Evolutionary dynamics of ovarian cancer microenvironments and tumour cells

by

Allen Wenyu Zhang

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Evolutionary dynamics of ovarian cancer microenvironments and tumour cells

submitted by Allen Wenyu Zhang in partial fulfillment for the requirements for

the degree of Doctor of Philosophy

in Bioinformatics

Examining Committee:

Sohrab Shah, Pathology and Laboratory Medicine
Co-supervisor

Wyeth Wasserman, Medical Genetics
Co-supervisor

Brad Nelson, Medical Genetics
Supervisory Committee Member

Wan Lam, Pathology and Laboratory Medicine
University Examiner

Gabriela Cohen Freue, Statistics
University Examiner

Additional Supervisory Committee Members:

Martin Hirst, Microbiology and Immunology
Supervisory Committee Member

Raymond Ng, Computer Science
Supervisory Committee Member

Daniel Renouf, Medicine
Supervisory Committee Member

Abstract

High-grade serous ovarian cancer (HGSC) is the most common and lethal histotype of epithelial ovarian cancer. Often presenting as multi-site disease, HGSC exhibits extensive malignant clonal diversity with widespread but non-random patterns of disease dissemination. The proclivity of HGSC toward clonally heterogeneous disease is thought to underlie the prevalence of treatment-resistant disease. Yet, the factors that influence the spatial distribution of cancer clones in HGSC remain largely uncharacterized. Hypothesizing that distinct peritoneal niches formed by microenvironmental cell types shape the observed patterns of clonal dynamics in HGSC, the primary aim of this thesis was to understand how microenvironmental factors influence malignant cell evolutionary dynamics.

To establish the experimental substrate for this thesis, I led the construction of a cohort of 148 tumour samples from 41 HGSC cases (Chapter **2**). In addition to coordinating clinical case identification, I oversaw and learned how to create patient-derived xenograft models and conduct single cell experiments from patient tumours. Leveraging this resource, I explored whether local immune microenvironment factors shape tumor progression properties at the interface of tumor-infiltrating lymphocytes and cancer cells (Chapter **3**). Through multi-region study with whole-genome sequencing, immunohistochemistry, image analysis, gene expression profiling, and T- and B-cell receptor sequencing, I identified three immunologic subtypes across samples associated with patterns of malignant clonal diversity. These findings were consistent with immunological pruning of tumor clones. Finally, in order to explore the non-lymphocytic components of the tumour microenvironment, I developed an automated approach to cell type identification from single cell RNA-seq data that eliminates the manual work involved in traditional workflows reliant on post-hoc expert annotation (Chapter **4**). I demonstrated how this method performs superiorly to state-of-the-art workflows for cell type identification and applied the method to profile the HGSC microenvironment.

Collectively, this work highlights multiple interfaces of evolutionary interplay between malignant and non-malignant cells in the HGSC microenvironment, identifying novel mechanisms by which tumour cells escape from immune recognition. These results will inform the interpretation of results from immunotherapy clinical trials and set the stage for comprehensive microenvironment profiling in large HGSC cohorts and other cancers.

Lay Summary

Ovarian cancer is one of the leading causes of death from cancer in the developed world. Over 50% of patients with the most common type of ovarian cancer, high-grade serous ovarian cancer, die within 5 years of diagnosis. While most patients get better initially with treatment, the disease eventually becomes resistant. These cancers often contain multiple distinct subpopulations of cancer cells. Treatment that works on one cancer cell population may not on others, allowing the cancer to survive. Tumours also contain non-cancerous cell types, including cells from the immune system. Some of these non-cancerous cell types are linked to how long patients survive. The goal of this work is to understand how these cell types affect cancer cell growth. In doing so, we may be able to change the way non-cancerous cells interact with cancer cells to treat ovarian cancer.

Preface

Under the guidance of my co-supervisors, Dr. Sohrab P. Shah and Dr. Wyeth W. Wasserman, I was involved in the conception and design of the work presented in this thesis. I was responsible for the experimental research, data analysis, interpretation, and presentation of the work. In addition, I learned how to perform the wet lab experiments for some components of Chapter **2**, including sample processing and patient-derived xenograft (PDX) construction. This work would not have been possible without the generous help of many close clinical and research collaborators, acknowledged below.

Chapter **2** is unpublished work describing the accrual of the largest currently published cohort of multi-site high-grade serous ovarian cancer (148 tumour samples from 41 patients). I led this project under the oversight of gynecologic oncology surgeon Dr. Jessica McAlpine and Dr. Sohrab Shah, coordinating a multidisciplinary team of clinical and research staff including surgeons, fellows, clinical research assistants, animal research staff, and technologists to identify, recruit, and collect from probable cases of multi-site HGSC. Dr. Jessica McAlpine and various clinical fellows were responsible for earmarking HGSC cases. I forwarded cases and directed sample collection for multiple experiments, including PDX construction, single cell processing, and bulk sequencing. Additionally, I worked with Jamie Lim, Dr. Ciara O'Flanagan, and Dr. Peter Eirew from BC Cancer to design the experimental protocol for single cell RNA-sequencing. Gayle Jagpal and Chriselle Mariz Serna from Vancouver General Hospital (VGH) prioritized research cases on surgical slates. Various surgeons at VGH and University of British Columbia (UBC) Hospital performed the surgeries. Khaye Rufin and Stephanie Lam from BC Cancer were responsible for obtaining consent and coordinating peripheral blood collection. Dr. Lien Hoang from VGH was responsible for pathologic evaluation. Nancy Ferguson and Martin Avancena from VGH were responsible for clinical processing of samples from the operating room. Jamie Lim and Clara Salamanca from BC Cancer additionally processed samples for research purposes and sequencing. Jamie Lim and Dr. Ciara O'Flanagan performed single cell RNA-seq experiments. Teresa Ruiz de Algara and So Ra Lee from BC Cancer transplanted patient tumours into PDX mouse models, and taught me how to do so. I performed the bioinformatics data analysis in this chapter.

Chapter **3** is a modified version of material published in "Zhang, AW *et al.* Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. *Cell* (2018)." [1] This project

was led by myself and co-supervised by Dr. Sohrab Shah and Dr. Brad Nelson, Director of the Deeley Research Centre at BC Cancer. This research leverages the experimental resource created in Chapter **3**. I was primarily responsible for the bioinformatics analysis, including the identification of immune infiltration patterns, the associative analysis between malignant clonal diversity and these patterns, processing and analysis of T-cell receptor and B-cell receptor sequencing data, neoantigen calling and integrative analysis, HLA loss-of-heterozygosity analysis, and integrative analysis of mutational signatures and immune patterns. Additionally, I formulated the hierarchical Bayesian probabilistic model for inferring subclonal HLA loss-of-heterozygosity, and helped formulate the clonal inference model with Dr. Andrew McPherson. Dr. Phineas Hamilton performed cell type identification from H&E (hematoxylin & eosin) images, and I performed the hotspot identification and integrative analysis with immune patterns. Dr. Katy Milne, Sonya Laan, Stacey LeDoux, and Dr. David Kroeger from the Deeley Research Centre performed the immunohistochemistry experiments. I was responsible for generating all of the figures and tables associated with the paper. Together, Dr. Sohrab Shah, Dr. Brad Nelson, Dr. Robert Holt, and I were responsible for conceiving the project, designing the experiments, interpreting the results, and writing the manuscript. For the full list of contributors and their important contributions, please refer to [1].

Chapter **4** is a modified version of material that is under peer review and was preprinted in BioRxiv [2]: "Zhang, AW *et al.* Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers". I led this project under the co-supervision of Dr. Kieran Campbell and Dr. Sohrab Shah. Together with Dr. Kieran Campbell and Dr. Sohrab Shah, I helped formulate the model and interpret the results. I performed the majority of the bioinformatics data analysis in the paper, including the model implementation, simulation analysis, Bayesian model fitting with `pymc3`, analysis of external datasets, and analysis of high-grade serous ovarian cancer and follicular lymphoma single cell RNA-seq data. I led the accrual of the high-grade serous ovarian cancer cohort for this chapter, as described in Chapter **2**. Jamie Lim and Dr. Ciara O'Flanagan processed ovarian samples for single cell RNA-sequencing. Dr. Sohrab Shah, Dr. Christian Steidl (BC Cancer), Dr. Anja Mottok (BC Cancer), and Dr. Clementine Sarkozy (BC Cancer) identified the follicular lymphoma cases. Elizabeth Chavez (BC Cancer) performed the experimental work for single cell RNA-seq of the follicular lymphoma and reactive lymph node samples. Matt Wiens, Pascale Walters, and Tim Chan, co-operative education students at BC Cancer I helped supervise, helped build infrastructure and perform other components of analysis related to the paper. I wrote the manuscript along with Dr. Kieran Campbell and Dr. Sohrab Shah. All other co-authors assisted in data collection, generation, and/or interpretation of the results.

Ethical approval for the content in Chapters 2, 3, and 4 was obtained from the University of British Columbia (UBC) Research Ethics Board (ethics numbers H08-01411, H14-02304, and H18-01090).

Table of Contents

List of Tables

List of Figures

List of Supplementary Materials

Supplementary Table A.1. Primers for deep amplicon sequencing.
Supplementary Table A.2. TIL densities, TIL subtypes, molecular subtypes, epithelial colocalization measures from histologic image analysis, somatic SNV and rearrangement counts, ITH measures, and TCR and BCR repertoire diversity.
Supplementary Table A.3. Neoepitope table
Supplementary Table A.4. HLA LOH table
Supplementary Table A.5. Mutation signature proportions and cluster assignments for multisite HGSC, OV-AU, and [3] cohorts.
Supplementary Table A.6. OV-AU differential gene expression analysis table
Supplementary Table A.7. TCGA sample information, foldback-HLAMP, and cytotoxicity clusters

Supplementary Table B.1. Performance measures on simulated data.
Supplementary Table B.2. Marker gene matrices.
Supplementary Table B.3. Pathway enrichment results.

List of Symbols and Abbreviations

BCR	B-cell receptor
BH	Benjamini-Hochberg
CNV	Copy number variant
DLP	Direct library preparation
DNA	Deoxyribonucleic acid
FACS	Fluorescence-activated cell sorting
FBI	Foldback inversion
FDR	False discovery rate
H&E	Hematoxylin & eosin
HLA	Human leukocyte antigen
HGSC	High-grade serous ovarian cancer
HRD	Homologous recombination deficiency
ITH	Intratumoural heterogeneity
LOH	Loss of heterozygosity
MCMC	Markov chain Monte Carlo
MHC	Major histocompatibility complex
PARP	Poly ADP-ribose polymerase
PDX	Patient-derived xenograft
RNA	Ribonucleic acid
scRNA-seq	Single-cell RNA sequencing
SNV	Single-nucleotide variant
SV	Structural variant
TCR	T-cell receptor
TIL	Tumour-infiltrating lymphocyte
TME	Tumour microenvironment
WGS	Whole-genome sequencing

Acknowledgements

I would like to extend my deepest gratitude to my supervisors Dr. Sohrab Shah and Dr. Wyeth Wasserman. I consider myself extremely fortunate to have been co-mentored by two of the world's leading experts in genome analysis who both took an acute interest in my training. Thank you both for providing me with numerous incredible research and career opportunities, and for not holding back despite my shorter timeline and obligations to clinical studies in the MD/PhD program. I was truly privileged to have had experience coordinating projects, designing experiments, performing wet lab techniques, and interacting with stellar cancer researchers at international workshops in your labs. Most of all, thank you for believing in and supporting me in my endeavours, even when they did not directly align with your own interests. I am forever indebted for these opportunities that have allowed me to grow, as both a scientist and an individual, over the last few years.

My MD/PhD has brought me close to many talented scientists who I have had the privilege of getting to know. Firstly, I would like to thank my committee members Dr. Brad Nelson, Dr. Daniel Renouf, Dr. Raymond Ng, and Dr. Martin Hirst for their advice, feedback, and enthusiasm on my research. I am especially grateful to Dr. Brad Nelson for going above and beyond to support me as a collaborator and mentor throughout my training, especially during the beginning of my graduate studies. I would also like to recognize the many incredible individuals from the Shah, Nelson, Holt, Aparicio, and Huntsman labs who were always willing to help me — you have made the research in this thesis possible. Specifically, I would like to thank my research and clinical collaborators Dr. Andrew McPherson, Dr. Phineas Hamilton, Dr. Alex Miranda Rodriguez, Dr. Katy Milne, Dr. Jessica McAlpine, and Dr. David Huntsman, who helped make Chapter **3** possible. Thank you to Dr. Kieran Campbell for supervising and co-leading Chapter **4** with me, and Dr. Ciara O'Flanagan and Dr. Samuel Aparicio for their help in advising and executing the project. Special thanks to Jamie Lim for supporting much of the work in this thesis through her diligent experimental work and enthusiasm to innovate. Also, thank you to the research staff and fellows who I had the pleasure of working with, including Dr. Camila de Souza and Dr. Daniel Lai. Big thanks to Cynthia Berry, Yessie Werner, Jad Maanaki, Dora Pak, Carolyn Lui, and Sogol Tahmasbi for their incredible administrative support. Thank you to Dr. Anthony Mathelier and Dr. Fong Chun Chan who provided invaluable guidance during the early phases of my MD/PhD, and to Dr. Torsten Nielsen, Dr. Lynn Raymond, and

the UBC MD/PhD program. And of course, none of this work would have been possible or meaningful without the patients and their families that support us.

I am grateful to the Canadian Institutes of Health Research, BC Children's Hospital, UBC MD/PhD Program, BC Cancer Foundation, and Canadian Cancer Society for generous financial support for Chapter **2**, Chapter **3**, and Chapter **4**. Additionally, thanks to the Vanier Canada Graduate Scholarship Program, the Michael Smith Foreign Study Supplement Program, the Canada Graduate Scholarship-Masters Program, the UBC Four Year Fellowship Program, the UBC Faculty of Medicine, the UBC Faculty of Science, the UBC Bioinformatics Program, and the Canadian Conference for Ovarian Cancer Research for providing scholarships and travel awards for my studies.

Finally, I would like to thank my family and friends for their unwavering support and patience. To my mother and father — thank you for always looking out for me, putting me first, and understanding when I needed time to work. None of what I have accomplished today and will achieve would be possible without the bold sacrifice you made to move to Canada. To my grandfather and grandmother — thank you for raising and loving me. And to my close friends — thank you for making the effort to stay connected and pulling me away from my books once in awhile.

Dedication

To my parents, grandfather, late grandmother, and loving family.

Chapter 1

Introduction

Cancer is a disease of the genome and one of the leading causes of death in Canada [4]. Most cancers develop from a single cell that undergoes successive cycles of expansion, diversification, and pruning [5]. Thus, cancers are non-homogeneous mixtures of genomically and phenotypically distinct populations of tumour cells called clones. The process by which these clones expand, shrink, and diversify over time is known as clonal evolution [5]. The phenotypic diversity of cancer cells generated by clonal evolution explains some cases of resistance, where pre-existing, treatment-resistant clones survive initial therapeutic assault and expand to give rise to tumour cells present at relapse [6].

Importantly, these processes do not occur in isolation. Malignant cells occupy niches shared by non-malignant cells, such as lymphocytes, macrophages, fibroblasts, adipocytes, and pericytes [7]. These cells can have tumour-promoting or inhibitory functions that alter the evolutionary trajectories of tumour cells. For example, tumour-infiltrating lymphocytes (TILs) can respond to tumour-associated antigens, mounting anti-tumour immune responses [8]. On the other hand, cancer-associated fibroblasts can promote metastasis, angiogenesis, and tumour cell proliferation through extracellular matrix remodelling and cytokine secretion [9]. Together, the dynamic network of interactions formed by malignant and non-malignant cells forms the tumour microenvironment (TME). Like tumour cells, the composition and properties of the microenvironment can change over carcinogenesis, progression, and treatment [7]. Understanding how the microenvironment shapes the evolutionary histories of tumours will aid in predicting how tumours will respond to treatment.

I investigate these phenomena in high-grade serous ovarian cancer (HGSC), a subtype of ovarian cancer characterized by rampant dissemination throughout the peritoneal cavity, forming an ideal substrate for investigating how tumour cells evolve in diverse microenvironmental contexts. In Section **1.1**, I review the epidemiology and pathophysiology of HGSC. In Section **1.2**, I discuss the current understanding of clonal diversity in HGSC. Section **1.3** provides a brief overview of the microenvironmental cell types most relevant to HGSC, and Section **1.4** reviews contemporary experimental and computational approaches to study tumour evolution and the

microenvironment. Finally, Section **1.5** outlines the central aims of this thesis – to characterize the TME of HGSC and understand how non-malignant cells impact malignant evolutionary dynamics.

1.1 High-grade serous ovarian cancer

1.1.1 Epidemiology

In North America, ovarian cancer is the leading cause of death from gynecological malignancies and the fifth most common cause of cancer death [10]. The most common histopathological subtype of ovarian cancer, high-grade serous ovarian cancer, makes up 70% of diagnoses and has a 5-year survival rate of under 50% [10]. HGSC patients often present with peritoneal spread to invasive foci in the omentum, small bowel, and other organs. The current standard-of-care treatment, primary debulking surgery followed by combination platinum-taxane chemotherapy, is effective at treating the primary tumour, but the disease almost always recurs (80%) [11]. Despite extensive research into developing new screening and therapeutic strategies, patient survival has not improved substantially over the last 3 decades [12].

HGSC is typically sporadic, with approximately 10-20% of cases being hereditary [13]. Most of these correspond to germline *BRCA1* or *BRCA2* variants, which are associated with superior response to platinum chemotherapy due to the impaired ability of BRCA-mutated tumours to repair platinum-induced DNA damage through homologous recombination (HR) [14, 15]. The recent introduction of poly ADP-ribose polymerase (PARP) inhibitors for HGSC, a class of drugs that exploits synthetic lethality by impairing compensatory DNA repair pathways, has demonstrated improved outcomes especially among HR-deficient (HRD) cases [16, 17]. Nevertheless, outcomes remain bleak, especially for cases without HRD which constitute approximately half of HGSC [3].

1.1.2 Pathophysiology

Traditionally, HGSC was thought to originate from the ovaries in cortical inclusion cysts of ovarian surface epithelium [18]. While some studies still support an ovarian origin of HGSC, most new evidence points toward the epithelium of the distal fallopian tube as the anatomic origin of HGSC in the majority of cases (**Figure 1.1**) [19, 20]. Serous tubal intraepithelial carcinoma (STIC) lesions found on the fallopian tubes of BRCA carriers share genomic features of HGSC [21]. Furthermore, phylogenetic analysis of mutations present in STICs and HGSC tumours from the same patients has established these lesions as the precursors to most occurrences of

HGSC [21].

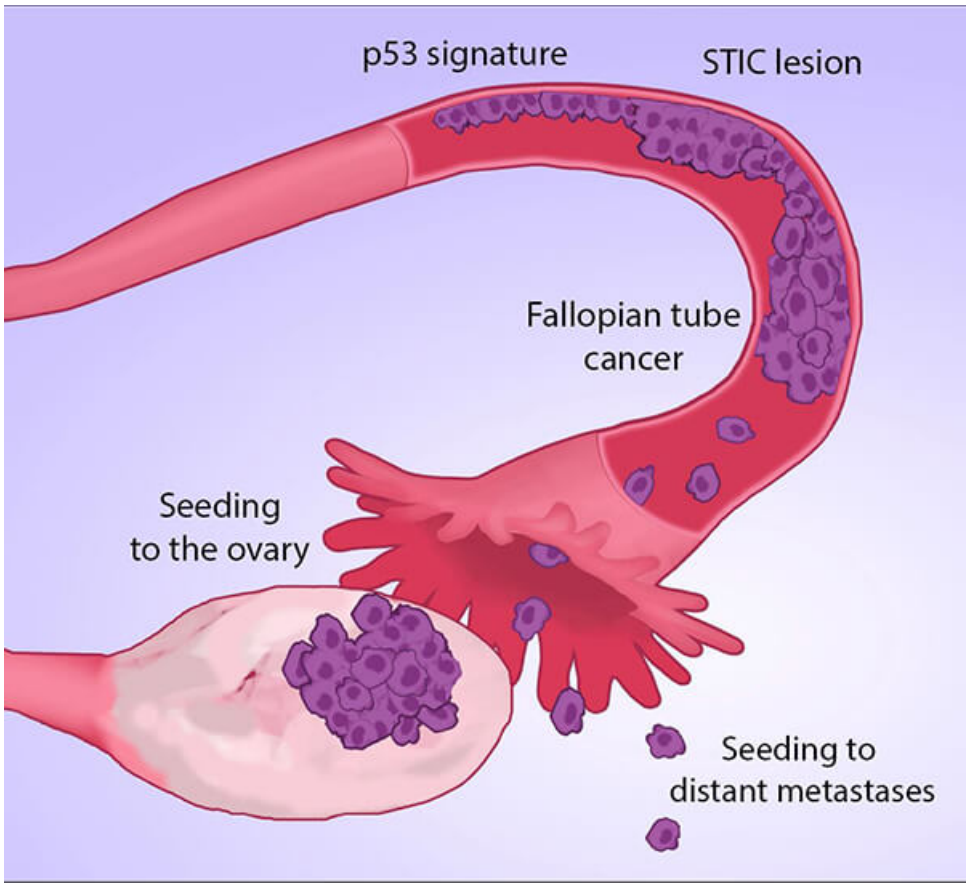


Figure 1.1: Fallopian tube origin hypothesis of HGSC pathogenesis. HGSC is thought to be derived from fallopian tube epithelial cells that acquire *TP53* mutations. These cells may eventually develop into histologically detectable STIC lesions. STICs seed the ovary and possibly also metastatic lesions elsewhere. Used with permission from https://www.cancer.gov/news-events/cancer-currents-blog/2017/ovarian-cancer-fallopian-tube-origins (Carolyn Hruban).

Almost all HGSC tumours harbour mutations in *TP53*, which occur as an early event in disease progression [14, 22]. Approximately 18% of cases have somatic *BRCA1* or *BRCA2* variants, which are largely exclusive with *CCNE1* amplification (20%); together, when combined with germline variants and epigenetic silencing, approximately half of all HGSC cases have HRD [14]. The prevalence of *TP53* mutations and HRD leads to incompetence in DNA repair. Thus, widespread chromosomal aberrations, aneuploidy, and severely disrupted karyotypes are defining characteristics of HGSC. Other genes affected by recurrent mutations and copy number events

include *MYC* (> 20%), *PIK3CA* (17%), *NF1* (12%), *RB1* (11%), *KRAS* (11%), *PTEN* (7%), and *CDK12* (3%) [14].

Recently, integrated genomic analyses of single-nucleotide (SNVs) and structural variants (SVs) derived from whole-genome sequencing studies of HGSC have revealed 2 major genomic subtypes of HGSC: homologous recombination-deficient (HRD) and foldback inversion-associated (FBI) [3, 23]. HRD cases, comprising approximately 50% of HGSC, are primarily defined by the presence of the HRD-associated SNV signature along with short deletions and tandem duplications. *BRCA1* and *BRCA2*-mutated tumours are included in this subgroup and are defined by distinct SV signatures (*BRCA1* by short tandem duplications and *BRCA2* by short deletions) [3]. FBI tumours make up most of the remainder of cases (40%) and harbour foldback inversions – duplicated sequences that face away from a breakpoint [3]. Foldback inversions are thought to arise through successive breakage-fusion-bridge cycles, co-occur with *CCNE1* amplification, and are associated with poor survival in HGSC [24]. A third tandem duplicator subgroup, associated with *CDK12* mutations, accounts for a final, minor fraction of cases (10%) [25] and has been linked to the worst outcomes in a recent preprint [26].

1.2 Intra- and inter-tumoural heterogeneity in HGSC

The clonal evolution theory of cancer cell populations posits that tumours arise from a single origin and diversify through acquisition of genomic alterations over time [5]. Ultimately, this process gives rise to genotypically distinct populations of cells, called clones, with corresponding phenotypes (**Figure 1.2**). Computational methods permit identification of clones based on somatic variants from bulk and single-cell DNA sequencing data (Section **1.4.1**).

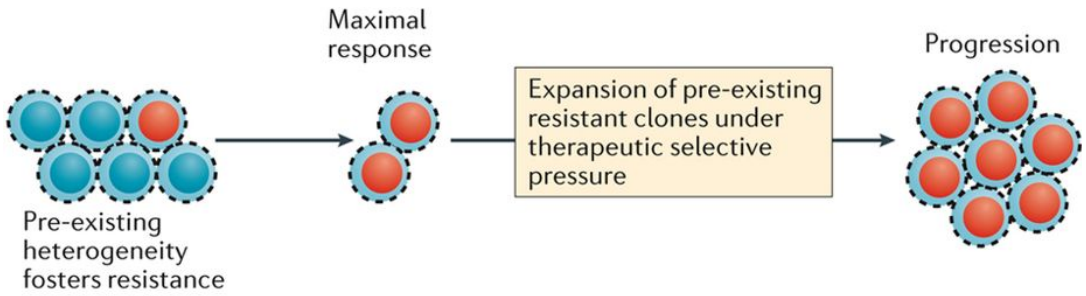


Figure 1.2: Pre-existing heterogeneity in a tumour increases the likelihood of treatment resistance due to the presence of a resistant clonal population. The red, resistant population expands following treatment due to a selective bottleneck. Used with permission from https://www.nature.com/articles/nrclinonc.2017.166.

Previous studies in HGSC have revealed a significant degree of clonal diversity in treatment-naïve tumours. On average, only 50% of mutations are shared across all samples from a tumour [27, 28]. Through detailed Bayesian probabilistic phylogenetic reconstruction of clonal lineages, our group has recently shown that tumours can harbour clones from divergent evolutionary lineages with markedly different mutational profiles [29]. Clonal diversity also exists between tumours from an individual patient, and there is significant cellular migration between metastatic sites [29]. Maximum parsimony reconstruction of clonal dissemination patterns in HGSC is consistent with monoclonal and polyclonal seeding from a single diverse site, typically an ovary or fallopian tube, in most cases [29]. However, recent organoid studies have established that multicellular aggregates (MCAs) are significantly more successful than single cells at invading ovarian mesothelium [30, 31]. These MCAs can contain phenotypically and morphologically distinct populations of cells [30]. Metastatic foci in mouse models implanted with phenotypically mixed tumour cells also contained phenotypically mixed populations [31]. Thus, successful metastatic spread likely involves multicellular aggregates rather than single cells, hinting at the possibility that polyclonal seeding and reseeding followed by pruning may be a common occurrence.

Contrary to the assumption that the intraperitoneal space allows for indiscriminate admixture of tumour cells, we have observed restricted clonal mixing in the majority of HGSC patients [29]. As such, local factors, such as the tumour microenvironment, may be involved in patterning clonal seeding and establishment. Understanding how these factors affect clonal migration may offer crucial insights into developing strategies to contend with the burden of metastatic disease in HGSC.

1.3 The tumour microenvironment in HGSC

Solid tumours are ecosystems populated by a milieu of malignant and non-malignant cell types, including tumour cells, fibroblasts, immune cells, endothelial cells, and adipocytes (**Figure 1.3**) [7]. Collectively, these cell types and the interactions that occur between them compose the tumour microenvironment (TME). Tumour cells can shape the composition of the TME, sustaining growth and proliferation while evading immune-mediated elimination [32, 33]. Likewise, the TME can impose extrinsic pressures such as hypoxia, altering the metabolic processes of tumour cells and contributing to the development of 'cancer hallmark' traits [32]. Inflammatory mediators in the TME can contribute to tumourigenesis through the pro-growth activity of cytokines released during inflammation [32]. These reciprocal interactions between tumour cells and the rest of the TME represent molecular targets that can potentially be

exploited by therapeutics.

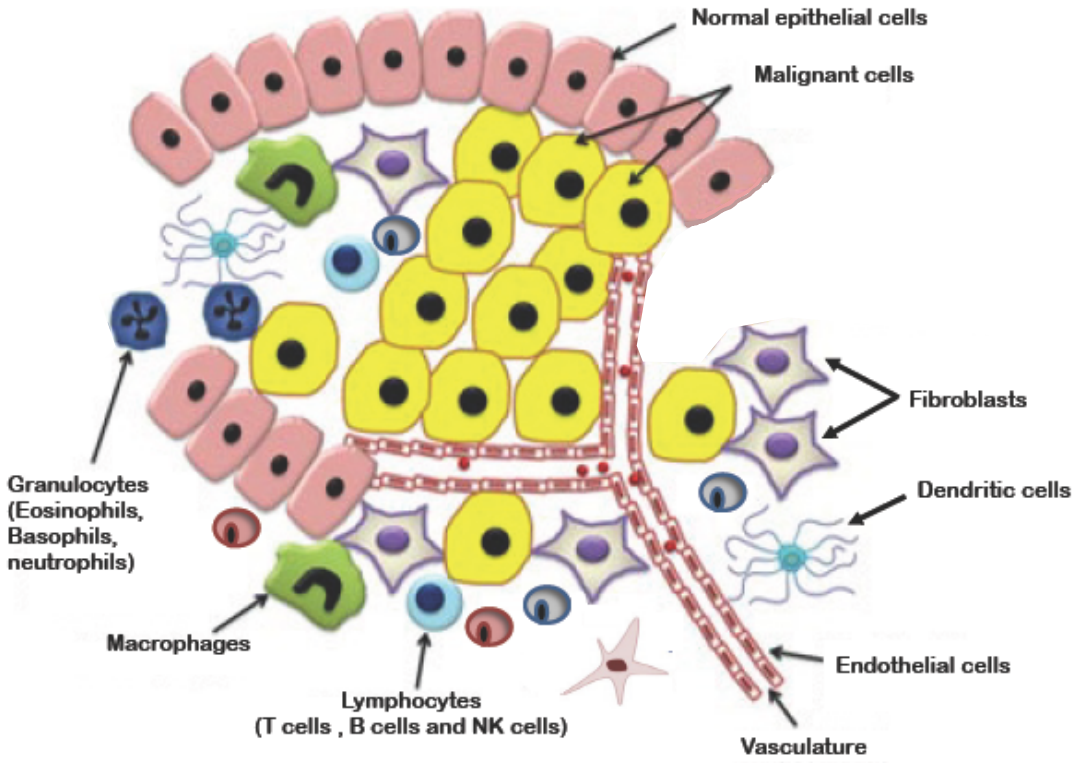


Figure 1.3: Repertoire of immune and non-immune cell types in the tumour microenvironment. Used with permission from http://tcr.amegroups.com/article/view/1549/2264.

1.3.1 The immune microenvironment

Populations of immune cells in the TME include cytotoxic T cells, helper T cells, regulatory T cells, B cells, NK cells, macrophages, and granulocytes (**Figure 1.3**) [32]. In HGSC, the immune microenvironment is dominated by T cells, B cells, and macrophages [1, 34, 35]. As in virtually every solid cancer type, cytotoxic (CD8+) tumour-infiltrating lymphocytes (TILs) have been associated with increased survival in HGSC [36–38]. Ordinarily, T cells recognize and respond to aberrant peptides presented by the Major Histocompatibility Complex (MHC/HLA). Recent studies have reported that CD8+ TILs can recognize somatically mutated peptides in HGSC, suggesting that TIL mount anti-tumour responses in part through neoantigen recognition [39]. Survival is more strongly associated with intraepithelial rather than stromal CD8+ TIL, implying that spatial localization of TIL is important for anti-tumour immunity [40]. Immunohistochemistical studies have identified that these intraepithelial CD8+ T cells

preferentially express CD103, an integrin subunit involved in epithelial localization of normal intraepithelial lymphocytes, as well as activation and cytolytic markers [40], supporting the interpretation that intraepithelial CD8+ T cells are involved in anti-tumour immunity.

However, CD8+ T cells do not function independently. In addition to CD8+ T cells, the presence of additional TIL types – B cells and plasma cells – is associated with superior outcomes in HGSC [38]. T and B cells spatially co-localized in tertiary lymphoid structures and other lymphoid aggregates may support interactions between these cell types in the context of anti-tumour immunity (**Figure 1.4**) [41]. B and plasma cells may be involved in anti-tumour immunity through autoantibody production, direct cytotoxicity, Th1/Th2 polarization, and antigen presentation (**Figure 1.4**) [42]. However, the relative contribution of each of these mechanisms in HGSC is poorly understood. On the other hand, regulatory CD4+CD25+FOXP3+ T cells curtail T effector functions by inhibiting type 1 cytokine (IFNγ, IL-2) production and T cell proliferation [34, 43, 44].

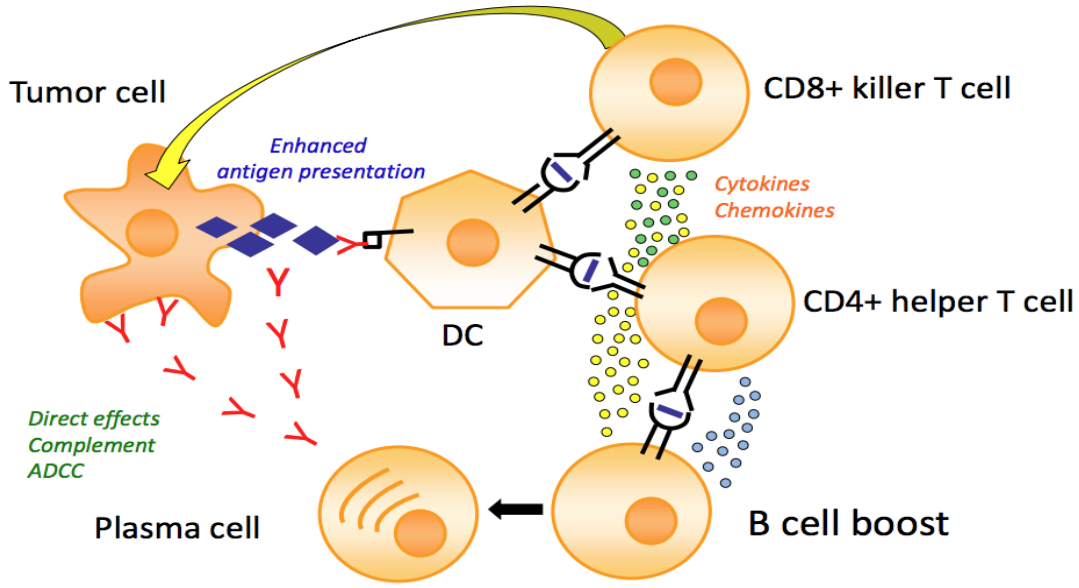


Figure 1.4: T cell- and B cell-mediated mechanisms of antitumour immunity. CD8+ T cells exert direct cytotoxic effects on tumour cells through granzyme and perforin secretion. CD4+ T cells can license dendritic cells to induce activation of CD8+ T cells, and activate B cells. Used with permission from Brad Nelson AACR 2017.

During acute infection, T cells differentiate into effector populations to mount antigen-specific responses. Following antigen clearance, these effector populations shrink to small memory T

cell pools capable of rapid reactivation upon future encounter with the same antigen. In the tumour microenvironment, however, prolonged exposure of effector T cells to antigen-depending signalling can lead to the development of an "exhausted" phenotype characterized by progressive impairment of effector activity [45]. In HGSC, CD8+ T cell infiltration is linked to expression of exhaustion markers PD-1 and CTLA4, and CD8+CD103+ TIL are associated with PD-1 positivity by immunohistochemistry [46, 47]. Nevertheless, CD8+PD-1+ TIL appear to retain some degree of effector functionality and are associated with superior patient survival [47]. The use of immune checkpoint inhibitors — monoclonal antibodies that block receptor-ligand interactions implicated in T cell exhaustion — to mobilize exhausted T cells has been associated with superior clinical trajectories compared to standard-of-care therapy in melanoma and some types of lung and colorectal cancer [8, 48]. However, despite the prevalence of PD-1+ and CTLA4+ TIL in HGSC, response rates to checkpoint inhibitor blockade in HGSC have been disappointing [49, 50]. As such, our understanding of immune checkpoints in HGSC remains incomplete.

Thus far, most studies of anti-tumour immunity have focused on the adaptive immune system. However, the observation that some patients show T cell priming to tumour-associated antigens prior to treatment ("spontaneous" T cell priming) has driven inquiry into understanding the innate immune mechanisms that ultimately give rise to intratumoural T cell infiltration. The major cell types involved in innate immunity include NK cells, macrophages, and dendritic cells [51]. NK cells are lymphocytes that express inhibitory receptors that bind to HLA complexes, and are thus thought to selectively target tumour cells that lack HLA expression. Consequently, tumour cells that downregulate HLA evade T cell recognition are vulnerable to NK-mediated cytotoxicity through granzyme and perforin release [52]. Macrophages, a type of leukocyte in the monocyte lineage, engulf abnormal cells or cellular debris through phagocytosis. The two main phenotypic subtypes of macrophages are pro-inflammatory, phagocytic M1 macrophages and tissue repair-associated M2 macrophages. In the context of the tumour microenvironment, M1 macrophages arise earlier in tumourigenesis, tend to promote inflammation and generally oppose tumour progression, whereas M2 macrophages dominate tumours at the time of diagnosis and are generally immunosuppressive and pro-tumourigenic [51]. However, the inflammatory response associated with M1 macrophages may also induce carcinogenesis [53]. In HGSC, M2-associated expression signatures have been linked to a stromal reorganization phenotype and inferior outcomes [54], whereas M1-associated genes including type I interferons are associated with superior patient survival [55]. Tumour transplantation into type I IFNR(-/-) mice resulted in reduced T cell responses against tumour antigens due to deficiencies in CD8+ T cell priming by antigen presenting cells (APCs), demonstrating that type I interferon signalling is necessary

for antitumour immunity. Studies to investigate factors upstream of type I interferon signalling in the tumour microenvironment have highlighted the STING (stimulator of DNA sensing genes) DNA sensing pathway [56]. STING is activated by the presence of cytosolic DNA – typically from intracellular pathogens, such as viruses and parasites – inducing innate immunity through type I interferon production. In tumours, STING pathway activation has been associated with the presence of tumour-derived DNA in APCs, and STING-deficient mice show defective T cell priming [57]. Thus, STING-dependent sensing of tumour DNA may lead to type I interferon production and T cell priming in cancers. However, this mechanism has yet to be shown in the context of HGSC specifically.

1.3.2 Other microenvironmental factors

Major non-immune cellular populations in the HGSC microenvironment include fibroblasts and endothelial cells. Fibroblasts are the major non-immune cellular component of the tumour stroma, involved in wound healing and responsible for extracellular matrix (ECM) deposition and maintenance through collagen and matrix metalloproteinase (MMP) production. In cancers, fibroblasts can be co-opted to facilitate tumour progression, transitioning to a myofibroblastic phenotype characteristic of cancer-associated fibroblasts (CAFs) [58]. CAFs are thought to promote tumour progression through a number of mechanisms including pro-growth signalling, vascular stabilization, and ECM remodelling [58]. Mechanistic studies have revealed a link between fibroblasts and platinum resistance in HGSC, demonstrating that fibroblasts can diminish intranuclear platinum accumulation in cancer cells *in vitro* [59].

Angiogenesis refers to the formation of new vasculature from existing blood vessels. Tumour cells rapidly proliferate, requiring the complementary development of vasculature to sustain increasing nutrient and oxygen demand. When angiogenesis cannot keep up with tumour growth, tumour regions that receive insufficient perfusion eventually become necrotic [60].

The inner lining of normal blood vessels is composed of a monolayer of squamous cells, called endothelial cells, that restrict influx and efflux between the vascular lumen and surrounding tissue. In contrast, tumour vasculature often displays an altered phenotype characterized by endothelial cell disorganization and abnormal branching [61]. The defective endothelium is associated with increased vascular permeability, faciltating nutrient extravasation and hematogenous metastasis [60]. The tumour endothelium can also act as a physical barrier to immune cell infiltration, preventing circulating lymphocytes from reaching certain tumour regions to mount anti-tumour immune responses [61]. Ultimately, the formation of tumour-associated endothelium is thought to be mediated by cancer cells. Tumour cells can alter normal endothelium by secreting pro-angiogenic and vasodilatory factors, such as vascular endothelial growth factor (VEGF), and

imposing biomechanical strain by encroaching on the vasculature itself [61]. Additionally, some groups have observed evidence of vasculogenic mimicry, where tumour cells are capable of trans-differentiating into endothelial-like cells to form "pseudo-vasculature" [62].

Cancer cell growth and metabolism can also be influenced by other microenvironmental factors. Lack of oxygenation – hypoxia – in poorly perfused regions of the tumour microenvironment leads to various tumour cell adaptations including pro-angiogenic signalling and an increased reliance on anaerobic respiration to maintain adenosine triphosphate (ATP) production [63]. Anaerobic respiration through glycolysis results in extracellular accumulation of tumour-derived lactate, impairing the ability of intratumoural T cells to perform glycolysis within hypoxic microenvironments [64]. Thus, tumour hypoxia can impair TIL activity. Moreover, lactate can induce angiogenesis and promote tumour cell migration [65]. Correspondingly, the expression of hypoxia-associated factors is associated with decreased overall survival and chemoresistance in HGSC [66].

1.4 Emerging approaches to study tumour heterogeneity and the microenvironment

1.4.1 Phylogenetic approaches for reconstructing tumour evolution

Genomic heterogeneity in cancers can be assayed with bulk or single cell sequencing. Bulk genomic sequencing generates sequencing reads from DNA derived from thousands to billions of cells that can be mapped onto a reference genome to identify germline and somatic variants (**Figure 1.5**) [67]. Currently, the most widely-used methods for bulk DNA sequencing are shotgun sequencing technologies, which generate millions of overlapping reads, providing a coarse view of genomic heterogeneity through analysis of the sequencing depth and allelic fraction of each variant (**Figure 1.5**) [67]. Recently developed single cell sequencing technologies utilize cellular barcodes that allow each read to be mapped back to its source cell so constituent genotypes of mixed cellular populations can be directly profiled [68]. However, these technologies were not available widely or at scale until recently. Furthermore, key tradeoffs between sequencing depth for variant calling and coverage uniformity for copy number profiling in single cell whole-genome sequencing, discussed in Section **1.4.3.1**, confound complete reconstruction of cell-level genotypes. Meanwhile, bulk whole-genome sequencing, which yields aggregate measurements of variant abundance for all sampled cells, has been successfully employed to characterize tumour evolution and profile driver mutations in many cancer types [1, 27, 29, 69, 70]. Evaluating genetic heterogeneity from bulk sequencing data demands statistical methods that can robustly

deconvolute clonal genotypes, abundances, and phylogenetic relationships in the presence of contaminating normal cells and aneuploidy.

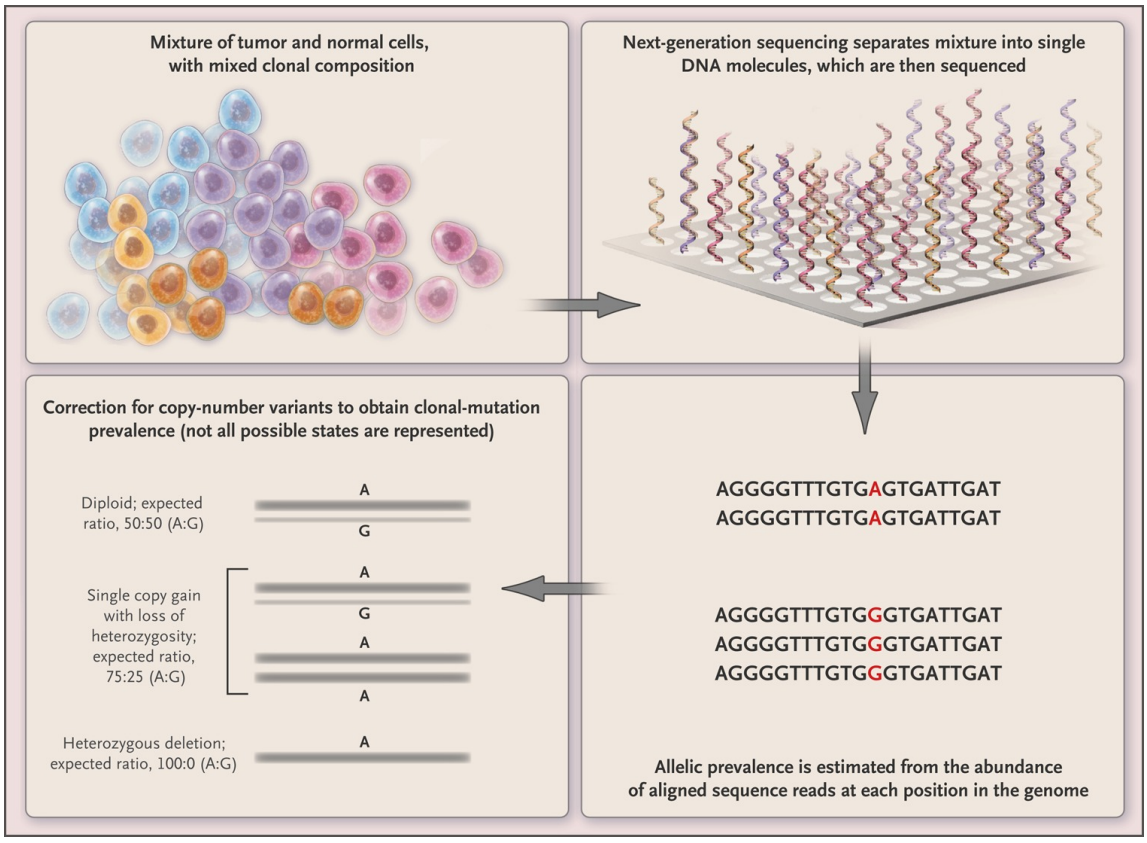


Figure 1.5: Whole-genome sequencing of heterogeneous cellular populations generates allelic read counts which are proportional to mutational prevalence (accounting for copy number and cellularity). Mutational prevalence values can be deconvolved, e.g. with PyClone [71], to yield clonal genotypes and prevalences. Used with permission from https://www.nejm.org/doi/full/10.1056/NEJMra1204892.. Reproduced with permission from [72], Copyright Massachusetts Medical Society.

Somatic variant calling from whole-genome sequencing data yields single-nucleotide variants (SNVs), short insertions and deletions (indels), copy number profiles (CNVs), and structural variant calls (SVs). In the majority of the discussion below, I focus on SNVs as these are typically the most common class of somatic variants. For each SNV, some reads map to the alternative (somatically mutated) allele, while the remainder map to the germline allele. The fraction of reads mapped to the alternative allele (henceforth referred to as the variant allele fraction, or VAF) can be used to compute the total cellular prevalence of clones harbouring the variant after accounting for the fraction of contaminating normal cells and allelic copy number

(**Figure 1.5**). Well-established methods such as PyClone [71] and SciClone [73] exploit allelic read counts and copy number profiles with Bayesian mixture models to identify mutational clusters: groups of SNVs present at similar cellular prevalence from shared clonal membership. However, clones present at similar proportions within a tumour can confound these models, due to significant overlap between clone-specific SNV clusters [71]. Data for multiple samples with shared clonal populations – for instance, from multiple metastatic sites within a patient or from different timepoints – provides greater resolution for resolving these scenarios [29, 71].

In order to construct a clonal phylogeny from SNV clusters, cluster acquisition events must be temporally ordered. Two computational models that tackle this problem are CITUP [74] and LiCHEE [75]. Fundamentally, these methods rely on a set of common assumptions: (1) the total cellular prevalence of SNV clusters from distinct lineages cannot exceed the prevalence of their most recent common ancestor (pigeonhole principle), (2) each SNV is acquired only once (infinite sites assumption) and cannot be lost, and (3) descendant SNV clusters must have lower cellular prevalence than their immediate ancestors. Other tools, such as PhyloWGS [76], perform mutation clustering and phylogeny inference simultaneously by leveraging distributions of trees to describe clonal mixing proportions. Intratumoural heterogeneity can be quantified following clonal decomposition. Heterogeneity can be expressed in terms of the relative proportions of clonal populations (mixture entropy) and in terms of the genotypic divergence between co-existing clones (phenotypic divergence) [29].

Each of the computational methods for mutation cluster inference and phylogeny reconstruction described above relies on key assumptions to reduce the search space of valid solutions. However, these assumptions may be invalid under certain realistic regimes. SciClone ignores non-diploid and copy number neutral loss-of-heterozygosity (LOH) regions of the genome [73], rendering it unsuitable for highly aneuploid cancer types such as HGSC. While PyClone uses information from copy number-altered regions, it does not allow for multiple variant clonal populations with different copy number profiles at any given locus [71]. Variants falling in regions of subclonal (non-integer) copy number may violate this assumption. Moreover, a small number of loci likely violate the infinite sites assumption imposed by phylogeny inference methods. Single cell sequencing, described in Section **1.4.3**, can help resolve clonal substructure in these regimes. For instance, a recent method called ddClone [77] jointly leverages bulk and single cell variant information to infer clonal subpopulation abundances based on genotypes informed by single cell sequencing.

1.4.2 T and B cell receptor sequencing

Lymphocyte differentiation from common lymphoid progenitors in the bone marrow proceeds through a ordered series of events marked by the acquisition of certain cell surface proteins (e.g. CD3, CD4, CD8, CD19, CD20) and antigen-specific receptor molecules [78, 79]. The eventual antigenic specificity of each lymphocyte is determined by the structure of its antigen-specific receptor – the T cell receptor (TCR) or B cell receptor (BCR) in T or B cells, respectively (**Figure 1.6**) [80]. To contend with the enormous space of potential antigens, a diverse repertoire of T and B cell receptor sequences are generated by somatic rearrangement of constituent germline variable (V), diversity (D), and joining (J) gene segments (**Figure 1.6**).

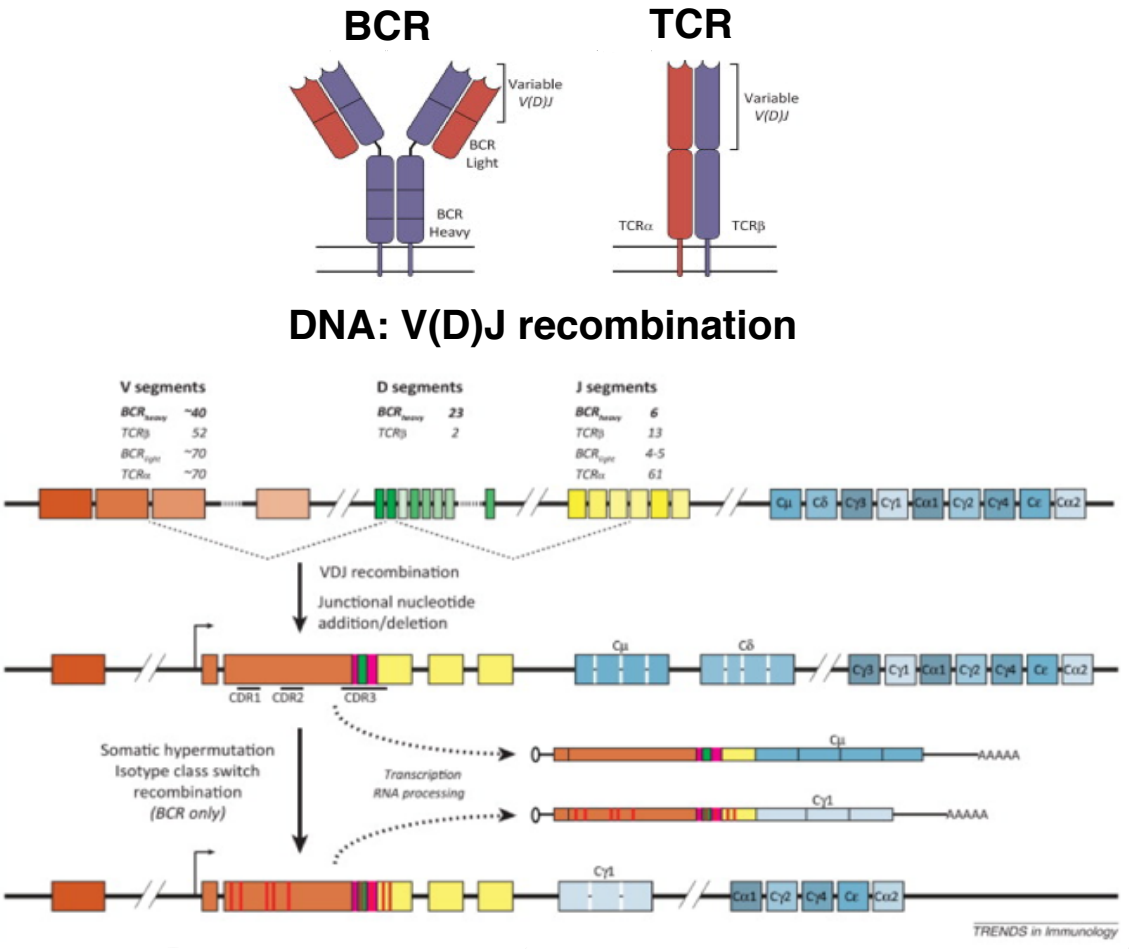


Figure 1.6: Top: Structure of a membrane-bound B cell receptor (BCR) and an α:β T cell receptor (TCR). The outer portions of the BCR/TCR, primarily composed of variable (V) chain sequence, are directly involved in epitope binding. Bottom: Depiction of the combinatorial diversity generated from V(D)J recombination and somatic hypermutation (for BCRs). Used with permission from https://www.cell.com/trends/immunology/fulltext/S1471-4906(14)00155-0.

In germline DNA, multiple V, D, and J variants are present (**Figure 1.6**). During somatic VDJ rearrangement, a D gene variant is first joined to a J gene, followed by V gene addition to the resulting D-J fragment. This process is faciltated by V(D)J recombinase, a collection of enzymes including *RAG1*, *RAG2*, TdT, and Artemis that bind to recombination signal sequences flanking V, D, and J genes [80]. For some TCR/BCR subunits, recombination occurs directly between V and J genes without the D segment. In α:β T cells – the dominant subpopulation of T cells – the TCR is composed of one VJ α subunit and one VDJ β subunit [80]. Similarly,

BCRs are composed of heavy (VDJ) and light (VJ) chains. Thus, the combinatorial space of V, D, and J germline genes forms the basis for T and B cell receptor diversity. Additional sequence diversity arises from terminal deletion and insertion events that occur at the ends of V, D, and J sequences during recombination (**Figure 1.6**) [80]. These junctional sequences, together with the end of the V and beginning of the J segments (and in TCR-β/BCR-heavy chains, the entire D gene) comprise the hypervariable portion of the TCR/BCR referred to as the CDR3 [80]. In B cells, somatic hypermutation introduces additional variants, primarily SNVs, concentrated in the CDR3 region (**Figure 1.6**) [81]. In total, the estimated sequence diversity of TCRs and BCRs generated through V(D)J recombination exceeds 10^{15} [82].

Due to the immense sequence diversity created through V(D)J recombination, the probability of two clonally unrelated T or B cells sharing the same receptor sequence is highly unlikely. Thus, TCR/BCR sequencing enables identification and quantification of clonally related families of T and B cells [83]. Expanded clonal families are thought to correspond to T and B cells that have been stimulated by antigen to proliferate. TCR/BCR repertoire profiling can be performed using either genomic DNA or RNA templates as starting material [83, 84]. Genomic DNA-based protocols enable direct quantification of lymphocyte abundance, as each T/B cell only produces one productive TCR/BCR species. However, genomic DNA protocols require V- and J-gene-specific primers that can result in PCR bias [84], and are prone to off-target priming of non-rearranged VDJ genes [85]. RNA-based protocols enrich for and amplify TCR/BCR-derived RNA or cDNA using sequence-specific primers to V and/or C genes [84]. Clone-specific read counts derived from RNA-based TCR/BCR sequencing roughly correspond to clonotype abundance, but are affected by variability in TCR/BCR expression levels [85]. Nevertheless, RNA-based methods generally capture more receptor sequence diversity than DNA-based protocols. Furthermore, RNA-based methods can utilize 5' rapid amplification of cDNA ends (RACE) PCR from constant region sequences, minimizing primer bias [84].

Following data generation, the readouts of TCR/BCR sequencing are processed by TCR/BCR clonotype calling methods to reconstruct the T/B cell clonotype repertoire. Several pipelines have been developed for TCR clonotype calling, including MiXCR [86], LymAnalyzer [87], and IMSEQ [88]. These methods rely on a similar framework: (1) initial alignment of input sequence reads to germline V, D, and J segments, (2) assignment of mapped sequence reads into clones according to sequence identity, usually by the CDR3 region, and (3) correction of sequence errors by merging clones with high sequence similarity. BCR clonotype calling can be similarly performed, but somatic hypermutation complicates clone assignment as BCR clonotypes from the same clonal family may harbour CDR3 sequences with multiple nucleotide mismatches. Thus, relaxing the similarity threshold for clonal merging may be more appropriate for calling

BCR clonotypes [86]. Alternatively, a recently developed approach for BCR repertoire inference uses Hidden Markov Models to describe the generative process of V(D)J recombination and somatic hypermutation [89, 90]. However, this approach is currently slow for large datasets (execution time of hours for datasets with $> 10^5$ reads).

Most TCR/BCR sequencing methods target the TCRβ and BCR-heavy chains, as these contain D genes and thus have greater diversity than TCRα and BCR-light chains. However, T cells with the same TCRβ chain can harbour different TCRα chains. A recently developed assay from Adaptive Biotechnologies, pairSEQ, allows for paired sequencing of TCRα and β chains from the same individual cells [91]. Single cell TCR/BCR sequencing also permits direct assessment of α:β pairing.

1.4.3 Single cell methods

1.4.3.1 Single cell DNA sequencing

Single cell DNA sequencing aims to bring the readouts of bulk genome sequencing – SNVs, CNVs, and SVs – to the cellular level. In the context of cancer genomics, single cell DNA sequencing offers distinct advantages over bulk sequencing in identifying rare clonal populations in heterogeneous tumours and resolving phylogenetically divergent clonal mixtures to understand the mechanistic bases of tumour progression (**Figure 1.7**). Analysis of single cell DNA sequencing data has provided insights into patterns of intratumoural genomic heterogeneity in patient-derived xenograft models [29, 92], spatial invasion of breast cancer clones [93], and chemoresistance in triple-negative breast cancer [94].

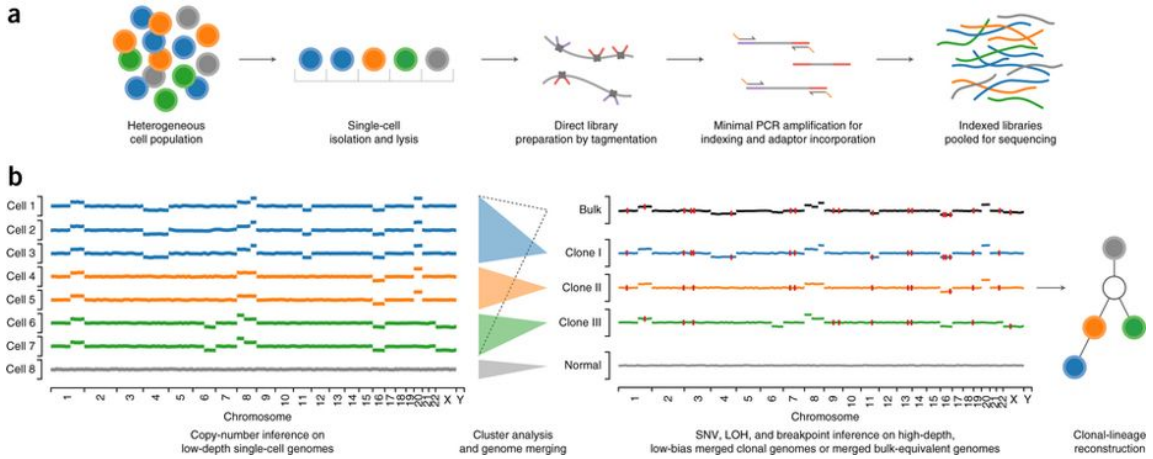


Figure 1.7: a) Depiction of single cell WGS library preparation by DLP. Single cells are isolated in individual wells, and lysed. The resulting DNA is tagmented prior to amplification to allow for computational identifiation of PCR duplicates and thus accurate recovery of copy number variants. b) Single cells can be clustered into clones with similar copy number profiles; the resulting clonal consensus copy number profiles can be used to build a phylogenetic tree. Used with permission from https://www.nature.com/articles/nmeth.4140/figures/introduction/1.

To date, no singular single cell DNA sequencing technology has been widely adopted across cancer genomics. Generally, single cell DNA sequencing technologies aim to optimize coverage depth, breadth, uniformity, and accuracy to faithfully recapitulate single cell genotypes. However, virtually all single cell DNA sequencing technologies exhibit key tradeoffs in one or more of these areas. Most single cell DNA sequencing workflows can be divided into 3 steps: (1) cell isolation, (2) DNA amplification, and (3) amplicon sequencing and interpretation (**Figure 1.7**) [95]. Differences in how DNA amplification is performed primarily underlie the key tradeoffs associated with single cell DNA sequencing technologies. Isothermal amplification methods such as multiple displacement amplification (MDA) are highly sensitive and generate high coverage depth, allowing for detailed interrogation of SNVs and indels at the cost of coverage uniformity [95]. Consequently, isothermal methods are unsuitable for CNV assessment. On the other hand, degenerate oligonucleotide primed PCR (DOP-PCR) employs thermocycling to recover amplification products with lower coverage bias but inferior depth [96]. Further improvements in coverage uniformity were provided by direct library preparation (DLP), a nanolitre-volume protocol carried out in a microfluidic device requiring no pre-amplification, designed for single cell WGS (**Figure 1.7**) [97]. DLP has been successfully employed to recover whole genome copy number profiles of cell lines and patient-derived xenograft samples [98]. "Pseudobulk" aggregation of single cell readouts from DLP faithfully recapitulates bulk whole

genome sequencing SNV profiles at similar coverage depth [97]. DLP can distinguish clonal populations of cancer cells defined by distinct copy number profiles, which can be leveraged to reconstruct pseudobulk SNV, CNV, and SV profiles at the clone level (**Figure 1.7**). A recently developed commercial assay from 10x Genomics also recovers whole genome CNV profiles at the single cell level [99], but the performance of this method has yet to be rigorously validated.

Single cell DNA sequencing can also be targeted to particular regions of the genome through target-specific amplification or capture, enabling superior coverage depth and breadth in targeted regions at the cost of uniformity [95]. This enables long-range variant phasing, which can be employed to validate clonal genotypes proposed by statistical deconvolution of bulk genomes [77, 92].

1.4.3.2 Single cell transcriptomics

Transcriptomics provides measurements of cellular phenotype through quantification of relative RNA abundance. Contemporary bulk transcriptome technologies such as microarrays and RNA sequencing (RNA-seq) have enabled quantitation of gene expression in tumour samples, establishing prognostically relevant transcriptomic subtypes and microenvironmental properties of many cancers [100–103]. However, the aggregate measurements provided by bulk transcriptomics are affected by cell type composition, making direct interrogation of malignant, immune, and stromal phenotypes difficult (**Figure 1.8**). Single cell RNA-sequencing generates whole transcriptomes at the single cell level, enabling direct assessment of individual cellular phenotypes, tissue composition, gene regulation, and cell state evolution throughout development and differentiation (**Figure 1.8**).

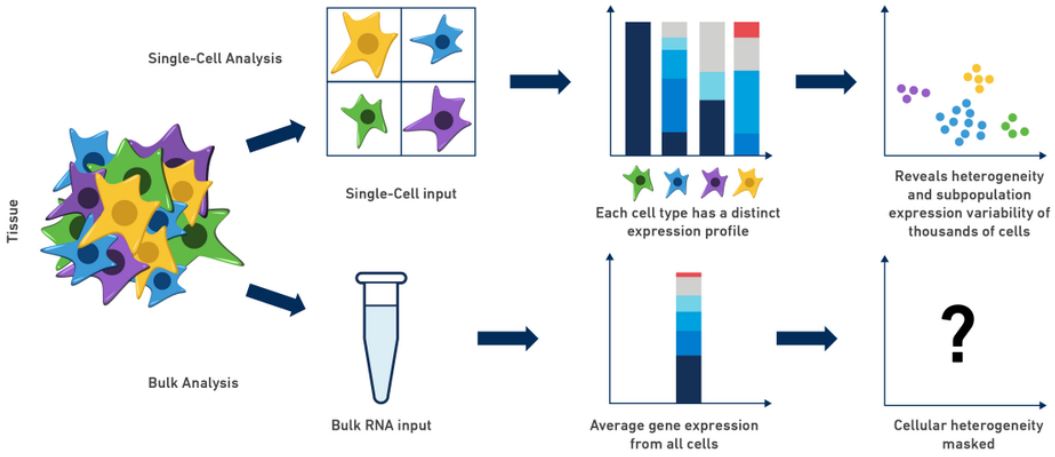


Figure 1.8: When applied to heterogeneous cellular populations, single cell RNA-sequencing can simultaneously recover single cell transcriptomes and cell type proportions in a nearly unbiased manner. In contrast, bulk RNA-seq recovers average expression profiles. Deconvolution methods can recover cell type proportions, but these usually require prior information on cell type expression profiles [104]. Image provided by 10x Genomics from https://community.10xgenomics.com/t5/10x-Blog/Single-Cell-RNA-Seq-An-Introductory-Overview-and-Tools-for/ba-p/547.

Over the past decade, single cell mRNA-seq technologies have matured from digital transcriptomic assessment of a single cell to droplet-based technologies capable of profiling thousands of cells per run [105]. Most single cell RNA-seq protocols can be divided into 4 main steps: (1) initial sample preparation, (2) single cell capture, (3) nucleic acid extraction and amplification, and (4) sequencing of amplified products; primarily differing from one another in (2) and (3). Fluorescence-activated cell sorting (FACS), microdroplet and microfluidic technologies enable high-throughput capture of hundreds to millions of single cells, but require dissociated tissue samples as input [105]. The enzymatic treatments used in tissue dissociation can introduce phenotypic changes marked by upregulation of immediate early genes (IEGs) such as *FOS* and *JUN* [106, 107]. Laser-capture microdissection (LCM) and micropipetting enable single cell capture from intact tissue specimens, but require manual isolation of single cells [105]. A recently developed method, SPLiT-seq, uses combinatorial barcoding to bypass cell capture altogether [108]. Following single cell capture and cell lysis, mRNA can be amplified by poly-T priming and second strand synthesis or 5' template switching synthesis. Template switching amplification, employed in the SmartSeq and STRT-Seq protocols, enables full-length transcript coverage with reduced bias while other methods suffer from 3' bias due to incomplete reverse transcription [109]. Recently, 10x Genomics has released a droplet-based commercial platform

for single cell RNA-seq that quantifies the abundance of 3' transcript fragments for thousands of cells per sample [110].

Imperfections in cell capture, reverse transcription, and amplification present unique technical challenges for single cell RNA-seq data analysis. During cell capture, multiple or dead cells may be collected in place of a single viable cell [105]. Cell lysis introduces ambient RNA that can contaminate libraries generated from live cells [110]. Due to Poisson sampling, many transcript species may not be reverse transcribed prior to amplification, leading to transcript dropout unremedied by increasing sequencing depth. Amplification bias can also lead to dropout for similar reasons [105]. Moreover, single cell RNA-seq data generated from different centers, reagent pools, or reaction chips is not directly comparable due to sizeable batch effects [111]. These batch effects can overwhelm or obscure subtle phenotypic differences between similar cell types or states.

As such, proper quality control is a critical step in single cell RNA-seq analysis. Ruptured cells are first removed by identifying libraries enriched for mitochondrial transcripts, indicating loss of cytoplasmic transcripts due to increased cell membrane permeability [112]. Doublets can be identified experimentally by imaging in plate- or well-based protocols or cell hashing with barcoded antibodies [113], and computationally by expression of mutually exclusive cell type markers. A recently developed tool, SoupX, removes signal from ambient RNA [114]. Some models for single cell RNA-seq data employ negative binomial distributions with zero-inflation to model transcript counts subject to dropout [115]. Alternatively, imputation approaches such as MAGIC attempt to correct for dropout by using information from similar cells [116]. Many different batch correction methods have been employed in single cell RNA-seq analysis. Batch effect correction by linear regression [117] can improve concordance between datasets generated from similar input material across different centers, but can also introduce biases when working with samples with different cellular composition. In these scenarios, more sophisticated batch effect correction methods that adjust transcript expression based on shared cellular populations across batches can be employed [118–120].

The readouts of single cell RNA-seq can be used to understand tissue composition, developmental trajectories, and gene networks. Single cell RNA-sequencing allows for virtually unbiased assessment of cell types by dimensionality reduction and subsequent unsupervised clustering. This approach has been employed in various tissue types and organisms to quantify known and novel cell types [110, 121–123]. In the cancer context, these methods have been used to profile the phenotypic subsets of immune and stromal cells in the microenvironment and their relationships with patient survival [124–126]. Algorithms that model continuous transitions between cell states – pseudotime algorithms [127–130] – have been developed to delineate phenotypic changes

that occur during cell differentiation from stem cells to terminally differentiated cells [131]. The wealth of individual measurements provided by contemporary single cell RNA-seq methods has enabled regulatory network reconstruction at unprecedented scale [132]. The networks identified by these algorithms could be feasibly used to improve orthogonally collected single cell RNA-seq data through imputation [133]. In summary, single cell RNA-seq enables simultaneous profiling of both cancer cell and microenvironment phenotypes to study cancer-microenvironment interplay.

1.5 Problem statement

Despite extensive effort being made to identify new therapeutic targets for HGSC, long-term outcomes have remained bleak for many patients. Crucially, most patients present with advanced stage disease characterized by within- and between-site heterogeneity [27, 29]. This genomic heterogeneity provides considerable substrate for selection to act on, and is thus thought to lead to the development of resistance. However, the observation that tumour-infiltrating lymphocyte abundance is associated with superior outcomes [36] hints at the tantalizing possibility that intratumoural immune infiltration may be able to contend with genomic heterogeneity.

Apart from T cells, the microenvironmental properties of HGSC and their associations with genomic and clinical features remain poorly understood. Initial efforts have been made to understand the role of B cells and plasma cells in HGSC [38, 41], but their antigenic targets and interacting partners in the HGSC microenvironment are unclear [42]. Even less is known about non-immune cell types such as fibroblasts and endothelial cells. Transcriptome-based subtyping of HGSC by independent groups identified 4 prognostically distinct subgroups largely distinguished by immune and stromal markers [14, 101], implying that stromal cell types also influence disease progression in HGSC.

In this thesis, I set out to investigate the evolutionary interplay between malignant and non-malignant cells in HGSC. Many HGSC cases present with disseminated disease, providing an incredible opportunity to study tumour evolution across distinct peritoneal microenvironments. In Chapter **2**, I describe the assembly of a collection of 148 tumours from 41 patients, the largest multi-site HGSC cohort to date I am aware of (**Figure 1.9**). I co-ordinated an integrated team of clinical and research personnel to identify and notify the team of potential HGSC cases for collection on a weekly basis, and helped devise methods for single cell processing and data curation from these samples. I outline the steps from clinical case identification to sample processing that generated the experimental substrate used in the following chapters. In addition, these samples will be used in future work involving drug testing of patient-derived xenograft

models and single cell whole-genome sequencing. While patient-derived xenografts are poor models for directly studying tumour-microenvironment interaction, they enable drug sensitivity testing of clones from different microenvironmental contexts. In Chapter **3**, I interrogate the interface of lymphocytic and malignant evolutionary dynamics in HGSC, described in a *Cell* publication (**Figure 1.9**) [1]. I was responsible for the bulk of computational work described in this chapter, including most of the integrative analysis, clonal inference, human leukocyte antigen loss-of-heterozygosity analysis, gene expression, TCR/BCR-seq analysis, and large portions of the histologic image analysis. The last section of this thesis extends the work of Chapter **3** to other non-malignant cell types using single cell RNA-sequencing (**Figure 1.9**). I describe a probabilistic method for identifying known cell types from single cell RNA-seq data in Chapter **4**, demonstrating its utility on simulated data. Finally, I apply this approach to single cell RNA-seq data from spatial samplings collected in Chapter **2** to comprehensively characterize the HGSC microenvironment. I led the work desribed in this chapter, helping formulate the model and conduct most of the analysis on simulated and real data.

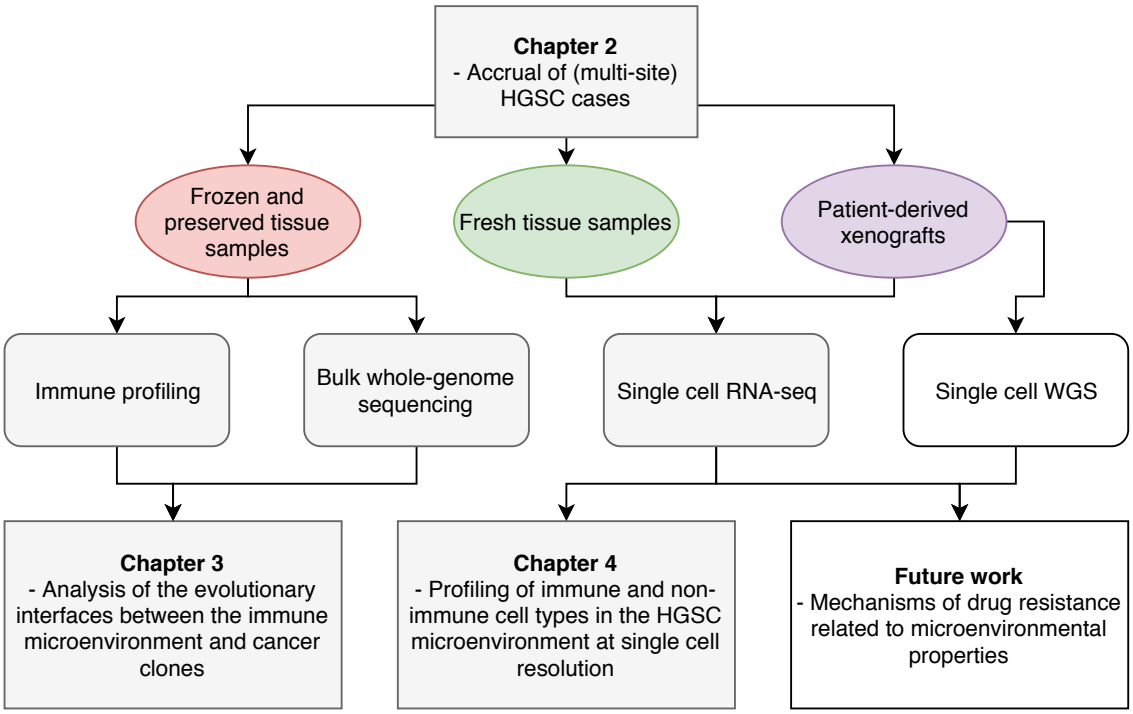


Figure 1.9: Outline of relationships between thesis chapters. White boxes (single cell WGS and future work) correspond to elements that were not performed as part of this thesis.

Chapter 2

Collection and processing of multi-site HGSC samples for high-throughput sequencing, PDX creation, and single cell experiments

2.1 Introduction

The high prevalence of multi-site disease and well-described site-to-site genomic, transcriptomic, and microenvironmental variation in HGSC necessitates study of multiple tumour foci from the same patient to understand disease pathogenesis. However, only a handful of groups have attempted to conduct multi-site studies of HGSC, and these studies have been restricted to small cohort sizes [27–29]. These initial studies have exemplified the degree of inter-site heterogeneity in HGSC and raised questions on how this heterogeneity affects prognostically relevant associations between genomic features and the tumour microenvironment [29]. Systematic collection of multi-site HGSC cases at scale will be required to decipher the evolutionary mechanisms by which HGSC tumours develop treatment resistance and thwart immunologic surveillance *in vivo*.

Patient-derived xenograft (PDX) models are laboratory mice transplanted, usually subcutaneously or subcapsularly, with human tumour cells. Under the assumption that these models faithfully recapitulate the phenotypic properties of their source tumours, PDXs serve as malleable systems for studying tumour evolution and drug response. Most PDXs are constructed from immunodeficient mice to prevent transplant rejection, but newer methods enable establishment of 'humanized' PDX models that contain human-like immune systems. Thus far, PDXs that recapitulate genomic properties of their source tumours have been established for several cancer types including ovarian cancer [134], breast cancer [135], and B cell lymphomas [136].

Despite extensive profiling of clonal heterogeneity in HGSC [27–29], the genomic and

transcriptomic properties of clones associated with treatment resistance remain unknown. Identifying the hallmarks of clones associated with treatment resistance and dissemination may provide critical insights into predicting response and personalizing therapeutic regimens for HGSC patients. One of the aims of this chapter will be to build PDXs for each tumour in a cohort of multi-site HGSC patients in order to study tumour evolution in response to treatment pressure. These PDXs will serve as an ideal substrate for interrogating the relative fitness of clonal genotypes and the reproducibility of clonal dynamics between clones derived from different tumour microenvironments in response to external selection pressure.

In order to supplement the tumour cell-focused view of tumour evolution provided by PDX modeling, another aim of this chapter is to create experimental substrates and methods for profiling the tumour microenvironment of HGSC. To date, most studies of the HGSC microenvironment utilize histologic image analysis for cell type quantification or bulk gene expression profiling for phenotypic analysis [14, 36, 38, 41, 101]. However, routine histologic image analysis and immunohistochemistry can only capture a limited number of cell types, and deconvolving cell type proportions and transcriptomes from bulk gene expression profiles is difficult. Single cell RNA sequencing, with technologies such as SMART-Seq [109], Drop-Seq [121], and 10x Chromium [110], enables simultaneous capture of cell type abundances and transcriptomes, but its use for studying solid tumours, especially ovarian cancers, is limited. Most single cell RNA-seq experiments thus far have utilized material from peripheral blood or mouse models, which yield high quality data due to the minimal extent of manipulation required to obtain viable single cell suspensions. In the context of profiling gross HGSC tissue specimens, methods for preparing single cell suspensions and libraries must be optimized to minimize technical effects on microenvironmental composition and phenotypes [106, 107].

With the goal of profiling the pre-treatment microenvironment and clonal dynamics of HGSC in response to treatment, we systematically collected a cohort of multi-site HGSC cases. In this chapter, I outline the process of sample accrual, from case identification to sample processing and PDX construction, that served as the basis for the work described in Chapter **3** and Chapter **4**.

2.2 Materials and Methods

2.2.1 Summary of accrual process

Surgeons and senior surgical residents identified potential HGSC cases at local hospitals, including Vancouver General Hospital (VGH) and the University of British Columbia (UBC) Hospital. Consents and peripheral blood were obtained for each patient prior to surgery, and cases were

prioritized if possible as first or second on surgical slates.

Surgeons sent debulking specimens for initial processing by hospital research assistants and medical technologists. Following initial pathologic assessment to confirm diagnosis of HGSC, specimens were transferred to the BC Cancer Research Center (BCCRC) for further laboratory work involving preparation and preservation of material for bulk sequencing, PDX establishment, and single cell RNA sequencing. Final pathologic assessment was performed by a trained pathologist at VGH and non-HGSC cases were retroactively removed from the study.

The overall accrual pipeline is shown in **Figure 2.1**. Each of these steps are described further in the sections below.

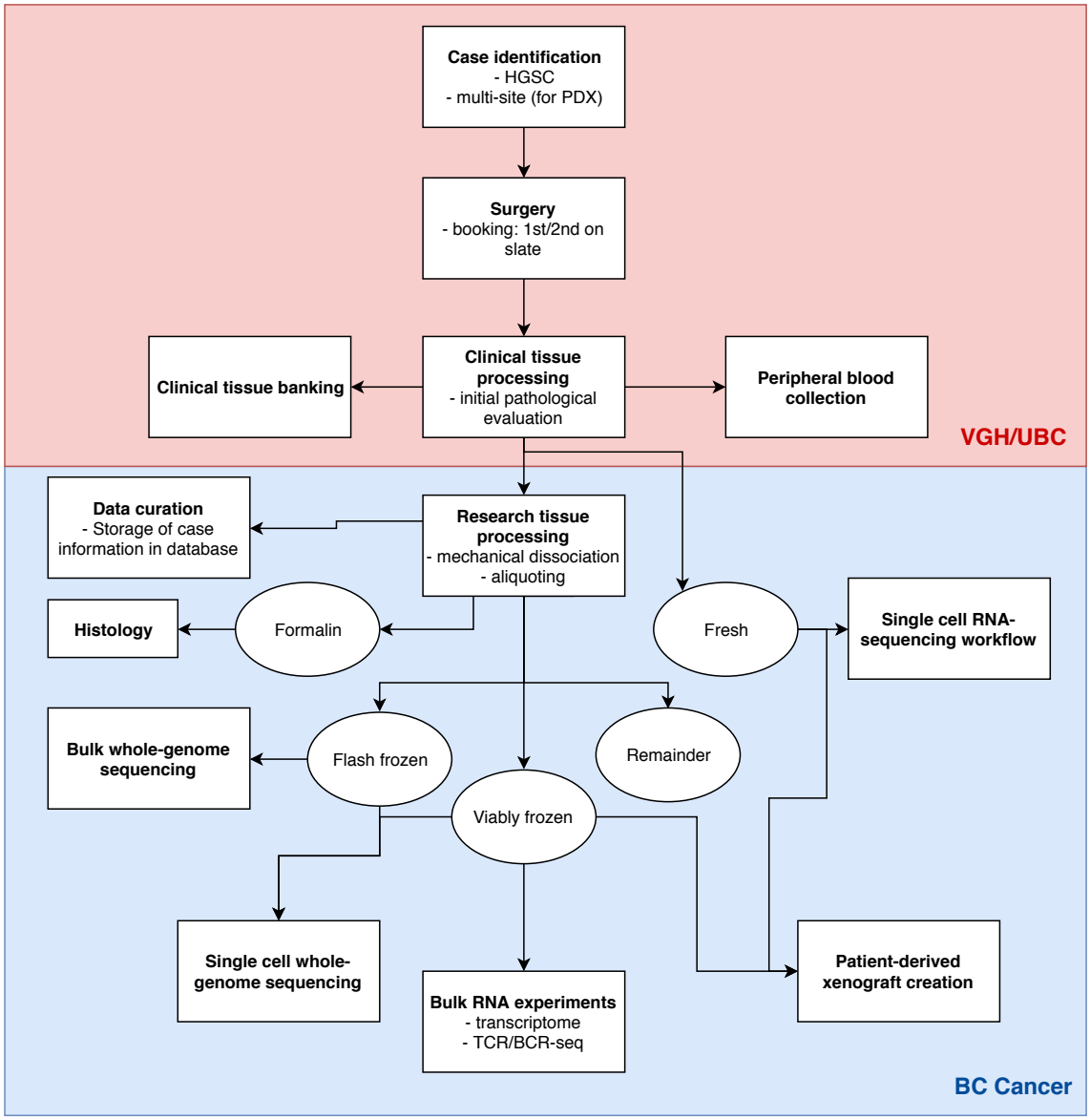


Figure 2.1: Clinical and research pipelines for processing high-grade serous ovarian cancer cases. Steps in red are executed at the hospital (either Vancouver General Hospital or the University of British Columbia Hospital) by clinical personnel; steps in red are carried out at BC Cancer by research personnel.

2.2.2 Patient cohort

Pre-operatively, patients were screened based on the following criteria: (1) clinical suspicion of HGSC based on history, imaging and blood work, (2) no prior treatment, i.e. chemo- or radiotherapy, and (3) patient consent. All cases – except for those dedicated for single cell RNA-seq pilot experiments – were additionally required to have at least 2 tumour foci from anatomically distinct masses or distal regions of a single mass that could be collected. Recurrence specimens were obtained from patients that presented for a subsequent operation related to their disease.

2.2.3 Collection of surgical specimens and peripheral blood

Patient consent was obtained prior to specimen collection and banking and documented at VGH or UBC Hospital Laboratory (Research Ethics Board numbers H08-01411 and H18-01090). Specimens of consented patients were placed into cold media and brought to the clinical laboratory by the messenger porter. Following this, each specimen was assigned a unique research identifier and processed as per VGH/UBC Anatomical Pathology specimen handling procedures (**Figure 2.1**).

Each case was initially assessed to determine whether or not the disease was HGSC and if sufficient material was available for research purposes. Specimens for cases where sufficient material was available from multiple tumour foci (or a single tumour focus for single cell RNA-seq experiments) were considered eligible. For cases with multiple sites, each site was sent out individually upon collection to minimize delay to sample processing.

Peripheral blood was separately collected in purple/pink top (plasma and buffy coat) and red top (serum) tubes (**Figure 2.1**). Blood components were spun down and transferred into labelled cryovials, snap frozen in liquid nitrogen and stored in the -80°C freezer.

2.2.4 Sample preparation

Each specimen was assigned a unique anonymous research identifier linked to a case identifier (Section **2.2.12**). Specimens were placed in a Petri dish and measured. One millilitre cryovials corresponding to each aliquot type to be created (formalin, flash frozen, transplant, viable frozen, and remainder) were prepared (**Figure 2.1**). A 1mm piece was first cut and placed in the formalin cryovial containing 1mL formalin. The remaining pieces were chopped finely on a cell culture dish and used to create the remaining aliquots.

A small portion of the finely chopped tissue was aliquoted into a stomacher bag containing 1mL of media, while the remaining tissue was set on ice for single cell dissociation (and single

cell RNA-seq). The stomacher machine was run for 60 seconds at normal settings to further dissociate the sample. One hundred microlitres of the supernatant was added to the transplant vial, with the rest aliquoted to the viable frozen vial. The remaining chunks of tissue in the stomacher bag were added to the remainder vial. Following this, the transplant and viable frozen vials were spun down and 1mL of freezing media was added to each vial. Transplant, viable frozen, and remainder vials were placed in a Mr. Frosty freezing container (Thermo Scientific) to be gradually frozen overnight, and transferred to the -80°C freezer the next day. Flash frozen vials were placed into a cryobox and stored directly in the -80°C freezer. Formalin vials were sent for embedding and hematoxylin and eosin (H&E) staining.

2.2.5 Patient-derived xenograft creation

For each specimen, the aliquot set aside for xenografting was used for PDX construction. When possible, PDXs were created from freshly processed aliquots; otherwise, aliquots set aside for xenografting were viably frozen for transplantation at a later date.

Each transplantation vial was divided into 4 equally-sized aliquots of approximately 250 microlitres each in Eppendorf tubes. Aliquots were spun down for 5 minutes at 1200rpm. Pellets were resuspended in 200 microlitres of 50% Matrigel and kept on ice until transplantation.

Each aliquot was subcutaneously injected into a NOD.Cg-$Prkdc^{scid}Il2rg^{tm1Wjl}$/SzJ (Nod-Scid-gamma, NSG) or NOD.Cg-$Rag1^{tm1Mom}Il2rg^{tm1Wjl}$/SzJ (Nod-Rag-gamma, NRG) mouse aged 5-12 weeks with a 21-gauge needle. Mice were placed in cages (up to 4 mice per cage, identified by ear punching) and monitored weekly initially and more frequently as humane or experimental endpoints were reached. Following euthanasia, mice were biopsied and tumours were collected and frozen.

2.2.6 Whole-genome sequencing of patient tumours

DNA was extracted from flash frozen tumour sample aliquots using the Qiagen Blood and Tissue Extraction Kit. DNA samples were submitted for sequencing at the BC Genome Sciences Centre (BCGSC). For all tumor and corresponding normal (blood) samples, sequencing was performed using Illumina HiSeq2500 whole genome shotgun v4 chemistry with paired-end 125bp reads. Samples were sequenced to an average of 96X coverage [1].

2.2.7 Single cell RNA-seq pilot project

For all suspected HGSC cases (single and multi-site), a portion of each specimen was set aside for single cell suspension creation and subsequent library preparation with the 10x Genomics

3' or 5' gene expression kits [110]. Various sample dissociation times, enzyme mixture, and viability assessment methods were piloted (**Table 2.2**). The workflow is shown in **Figure 2.2**.

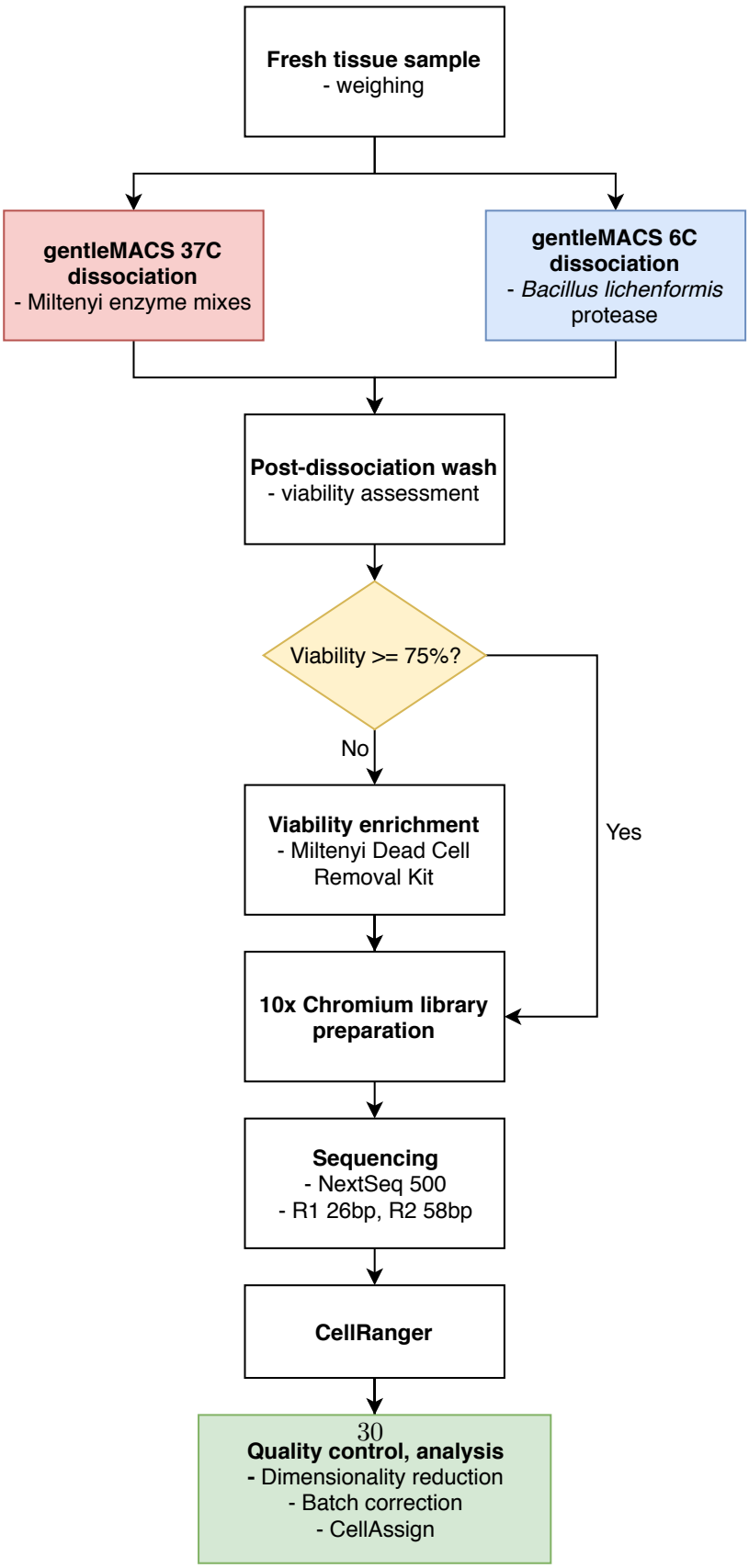


Figure 2.2: Single cell RNA-sequencing workflow from fresh tissue samples to analysis.

2.2.8 Single cell dissociation

For each specimen, the portion set aside for single cell RNA-seq was used to prepare a single cell suspension.

2.2.8.1 37C protease

After weighing in a cell culture dish, tissue was transferred into a gentleMACS C tube and pipetted up and down using a wide bore pipette tip. GentleMACS programs h_tumour_01, h_tumour_02, and h_tumour_03 were run, with samples incubated for 30 minutes at 37°C under continuous rotation using the MACSmix Tube Rotator between programs. Miltenyi Biotec enzymes H (200μl), R (100μl) and A (25μl) were used for dissociation. Following dissociation, cells were assessed for viability using the cell counter (5μl cells + 5μl trypan blue) under a microscope. Cells were then pelleted by centrifugation for 5 minutes at 4°C, resuspended in freezing media, placed in Mr. Frosty overnight, and frozen at -80°C.

2.2.8.2 6C protease

After weighing in a cell culture dish, tissue was transferred into a gentleMACS C tube, and one millilitre of 10 mg/mL *Bacillus lichenformis* protease (Creative Enzymes NATE-0633; henceforth referred to as 6C protease) was added to each 25 mg of tissue. The resulting solution was incubated and mechanically disrupted at 6°C. Depending on the sample, two different protocols were used for mechanical disruption. The first protocol involved pipetting up and down for 15 seconds every minute for a total of 15 minutes. The second protocol utilized the Miltenyi Biotec MACS Separator (programs h_tumour_01, h_tumour_02, h_tumour_03) with the 6C protease for 30 minutes or 1 hour. Dissociation specifications for each sample are listed in **Table 2.1**. Following dissociation, cells were assessed for viability using the cell counter (5μl cells + 5μl trypan blue) under a microscope. Cells were then pelleted by centrifugation for 5 minutes at 4°C, resuspended in freezing media, placed in Mr. Frosty overnight, and frozen at -80°C.

Patient	**Sample**	**Anatomic site**	**Digestion**	**10x Method**
62	VOA11019SA-37	RLQ site metastasis	37C collagenase 4h	3' GE
62	VOA11019SA-CD3	RLQ site metastasis	37C collagenase 4h	3' GE
62	VOA11019SA-CD45	RLQ site metastasis	37C collagenase 4h	3' GE

62	VOA11019SA-6	RLQ site metastasis	o/n 6C	3' GE
63	VOA11095SA	Posterior Cul de Sac	o/n 6C	
63	VOA11095SB	Splenic Stricture	o/n 6C	
63	VOA11095SC	Left Ovary	o/n 6C	
63	VOA11095SD	Epiploica Sigmoid	o/n 6C	
63	VOA11095SE	Right Ovary	o/n 6C	
63	VOA11095SF	Omentum	o/n 6C	
64	VOA11213SA	Ovary	MACS 37C 1h	5' GE
64	VOA11213SB	Omentum	MACS 37C 1h	5' GE
64	VOA11213SC	Bowel	MACS 37C 1h	5' GE
64	VOA11213SC	Bowel	37C collagenase 1h	
65	VOA11083A	Pelvic	o/n 6C	
65	VOA11083B	Pelvic	o/n 6C	
65	VOA11083C	Right Ovary	o/n 6C	
65	VOA11083D	Omentum	o/n 6C	3' GE
65	VOA11083E	Cecum	o/n 6C	
66	VOA11088A	Left Fallopian Tube	o/n 6C	
66	VOA11088B	Omentum	o/n 6C	
67	VOA11220SA	Right Ovary	MACS 37C 30min	5' GE
67	VOA11220SB	Gastric Nodule	MACS 37C 30min	
67	VOA11220SC	Omental Nodule	MACS 37C 30min	5' GE
67	VOA11220SD	Rectal Sigmoid	MACS 37C 30min	5' GE
68	VOA11243SA	Uterus Surface	6C 1hr	
68	VOA11243SB	Right Ovary	6C 1hr	
68	VOA11243SC	Right Tube	6C 1hr	
68	VOA11243SD	Left Ovary	6C 1hr	
68	VOA11243SE	Omentum	6C 1hr	
68	VOA11243SF	Pouch of Douglas	6C 1hr	
68	VOA11243SA	Uterus Surface	6C O/N	

68	VOA11243SB	Right Ovary	6C O/N	
68	VOA11243SC	Right Tube	6C O/N	
68	VOA11243SD	Left Ovary	6C O/N	
68	VOA11243SE	Omentum	6C O/N	
68	VOA11243SF	Pouch of Douglas	6C O/N	
69	VOA11265SA	Omentum	MACS 6C 1hr	
69	VOA11265SB	Left fallopian tube nodule	MACS 6C 1hr	
70	VOA11267-6	Left adnexal mass	MACS 6C 1hr	5' and 3' GE
70	VOA11267-37	Left adnexal mass	MACS 37C 1hr	5' and 3' GE
71	VOA11294SA	Left Ovary	MACS 6C 1hr	
71	VOA11294SA	Left Ovary	MACS 37C 1hr	
71	VOA11294SB	Small Bowel Tumour	MACS 6C 1hr	
71	VOA11294SC	Right Ovary	MACS 6C 1hr	
71	VOA11294SC	Right Ovary	MACS 37C 1hr	
71	VOA11294SD	Left Ovarian Tumour	MACS 6C 1hr	
71	VOA11294SD	Left Ovarian Tumour	MACS 37C 1hr	
71	VOA11294SE	Surface Uterine Tumour	MACS 6C 1hr	
71	VOA11294SF	Omentum	MACS 6C 1hr	
71	VOA11294SF	Omentum	MACS 37C 1hr	
72	VOA11558SA	Omentum	MACS 6C 1hr	
73	VOA11543SA	Left Ovary	MACS 6C 1hr	5' GE
73	VOA11543SB	Right Ovary	MACS 6C 1hr	5' GE
C2	VOA11229	Right Ovary	MACS 6C 30mins	
C2	VOA11229	Right Ovary	MACS 6C 1 hour	
E2	VOA11520SA	Right Ovary	MACS 6C 1hr	
E2	VOA11520SA	Right Ovary	MACS 37C 1hr	

Table 2.1: Identifiers of samples used for 10x Chromium library preparation and/or sequencing. The tissue dissociation protocols and types of 10x Chromium library preparation are listed for each sample.

2.2.8.3 Post-dissociation wash

Enzymatically dissociated samples were thawed, spun down, and washed with 1mL PBS twice to remove the dimethyl sulfoxide (DMSO) present in freezing media. Samples were then diluted with cold HFN and washed with trypsin, dispase, and DNAse while gently pipetting up and down. Cold ammonium chloride was added to bloody samples. Cells were assessed for viability using the cell counter (5μl cells + 5μl trypan blue) under a microscope, and kept on ice.

2.2.9 Viability sorting and assessment

Viability sorting was performed for samples with <75% viability after post-dissociation wash, with a target viability of ≥75% viability (**Figure 2.2**). Cells were spun down and the pellet resuspended in 100μl of Miltenyi Dead Cell Removal MicroBeads and incubated at room temperature for 15 minutes. Viable cell enrichment was performed using the positive selection column type MS with a MACS Separator. Cells were then placed on ice for 10X Genomics scRNAseq library preparation.

2.2.10 Single cell RNA-seq library preparation and quality control

Single cell RNA-seq libraries were prepared following the 10x Genomics protocols for 3' or 5' gene expression library construction [110]. The concentration and amount of cells and reagents added corresponded to the protocol requirements for obtaining 3000 cells with data [137].

2.2.11 Sequencing of single cell RNA-seq libraries

Sequencing of 10x Genomics 3' single cell RNA-seq or 5' single cell RNA-seq libraries was performed on an Illumina NextSeq 500 at high throughput with 75bp paired-end reads at the UBC Biomedical Research Centre (sequencing the terminal 58bp of R2). The target sequencing depth for each sample was 50,000 read pairs per cell, as recommended by 10x Genomics [137]. However, as fewer cells (than targeted during library preparation) were recovered for several samples, the actual sequencing depth per cell varied from 50,000 to 1,000,000 read pairs/cell.

2.2.12 Data curation

Each patient enrolled in the study was assigned a unique anonymous research identifier. Samples from the same patient were assigned unique identifiers associated with a common patient identifier. To preserve patient privacy, the corresponding confidential patient identifiers, patient consents, and clinical data (survival status, treatment, age, name, etc.) were stored in a clinical database inaccessible to research personnel.

A structured query language (SQL) database was used to store sample information (collection date, anatomic site of collection, matched normal available) associated with each research identifier. Additionally, PDX model information associated with each specimen (transplantation date, transplantation site, mouse model type, PDX identifier, mouse date-of-birth, passage number, euthanasia/termination date, necropsy findings, and tumour size) was stored in the SQL database. Physical sample locations were tracked in an OpenSpecimen database.

2.3 Results

2.3.1 Accrual of 41 HGSC cases

Forty-eight ovarian cancer cases – 8 single-site and 40 multi-site – were accrued from May 2015 to September 2018 (**Table 2.2**). On confirmatory pathologic assessment, 41 were HGSC, 2 clear cell ovarian cancer, 2 endometroid, 1 serous borderline, 1 carcinoma with sarcoma elements, and 1 Krukenberg (**Table 2.2**). In total, 148 HGSC samples have been collected from these cases (**Figure 2.3**). Whole-genome sequencing has been performed on 56 samples from 14 of these HGSC cases (**Table 2.3**). Library construction for single cell RNA-seq has been performed on 43 samples from 17 HGSC cases; of these, 11 samples from 6 patients have been sequenced (**Table 2.1**).

Patient	**Histotype**	**# samples**	**# transplanted**
24	HGSC	1	1
25	HGSC	4	4
26	HGSC	1	1
28	HGSC	3	3
29	HGSC	7	6
30	HGSC	7	7
31	HGSC	9	9
32	HGSC	6	6

37	HGSC	4	4
38	HGSC	3	3
41	HGSC	4	4
42	HGSC	5	5
43	HGSC	5	0
44	HGSC	4	4
45	HGSC	3	3
46	HGSC	4	4
47	HGSC	5	5
48	HGSC	4	3
49	HGSC	4	4
50	HGSC	4	4
51	HGSC	3	3
52	HGSC	4	4
53	HGSC	3	3
56	HGSC	1	0
57	HGSC	4	3
59	HGSC	2	2
60	HGSC	2	2
61	HGSC	1	0
62	HGSC	1	0
63	HGSC	6	6
64	HGSC	3	3
65	HGSC	5	0
66	HGSC	2	2
67	HGSC	4	4
68	HGSC	6	6
69	HGSC	2	2
70	HGSC	1	0
71	HGSC	6	6
72	HGSC	1	0
73	HGSC	2	0
74	HGSC	2	2

B1	Serous borderline	4	4
C2	CCOC	1	0
CCOC1	CCOC	4	4
CS1	Carcinoma w/ sarcoma-like	4	4
E1	Endometroid	5	5
E2	Endometroid	2	2
K1	Krukenberg	2	0

Table 2.2: Patient identifiers, histotype as determined by final pathologic evaluation, and the number of tumour samples collected per case. The number of tumour samples transplanted (to create PDX models) is also shown.

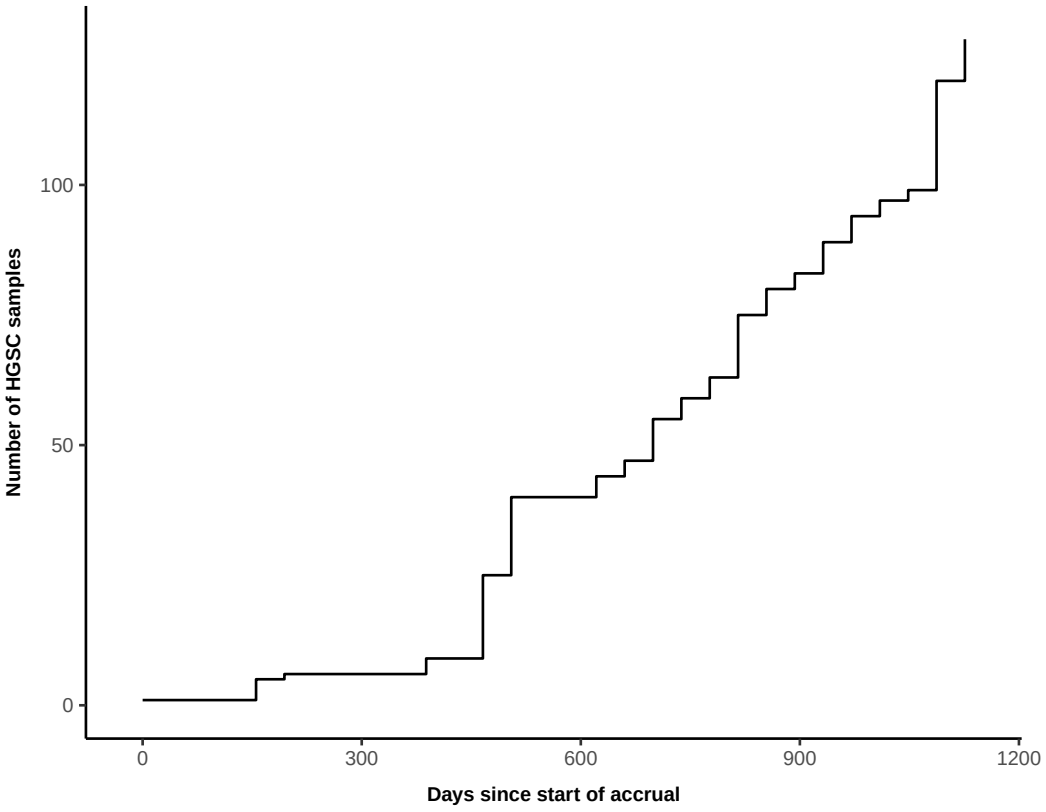


Figure 2.3: HGSC sample accrual since the beginning of the study (first collected sample).

Patient	Sample	Anatomic site
25	VOA6428AX	omentum site 1
25	VOA6428BX	right ovary site 1
25	VOA6428CX	right ovary site 2
25	VOA6428DX	right ovary site 3
26	VOA6491X	left ovary site 1
28	VOA7640CX	left fallopian tube distal site 1
28	VOA7640AX	omentum site 1
28	VOA7640BX	left fallopian tube proximal site 1
29	VOA7648DX	omentum site 1
29	VOA7648CX	ascites site 1
29	VOA7648BX	cul-de-sac site 1
29	VOA7648EX	sigmoid colon site 1
29	VOA7648FX	round ligament site 1
29	VOA7648AX	diaphragm site 1
30	VOA7652EX	omentum site 1
30	VOA7652GX	left pelvic wall site 1
30	VOA7652DX	right fallopian tube site 1
30	VOA7652BX	left ovary site 1
30	VOA7652FX	sigmoid site 1
30	VOA7652AX	right ovary site 1
30	VOA7652CX	left fallopian tube site 1
31	VOA7668EX	ileal tumour site 1
31	VOA7668JX	left fallopian site 1
31	VOA7668AX	pelvic sidewall site 1
31	VOA7668BX	right ovary site 1
31	VOA7668DX	small bowel serosa site 1
31	VOA7668HX	uterine surface site 1
31	VOA7668CX	right fallopian tube site 1
31	VOA7668GX	left ovary site 1
31	VOA7668FX	anterior cul-de-sac site 1
32	VOA7685FX	omentum site 1
32	VOA7685AX	left ovary site 1
32	VOA7685EX	right ovary site 1

32	VOA7685DX	omental nodule site 1
32	VOA7685BX	left ovary site 2
32	VOA7685CX	left ovary site 3
37	VOA8841ax	right ovary site 1
37	VOA8841bx	retrosigmoid serosa site 1
37	VOA8841cx	omentum site 1
38	VOA9127ax	right ovary site 1
38	VOA9127bx	left ovary site 1
41	VOA9465ax	cul de sac site 1
41	VOA9465bx	uterine serosa site 1
41	VOA9465cx	omentum site 1
41	VOA9465dx	left fallopian tube site 1
43	VOA7255ax	right ovary site 1
43	VOA7255bx	left ovary site 1
43	VOA7255cx	right fallopian tube site 1
43	VOA7255ex	peritoneal nodule site 1
44	VOA9655ax	right ovary site 1
44	VOA9655bx	left ovary site 1
44	VOA9655cx	left ovarian cyst site 1
46	VOA9921ax	left ovary site 1
46	VOA9921bx	omentum site 1
46	VOA9921cx	left fallopian tube site 1
47	VOA9955cx	right ovary site 1

Table 2.3: Sample and patient identifiers of HGSC samples used for whole genome sequencing.

2.3.2 Construction of patient-derived xenograft models

Thus far, 128 samples from 38 HGSC cases have been engrafted in PDXs (total of 275 PDXs, see **Table 2.4** for a full list and **Figure 2.4** for established PDXs). Up to 4 models per passage were created from each primary tumour. In total, 52 samples from 19 patients have grown macroscopically visible tumours. Engraftment (growth/establishment) rates of HGSC tumours at the model, sample, and patient level for NSG vs. NRG strains are shown in **Table 2.5**. Engraftment rates for NRG mice were lower than those for NSG mice (**Table 2.5**). Engraftment rates as a function of time since transplant are shown in **Figure 2.5**. Approximately 70% of

models that eventually engrafted had already established a mass by 300 days (**Figure 2.5**A). Out of all models that were euthanized, approximately 50% had grown tumours after 100 days since surgery (**Figure 2.5**B). The rates shown in **Figure 2.5**B appear to decrease over time because the humane endpoint for most models that do not grow tends to occur later than for those that do (**Figure 2.6**); models that do not grow are euthanized based on health/age, while those that do are euthanized based on health/tumour size.

Patient	**Sample**	**PDX ID**	**Strain**	**Grown**
24	VOA5576	Y55761	NSG	0
24	VOA5576	Y55762	NSG	1
24	VOA5576	Y55763	NSG	1
24	VOA5576	Y55764	NSG	1
24	VOA5576	Y557631	NSG	1
24	VOA5576	Y557632	NSG	1
24	VOA5576	Y557633	NSG	0
24	VOA5576	Y557634	NSG	1
25	VOA6428A	Y6428A1	NSG	1
25	VOA6428A	Y6428A2	NSG	0
25	VOA6428A	Y6428A3	NSG	1
25	VOA6428B	Y6428B1	NSG	0
25	VOA6428B	Y6428B2	NSG	0
25	VOA6428B	Y6428B3	NSG	1
25	VOA6428C	Y6428C1	NSG	1
25	VOA6428C	Y6428C2	NSG	0
25	VOA6428C	Y6428C3	NSG	1
25	VOA6428D	Y6428D1	NSG	1
25	VOA6428D	Y6428D2	NSG	0
25	VOA6428B	Y6428B31	NSG	1
25	VOA6428B	Y6428B32	NSG	1
25	VOA6428B	Y6428B33	NSG	1
25	VOA6428B	Y6428B34	NSG	1
25	VOA6428B	Y6428B341	NSG	1
25	VOA6428B	Y6428B342	NSG	0
25	VOA6428B	Y6428B343	NSG	1
25	VOA6428B	Y6428B344	NSG	1

26	VOA6491	Y64911	NSG	1
26	VOA6491	Y64912	NSG	1
26	VOA6491	Y64913	NSG	1
26	VOA6491	Y64914	NSG	1
28	VOA7640A	Y7640A1	NSG	1
28	VOA7640A	Y7640A2	NSG	1
28	VOA7640B	Y7640B1	NSG	0
28	VOA7640B	Y7640B2	NSG	0
28	VOA7640C	Y7640C1	NSG	1
28	VOA7640C	Y7640C2	NSG	1
29	VOA7648A	Y7648A1	NSG	1
29	VOA7648A	Y7648A2	NSG	1
29	VOA7648D	Y7648D1	NSG	1
29	VOA7648D	Y7648D2	NSG	1
29	VOA7648B	Y7648B1	NSG	1
29	VOA7648B	Y7648B2	NSG	1
29	VOA7648E	Y7648E1	NSG	1
29	VOA7648E	Y7648E2	NSG	1
29	VOA7648C	Y7648C1	NSG	1
29	VOA7648C	Y7648C2	NSG	1
29	VOA7648F	Y7648F1	NSG	1
29	VOA7648F	Y7648F2	NSG	1
30	VOA7652A	Y7652A1	NSG	1
30	VOA7652A	Y7652A2	NSG	1
30	VOA7652B	Y7652B1	NSG	1
30	VOA7652B	Y7652B2	NSG	1
30	VOA7652C	Y7652C1	NSG	0
30	VOA7652C	Y7652C2	NSG	1
30	VOA7652D	Y7652D1	NSG	0
30	VOA7652D	Y7652D2	NSG	1
30	VOA7652E	Y7652E1	NSG	0
30	VOA7652E	Y7652E2	NSG	0
30	VOA7652F	Y7652F1	NSG	0

30	VOA7652F	Y7652F2	NSG	0
30	VOA7652G	Y7652G1	NSG	1
30	VOA7652G	Y7652G2	NSG	0
31	VOA7668A	Y7668A1	NSG	1
31	VOA7668A	Y7668A2	NSG	1
31	VOA7668B	Y7668B1	NSG	1
31	VOA7668B	Y7668B2	NSG	1
31	VOA7668C	Y7668C1	NSG	1
31	VOA7668C	Y7668C2	NSG	1
31	VOA7668D	Y7668D1	NSG	1
31	VOA7668D	Y7668D2	NSG	1
31	VOA7668E	Y7668E1	NSG	1
31	VOA7668E	Y7668E2	NSG	1
31	VOA7668F	Y7668F1	NSG	1
31	VOA7668F	Y7668F2	NSG	1
31	VOA7668G	Y7668G1	NSG	0
31	VOA7668G	Y7668G2	NSG	0
31	VOA7668H	Y7668H1	NSG	0
31	VOA7668H	Y7668H2	NSG	1
31	VOA7668J	Y7668J1	NSG	1
31	VOA7668J	Y7668J2	NSG	1
32	VOA7685A	Y7685A1	NSG	0
32	VOA7685A	Y7685A2	NSG	0
32	VOA7685B	Y7685B1	NSG	1
32	VOA7685B	Y7685B2	NSG	1
32	VOA7685C	Y7685C1	NSG	0
32	VOA7685C	Y7685C2	NSG	0
32	VOA7685D	Y7685D1	NSG	1
32	VOA7685D	Y7685D2	NSG	1
32	VOA7685E	Y7685E1	NSG	1
32	VOA7685E	Y7685E2	NSG	1
32	VOA7685F	Y7685F1	NSG	1
32	VOA7685F	Y7685F2	NSG	1

37	VOA8841A	Y8841A1	NSG	1
37	VOA8841A	Y8841A2	NSG	0
37	VOA8841B	Y8841B1	NSG	0
37	VOA8841B	Y8841B2	NSG	0
37	VOA8841C	Y8841C1	NSG	1
37	VOA8841C	Y8841C2	NSG	0
37	VOA8841D	Y8841D1	NSG	0
37	VOA8841D	Y8841D2	NSG	0
38	VOA9127A	Y9127A1	NSG	0
38	VOA9127A	Y9127A2	NSG	0
38	VOA9127B	Y9127B1	NSG	0
38	VOA9127B	Y9127B2	NSG	0
38	VOA9127C	Y9127C1	NSG	0
38	VOA9127C	Y9127C2	NSG	0
41	VOA9465A	Y9465A1	NRG	0
41	VOA9465A	Y9465A2	NRG	1
41	VOA9465B	Y9465B1	NRG	0
41	VOA9465B	Y9465B2	NRG	0
41	VOA9465C	Y9465C1	NRG	1
41	VOA9465C	Y9465C2	NRG	1
41	VOA9465D	Y9465D1	NRG	0
41	VOA9465D	Y9465D2	NRG	0
42	VOA10243SA	Y10243SA1	NSG	1
42	VOA10243SA	Y10243SA2	NSG	1
42	VOA10243SB	Y10243SB1	NSG	1
42	VOA10243SB	Y10243SB2	NSG	1
42	VOA10243SD	Y10243SD1	NSG	1
42	VOA10243SD	Y10243SD2	NSG	1
42	VOA10243SC	Y10243SC1	NSG	0
42	VOA10243SC	Y10243SC2	NSG	1
42	VOA10243SE	Y10243SE1	NRG	1
42	VOA10243SE	Y10243SE2	NRG	1
44	VOA9655A	Y9655A1	NRG	0

44	VOA9655A	Y9655A2	NRG	0
44	VOA9655B	Y9655B1	NRG	0
44	VOA9655B	Y9655B2	NRG	0
44	VOA9655C	Y9655C1	NSG	0
44	VOA9655C	Y9655C2	NSG	0
44	VOA9655D	Y9655D1	NSG	0
44	VOA9655D	Y9655D2	NSG	0
45	VOA9907A	Y9907A1	NSG	1
45	VOA9907A	Y9907A2	NSG	1
45	VOA9907B	Y9907B1	NSG	0
45	VOA9907B	Y9907B2	NSG	1
45	VOA9907C	Y9907C1	NSG	0
45	VOA9907C	Y9907C2	NSG	0
46	VOA9921A	Y9921A1	NSG	1
46	VOA9921A	Y9921A2	NSG	1
46	VOA9921B	Y9921B1	NSG	0
46	VOA9921B	Y9921B2	NSG	1
46	VOA9921C	Y9921C1	NSG	1
46	VOA9921C	Y9921C2	NSG	1
46	VOA9921D	Y9921D1	NSG	0
46	VOA9921D	Y9921D2	NSG	0
47	VOA9955A	Y9955A1	NSG	0
47	VOA9955A	Y9955A2	NSG	0
47	VOA9955B	Y9955B1	NSG	0
47	VOA9955B	Y9955B2	NSG	0
47	VOA9955C	Y9955C1	NSG	0
47	VOA9955C	Y9955C2	NSG	0
47	VOA9955D	Y9955D1	NSG	0
47	VOA9955D	Y9955D2	NSG	0
47	VOA9955E	Y9955E1	NSG	1
47	VOA9955E	Y9955E2	NSG	0
48	VOA7294A	Y7294A1	NSG	0
48	VOA7294A	Y7294A2	NSG	0

48	VOA7294B	Y7294B1	NSG	0
48	VOA7294B	Y7294B2	NSG	0
48	VOA7294C	Y7294C1	NSG	0
48	VOA7294C	Y7294C2	NSG	0
49	VOA9186A	Y9186A1	NSG	0
49	VOA9186A	Y9186A2	NSG	0
49	VOA9186B	Y9186B1	NSG	0
49	VOA9186B	Y9186B2	NSG	0
49	VOA9186C	Y9186C1	NSG	1
49	VOA9186C	Y9186C2	NSG	0
49	VOA9186D	Y9186D1	NSG	0
49	VOA9186D	Y9186D2	NSG	0
50	VOA9453a	Y9453a1	NRG	1
50	VOA9453a	Y9453a2	NRG	1
50	VOA9453b	Y9453b1	NRG	0
50	VOA9453b	Y9453b2	NRG	0
50	VOA9453c	Y9453c1	NRG	0
50	VOA9453c	Y9453c2	NRG	1
50	VOA9453d	Y9453d1	NRG	0
50	VOA9453d	Y9453d2	NRG	1
51	VOA10288SA	Y10288SA1	NRG	0
51	VOA10288SA	Y10288SA2	NRG	0
51	VOA10288SB	Y10288SB1	NRG	1
51	VOA10288SB	Y10288SB2	NRG	1
51	VOA10288SC	Y10288SC1	NRG	0
51	VOA10288SC	Y10288SC2	NRG	0
52	VOA10429SA	Y10429SA1	NRG	0
52	VOA10429SA	Y10429SA2	NRG	0
52	VOA10429SB	Y10429SB1	NRG	0
52	VOA10429SB	Y10429SB2	NRG	0
52	VOA10429SC	Y10429SC1	NRG	0
52	VOA10429SC	Y10429SC2	NRG	0
52	VOA10429SD	Y10429SD1	NRG	0

52	VOA10429SD	Y10429SD2	NRG	0
53	VOA10471SA	Y10471SA1	NRG	0
53	VOA10471SA	Y10471SA2	NRG	0
53	VOA10471SB	Y10471SB1	NRG	0
53	VOA10471SB	Y10471SB2	NRG	0
53	VOA10471SC	Y10471SC1	NRG	1
53	VOA10471SC	Y10471SC2	NRG	1
57	VOA10863SB	Y10863SB1	NRG	0
57	VOA10863SB	Y10863SB2	NRG	0
57	VOA10863SC1	Y10863SC11	NRG	0
57	VOA10863SC1	Y10863SC12	NRG	0
57	VOA10863SC2	Y10863SC21	NRG	0
57	VOA10863SC2	Y10863SC22	NRG	0
59	VOA10439SA	Y10439SA1	NRG	0
59	VOA10439SA	Y10439SA2	NRG	0
59	VOA10439SB	Y10439SB1	NRG	0
59	VOA10439SB	Y10439SB2	NRG	0
60	VOA10497SA	Y10497SA1	NRG	0
60	VOA10497SA	Y10497SA2	NRG	0
60	VOA10497SB	Y10497SB1	NRG	0
60	VOA10497SB	Y10497SB2	NRG	0
63	VOA11095SA	Y11095SA1	NSG	0
63	VOA11095SA	Y11095SA2	NSG	0
63	VOA11095SB	Y11095SB1	NSG	0
63	VOA11095SB	Y11095SB2	NSG	0
63	VOA11095SC	Y11095SC1	NSG	0
63	VOA11095SC	Y11095SC2	NSG	0
63	VOA11095SD	Y11095SD1	NSG	0
63	VOA11095SD	Y11095SD2	NSG	0
63	VOA11095SE	Y11095SE1	NRG	0
63	VOA11095SE	Y11095SE2	NRG	0
63	VOA11095SF	Y11095SF1	NRG	0
63	VOA11095SF	Y11095SF2	NRG	0

64	VOA11213SA	Y11213SA1	NRG	0
64	VOA11213SA	Y11213SA2	NRG	0
64	VOA11213SB	Y11213SB1	NRG	0
64	VOA11213SB	Y11213SB2	NRG	0
64	VOA11213SC	Y11213SC1	NRG	0
64	VOA11213SC	Y11213SC2	NRG	0
66	VOA11088A	Y11088A1	NRG	0
66	VOA11088A	Y11088A2	NRG	0
66	VOA11088B	Y11088B1	NRG	0
66	VOA11088B	Y11088B2	NRG	0
67	VOA11220SA	Y11220SA1	NRG	0
67	VOA11220SA	Y11220SA2	NRG	0
67	VOA11220SB	Y11220SB1	NRG	0
67	VOA11220SB	Y11220SB2	NRG	0
67	VOA11220SC	Y11220SC1	NSG	0
67	VOA11220SC	Y11220SC2	NSG	0
67	VOA11220SD	Y11220SD1	NSG	0
67	VOA11220SD	Y11220SD2	NSG	0
68	VOA11243SA	Y11243SA1	NRG	0
68	VOA11243SA	Y11243SA2	NRG	0
68	VOA11243SB	Y11243SB1	NRG	0
68	VOA11243SB	Y11243SB2	NRG	0
68	VOA11243SC	Y11243SC1	NSG	0
68	VOA11243SC	Y11243SC2	NSG	0
68	VOA11243SD	Y11243SD1	NSG	0
68	VOA11243SD	Y11243SD2	NSG	0
68	VOA11243SE	Y11243SE1	NRG	0
68	VOA11243SE	Y11243SE2	NRG	0
68	VOA11243SF	Y11243SF1	NRG	0
68	VOA11243SF	Y11243SF2	NRG	0
69	VOA11265SA	Y11265SA1	NRG	0
69	VOA11265SA	Y11265SA2	NRG	0
69	VOA11265SB	Y11265SB1	NRG	0

69	VOA11265SB	Y11265SB2	NRG	0
71	VOA11294A	Y11294A1	NRG	0
71	VOA11294A	Y11294A2	NRG	0
71	VOA11294B	Y11294B1	NRG	0
71	VOA11294B	Y11294B2	NRG	0
71	VOA11294C	Y11294C1	NRG	0
71	VOA11294C	Y11294C2	NRG	0
71	VOA11294D	Y11294D1	NRG	0
71	VOA11294D	Y11294D2	NRG	0
71	VOA11294E	Y11294E1	NRG	0
71	VOA11294E	Y11294E2	NRG	0
71	VOA11294F	Y11294F1	NRG	0
71	VOA11294F	Y11294F2	NRG	0
74	VOA11258SA	Y11258SA1	NRG	0
74	VOA11258SA	Y11258SA2	NRG	0
74	VOA11258SB	Y11258SB1	NRG	0
74	VOA11258SB	Y11258SB2	NRG	0
B1	VOA7618A	Y7618A1	NSG	0
B1	VOA7618A	Y7618A2	NSG	0
B1	VOA7618B	Y7618B1	NSG	0
B1	VOA7618B	Y7618B2	NSG	0
B1	VOA7618C	Y7618C1	NSG	0
B1	VOA7618C	Y7618C2	NSG	1
B1	VOA7618D	Y7618D1	NSG	0
B1	VOA7618D	Y7618D2	NSG	1
CCOC1	VOA6851A	Y6851A1	NSG	0
CCOC1	VOA6851A	Y6851A2	NSG	0
CCOC1	VOA6851B	Y6851B1	NSG	0
CCOC1	VOA6851B	Y6851B2	NSG	0
CCOC1	VOA6851C	Y6851C1	NSG	0
CCOC1	VOA6851C	Y6851C2	NSG	0
CCOC1	VOA6851D	Y6851D1	NSG	0
CCOC1	VOA6851D	Y6851D2	NSG	0

CS1	VOA6873A	Y6873A1	NSG	0
CS1	VOA6873A	Y6873A2	NSG	0
CS1	VOA6873B	Y6873B1	NSG	0
CS1	VOA6873B	Y6873B2	NSG	0
CS1	VOA6873C	Y6873C1	NSG	0
CS1	VOA6873C	Y6873C2	NSG	0
CS1	VOA6873D	Y6873D1	NSG	1
CS1	VOA6873D	Y6873D2	NSG	1
E1	VOA7298A	Y7298A1	NSG	0
E1	VOA7298A	Y7298A2	NSG	0
E1	VOA7298B	Y7298B1	NSG	0
E1	VOA7298B	Y7298B2	NSG	0
E1	VOA7298C	Y7298C1	NSG	0
E1	VOA7298C	Y7298C2	NSG	0
E1	VOA7298D	Y7298D1	NSG	0
E1	VOA7298D	Y7298D2	NSG	0
E1	VOA7298E	Y7298E1	NSG	0
E1	VOA7298E	Y7298E2	NSG	0
E2	VOA11520SA	Y11520SA1	NSG	0
E2	VOA11520SA	Y11520SA2	NSG	0
E2	VOA11520SB	Y11520SB1	NSG	0
E2	VOA11520SB	Y11520SB2	NSG	0

Table 2.4: Inventory of ovarian PDXs created from patient primary tumours. The strain of mouse used (NSG or NRG) and whether or not a macroscopically visible tumour was grown and harvested from each model (1 = grown, 0 = not grown) are indicated.

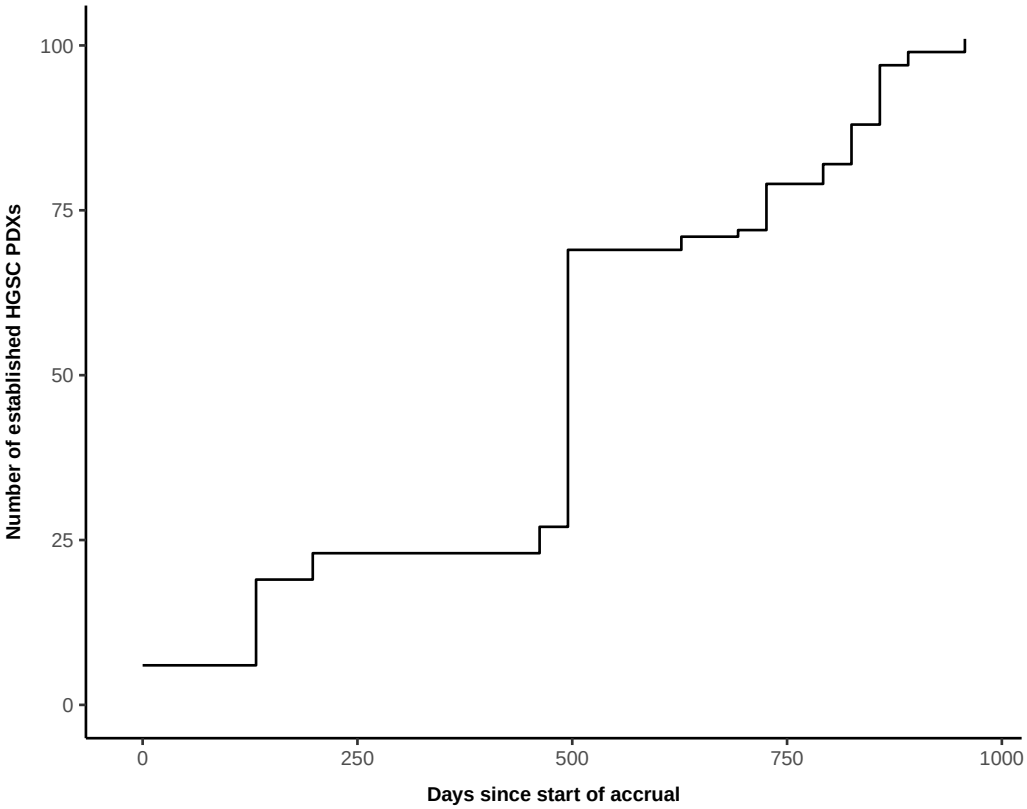


Figure 2.4: Accrual of established PDXs since the beginning of the study (first collected sample).

	NRG	**NSG**
Models	102	173
Samples	51	77
Patients	18	20
Models grown	13	88
Patients w/ grown models	5	14
Samples w/ grown models	8	44
Grown models (1 yr)	35%	58%
Grown samples (1 yr)	50%	64%
Grown patients (1 yr)	67%	77%

Table 2.5: Summary statistics for PDX collection by strain for HGSC tumours. The number patients and samples with at least 1 grown PDX, along with the engraftment rate (by model, sample, and patient) for models collected ≥ 1 year ago are shown.

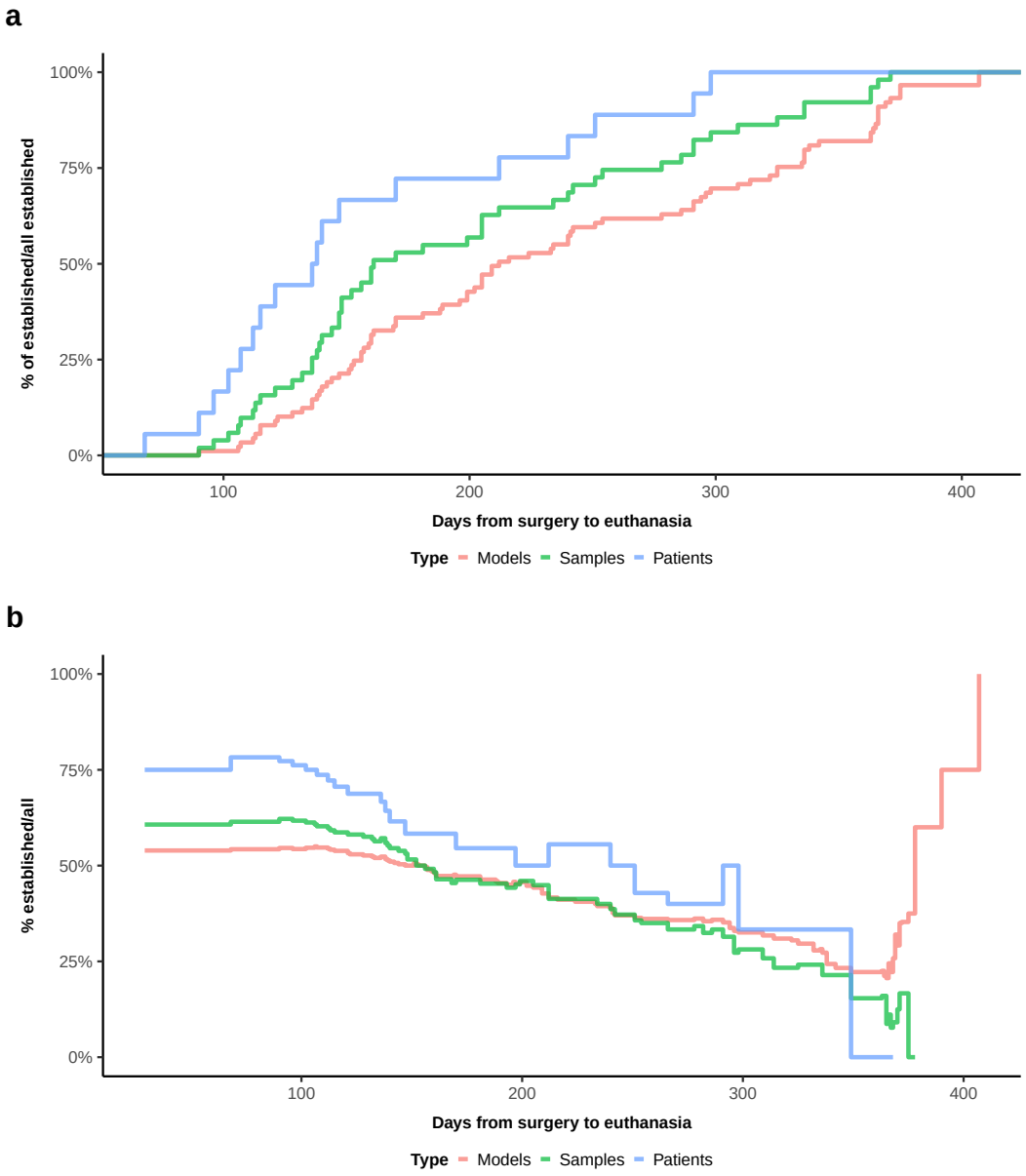


Figure 2.5: (a) Cumulative distribution function of engrafted and euthanized tumours for HGSC PDXs at the level of models, samples, and patients. (b) Percent of engrafted and euthanized tumours (out of all euthanized tumours) as a function of engraftment time. Statistics summarized at the level of models, samples, and patients.

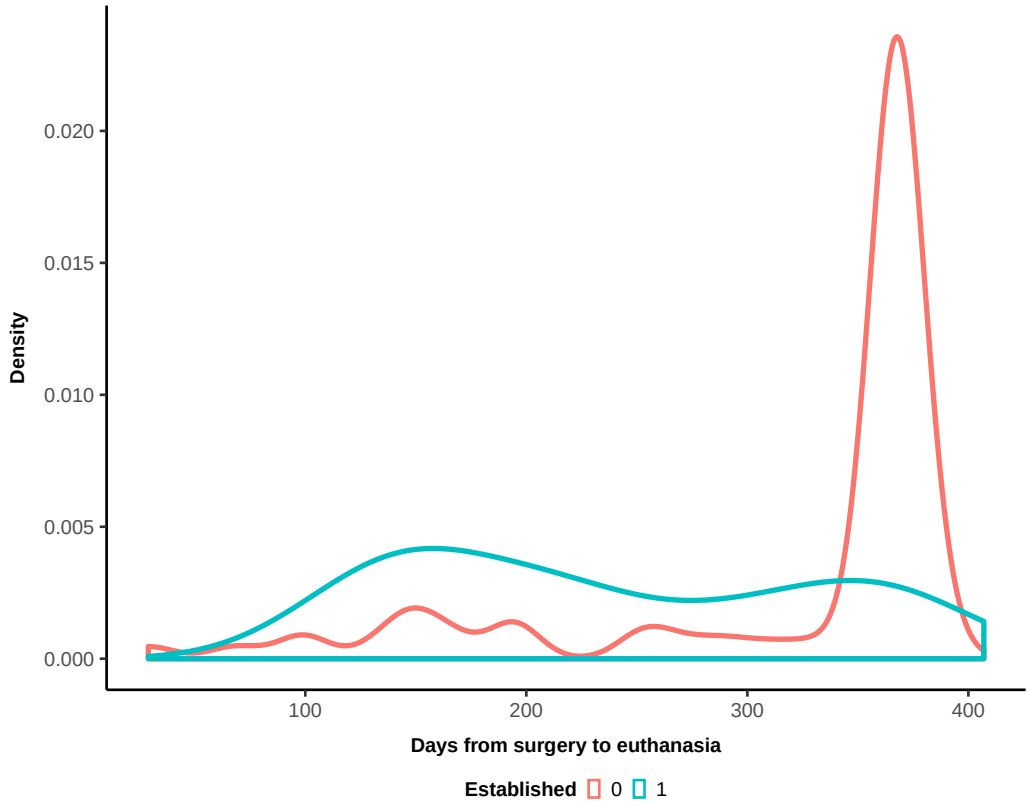


Figure 2.6: Time from surgery to euthanasia for PDX models that do and do not grow tumours.

2.4 Discussion

Prognosis for HGSC patients has remained poor (5-year survival approx. 35%) over the last few decades [138]. Despite the development of many new therapies including PARP inhibitors [16], platinum resistant HGSC remains difficult to manage [139]. Extensive clonal heterogeneity across space is thought to be a key factor that engenders therapeutic resistance in this devastating disease. In some cases, tumour-intrinsic mechanisms of resistance such as BRCA mutation reversion [15, 140] and upregulation of drug efflux transporters have been observed. Initial *in vitro* work has revealed a microenvironment-mediated mechanism of resistance involving fibroblasts and lymphocytes [59]. Yet, the mechanistic underpinnings of resistance in HGSC in the majority of cases – especially in platinum-refactory foldback inversion-subtype tumours

[3] – remain unknown. Patient-derived xenografts that faithfully recapitulate tumour histology, genomics, and expression patterns offer ideal model systems for studying drug response in HGSC. The combination of single cell-resolution assays and PDX models enables accurate characterization of rare clonal genotypes and phenotypes that underlie treatment resistance. Tracking clonal prevalence trajectories in PDXs will serve as the basis for understanding the contribution of genomic subtypes to platinum resistance and for predictive modeling of clonal dynamics to inform therapeutic regimen choices for patients.

The multi-site nature of HGSC necessitates collection of multiple tumour sites per patient to obtain a comprehensive view of the treatment-naïve clonal repertoire. To this end, our cohort constitutes the largest set of multi-site high-grade serous ovarian cancer patients we are aware of, with high-quality specimens for genomic, transcriptomic, and proteomic analysis and matched patient-derived xenograft models. Our model-level engraftment rate of approximately 50% after 100 days (**Figure 2.5**A) is within the range of previously described values (48% to 90%) [141–143]. In addition, we have demonstrated that high-quality single cell copy number profiles can be derived from similarly constructed PDX models [97, 98]. Thus, our PDXs can be leveraged to investigate drug response and clonal dynamics in HGSC over time and space. As the determinants of successful HGSC tumour engraftment are largely unknown, we note that certain aspects of the engrafted cohort may not be completely representative of all HGSC tumours. Moreover, despite the similarity between model types, engraftment rates appeared to be higher in NSG mice than NRG mice. Thus, analyses using these models will have to account for possible biases in cohort composition and differences between model types. Engraftment rates may be improved by subcapsular transplantation in the kidney. Our future work will entail single cell whole-genome sequencing of carboplatin-treated PDX models to reveal and develop predictive models for clonal dynamics under treatment selection pressure. Additionally, we have established single cell dissociation protocols for single cell RNA-sequencing that will enable microenvironment decomposition.

The cohort established in this chapter sets the groundwork for studying interactions between malignant and immune cells in HGSC in Chapter **3** and studying single cell properties and microenvironment composition in Chapter **4**.

Chapter 3

The evolutionary interface between tumour-infiltrating lymphocytes and cancer cells in multi-site HGSC

3.1 Introduction

High-grade serous ovarian cancer (HGSC) exhibits the highest disease mortality among gynecologic cancers. Despite recent progress with poly ADP-ribose polymerase (PARP) inhibitor-based synthetic lethal approaches exploiting homologous recombination deficiency [16], HGSC remains incurable in most cases. Characterized by profound genomic instability and clonal diversity, HGSC often presents with widespread peritoneal dissemination. Multi-site studies have revealed genomic intratumoral heterogeneity (ITH) as a correlate to poor survival [28], as well as specific patterns of malignant cell spread within the peritoneal cavity [27]. Importantly, the physical distribution of malignant clones across the peritoneal cavity is non-random, with the majority of sites exhibiting clonal homogeneity and a minority of sites harboring diverse clones [29]. This raises the hypothesis that region-specific properties, including immunologic components of the tumor microenvironment, may modulate malignant cell invasion and expansion, thereby shaping evolutionary selection.

HGSC patients with abundant CD8+, CD4+, CD20+, and plasma cell tumor-infiltrating lymphocytes (TILs) are associated with favorable clinical outcomes [36, 38, 41, 144]. TILs can respond to and temporally track neoantigens [39] and mitigate resistance to platinum chemotherapy [59]. However, much of our understanding of the immune response in HGSC derives from single biopsies; far less is known about spatial immunologic variation across distal tumor foci. Histologic imaging has revealed that lymphocyte abundance can vary between tumor foci in HGSC [145]. Furthermore, lymphocyte expression signatures are linked to patterns of metastasis [146]. A single case report has described immunologic variation across relapse specimens [147]; however, given the immunomodulatory effects of chemotherapy [148],

understanding of pre-treatment spatial variation is still lacking.

Beyond immunologic features, prognostic mutational processes in HGSC through analysis of point mutation, copy number, and rearrangement features has indicated a prominent association between foldback inversions (FBIs) and poor response to platinum-based chemotherapy [3]. FBI-dominated tumors, which comprise approximately 40% of HGSC, tend to be exclusive to homologous-recombination-deficient (HRD) cases and bear a distinct pattern of high-level amplifications colocalized with foldback rearrangements typical of breakage-fusion-bridge processes [3, 24]. How mutational processes co-vary with immune response characteristics in HGSC remains poorly understood. This will become of central importance as clinical trials assaying synthetic lethal compounds targeting DNA repair processes combined with immune-modulation therapies read out.

We surmised that localized selective pressures imposed by immune microenvironments shape the distribution of malignant clones during disease progression. Thus, we systematically profiled the inter-relationship of clonal diversity, mutational processes, and immunologic response across a cohort of patients and multi-region samples. Genome-sequencing-based clonal decomposition, transcriptome-based T and B cell receptor sequencing, multicolor immunohistochemistry (IHC), and histologic image analyses were applied. Our results elucidate the landscape of cell-type interactions at the interface of malignant and immune cells across 212 samples from 38 patients. We show that samples robustly segregate into three distinct TIL subtypes, reflecting little or no immune infiltration, stromal infiltration, and combined epithelial and stromal infiltration. We reveal an association between these classes and malignant clone diversity properties. Regions with highest levels of epithelial immune infiltration exhibit the lowest malignant clone diversity, neoantigen depletion, and subclonal loss of heterozygosity (LOH) at human leukocyte antigen (HLA) loci as evidence of purifying selection. Moreover, T cell clonotypes, but not B cell clonotypes, spatially track with tumor clones in patients with heavily infiltrated tumors. Finally, we show combinatorial prognostic effects between mutational processes and immune infiltration with foldback inversions exhibiting high risk even in the presence of high cytotoxicity. In aggregate, our findings illuminate molecular and evolutionary properties at the immune-malignant interface in HGSC with new insights on how tumor progression and clonal dissemination are driven by immune-related selective pressures.

3.2 Materials and Methods

3.2.1 Experimental Model and Subject Details

3.2.1.1 Sample acquisition, consent, & surgery

Ethical approval for this study was obtained from the University of British Columbia (UBC) Research Ethics Board. Women (biological sex: XX) undergoing debulking surgery (primary or recurrent) for carcinoma of ovarian/peritoneal/fallopian tube origin were approached for informed consent to bank tumor tissue. Cases of high-grade serous carcinoma where more than one sample was collected were chosen for this analysis. Clinicopathologic and outcome data were collected by chart review. Consistent with the practice at UBC and BC Cancer, all patients with high-grade serous ovarian cancer (HGSC) are referred to the hereditary cancer clinic and offered genetic testing for *BRCA1* and *BRCA2* mutations (`http://www.bccancer.bc.ca/screening/Documents/HCP_GuidelinesManuals-HBOCCriteria.pdf`).

For consented patients, when multiple tumor sites were encountered intraoperatively, effort was made to bank as many sites as possible. Samples were flash frozen and stored according to conditions outlined below. For cases where multiple tumor sites were encountered but not all anatomic sites could be frozen (e.g., due to unavailability of trained staff), archival specimens stored within our pathology department were used. All samples were from removed structures during attempts at optimal debulking; hence the majority of samples were from omentum and ovarian sites.

Platinum sensitive is defined as no relapse within 6 months of the chemotherapy stop date.

3.2.1.2 Sample preservation & histologic evaluation

When adequate tumor volume was available, multiple tissue samplings were obtained from each tissue specimen. Up to 5 samplings were taken from a given tumor, with effort made to equally space samples while staying within grossly apparent tumor tissue. Each sampling was cut into three pieces, yielding two end-pieces for cryovials and a middle portion placed in 10% buffered formalin. End pieces were homogenized manually and with a paddle blender (Stomacher). All paraffin-embedded blocks, including formalin-fixed tumor samples and molecular-fixed fallopian tubes, were sectioned and stained with hematoxylin and eosin prior to expert histopathological review to confirm the presence of high-grade serous carcinoma. Pieces from the same sampling were given the same sample identifier for the analysis steps described below.

3.2.2 Method Details

3.2.2.1 WGSS library construction & sequencing

Frozen tumor samples from 14 patients (patients 11-17, 25, 26, 28-32, total 71 samples) were submitted for library construction and sequencing. Sample size was determined by availability of resectable, cryopreserved tissue, and DNA quality. For all tumor and normal samples, DNA extraction was followed by library construction and sequencing using Illumina HiSeq2500 whole genome shotgun v4 chemistry with paired-end 125bp reads. Samples were sequenced to an average of 96X coverage. Patients 1-4, 7, 9, and 10 were previously sequenced according to specifications described in [29].

3.2.2.2 Targeted bulk sequencing analysis

Target selection For each patient we performed targeted sequencing on (11-17), a total of 192 positions were deeply sequenced, including 4 experimental controls, a TP53 variant, heterozygous germline SNPs lost in dominant loss of heterozygosity (LOH) events, lost SNVs that could and could not be explained by copy number events, and SNVs inferred to originate at each node of the sample phylogeny obtained by applying the stochastic Dollo approach (infinite sites with loss model) [29] (**Supplemental Table A.1**). SNVs were sampled as evenly as possible across nodes.

Data for patients 1-4, 7, 9, and 10 was obtained from [29], and used as input for section Clonal analysis onward.

Primer design Primers targeting the positions described above were designed using primer3. The full list of primers is included in **Supplemental Table A.1**. Optimal primer length was 27nt (18-30nt) and products were designed to be 150-250nt long with 53-61°C melting temperature. Breslauer thermodynamic correction and Schildkraut and Lifson salt correction settings in primer3 were used. Additionally, primers targeting SNVs were required to pass the following preliminary filters: minimum of 5 alignments to the genome as given by BLAT for each primer, and each primer position at least 30nt away from the target SNV.

Primers were additionally tested using a combination of UCSCs in silico PCR tool (`http://genome.ucsc.edu/cgi-bin/hgPcr`) aligned against the reference hg19 genome and custom in-house code (Canadas Michael Smith Genome Sciences Centre) to verify a unique hit and check that the variant was located within 150bp of the nearest end of the amplicon to ensure coverage in an Illumina NextSeq 150bp paired end read. The primers were tagged with Illumina adapters to enable a direct sequencing approach that precludes the need for adaptor ligation

during sample preparation. The Illumina adaptor tags were: `5'-CGCTCTTCCGATCTCTG-3'` on the forward amplicon primer and `5'-TGCTCTTCCGATCTGAC-3'` on the reverse amplicon primers.

PCR and Illumina sequencing Genomic DNA templates were used as starting material to generate PCR products. PCR was set up using Phusion DNA polymerase (Fisher Scientific, USA) according to the manufacturers specifications. The standard PCR conditions used were an initial denaturation at 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 60°C for 15 s and 72°C for 8 s, and a final extension at 72°C for 10 minutes.

Amplicons were pooled by template for sequencing sample preparation. Sample preparation involved a second round of amplification using Phusion DNA polymerase with 6 PCR cycles using PE primer 1.0-DS (`5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCTG-3'`) and a custom PCR Primer (`5'-CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAC-3'`) that contains a unique six-nucleotide 'index' shown as N's. PCR products were cleaned up using PCRClean DX beads (Aline Biosciences, USA). DNA quality was assessed using the Caliper LabChip GX High Sensitivity Assay (Caliper Life Sciences, USA) and DNA quantity was measured using a Qubit dsDNA HS assay kit on a Qubit fluorometer (Life Technologies, USA).

The indexed libraries were pooled together and sequenced on the Illumina NextSeq500 platform with paired-end 150bp reads using v2 chemistry reagents.

3.2.2.3 Immunohistochemistry

All reagents were from Biocare Medical (Pacheco, CA) unless otherwise stated. Slides of formalin-fixed, paraffin embedded tissue were deparaffinized and rehydrated through xylene and graded alcohols. Antigen retrieval was performed using Diva Decloaker in a Biocare decloaking chamber at 125°C for 30 s. Slides were then rinsed with water, marked with PAP pen and loaded into the Biocare Intellipath FLX autostainer. Slides were blocked with peroxidazed-1 and background sniper for 5 minutes and 10 minutes respectively then a cocktail of either CD8 (1/250, clone C8/144B, Cell Marque, Rocklin, CA) and CD3 (1/500, clone SP7, Spring Biosciences, Pleasanton, CA), or CD79a (1/400, clone SP18, Spring Biosciences, Pleasanton, CA) and CD138 (1/200, clone B-A38, Biocare Medical, Pacheco, CA) in Da Vinci Green diluent was added for 30 minutes at room temperature. Following a wash step, Mach2 Doublestain #2 polymer was added for 30 minutes at room temperature and then antigens detected with IP Ferengi Blue chromogen for 7 minutes followed by IP DAB chromogen for 5 minutes. To denature the first round of antibodies, slides were removed from the autostainer and placed in pre-warmed SDS-glycine pH 2.0 solution for 45 minutes at 50°C with periodic agitation. Slides

were then washed with water and replaced in the autostainer for the 2nd round of staining. CD20 (1/300, clone L26, Biocare Medical, Pacheco, CA) diluted in Da Vinci Green diluent was added to the slides and incubated for 30 minutes at room temperature. Mach2 Mouse-AP polymer or Mach2 Rabbit-AP polymer was added for 30 minutes at room temperature to detect CD20. Warp red chromogen was added to the slide for 7 minutes, hematoxylin at a 1/5 dilution was then added for 5 minutes. The slides were then washed, air-dried and coverslipped with Ecomount coverslipping medium.

3.2.2.4 Nanostring gene expression

FFPE samples were deparaffinised with xylene and washed with 100% ethanol. Tissue was then extracted using QIAGEN miRNeasy FFPE Kit, following the protocol for purification of total RNA (including miRNA) from FFPE tissue sections. RNA quality was assessed with Nanodrop. 500ng of high quality RNA (260/280 ratio of 1.7-2.3 and A260/230 ratio of 1.8-2.3) for each sample was used in the Nanostring assay (PanCancer Immune Profiling panel [149] additionally containing markers for high-grade serous ovarian cancer subtypes C1, C2, C4, and C5 [150]). Data was normalized with the voom function from the R package limma and TMM normalization. Samples flagged by nSolver (Nanostring Technologies) were removed from further analysis.

3.2.2.5 TCR & BCR sequencing

In the text below, *TRB* and *IGH* refer to TCR-β chain and Ig-heavy chain, respectively.

RNA was extracted from frozen tissue using the miRNeasy Mini kit. Quality (260/280) and quantity were determined using Nanodrop. Total RNA samples were also QC checked using the Caliper HT RNA HiSens assay (Caliper Life Sciences, USA). Samples ranging from 60-255ng RNA were re-arrayed into a 96-well plate. First-strand cDNA was synthesized from the total RNA samples using the SMARTScribe Reverse Transcriptase from Clontech, BNA oligo, TRB and IGH gene specific primers at a concentration of 0.5uM. Reactions were incubated on a tetrad using the following program: 90mins at 42°C, 15mins at 70°C and 2mins at 4°C. Using cDNA as a template, first round PCR for TRB and IGH was set up using Phusion DNA polymerase (Fisher Scientific, USA) according to manufacturers specifications. The gene specific primers used were *TRB* `5'-TCTCTGCTTCTGATGGCTCAAAC-3'` and *IGH* `5'-ACACCGTCACCGGTTCG G-3'`. The PCR conditions used were an initial denaturation of 98°C for 30 s, followed by 35 cycles of 98°C for 10 s, 55°C for 10 s and 72°C for 20 s, and a final extension at 72°C for 5 minutes. PCR products were size selected and cleaned up using PCRClean DX beads (Aline

Biosciences, USA). Using first round PCR product as a template, a nested round of PCR for *TRB* and *IGH* was set up using Phusion DNA polymerase (Fisher Scientific, USA) according to manufacturers specifications. The gene specific primers used were *TRB* `5'-TGCTCTTCCGATCT GACAGCGACCTCGGGTGGGAACA-3'` and *IGH* `5'-TGCTCTTCCGATCTGACAAGACSGATGGGCCCTTGGT-3'`. The PCR conditions used were an initial denaturation of 98°C for 30 s, followed by 10 cycles of 98°C for 10 s, 65°C for 10 s and 72°C for 20 s, and a final extension at 72°C for 5 minutes. PCR products were cleaned up using PCRClean DX beads (Aline Biosciences, USA).

TRB and *IGH* amplicons were pooled by template for sequencing sample preparation. Sample preparation involved a second round of amplification using Phusion DNA polymerase with 6 PCR cycles using PE primer 1.0-DS (`5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA CGCTCTTCCGATCTCTG-3'`) and a custom PCR Primer (`5'-CAAGCAGAAGACGGCATACGAGATNNNNNN GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAC-3'`) that contains a unique six-nucleotide 'index' shown here as N's. Products were cleaned up using PCRClean DX beads (Aline Biosciences, USA). DNA quality was assessed using the Caliper LabChip GX High Sensitivity Assay (Caliper Life Sciences, USA) and DNA quantity was measured using a Qubit dsDNA HS assay kit on a Qubit fluorometer (Life Technologies, USA).

The indexed libraries were pooled together and sequenced on the Illumina HiSeq platform with paired-end 250bp reads using v2 chemistry reagents.

3.2.3 Quantification and Statistical Analysis

3.2.3.1 WGSS analysis

Alignment Reads were aligned to the hg19 reference genome downloaded from `http://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/GRCh37-lite.fa`.

Alignments were performed using bwa [151] using the aln and sampe commands. Duplicates were flagged with Picard `http://broadinstitute.github.io/picard/`.

SNV and indel calling Somatic SNVs were called using both Strelka 1.0.14 [152] and MutationSeq 4.2.0 [153] with default parameters. Somatic indels were additionally called with Strelka. We considered a somatic SNV high quality if it was predicted by both MutationSeq and Strelka to be present in any sample from a patient, not necessarily the same sample for each program. Germline SNVs and indels were called with samtools mpileup and bcftools call 1.4.1, with default parameters.

Gene name, predicted effect and impact of SNVs and indels were annotated using SnpEff 4.0e. Mappability scores were annotated for each position using precomputed values down-

loaded from UCSC (`http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/release3/wgEncodeCrgMapabilityAlign50mer.bigWig`). For downstream analysis we only considered variants with a mappability score > 0.99.

Breakpoint calling We used deStruct [154] and lumpy [155] to call breakpoints from WGS data. deStruct breakpoints were filtered for those with at least 2 discordant reads, and at least 2 split reads. Additional filters removed breakpoints for which the reconstructed sequence was less than 120nt, and removed breakpoints with read data likelihood less than 20. Following this, the intersection of deStruct and lumpy predictions was taken, and events lying within poor mappability regions, with break distance $\leq$ 30bp, and deletions with breakpoint size $<$ 1000bp were excluded [3]. Furthermore, breakpoints overlapping germline structural variation as determined from the database of genomic variants or identification of a similar event in the matched normal sample. Classification of breakpoint and rearrangement type was performed according to [3].

Copy number calling We applied ReMixT [156] to predict allele and clone-specific copy number from WGS samples. ReMixT jointly infers clone and allele specific copy number of both segments and breakpoints, allowing for increased statistical strength for detecting subclonal rearrangements associated with subclonal copy number changes. Additionally, ReMixT uses haplotype blocks obtained from phased SNPs to increase the power for detecting small allelic imbalances resulting from subclonal copy number changes. ReMixT was run on each patients full set of WGS samples with default parameters. Accurately inferred clone specific segment copy number was used to calculate the length-normalized proportion of segments predicted with divergent clonal copy number.

In order to call high-level amplification (HLAMP), we employed identical methods to [3]. We ran TITAN [157] on WGS data to infer logR values; HLAMP was called for segments with median logR values > 1.

Identifying BRCA variants Point mutations and indels in *BRCA1* and *BRCA2* were called from germline and somatic WGS data, as described above. Variants with high SnpEff-annotated impact were used. Somatic BRCA status was determined from variant calls. Where available, clinical test results were used to determine germline BRCA status; germline variant calls were used for patients that did not consent to clinical testing. Clinically-determined BRCA status is shown in **Table 3.1**.

3.2.3.2 Clonal analysis

Mutation cluster inference We ran PyClone 0.13.0 [71] in multi-sample mode to perform initial clonal analysis. Parental copy number and tumor content estimates from ReMixT along with reference and alternative allele counts from deep sequencing data of SNVs (PCR and Illumina sequencing) were used as input for PyClone. The following SNVs were filtered out for clonal analysis: germline SNVs, SNVs absent (probability < 0.01) in all samples in a patient (probabilities computed from a binomial test, assuming a sequence error rate of 0.001), and SNVs on sex chromosomes. The MCMC chain was run for 100,000 iterations, with a burn-in of 50,000. Posterior plots were visually inspected to confirm convergence. Flat cluster assignments were produced from posterior similarity matrices using the MPEAR method described in [71]. SNVs with broad posterior cellular prevalence distributions (width of 95% credible interval $\geq$ 0.2) far from the corresponding cluster median (difference of ≤ 0.05) were excluded from further analysis. Additionally, clusters absent or present at low prevalence in all samples (median cluster prevalence across SNVs ≤ 0.05 in all samples), with only one SNV, or with $\geq 50\%$ SNVs lost were filtered out.

Archival samples without a corresponding flash frozen sample (i.e., no copy number predictions) were excluded from this initial analysis. They are reintroduced in section Clonal phylogenies & postprocessing.

Clonal phylogenies & postprocessing Filtered PyClone results were provided as input to LICHeE, a multi-sample cancer lineage inference method [75], to elucidate clonal phylogenies. LICHeE was run in cellular prevalence mode (-cp), with additional options -completeNetwork -sampleProfile. Other parameters were set to the defaults. The top ranking lineage tree from LICHeE was kept.

To remove artifacts (e.g., falsely called low prevalence clones) and obtain clonal prevalences for archival samples, clonal prevalences were refined by resampling alternative and reference allele counts for deeply sequenced tumor samples and matched peripheral blood (normal) according to the following Bayesian generative model, adapted from [29]. We suppress indices for samples as these can be treated independently.

We assume that the alternative allele counts of SNV n in the matched normal and tumor samples, b^n_{normal} and b^n_{tumor}, respectively, are distributed as:

$$b^n_{normal} | p^n_{normal} \sim \text{Binomial}(d^n_{normal}, p^n_{normal}) \tag{3.1}$$

and

$$b^n_{tumor}|\psi^n, Z_n = c, p^n_{normal} \sim \text{BetaBinomial}(d^n_{tumor}, \xi(\psi^n, \phi^c, t, p^n_{normal}), \sigma_{tumor}) \tag{3.2}$$

where d^n_{normal} and d^n_{tumor} correspond to the total read depth of SNV n in the normal and tumor sample, respectively, p^n_{normal} is the probability of observing the alternative allele of SNV n in the normal sample, σ_{tumor} is the dispersion parameter, Z_n is the cluster membership of SNV n, and $\xi(\psi^n, \phi^c, t, p^n_{normal})$, using similar notation to [71], is given by:

$$\xi(\psi^n, \phi^c, t, p^n_{normal}) = \frac{(1-t)c(g_N)}{T^n} p^n_{normal} + \frac{t\phi^c\psi^n}{T^n}$$

where ψ^n is the copy number genotype of SNV n in the tumor variant population, t is tumor content, $c(g_N) = 1$ is the copy number genotype of the alternative allele in the normal population, total copy number $T^n = 2(1-t) + \psi^n t$, and ϕ^c is the cellular prevalence of PyClone cluster c, which can be expressed as the summation of clonal prevalences f_j over clones that contain PyClone cluster c. That is:

$$\phi^c = \sum_{j:G^c_j=1} f_j$$

where G^c_j is a binary indicator of whether clone j contains PyClone cluster c. We then assume the following distributions over the parameters in equations 3.1 and 3.2:

$$\begin{aligned} \boldsymbol{f} &\sim \text{Dirichlet}(\boldsymbol{\kappa}) \\ \psi^n &\sim \text{Categorical}(\mathbf{1}) \\ p^n_{normal} &\sim \text{Beta}(\zeta * \sigma_{normal}, (1-\zeta) * \sigma_{normal}) \end{aligned}$$

with $\boldsymbol{\kappa}$ the Dirichlet parameter as defined in [29], and σ_{normal} the dispersion parameter. The value ζ corresponds to twice the mean allelic fraction of alternative alleles in the normal sample (twice because we model $c(g_N) = 1$). In essence, our model is analogous to that of [29], but we now consider the probability of sampling a variant allele from non-tumor cells to be nonzero, equal to p^n_{normal}, rather than 0.

Informally, the model can be described as follows. For each tumor sample:

1. Generate clonal prevalences

2. Compute the cellular prevalence of a mutation n by summing the prevalences of all clones containing the PyClone cluster associated with n
3. Generate the SNV-specific normal contamination fraction p^n_{normal} and allelic count data for the matched normal sample
4. Based on the contamination fraction, apply a modified PyClone likelihood model to simulate allelic count data in the tumor sample

The normal contamination fraction can be interpreted as the allelic fraction of SNV n in the matched normal, likely due to sequence errors or contamination. Samples with low tumor purity are particularly confounded by these issues; the addition of step 3 and modification of step 4 relative to [29] helps eliminate erroneously identified rare clones in these samples.

We set the following hyperparameter values: $\sigma_{tumor} = \sigma_{normal} = 200$ and $\boldsymbol{\kappa}$ as a repeating vector of 0.01. The effect of our setting for $\boldsymbol{\kappa}$ is to assume clonal purity unless there is substantial evidence for the contrary.

The Hamiltonian Monte Carlo chain was run for 10,000 iterations, with an additional burnin of 5000. Posterior plots were visually inspected for convergence. Clones falling below a prevalence threshold ($< 90\%$ of the posterior distribution of clonal prevalence > 0.01) were removed.

Due to difficulties in lineage construction for patients with several samples composed of divergent clonal lineages [29], results for patients 3 and 9 were taken from previously analyzed single-cell sequencing data [29].

Clonal architecture distance Pairwise similarity between clonal compositions (within a given patient) was computed using a modified version of the weighted uniFrac measure, to simultaneously incorporate clonal architecture and phylogeny information. First, clonal phylogenies from Section **3.2.3.2** were taken as ground truth and used to recompute cellular prevalences for all SNVs ($\boldsymbol{\psi}_a$ and $\boldsymbol{\psi}_b$) determined by WGS, where a and b denote the samples being compared. Clonal distance was computed as the summation of the differences in cellular prevalences across SNVs, or equivalently $||\boldsymbol{\psi}_a - \boldsymbol{\psi}_b||_1$.

Measures of intratumoral heterogeneity Sample mixture entropy and clone divergence were defined as in [29]. In order to compute divergence, SNVs from WGS data were assigned to PyClone clusters - and transitively, clones - by maximum likelihood according to the PyClone likelihood model [71]. Proportion subclonality (copy number based measure) was computed as the proportion of the genome with subclonal copy number according to results from ReMixT. Heterogeneity index, a combined measure of intratumoral heterogeneity incorporating both clone prevalences and phylogenetic relationships, was computed as the sum of relative phylogenetic

divergence between all pairs of distinct clones, weighted by clonal prevalence. The heterogeneity index is the mean phylogenetic divergence between a randomly selected pair of tumor cells from a sample (based on inferred clonal composition). Formally, for a sample A with clone set $C(A) = \{c_i\}$ and corresponding prevalences p_i (where $0 < p_i < 1, \sum_i p_i = 1$):

$$HI(A) = \sum_{c_j, c_k \in C(A)} p_j p_k D(c_j, c_k)$$

where $D(c_j, c_k)$ is the relative phylogenetic divergence between clones c_j and c_k, defined as:

$$D(c_j, c_k) = \frac{|S_{c_j} \cup S_{c_k}| - |S_{c_j} \cap S_{c_k}|}{|S_{c_j} \cup S_{c_k}|}$$

where S_{c_i} is the set of WGS SNVs assigned to clone c_i. By construction, the heterogeneity index obtains values between 0 and 1. Intratumoral heterogeneity values for each sample are listed in **Supplemental Table A.2**.

Samples were also assigned to clonal mixture classes (pure, chain, branched) based on the phylogenetic relationships between constituent clones. Pure samples contained a single clone; chain samples contained clones along a single lineage (in other words, the minimal spanning tree is a line); branched samples contained at least 2 clones that were not ancestors/descendants of each other (in other words, the minimal spanning tree contains a bifurcation).

The significance of differences in the 3 clone-derived intratumoral heterogeneity measures (entropy, clone divergence, heterogeneity index) between the 3 TIL subtypes was assessed with the Kruskal-Wallis test (**Figure 3.7**A). Post hoc comparisons were made with Dunns test (P-values were BH corrected).

To assess the significance of differences in subclonal copy number proportion between the 3 TIL subtypes, ANOVA was performed (aov function in R) with subclonal CN proportion as the dependent variable (logit-transformed, as subclonal CN proportion values lie between 0 and 1, exclusive), TIL subtype and cellularity as independent variables (to control for tumor cellularity). The residual plot did not indicate any substantial deviations from normality, with relatively constant variance across the fitted range. Post hoc comparisons were made with Tukeys range test (P-values were BH corrected).

3.2.3.3 RNA-seq analysis

RNA-seq raw counts for 54 primary HGSC tumors from the Australian Ovarian Cancer Study (OV-AU) [158] were downloaded from the International Cancer Genome Consortium (ICGC) Data Portal. Ensembl Gene IDs were mapped to gene symbols using biomaRt. Duplicate entries

were summarized by taking the mean of expression values. Raw counts were normalized using voom from the R package limma with quantile normalization.

3.2.3.4 Mutation signature analysis

Data Mutation signatures were jointly inferred for 102 multi-site HGSC tumors (21 patients), 62 primary HGSC tumors from the Australian Ovarian Cancer Study with BAM files [158], and 133 additional ovarian tumors (59 HGSC, 35 clear cell, 10 germinal cell, and 29 endometrioid) [3] (**Supplemental Table A.5**). Note that a POLE hypermutant (one of the endometrioid cases) was excluded from the original set of 133 cases described in [3], and while 93 cases were available from the Australian Ovarian Cancer Study, only 62 had BAM files on the data portal. Similarly processed variant calls to WGSS analysis were obtained from [3]. In order to avoid counting the same variant more than once, the union of SNVs from all samples for each multi-site HGSC patient was analyzed together as a 'meta-sample'.

Signature inference & clustering Signatures and proportions were inferred from WGS SNV and rearrangement (structural variation, SV) calls (section WGSS analysis) by applying the multimodal correlated topic model method [26]. For SNVs, the pentanucleotide context of each variant is considered. Rearrangements (deletions, duplications, inversions, and foldback inversions) were binned by breakpoint distance (<10kb, 10kb-100kb, 100kb-1Mb, 1Mb-10Mb, >10Mb) and microhomology length [26, 159]. The optimal number of SNV and SV clusters was determined using the elbow method on model log-likelihoods [26]. The probable identity of each point mutation signature is as follows: P-MMR-1 mismatch repair (MMR), P-HRD homologous recombination deficiency (HRD), P-UM ultramutator-associated mutation signature (present at very low levels in the HGSC samples; primarily observed because of an endometrial sample from [3]), P-APOBEC APOBEC, P-AGE age signature, and P-MMR-2 uncertain, but with a strikingly similar T→C substitution pattern to the MMR signature. Sample-specific and non-ancestral mutation signatures were calculated by adding signature assignment weights for all constituent variants. For non-ancestral analysis (**Figure 3.12**D), non-ancestral SNVs were defined as those not present (and not called as ancestral) in all samples from that patient, and samples with fewer than 50 non-ancestral SNVs or SVs were excluded. Prior to clustering (**Figure 3.12**A, **Figure 3.11**C,D), signature proportions were scaled across the entire pooled cohort to a standard Gaussian distribution. Hierarchical clustering was performed with Wards method and a Pearson correlation-based distance measure ($d = (1 - r)/2$, where r is the Pearson correlation coefficient). For patients in the discovery cohort with more than 2 samples, molecular subtype annotations on the heatmap correspond to the mode of subtype assignments for each

patient. The 4 described subtypes (HRD-DEL, HRD-DUP, FBI, and TD) were recovered using the dynamicTreeCut R package (or equivalently, by cutting the dendrogram into 4 clusters).

Association with immune markers RNA-seq expression data (see RNA-seq analysis, Nanostring analysis) from a set of 54 untreated primary OV-AU cases was used for the comparison depicted in **Figure 3.12**C.

Differential gene expression Differential gene expression analysis between mutation signature clusters for ICGC OV-AU cases (see RNA-seq analysis) was carried out using the limma method (R package). limma results for HRD versus FBI, TD versus FBI, and HRD versus TD contrast matrices were fed as input to the R package GAGE for gene set enrichment analysis using KEGG pathways. Pathways significantly up- or downregulated with Q ¡ 0.01 were regarded as significant. Results of differential expression analysis are shown in **Figure 3.13** and **Supplemental Table A.6**.

TCGA foldback inversions A set of $n = 433$ TCGA ovarian serous cystadenocarcinoma cases with complete copy number, clinical, hg19 exome BAM files, and array-based gene expression data was selected for analysis [14]. Selected TCGA cases are listed in **Supplemental Table A.7**. Expression data was downloaded from the TCGA data portal and clinical data was downloaded from the TCGA Pancancer project under Synapse (ID: syn1461171).

Array gene expression data was preprocessed with the voom function from limma (R package), using quantile normalization. The median of normalized expression values for genes associated with cytotoxicity (derived from Nanostring PanCancer Immune Profiling Panel annotations [149]) was computed. Samples were stratified into immune-high and immune-low classifications by thresholding on median cytotoxicity score across the cohort (**Supplemental Table A.7**). To threshold on FBI status, foldback-amplification colocalization status (FBI-AMP High, FBI-AMP Low, No AMP) for all cases was retrieved from [3]. We performed a survival analysis on FBI groups after subsetting by immune cluster. The log-rank test was used to compare survival outcomes between subgroups.

A Cox proportional hazards model was also fit to the overall survival data, using foldback-amplification colocalization status as a discrete explanatory variable, interaction terms between cytotoxicity score and FBI-HLAMP status, along with control variables for age of pathologic diagnosis and treatment regimen (columns immunotherapy, additional immunotherapy, additional drug therapy, and additional chemotherapy in the Synapse table). Age of diagnosis was binned into < 50, 50-70, and > 70 categories, and along with immunotherapy and additional

chemotherapy used as stratification variables (as these originally violated the proportionality assumption). Patients without available data for age of diagnosis (5) were excluded. To assess the validity of the proportional hazards assumption, the `cox.zph` function the survival R package was used. None of the individual proportionality assumption tests or the global test were violated.

The R formula for the model was:

```
coxph(survival ~ mutation_signature_subgroup + cytotoxicity:mutation_signature_
    subgroup + strata(age_binned) + strata(immuno_therapy) + strata(additional_
    chemo_therapy) + additional_drug_therapy + additional_immuno_therapy, data)
```

To evaluate the significance of the model including the cytotoxicity × FBI-HLAMP interaction term, we constructed an identical model, but with a cytotoxicity score as an explanatory variable without the interaction terms with FBI-HLAMP. A likelihood ratio test was performed on the resulting fits of the 2 models.

3.2.3.5 Immunohistochemistry analysis

Tissue segmentation & cell counting Slides were scanned using the Vectra Multispectral Imaging System (Perkin Elmer) and 20 random 20× images (high-powered fields, HPFs) collected for each sample. The resulting multispectral images were then analyzed using Inform software (Perkin Elmer) with the resulting cell segregation data consolidated using Spotfire (Tibco). Phenotyping algorithms were created by 2 independent researchers (K.M., S.L.) and the results validated by a 3rd researcher (A.W.Z.). Briefly a training set of 10 images, selected to be histologically diverse on visual inspection, was used by each of the researchers to train Inform to recognize the different phenotypes of interest in each image. Training was run until at least 98% validation accuracy was achieved. The 2 algorithms were compared and visual inspection used to confirm the cell counts. TIL densities for each image were calculated by normalizing validated TIL counts by total area covered by tissue in the image (in units of cells/HPF). Overall TIL densities for each slide were similarly calculated, but using the summation of TIL counts and area across all constituent images. Epithelial and stromal TIL densities employed similar calculations, with counting and area restricted to epithelial/stromal regions identified by tissue segmentation (Inform). Thus, a cell was called epithelial if it fell within epithelial regions identified by Inform, and stromal if it fell within identified stromal regions.

Correlations between TIL densities Correlations between TIL densities (epithelial and stromal CD8+, CD4+, CD20+, and plasma cell) were quantified with Spearmans correlation coefficient (**Figure 3.9**A) and *P values* of their significance were adjusted for multiple testing with the Benjamini-Hochberg method.

Clustering Hierarchical clustering of TIL density profiles was performed using Wards method with Euclidean distance. Heatmap values were obtained by normalizing (to a standard Gaussian distribution) across samples for each TIL type. For **Figure 3.1**B,C, only samples with valid epithelial and stromal TIL densities (i.e., non-zero epithelial and stromal tissue area) are shown. Additionally, for **Figure 3.1**B, only samples with both TIL density and Nanostring expression data are shown. The optimal number of clusters (3) was determined with the Dunn index.

Malignant clone similarity and TIL subtype To compare whether samples from the same TIL subtype were more clonally similar (within patients), we used a nested ranks test (nestedRanksTest R package), treating patient as a random effect. Specifically, for each pair of samples within a patient, we (1) categorize them as belonging to the same, or different TIL subtypes (`til_cluster_comparison`); and (2) compute clonal composition similarity as per Clonal architecture distance (`clonal_similarity`). Then, we run:

```
nestedRanksTest(clonal_similarity ~ til_cluster_comparison | patient_id, data)
```

3.2.3.6 Nanostring analysis

Molecular subtyping Ground truth molecular subtypes for a training set of 62 primary HGSC tumors from [158] were obtained from the authors. Matched RNA-seq data for these tumors was obtained from the International Cancer Genome Consortium (project OV-AU) and normalized according to section RNA-seq analysis. The resulting expression profiles were pooled with Nanostring-derived expression profiles, and subjected to batch effect correction with the ComBAT R package. To confirm the effectiveness of batch correction, expression profiles from all samples were hierarchically clustered. Samples from different batches were not clearly segregated.

Following this, a k-nearest neighbors classifier ($k = 5$) was trained and applied to the data using the [158] molecular subtypes as ground truth. Six-fold cross-validation accuracy of 85.8% on ground truth data was obtained, similar to that reported in [150]. As comparison, the diagonal LDA classifier attained an inferior 80.9% cross-validation accuracy and was thus not

used. To further test these molecular subtypes, a subset of 62 tumors was additionally profiled with the Affymetrix U133A2 microarray platform. As described in [27], the expression data from these tumors was normalized with RMA and quantile normalization, corrected for batch effects with ComBAT, pooled with TCGA array expression data (see TCGA foldback inversions), and subjected to another level of batch effect correction with ComBAT. Following the methods of TCGA [14], consensus non-negative matrix factorization (NMF) was applied to determine molecular subtypes ($k = 4$). NMF-derived subtypes and k-nearest neighbor-derived subtypes were largely concordant (mutual information: 0.74).

Overrepresentation of each molecular subtype or set of molecular subtypes within each IHC-based subgroup (N-TIL, S-TIL, ES-TIL) was computed relative to the other 2 subgroups and other molecular subtypes with Fishers exact test.

Pathway signature analysis Genes were grouped on the basis of pathway annotations from the Nanostring PanCancer Immune Profiling panel [149]. Metagene expression values were constructed by taking the median of expression values for constituent genes in each pathway.

3.2.3.7 TCR/BCR-seq analysis

Alignment and clonotype calling Alignment to germline TCR and BCR segments was performed with mixcr align from MiXCR 2.0 [86], using the human IMGT reference (`https://github.com/repseqio/library-imgt/releases`, commit `d993d704553c0a1e905c702ab93c99c0001b30d9`). Reads mapping to the same clonotype were clustered using `mixcr assemble`, and the resulting TRB and IGH clonotypes were exported with `mixcr export`. Clonotypes were identified by V and J germline gene names and CDR3 nucleotide sequence. All other mixcr parameters were set to the defaults.

Decontamination and quality control Clonotypes with fewer than 5 assigned reads were immediately removed. In order to filter out potential cross-sample contamination, clonotypes shared between samples from different patients were identified. Clonotypes present at an absolute prevalence (read count) in one sample > 25 times lower than in another sample from a different patient were removed (from the former sample). Consistent with contamination, samples (from different patients) arranged close by on each 96-well PCR plate contained a larger number of shared clonotypes. Finally, clonotypes that produced non-functional (frameshift or premature stop) receptor sequences were removed.

Prior to computing repertoire diversity or similarity, TCR/BCR reads were randomly downsampled (using the minimal nonzero library size across the cohort, for TCR/BCR separately)

were randomly downsampled (10 times) with replacement from each sample to account for differences in library size. Mean clonotype abundances across these resamplings were used for the computations described below, and the corresponding statistics are reported in **Supplemental Table A.2**.

Calculating repertoire diversity The following indices of diversity were calculated:

- Number of unique clonotypes
- Shannon's entropy
- Efron-Thisted index
- D50 index (`https://patents.google.com/patent/WO2012097374A1/en`)

The Efron-Thisted index estimates the total repertoire diversity (by estimating the number of unseen clonotypes), and the D50 index quantifies the preponderance of rare clonotypes in a repertoire.

Correlations between repertoire diversity and ITH were computed as Spearmans rank correlation, using the first 2 measures listed above.

Repertoire similarity analysis Pairwise similarity between TCR/BCR repertoires A and B was calculated with the Morisita-Horn index (R package vegan):

$$S(A,B) = \frac{2\sum_{i=1}^{N} A_i B_i}{|A||B|(\frac{\sum_{i=1}^{N} A_i^2}{|A|^2} + \frac{\sum_{i=1}^{N} B_i^2}{|B|^2})}$$

where A_i denotes the number of reads associated with clonotype i in repertoire A, $|A|$ and $|B|$ are the total number of clonotype reads in A and B, respectively, and N is the number of unique clonotypes in $A \cup B$.

Correlation with clonal composition TCR repertoire and clonal dissimilarity matrices were computed as described above. These dissimilarities were correlated with Mantels test. Uncorrected P-values are reported in **Figure 3.10** and **Figure 3.5**.

3.2.3.8 TCR clonotype classification

Previous studies have revealed differences in the physicochemical properties of CDR3 sequences [160] and VJ (Vβ-Jβ) gene usage [161] between CD8+ and CD4+ T cells. We designed a binary classifier to predict the class (CD8+ or CD4+) of a T cell receptor based on both germline VJ genotype and physicochemical properties of the TCR CDR3 sequence.

Training data To train the classifier, unprocessed TCR sequence data from flow-sorted naive CD8+ and CD4+ mononuclear cells derived from 18 unrelated healthy donors were obtained from a previous study [162]. We made an effort to obtain TCR-sequence data of flow sorted CD8+ and CD4+ T cells from other sources as well [160, 161], but these data were short-read or had been preprocessed (with no raw sequence files available), and thus not amenable to uniform downstream analysis. While these training data were derived from naive T cells, [161] have reported that there are no significant differences in Vβ and Jβ usage between naive and memory T cells (for both CD4+s and CD8+s separately). For the analysis described below, we operated under the assumptions that differences in VJ gene usage patterns and CDR3 physicochemical features between CD4+ and CD8+ T cells are similar in the training and multi-site HGSC datasets. We later assessed the validity of these assumptions by comparing predicted CD8/CD4 abundance with results from immunohistochemistry (see Classifier). Alignment and clonotype calling were carried out according to the methods described in Alignment and clonotype calling. Twenty percent of the data, stratified by class, was randomly split off for testing; 5-fold cross-validation was carried out on the remaining 80%.

Features V and J genotypes were binarized (80 features). Additionally, Atchley factors (R package HDMD) quantifying the physicochemical properties of amino acids at each position in the CDR3 were used ($5n$ features, where n is the CDR3 amino acid length). Separate classifiers were trained for each length category between 11 and 18 amino acids (0.70 of all clonotypes). The distribution of V and J gene usage was comparable between training and test data.

Classifier A binary gradient-boosted tree classifier was trained on the data described in section Alignment and clonotype calling. Training with 5-fold cross-validation was allowed to proceed until 100 consecutive rounds of no improvement in validation accuracy. Based on area under the receiver operating characteristic curve, the gradient-boosted tree classifier outperformed random forest, logistic regression, support vector machine (SVM), and extreme value regression classifiers. The classifier was then applied to clonotype calls from TCR-seq data of multisite HGSC samples to predict whether each clonotype was CD8-type or CD4-type. Clonotypes assigned to either class with >80% probability were kept.

Clonotype distribution broadness across tumor samples within each patient was computed with Simpsons diversity index on the vector of per-sample relative clonotype prevalence values (R package vegan). The significance of differences in the distribution broadness between CD4+ and CD8+ associated TCRs was evaluated by computing the average of CD4+ and CD8+ TCR distribution broadness values within each patient, and applying the Wilcoxon signed-rank test

for paired data between the two groups.

3.2.3.9 Neoantigen analysis

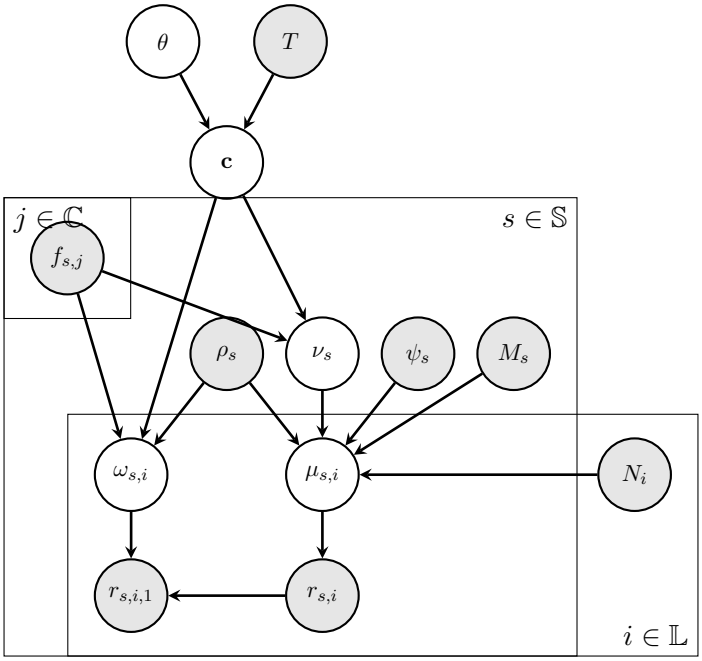
HLA typing Four-digit HLA class I types were determined from WGS data for each multisite and background patient (see Neoantigen depletion score) using OptiType [163]. OptiType was run on the WGS bam of the normal sample.

Sample-level HLA LOH prediction For OV-AU and [3] patients, HLA class I loss-of-heterozygosity (LOH) was called from tumor and matched normal bams as well as OptiType 4-digit HLA types using LOHHLA [52]. HLA LOH was called for an allele if the estimated copy number (with binning and B-allele frequency settings) was < 0.5 and the significance of allelic imbalance $p < 0.1$ (paired t test, no duplicate counts). A less stringent P-value threshold (compared to [52]) was used due to the lower depth of the input bams.

Clone-level HLA LOH prediction We devised a Bayesian statistical extension to call clone-level HLA LOH from multi-sample WGS data leveraging clonal phylogenies and clonal compositions inferred from Clonal phylogenies & postprocessing as input. Inference is done separately for each heterozygous HLA locus and patient. We define:

T	Tumor clone phylogeny
$\mathbf{c} = \{c_j : j \in \mathbb{C}\}$	Set of HLA locus copy number genotypes, one for each clone
θ	"Stay" rate between copy number states
$f_{s,j}$	Prevalence of clone j in tumor sample s
$r_{s,i,1} \in \mathbb{N}_0$	Read depth at polymorphic site i for allele 1 in sample s
$r_{s,i} \in \mathbb{N}_0$	Total read depth at polymorphic site i in sample s (sum of allele 1 and 2)
ρ_s	Cellularity/tumor content of tumor sample s
ψ_s	Ploidy of tumor sample s
$\omega_{s,i}$	Allele 1 fraction at polymorphic site i in sample s
ν_s	Total copy number of HLA locus in sample s
$\mu_{s,i}$	Mean parameter for total read depth at site i in sample s
M_s	Multiplicative factor between WGS library sizes of tumor sample s and the matched normal sample
$N_i \in \mathbb{N}_0$	Observed read depth at site i in matched normal sample
$\mathbb{L}$	Set of all polymorphic sites between the 2 alleles at a given HLA locus
$\mathbb{S}$	Set of all tumor samples for a given patient
$\mathbb{C}$	Set of all clones in a given patient

Ploidy and cellularity estimates are assumed to be known and equal to the estimates from ReMixT [156]. We present our graphical model:



We begin by defining the clone-specific copy number genotype at a given HLA locus c_j as a $(c_{j,1},\ c_{j,2})$ tuple (allele 1 copy number and allele 2 copy number, respectively), where allele 1 can be arbitrarily assigned to either one of the 2 HLA alleles at a heterozygous locus without loss of generality. Given a clonal phylogeny T, we assume that the latent clone-specific copy number genotype at a given HLA locus evolves according to a Markov chain with transition rate $1-\theta$, "stay rate" θ and the initial state distribution defined to be uniform across all possible genotypes. The transition and stay rates can be described by an n-by-n transition matrix P (n is the total number of genotype states) with diagonal entries $P_{ii} = \theta$ and non-diagonal entries satisfying $\sum_{j,j\neq i} P_{ij} = 1-\theta$. In addition, the total transition probability $1-\theta$ is divided evenly amongst all valid transitions (transitions from zero to non-zero allelic copy number are deemed invalid, as an allele cannot be acquired from nothing).

We use Markov chain Monte Carlo (MCMC) to sample from the posterior of $\mathbf{c}$, the assignment of genotypes to clones described above. In what follows we describe our proposed distributions for the observed data given $\mathbf{c}$. We assume that, given $\mathbf{c}$, the observed read depth of allele 1, $r_{s,i,1}$, is distributed as:

$$r_{s,i,1}|\mathbf{c} \sim \text{BetaBinomial}(r_{s,i}, \omega_{s,i}, \sigma),$$

where σ is the dispersion parameter and $\omega_{s,i}$, the fraction of allele 1 in tumor sample s (accounting

for normal contamination), is given by:

$$\omega_{s,i} = \sum_{j \in \mathbb{C}} \rho_s f_{s,j} c_{j,1} + (1 - \rho_s).$$

To then anchor the total copy number estimates, we use data from the matched normal bam. Given $\mathbf{c}$, we assume that the total observed read depth at site i in sample s, $r_{s,i}$, follows:

$$r_{s,i} | \mathbf{c} \sim \text{NegBinomial}(\mu_{s,i}, \alpha),$$

where α is the hyperparameter of the Gamma-distributed rate parameter in the negative binomial, and $\mu_{s,i}$, the expected read depth of polymorphic site i, can be computed as:

$$\mu_{s,i} = \frac{\nu_s}{\psi_s \rho_s + 2(1 - \rho_s)} \times M_s \times N_i,$$

with ν_s, the total copy number at the HLA locus under consideration for sample s, accounting for normal contamination, given by:

$$\nu_s = \sum_{j \in \mathbb{C}} \rho_s f_{s,j} (c_{j,1} + c_{j,2}) + 2(1 - \rho_s).$$

The space of possible clonal genotypes c_j is restricted to those with total copy number ≤ 6. The dispersion parameter σ for the beta binomial distribution is set to 200, and α for the negative binomial distribution is set to 0.5.

We consider the following prior distribution for the stay rate of the genotype Markov chain:

$$\theta \sim \text{TruncNormal}(\pi, \delta, 0, 1),$$

where 0 and 1 correspond to the lower and upper bounds of the truncated normal distribution, and the mean and standard deviation π and δ were set to be relatively uninformative (0.75 and 0.4, respectively).

MCMC was run for 100,000 iterations, using 50,000 additional tuning iterations. HLA LOH for a given clone j and allele a was called when $\geq 90\%$ of the posterior trace supported $c_{j,a} = 0$.

Identification of putative neoepitopes All 8 to 11-mer peptides overlapping nonsynonymous SNVs were considered candidate epitopes. MHC-I binding affinity was computed for every mutant and corresponding wild-type allele using netMHCpan-3.0 [164]. Percentile binding scores of $\leq 2\%$, where the mutant epitope had equal or better affinity than the wild-type epitope, were

considered as putative neoepitopes. In cases of HLA LOH, predicted neoepitopes associated with the lost HLA allele were excluded (for subclonal HLA LOH, a neoepitope was only excluded if all clones containing the neoepitope also exhibited loss of the corresponding HLA allele).

Neoantigen depletion score Neoepitopes were predicted from nonsynonymous SNVs in a background set of ovarian tumors consisting of 62 primary HGSC tumors from the Australian Ovarian Cancer Study [158] and 59 additional HGSC tumors [3], following the methods described above. Following similar methods to [165], the probability of generating at least one overlapping neoepitope from each trinucleotide pattern was determined.

For each considered tumor sample (from the multi-site HGSC cohort), the expected rate of neoepitope-generating SNVs was calculated from the trinucleotide context of synonymous SNVs and the expected rate of nonsynonymous SNVs per synonymous SNV for each trinucleotide pattern. Mathematically, define $\bar{N}_s$ to be the expected number of nonsynonymous SNVs per synonymous SNV with trinucleotide pattern s and $\bar{B}_s$ to be the expected number of neoepitope-generating SNVs per nonsynonymous SNV with pattern s. Then, for a given sample i, define Y_i as the set of synonymous SNVs and N_i the set of nonsynonymous SNVs. We can write:

$$N_{pred,i} = \sum_{m}^{Y_i} \bar{N}_{s(m)}$$
$$B_{pred,i} = \sum_{m}^{Y_i} \bar{N}_{s(m)} \bar{B}_{s(m)}$$

where $N_{pred,i}$ and $B_{pred,i}$ are the expected number of nonsynonymous SNVs and neoepitope-generating SNVs in sample i under the null model, respectively. $s(m)$ is the trinucleotide pattern for synonymous SNV m. Denote $B_{obs,i}$ to be the observed number of neoepitope-generating SNVs in i, and $N_{obs,i} = |N_i|$ the observed number of nonsynonymous SNVs in i. We then define the neoantigen depletion score is:

$$E_i = \frac{\frac{B_{obs,i}}{N_{obs,i}}}{\frac{B_{pred,i}}{N_{pred,i}}}$$

Lower values of this score were interpreted as evidence of higher neoantigen depletion.

The within-patient relationship between the response, neoantigen depletion score and the covariate, epithelial CD8+ TIL density was modeled with a Bayesian linear mixed model with patient-specific random intercepts. Samples with fewer than 3 nonsynonymous mutations were

excluded. The corresponding R code (using the MCMCglmm R package) was:

```
MCMCglmm(log(observed_neoantigen_ratio/expected_neoantigen_ratio) ~ E_CD8_
    rescaled, random=~patient_id, data=data, family = "gaussian", nitt = 500000,
    thin = 500, burnin = 50000, prior  = prior)
```

where `observed_neoantigen_ratio/expected_neoantigen_ratio` corresponds to E_i, epithelial CD8+ TIL density values were rescaled between 0 and 1, the residual covariance prior was set to be relatively uninformative ($V = 1$ and $nu = 0.002$ in R), and likewise for the random effect prior ($V = 1$, $nu = 1$, `alpha.mu` $= 0$, `alpha.V` $= 1000$ in R). For the fixed effect coefficient, an uninformative prior with mean 0 and variance 10^{10} was used. Lack of autocorrelation in the MCMC traces was confirmed with `autocorr` from the coda R package. Posterior densities of parameter estimates were checked to ensure certain assumptions of the model (e.g. fixed effect being Gaussian-distributed) were met. Reported significance values correspond to area under the (right) tail of the posterior distribution of the fixed effect coefficient.

The across-patient relationship was computed similarly, but with no patient-specific intercept term. To compute subclonal- or clonal-specific correlations, observed nonsynonymous mutations (and transitively, neoepitopes) were classified based on the clonal phylogenies inferred in Clonal phylogenies & postprocessing. Similar correlations between subclonal neoantigen depletion and epithelial CD8+ TIL densities were observed using multilevel analysis (intrapatient Spearman's correlation $p = 0.034$ across the cohort and $p = 6.110^5$ in patients containing samples with highest epithelial CD8+ TIL densities; all between-patient $p > 0.2$).

Lymphocyte marker expression and HLA LOH *CD3D*, *CD8A*, and *CD8B* expression values was extracted from Nanostring expression data for HGSC cases from [3] and RNA-seq expression data from OV-AU cases (see RNA-seq analysis). As expression data from few genes was available from the Nanostring data, expression values were modeled as a function of HLA LOH using the nested ranks test (nestedRanksTest R package; gene expression as the dependent variable, HLA LOH status as the explanatory variable, and cohort as a random effect). *P*-values representing significance of the HLA LOH coefficient are shown in **Figure 3.7**H.
The corresponding R code for the nested ranks test is:

```
nestedRanksTest(expression ~ loh_status | cohort, data = data)
```

3.2.3.10 Histologic image analysis

Cell classification and tissue segmentation QuPath v.0.1.2 (`https://qupath.github.io/`) was used to detect epithelial tissue and presumptive lymphocytes on hematoxylin & eosin (H&E) pathology slides. Briefly, slides were subjected to superpixel segmentation following automated tissue detection, and intensity features calculated for superpixels. A random trees classifier was trained (by P.T.H.) to distinguish epithelial (tumor) and stromal regions from whitespace and other tissues using small sub-regions from 10 slides on the basis of 145 superpixel features to produce tissue segmentation masks. QuPaths cell detection algorithm was subsequently used to detect individual cells, and an additional random trees classifier trained to distinguish putative immune cells on the basis of 22 cellular features. Trained classifiers are available from the authors.

Hotspot identification Cell location (coordinate) data for tumor epithelial regions from the classifier were used as input for Getis-Ord Gi hotspot detection [166]. Getis-Ord Gi hotspots denote regions with statistically significant clustering of a variable of interest. Getis-Ord Gi hotspots were identified for each cell type (cancer and lymphocyte).

To identify hotspots, a grid composed of squares with side length $s = 30$ pixels was first applied to each tissue section image. Only epithelial regions of each image were considered. Grid squares devoid of cells were excluded from further analysis by applying a binary mask. Neighborhood weights were computed using a neighborhood size of $4s$ [167]. Getis-Ord Gi G_i^* values for each grid square i were computed using `localG` from spdep. For each image, permutation testing (400 random permutations of grid point counts) was applied to compute empirical P-values of G_i^*. Regions with associated $p_i < 0.05$ were called as hotspots.

Samples with no identifiable epithelial regions from which to call hotspots were excluded.

Cancer-immune hotspot colocalization Spatial colocalization between cancer and immune hotspots was computed with the following statistics [167]:

- f_C = proportion of cancer cell hotspots that are also lymphocyte hotspots
- f_I = proportion of lymphocyte hotspots that are also cancer cell hotspots
- f_{CI} = fractional area of tumor occupied by colocalized cancer-lymphocyte hotspots

3.2.4 General statistical methods

Unless otherwise indicated, correlations between continuous data types were computed using Spearman's correlation coefficient and hierarchical clustering was performed with Ward's method

on pairwise Euclidean distances. Sample sizes (n) for statistical comparisons are shown in the respective figures and supplemental figures. $p < 0.05$ was considered statistically significant (after adjusting for multiple testing with the BH method). All Dunns test P-values were BH-adjusted. All boxplot whisker ends correspond to Q1 (first quartile) - 1.5IQR (interquartile range) and Q3 + 1.5IQR. Sample size estimation was not performed.

3.3 Results

3.3.1 High-Resolution Multi-site Profiling of Immune and Malignant Populations in the HGSC Tumor Microenvironment

We assembled a cohort of 212 tumor samples from 38 HGSC patients (**Figure 3.1**). Multiple samples per patient were collected via primary debulking surgery from ovary, omentum, and other distant metastatic sites (except some relapse samples from patients 7, 11, and 23; **Table 3.1**). TIL densities were measured by multicolor IHC, cell-type colocalization with 20× histologic images, clonotype diversity in T and B cell populations with T and B cell receptor sequencing (TCR-/BCR-seq), total mRNA gene expression from the 770-gene Nanostring PanCancer Immune Profiling Panel (Cesano, 2015) augmented with 39 molecular subtyping probes [150], mutational signatures and clonal diversity of malignant cells from whole-genome sequencing (WGS; mean depth: 86×), and deep amplicon sequencing (mean depth: 16 278×, median number of loci: 188, **Supplemental Table A.1**) (**Figure 3.2**). Both WGS and immune data (IHC, TCR/BCR-seq, or Nanostring) were obtained for 101 samples from 21 of 38 patients.

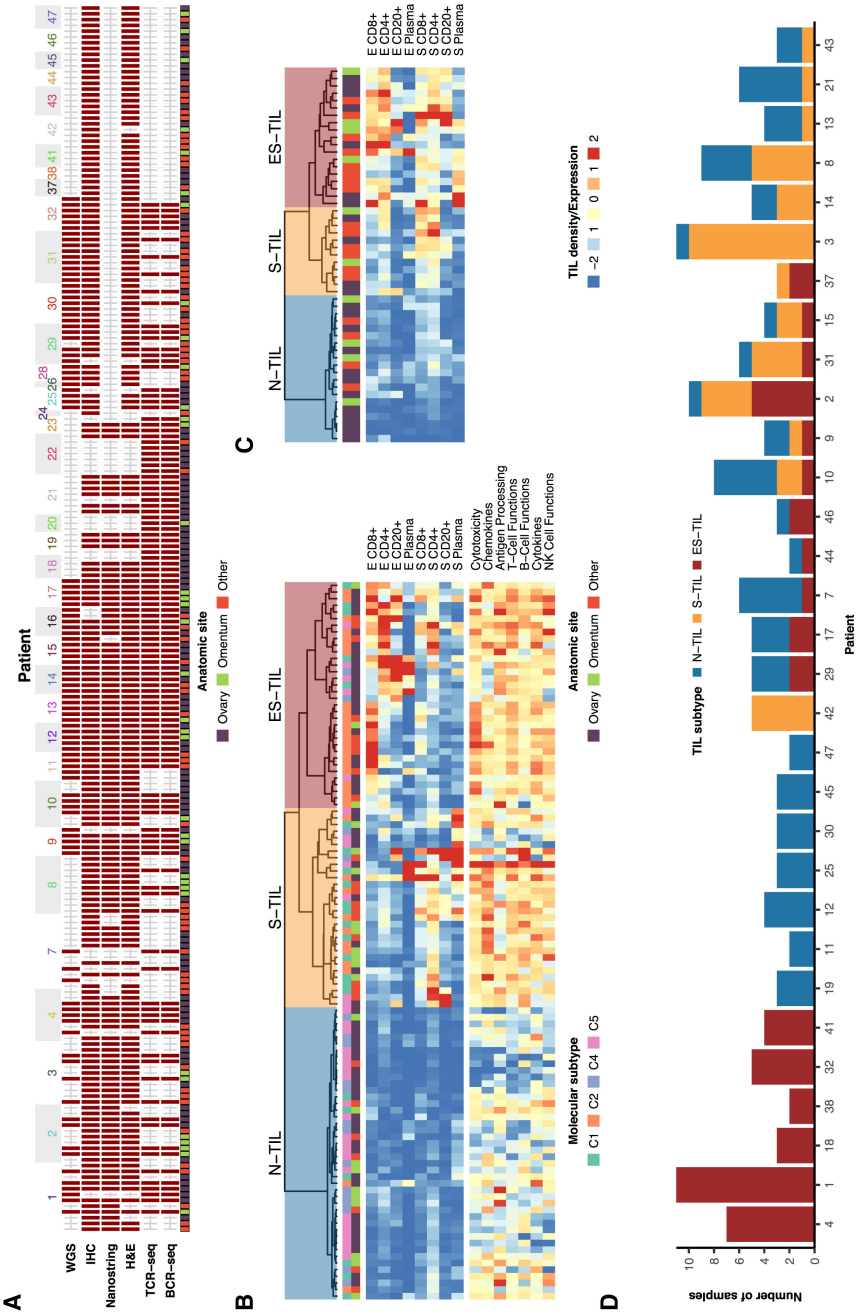


Figure 3.1: (A and B) (A) Experiments conducted on each tumor sample. Hierarchical clustering (Wards method on L2-distances) of TIL densities from (B) discovery cohort of 119 samples from 20 patients. (C) Additional cohort of 69 samples from 17 patients. Median expression of select immune pathways also shown in (B). Heatmap values standardized and clipped between 2 and 2. Samples with zero epithelial/stromal areas were removed. (D) Distribution of TIL subtypes by patient.

Patient	**Age**	**Stage**	**Recurrence**	**RFS**	**Status**	**OFS**	**BRCA status**
1	72	IIIC	no	N/A	NED	71	screen negative
2	76	IIIC	yes	12	DOD	45	screen negative
3	69	IIIC	yes	25	AWD	73	screen negative
4	53	IIIA	yes	50	AWD	71	screen negative
7(a)	47	IIIC	yes	8	DOD	52	screen negative
8	62	IIIC	no	N/A	NED	65	BRCA1 mut and unclassified BRCA2 variant
9	53	IIIB	yes	5	DOD	32	unknown
10	74	IIIC	no	N/A	NED	59	unknown
11(b)	53	IIIB	yes	32	AWD	174	BRCA2 mut
12	62	IIIC	yes	15	DOD	44	screen negative
13	80	IV	no	N/A	NED	40	screen negative
14	58	IIIC	yes	7	DOD	36	screen negative
15	61	IIIC	no	N/A	NED	38	BRCA1 VUS
16	72	IIIC	yes	23	AWD	35	screen negative
17	56	IIIC	yes	19	AWD	32	BRCA2 and MUTYH variant
18	56	IIIC	yes	19	DOD	34	unknown
19	59	IIIA	no	N/A	NED	32	screen negative
20	64	IIIA	no	N/A	NED	10	unknown
21	79	IIIC	yes	4	DOD	45	screen negative
22	73	IIIC	yes	22	AWD	75	rare BRCA2 variant (2680GA) likely benign
23(c)	65	IIIC	yes	9	DOD	75	screen negative
24	40	IIIB	yes	22	DOD	66	screen negative
25	46	IIIC	yes	6	AWD	23	screen negative
26	55	IB	no	N/A	NED	14	unknown
28	83	IIIC	yes	4	AWD	16	unknown
29	19	IV	yes	5	AWD	16	BRCA1 mut
30	38	IIIC	no	N/A	NED	13	screen negative

31	38	IIIC	yes	10	AWD	16	BRCA1 mut
32	46	IIIC	yes	1	AWD	14	screen negative
37	81	IIIB	no	N/A	NED	12	unknown
38	80	IIC	no	N/A	NED	6	unknown
41	68	IIIC	no	N/A	NED	8	screen negative
42	54	IIIC	no	N/A	NED	4	unknown
43	70	IIIC	yes	5	AWD	20	screen negative
44	35	IIIC	no	N/A	NED	6	unknown
45	79	IIIC	no	N/A	NED	5	unknown
46	77	IIIC	no	N/A	NED	6	unknown
47	45	IIIC	no	N/A	NED	4	unknown

Table 3.1: Studied patients and samples. Age refers to age (in years) at diagnosis. Recurrence-free survival (RFS) and overall survival (OFS) are indicated in months. BRCA status was determined through clinical testing. Current disease status: NED, no evidence of disease; AWD, alive with disease; DOD, dead of disease. (a): BrnM and BrnMA1 14 months, RPvM and BwlImA6 33 months post-diagnosis; (b): Pv1, Rct1, Rct2 139 months post-diagnosis; (c): LOv1 14 months post-diagnosis

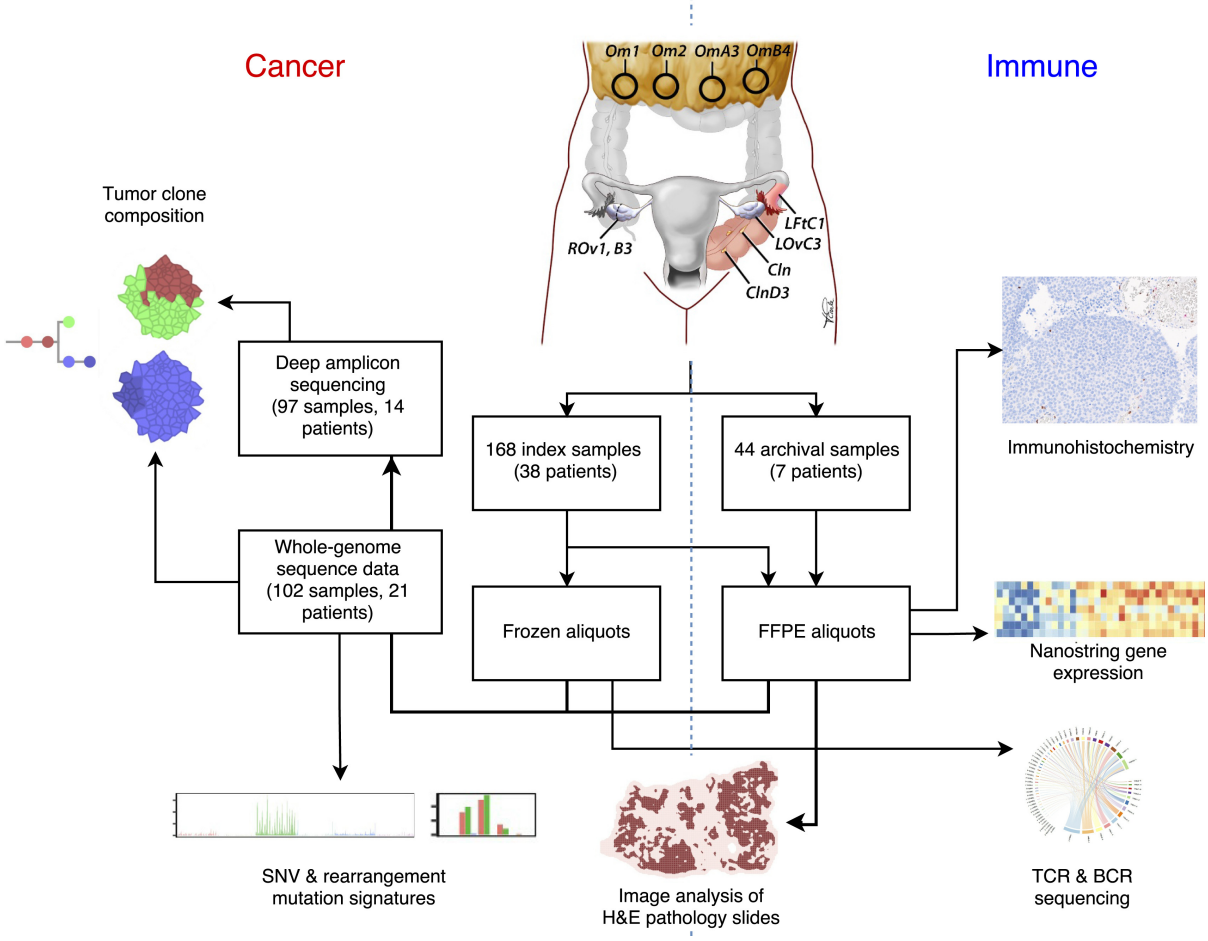


Figure 3.2: Schematic Diagram Depicting Sample Collection, Experimental Modalities, and Analysis Workflows Applied to the Data.

3.3.2 Tumor-Infiltrating Lymphocyte Subtypes Reveal Extensive Intrapatient Variation in Immune Responses across Peritoneal Sites

We began by profiling 188 tumor samples from 37 patients with multicolor IHC for CD8+ T cells (CD3+CD8+), CD4+ T cells (CD3+CD8-), CD20+ B cells (CD20+), and plasma cells (CD79a+CD138+). All but three patients were surveyed at multiple sites, providing an unprecedented view of intrapatient spatial variation. CD8+ T cells were the most abundant TIL type (0-1125.65 cells per high-powered field [HPF], median: 53.08), while CD20+ B cells were the rarest (0-136.77 cells per HPF, median: 2.74). Densities of all TIL types were correlated (**Figure 3.3**), with extensive variation across the cohort (**Figure 3.3**).

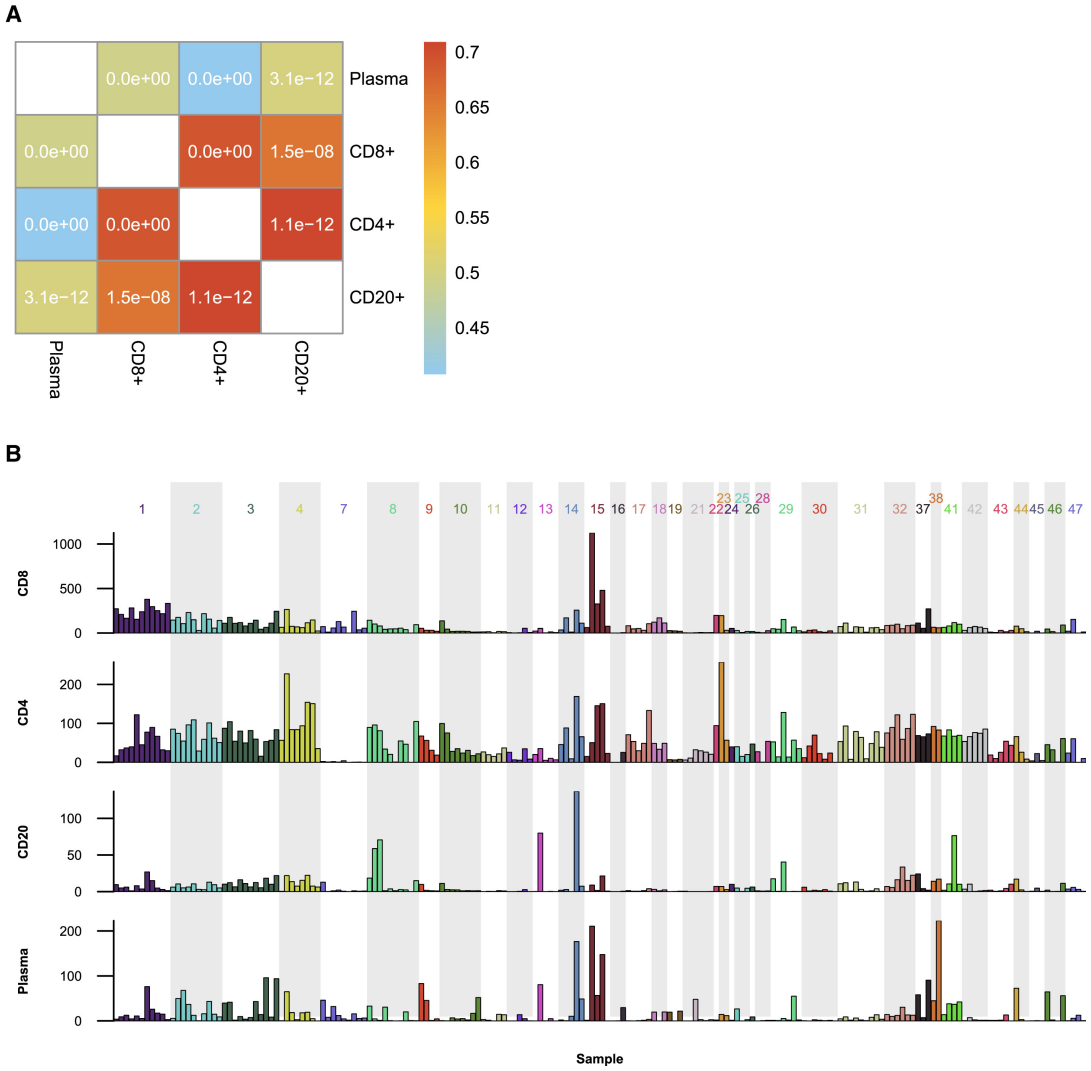


Figure 3.3: (A) Correlations between overall TIL densities. Color indicates Spearmans , P-values shown inside each cell. (B) Overall CD8+, CD4+, CD20+, and plasma cell densities across the cohort. Bars colored by patient.

Using TIL densities as input features, we first analyzed a discovery cohort of 119 samples from 20 patients. Hierarchical clustering revealed three major TIL subtypes: N-TIL (tumors sparsely infiltrated by TILs), S-TIL (tumors dominated by stromal TILs), and ES-TIL (tumors with substantial levels of both epithelial and stromal TILs) (**Figure 3.1** and **Supplemental Table A.2**). Based on orthogonal Nanostring probe counts, gene expression values for immune-associated pathways, including cytotoxicity, cytokines, and T cell- and B cell-associated genes,

were comparable between S-TIL and ES-TIL but lower in N-TIL (**Figure 3.1**). The three TIL subtypes mapped to previously described gene expression subtypes (C1, C2, C4, and C5) of HGSC [150]. N-TIL was enriched for C4 and C5 tumors ($p < 10^{-5}$, Fishers exact test), while S-TIL was overrepresented for C1 tumors ($p < 0.01$, Fishers exact test) and ES-TIL for C2 tumors, respectively ($p < 10^{-5}$, Fishers exact test; **Figure 3.1** and **Supplemental Table A.2**), suggesting previously reported HGSC gene expression subtypes [14, 101] largely reflect immune cell content. We analyzed IHC data from an additional cohort of 69 samples from 17 patients and observed a similar N-TIL, S-TIL, and ES-TIL distribution (**Figure 3.1**), indicating reproducibility of the TIL subtypes. Among patients with $\geq$2 treatment-naive samples, 14 of 31 patients harbored only one TIL subtype: seven were N-TIL only, six were ES-TIL only, and one was S-TIL only. The remaining 17 of 31 patients harbored tumors from more than one TIL subtype (**Figure 3.1**), and five patients harbored samples from all three subtypes, indicating extensive variation in immune response within patients.

While the ES-TIL pattern suggests active cytolytic TIL response against tumor cells, the presence of TILs in an epithelial region does not necessarily indicate active engagement with malignant cells. We therefore used histologic image analysis to profile microscopic spatial relationships between cancer cells and TILs. For each sample, we leveraged hematoxylin and eosin (H&E) images to identify cancer cell and lymphocyte "hotspots" within the tumor epithelium–i.e., regions of local aggregation relative to epithelial cellular density (**Figure 3.4**). We computed three measures of cancer-lymphocyte hotspot colocalization [167]: f_C (the fraction of cancer cell hotspots that are lymphocyte hotspots); f_I (the fraction of lymphocyte hotspots that are cancer cell hotspots), and f_{CI} (fractional tissue area occupied by colocalized cancer-lymphocyte hotspots) (**Supplemental Table A.2**). ES-TIL tumors exhibited high levels of overlap between cancer and lymphocyte hotspots, while S-TIL samples contained relatively low overlap (all $p < 0.05$, Kruskal-Wallis test, **Figure 3.4**). Thus, in S-TIL tumors, the rare immune cells that enter epithelial compartments appear to fail to engage with tumor cells, possibly due to lack of recognition. Although N-TIL tumors have negligible levels of TIL, they nonetheless showed occasional immune cells that could be evaluated by hotspot analysis. Where measureable, N-TIL tumors showed similar levels of colocalization as ES-TIL (**Figure 3.4**).

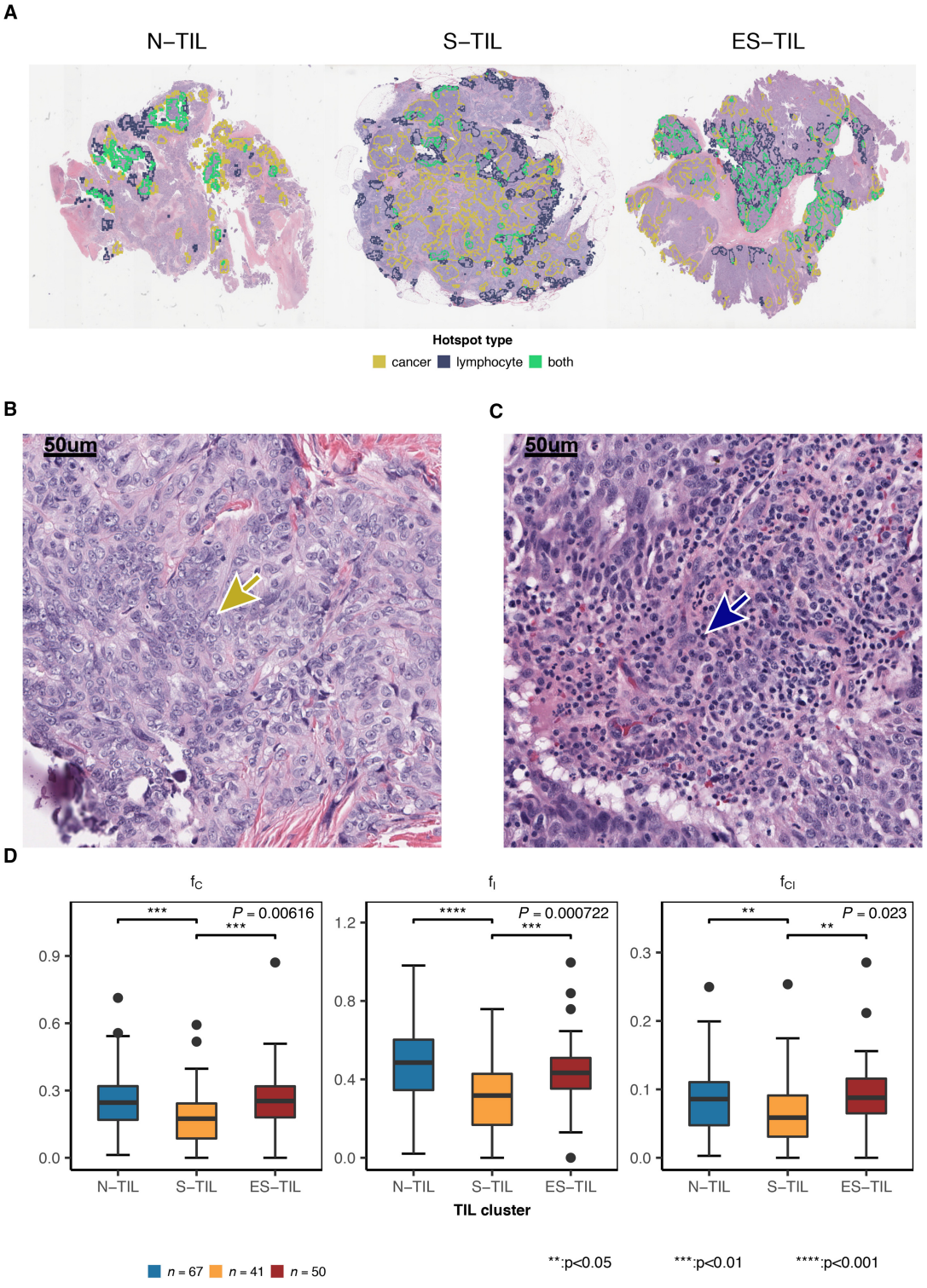


Figure 3.4: (A) Epithelial cancer cell and lymphocyte hotspots for representative N-, S-, and ES-TIL examples. (B) Histology of a cancer cell hotspot. Yellow arrow: cancer cell. (C) Histology of a colocalized cancer-lymphocyte hotspot. Blue arrow: lymphocyte. (D) Comparison of epithelial cancer-lymphocyte hotspot colocalization between TIL subtypes. *P*-values from Kruskal-Wallis test; post hoc comparisons (Benjamini-Hochberg adjusted) from Dunn's test. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR.

3.3.3 Evidence for Purifying Malignant Clonal Selection at Tumor Sites with High Epithelial Lymphocyte Infiltration

We next evaluated whether regional variation in TIL subtypes provided insight into the evolutionary trajectories and dissemination patterns of malignant clones. Using WGS on cryopreserved tissues (102 samples from 21 patients, of which 31 from 7 patients were previously described in [29]), we profiled somatic single-nucleotide variants (SNVs), allele-specific copy number, and rearrangements (**Supplemental Table A.2**) as markers of malignant clones. In addition, we performed deep amplicon sequencing on 97 samples from 14 of these patients (66 frozen and 31 formalin-fixed samples) to calculate clonal phylogenies and the clonal composition of each sample (**Figure 3.5**). We then related quantitative attributes of malignant clone composition to the N-TIL, S-TIL, and ES-TIL subtypes.

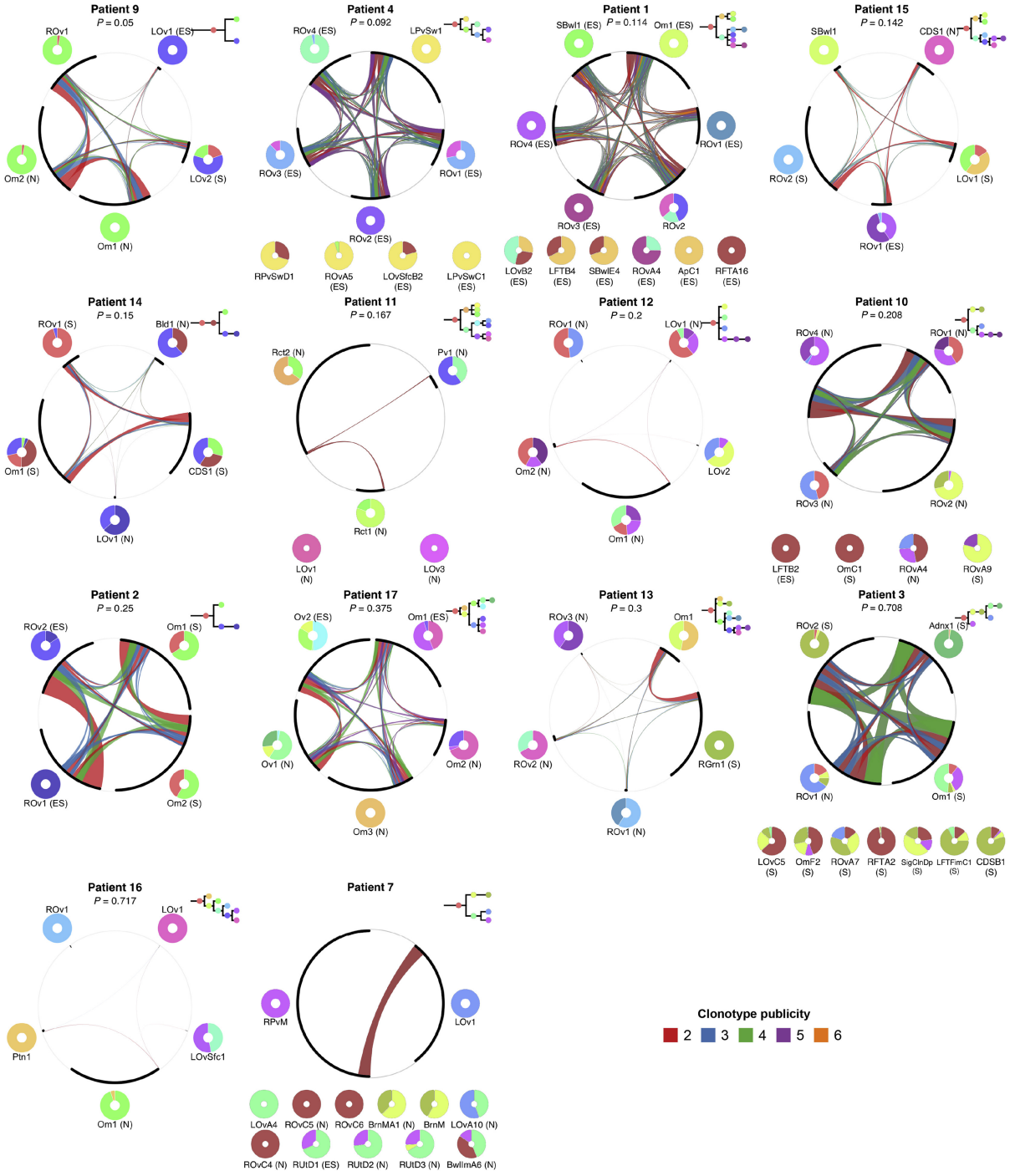


Figure 3.5: Patients are ordered by significance of the association between BCR repertoire and clonal composition dissimilarities. Chords denote shared clonotypes, width proportional to clonotype count, colored by publicity (number of samples containing a clonotype). Shared clonotypes: publicity≥2, private clonotypes: publicity = 1. Arc length (along circumference) is proportional to total clonotype count. Tumor clone composition and phylogenies shown external to each circle. Samples without BCR-seq data shown separately below each circle. TIL subtypes indicated by N (N-TIL), S (S-TIL), and ES (ES-TIL) labels. Uncorrected Mantel's test *P*-values between BCR repertoire dissimilarity and clonal dissimilarity shown below patient labels.

For each sample, we computed three continuous measures of malignant clone complexity: mixture entropy (the mixture distribution of clones present within a sample), clone divergence (the maximum phylogenetic distance between clones present within a sample; see [29]), and heterogeneity index (the mean phylogenetic distance between a randomly selected pair of clones within a sample, weighted by abundance). We also computed an orthogonal measure from WGS directly with copy-number analysis ([156]; **Supplemental Table A.2**). All four measures of ITH were correlated (all p ¡ 0.1, significance of Spearman ρ; **Figure 3.6**). For quality control, we confirmed entropy, clone divergence, and heterogeneity index were not correlated with tumor purity (all $p > 0.2$; **Figure 3.6**). We evaluated the associations between measures of malignant clone complexity and the three TIL subtypes over all treatment-naïve samples. ES-TIL samples were lower for all four ITH measures relative to S-TIL and N-TIL samples (**Figure 3.7**; accounting for tumor purity in the subclonal copy-number comparison) with mixture entropy, heterogeneity index, and subclonal copy number statistically significant. Accordingly, clonally pure tumors had the highest epithelial CD8+ TIL densities (**Figure 3.6**). Despite the association between TIL and ITH, clonal similarity between intrapatient sites was not associated with TIL subtype ($p > 0.3$, nested ranks test; **Figure 3.6**). For example, omentum sites 1 and 2 from patient 17 had comparable clonal composition, while ovary site 1 contained different clones (**Figure 3.5**); however, omentum site 1 was ES-TIL subtype, whereas omentum site 2 and ovary site 1 were N-TIL subtype (**Supplemental Table A.2**). Together, these data are consistent with epithelial TIL abundance as a negative determinant of regional malignant clonal complexity.

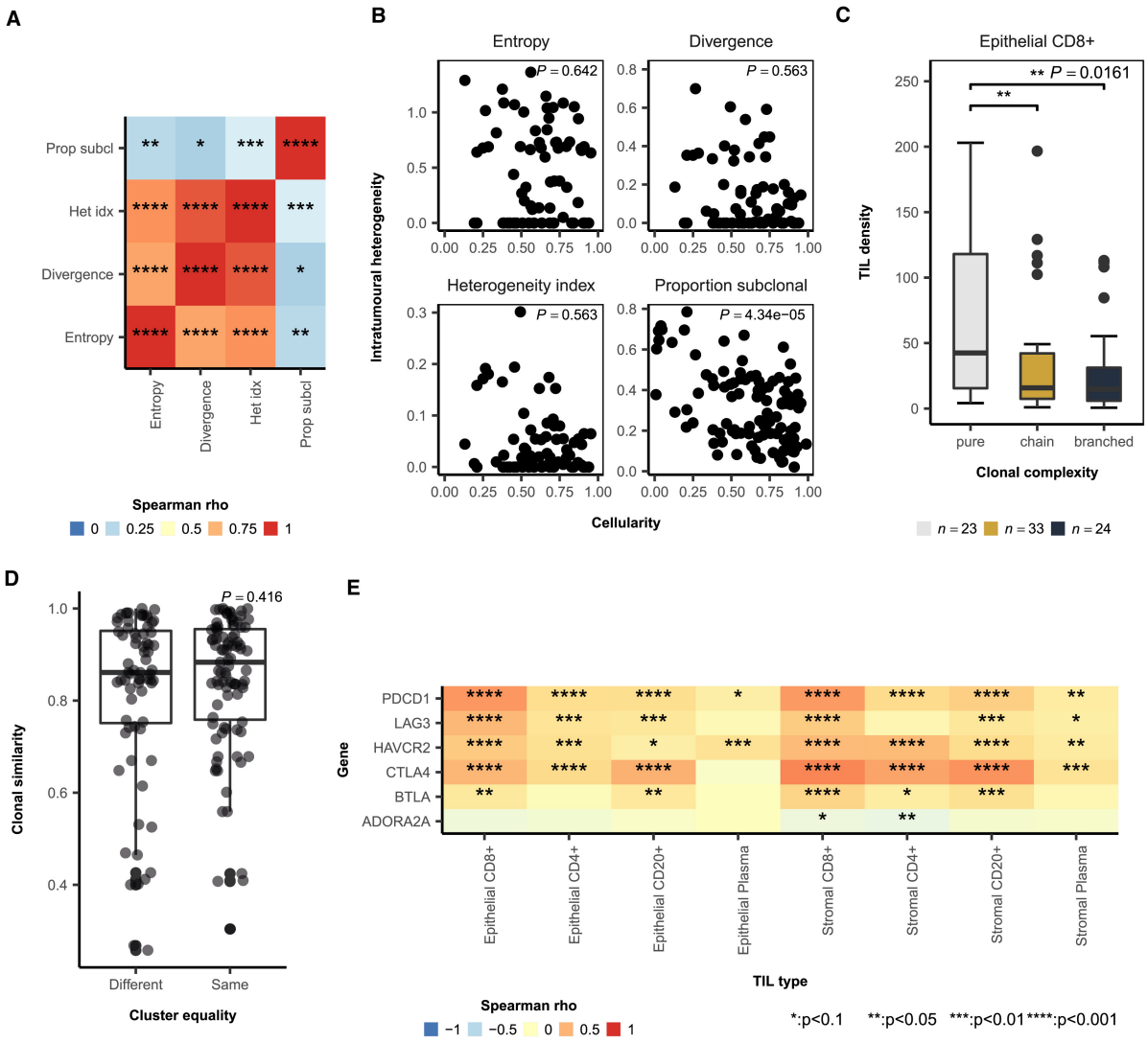


Figure 3.6: (A) Correlations between ITH measures. Asterisks indicate significance of Spearmans correlation (legend shown in D). (B) Correlations between tumor cellularity and ITH. *P*-values of Spearmans correlation are shown. (C) Epithelial CD8+ TIL densities for pre-treatment samples, stratified by clonal mixture type. P-value from the Kruskal-Wallis test shown. Whisker ends correspond to Q1 - 1.5*IQR and Q3 + 1.5*IQR. Significance of post hoc Dunn's test shown (legend in D). (D) Degree of similarity in tumor clone composition for pre-treatment samples with different or identical TIL subtypes. Subtype comparisons were made within patients; mean similarity across all comparisons was used. Lines connect comparisons made within the same patient. Nested ranks test *P*-value is shown. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR. (E) Correlation matrix between TIL densities and Nanostring-derived expression of inhibitory immune checkpoint genes. Asterisks indicate significance of Spearmans correlation.

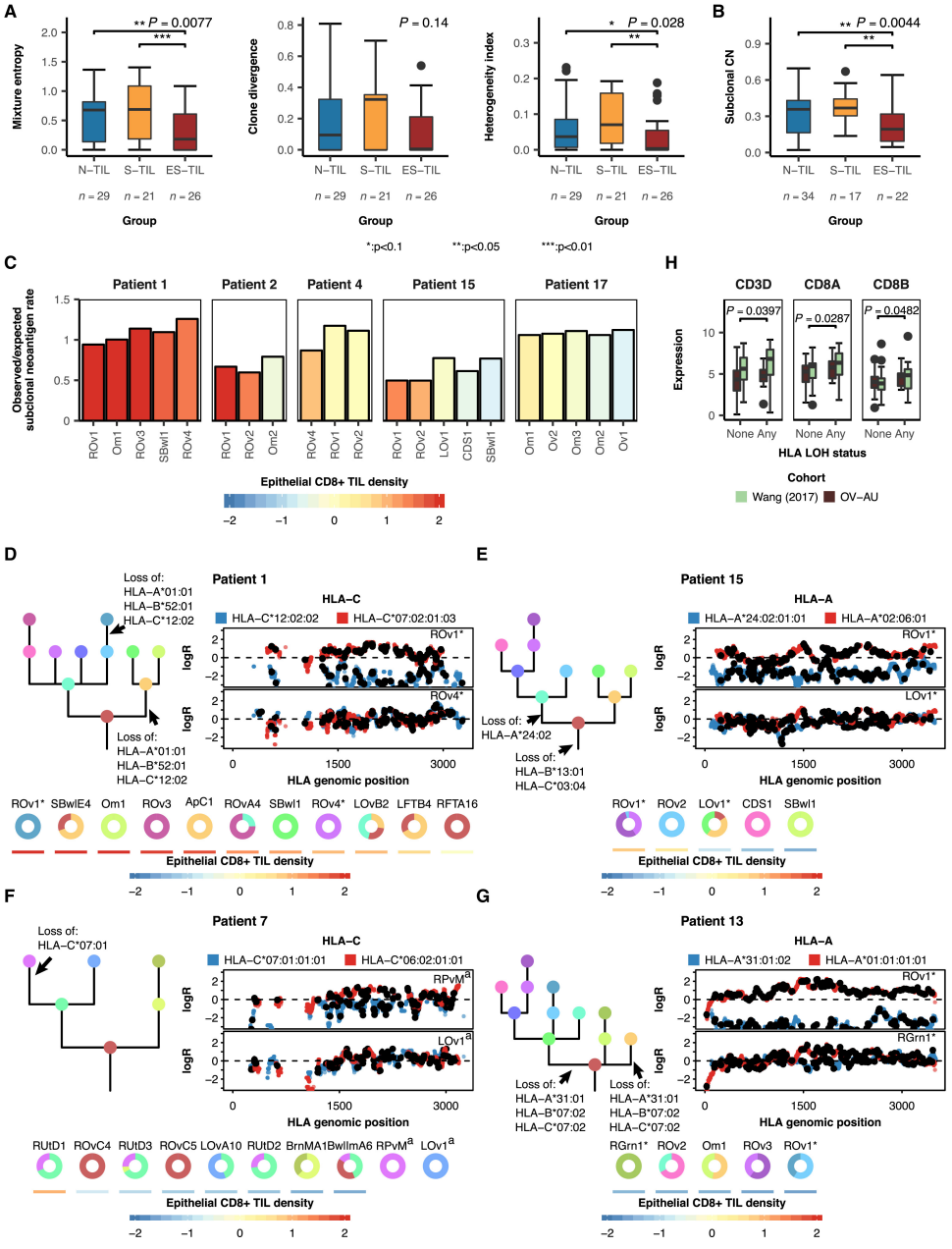

A
Mixture entropy
** P = 0.0077
***
Clone divergence
P = 0.14
Heterogeneity index
* P = 0.028
**
N−TIL S−TIL ES−TIL
n = 29 n = 21 n = 26
Group
B
Subclonal CN
** P = 0.0044
**
N−TIL S−TIL ES−TIL
n = 34 n = 17 n = 22
Group
*:p<0.1 **:p<0.05 ***:p<0.01
C
Observed/expected subclonal neoantigen rate
Patient 1
ROv1 Om1 ROv3 SBwl1 ROv4
Patient 2
ROv1 ROv2 Om2
Patient 4
ROv4 ROv1 ROv2
Patient 15
ROv1 ROv2 LOv1 CDS1 SBwl1
Patient 17
Om1 Ov2 Om3 Om2 Ov1
Epithelial CD8+ TIL density
H
CD3D CD8A CD8B
P = 0.0397 P = 0.0287 P = 0.0482
Expression
None Any
HLA LOH status
Cohort
Wang (2017) OV−AU
D
Loss of:
HLA−A*01:01
HLA−B*52:01
HLA−C*12:02
Patient 1
HLA−C
HLA−C*12:02:02 HLA−C*07:02:01:03
ROv1*
ROv4*
logR
HLA genomic position
ROv1* SBwlE4 Om1 ROv3 ApC1 ROvA4 SBwl1 ROv4* LOvB2 LFTB4 RFTA16
Epithelial CD8+ TIL density
E
Patient 15
HLA−A
HLA−A*24:02:01:01 HLA−A*02:06:01
Loss of:
HLA−A*24:02
Loss of:
HLA−B*13:01
HLA−C*03:04
ROv1* LOv1*
ROv1* ROv2 LOv1* CDS1 SBwl1
F
Loss of:
HLA−C*07:01
Patient 7
HLA−C
HLA−C*07:01:01:01 HLA−C*06:02:01:01
RPvM[a]
LOv1[a]
RUtD1 ROvC4 RUtD3 ROvC5 LOvA10 RUtD2 BrnMA1BwlImA6 RPvM[a] LOv1[a]
G
Patient 13
HLA−A
HLA−A*31:01:02 HLA−A*01:01:01:01
Loss of:
HLA−A*31:01
HLA−B*07:02
HLA−C*07:02
Loss of:
HLA−A*31:01
HLA−B*07:02
HLA−C*07:02
ROv1* RGrn1*
RGrn1* ROv2 Om1 ROv3 ROv1*

Figure 3.7: (A) Clonal measures of ITH by TIL subtype. p values from Kruskal-Wallis tests; asterisks indicate post hoc significance (Benjamini-Hochberg adjusted) from Dunns test. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR. (B) Subclonal copy-number proportion by TIL subtype. *P* value from ANOVA, controlling for cellularity. Asterisks indicate post hoc significance from Tukey's range test. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR. (C) Ratio between observed and expected neoantigen rates for pre-treatment samples in patients with highest sample-level epithelial CD8+ densities (indicated by bar color). (D-G) For patients with subclonal HLA class I LOH, (left) clonal phylogeny showing HLA LOH events and (right) logR values for samples with and without HLA LOH based on clonal composition. a:RPvM and LOv1 did not have IHC data. (Bottom) Clonal composition and epithelial CD8+ density of each sample. (D) Patient 1. (E) Patient 15. (F) Patient 7. HLA-C*07:01:01:01 was not as visually depleted in RPvM due to low cellularity (38%). (G) Patient 13. Sample labels defined in **Supplemental Table A.2**. (H) Expression of lymphocyte markers in cases with none or any HLA LOH for [3] (Nanostring) and OV-AU (RNA sequencing) cohorts. *P* values from nested ranks test. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR.

The negative association between epithelial TIL densities and malignant clone diversity could be explained by clonally complex tumors suppressing development of ES-TIL microenvironments and/or tumor clones undergoing immune-mediated purifying selection in the presence of high epithelial TIL density. In the latter scenario, subclonal (non-ancestral) neoepitopes might serve as targets of T cell recognition and hence show evidence of depletion at ES-TIL sites. To test this, we used NetMHCpan [164] to computationally predict neoepitopes from nonsynonymous somatic SNVs (**Supplemental Table A.3**), categorizing each neoepitope as clonal or subclonal through phylogenetic analysis. For each sample, we then quantified neoantigen depletion by comparing observed to expected (computed on an independent cohort of 121 primary HGSC samples) neoantigen rates. Within patients, samples with higher epithelial CD8+ density exhibited higher levels of subclonal neoantigen depletion (lower observed/expected subclonal neoantigen rate, $p = 0.09$, linear mixed model; **Supplemental Table A.3**), but not clonal neoantigen depletion ($p > 0.3$), compared to other samples from the same patient. This association was pronounced in patients containing samples with the highest epithelial CD8+ TIL densities ($p = 0.001$, linear mixed model; **Figure 3.7**). In contrast, no significant association was observed between stromal CD8+ TIL density and clonal or subclonal neoantigen depletion (all $p > 0.2$, linear mixed model). Thus, samples with high epithelial CD8+ TILs show evidence of immune editing of subclonal neoantigens, raising the possibility that immune-driven purifying selection underlies the observed reduction in malignant cell diversity at TIL-rich sites.

In tumors with high epithelial CD8+ TIL densities, we postulated that the few remaining tumor clones might have avoided immune-related negative selection through clonal expansion of cells lacking neoantigen- or other tumor antigen-presenting HLA alleles. We used a Bayesian statistical extension of the LOHHLA algorithm [52] to analyze WGS data for clone-specific HLA

class I allele loss. Of 14 patients evaluated, we identified four patients harboring clonal HLA LOH and four with subclonal HLA LOH (one patient had both; **Supplemental Table A.4**). In three out of four patients with subclonal HLA LOH, the samples with the highest epithelial CD8+ TIL densities contained tumor clones with subclonal HLA LOH (**Figure 3.7**), including two of the patients (1 and 15) that demonstrated subclonal neoantigen depletion. An exception was patient 13, where subclonal HLA LOH was observed despite all samples having low epithelial CD8+ TIL density (**Figure 3.7**; no samples were ES-TIL). Nevertheless, these findings suggest that tumor clones at ES-TIL sites have, in some cases, escaped immune clearance by somatic genomic loss of HLA haplotypes. We next examined the prevalence of HLA LOH in orthogonal WGS external cohorts [3, 158]. HLA LOH was found in 33.3% of samples (OV-AU: 34.7%, Wang: 32.1%) and was associated with significantly higher expression of lymphocyte markers (**Figure 3.7**), establishing a link between HLA LOH and higher TIL levels.

To provide context, we also considered other known mechanisms of immune escape, including anatomic site, disruption of antigen presentation machinery [168], and expression of immunosuppressive factors [165, 169]. TIL subtype was not significantly associated with any specific anatomic location (Fishers exact test, $p > 0.05$, **Supplemental Table A.2**), and no point mutations, indels, or copy-number losses in antigen presentation machinery molecules were observed in ES-TIL samples. However, consistent with expectation from previous reports [165], we found that inhibitory immune checkpoint molecules were generally upregulated in tumors with high epithelial CD8+ TIL density (**Figure 3.6**).

3.3.4 T Cell, but Not B Cell, Clonotypes Show Evidence of Tumor Clone Tracking

We next investigated whether T and B cell clonotypes associate with tumor clones. We applied TCR β chain and BCR heavy-chain sequencing to total RNA from 116 samples (27 patients) and defined the clonotype-level composition of T and B cell populations in each sample (**Figure 3.8** and **Figure 3.9**, **Supplemental Table A.2**). TCR diversity was positively correlated with IHC-based CD8+ and CD4+ TIL densities (all Spearman $p < 10^{-5}$; **Figure 3.9**). Similarly, BCR diversity was positively correlated with CD20+ and plasma cell densities (all Spearman $p < 0.01$; **Figure 3.9**). S-TIL and ES-TIL tumors had significantly more diverse TCR and BCR repertoires than N-TIL tumors (**Figure 3.8** and **Figure 3.9**) and a higher proportion of rare clonotypes (**Figure 3.9**). None of the four ITH measures were significantly associated with TCR or BCR diversity across treatment-naive samples (all Spearman $p > 0.3$), indicating that diverse malignant populations do not recruit similarly diverse TIL repertoires.

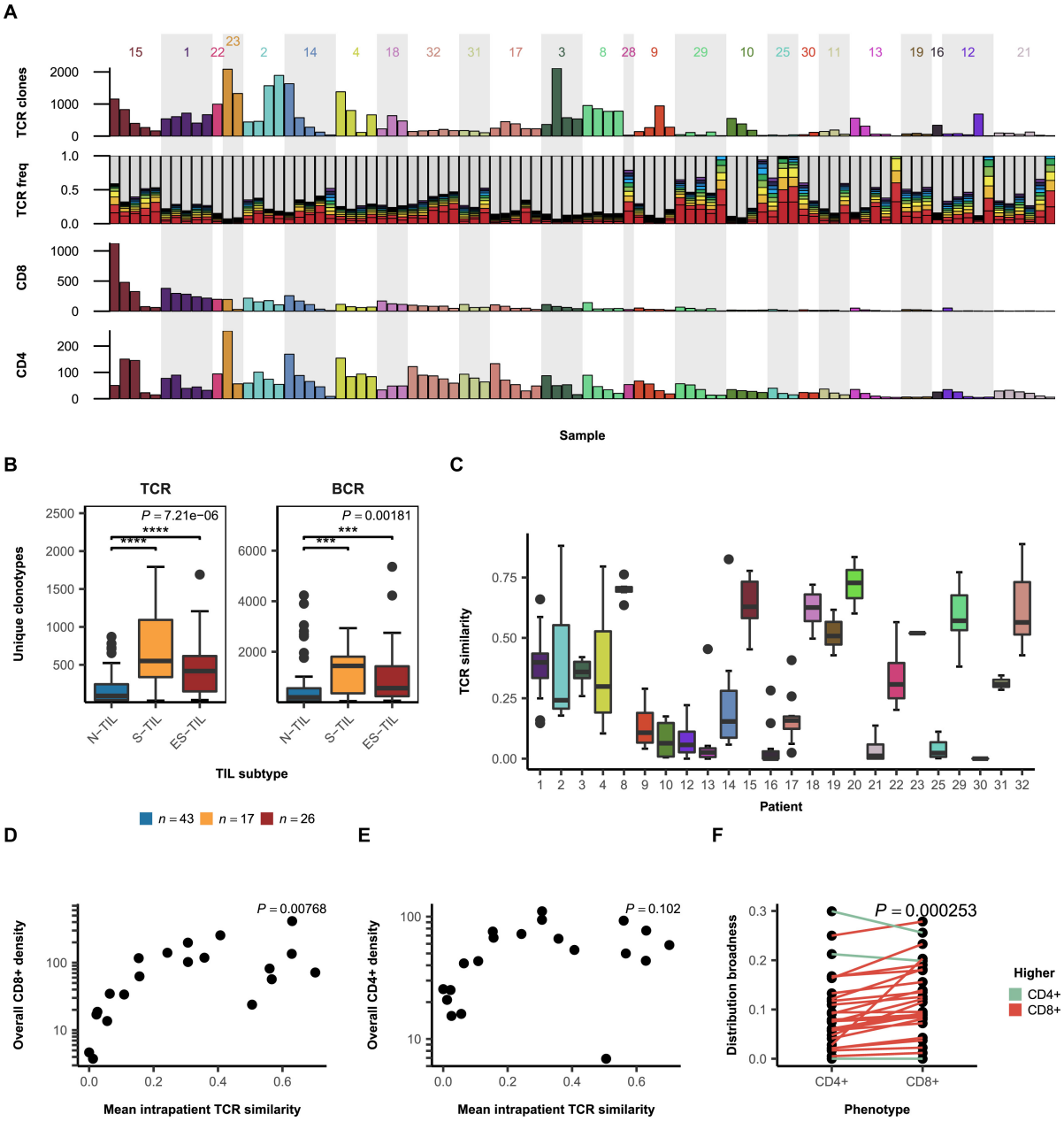


Figure 3.8: (A) Number of unique TCR clonotypes, prevalences of top 10 clonotypes (gray = all others), and CD8+ and CD4+ TIL density for each sample. (B) Comparison of unique TCR and BCR clonotype counts between TIL subtypes. p values from Kruskal-Wallis tests; asterisks indicate post hoc significance (Benjamini-Hochberg adjusted) from Dunn's test. (C) Distribution of pairwise TCR similarity for each patient. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR. (D and E) Scatterplot of mean intrapatient TCR similarity and (D) CD8+ TIL density and (E) CD4+ TIL density. P value of Spearman ρ shown. (F) Mean repertoire broadness for CD8+ and CD4+ type clonotypes in each patient. P value from Wilcoxon signed-rank test. Post-treatment tumors excluded in (B), (C), (D), (E), and (F).

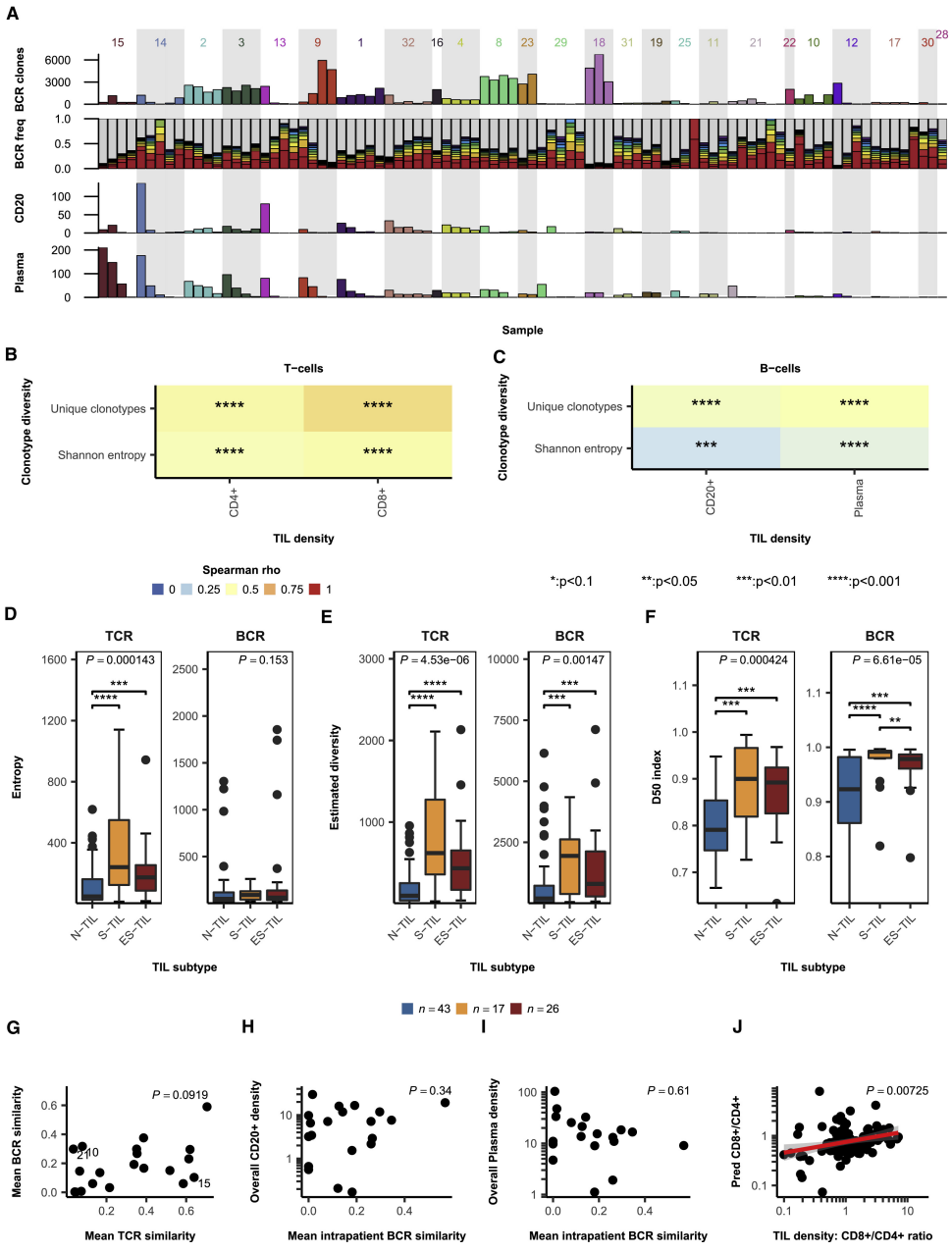

A
BCR clones
BCR freq
CD20
Plasma
Sample
B
T−cells
Clonotype diversity
Unique clonotypes
Shannon entropy
CD4+
CD8+
TIL density
C
B−cells
CD20+
Plasma
Spearman rho
*:p<0.1
**:p<0.05
***:p<0.01
****:p<0.001
D
TCR
BCR
Entropy
P = 0.000143
P = 0.153
N−TIL
S−TIL
ES−TIL
TIL subtype
E
Estimated diversity
P = 4.53e−06
P = 0.00147
F
D50 index
P = 0.000424
P = 6.61e−05
n = 43
n = 17
n = 26
G
Mean BCR similarity
P = 0.0919
Mean TCR similarity
H
Overall CD20+ density
P = 0.34
Mean intrapatient BCR similarity
I
Overall Plasma density
P = 0.61
J
Pred CD8+/CD4+
P = 0.00725
TIL density: CD8+/CD4+ ratio

Figure 3.9: (A) Unique BCR clonotype count, relative frequencies of the top 10 BCR clonotypes (gray = all other clonotypes), CD20+ and plasma TIL density for each sample. (B and C) Correlations between (B) overall CD8+ and CD4+ densities and TCR diversity; (C) overall CD20+ and plasma densities and BCR diversity. Diversity was quantified as the (1) number of unique clonotypes and (2) the entropy of the clonotype abundance distribution. BH-adjusted P-values of Spearmans are shown. (DF) TCR and BCR repertoire diversity across TIL subtypes. Diversity was measured by (D) Shannon entropy of clonotype prevalences, (E) Efron-Thisted index, and (F) D50 index. Whisker ends correspond to Q1 - 1.5*IQR and Q3 + 1.5*IQR. P-value from Kruskal-Wallis tests shown; asterisks indicate post hoc Dunn's test significance. (G) Correlation between mean intrapatient TCR and BCR repertoire similarity. Spearman correlation P-value is shown. (H) Correlation between mean intrapatient BCR repertoire similarity and CD20+ TIL density. P-value of Spearman ρ is shown. (I) Correlation between mean intrapatient BCR repertoire similarity and plasma cell density. P-value of Spearman ρ is shown. (J) Consistency between CD8+/CD4+ ratios from immunohistochemistry and from TCR-based prediction. P-value of Spearman ρ is shown.

We next ascertained the degree of homogeneity (similarity) between TCR and BCR repertoires across spatial samples within patients. This revealed marked variation in both intrapatient TCR and BCR similarity across the cohort (**Figure 3.8** and **Figure 3.9**). Considering patients with at least three samples, the extent of intrapatient TCR and BCR repertoire similarities were correlated (Spearman $p < 0.1$), but with notable exceptions (**Figure 3.9**). Patient 15 had high TCR similarity (ranked 2nd out of 20 patients), but not BCR similarity (14th), while patients 10 and 21 had high BCR similarity (3rd and 5th), but not TCR similarity (15th and 20th). Mean intrapatient BCR similarity was not significantly correlated with IHC-based CD20+ or plasma cell density (all Spearman $p > 0.2$, **Figure 3.9**). However, mean intrapatient TCR similarity was strongly associated with CD8+ (Spearman $p < 0.01$), but not CD4+, TIL density (**Figure 3.8**), suggesting that CD8+ TILs were more broadly distributed (shared) across tumor sites compared to CD4+ TILs. To test this, we trained a classifier to separate TCRs as CD8+ type or CD4+ type on the basis of V/J genes and physicochemical properties of the hypervariable domain. The ratio of CD8+-/CD4+-type TCRs was correlated with the ratio of CD8+/CD4+ densities by IHC (Spearman $p < 0.01$; **Figure 3.9**). Corroborating our predictions, CD8+-type TCRs were significantly more broadly distributed than CD4+-type TCRs ($p < 0.001$; **Figure 3.8**).
Having established that TCR-/BCR-based immune profiles vary across space, we asked how this variation is related to the spatial distribution of tumor clones. Pairwise T cell repertoire similarity was significantly correlated with malignant clone composition similarity in 7 out of 13 patients (**Figure 3.10**). Importantly, this relationship was significant in 5 of 6 patients with the highest epithelial CD8+ TIL densities (patients 1, 2, 9, 15, and 17), consistent with T cell clonotypes spatially tracking tumor clones in patients with high epithelial CD8+ TILs. This association held in the same six patients when considering only major TCR clonotypes (most abundant clonotypes constituting the top 50% of reads within each patient), but was only significant in

patients 2, 9, and 12 when considering minor clonotypes (all other clonotypes), indicating that the most abundant clonotypes drove this effect. In contrast, pairwise BCR similarity was not significantly correlated with tumor clone similarity in any patient (**Figure 3.5**), suggesting an absence of spatial tracking between B cells and tumor clones.

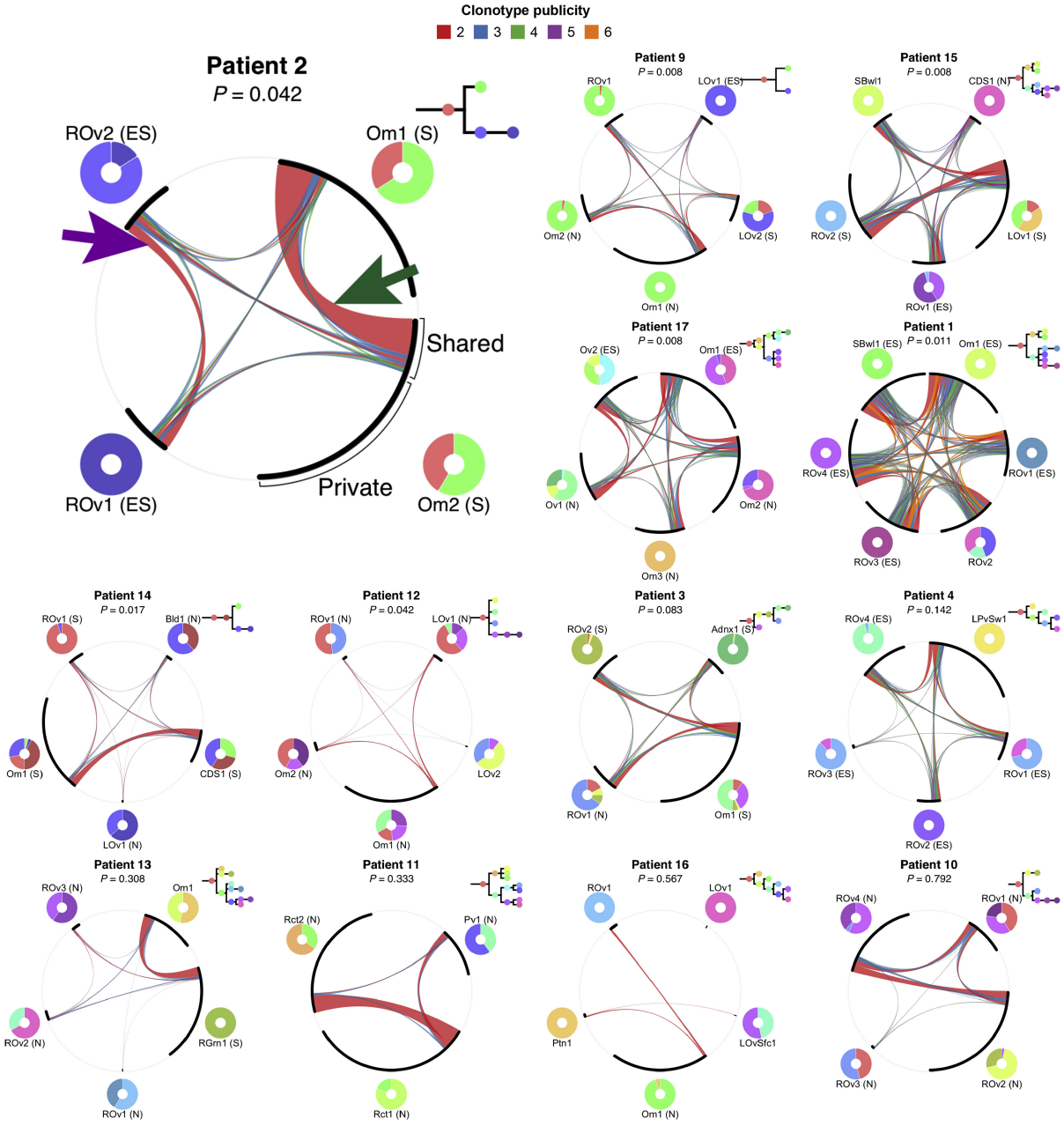


Figure 3.10: Patient 2 aside, cases ordered by significance of association between TCR repertoire and clonal composition dissimilarities (uncorrected Mantel's test p values). Chords denote shared clonotypes, width proportional to clonotype count, colored by publicity (number of samples containing a clonotype). Shared: publicity ≥ 2; private: publicity $= 1$. Purple arrow: chord denoting clonotypes shared only between right ovary sites 1 and 2. Green arrow: clonotypes shared only between omentum sites 1 and 2. Tumor clone composition and phylogenies next to each circle. TIL subtypes indicated as N (N-TIL), S (S-TIL), and ES (ES-TIL). Patient 7 excluded, as only two samples had TCR and tumor clone data.

3.3.5 Mutation Signatures Prognostically Associate with Patient-Level Immunologic Features

We next investigated interaction of malignant and immune infiltration from the perspective of mutational processes operating in HGSC. We previously identified two prognostically relevant mutation signature-associated subtypes: H-HRD and H-FBI [3]. Here, we explored whether those subtypes could explain the observed variation in immune infiltration within and between patients. We pooled WGS data from our 21 cases with 195 additional single-site ovarian cancer cases (133 from [3] and 62 from OV-AU in the International Cancer Genome Consortium [ICGC]) and applied a novel multimodal correlated topic model (MMCTM) [26], identifying six SNV and seven rearrangement signatures (**Figure 3.11** and **Supplemental Table A.5**). Hierarchical clustering by signature proportions identified four major clusters (**Figure 3.12**, **Supplemental Table A.5**): one subtype (HRD-DEL) dominated by the point mutation signature associated with homologous recombination deficiency (P-HRD) along with a short deletion signature (R-SDEL) associated with BRCA2 mutations [159], a second subtype (HRD-DUP) with P-HRD and a short tandem duplication signature (R-SDUP) associated with BRCA1 mutations [159], a third subtype (FBI) characterized by an FBI rearrangement signature (R-FB) associated with breakage-fusion-bridge [3], and a fourth, minor subtype distinguished by medium and large tandem duplications (TDs) (R-MDUP and R-LDUP, respectively) associated with CDK12 point mutations [25, 26].

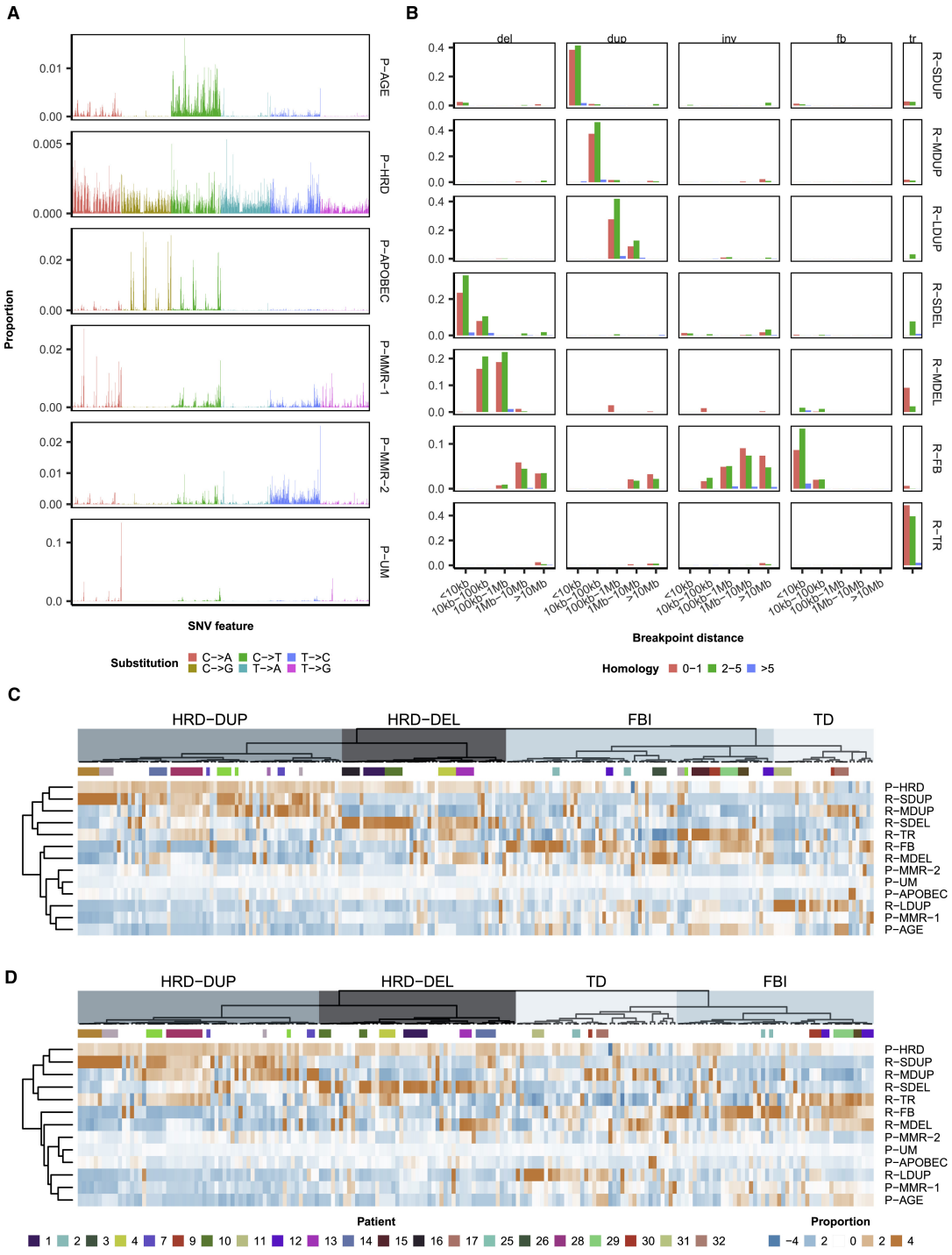

A
B
C
D
P-AGE
P-HRD
P-APOBEC
P-MMR-1
P-MMR-2
P-UM
Proportion
SNV feature
Substitution
C->A
C->G
C->T
T->A
T->C
T->G
del
dup
inv
fb
tr
R-SDUP
R-MDUP
R-LDUP
R-SDEL
R-MDEL
R-FB
R-TR
<10kb
10kb-100kb
100kb-1Mb
1Mb-10Mb
>10Mb
Breakpoint distance
Homology
0-1
2-5
>5
HRD-DUP
HRD-DEL
FBI
TD
Patient
Proportion

Figure 3.11: (A and B) Jointly inferred SNV and rearrangement signature profiles from MMCTM. Point mutation signatures: P-AGE, age associated; P-HRD, homologous recombination deficiency; P-APOBEC, APOBEC associated; P-MMR-1 + P-MMR-2, mismatch-repair associated. P-UM: ultramutator-associated (virtually absent in HGSC). Rearrangement signatures: R-TDUP, tandem duplications; R-SDUP, short duplications; R-MDUP, medium length duplications; R-LDUP, long duplications; R-SDEL, short deletions; R-MDEL, medium length deletions; R-FB, foldback inversions; R-TR, translocations. Pentanucleotide contexts are shown for each SNV signature and relative prevalences of deletions, duplications, inversions, foldback inversions, and translocations are shown for each rearrangement signature. For rearrangements, microhomology length is labeled. (C) Standardized proportions of each mutation signature for multisite HGSC, OV-AU, and HGSC [3] samples, showing clustering of samples from the same patient. (D) Standardized proportions of each mutation signature for multisite HGSC, OV-AU, and HGSC [3] samples, where only non-ancestral mutations were considered for the multisite HGSC cohort. Heatmap values were clipped between 4 and 4.

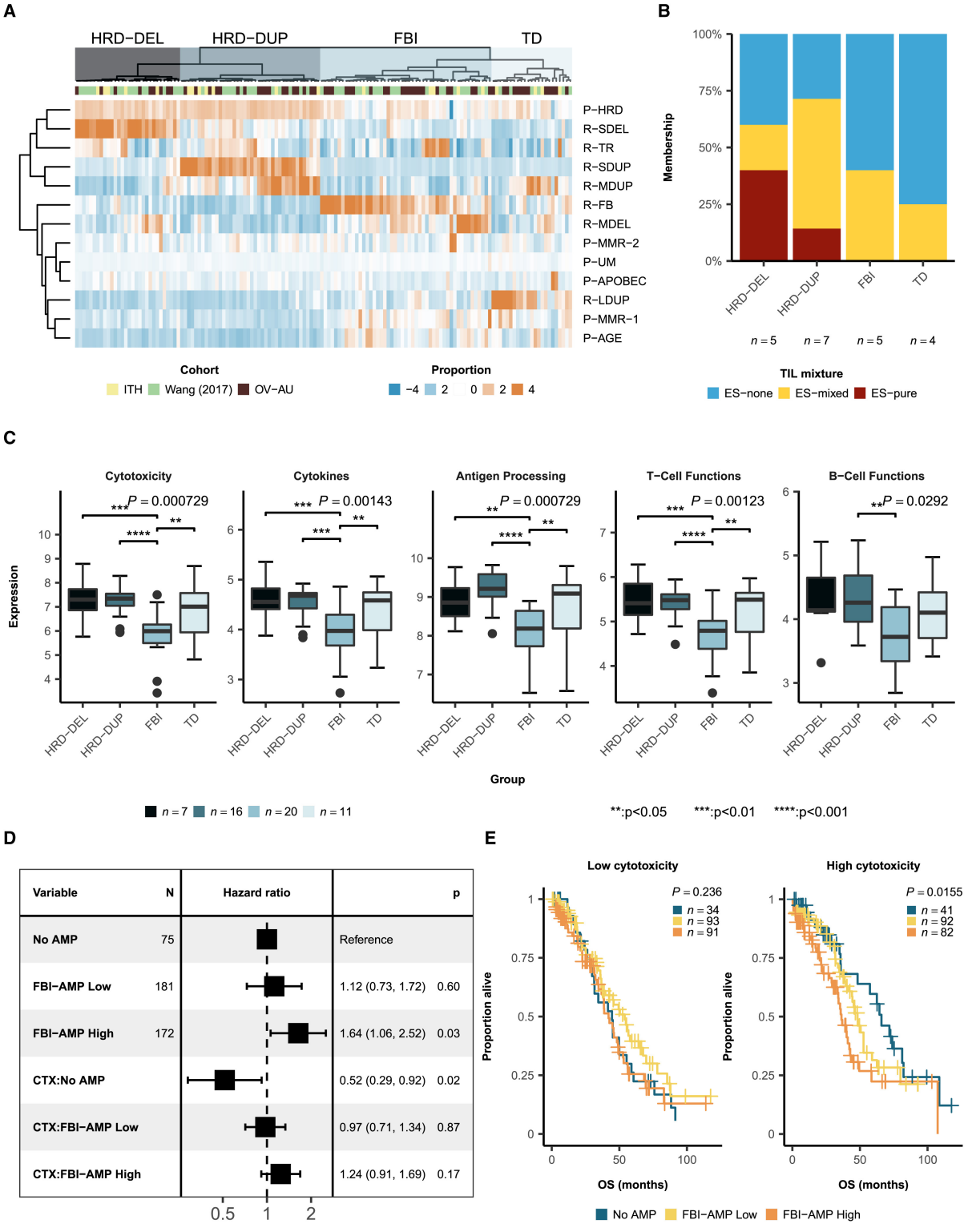

A
HRD−DEL
HRD−DUP
FBI
TD
P−HRD
R−SDEL
R−TR
R−SDUP
R−MDUP
R−FB
R−MDEL
P−MMR−2
P−UM
P−APOBEC
R−LDUP
P−MMR−1
P−AGE
Cohort
ITH
Wang (2017)
OV−AU
Proportion
−4
2
0
2
4
B
100%
75%
50%
25%
0%
Membership
HRD−DEL
HRD−DUP
FBI
TD
n = 5
n = 7
n = 5
n = 4
TIL mixture
ES−none
ES−mixed
ES−pure
C
Cytotoxicity
P = 0.000729
Cytokines
P = 0.00143
Antigen Processing
P = 0.000729
T−Cell Functions
P = 0.00123
B−Cell Functions
P = 0.0292
Expression
Group
n = 7
n = 16
n = 20
n = 11
**:p<0.05
***:p<0.01
****:p<0.001
D
Variable
N
Hazard ratio
p
No AMP
75
Reference
FBI−AMP Low
181
1.12 (0.73, 1.72)
0.60
FBI−AMP High
172
1.64 (1.06, 2.52)
0.03
CTX:No AMP
0.52 (0.29, 0.92)
0.02
CTX:FBI−AMP Low
0.97 (0.71, 1.34)
0.87
CTX:FBI−AMP High
1.24 (0.91, 1.69)
0.17
0.5
1
2
E
Low cytotoxicity
P = 0.236
n = 34
n = 93
n = 91
High cytotoxicity
P = 0.0155
n = 41
n = 92
n = 82
Proportion alive
OS (months)
No AMP
FBI−AMP Low
FBI−AMP High

Figure 3.12: (A) Signature proportions in HGSC cases standardized and clipped from 4 to 4. Dendrogram computed with Ward's method on Pearson correlation dissimilarities. ITH: multisite cohort from this study. (B) Fractions of ES-none, ES-mixed, and ES-pure patients across mutational subtypes. (C) Expression of select immune-associated pathways across mutational subtypes in OV-AU. P values (Benjamini-Hochberg adjusted) from Kruskal-Wallis test; asterisks indicate post hoc significance (Benjamini-Hochberg adjusted) from Dunn's test. Survival analysis of 433 TCGA patients. Whisker ends correspond to Q1 -1.5*IQR and Q3 +1.5*IQR. (D) Hazard ratios, 95% confidence intervals, and P values from Cox regression of overall survival. Interaction terms indicated by colons; e.g., CTX:No AMP: effect of cytotoxicity in No AMP subtype. (E) Differences in overall survival between FBI-HLAMP subgroups for tumors with low/high cytotoxicity. P values from log-rank test.

Using this grouping of samples, we asked how immune response characteristics co-segregated with mutational signatures. Unlike TIL subtypes, mutational subtypes were largely invariant within patients (**Figure 3.11**), indicating that mutational processes cannot explain intrapatient heterogeneity in TIL subtypes. We next asked whether mutational subtypes related to the mixture of TIL subtypes within each patient. Focusing on the ES-TIL subtype, we categorized patients with multi-sample IHC data as ES-none (no ES-TIL samples), ES-mixed (both ES-TIL and N-TIL/S-TIL samples), or ES-pure (all samples ES-TIL). The HRD subtypes contained the only three ES-pure patients (out of 12 HRD patients), although this did not reach significance with respect to the other mutational subtypes (Fishers exact test, $p = 0.23$; **Figure 3.12**). Expression values of immune-associated pathways [149] for 54 OV-AU cases revealed that cytotoxicity, antigen processing, cytokine, and T cell markers were highest among HRD tumors (**Figure 3.12**), concordant with similar findings in ER+ breast cancer [170] and among BRCA1-mutated tumors in HGSC [34]. Relative to HRD tumors, TD tumors had similar expression of immune markers, whereas FBI tumors were significantly depleted of these (**Figure 3.12**). Corroborating these findings, differential expression analysis of OV-AU cases revealed that antigen processing, TCR/BCR signaling, cytotoxicity, and cytokine pathways were upregulated in HRD and TD relative to FBI ($Q < 0.01$), while none of these were differentially expressed between HRD and TD (**Figure 3.13** and **Supplemental Table A.6**).

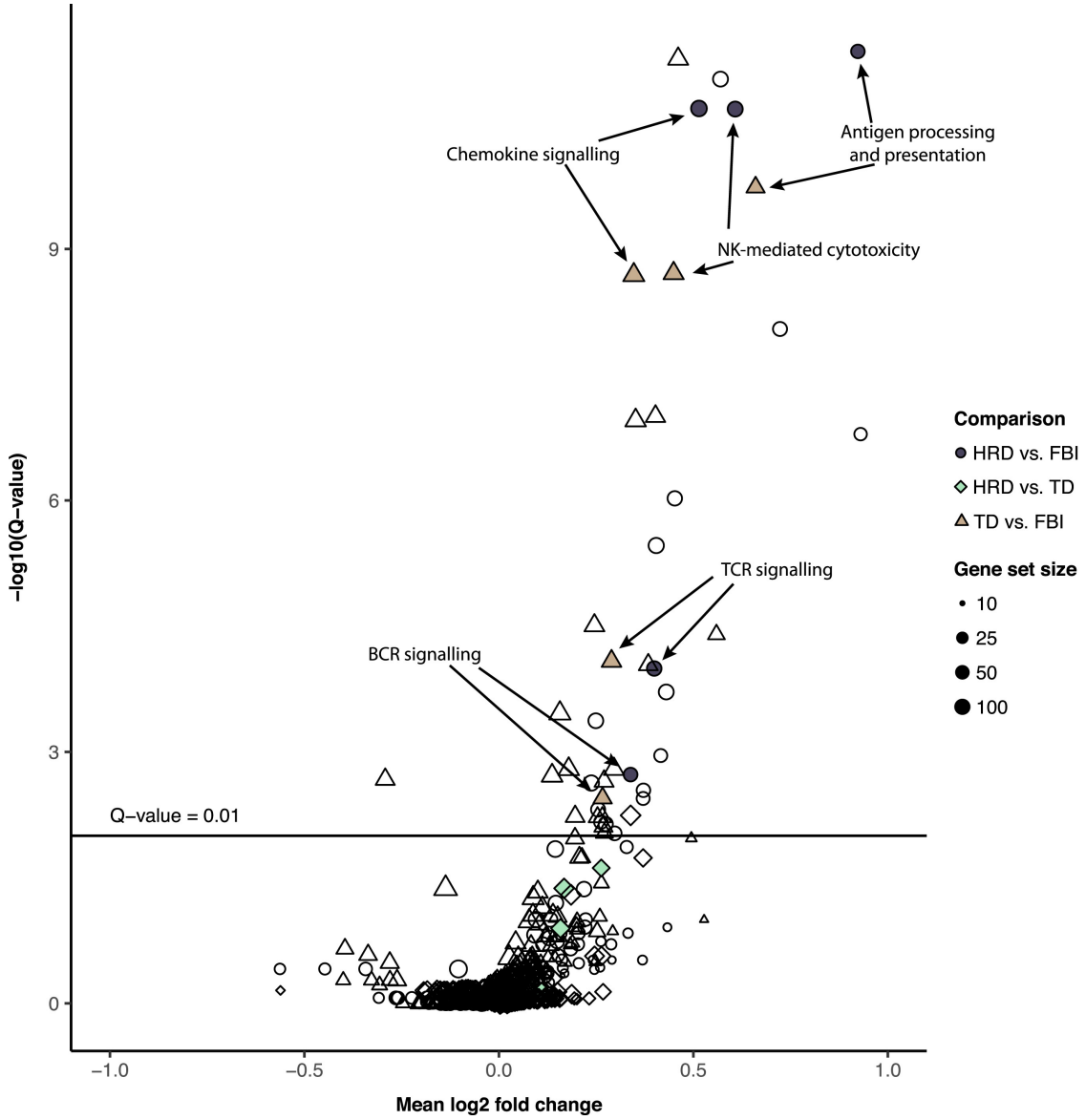


Figure 3.13: Pathway annotations derived from KEGG. Fold change indicated as, e.g., HRD versus TD: mean log2 fold change in expression from HRD to TD (> 0 = higher in HRD). Q-values computed with the BH procedure; $Q < 0.01$ was considered significant. Selected immunologic pathways are highlighted; significant hits are additionally labeled.

Colocalized foldback inversions and focal high-level amplifications (HLAMPs), thought to be reflective of breakage-fusion-bridge, have been associated with poor outcomes in HGSC [3]. We asked whether immune activity could be used to further stratify foldback-enriched tumors into

subgroups with distinct survival outcomes. Using gene expression data for 433 ovarian cystadenocarcinoma cases from the Cancer Genome Atlas (TCGA [14]; **Supplemental Table A.7**), we jointly modeled the effects of colocalized foldback-HLAMP events and cytotoxicity expression with a Cox proportional hazards model, controlling for age of diagnosis and therapeutic regimen. In agreement with [3], high levels of colocalized foldback-HLAMP events were associated with significantly shorter overall survival (hazard ratio: 1.64, 95% CI: 1.06–2.52, $p < 0.05$; **Figure 3.12**). The association between cytotoxicity and survival differed between FBI-HLAMP groups ($p < 0.05$, likelihood ratio test between Cox models with and without cytotoxicity $\times$ FBI-HLAMP interaction). In cases with no HLAMP events, cytotoxicity was significantly associated with a decreased hazard ratio (0.52, 95% CI: 0.29–0.92, $p < 0.05$; **Figure 3.12**). However, among cases with colocalized foldback-HLAMP events, the hazard ratio for cytotoxicity was not significant (FBI-AMP low: 0.97, 95% CI: 0.71–1.34, $p > 0.3$; FBI-AMP high: 1.24, 95% CI: 0.91–1.69, $p > 0.1$; **Figure 3.12**), suggesting that HLAMP-positive foldback-containing tumors harbor prognostic effects that are independent of immune response. We then median-stratified cases into low- and high-cytotoxicity groups. Low FBI was associated with significantly longer overall survival among tumors with high cytotoxicity (log-rank $p < 0.05$; **Figure 3.12**), but not low cytotoxicity (log-rank $p > 0.2$; **Figure 3.12**). Together, the covarying effects of immune activity and mutational processes suggest a combinatorial prognostic effect with high immune activity and low prevalence of FBIs leading to the best outcomes, while FBI-bearing patients have poor outcomes even in the presence of high immune activity.

3.4 Discussion

Our results illuminate evolutionary properties at the malignant-immune interface of HGSC. In patients with the highest epithelial TIL densities, our data are consistent with active pruning of malignant cell diversity by TIL through subclonal neoepitope recognition, resulting in expansion of clones harboring neoantigen loss and/or HLA LOH. The underlying mechanism likely involves tracking of tumor clones across peritoneal space by T cell clones, but not B cell clones. As such, immune infiltrates impose selective constraints, shaping patterns of malignant spread and clonal diversity in HGSC. Our findings do not exclude the possibility that T cells can also recognize clonal neoepitopes [171]; however, subclonal neoepitopes, which have been reported to have higher predicted immunogenicity than clonal neoepitopes [147], may be under stronger negative selection. Moreover, depletion of clonal neoantigens could result in complete tumor elimination and therefore go clinically undetected.

The presence of extensive intrapatient immune variation prior to treatment highlights potential

shortcomings of prognostic stratification and study of the immune microenvironment from single biopsies. The widespread multi-site variation we observed suggests that even a single site harboring relative immune privilege may be sufficient to engender resistant disease, regardless of active immune responses in distal intraperitoneal regions. We suggest immunologically sheltered havens may plausibly act as reservoirs of clonal diversity from which malignant clones impacting disease relapse might emerge. As a preliminary illustrative example, ES-pure patients had better outcomes (5 of 6 no evidence of disease [NED] or alive with disease [AWD], 5 of 6 platinum sensitive, median progression-free survival [PFS] for relapsed patients was 19 months) than ES-mixed and ES-none patients (8 of 11 and 11 of 14 NED or AWD, 7/9 and 7/11 platinum sensitive, median PFS for relapsed patients was 9.3 and 7.1 months, respectively).

Our data show for the first time a prognostically relevant interaction between mutational processes and immune response in HGSC. Notably, foldback inversions associate with poor outcomes, even in highly cytotoxic tumor microenvironments. Thus, in contrast to point mutations resulting from mismatch repair deficiency [48], FBIs likely represent a class of non-immunogenic genomic aberrations. Conversely, our findings also provide context for explaining superior outcomes observed in BRCA1- and BRCA2-mutated HGSC [34]. In contrast to previous reports that BRCA1 disruption, but not BRCA2 disruption, is associated with elevated TILs [34, 172], we observe comparably high immune activity between BRCA1-associated (HRD-DUP), BRCA2-associated (HRD-DEL), and TD subtypes. Shared deficiencies in homologous recombination between HRD and TD subtypes [173] may result in patterns of rearrangements or point mutations [3] responsible for eliciting these immune responses [170].

Our study provides context for clinical trials investigating various classes of immunotherapy in ovarian cancer (e.g., immune checkpoint blockade, adoptive T cell transfer, neoepitope vaccination, combination immunotherapy with PARP inhibition). A recent case study tracking immune response over time in a HGSC patient with a remarkable clinical trajectory [147] demonstrated that spatiotemporal variation of the immune microenvironment relates specifically to treatment sensitivity of malignant clones. We reveal that immune-microenvironment spatial variation exists prior to treatment and is prevalent in the HGSC patient population. Given that efficacy of PD-1 axis blockade hinges on pre-existing adaptive immunity [174], immunologically privileged sites on an otherwise highly infiltrated background may explain the limited success of immunotherapy in HGSC to date [49, 50]. While some tumors contain abundant TILs, lack of cancer cell-lymphocyte colocalization and reduced tumor-immune engagement in S-TIL sites may result from a failure of immune recognition or region-specific barriers to infiltration. Consequently, TIL abundance alone is an insufficient predictor of active immune response. Even at sites patterned by extensive epithelial TILs, neoantigen depletion and apparent positive

selection of clones harboring HLA LOH may render checkpoint blockade ineffective.

Despite these challenges, our findings inform on several potential therapeutic solutions. While FBI cases exhibit poor prognostic profiles independently of immune properties, HRD cases, typically associated with fewer foldback inversions, likely represent optimal candidates for immunotherapy approaches. Thus, mutational processes considered in conjunction with immune properties will aid in interpretation of newly initiated clinical trials examining combination PARP inhibition with checkpoint-blockade approaches. Furthermore, if obstacles to infiltration at immunologically privileged sites can be surmounted, our findings hint at the tantalizing potential that such tumor sites may represent targetable cancer cell populations, owing to their limited neoantigen and HLA depletion at baseline.

As the cancer evolution field progresses toward a more rigorous understanding of the fitness of heterogeneous clones within disease spectra and over temporal dimensions [175], external selective pressures imposed by the immune system must be considered as highly relevant factors. Here we show that high-resolution measurement of the immune microenvironment together with clonal decomposition analysis is tractable and yields novel insight into forces shaping malignant cell diversity and intraperitoneal spread. Broadly disseminated intraperitoneal disease at diagnosis in HGSC remains a formidable clinical problem. Our study informs on how regional variation at the interface of immunological and cancer cells controls dissemination and diversification of clones and simultaneously identifies microenvironmental and malignant cell properties to exploit in future immuno-oncologic therapeutic strategies for HGSC.

Chapter 4

Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironmental dynamics in human cancers

4.1 Introduction

Gene expression observed at the single-cell resolution in human tissues enables studying the cell type composition and dynamics of mixed cell populations in a variety of biological contexts, including cancer progression. Cell types inferred from single-cell RNA-seq (scRNA-seq) data are typically annotated in a two-step process, whereby cells are first clustered using unsupervised algorithms and then clusters are labeled with cell types according to aggregated cluster-level expression profiles [176]. Myriad methods for unsupervised clustering of scRNA-seq have been proposed, such as SC3 [177], Seurat [178], PCAReduce [179], and PhenoGraph [180], along with studies evaluating their performance across a range of settings [181, 182]. However, clustering of low-dimensional projections may limit biological interpretability due to i) low-dimensional projections not encoding variation present in high-dimensional inputs [183] and ii) overclustering of populations that are not sufficiently variable.

Furthermore, even in the context of robust clustering which recapitulates biological cell states or classes, few principled methods for annotating clusters of cells into known cell types exist. In contrast to unsupervised statistical frameworks, this latter step is a supervised, or classification problem. Typical workflows employ differential expression analysis between clusters to manually classify cells according to highly differentially expressed markers, aided by recent databases

linking cell types to canonical gene-based markers [184]. In situations where investigators wish to identify and quantify specific cell types of interest with known marker genes across multiple samples or replicates, such workflows can be cumbersome, and differences in clustering strategies can affect downstream interpretation [181]. Alternatively, cell types may be assigned by gating on marker gene expression, but this strategy is difficult to implement in practice as (i) gating is difficult for more than a few genes and relies on knowledge of marker gene expression levels and (ii) cells that fall outside these gates will not be assigned to any type, rather than being probabilistically assigned to the most likely cell type.

Another approach to cell type annotation is to leverage ground-truth single-cell transcriptomic data from labeled and purified cell types to establish robust profiles against which new data can be compared and classified. For example, scmap-cluster [185] calculates the medioid expression profile for each cell type in the known transcriptomic data, and then assigns input cells based on maximal correlation to those profiles. However, this approach requires existing scRNA-seq data for purified cell populations of interest. Given the technical effects associated with differences in experimental design and processing, expression profiles for reference populations may not be directly comparable to those for other single-cell RNA-seq experiments [186].

We assert that statistical cell type classification approaches leveraging prior knowledge in the literature (or from experiments) will be an effective complement to unsupervised approaches for quantitative decomposition of heterogeneous tissues from scRNA-seq data. Therefore, to address the analytical challenges inherent in both clustering and mapping approaches, we developed CellAssign, a scalable statistical framework that annotates and quantifies both known and *de novo* cell types in scRNA-seq data. CellAssign automates the process of annotation by encoding a set of *a priori* marker genes for each cell type. The statistical model then classifies the most likely cell type for each cell in the input data, using a marker gene matrix (cell type-by-gene). The model allows for flexible expression of marker genes, assuming that marker genes are more highly expressed in the cell types they define relative to others. Implemented in Google's Tensorflow framework, CellAssign is highly scalable, capable of annotating thousands of cells in seconds while controlling for inter-batch, patient and site variability. We evaluated CellAssign across a range of simulation contexts and on ground truth data for FACS-purified H7 human embryonic stem cells (HSCs) at various differentiation stages [187], showing that CellAssign outperforms both clustering and correlation based methods—more readily discriminating closely related cell types—and is robust to errors in marker gene specification. In addition, we applied CellAssign to two novel datasets generated to profile spatiotemporal tumor microenvironment (TME) dynamics in human cancers. Using the CellAssign approach, we demonstrated tumor 'ecosystem' spatial diversity in untreated high-grade serous ovarian cancer through variable

composition in stromal and immunologic cell types comprising the TME and variation in key pathways across malignant cell populations including immune evasion, epithelial-mesenchymal transition and hypoxia. Temporal dynamics were also exemplified using the CellAssign approach. We generated scRNA-seq libraries from matched diagnostic and relapsed pairs of follicular lymphoma samples, with one case having undergone histologic transformation to an aggressive lymphoma. We show compositional and phenotypic changes, including T-cell activation and HLA downregulation in cancer cells upon transformation, pointing towards an evolutionary interplay with cancer cells escaping immune recognition following transformation. In aggregate we conclude the CellAssign approach provides a robust new statistical framework through which disease dynamics in tissues comprised of mixed cell populations can be quantified and interpreted to ultimately uncover new properties and understanding of disease progression.

4.2 Methods

4.2.1 The CellAssign model

4.2.1.1 Model description

Let $\mathbf{Y}$ be a cell-by-gene expression matrix of raw counts for N cells and G genes. Suppose among those cells we have C total cell types, each of which is defined by high expression of several "marker" genes. We encode the relationship between cells and marker genes through a binary matrix $\boldsymbol{\rho}$, where $\rho_{gc} = 1$ if gene g is a marker for cell type c and 0 otherwise. To relate cells to cell types, we introduce an indicator vector $\mathbf{z} = \{z_n\}$ that encodes to which of the C cell types each cell belongs -

$$z_n = c \text{ if cell } n \text{ of type } c.$$

In order to assign cells to cell types we perform statistical inference of the probability that each cell is of a given cell type for which we must compute the quantity $p(z_n = c|\mathbf{Y}, \hat{\boldsymbol{\Theta}})$, where $\hat{\boldsymbol{\Theta}}$ are the MAP estimates of the model parameters.

Let s_n be the size factor for cell n and $\mathbf{X}$ be a $P \times N$ matrix of P covariates (such as patient of origin). Then our model is

$$\mathbb{E}[y_{ng}|z_n = c] = \mu_{ngc}$$

where

$$\overbrace{\log \mu_{ngc}}^{\text{Mean log expression}} = \underbrace{\log s_n}_{\text{Cell size factor}} + \overbrace{\delta_{gc}\rho_{gc}}^{\text{Cell type}} + \underbrace{\beta_{g0}}_{\text{Base expression}} + \overbrace{\sum_{p=1}^{P} \beta_{gp} x_{pn}}^{\text{Other covariates (incl. batch)}}$$

with the constraint that $\delta_{gc} > 0$.

The intuition here is that we expect the expression of gene g in cell type c is multifactorial, influenced by a cell type specific factor δ_{gc} if gene g is a marker for cell type c, combined with covariates expected from batch effects and other arbitrary sources. In this way we put no restriction that marker genes can't be expressed in other cell types and that they must be highly expressed in their cell type, only that they exhibit higher expression in the cells of type for which they are a marker. The quantity δ_{gc} corresponds to the average log fold change that gene g is over-expressed in cell c, which only occurs for marker genes for cell types since ρ_{gc} must equal 1 for this to contribute to the likelihood. By default we impose a lower bound such that $\delta > \log 2$, making the interpretation that a marker gene must be over-expressed by a factor of 2 relative to cells for which it is not a marker, but this is left as an option for the user. We also control for technical or sample effects through the matrix $\mathbf{X}$.

The user can specify whether or not to put a lognormal shrinkage prior over δ_{gc} values, where the mean and variance parameters of the lognormal are initialized to 0 and 1, respectively. In plot labels, `cellassign_shrinkage` refers to the version of CellAssign with this option turned on.

4.2.1.2 Inference

The likelihood is given by

$$y_{ng}|z_n = c \sim \mathcal{NB}(\mu_{ngc}, \widetilde{\phi}_{ngc})$$

where $\mathcal{NB}$ is the negative binomial distribution parametrized by a mean μ and a μ-specific dispersion $\widetilde{\phi}_{ngc}$. We define $\widetilde{\phi}_{ngc}$ as a sum of radial basis functions dependent on the modelled mean μ_{ngc} as proposed by a recent publication [188]:

$$\widetilde{\phi}_{ngc} = \sum_{i=1}^{B} a_i \times \exp(-b_i \times (\mu_{ngc} - x_i)^2)$$

where a_i and b_i represent RBF parameters to be fitted, B is the total number of *centers* in the RBF, and x_i is center i. The centers are set to be equally spaced apart from 0 to the maximum

number of counts $\max y_{ng}$.
Using EM for inference, the latent variables are $\mathbf{z} \equiv \{z_n\}$ while the model parameters to be maximized are $\boldsymbol{\delta} = \{\delta_{gc}\}$, $\boldsymbol{\beta} = \{\beta_{g0}, \beta_{gp}\}$, $\mathbf{a} = \{a_i\}$, and $\mathbf{b} = \{b_i\}$.

E-step Compute

$$\gamma_{nc} \coloneqq p(z_n = c|\mathbf{y}_n, \boldsymbol{\delta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \mathbf{a}^{(t-1)}, \mathbf{b}^{(t-1)}) = \frac{\prod_g \mathcal{NB}(\mu_{ngc}, \widetilde{\phi}_{ngc})}{\sum_{c'} \prod_{g'} \mathcal{NB}(\mu_{ng'c'}, \widetilde{\phi}_{ng'c'})},$$

where $\boldsymbol{\theta}^{(t)}$ is the value of some parameter $\boldsymbol{\theta}$ at iteration t. We then form the Q function

$$\begin{aligned} Q(\boldsymbol{\delta}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{a}^{(t)}, \mathbf{b}^{(t)}|\boldsymbol{\delta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, &\mathbf{a}^{(t-1)}, \mathbf{b}^{(t-1)}) \\ &= \mathbb{E}_{\mathbf{z}|\mathbf{Y}, \boldsymbol{\delta}^{(t-1)}, \boldsymbol{\beta}^{(t-1)}, \boldsymbol{\phi}^{(t-1)}} \left[\log p(\mathbf{Y}|\boldsymbol{\pi}, \boldsymbol{\delta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\phi}^{(t)})\right] \\ &= \sum_{n=1}^{N} \sum_{c=1}^{C} \gamma_{nc} \sum_{g=1}^{G} \log \mathcal{NB}(y_{ng}|\mu_{ngc}, \widetilde{\phi}_{ngc}) \end{aligned}$$

M-step During the M-step we optimize the above Q-function using the ADAM optimizer [189] as implemented in Google's Tensorflow [190]. By default we use a learning rate of 0.1, allow a maximum of 10^5 ADAM iterations per M-step, and consider the M-step converged when the change in the Q function value falls below $10^{-4}\%$. By default we consider the EM algorithm converged when the change in the marginal log likelihood falls below $10^{-4}\%$.

Initialization The following initializations are used for model parameters:

- β_{gp} is drawn from a $\mathcal{N}(0, 1)$ distribution
- $\log \delta_{gc}$ is drawn from a $\mathcal{N}(0, 1)$ distribution truncated at $[\log(\delta_{\min}), 2]$
- a is initialized to 0
- b is initialized to twice the square difference between successive spline bases

To deal with convergence to local optima, multiple random initializations of $\log \delta_{gc}$ and β_{gp} can be used for each run (5 by default). The number of spline bases is set to 20 by default, but the model appears to be fairly insensitive to this setting in the tested range of 5 to 40.

4.2.2 Simulation

4.2.2.1 Model description and rationale

Initially, we attempted to simulate multi-group data from the `splatter` model. We employed 10x Chromium data for peripheral blood mononuclear cells (PBMC) [110] with cell type labels derived from [191] to determine realistic parameter estimates for the differential expression component of the model (see below). In order to do so, group-specific log fold-change (logFC) values were drawn from a mixture distribution of a central, narrow Gaussian-Laplace mixture (representing non-differentially expressed genes) and two flanking, absolute value-transformed Gaussians (representing downregulated/upregulated genes). This mixture distribution was fitted to logFC values derived from differential expression analysis (see below).
However, inspection of posterior predictive samples for multiple fits, using labeled single cell RNA-seq data from [110] and FACS-purified data from Koh et al. [187] (**Figure 4.1**A,B, **Figure 4.2**A,B), revealed that this model systematically underestimates extreme logFC values (**Figure 4.1**C, **Figure 4.2**C). Thus, to accommodate the heavier tails present in observed data, we augmented the `splatter` model by replacing the flanking absolute value-transformed Gaussian components with bounded Student's t distributions. Posterior predictive logFC distributions from this modified model better fit the observed data (**Figure 4.1**D, **Figure 4.2**D). Consequently, we used this model to perform simulation analysis.

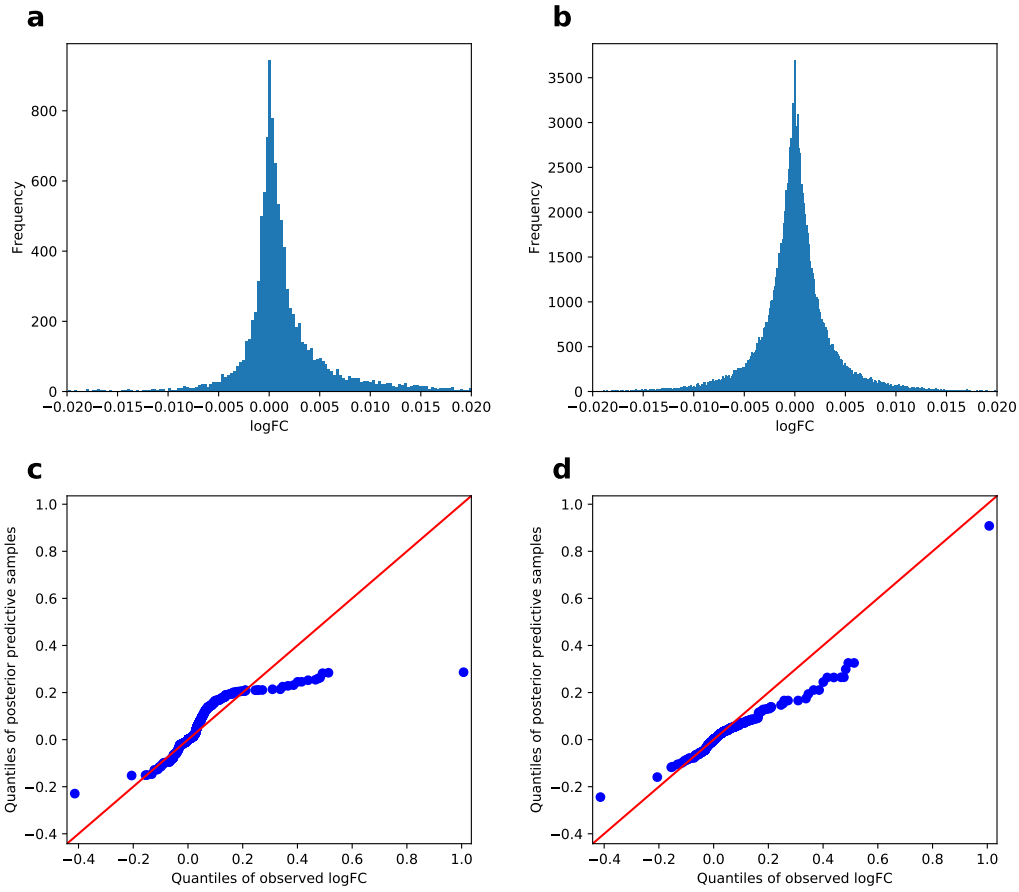


Figure 4.1: Fitting single cell RNA-seq simulation models to the Zheng PBMC 68k dataset, using cell type annotations provided in [191]. (a) Log fold change values computed from differential expression analysis between naive CD8+ and naive CD4+ T cells. (b) 'Null' log fold change values computed by randomly splitting naive CD8+ T cells into equally sized halves 10 times. (c) Quantile-quantile (QQ) plot comparing observed log fold change values between naive CD8+ and naive CD4+ T cells and posterior predictive samples from the `splatter` model (**Methods**). (d) Quantile-quantile (QQ) plot comparing observed log fold change values between naive CD8+ and naive CD4+ T cells and posterior predictive samples from the modified model (**Methods**).

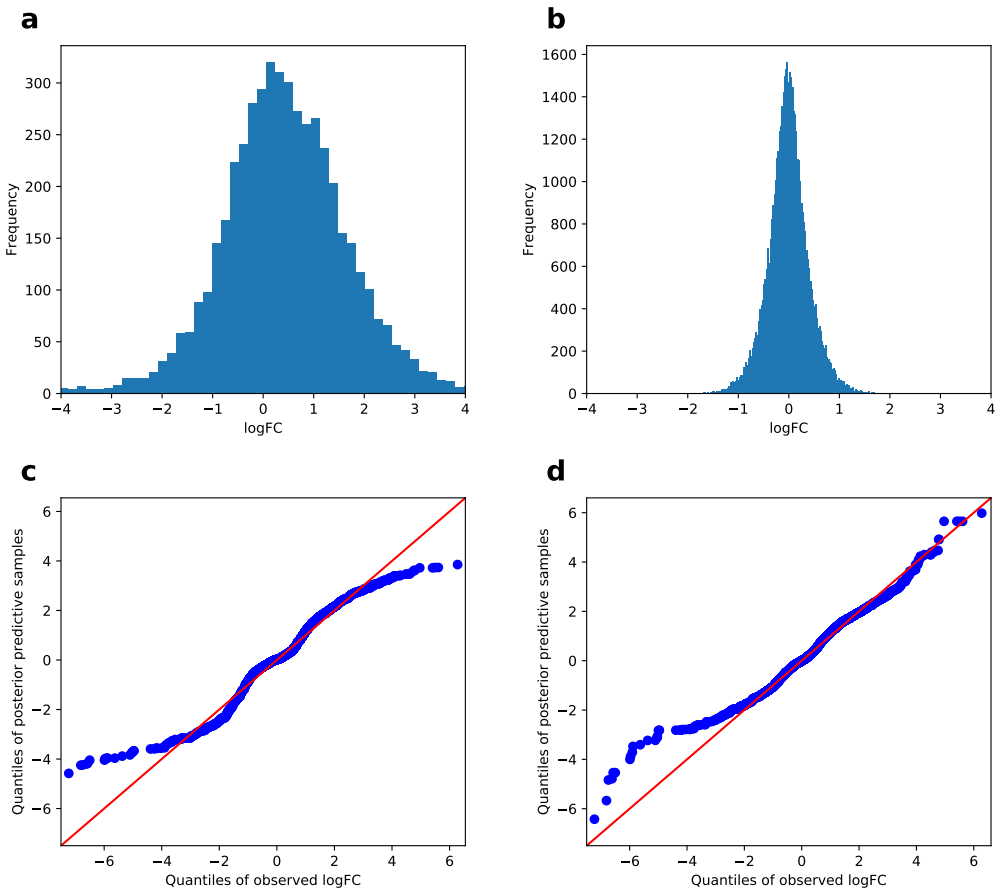


Figure 4.2: Fitting single cell RNA-seq simulation models to the Koh et al. [187] dataset of FACS-purified cell types. (a) Log fold change values computed from differential expression analysis between human embryonic stem cells (hESCs) and day 3 somite cells (ESMT). (b) 'Null' log fold change values computed by randomly splitting naive anterior primitive streak cells into equally sized halves 10 times. (c) Quantile-quantile (QQ) plot comparing observed log fold change values between hESC and ESMT cells and posterior predictive samples from the `splatter` model (**Methods**). (d) Quantile-quantile (QQ) plot comparing observed log fold change values between hESC and ESMT cells and posterior predictive samples from the modified model (**Methods**).

4.2.2.2 Model fitting

The models described above were fit to logFC values derived from real data. Using the labeled 10x Chromium data for 68k PBMCs [110], differential expression was performed with the `findMarkers` function from the R package `scran` [192]. To generate corresponding null

distributions of logFC values for non-differentially expressed genes, we split data for each cell type into equally sized halves 10 times, running `findMarkers` to compare the resulting halves. A central Gaussian-Laplace mixture ($\mu = 0$) was first fit to the null logFC values. The distribution of posterior predictive logFC values appeared to be consistent with observed logFC values for this null component (**Figure 4.1**D). Following this, the entire mixture distribution was fitted to logFC values for pairs of distinct cell types, using *maximum a posteriori* (MAP) estimates of parameters for the central Gaussian-Laplace component. Posterior distributions of model parameters were inferred using the no U-turn sampler (NUTS) in `pymc3`, using 4 independent chains, 1000 tuning iterations, and 2500 additional iterations per chain. Trace plots and the Gelman-Rubin diagnostic were used to assess convergence.

4.2.2.3 Simulating multi-group data

Expression count matrices were simulated using a modified version of the `splatter` package. Log fold change values were simulated according to our model instead of the `splatter` model. Other settings were kept identical. We used MAP estimates of μ_+, μ_-, σ_+, σ_-, ν_+, and ν_-, determined by fitting our simulation model to (1) logFC values between naive CD4+ and naive CD8+ T cells (**Figure 4.1**A); and (2) logFC values between B cells and CD8+ T cells (Section **4.2.2.1**) for the differential expression component. The proportion of downregulated genes out of differentially expressed genes was set to 0.5 (i.e. equally probable for a differentially expressed gene to be downregulated vs. upregulated). Three "groups" (cell types) were simulated at equal proportions. Other parameters for splatter were fitted from 10x Chromium data for 4,000 T cells available from 10x Genomics.

To assess the performance of CellAssign relative to other clustering methods across a range of p_d values (proportion of genes differentially expressed between each pair of cell types), p_d was chosen from $\{0.05, 0.15, 0.25, 0.35, 0.45, 0.55\}$. (The true MAP estimate of p_d was 0.0746 for naive CD4+ vs. naive CD8+ T cells, and 0.153 for B vs. CD8+ T cells.) The number of simulated cells, n, was set to 2000, and 1000 were randomly set aside for training (for scmap and correlation-based supervised clustering).

To assess the robustness of CellAssign to misspecification of the marker gene matrix ρ, p_d was set to 0.25 and the number of simulated cells n to 1500.

Simulations were run 9 times with unique random seeds for each combination of parameter settings.

4.2.2.4 Clustering multi-group data

Count matrices were normalized with `scater normalize` and the top 50 principal components were computed from the top 1000 most variable genes. For `phenograph`, `Seurat` (resolution $\in \{0.4, 0.8, 1.2\}$), k-means, densitycut, and dynamicTreeCut, unsupervised clustering was performed on the values of these top 50 PCs. For `SC3`, the entire normalized SingleCellExperiment object was passed as input instead. For supervised methods (`scmap-cluster` [185] and correlation-based [110]), expression data for both training and evaluation sets was provided. For CellAssign, the raw count matrix was provided as input, along with a set of marker genes selected based on simulated log fold change and mean expression values. Specifically, a gene was defined as a marker gene if it was in the top 5th percentile of differentially expressed genes according to logFC and the top 10th percentile of differentially expressed genes according to mean expression. In simulations of robustness to marker gene misspecification, a proportion of randomly selected entries in the marker gene matrix ρ were flipped from 0 to 1 (or vice versa).

4.2.2.5 Mapping clusters to true groups

For assignments derived from unsupervised clustering, clusters were mapped to simulated groups by first performing differential expression between each cluster and the remaining cells. Following this, we computed the Spearman correlation between these logFC values and the simulated (true) logFC values for each simulated group. Each inferred cluster was mapped to most highly correlated simulated group based on Spearman's ρ where $\rho > 0$ and $P \leq 0.05$. Clusters that could not be mapped based on these criteria were marked as 'unassigned'.

4.2.2.6 Benchmarking

We generated synthetic datasets for benchmarking from the modified `splatter` model (Section **4.2.2.1**) with Student's t parameters $\mu = 0.1$, $\sigma = 0.1$, $\nu = 1$ and the proportion of differentially expressed genes per cell type set to 20%. Synthetic datasets of various sizes (number of cells $N \in \{1000, 2000, 4000, 8000, 10000, 20000, 40000, 80000\}$ and number of cell types $C \in \{2, 4, 6, 8\}$ with a balanced number of cells per type were generated. Markers for CellAssign were selected from genes in the top 20th percentile in terms of log fold change among differentially upregulated genes and the top 10th percentile in terms of expression. CellAssign was run with 2, 4, 6, and 8 markers per cell type, with a maximum minibatch size of 5000 cells. Five separate CellAssign runs were timed for each combination of parameters.

4.2.3 Koh *et al.* dataset

This section refers to the scRNA-seq dataset from [187].

4.2.3.1 Preprocessing and normalization of single cell RNA-seq data

Preprocessed data was obtained from the R package `DuoClustering2018` [181, 187]. Celltypes with both single cell RNA-seq data and bulk RNA-seq data were used: hESC (day 0 human embryonic stem cell), APS (day 1 anterior primitive streak), MPS (day 1 mid primitive streak), DLL1pPXM (day 2 DLL1+ paraxial mesoderm), ESMT (day 3 somite), Sclrtm (day 6 sclerotome), D5CntrlDrmmtm (day 5 dermomyotome), D2LtM (day 2 lateral mesoderm). Normalization and dimensionality reduction was performed with `scater` `normalize`, `runPCA`, `runTSNE`, and `runUMAP`. The top 500 most variable genes were used to compute the top 50 principal components, and the top 50 PCs were used as input for t-SNE and UMAP.

4.2.3.2 Identification of marker genes from bulk RNA-seq data

Differential expression analysis results for bulk RNA-seq data for the same cell types was used to compute the relative expression of each gene in each cell type. Briefly, bulk RNA-seq log fold change values obtained from [187] were used to compute log-scale relative gene expression levels. Next, we identified gene-specific thresholds for defining the cell types in which each gene is a marker. For each gene, relative expression levels across cell types were sorted in ascending order, denoted as $E_1, ..., E_C$, where C is the total number of cell types. The maximum difference between sorted expression levels, $\max_{1 \leq i < C}(E_{i+1} - E_i)$, was then computed. Denote the index i for gene g in which this difference is maximal i_g. For gene g, cell types in which relative expression values were equal to or greater than E_{i_g+1} were considered cell types with gene g as a marker. Genes with a maximum difference value in the the top 20th percentile were used as marker genes.

4.2.3.3 CellAssign

CellAssign was run on count data using the marker gene matrix defined from bulk RNA-seq data described above. Three random initializations of expectation-maximization were used with shrinkage priors on δ_{gc} turned on (Section **4.2.1.1**). Results from the run that reached the highest marginal log-likelihood at convergence were kept.

4.2.3.4 Unsupervised clustering

Unsupervised clustering was performed on the top 50 PCs with phenograph [193] and Seurat [178] (resolution $\in \{0.4, 0.8, 1.2\}$ and on the `SingleCellExperiment` object of raw and normalized counts with SC3 [177]. Inferred clusters were mapped to true (FACS-purified) cell types by computing the pairwise Spearman correlation between mean expression vectors for each cluster and each true cell type. Each cluster was treated as the cell type it was most strongly positively associated with by Spearman's ρ.

4.2.4 High-grade serous ovarian cancer

4.2.4.1 Sample preparation, library preparation, and sequencing

Sample preparation, library preparation, and sequencing steps are described in Chapter **2** (see Section **2.2.8**). Cell dissociation was carried out at 6°C to maximize lymphocyte yield (O'Flanagan et al., unpublished). The 10x Chromium 5' gene expression kit was used for single cell RNA-seq library preparation.

4.2.4.2 Preprocessing and normalization of single cell RNA-seq data

Raw sequence files were processed with CellRanger v2.1.0. The resulting filtered count matrices were read into `SingleCellExperiment` objects. Outlier cells according to quality control parameters (≥ 3 median absolute deviations from the median) were filtered out using the `scater` R package. Additionally, cells with $\geq 20\%$ mitochondrial UMIs or $\geq 50\%$ ribosomal UMIs were removed. Size factors were computed using `quickCluster` and `computeSumFactors` from the `scran` R package. Following this, data normalization was performed using `scater normalize`. Principal components analysis was performed on the resultant normalized logcounts for the top 1000 most variable genes. The first 50 PCs were used as input for t-SNE and UMAP.

4.2.4.3 CellAssign

The following marker gene list was used for CellAssign [194–197]:

- B cells: *VIM*^{c}, *MS4A1*^{c}, *CD79A*^{c}, *PTPRC*^{c}, *CD19*^{c}, *BANK1* [194]
- CD4 T cells: *VIM*^{c}, *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *CD28*^{c}, *PTPRC*^{c}, *CD4*^{c}
- Cytotoxic T cells: *VIM*^{c}, *CD2*^{c}, *CD3D*^{c}, *CD3E*^{c}, *CD3G*^{c}, *CD28*^{c}, *PTPRC*, *CD8A*^{c}, *CD8B*^{c}, *PRF1*^{c}, *GNLY*^{c}, *NKG7*^{c}, *KLRC1*^{c}
- Monocyte/Macrophage: *VIM*^{c}, *CD14*^{c}, *FCGR3A*^{c}, *CD33*^{c}, *ITGAX*^{c}, *ITGAM*^{c}, *CD4*^{c}, *PTPRC*^{c}, *LYZ*^{c}

- Epithelial/cancer cell: *EPCAM*[c], *MUC1*[c], *CDH1*[c], *MYC*[c]
- Stromal cell: *VIM*[c], *ECEL1*[198], *KLHDC8A*[198], *MUM1L1*[198], *ARX*[198], *ACTA2*[c]
- Endothelial cells: *VIM*[c], *EMCN*[c], *CLEC14A* [199], *CDH5*[c], *PECAM1*[c]

[c]: canonical marker

CellAssign was run with default parameters, the shrinkage prior on δ_{gc} values turned on, and 5 random initializations. Patient was added as an additional covariate into the design matrix X (Section **4.2.1.1**). The best result according to marginal log-likelihood at convergence was kept. Optimization was considered converged after 3 consecutive rounds of no improvement (relative change in log-likelihood $< 10^{-5}$). MAP assignments from CellAssign were used for downstream analysis.

4.2.4.4 Unsupervised clustering

Cells with a total probability of at least 0.99 for the stromal cell type were subsetted. The top 50 PCs from preprocessing were provided as input to `densitycut`, which was run with default parameters.

4.2.5 Follicular lymphoma

4.2.5.1 Sample preparation

Leftovers from clinical flowed samples were collected and frozen in fetal calf serum containing 10% DMSO. Cells were thawed and washed according to the steps outlined in the 10X Genomics Sample Preparation Protocol. Cells were stained with PI for viability and sorted in the BD FACSAria Fusion using a 85um nozzle. Sorted cells were collected in 0.5 ml of medium, centrifuged and diluted in 1X PBS with 0.04% bovine serum albumin.

4.2.5.2 Library preparation and sequencing

Cell concentration was determined by using a Countess II Automated Cell Counter and approximately 3,500 cells were loaded per well in the Single Cell 3' Chip. Single cell libraries were prepared according to the Chromium Single Cell 3'Reagent Kits V2 User Guide. Single cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane.

4.2.5.3 Preprocessing and normalization of single cell RNA-seq data

Raw sequence files were processed with CellRanger v2.1.0. The resulting filtered count matrices were read into `SingleCellExperiment` objects. Outlier cells according to quality control

parameters (≥ 3 median absolute deviations from the median) were filtered out using the `scater` R package. Additionally, cells with $\geq 10\%$ mitochondrial UMIs or $\geq 60\%$ ribosomal UMIs were removed. Size factors were computed using `quickCluster` and `computeSumFactors` from the `scran` R package. Following this, data normalization was performed using `scater normalize`. Principal components analysis was performed on the resultant normalized logcounts for the top 1000 most variable genes. The first 50 PCs were used as input for t-SNE and UMAP.
Cell cycle scores were computed with `cyclone` from the `scran` package [192, 200].

4.2.5.4 scvis analysis

scvis train (v0.1.0) [201] was run with default settings on the top 50 PCs to produce a 2-dimensional embedding of the follicular lymphoma data. Early stopping was added to scvis, so that the model would terminate after 3 successive iterations of no improvement (relative improvement in ELBO $< 10^{-5}$). The resultant model was saved and used for mapping in Section **4.2.6.4**.

4.2.5.5 CellAssign

The following marker gene list was used for CellAssign:

- B cells: *CD19*[c], *MS4A1*[c], *CD79A*[c], *CD79B*[c], *CD74*[c], *CXCR5* [202]
- Cytotoxic T cells: *CD2*[c], *CD3D*[c], *CD3E*[c], *CD3G*[c], *TRAC*[c], *CD8A*[c], *CD8B*[c], *GZMA*[c], *NKG7*[c], *CCL5*[c], *EOMES*[c]
- Follicular helper T cells: *CD2*[c], *CD3D*[c], *CD3E*[c], *CD3G*[c], *TRAC*[c], *CD4*[c], *CXCR5*[c], *PDCD1*[c], *TNFRSF4* [194], *ST8SIA1* [194], *ICA1* [194], *ICOS* [194]
- Other CD4+ T cells: *CD2*[c], *CD3D*[c], *CD3E*[c], *CD3G*[c], *TRAC*[c], *CD4*[c], *IL7R* [194]

[c]: canonical marker
CellAssign was run with default parameters, the shrinkage prior on δ_{gc} values turned on, and 5 random initializations. Patient was added as an additional covariate into the design matrix X (Section **4.2.1.1**). The best result according to marginal log-likelihood at convergence was kept. Optimization was considered converged after 3 consecutive rounds of no improvement (relative change in log-likelihood $< 10^{-5}$). MAP assignments from CellAssign were used for downstream analysis.

4.2.5.6 Classifying B cells

B cells from CellAssign were further subclassified into 'malignant' or 'nonmalignant' groups according to expression of the constant region of the immunoglobulin light chain (kappa or

lambda type) and the results of PCA. Seurat [178] (resolution = 0.8) was used to separate B cells into clusters, based on the top 50 PCs. Following this, the sole cluster associated with *IGKC* (immunoglobulin light chain kappa-type constant region) expression was designated as nonmalignant. We further reasoned this was the case based on the cluster containing a mixture of T1 and T2 cells and constituting only a minor subset of the B cells.

4.2.5.7 Differential expression between timepoints

Differential expression analysis between timepoints for a given celltype and patient was performed using voom from the limma package for each patient and cell type separately, with timepoint as the independent variable. Genes with low expression (< 500 UMIs in total across all cells) were removed. P-values were adjusted with the Benjamini-Hochberg method, and genes with $Q \leq 0.05$ were considered differentially expressed. Differential expression between malignant and nonmalignant B cells was performed similarly, but using the formula `~malignant_status + timepoint + malignant_status:timepoint` to control for timepoint and any interactions.

4.2.5.8 Reactome pathway enrichment analysis

Pathway analysis was performed for the top 50 most upregulated and top 50 most downregulated genes (separately) by log fold change from limma (where $Q \leq 0.05$, filtering out ribosomal and mitochondrial genes). Overrepresentation of Reactome [203] pathways was assessed using the R package ReactomePA. Pathways were considered significantly overrepresented if the adjusted P-value ≤ 0.05 and at least 2 genes from the pathway were present.

4.2.5.9 Comparing malignant cells between timepoints

Log fold change values from the `findMarkers` function (filtering out ribosomal and mitochondrial genes) from `scran` were used as input for gene set enrichment analysis with the fGSEA R package, using default parameters with 10,000 permutations, and the hallmark pathway gene set [204]. Annotations for cell cycle-associated pathways (E2F targets, G2M checkpoint, and mitotic spindle) were taken from [204]. BH-adjusted P-values for differences in proliferation marker expression (*MKI67* and *TOP2A*) were also computed with the `findMarkers` function from `scran`, using default parameters.

4.2.5.10 Somatic variant calling

Somatic single-nucleotide variants (SNVs), indels, and breakpoints for both cases were obtained from [205]. Annotations from the Nanostring PanCancer Immune Profiling panel were used to

identify antigen processing and presentation genes [149].

4.2.5.11 HLA loss-of-heterozygosity analysis

HLA class I typing was performed using matched normal bams [205] with OptiType [163]. Following this, HLA class I loss-of-heterozygosity (LOH) was called from tumor and matched normal bams as well as OptiType 4-digit HLA types using LOHHLA [52]. HLA LOH was called for an allele if the estimated copy number (with binning and B-allele frequency settings) was < 0.5 and the significance of allelic imbalance $p < 0.05$ (paired t test, no duplicate counts).

4.2.6 Reactive lymph node data

4.2.6.1 Sample preparation

Cell suspensions from patients with reactive lymphoid hyperplasia but no evidence of malignant disease and collagen disease were used. Leftovers from clinical flowed samples were collected and frozen in FCS containing 10%DMSO. The day of the experiment cell suspensions were rapidly thawed at 37°C, and washed according to the steps outlined in the 10X Genomics Sample Preparation Protocol. Cells were stained with DAPI and viable cells (DAPI negative) were sorted on a FACS ARIAIII or FACS Fusion (BD Biosciences) instrument.

4.2.6.2 Library preparation and sequencing

Approximately 8,700 cells per sample were loaded into a Chromium Single Cell 3' Chip kit v2 (PN-120236) and processed according to the Chromium Single Cell 3'Reagent kit v2 User Guide. Libraries were constructed using the Single 3' Library and Gel Bead Kit v2 (PN-120237) and Chromium i7 Mulitiplex Kit v2 (PN-120236). Single cell libraries from two samples were pooled and sequenced on one HiSeq 2500 125 base PET lane.

4.2.6.3 Preprocessing and normalization of single cell RNA-seq data

Preprocessing steps for the reactive lymph node data were identical to those for single cell RNA-seq data, described in Section **4.2.5.3**.

4.2.6.4 scvis analysis

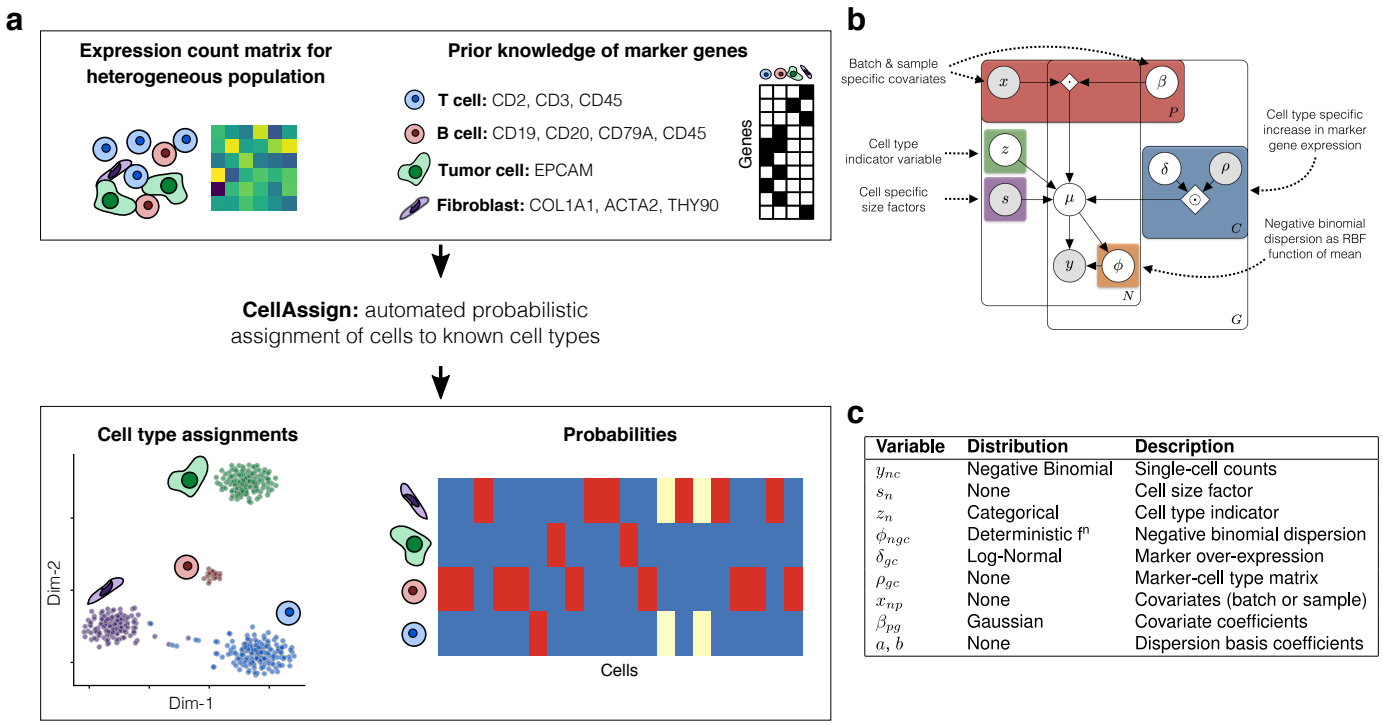
The identities of the top 1000 most variable genes and PCA loadings from follicular lymphoma data analysis were used to compute a 50-dimensional embedding for the reactive lymph node

data. Following this, the resultant 50 PCs were provided as input to scvis map [201], using the model trained in Section **4.2.5.4** and default settings.

4.3 Results

4.3.1 Automated assignment of cell types with CellAssign

The CellAssign statistical framework (**Figure 4.3**) models observed gene expression as a composite of cell type-specific, library size, and batch effects, using raw single cell RNA-seq counts from a heterogeneous cellular population as input. To enable automated cell type classification, marker gene information is provided *a priori* to CellAssign in the form of a set of marker genes for each modeled cell type. The sole assumption for a marker gene to be indicative of a cell type is that it is should be over-expressed in that cell type relative to all others - it may still be expressed in all cells and variable between others. Information on other experimental and biological covariates - such as batch and patient-of-origin - can also be encoded as a standard design matrix. Using this information, CellAssign employs a hierarchical Bayesian statistical framework to determine the probability that each cell belongs to each of the modeled cell types, along with estimates of the model parameters including the relative expression of marker genes in each cell type and the systematic effects of other covariates on marker gene expression patterns. To prevent misclassification when unknown cell types are present, CellAssign can assign cells that do not belong to any of the provided cell types to an 'unassigned' group.



Variable	Distribution	Description
y_{nc}	Negative Binomial	Single-cell counts
s_n	None	Cell size factor
z_n	Categorical	Cell type indicator
ϕ_{ngc}	Deterministic f^n	Negative binomial dispersion
δ_{gc}	Log-Normal	Marker over-expression
ρ_{gc}	None	Marker-cell type matrix
x_{np}	None	Covariates (batch or sample)
β_{pg}	Gaussian	Covariate coefficients
a, b	None	Dispersion basis coefficients

Figure 4.3: (**a**) Overview of CellAssign. CellAssign takes raw count data from a heterogeneous single-cell RNA-seq population, along with a set of known markers for various cell types understudy. Using CellAssign for inference, each cell is probabilistically assigned to a given cell type without any need for manual annotation or intervention, accounting for any batch or sample-specific effects. (**b**) An overview of the CellAssign probabilistic graphical model. (**c**) The random variables and data that form the model, along with the distributional assumptions and description.

4.3.2 Performance of CellAssign relative to state-of-the-art unsupervised and supervised classification methods

While CellAssign is the only method to-date that can automatically assign cells to cell types based on prior marker gene associations, we sought to demonstrate its performance was competitive compared to standard workflows including unsupervised clustering followed by manual curation and methods that map cells to existing data from purified populations. We employed an adapted version of the `splatter` model to simulate single cell RNA-seq data for multiple cell populations (**Methods**). On simulated data for 80000 cells from 2 cell types, CellAssign completed in under 2 minutes, and appeared to scale at worst linearly in the number of cell types and marker genes used per cell type (**Figure 4.4**). In order to select realistic parameter settings for simulation, we fitted the `splatter` model to data for naïve CD8+ and CD4+ T cells from peripheral blood mononuclear cell (PBMC) data. Simulations were conducted across a wide range of values for the fraction of differentially expressed genes (0.05 to 0.55), to represent cellular mixtures of similar and distinct cell types. Following this, we evaluated the performance of each unsupervised (Seurat [178], SC3 [177], phenograph [193], densitycut [206], dynamicTreeCut [207], k-means) and supervised (scmap-cluster [185], correlation-based [110]) clustering methods for single cell RNA-seq data (**Methods**). Half of the simulated cells (n=2000 total, n=1000 training, n=1000 evaluation) were set aside for training the supervised methods. Marker genes for CellAssign were selected based on simulated log-fold change values and mean expression (**Methods**), and *maximum a posteriori* (MAP) cell type probability estimates were treated as cell type assignments. For all values of the fraction of differentially expressed genes, CellAssign performed comparably or superior to alternative workflows in terms of accuracy and F1 score (**Figure 4.5**A, **Supplemental Table B.1**). As expected, supervised methods generally performed better than unsupervised methods (**Figure 4.5**A).

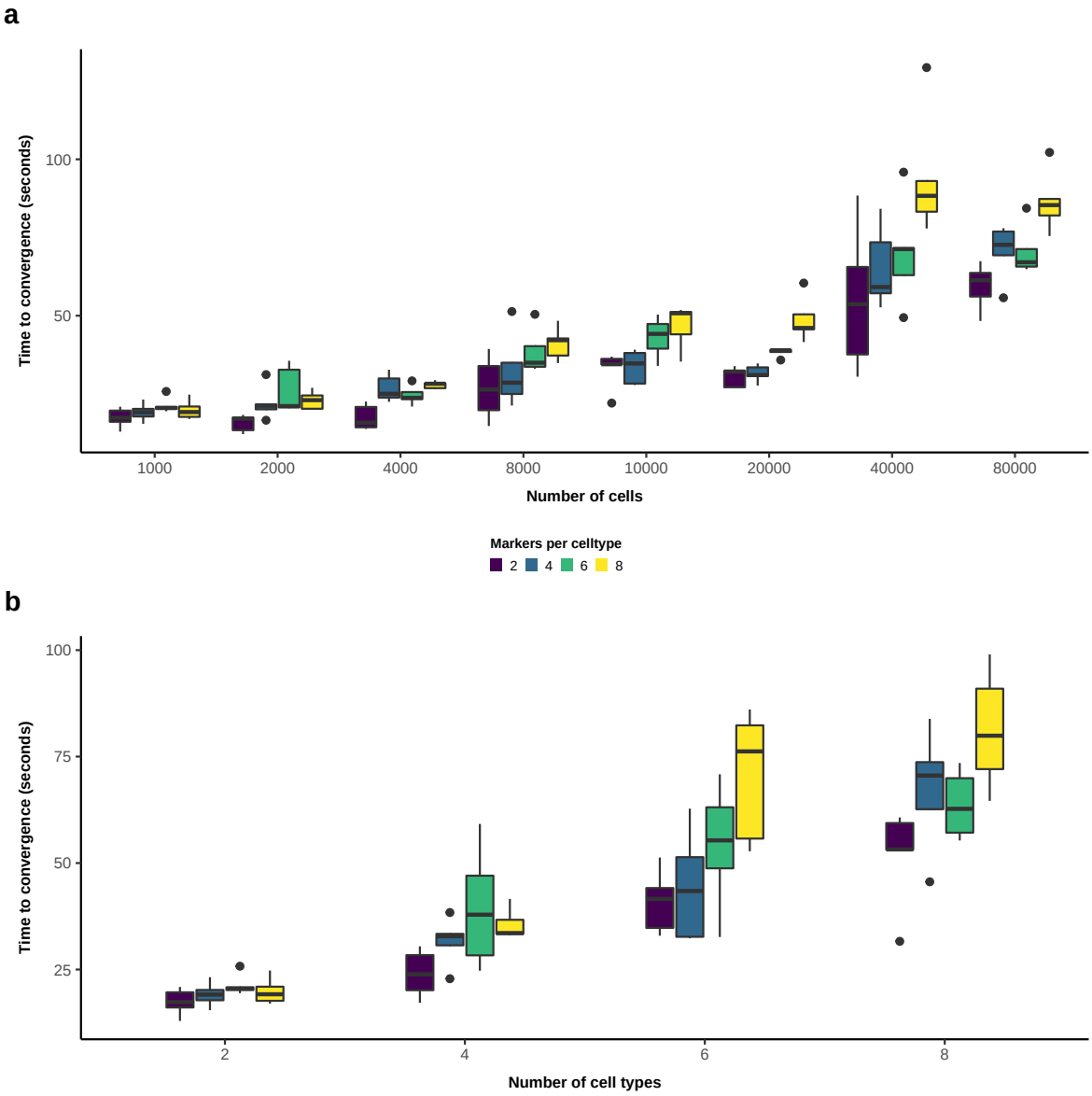


Figure 4.4: Benchmarking results for CellAssign across a range of simulated data set sizes (number of cells), number of cell types being inferred, and number of marker genes per cell type. (a) Runtime (to convergence, defined as a relative change in log-likelihood $< 10^{-3}$ between successive iterations, as a function of data set size and the number of marker genes used per cell type, on simulated data (**Methods**). Two cell types were used. (b) Runtime (to convergence, defined as a relative change in log-likelihood $< 10^{-3}$ between successive iterations, as a function of the number of cell types and the number of marker genes used per cell type, on simulated data. One thousand cells were used.

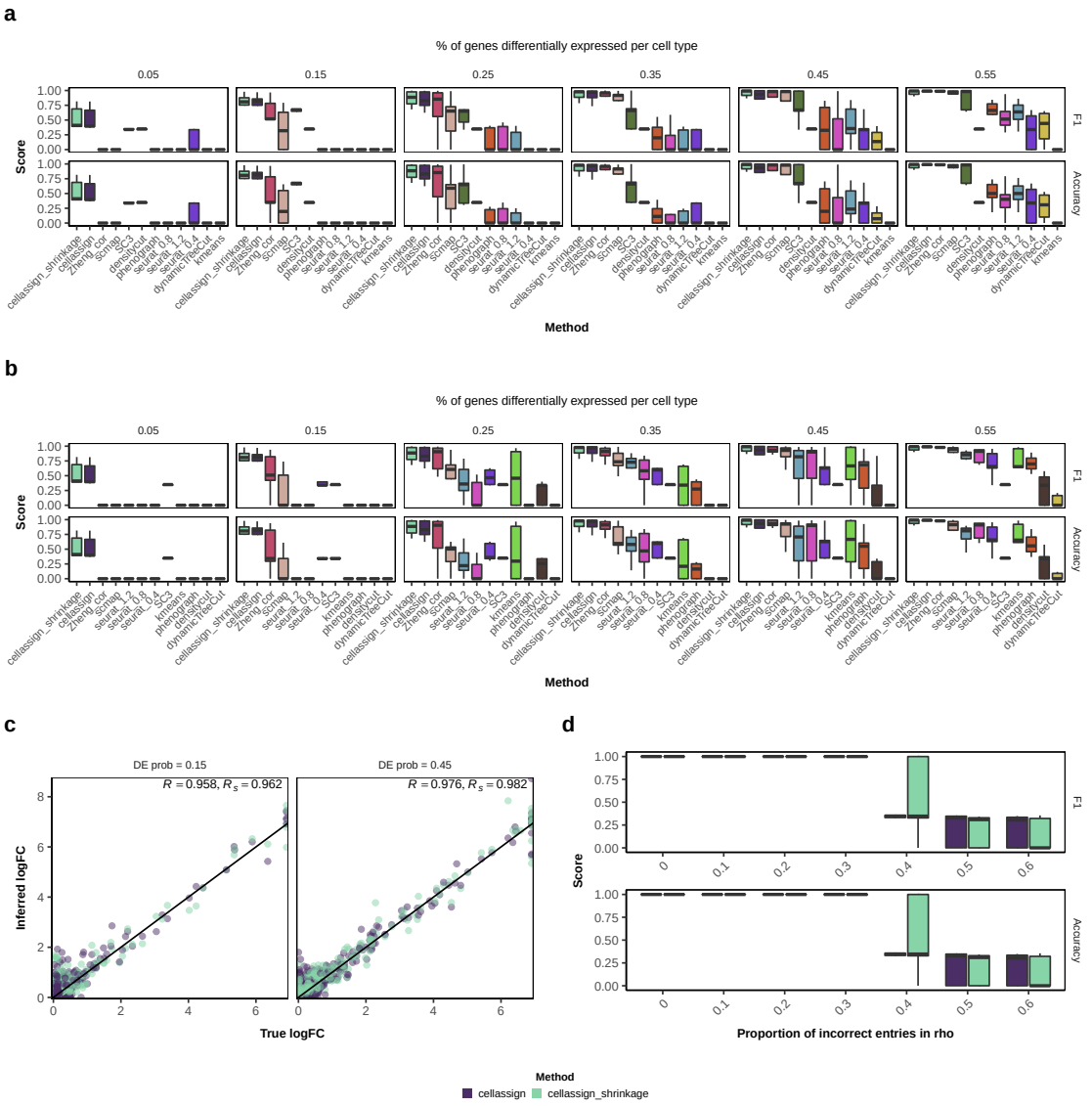


Figure 4.5: Performance of CellAssign on simulated data. (**a**) Accuracy and cell-level F1 score (**Methods**) for varying proportions of differentially expressed genes per cell type, with other differential expression parameters set to MAP estimates determined from comparing naive CD8+ and naive CD4+ T cells (**Methods**). CellAssign was provided with a set of marker genes (**Methods**); all other methods were provided with all genes. cellassign_shrinkage refers to a version of CellAssign with a shrinkage prior on δ (**Methods**). (**b**) Accuracy and cell-level F1 score for varying proportions of differentially expressed genes per cell type, with other differential expression parameters set to MAP estimates determined from comparing naive CD8+ and naive CD4+ T cells. All methods were provided with the same set of marker genes. (**c**) Correspondence between true simulated log fold change values and log fold change (δ) values inferred by CellAssign. R and R_s refer to the Pearson correlation between true and inferred logFC values for cellassign and cellassign_shrinkage, respectively. (**d**) Performance of CellAssign where a certain proportion of entries in the marker gene matrix are flipped at random. Differential expression parameters used for these simulations were based on those determined from comparing B and CD8+ T cells.

In case CellAssign's superior performance was due to being provided solely with informative marker genes compared to transcriptome-wide data provided to other methods, we repeated our simulations providing other methods with exactly the same data as CellAssign. Nonetheless, CellAssign performed superiorly to the other tested methods (**Figure 4.5**B). Similar results were obtained on data simulated from parameter estimates fitted to B cells and CD8+ T cells (**Figure 4.6**A,B, **Supplemental Table B.1**). Moreover, CellAssign accurately infers the relative expression of marker genes in each cell type (all $R > 0.958$; **Figure 4.5**C, **Figure 4.6**C).

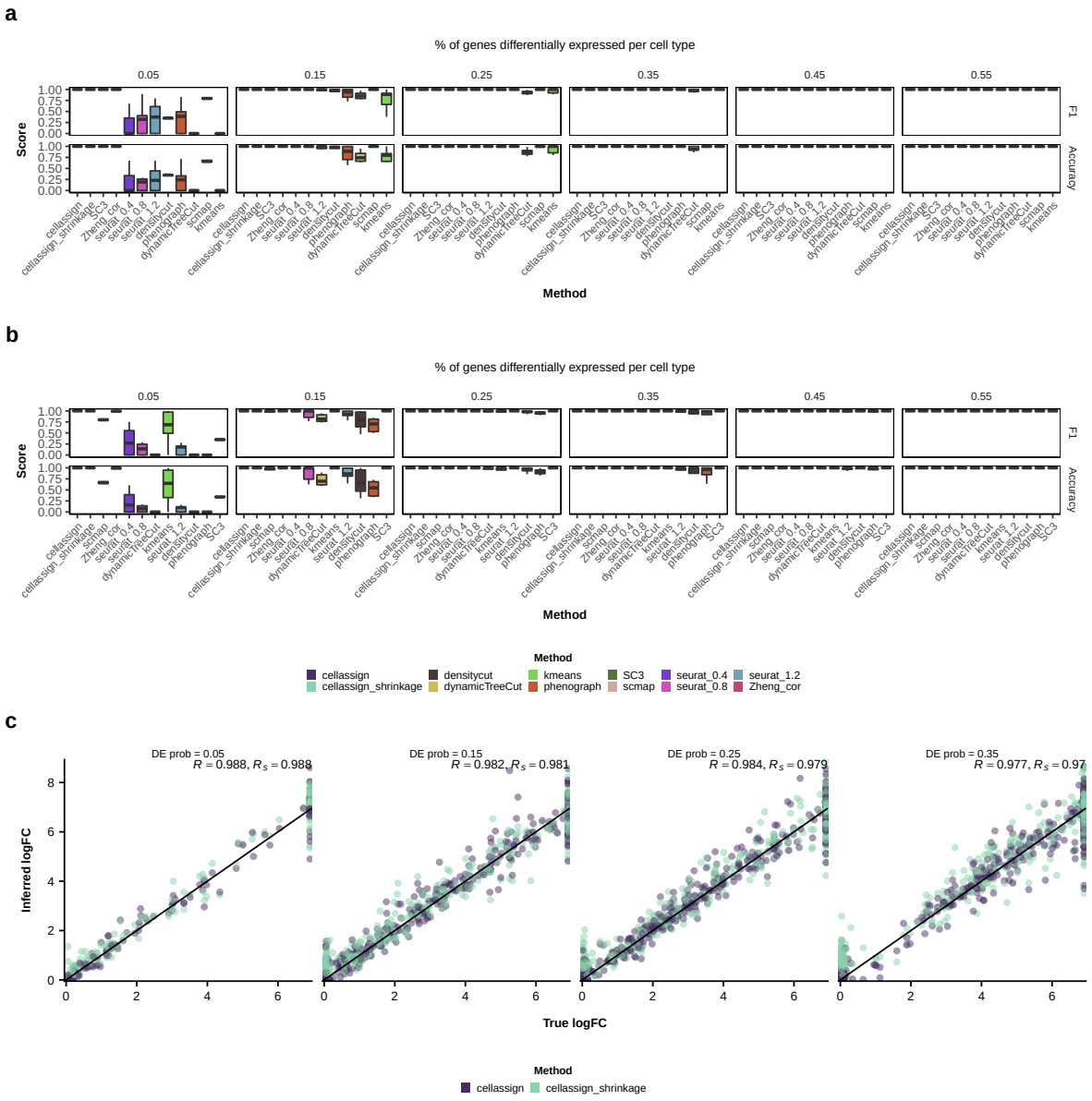


Figure 4.6: Simulation performance across a range of proportions of differentially expressed genes, using differential expression parameters derived from comparing B and CD8+ T cells. (a) Accuracy and cell-level F1 score (**Methods**) for varying proportions of differentially expressed genes per cell type. CellAssign was provided with a set of marker genes (**Methods**); all other methods were provided with all genes. (b) Accuracy and cell-level F1 score for varying proportions of differentially expressed genes per cell type. All methods were provided with the same set of marker genes. (c) Correspondence between true simulated log fold change values and log fold change (δ) values inferred by CellAssign. R and R_s refer to the Pearson correlation between true and inferred logFC values for cellassign and cellassign_shrinkage, respectively.

We next assessed the robustness of CellAssign to mis-specification of the association of marker genes to cell types. While CellAssign assumes information for user-provided marker genes is complete and correct, in reality this may not always be the case. For example, a shared marker gene may be incorrectly specified as a cell type-specific marker gene due to incomplete prior information or human error. Thus, we tested the robustness of CellAssign to marker gene mis-specification by changing a proportion of entries in the binary marker gene matrix from 0 to 1 or vice-versa at random. Supplied with data for 5 marker genes per cell type, CellAssign maintained comparable performance in scenarios where up to 30% of matrix entries were mis-specified (**Figure 4.5**D, **Supplemental Table B.1**). This robustness to marker mis-specification is maintained even when the simulated cells belong to transcriptionally similar cell types containing few highly differentially expressed genes. For example, when cells were simulated based on the degree of dissimilarity between naïve CD4+ and naïve CD8+ T cells, the accuracy of CellAssign predictions was comparably high in scenarios where 20% of marker gene matrix entries were mis-specified (**Figure 4.7**, **Supplemental Table B.1**).

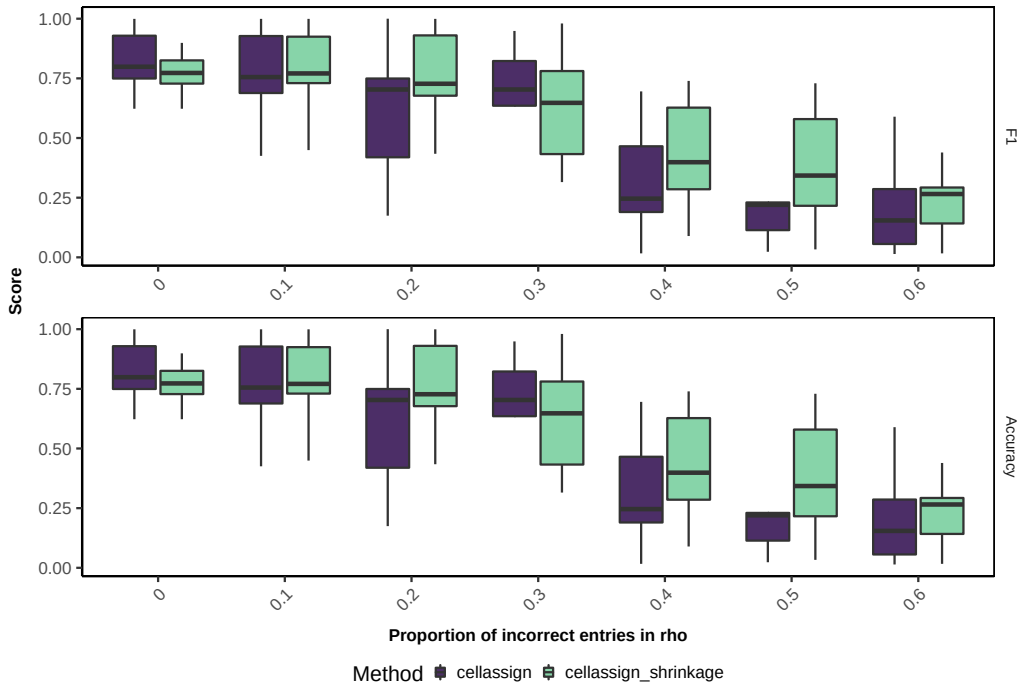


Figure 4.7: Simulation performance across a range of proportions of randomly flipped entries in the binary marker gene matrix, using differential expression parameters derived from comparing naive CD8+ and naive CD4+ T cells.

We also assessed the performance of CellAssign on real single cell RNA-seq data. We first applied CellAssign to data for FACS-purified H7 human embryonic stem cells in various stages of differentiation [187]. Using bulk RNA-seq data from the same cell types, we defined a set of 84 marker genes for CellAssign based on differential expression results (**Supplemental Table B.2**; **Methods**). CellAssign performed superiorly to the most competitive unsupervised methods from systematic analysis (SC3, Seurat) [181] in terms of accuracy and cell type-level F1 score (**Figure 4.8**A-D,F; **Methods**). Similar results were obtained for comparisons using only expression data for the marker genes (**Figure 4.8**E,G). Crucially, CellAssign was able to distinguish anterior primitive streak (APS) and mid primitive streak (MPS) cells, while no other method could reliably do so (**Figure 4.8**).

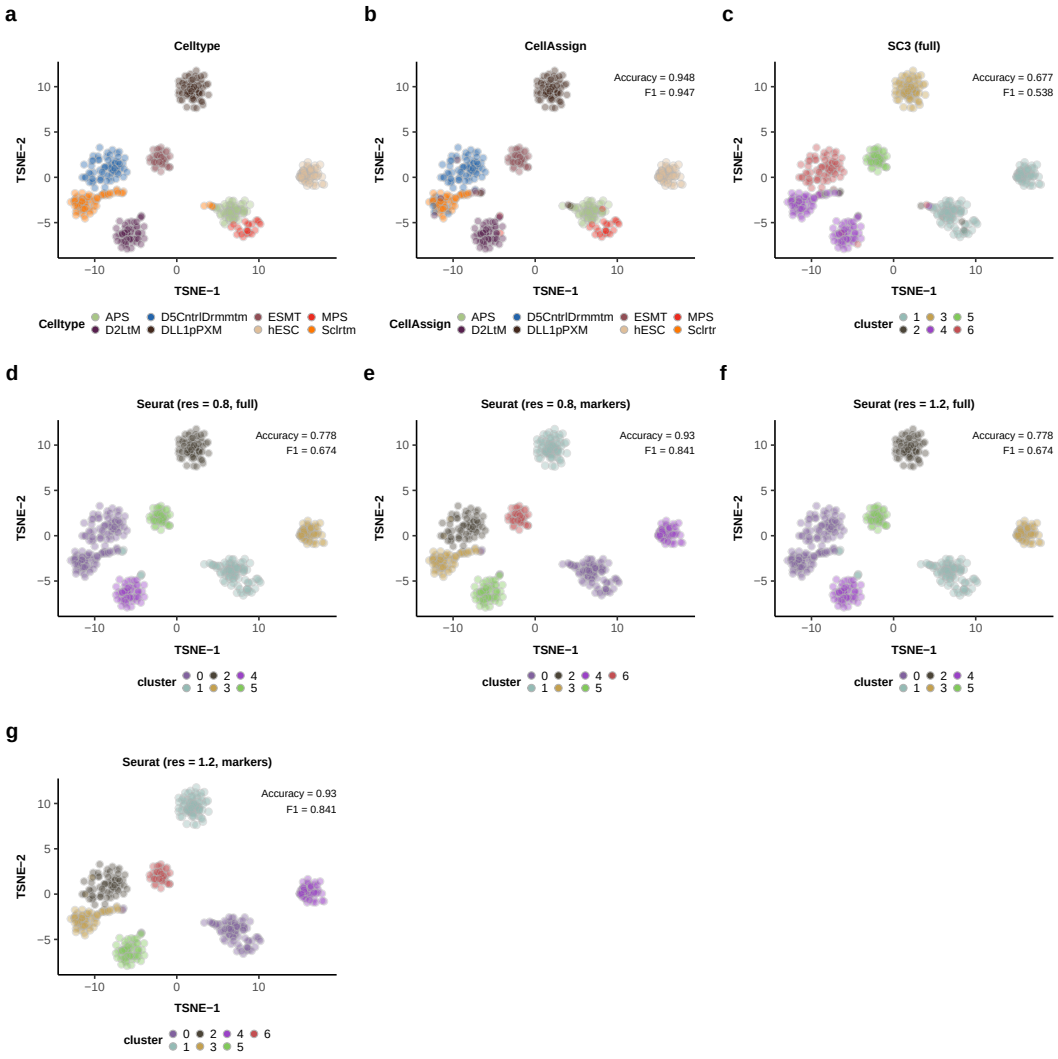


Figure 4.8: Performance (accuracy and cell type-level F1 score, **Methods**) of CellAssign and the best-performing clustering methods evaluated by [181] on FACS-purified H7 human embryonic stem cells in various stages of differentiation. t-SNE plots of (a) ground-truth FACS annotations; (b) CellAssign-derived annotations; (c) SC3 clusters (using all genes); (d) Seurat clusters (resolution = 0.8, using all genes); (e) Seurat clusters (resolution = 0.8, using the same marker gene set used by CellAssign); (f) Seurat clusters (resolution = 1.2, using all genes); (g) Seurat clusters (resolution = 1.2, using the same marker gene set used by CellAssign).

4.3.3 Profiling the malignant and nonmalignant composition of high-grade serous ovarian cancer

To profile the microenvironment of HGSC, we sequenced the transcriptomes of 6298 cells from 4 spatially collected pre-treatment biopsies of 3 patients with HGSC (**Table 4.1**; sample identifiers correspond to those from Chapter **2**). Following data preprocessing, we used CellAssign to identify 7 known cell types including epithelial cells, endothelial cells, other stromal cells, CD4 T cells, cytotoxic T cells, B cells, and monocytes, and visualized the normalized data using principal components analysis (PCA) and t-distributed stochastic neighbour embedding (t-SNE) (**Figure 4.9**A,B, **Figure 4.10**A,B, **Methods**). These cell type assignments appeared to be consistent with the expression patterns of well-known marker genes [104, 195–197] (**Figure 4.9**C). Overall, epithelial cells were the dominant cell type (52.8% of assigned cells), composing 20% to 86% of assigned cells across samples (**Figure 4.10**A). Multiple clusters of epithelial cells were observed in both samples from patient 73, suggestive of clonally distinct malignant cell subpopulations or a mixture of malignant and nonmalignant epithelial populations (**Figure 4.9**A,B). While most epithelial cell clusters in t-SNE space were largely patient-specific, most nonmalignant clusters, such as endothelial cells, stromal cells, and monocytes, contained cells from multiple patients (**Figure 4.9**A,B). Lymphocytes were rare, composing only 4.3% of all cells (B cells: 1.4%, CD4 T cells: 2.2%, cytotoxic T cells: 0.7%). Most B cells appeared to express *CD79A* and *SDC1* (*CD138*) but not *MS4A1* (*CD20*), consistent with plasma cells [208] (**Figure 4.9**C).

Patient	Sample	ID	Site	Raw	Filtered	Temperature
70	70LAdnx6	VOA11267	Left adnexal mass	506	280	6
72	72Om	VOA11558SA	Omentum	559	282	6
73	73LOv	VOA11543SA	Left ovary	2818	2707	6
73	73ROv	VOA11543SB	Right ovary	2415	2125	6

Table 4.1: HGSC samples profiled by single cell RNA-seq. Raw and filtered correspond to raw and preprocessed cell counts, respectively.

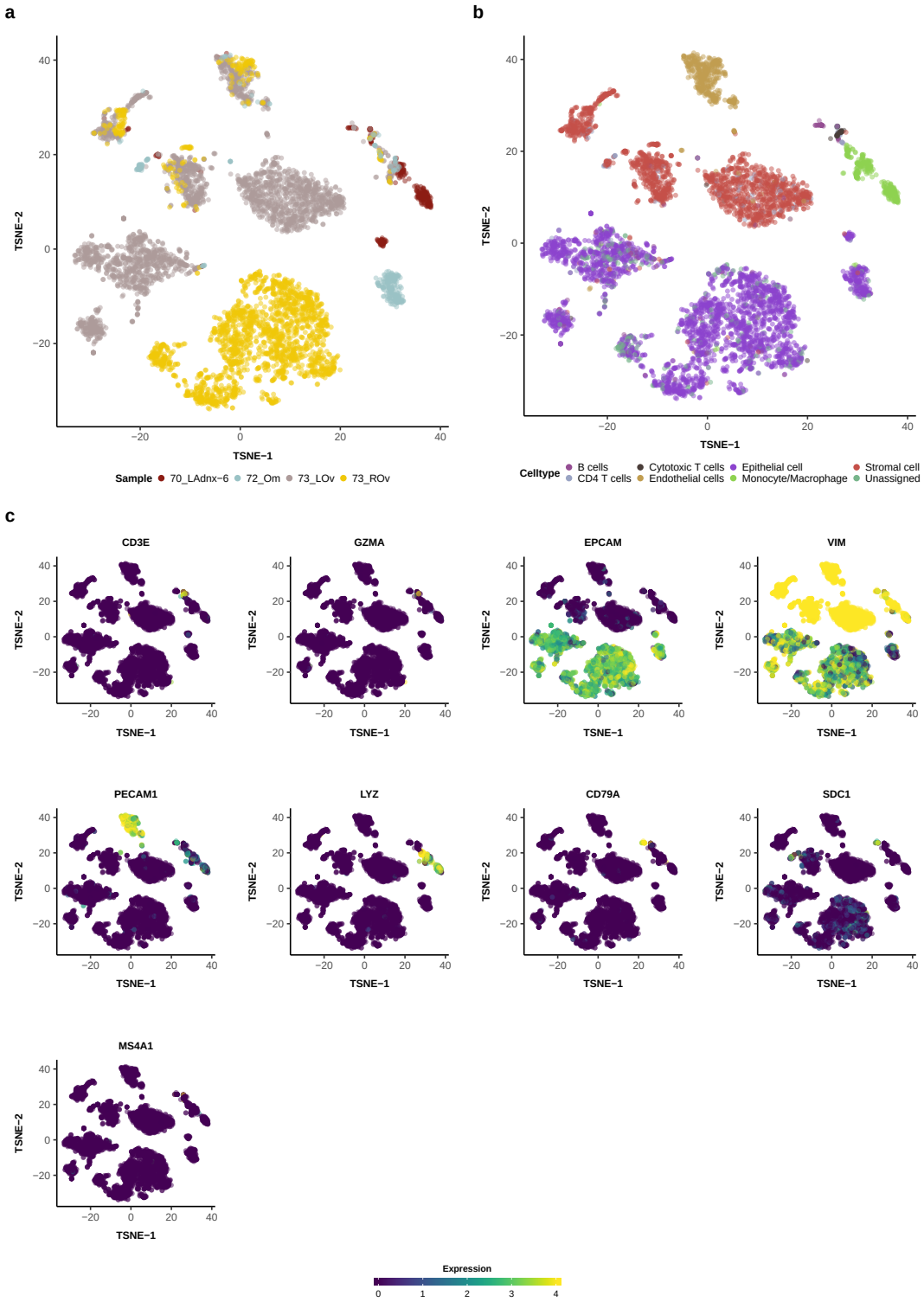


Figure 4.9: CellAssign infers the composition of the HGSC microenvironment. (**a**) t-SNE plot of HGSC single cell expression data, labeled by sample. (**b**) t-SNE plot of HGSC single cell expression data, labeled by maximum probability assignments from CellAssign. (**c**) Expression of select marker genes *CD3E* (for T cells [104]), *GZMA* (for CD8 T cells [104]), *EPCAM* (for epithelial cells [195]), *VIM* (for mesenchymal cells), *PECAM1* (for endothelial cells [197]), *CD79A* (for B cells [104]), *LYZ* (for monocytes [104]), *SDC1* (for plasma cells [208]), and *MS4A1* (for non-plasma B cells).

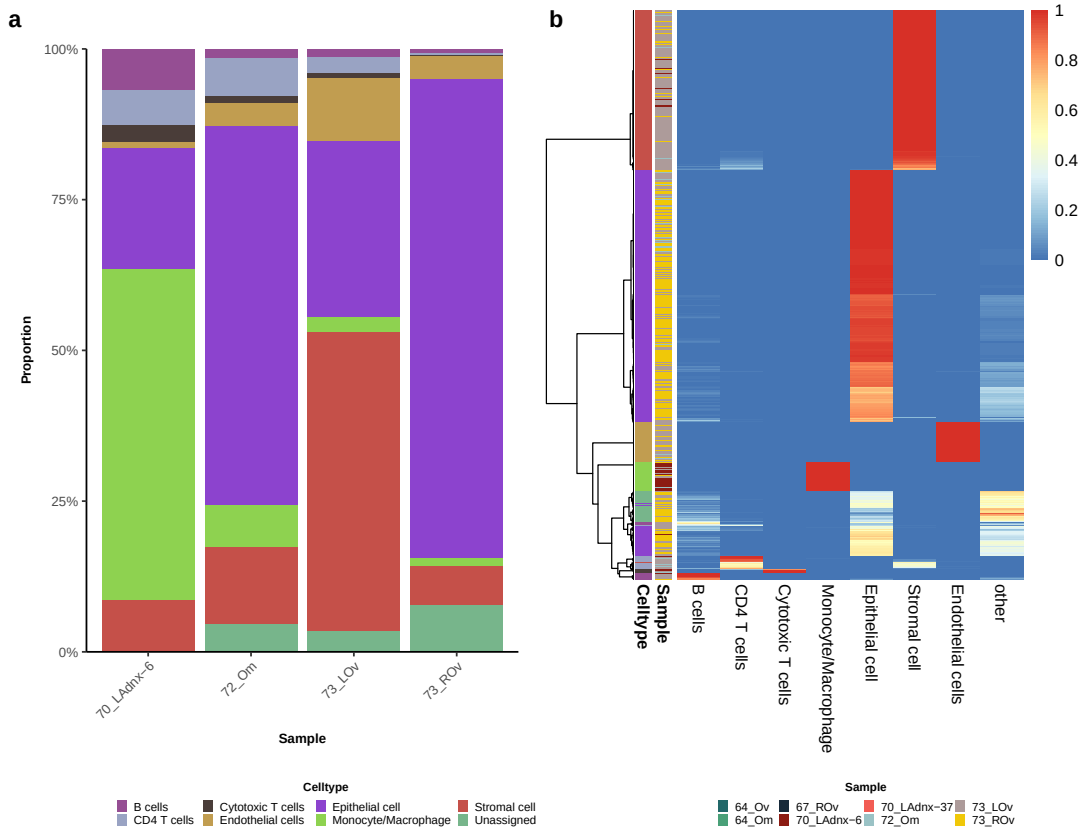


Figure 4.10: Proportions and probabilities of cell type assignments. (**a**) Proportions of each CellAssign-assigned celltype in each sample. (**b**) Cell-level probabilities from the CellAssign model, labeled by maximum probability celltypes and sample.

4.3.4 Stromal subpopulations in the ovarian cancer microenvironment

Having profiled the immune microenvironment in Chapter **3**, we next surveyed stromal cell populations in the HGSC microenvironment. Considering cells assigned to the stromal cell type (i.e. stromal cells with the exception of endothelial cells) with a probability of at least 0.99 by CellAssign, we performed unsupervised clustering with `densitycut` [206] (**Figure 4.11**A,B), revealing 6 clusters. Reasoning that stromal cells from different anatomic sites may express different markers, we interrogated the expression profiles of ovarian stroma-specific markers. Based on the expression of *MUM1L1* and *ARX* [198], we annotated clusters 1 and 2 as ovarian stromal cells, consistent with the ovarian or adnexal origin of cells from these clusters (**Figure 4.11**A-C). In contrast, cluster 5, which corresponds to cells from an omental sample (72_Om) did not express these markers (**Figure 4.11**A-C). Following this, we explored genes

associated with other microenvironmental populations, including pericytes, myofibroblasts, and smooth muscle cells. Based on the expression of *PLN*, *MYH11*, and *ACTA2* [209–211], we putatively labeled cluster 4 as vascular smooth muscle (**Figure 4.11**D). Most smooth muscle cells are *VIM* (vimentin) negative and desmin positive, but vascular smooth muscle contains a predominance of vimentin [212] (**Figure 4.9**C). While cluster 3 also contained *MYH11*-expressing cells, the pattern of expression in t-SNE space was not homogeneous. *MYH11* expression in cluster 3 negatively correlated with the expression of pericyte markers *THY1*, *CD248*, and *PDGFRB* [213] (**Figure 4.11**D). As such, cluster 3 likely contains a mixture of pericytes and myofibroblasts. Thus, the HGSC microenvironment contains multiple phenotypically-distinct stromal subpopulations that resemble ovarian stromal cells, extraovarian stromal cells, myofibroblasts, pericytes, and vascular smooth muscle cells.

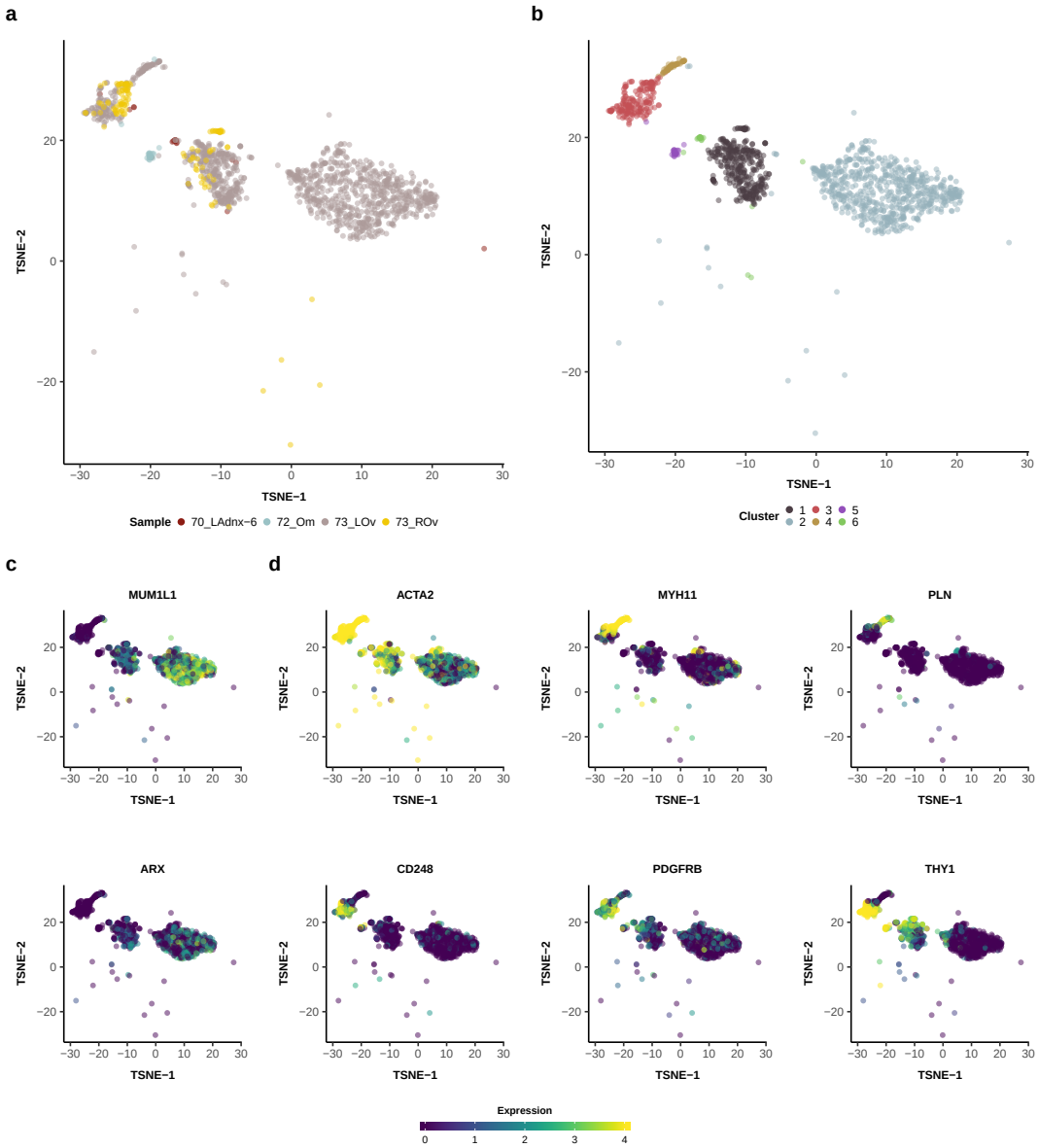


Figure 4.11: Stromal subpopulations in the HGSC microenvironment. (**a**) t-SNE projection of stromal (and non-endothelial) populations in the HGSC microenvironment, labeled by sample. (**b**) t-SNE projection of stromal populations in the HGSC microenvironment, labeled by CellAssign-assigned cell type. (**c**) Expression of ovarian stromal marker genes *MUM1L1* and *ARX* in the HGSC microenvironment [198]. (**d**) Expression of various stromal, pericyte, and muscle-associated genes in the HGSC microenvironment.

4.3.5 Dissecting the lymphocyte composition of follicular lymphoma

To demonstrate the utility of CellAssign for microenvironment analysis in another cancer type, we sequenced the transcriptomes of 9754 cells from serially collected lymph node biopsies of 2 patients with follicular lymphoma (FL1018: 4321 cells; FL2001: 5433 cells). Histopathological transformation to an aggressive subtype of B cell lymphoma, diffuse large B cell lymphoma (DLBCL), occurred in one patient (FL1018), while early progression (4 years after initial treatment with rituximab) occurred in the other (FL2001) (**Figure 4.12**A). Following data preprocessing and normalization, we applied principal components analysis (PCA) and uniform manifold approximation and projection (UMAP, [214]), revealing 5 major clusters in the reduced-dimension projections (**Figure 4.12**B). Three clusters appeared to be relatively pure for cells from a single patient, while the other 2 clusters comprised a mixture of cells from both patients. Leveraging marker gene information derived from the literature (**Supplemental Table B.2**; **Methods**), we applied CellAssign to identify 4 major B and T cell populations across these clusters (**Figure 4.12**C). One of the mixed clusters exclusively contained T cells, while the other 4 clusters were largely B cell-specific (**Figure 4.12**C). These cell type assignments appeared to be consistent with the expression patterns of well-known marker genes, such as *CD3D* and *CD2* for T cells, *CD79A* and *MS4A1* (*CD20*) for B cells, *CCL5*, *CD8A*, and *GZMA* for CD8+ T cells, *CD4*, *CXCR5*, and *ICOS* for T follicular helper cells, and *CD4* for other CD4+ T cells (**Figure 4.12**D, **Figure 4.13**, **Methods**). No evidence of regulatory T cells (*FOXP3* and *IL2RA* expression), NK cells (*NCAM1* expression), and myeloid cells (*CD14*/*CD16* and *LYZ* expression) was detected.

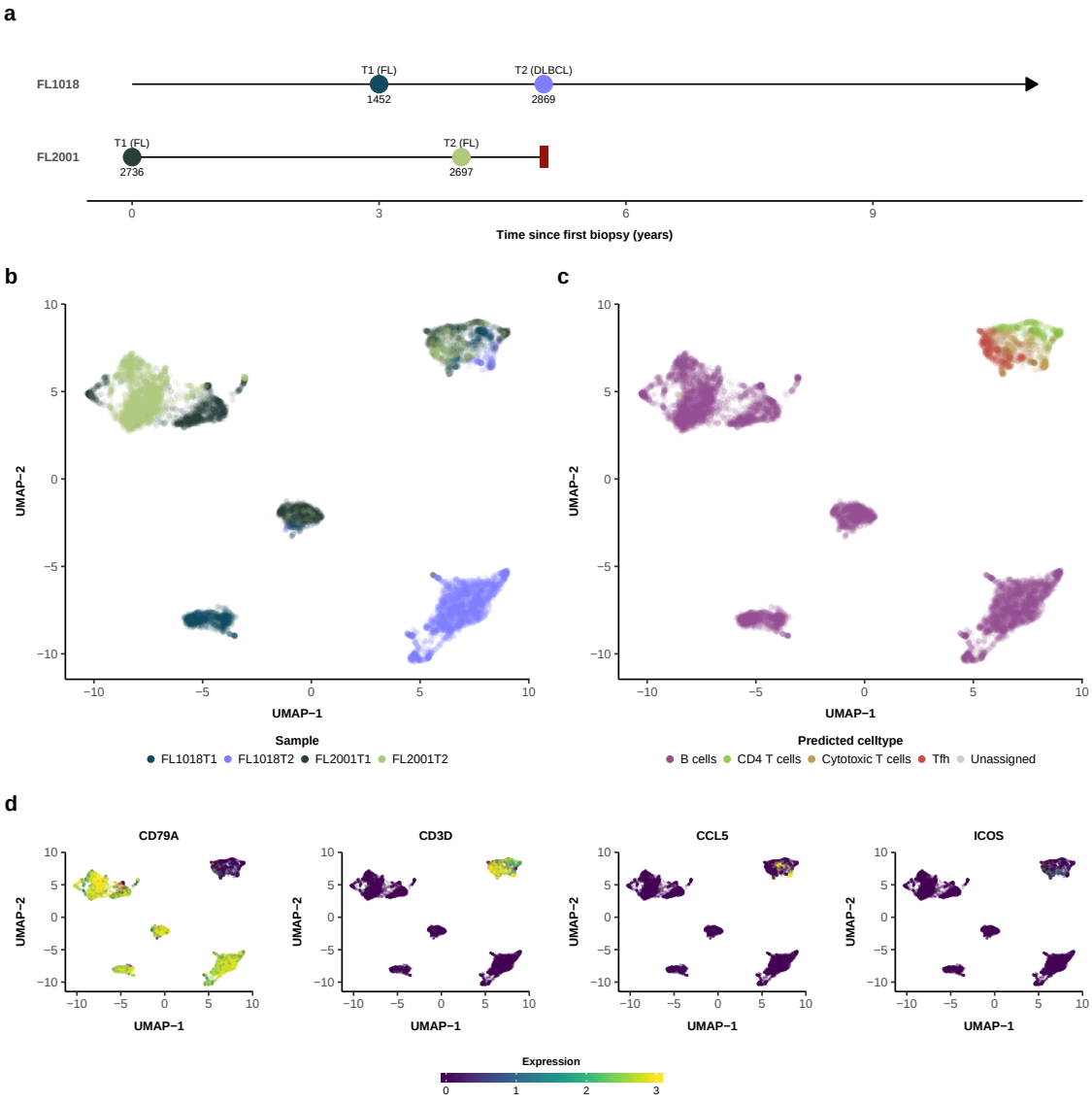


Figure 4.12: CellAssign infers the composition of the follicular lymphoma microenvironment. (**a**) Sample collection times for FL1018 (transformed FL) and FL2001 (progressed FL). FL1018 is alive while FL2001 was lost to followup (indicated by the red rectangle). The number of cells collected for each sample is indicated. (**b**) UMAP plot of follicular lymphoma single cell expression data, labeled by sample. (**c**) UMAP plot of follicular lymphoma single cell expression data, labeled by maximum probability assignments from CellAssign. (**d**) Expression of select marker genes *CD79A* (for B cells), *CD3D* (for T cells), *CCL5* (for CD8+ T cells), and *ICOS* (for T follicular helper cells).

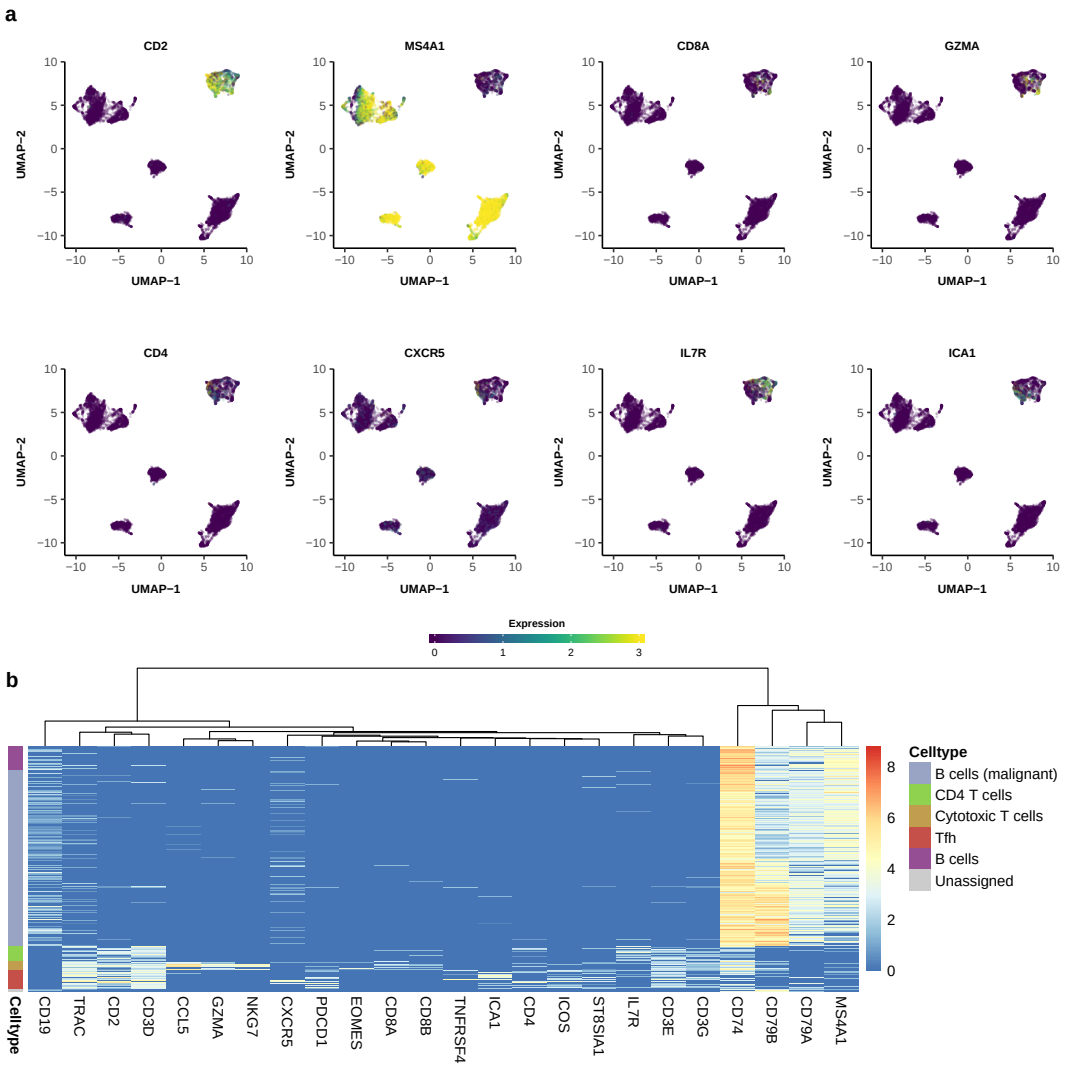


Figure 4.13: (a) Expression of select marker genes *CD2* (for T cells), *MS4A1* (for B cells), *CD8A* and *GZMA* (for CD8+ T cells), *CD4* (for CD4+ T cells and T follicular helper cells) and *CXCR5* and *ICOS* (for T follicular helper cells). (b) Heatmap of marker gene expression, labeled by maximum probability CellAssign-inferred cell types.

We next interrogated the identity of each B cell cluster. Of the B cell clusters, three were almost exclusively comprised of cells from a single patient, while one was a mixture of cells from both patients (**Figure 4.14**A). Reasoning that nonmalignant B cells were likely more phenotypically similar across timepoints than cancer cells, we hypothesized that the mixed cluster contained nonmalignant B cells. To explore this further, we examined immunoglobulin light chain constant domain expression across these clusters (**Figure 4.14**B). Each clonally identical population of B

cells produces immunoglobulins containing a single class of immunoglobulin light chain ($\kappa/IGKC$ or $\lambda/IGLC$), created through V(D)J recombination. Whereas normal lymphoid organs typically contain a 60:40 ratio of κ- to λ-expressing B cells [215], we observed a substantial departure from this ratio among all B cells in both patients (**Figure 4.14**B). Notably, the majority of the three patient-specific clusters were exclusively IGLC+, consistent with expansion of a malignant IGLC+ cell, while the mixed cluster contained both IGKC+ and IGLC+ cells (**Figure 4.14**B). Hypothesizing that the $IGKC$:$IGLC$ ratio in nonmalignant B cells should be similar to that for normal lymphoid organs, we applied CellAssign to the mixed cluster, using $IGKC$ as a marker of IGKC+ cells and $IGLC2$ and $IGLC3$ as markers of IGLC+ cells (*IGLC1* and *IGLC7* were not expressed; **Supplemental Table B.2**). Out of the 774 cells that were assigned to either group, 456 cells (58.9%) were classified as IGKC+ (FL1018: 67/106 (63.2%), FL2001: 389/668 (58.2%)), consistent with results for normal lymphoid organs (**Figure 4.15**). Finally, we attempted to delineate nonmalignant B cells by comparing B cell expression patterns to those derived from lymph node B cells from healthy donors. Using scvis [201], we first trained a variational autoencoder to produce a 2-dimensional embedding of the follicular lymphoma single cell RNA-seq data. Following this, we applied scvis to map similarly processed single cell RNA-seq data for reactive lymph node (RLN) B and T cells from four healthy donors onto this embedding. Concordant with our other predictions, RLN-derived T cells mapped to follicular lymphoma-derived T cells and RLN-derived B cells mapped to the mixed cluster of follicular lymphoma-derived B cells (**Figure 4.14**C, **Figure 4.16**). Thus, the mixed cluster is comprised of nonmalignant B cells, while the other 2 clusters represent malignant B cells. Corroborating this, differential expression analysis revealed that these malignant B cells express significantly higher levels of follicular lymphoma-associated markers, such as *BCL2* and *BCL6* [216–218], than nonmalignant B cells (all $Q < 1.8\text{e-}07$; **Figure 4.17**, **Supplemental Table B.3**).

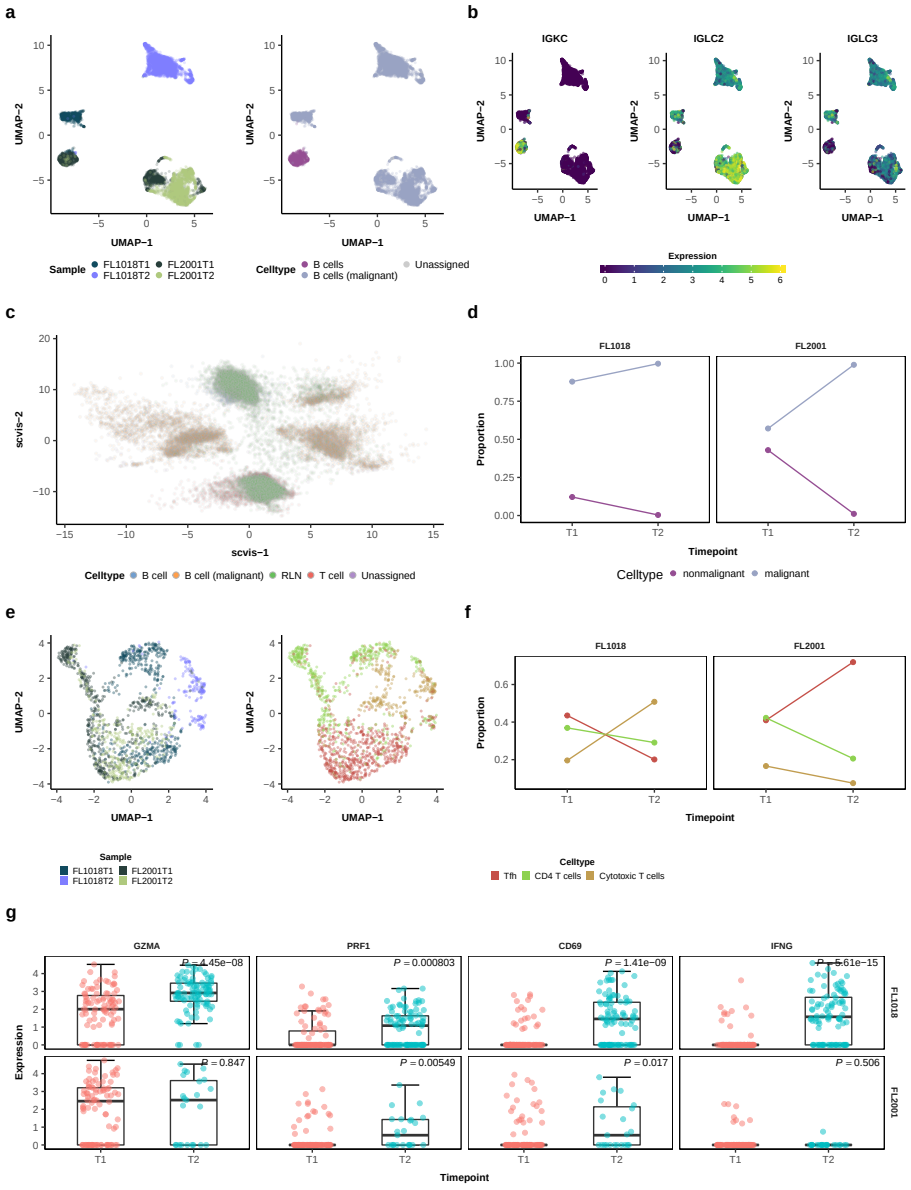


Figure 4.14: Temporal changes in nonmalignant cells in the follicular lymphoma microenvironment. (**a**) Left: UMAP plot of CellAssign-inferred B cells, labeled by sample. Right: UMAP pot of CellAssign-inferred B cells, labeled by putative malignant/nonmalignant status. (**b**) Expression of κ (*IGKC*) and λ (*IGLC2* and *IGLC3*) light chain constant region genes. (**c**) Scvis plot of follicular lymphoma data and single cell RNA-seq data of lymphocytes from reactive lymph nodes from healthy patients. The follicular lymphoma data was used to train the variational encoder and produce the two-dimensional embedding. Indicated cell types are B cell (nonmalignant B cell from FL), B cell (malignant) (malignant B cell from FL), T cell (T cell from FL), RLN (reactive lymph node cell). (**d**) Relative proportion of B cell subpopulations over time. (**e**) UMAP plots of FL T cells, labeled by sample and CellAssign-inferred celltype. (**f**) Relative proportion of T cell subpopulations over time. (**g**) Normalized expression of CD8+ T cell activation markers over time. *P*-values computed with the Wilcoxon rank-sum test and adjusted with the Benjamini-Hochberg method.

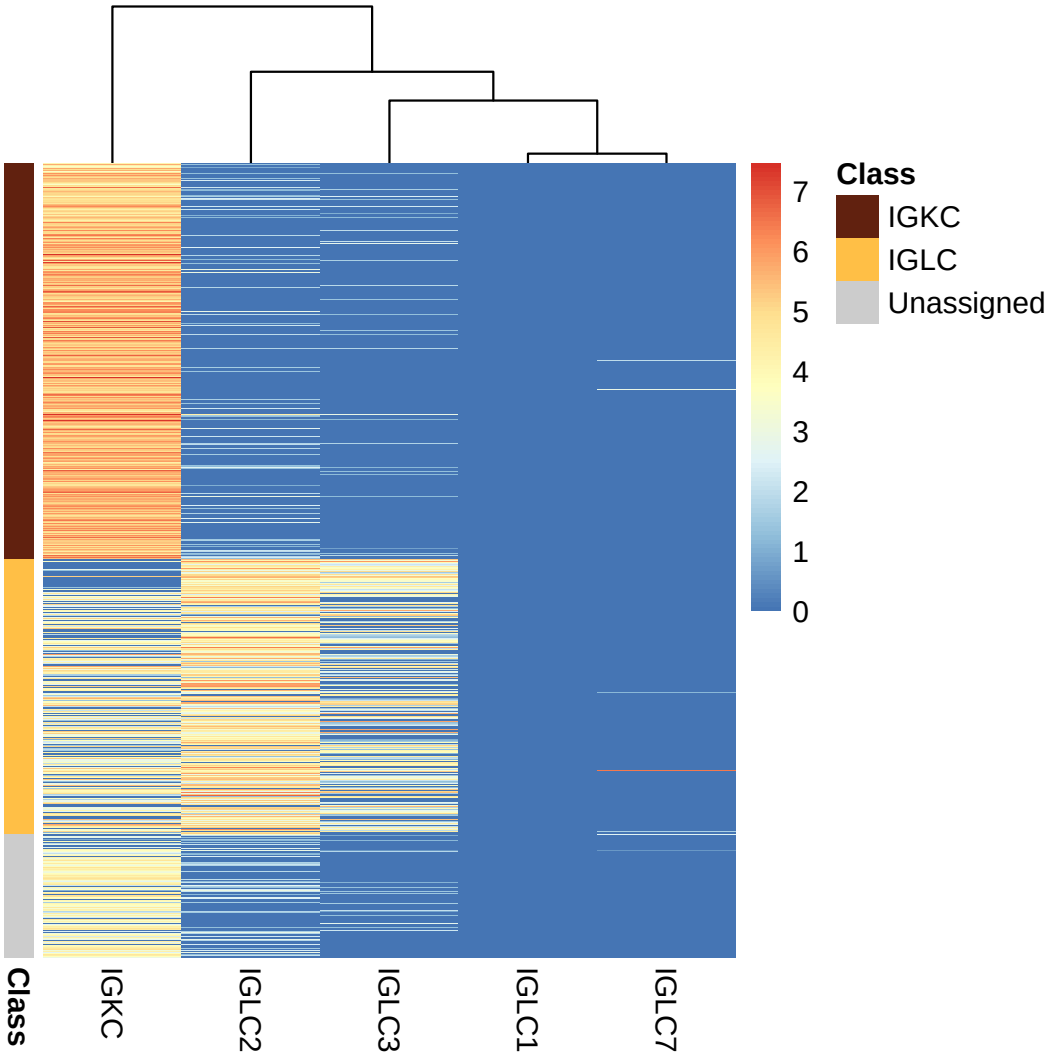


Figure 4.15: Expression of κ and λ light chain constant region genes in nonmalignant B cells. Class assignments were determined by CellAssign (**Methods**).

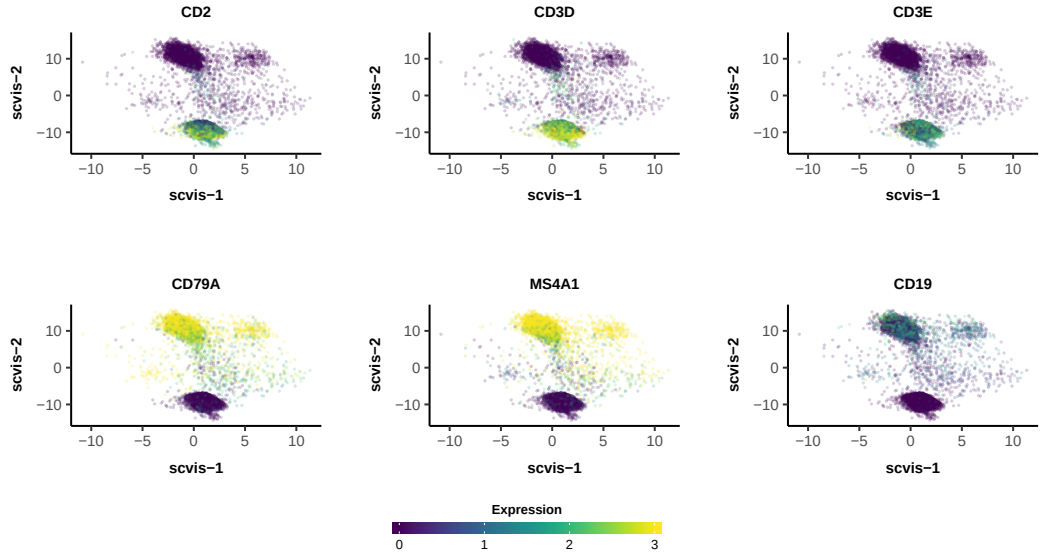


Figure 4.16: Expression of selected marker genes (*CD2*, *CD3D*, and *CD3E* for T cells; *CD79A*, *MS4A1*, and *CD19* for B cells) in scvis embedding of reactive lymph node data.

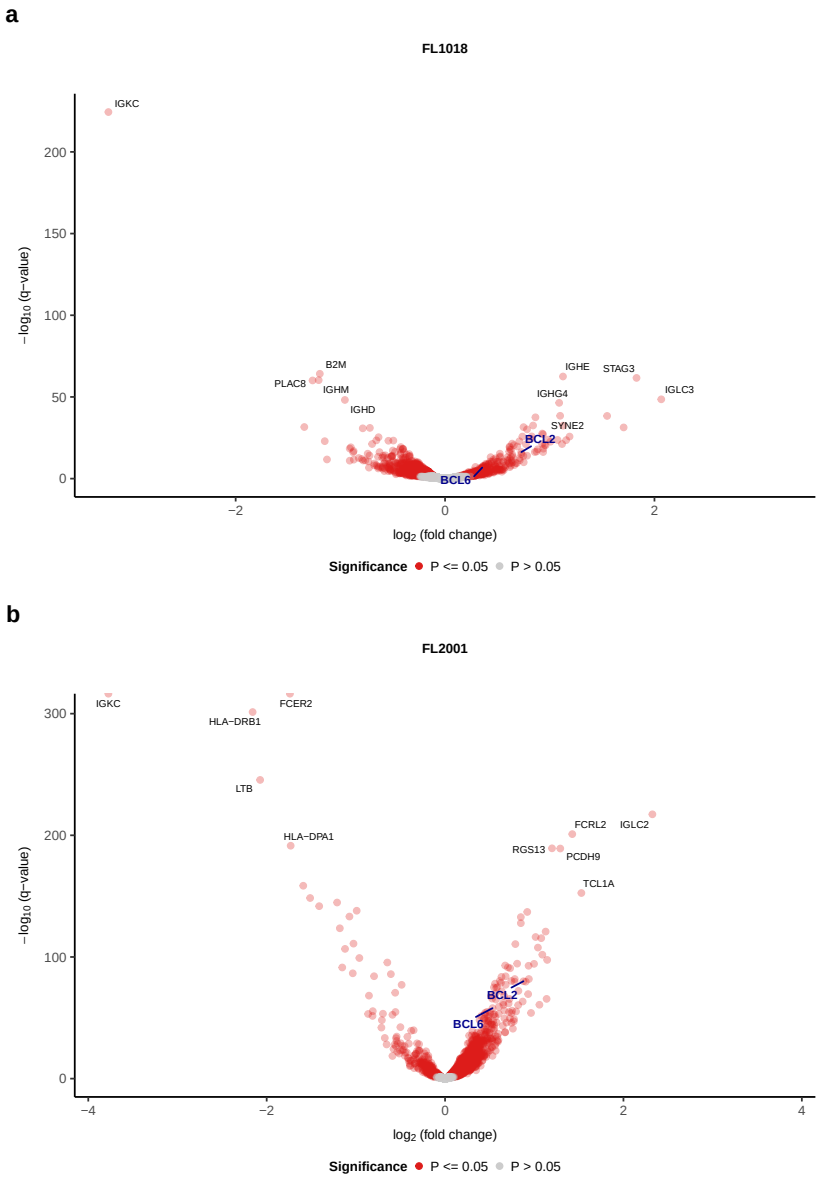


Figure 4.17: Differential expression results for malignant vs. nonmalignant B cells in (a) FL1018 and (b) FL2001. Comparisons was performed accounting for timepoint and potential interactions between malignant status and timepoint using a multivariate linear model described in **Methods**. Genes upregulated among malignant cells have logFC values > 0. *P*-values were adjusted with the Benjamini-Hochberg method.

4.3.6 CellAssign uncovers compositional and phenotypic changes in the follicular lymphoma microenvironment

We next asked how the relative abundance of each cell type differed after transformation or early progression. While the overall proportion of B cells in both cases was higher in the second timepoint, the relative proportion of nonmalignant B cells decreased dramatically (FL1018: 12.2% to 1.4%; FL2001: 44.4% to 1.4%) (**Figure 4.14**D). Thus, malignant B cells appear to dominate the B cell population upon transformation or early progression. Among T cells, the relative proportions of each cell subtype were comparable in FL1018 and FL2001 at the first timepoint, with T follicular helper cells and CD4+ T cells composing the majority of T cells and cytotoxic T cells the minority (**Figure 4.14**E,F). However, compared to the consistent pattern of B cell dynamics seen across both patients, T cell compositional dynamics upon transformation or early progression appeared to be divergent (**Figure 4.14**F). Cytotoxic T cells dominated the recurrence sample in FL1018, whereas T follicular helper cells became the major T cell population following progression.

Additionally, we looked at whether these compositional changes in the follicular lymphoma microenvironment were accompanied by phenotypic changes in nonmalignant cell populations between timepoints. To address this, we performed differential expression analysis across timepoints for each patient and cell type separately (**Supplemental Table B.3**). In FL1018, differential expression analysis revealed upregulation of immune-associated pathways [203], such as cytokine signalling and toll-like receptor pathways, among cytotoxic T cells in the transformed sample (**Figure 4.18**A). Similar results were observed for T follicular helper and CD4+ T cells (**Supplemental Table B.3**). T cell activation and effector molecules including *CD69*, *IFNG*, *GZMA*, and *PRF1* were also significantly upregulated in the second timepoint among cytotoxic T cells in FL1018 [219] (**Figure 4.14**G). Likewise, *CD69* was significantly upregulated among T follicular helper and CD4+ T cells in the recurrence sample (all $Q <$ 9.1e-07; **Figure 4.19**). Among cytotoxic T cells in FL2001, other immune related pathways such as antigen presentation and TCR/BCR signalling were upregulated in the early progressed sample, but *GZMA* and *IFNG* were not significantly differentially expressed between timepoints (**Figure 4.18**A, **Figure 4.14**G). Ubiquitin-associated genes and pathways were significantly upregulated in T follicular helper cells and CD4+ T helper cells in FL2001 as well (**Supplemental Table B.3**). In nonmalignant B cells, no significantly differentially expressed pathways were observed in either patient apart from upregulation of general adaptive immune system genes encompassing canonical B cell markers such as *CD79A*, *CD79B*, and HLA molecules in FL2001 (**Supplemental Table B.3**). Thus, transformation, and to a lesser extent early

progression, appears to be accompanied by T cell activation.

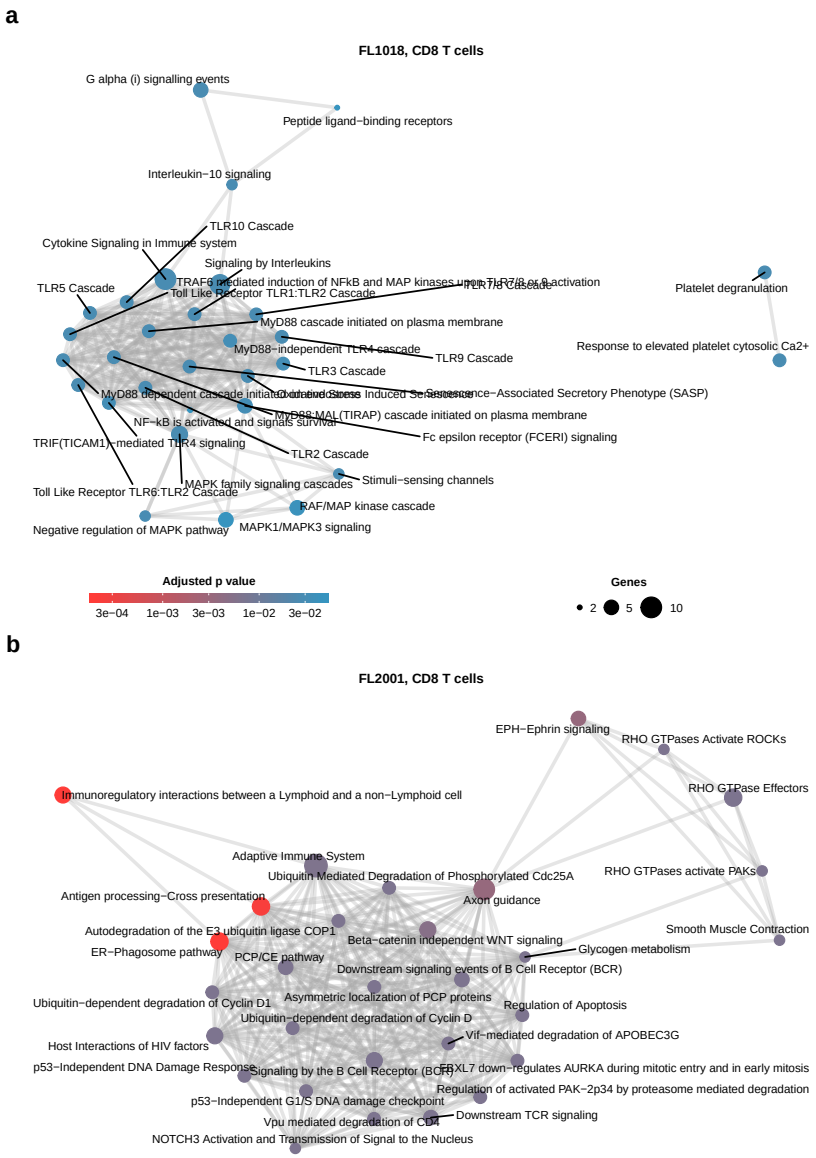


Figure 4.18: Significantly enriched Reactome pathways (BH-adjusted P-value ≤ 0.05) among the top 50 most highly upregulated genes (ranked by log fold change) in (a) FL1018 and (b) FL2001. Up to 30 pathways are shown in either plot (**Methods**).

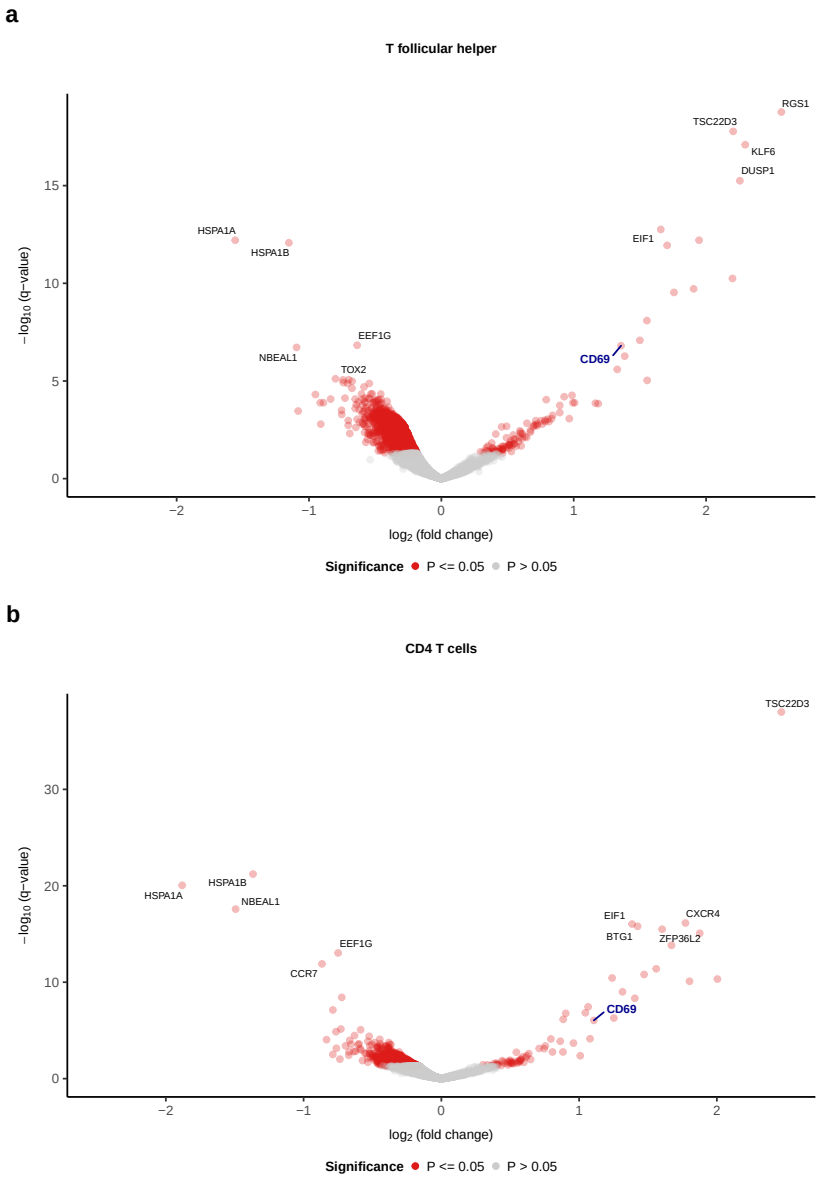


Figure 4.19: Differentially expressed genes for (a) T follicular helper and (b) other CD4 T cells between T2 vs. T1. Genes upregulated in T2 have log fold change values > 0. The activation marker *CD69* is highlighted. *P*-values were adjusted with the Benjamini-Hochberg method.

4.3.7 Malignant cell dynamics associated with early progression and transformation

Having explored the temporal dynamics of nonmalignant cells, we next investigated transcriptomic changes in malignant B cells. At a high level, malignant B cells from FL1018T1 and FL1018T2 were more distinct in UMAP space and were less concordant than those from FL2001T1 and FL2001T2 (Pearson's correlation coefficient between mean sample expression profiles = 0.97 and 0.982 in FL1018 and FL2001, respectively), suggesting higher levels of transcriptomic divergence upon transformation compared to early progression (**Figure 4.14**A). To analyze the nature of these differences in malignant cell transcriptomes, we performed differential expression and gene set enrichment analysis of malignant B cells across timepoints using cancer hallmark pathways (**Supplemental Table B.3**). Proliferation and cell cycle-associated pathways, including MYC targets, E2F targets, and G2M checkpoint-associated genes, were significantly upregulated in the recurrence sample of FL1018 (all Q <0.0016), suggesting an increase in proliferative potential following transformation [204] (**Figure 4.20**A). While MYC targets were also upregulated following recurrence in FL2001 (Q =0.0043), the cell cycle-associated E2F and G2M pathways were not (all Q >0.34), and the cell cycle-associated mitotic spindle pathway was significantly downregulated (Q = 0.0043; **Figure 4.20**B). Based on these findings, we explored the distribution of cell cycle-associated genes in malignant cells. Overlaying the expression of *MKI67* and *TOP2A* onto the UMAP embedding for malignant B cells revealed putative replicative clusters in both patients (**Figure 4.20**C,E). In the transformed but not early progressed case, a larger proportion of malignant cells from the recurrence sample appeared to be associated with these replicative clusters (**Figure 4.20**C,E), and differential expression analysis showed significant upregulation of *MKI67* and *TOP2A* in the recurrence sample (all Q < 7e-07. Correspondingly, cell cycle analysis with cyclone [200] revealed that a higher proportion of cycling (S or G2/M phase) malignant cells was present following transformation in FL1018 (**Figure 4.20**D). In contrast, we observed a reduced proportion of cycling T follicular helper, cytotoxic, and CD4+ T cells in FL1018 (**Figure 4.20**D). Consistent with the findings from pathway analysis, there were fewer cycling malignant cells in the recurrence sample of FL2001 (**Figure 4.20**F). Therefore, transformation, but not progression, appears to be associated with increased replication among malignant cells.

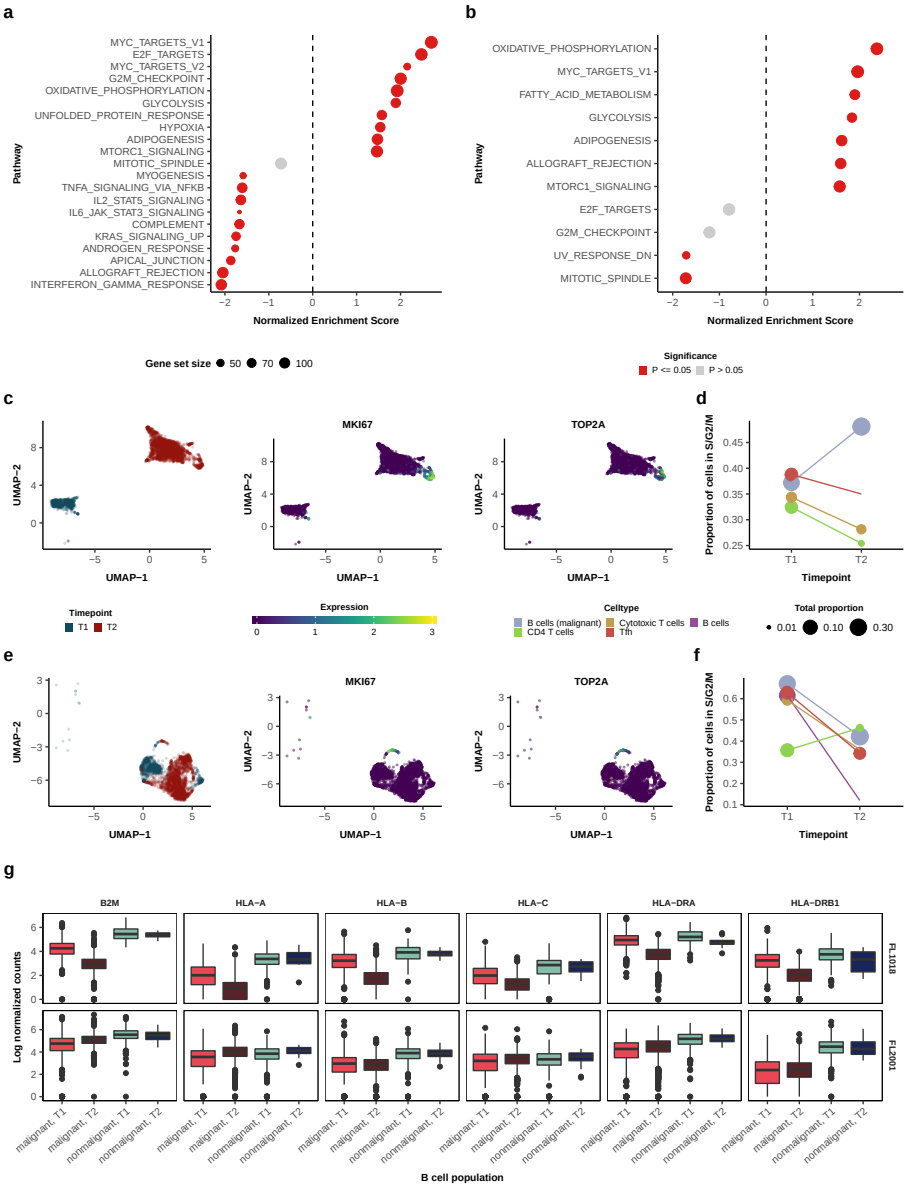


Figure 4.20: Temporal changes in malignant cells in the follicular lymphoma microenvironment. (**a,b**) Pathway enrichment scores computed by fGSEA for differentially enriched (adjusted $P \leq 0.05$) and cell cycle-associated pathways among malignant cells between timepoints for (a) FL1018 and (b) FL2001 (**Methods**). Pathways with a positive enrichment score are upregulated in T2 compared to T1 samples. *P*-values were adjusted with the Benjamini-Hochberg method. (**c,e**) UMAP plots, labeled by sample and proliferation marker expression (*MKI67* and *TOP2A*), for (c) FL1018 and (e) FL2001. (**d,f**) Proportion of cells assigned to be in non-G1 cell cycle phases (S/G2/M) by `cyclone` across timepoints in (d) FL1018 and (f) FL2001. (**g**) Normalized expression of HLA class I genes and select HLA class II genes across timepoints in FL1018 and FL2001.

Several immune-associated pathways, including complement and interferon gamma response, were downregulated in the recurrence sample of FL1018 (**Figure 4.20**A, **Supplemental Table B.3**). To interpret these findings, we looked at genes that were most downregulated in the recurrence sample based on effect size and significance (**Figure 4.21**). Among these were HLA class I and II genes, including *HLA-A*, *HLA-B*, *HLA-C*, *B2M*, *HLA-DRA*, and *HLA-DRB1*. In order to further investigate the temporal dynamics of HLA expression, we analyzed HLA expression levels in both nonmalignant and malignant B cells across timepoints. In both patients, HLA class I and II genes were expressed at lower levels in malignant B cells compared to nonmalignant B cells (all $Q < 0.037$; **Figure 4.20**G,H; **Methods**). Moreover, HLA expression levels in nonmalignant B cells were similar between timepoints (**Figure 4.20**G,H). However, while the expression of HLA genes in malignant cells from FL2001 was comparable across timepoints, malignant cells in the transformed case expressed significantly lower levels of HLA genes at recurrence (all $Q <$ 9.6e-24; **Figure 4.20**G,H). Corroborating these findings, differential expression and pathway analysis revealed that the HLA class I antigen presentation pathway was downregulated in malignant cells from FL1018 upon transformation ($Q = 0.019$; **Figure 4.22**, **Supplemental Table B.3**). Coupled with the increase in cytotoxic proportion among T cells and upregulation of T cell activation markers in FL1018T2 compared to FL1018T1, these results are consistent with immune escape in response to T cell activation following histologic transformation. Asking whether these results could be explained by genomic changes in antigen processing or presentation genes, we analyzed whole-genome sequencing data to profile somatic single nucleotide variants (SNVs), indels, and copy number alterations. No variants in these genes or loss-of-heterozygosity of HLA class I genes were detected in either sample from FL1018. However, coding mutations in the histone acetyltransferase *CREBBP*, recently reported to be associated with HLA class II downregulation [220], were detected in all samples from both patients, providing a possible explanation for the lower HLA class II expression observed among malignant cells.

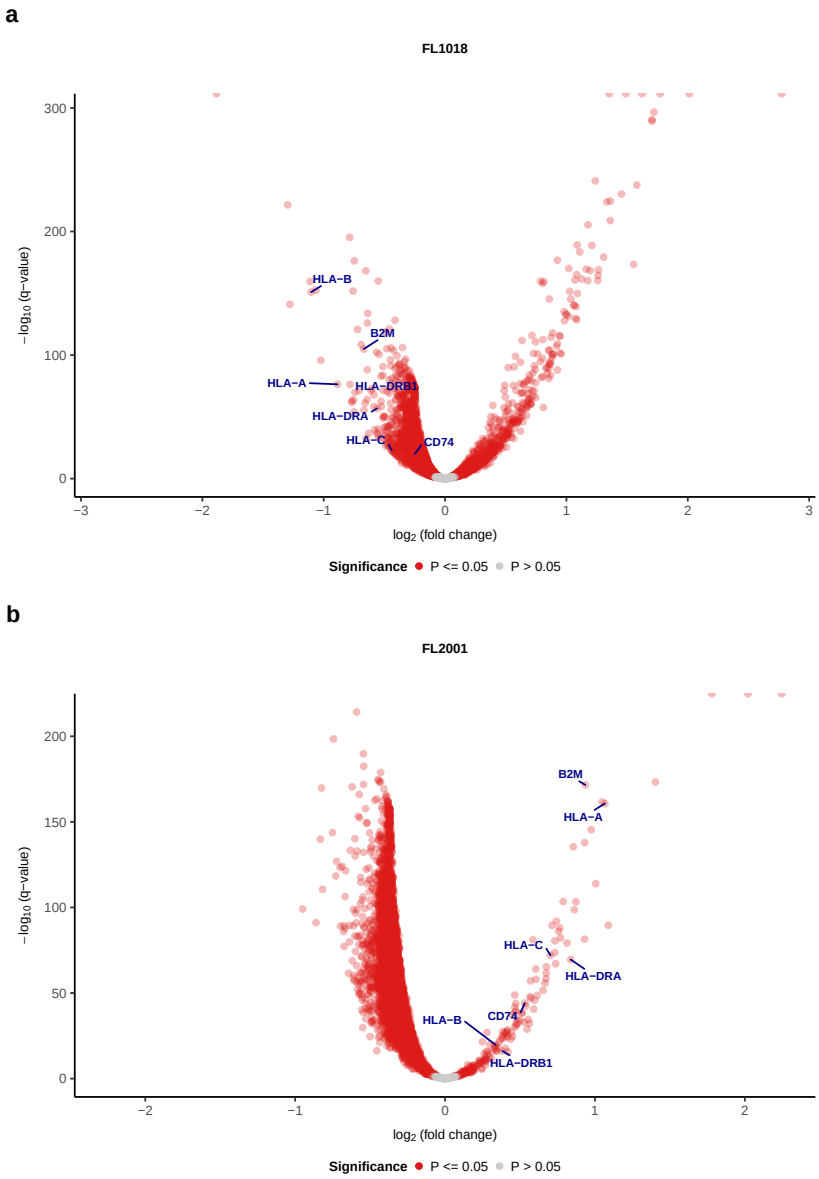


Figure 4.21: Differentially expressed genes between malignant B cells from T2 vs. T1 in (a) FL1018 and (b) FL2001. Genes upregulated in T2 have log fold change values > 0. HLA class I genes and select HLA class II genes are highlighted. P-values were adjusted with the Benjamini-Hochberg method.

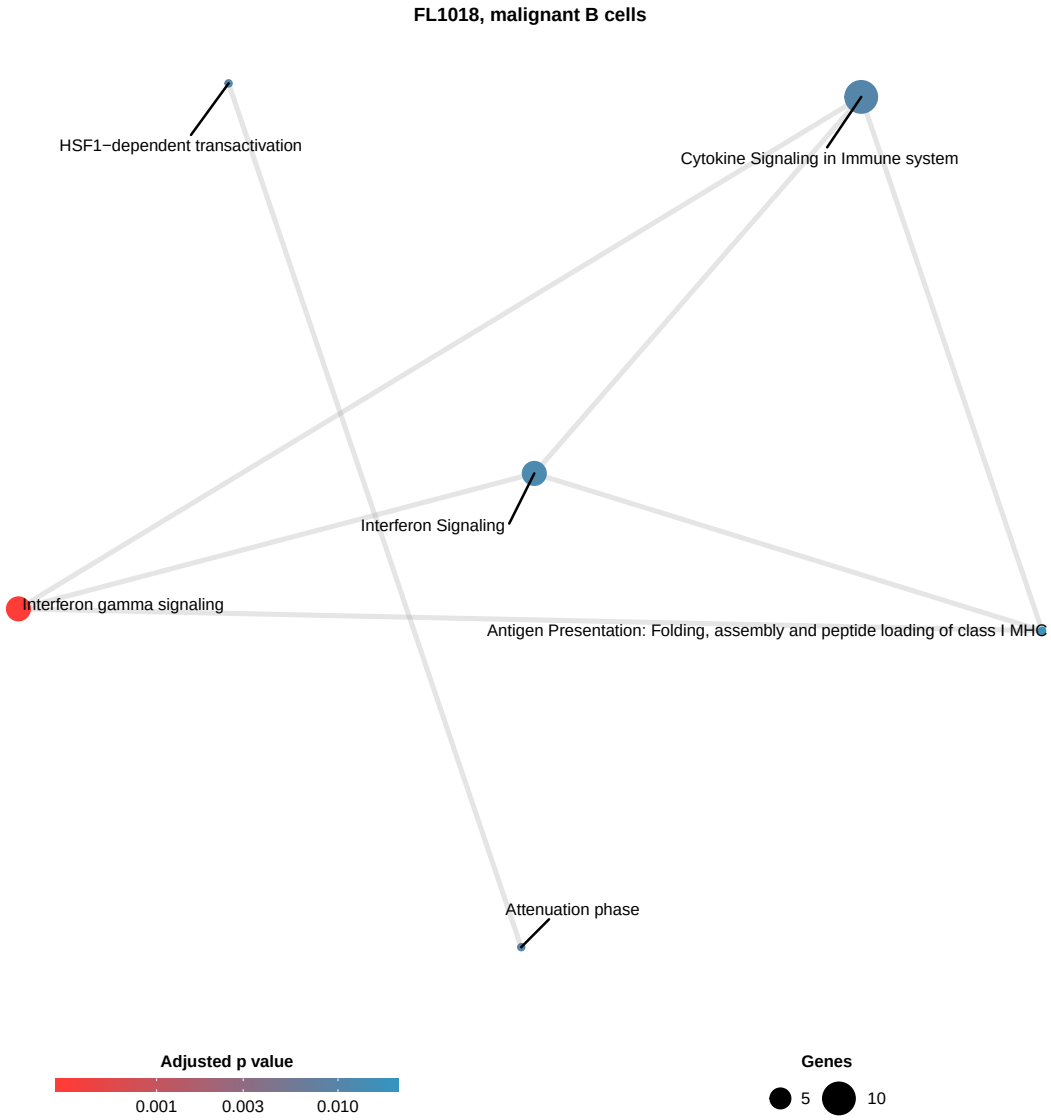


Figure 4.22: Significantly enriched Reactome pathways (BH-adjusted P-value ≤ 0.05) among the top 50 most downregulated genes (ranked by log fold change) in FL1018. No pathways were significantly downregulated in FL2001.

4.4 Discussion

We developed a computational method to automatically annotate single cell RNA sequencing data into cell types based on pre-defined marker gene information. Our approach systematically determines cell type expression patterns and assignment probabilities based solely on the assumption that marker genes are highly expressed in their respective cell types, eliminating the need for manual cluster annotation or existing training data for cell type mapping methods. In simulations and on real scRNA-seq data from purified populations, CellAssign's accuracy was comparable or superior to state-of-the-art workflows based on unsupervised clustering and mapping methods, and ran in a minute on datasets of tens of thousands of cells. We additionally demonstrate how bulk RNA-seq data can enable marker gene identification for accurate discrimination of phenotypically similar cell types with CellAssign.

We subsequently applied CellAssign to dissect the microenvironment composition of spatially- and temporally-collected samples from HGSC and follicular lymphoma. We show how CellAssign can not only delineate multiple malignant and nonmalignant epithelial, stromal, and immune cell types, but also identify subpopulations defined by arbitrary marker genes, uncovering $IGKC$:$IGLC$ ratios among nonmalignant B cells in follicular lymphoma consistent with those for normal lymphoid structures [215]. While these analyses are constrained by restricted cohort size, they provide first-of-kind examples of spatiotemporal dynamics and microenvironment interplay interpreted through leveraging prior knowledge of cell types in a prinicipled statistical approach.

We note that CellAssign is intended for scenarios where well understood marker genes exist. Poorly characterized cell types (or unknown cell types or cell states) may be invisible to the CellAssign approach. Furthermore, we make no *a priori* distinction between "medium" or "high" expression of the same marker in two different cell types, though these could be incorporated by extending the model to accommodate constraints between different δ parameters. Nevertheless, we suggest a large proportion of clinical applications profiling complex tissues start with hypotheses relating the composition of known cell types to disease states. As such, CellAssign fills an important role in the scRNA-seq analysis toolbox, providing interpretable output from biologically motivated prior knowledge that is immune to common issues plaguing unsupervised clustering approaches [183].

The volume of scRNA-seq data will increase over time in two important ways: (i) the number of cell types profiled will increase, thereby expanding databases of known marker genes and (ii) scRNA-seq data will become more widely available in research and clinical settings [221]. CellAssign is therefore poised to provide scalable, systematic and automated classification of

cells based on known parameters of interest, such as cell type, clone-specific markers, or genes associated with drug response. Furthermore, by appropriately modifying the observation model CellAssign can easily be extended to annotate cell types in data generated by other single-cell measurement technologies such as mass cytometry. We anticipate the CellAssign approach will help unlock the potential for large scale population-wide studies of cell composition of human disease and other complex tissues through encoding biological prior knowledge in a robust probabilistic framework.

Chapter 5

Conclusions and Future Directions

Despite extensive efforts to find effective therapies, prognosis for patients with high-grade serous ovarian cancer remains bleak. High-grade serous ovarian cancer patients often present with clonally heterogeneous disease with spread to peritoneal sites including the contralateral ovary, omentum, and pelvic wall. The extent of intratumoural heterogeneity in HGSC is thought to contribute to the prevalence of recurrence following initial treatment with standard-of-care combination platinum-taxane chemotherapy, with treatment resistant-clones escaping elimination. Nevertheless, the presence of TILs is associated with superior outcomes, hinting at the tantalizing possibility that the immune system may be able to contend with intratumoural heterogeneity in HGSC.

Consequently, the primary goal of this thesis was to profile the immune microenvironment, and more broadly, the tumour microenvironment in HGSC. This would improve understanding of the underlying spatial characteristics driving differential patterns of clonal seeding and proliferation in HGSC. In addition, given the recent success of immunotherapies in other cancer types, including checkpoint inhibitors in melanoma [222, 223] and CAR T-cells in B cell lymphomas [224], this work may help inform immunotherapeutic strategies for HGSC. The broad implications of this work are summarized below.

Assembly of an extensively profiled and largest published multi-site cohort of HGSC cases to date. Chapter **2** described the construction of our multi-site HGSC cohort which I led, encompassing clinical case identification and sample processing for genomic, imaging, single cell, and patient-derived xenograft studies. The patient-derived xenografts will be used to study clonal evolution in response to drug perturbation with cytotoxic and DNA damaging agents such as platinum compounds and PARP inhibitors, and the single cell materials will set the foundation for biological studies that leverage the CellAssign method described in Chapter **4**. This resource acts as the bedrock for the multimodal profiling study described in Chapter **3**, the single cell RNA-sequencing analysis of HGSC described in Chapter **4**, and planned future work (Section **5.1**).

Deciphering multiple interfaces of evolutionary interplay between tumour and im-

mune cells in HGSC. To investigate the interplay between immune and malignant cells underlying the spatial distribution of clones in HGSC, we profiled 212 multi-site HGSC samples from 38 patients, including cases collected as part of Chapter **2** with whole-genome sequencing, targeted sequencing and clonal deconvolution, Nanostring expression assays, histologic image analysis, immunohistochemistry, and T- and B-cell receptor sequencing. From this work, I uncovered 3 major subtypes of HGSC based on the spatial distribution of lymphocytes between tumour epithelium and stroma, with imaging and sequencing-based evidence of tumour-immune interaction in extensively infiltrated (ES-TIL) samples. I extended and applied probabilistic methods for clonal decomposition, revealing lower intratumoural heterogeneity among highly infiltrated samples, with some cases exhibiting loss-of-heterozygosity of HLA class I loci. The findings from this work include the novel discovery of immune escape mechanisms in HGSC (HLA loss-of-heterozygosity) despite T cell tracking of tumour clones among highly infiltrated tumour samples, providing an explanation for the poor responses to immune checkpoint inhibitors observed in HGSC to date.

Discovery of prognostically relevant associations between mutational processes and immunologic signatures. In Chapter **3**, I also applied a novel topic model-based approach developed by [26] to quantify mutational signatures associated with defective DNA damage repair in HGSC. Building on first-of-kind work by [3] identifying 2 major genomic subtypes in HGSC, we identified 4 major subtypes of HGSC, further subdividing the homologous recombination-deficient subtype and introducing a new tandem duplicator subtype with distinct survival outcomes and immunologic infiltration patterns [26] previously classified as foldback inversion-type [1]. Additionally, I describe a novel association between low immunologic infiltration and foldback inversion-harbouring tumours, along with a prognostic association between immune infiltration and mutational processes whereby immune infiltration is associated with superior overall survival in HRD but not FBI cases. In light of recent success with PARP inhibitor maintenance therapy in HGSC [17], these findings provide context for combination immunotherapy-PARP inhibitor therapies in clinical trials.

Developing an automated, scalable method for microenvironmental cell type identification from single cell transcriptomic data. Extending the work I led on exploring the immune composition of HGSC in Chapter **3**, I developed a novel marker gene-based probabilistic approach to identifying known cell types from single cell RNA-seq data (Chapter **4**). The CellAssign model uses a hierarchical Bayesian framework to systematically assign input cells to known cell types while accounting for any additional experimental covariates such as batch effects in a scalable, automated fashion. CellAssign only requires binary marker gene information, rather than the purified single cell RNA-seq data required by supervised methods, and

performs superiorly or comparably to state-of-the-art methods. I apply CellAssign to high-grade serous ovarian cancer and follicular lymphoma, demonstrating the utility of CellAssign for microenvironmental deconvolution and revealing interplay between malignant and immune cells. CellAssign sets the stage for scalable inference of cell types in large-scale single cell RNA-seq studies of cancer that are beginning to emerge, including our ongoing single cell RNA-seq work in high-grade serous ovarian cancer.

5.1 Future Directions

5.1.1 Understanding the molecular basis of immunologic infiltration patterns in HGSC

Chapter **3** identified 3 major patterns of immunologic infiltration in HGSC characterized by the absence of TIL (N-TIL), stromally-restricted TIL (S-TIL), and epithelial and stromal TIL (ES-TIL). While our work identified potential mechanisms by which ES-TIL tumours may escape from immune recognition, it does not explain the lack of epithelial immune infiltration in N-TIL and S-TIL tumours. Tumours with lower immune infiltration generally harbour higher subclonal neoantigen loads, suggesting that antigen deficiency does not underlie the lack of observed immune infiltrate. Furthermore, anatomic location was not significantly associated with particular patterns of immune infiltration. How S-TIL tumours manage to exclude TIL from epithelial areas remains unknown. One possibility is that S-TIL tumours contain stromal barriers that physically prevent TIL entry into epithelial regions. The predominance of fibroblasts in S-TIL tumours may inhibit T cell effector function through TGFβ production [225]. Recent work by [226] proposes an alternative cancer cell-mediated mechanism for epithelial TIL exclusion in triple-negative breast cancer, whereby immune infiltration in S-TIL patterns is more consistent with the presence of an TIL repellent produced by tumour cells than with physical blockade by desmoplastic elements. Moreover, the mechanisms that N-TIL tumours employ to escape from immune recognition remain a mystery. Laser capture microdissection-assisted RNA-seq and single cell RNA-sequencing may provide crucial insights into malignant or nonmalignant phenotypes that contribute to immune cell exclusion. These methods enable investigation of site-specific, clone-level phenotypes associated with N-TIL and S-TIL patterns. They also allow for deeper investigation of other microenvironmental cell types classically associated with immune regulation, including macrophages and fibroblasts. Multi-site cohorts, which intrinsically control for patient-specific factors and minimize batch effects associated with single cell processing, provide ideal substrate for these types of studies. Understanding the mechanisms underlying

immune exclusion in antigen-harbouring tumours may open therapeutic avenues for improving the delivery and efficacy of cancer immunotherapies.

5.1.2 Deciphering the mechanisms of treatment resistance in HGSC

The patterns of clonal dynamics across space explored in Chapter **3** and [29] highlight the extensive intra- and intertumoural heterogeneity that exists at the time of diagnosis in HGSC. While it is believed that clonal heterogeneity underlies treatment resistance by providing genetic substrate for evolution to act upon, the molecular mechanisms that are ultimately responsible for recurrent disease in the context of treatment with standard-of-care chemotherapy, apart from secondary mutations in *BRCA* genes [15], remain largely unknown. In Chapter **3**, we present evidence for subclonal immune escape, including neoantigen depletion and HLA loss-of-heterozygosity, that may allow certain clones to escape immune-mediated elimination. This may help explain some cases of resistance, where platinum-taxane chemotherapy leads to cell death, inducing a systemic abscopal-like anticancer response. More generally however, phenotypic alterations in cancer cells that affect drug influx and efflux, metabolism, and proliferation may lead to chemotherapeutic resistance. Drug challenge studies in model systems, such as patient-derived xenografts (PDXs), provide controlled systems in which to study clonal dynamics in response to treatment. Chapter **2** describes the construction of a first-of-kind multi-site HGSC PDX cohort intended for this purpose.

Broadly speaking, resistance mechanisms can be classified as intrinsic—encoded in the genome—or adaptive—mediated by epigenetic or context-specific factors, such as the microenvironment, and reflected in the transcriptome and proteome. The advent of scalable, low-bias single cell whole-genome sequencing technologies for clonal reconstruction and single cell transcriptome sequencing technologies for malignant cell phenotyping can help identify rare resistant clones and distinguish these two categories of resistance. This can be accomplished through experimental integration of single cell data from multiple modalities through combined genome-transcriptome-proteome sequencing or computational integration with methods such as clonealign [227]. Clonal populations that persist in drug-treated PDX models across multiple replicates and models derived from multiple diagnostic metastases may contain variants that can be profiled with these methods. Candidate variants could be further studied in cell lines with CRISPR perturbation. Emerging methods for fitness modeling from the population genetics literature can be leveraged to quantify the selective advantage conferred by each variant and genotype and predict future clonal dynamics. Humanized mouse models can also be used to transplant human hematopoietic stem cells in PDXs, enabling study of clonal dynamics in similar microenvironmental contexts [228, 229].

5.1.3 Deconvolution of the HGSC microenvironment

Despite extensive literature profiling lymphocytes in HGSC [36, 38, 41] including our own (Chapter **3**, [1]), few studies have characterized the cell type composition of the entire HGSC microenvironment. Other cell types, including macrophages and fibroblasts, are major components of the tumour microenvironment but remain largely uncharacterized in HGSC. One recent study attempted to address this question in a small cohort of epithelial ovarian cancers including 5 HGSC cases [230], but this sample size was insufficient to capture single cell phenotypes reflecting the complete diversity of microenvironmental profiles according to bulk expression profiling [101]. Indeed, our initial investigation of the HGSC microenvironment using single cell RNA-seq revealed several additional populations, such as pericytes and non-collagen expressing ovarian stromal cells (Chapter **4**). Larger cohorts are needed to phenotypically and prognostically characterize these rare microenvironmental subpopulations. Given mounting evidence implicating the immune and non-immune microenvironment along with interactions between microenvironmental cell types in HGSC disease progression [1, 59, 147], treatment decisions informed by mathematical modeling of cancer clones (Section **5.1.2**) will have to be interpreted in the context of the microenvironment. In Chapter **4**, we developed an automated method for systematically identifying cell types from single cell RNA-seq data. As single cell transcriptome studies on large cohorts of HGSC tumours begin to emerge, our method will enable cell type identification at scale while controlling for batch effects.

Furthermore, the roles of cell types that have been profiled in HGSC, particularly B cells and plasma cells, are unclear. While the presence of B and plasma cells is associated with superior outcomes in tumours that contain T cells [38, 41], the mechanisms by which these cell types act are largely unknown. In a preprint related to Chapter **3** [231], I explored the phylogeographics of B cell clones across space in multi-site HGSC, uncovering evidence suggesting that B cells are active participants in the anticancer immune response. B cells may act as antigen presenting cells for T cells, produce antibodies against tumour antigens, or exert direct cytotoxic effects [42]. Single cell transcriptomics may help to resolve the role of individual B cell clones. The 10X Chromium technology enables paired recovery of single cell transcriptomes and B cell receptor sequences, allowing B cell clones to be matched up to cellular phenotypes. This will enable phenotypic profiling of tumour-reactive B cell clones identified through clonal frequency analysis or antibody reactivity assays.

Finally, molecular subtyping from bulk expression data has yielded 4 major subtypes of HGSC that appear to be associated with microenvironmental features [14, 101]. Certain subtypes (C1 [101], mesenchymal [14]) are generally associated with worse outcomes, though the significance

of this finding varies between studies. Single cell transcriptomics will uncover the cell types and cell type-specific phenotypes that ultimately underlie each subtype. Furthermore, single cell RNA-seq applied to large cohort studies will decode phenotypically-distinct subtypes of each cell type, including cancer cells.

5.1.4 *In situ* profiling of the tumour microenvironment

Multi-site studies of HGSC have reproducibly shown that the properties of both malignant cells and the surrounding microenvironment can differ appreciably between peritoneal foci, even within the same macroscopic tumour [1, 27, 29, 147, 232]. Tumours may contain both immune-privileged niches capable of supporting diverse clonal populations and highly infiltrated areas in which clonal pruning has occurred [1]. Beyond microenvironmental cell type composition (Chapter **3**, Chapter **4**), *in situ* spatial profiling of single cell phenotypes may yield important insights into microenvironment-malignant cell interactions. For example, the chemotherapeutic resistance properties associated with fibroblasts and abrogated by T cells *in vitro* [59] may be dependent on spatial proximity between fibroblasts, cancer cells, and T cells. Likewise, clones bearing HLA loss-of-heterozygosity may reside in adjacent regions to cytotoxic T cells. These studies may also help decipher the mechanisms behind TIL exclusion in N-TIL and S-TIL tumours. Spatial transcriptomic profiling can be performed with techniques such as merFISH [233]. A recent study has described 3D intact-tissue spatial transcriptomics with a novel method called STAR-MAP that leverages DNA barcoding with SEDAL sequencing to simultaneously obtain readouts of up to 1000 genes [234]. These technologies allow for deep investigation of spatial interactions between microenvironmental cell types.

Alternatively, spatial patterns in the tumour microenvironment can be studied at lower depth but at scale. Cell type identification and pattern recognition from histologic images has been used to prognostically stratify cancers [145, 167]. In Chapter **3**, we profiled histologic images to identify cancer-immune cell hotspots associated with highly infiltrated tumours, suggesting direct cell type interaction. Cell type interactions can be further studied using spatial statistics. For instance, Gibbs point process models model pairwise interactions between collections of points, inferring the relative attractive or repulsive force between each pair of cell types. Immune recognition of cancer cells may be read out as attractive forces between immune and cancer cells. In summary, spatial profiling of histologic images from large cohorts may yield additional insights beyond those associated with lymphocyte abundance or TIL cluster (N-TIL, S-TIL, ES-TIL).

5.1.5 Guiding precision immunotherapies for HGSC

Cell-based immunotherapies, such as CAR T cell therapies, rely on selective expansion of immune-reactive T cell clones. Understanding the properties of these clones prior to perturbation, such as immune exhaustion marker expression, may be important to stratifying patients likely to respond to cell-based immunotherapies. Reactive T cell clones can be isolated with traditional MHC multimer assays, but transcriptome sequencing can only be done post-expansion when phenotypes have likely changed substantially. Paired single cell RNA-sequencing and TCR-seq presents a unique opportunity for phenotypic profiling of tumour-reactive T cells at diagnosis. Reactive T cell clones identified by MHC multimer or ELISPOT assays can be mapped to single cell RNA-seq data using the cell-specific barcodes in the 10X Chromium protocol. This may reveal unique properties of tumour-reactive T cells that can be deconvolved from bulk RNA-seq data to stratify patients and inform immunotherapeutic options at the time of diagnosis.

5.2 Concluding Remarks

The work presented in this thesis advances our understanding of clonal dynamics and the tumour microenvironment in HGSC. At the beginning of my thesis, I set out to understand the factors influencing evolutionary dynamics across space in ovarian cancer. These studies leverage first-of-kind multimodal experimental design with spatial sampling to reveal some of the first evidence of immune-cancer evolution in ovarian cancer, and set the stage for further systematic investigation of the microenvironment in HGSC and other cancer types. The next major advances will be enabled by *in situ* methods that can provide spatial evidence of interaction within individual tumour samples, and temporal sampling to profile malignant-immune evolutionary dynamics in recurrent disease. Resolving the tumour-microenvironment interface in HGSC will pave the way for therapeutic options that exploit non-malignant cell types to overcome the clonal heterogeneity pervasive to the disease.

Bibliography

[1] Allen W. Zhang, Andrew McPherson, Katy Milne, David R. Kroeger, Phineas T. Hamilton, Alex Miranda, Tyler Funnell, Nicole Little, Camila P.E. de Souza, Sonya Laan, Stacey LeDoux, Dawn R. Cochrane, Jamie L.P. Lim, Winnie Yang, Andrew Roth, Maia A. Smith, Julie Ho, Kane Tse, Thomas Zeng, Inna Shlafman, Michael R. Mayo, Richard Moore, Henrik Failmezger, Andreas Heindl, Yi Kan Wang, Ali Bashashati, Diljot S. Grewal, Scott D. Brown, Daniel Lai, Adrian N.C. Wan, Cydney B. Nielsen, Curtis Huebner, Basile Tessier-Cloutier, Michael S. Anglesio, Alexandre Bouchard-Côté, Yinyin Yuan, Wyeth W. Wasserman, C. Blake Gilks, Anthony N. Karnezis, Samuel Aparicio, Jessica N. McAlpine, David G. Huntsman, Robert A. Holt, Brad H. Nelson, and Sohrab P. Shah. Interfaces of Malignant and Immunologic Clonal Dynamics in Ovarian Cancer. *Cell*, 173(7):1755–1769, 6 2018.

[2] Allen W Zhang, Ciara O'Flanagan, Elizabeth Chavez, Jamie LP Lim, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, Daniel Lai, Anja Mottok, Clementine Sarkozy, Lauren Chong, Tomohiro Aoki, Xuehai Wang, Andrew P Weng, Jessica N McAlpine, Samuel Aparicio, Christian Steidl, Kieran R Campbell, and Sohrab P Shah. Probabilistic cell type assignment of single-cell transcriptomic data reveals spatiotemporal microenvironment dynamics in human cancers. *bioRxiv*, page 521914, 1 2019.

[3] Yi Kan Wang, Ali Bashashati, Michael S Anglesio, Dawn R Cochrane, Diljot S Grewal, Gavin Ha, Andrew McPherson, Hugo M Horlings, Janine Senz, Leah M Prentice, Anthony N Karnezis, Daniel Lai, Mohamed R Aniba, Allen W Zhang, Karey Shumansky, Celia Siu, Adrian Wan, Melissa K McConechy, Hector Li-Chang, Alicia Tone, Diane Provencher, Manon de Ladurantaye, Hubert Fleury, Aikou Okamoto, Satoshi Yanagida, Nozomu Yanaihara, Misato Saito, Andrew J Mungall, Richard Moore, Marco A Marra, C Blake Gilks, Anne-Marie Mes-Masson, Jessica N McAlpine, Samuel Aparicio, David G Huntsman, and Sohrab P Shah. Genomic consequences of aberrant DNA repair mechanisms stratify ovarian cancer histotypes. *Nature Genetics*, 4 2017.

[4] Roger Collier. Half of Canadians can expect cancer diagnosis during lifetime. *CMAJ :*

Canadian Medical Association journal = journal de l'Association medicale canadienne, 189(27):E920, 7 2017.

[5] Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 1 2012.

[6] Nicholas A Saunders, Fiona Simpson, Erik W Thompson, Michelle M Hill, Liliana Endo-Munoz, Graham Leggatt, Rodney F Minchin, and Alexander Guminski. Role of intratumoural heterogeneity in cancer drug resistance: molecular and clinical perspectives. *EMBO molecular medicine*, 4(8):675–84, 8 2012.

[7] Christian Frantz, Kathleen M Stewart, and Valerie M Weaver. The extracellular matrix at a glance. *Journal of cell science*, 123(Pt 24):4195–200, 12 2010.

[8] Padmanee Sharma and James P Allison. Immune checkpoint targeting in cancer therapy: toward combination strategies with curative potential. *Cell*, 161(2):205–14, 4 2015.

[9] Stephanie C Casey, Amedeo Amedei, Katia Aquilano, Asfar S Azmi, Fabian Benencia, Dipita Bhakta, Alan E Bilsland, Chandra S Boosani, Sophie Chen, Maria Rosa Ciriolo, Sarah Crawford, Hiromasa Fujii, Alexandros G Georgakilas, Gunjan Guha, Dorota Halicka, William G Helferich, Petr Heneberg, Kanya Honoki, W Nicol Keith, Sid P Kerkar, Sulma I Mohammed, Elena Niccolai, Somaira Nowsheen, H P Vasantha Rupasinghe, Abbas Samadi, Neetu Singh, Wamidh H Talib, Vasundara Venkateswaran, Richard L Whelan, Xujuan Yang, and Dean W Felsher. Cancer prevention and therapy through the modulation of the tumor microenvironment. *Seminars in cancer biology*, 35 Suppl:199–223, 12 2015.

[10] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2016. *CA: A Cancer Journal for Clinicians*, 66(1):7–30, 1 2016.

[11] Martine J. Piccart. Response: Re: Randomized Intergroup Trial of CisplatinPaclitaxel Versus CisplatinCyclophosphamide in Women With Advanced Epithelial Ovarian Cancer: Three-Year Results. *JNCI: Journal of the National Cancer Institute*, 92(17):1446–1447, 9 2000.

[12] Rosemary D. Cress, Yingjia S. Chen, Cyllene R. Morris, Megan Petersen, and Gary S. Leiserowitz. Characteristics of Long-Term Survivors of Epithelial Ovarian Cancer. *Obstetrics & Gynecology*, 126(3):491–497, 9 2015.

[13] Mirjana Kessler, Christina Fotopoulou, and Thomas Meyer. The molecular fingerprint of high grade serous ovarian cancer reflects its fallopian tube origin. *International journal of molecular sciences*, 14(4):6571–96, 3 2013.

[14] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, P. Glenn, A. K. Godwin, J. Gross, L. Hartmann, M. Huang, D. G. Huntsman, M. Iacocca, M. Imielinski, S. Kalloger, B. Y. Karlan, D. A. Levine, G. B. Mills, C. Morrison, D. Mutch, N. Olvera, S. Orsulic, K. Park, N. Petrelli, B. Rabeno, J. S. Rader, B. I. Sikic, K. Smith-McCune, A. K. Sood, D. Bowtell, R. Penny, J. R. Testa, K. Chang, H. H. Dinh, J. A. Drummond, G. Fowler, P. Gunaratne, A. C. Hawes, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. Santibanez, J. G. Reid, L. R. Trevino, Y.-Q. Wu, M. Wang, D. M. Muzny, D. A. Wheeler, R. A. Gibbs, G. Getz, M. S. Lawrence, K. Cibulskis, A. Y. Sivachenko, C. Sougnez, D. Voet, J. Wilkinson, T. Bloom, K. Ardlie, T. Fennell, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, D. C. Koboldt, M. D. McLellan, T. Wylie, J. Walker, M. OLaughlin, D. J. Dooling, L. Fulton, R. Abbott, N. D. Dees, Q. Zhang, C. Kandoth, M. Wendl, W. Schierding, D. Shen, C. C. Harris, H. Schmidt, J. Kalicki, K. D. Delehaunty, C. C. Fronick, R. Demeter, L. Cook, J. W. Wallis, L. Lin, V. J. Magrini, J. S. Hodges, J. M. Eldred, S. M. Smith, C. S. Pohl, F. Vandin, B. J. Raphael, G. M. Weinstock, E. R. Mardis, R. K. Wilson, M. Meyerson, W. Winckler, G. Getz, R. G. W. Verhaak, S. L. Carter, C. H. Mermel, G. Saksena, H. Nguyen, R. C. Onofrio, M. S. Lawrence, D. Hubbard, S. Gupta, A. Crenshaw, A. H. Ramos, K. Ardlie, L. Chin, A. Protopopov, Juinhua Zhang, T. M. Kim, I. Perna, Y. Xiao, H. Zhang, G. Ren, N. Sathiamoorthy, R. W. Park, E. Lee, P. J. Park, R. Kucherlapati, D. M. Absher, L. Waite, G. Sherlock, J. D. Brooks, J. Z. Li, J. Xu, R. M. Myers, P. W. Laird, L. Cope, J. G. Herman, H. Shen, D. J. Weisenberger, H. Noushmehr, F. Pan, T. Triche Jr, B. P. Berman, D. J. Van Den Berg, J. Buckley, S. B. Baylin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, L. R. Jakkula, S. Durinck, J. Han, S. Dorton, H. Marr, Y. G. Choi, V. Wang, N. J. Wang, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, D. A. Levine, N. D. Socci, Y. Liang, B. S. Taylor, N. Schultz, L. Borsu, A. E. Lash, C. Brennan, A. Viale, C. Sander, M. Ladanyi, K. A. Hoadley, S. Meng, Y. Du, Y. Shi, L. Li, Y. J. Turman, D. Zang, E. B. Helms, S. Balu, X. Zhou, J. Wu, M. D. Topal, D. N. Hayes, C. M. Perou, G. Getz, D. Voet, G. Saksena, Junihua Zhang, H. Zhang, C. J. Wu, S. Shukla, K. Cibulskis, M. S. Lawrence, A. Sivachenko, R. Jing, R. W. Park, Y. Liu, P. J. Park, M. Noble, L. Chin, H. Carter, D. Kim, R. Karchin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, S. Durinck, J. Han, J. E. Korkola, L. M. Heiser, R. J. Cho, Z. Hu, B. Parvin, T. P. Speed, J. W. Gray,

N. Schultz, E. Cerami, B. S. Taylor, A. Olshen, B. Reva, Y. Antipin, R. Shen, P. Mankoo, R. Sheridan, G. Ciriello, W. K. Chang, J. A. Bernanke, L. Borsu, D. A. Levine, M. Ladanyi, C. Sander, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Z. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, K. A. Hoadley, Y. Du, S. Balu, D. N. Hayes, C. M. Perou, M. D. Wilkerson, N. Zhang, R. Akbani, K. A. Baggerly, W. K. Yung, G. B. Mills, J. N. Weinstein, R. Penny, T. Shelton, D. Grimm, M. Hatfield, S. Morris, P. Yena, P. Rhodes, M. Sherman, J. Paulauskis, S. Millis, A. Kahn, J. M. Greene, R. Sfeir, M. A. Jensen, J. Chen, J. Whitmore, S. Alonso, J. Jordan, A. Chu, Jinghui Zhang, A. Barker, C. Compton, G. Eley, M. Ferguson, P. Fielding, D. S. Gerhard, R. Myles, C. Schaefer, K. R. Mills Shaw, J. Vaught, J. B. Vockley, P. J. Good, M. S. Guyer, B. Ozenberger, J. Peterson, and E. Thomson. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, 6 2011.

[15] Elizabeth M Swisher, Wataru Sakai, Beth Y Karlan, Kaitlyn Wurz, Nicole Urban, and Toshiyasu Taniguchi. Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance. *Cancer research*, 68(8):2581–6, 4 2008.

[16] Mansoor R. Mirza, Bradley J. Monk, Jrn Herrstedt, Amit M. Oza, Sven Mahner, Andrs Redondo, Michel Fabbro, Jonathan A. Ledermann, Domenica Lorusso, Ignace Vergote, Noa E. Ben-Baruch, Christian Marth, Radosaw Mądry, Ren D. Christensen, Jonathan S. Berek, Anne Dørum, Anna V. Tinker, Andreas du Bois, Antonio González-Martín, Philippe Follana, Benedict Benigno, Per Rosenberg, Lucy Gilbert, Bobbie J. Rimel, Joseph Buscema, John P. Balser, Shefali Agarwal, and Ursula A. Matulonis. Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. *New England Journal of Medicine*, 375(22):2154–2164, 12 2016.

[17] Kathleen Moore, Nicoletta Colombo, Giovanni Scambia, Byoung-Gie Kim, Ana Oaknin, Michael Friedlander, Alla Lisyanskaya, Anne Floquet, Alexandra Leary, Gabe S. Sonke, Charlie Gourley, Susana Banerjee, Amit Oza, Antonio González-Martín, Carol Aghajanian, William Bradley, Cara Mathews, Joyce Liu, Elizabeth S. Lowe, Ralph Bloomfield, and Paul DiSilvestro. Maintenance Olaparib in Patients with Newly Diagnosed Advanced Ovarian Cancer. *New England Journal of Medicine*, page NEJMoa1810858, 10 2018.

[18] Robert E. (Robert Edward) Scully, Robert H. (Robert Henry) Young, Philip B. Clement, Armed Forces Institute of Pathology (U.S.), and Universities Associated for Research and Education in Pathology. *Tumors of the ovary, maldeveloped gonads, fallopian tube, and broad ligament.* Armed Forces Institute of Pathology, 1998.

[19] Joseph W. Carlson, Alexander Miron, Elke A. Jarboe, Mana M. Parast, Michelle S. Hirsch, Yonghee Lee, Michael G. Muto, David Kindelberger, and Christopher P. Crum. Serous Tubal Intraepithelial Carcinoma: Its Potential Role in Primary Peritoneal Serous Carcinoma and Serous Cancer Prevention. *Journal of Clinical Oncology*, 26(25):4160–4165, 9 2008.

[20] Michael H. Roh, David Kindelberger, and Christopher P. Crum. Serous Tubal Intraepithelial Carcinoma and the Dominant Ovarian Mass. *The American Journal of Surgical Pathology*, 33(3):376–383, 3 2009.

[21] S. Intidhar Labidi-Galy, Eniko Papp, Dorothy Hallberg, Noushin Niknafs, Vilmos Adleff, Michael Noe, Rohit Bhattacharya, Marian Novak, Sin Jones, Jillian Phallen, Carolyn A. Hruban, Michelle S. Hirsch, Douglas I. Lin, Lauren Schwartz, Cecile L. Maire, Jean-Christophe Tille, Michaela Bowden, Ayse Ayhan, Laura D. Wood, Robert B. Scharpf, Robert Kurman, Tian-Li Wang, Ie-Ming Shih, Rachel Karchin, Ronny Drapkin, and Victor E. Velculescu. High grade serous ovarian carcinomas originate in the fallopian tube. *Nature Communications*, 8(1):1093, 12 2017.

[22] Ahmed Ashour Ahmed, Dariush Etemadmoghadam, Jillian Temple, Andy G Lynch, Mohamed Riad, Raghwa Sharma, Colin Stewart, Sian Fereday, Carlos Caldas, Anna Defazio, David Bowtell, and James D Brenton. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *The Journal of pathology*, 221(1):49–56, 5 2010.

[23] Geoff Macintyre, Teodora E. Goranova, Dilrini De Silva, Darren Ennis, Anna M. Piskorz, Matthew Eldridge, Daoud Sie, Liz-Anne Lewsley, Aishah Hanif, Cheryl Wilson, Suzanne Dowson, Rosalind M. Glasspool, Michelle Lockley, Elly Brockbank, Ana Montes, Axel Walther, Sudha Sundar, Richard Edmondson, Geoff D. Hall, Andrew Clamp, Charlie Gourley, Marcia Hall, Christina Fotopoulou, Hani Gabra, James Paul, Anna Supernat, David Millan, Aoisha Hoyle, Gareth Bryson, Craig Nourse, Laura Mincarelli, Luis Navarro Sanchez, Bauke Ylstra, Mercedes Jimenez-Linan, Luiza Moore, Oliver Hofmann, Florian Markowetz, Iain A. McNeish, and James D. Brenton. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, 50(9):1262–1270, 9 2018.

[24] Peter J. Campbell, Shinichi Yachida, Laura J. Mudie, Philip J. Stephens, Erin D. Pleasance, Lucy A. Stebbings, Laura A. Morsberger, Calli Latimer, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena A. Nik-Zainal, Catherine Leroy, Mingming Jia,

Andrew Menzies, Adam P. Butler, Jon W. Teague, Constance A. Griffin, John Burton, Harold Swerdlow, Michael A. Quail, Michael R. Stratton, Christine Iacobuzio-Donahue, and P. Andrew Futreal. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–1113, 10 2010.

[25] Tatiana Popova, Elodie Manié, Valentina Boeva, Aude Battistella, Oumou Goundiam, Nicholas K. Smith, Christopher R. Mueller, Virginie Raynal, Odette Mariani, Xavier Sastre-Garau, and Marc-Henri Stern. Ovarian Cancers Harboring Inactivating Mutations in CDK12 Display a Distinct Genomic Instability Pattern Characterized by Large Tandem Duplications. *Cancer Research*, 76(7):1882–1891, 4 2016.

[26] Tyler Funnell, Allen Zhang, Yu-Jia Shiah, Diljot Grewal, Robert Lesurf, Steven McKinney, Ali Bashashati, Yi Kan Wang, Paul Boutros, and Sohrab Shah. Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers. *bioRxiv*, 267500, 2 2018.

[27] Ali Bashashati, Gavin Ha, Alicia Tone, Jiarui Ding, Leah M Prentice, Andrew Roth, Jamie Rosner, Karey Shumansky, Steve Kalloger, Janine Senz, Winnie Yang, Melissa McConechy, Nataliya Melnyk, Michael Anglesio, Margaret T Y Luk, Kane Tse, Thomas Zeng, Richard Moore, Yongjun Zhao, Marco A Marra, Blake Gilks, Stephen Yip, David G Huntsman, Jessica N McAlpine, and Sohrab P Shah. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *The Journal of pathology*, 231(1):21–34, 9 2013.

[28] Roland F. Schwarz, Charlotte K. Y. Ng, Susanna L. Cooke, Scott Newman, Jillian Temple, Anna M. Piskorz, Davina Gale, Karen Sayal, Muhammed Murtaza, Peter J. Baldwin, Nitzan Rosenfeld, Helena M. Earl, Evis Sala, Mercedes Jimenez-Linan, Christine A. Parkinson, Florian Markowetz, James D. Brenton, PC Nowell, DL Dexter, HM Kowalski, BA Blazar, Z Fligiel, R Vogel, L Khalique, A Ayhan, ME Weale, IJ Jacobs, SJ Ramus, L Khalique, A Ayhan, JC Whittaker, N Singh, IJ Jacobs, SP Shah, RD Morin, J Khattra, L Prentice, T Pugh, N Navin, A Krasnitz, L Rodgers, K Cook, J Meth, PJ Campbell, S Yachida, LJ Mudie, PJ Stephens, ED Pleasance, N Navin, J Kendall, J Troge, P Andrews, L Rodgers, A Marusyk, V Almendro, K Polyak, JS Vermaat, IJ Nijman, MJ Koudijs, FL Gerritse, SJ Scherer, X Wu, PA Northcott, A Dubuc, AJ Dupuy, DJH Shih, M Gerlinger, AJ Rowan, S Horswell, J Larkin, D Endesfelder, SP Shah, A Roth, R Goya, A Oloumi, G Ha, EC de Bruin, N McGranahan, R Mitter, M Salm, DC Wedge, S Nik-Zainal, P Van Loo, DC Wedge, LB Alexandrov, CD Greenman, LMF Merlo, JW Pepper, BJ Reid,

CC Maley, M Greaves, CC Maley, K Anderson, C Lutz, FW van Delft, CM Bateman, Y Guo, DA Landau, SL Carter, P Stojanov, A McKenna, K Stevenson, AA Ahmed, D Etemadmoghadam, J Temple, AG Lynch, M Riad, KL Gorringe, S Jacobs, ER Thompson, A Sridhar, W Qiu, N Sangha, R Wu, R Kuick, S Powers, D Mu, SL Carter, K Cibulskis, E Helman, A McKenna, H Shen, A Bashashati, G Ha, A Tone, J Ding, LM Prentice, J Zhang, Y Shi, E Lalonde, L Li, L Cavallone, M Hoogstraat, MS de Pagter, GA Cirkel, MJ van Roosmalen, TT Harkins, SL Cooke, CKY Ng, N Melnyk, MJ Garcia, T Hardcastle, SL Cooke, JD Brenton, ZC Wang, NJ Birkbak, AC Culhane, R Drapkin, A Fatima, PA Cowin, J George, S Fereday, E Loehrer, P Van Loo, RF Schwarz, A Trinh, B Sipos, JD Brenton, N Goldman, CD Greenman, G Bignell, A Butler, S Edkins, J Hinton, CKY Ng, SL Cooke, K Howe, S Newman, J Xian, H Li, R Durbin, A Untergasser, I Cutcutache, T Koressaar, J Ye, BC Faircloth, T Forshew, M Murtaza, C Parkinson, D Gale, DWY Tsui, JT Robinson, H Thorvaldsdóttir, W Winckler, M Guttman, ES Lander, KM Archibald, H Kulbe, J Kwong, P Chakravarty, J Temple, E Sala, MY Kataoka, AN Priest, AB Gill, MA McLean, DJ McBride, D Etemadmoghadam, SL Cooke, K Alsop, J George, D Bryant, V Moulton, M Castellarin, K Milne, T Zeng, K Tse, and M Mayo. Spatial and Temporal Heterogeneity in High-Grade Serous Ovarian Cancer: A Phylogenetic Analysis. *PLOS Medicine*, 12(2):e1001789, 2 2015.

[29] Andrew McPherson, Andrew Roth, Emma Laks, Tehmina Masud, Ali Bashashati, Allen W Zhang, Gavin Ha, Justina Biele, Damian Yap, Adrian Wan, Leah M Prentice, Jaswinder Khattra, Maia A Smith, Cydney B Nielsen, Sarah C Mullaly, Steve Kalloger, Anthony Karnezis, Karey Shumansky, Celia Siu, Jamie Rosner, Hector Li Chan, Julie Ho, Nataliya Melnyk, Janine Senz, Winnie Yang, Richard Moore, Andrew J Mungall, Marco A Marra, Alexandre Bouchard-Côté, C Blake Gilks, David G Huntsman, Jessica N McAlpine, Samuel Aparicio, and Sohrab P Shah. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*, 48(7):758–767, 5 2016.

[30] Yuliya Klymenko, Jeffrey Johnson, Brandi Bos, Rachel Lombard, Leigh Campbell, Elizabeth Loughran, and M. Sharon Stack. Heterogeneous Cadherin Expression and Multicellular Aggregate Dynamics in Ovarian Cancer Dissemination. *Neoplasia*, 19(7):549–563, 7 2017.

[31] Sara Al Habyan, Christina Kalos, Joseph Szymborski, and Luke McCaffrey. Multicellular detachment generates metastatic spheroids during intra-abdominal dissemination in epithelial ovarian cancer. *Oncogene*, 37(37):5127–5135, 9 2018.

[32] Daniela F Quail and Johanna A Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine*, 19(11):1423–37, 11 2013.

[33] Costas A Lyssiotis and Alec C Kimmelman. Metabolic Interactions in the Tumor Microenvironment. *Trends in cell biology*, 27(11):863–875, 11 2017.

[34] Brad H Nelson, Philip D Greenberg, and Hans Schreiber. New insights into tumor immunity revealed by the unique genetic and genomic aspects of ovarian cancer. *Current Opinion in Immunology*, 33:93–100, 2015.

[35] Jean M. Hansen, Robert L. Coleman, and Anil K. Sood. Targeting the tumour microenvironment in ovarian cancer. *European Journal of Cancer*, 56:131–143, 3 2016.

[36] Lin Zhang, Jose R Conejo-Garcia, Dionyssios Katsaros, Phyllis A Gimotty, Marco Massobrio, Giorgia Regnani, Antonis Makrigiannakis, Heidi Gray, Katia Schlienger, Michael N Liebman, Stephen C Rubin, and George Coukos. Intratumoral T cells, recurrence, and survival in epithelial ovarian cancer. *The New England journal of medicine*, 348(3):203–13, 1 2003.

[37] Sine Hadrup, Marco Donia, and Per Thor Straten. Effector CD4 and CD8 T cells and their role in the tumor microenvironment. *Cancer microenvironment : official journal of the International Cancer Microenvironment Society*, 6(2):123–33, 8 2013.

[38] Julie S. Nielsen, Rob A. Sahota, Katy Milne, Sara E. Kost, Nancy J. Nesslinger, Peter H. Watson, and Brad H. Nelson. CD20+ Tumor-Infiltrating Lymphocytes Have an Atypical CD27- Memory Phenotype and Together with CD8+ T Cells Promote Favorable Prognosis in Ovarian Cancer. *Clinical Cancer Research*, 18(12), 2012.

[39] Darin A Wick, John R Webb, Julie S Nielsen, Spencer D Martin, David R Kroeger, Katy Milne, Mauro Castellarin, Kwame Twumasi-Boateng, Peter H Watson, Rob A Holt, and Brad H Nelson. Surveillance of the tumor mutanome by T cells during progression from primary to recurrent ovarian cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 20(5):1125–34, 3 2014.

[40] J. R. Webb, K. Milne, P. Watson, R. J. deLeeuw, and B. H. Nelson. Tumor-Infiltrating Lymphocytes Expressing the Tissue Resident Memory Marker CD103 Are Associated with Increased Survival in High-Grade Serous Ovarian Cancer. *Clinical Cancer Research*, 20(2):434–444, 1 2014.

[41] David R Kroeger, Katy Milne, and Brad H Nelson. Tumor-Infiltrating Plasma Cells Are Associated with Tertiary Lymphoid Structures, Cytolytic T-Cell Responses, and Superior Prognosis in Ovarian Cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 22(12):3005–15, 6 2016.

[42] Brad H Nelson. CD20+ B cells: the other tumor-infiltrating lymphocytes. *Journal of immunology (Baltimore, Md. : 1950)*, 185(9):4977–82, 11 2010.

[43] Tyler J Curiel, George Coukos, Linhua Zou, Xavier Alvarez, Pui Cheng, Peter Mottram, Melina Evdemon-Hogan, Jose R Conejo-Garcia, Lin Zhang, Matthew Burow, Yun Zhu, Shuang Wei, Ilona Kryczek, Ben Daniel, Alan Gordon, Leann Myers, Andrew Lackner, Mary L Disis, Keith L Knutson, Lieping Chen, and Weiping Zou. Specific recruitment of regulatory T cells in ovarian carcinoma fosters immune privilege and predicts reduced survival. *Nature Medicine*, 10(9):942–949, 9 2004.

[44] Claudia C. Preston, Matthew J. Maurer, Ann L. Oberg, Daniel W. Visscher, Kimberly R. Kalli, Lynn C. Hartmann, Ellen L. Goode, and Keith L. Knutson. The Ratios of CD8+ T Cells to CD4+CD25+ FOXP3+ and FOXP3- T Cells Correlate with Poor Clinical Outcome in Human Serous Ovarian Cancer. *PLoS ONE*, 8(11):e80063, 11 2013.

[45] Y Jiang, Y Li, and B Zhu. T-cell exhaustion in the tumor microenvironment. *Cell Death & Disease*, 6(6):e1792–e1792, 6 2015.

[46] Fenne L Komdeur, Maartje C A Wouters, Hagma H Workel, Aline M Tijans, Anouk L J Terwindt, Kim L Brunekreeft, Annechien Plat, Harry G Klip, Florine A Eggink, Ninke Leffers, Wijnand Helfrich, Douwe F Samplonius, Edwin Bremer, G Bea A Wisman, Toos Daemen, Evelien W Duiker, Harry Hollema, Hans W Nijman, and Marco de Bruyn. CD103+ intraepithelial T cells in high-grade serous ovarian cancer are phenotypically diverse TCR$\alpha\beta$+ CD8$\alpha\beta$+ T cells that can be targeted for cancer immunotherapy. *Oncotarget*, 7(46):75130–75144, 11 2016.

[47] John R. Webb, Katy Milne, David R. Kroeger, and Brad H. Nelson. PD-L1 expression is associated with tumor-infiltrating T cells and favorable prognosis in high-grade serous ovarian cancer. *Gynecologic Oncology*, 141(2):293–302, 5 2016.

[48] Dung T. Le, Jennifer N. Uram, Hao Wang, Bjarne R. Bartlett, Holly Kemberling, Aleksandra D. Eyring, Andrew D. Skora, Brandon S. Luber, Nilofer S. Azad, Dan Laheru, Barbara Biedrzycki, Ross C. Donehower, Atif Zaheer, George A. Fisher, Todd S. Crocenzi, James J.

Lee, Steven M. Duffy, Richard M. Goldberg, Albert de la Chapelle, Minori Koshiji, Feriyl Bhaijee, Thomas Huebner, Ralph H. Hruban, Laura D. Wood, Nathan Cuka, Drew M. Pardoll, Nickolas Papadopoulos, Kenneth W. Kinzler, Shibin Zhou, Toby C. Cornish, Janis M. Taube, Robert A. Anders, James R. Eshleman, Bert Vogelstein, and Luis A. Diaz. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, 372(26):2509–2520, 6 2015.

[49] Krisztian Homicsko and George Coukos. Targeting Programmed Cell Death 1 in Ovarian Cancer. *Journal of Clinical Oncology*, 33(34):3987–3989, 12 2015.

[50] Stphanie L. Gaillard, Angeles A. Secord, and Bradley Monk. The role of immune checkpoint inhibition in the treatment of ovarian cancer. *Gynecologic Oncology Research and Practice*, 3(1):11, 12 2016.

[51] Thomas F Gajewski, Hans Schreiber, and Yang-Xin Fu. Innate and adaptive immune cells in the tumor microenvironment. *Nature Immunology*, 14(10):1014–1022, 9 2013.

[52] Nicholas McGranahan, Rachel Rosenthal, Crispin T. Hiley, Andrew J. Rowan, Thomas B.K. Watkins, Gareth A. Wilson, Nicolai J. Birkbak, Selvaraju Veeriah, Peter Van Loo, Javier Herrero, Charles Swanton, Charles Swanton, Mariam Jamal-Hanjani, Selvaraju Veeriah, Seema Shafi, Justyna Czyzewska-Khan, Diana Johnson, Joanne Laycock, Leticia Bosshard-Carter, Rachel Rosenthal, Pat Gorman, Robert E. Hynds, Gareth Wilson, Nicolai J. Birkbak, Thomas B.K. Watkins, Nicholas McGranahan, Stuart Horswell, Richard Mitter, Mickael Escudero, Aengus Stewart, Peter Van Loo, Andrew Rowan, Hang Xu, Samra Turajlic, Crispin Hiley, Christopher Abbosh, Jacki Goldman, Richard Kevin Stone, Tamara Denner, Nik Matthews, Greg Elgar, Sophia Ward, Marta Costa, Sharmin Begum, Ben Phillimore, Tim Chambers, Emma Nye, Sofia Graca, Maise Al Bakir, Kroopa Joshi, Andrew Furness, Assma Ben Aissa, Yien Ning Sophia Wong, Andy Georgiou, Sergio Quezada, John A. Hartley, Helen L. Lowe, Javier Herrero, David Lawrence, Martin Hayward, Nikolaos Panagiotopoulos, Shyam Kolvekar, Mary Falzon, Elaine Borg, Teresa Marafioti, Celia Simeon, Gemma Hector, Amy Smith, Marie Aranda, Marco Novelli, Dahmane Oukrif, Sam M. Janes, Ricky Thakrar, Martin Forster, Tanya Ahmad, Siow Ming Lee, Dionysis Papadatos-Pastos, Dawn Carnell, Ruheena Mendes, Jeremy George, Neal Navani, Asia Ahmed, Magali Taylor, Junaid Choudhary, Yvonne Summers, Raffaele Califano, Paul Taylor, Rajesh Shah, Piotr Krysiak, Kendadai Rammohan, Eustace Fontaine, Richard Booton, Matthew Evison, Phil Crosbie, Stuart Moss, Faiza Idries, Leena Joseph, Paul Bishop, Anshuman Chaturved, Anne Marie Quinn, Helen Doran, Angela Leek, Phil Harrison, Katrina

Moore, Rachael Waddington, Juliette Novasio, Fiona Blackhall, Jane Rogan, Elaine Smith, Caroline Dive, Jonathan Tugwood, Ged Brady, Dominic G. Rothwell, Francesca Chemi, Jackie Pierce, Sakshi Gulati, Babu Naidu, Gerald Langman, Simon Trotter, Mary Bellamy, Hollie Bancroft, Amy Kerr, Salma Kadiri, Joanne Webb, Gary Middleton, Madava Djearaman, Dean Fennell, Jacqui A. Shaw, John Le Quesne, David Moore, Apostolos Nakas, Sridhar Rathinam, William Monteiro, Hilary Marshall, Louise Nelson, Jonathan Bennett, Joan Riley, Lindsay Primrose, Luke Martinson, Girija Anand, Sajid Khan, Anita Amadi, Marianne Nicolson, Keith Kerr, Shirley Palmer, Hardy Remmen, Joy Miller, Keith Buchan, Mahendran Chetty, Lesley Gomersall, Jason Lester, Alison Edwards, Fiona Morgan, Haydn Adams, Helen Davies, Malgorzata Kornaszewska, Richard Attanoos, Sara Lock, Azmina Verjee, Mairead MacKenzie, Maggie Wilcox, Harriet Bell, Allan Hackshaw, Yenting Ngai, Sean Smith, Nicole Gower, Christian Ottensmeier, Serena Chee, Benjamin Johnson, Aiman Alzetani, Emily Shaw, Eric Lim, Paulo De Sousa, Monica Tavares Barbosa, Alex Bowman, Simon Jordan, Alexandra Rice, Hilgardt Raubenheimer, Chiara Proli, Maria Elena Cufari, John Carlo Ronquillo, Angela Kwayie, Harshil Bhayani, Morag Hamilton, Yusura Bakar, Natalie Mensah, Lyn Ambrose, Anand Devaraj, Silviu Buderi, Jonathan Finch, Leire Azcarate, Hema Chavan, Sophie Green, Hillaria Mashinga, Andrew G. Nicholson, Kelvin Lau, Michael Sheaff, Peter Schmid, John Conibear, Veni Ezhil, Babikir Ismail, Melanie Irvin-sellers, Vineet Prakash, Peter Russell, Teresa Light, Tracey Horey, Sarah Danson, Jonathan Bury, John Edwards, Jennifer Hill, Sue Matthews, Yota Kitsanta, Kim Suvarna, Patricia Fisher, Allah Dino Keerio, Michael Shackcloth, John Gosney, Pieter Postmus, Sarah Feeney, Julius Asante-Siaw, Hugo J.W.L. Aerts, Stefan Dentro, and Christophe Dessimoz. Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution. *Cell*, 171(6):1259–1271, 11 2017.

[53] Yulei Chen and Xiaobo Zhang. Pivotal regulators of tissue homeostasis and cancer: macrophages. *Experimental Hematology & Oncology*, 6(1):23, 12 2017.

[54] Florian Finkernagel, Silke Reinartz, Sonja Lieber, Till Adhikary, Annika Wortmann, Nathalie Hoffmann, Tim Bieringer, Andrea Nist, Thorsten Stiewe, Julia M Jansen, Uwe Wagner, Sabine Müller-Brüsselbach, and Rolf Müller. The transcriptional signature of human ovarian carcinoma macrophages is associated with extracellular matrix reorganization. *Oncotarget*, 7(46):75339–75352, 11 2016.

[55] Till Adhikary, Annika Wortmann, Florian Finkernagel, Sonja Lieber, Andrea Nist, Thorsten Stiewe, Uwe Wagner, Sabine Müller-Brüsselbach, Silke Reinartz, and Rolf Müller. Interferon

signaling in ascites-associated macrophages is linked to a favorable clinical outcome in a subgroup of ovarian carcinoma patients. *BMC Genomics*, 18(1):243, 12 2017.

[56] Samuel F. Bakhoum, Bryan Ngo, Ashley M. Laughney, Julie-Ann Cavallo, Charles J. Murphy, Peter Ly, Pragya Shah, Roshan K. Sriram, Thomas B. K. Watkins, Neil K. Taunk, Mercedes Duran, Chantal Pauli, Christine Shaw, Kalyani Chadalavada, Vinagolu K. Rajasekhar, Giulio Genovese, Subramanian Venkatesan, Nicolai J. Birkbak, Nicholas McGranahan, Mark Lundquist, Quincey LaPlant, John H. Healey, Olivier Elemento, Christine H. Chung, Nancy Y. Lee, Marcin Imielenski, Gouri Nanjangud, Dana Peer, Don W. Cleveland, Simon N. Powell, Jan Lammerding, Charles Swanton, and Lewis C. Cantley. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature*, 553(7689):467–472, 1 2018.

[57] Seng-Ryong Woo, Leticia Corrales, and Thomas F. Gajewski. Innate Immune Recognition of Cancer. *Annual Review of Immunology*, 33(1):445–474, 3 2015.

[58] Fei Xing, Jamila Saidou, and Kounosuke Watabe. Cancer associated fibroblasts (CAFs) in tumor microenvironment. *Frontiers in bioscience (Landmark edition)*, 15:166–79, 1 2010.

[59] Weimin Wang, Ilona Kryczek, Lubomr Dostál, Heng Lin, Lijun Tan, Lili Zhao, Fujia Lu, Shuang Wei, Tomasz Maj, Dongjun Peng, Gong He, Linda Vatan, Wojciech Szeliga, Rork Kuick, Jan Kotarski, Rafa Tarkowski, Yali Dou, Ramandeep Rattan, Adnan Munkarah, J Rebecca Liu, and Weiping Zou. Effector T Cells Abrogate Stroma-Mediated Chemoresistance in Ovarian Cancer. *Cell*, 165(5):1092–105, 5 2016.

[60] Naoyo Nishida, Hirohisa Yano, Takashi Nishida, Toshiharu Kamura, and Masamichi Kojiro. Angiogenesis in cancer. *Vascular health and risk management*, 2(3):213–9, 2006.

[61] Andrew C Dudley. Tumor endothelial cells. *Cold Spring Harbor perspectives in medicine*, 2(3):a006536, 3 2012.

[62] Stuart C. Williamson, Robert L. Metcalf, Francesca Trapani, Sumitra Mohan, Jenny Antonello, Benjamin Abbott, Hui Sun Leong, Christopher P. E. Chester, Nicole Simms, Radoslaw Polanski, Daisuke Nonaka, Lynsey Priest, Alberto Fusi, Fredrika Carlsson, Anders Carlsson, Mary J. C. Hendrix, Richard E. B. Seftor, Elisabeth A. Seftor, Dominic G. Rothwell, Andrew Hughes, James Hicks, Crispin Miller, Peter Kuhn, Ged Brady, Kathryn L. Simpson, Fiona H. Blackhall, and Caroline Dive. Vasculogenic mimicry in small cell lung cancer. *Nature Communications*, 7:13322, 11 2016.

[63] K L Eales, K E R Hollinshead, and D A Tennant. Hypoxia and metabolic adaptation of cancer cells. *Oncogenesis*, 5(1):e190, 1 2016.

[64] Ying Zhang and Hildegund C. J. Ertl. Starved and Asphyxiated: How Can CD8+ T Cells within a Tumor Microenvironment Prevent Tumor Progression. *Frontiers in Immunology*, 7:32, 2 2016.

[65] Dong Chul Lee, Hyun Ahm Sohn, Zee-Yong Park, Sangho Oh, Yun Kyung Kang, Kyoung-Min Lee, Minho Kang, Ye Jin Jang, Suk-Jin Yang, Young Ki Hong, Hanmi Noh, Jung-Ae Kim, Dong Joon Kim, Kwang-Hee Bae, Dong Min Kim, Sang J Chung, Hyang Sook Yoo, Dae-Yeul Yu, Kyung Chan Park, and Young Il Yeom. A lactate-induced response to hypoxia. *Cell*, 161(3):595–609, 4 2015.

[66] Emma Williams, Stewart Martin, Robert Moss, Lindy Durrant, and Suha Deen. Co-expression of VEGF and CA9 in ovarian high-grade serous carcinoma and relationship to survival. *Virchows Archiv*, 461(1):33–39, 7 2012.

[67] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 1 2016.

[68] Yong Wang and Nicholas E Navin. Advances and applications of single-cell sequencing technologies. *Molecular cell*, 58(4):598–609, 5 2015.

[69] Noemi Andor, Trevor A Graham, Marnix Jansen, Li C Xia, C Athena Aktipis, Claudia Petritsch, Hanlee P Ji, and Carlo C Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, 11 2015.

[70] Luc G T Morris, Nadeem Riaz, Alexis Desrichard, Yasin Şenbabaoğlu, A Ari Hakimi, Vladimir Makarov, Jorge S Reis-Filho, and Timothy A Chan. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, 7(9):10051–63, 3 2016.

[71] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 3 2014.

[72] Samuel Aparicio and Carlos Caldas. The Implications of Clonal Genome Evolution for Cancer Medicine. *New England Journal of Medicine*, 368(9):842–851, 2 2013.

[73] Christopher A Miller, Brian S White, Nathan D Dees, Malachi Griffith, John S Welch, Obi L Griffith, Ravi Vij, Michael H Tomasson, Timothy A Graubert, Matthew J Walter, Matthew J Ellis, William Schierding, John F DiPersio, Timothy J Ley, Elaine R Mardis, Richard K Wilson, and Li Ding. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology*, 10(8):e1003665, 8 2014.

[74] Salem Malikic, Andrew W. McPherson, Nilgun Donmez, and Cenk S. Sahinalp. Clonality inference in multiple tumor samples using phylogeny. *Bioinformatics*, 31(9):1349–1356, 5 2015.

[75] Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. *Genome Biology*, 16(1):91, 12 2015.

[76] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2 2015.

[77] Sohrab Salehi, Adi Steif, Andrew Roth, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. ddClone: joint statistical inference of clonal populations from single cell and bulk tumour sequencing data. *Genome Biology*, 18(1):44, 12 2017.

[78] E L Reinherz and S F Schlossman. The differentiation and function of human T lymphocytes. *Cell*, 19(4):821–7, 4 1980.

[79] Kathrin Pieper, Bodo Grimbacher, and Hermann Eibel. B-cell biology and development. *Journal of Allergy and Clinical Immunology*, 131(4):959–971, 4 2013.

[80] Craig H Bassing, Wojciech Swat, and Frederick W Alt. The Mechanism and Regulation of Chromosomal V(D)J Recombination. *Cell*, 109(2):S45–S55, 4 2002.

[81] Heinz Jacobs and Linda Bross. Towards an understanding of somatic hypermutation. *Current Opinion in Immunology*, 13(2):208–218, 4 2001.

[82] Daniel J Laydon, Charles R M Bangham, and Becca Asquith. Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1675), 8 2015.

[83] Ren L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome research*, 21(5):790–7, 5 2011.

[84] Daniel J Woodsworth, Mauro Castellarin, and Robert A Holt. Sequence analysis of T-cell repertoires in health and disease. *Genome medicine*, 5(10):98, 2013.

[85] Elisa Rosati, C Marie Dowds, Evaggelia Liaskou, Eva Kristine Klemsdal Henriksen, Tom H Karlsen, and Andre Franke. Overview of methodologies for T-cell receptor repertoire analysis. *BMC biotechnology*, 17(1):61, 7 2017.

[86] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. MiXCR: software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5):380–381, 4 2015.

[87] Yaxuan Yu, Rhodri Ceredig, and Cathal Seoighe. LymAnalyzer: a tool for comprehensive analysis of next generation sequencing data of T cell receptors and immunoglobulins. *Nucleic Acids Research*, 44(4):e31–e31, 2 2016.

[88] Leon Kuchenbecker, Mikalai Nienen, Jochen Hecht, Avidan U. Neumann, Nina Babel, Knut Reinert, and Peter N. Robinson. IMSEQa fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics*, 31(18):2963–2971, 9 2015.

[89] Duncan K. Ralph and Frederick A. Matsen. Likelihood-Based Inference of B Cell Clonal Families. *PLOS Computational Biology*, 12(10):e1005086, 10 2016.

[90] Duncan K. Ralph, Frederick A. Matsen, BJ Huntly, R Rance, GS Vassiliou, and GA Follows. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLOS Computational Biology*, 12(1):e1004409, 1 2016.

[91] Bryan Howie, Anna M Sherwood, Ashley D Berkebile, Jan Berka, Ryan O Emerson, David W Williamson, Ilan Kirsch, Marissa Vignali, Mark J Rieder, Christopher S Carlson, and Harlan S Robins. High-throughput pairing of T cell receptor α and β sequences. *Science translational medicine*, 7(301):301ra131, 8 2015.

[92] Andrew Roth, Andrew McPherson, Emma Laks, Justina Biele, Damian Yap, Adrian Wan, Maia A Smith, Cydney B Nielsen, Jessica N McAlpine, Samuel Aparicio, Alexandre

Bouchard-Côté, and Sohrab P Shah. Clonal genotype and population structure inference from single-cell tumor sequencing. *Nature methods*, 13(7):573–6, 7 2016.

[93] Anna K. Casasent, Aislyn Schalck, Ruli Gao, Emi Sei, Annalyssa Long, William Pangburn, Tod Casasent, Funda Meric-Bernstam, Mary E. Edgerton, and Nicholas E. Navin. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell*, 172(1-2):205–217, 1 2018.

[94] Charissa Kim, Ruli Gao, Emi Sei, Rachel Brandt, Johan Hartman, Thomas Hatschek, Nicola Crosetto, Theodoros Foukakis, and Nicholas E. Navin. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell*, 173(4):879–893, 5 2018.

[95] Charles Gawad, Winston Koh, and Stephen R. Quake. Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, 17(3):175–188, 3 2016.

[96] Konstantin A. Blagodatskikh, Vladimir M. Kramarov, Ekaterina V. Barsova, Alexey V. Garkovenko, Dmitriy S. Shcherbo, Andrew A. Shelenkov, Vera V. Ustinova, Maria R. Tokarenko, Simon C. Baker, Tatiana V. Kramarova, and Konstantin B. Ignatov. Improved DOP-PCR (iDOP-PCR): A robust and simple WGA method for efficient amplification of low copy number genomic DNA. *PLOS ONE*, 12(9):e0184507, 9 2017.

[97] Hans Zahn, Adi Steif, Emma Laks, Peter Eirew, Michael VanInsberghe, Sohrab P Shah, Samuel Aparicio, and Carl L Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods*, 14(2):167–173, 2 2017.

[98] Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Adi Steif, Jazmine Brimhall, Justina Biele, Beixi Wang, Tehmina Masud, Diljot Grewal, Cydney Nielsen, Samantha Leung, Viktoria Bojilova, Maia Smith, Oleg Golovko, Steven Poon, Peter Eirew, Farhia Kabeer, Teresa Ruiz de Algara, So Ra Lee, M. Jafar Taghiyar, Curtis Huebner, Jessica Ngo, Tim Chan, Spencer Vatrt-Watts, Pascale Walters, Nafis Abrar, Sophia Chan, Matt Wiens, Lauren Martin, R. Wilder Scott, Michael T. Underhill, Elizabeth Chavez, Christian Steidl, Daniel Da Costa, Yusanne Ma, Robin J. N. Coope, Richard Corbett, Stephen Pleasance, Richard Moore, Andy J. Mungall, CRUK IMAXT Consortium, Marco A. Marra, Carl Hansen, Sohrab Shah, and Samuel Aparicio. Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. *bioRxiv*, page 411058, 9 2018.

[99] Single Cell CNV - 10x Genomics.

[100] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 1 2009.

[101] R. W. Tothill, A. V. Tinker, J. George, R. Brown, S. B. Fox, S. Lade, D. S. Johnson, M. K. Trivett, D. Etemadmoghadam, B. Locandro, N. Traficante, S. Fereday, J. A. Hung, Y.-E. Chiew, I. Haviv, D. Gertig, A. deFazio, D. D.L. Bowtell, and David D L Bowtell. Novel Molecular Subtypes of Serous and Endometrioid Ovarian Cancer Linked to Clinical Outcome. *Clinical Cancer Research*, 14(16):5198–5208, 8 2008.

[102] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S. N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Hye-Jung E. Chun, Robin J. N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Erin Pleasance, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J. M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, Nam H. Pho, Scott L. Carter, Steven E. Schumacher, Barbara Tabak, Bryan Hernandez, Jeff Gentry, Huy Nguyen, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Gad Getz, Stacey B. Gabriel, Matthew Meyerson, Lynda Chin, Peter J. Park, Raju Kucherlapati, Katherine A. Hoadley, J. Todd Auman, Cheng Fan, Yidi J. Turman, Yan Shi, Ling Li, Michael D. Topal, Xiaping He, Hann-Hsiang Chao, Aleix Prat, Grace O. Silva, Michael D. Iglesia, Wei Zhao, Jerry Usary, Jonathan S. Berg, Michael Adams, Jessica Booker, Junyuan Wu, Anisha Gulabani, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Simeen Malik, Swapna Mahurkar, Hui Shen, Daniel J. Weisenberger, Timothy Triche Jr, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Benjamin P. Berman, David J. Van Den Berg, Stephen B. Baylin, Peter W. Laird, Chad J. Creighton, Lawrence A. Donehower, Gad Getz, Michael Noble, Doug Voet, Gordon Saksena, Nils Gehlenborg, Daniel DiCara, Juinhua Zhang, Hailei Zhang, Chang-Jiun Wu, Spring Yingchun Liu, Michael S. Lawrence, Lihua Zou, Andrey Sivachenko, Pei Lin, Petar Stojanov, Rui Jing, Juok Cho, Raktim Sinha, Richard W. Park, Marc-Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Peter J. Park, Lynda Chin, Sheila Reynolds, Richard B. Kreisberg, Brady Bernard, Ryan Bressler, Timo Erkkila, Jake

Lin, Vesteinn Thorsson, Wei Zhang, Ilya Shmulevich, Giovanni Ciriello, Nils Weinhold, Nikolaus Schultz, Jianjiong Gao, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Marc Ladanyi, Chris Sander, Pavana Anur, Paul T. Spellman, Yiling Lu, Wenbin Liu, Roel R. G. Verhaak, Gordon B. Mills, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod D. Casasent, Chris Wakefield, Anna K. Unruh, Keith Baggerly, Kevin Coombes, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, Jingchun Zhu, Christopher C. Szeto, Gary K. Scott, Christina Yau, Evan O. Paull, Daniel Carlin, Christopher Wong, Artem Sokolov, Janita Thusberg, Sean Mooney, Sam Ng, Theodore C. Goldstein, Kyle Ellrott, Mia Grifford, Christopher Wilks, Singer Ma, Brian Craft, Chunhua Yan, Ying Hu, Daoud Meerzaman, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron D. Black, Robert E. XPATH ERROR: unknown variable "tname"., Peter White, Erik J. Zmuda, Jessica Frick, Tara M. Lichtenberg, Robin Brookens, Myra M. George, Mark A. Gerken, Hollie A. Harper, Kristen M. Leraas, Lisa J. Wise, Teresa R. Tabler, Cynthia McAllister, Thomas Barr, Melissa Hart-Kothari, Katie Tarvin, Charles Saller, George Sandusky, Colleen Mitchell, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Oleg Dolzhansky, Mikhail Abramov, Olga Voronina, Olga Potapova, Jeffrey R. Marks, Wiktoria M. Suchorska, Dawid Murawa, Witold Kycler, Matthew Ibbs, Konstanty Korski, Arkadiusz Spychała, Pawe Murawa, Jacek J. Brzeziński, Hanna Perz, Radosaw Łaźniak, Marek Teresiak, Honorata Tatka, Ewa Leporowska, Marta Bogusz-Czerniewicz, Julian Malicki, Andrzej Mackiewicz, Maciej Wiznerowicz, Xuan Van Le, Bernard Kohl, Nguyen Viet Tien, Richard Thorp, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Tran Viet The Phuong, Huynh Quyet Thang, Khurram Zaki Khan, Robert Penny, David Mallery, Erin Curley, Candace Shelton, Peggy Yena, James N. Ingle, Fergus J. Couch, Wilma L. Lingle, Tari A. King, Ana Maria Gonzalez-Angulo, Gordon B. Mills, Mary D. Dyer, Shuying Liu, Xiaolong Meng, Modesto Patangan, Frederic Waldman, Hubert Stöppler, W. Kimryn Rathmell, Leigh Thorne, Mei Huang, Lori Boice, Ashley Hill, Carl Morrison, Carmelo Gaudioso, Wiam Bshara, Kelly Daily, Sophie C. Egea, Mark D. Pegram, Carmen Gomez-Fernandez, Rajiv Dhir, Rohit Bhargava, Adam Brufsky, Craig D. Shriver, Jeffrey A. Hooke, Jamie Leigh Campbell, Richard J. Mural, Hai Hu, Stella Somiari, Caroline Larson, Brenda Deyarmin, Leonid Kvecher, Albert J. Kovatich, Matthew J. Ellis, Tari A. King, Hai Hu, Fergus J. Couch, Richard J. Mural, Thomas Stricker, Kevin White, Olufunmilayo Olopade, James N. Ingle, Chunqing Luo, Yaqin Chen, Jeffrey R. Marks, Frederic Waldman, Maciej Wiznerowicz, Ron Bose, Li-Wei Chang, Andrew H. Beck, Ana Maria Gonzalez-

Angulo, Todd Pihl, Mark Jensen, Robert Sfeir, Ari Kahn, Anna Chu, Prachi Kothiyal, Zhining Wang, Eric Snyder, Joan Pontius, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique Berton, Matthew Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter Kigonya, Shelley Alonso, Rashmi Sanbhadti, Sean Barletta, David Pot, Margi Sheth, John A. Demchok, Kenna R. Mills Shaw, Liming Yang, Greg Eley, Martin L. Ferguson, Roy W. Tarnuzzer, Jiashan Zhang, Laura A. L. Dillon, Kenneth Buetow, Peter Fielding, Bradley A. Ozenberger, Mark S. Guyer, Heidi J. Sofia, and Jacqueline D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 9 2012.

[103] The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 7 2012.

[104] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 3 2015.

[105] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell RNA sequencing. *Molecular cell*, 58(4):610–20, 5 2015.

[106] D P Bartel, M Sheng, L F Lau, and M E Greenberg. Growth factors and membrane depolarization activate distinct programs of early response gene expression: dissociation of fos and jun induction. *Genes & development*, 3(3):304–13, 3 1989.

[107] Benjamin Lacar, Sara B Linker, Baptiste N Jaeger, Suguna R Krishnaswami, Jerika J Barron, Martijn J E Kelder, Sarah L Parylak, Apu C M Paquola, Pratap Venepally, Mark Novotny, Carolyn O'Connor, Conor Fitzpatrick, Jennifer A Erwin, Jonathan Y Hsu, David Husband, Michael J McConnell, Roger Lasken, and Fred H Gage. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nature communications*, 7:11022, 2016.

[108] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science (New York, N.Y.)*, 360(6385):176–182, 3 2018.

[109] Simone Picelli, Omid R Faridani, sa K Björklund, Gsta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, 9(1):171–181, 1 2014.

[110] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 1 2017.

[111] Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific Reports*, 7(1):39921, 12 2017.

[112] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C Marioni, and Sarah A Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome biology*, 17:29, 2 2016.

[113] Marlon Stoeckius, Shiwei Zheng, Brian Houck-Loomis, Stephanie Hao, Bertrand Yeung, Peter Smibert, and Rahul Satija. Cell "hashing" with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *bioRxiv*, page 237693, 12 2017.

[114] Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv*, page 303727, 4 2018.

[115] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):284, 12 2018.

[116] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729, 7 2018.

[117] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 4 2015.

[118] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 4 2018.

[119] Brian L Hie, Bryan Bryson, and Bonnie Berger. Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv*, page 371179, 7 2018.

[120] Jong-Eun Park, Krzysztof Polanski, Kerstin Meyer, and Sarah A Teichmann. Fast Batch Alignment of Single Cell Transcriptomes Unifies Multiple Mouse Cell Atlases into an Integrated Landscape. *bioRxiv*, page 397042, 8 2018.

[121] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, Steven A McCarroll, A.D. Amir, K.L. Davis, M.D. Tadmor, E.F. Simonds, J.H. Levine, S.C. Bendall, D.K. Shenfeld, S. Krishnaswamy, G.P. Nolan, D. Peer, N.R. Beer, E.K. Wheeler, L. Lee-Houghton, N. Watkins, S. Nasarabadi, N. Hebert, P. Leung, D.W. Arnold, C.G. Bailey, B.W. Colston, G.J. Berman, D.M. Choi, W. Bialek, J.W. Shaevitz, P. Brennecke, S. Anders, J.K. Kim, A.A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S.A. Teichmann, J.C. Marioni, M.G. Heisler, R.J. Britten, D.E. Kohne, N.C. Chung, J.D. Storey, F.J. Descamps, E. Martens, P. Proost, S. Starckx, P.E. Van den Steen, J. Van Damme, G. Opdenakker, M. Ester, H.P. Kriegel, J. Sander, X. Xu, E.V. Famiglietti, S.J. Sundquist, A. Feigenspan, B. Teubner, K. Willecke, R. Weiler, T. Hashimshony, F. Wagner, N. Sher, I. Yanai, S. Haverkamp, H. Wässle, B.J. Hindson, K.D. Ness, D.A. Masquelier, P. Belgrader, N.J. Heredia, A.J. Makarewicz, I.J. Bright, M.Y. Lucero, A.L. Hiddessen, T.C. Legler, et al., S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, S. Linnarsson, D.A. Jaitin, E. Kenigsberg, H. Keren-Shaul, N. Elefant, F. Paul, I. Zaretsky, A. Mildner, N. Cohen, S. Jung, A. Tanay, I. Amit, C.J. Jeon, E. Strettoi, R.H. Masland, J.N. Kay, P.E. Voinescu, M.W. Chu, J.R. Sanes, T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, J. Taipale, A.M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D.A. Weitz, M.W. Kirschner, L. Luo, E.M. Callaway, K. Svoboda, R.H. Masland, A. McDavid, G. Finak, P.K. Chattopadyay, M. Dominguez, L. Lamoreaux, S.S. Ma, M. Roederer, R. Gottardo,

G.A. Ascoli, L. Alonso-Nanclares, S.A. Anderson, G. Barrionuevo, R. Benavides-Piccione, A. Burkhalter, G. Buzsáki, B. Cauli, J. Defelipe, A. Fairén, Petilla Interneuron Nomenclature Group, et al., S. Picelli, A.K. Björklund, O.R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, J.R. Sanes, R.H. Masland, J.R. Sanes, S.L. Zipursky, R. Satija, J.A. Farrell, D. Gennert, A.F. Schier, A. Regev, A.K. Shalek, R. Satija, X. Adiconis, R.S. Gertner, J.T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, et al., A.K. Shalek, R. Satija, J. Shuga, J.J. Trombetta, D. Gennert, D. Lu, P. Chen, R.S. Gertner, J.T. Gaublomme, N. Yosef, et al., K. Shekhar, P. Brodin, M.M. Davis, A.K. Chakraborty, S. Siegert, E. Cabuy, B.G. Scherf, H. Kohler, S. Panda, Y.Z. Le, H.J. Fehling, D. Gaidatzis, M.B. Stadler, B. Roska, N.T. Sweeney, H. Tierney, D.A. Feldheim, F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B.B. Tuch, A. Siddiqui, et al., T. Thorsen, R.W. Roberts, F.H. Arnold, S.R. Quake, P.B. Umbanhowar, V. Prasad, D.A. Weitz, A.S. Utada, A. Fernandez-Nieves, H.A. Stone, D.A. Weitz, L. van der Maaten, G. Hinton, B. Vogelstein, K.W. Kinzler, J.G. Wetmur, N. Davidson, M.L. Whitfield, G. Sherlock, A.J. Saldanha, J.I. Murray, C.A. Ball, K.E. Alexander, J.C. Matese, C.M. Perou, M.M. Hurt, P.O. Brown, D. Botstein, Y. Yang, A. Cvekl, Y.Y. Zhu, E.M. Machleder, A. Chenchik, R. Li, and P.D. Siebert. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–14, 5 2015.

[122] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33(2):155–160, 1 2015.

[123] Angelo Duò, Mark D. Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, 9 2018.

[124] Elham Azizi, Ambrose J. Carr, George Plitas, Andrew E. Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, Kristy Choi, Rachel M. Fromme, Phuong Dao, Peter T. McKenney, Ruby C. Wasti, Krishna Kadaveru, Linas Mazutis, Alexander Y. Rudensky, and Dana Peer. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 0(0), 6 2018.

[125] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan,

Wonshik Han, and Woong-Yang Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8:15081, 5 2017.

[126] Diether Lambrechts, Els Wauters, Bram Boeckx, Sara Aibar, David Nittner, Oliver Burton, Ayse Bassez, Herbert Decaluwé, Andreas Pircher, Kathleen Van den Eynde, Birgit Weynand, Erik Verbeken, Paul De Leyn, Adrian Liston, Johan Vansteenkiste, Peter Carmeliet, Stein Aerts, and Bernard Thienpont. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nature Medicine*, 24(8):1277–1289, 8 2018.

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

[128] Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):637–645, 6 2016.

[129] Zhicheng Ji and Hongkai Ji. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Research*, 44(13):e117–e117, 7 2016.

[130] Kieran R Campbell, Christopher Yau, and Inanc Birol. A descriptive marker gene approach to single-cell pseudotime inference. *Bioinformatics*, 6 2018.

[131] Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science (New York, N.Y.)*, 360(6391):eaaq1723, 4 2018.

[132] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru S H Ko, Shigeru B H Ko, Norio Gouda, Tetsutaro Hayashi, Itoshi Nikaido, and Ziv Bar-Joseph. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 8 2017.

[133] Jonathan Ronen and Altuna Akalin. netSmooth: Network-smoothing based imputation for single cell RNA-seq. *F1000Research*, 7:8, 2018.

[134] Nicolette G. Alkema, Tushar Tomar, Evelien W. Duiker, Gert Jan Meersma, Harry Klip, Ate G. J. van der Zee, G. Bea A. Wisman, and Steven de Jong. Biobanking of patient and patient-derived xenograft ovarian tumour tissue: efficient preservation with low and high fetal calf serum based methods. *Scientific Reports*, 5(1):14495, 11 2015.

[135] Peter Eirew, Adi Steif, Jaswinder Khattra, Gavin Ha, Damian Yap, Hossein Farahani, Karen Gelmon, Stephen Chia, Colin Mar, Adrian Wan, Emma Laks, Justina Biele, Karey Shumansky, Jamie Rosner, Andrew McPherson, Cydney Nielsen, Andrew J. L. Roth, Calvin Lefebvre, Ali Bashashati, Camila de Souza, Celia Siu, Radhouane Aniba, Jazmine Brimhall, Arusha Oloumi, Tomo Osako, Alejandra Bruna, Jose L. Sandoval, Teresa Algara, Wendy Greenwood, Kaston Leung, Hongwei Cheng, Hui Xue, Yuzhuo Wang, Dong Lin, Andrew J. Mungall, Richard Moore, Yongjun Zhao, Julie Lorette, Long Nguyen, David Huntsman, Connie J. Eaves, Carl Hansen, Marco A. Marra, Carlos Caldas, Sohrab P. Shah, and Samuel Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, 11 2014.

[136] Bjoern Chapuy, Hongwei Cheng, Akira Watahiki, Matthew D Ducar, Yuxiang Tan, Linfeng Chen, Margaretha G M Roemer, Jing Ouyang, Amanda L Christie, Liye Zhang, Daniel Gusenleitner, Ryan P Abo, Pedro Farinha, Frederike von Bonin, Aaron R Thorner, Heather H Sun, Randy D Gascoyne, Geraldine S Pinkus, Paul van Hummelen, Gerald G Wulf, Jon C Aster, David M Weinstock, Stefano Monti, Scott J Rodig, Yuzhuo Wang, and Margaret A Shipp. Diffuse large B-cell lymphoma patient-derived xenograft models capture the molecular and biological heterogeneity of the disease. *Blood*, 127(18):2203–13, 2016.

[137] Library Prep - Single Cell Gene Expression - Official 10x Genomics Support.

[138] Els M J J Berns and David D Bowtell. The Changing View of High-Grade Serous Ovarian Cancer. 2012.

[139] Daniela Luvero, Andrea Milani, and Jonathan A Ledermann. Treatment options in recurrent ovarian cancer: latest evidence and clinical potential. *Therapeutic advances in medical oncology*, 6(5):229–39, 9 2014.

[140] Wataru Sakai, Elizabeth M. Swisher, Beth Y. Karlan, Mukesh K. Agarwal, Jake Higgins, Cynthia Friedman, Emily Villegas, Cline Jacquemont, Daniel J. Farrugia, Fergus J. Couch, Nicole Urban, and Toshiyasu Taniguchi. Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers. *Nature*, 451(7182):1116–1120, 2 2008.

[141] Eun Jin Heo, Young Jae Cho, William Chi Cho, Ji Eun Hong, Hye-Kyung Jeon, Doo-Yi Oh, Yoon-La Choi, Sang Yong Song, Jung-Joo Choi, Duk-Soo Bae, Yoo-Young Lee, Chel Hun Choi, Tae-Joong Kim, Woong-Yang Park, Byoung-Gie Kim, and Jeong-Won Lee. Patient-Derived Xenograft Models of Epithelial Ovarian Cancer for Preclinical Studies. *Cancer research and treatment : official journal of Korean Cancer Association*, 49(4):915–926, 10 2017.

[142] Clare L Scott, Helen J Mackay, Paul Haluska, and Jr. Patient-derived xenograft models in gynecologic malignancies. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, pages 258–66, 2014.

[143] Ruifen Dong, Wenan Qiang, Haiyang Guo, Xiaofei Xu, J Julie Kim, Andrew Mazar, Beihua Kong, and Jian-Jun Wei. Histologic and molecular analysis of patient derived xenografts of high-grade serous ovarian carcinoma. *Journal of hematology & oncology*, 9(1):92, 9 2016.

[144] Wei-Ting Hwang, Sarah F. Adams, Emin Tahirovic, Ian S. Hagemann, and George Coukos. Prognostic significance of tumor-infiltrating T cells in ovarian cancer: A meta-analysis. *Gynecologic Oncology*, 124(2):192–198, 2 2012.

[145] Andreas Heindl, Chunyan Lan, Daniel Nava Rodrigues, Konrad Koelble, Yinyin Yuan, Andreas Heindl, Chunyan Lan, Daniel Nava Rodrigues, Konrad Koelble, and Yinyin Yuan. Similarity and diversity of the tumor microenvironment in multiple metastases: critical implications for overall and progression-free survival of high-grade serous ovarian cancer. *Oncotarget*, 7(44):71123–71135, 9 2016.

[146] Katharina Auer, Anna Bachmayr-Heyda, Nyamdelger Sukhbaatar, Stefanie Aust, Klaus G. Schmetterer, Samuel M. Meier, Christopher Gerner, Christoph Grimm, Reinhard Horvat, Dietmar Pils, Katharina Auer, Anna Bachmayr-Heyda, Nyamdelger Sukhbaatar, Stefanie Aust, Klaus G. Schmetterer, Samuel M. Meier, Christopher Gerner, Christoph Grimm, Reinhard Horvat, and Dietmar Pils. Role of the immune system in the peritoneal tumor spread of high grade serous ovarian cancer. *Oncotarget*, 7(38):61336–61354, 9 2016.

[147] Alejandro Jiménez-Sánchez, Danish Memon, Stephane Pourpe, Harini Veeraraghavan, Yanyun Li, Hebert Alberto Vargas, Michael B Gill, Kay J Park, Oliver Zivanovic, Jason Konner, Jacob Ricca, Dmitriy Zamarin, Tyler Walther, Carol Aghajanian, Jedd D Wolchok, Evis Sala, Taha Merghoub, Alexandra Snyder, and Martin L Miller. Heterogeneous Tumor-

Immune Microenvironments among Differentially Growing Metastases in an Ovarian Cancer Patient. *Cell*, 170(5):927–938, 8 2017.

[148] Charlotte S. Lo, Sanaz Sanii, David R. Kroeger, Katy Milne, Aline Talhouk, Derek S. Chiu, Kurosh Rahimi, Patricia A. Shaw, Blaise A. Clarke, and Brad H. Nelson. Neoadjuvant Chemotherapy of Ovarian Cancer Results in Three Patterns of Tumor-Infiltrating Lymphocyte Response with Distinct Implications for Immunotherapy. *Clinical Cancer Research*, 23(4):925–934, 2 2017.

[149] Alessandra Cesano. nCounter(®) PanCancer Immune Profiling Panel (NanoString Technologies, Inc., Seattle, WA). *Journal for immunotherapy of cancer*, 3:42, 2015.

[150] Huei San Leong, Laura Galletta, Dariush Etemadmoghadam, Joshy George, Martin Köbel, Susan J Ramus, David Bowtell, and David Bowtell. Efficient molecular subtype classification of high-grade serous ovarian cancer. *The Journal of Pathology*, 236(3):272–277, 7 2015.

[151] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 7 2009.

[152] Christopher T. Saunders, Wendy S. W. Wong, Sajani Swamy, Jennifer Becq, Lisa J. Murray, and R. Keira Cheetham. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817, 7 2012.

[153] Jiarui Ding, Ali Bashashati, Andrew Roth, Arusha Oloumi, Kane Tse, Thomas Zeng, Gholamreza Haffari, Martin Hirst, Marco A. Marra, Anne Condon, Samuel Aparicio, and Sohrab P. Shah. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics*, 28(2):167–175, 1 2012.

[154] Andrew McPherson, Sohrab P Shah, and S Cenk Sahinalp. deStruct: Accurate Rearrangement Detection using Breakpoint Specific Realignment. *bioRxiv*, 117523, 3 2017.

[155] Ryan M Layer, Colby Chiang, Aaron R Quinlan, and Ira M Hall. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*, 15(6):R84, 6 2014.

[156] Andrew W. McPherson, Andrew Roth, Gavin Ha, Cedric Chauve, Adi Steif, Camila P. E. de Souza, Peter Eirew, Alexandre Bouchard-Côté, Sam Aparicio, S. Cenk Sahinalp, and Sohrab P. Shah. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biology*, 18(1):140, 12 2017.

[157] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A. Marra, C. Blake Gilks, David G. Huntsman, Jessica N. McAlpine, Samuel Aparicio, and Sohrab P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, 11 2014.

[158] Ann-Marie Patch, Elizabeth L. Christie, Dariush Etemadmoghadam, Dale W. Garsed, Joshy George, Sian Fereday, Katia Nones, Prue Cowin, Kathryn Alsop, Peter J. Bailey, Karin S. Kassahn, Felicity Newell, Michael C. J. Quinn, Stephen Kazakoff, Kelly Quek, Charlotte Wilhelm-Benartzi, Ed Curry, Huei San Leong, Anne Hamilton, Linda Mileshkin, George Au-Yeung, Catherine Kennedy, Jillian Hung, Yoke-Eng Chiew, Paul Harnett, Michael Friedlander, Michael Quinn, Jan Pyman, Stephen Cordner, Patricia O?Brien, Jodie Leditschke, Greg Young, Kate Strachan, Paul Waring, Walid Azar, Chris Mitchell, Nadia Traficante, Joy Hendley, Heather Thorne, Mark Shackleton, David K. Miller, Gisela Mir Arnau, Richard W. Tothill, Timothy P. Holloway, Timothy Semple, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Suzanne Manning, Senel Idrisoglu, Timothy J. C. Bruxner, Angelika N. Christ, Barsha Poudel, Oliver Holmes, Matthew Anderson, Conrad Leonard, Andrew Lonie, Nathan Hall, Scott Wood, Darrin F. Taylor, Qinying Xu, J. Lynn Fink, Nick Waddell, Ronny Drapkin, Euan Stronach, Hani Gabra, Robert Brown, Andrea Jewell, Shivashankar H. Nagaraj, Emma Markham, Peter J. Wilson, Jason Ellul, Orla McNally, Maria A. Doyle, Ravikiran Vedururu, Collin Stewart, Ernst Lengyel, John V. Pearson, Nicola Waddell, Anna DeFazio, Sean M. Grimmond, David D. L. Bowtell, and David D. L. Bowtell. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*, 521(7553):489–494, 5 2015.

[159] Serena Nik-Zainal, Helen Davies, Johan Staaf, Manasa Ramakrishna, Dominik Glodzik, Xueqing Zou, Inigo Martincorena, Ludmil B Alexandrov, Sancha Martin, David C Wedge, Peter Van Loo, Young Seok Ju, Marcel Smid, Arie B Brinkman, Sandro Morganella, Miriam R Aure, Ole Christian Lingjærde, Anita Langerød, Markus Ringnér, Sung-Min Ahn, Sandrine Boyault, Jane E Brock, Annegien Broeks, Adam Butler, Christine Desmedt, Luc Dirix, Serge Dronov, Aquila Fatima, John A Foekens, Moritz Gerstung, Gerrit K J Hooijer, Se Jin Jang, David R Jones, Hyung-Yong Kim, Tari A King, Savitri Krishnamurthy, Hee Jin Lee, Jeong-Yeon Lee, Yilong Li, Stuart McLaren, Andrew Menzies, Ville Mustonen, Sarah O'Meara, Iris Pauporté, Xavier Pivot, Colin A Purdie, Keiran Raine, Kamna Ramakrishnan, F Germn Rodríguez-González, Gilles Romieu, Anieta M Sieuwerts, Peter T

Simpson, Rebecca Shepherd, Lucy Stebbings, Olafur A Stefansson, Jon Teague, Stefania Tommasi, Isabelle Treilleux, Gert G Van den Eynden, Peter Vermeulen, Anne Vincent-Salomon, Lucy Yates, Carlos Caldas, Laura van't Veer, Andrew Tutt, Stian Knappskog, Benita Kiat Tee Tan, Jos Jonkers, ke Borg, Naoto T Ueno, Christos Sotiriou, Alain Viari, P Andrew Futreal, Peter J Campbell, Paul N Span, Steven Van Laere, Sunil R Lakhani, Jorunn E Eyfjord, Alastair M Thompson, Ewan Birney, Hendrik G Stunnenberg, Marc J van de Vijver, John W M Martens, Anne-Lise Børresen-Dale, Andrea L Richardson, Gu Kong, Gilles Thomas, and Michael R Stratton. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, 2016.

[160] H. M. Li, T. Hiroi, Y. Zhang, A. Shi, G. Chen, S. De, E. J. Metter, W. H. Wood, A. Sharov, J. D. Milner, K. G. Becker, M. Zhan, and N.-p. Weng. TCRB repertoire of CD4+ and CD8+ T cells is distinct in richness, distribution, and CDR3 amino acid composition. *Journal of Leukocyte Biology*, 99(3):505–513, 3 2016.

[161] Ryan Emerson, Anna Sherwood, Cindy Desmarais, Sachin Malhotra, Deborah Phippard, and Harlan Robins. Estimating the ratio of CD4+ to CD8+ T cells using high-throughput sequence data. *Journal of Immunological Methods*, 391(1):14–21, 2013.

[162] Paul L Klarenbeek, Marieke E Doorenspleet, Rebecca E E Esveldt, Barbera D C van Schaik, Neubury Lardy, Antoine H C van Kampen, Paul P Tak, Robert M Plenge, Frank Baas, Paul I W de Bakker, and Niek de Vries. Somatic Variation of T-Cell Receptor Genes Strongly Associate with HLA Class Restriction. *PloS one*, 10(10):e0140815, 2015.

[163] Andrs Szolek, Benjamin Schubert, Christopher Mohr, Marc Sturm, Magdalena Feldhahn, and Oliver Kohlbacher. OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics (Oxford, England)*, 30(23):3310–6, 12 2014.

[164] Nicola Ternette, Hongbing Yang, Thomas Partridge, Anuska Llano, Samandhy Cedeño, Roman Fischer, Philip D. Charles, Nadine L. Dudek, Beatriz Mothe, Manuel Crespo, William M. Fischer, Bette T. M. Korber, Morten Nielsen, Persephone Borrow, Anthony W. Purcell, Christian Brander, Lucy Dorrell, Benedikt M. Kessler, and Tom Hanke. Defining the HLA class I-associated viral antigen repertoire from HIV-1-infected human cells. *European Journal of Immunology*, 46(1):60–69, 1 2016.

[165] MichaelS. Rooney, SachetA. Shukla, CatherineJ. Wu, Gad Getz, and Nir Hacohen. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell*, 160(1-2):48–61, 1 2015.

[166] Arthur Getis and J. K. Ord. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, 24(3):189–206, 9 1992.

[167] Sidra Nawaz, Andreas Heindl, Konrad Koelble, and Yinyin Yuan. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 28(6):766–77, 6 2015.

[168] Sayuri Yoshihama, Jason Roszik, Isaac Downs, Torsten B. Meissner, Saptha Vijayan, Bjoern Chapuy, Tabasum Sidiq, Margaret A. Shipp, Gregory A. Lizee, and Koichi S. Kobayashi. NLRC5/MHC class I transactivator is a target for immune evasion in cancer. *Proceedings of the National Academy of Sciences*, 113(21):5999–6004, 5 2016.

[169] S. Spranger, R. M. Spaapen, Y. Zha, J. Williams, Y. Meng, T. T. Ha, and T. F. Gajewski. Up-Regulation of PD-L1, IDO, and Tregs in the Melanoma Tumor Microenvironment Is Driven by CD8+ T Cells. *Science Translational Medicine*, 5(200):116–200, 8 2013.

[170] Marcel Smid, F Germn Rodríguez-González, Anieta M Sieuwerts, Roberto Salgado, Wendy J C Prager-Van der Smissen, Michelle van der Vlugt-Daane, Anne van Galen, Serena Nik-Zainal, Johan Staaf, Arie B Brinkman, Marc J van de Vijver, Andrea L Richardson, Aquila Fatima, Kim Berentsen, Adam Butler, Sancha Martin, Helen R Davies, Reno Debets, Marion E Meijer-Van Gelder, Carolien H M van Deurzen, Gatan MacGrogan, Gert G G M Van den Eynden, Colin Purdie, Alastair M Thompson, Carlos Caldas, Paul N Span, Peter T Simpson, Sunil R Lakhani, Steven Van Laere, Christine Desmedt, Markus Ringnér, Stefania Tommasi, Jorunn Eyford, Annegien Broeks, Anne Vincent-Salomon, P Andrew Futreal, Stian Knappskog, Tari King, Gilles Thomas, Alain Viari, Anita Langerød, Anne-Lise Børresen-Dale, Ewan Birney, Hendrik G Stunnenberg, Mike Stratton, John A Foekens, and John W M Martens. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nature communications*, 7:12910, 9 2016.

[171] N. McGranahan, A. J. S. Furness, R. Rosenthal, S. Ramskov, R. Lyngaa, S. K. Saini, M. Jamal-Hanjani, G. A. Wilson, N. J. Birkbak, C. T. Hiley, T. B. K. Watkins, S. Shafi, N. Murugaesu, R. Mitter, A. U. Akarca, J. Linares, T. Marafioti, J. Y. Henry, E. M. Van Allen, D. Miao, B. Schilling, D. Schadendorf, L. A. Garraway, V. Makarov, N. A. Rizvi, A. Snyder, M. D. Hellmann, T. Merghoub, J. D. Wolchok, S. A. Shukla, C. J. Wu, K. S. Peggs, T. A. Chan, S. R. Hadrup, S. A. Quezada, and C. Swanton. Clonal

neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351(6280):1463–1469, 3 2016.

[172] Ovarian Tumor Tissue Analysis Consortium, Ellen L Goode, Matthew Block, Kimberly R Kalli, Robert A Vierkant, Wenqian Chen, Zachary Fogarty, Aleksandra Gentry-Maharaj, Aleksandra A Tołoczko, Alexander Hein, Aliecia L Bouligny, Allan Jensen, Ana Osorio, Andreas Hartkopf, Andy Ryan, Anita Chudecka-Głaz, Anthony M Magliocco, Arndt Hartmann, Audrey Y Jung, Bo Gao, Brenda Y Hernandez, Brooke L Fridley, Bryan M McCauley, Catherine J Kennedy, Chen Wang, Chloe Karpinskyj, Christiani B de Sousa, Daniel G Tiezzi, David L Wachter, Esther Herpel, Florin Andrei Taran, Francesmary Modugno, Gregg Nelson, Jan Lubiński, Janusz Menkiszak, Jennifer Alsop, Jenny Lester, Jess García-Donas, Jill Nation, Jillian Hung, Jos Palacios, Joseph H Rothstein, Joseph L Kelley, Jurandyr M de Andrade, Luis Robles-Díaz, Maria P Intermaggio, Martin Widschwendter, Matthias W Beckmann, Matthias Ruebner, Mercedes Jimenez-Linan, Naveena Singh, Oleg Oszurek, Paul R Harnett, Peter F Rambau, Peter Sinn, Philipp Wagner, Prafull Ghatage, Raghwa Sharma, Robert P Edwards, Roberta B Ness, Sandra Orsulic, Sara Y Brucker, Sharon E Johnatty, Teri A Longacre, Eilber Ursula, Valerie McGuire, Weiva Sieh, Yanina Natanzon, Zheng Li, Alice S Whittemore, DeFazio Anna, Annette Staebler, Beth Y Karlan, Blake Gilks, David D Bowtell, Estrid Høgdall, Francisco J Candido dos Reis, Helen Steed, Ian G Campbell, Jacek Gronwald, Javier Benítez, Jennifer M Koziak, Jenny Chang-Claude, Kirsten B Moysich, Linda E Kelemen, Linda S Cook, Marc T Goodman, Mara Jos García, Peter A Fasching, Stefan Kommoss, Suha Deen, Susanne K Kjaer, Usha Menon, James D Brenton, Paul DP Pharoah, Georgia Chenevix-Trench, David G Huntsman, Stacey J Winham, Martin Köbel, and Susan J Ramus. Dose-Response Association of CD8+ Tumor-Infiltrating Lymphocytes and Survival Time in High-Grade Serous Ovarian Cancer. *JAMA oncology*, 3(12):e173290, 2017.

[173] Poorval M Joshi, Shari L Sutor, Catherine J Huntoon, and Larry M Karnitz. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *The Journal of biological chemistry*, 289(13):9247–53, 3 2014.

[174] Roy S. Herbst, Jean-Charles Soria, Marcin Kowanetz, Gregg D. Fine, Omid Hamid, Michael S. Gordon, Jeffery A. Sosman, David F. McDermott, John D. Powderly, Scott N. Gettinger, Holbrook E. K. Kohrt, Leora Horn, Donald P. Lawrence, Sandra Rost, Maya Leabman, Yuanyuan Xiao, Ahmad Mokatrin, Hartmut Koeppen, Priti S. Hegde, Ira

Mellman, Daniel S. Chen, and F. Stephen Hodi. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. *Nature*, 515(7528):563–567, 11 2014.

[175] Kamil A Lipinski, Louise J Barber, Matthew N Davies, Matthew Ashenden, Andrea Sottoriva, and Marco Gerlinger. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. *Trends in cancer*, 2(1):49–63, 1 2016.

[176] Tabula Muris Consortium and others. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 2018.

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

[178] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 2018.

[179] Justina Zurauskiene and Christopher Yau. pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC bioinformatics*, 17(1):140, 2016.

[180] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, and others. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

[181] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7, 2018.

[182] Saskia Freytag, Luyi Tian, Ingrid Lönnstedt, Milica Ng, and Melanie Bahlo. Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, 7:1297, 8 2018.

[183] Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, page 1, 1 2019.

[184] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, Yanyan Ping, Feng Li, Aiai Shi, Jing Bai, Tingting

Zhao, Xia Li, and Yun Xiao. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Research*, 10 2018.

[185] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nature methods*, 15(5):359, 2018.

[186] Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 2017.

[187] Pang Wei Koh, Rahul Sinha, Amira A. Barkal, Rachel M. Morganti, Angela Chen, Irving L. Weissman, Lay Teng Ang, Anshul Kundaje, and Kyle M. Loh. An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Scientific Data*, 3:160109, 12 2016.

[188] Sylvia Richardson John C Marioni Catalina A Vallejos Nils Eling Arianne C. Richard. Correcting the Mean-Variance Dependency for Differential Variability Testing Using Single-Cell RNA Sequencing Data. *Cell Systems*, 7, 2018.

[189] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[190] Mart\'\in~Abadi, Ashish~Agarwal, Paul~Barham, Eugene~Brevdo, Zhifeng~Chen, Craig~Citro, Greg~S.~Corrado, Andy~Davis, Jeffrey~Dean, Matthieu~Devin, Sanjay~Ghemawat, Ian~Goodfellow, Andrew~Harp, Geoffrey~Irving, Michael~Isard, Yangqing Jia, Rafal~Jozefowicz, Lukasz~Kaiser, Manjunath~Kudlur, Josh~Levenberg, Dandelion~Mané, Rajat~Monga, Sherry~Moore, Derek~Murray, Chris~Olah, Mike~Schuster, Jonathon~Shlens, Benoit~Steiner, Ilya~Sutskever, Kunal~Talwar, Paul~Tucker, Vincent~Vanhoucke, Vijay~Vasudevan, Fernanda~Viégas, Oriol~Vinyals, Pete~Warden, Martin~Wattenberg, Martin~Wicke, Yuan~Yu, and Xiaoqiang~Zheng. {TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems, 2015.

[191] Debajyoti Sinha, Akhilesh Kumar, Himanshu Kumar, Sanghamitra Bandyopadhyay, and Debarka Sengupta. dropClust: efficient clustering of ultra-large scRNA-seq data. *Nucleic Acids Research*, 46(6):e36–e36, 4 2018.

[192] Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*, 5:2122, 10 2016.

[193] Jacob H. Levine, Erin F. Simonds, Sean C. Bendall, Kara L. Davis, El-ad D. Amir, Michelle D. Tadmor, Oren Litvin, Harris G. Fienberg, Astraea Jager, Eli R. Zunder, Rachel Finck, Amanda L. Gedman, Ina Radtke, James R. Downing, Dana Peer, and Garry P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 7 2015.

[194] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods*, 12(5):453–457, 5 2015.

[195] Monika Trzpis, Pamela M J McLaughlin, Lou M F H de Leij, and Martin C Harmsen. Epithelial cell adhesion molecule: more than a carcinoma marker and adhesion molecule. *The American journal of pathology*, 171(2):386–95, 8 2007.

[196] Melissa G Mendez, Shin-Ichiro Kojima, and Robert D Goldman. Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 24(6):1838–51, 6 2010.

[197] Jamie R Privratsky and Peter J Newman. PECAM-1: regulator of endothelial junctional integrity. *Cell and tissue research*, 355(3):607–19, 3 2014.

[198] Mathias Uhlen, Cheng Zhang, Sunjae Lee, Evelina Sjöstedt, Linn Fagerberg, Gholamreza Bidkhori, Rui Benfeitas, Muhammad Arif, Zhengtao Liu, Fredrik Edfors, Kemal Sanli, Kalle von Feilitzen, Per Oksvold, Emma Lundberg, Sophia Hober, Peter Nilsson, Johanna Mattsson, Jochen M. Schwenk, Hans Brunnström, Bengt Glimelius, Tobias Sjöblom, Per-Henrik Edqvist, Dijana Djureinovic, Patrick Micke, Cecilia Lindskog, Adil Mardinoglu, and Fredrik Ponten. A pathology atlas of the human cancer transcriptome. *Science*, 357(6352):eaan2507, 8 2017.

[199] M Mura, R K Swain, X Zhuang, H Vorschmitt, G Reynolds, S Durant, J F J Beesley, J M J Herbert, H Sheldon, M Andre, S Sanderson, K Glen, N-T Luu, H M McGettrick, P Antczak, F Falciani, G B Nash, Z S Nagy, and R Bicknell. Identification and angiogenic role of the novel tumor endothelial marker CLEC14A. *Oncogene*, 31(3):293–305, 1 2012.

[200] Antonio Scialdone, Kedar N Natarajan, Luis R Saraiva, Valentina Proserpio, Sarah A Teichmann, Oliver Stegle, John C Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[201] Jiarui Ding, Anne Condon, and Sohrab P. Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature Communications*, 9(1):2002, 12 2018.

[202] D Payne, S Drinkwater, R Baretto, M Duddridge, and M J Browning. Expression of chemokine receptors CXCR4, CXCR5 and CCR7 on B and T lymphocytes from patients with primary antibody deficiency. *Clinical and experimental immunology*, 156(2):254–62, 5 2009.

[203] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 1 2018.

[204] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, JillP. Mesirov, and Pablo Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 12 2015.

[205] Robert Kridel, Fong Chun Chan, Anja Mottok, Merrill Boyle, Pedro Farinha, King Tan, Barbara Meissner, Ali Bashashati, Andrew McPherson, Andrew Roth, Karey Shumansky, Damian Yap, Susana Ben-Neriah, Jamie Rosner, Maia A. Smith, Cydney Nielsen, Eva Giné, Adele Telenius, Daisuke Ennishi, Andrew Mungall, Richard Moore, Ryan D. Morin, Nathalie A. Johnson, Laurie H. Sehn, Thomas Tousseyn, Ahmet Dogan, Joseph M. Connors, David W. Scott, Christian Steidl, Marco A. Marra, Randy D. Gascoyne, and Sohrab P. Shah. Histological Transformation and Progression in Follicular Lymphoma: A Clonal Evolution Study. *PLOS Medicine*, 13(12):e1002197, 12 2016.

[206] Jiarui Ding, Sohrab Shah, and Anne Condon. densityCut: an efficient and versatile topological approach for automatic clustering of biological data. *Bioinformatics*, 32(17):2567–2576, 9 2016.

[207] Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 3 2008.

[208] Fionnuala P. O'Connell, Jack L. Pinkus, and Geraldine S. Pinkus. CD138 (Syndecan-1), a Plasma Cell Marker Immunohistochemical Profile in Hematopoietic and Nonhematopoietic Neoplasms. *American Journal of Clinical Pathology*, 121(2):254–263, 2 2004.

[209] R J Paul. The role of phospholamban and SERCA3 in regulation of smooth muscle-endothelial cell signalling mechanisms: evidence from gene-ablated mice. *Acta physiologica Scandinavica*, 164(4):589–97, 12 1998.

[210] Henrik Lindskog, Elisabet Athley, Erik Larsson, and Samuel Lundin. New Insights to Vascular Smooth Muscle Cell and Pericyte Differentiation of Mouse Embryonic Stem Cells In Vitro. 2006.

[211] O Skalli, M F Pelte, M C Peclet, G Gabbiani, P Gugliotta, G Bussolati, M Ravazzola, and L Orci. Alpha-smooth muscle actin, a differentiation marker of smooth muscle cells, is present in microfilamentous bundles of pericytes. *Journal of Histochemistry & Cytochemistry*, 37(3):315–321, 3 1989.

[212] G Gabbiani, E Schmid, S Winter, C Chaponnier, C de Ckhastonay, J Vandekerckhove, K Weber, and W W Franke. Vascular smooth muscle cells differ from other smooth muscle cells: predominance of vimentin filaments and a specific alpha-type actin. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1):298–302, 1 1981.

[213] Aline Lopes Ribeiro and Oswaldo Keith Okamoto. Combined effects of pericytes in the tumor microenvironment. *Stem cells international*, 2015:868475, 2015.

[214] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[215] Roy Jefferis and Marie-Paule Lefranc. Human immunoglobulin allotypes: possible implications for immunogenicity. *mAbs*, 1(4):332–8, 2009.

[216] O Hermine, C Haioun, E Lepage, M F d'Agay, J Briere, C Lavignac, G Fillet, G Salles, J P Marolleau, J Diebold, F Reyas, and P Gaulard. Prognostic significance of bcl-2 protein expression in aggressive non-Hodgkin's lymphoma. Groupe d'Etude des Lymphomes de l'Adulte (GELA). *Blood*, 87(1):265–72, 1 1996.

[217] Keni Gu, Kai Fu, Smrati Jain, Zhongfen Liu, Javeed Iqbal, Min Li, Warren G Sanger, Dennis D Weisenburger, Timothy C Greiner, Patricia Aoun, Bhavana J Dave, and Wing C Chan. t(14;18)-negative follicular lymphomas are associated with a high frequency of BCL6 rearrangement at the alternative breakpoint region. *Modern Pathology*, 22(9):1251–1257, 9 2009.

[218] Katerina Hatzi and Ari Melnick. Breaking bad in the germinal center: how deregulation of BCL6 contributes to lymphomagenesis. *Trends in molecular medicine*, 20(6):343–52, 6 2014.

[219] Bailey E Freeman, Erika Hammarlund, Hans-Peter Raué, and Mark K Slifka. Regulation of innate CD8 + T-cell activation mediated by cytokines.

[220] Michael R. Green, Shingo Kihira, Chih Long Liu, Ramesh V. Nair, Raheleh Salari, Andrew J. Gentles, Jonathan Irish, Henning Stehr, Carolina Vicente-Dueñas, Isabel Romero-Camarero, Isidro Sanchez-Garcia, Sylvia K. Plevritis, Daniel A. Arber, Serafim Batzoglou, Ronald Levy, and Ash A. Alizadeh. Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proceedings of the National Academy of Sciences*, 112(10):E1116–E1125, 3 2015.

[221] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, 50(8):96, 8 2018.

[222] Julie R. Brahmer, Scott S. Tykodi, Laura Q.M. Chow, Wen-Jen Hwu, Suzanne L. Topalian, Patrick Hwu, Charles G. Drake, Luis H. Camacho, John Kauh, Kunle Odunsi, Henry C. Pitot, Omid Hamid, Shailender Bhatia, Renato Martins, Keith Eaton, Shuming Chen, Theresa M. Salay, Suresh Alaparthy, Joseph F. Grosso, Alan J. Korman, Susan M. Parker, Shruti Agrawal, Stacie M. Goldberg, Drew M. Pardoll, Ashok Gupta, and Jon M. Wigginton. Safety and Activity of Anti-PD-L1 Antibody in Patients with Advanced Cancer. *New England Journal of Medicine*, 366(26):2455–2465, 6 2012.

[223] Jedd D Wolchok, Harriet Kluger, Margaret K Callahan, Michael A Postow, Naiyer A Rizvi, Alexander M Lesokhin, Neil H Segal, Charlotte E Ariyan, Ruth-Ann Gordon, Kathleen Reed, Matthew M Burke, Anne Caldwell, Stephanie A Kronenberg, Blessing U Agunwamba, Xiaoling Zhang, Israel Lowy, Hector David Inzunza, William Feely, Christine E Horak, Quan Hong, Alan J Korman, Jon M Wigginton, Ashok Gupta, and Mario Sznol. Nivolumab plus ipilimumab in advanced melanoma. *The New England journal of medicine*, 369(2):122–33, 7 2013.

[224] Stephen J. Schuster, Michael R. Bishop, Constantine S. Tam, Edmund K. Waller, Peter Borchmann, Joseph P. McGuirk, Ulrich Jäger, Samantha Jaglowski, Charalambos Andreadis, Jason R. Westin, Isabelle Fleury, Veronika Bachanova, S. Ronan Foley, P. Joy Ho, Stephan Mielke, John M. Magenau, Harald Holte, Serafino Pantano, Lida B. Pacaud,

Rakesh Awasthi, Jufen Chu, zlem Anak, Gilles Salles, and Richard T. Maziarz. Tisagenlecleucel in Adult Relapsed or Refractory Diffuse Large B-Cell Lymphoma. *New England Journal of Medicine*, page NEJMoa1804980, 12 2018.

[225] Jennifer L Barnas, Michelle R Simpson-Abelson, Sandra J Yokota, Raymond J Kelleher, Richard B Bankert, and Richard B. Bankert. T cells and stromal fibroblasts in human tumor microenvironments represent potential therapeutic targets. *Cancer microenvironment : official journal of the International Cancer Microenvironment Society*, 3(1):29–47, 3 2010.

[226] Xuefei Li, Tina Gruosso, Dongmei Zuo, Atilla Omeroglu, Sarkis Meterissian, Marie-Christine Guiot, Adam Salazar, Morag Park, and Herbert Levine. Infiltration of CD8+ T cells into tumor-cell clusters in Triple Negative Breast Cancer. *bioRxiv*, page 430413, 10 2018.

[227] Kieran R Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew McPherson, Hossein Farahani, Farhia Kabeer, Ciara O'Flanagan, Justina Biele, Jazmine Brimhall, Beixi Wang, Pascale Walters, IMAXT Consortium, Alexandre Bouchard-Côté, Samuel Aparicio, and Sohrab P Shah. clonealign: statistical integration of independent single-cell RNA & DNA-seq from human cancers. *bioRxiv*, page 344309, 6 2018.

[228] Richard B. Bankert, Sathy V. Balu-Iyer, Kunle Odunsi, Leonard D. Shultz, Raymond J. Kelleher, Jennifer L. Barnas, Michelle Simpson-Abelson, Robert Parsons, and Sandra J. Yokota. Humanized Mouse Model of Ovarian Cancer Recapitulates Patient Solid Tumor Progression, Ascites Formation, and Metastasis. *PLoS ONE*, 6(9):e24420, 9 2011.

[229] Nicole C Walsh, Laurie L Kenney, Sonal Jangalwe, Ken-Edwin Aryee, Dale L Greiner, Michael A Brehm, and Leonard D Shultz. Humanized Mouse Models of Clinical Disease. *Annual review of pathology*, 12:187–215, 1 2017.

[230] Andrew J. Shih, Andrew Menzin, Jill Whyte, John Lovecchio, Anthony Liew, Houman Khalili, Tawfiqul Bhuiya, Peter K. Gregersen, and Annette T. Lee. Identification of grade and origin specific cell populations in serous epithelial ovarian cancer by single cell RNA-seq. *PLOS ONE*, 13(11):e0206785, 11 2018.

[231] Allen W. Zhang, Andrew McPherson, Katy Milne, David R. Kroeger, Phineas T. Hamilton, Alex Miranda, Tyler Funnell, Sonya Laan, Dawn R. Cochrane, Jamie L. P. Lim, Winnie Yang, Andrew Roth, Maia A. Smith, Camila de Souza, Julie Ho, Kane Tse, Thomas Zeng, Inna Shlafman, Michael R. Mayo, Richard Moore, Henrik Failmezger, Andreas

Heindl, Yi Kan Wang, Ali Bashashati, Scott D. Brown, Daniel Lai, Adrian N. C. Wan, Cydney B. Nielsen, Alexandre Bouchard-Cote, Yinyin Yuan, Wyeth W. Wasserman, C. Blake Gilks, Anthony N. Karnezis, Samuel Aparicio, Jessica N. McAlpine, David G. Huntsman, Robert A. Holt, Brad H. Nelson, and Sohrab P. Shah. The interface of malignant and immunologic clonal dynamics in high-grade serous ovarian cancer. *bioRxiv*, page 198101, 10 2017.

[232] Alejandro Jiménez-Sánchez, Paulina Cybulska, Katherine Lavigne, Tyler Walther, Ines Nikolovski, Yousef Mazaheri, Britta Weigelt, Dennis S Chi, Kay J Park, Travis Hollmann, Dominique-Laurent Couturier, Alberto Vargas, James D Brenton, Evis Sala, Alexandra Snyder, and Martin L Miller. Unraveling Tumor-Immune Heterogeneity in Advanced Ovarian Cancer Uncovers Immunogenic Effect of Chemotherapy.

[233] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090–aaa6090, 4 2015.

[234] Xiao Wang, William E Allen, Matthew A Wright, Emily L Sylwestrak, Nikolay Samusik, Sam Vesuna, Kathryn Evans, Cindy Liu, Charu Ramakrishnan, Jia Liu, Garry P Nolan, Felice-Alessio Bava, and Karl Deisseroth. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science (New York, N.Y.)*, 361(6400):eaat5691, 6 2018.

Appendices

Appendix A

Chapter 3 Supplementary Materials

Supplementary Table A.1 Related to: **Figure 3.10**. Primers used for deep amplicon sequencing and clonal inference.

Supplementary Table A.2 Related to: **Figure 3.1**, **Figure 3.8**, and **Figure 3.4**. TIL densities, TIL-based clusters, molecular subtypes, epithelial colocalization measures from histologic image analysis, somatic SNV and rearrangement counts, ITH measures, and TCR and BCR repertoire diversity. NA: cannot be computed/data not available.

Supplementary Table A.3 Nonsynonymous SNVs and the highest predicted affinity neoepitope for each neoantigen, after filtering for HLA LOH. Observed and expected subclonal neoantigen rates, and subclonal neoantigen depletion indices for each multisite HGSC sample.

Supplementary Table A.4 Related to: **Figure 3.7**. HLA-A, HLA-B, and HLA-C germline calls and LOH predictions for multisite HGSC patients. The "clonality" column indicates whether the LOH event is clonal or subclonal.

Supplementary Table A.5 Related to: **Figure 3.12**. Mutation signature proportions and mutational subtype assignments for multisite HGSC (labeled as ITH), OV-AU, and [3] (labeled as OV133) patients.

Supplementary Table A.6 Related to: **Figure 3.12**. Differentially expressed genes between HRD (HRD-DUP + HRD-DEL), FBI, and TD groups in the OV-AU cohort.

Supplementary Table A.7 Related to: **Figure 3.12**. Foldback-HLAMP status and cytotoxicity expression values for TCGA ovarian serous cystadenocarcinoma samples.

Appendix B

Chapter 4 Supplementary Materials

Supplementary Table B.1 Performance measures on simulated data.

Supplementary Table B.2 Marker gene matrices used in analysis.

Supplementary Table B.3 Pathway enrichment results for follicular lymphoma data, by celltype.