later produced HGSOC scRNA-seq libraries (i.e. no evidence of fibroblasts, endothelial or mesenchymal cells).

Therefore, the analysis taking place in this section omit the sc-RNAseq libraries from sequencing batch 1 (HGSOC scRNA-seq datasets featured in figure 3.2) and focus primarily on the libraries in sequencing batches 2 and 3.
These UMAPs depict the clusters uncovered in each of the 5 HGSOC datasets in the first sequencing batch. The first (S1) HGSOC dataset features T cells, B cells and neural cells. These neural cells were defined by their differential expression of \textit{BCAN}, \textit{PTPRZ1} and \textit{NCAM}. The S2 HGSOC dataset featured cells predominately expressing ribosomal subunit genes, some T cells, ciliated cells, myeloid cells and a cluster of cells undefined by ClusterProfiler. DEGs in this cluster included \textit{PLD3}, \textit{PAGE2}, \textit{FBXO2}, \textit{MEG3} and \textit{ECH1}. S3 featured epithelial cells and a cluster of cells expressing ribosomal subunit genes. Clustering in S4 resulted in no distinct clusters being found. In S5 the no DEGs cluster returned no DEGs at the specified thresholds (see section 3.2.4). Relaxing the threshold uncovered 4 ribosomal subunit genes, \textit{FOLR1} and \textit{EEF1A1}. The undefined cluster featured genes including \textit{JUNB}, \textit{IER2}, \textit{JUN}, \textit{FOS} and \textit{ZFP36}. 
3.3.2 Single cell landscape across HGSOC

A total of 98,252 number of cells from 4 HGSOC patients, across 8 HGSOC tumour specimens were successfully sequenced using the 10X 3’ single-cell RNA seq platform. Across these HGSOC tumours shared cell types are present, the most frequently occurring cell types were cancer epithelial cells, fibroblasts, T-cells, myeloid lineage cells and endothelial cells (figures 3.3 and 3.4).

Novel cell types include ciliated epithelial cells, evident in patient 1 and 4’s primary tumour, which share markers with the cancer epithelial cells but are differentiated by their robust expression of the marker genes *TPPP3*, *PIFO* and *FOXJ1*, see figure 3.4 A.

B cells were present in patient 2 and patient 4’s metastatic samples and patient 3’s primary tumour. These novel cell types constitute a minority of all the cells present in their respective samples, typically comprising the smallest recognised cluster.

Red blood cells were evident in patient 2 and all of patient 4’s metastatic samples, in spite of efforts to remove them during sample processing. The selection of genes these clusters express was sparse, featuring genes only from the globin family and *ALAS2*, an erythrocyte specific gene.

The clustering also rendered two distinct stromal cell populations in patients 1, 2 & 4’s peritoneal sample, referred to as fibroblasts and mesenchymal. Marker genes that differentiate these cell types are shown in figure 3.4 C. Most of the collagenase and matrix metalloproteinase genes are preferentially expressed in the fibroblasts, while the mesenchymal cells upregulate complement genes *C3* & *C7*, *TIMP3* and *MGP* across the patient datasets.

In the case of patient 2, there is evidence of the impact of their chemotherapy regime present in the dataset. This patient received neo-adjuvant chemotherapy prior to their debulking surgery where the tumour specimen was delivered to the laboratory for dissociation and sequencing. Besides the total absence of cancer epithelial cells (figure 3.3, top right UMAP ‘Patient 2 CTx Peritoneal UMAP) there are a population of T-cells classified as ‘MT-tcells’ which differ from regular T-cells by their additional expression of metallothionein, heat shock response and DNA-repair genes, possibly as a direct response to exposure of the chemotherapeutics.
The ‘Immune’ cell population from patient 1 contains upregulated genes which indicate the presence of both myeloid and lymphoid lineage cells. It is likely the sparsity of either cell type in the dataset is responsible for the lack of distinction during clustering. Figure 3.8 provides evidence via nearest neighbour voting to suggest the constituents of this cluster are predominately myeloid and as does figure 3.9 by way of Pearson correlation to other myeloid cell clusters.
Cells were clustered and coloured according to the cell type key at the bottom left-hand of the figure. Two library exclusive ‘quasi’ cell types are labelled in their respective datasets (Patient 1 ‘Immune cluster’ & Patient 2 ‘Metallothionein expressing-Tcell’).
Figure 3.5 UMAPs of 4 HGSOC scRNA-seq libraries from sequencing batch 3

Cells were clustered and coloured according to the cell type key at the bottom left-hand of the figure.
Figure 3.6 Cell type composition (%) across the 8 HGSOC scRNA-seq libraries (batches 2 & 3)
3.3.3 Differential gene expression across similar cell phenotypes in HGSOC.

Similar cell phenotypes are cell types that are closely related but when clustered on their transcriptomic data results in their separation. Examples featured here are ciliated epithelial and non-ciliated epithelial cells observed in patient 1 and 4 primary sites, sample IDs 6 and 14. After batch correction, most of the differentially expressed genes were in the ciliated epithelial cells, the most highly differentiated genes including DNFAA1, PIFO, TPPP3 and C1orf194. The non-ciliated epithelial cells exhibited few differentially expressed genes, most were ribosomal protein subunits (figure 3.6 A). Canonical markers for HGSOC epithelial cells were highly but non-differentially expressed across these cell types (figure 3.6 B).

Fibroblasts and mesenchymal cells cluster separately in patient 1, 2 and 4, sample IDs 6, 7 and 15. Differential gene expression following batch correction reveals the genes that separate these cell-types. Fibroblasts feature more genes from the collagenase family, such as COL11A1, COL6A3, COL12A1 and COL5A1. The highest differentially expressed genes in the mesenchymal cells include MGP, GPX3, PLAC9, CDF and TXNIP (figure 3.7 A).

A distinction between T-cells and Cytotoxic T-cells was only observed in patient 3, in both the primary and peritoneal sites (sample IDs 9 and 10). T-cells exhibited greater differential expression of generic T-cell markers such as CD3D, CD2 and CD3G while the Cytotoxic T-cells expressed CCL5, GNLY among other genes (figure 3.7 B).
Figure 3.7 Plots of differentially and non-differentially expressed genes between epithelial cells and ciliated (epithelial) cells

A: Volcano plot of the differentially expressed genes comparing the epithelial cells and ciliated cell clusters from the datasets that feature both these cell-types. The primary sites from patient 1 and 4, sample IDs 6 and 14. Wilcoxon Rank Sum Test was used to determine the significance of the differential expression observed between the epithelial and ciliated cells.

B: Violin plots of the expression of 3 canonical markers for HGSOC epithelial cells in the epithelial and ciliated cell clusters. These genes were not statistically differentially expressed between the two clusters when testing using the Wilcoxon Rank Sum test.
A: Volcano plot of the differentially expressed genes comparing the fibroblast and mesenchymal cell clusters from the datasets that feature both these cell-types. Patient 1 primary site, patient's 2 and 4 peritoneal site (Sample IDs 6, 7 and 15). A Wilcoxon Rank Sum Test was used to determine the significance of the differential expression observed between the fibroblast and mesenchymal clusters.

B: Volcano plot of the differentially expressed genes comparing the T-cell and cytotoxic T-cell clusters from the datasets that feature both these cell-types. Only Patient 3 exhibited the T-cell and cytotoxic T-cell distinction in both their primary site and peritoneal site (Sample ID 9 and 10).
A Wilcoxon Rank Sum Test was used to determine the significance of the differential expression observed between the T-cell and cytotoxic T-cell clusters.
3.3.4 Cell types transcriptomes across HGSOC tumour sites, condition and patients are highly correlated.

Examination of the 8 HGSOC tumours can be impacted by a mixture of both intra-patient and inter-patient heterogeneity. The sample from patient 1 was processed after a period of being cryopreserved, patient 2 subsequent to 1 round of neo-adjuvant platinum chemotherapy and samples from patients 1, 2 and 3 were sequenced on a 10X chip and sequence run separate from patient 4. Canonical marker genes characterising cell types are robust to these aforementioned sources of variation as these genes represent the strongest signal inherent to a cluster's cell type identity (see figure 3.8 and figure 3.9). Beyond canonical marker genes, it is evident below that generic variable genes (1,218 genes for figure 3.8 and 418 in figure 3.9) demonstrate that across the varied sites and patients, the cell-type classification strategy does find highly similar cells within the HGSOC scRNA-seq datasets.

The immune cluster unique to patient 1 appears to be chiefly myeloid cells, but some votes in the AUROC score indicate there are lymphoid lineage cells present in this cluster as well. This observation is made in figure 3.9, whereby Pearson correlation groups the immune cluster among other myeloid cell populations. The differences between the approaches taken to produce figures 3.8 and 3.9 are that the MetaNeighbor approach considers individual cells based on their highly variable gene (HVG) expression, but the process is slow and was therefore limited to analysing 1,750 random cells across the 8 datasets. The Z-score Pearson correlation approach considers fewer HVGs (411) across each of the clusters in the 8 datasets. Both heatmaps group cells based on phenotype and then in a non-coincidental manner, on cell compartments. In figure 3.8 the rows/columns are ordered epithelial and ciliated (tumour compartment), mesenchymal, fibroblast and endothelial (stromal compartment) and then immune, B cell, MT-T cell, cytotoxic T cells and T cells. This pattern is present in figure 3.9, where the ciliated and epithelial cells are grouped together (tumour compartment), the endothelial clusters are grouped next to fibroblast and mesenchymal cells (stromal compartment) followed by T cells, B cells and myeloid cells (immune compartment).
Figure 3.9 MetaNeighbor AUROC correlation of 1,750 random labelled cells across the HGSOC scRNA-seq datasets.

The neighbour voting algorithm in MetaNeighbor necessitates the use of a small subset of cells, an AUROC of > 0.9 means the cell-types across dataset are near equivalent for the purposes of consensus labelling between datasets.
Figure 3.10 Z-score Pearson correlation of 411 shared HGVs across 10 cell types uncovered across 8 HGSOC scRNA-seq tumour microenvironments

The colour on the heatmap is based on the Pearson correlation of the cell type Z-scores factoring 411 shared HGVs, the genes most responsible for distinguishing cell-type.
3.3.5 Using CNV estimation to distinguish between malignant and non-malignant epithelial cells in the HGSOC datasets

6 of the 8 HGSOC tumour scRNA-seq libraries featured epithelial cells. The 2 HGSOC tumour libraries that did not feature epithelial cells include sample ID 7, the neo-adjuvant chemotherapy sample and sample ID 16 taken from a small bowel nodule site of patient 4. The majority of epithelial cells in each of the other datasets with epithelial cells were predicted to be aneuploid (Table 3.3). Sample IDs 9, 10, 14 and 15 each have >90% epithelial cells predicted to be aneuploid. Sample ID 6 and 17 epithelial cell population had predicted 73% and 64% aneuploid cell population respectively.

The epithelial cells were merged in a separate dataset after using the CopyKAT algorithm in each scRNA-seq dataset to determine their ploidy status, shown in the UMAPs of figure 3.10 A. The differential gene expression analysis was done to determine which genes in these epithelial cells drive the separation between the predicted diploid and aneuploid status. It is apparent that the CNV analysis distinguishes further the (aneuploid) epithelial cells which express the genes commonly associated with HGSOC, such as WFDC2, MUC16 and features other genes of potential interest associated with HGSOC clinicopathology such as GPX3, LY6E, SLPI and IFITM3 (figure 3.10 B). It is suspected that malignant cells would feature a more active transcriptome, one hallmark being greater mitochondrial activity. The CNV analysis managed to parses cells within the epithelial and ciliated cell cluster to produce two statistically distinct cell profiles based on their average mitochondrial gene expression and number of genes expressed (figure 3.10 C).
### Table 3.3 Proportions of epithelial cells classified as either aneuploid or diploid by CopyKAT

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Library/UMAP name</th>
<th>N cells - aneuploid (%)</th>
<th>N cells - diploid (%)</th>
<th>Total Epithelial</th>
</tr>
</thead>
<tbody>
<tr>
<td>S6</td>
<td>Patient Cryopreserved Primary</td>
<td>1,753 (73%)</td>
<td>640 (27%)</td>
<td>2,393</td>
</tr>
<tr>
<td>S9</td>
<td>Patient Primary</td>
<td>3,039 (&gt;99%)</td>
<td>11 (&lt;1%)</td>
<td>3,050</td>
</tr>
<tr>
<td>S10</td>
<td>Patient Peritoneal</td>
<td>5,002 (99%)</td>
<td>39 (1%)</td>
<td>5,041</td>
</tr>
<tr>
<td>S14</td>
<td>Patient Primary</td>
<td>5,054 (95%)</td>
<td>248 (5%)</td>
<td>5,302</td>
</tr>
<tr>
<td>S15</td>
<td>Patient Peritoneal</td>
<td>1,107 (97%)</td>
<td>34 (3%)</td>
<td>1,141</td>
</tr>
<tr>
<td>S17</td>
<td>Patient Omentum</td>
<td>2,783 (64%)</td>
<td>1,539 (36%)</td>
<td>4,322</td>
</tr>
</tbody>
</table>

### 3.3.6 Feature differences between aneuploid and diploid epithelial cells

Of the 21,467 epithelial cells identified across 8 HGSOC tumour scRNA-seq libraries, 18,887 (88%) of these were predicted to be aneuploid with the remaining 2,580 (12%) cells classified as diploid. A notable difference between the aneuploid and diploid cells was the percentage of counts attributed to mitochondrial genes and number of features present in these cells. In aneuploid cells, a median of 17% of counts were from mitochondrial genes, while it was 10% in diploid cells, this difference in distribution was found to be statistically significant using a Wilcoxon test ($P < 2.22e^{-16}$). The same finding was uncovered when considering number of features (median 1,522 in aneuploid cells, 120...
1,243 in diploid). Gene expression analysis indicate genes frequently associated with HGSOC are differently expressed in the aneuploid cells relative to the diploid, such as WFDC2 and MUC16 (average log₂ fold change of 1.25 and 1.15 respectively). Other genes which exhibited greater differential expression include CLU and IFI27 (average log₂ fold change of 2.12 and 1.41 respectively).

Figure 3.11 UMAPs, volcano and violin plots of distinguished aneuploid and diploid epithelial cells

A: UMAPs of merged epithelial cells from patients 1 (S6), patient 3 (S9 & S10) and patient 4 (S14, S15 & S17). The left UMAP features the cells coloured by patient number. The right UMAP features these same cells coloured by CNV prediction by CopyKAT.

B: Volcano plot of the DEGs uncovered between the diploid and aneuploid epithelial cells. The statistical test used to determine p-values was the Wilcoxon Rank Sum Test.

C: Violin plot of the distributions of the mitochondrial proportion of read counts and number of features. P-values determined using Wilcoxon Rank Sum Test.
3.3.7 Chromosomal CNV heterogeneity across HGSOC patients

The CNV signatures across the genomes reveal some common regions of CNV gain across these 3 HGSOC patients and their 6 sequenced tumours (figure 3.11). The shared CNV gain include regions in chromosomes 1, 7, 8 and 11. There was not strong indication of shared regions of CNV loss across these 3 HGSOC patients. Broadly, the CNV signatures exhibit heterogeneity across the 3 HGSOC patients, but within these patients the CNV signature is similar across their tumour sites. The CNV signatures across patient 4, sample IDs 14,15, and 17 share many regions of CNV gain (chromosomes 1,3,4,6,7 and 11) and loss (chromosomes 4,5,6,9 and 12). This trend is also observed in patient 3, sample IDs 9, sharing regions of CNV gain (chromosomes 3,4,8,9,11,17,19,20 and X) and loss (chromosomes 6,10,13,16,18 and 21) and the same 2 CNV subclones between the primary and peritoneal site. Patient 1, sample ID 6 shows evidence of CNV gain in regions of chromosomes 1,2,7,8,11,12 and X and loss in regions of chromosomes 3,4,6,9,10,13,15 and 19.
Figure 3.11 CNV heatmaps across the transcriptome of diploid and aneuploid cells from the scRNA-seq libraries featuring epithelial cells.

The heatmap is plotted across a single-cell (row) x gene bin (column) axis. The gene bins (featuring a minimum of 25 genes) aggregate to form chromosomal bins. The diploid / background CNV heatmaps were produced from the myeloid cell population from the respective scRNA-seq library (Patient – Sample ID). Grey chromosome bins represent the odd numbered chromosomes (1-21, X) and the black represent the even numbered chromosomes (2-20, Y). The resolution of each chromosome bin is 220Kb, The genome wide resolution for each single-cell is 5Mb.
3.3.8 CNV subclone differential expression analysis

Patient 3 sample IDs 9 and 10 revealed subclones in their CNV signatures across their genomes. Both these subclones are present in the primary (sample ID 9) and peritoneal (sample ID 10) sites, at a proportion of 65% and 28% for CNV subclone 1 across the primary and peritoneal site and 35% and 72% for CNV subclone 2 respectively (Figure 3.12, B). Projected on a UMAP, the cells of these samples separated on CNV subclone identity rather than tumour site, prompting further investigation (Figure 3.12, A). Gene expression analysis reveals that CNV subclone 1 differentially expresses genes such as VIM, SCX, IGLC2 and UQCRH and CNV subclone 2 upregulated genes more associated with malignancy, including MUC16, GPX3, ICAM1 and XIST.
Figure 3.12 Patient 3 CNV subclone distribution across tumour sites and differential gene analysis

A: UMAPs of patient 3 (S9 and S10) aneuploid epithelial cells. The left UMAP features the aneuploid cells coloured by tumour site (S9 = primary site, S10 = metastatic (peritoneal) site). The right UMAP are the aneuploid cells coloured by CNV subclone status, identified from CopyKAT CNV.

B: Bar plot of the distribution of CNV subclones across patient 3 sample sites (S9 and S10).

C: Scatterplot of the log_{10} RNA expression of genes found in patient 3 aneuploid cells. The most differentially expressed genes are labelled on the plot in colour representative of the CNV subclone.
3.3.9 Integrative analysis of individual cell types pooled across HGSOC tumour samples

To take the investigation of the malignant cells further, 18,851 cells across 6 HGSOC tumours were integrated into a combined dataset. The integration means there are common anchors found between the dataset to perform batch correction on the expression data of these cells. This dataset was re-clustered to identify novel sub-types of the malignant cell population. The annotations of these sub-clusters are thought not to be of their cell-type, but rather cell activity or state. Unlike earlier clustering to identify cell-type, these cell states may be more fluid between the sub-clusters. ClusterProfiler was used to annotate these cells, but instead using canonical markers to determine identity, the DEGs found in these subclusters were input into enrichr to conduct gene enrichment analysis against the Molecular Signature Database (MsigDB), ‘Hallmark’ (of cancer) terms and Gene Ontology: (GO) Biological Processes terms to determine broad a biological description of the transcriptome activity of these sub-clusters.

Figure 3.13 features a UMAP of the malignant epithelial cell subclusters after re-clustering. 61% (11,577 cells) of the malignant epithelial cells do not exhibit any DEGs, 22% (4,174 cells) are labelled malignant, 12% (2,167 cells) ciliated, 3% (562 cells) MT/respiratory and 2% (371 cells) Immuno-reactive. The differentially expressed genes of each of the subclusters are submitted for enrichment analysis to determine if there is any recognisable cancerous or biological themes that emerge. Using the MsigDB ‘Hallmark of cancer’ genes list for enrichment analysis, 3 of the subclusters return multiple statistically significant associated terms (figure 3.14). The malignant subcluster returns a variety of terms including hypoxia, apoptosis, Epithelial Mesenchymal Transition (EMT) and P53 Pathway perturbation, a key feature in >90% HGSOC tumours. The immuno-reactive subcluster returns terms associated with the inflammatory response and response to interferons. GO Biological Processes terms were used to attempt to describe the biological functions of these subclusters (figure 3.15). The malignant subcluster returns terms associated with negative regulation of peptidases, due to exaggerated \textit{SLPI}, \textit{SERPINA1}, \textit{SERPINA5}, \textit{GPX3} and \textit{FTH} expression. The ciliated cell subcluster re-emerges, despite most ciliated
cells being excluded from the malignant cell integrative analysis. It is unknown if this is related back to HGSOC being a disease originating from distal epithelium from the fallopian tubes, or an independent phenomenon. The genes distinguishing this cluster are similar to the genes used to distinguish the ciliated cells from epithelial cells in HGSOC scRNA-seq sample IDs 6 and 14, *TPPP3*, *MORN2* and *TUBA1A*. The MT/respiratory subcluster DEGs are associated with ATP metabolic processes and the electron transport chain. Key DEGs of this subcluster include CENPF, TOP2A, CCNB1, CCNA2, UBE2C and MKI67, the latter being used as a marker to indicate proliferation in HGSOC\textsuperscript{155}. Lastly, the immuno-reactive subcluster shares some of the statistically significant terms of the malignant subcluster while being mainly associated with the viral response. Key genes in the immuno-reactive subcluster are the *IFIT* family of genes, *CXCL10*, *OAS* genes and *WARS*. 
The cell labels after re-clustering were assigned based on 2 GSEAs (Gene-Set Expression Analysis). The null-clustered returned no differentially expressed genes using a Wilcoxon Rank Sum test of gene-expression.
Figure 3.14 Dot plot of MSigDb terms associated with aneuploid subclusters from GSEA

The number in the boxes at the bottom of the dot plot corresponds how many genes from the subcluster were involved in the enrichment analysis. All enrichment analysis associations were adjusted for FDR.
Figure 3.14 Dot plot of GO (Gene Ontology): Biological Processes associated with aneuploid subclusters determined from GSEA

The number in the boxes at the bottom of the dot plot corresponds how many genes from the subcluster were involved in the enrichment analysis. All enrichment analysis associations were adjusted for FDR
3.4 Discussion

The characterisation of the HGSOC tumour environment using single cell RNA-seq in this investigation is concordant with previous work using single cell RNA-seq to investigate HGSOC. According to the results from other HGSOC scRNA-seq papers, the TME found in HGSOC scRNA-seq datasets tends to separate into 3 key compartments\textsuperscript{137,156}. As suggested in Figures 3.8 and 3.9, the ciliated and epithelial cells exhibit an association that can be described as the tumour compartment. The lymphoid and myeloid cell populations in the HGSOC scRNA-seq datasets coalesce to form the immune compartment, leaving the fibroblasts, endothelial cells, and mesenchymal cells to form the stromal compartment. This study analyses each of these compartments on a dataset-by-dataset basis to find the individual cell type cluster identities were highly correlated between sequencing batches, patient identity, and specimen origin site (Figures 3.7, 3.8). Furthermore, this study leverages its number of sequenced cells to perform CNV and integrative analyses of the tumour component in an effort to describe this component in a more general way.

The approach taken to cluster the cells using SCTransform from the Seurat package\textsuperscript{64}, modularity optimisation, and ClusterProfiler\textsuperscript{151} made it easy and consistent to apply informative labels to the clusters in the 13 HGSOC libraries (Figures 3.1,3.3-3.4). Notably, however, the 5 libraries from the first sequencing batch did not resemble the later libraries in terms of structure or cell types uncovered. There were clear data quality issues with the first five pilot HGSOC libraries (Figure 3.2), many of the cells were expressing only ribosomal and mitochondrial genes, an indicator of low quality or highly stressed sequenced cells\textsuperscript{157}. Additionally, many of the cells did not express any differentially expressed genes or were not reconcilable to any one cell type when using the ClusterProfiler tool. Recognisable cell types uncovered in these datasets included epithelial cells, ciliated cells myeloid, T cell and B cells. However, due to the limited cell number, the lack of the variety of genes expressed in the clusters that were recognisable, and other ambiguities present in these datasets, it was decided not include cells from these pilot datasets in any analysis involving the later 8 HGSOC datasets.

Clustering the 8 HGSOC scRNA-seq datasets from sequencing batches 2 and 3 revealed 11 distinct cell types that coalesce into 3 TME compartments. Small numbers of platelet and RBC cells were found after clustering in some of these datasets and were removed.
Across 7 of the 8 HGSOC scRNA-seq libraries, the stromal cells (fibroblasts and endothelial cell types) were a minority components. The tumour and immune compartments were the majority of cell types identified in the other HGSOC scRNA-seq datasets. Based on the earlier LGSOC duplicate experiment in Chapter 2, these observations are unlikely to be based solely on the sampling of the input dissociated tumour specimen. However, all scRNA-seq libraries presented in this work are derived from 1g of solid tumour material. There is evidence to suggest that the part of the tumour that is processed and made into a scRNA-seq library does affect the cell populations and proportions uncovered. Zhang et al.\textsuperscript{138} featured 2 HGSOC scRNA-seq datasets from 1 HGSOC ovary specimen. 1 scRNA-seq library was produced from a sample taken from the left side of the ovary and the other scRNA-seq library from a sample taken from the right side of the ovary. While the cell types were uncovered at both sites were the same, the proportions were not. The authors also provided evidence indicating the epithelial cells of these sites were different by virtue of them clustering separately on a shared UMAP. A similar finding is noted from Shih et al.\textsuperscript{137} whereby the authors identified epithelial cells from one patient (their patient 3) clustered distinctly from the rest of the epithelial cells found in the other 13 samples. In our data, using highly variable genes (HVGs), cell types between patients and tumour site clustered together without the use of any specific batch correlation of the RNA-expression data using Nearest Neighbour Voting and correlation analysis (Figures 3.8 and 3.9). When assessing correlation on the entire transcriptome (i.e., not limiting to only HVGs) there is evidence of patient specific clustering of cell types. This highlights an important aspect when interrogating scRNA-seq data, that there are multiple orders of scale within the data to consider when inspecting and analysing the data\textsuperscript{158}.

The ciliated cells of the tumour compartment were mostly non-malignant as identified by CNV analysis. The CNV approach taken here infers the CNV of a target population of cells, which in this work was the tumour compartment of the latest HGSOC datasets; by comparing their genomic profile to selected known normal cells. In this case, the known normal was the myeloid cell population from the respective HGSOC datasets. This is a
fascinating finding, given the mainstream idea of the pathogenesis for HGSOC is that the disease arises from cells migrating from the fallopian tubes\textsuperscript{103}. This finding supports that idea in conjunction with new research, which indicates that the early progenitor cells of the disease in HGSOC do not exhibit the distinct copy number signature of the later stage disease\textsuperscript{159}. The main DEGs of the ciliated cells found in the HGSOC datasets (S6 and S14) were \textit{TPPP3, PIFO, FOXJ1, OVGP1} and \textit{TUBB4B}, the last three being markers for the identification of fallopian cells in publications\textsuperscript{160–162}. These cells also share the key marker genes of HGSOC, found in the epithelial cells in the rest of the datasets such as \textit{MUC16, WFDC2} and \textit{EPCAM} (Figure 3.6). The aneuploid epithelial cells (88\% of all epithelial cells, Table 3.3) were classified as malignant. These cells, in contrast to the diploid epithelial cells, expressed \textit{MUC16} and \textit{WFDC2} at a significantly higher intensity (Figure 3.10 B). The observation of the omental site of patient 4’s (S17) contained 63\% malignant epithelial cells in contrast to their primary (S14) and peritoneal (S15) sites consisting of >95\% malignant cells raise questions about how the TME of different metastatic sites in HGSOC can vary and what this could mean for patient treatment and how different sites respond to treatment.

The integrative analysis of the malignant epithelial cells identified 4 distinct sub-clusters, and 1 null cluster featuring no discernible DEGS. The approach of sub-clustering similar cells is limited by the sequencing depth of the individual cells in these datasets. Of the 4 subclusters identified, 3 of them exhibited DEGs that were associated with hallmark of cancer terms from the molecular signature database (Figure 3.14). This provides some indication as to the mechanisms at play and the biological behaviours of the malignant cells as they were in the patient’s tumour. One of the sub-clusters was labelled ‘malignant’ as it returned the most hallmark of cancer terms, including the P53 pathway perturbation. Another rationale for designating this subcluster as especially malignant was its unique upregulation of key genes such as \textit{GPX3, SLPI}, and PI3. GPX3 belongs to the glutathione peroxidase family, whose function is to protect cells from damage by peroxides. There is published evidence that suggests that GPX3 aids in the progression of OC by altering an hostile extracellular environment to benefit the survival of malignant cells\textsuperscript{163,164}. James et al.\textsuperscript{165} conducted a bioinformatic investigation of 378 HGSOC RNA-seq samples that showed that SLPI, an leukocyte peptidase inhibitor; is correlated with the expression of \textit{WFDC2}, a well-established marker in HGSOC. The same authors identified a \textit{PI3 / WFDC14} association. Although there was no expression of \textit{WFDC14} in our scRNA-seq
datasets, PI3 is another peptidase inhibitor, the fifth most upregulated gene in the ‘malignant’ subcluster of malignant epithelial cells. This work provides further evidence that the tumour cells of HGSOC display biological activity to modify their microenvironment to promote their own survival and proliferation, specifically to undermine harmful peptides and peroxides. Another facet of the tumour cells discovered in this integrative analysis were hallmarks ‘Oxidative phosphorylation’ and ‘Adipogenesis’ and the biological functionality of enhanced cellular energetics of the ‘MT/respiratory’ subcluster. Glasgow et al.,\textsuperscript{166} found that expression of genes related to cell-cycle regulation (CHEK2, E2F7, E2F1 found in the subcluster), mitosis regulators (CENPF, BUB1, AURKB found in the subcluster) and cell division genes (CDC20, CDC25B found in the subcluster) were enriched in HGSOC RNA-seq samples prior to neo-adjuvant chemotherapy. This fits the concept that HGSOC is a highly proliferative disease prior to the initiation of cisplatin or taxane chemotherapy. The ‘MT/respiratory’ subcluster also highly expressed PCNA and MK167 (eleventh most highly expressed gene in the subcluster) which are both recognised markers of proliferation\textsuperscript{155,166}. The other malignant epithelial subcluster to return associations with both hallmark terms and GO: Biological processes was the immuno-reactive subcluster. The hallmark terms were inflammation response and allograft rejection, with GO biological processes suggesting viral defence responses mediated by interferons. The DEGs of this subcluster driving these associations were CXCL10, ICAM1, the IFIT family of genes (IFIT3,IFIT1,IFIT2,IFITM3,IFITM2 and IFITM1), TNFSF10 and GBP1. There is a report of a subset of immunoreactive ovarian tumours that overexpress CXCL10, paradoxically predicting poorer patient survival, despite the function of the chemokine to recruit T cells to the tumour site\textsuperscript{167}. The authors of the report hypothesised that the tumour cells expressed a modified CXCL10 chemokine which serves to antagonise and hinder the recruitment and antitumour immunity of T cells. The immuno-reactive subcluster in addition to CXCL10, overexpressed ICAM1, a glycoprotein that is usually expressed on the surface of endothelial cells or cells of the immune system. Similar to CXCL10, the overexpression of ICAM1 in HGSOC was found to confer a poorer prognosis for patient survival\textsuperscript{168,169}. A review of the role of IFIT genes in cancer suggest
that this family of genes can promote invasion, EMT and resistance to apoptosis and platinum-based chemotherapies (including cisplatin, carboplatin, and oxaliplatin)\textsuperscript{170}.

Many of the works cited in this section investigate the association and role of 1 or 2 genes in cancer. These studies provide compelling evidence for the associations they report, although these associations alone provide only limited clarity on the role these genes play in cancer progression. One of the main advantages of this scRNA-seq data is the possibility of considering and investigating the expression and action of multiple genes (hundreds to thousands) that act along many similar and seemingly unrelated pathways, finding connections that may be overlooked when investigated in isolation. The observation that the malignant cells in this study exhibit distinct patterns involving the expression of multiple genes related to the modification of the extracellular redox environment (the malignant subcluster), proliferation through the cell cycle, cell division and regulation of mitosis (the MT/respiratory subcluster). In addition to activity that suggests the co-opting of, and undermining components of the host-immune system (the immuno-reactive subcluster). Considered together, the integrative analysis of the tumour cells provides a unique vantage point to examine the role these groups of genes and their possible implications for the progression of HGSOC.
Chapter 4: Application of HGSOC single-cell RNA-sequencing signatures onto HGSOC bulk-RNA sequenced data.

4.1 Introduction

4.1.1 Traditional techniques to assay complex tissues.

The structure and function of tissues are emergent properties made possible by the ecosystem of cells and noncellular material (e.g. fibres and collagen) that make up the constituents of tissue. The development of strategies for the in-depth study of the characteristics and composition of tissues has yielded massive insight that has propelled advances in the biological and medical sciences. Traditional techniques include visualisation and characterisation of a tissue’s contents using microscopy with immunohistochemical staining to highlight the cellular constituents of interest based on cell-type markers (in the retina, the developing human kidney and prostate; human bone marrow and in many other tissues). Flow cytometry, a higher performance method that sorts cells based on the fluorescent labelling of cells in a suspension based on antibody staining, can be used to characterise the cells in liquid suspension, routinely used to study immune diseases. These techniques continue to deliver valuable insights that expand our understanding of biology and the disease states investigated with these methods. However, since there is a growing emphasis on precision medicine to leverage information from biological specimens with several forms of accompanying meta-data, that is, quantifications of RNA, DNA and protein content and the clinical characteristics for individual patient resolution. Conventional visualisation methods are limited in their capacity for to scale and plex information. The task of systematically evaluating and annotating visual images of tissues requires extensive clinical training, cross-validation among clinicians, and is
overall very time consuming\textsuperscript{84,183–185}. There is active research into automating these processes using machine learning and AI to address the scalability limitation\textsuperscript{184,185}. This emergent field of research is called digital pathology\textsuperscript{82}. Regarding the characterisation tissue cells using flow cytometry in the era precision medicine, it is constrained by the reliance on a limited number of markers that can be used simultaneously to identify and further characterise the cells of interest\textsuperscript{85,172,186,187} and prerequisite knowledge on what markers to use\textsuperscript{188,189}.

4.1.2 RNA sequencing of complex tissues.

Bulk RNA sequencing (bulk RNA-seq) is over a decade old\textsuperscript{190} and remains a standard technique used to attain insight into the transcriptional variation that exists within various biological entities, for example, disease states between individuals\textsuperscript{53,189,191}. This method can be performed on the RNA extracted from any collection of cells. These cells can be derived from a culture plate/broth, blood, liquid, and solid tissues. Ultimately, the expression data derived from the bulk RNA-seq methodology represents a weighted (by proportion) average of the transcriptomic profiles of the input cellular material from which the RNA was extracted\textsuperscript{53,54}. In the case of tissues, bulk RNA-seq captures the average transcriptomic landscape of the complex ecosystem of cells that constitute the tissue\textsuperscript{53,54,192}. When applied to the study of cancer, these average signals are used to develop expression signatures that can relate to the clinical condition of a patient. In HGSOC, analysis of microarray data has yielded a prognostic molecular classification of patient tumours\textsuperscript{124,125,193,194}, a diagnostic signature in thyroid cancer\textsuperscript{195} and prognostic signatures in lung adenocarcinoma\textsuperscript{196} and neuroblastoma\textsuperscript{197}. There have been several initiatives to expand the utility of these RNA-seq signatures to estimate the cell-type constituents of the sequenced tissue, with the publication of tools such as CIBERSORT\textsuperscript{189}, xCell\textsuperscript{188}, and EPIC\textsuperscript{187}. These tools have been designed for and heavily applied in the study of cancer, primarily with the aim of characterising the complexity of the tumour microenvironment. The utility of understanding the cell-type constituents in the study of tumour biology is to shed light on important facets of the dynamic and heterogeneous tumour micro environment; such as the host immune response to cancer\textsuperscript{140,198}, mechanisms driving malignancy in the tumour\textsuperscript{188,192,199}, the cells of origin\textsuperscript{103,200} and progression of the cancer in the patient\textsuperscript{192}. 
4.1.3 Leveraging single-cell RNAseq resolution to decompose RNA-seq HGSOC solid tissue specimens.

The recent and rapid development of single cell RNA-seq technology has made possible the transcriptomic profiling of hundreds to hundred thousands of individual cells from solid tissues or blood. The transcriptomic profiles of these cells can be clustered by similarity of expression to discern distinct communities of cells, referred to as cell types or cell phenotypes. ScRNA-seq enables a bottom-up approach of elucidating the cellular composition of solid tissues while providing the high-resolution transcriptome profile for each of those cells. In the same vein, bulk-RNA sequencing can be considered as a top-down approach method for elucidating the cellular constituents of solid tissues, with this information essentially encoded in the structure of the data. The issue with the bulk RNA-seq data is that without a key or map (i.e. additional information), decoding the deconvoluted signal into its cell type constituents is impossible. As mentioned earlier, prior to the use of scRNA-seq; tools have been developed which aim to estimate cell-type constituents in bulk RNA-seq samples. They achieve this by leveraging existing knowledge from widely repudiated resources (i.e. TCGA, ENCODE, FANTOM5) and generate their own ‘keys’ in the form of cell-type signature matrixes / expression profiles (the LM22 matrix in CIBERSORT\textsuperscript{189}, the xCell matrix in xCell\textsuperscript{188}, the tumor infiltrating or blood circulating immune expression profiles available for EPIC\textsuperscript{187}). Each of these tools also accommodates the users own curated signature matrix. With the current wide-spread adoption of single cell RNA-seq, producing in-depth and highly relevant cell-type signature matrixes for the cancer of interest/study has been made relatively straightforward. It is now a matter of running a few single cell RNA-seq experiments or using available single cell RNA-seq data on a given cancer type and clustering these data to produce a cell-type by gene matrix and feeding this information into a deconvolution algorithm. This work assesses the performance of four deconvolution algorithms (CIBERSORTx\textsuperscript{172}, MuSiC\textsuperscript{53}, SCDC\textsuperscript{54}, BisqueRNA\textsuperscript{191}) designed to work directly with scRNA-seq data using our own and published scRNA-seq data on HGSOC (Shih 2018\textsuperscript{137} and Zhang 2019\textsuperscript{138}). The goal is
to determine which algorithm and HGSOC scRNA-seq dataset will most accurately deconvolve bulk RNA-seq HGSOC samples where cell type proportions are unknown.

Figure 4.1 Introductory illustration of the concepts for leveraging scRNA-seq information to deconvolve bulk RNA-seq samples

**A:** A tumour specimen used for the either the bulk RNA-seq or scRNA-seq process.

**B:** Is the product of the bulk-RNAseq process. Given multiple patient samples; a bulk mixture matrix ($Y$) is produced.
C: Outlines the product of the scRNA-seq process. A basis matrix (\( \mathbb{B} \)) which characterises the expression of cell-types found in the initial tumour specimen is cell type K by N genes.

D: Further outlines the premise of how the products from B and C are used by deconvolution algorithms to yield the desired information from the bulk mixture samples in B.

4.2 Aims

The aims of the deconvolving HGSOC bulk-RNA sequencing samples using HGSOC scRNA-seq datasets include the following:

1) Establish a framework to assess the accuracy of available algorithms that learn features from scRNA-seq data to deconvolve bulk RNA-seq samples.
2) Arrive at an optimal approach for deconvolving bulk RNA-seq HGSOC samples using scRNA-seq HGSOC material where the true proportions of the bulk RNA-seq material are not known.
3) Evaluate the results of deconvolution of bulk RNA-seq HGSOC samples and assess how this practice improves our understanding of the disease.

4.3 Material and methods

4.3.1 Acquisition and processing of C.Gourley HGSOC tumour samples

HGSOC tumour specimens were acquired from recruited patients diagnosed with FIGO stage III-IV disease. Where possible, multiple tumour sites were acquired from the same patient. The study for which these samples were acquired for was approved by the Lothian Local Research Ethics Committee (approval number SR865/15/ES/0094).

In the case of the patient’s primary debulking surgery, tumour specimens from the primary ovary site and other metastatic sites were excised and placed in MACs tissue storage solution on ice. These specimens were promptly transported from the
operating theatre to a laboratory where the specimens were dissociated using a Miltenyi Biotec tumour dissociation kit on the gentleMACS Octo dissociator. Subsequently, the dissociated cell suspension was passed through a 70µm MACS smart strainer. The next step involved the removal of contaminating red blood cells (RBC); this was achieved by incubating the dissociated cell suspension with equal part RBC lysis buffer for 15 minutes chilled on ice. Afterward, the cells were counted and their viability was assessed using Trypan blue staining with a Countess II automated cell counter. Following this, the live cells were enriched by eliminating the dead cell fraction using a Miltenyi Dead Cell Removal kit and MS columns. Cells were counted again using the Countess II device and if the viability was >70% the cell suspension would be further processed for single-cell sequencing.

4.3.2 Single-cell RNA sequencing

Single cell suspensions passing the >70% viability threshold from dissociation were processed using the 10X Genomics Chromium system. This is a microfluidic platform for single-cell RNA sequencing that captures single cells and encapsulates them in a gel bead-in emulsion (GEM). Each GEM contains a single cell with a 16bp unique barcode, a 10bp UMI and the necessary molecular ingredients for reverse transcription, i.e., sequencing primers and an anchored 30bp oligo-DT. Reverse transcription is performed, followed by the breaking of the GEMs and PCR amplification of the captured transcripts. Following this, the 3’ end libraries were constructed by following the 10X library construction protocol. The finished libraries were sent to Edinburgh Genomics and sequenced on an Illumina NovaSeq system.

4.3.3 Single-cell RNA sequencing data processing.

Following sequencing, the FASTQ files were retrieved from the Edinburgh Genomics delivery server, the MD5 sums were checked, and summary .json files from the sequencer were viewed to make sure sequencing was performed as expected. These FASTQ files for each library were aligned to GRCh38 using Cellranger. After alignment was finished, the 10X .html summaries were checked to assess for the absence of any warnings or errors regarding library construction or alignment.
For each sequenced library; filtered cell, genes, and count matrixes produced from the Cellranger alignment were loaded into R using Seurat version 3. Cells were further filtered if their number of counts, genes, or percentage of counts from mitochondrial genes fall outside of 3 MADs as described in Lun et al., 201. The counts were normalised using SCTransform while regressing mitochondrial expression and predicted cell-cycle state 64. Following normalisation, PCA was performed and the first 30 principal components were used to construct an SNN graph object using the Seurat function ‘FindNeighbors’. Using the embedded SNN graph, clusters were obtained using the Louvain modularity optimisation function and the umap-learn method to produce the UMAPs. The resolution parameters used for the Louvain algorithm ranged from 0.1 to 0.3 in these datasets. From this point on, canonical markers were used to identify the cell types across each dataset.

4 fresh tumour samples from 2 patients were integrated into a single dataset for the purposes of this study (1 primary and peritoneal tumour specimen). This dataset was normalised for technical differences between sequencing runs and patient ID.
Figure 4.2 Composition of the C.Gourley HGSOC scRNA-seq dataset.

Top figure is the UMAP featuring 24,357 cells across 4 HGSOC tumour specimens. Below are stacked bar plots detailing the raw numbers of each cell type in a given dataset and the corresponding proportions beneath.
4.3.4 Shih 2018 HGSOC single-cell dataset

The data from Shih et al.\textsuperscript{137} publication was publicly available from NCBI’s Gene Expression Omnibus, accessed through the GEO series accession number GSE118828.

Table 4.1 Sample accession numbers and name of the samples under the GSE118828 included in the Shih 2018 dataset for this report.

<table>
<thead>
<tr>
<th>Sample Accession Number</th>
<th>Sample Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSM3348320</td>
<td>HG1-M</td>
</tr>
<tr>
<td>GSM3348319</td>
<td>HG1-P</td>
</tr>
<tr>
<td>GSM3348310</td>
<td>HG2F-M</td>
</tr>
<tr>
<td>GMS3348309</td>
<td>HG2F-P</td>
</tr>
<tr>
<td>GMS3348305</td>
<td>HG3-M1</td>
</tr>
<tr>
<td>GMS3348312</td>
<td>HG3-M2</td>
</tr>
<tr>
<td>GMS3348311</td>
<td>HG3-P</td>
</tr>
<tr>
<td>GMS3348314</td>
<td>HG4-P2</td>
</tr>
<tr>
<td>GMS3348315</td>
<td>HG4-P3</td>
</tr>
<tr>
<td>GMS3348316</td>
<td>HG4-P4</td>
</tr>
</tbody>
</table>

Low-grade, benign and normal samples were omitted from inclusion as they were not explicitly high grade serous.

The downloaded samples were .csv files of UMI counts of individual cells (rows) across a set of genes (columns).

As these cells were not annotated, the Seurat SCTransform workflow was applied as described above in 4.3.1.3 for the C.Gourley datasets. Briefly, variable features (N=3,000) were identified in the dataset that were used to conduct PCA. Following from the PCA the first 30 principal components make up the dimension to which graph based clustering is performed. From here an SNN graph is produced. This SNN graph provides parameters for running a UMAP on the cells in the dataset. Clusters were uncovered using the Louvain modularity optimisation algorithm. These clusters were
labelled using canonical markers to positively identify the cell type, i.e., the biological dimension which drives the distinction of clusters. In total 1,803 cells from Shih were retained for training the deconvolution algorithms. The labelling of the cells was done to mirror as accurately as possible what was uncovered in the original publication of this data\textsuperscript{137}.  

*Shih et al 2018*  
1,803 cells

![Graph showing cell types and their distribution](image)
Figure 4.3 Composition of Shih et al. scRNA-seq dataset

Top figure is the UMAP featuring all 1,803 cells across the 10 HGSOC tumour specimens. Below are stacked bar plots detailing the raw numbers of each cell type in a given dataset and the corresponding proportions beneath.

4.3.5 Zhang 2019 HGSOC single-cell dataset

The raw count data from Zhang 2019 were obtained from Zendo by searching for the title of their paper `Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling` and downloading the R object file `sce_hgsc.annotated_final.rds`.

This R object was a single cell experiment (SCE) object which features 4,848 cells from a single HGSOC tumour specimen cut into two sections (left and right) and sequenced separately. For the purposes of this work, the existing cell labels were used when converting the SCE object into a Seurat object by using the Seurat function `As.Seurat()`. This dataset was treated in a manner identical to the C.Gourley in-house scRNA-seq datasets.
Figure 4.4 Composition of Zhang et al. scRNA-seq dataset

Top figure is the UMAP featuring 4,848 cells from 2 HGSOCC tumour specimens, the left and right ovary. Below are stacked bar plots detailing the raw numbers of each cell type in each dataset and the corresponding proportions beneath.
4.3.6 Pseudo-bulk RNA sequence data from Single cell RNA-sequencing data.

Pseudo-bulk RNA samples used to assess the deconvolution algorithms were constructed by summing up the counts of labelled cells (columns) from the single cell dataset and merging them into 1 vector. This new vector represents the convoluted signal of counts attained from sampling various cell types in a bulk RNA-seq sample. The proportions of cell types contributing counts into the new vector are known and retained as meta-data for each pseudo-bulk RNA sample as its `true-proportion`.

There are two ways in which pseudo-bulk RNA samples were generated. One method called `Random-sampling` aimed to generate random proportions pulled from a normal distribution N(µ, σ) for each cell type prior to pseudo-bulk creation. This results in a normal distribution for each cell type across all 50 pseudo-bulk samples for each single-cell RNAseq training dataset.

The second method called `Sc-representative` (Single cell representative) takes a random number of columns from the single cell dataset and produces a pseudo-bulk sample by summing all the counts together. This approach tends to produce pseudo-bulk RNA samples of proportions which represent the single-cell training dataset they were constructed from. The premise behind the two methods of constructing pseudo-bulk samples is to assess the algorithms in their capability to deconvolve mixtures which are both biased in their cell type composition (e.g. relatively abundant for epithelial cell types, but sparse for T-cells) and accurately detect the cell type signatures present in a pseudo-bulk RNA sample from nonbiased mixtures of cell types.

4.3.7 Bulk RNA-sequencing data processing

There are 3 HGSOC cohorts featured in the bulk RNA-seq data. The Australian Ovarian Cancer Study\(^{28}\) (AOC, n = 80) comprises chemo-resistant and relapsed tumours. Treatment naïve, primary ovarian tumours from The Cancer Genome Atlas (TCGA, n = 44)\(^{123}\) and primary tumours from the Scottish High-Grade Serous Ovarian
In total, there was 205 bulk RNA-samples with accompanying clinical annotation that were deconvolved in this work. The RNA sequencing reads were downloaded in aligned BAM format for the AOC and TCGA cohorts from the European Genome/Phenome Archive and Bionimbus Protected Data Cloud, respectively. For the SHGSOC cohort, the reads were retrieved in FASTQ format directly from the sequencing facility. The framework to analyse these data followed the Illumina RNA-seq best practice template. This involved alignment of the reads to the GRCh38 human reference genome with subsequent quality control steps as needed. Next, Salmon quant was used for transcript quantification against the GRCh38 RefSef transcript database, setting the salmon kmer index to 31. Afterwards, for each cohort, the transcript level read-counts were imported into R, converted into TPM and made into expression set objects.

### 4.3.8 Acquisition of patient clinical data for the bulk RNA-sequencing data.

For the AOC and TCGA cohorts, information of donor age, stage at diagnosis, and survival endpoints including the occurrence of an overall survival event, progression free survival event and the survival times are available from the PCAWG project\textsuperscript{202}. The same clinical features for the SHGSOC cohort were retrieved from the Edinburgh Ovarian Cancer Database and the CRUK Clinical Trials unit in Glasgow\textsuperscript{106}.

### 4.3.9 Deconvolution algorithms used

4 deconvolution algorithms were evaluated for their applicability and accuracy in deconvolving HGSOC bulk RNA-seq samples from learned properties of scRNA-seq training datasets.

#### 4.3.9.1 BisqueRNA

BisqueRNA\textsuperscript{191} is a tool designed for efficient deconvolution of bulk RNA-seq samples while accounting for the biases of various single-cell sequencing methodologies. BisqueRNA has the option to improving the estimated cell proportions if the user can provide single-cell or bulk measurements from the same sample. This option was not
explored in this work due to the unavailability of matched single cell and bulk RNA samples.

The framework used by BisqueRNA to decompose bulk RNA-seq samples requires single cell data labelled with their cell type. In the first instance, a reference profile $Z \in \mathbb{R}^{m \times k}$ is produced by averaging read count abundances of $m$ genes for each $k$ cell type in the supplied single cell data. BisqueRNA discards genes that exhibit zero variance, are not expressed and are mitochondrial from the reference profile. The remaining genes in the scRNA-seq data are converted to counts per million reads (CPM) to account for the depth of sequencing between cells. To estimate cell type proportions $P \in \mathbb{R}^{k \times n'}$ with the $n'$ term being individuals from the single-cell data. BisqueRNA generates pseudo-bulk single cell samples as $Y = Z \cdot P$ where $Y \in \mathbb{R}^{m \times n'}$ with the goal of transforming each gene $m$ for all $n$ samples of the observed bulk expression data $X_m \in \mathbb{R}^{n}$ to match the mean and variance of the $m$ genes in the single cell data from $n'$ individuals $Y_m \in \mathbb{R}^{n}$.

The assumptions used for the linear transformation of the input bulk RNA expression samples to be deconvolved are that the mean of $Y_m$ is the true mean of the goal distribution for the transformed $X_m$ and that the variance of $Y_m$ is smaller than the goal distribution, so a shrinkage estimator is applied to minimize the mean squared error to reduce the estimated variance. Following the bulk RNA-seq transformation to $X_{madj}$, nonnegative least-squares (NNLS) regression with a sum-to-one constraint is used to estimate the cell proportions present in the bulk RNA samples. The logic of the framework for deconvolution as explained by the authors of BisqueRNA is to maximize the global linear relationship between all of the genes in the bulk RNA-seq samples and those uncovered in the scRNA-seq data. The aim of the bulk RNA transformation is to recover a one-to-one relationship between the $m$ genes of the bulk $X_{madj}$ and pseudo-constructs of $Y_m$ to optimise the NNLS regression.
4.3.9.2 MuSiC

MUlti-subject SIngle Cell deconvolution (MuSiC)\textsuperscript{53} leverages the single cell expression profiles obtained by multiple subjects to achieve marker gene consistency to guide bulk RNA-seq deconvolution. MuSiC uses W-NNLS (weighted nonnegative least squares) following the below equation to obtain \( P_j = \{p_{jk}^m, k = 1 ..., K \} \) where \( P_j \) is the cell-type proportions of bulk sample \( j \) composed of \( k \) cell types.

\[
Y = C_j \cdot (\alpha_{jg} + \sum_{k=1}^{K} p_{jk}^m S_k \theta_{jg}^k + \epsilon_{jg}) \tag{2}
\]

Equation 2

Given in\textsuperscript{53}

Where \( Y \in \mathbb{R}^{jg} \) represents the bulk tissue relative abundance matrixes of subjects \( j \) across \( g \) genes supplied by the user.

\( C_j \) is a normalising constant that enforces the deconvolution to satisfy the two constraints imposed for the W-NNLS regression, being the non-negativity of any \( p_{jk}^m \) and sum-to-one. \( \alpha_{jg} \) is a gene and subject specific intercept, \( S_k \) is a term denoting cell size of cell type \( k \). The term \( \theta_{jg}^k \) denotes \( k \) cell type specific relative abundance of mRNA for gene \( g \) in subject \( j \) and the final \( \epsilon_{jg} \) is a bulk RNA noise capturing term. The authors\textsuperscript{53} outline in detail the proofs and additional assumptions to derive the terms described in the above equation. Simply put, the objective of the MuSiC framework when performing W-NNLS is to account for two fundamental consistencies across the input scRNA-seq data, these being the `cross-subject` and `cross-cell` consistencies. The goal of reflecting these consistencies is to improve the weights in the W-NNLS. For instance, cross-subject consistency is accounted for by up-weighting genes with low cross-subject variance and penalising (down-weighting) genes with high cross-subject variance. Cross-cell marker consistency is accounted for by up-weighting genes that are highly differential (i.e., exhibit high variance) across cell-types in the single cell data and down-weighting genes which have low variance across cell types. The effect of weighing genes in this manner for the linear regression is to give greater leverage to informative genes while diminishing the leverage of uninformative genes, yielding more accurate deconvolution results. To this aim, MuSiC also uses a recursive tree-guided procedure to aid in the deconvolution of similar and/or rare cell types, which may exhibit collinearity in complex solid tissues in their single cell expression.
profiles. This procedure involves the hierarchical clustering of the cell types in the learned design matrix. Closely related cell-types cluster together, but their proportions can be estimated using cell type consistent genes identified in the earlier W-NNLS. This procedure can be repeated recursively until each cell type is separated from a higher order cluster level with an estimated proportion.

The input datasets include the read counts of the bulk RNA-seq samples to deconvolve ($\mathbf{Y}$) and read counts of cell type ($k$) labelled, multi-subject ($j$) scRNA-seq data ($\mathbf{S}$), effectively the $S_k \theta^k_{ij}$ term in equation 1.

### 4.3.9.3 SCDC

SCDC$^{54}$ presents a framework that enables the use and quality control of multiple single-cell RNA sequence datasets for bulk RNA sample deconvolution. In addition to the deconvolution tool featured in SCDC, the framework has an additional ENSEMBLE method. Briefly, one can combine individual deconvolution results from SCDC and other algorithms (i.e., BisqueRNA, MuSiC) into a final weighted result. For the purposes of assessing individual deconvolution algorithms, the ENSEMBLE feature was not used in this work. Only the deconvolution algorithm in the package was evaluated in this work. SCDC prior to performing W-NNLS regression based deconvolution on the bulk RNA-seq samples implements a quality control step for the clustering of the input scRNA-seq data. This step seeks to remove single cells with low total read count, questionable cell type assignment, and sequencing quality. This is determined by treating each single cell as a bulk sample and running SCDC using the single cell data as both the bulk expression and single cell training data. Each single cell should return estimate proportions that are sparse for all except one entry which is one or near one. Only single cells that return an estimated proportion for their original cell type label of >0.8 (recommended threshold) as determined by SCDC were retained for bulk sample deconvolution estimation.
Concerning the deconvolution as described in their paper, the goal of the deconvolution is to obtain two nonnegative matrices. The basis matrix $\mathbf{B} \in \mathbb{R}^{N \times K}$ with $N$ number of genes across $K$ different cell types; representing the average gene expression for each cell type to be found, and the mixing proportion matrix $\mathbf{P} \in \mathbb{R}^{K \times M}$ featuring $K$ different cell type proportions across $M$ number of bulk RNA mixture samples, the set of values attempted to be estimated. These matrixes are such that

$$\mathbf{Y} = \mathbf{B}\mathbf{P}, \quad \text{and} \quad \tilde{\mathbf{Y}} = \tilde{\mathbf{B}} \ast \tilde{\mathbf{P}}$$

Equation 3

Given in\textsuperscript{54}

Where $\mathbf{Y}$ represents an observed bulk expression matrix $\mathbf{Y} \in \mathbb{R}^{N \times M}$ while $\tilde{\mathbf{Y}}$, $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{P}}$ are estimations of the real, afore described matrices. SCDC uses a W-NNLS framework as used in MuSiC but differs in their approach to create $\mathbf{B}$.

For instance, SCDC scales the single cell count data by a gene specific variance weight $[\sigma_{gkd}^2]$ and subject specific variance weight $[\sigma_{gd}^2]$ to construct a maximal variance weight (MVW) intended to reflect the scRNA-seq data quality. Where $\sigma_{gkd}^2$ reflects cross cell variation for gene $g$ of cell type $k$ within $d$ individual and $\sigma_{gd}^2$ captures the cross subject variation of gene $g$ for individual $d$. The MVM is produced by:

$$\Delta_{gd} = \frac{\sigma_{gd}^2}{\text{median}(\sigma_{gd}^2)}$$

Equation 4

Given in\textsuperscript{54}

Then SCDC scales the raw single-cell read count matrix by $\sqrt{\Delta_{gd}}$ which is designed to diminish the impact of genes with high residuals on the cell type composition estimates within the basis matrix $\mathbf{B}$.

Additional differences between the framework of SCDC and MuSiC are that SCDC can run using single subject scRNA-seq data, whereas MuSiC can not\textsuperscript{53,54}. Finally, the initial single cell quality control step exerts an indirect effect on the basis matrix construction by eliminating the presence of low-quality single cell available for the basis matrix construction.
CIBERSORTx\textsuperscript{172} is an extension of the framework produced by the team who made CIBERSORT\textsuperscript{172,189}. For the purposes of this investigation, 2 of the new modes introduced in CIBERSORTx (group-based expression and high resolution) are not evaluated or discussed here.

To obtain estimates of cell type composition from bulk RNA sequenced samples, CIBERSORTx uses the CIBERSORT support vector regression (SVR) to solve the equation for $\mathbf{F}$

$$\mathbf{B} \times \mathbf{F}_j = \mathbf{M}'_j, 1 \leq j \leq k$$

Equation 5

Given in\textsuperscript{172}

Where $\mathbf{H} \in \mathbb{R}^{N \times C}$ features $n$ number of genes and $c$ distinct cell types. $\mathbf{B}$ is a subset of $\mathbf{H}$ featuring $n_B$ genes that are discriminatory markers of $c$ cell-types. $\mathbf{M} \in \mathbb{R}^{N \times K}$ consists of $n$ number of genes and $k$ mixture GEPs (Gene Expression Profiles), i.e., the convoluted bulk samples. $\mathbf{M}'$ is a subset of $\mathbf{M}$ containing only the $n_B$ genes featured in $\mathbf{B}$. $\mathbf{F}$ is vector made from $f$ estimated cell frequencies of length $c$ by $k$ mixtures $\mathbf{F} = [f_1,f_2,f_3,...,f_k]$. The framework of CIBERSORTx enables the automated learning of what $n_B$ are in $\mathbf{B}$ whereas previously in CIBERSORT these were supplied by the user\textsuperscript{189}. SVR is a machine learning technique that is suited for deconvolution as the method addresses the multicollinearity of deconvolving similar cell-types by utilising a $L_2$-norm penalty function which favours less complex modelling while minimising the variance in weights assigned to closely related cell types. Further, SVR is robust to the noise of the mixture samples and resistant to overfitting owing to the linear $\varepsilon$-insensitive loss function.

S-mode batch correction: When using CIBERSORTx in this investigation, the S-mode batch correction feature was enabled to improve the deconvolution results obtained from droplet based scRNA-seq capture technologies. This batch correction step is designed to account for the excessive technical variation within scRNA-seq data.
obtained from 3’/5’ UMI-based transcript capture methods. Briefly, given a basis-matrix of \( \mathbb{B} \in \mathbb{R}^{N \times K} \) and a set of single cell reference profiles \( \mathbb{S} \in \mathbb{R}^{N \times S} \) where \( s \) represents the gene expression of a single-cell across \( n \) genes. A series of artificial \( k \) mixtures \( \mathbb{M}^* \) from single cell GEPs within \( \mathbb{S} \) are created. The mixing coefficients \( \mathbb{F}^* \) are known, sampled from a normal distribution \( \mathcal{N}(\mu, \sigma) \) for each cell type. These artificial \( k \) mixtures of \( \mathbb{M}^* \) are normalised in TPM space and are adjusted using ComBat\(^{203} \) to yield \( \mathbb{M}^{*\text{adj}} \). These \( \mathbb{F}^* \) and \( \mathbb{M}^{*\text{adj}} \) are used as input for NNLS to reconstruct cell type coefficients for each gene in \( \mathbb{B} \) to make \( \mathbb{B}^{\text{adj}} \). Finally, \( \mathbb{B}^{\text{adj}} \) is used in the equation to estimate \( \mathbb{F} \) from the supplied \( \mathbb{M} \) to be deconvolved.

### 4.3.10 Assessment metrics for deconvolution evaluation

To quantify the performance of the deconvolution algorithms on the pseudo-bulk RNA samples with known proportions the following metrics are used to measure accuracy and error.

A) Measures accuracy

I) Pearson correlation: \( R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \)

B) Measures error

I) Root mean squared deviation: \( \text{RMSD} = \sqrt{\frac{\sum(p - \bar{p})^2}{n}} \)

II) Mean absolute deviation: \( \text{mAD} = \frac{\sum|p - \bar{p}|}{n} \)

III) 95\% confidence interval of \( R = (x \pm \mu)/\sigma \)
Figure 4.5 Deconvolution evaluation framework

A: The leftmost panel is a heatmap of the simulated proportions from the pseudo-bulk samples, these values are the target simulated (i.e., ‘true’/‘real’) values for the purpose of evaluating the deconvolution algorithms.

B: These heatmaps are the estimated proportions from the deconvolution algorithms (BisqueRNA, CIBERSORTx, MuSiC, SCDC) and a non-weighted NNLS regression (from the MuSiC package). The closer these heatmaps in this panel resemble the target heatmap in panel A, the better the algorithm performance. Algorithm performance is quantified through measurements captured in panels C and D.

C: Row-wise correlation assesses sample-specific mixture deconvolution accuracy. As each sample is only deconvolved once by each algorithm, measurements to quantify the error include mean absolute deviation (mAD) and the root mean squared deviation (RMSD).

D: Column-wise correlation values are used to assess cell-type specific accuracy for each algorithm. As each cell type being deconvolved is being estimated through 50 pseudo-bulk samples, confidence intervals for cell type specific R for each algorithm can be calculated.
4.3.11 Statistical Analysis

All statistical analysis was carried out using R version 4.0.3 ‘Bunny-Wunnies Freak Out’ (R Foundation for Statistical Computing, Vienna, Austria). The survival analysis consisted of building an initial multivariable model using Cox proportional hazards regression, stratifying by cohort (AOC, TCGA, SHGSOC) and adjusting for donor age at diagnosis and stage of disease. Using this initial model, inferred features from deconvolution analysis were added in a step-wise manner. At first, this involved the simple deconvolution proportions (Tumour, Immune and Stromal) followed by the addition of the higher resolution deconvolution results, i.e., the cell type proportions. Collinearity was assessed between the simple deconvolution proportions and the cell type proportions, preferring to retain the feature which contributed more information to the overall model.

4.4 Results

4.4.1 Pseudo-bulk data and algorithm evaluation

For each single cell dataset available in the investigation, 2 sets of pseudo-bulk samples were produced. One set, produced by `random-sampling`, consisted of pseudo-bulk samples with cell type proportions pulled from a normal distribution $N(\mu, \sigma)$. This is designed to test how well each prospective algorithm can deconvolve the cell type signal in a manner independent of sample composition bias. The second set of pseudo-bulk samples were produced in a way which has cell type proportions that reflect the proportions found in the scRNA-seq data. Assessing the algorithms on these pseudo-bulk samples tests how well the algorithms can deconvolve mixtures, which represent the complexity of solid tissues, which may be abundant in with some cell types (e.g., epithelial cells) but sparse in others (e.g., endothelial, B-cells, etc.).
4.4.1.1 Composition of the single-cell pseudo-bulk RNAseq samples

In the normally distributed pseudo-bulk RNA-seq datasets, 50 bulk RNA-seq samples were generated for each scRNA-seq dataset (C.Gourley, Shih et al and Zhang et al). These pseudo-bulk RNAseq samples are comprised of reads in approximately equal proportions of all the cell types featured in their respective scRNA-seq datasets (see: figures 4.2 to 4.5). The left-hand side of Figures 4.6 to 4.8 illustrate this approach produced pseudo-bulk data with the intended proportions. The `single cell representative` pseudo-bulk RNAseq samples were more heterogeneous in their composition of cell types, reflecting the uneven distribution of cell types uncovered in the scRNA-seq datasets (figures 4.6 to 4.8, right-hand side).
Figure 4.6 Boxplots of the simulated proportions of 100 C.Gourley pseudo-bulk RNA-seq samples.

Left facet are the proportions of the cell-types of the pseudo-bulk RNA-seq samples (N=50) derived from a normal distribution of cells in the C.Gourley scRNA-seq dataset. Right facet are these proportions sampled from a distribution representative of the C.Gourley scRNA-seq dataset.
Figure 4.7 Boxplots of the simulated proportions of 100 Shih et al pseudo-bulk RNAseq samples.

Left facet are the proportions of the cell-types of the pseudo-bulk RNAseq samples (N=50) derived from a normal distribution of cells in the Shih et al scRNA-seq dataset. Right facet are these proportions sampled from a distribution representative of the Shih et al scRNA-seq dataset.
Figure 4.8 Boxplots of the simulated proportions of 100 Zhang et al pseudo-bulk RNAseq samples.

Left facet are the proportions of the cell-types of the pseudo-bulk RNAseq samples (N=50) derived from a normal distribution of cells in the Zhang et al scRNA-seq dataset. Right facet are these proportions sampled from a distribution representative of the Zhang et al scRNA-seq dataset.
4.4.1.2 Evaluation of 4 published deconvolution algorithms against pseudo-bulk RNA-seq samples generated from C.Gourley scRNA-seq data

Evaluating the performance of the 4 contending deconvolution algorithms in a sample-wise manner shows that SCDC and MuSiC outperform the other algorithms when testing against the `single cell representative` and normally distributed pseudo-bulk RNA-seq samples (Figure 4.9 A and B, for values, see Table 4.2), respectively. Considering the single cell representative pseudo-bulk deconvolution, SCDC scored the highest sample-wise Pearson correlation (R = 0.86) and the lowest median absolute deviation (0.052) and the root mean squared difference (0.07). This means that on average, across the 50 single-cell representative pseudo-bulk samples, SCDC deconvolution estimates were the closest to the simulated values (Figure 4.9 A).

MuSiC performed best when the pseudo-bulk samples were compiled from a normal distribution of single cells from the C.Gourley scRNA-seq dataset. The Person correlation for this test was the highest of all tests on C.Gourley pseudo-bulk RNA-seq samples at R = 0.97 (Table 4.2).

Taking into account cell type deconvolution performance MuSiC outperformed SCDC in both pseudo-bulk RNAseq datasets (`single cell representative` and normally distributed, see Figure 4.10). For the `single cell representative` pseudo-bulk data, MuSiC achieved the highest average Pearson correlation for 8 cell types (R = 0.7), SCDC scored (R = 0.67). The main difference in these scores is the trade-off in accuracy and error made regarding the CD4-Tcells and CD8-Tcells specifically. MuSiC scored the lowest correlation for CD4-Tcells (R = 0.36) while the next highest, CIBERSORTx achieved (R = 0.62) and SCDC (R = 0.72). While all algorithms made errors, i.e., attained a negative Pearson correlation when estimating CD8-Tcell proportion in mixtures. MuSiC scored (R = -0.12) and SCDC scored (R = -0.46) with CIBERSORTx and BisqueRNA attaining marginally worse scores (R = -0.55). All algorithms, with the exception of BisqueRNA, performed well in their cell-type
estimations on the normally distributed pseudo-bulk RNAseq samples.

Overall, MuSiC performed best when using the C.Gourley scRNA-seq data to deconvolve pseudo-bulk mixtures, and would be, based on these test; the recommended algorithm to deploy when using C.Gourley scRNA-seq data on mixtures of unknown proportions.
Figure 4.9 Heatmap of deconvolution estimates of C.Gourley scRNA-seq pseudo-bulk data by 5 scRNA-seq to bulk RNA-seq deconvolution algorithms.

A: Deconvolution estimates of C.Gourley scRNA-seq pseudo-bulk samples resembling the cell type proportion found in C.Gourley scRNA-seq data.

B: Deconvolution estimates of C.Gourley scRNA-seq pseudo-bulk samples with cell type proportions normally distributed from C.Gourley scRNA-seq data.
Table 4.2 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on C.Gourley single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples).

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>mAD</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BisqueRNA</td>
<td>0.82</td>
<td>0.093</td>
<td>0.127</td>
</tr>
<tr>
<td>CIBERSORTx</td>
<td>0.79</td>
<td>0.067</td>
<td>0.085</td>
</tr>
<tr>
<td>MuSiC</td>
<td>0.77</td>
<td>0.07</td>
<td>0.102</td>
</tr>
<tr>
<td>SCDC*</td>
<td>0.86</td>
<td>0.052</td>
<td>0.07</td>
</tr>
<tr>
<td>NNLS</td>
<td>0.68</td>
<td>0.099</td>
<td>0.145</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>mAD</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BisqueRNA</td>
<td>0.69</td>
<td>0.116</td>
<td>0.165</td>
</tr>
<tr>
<td>CIBERSORTx</td>
<td>0.88</td>
<td>0.037</td>
<td>0.046</td>
</tr>
<tr>
<td>MuSiC*</td>
<td>0.97</td>
<td>0.014</td>
<td>0.018</td>
</tr>
<tr>
<td>SCDC</td>
<td>0.76</td>
<td>0.036</td>
<td>0.048</td>
</tr>
<tr>
<td>NNLS</td>
<td>0.91</td>
<td>0.099</td>
<td>0.145</td>
</tr>
</tbody>
</table>

*Best performing algorithm for that specific pseudo-bulk deconvolution. ‡ In reference to figure 4.5 C: deconvolution evaluation framework. The row-wise estimates.
Figure 4.10 Line plots of the Pearson correlation (+ 95% CI) recorded in 4 deconvolution algorithms for each cell-type in C.Gourley scRNA-seq bulk RNA-seq data.

A: Cell-wise correlation from each deconvolution algorithm against single cell representative simulated pseudo-bulk generated from C.Gourley scRNA-seq data.

B: Cell-wise correlation from each deconvolution algorithm against normally distributed simulated pseudo-bulk generated from C.Gourley scRNA-seq data.
4.4.1.3 Evaluation of 4 published deconvolution algorithms against pseudo-bulk RNA-seq samples generated from Shih et al. scRNA-seq data

CIBERSORTx performed the most accurate deconvolution of pseudo-bulk RNAseq samples produced from the published scRNA-seq data of Shih et al.\textsuperscript{137} (Figure 4.11 A and B). Sample-wise, for both `single cell representative` and normally distributed pseudo-bulk RNAseq samples (values in Table 4.3), CIBERSORTx scored the highest Pearson correlation ( $R = 0.97$ and $R = 0.92$) and the lowest overall error in terms of mAD (0.027 and 0.023) and RMSD (0.042 and 0.028). MuSiC performed comparably well sample-wise to CIBERSORTx, achieving high Pearson correlations ($R = 0.94$ and $R = 0.9$) and low errors.

CIBERSORTx outperformed MuSiC when evaluating cell type-wise correlation in the Shih et al. pseudo-bulk RNA-seq data (Figure 4.12). CIBERSORTx attained the highest Pearson correlation across the cell types in both the `single cell representative` and normally distributed pseudo-bulk RNA-seq samples ($\bar{R} = 0.97$ and $\bar{R} = 0.98$). No other algorithm was able to perform on-par with CIBERSORTx regarding the cell-type Pearson correlation from the normally distributed pseudo-bulk RNAseq samples.
Figure 4.11 Heatmap of deconvolution estimates of Shih et al. scRNA-seq pseudo-bulk data by 5 scRNA-seq to bulk RNA-seq deconvolution algorithms

A: Deconvolution estimates of Shih et al. scRNA-seq pseudo-bulk samples resembling the cell type proportion found in Shih et al. scRNA-seq data.

B: Deconvolution estimates of Shih et al. scRNA-seq pseudo-bulk samples with cell-type proportions normally distributed from Shih et al. scRNA-seq data.
Table 4.3 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on Shih et al. single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples).

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>mAD</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BisqueRNA</td>
<td>0.92</td>
<td>0.041</td>
<td>0.066</td>
</tr>
<tr>
<td>CIBERSORTx*</td>
<td>0.97</td>
<td>0.027</td>
<td>0.042</td>
</tr>
<tr>
<td>MuSiC</td>
<td>0.94</td>
<td>0.042</td>
<td>0.066</td>
</tr>
<tr>
<td>SCDC*</td>
<td>0.53</td>
<td>0.01</td>
<td>0.163</td>
</tr>
<tr>
<td>NNLS</td>
<td>0.88</td>
<td>0.048</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Pearson</th>
<th>mAD</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>BisqueRNA</td>
<td>0.65</td>
<td>0.068</td>
<td>0.098</td>
</tr>
<tr>
<td>CIBERSORTx*</td>
<td>0.92</td>
<td>0.023</td>
<td>0.028</td>
</tr>
<tr>
<td>MuSiC</td>
<td>0.9</td>
<td>0.022</td>
<td>0.028</td>
</tr>
<tr>
<td>SCDC</td>
<td>0.24</td>
<td>0.103</td>
<td>0.132</td>
</tr>
<tr>
<td>NNLS</td>
<td>0.85</td>
<td>0.026</td>
<td>0.034</td>
</tr>
</tbody>
</table>
Figure 4.12 Line plots of the Pearson correlation (+ 95% CI) recorded in 4 deconvolution algorithms for each cell-type in Shih et al. scRNA-seq pseudo-bulk RNA-seq data.

A: Cell-wise correlation from each deconvolution algorithm against single cell representative simulated pseudo-bulk generated from Shih et al. scRNA-seq data.

B: Cell-wise correlation from each deconvolution algorithm against normally distributed simulated pseudo-bulk generated from Shih et al. scRNA-seq data.
4.4.1.4 Evaluation of 4 published deconvolution algorithms against pseudo-bulk RNA-seq samples generated from Zhang et al scRNA-seq data

MuSiC cannot estimate the cell type proportions from scRNA-seq training data without $n > 1$ patients. As both scRNA-seq datasets in Zhang et al$^{138}$. are derived from the left and right section of the same patient’s HGSOC ovary specimen, MuSiC was not ran on these data.

The three applicable algorithms (CIBERSORTx, BisqueRNA and SCDC) each attained high sample-wise Pearson correlations for the `single cell representative` pseudo-bulk RNA-seq samples ($R = 0.99$, $R = 0.98$, and $R = 0.94$ respectively, table 4.4). These correlations dropped considerably when tested on pseudo-bulk RNA-seq samples composed of normally distributed cells ($R = 0.54$, $R = 0.18$, and $R = 0.66$ respectively, Table 4.4). Figure 4.13 illustrates the contrast of how concordant estimation of each of the algorithms were for the `single cell representative` pseudo-bulk RNA-seq data (Figure 4.13 A), then the dissimilarity of the algorithm estimations were for the normally distributed pseudo-bulk RNA-seq samples (Figure 4.13 B).

Cell type estimation accuracy was mixed for the algorithms on the Zhang et al. scRNA-seq pseudo-bulk RNA-seq samples (Figure 4.14). SCDC performed best overall, attaining an $\bar{R} = 0.75$ for all cell-types. BisqueRNA achieved $\bar{R} = 0.68$ and CIBERSORTx $\bar{R} = -0.003$. 
Figure 4.13 Heatmaps and row-wise summaries of the estimated proportions from 3 deconvolution algorithms using pseudo-bulk RNA-seq samples produced from Zhang et al. scRNA-seq data.

A: Heatmap of estimated proportions of pseudo-bulk sample produced via ‘single cell representative’ sampling of the Zhang et al. scRNA-seq dataset.

B: Heatmap of the estimated proportions of pseudo-bulk sample provided via random sampling of the Zhang et al. scRNA-seq dataset.
Table 4.4 Tabulated deconvolution evaluation‡ of 5 scRNA-seq to bulk RNA-seq deconvolution algorithms on Zhang et al. single cell pseudo-bulk RNA-seq dataset (N=100 pseudo-bulk samples)

<table>
<thead>
<tr>
<th>Method</th>
<th>Single cell representative pseudo-bulk RNA-seq</th>
<th>Normally distributed pseudo-bulk RNA-seq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pearson</td>
<td>mAD</td>
</tr>
<tr>
<td>BisqueRNA</td>
<td>0.98</td>
<td>0.026</td>
</tr>
<tr>
<td>CIBERSORTx*</td>
<td>0.99</td>
<td>0.019</td>
</tr>
<tr>
<td>SCDC*</td>
<td>0.94</td>
<td>0.019</td>
</tr>
</tbody>
</table>
Figure 4.14 Line plots and summary tables of the column-wise Pearson correlation from each deconvolution algorithm with 95% confidence intervals for each cell type on the Zhang et al. pseudo-bulk samples and training data.

A: Line plots with 95% confidence interval of the correlation for the estimates of every Zhang et al. cell type for each deconvolution algorithm from the `single cell representative` pseudo-bulk data.

B: Line plots with 95% confidence interval of the correlation for the estimates of every Zhang et al. cell type for each deconvolution algorithm from the ‘normally distributed’ pseudo-bulk data.
Based on the assessments conducted above, the algorithm and training data that was selected to estimate the composition of real HGSOC bulk RNA-seq samples where the true proportions are not known was the CIBERSORTx algorithm with Shih et al. scRNA-seq data. This combination of algorithm and training data performed the best when deconvolving pseudo-bulk RNA-seq samples with known mixture proportions sampled in a way that represents the scRNA-seq dataset ($R = 0.97$, Table 4.3) and achieved consistently higher correlations across each cell type in pseudo-bulk samples constructed in a representative and normally distributed manner than the other algorithms (Figure 4.12).

In other cases where different algorithms and dataset combinations performed better on deconvolving in a specific instance than CIBERSORTx and Shih et al., they generally performed worse in other regards. For example, the sample-wise deconvolution of `single cell representative` pseudo-bulk RNA-seq samples from Zhang et al. scRNA-seq data using CIBERSORTx and BisqueRNA were $r = 0.99$ and 0.98 respectively (Table 4.4), marginally more accurate than CIBERSORTx and Shih et al. for that same test ($R = 0.97$, Table 4.3). However, in the normally distributed pseudo-bulk RNA-seq samples, CIBERSORTx and BisqueRNA performed considerably worse for the Zhang et al. datasets than CIBERSORTx for Shih et al. ($R = 0.54$ and $R = 0.16$, Table 4.4) compared to ($R = 0.92$, Table 4.3).

Therefore, the CIBERSORTx algorithm and Shih et al. scRNA-seq data forming the basis matrix ($\mathbb{B}$, see Figure 4.15) was used to deconvolve HGSOC bulk RNA-seq samples from clinical cohort studies to evaluate whether there are any latent associations between the estimated composition of the tumour microenvironment and clinical outcomes.

### 4.4.2 Deconvolution estimates on 206 HGSOC RNAseq samples data where the true proportions are not known.

206 HGSOC RNAseq samples in total were deconvolved using CIBERSORTx and Shih et al. scRNA-seq data (Figure 4.16). The basis matrix used in the deconvolution is featured in Figure 4.15 which informs which genes and at what weights cell types are estimated in a mixture (bulk RNA-seq) sample.
The TCGA HGSOC samples scored significantly higher tumour cellularity as determined by using ESTIMATE\textsuperscript{204} on average compared to the other 3 clinical cohorts. The average tumour cellularity in the TCGA cohort was 81%, compared to the 63% in the AOC cohort (Wilcoxon rank sum test: \( P_{\text{adj}} = 1.6 \times 10^{-7} \)), 61% in the SHGSOC cohort (Wilcoxon rank sum test: \( P_{\text{adj}} = 8.2 \times 10^{-7} \)) and 53% in the MDA cohort (Wilcoxon rank sum test: \( P_{\text{adj}} = 2.8 \times 10^{-6} \)). All Wilcoxon tests were corrected for multiple hypothesis testing using the Bonferroni-Holm p-value adjustment method. No other comparisons in the clinical cohorts yielded statistically significant differences between mean tumour cellularity (Figure 4.17, top-left cellularity panel).

The samples from the Scottish HGSOC cohort (SHGSOC) scored significantly higher in the deconvolved tumour cell compartment relative to the other clinical HGSOC cohorts (Figure 4.17, top-right tumour panel). This tumour score is a linear combination of the estimated cell type proportions of epithelial, HG3-epithelial and fallopian cells from CIBERSORTx using Shih et al. scRNA-seq data as training data. The average tumour proportion in the SHGSOC RNA-seq samples was 52%, greater than the average tumour proportion of the other HGSOC cohorts. The AOC cohort average 45% (Wilcoxon rank sum test: \( P_{\text{adj}} = 4.8 \times 10^{-5} \)), the TCGA average 47% (Wilcoxon rank sum test: \( P_{\text{adj}} = 0.008 \)) and the MDA cohort average of 41% (Wilcoxon rank sum test: \( P_{\text{adj}} = 0.00023 \)). No other comparisons of tumour proportions across the clinical cohorts yielded statistically significant differences.

The constituent of the tumour proportion consists of epithelial, HG3 epithelial, and fallopian cell type proportions (Figure 4.18). Each of these was analysed using a one-way ANOVA, followed by TukeyHSD (Honest Statistical Difference) to detect where the variance was present within these cell type proportions. Considering the proportions of the epithelial cells, the one-way ANOVA indicated variance between HGSOC cohorts (\( Df = 3, p = 9.61 \times 10^{-6} \)), Tukey pairwise comparison indicates that variance is from the cohort pairs SHGSOC-AOC (+6.4%, 95% CI: 2.3-11%, \( P_{\text{adj}} = 0.0004 \)) and SHGSOC-MDA (+10%, 95% CI: 4.6-16%, \( P_{\text{adj}} = 0.00003 \)). A one-way ANOVA of the HG3 Epithelial cell type indicates variance between HGSOC cohorts 176
(Df = 3, P = 0.002), the Tukey pairwise comparison reveals that this variance is between the cohort pairs AOC-MDA (+3%, 95% CI: 0.8-7%, $P_{adj} = 0.006$) and SHGSOC-MDA (+4%, 95% CI: 1.2-7.3%, $P_{adj} = 0.001$). Lastly, the one-way ANOVA of the fallopian proportions also indicates variance between HGSOC cohorts (Df = 3, $p = 9.9*10^{-10}$), where the MDA cohort exhibits statistically higher proportions ($\bar{\mu} = 9.2\%$) of the fallopian cell type in pairwise comparisons with the other cohorts. MDA-AOC (+3.6%, 95% CI: 2.2-5.1%, $P_{adj} = <0.0001$), MDA-SHGSOC (+3.7%, 95% CI: 2.2-7.3%, $P_{adj} = <0.0001$) and MDA-TCGA (+3.6%, 95% CI: 1.6-5%, $P_{adj} = <0.0001$).

The AOC cohort featured RNA-seq samples that had the highest proportion of immune cells compared to the other HGSOC clinical cohorts (Figure 4.17, bottom-right Immune panel). The AOC cohort had an average immune proportion of 34%, compared to the SHGSOC cohort immune average of 25% ($P_{adj} = 5.4*10^{-14}$), The TCGA cohort immune average of 31% ($P_{adj} = 0.004$) and the MDA cohort immune average of 27% ($P_{adj} = 1.2*10^{-5}$). The T-cell proportions exhibit many statistically significant differences in their means between HGSOC cohorts (Figure 4.18). The one-way ANOVA for the T-cells returns a p-value of $<2*10^{-16}$. The pairwise comparisons consistently indicate that the AOC cohort has higher proportions of T-cells on average ($\bar{\mu} = 17\%$) than the rest of the HGSOC cohorts. AOC-SHGSOC (+6.5%, 95% CI: 5.2-7.7%, $P_{adj} = <0.0001$), AOC-TCGA (+3.6%, 95% CI: 2-5.2%, $P_{adj} = <0.00001$), AOC-MDA (+7%, 95% CI: 5%-8.4%, $P_{adj} = <0.0001$) and TCGA-SHGSOC (+2.8%, 95% CI: 1.2-4.5%, $P_{adj} = <0.0001$), TCGA-MDA (+3%, 95% CI: 1-5%, $P_{adj} = 0.0007$). There were no statistical difference found between the SHGSOC and MDA cohorts in their T-cell proportions.

The B-cell proportions also exhibit wide a variation in means across the HGSOC cohorts (Figure 4.18). The one-way ANOVA P value = 2.4*10-9. Pairwise comparisons indicate that MDA and AOC have the highest proportion of B-cells, respectively (MDA B-cell $\bar{\mu} =10\%$, AOC B-cell $\bar{\mu} =8\%$). MDA-AOC (+2.5%, 95% CI: 0.8-4.2%, $P_{adj} = 0.0008$), MDA-SHGSOC (+4.5%, 95% CI: 2.7-6.2%, $P_{adj} = <0.0001$), MDA-TCGA (+3.5%, 95% CI: 1.5-5.5%, $P_{adj} = <0.0001$), AOC-SHGSOC (+2%, 95% CI: 0.7-3.2%, $P_{adj} = 0.004$).

The HGSOC RNA-seq samples of the MDA cohort consist of a higher stromal content compared to the RNA-seq samples of the other HGSOC cohorts (Figure 4.17, bottom-left Stromal panel). The average stromal proportion in the MDA cohort was 31%, in
contrast to the stromal proportions of the AOC cohort at 21% (Wilcoxon rank sum test: $P_{\text{adj}} = 2.7 \times 10^{-5}$), the SHGSOC cohort at 23% (Wilcoxon rank sum test: $P_{\text{adj}} = 0.001$) and the TCGA cohort at 22% (Wilcoxon rank sum test: $P_{\text{adj}} = 0.0009$). The cell type proportions that constitute the stromal content are the endothelial cell, primary and metastatic fibroblast cells. A one-way ANOVA on the primary fibroblasts across the HGSOC cohorts suggests statistically significant variance between HGSOC cohorts ($\text{Df} = 3$, $P = 0.002$). Tukey pairwise comparisons across the cohorts reveal that this variance is from the MDA cohort RNA-seq samples exhibiting higher proportions of primary fibroblasts ($\bar{\mu} = 21\%$) compared to the AOC cohort (MDA-AOC (+7.2%, 95% CI: 2.1-12.3%, $P_{\text{adj}} = 0.001$) and the SHGSOC cohort (MDA-SHGSOC (+7.3%, 95% CI: 1.3 - 20%, $P_{\text{adj}} = 0.002$) cohorts. A one-way ANOVA of the metastatic fibroblast proportion does return a marginal statistically significant result ($p = 0.45$). However, the TukeyHSD does not return any statistically significant pairwise comparisons of the proportions of these cells. Lastly, the one-way ANOVA on the endothelial cell proportions indicates variance between HGSOC cohorts ($\text{Df} = 3$, $p = 9.63 \times 10^{-16}$). Pairwise comparisons suggest that this variance is from the MDA RNA-seq samples having higher endothelial cell proportions ($\bar{\mu} = 7\%$) compared to the rest of the cohorts. MDA-AOC (+3.7%, 95% CI: 2.6-4.7%, $P_{\text{adj}} = <0.0001$), MDA-SHGSOC (+2.2%, 95% CI: 1.1-3.2%, $P_{\text{adj}} = <0.0001$), MDA-TCGA (+2.9%, 95% CI: 1.6-4.1%, $P_{\text{adj}} = <0.0001$) and SHGSOC-AOC (+1.5%, 95% CI: 0.7-2.3%, $P_{\text{adj}} = <0.0001$).

The deconvolution estimates of the tumour proportion are correlated with the tumour-cellularity estimates from ESTIMATE in each HGSOC cohort except for the MDA cohort. The Pearson correlation of the tumour proportion and tumour cellularity of each HGSOC cohort was as follows. AOC ($R = 0.47$, $p = 1.1 \times 10^{-5}$), SHGSOC ($R = 0.52$, $p = 5.6 \times 10^{-6}$), TCGA ($R = 0.46$, $p = 0.008$) and MDA ($R = 0.29$, $p = 0.16$). Across all the HGSOC cohorts, the correlation was $R = 0.4$, $p = 2.9 \times 10^{-9}$ (Figure 4.19). The estimates of immune proportion did not show a correlation with the tumour cellularity from ESTIMATE at the global level ($R = -0.09$, $p = 0.2$, Figure 4.20), although at the individual cohort level, negative correlations between immune proportion and tumour
cellularity were observed in the SHGSOC (R = -0.34, p = 0.004) and MDA (R = -0.47, p = 0.01) cohorts. The estimates of stromal proportion and tumour cellularity was negatively correlated globally across all HGSOC cohorts (R = -0.35, P = 1.7*10^{-7}, Figure 4.21) and this trend was exhibited in each of the individual HGSOC cohorts except the MDA cohort (R = 0.006, p = 0.9). AOC cohort (R = -0.48, p = 5.2*10^{-6}), SHGSOC cohort (R = -0.37, p = 0.001) and the TCGA cohort (R = -0.45, p = 0.01).

Figure 4.15 Basis matrix (\(B\)) of Shih et al. scRNA-seq data constructed by CIBERSORTx

Dendrogram on the left of the matrix are genes that cluster which are differentially expressed in each of the cell types (columns). The components of the TME (Immune; blue, Tumour; red and Stromal; green) are represented in the coloured column-wise dendrogram.
Figure 4.16 Simplified deconvolution decomposition of 206 deconvolved HGSOC RNA-seq samples and their cellularity* estimates

Simplified deconvolution refers to the representation of the 9 distinct estimated cell types into 3 main compartments. These compartments are split into i) Tumour = linear combination of Epithelial, HG3.Epithelial and Fallopian cell proportions. ii) Stromal = linear combination of Primary, Metastatic and Endothelial cell proportion and iii) Immune = linear combination of T-cell, B-cell and Myeloid cell proportions.

*Cellularity estimates are of tumour cellularity in a sequenced RNA-seq sample. Tool used to estimate tumour cellularity (or tumour purity) is ESTIMATE.
Figure 4.17 Boxplot of simplified deconvolution estimates of 206 deconvolved HGSOC RNA-seq samples from 4 clinical cohorts.

MHTC: Multiple Hypothesis Test Correction, using Bonferroni-Holm correction. Cellularity estimate were made using the ESTIMATE algorithm.
Figure 4.18 Boxplot of deconvolution estimates of 9 cell types found in Shih et al. scRNA-seq data in 206 deconvolved HGSOC RNA-seq samples from 4 clinical cohorts.
Figure 4.19 Scatter plot of tumour proportion against estimated cellularity of 206 deconvolved HGSOC RNA-seq samples
Figure 4.20 Scatter plot of immune proportion against estimated cellularity of 206 deconvolved HGSOC RNA-seq samples
4.4.3 Survival analysis of HGSOC patients

The approach used to investigate the role of inferred cell type proportions on survival in HGSOC was to build two independent models, one for OS (Overall Survival) and the other for PFS (Progression Free Survival). These models are where cohort and patient specific variables are controlled and adjusted prior to introducing the deconvolved cell type proportions in the survival analysis. Both OS and PFS models stratified survival/relapse time, by cohort and adjusted for patient age and stage. For the analysis, FIGO stages 1 & 2 were collapsed into a single variable to bring the group size to 15 (see Table 4.5). FIGO stage 4 (most advanced) was used as the reference group to estimate the effect of stage on OS & PFS. For the PFS model, the TCGA RNA-seq samples were omitted from the analysis because to these samples...
did not have the necessary clinical annotation for inclusion. The MDA HGSOC RNA-seq samples were omitted for the entire analysis, as no clinical annotation was available for this cohort.

The base model for OS (Table 4.6) showed that increased patient age was weakly detrimental to survival (HR 1.02, 95% CI: 1.01-1.04, P = 0.005). The model also suggests that patients diagnosed at the earlier FIGO stages are conferred a benefit on their survival, although this association was borderline insignificant (HR 0.34, 95%CI: 0.11 -1.02, p = 0.055).

As for PFS (Table 4.7) similar to OS; earlier FIGO stage at diagnosis conferred a beneficial effect, although to a lesser extent to than what was demonstrated in the OS model. Unlike the OS model, the age of the patient at diagnosis does not appear to not contribute any impact on PFS.

Using the OS and PFS models in survival analysis reveals differences in patient OS between the HGSOC cohort (p = 0.012, Figure 4.22 A) and a larger difference in patient PFS between the SHGSOC and AOC cohort (p < 0.0001, Figure 4.22 B).
Table 4.5 Clinical characteristics of the deconvolved HGSOC RNA-seq ovarian cancer samples by cohort.

<table>
<thead>
<tr>
<th>Cohort (N)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AOCS (80)</td>
</tr>
<tr>
<td><strong>Age (years)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Mean (SD)</strong></td>
<td>60 (8.6)</td>
</tr>
<tr>
<td><strong>FIGO stage at diagnosis</strong></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>II</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>III</td>
<td>68 (85%)</td>
</tr>
<tr>
<td>IV</td>
<td>12 (15%)</td>
</tr>
<tr>
<td><strong>Neoadjuvant chemotherapy</strong></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>5</td>
</tr>
<tr>
<td>No</td>
<td>75</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
</tr>
</tbody>
</table>

AOC, Australian Ovarian Cancer Study; SHGSOC, Scottish High Grade Serous Ovarian Carcinoma; TCGA, The Cancer Genome Atlas; FIGO, International Federation of Obstetrics and Gynaecology; SD, Standard deviation; NA, Not available.

aKruskal-Wallis test across the 3 HGSOC cohorts.

bChi-square test between the AOCS and SHGSOC cohort.
Figure 4.22 Kaplan Meier curves of HGSOC patient overall survival and progression free survival stratified across HGSOC RNA-seq cohorts.

A: Kaplan Meier curve of HGSOC patient overall survival stratified by 3 HGSOC RNA-seq cohorts.
**B:** Kaplan Meier curve of HGSOC patient progression free survival stratified by 2 HGSOC RNA-seq cohorts.

Table 4.6 Base model of Overall Survival of HGSOC patients in the SHGSOC, AOCS and TCGA cohorts

**OS base model (prior to deconvolution estimates) – Stratified by cohort**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>HR (95% CI)</th>
<th>HR S.E</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1 &amp; 2</td>
<td>0.34 (0.11 -1.02)</td>
<td>0.55</td>
<td>0.055</td>
</tr>
<tr>
<td>Stage 3</td>
<td>1.26 (0.72-2.19)</td>
<td>0.28</td>
<td>0.41</td>
</tr>
<tr>
<td>Age</td>
<td>1.02 (1.01-1.04)</td>
<td>0.01</td>
<td>0.005</td>
</tr>
</tbody>
</table>

HR, Hazard ratio, 95% CI, 95% confidence interval; S.E, Standard error. N = 168, N of events = 130. Concordance = 0.6, Wald test = 14.83, p = 0.002

Table 4.7 Base model for Progression Free Survival of HGSOC patients from the AOCS and SHGSOC cohorts.

**PFS base model (prior to deconvolution estimates) – Stratified by cohort**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>HR (95% CI)</th>
<th>HR S.E</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1 &amp; 2</td>
<td>0.41 (0.15 -1.07)</td>
<td>0.49</td>
<td>0.069</td>
</tr>
<tr>
<td>Stage 3</td>
<td>1.35 (0.74-2.39)</td>
<td>0.29</td>
<td>0.3</td>
</tr>
<tr>
<td>Age</td>
<td>1.01 (0.99-1.03)</td>
<td>0.01</td>
<td>0.31</td>
</tr>
</tbody>
</table>

HR, Hazard ratio, 95% CI, 95% confidence interval; S.E, Standard error. N = 140, N of events = 121. Concordance = 0.57, Wald test = 10.47, p = 0.01
4.4.4 Cox proportional hazard model of HGSOC using inferred deconvolution proportions.

Using the inferred proportions of the tumour compartment in a cox proportional hazard model, adjusted for patient stage and age; conferred a modest, but non-statistically significant protective effect on both OS (HR: 0.24, 95% CI: 0.03-1.91, \( p = 0.17 \)) and PFS (HR: 0.21, 95% CI: 0.03-1.55, \( p = 0.12 \)). The direction of effect for the immune compartment was the same as that of the tumour for OS (HR: 0.09, 95% CI: 0.002-3.15, \( p = 0.18 \)) and stronger for PFS (HR: 0.04, 95% CI: 0.001-1.45, \( p = 0.08 \)), although none of these results did met the \( p \)-value threshold of < 0.05 for statistical significance.

The inferred stromal content of the patient samples was negatively associated with OS and PFS. For OS, the scaled values of the proportion of metastatic fibroblasts was the component driving the weak association with worsening survival (HR: 1.3, 95% CI:1.01-1.57, \( p = 0.0039 \)). In terms of PFS, the scaled proportions of primary fibroblasts was the stromal component responsible for the impact (HR: 1.25, 95% CI: 1.04 – 1.54, \( p = 0.016 \)).

Diagnostic Schoenfeld tests for both (OS and PFS) cox proportional hazard models are included to demonstrate that the assumptions of the cox proportional were not violated (Figures 4.23 A and 4.24 A). These Schoenfeld tests plots demonstrate that the variables included in the model do not change in their risk over time. The plots in Figures 4.23 B and 4.24 B are reasonably symmetric around 0, indicating that no specific data points are exerting an outsized influence on the model.

As the absolute proportions of the metastatic fibroblasts found in the deconvolution were small (mean proportion is 3%). The values were scaled and centred at 0 to identify boundaries for patients with relatively less or greater metastatic fibroblast burden (Figure 4.25) for an independent variable survival analysis which suggests patients with more metastatic fibroblasts in their tumours experienced worsen survival (Figure 4.26).
The genes that define in metastatic fibroblasts in this work are outlined in the Table 4.8 and their differential expression contrasted to primary fibroblasts are plotted in Figure 4.27. These differentially expressed genes were submitted to a GO Biological Processes enrichment analysis in an effort to discern what biological functions are distinguishing these metastatic fibroblasts from primary fibroblasts.
Figure 4.23 Diagnosis plots of the final OS Cox-proportional hazard models for the HGSOC RNA-seq cohorts.
A: Schoenfeld residuals for each variable included in the final OS Cox-proportional hazard model. Including each variable model’s global Schoenfeld test p-value.

B: Deviance residual plot for the final OS Cox-proportional hazard model. The global standard deviation of these deviance residuals is displayed on top of the figure.
A: Schoenfeld residuals for each variable included in the final PFS Cox-proportional hazard model. Including each variable model’s global Schoenfeld test p-value.

B: Deviance residual plot for the final PFS Cox-proportional hazard model. The global standard deviation of these deviance residuals is displayed on top of the figure.

Figure 4.25 Histogram of the distribution of the scaled metastatic fibroblast proportion across 180 HGSOC RNA-seq samples.

The centre value 0.0 was used to produce a binary categorisation for patients based on their metastatic fibroblast proportion. Patients at or below the 0.0 value were classified as having a ‘low metastatic fibroblast population’ and patients beyond 0.0 were classified as having a ‘high metastatic fibroblast population’. This binary categorisation was used for a survival analysis.
Figure 4.26 Kaplan Meier curve of overall survival of HGSOC patients stratified by their metastatic fibroblast proportion.

The bulk RNA-seq samples featured in this survival analysis were from the SHGSOC, AOC and TCGA cohorts.
Figure 4.27 Volcano plot of Primary and Metastatic Fibroblasts DEGs

The Log₂ Fold Change threshold was 0.5 relative to the other cluster, the negative Log₁₀ $P_{adj}$ threshold was < 0.05. An additional criteria was the gene must be expressed in at least 10% more cells than the other cluster.
Table 4.8 Top 15 genes expressed in the metastatic fibroblast cluster

<table>
<thead>
<tr>
<th>Gene</th>
<th>Average log$_2$FC</th>
<th>Pct in cluster (%)</th>
<th>Pct out of cluster (%)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGS5</td>
<td>4.36</td>
<td>0.84</td>
<td>0.02</td>
<td>4.8e-177</td>
</tr>
<tr>
<td>COL4A1</td>
<td>3.45</td>
<td>0.78</td>
<td>0.08</td>
<td>1.3e-72</td>
</tr>
<tr>
<td>ACTA2</td>
<td>3.33</td>
<td>0.7</td>
<td>0.08</td>
<td>7.9e-57</td>
</tr>
<tr>
<td>IGFBP7</td>
<td>3.16</td>
<td>1</td>
<td>0.28</td>
<td>1.4e-48</td>
</tr>
<tr>
<td>TAGLN</td>
<td>2.8</td>
<td>0.72</td>
<td>0.13</td>
<td>5.2e-41</td>
</tr>
<tr>
<td>COL4A2</td>
<td>2.8</td>
<td>0.78</td>
<td>0.08</td>
<td>3.3e-69</td>
</tr>
<tr>
<td>MYL9</td>
<td>2.51</td>
<td>0.83</td>
<td>0.13</td>
<td>1.6e-52</td>
</tr>
<tr>
<td>SPARC</td>
<td>2.45</td>
<td>0.89</td>
<td>0.22</td>
<td>1.7e-36</td>
</tr>
<tr>
<td>CALD1</td>
<td>2.22</td>
<td>0.8</td>
<td>0.15</td>
<td>2.2e-44</td>
</tr>
<tr>
<td>COL18A1</td>
<td>2.1</td>
<td>0.78</td>
<td>0.24</td>
<td>2.8e-27</td>
</tr>
<tr>
<td>MCAM</td>
<td>1.95</td>
<td>0.78</td>
<td>0.03</td>
<td>1.9e-142</td>
</tr>
<tr>
<td>NOTCH3</td>
<td>1.81</td>
<td>0.72</td>
<td>0.05</td>
<td>4.1e-85</td>
</tr>
<tr>
<td>BGN</td>
<td>1.77</td>
<td>0.8</td>
<td>0.12</td>
<td>4.6e-51</td>
</tr>
<tr>
<td>TIMP3</td>
<td>1.74</td>
<td>0.65</td>
<td>0.07</td>
<td>2.5e-49</td>
</tr>
<tr>
<td>PDGFRB</td>
<td>1.5</td>
<td>0.56</td>
<td>0.04</td>
<td>1.02e-66</td>
</tr>
</tbody>
</table>

Average log2 fold change values of genes in the metastatic fibroblasts cluster found in the Shih et al. scRNA-seq dataset were compared to the expression of genes from all other clusters in the dataset.

Pct in cluster: Percentage of cells in the metastatic fibroblast cluster that express the gene

Pct out of cluster: Percentage of cells outside the cluster (i.e., the rest of the dataset) that express the gene

p-value: Determined using Wilcoxon Rank Sum Test
Figure 4.28 GO-plot of Metastatic Fibroblast DEGs

GO enrichment analysis of the metastatic fibroblast gene set (N = 413 genes), the terms returned are statistically significant after controlling for false discovery rate (FDR). The GO-plot illustrates where in the hierarchy of GO terms the biological process terms associated with the metastatic fibroblasts belong.
4.5 Discussion

This multifaceted deconvolution study has provided an account of the strengths and limitations of current deconvolution algorithms which seek to leverage single-cell RNAseq data in their estimations of cell type proportions that exist in bulk-RNAseq samples.

The overall aims of this work were 1) Establish a framework to assess the accuracy of available algorithms that learn features from scRNA-seq data to deconvolve bulk RNA-seq samples. 2) Apply the evaluations from the first aim to select an optimal approach for applying sc-RNAseq deconvolution to HGSOC RNA-seq samples where the true proportions of the bulk RNA-seq materials are not known. 3) Evaluate the deconvolution results of HGSOC RNA-seq samples and investigate if this approach furthers our understanding of the disease.

The framework of the first aim is set out in Figure 4.5. This work uses different HGSOC scRNA-seq datasets (C.Gourley, Shih et al., and Zhang et al.) and produces two sets of pseudo-bulk samples (normally distributed, representative distribution) from each scRNAseq dataset to evaluate the performance four scRNA-seq to bulk RNA-seq deconvolution algorithms (BisqueRNA, CIBERSORTx, MuSiC, and SCDC). The key premise behind the combinatorial approach between HGSOC scRNA-seq dataset and deconvolution algorithm is to find the training dataset-algorithm pair which would be expected to perform the most accurate deconvolution on bulk RNA-seq where the true proportions are not known. The framework evaluates the accuracy and error of each algorithm’s sample-wise deconvolution of pseudo-bulk RNAseq samples in addition to establishing cell-type-wise correlation and confidence intervals. The purpose of conducting this evaluation on two sets of pseudo-bulk samples was to see how well each algorithm could discern sample composition when cell type distributions were sampled from different distributions. The composition of the two sets of pseudo-bulk samples generated from each scRNA-seq data (C.Gourley, Shih et al., and Zhang et al.) are presented from Figures 4.6-4.8. The pseudo-bulk samples composed of an even proportion (i.e., not biased towards cell types) of single cells serves as an indicator to how well the cell type signal is teased from the training data by the algorithm. The pseudo-bulk samples produced from single cells pulled in proportion to how they are in the scRNA-seq data indicate how robust the deconvolution is to
mixture samples with heterogeneous RNA signals. As the true proportions of the prospective HGSOC bulk RNA-seq samples are not known, an algorithm that can perform well on both pseudo-bulk sample sets is preferred. An illustration of using how evaluating algorithms against 2 pseudo-bulk RNA seq datasets is helpful can be seen by looking at BisqueRNA sample-wise deconvolution accuracy of Zhang et al. scRNA-seq pseudo-bulk RNA-seq data from Table 4.4. The BisqueRNA algorithm attains a high sample-wise deconvolution accuracy of $R = 0.98$ when deconvolving the single cell representative Zhang et al. dataset. These samples typically feature epithelial cell proportions ranging between 30-80% while some cell types such as B-cells and T-cells are as scarce as <5% in a given ‘representative’ pseudo-bulk RNA-seq sample (Figure 4.8). Yet, when deconvolving pseudo-bulk RNA-seq samples made with even proportions of all the cell types found in Zhang et al., the BisqueRNA algorithm attains an low sample-wise $R$ of 0.18. Checking Figure 4.14 B it is apparent that BisqueRNA performs exceptionally well at deconvolving the epithelial cell signal to the exclusion of most other cell types in the scRNA-seq dataset.

The framework to evaluate the algorithms and the scRNA-seq training data is not a trivial matter. The authors of the deconvolution algorithms evaluate their own algorithms to those of similar purpose. Generally, the evaluations conducted and presented in their publications are done in such a way that demonstrates the strengths of the optimisation feature(s) of their debut algorithm. Many of these differentiating features were not exploited in the evaluations of this work. For example, in Wang\textsuperscript{53} MuSiC produces pseudo-bulk RNA-seq data from single cell data that feature many patients (10 subjects and 18 subjects). Training scRNA-seq data that feature many subjects takes advantage of MuSiC’s modelling of cross-subject mean and variance to produce accurate weights for each gene in their W-NNLS. In this work, the scRNA-seq training data from Shih et al., featured the most subjects at $N = 8$. Followed by C.Gourley subject $N = 2$ and Zhang et al., $N = 1$. MuSiC did perform its best deconvolution on the Shih et al., scRNA-seq data, only being marginally outperformed by CIBERSORTx in our tests. Further, for this framework, the algorithms were
provided each of the scRNA-seq reference datasets one at a time for pseudo-bulk RNA-seq deconvolution. One of the main differentiating features of the SCDC algorithm is their ENSEMBLE framework that uses multiple deconvolution estimates from multiple scRNA-seq references when estimating proportions \( (\mathbb{P})^{54} \). The authors of the SCDC publication highlight SCDC’s favourable performance on pseudo-bulk RNA-seq sample deconvolution when using their ENSEMBLE framework. The evaluation performed in this work did not utilise one of the features of the BisqueRNA algorithm, whereby the user can provide a reference scRNA-seq and bulk RNA-seq data from the same tissue for the algorithm to improve deconvolution\(^{191}\). Additionally, BisqueRNA can accommodate snRNA-seq (single nuclei RNA-sequencing) to identify cell populations in bulk RNA-seq samples. Despite the evaluation framework presented in this work not fully exploiting each algorithm. The framework aimed to set as fair a test as possible for each algorithm, not curating the scRNA-seq or pseudo-bulk RNAseq data in such a way that lends itself to specific optimisations of any select algorithm. However, the framework did allow for the adjustment of the datasets if provided ‘out of the box’ by the algorithms package. Examples of dataset adjustments include using the ‘S-mode’ batch correction from CIBERSORTx, no special operations were required to use this feature on the part of the user. Another example is using the single cell filtering QC tools for the scRNA-seq reference data by the SCDC package. This framework may better emulate how these algorithms are used ‘in the wild’, that is, not by the authors of the respective algorithms or researchers with particular interests in deconvolution, but rather by researchers and other individuals who want to use them to meet their research needs.

Taken together, this evaluation indicates that the best overall algorithms on pseudo-bulk RNA-seq samples were CIBERSORT\(^{x172}\) and MuSiC\(^{53}\). This indicates that both approaches for making proportion estimations in deconvolution are valid. That is, using either W-NNLS as seen in the MuSiC, SCDC and BisqueRNA algorithms\(^ {53,54,191}\) and SVR used in CIBERSORT\(^ {x172,189}\) can arrive at sensible estimations. Indeed, in a benchmarking review of 20 deconvolution algorithms from Cobos et al.\(^ {205}\) the authors concluded by recommending regression based deconvolution. The main difference between MuSiC and other W-NNLS algorithms was how MuSiC contends with the collinearity problem of highly concordant cell types, using a recursive tree procedure to delineate between the immune cell subtypes in the C.Gourley data (Table 4.2).
Furthermore, the effectiveness of this procedure is also illustrated in Figures 4.10 and 4.11 showing MuSiC achieving the highest sample-wise R values in samples where all cell-types are pulled from a normal distribution, beating out CIBERSORTx. These results suggest that MuSiC excels at detecting the cell-type signals embedded in the pseudo-bulk samples. Overall, CIBERSORTx tended to outperform the W-NNLS methods, it achieved marginal gains in sample-wise R values in the `single cell representative` pseudo-bulk samples in the Shih et al. scRNA-seq data (Table 4.3) and the Zhang et al., scRNA-seq data (Table 4.4). The key feature of CIBERSORTx is their S-mode batch correction, which uses a form of NNLS regression to adjust their basis-matrix for UMI/droplet based scRNA-seq data. This indicates outside the machine learning method used; other operations such as input data transformation, normalisation and scaling impact the accuracy of the estimations. A conclusion also drawn in the Cobos et al. review.

Concerning the limitations of deconvolution methods. It is important to remember the assumptions made by these models. Each model evaluated in this work (MuSiC, SCDC, CIBERSORTx, and BisqueRNA) constrains their regression with a nonnegativity of estimated values and a sum-to-one constraint. Nonnegativity is the least contentious constraint, as a mixture sample cannot have a negative estimated proportion of a cell type. One consideration when designing algorithms to contend with this constraint is how the regression values are adjusted to enforce nonnegativity. The implication of the sum-to-one constraint on the deconvolution estimates presupposes that all cell types are present in the input basis matrix, and that there are no other cell types in the mixture samples. This is a clear shortcoming of many present deconvolution approaches. Although scRNA-seq provides unparalleled resolution of the cellular constituents of solid tissue, there are still new cell types being discovered and to be discovered. It is unreasonable to expect any input basis matrix to contain the signatures of all cell types in a mixture bulk RNA-seq sample. Indeed, the impact of omitting cell types from the basis matrix in bulk RNA-seq deconvolution has been investigated by the publisher of the MuSiC algorithm and in
the benchmarking review by Corbos et al. Briefly, Wang et al.\textsuperscript{53} found that removing one cell type at a time from the reference scRNA-seq data of six pancreatic cell types (alpha, beta, delta, gamma, ductal, and acinar) resulted in less accurate estimations in the remaining cell types compared to estimates using the most comprehensive scRNA-seq reference data. Cobos et al\textsuperscript{205} performed a similar leave one cell type out investigation using the immune cell types in a PBMC dataset. They found mixed effects when certain cell types were omitted. For example, when using the DWLS algorithm, removing CD14\textsuperscript{+} monocytes from the scRNA-seq reference data resulted in the algorithm estimating more dendritic cells and T-cells. However, the removal of T-cells resulted in reduced estimation of dendritic cells, but not CD14\textsuperscript{+} monocytes on average. One bulk RNA-seq deconvolution algorithm, EPIC\textsuperscript{209}; does attempt to account for and provide an estimation of unaccounted cell types in the reference dataset. In the future, as more scRNA-seq cell atlases and datasets of niche tissue are published, aggregation of this information will result in more representative scRNA-seq reference data that can be used to improve deconvolution accuracy. Furthermore, the total error of algorithmic approaches may continue to diminish as approaches that leverage multiple scRNA-seq reference data (such as the SCDC EMSEMBLE framework\textsuperscript{54,205}) become more available.

Applying the most appropriate scRNA-seq dataset (Shih et al.) and algorithm (CIBERSORTx) to bulk RNA-seq HGSOC samples provided valuable insight as to what may be learned from this approach and how it can be improved. First, thought that the general trend of the proportions of cell type compartments were similar across all the HGSOC bulk RNA-seq cohorts (i.e., the largest compartment being the tumour, followed by the immune and stromal compartments). The data from Figure 4.17 show that each of the HGSOC cohorts exhibited a higher proportion for certain cell compartments relative to other cohorts. This raises some questions, such as: Are the compositions of tumours of HGSOC patients from separate cohorts truly different? Is this finding a consequence of some batch effects or other technical artefacts? Given that the deconvolution itself may not be entirely accurate, the observed heterogeneity in cell composition across the HGSOC cohort could be in part a reflection of the biases in the training data. However, for what biases may exist in this approach, these should be applied evenly across the HGSOC cohorts, so the differences from Figure 4.17 must be in part from something else.
The tumour compartment was the most dominant compartment (compared to stromal and immune) in terms of percentage across all the HGSOC bulk RNA-seq samples (Figure 4.17). Across all of the cohorts, the average tumour proportion was 47%, while the average immune proportion was 30% and the stromal proportion was 23%. These proportions are a little dissimilar from the original Shih et al. scRNA-seq dataset (tumour = 38%, immune = 37% and stromal = 24%). The in-house scRNA-seq datasets of this thesis (C.Gourley) that were profiled extensively in chapter 3, exhibit a similar pattern to the Shih et al. scRNA-seq dataset, but on a larger scale. These datasets tend to find that the tumour or immune compartment is dominant or equally represented throughout the TME, with the stromal compartment consistently a minority compartment. The deconvolution shows that the tumour compartment is dominant (≥40%) in 151 of 206 (73%) bulk RNA-seq samples. The proportions of the tumour compartment were positively correlated with the ESTIMATE cellularity estimates of the RNA-seq samples (R = 0.4, p = 2.9e-09, Figure 4.19). This finding would suggest that there is some concordance between this deconvolution approach and other recognised computational approaches that aim to learn more about clinical material from their RNA-seq expression data.

The survival analysis first seeks to find if any differences in overall survival and progression free survival differs between the HGSOC bulk RNA-seq cohorts prior to using any learned features from the deconvolution. Clinical data on overall survival were available only for the SHGSOC, AOC and TCGA cohorts and progression free survival for the SHGSOC and AOC cohorts. The baseline characteristics of the patients in these cohorts (i.e., age, FIGO stage at diagnosis and neoadjuvant therapy) were not statistically different (Table 4.5). There was a statistical difference in overall survival between the 3 HGSOC cohorts in overall survival (figure 4.22 A, p = 0.012) and between the SHGSOC and AOC in progression free survival (Figure 4.22 B, p < 0.0001). When applying the deconvolution features in the cox proportional hazards modelling, these models were stratified by cohort to isolate the effects of only the deconvoluted features.
Using the cell compartments in the cox proportional hazard models indicated that the proportions of the tumour and immune compartment were noninformative or very weakly associated with improved overall survival and progression free survival, though this is was not statistically significant. Interestingly, there was statistical evidence to suggest that the proportion of the stromal compartment imparts a worsening in overall survival for HGSOC patients (HR = 1.3, 95% CI = 1.1-1.5, p =0.0039). A similar but not statistically significant effect was observed for progression free survival using the stromal compartment. A follow-up survival analysis was performed on the components of the stromal compartment (primary & metastatic fibroblasts and endothelial cells) to identify if this effect was isolated in any particular identified cell type. There was a statistically significant worsening in overall survival according to whether HGSOC patients had proportionally higher burdens of the metastatic fibroblasts (p = 0.00037, n = 180). This finding supports the other publications that suggest that fibroblasts, or CAFS (Cancer Associated Fibroblasts) in HGSOC negatively affect patient outcomes\textsuperscript{210,211}. The evidence provided in this work is novel because this result was found using deconvolution using scRNA-seq data to estimate the cell types that constituted the samples of earlier RNA-seq datasets. Previous work that provides evidence of the deleterious effects of CAFS on HGSOC survival was done using IHC and RNA-seq markers (such as FAP, TGFβ, COLA11A1) to enumerate CAFs. The work of Hussian et al\textsuperscript{210} also used the ABSOLUTE and ESTIMATE algorithms to estimate the tumour cellularity of HGSOC RNA-seq samples. They found that their markers for CAFS (FAP, COL11A1, SULF1) and the mesenchymal TCGA molecular subtype exhibited the lowest tumour purity. This observation is also reflected using the deconvolution approach as well. Using the ESTIMATE cellularity scores of all the HGSOC cohorts RNA-seq samples there was a positive correlation between cellularity and tumour cell proportion (Figure 4.19), there was also a negative correlation between cellularity and the stromal compartment (r = -0.35, p = 1.71e\textsuperscript{-07}, Figure 4.21). This negative correlation is stronger when using deconvoluted estimations of the tumour compartment and the stromal compartment (r = -0.8, p = 2.2e\textsuperscript{-16}). Studies indicate that the presence of CAFs in HGSOC exhibit immunosuppressive functions in HGSOC\textsuperscript{212} and Breast cancer\textsuperscript{213}. There is a negative global correlation between the stromal and the immune compartment (r = -0.26, p = 0.0001), but this should not to be considered illustrative of any direct immunosuppressive function. In fact, there is a
stronger negative correlation between the tumour compartment and the immune (r = -0.35, p = 1.3e-07). This is to highlight that there are limitations to the utility of the deconvolution approach for studying the interactions between the inferred compartments and cell types in RNA-seq samples.

Although the opportunity to study biological interactions from deconvolved RNA-seq samples is limited. The source scRNA-seq data are rich datasets that can describe the cell-types in the deconvolution and can be explored extensively. Table 2.8 and Figures 2.27 and 2.28 describe the differentially expressed genes (DEGS) of the metastatic fibroblasts identified as detrimental to survival. As described in the original scRNA-seq paper from Shih et al.\textsuperscript{137}, the metastatic fibroblasts were more active (by way of possessing more differentially expressed genes) than the primary fibroblasts. This is supported by the clustering in this work, as the number of DEGs in the metastatic fibroblast cluster outnumber those found in the primary fibroblast cluster (Figure 2.27). The original authors of the data mentioned that the genes they found that describe metastatic fibroblasts are largely involved in the transcription of soluble factors. Using ClusterProfiler and GO (Biological Processes) terminology to describe the biological activity of the metastatic fibroblast cluster reveals that many of the DEGs are involved in protein transport around the cell’s organelles, membrane and endoplasmic reticulum. This is unsurprising as fibroblasts are well known for their role in modifying the extracellular matrix (ECM). There is also activity related to viral activity and mRNA catabolism that may in part be of clinical interest when studying how these cells exist in the patient TME.

The advantages of this deconvolution approach is the scalability of performing the analysis; given a basis matrix and algorithm, hundreds, if not thousands of HGSOC RNA-seq samples can be deconvolved in a matter of minutes and the approach is not constrained to making classifications based on a limited number of marker genes. This allows for a deeper subclassification of cells that express marker genes of interest, i.e., primary and metastatic fibroblasts. This approach also considers the quantitative expression of all the RNA transcripts of the cells captured in the preparation of the 206
RNA-seq libraries, illuminating more characteristics of the entire tumour environment rather than a narrow focus limited to how many markers can be multiplexed in IHC or in flow cytometry. Lastly, there is a lot of scope for the deconvolution approach to improve when high-quality, general-purpose cell atlases and niche scRNA-seq datasets are inevitably published. Alongside with more sophisticated algorithms, which are better able to leverage these datasets to reduce error in estimation. The unison of these developments make a good case for the usage of similar deconvolution approaches to investigate RNA-seq datasets in other clinical settings to identify cell types of interest for more targeted research.
Chapter 5: Discussion and conclusion

HGSOC continues to be the most deadly gynaecological malignancy worldwide. The disease exhibits extensive heterogeneity at the molecular level; a key trait of the disease believed to lead to the common recurrence of the disease after initially effective chemotherapy treatment\textsuperscript{214}. As such, there is a pressing unmet need to further elucidate the precise biological mechanisms that underlie the adaptive chemoresistance phenotype that is responsible for the progression and lethality of the disease.

As our ability to assay and examine the molecular landscape of biological systems such as ovarian tumours advances, so too does our understanding of the condition of OCs. This in turn drives innovation in interventions for the disease. Platinum chemotherapy exploits the initially compromised DNA repair mechanisms of tumour cells, leading to greater lethality in the tumour cells than in the host cells\textsuperscript{214}. Identifying and screening for mutations in key genes may indicate the usefulness of other therapies. In fact, a class of drugs called PARP inhibitors (poly ADP-ribose polymerase) have been shown to benefit the HGSOC patients with BRCA mutations. Mutations in BRCA indicate that the cancer exhibits an HRD (homologous repair deficient) phenotype, which these PARP inhibitors exploit to induce selective tumour cell lethality\textsuperscript{215}.

Regarding the tumour microenvironment of HGSOC, understanding of the importance of TILs (tumour-infiltrating lymphocytes) in HGSOC is used to introduce immunotherapy in the treatment of patients with HGSOC. The programmed death cell protein-1 (PD-1) and its ligand (PD-L1/2) suppress the antitumour response of the T-cells in the tumour microenvironment\textsuperscript{214,215}. As with PARP inhibitors, HGSOC patients are considered potential candidates for anti-PDL1 treatment, often in combined targeted therapy with PARP inhibitors and platinum chemotherapy. These combined targeted therapies not only aim to reduce the population of malignant cells, but influence the tumour microenvironment to promote T-cell activation, increase the
presence of key effector cells (CD8+ T cells and NK-cells) and decrease the immune suppressive activity of FOXP3 immunoregulatory T-cells\textsuperscript{215}.

Future work involving scRNA-seq in the investigation of HGSOC should include a larger cohort of patients and tumour samples (from primary and distal sites) at time points prior to, and after relapse. A larger cohort of patients is critical to determine which trends observed in the scRNA-seq data are potentially generalisable across patients that may be clinically translational. Additionally, an expanded cohort of HGSOC scRNA-seq data can elucidate disease features that are patient-specific. An ideal HGSOC scRNA-seq cohort would feature patients whose tumour samples are available both before and after the chemo-resistant phenotype emerges. This would enable novel investigations such as trajectory analysis of differential gene expression between states of chemosensitivity and chemoresistance\textsuperscript{216,217}. Identifying exploitable mechanisms along the trajectory from chemosensitivity to the acquisition of chemoresistance would be valuable for designing treatments tailored for HGSOC patients. Furthermore, as scRNA-seq resolves a TME viewpoint of these tumours, the receptor-ligand signalling between cancerous cells and the host environment may prove pivotal in understanding the current state of a tumour and its trajectory\textsuperscript{97}. As seen in microarray and RNA-seq publications, as more studies exploit the resolution afforded by scRNA-seq become published, it can be expected that new treatment modalities may emerge to treat HGSOC. Examples include identifying markers predicting responsivity to immunotherapies\textsuperscript{218}, targetable somatic mutations that arise during tumour / disease progression\textsuperscript{158}, and discovering neoantigens for vaccine development which stimulate highly active T-cells and antitumour immunity while reducing the risk of autoimmunity\textsuperscript{219}. Ultimately, the more research on the aspects of the TME related to the progression of HGSOC, the more clinical options that can be developed that serve to improve HGSOC survival and curtail disease progression.

The work done to develop the content of Chapter 2 is valuable for researchers wishing to expand and further develop scRNA-seq in HGSOC studies. The work details the development and optimisation of the protocol for the liberation of a high-viable cell fraction directly from freshly acquired and cryogenically preserved HGSOC tumours. This protocol would benefit any scRNA-seq technology, though it was developed specifically for the high cell throughput, microfluidic approaches. The approach used
in this work shared many similarities with how HGSOC scRNA-seq data were produced in earlier studies\textsuperscript{136–138}. However, the content of this chapter demonstrates how well and consistently the developed protocol produces single cell suspensions from dissociated HGSOC tissues. The cells recovered from the HGSOC tumours are comparable to the PBMC of the same patients in terms of cellular viability before sequencing. We demonstrate the scRNA-seq libraries from dissociated HGSOC material following the optimised protocol can be equally high quality to libraries from non-dissociated blood tissue, the ideal input material for 10X scRNA-seq. Another valuable insight that is often not evaluated in scRNA-seq publications is the high concordance observed between duplicated 10X scRNA-seq libraries. Two LGSOC libraries featuring \textasciitilde 13K and \textasciitilde 7K sequenced cells exhibited not only all the same cell types, but in very similar proportions as well. This provides some evidence that cell populations sampled only once in 10X scRNA-seq libraries can represent the heterogeneity of that sample.

The work in Chapter 3 presents the characterisation of the TME of 8 HGSOC scRNA-seq libraries. After the cells from these scRNA-seq libraries were labelled, the observations of the TME coincide with earlier scRNA-seq research of HGSOC. The work in Shih et al.\textsuperscript{137} identified cell populations in 14 HGSOC samples (2,911 cells total) and, when merging their disperse datasets together, identified three main components that clustered in distinct regions of their TNSE plot (Epi/Mesothelial, Fibroblasts/Stoma and Leukocytes). These were the key cell populations found in the HGSOC dataset featured in Zhang et al.\textsuperscript{138}. In this work, the cell components are defined as tumour, stromal and immune and are apparent in 6 HGSOC datasets. The only datasets that did not feature each of the TME components was the sample from the patient who had neo-adjunct chemotherapy, their TME featured no cells from the tumour compartment and a tumour from the a patient 4’s small bowel nodule. The reason why this site lacked the tumour compartment are unclear. The absence of the tumour compartment in the neo-adjunct chemotherapy dataset is likely due to the chemotherapy selectively targeting the cancerous cells. Positive evidence of the
patient’s exposure to the chemotherapy was evident in their T-cells, where a subset expressed metallothionein (MT) genes. A study of the expression of MT genes in OC treated with cisplatin and paclitaxel found that there was no association between MT expression and resistance to treatment. The same study indicated that Ki67 (MKI67), a marker for proliferation; decreased after chemotherapy\textsuperscript{155}. With only one chemo-exposed scRNA-seq dataset, no conclusion can be drawn in this work, but the distinct lack of the tumour compartment in the dataset may be evidence of this reduction in proliferation. Furthermore, there is expression of MKI67 in the epithelial cells in the chemo-naïve scRNA-seq datasets, especially in the MT/respiratory subcluster of the malignant epithelial cells. In contrast to the neo-adjunct chemotherapy dataset, the gene MKI67 is not expressed at all. There is other published research that suggests MT expression as a consequence to cisplatin chemotherapy is associated with chemotherapy resistance in OC\textsuperscript{220,221}. A paper from Surowiak et al.\textsuperscript{221} presented evidence of MT expression in the nuclei of the cells of the tumour. In this work it is identified that the MT expression is predominately in the T-cells in the tumours. It remains unclear whether MT expression in T-cells would lead to the frequent observed chemo-resistance in HGSOC, or if it is an adaptation of a diminished population of tumour cells.

The work in this chapter goes on to further divide the tumour compartment (epithelial and ciliated cell populations) using a CNV analysis approach to delineate between epithelial cells likely to be truly cancerous and noncancerous and present differential analysis between these cells. This work suggests that ciliated cells, suspected to be of fallopian origin, do not exhibit the extensive copy number variation that is typically observed in the tumour cells of HGSOC\textsuperscript{28}. Yet, when only the malignant cells are integrated into a new dataset and re-clustered, a ciliated cluster appears to re-emerge within the population of malignant cells. These observations are interpreted to mean that the early lesions that arise from the fallopian tubes do not exhibit extensive copy number variation, these are cells found in the ciliated cell clusters featured in HGSOC datasets S6 and S12. As the disease advances, the cells maladapt to a greater mutation load and aberration of their chromosomes and copy number variation. Meanwhile, the ciliated phenotype becomes less prominent amongst the other malignant cell functions which were uncovered in the integrative analysis (i.e. redox environment modification, proliferation, and immunomodulation).
The work in Chapter 4 features a novel analysis of RNA-seq HGSOC datasets using the granularity of scRNA-seq data and deconvolution algorithms to learn new features in older datasets. This approach to analysing earlier sequenced RNA-seq samples is still nascent and has a lot of promise for further development. This may be the most scalable way to compare the TMEs of thousands of HGSOC patients while the scRNA-seq datasets are still being generated. Furthermore, the principles used in this study can be applied to other cancer types. We show that there is still a lot to consider when attempting to apply the resolution of scRNA-seq datasets to bulk RNA-seq samples. Examples include considering to what extent the input scRNA-seq data are suitable for the bulk RNA-seq samples to be deconvolved, and how the models used in the algorithms can utilise the features present in the scRNA-seq dataset to provide accurate estimations in the RNA-seq data. Careful consideration when attempting to deconvolve RNA-seq samples can be rewarded with novel insight into those samples. This work demonstrates that deconvolved features can be associated with clinical variables such as survival using survival analysis and cox proportional hazard models. This work provides evidence that the stromal compartment of the HGSOC TME is negatively associated with survival, and that this effect is mainly from the metastatic fibroblast population. It is known that components of the stroma, particularly cancer-associated fibroblasts (CAFs), confer a pro-tumour function in many TMEs\textsuperscript{222,223}, including ovarian\textsuperscript{210}. There are aspirations to develop therapeutics in HGSOC that specifically target CAFs, yet present attempts are limited by not being fully able to fully elucidate and reach consensus on what signals consistent CAF signalling in a given TME\textsuperscript{222}.

In summary, this body of work uses scRNA-seq technology to contribute rich scRNA-seq data to the substantial research effort to understand the biological characteristics of HGSOC tumour microenvironment. This work features an optimised and demonstrably robust protocol to promote the use of fresh or preserved clinical material in the study of HGSOC in the laboratory that will help accelerate the generation of future scRNA-seq datasets. The analysis procedures presented here are reproducible.
and present a step forward in the profiling of the eight separate HGSOC TMEs. The surrounding discussion provides guidance on how gene sets of novel cell types and emergent subclusters can be evaluated and interpreted. This in turn accelerates the translation of these pioneering studies into clinical utility for the patient who participates in studies such as this. A step forward in characterising the TME of HGSOC and discovering the transcriptomic profiles of the malignant cells that these tumours harbour. This work concludes by capitalising on the level of detail afforded by scRNA-seq data, extending the clinical utility of older HGSOC datasets. The novel approach of deconvolving 206 bulk RNA-seq HGSOC datasets to provide new evidence of the deleterious affect metastatic fibroblasts have on patient survival. Returning to the source of the scRNA-seq data to characterise these cells of interest, we investigated possible marker genes for these cells and the genes that elude to their pathologic function in this aggressive cancer.
6 Limitations of Thesis Work and Priorities for Future Work

6.1 Limitations of Thesis work

6.1.1 Sample size and clinical diversity

The present thesis featured a total of 18 single-cell RNA sequencing (scRNA-seq) datasets across 3 sequencing batches to explore the intricacies of the tumour microenvironment (TME). These datasets encompassed 8 comprehensively characterised HGSOC tumour specimens, derived from a cohort of 4 patients. The specimens included 3 primary ovarian sites, 3 peritoneal sites, 1 omental metastasis and, one metastatic nodule on a small bowel.

While these datasets represent a significant contribution to the field of scRNA-seq investigations into HGSOC, particularly in terms of the number of single cell sequenced. It is imperative to acknowledge its limitations. The sample size, and the single-cell resolution and in the case of this thesis, the diversity of anatomical locations further constrains the statistical power of this study, making it challenging to present a generalisable ‘Characterisation of the TME of HGSOC’ that would describe a broader HGSOC patient population.

6.1.2 Lack of Patient follow up and longitudinal data

One of the most salient limitations of this thesis, particularly with respect to the objectives outlined in Chapter 3, is the absence of longitudinal HGSOC scRNA-seq datasets. The lack of follow up of the chemo-naïve and likely chemo sensitive patients over their exposure to treatment constrains the study’s capacity to delve into the
dynamics of chemoresistance in HGSOC. In HGSOC, most patients experience an initial response to chemotherapy and then go on to relapse with a chemo resistant HGSOC\textsuperscript{101,224}. The phenomena of chemoresistance in HGSOC is not static, it evolves in response to the patient’s chemotherapy regimens, manifesting as a complex interplay of genetic and epigenetic changes within the TME\textsuperscript{225}.

The absence of scRNA-seq data from HGSOC tumours that have developed chemoresistance poses a challenge in inferring the underlying mechanisms that contribute to this phenotype. The datasets featured in this thesis primarily consist of cells from chemo naïve tumours, with a sole expectation of S7- Patient 2 CTx Peritoneal sample (See in table 3.2) – this patient received 1 treatment of chemotherapy prior to primary debulking and featured no identifiable ‘HGSOC tumour’ cells (Figures 3.4 ‘Patient 2 CTx Peritoneal UMAP’ and in Figure 3.6 ‘Patient_2’). Consequently, the approach taken to identify subpopulations of HGSOC tumour cells to elucidate potential mechanisms of chemoresistance are inherently speculative and lack the empirical foundation that longitudinal data would provide.

\section*{6.1.3 Methodological Constraints}

While scRNA-seq has revolutionised our understanding of cellular heterogeneity and the TME, it is not without its limitations, some of which were encountered in this thesis. One source of methodological bias is that the ‘entire’ TME of a HGSOC specimen is based on a 1g sample from each patient’s tumour. This approach inherently loses the spatial context of the cells within the tumour, providing only a ‘snapshot’ of the transcriptome at the time of sample preparation. The ‘snapshot’ fails to capture the dynamic changes in gene expression that occur over time within the patients\textsuperscript{226}, limiting the scope of inferences the data from this thesis can reasonably substantiate.

Technical variation also poses a challenge in scRNA-seq studies. As outlined in throughout Chapter 2, scRNA-seq is sensitive to batch effect (see Figure 2.7 for outline of how viability varied across the sequencing batches in this study), which can confound downstream analysis, but can be partially ameliorated with rigorous optimisation. See as sequencing batches 2 and 3 did not exhibit a statistical difference in viability (Figure 2.7), and were both prepared as outlined in Chapter 2 methods 2.3.5 Automated dissociation of HGSOC tumours. Batch effect can arise from multiple
sources, including differences in sample preparation and dissociation (Table 2.4 & Figure 2.7), sequencing runs (Table 2.6 and Figure 2.8) and data processing, thereby introducing additional layers of complexity in the interpretation of the data.

An inherent bias was observed in the selection of HGSOC patient tumour specimens that were amenable to scRNA-seq library construction and subsequent analysis. This bias can be partly attributed to the varying efficacy of the tissue dissociation methods employed in this study. Specifically, the manual dissociation method (described in Chapter 2, method section 2.3.4) demonstrated inferior performance in terms of viable cell recovery and granularity of downstream analysis when compared to the scRNA-seq libraries produced using the automated dissociation method (described in Chapter 2, method section 2.3.5). Furthermore, there were tumour samples from patients that yielded low viability following the optimised automated dissociation method, baring a subset of HGSOC tumour samples from featuring in this study.

6.2 Priorities for Future Work

6.2.1 Increasing HGSOC scRNA-seq dataset sample size

The scRNA-seq dataset generated in this thesis offer a valuable foundation for research endeavours in the field of HGSOC tumour heterogeneity. One of the most pressing priorities for subsequent studies is the expansion of the number of HGSOC tumour specimens are sequenced using scRNA-seq. The limited number of samples in this thesis, although informative, restricts the statistical power and generalisability of the findings. Studies featuring more HGSOC scRNA-seq tumour can validate the cell types, transcriptomic subtypes, and other key findings presented here, thereby enhancing the robustness of the conclusions drawn about the TME of HGSOC.

A larger and more diverse HGSOC patient pool would enable more nuanced analysis that could significantly advance our understanding of the HGSOC TME. For instance,
stratification based on disease FIGO stage and treatment regimen could provide insight into how the TME evolves over the course of the disease and in response to different therapeutic interventions. Such stratification could also facilitate the identification of novel biomarkers or therapeutic targets that are stage or treatment-specific, potentially expanding the avenues for precision medicine in HGSOC.

The more patients enrolled in HGSOC scRNA-seq studies would also permit a greater exploration of inter-patient heterogeneity. This would both provide a more comprehensive view of the HGSOC TME and potentially reveal subgroups of patients who may benefit from targeted or specific therapies.

6.2.2 Longitudinal Studies to provide temporal resolution

Expanding the patient pool for scRNA-seq studies in HGSOC would be valuable, the incorporation of longitudinal follow up of patients recruited in HGSOC scRNA-seq studies would offer unique advantages that are particularly pertinent to the study of chemoresistance. Longitudinal studies allow for re-evaluation and tracking of any changes of the TME over the course of that patient’s treatment. This temporal resolution is crucial for elucidating the evolved mechanisms that underlie chemoresistance in HGSOC.

By sampling HGSOC patient tumours at multiple time points- before, during and after chemotherapy can identify specific cellular and molecular changes associated with the development of chemoresistance in HGSOC. This could potentially be a new lineage of tumour cells, or existing cells adapting their gene expression profiles that are temporally correlated with the onset of chemoresistance. Conversely, the patient’s HGSOC TME cellular proportions may be found to be associated with their response to therapy or prognosis, such as the designation of ‘hot’ and ‘cold’ TMEs.

As such, of the discussed future priorities expanding on the work of this thesis. Longitudinal studies offer the unique and indispensable perspective for studying chemoresistance in HGSOC, a pressing clinical need. These studies should be considered a high priority for future research in this area, complementing the efforts to expand the patient pool in scRNA-seq studies.
6.2.3 Mulit-omic integration

The advent of scRNA-seq has revolutionised our understanding of cellular heterogeneity in many cancers, including the efforts of this thesis for HGSOC and others. However, transcriptomic data alone may not sufficiently represent the full complexity of the TME of HGSOC. There are rapidly developing new methods and workflows that address this limitation, the integration of multiple ‘omics’ data types—such as proteomics, genome sequencing, epigenetics and metabolomics—which offer promising avenues for future research.

Multi-omic integration at the single-cell level adds additional dimensions to the cellular landscape of HGSOC. The scRNA-seq data in this work revealed the gene expression profiles of individual cells that were sequenced, adding proteomic data could serve to validate the gene expression profile by providing affirmative evidence of the transcript protein product in these cells, or conversely, demonstrate that although there is gene expression, some underlying mechanisms otherwise prevent protein synthesis.

Moreover, the integration of multi-omic data can enhance the robustness and reproducibility of each of their findings. It allows for cross-validation between the different ‘omic’ layers, reducing the likelihood of spurious results and increasing the confidence in identified targets or markers found to be associated with pertinent clinical features.
7 Bibliography


55. Potter, S. S. Single-cell RNA sequencing for the study of development,


66. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-


132. Proserpio, V. & Mahata, B. Single-cell technologies to study the immune system. *Immunology* 147, 133–140 (2016).


183. Van Herck, Y. *et al.* Multiplexed Immunohistochemistry and Digital Pathology as the Foundation for Next-Generation Pathology in Melanoma: Methodological


<table>
<thead>
<tr>
<th>Sample ID</th>
<th>cDNA</th>
<th>Indexed Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td><img src="image1" alt="Bioanalyzer trace" /></td>
<td><img src="image2" alt="Bioanalyzer trace" /></td>
</tr>
<tr>
<td>S2</td>
<td><img src="image3" alt="Bioanalyzer trace" /></td>
<td><img src="image4" alt="Bioanalyzer trace" /></td>
</tr>
<tr>
<td>S3</td>
<td><img src="image5" alt="Bioanalyzer trace" /></td>
<td><img src="image6" alt="Bioanalyzer trace" /></td>
</tr>
<tr>
<td>S4</td>
<td><img src="image7" alt="Bioanalyzer trace" /></td>
<td><img src="image8" alt="Bioanalyzer trace" /></td>
</tr>
<tr>
<td>S5</td>
<td><img src="image9" alt="Bioanalyzer trace" /></td>
<td><img src="image10" alt="Bioanalyzer trace" /></td>
</tr>
</tbody>
</table>

Figure S1: Bioanalyzer traces of the cDNA and Indexed libraries of sequencing batch 1. The traces were produced following the reverse transcription step following GEM formation during the 10X step (cDNA) and after the library had been constructed (Indexed Library).
Figure S2 Bioanalyzer traces of the cDNA and Indexed libraries of sequencing batch 2. The traces were produced following the reverse transcription step following GEM formation during the 10X step (cDNA) and after the library had been constructed (Indexed Library).
Figure S3 Bioanalyzer traces of pooled library batches #2 and #3 prior to submission for sequencing.

**A**: Bioanalyzer trace of 8 uniquely indexed, pooled 10X libraries prior to being sequenced on a S2 chipset on an Illumina Nova-Seq at Edinburgh Genomics. **B**: Bioanalyzer trace of 4 distinctly indexed, pooled 10X libraries prior to being sequenced on a S1 chipset on an Illumina Nova-Seq at Edinburgh Genomics.
Table S1: Batch 1 Sample preparation [cells to indexed 10X library] quality control metrics.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Specimen type</th>
<th>Total Cell Count (cells/mL)</th>
<th>Viability (%)</th>
<th>Input cell vol. (µL)</th>
<th>cDNA yield (ng)</th>
<th>PCR Cycle no. (cycles)</th>
<th>Library yield (ng/µL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HGSOC Fresh</td>
<td>1,260,000</td>
<td>63</td>
<td>13.8</td>
<td>16</td>
<td>16</td>
<td>12.4</td>
</tr>
<tr>
<td>2</td>
<td>HGSOC Fresh</td>
<td>686,000</td>
<td>54</td>
<td>27.5</td>
<td>190.1</td>
<td>16</td>
<td>3.2</td>
</tr>
<tr>
<td>3</td>
<td>HGSOC Cryo</td>
<td>1,700,000</td>
<td>73</td>
<td>9.7</td>
<td>18</td>
<td>16</td>
<td>1.1</td>
</tr>
<tr>
<td>4</td>
<td>HGSOC Cryo</td>
<td>322,750</td>
<td>80</td>
<td>38.5</td>
<td>45</td>
<td>16</td>
<td>1.7</td>
</tr>
<tr>
<td>5</td>
<td>HGSOC Cryo</td>
<td>2,060,000</td>
<td>84</td>
<td>8.3</td>
<td>298.9</td>
<td>16</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Sequencing batch 1 single-cell RNAseq libraries quality control metrics include total cell count, viability and input cell volume (cell volume loaded into the 10X Chromium device). The input cell volumes were based off the 10X protocol with the desire to achieve 10,000 recovered cells (maximum recommended by manufacturer). Sample 4 represents the maximum input volume of cells recommended, with the scope to recover 7,000 cells. The cDNA and Library yields were measured using the Agilent 2100 Bioanalyzer Expert Software using the dsDNA high sensitivity assay.

Fresh: Library was produced from a freshly acquired tumour specimen
Cryo: Library was produced from a tumour specimen that underwent cryopreservation and cryo-recovery.
<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Specimen type</th>
<th>Total Cell Count (cells/mL)</th>
<th>Viability (%)</th>
<th>Input cell vol. (uL)</th>
<th>cDNA yield QB: (ng/µL) BA: (ng)</th>
<th>PCR Cycle no. (cycles)</th>
<th>Library yield QB: (ng/µL) BA: (ng/µL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>HGSOC Cryo</td>
<td>1,440,000</td>
<td>83</td>
<td>18.7</td>
<td>4.9</td>
<td>389.7</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>HGSOC Fresh</td>
<td>1,850,000</td>
<td>81</td>
<td>14.8</td>
<td>13.1</td>
<td>774.8</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>HGSOC Blood</td>
<td>4,640,000</td>
<td>91</td>
<td>8</td>
<td>2.5</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>9</td>
<td>HGSOC Fresh</td>
<td>3,000,000</td>
<td>95</td>
<td>10</td>
<td>49.9</td>
<td>694</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>HGSOC Fresh</td>
<td>1,985,000</td>
<td>96</td>
<td>10</td>
<td>33.6</td>
<td>1913</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>HGSOC Blood</td>
<td>4,000,000</td>
<td>93</td>
<td>10</td>
<td>44.1</td>
<td>280</td>
<td>12</td>
</tr>
<tr>
<td>12</td>
<td>LGSOC Fresh</td>
<td>5,405,000</td>
<td>74</td>
<td>6.4</td>
<td>53.9</td>
<td>929</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>LGSOC Fresh</td>
<td>5,405,000</td>
<td>74</td>
<td>6.4</td>
<td>36.4</td>
<td>787</td>
<td>11</td>
</tr>
</tbody>
</table>

Fresh: Library was produced from a freshly acquired tumour specimen
Cryo: Library was produced from a tumour specimen that underwent cryopreservation and cryo-recovery.
Table S 3 Batch 3 sample preparation quality control metrics

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Specimen type</th>
<th>Total Cell Count (cells/mL)</th>
<th>Viability (%)</th>
<th>Input cell vol. (uL)</th>
<th>cDNA yield QB: (ng/µL)</th>
<th>PCR Cycle no. (cycles)</th>
<th>Library yield QB: (ng/µL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>HGSOC Fresh</td>
<td>4,730,000</td>
<td>94</td>
<td>6.4</td>
<td>5.4 389.7</td>
<td>11</td>
<td>46.9</td>
</tr>
<tr>
<td>15</td>
<td>HGSOC Fresh</td>
<td>1,270,000</td>
<td>98</td>
<td>20</td>
<td>9.3 774.8</td>
<td>11</td>
<td>25.7</td>
</tr>
<tr>
<td>16</td>
<td>HGSOC Fresh</td>
<td>1,430,000</td>
<td>82</td>
<td>20</td>
<td>11 l 30</td>
<td>11</td>
<td>34.2</td>
</tr>
<tr>
<td>17</td>
<td>HGSOC Fresh</td>
<td>5,230,000</td>
<td>99</td>
<td>6.4</td>
<td>21 l 694</td>
<td>11</td>
<td>28.5</td>
</tr>
</tbody>
</table>

Sequencing batch 3 single-cell RNAseq libraries quality control metrics include total cell count, viability and input cell volume (cell volume loaded into the 10X Chromium device). The input cell volumes were based off consultations with 10X technicians suggesting an adjustment to the protocol encouraging aiming for recovery of between 15,000-20,000 cells/library. The cDNA and Library yields were measured a Qubit dsDNA high-sensitivity assay.