Somatic Mutations in Primary Sjögren's Syndrome

Aleksandra Ivovic
Clare College
University of Cambridge

November 2020

Dissertation submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or is being concurrently submitted, for a degree or other qualification at the University of Cambridge or any other university. It does not exceed the prescribed limit of 60,000 words.

Summary

Somatic Mutations in Primary Sjögren's Syndrome

Aleksandra Ivovic

Despite decades of research and many insightful findings, a complete understanding of the pathogenesis of autoimmune diseases continues to elude us. In this thesis, I explore the hypothesis that somatic mutations may underpin the development and progression of autoimmune disease, specifically primary Sjögren's syndrome (PSS). PSS is a chronic, systemic autoimmune disease characterized by dysfunction of salivary and lacrimal glands, which is mediated by immune cell infiltration into these tissues.

As cells age and divide, they accumulate mutations and structural changes in DNA. These somatic mutations are often inconsequential but sometimes result in cell death, cancerous transformation, or other phenotypes. To describe the somatic mutational landscape of relevant cell types in PSS and investigate whether somatic mutations may have a role in the pathogenesis of this disease, we performed targeted and whole genome sequencing of glandular epithelial cells and tissue-infiltrating T and B lymphocytes from biopsies of minor salivary glands from PSS patients and controls. The results illustrate the mutational trends and clonal dynamics present in immune cells and glandular epithelial cells, suggesting ways in which somatic mutations in key cell types may be affecting disease pathogenesis. To complement the genomic studies and interrogate the lymphocyte cell types present in the salivary glands, I also a performed single cell transcriptome analysis of tissue-infiltrating immune cells. Based on the results of this transcriptomic investigation, I highlight new insights about the phenotypes and activation states of immune cells in the inflamed salivary glands.

To our knowledge, this is the first study to perform genomic sequencing of affected tissue and tissue-infiltrating immune cells to investigate the presence of somatic mutations in autoimmune disease. Additionally, this is the first single cell genome-wide transcriptomic investigation of lymphocytes infiltrating minor salivary glands in PSS. The genomic and transcriptional findings of this research contribute novel insights to understanding the molecular origins of primary Sjögren's syndrome.

Acknowledgements

For the mentorship that has guided my doctoral work over the past four years, I would like to express my sincere gratitude towards my supervisors, Peter Campbell and Richard Siegel. As a student of the NIH-Cambridge partnership program, I undertook a research project split between two laboratories and two continents. This presented an exciting opportunity as well as a unique challenge, one which my mentors helped me navigate with the utmost support and investment in my success. The three of us spent many hours discussing the details of my preliminary findings over video calls, from which I would emerge more energized to continue my work. For the past two years at the Sanger Institute, I would especially like to thank Peter for his guidance and unwavering encouragement, and for the opportunities he provided for me to explore the various avenues of my project. To Richard, I would like to say a special thanks for encouraging me to apply to this PhD program, for believing in me since I was a fledgling post-bac, and for guiding me through the whole graduate school experience together with Peter.

My PhD project would not have been possible without our collaborators Matthew Collin and Paul Milne, who set the study of primary Sjögren's syndrome in motion and did the groundwork which enabled me to perform the studies in my dissertation. I immensely appreciate our seamless collaboration and the trust placed in me to conduct this research. I also extend thanks to the patients who provided samples and the clinicians who helped procure them. Furthermore, I'd like to thank my colleagues in the Campbell group and others at the Wellcome Sanger Institute who taught me a great deal about the human genome over the past two years, in addition to providing a network of friendship and support. Likewise, I extend my gratitude to my colleagues at the US National Institute of Arthritis, Musculoskeletal and Skin Diseases, and in particular to the members of the Siegel group, for their support and lasting friendship. I'm thankful for the invaluable amount I learned about immunology and rheumatic diseases during my time there, along with the opportunity to witness a remarkable research environment that integrates the bench and the bedside. I extend a special thanks to Zuoming Deng, who helped me start my journey into bioinformatics and the study of somatic mutations.

I owe my gratitude to the National Institute of Arthritis, Musculoskeletal Disease, and Skin for funding the four years of my PhD, and the Wellcome Trust Sanger Institute for their support and supplementary funding. I thank the NIH-Oxford/Cambridge PhD program and the International Biomedical Research Alliance, for their effective management of the geographical challenges of this program, and for the travel funding, career-building opportunities, and additional research grants they provided.

Finally, this doctorate would not have happened without the support of my loved-ones. My parents, Vera and Dejan, who poured all their love and support into helping me achieve my goals, and who shared my excitements and disappointments along the way. My grandmother Vukosava, who followed every step of my journey and who lovingly raised me to believe that the most important thing for a girl to strive for is an education. My partner, John, who crossed an ocean with me to pursue our adventures, and who shared in my day-to-day successes and failures, always there to offer his unreserved support. To them I dedicate this thesis; it is as much theirs as it is mine.

Овај докторат не би био могућ без подршке мојих најмилијих. Мојих родитеља, Вере и Дејана, који су пружили сву своју љубав и подршку да ми помогну да остварим своје циљеве, и који су сва успутна узбуђења и разочарења проживљавали заједно са мном. Моје баке Вукосаве, која ме је с љубављу одгајила да верујем да је најважнија ствар за жену да стекне образовање, и која је пратила цео ток мог школовања и мојих авантура. Мог партнера Џона, који је са мном прешао океан да како бисмо заједно стварали своје авантуре, са којим сам делила свакодневне успоне и падове, и који је увек био спреман да ми пружи безрезервну подршку. Ову докторску дисертацију посвећујем њима; њихова је колико је и моја.

Table of Contents

Chapter 1: Introduction 1

Somatic mutations beyond cancer 1

I. SOMATIC EVOLUTION 3

I.1 Somatic mutations in non-cancerous tissue 3
I.2 Clonal haematopoiesis & somatic mutations in blood 3
I.3 Somatic mutations and clonal evolution in tissue 6
I.4 Mutational signatures in normal tissues 8
I.5 Somatic mutations in chronic disease 10
I.6 Methods and challenges of somatic mutation detection in normal tissue 12

II. AUTOIMMUNE DISEASE 14

II.1 Primary Sjögren's syndrome: overview 16
II.2 Diagnosis of primary Sjögren's syndrome 16
II.3 Association studies: genes and epigenomes 17
II.4 B cells and plasma cells in PSS 18
II.5 Interferons and the innate immune response 20
II.6 Quantitative immunophenotyping: identifying key cell types in PSS 21
II.7 Salivary epithelium, microbial defence, and hormones 21
II.8 Clonal evolution and B cell lymphomas complicating Primary Sjögren's syndrome 23
II.9 Disease pathogenesis: hypotheses and future directions 24
II.10 Somatic mutations in autoimmune disease 26

Thesis aims 29

Chapter 2: Methods 31

Contributions 31

I. SAMPLES AND WET LAB METHODS 32

I.1 Minor salivary gland biopsies and blood samples 32
I.2 Processing of tissue and blood 33
I.3 Isolating lymphocyte populations from minor salivary gland tissue 33
I.4 Histology and immunohistochemistry of minor salivary gland biopsies 34
I.5 Laser capture microdissection 35
I.6 Library preparation and sequencing 36

II. DRY LAB METHODS 37

II.1 Design of a bait set for targeted sequencing 37
II.2 Alignment of sequencing data 38
II.3 Caveman single nucleotide substitution calling and filtering 39
II.4 Pindel insertion/deletion calling and filtering 40
II.5 Shearwater variant calling and filtering 41
II.6 Detection of driver mutations and selection 43
II.7 Analysis of mutational burden 44
II.8 Copy number and structural variant calling 45
II.9 V(D)J repertoire assembly 45
II.10 Phylogenetic tree construction 46
II.11 Mutational signature extraction 46
II.12 Detection of viral elements 47
II.13 Single cell RNA expression analysis 48

Chapter 3: DNA sequencing of lymphocytes from minor salivary gland biopsies 51

STUDY AIMS 52

I. SEQUENCING OF SORTED LYMPHOCYTE POPULATIONS FROM MINOR SALIVARY GLANDS 54

I.1 Patient cohort 55
I.2 Cell sorting of lymphocyte compartments 55
I.3 Library preparation 57
I.4 Targeted sequencing (TGS) of bulk sorted lymphocytes 58
I.5 Mutational findings in TGS dataset 59
I.6 Copy number alterations in TGS dataset 62
I.7 Receptor sequences of tissue-derived lymphocytes 68

II. Whole genome sequencing (WGS) of bulk sorted lymphocytes 71
II.1 WGS findings 72

III. Laser-capture microdissection approach for sequencing lymphocytes from PSS biopsies 74
III.1 Histology and immunohistochemistry of minor salivary glands in PSS 76
III.2 WGS of LCM-derived lymphocytes 77

IV. Integrated analysis of minor salivary gland lymphocytes 81
IV.1 Mutational signature analysis 81
IV.2 Analysis of genes under selection 83

V. Discussion 84

Chapter 4: Single-cell expression profiling of lymphocytes infiltrating minor salivary glands 92

STUDY AIMS 93

STUDY DESIGN 94

I. Results of plate-based full-length cDNA single cell sequencing (SmartSeq2 protocol) 96
I.1 Clustering and cell type detection 98
I.2 T cell dynamics 101
I.3 B cell and plasma cell dynamics 104
I.4 V(D)J expression in single cells 106

II. Results of droplet-based 5' expression with V(D)J enrichment single cell sequencing (10X Genomics protocol) 109

III. Discussion 113
III.1 Summary of approach 113
III.2 B cell findings 114
III.3 T cell findings 115
III.4 Limitations and future directions 116

CHAPTER 5: DNA sequencing of salivary gland epithelial cells 118

STUDY AIMS 120

STUDY DESIGN 120

RESULTS 122
I.1 Whole genome and exome sequencing of glandular epithelium 122
I.2 Clonality and phylogenetic relationships 123
I.3 Mutational burden and driver detection 125
I.4 Mutational signature analysis 128
I.5 Metagenomic analysis of viral elements 130
I.6 Copy number variation 132

DISCUSSION 134
II.1 Findings related to normal minor salivary gland epithelium 134
II.2 Findings related to minor salivary gland epithelium in PSS 137
II.3 Limitations and future directions 138

Chapter 6: Conclusion 141

I. B cells **141**
II. T cells **142**
III. Epithelial cells **144**
Original publications ***147***
References ***148***
Appendix Table A1 ..... ***162***

Chapter 1: Introduction

Somatic mutations beyond cancer

The study of somatic mutations has long been relegated solely to the domain of cancer research. The somatic evolution of malignant clones has been extensively studied, leading to our current understanding of the origin and progression of cancer. In essence, individual cells within an organism acquire mutations and changes in DNA throughout their lifetime, most of which are phenotypically inconsequential; occasionally, these changes result in a selective advantage over other cells, leading to sustained survival and proliferation - an early event in the trajectory leading to cancer[1]. Additional mutations subsequently occurring in the cell can provide the definitive push towards malignant transformation. We therefore understand cancer to be an acquired genetic disease, centred on the accumulation and Darwinian selection of somatic mutations within an organism – evolution on a microscale.

In recent years, the study of somatic mutations has begun to explore healthy tissues to better understand the early mutagenic events that precede the development of cancer. These sequencing studies began to shed light on the mutational burden, patterns, and mutagenic processes operative across different normal, non-cancerous cell types. As it turned out, somatic mutations are prevalent and diverse in normal tissues, and they accumulate consistently with age[2–5]. Mutations previously associated with cancer are found to exist in healthy cells and drive the clonal dynamics of the tissue[2–5]. A clone, which consists of all the cells that derive from a common progenitor and share the same mutations, can exist in various sizes, and it can shrink and expand over time in response to selection pressures in the microenvironment. The majority of mutated clones will not go on to drive cancer, but they will shape the mosaic structure of the tissue. This perhaps unexpected finding has imparted more complexity to the classical paradigm of cancer evolution, suggesting a fluid and malleable process rather than a linear progression, as was previously implied.

If pathogenic somatic mutations exist in phenotypically healthy tissue and shape its clonal landscape, do they also play a role in non-cancerous diseases? If mutations accumulate with

cellular ageing, could they contribute to diseases that occur sporadically in the lifetime of an organism? Based on current knowledge of the role of somatic mutations in cancer and healthy tissue, we sought to apply this line of inquiry to the realm of chronic and age-related diseases, specifically autoimmune disease.

Autoimmune disease occurs when the immune system loses tolerance toward antigens of the body and begins to attack its own cells and tissues. There are over 80 different known autoimmune diseases, and their cumulative burden in the population is around 5%[6]. These chronic diseases range from mild to life-threatening with a diverse set of symptoms. There are no cures, and until recently treatment options were limited and often associated with considerable toxicity. With the advent of modern biological treatments, we are now better equipped than previously to ease the burden of suffering in patients, yet there still remains a significant unmet need not addressed by current therapies. For most autoimmune diseases, there is a sparse understanding of the genetic basis and external factors that are thought to drive the pathogenesis.

We hypothesized that somatic mutations might play a role in autoimmune disease pathogenesis, building on previous work that has suggested this idea[7,8]. Whether functionally relevant somatic mutations are present in immune cells is a concept that has not been experimentally explored until recently, even though the connection between somatic mutations and autoimmunity was first proposed by MacFarlane Burnet in 1965[7]. Somatic mutations in lymphocytes may be a driving factor in the loss of immune tolerance and development of autoimmune disease, or they may be an occurrence secondary to the disease itself and lead to subsequent complications such as lymphoma development. To explore these concepts, we undertook a study of somatic mutations in the autoimmune disease primary Sjögren's syndrome. While most investigations of autoimmune disease have relied on profiling readily available blood samples to infer systemic changes in the immune response, examining affected tissue, which is often harder to come by, presents an opportunity to study the localized tissue-specific autoimmune response. By obtaining biopsies of affected salivary glands from patients with primary Sjögren's syndrome, we were able to analyse infiltrating immune cells as mediators of the localized autoimmune response and somatic mutations which may be enriched in cells at the site of tissue inflammation. In this dissertation, I will

discuss our exploration of the landscape and putative role of somatic mutations in primary Sjögren's syndrome.

I. SOMATIC EVOLUTION

I.1 Somatic mutations in non-cancerous tissue

The study of somatic mutations in cancer exploded in the twenty-first century with the advent of next generation sequencing, which enabled whole exome and whole genome sequencing at massive scale. These studies led to a comparative understanding of mutational burden, rates of mutagenesis, structural variation, and other features across a wide array of cancer types[9]. Specific patterns of mutation, known as mutational signatures, were attributed to specific mutagens, explaining some of the driving forces behind mutation acquisition[10–12]. The mutational landscape and genetic patterns of cancer thus became understood on a new level.

The logical extension in this process of inquiry then became to understand early events of somatic mutation accumulation in normal cells that sets them on the road to cancerous transformation. A few early landmark studies of normal tissue shifted the focus of the cancer genomics field, and the study of somatic mutations in normal cells took off.

I.2 Clonal haematopoiesis & somatic mutations in blood

Clonal haematopoiesis is the occurrence of a disproportionally large fraction of mature blood cells deriving from a single haematopoietic stem cell, thus forming a "clone" of blood cells. In 2014, two parallel studies by Jaiswal *et al* and Genovese *et al* found that clones harbouring leukaemia-associated somatic mutations accumulate in the blood of healthy individuals over time, indicating clonal haematopoiesis that is likely driven by somatic mutations[13,14]. This unexpected finding in apparently healthy people is a phenomenon now known as "clonal haematopoiesis of indeterminate potential" ("CHIP") or "age-related clonal haematopoiesis". Based on whole exome sequencing data analysed in these studies, the prevalence of clonal haematopoiesis increased with age, with 10% of individuals older than 65 years harbouring a clone with a somatic mutation, compared to only 1% of those younger than 50 years[13]. Those

over the age of 90 had detectable somatic mutations in 18.4% of cases[14]. In subsequent studies, where deeper sequencing was implemented for improving the sensitivity of detection, these proportions were even higher[15].

Clonal haematopoiesis increases the risk of developing haematological malignancy and is also associated with a higher risk of cardiovascular disease and all-cause mortality[13,14,16,17]. Across all age groups, the most commonly mutated are three genes encoding for epigenetic regulators which are associated with myeloid malignancy: *DNMT3A, ASXL1*, and *TET2.* Stratification of individuals with clonal haematopoiesis by number of mutations, clone size, and enrichment of specific genes demonstrated that it is possible to quantify cancer risk and identify those more likely to develop acute myeloid leukaemia in subsequent years versus those who have benign age-related clonal haematopoiesis[18]. The presence of *TET2* mutations in particular is correlated with adverse cardiovascular outcomes and has been shown to contribute to atherosclerosis in mouse models[16,17].

While the concept of clonal haematopoiesis has been known for many years, mainly through studies of non-random X-inactivation in blood[19], the detection of it through somatic mutations has greatly expanded this field of research. The aforementioned sequencing studies have illustrated the prevalence of clonal haematopoiesis in healthy individuals' blood and demonstrated its utility in predicting risk of blood cancer and other conditions.

Clonal haematopoiesis has been significantly associated with aplastic anaemia, an autoimmune disease in which the immune system destroys hematopoietic cells and leads to bone marrow failure. Aplastic anaemia has a high incidence of clonal haematopoiesis, occurring in 47% of 439 patients analysed in a study by Yoshizato *et al*[20]*.* Clones with somatic mutations in *DNMT3A* and *ASXL1* were found to increase in size over time and were associated with worse outcomes and higher rate of progression to acute myeloid leukaemia and myelodysplastic syndromes. Conversely, clones harbouring mutations in *PIGA, BCOR*, and *BCORL1* were correlated with better response to immunosuppressive treatment and longer progression-free survival[20]. The differential outcomes associated with these two sets of mutations may indicate distinct selective pressures in the bone marrow environment. Clones with *DNMT3A* and *ASXL1* mutations have a proliferative advantage over other cells, as they

do in healthy individuals, while clones carrying *PIGA, BCOR*, and *BCORL1* mutations are likely an adaptation selected to confer protection from the pathogenic immune response.

Most studies investigating clonal haematopoiesis of indeterminate potential have focused exclusively on myeloid cells and the presence of known myeloid drivers in the blood. However, examples of pre-symptomatic lymphocyte expansion are known as well, most notably monoclonal gammopathy of undetermined significance ("MGUS"), where an otherwise healthy individual has a monoclonal component in the blood (an antibody or paraprotein), indicative of monoclonal plasma cell expansion and imparting a predisposition to myeloma[21]. Individuals with monoclonal gammopathy of undetermined significance have been shown to harbour cancerous mutations associated with myeloma, though a notably lower mutational burden has been observed than that of myeloma[22]. A related pre-malignant disorder is clonal B-cell lymphocytosis, where a monoclonal B lymphocyte population exists in the blood of an asymptomatic individual and imparts a predisposition to B-cell malignancy, most often B-cell chronic lymphocytic leukaemia[23]. The significance of lymphoid and myeloid clonal expansions is not only in their ability to predict risk of developing haematological cancers, but increasingly in correlating the pre-malignant state with other morbidities as well, such as myeloid *TET2* mutations with cardiovascular disease[17] or thrombosis and osteoporosis with monoclonal gammopathy of unknown significance[24].

Ongoing research into the clonal evolution of blood cells continues to elucidate the biological mechanisms of healthy haematopoiesis. Somatic mutations in haematopoietic stem cells can be traced through their progeny of differentiated blood cells to quantify the contribution of respective progenitors. This approach has been used to construct phylogenetic trees of hematopoietic differentiation based on shared somatic mutations[25,26]. The authors identified stem cell clones that generated multilineage progeny, both myeloid and lymphoid. Additionally, Lee-Six *et al* inferred the number of active haematopoietic stem cells present in a healthy donor to be in the range 50,000 to 200,000 cells[25]. Lineage tracing using somatic mutations as cellular barcodes can thus be an important tool for studying clonal dynamics in normal tissue and disease states.

I.3 Somatic mutations and clonal evolution in tissue

Akin to the study of clonal haematopoiesis in healthy blood, there has been a surge of interest in the study of somatic mutations in other healthy tissues in the past several years. The first study to launch this wave was an investigation of the somatic mutational landscape of healthy, sun-exposed skin by Martincorena *et al* in 2015[2]. Deep sequencing of 234 skin micro-biopsies from four donors revealed that ~25% of all cells contained at least one "driver" mutation previously associated with cancer, namely with cutaneous squamous cell carcinoma. Clones carrying driver mutations were numerous and expanded to varying sizes in the tissue. The high prevalence and clonal expansion of cells carrying driver mutations indicates that the mutated genes were under positive selection in normal skin. The total burden of mutations per cell was high, comparable to or higher than that seen in many cancers[2]. The main mutagenic force associated with the excess burden of mutations was UV-light, recognizable by its characteristic mutational signature. These surprising findings suggested that hidden clonal competition is continually taking place in otherwise normal skin, reshaping the tissue microenvironment and, in turn, our understanding of the early inciting mutations that lead to skin cancer.

To understand whether similar mutational burden is found in tissues that are not exposed to a strong mutagen such as UV light, the authors undertook a study of normal human oesophageal epithelium[4]. The skin and the oesophagus have a similar histological structure consisting of a layered epithelium that has a high rate of shedding and turn-over, as well as an association with respective squamous cell carcinomas, so they are apt tissues for comparison. As expected, the mutational burden in the oesophagus was lower than in skin, however the density of cancer-associated driver mutations was surprisingly high, indicating a strong positive selection of clones carrying these mutations. The burden of mutations increased with age and was attributed largely to endogenous cellular process associated with age or transcription. The size of mutated clones also increased with age, where in middle age more than half of the oesophageal epithelium was colonized by mutant clones. Fourteen cancer-associated genes were found to be under positive selection in normal oesophagus, at least 11 of which are canonical drivers of oesophageal squamous cell carcinoma (ESCC). Interestingly however, there was a higher prevalence of *NOTCH1* mutations in normal

oesophagus than is found in ESCC, with 30-80% of normal oesophageal epithelium being colonized by a *NOTCH1* clone. This suggests that *NOTCH1* mutations may drive benign clonal expansions but are less likely than other mutations to drive evolution towards ESCC. Conversely, *TP53* mutations were less common in normal oesophagus but highly prevalent in ESCC, indicating that selection of *TP53* clones occurs in the process of malignant transformation. In this comparative fitness model of normal tissue, different mutant clones appear to have varying degrees of cancer progression risk.

The findings of the oesophageal epithelium study have important implications for understanding the events that precede carcinogenesis, as well as the biology of ageing. The increasing appropriation of tissue by mutant clones with age may play a role in its physiological decline, supporting the long-hypothesized somatic theory of ageing. Furthermore, the finding that some expanded clones are less likely than others to evolve into cancer, e.g. *NOTCH1* clones appearing more benign than *TP53* clones, opens a possibility of intervention whereby the benign clones could be stimulated to outcompete the higher-risk clones and thus reduce cancer risk. In a recent study, Fernandez-Antoran *et al* tested the effects of oxidative stress from low-dose ionizing radiation on the fitness of wild-type and *TP53* mutant cells in the transgenic mouse oesophagus[27]. While exposure to low-dose ionizing radiation induced oxidative stress and caused *TP53* mutant cells to proliferate and outcompete wild-type cells, the addition of an antioxidant reversed this effect, causing proliferation of wild-type cells and reduction of *TP53* mutant clones. The external modification of selection pressures by redox manipulation demonstrates that intervention to deplete high-risk clones in tissues is possible. This example of low-dose ionizing radiation draws parallels with patients undergoing frequent CT scans, which may translate to a clinical opportunity for intervention with antioxidizing agents.

In addition to the studies of blood, skin, and oesophagus, further genomic investigations have characterized the landscape of somatic mutations in healthy colon[5], endometrium[3,28], liver[29], bronchial epithelium[30], and brain[31], yielding new tissue-specific observations as well as confirming trends that occur across many cell types. New insights were also found from clever repurposing of transcriptomic data to detect somatic mutations. An RNA sequencing meta-analysis of 6,700 samples from 29 normal tissues revealed macroscopic clonal expansions

across many of the tissues[32]. While lower in resolution compared to DNA sequencing analyses, the findings of the transcriptional dataset aligned with what is known so far about mutations in normal tissue: the highest burden was found in sun-exposed skin, oesophagus, and lung – tissues that are exposed to exogenous mutagens. Mutations accumulated with age, and cancer driver mutations were commonly seen across various tissues. Importantly, the mutational burden of a tissue was associated with its rate of cell proliferation and turn-over, and consequently its propensity to form macroscopic clones[32,33].

I.4 Mutational signatures in normal tissues

While the concept of mutational signatures has been known for some time, cancer genome sequencing has greatly broadened the spectrum of signatures attributable to distinct carcinogens. An early landmark study in 1991 compared the types of *TP53* mutations across different cancers and found that some tumour types such as colon and brain are dominated by T to C transitions, while others like lung and breast contain mostly A to G transitions in *TP53*[34]. Subsequent studies elucidated the predilection of mutagens towards certain base changes, such as the high rate of C > T and CC > TT mutations caused by UV light as compared to G >T mutations caused by tobacco[35]. UV light and tobacco smoke were some of the earliest mutagens characterized molecularly, and it was not until the implementation of wide-scale genomic sequencing that the mutational signatures of many others became known.

Single nucleotide changes can be categorized into six classes, C>A, C>G, C>T, T>A, T>C, and T>G, which can further be expanded into 96 categories by the inclusion of the 3' and 5' adjacent base to establish a trinucleotide context. Several large-scale studies in the last decade compared the mutation distribution in various cancers among these categories and discovered associations with numerous mutagens, both endogenous to the cell and those in the external environment[10–12,36–39]. Signatures SBS1 (Single Base Substitution 1) and SBS5 defined in the COSMIC database (Catalogue of Somatic Mutations in Cancer)[40] were found in most tissues and are the result of cellular ageing and replication, thus are appropriately termed "clock-like" signatures[36]. Other novel signatures identified included those associated with defective homologous recombination, defective mismatch repair, cytidine deaminases AID and APOBEC[39], damage by reactive oxygen species[41], aristolochic acid[42], and more.

Additionally, signatures of dinucleoutide substitutions, insertion and deletions, and structural variants have been identified[11,43]. A significant portion of identified signatures still do not have a known aetiology.

Recent genomic studies of normal tissues have investigated the presence of known and novel mutational signatures in non-cancerous cells. As in cancer, the largest proportion of single base substitutions in healthy cells is attributable to ubiquitous "clock-like" signatures SBS1 and SBS5, which have been identified in normal blood[25,26], colon[5], liver[29,33], brain[31], endometrium[3,28], lung[30], and other tissues. Tissue-specific endogenous signatures were also found, such as the AID cytidine deaminase signature in mature B cells[44] and a novel signature found across blood cells[25,26]. Higher rates of SBS1 were found in tissues with higher turnover rate such as small intestine and colon[5,33], compared to tissues with lower turnover such as liver[33,29]. The contribution of SBS5 increased with age in a linear trend across many tissues, suggesting an intrinsic mutational mechanism independent of cell type or proliferation rate[33]. The mutational spectra of driver genes in cancer types corresponding to the normal tissues surveyed showed highly similar trends, indicating that endogenous processes driving mutagenesis in normal tissue contribute to tumorigenesis[11,33,36].

In addition to signatures of endogenous origin, various signatures of environmental exposure were found in normal tissues. As mentioned, a high burden of UV-light signature was found in healthy sun-exposed skin[2]. Tobacco smoke significantly increased the burden of C>A mutations in bronchial epithelium of healthy donors[30] and has also been found in samples of healthy and non-cancerous cirrhotic liver[29]. Non-cancerous hepatocytes in the study by Brunner *et al* also demonstrated the presence of signature SBS22, characteristic of aristolochic acid[29,42] as well as SBS24 associated with aflatoxin-B1 exposure[29,45], in certain individuals whose personal histories confirmed those exposures. Ongoing research continues to investigate novel signatures in cancer and healthy tissue to find further environmental exposures contributing to tumorigenesis, with the hope of identifying preventable causes of cancer. The sum of mutational signatures in a cell represent a record of its exposure to both exogenous and endogenous mutational forces, providing a retrospective summary of its life history, which is highly pertinent to the study of cancer and chronic diseases.

I.5 Somatic mutations in chronic disease

Somatic mutational findings in normal tissue have implications for ageing and chronic diseases. The expansion of a clone with a single driver mutation can be an inciting event in the trajectory leading to cancer, but it might also be an early molecular event leading to physiological decline in the context of ageing or a pathological manifestation in the context of chronic disease. Support for the latter is the connection of certain chronic diseases with cancer, such as that of liver cirrhosis with hepatocellular carcinoma[33,46] or autoimmune diseases with non-Hodgkins lymphoma[47]. Thus, the aetiology of many chronic diseases may be clonal in origin.

There are examples of rare diseases caused by early embryonic mutations such as the overgrowth disorders Proteus syndrome and melorheostosis. During my PhD, I collaborated on a study that discovered that somatic mutations in the *MAP2K1*[48] or *SMAD3*[49] genes cause the rare sporadic bone overgrowth condition Melorheostosis. In a similar way, Proteus syndrome is caused by a single somatic mutation in the *AKT1* gene that leads to overgrowth of skin and connective tissues[50]. These overgrowth disorders are caused by somatic activating mutations in growth-promoting genes, with the mutated clones being sequestered in mosaic regions of the body.

In some rare disorders of the immune system, embryonic mutations in immune-activating genes are the cause of disease, and these are mimics of Mendelian disorders involving those same genes[51–53]. In this scenario, the mutations promote immune stimulation by constitutive activation of inflammasome complexes or similar mechanisms in a subset of mutated immune cells, but they do not necessarily have a selective advantage by means of a growth-promoting cellular phenotype. Early embryonic mutations with significant functional consequences such as these are one way in which somatic mutations play a direct role in rare chronic diseases.
In common chronic diseases, the situation is more complex and likely due to multiple molecular events occurring throughout the lifetime of an individual. There is now early evidence in favour of this hypothesis. As previously mentioned, the genomic analysis of healthy liver as compared to chronic liver disease portrays a vastly different mutational landscape in the disease state, manifesting as higher mutational burden, more frequent driver

mutations, more structural variation, and larger clonal expansions[29]. As chronic liver disease predisposes to hepatocellular carcinoma, this indicates that the genomic instability leading to malignancy is gradually acquired throughout the progression of chronic liver disease. Unexpectedly, there is also an enrichment of mutations not associated with cancer in regenerative nodules of cirrhotic liver, suggesting selection of clones that appear to promote hepatocyte fitness and a decrease in fibrosis[54].

Additionally, an enrichment of mutations in genes involved in insulin signalling are observed in non-alcoholic fatty liver disease that are not observed in hepatocellular carcinoma (Ng et al, submitted 2020). The recurrent mutations were observed in conserved regions of insulin signalling genes and were shown to disrupt downstream metabolic pathways. Non-alcoholic fatty liver disease is common in the population and is highly associated with metabolic syndrome, in particular insulin resistance[55]. The findings of this study suggest a novel hypothesis linking fatty liver disease and metabolic syndrome: selection of hepatocyte clones with reduced insulin reactivity allows them to escape death from lipotoxicity and oxidative stress in a fatty liver environment – a survival advantage that comes at a metabolic cost to the organism.

Perhaps more related in pathophysiology to Sjögren's syndrome is a recent set of studies on somatic mutations in inflammatory bowel disease. Sequencing of colonic crypts and organoids from Crohn's disease and ulcerative colitis patients identified expansions of clones with mutations that converge on the IL-17a signalling pathway (including *NKFBIZ, TRAF3IP2, ZC3H12A* and *PIGR* genes)[56–58]. This suggests a potential adaptive mechanism for escaping inflammatory damage by IL-17 activation, as most of the mutations conferred protection from IL-17 mediated cytotoxicity. Alternatively, mutations disrupting IL-17 signalling, specifically *PIGR* mutations, may be directly contributing to pathogenesis by causing commensal dysbiosis which recruits a sustained immune response and positive feedback loop, leading to expansion of the mutated clone[57,58]. The latter, more provocative hypothesis warrants further mechanistic investigation but is supported by GWAS findings associating loci in *PIGR* and other IL-17 pathway genes with susceptibility to inflammatory bowel disease[59–61]. Overall, the enrichment of mutations in immune-signalling genes demonstrates a distinct mechanism of positive selection in inflammatory bowel disease that does not necessarily contribute to

neoplastic transformation, as the aforementioned IL-17 related genes are not found in colon cancers associated with inflammatory bowel disease.

I.6 Methods and challenges of somatic mutation detection in normal tissue

Next generation sequencing advances have enabled the study of somatic mutations in normal tissue, which was previously precluded by significant technical challenges. In contrast to neoplastic clones which are often macroscopic and easily detectable, clones found in normal tissue are smaller and more difficult to identify. It is straightforward to detect somatic mutations in a clonal population of cells but less so in a mixed polyclonal population in which each mutation is represented by fewer sequencing reads. To increase the sensitivity of somatic mutation detection in polyclonal samples, high depth sequencing can be used to detect mutations present at low allele frequency. Whilst this approach is useful for detecting mutations in a targeted set of genes, sequencing whole exomes and genomes at very high depth quickly becomes cost-prohibitive. Additionally, in a polyclonal sample it is difficult to assess if multiple mutations derive from the same clone or multiple clones. To overcome some of these challenges and allow somatic mutation assessment across normal tissues, several approaches have been adopted.

Since clonal composition is an important factor for mutation detection, it is worthwhile to consider the origin of normal and neoplastic clonal structures. It is well established that cancers arise from long-lived cells with proliferative potential, i.e. progenitor and stem cells[33,62], as opposed to differentiated cells which are short-lived and senescent. Moreover, it has been shown that the organ-specific risk of acquiring cancer is directly related to the number of stem cell divisions, explaining the differential incidence of cancer across tissues[63]. Similarly, the rate of stem cell division in a tissue determines the size of non-neoplastic clones and the resulting tissue microarchitecture. For example, colonic crypts are a clonal unit, deriving from a stem cell at the bottom of the crypt which divides to create daughter cells that populate the walls of the crypt as they differentiate into a squamous epithelium[5,64]. Characteristics like these can be taken advantage of when designing studies, such that dissection of a colonic crypt will provide a clonal population which carries all the mutations present in the basal stem cell. This allows genome-wide assessment of the mutational

processes that occurred in the lifetime of that stem cell, independently of surrounding cells. The clonal composition of a tissue can be modified by selective pressures, as seen in inflammatory bowel disease by spreading of clones between multiple colonic crypts[56–58]. Therefore, tissue composition and clonality are important considerations for designing sequencing studies.

To enrich for the presence of clonal populations, small biopsies have been the approach of choice in many studies. Dividing a tissue into evenly-sized small biopsies to limit the number of clones present was performed in the study of sun-exposed epithelium[2]. More recently, laser capture microdissection has been paired with an ultra-low input DNA library preparation method to make high quality sequencing libraries from as few as 100 cells[3,5]. This approach enables microdissection of specific histological structures such as colonic crypts or endometrial acini, which contain few cells but are highly clonal.

While microbiopsies work well for easily dissected tissue types, other tissues such as blood or brain require different approaches. Haematopoietic stem cells have been successfully expanded into colonies *in vitro*, enabling study of mutational dynamics of the single clone while providing abundant genetic material to make sequencing libraries[25,26]. A similar idea is applied to single cell derived organoids, such as those used in the study of healthy bronchial epithelium organoids[30]. The replication machinery of the cell is less error-prone than whole genome amplification methods, and although mutations arise *in vitro,* they can be largely corrected by removal of those with low variant allele frequency.

Ultimately, single cell genomic sequencing is the ideal solution for querying mutations in single cells within a mosaic population. While ongoing efforts are being made to optimize this technique, the drawback of the approach is the difficulty of whole genome amplification of DNA from a single cell, which is necessary to generate enough material to create a sequencing library. Commonly, a technique called multiple displacement amplification is used to amplify the genomic material from a single cell. However, this process generates many artefacts which are later difficult to distinguish from true mutations, and it also amplifies the genome unevenly leading to inadequate coverage of some regions[65]. To reduce these artefacts, some investigators have used *in silico* error correction methods to identify true variant calls in the

data. Lodato *et al* used a linkage-based approach to identify true variants in single neurons[31], while Zhang *et al* paired an optimized DNA amplification protocol with a mutation caller that adjusts for local amplification bias[44]. While these implementations do reduce the burden of artefacts, further improvement is needed before single cell sequencing can become the gold standard method for somatic mutation detection in polyclonal cell populations.

Finally, ultra-accurate sequencing methods enable detection of rare somatic variants with high certainty. The duplex sequencing method labels both strands of a DNA molecule with a unique barcode, allowing the sequences of the two strands to be sequenced separately and then compared in the resulting data[66]. A true mutation will be present on both the 5' and the 3' strand, while artefacts resulting from PCR amplification will presumably only be present on one strand, allowing for distinction of true variants. However, the probability of both strands being amplified, sequenced, and represented in the downstream data is low and requires a significant amount of sequencing depth for a given genomic region. Therefore, the efficiency of this approach is low and its application is limited to small genomic targets[67]. A related method called BotSeqS (bottleneck sequencing system) aims to reduce this difficulty by introducing a simple dilution step of the barcoded library prior to PCR amplification[68], which produces a random sampling of template molecules while increasing the likelihood of retrieval of the 5' and 3' sequences from each of them. This results in a set of high-confidence variant calls, but it can only be used to survey the mutational landscape rather than to thoroughly assess the mutations present in a given genomic region.

For the genomic profiling of lymphocytes and glandular epithelium from primary Sjögren's syndrome patient biopsies discussed in this thesis, we enriched for clonal populations by laser capture microdissection and by fluorescence activated cell sorting (described in Methods section).

II. AUTOIMMUNE DISEASE

Autoimmune diseases are a diverse set of disorders ranging from mild to life-threatening that occur when the immune system attacks the host's own tissues. Collectively, they affect about

5% of the population[6] and are the third most common cause of morbidity after cancer and heart disease in the West[69]. Despite numerous interrogations into the pathophysiology, the mechanisms by which immune tolerance is lost and autoimmunity develops remain obscure. It has long been hypothesized that autoimmune diseases occur as a combination of genetic predisposition and environmental triggers such as infection, though this has not been proven. Genome-wise association studies (GWAS) have identified variants associated with autoimmune diseases, most prominently in the HLA locus, but also in other genes converging on key immune system pathways[70]. They have identified overlapping patterns of inheritance associated with distinct clusters of autoimmune disease, but the inherited landscape has been described as distinctly polygenic, with individual variants contributing low odds ratios[70,71]. Therefore, with the exception of certain rare early-onset disorders for which there is a monogenic cause[72], the heredity of most common autoimmune diseases remains complex and difficult to translate into a functional understanding of pathophysiology. Therefore, in this thesis I examine the that hypothesis that, in addition to the effect of predisposing inherited variants, somatic mutations may play an important role in the pathogenesis of autoimmune disease.

An important feature of many autoimmune diseases that is not well understood is the distinctly higher prevalence in women. Differences in sex hormones and expression of genes from the X chromosome have been suspected[73], but still no definitive explanation exists. Additionally, many autoimmune diseases confer an increased risk of lymphoma development[74–76]. It is suspected that lymphocytes, the primary cells driving autoimmune pathology, accumulate somatic mutations over the course of the disease and are the ones to develop into malignancy. Primary Sjögren's syndrome is the autoimmune disease we have chosen for investigating the somatic mutation hypothesis due to its high predisposition to lymphoma and pronounced female predominance, in addition to the availability of tissue biopsies. The overview below will outline the current understanding of the aetiology of primary Sjögren's syndrome to contextualize the aims of this thesis.

II.1 Primary Sjögren's syndrome: overview

Primary Sjögren's syndrome (PSS) is a chronic, systemic autoimmune disease that is characterized by immune-mediated destruction of exocrine glands, also known as autoimmune epithelitis. It manifests with hallmark "sicca symptoms", i.e. severe dryness of the eyes and mouth known as keratoconjunctivitis sicca and xerostomia. The implications of oral and ocular dryness include difficulty swallowing and susceptibility to infections and tooth cavities[77]. Dryness, joint pain, and fatigue, which can be debilitating, are a characteristic clinical triad present in around 80% of PSS patients[78]. Systemic complications are observed in 30-40% of patients and include peripheral neuropathy, kidney disease, vasculitis, and lung disease[78,79]. Primary Sjögren's syndrome is defined as occurring on its own, whereas secondary Sjögren's syndrome co-occurs with another autoimmune disorder, such as rheumatoid arthritis or systemic lupus erythematosus[77]. PSS has a peak incidence around 50 years of age and occurs with a highly skewed sex ratio of 9:1 in women versus men[75]. It is one of the most common systemic autoimmune diseases, second only to rheumatoid arthritis, with a population prevalence of around 0.5%[74,76]. PSS confers the highest risk of lymphoma among the common autoimmune diseases, constituting a 44-fold increased risk[80] or 5-10% lifetime chance of developing the malignancy[74,81]. The lymphomas are mainly marginal zone B-cell non-Hodgkin lymphomas occurring in the mucosa-associated lymphoid tissues (MALT lymphoma), most often in the major salivary glands[80].

While there is no cure, treatment for PSS includes alleviation of sicca symptoms, and in more advanced cases, immunosuppressive therapy. Biological therapies targeting B cell activity and other inflammatory pathways are under investigation in clinical trials.

II.2 Diagnosis of primary Sjögren's syndrome

Diagnostic criteria for PSS include clinical confirmation of sicca symptoms, serological testing for systemic autoimmunity, and histopathologic evidence of lymphocytic infiltration in labial salivary glands. Diagnosis does not require all the criteria to be fulfilled; presence of clinical symptoms paired with significant autoantibody findings in the blood may be sufficient for diagnosis, obviating the need for minor salivary gland biopsy[82]. PSS is a seropositive

autoimmune disease, commonly occurring with characteristic autoantibodies in the blood, including antibodies to Ro/SSA and La/SSB antigen (present in 60-80% of patients), as well as rheumatoid factor and antinuclear antibodies. In the absence of a strongly positive antibody test, patients with suspected PSS frequently undergo biopsies of the minor (labial) salivary glands for histopathologic analysis to determine the presence of lymphocytic infiltrates. The degree of infiltration is assigned a focus score by the number of lymphocytic foci (dense aggregates of lymphocytes) per 4 mm^2 of tissue, each of which is defined as 50 or more mononuclear lymphoid cells adjacent to normal-appearing mucous acini. The histopathologic criterion for a positive biopsy is a focus score ≥1, termed focal lymphocytic sialadenitis[82]. Minor salivary gland biopsies from PSS patients also show progressive atrophy and destruction of glandular acini, fibrosis, and duct dilation.

II.3 Association studies: genes and epigenomes

The last decade has seen many insightful findings contributing to the understanding of PSS. Genome-wide association studies (GWAS) have identified several key pathways, most notably in the HLA region (*HLA-DQB1, HLA-DRA, and HLA-DQA1)*[83]. Multiple independent studies have highlighted associations of PSS with polymorphisms in the immune regulating genes *IRF5, STAT4,* and *TNFAIP3*[84,85]. Furthermore, loci in other immune-relevant genes such as *BLK, IL12A, CXCR5, TNIP1, PRDM1, GTF2I, KLRG1, SH2D2a,* and *NFAT5*[83] have been identified*. BLK* mediates activation of B cells through the B cell receptor; *PRDM1* promotes plasma cell differentiation; HLA/MHC implies antigen-presentation and T cell involvement, and *TNFAIP3* is critical for control of the NF-kB signalling pathway which is activated in multiple immune cell types. Additionally, genome-wide methylation studies have found interferon-regulated genes to be hypomethylated in PSS, which correlated with their increased gene expression in B cells[86]. Other genes identified by GWAS were commonly affected by hypomethylation as well. Recently, miRNAs have come into focus as being associated with PSS, especially miR-30b-5p[87], which has been found to have a role in stimulating B cells through the B cell activating factor (BAFF).

II.4 B cells and plasma cells in PSS

B cells have long been considered key players in autoimmune disease. They have several roles including the production of cytokines, antigen presentation, and secretion of autoantibodies. Autoantibody production is a characteristic feature of PSS, similarly to rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE). PSS patients also tend to have higher counts of plasmablasts and immunoglobulins in peripheral blood (even though lower total lymphocyte counts are common)[78], suggesting a role for antibody production in disease. For this reason, much attention has been focused on B cells and CD4 helper T cells as mediators of the autoantigen-driven humoral immune response. Additionally, there are indications of T independent B cell activation, as demonstrated by the increased presence of marginal zone type B cells in blood and salivary glands of PSS patients[88]. Lymphomas arising as a complication of PSS in salivary glands are often of the marginal zone type[89], underscoring the importance of this B cell subset. In addition to infiltration of B cells and plasma cells, elevated expression of the B cell attracting chemokine CXCL13 is also observed in PSS minor salivary gland tissue[90].

In PSS, the two most commonly found autoantibodies target the ubiquitous ribonucleoproteins Ro/SSA and LA/SSB (Sjögren's-syndrome-related antigens A and B). A shared, or “public,” clonotypic autoantibody against Ro/SSA has been identified in the serum of several patients with PSS, consisting of an IGHV3-23 immunoglobulin heavy chain paired with a Vk3-20 light chain[91]. A longitudinal study of this clonotype in patients showed that it is constantly replenished by new clonal variants, suggesting sustained, cyclic B cell activation rather than production of autoantibodies by long-lived plasma cells activated at the onset of disease[92]. However, a direct functional role for these autoantibodies in the pathogenesis of PSS has not been established. It is thought they may contribute to disease indirectly through the formation of immune complexes. The presence of immune complexes can stimulate the production of interferons, which in turn additionally stimulate B cells, resulting in a sustained feedback loop of immune activation[74].

Another common autoantibody in PSS is rheumatoid factor (RF), which was originally discovered in rheumatoid arthritis[93]. It is reactive against the Fc portion of IgG

immunoglobulins. Production of RF by marginal zone B cells bearing RF surface Ig is stimulated by the presence of immune complexes[94]. RF is pathogenic in cryoglobulinemia that can develop as a complication of Sjögren's syndrome, where it forms a crystallizable immune complex that precipitates in tissues at low temperature, often resulting in vasculitis or kidney disease[95]. There are two known public idiotypes of RF, with the "Wa" idiotype (IGHV1-69/IGHJ4 heavy chain paired with a IGKV3-20) being more common than the "Po" idiotype (IGHV3-7/IGHJ3 and light chain IGKV3-15)[96,97]. The presence of RF factor has been associated with higher PSS disease activity and increased risk of lymphoma[98], and it is thought that RF-positive B cells are frequently the ones to undergo lymphomatous transformation[99].

Salivary glands are thought to be a site of significant antibody production in PSS[74]. The infiltration of salivary glands by B cells and plasma cells increases with the progression of disease, and in about 25% of PSS biopsies, ectopic germinal centre-like structures are found in the glands[90,100]. These structures are characterized by lymphoid organization into a dark zone and a light zone, along with a follicular dendritic cell network supportive separation into B and T enriched areas and T-B cell interactions conducive to the processes of antigen-driven somatic hypermutation, receptor editing, and isotype-switching. Some studies have associated the presence of germinal centre-like structures with more advanced disease, higher autoantibody production, and higher risk of lymphoma[81].

BAFF is a crucial cytokine that promotes proliferation and differentiation of B cells and has been repeatedly associated with PSS[74]. Increased levels of BAFF are found in PSS patient serum and salivary gland biopsies[101], and higher serum BAFF levels correlate with higher levels of autoantibodies[102]. Additionally, transgenic mice that overexpress BAFF develop autoimmune disease with sialadenitis (inflammation of salivary glands) and are predisposed to lymphoma[103]. BAFF is characteristically produced by myeloid cells in response to type I and type II interferons (IFNs). However, it can also be produced by other cell types including infiltrating lymphocytes and salivary epithelial cells[101], the latter suggesting involvement of the affected tissue in promoting an aberrant immune response. Further evidence of B cell involvement is suggested by elevated levels of additional B cell activating cytokines in serum or salivary glands of PSS patients, such as IL-14, IL-21, and others[74].

Despite much optimism about treatment of PSS by B cell depletion with rituximab, a monoclonal antibody targeting the CD20 B cell surface marker, the therapeutic effects observed in clinical trials were disappointing[104,105]. Reasons for this are complex and likely due to the involvement of multiple different cell types, including pathogenic plasma cells lacking CD20, which are not targeted by the drug. Future clinical trials may focus on administration of B cell-targeting therapies earlier in the course of disease, as well as in combination with therapeutics targeting other cell types and pathways.

II.5 Interferons and the innate immune response

Importantly, an 'interferon signature' has been observed in peripheral blood and minor salivary glands of PSS patients[106]. Biochemical and transcriptomic studies have characterized the presence of type I and type II IFNs in blood[107] and tissue[108,109], and cells that produce type I IFN (IFN-α) have been detected in salivary gland biopsies[110]. Elevated levels of interferon indicate activation of the innate immune system, which can in turn incite an adaptive immune response. The physiological role of interferons is primarily in the defence against viral infection. IFN-α can be produced by a virally infected cell to protect surrounding cells in a paracrine manner, or it can be produced in large quantities by plasmacytoid dendritic cells (PDCs) which detect microbial nucleic acids and activate Toll-like receptor (TLR) signalling to produce IFN-α. PDCs have been detected in minor salivary glands of PSS patients[110], as well in affected tissues of other autoimmune diseases[106].

Interferons can induce the production of chemokines, which recruit immune cells to the site of inflammation. They can also activate dendritic cells, T cells, and B cells and thereby initiate an adaptive immune response. Type I and type II IFNs have been shown to promote the production of B-cell stimulating cytokine (BAFF), which is why BAFF is considered to be an important link between the innate and adaptive immune response[78,94]. Interestingly, it has been observed that individuals given IFN-α for the treatment of viral infection or cancer frequently develop autoantibodies and sometimes distinct autoimmune disease[111], which is another line of evidence pointing to the role of the interferon system in the loss of immune tolerance.

II.6 Quantitative immunophenotyping: identifying key cell types in PSS

In a recent immunophenotyping study, Mingueneau *et al* used cytometry by time-of-flight to profile immune cell populations in paired blood and minor salivary gland biopsies of PSS patients[112]. This approach was more quantitative and unbiased than previous studies using histology and immunohistochemistry (IHC) to study cell types in affected tissue. The conclusions were two-fold: a blood cell-type signature associated with PSS, and a description of predominant cells types in tissue that correlate with disease activity. In the blood, total numbers of CD4 T cells, memory B cells, and plasmacytoid dendritic cells were decreased, while the numbers of activated HLA-DR+ CD4, activated HLA-DR+ CD8 cells, and total plasmablasts were increased in PSS patients (markedly in those with anti-SSA antibodies) as compared to non-PSS controls. Analysis of PSS minor salivary gland biopsies revealed that most infiltrating cells are CD3 T cells, which are present in non-PSS glands as well but at much lower abundance. The infiltrating T cells comprised both the CD4 and CD8 subtypes, but only the CD8 compartment had a significant number of HLA-DR+ activated T cells. This suggests that perhaps cytotoxic CD8 T cells have a more prominent role in disease than previously appreciated and require further investigation. Patient biopsies also had significantly elevated levels of fully differentiated plasma cells. Additionally, the salivary epithelial cells of PSS patients upregulated HLA-DR in comparison to non-PSS control epithelium, indicating a possible antigen presenting role of these cells.

II.7 Salivary epithelium, microbial defence, and hormones

Epithelial cells that form acini and ducts of the salivary glands are targets of the autoimmune response, but they are also emerging as active participants in the immunopathology. The epithelium is a first line of defence against microbial pathogens and as such has to carefully balance the induction of a sufficient immune response in the event of pathogen invasion without inducing excessive cytotoxicity that would significantly damage the tissue. If this balance is perturbed, it is plausible that immune autoreactivity could develop as a consequence. Indeed, salivary epithelial cells proximal to heavy immune infiltration in PSS have been shown to express high levels of molecules involved in immune stimulation and recruitment. Studies using in-situ expression and long-term cultured PSS salivary gland

epithelial cells have shown constitutive upregulation of MHC class I and II molecules, costimulatory molecules involved in T cell activation, BAFF, Toll-like receptors (TLR-3, TLR-7, TLR-9), proinflammatory cytokines (IL-1, IL-6, TNFα, and other), and various chemokines (CCL3, CXCL13, CXCL-9, and others)[90,113]. These findings suggest that the perpetual abundance of cytokines observed in salivary glands of PSS patients comes in part from the epithelium itself, which is likely an active participant, not a passive target, in the process of constitutive antigen presentation and immune activation.

Elevated levels of Toll-like receptors in the salivary glands of patients point to activation of the innate immune system. In particular, constitutive expression of TLR3 is markedly increased in PSS salivary gland epithelial cells[113,114]. The primary role of TLR3 is response to viral infection by detection of viral dsRNA molecules, which leads to activation of the innate immune response and type I IFN production. In the NZB/W F1 mouse model of lupus, TLR3 stimulation has been shown to promote accelerated sialadenitis[115], suggesting a possible viral-driven, TLR3-mediated route for initiation of autoimmunity. This complements a long-standing theory regarding the pathogenesis of PSS, which suspects viral infection as a trigger for disease. Much attention has been focused on Epstein-Barr virus (EBV) as a likely candidate, which has been demonstrated to promote the release of Ro/SSA and La/SSB ribonuclear proteins, the targets of canonical autoantibodies in PSS, by apoptosis of epithelial cells[74,116]. EBV small RNA can also form complexes with La/SSB that lead to interferon production through TLR3 activation[109]. Though these findings point to the involvement of EBV as a viral trigger, no causal link to PSS has ever been proven.

Multiple studies have tested the association of PSS with other viruses as well, including cytomegalovirus, hepatitis C, and Coxsackie A virus, to name a few, but none of the candidates held up to replication and validation[67]. The difficultly may lie in the inability to detect the virus causing initial infection by the time PSS has reached a clinical course. Evidence for this is perhaps in the infection of Fas-deficient LPR mice, which have lupus-like autoimmunity but not salivary gland involvement, with cytomegalovirus (CMV). The mice developed sialadenitis three months after CMV infection, but the virus was no longer detectable by that point[117]. Alternatively, it is plausible that the trigger for PSS is the activation of endogenous

retroviruses rather than an exogenous pathogen[118]. This possibility deserves further examination, as it could be an overlooked stimulus for innate immune activation.

The stark female preponderance of PSS is poorly understood. Given the average age of onset at 40-50 years, hormonal factors that change significantly during menopause could play a role. Since salivary epithelial cells express oestrogen receptors, the functional responsiveness to oestradiol was assessed in cultured salivary gland epithelial cells from PSS patients and controls[119]. The study found that pre-treatment of normal epithelial cells with 17-beta-oestradiol impeded the downstream effects of IFN-γ, namely the upregulation of CD54, and that this functionality was significantly reduced in PSS epithelial cells, where IFN-γ signalling remained unimpeded. It therefore seems that in this context, oestrogen has an anti-inflammatory role in the normal salivary gland. In a separate study, female wildtype mice after ovariectomy began to show apoptosis of salivary gland epithelial cells with lymphocytic infiltration similar to that in Sjögren's syndrome[120], suggesting a role of oestrogen in the maintenance of normal salivary gland function and a link to autoimmune disease.

II.8 Clonal evolution and B cell lymphomas complicating Primary Sjögren's syndrome

The most serious complication of PSS is lymphoma. Most of the lymphomas associated with PSS are non-Hodgkin B cell lymphomas that develop in salivary glands (MALT lymphoma), which are usually indolent and low-grade but sometimes transition into more fulminant diffuse large B-cell lymphoma (DLBCL)[89]. Clinical and biochemical predictors of lymphoma in PSS patients include salivary gland swelling, lymphadenopathy, lymphocytopenia (usually of CD4 T cells), low complement levels, cryoglobulinemia, monoclonal gammopathy or paraprotein, and others[89]. It is suspected that lymphoma develops as a consequence of constitutive immune stimulation with autoantigens in salivary glands, in the very same B cells that are the effectors of the autoimmune response. In other words, it is likely that the autoimmune B cell that is constantly activated gains oncogenic mutations that allow it to escape from proliferative control.

Lymphomas are largely of marginal zone (MZ) B cell origin, which means that maturation of these B cells occurs independently of T cell help and outside of a germinal centre. Rheumatoid factor-producing B cells are often the MZ type, producing IgM antibodies in response to immune complexes presumably formed from other autoantibodies such as anti-Ro/SSA and anti-La/SSB, which are produced through antigen-driven B cell selection and T cell help[94]. The idea that RF+ B cells are the ones that transition to lymphoma is corroborated by evidence of rheumatoid factor-positivity in a large proportion of MALT lymphomas in PSS[99]. Autoreactivity of lymphomas to Ro/SSA or La/SSB has not been observed.

The most strongly associated gene with PSS-related lymphomas is *TNFAIP3*, which encodes for the TNFAIP3 (A20) protein, a key regulator of NF-kB activation downstream of TNF-family receptors. The NF-kB pathway is central to the activation of B cells[121] and other immune cell types, and germline mutations in *TNFAIP3* are associated with several autoimmune and inflammatory diseases[122,123]. TNFAIP3 has been associated with PSS by GWAS studies[83], and salivary glands from PSS patients were found to have lower levels of A20 and correspondingly higher NF-kB activity, compared to controls[124]. Importantly, *TNFAIP3* mutations are commonly seen in MALT lymphoma, and in a recent study 77% of PSS-associated lymphomas analysed had either germline or somatic mutations in *TNFAIP3* predicted to be functionally deleterious[125].

It is speculated that the evolution of the B cell response in salivary glands begins with a polyclonal infiltration, which over time may become increasingly monoclonal as antigen-specific B cells are selected to proliferate[88]. The line between monoclonal B cell infiltration and lymphoma is a blurry one, as it can be difficult to differentiate a benign lymphoproliferative lesion from low-grade lymphoma. Therefore, this process could be viewed as a spectrum of B cell progression: from polyclonal, to monoclonal, to lymphomatous.

II.9 Disease pathogenesis: hypotheses and future directions

A proposed hypothesis of disease aetiology in PSS suggested by current knowledge is as follows. The initial stimulus for disease in a predisposed individual is likely to be a viral

infection, which induces the production of interferons and apoptosis of glandular epithelial cells leading to expulsion of autoantigens such as Ro/SSA and La/SSB. Interferons activate the adaptive immune response, which includes the production of antibodies against the released autoantigens and their formation into immune complexes that further perpetuate immune activation. Due to the constitutive presence of self-antigens, and perhaps a genetic background and hormonal environment that enhance immune function, the immune response continues to be perpetuated. What is initially a polyclonal B cell infiltration into salivary glands becomes a more evolved immune response favouring antigen-selected B cells, in some cases by germinal centre-like affinity maturation, leading to a more monoclonal population of cells over time. On this background of chronic stimulation and selection, it is likely that lymphomatous transformation occurs upon mutation accumulation and immune checkpoint dysfunction in the constitutively activated autoimmune B cells.

Although clinical, molecular, and *in vivo* studies have yielded ample valuable insights, there remain many unknowns in the proposed PSS pathogenesis model. The presence of antibodies against Ro/SSA and La/SSB are a hallmark of PSS, yet a direct role for antibodies in disease pathogenesis has not been proven. It is unknown why autoantigens that are present in all cells are associated with tissue-specific autoimmunity. A viral trigger for disease, though heavily speculated, has not been proven. There has been interest in the effects of oestrogen on the immune system and on exocrine glands, yet its role remains complex and elusive. Understanding of the immune cell types involved in pathogenesis is rapidly evolving with new technology. It is increasingly appreciated that the affected tissue is a complex glandular milieu where various cell types of the innate and adaptive immune system interact, and further studies are needed to understand their contributions. GWAS studies have highlighted some immune pathways, however genetic predisposition has been difficult to pinpoint and most instances of PSS remain seemingly sporadic and late-onset (around middle-age)[78].

It is this latter concept that we sought to explore through the work described in this thesis. Given a seemingly sporadic midlife onset in (mainly female) individuals with little evidence of heredity, can this phenomenon be, at least to some extent, attributable to stochastic mutational processes that lead to the breakdown of immune tolerance? Early evidence in

favour of this hypothesis from recent studies of somatic mutations in related autoimmune disorders is described below.

II.10 Somatic mutations in autoimmune disease

The concept of this thesis, i.e. the examination of somatic mutations as a contributor to primary Sjögren's syndrome, is situated in a wider context of emerging interest in the role of somatic mutations in health and disease outside of cancer. As an exploration into chronic immune disorders, several groups have made promising early findings of somatic mutations and clonal expansions in autoimmune and inflammatory diseases.

A study by Savola *et al* in 2016 investigated T cells in the peripheral blood of newly diagnosed rheumatoid arthritis (RA) patients and detected somatic mutations and distinct clonal expansions[126]. By flow cytometry-based T cell receptor (TCR) assessment, it was apparent that CD8 T cells were more clonal than CD4 T cells, in the blood of both patients and controls, with patient clones tending to be larger and increasing with age. In 5 out 25 patients, somatic mutations were detected in CD8 clones, but none were found in CD4 clones. The mutated genes included *SLAMF6* and *IRF5*, which have been previously associated with autoimmune disease, as well as several proliferation-associated genes. Transcriptomic profiling by RNA sequencing showed that the mutated clones upregulated cellular proliferation pathways, indicating activation of these cells. This study not only demonstrates the presence of CD8+ T cell clonal expansions associated with somatic mutations which may promote cellular survival and/or proliferation, but also highlights a possibly important and overlooked role of CD8 T cells in autoimmune disease, which warrants further investigation.

Recently, an investigation of rheumatoid factor-producing cells by Singh *et al* discovered lymphoma-associated somatic mutations in these autoimmune B cell clones[127]. The study followed four patients with Sjögren's syndrome who developed mixed-type cryoglobulinemia as a complication. While the presence of RF is thought to often be benign, in mixed cryoglobulinemia it is directly pathogenic by formation of precipitable immune complexes. By autoantibody peptide sequencing, the investigators found that three of the patients had the public "Wa" idiotype of RF, which has been previously described and consists of IGHV1-

69/IGHJ4 heavy chain paired with a IGKV3-20 light chain. The fourth patient had the less common "Po" idiotype consisting of IGHV3-7/IGHJ3 and light chain IGKV3-15. Memory B cells with these receptor rearrangements were enriched from peripheral blood and processed for single cell DNA and RNA sequencing. The findings revealed somatic mutations in lymphoma-associated genes in all four patients, as well as V(D)J somatic mutations that showed evolution from benign RF-producing clones to those producing pathogenic RF, as shown in functional assays. The mutated genes included *CARD11, TNFAIP3,* and *KLHL6.* The significance of CARD11 in autoimmunity has previously been suggested by a mouse study where gain-of-function mutations in *CARD11,* analogous to those seen in DLBCL lymphomas, were found to be a switch that allows autoreactive B cells to evade anergy and immune checkpoints, resulting in activation and proliferation[128]. *TNFAIP3* loss-of-function mutations observed were similar to those found in lymphoma and disinhibited downstream NF-κB activation. The driver mutations mostly preceded V(D)J somatic mutations in B cells, and *KLHL6* mutations were especially found to increase the accumulation of V(D)J mutations, suggesting a KLHL6 mutator phenotype and raising the possibility of regulation of somatic hypermutation by KLHL6[127].

The relevant and important findings of this study demonstrate a long-suspected connection between the pathogenesis of autoimmune disease and that of lymphoma, lending further credence to the somatic mutation hypothesis proposed in this thesis. Identification and purification of autoreactive lymphocyte clones is a key challenge that the authors were able to overcome by selecting B cells with public clonotypes and performing single cell genomic sequencing. However, identifying pathogenic B and T cells remains a challenge in many autoimmune disease contexts where disease-associated clonotypes are unknown or difficult to isolate. In this dissertation, we focus on resident lymphocytes in affected tissue as a potential reservoir enriched in pathogenic cells.

That somatic mutations should be found more commonly in B cells than other immune cell types is suggested by evidence of off-target activity of the physiological hypermutation machinery, AID (activation induced cytidine deaminase). AID is activated as part of the germinal centre reaction in order to introduce variation to the B-cell receptor complementarity determining regions (CDRs), and in doing so creates B cells with higher affinity for binding antigen. However, this process has been demonstrated to promiscuously

introduce mutations outside of the immunoglobulin locus, genome-wide[129]. These off-target AID mutations have been found in lymphoma-associated genes, and AID mutagenesis is associated with malignant transformation in B cells[130]. As previously mentioned, a specific mutational signature dominated by T>G transversions is ascribed to AID and has been observed in B-cell lymphomas as well as in non-malignant post-germinal centre B-cells[44].

Disease-relevant somatic mutations in autoimmune conditions needn't be confined to immune cells. Recent studies have demonstrated somatic mutations in the IL-17 signalling pathway in colonic epithelial cells from inflammatory bowel disease (IBD) patients[56–58]. In this context, the expansion of mutated clones that are less sensitive to inflammatory stimuli suggests an adaptation that evolved under the selective pressure of inflammatory damage. A similar mechanism may be operative in the synovial microenvironment in rheumatoid arthritis (RA). A study in 1997 by Firestein *et al* first observed recurrent *TP53* somatic mutations in RA synovial tissue[131], a finding reproduced in subsequent years[132], though not with less biased NGS methods. These mutations are thought to arise in an environment of inflammatory oxidative stress and induce a proliferative synovial phenotype with potentially bone-invasive properties characteristic of the joint damage observed clinically. Once established, the selective advantage and invasive properties of these mutated synoviocytes could allow them to propagate joint damage independently of inflammation. This hypothesis lends itself to other inflammatory diseases which may operate through a similar mechanism of clonal selection in affected tissue. If proven, the mechanistic scenarios proposed in the IBD and RA studies have the potential to transform our understanding and approach to treatment of inflammatory disorders.

Thesis aims

Guided by current evidence of somatic mutation dynamics in normal tissue and chronic diseases, we sought to examine the somatic mutational landscape of immune cells and affected tissue in primary Sjögren's syndrome. The main aims of this study that will be described in the three results chapters include:

1. Examination of infiltrating lymphocytes in minor salivary glands to assess clonality and presence of lymphoma-associated driver mutations
2. Transcriptomic profiling of infiltrating lymphocytes to examine population trends and correlations with somatic mutation findings
3. Comparison of the genomic landscape of epithelial cells in minor salivary glands from PSS patients and non-PSS controls

Chapter 2: Methods

Contributions

The project described in this thesis was collaborative in nature. Here I provide an overview of the contributions of others and myself, which will be outlined further in the corresponding method descriptions to follow.

The idea to do deep sequencing on sorted lymphocyte populations from minor salivary gland biopsies to look for somatic mutations was conceived by Matthew Collin and Peter Campbell. The design of a bait set targeting key lymphoma and immune system genes was done by Paul Milne, Anthony Fullam, Matthew Collin, and Peter Campbell. Fresh tissue and blood samples were obtained from the Newcastle University clinic, courtesy of Fai Ng. Snap frozen minor salivary gland tissue for laser capture microdissection was obtained from the Newcastle University biobank. All sample processing, flow cytometry, and cell sorting was performed by Paul Milne. Tissue sectioning was performed by Yvette Hooks at the Sanger Institute. Tissue staining, immunohistochemistry, and laser capture microdissection were performed by me.

Library preparation was done by the Sanger Institute Core Pipelines, and standard alignment to the genome was done by the Sanger Institute Core Informatics team (with the exception of SmartSeq2 single cell data alignment, which I performed myself). Library preparation, sequencing, and alignment of 10X Genomics single cell RNA data was done at Newcastle University by Paul Milne, Jason Lam, and Anastasia Resteu. Whole genome and whole exome somatic variant calling were done through the Cancer, Ageing, and Somatic Mutation standard pipeline facilitated by the core bioinformatics team. Filtering to remove bovine DNA contamination was done by Kathryn Beale and Mark Emery in the core bioinformatics team, and myself.

All other bioinformatic analyses were carried out by me, unless specified otherwise in the following text. Statistical models, scripts, and advice were generously provided by colleagues in the group to assist with analysis. In particular, a filter to remove artefacts specific to the low input library preparation protocol was written by Mathijs Sanders. A variant filtering

approach using binomial models was written by Tim Coorens, as was a sensitivity adjustment model for mutational burden. An algorithm to map mutations onto phylogenetic trees was written by Nick Williams. Filters for Shearwater variant calls and several plotting scripts were provided by Federico Abascal and Inigo Martincorena. A pipeline for processing and analysing single cell RNA sequencing data was provided by Raheleh Rahbari.

Interpretation of findings and design of subsequent experiments was done by me, through discussion with my PhD supervisors, Peter Campbell and Richard Siegel, and our Newcastle University collaborators, Paul Milne and Matthew Collin.

Sample management and coordination of sequencing was done by the Cancer, Ageing, and Somatic Mutation administrative and lab support teams, with special thanks to Laura O'Neill and James Hewinson.

I. SAMPLES AND WET LAB METHODS

I.1 Minor salivary gland biopsies and blood samples

Minor salivary gland tissue was obtained from individuals undergoing biopsies based on suspicion of primary Sjögren's syndrome (PSS) at the Newcastle University clinic. Upon biopsy, individuals were diagnosed with either PSS, early or possible PSS, or non-PSS sialadenitis. Part of each biopsy was used for diagnostic purposes and part was donated to our research endeavours, with approval from the UK Research Ethics Committee and written informed consent from patients. Fresh biopsies were obtained from 55 patients for DNA sequencing and 16 patients for single cell RNA sequencing. Additionally, 23 snap frozen biopsies were obtained from the Newcastle Biobank for laser capture microdissection. Matched peripheral blood samples were obtained from a subset of patients. The sex and age characteristics of the cohort matched the general demographics of PSS patients, who are predominantly female and middle-aged. All samples were obtained in collaboration with Matthew Collin and Fai Ng at Newcastle University.

I.2 Processing of tissue and blood

To disaggregate minor salivary gland biopsies, an enzymatic digestion protocol was used by Paul Milne. The tissue was minced with a scalpel, then incubated with a 1:1000 dilution of collagenase (type 4, Worthington Biochemical Corp.) for 3 hours to digest. The disaggregated cell suspension was filtered (Sysmex CellTrics 100 µm filter) before the downstream application of cell sorting.

Where matched peripheral blood was available, mononuclear cells were isolated by density gradient centrifugation (Lymphoprep, StemCell) before cell sorting.

I.3 Isolating lymphocyte populations from minor salivary gland tissue

Fluorescence activated cell sorting (FACS) of single cell suspensions from minor salivary gland biopsies was performed by Paul Milne. After extensive prior flow cytometry analysis of PSS minor salivary glands with numerous cell markers, a subset of the markers was selected for sorting the samples used for this project. The antibodies used to sort lymphocytes were against the following cell surface markers: CD45, CD3, CD19, CD4, CD8, CD38, and HLA-DRA. The lymphocyte subsets defined by these markers included the following:

B cells: $CD45^{+}CD3^{-}CD19^{+}CD38^{-}$

Plasmablasts: $CD45^{+}CD3^{-}CD19^{+}CD38^{+}$

Plasma cells: $CD45^{+}CD3^{-}CD19^{-}CD38^{hi}$

CD4 T cells: $CD45^{+}CD3^{+}CD19^{-}CD4^{+}$

CD8 T cells: $CD45^{+}CD3^{+}CD19^{-}CD8^{+}$

Antigen-presenting cells: $CD45^{+}CD3^{-}CD19^{-}HLADRA^{+}$

Sorting was done on a BD FACSAria Fusion instrument, with up to 5,000 cells per lymphocyte subset sorted into tubes for the low-input DNA sequencing protocol. The cells were then centrifuged, sorting medium was aspirated, and cell pellets were directly resuspended in Arcturus PicoPure proteinase solution for cell lysis. Cells were incubated in lysis buffer for 3 hours at 65C and 10 minutes at 75C, per manufacturer instructions. Cell lysates were then frozen and shipped to the Sanger Institute for library preparation and DNA sequencing.

For plate-based single cell RNA sequencing, cell sorting was done by FACS with the same antibodies used above, except the cells were sorted into single wells of a 96-well plate. Cells were sorted directly into a bespoke lysis buffer developed at the Sanger Institute single cell facility, which allows for the extraction of both DNA and RNA from the same cells and contains the internal ERCC gene control spike-in. A total of 50 plates from 10 minor salivary gland biopsies were sorted in this way.

For droplet-based single cell RNA sequencing, $CD45^+$ immune cells were enriched from the digested minor salivary gland biopsies by magnetic bead selection, prior to single cell library preparation with the 10X Genomics protocol.

I.4 Histology and immunohistochemistry of minor salivary gland biopsies

For the purposes of laser capture microdissection, 23 snap frozen biopsies of minor salivary gland tissue were obtained from Newcastle University, courtesy of Fai Ng and Paul Milne. The biopsies were fixed in PAXgene ethanol-based solution (PAXgene Tissue FIX, Qiagen; contains no formalin). The fixed tissue was paraffin-embedded and cut into 10µm thick sections by Yvette Hooks at the Sanger Institute. Tissue sections were mounted onto slides covered with a polyethylene naphthalate membrane (required for laser capture microdissection) and left to dry at room temperature overnight.

To stain the tissue sections for morphological features, I used a standard haematoxylin and eosin protocol. Slides were sequentially immersed in xylene twice for 2 minutes, 100% ethanol twice for 1 minute, deionized water for 1 minute, Gill's haematoxylin for 15 seconds, tap water twice for 20 seconds, eosin for 10 seconds, tap water for 15 seconds, 70% ethanol twice for 20 seconds, and neo-clear xylene substitute twice for 15 seconds. Once dry, slides were mounted with temporary cover slips for high resolution scanning, which was done on a Hamamatsu NanoZoomer S60 instrument. The temporary cover slip was removed before laser capture microdissection.

To stain subsets of infiltrating immune cells with chromogens, I used a dual-stain immunohistochemistry approach (ImmPRESS Duet Double Staining HRP/AP Polymer Kit,

Vector Labs). Red and brown stains were used to identify lymphocyte markers on 10µm tissue sections mounted on polyethylene naphthalate membrane slides. Tissue sections were stained to highlight the following combinations of markers: CD4/CD8, CD8/CD20, and CD20/CD38. The Vector Labs ImmPRESS Duet protocol was followed, with some slight adaptations. The antigen retrieval step, which involves a high temperature incubation that would have negative downstream implications for sequencing, was omitted. Secondly, the ImmPRESS Duet DAB stain was replaced with a slightly less sensitive DAB stain from Vector Labs (SK-4100) to resolve the issue of non-specific background staining. Primary antibodies were incubated for 3 hours at room temperature. They were titrated to find the optimal dilutions, which were the following:

CD4 (MA5-12259, ThermoFisher)	**1:100**
CD8 (RM-9116-S0, ThermoFisher)	**1:200**
CD20 (NCL-L-CD20-L26, Leica Biosystems)	**1:200**
CD38 (ab108403, Abcam)	**1:200**

After dual staining was complete, tissue sections were counterstained with diluted haematoxylin (Vector Labs) to visualize morphological structures. A temporary cover slip was appended, and slides were scanned at high resolution on a Hamamatsu NanoZoomer S60 scanner.

I.5 Laser capture microdissection

Laser capture microdissection (LCM) was used to dissect small features from polyethylene naphthalate membrane slides with H&E and immunohistochemistry staining, using a Leica LMD7 instrument. Individual acini and ducts of the minor salivary gland were dissected from H&E stained sections of 10µm thickness. Acini and ducts are small structures that required Z-stacking of cuts from multiple adjacent sections to achieve the goal of minimum ~100 cells per well, an approximate threshold of success for the low-input DNA library preparation protocol. Lymphocyte aggregates were dissected from immuno-stained sections to enrich populations of CD20 B cells, CD38 plasma cells, CD4 T cells, and CD8 T cells.

Individual cuts fell into wells of a 96-well plate; images were taken before and after each cut. The dry dissected samples were then resuspended in 20µl of Arcturus PicoPure lysis and incubated for 3 hours at 65C and 10 minutes at 75C, per manufacturer instructions. Cell lysates were frozen and stored for library preparation.

I.6 Library preparation and sequencing

A bespoke, low-input library preparation protocol for DNA sequencing, using as little as 100 cells, was developed at the Sanger by Peter Ellis and team[133],[3,5]. This protocol was used for all DNA sequencing, including sorted lymphocyte populations and samples obtained by laser capture microdissection, and was carried out by the Sanger Institute Core Pipelines team. Starting from cell lysates obtained by incubation with Arcturus PicoPure proteinase buffer (intended for low-input samples), downstream library preparation steps were carefully optimised to maximise yield. This involved digestion with restriction enzymes instead of acoustic shearing to fragment the genomic material, which has been shown to decrease loss of DNA[133]. Additionally, genomic DNA was not initially purified from cell lysates, instead DNA in the lysate was bound to magnetic beads and enzymatic fragmentation were carried out directly on the bead-DNA mixture to minimise loss[133]. Standard Illumina sequencing adapters were then ligated and PCR amplification was performed in a standard way. DNA quantification was done after library preparation, and depending on sufficient concentration of libraries, samples were taken forward for targeted gene pull-down by hybridization. The recommended minimum library concentration needed for pull-down was 10 ng/µl, while the recommended minimum library concentration for whole genome sequencing was 5 ng/µl to obtain 10-15X coverage.

Paired-end 2x75bp sequencing of targeted DNA libraries was done on an Illumina HiSeq2000 or HiSeq4000 instruments, with 8 samples per lane. Paired-end 2x150bp sequencing of whole genomes was done on Illumina HiSeq4000 or Illumina NovaSeq instruments, with a target depth of 30X per genome.

For plate-based single cell RNA sequencing, library preparation was done from single cell lysates in bespoke buffer developed at the Sanger Institute Single Cell Facility for stabilization and dual extraction of RNA and DNA. This buffer contained ERCC synthetic spike-ins which

serve as internal gene controls. For the purposes of this project, we sequenced only the RNA of single cells, not DNA. Reverse transcription of mRNA was performed from single cell lysates using the Smart-Seq2 template-switching protocol that allows subsequent PCR amplification of full-length cDNA molecules[134]. Library preparation was then performed using the Nextera XT protocol (Illumina). Single cell processing from lysates and library preparation was done by the Sanger Institute Single Cell Facility. Paired end 2x75bp sequencing was performed on an Illumina HiSeq4000 with one half-plate per lane, or equivalent on Illumina NovaSeq.

For droplet-based single cell RNA sequencing, the 10X Chromium 5' Gene Expression with V(D)J enrichment protocol (10X Genomics) was used for library preparation of CD45+ sorted immune cells from minor salivary gland biopsies. This was done by Paul Milne and Jason Lam at Newcastle University. In brief, individual cells were combined with gel beads containing unique cellular barcodes and enzymes by a droplet-generating approach. The uniquely barcoded single cells underwent cDNA amplification. The cDNA libraries were split in two, with one half used for PCR-based V(D)J enrichment and subsequent library preparation, and the other half used for library preparation of 5' enriched total transcripts. V(D)J and 5' gene expression libraries were sequenced separately, on an Illumina NextSeq instrument, at Newcastle University.

II. DRY LAB METHODS

II.1 Design of a bait set for targeted sequencing

A targeted gene panel was designed to include known lymphoma driver genes, immune regulatory and checkpoint genes, V(D)J regions, HLA loci, polymorphic sites, and microRNAs associated with primary Sjögren's syndrome. This list of genes was curated from the literature by Matthew Collin, Paul Milne, Peter Campbell, and Anthony Fullam, and designed by Agilent SureDesign software (https://earray.chem.agilent.com/suredesign/). A pool of biotinylated probes for hybridization-based pull-down of these genes was made by Agilent Technologies. The full list of genes can be found in Appendix Table 1.

II.2 Alignment of sequencing data

All DNA sequencing data, including whole genome, whole exome, and targeted sequencing was aligned to the NCBI build37 human genome by the Sanger Institute Core Informatics team, using the BWA mem algorithm[135]. Reads were inspected for quality and coverage. After processing data from bulk sorted lymphocyte samples, I discovered a large excess of somatic variants in some samples. By performing a BLAST (Basic Local Alignment Search Tool)[136] query of reads harbouring multiple variants, it became apparent that the reads strongly aligned to bovine DNA sequences, suggesting contamination with genetic material from cow. The source of this contamination was traced back to the foetal bovine serum added to the buffer used for fluorescence-activated cell sorting. To remove this artefact, we implemented the Xenome algorithm, which is designed to remove xenograft contamination[137]. Xenome realigned all reads from a given sample to human and cow genomes and separated them into the categories “human”, “cow”, “ambiguous”, or “neither”. For downstream analysis, only reads aligning exclusively to the human genome (NCBI build37) were kept. The Xenome algorithm was run by Kathryn Beale and Mark Emery in the Cancer, Ageing, and Somatic Mutation core bioinformatics team.

For in-depth quantification of sequencing depth and breadth (which refers to the fraction of regions covered at a given depth, for example 65% of a genome being covered at 15X), I used the Mosdepth tool[138]. For targeted sequencing and whole exome sequencing data, I used Mosdepth to calculate the average depth per bait-targeted region. For whole genome data, average depth was calculated across 1 kilobase regions of the genome.

Single cell RNA sequencing data from the plate-based Smart-Seq2 platform was processed by me, after initial demultiplexing of samples by Sanger Institute Core Informatics. Fastq files were trimmed to remove adapter sequences using the Cutadapt algorithm[139]. Reads were then aligned to the NCBI build37 genome using splice-aware mapping with the STAR aligner[140]. A count matrix of reads per gene per cell was generated with the FeatureCounts tool[141].

Single cell RNA sequencing data from the droplet-based 10X Genomics platform was initially processed at Newcastle University by Anastasia Resteu. Raw fastq files were demultiplexed and aligned to the NCBI build37 genome genome build in a splice-aware manner, and a count matrix was generated, both using the accompanying 10X software, CellRanger. V(D)J data was also analysed by CellRanger to identify the B and T cell repertoires of individual cells.

II.3 Caveman single nucleotide substitution calling and filtering

Calling of somatic single nucleotide substitutions in whole genome and whole exome sequencing data was performed using the Cancer Variants through Expectation Maximization (CaVEMan) algorithm[142]. Caveman uses an expectation maximization approach to call variants by comparing "tumour" (in this case, lymphocyte) and normal reads to a reference genome and calculating a probability of the genotype at each base, assuming given copy number for tumour and normal samples. To maximize sensitivity, major copy number was set to 5 and minor copy number set to 2, as this empirically generated the best results in normal (non-cancerous) samples. The algorithm also integrates base quality, read position, and read orientation to calculate sequence error rates at each position. Caveman is part of the Cancer, Ageing, and Somatic Mutation standard variant calling pipeline and was run through an interface facilitated by the core bioinformatics team.

For laser capture microdissection samples, Caveman variant calling was performed using a synthetic "normal" sample and subsequently removing germline variants shared by all samples from a given donor. For bulk sorted lymphocyte samples, Caveman was run against a matched fibroblast normal sample.

To filter the variant calls from Caveman, firstly BWA mem mapping artefacts were removed by filtering the median alignment score (ASMD ≥ 140) and soft clipping score (CLMP=0, meaning fewer than half the reads are clipped). Secondly, all DNA samples in this project underwent library preparation using the bespoke low-input protocol, which uses an enzymatic fragmentation step (instead of acoustic sonication) to cut DNA into smaller fragments. This enzymatic digestion step was found to introduce a specific mutational artefact to the sequence data, which results from the processing of cruciform DNA structures

by the enzymatic digestion and manifests as an excess of variants in regions of inverted repeats capable of forming hairpin structures. To remove this artefact, Mathijs Sanders devised a filtering method based on proximity of the variant to the alignment start site and standard deviation or median absolute deviation of the variant position in the supporting read[3,5], features that correlated strongly with presence of the artefact. Thirdly, because the enzymatic library preparation protocol produced shorter reads than standard sonication protocols, paired end reads sometimes overlapped and resulted in the same variant being counted twice. To overcome this, Mathijs' filtering method calculated fragment-based statistics instead of read-based statistics (after marking PCR duplicates), and retained variants supported by at least 3 high quality fragments (alignment score ≥ 40 and base scores ≥ 30). Variants were annotated using ANNOVAR[143]. Variant allele frequencies were recalculated based on number of mutant and wildtype reads at each site, across all samples from a given donor, using a pile-up method developed at Sanger, vafCorrect (https://github.com/cancerit/vafCorrect).

For Caveman variants from laser capture microdissection samples, which were called against an unmatched synthetic normal sample, additional filters were required to remove germline variants. A germline filter was devised by Tim Coorens, using an exact binomial model which classified variants as germline or somatic based on variant allele frequency across all samples from a given donor. An additional filter created by Tim removed false positive variants by a beta binomial over-dispersion model, which typically removed sequencing artefacts present at low frequency in a majority of samples in genomic regions prone to noise.

II.4 Pindel insertion/deletion calling and filtering

Calling of somatic small insertion and deletion events in whole genome and whole exome sequencing data was performed using the cgpPindel approach[144]. This entails detection of read pairs where one is mapped and the other is unmapped or a split read, then performing remapping of query reads to identify putative indel sites. Empirically derived post-processing filters are then applied, which include filters for strand bias, existence in normal sample, highly repetitive small repeats, and sufficient depth. Variants are annotated as passing or failing the respective filters. For laser capture microdissection samples, cgpPindel was called

against an unmatched normal, then subsequently filtered to remove germline variants, as done with Caveman calls. For bulk sorted samples, cgpPindel was run against matched fibroblast samples. cgpPindel is part of the Cancer, Ageing, and Somatic Mutation standard variant calling pipeline and was run through an interface facilitated by the core bioinformatics team.

For subclonal samples with modest depth, such as many bulk sorted lymphocytes and some laser capture microdissection samples, one of the default cgpPindel filters proved too stringent. This filter, which requires at least 5 reads supporting each indel call, was found to remove many true somatic indels, and therefore calls failing this requirement were rescued. Through previous analyses of normal tissue, others in the group discovered that a significant proportion of indel calls passing the aforementioned filters were artefacts comprised of 1 base pair indels in homopolymer regions of 10 or more bases. A filter to remove these indel calls was applied.

To remove germline indel calls in laser capture microdissection samples which were run against an unmatched control, a similar approach to filtering single nucleotide substitutions was applied. To determine accurate variant allele frequencies in all samples from a given donor, vafCorrect (https://github.com/cancerit/vafCorrect) was used. An exact binomial model, developed by Tim Coorens, was then applied to classify indels as somatic or germline based on variant allele frequency of a site across all samples.

II.5 Shearwater variant calling and filtering

To call single nucleotide substitutions and small insertion/deletion events in targeted sequencing data from bulk sorted lymphocytes, the Shearwater algorithm was used[145]. Shearwater has a higher sensitivity for calling subclonal variants than Caveman and has been shown to perform well with targeted deep sequencing data[2]. The Shearwater algorithm uses a beta binomial model to compute local error rates from a panel of normal samples and prior knowledge, which are then used to derive the likelihoods of true variants in samples of interest. The Shearwater algorithm is based on the deepSNV R package

(https://github.com/im3sanger/deepSNV)[145]. Advise and scripts for running Shearwater were provided by Federico Abascal and Inigo Martincorena.

To run Shearwater on targeted sequencing of bulk lymphocytes, I attempted two configurations. In the first, I used all biopsy-derived fibroblast samples as a bulk normal panel for calling mutations in tissue lymphocytes and blood lymphocytes. In the second, I used both fibroblasts and blood samples as the normal panel, calling mutations in only tissue lymphocytes. By comparing the mutations in tissue lymphocytes from the two approaches and visually inspecting variants using the JBrowse read viewer[146], I found that the second approach yielded a more accurate list of variant calls, with fewer artefacts. The reason for this is that having a larger panel of normal samples provides a more robust local error model for calling variants. To ensure that I was not missing variants that might be present both in tissue and in blood lymphocytes from a donor, I called variants in blood samples separately, using only fibroblasts as the normal panel. The overlap of the variants in blood and tissue from a given donor almost entirely consisted of germline variants that were not properly removed when calling against the fibroblast normal panel. Therefore, I proceeded with the cleaner set of variants in tissue lymphocytes produced by calling against both fibroblasts and blood as the normal panel.

BWA mem mapping artefacts were removed by filtering reads on their median alignment score (ASMD ≥ 140) and soft clipping score (CLMP=0, meaning fewer than half the reads are clipped), which was shown to remove erroneous variant calls. Each variant called by Shearwater also had an associated p-value based on the local error model; these p-value were adjusted by a Benjamini-Hochberg multiple testing correction, and only those with a corrected p-value < 0.05 were considered further. Additionally, variants were required to have one supporting read from each strand. A *KDM6A* variant that failed the adjusted p-value cut-off (likely due to having only two supporting reads) was rescued post-hoc as a likely real variant during manual inspection of the *KDM6A* gene, which I did after discovering truncating variants in this gene in other samples. Similarly, two synonymous variants in *KDM6A* were removed on manual inspection due to alignment of the reads to the cow genome; these reads constituted a small fraction that were missed by the Xenome algorithm used to filter bovine DNA contamination. Finally, some germline variants were carried through these filtering

steps, so an additional filter based on an exact binomial model was used to remove them (as described in section II.3 above, written by Tim Coorens).

II.6 Detection of driver mutations and selection

For discovery of cancer driver mutations, two approaches were used: manual annotation and selection analysis using a dN/dS approach. To be considered a cancer driver by manual annotation, coding variants were required to be found in multiple tumour samples in the COSMIC (Catalogue of Somatic Mutations in Cancer)[147] database, have a high predicted deleteriousness by CADD score (Combined Annotation Dependent Depletion)[148], and not be located at a common polymorphic site[149]; alternatively, any protein-truncating mutation (nonsense, frameshift, or essential splice) in a known tumour suppressor gene described in the literature was considered a likely driver.

Selection of genes across a set of samples was inferred by the ratio of nonsynonymous to synonymous variants (dN/dS) using the dNdScv algorithm[150] (https://github.com/im3sanger/dndscv). The dNdScv method compares the observed ratio of nonsynonymous to synonymous variants to the expected ratio under no selection pressure, adjusting for local mutation rates and processes across coding regions. The algorithm calculates the likelihood of missense, nonsense, and essential splice variants compared to a neutral model, using trinucleotide mutation contexts to build a robust local mutation rate model. Positive selection of a gene is indicated by an excess of nonsynonymous variants, while negative selection (which is less common) is indicated by a higher proportion of synonymous variants.

The dNdScv method was implemented for both targeted and whole genome datasets. In the targeted lymphocyte dataset, dNdScv was used to evaluate selection of only the genes sequenced. In the whole genome lymphocyte dataset, dNdScv was used to detect selection both genome-wide, across all coding genes, as well as restricted to the same genes sequenced in the lymphocyte targeted dataset. For the epithelial sample dataset, variants found by whole genome and whole exome sequencing were evaluated for selection across all coding

genes as well as restricted to a set of 892 known cancer genes curated from TCGA (The Cancer Genome Atlas)[151], which can be found in Appendix Table 2.

II.7 Analysis of mutational burden

Genome-wide mutational burden is a metric that has been shown to vary across tissues[2,5], with exposures[10,30], and with disease state[11,58]. I used a linear mixed effects model to evaluate the effects of age, sex, and diagnosis on mutation burden in the genomes of salivary epithelial samples.

The number of variants detected in a sample is dependent on the coverage and the clonality of the sample. A sample with low coverage and low clonality will miss many true somatic variants due to low sensitivity of detection. To adjust for the variability in coverage and clonality of samples in my dataset, I used an adjustment calculation devised by Tim Coorens. A sensitivity parameter for each sample was calculated based on its median variant allele frequency (VAF) and mean coverage. Given that the Caveman variant calling algorithm requires four reads to call a variant, the sensitivity is the probability of observing four or more mutant reads at a given VAF based on a Poisson-distributed coverage given the mean coverage. This was calculated using the Poisson distribution function in R, "rpois", with 100,000 iterations. The observed number of variants in a sample was divided by the sensitivity parameter for the sample to obtain an adjusted number of variants, which was used in the linear mixed effects model.

The linear mixed effects model was used to evaluate effects of age, sex, and diagnosis on mutational burden, while accounting for interpatient variation (non-independent sampling per patient). Patient samples with missing metadata were omitted. The model was first constructed using all three variables, then sequentially dropping those variables determined not to have a significant effect, until only the significant variables remained. This was done using the "lmer" package in R.

II.8 Copy number and structural variant calling

Somatic copy number alterations in whole genome data were called using the ASCAT algorithm (Allele-Specific Copy number Analysis of Tumours)[152]. ASCAT calls copy number changes in a tumour sample using a matched normal as a control and accounting for tumour aneuploidy and tumour purity (which refers to "contamination" of a tumour sample with normal tissue). For sorted lymphocyte samples, ASCAT was run using a matched fibroblast control. For epithelial samples, a matched lymphocyte control was used if available, otherwise a different epithelial sample was used.

Additionally, the Battenberg algorithm[153] was also used to profile genome-wide copy number changes. Battenberg is especially useful for detecting subclonal copy number changes, making it appropriate for our data. However, Battenberg failed in a number of samples due to low tumour purity and coverage.

For targeted and whole exome sequencing data, CNVkit[154] was used to detect copy number changes. CNVkit was run using a panel of normal fibroblast samples as a reference to evaluate changes in the sorted lymphocyte populations. For the whole exome epithelial samples, which did not have matched controls, copy number calling was done using a "flat" reference panel of neutral copy number for each target gene region.

Structural variants in whole genome data were called against a matched normal, where available, using the BRASS (Breakpoints via Assembly) algorithm, developed at Sanger (https://github.com/cancerit/BRASS).

II.9 V(D)J repertoire assembly

For V(D)J assembly of T and B cell receptor sequences in whole genome, targeted sequencing, and single cell Smart-Seq2 RNA sequencing data, I used the MiXCR algorithm[155]. MiXCR has analysis modules specific to the input data type, allowing assembly from genomic and targeted DNA, bulk and single cell RNA, or PCR-enriched V(D)J sequences (https://mixcr.readthedocs.io/en/master/#). MiXCR reconstructs fragmented V(D)J sequences, assembles clonotypes, identifies productive rearrangements, detects mutations

in germline sequences, and corrects PCR errors. The number of unique rearrangements detected in bulk sorted lymphocyte populations correlated to the estimated oligoclonality observed by mutation detection, and the single clones observed in single cell RNA data validated the ability of the algorithm to detect an accurate number of rearrangements.

T and B cell repertoires from 10X Genomics single cell RNA sequencing data were obtained using the accompanying 10X software, CellRanger.

II.10 Phylogenetic tree construction

Phylogenetic trees of single nucleotide substitutions called in laser capture microdissection samples were constructed using the MPBoot software[156]. MPBoot uses the maximum parsimony principle to compute branch support, with an adaptation allowing for ultrafast bootstrapping. To construct phylogenies, mutations from all samples in a given donor were re-genotyped by the VAFcorrect approach, as described in the Caveman filtering protocol. Variants with variant allele frequency (VAF) > 0.3 were denoted as present and annotated "1", those with VAF < 0.1 were denoted as absent ("0"), and those in between 0.1 and 0.3 were denoted as ambiguous; this excluded private subclonal variants from the tree building process. The input variants were bootstrapped 1,000 times to construct trees, and nodes with confidence less than 50 were collapsed into polytomies. This was conducted with advice and scripts from Tim Coorens and is similar to an approach used in previous work[3]. Branch lengths were determined by the number of assigned substitutions, through a script written by Nick Williams.

II.11 Mutational signature extraction

Mutational signatures are characteristic patterns of mutation created by different mutagens, which can be distinguished by their nucleotide patterns in DNA. The COSMIC[40] database (Catalogue of Somatic Mutations in Cancer) defines multiple signatures of exposure and endogenous mutagens based on the mutation type and trinucleotide context, meaning the adjacent 5' and 3' bases, resulting in 96 mutational categories.

To identify the single nucleotide substitution signatures present in my dataset, mutational signature analysis was performed using a method based on the Hierarchical Dirichlet Process[157] (https://github.com/nicolaroberts/hdp) to estimate underlying signatures and their contribution to a sample.

This approach is a nonparametric Bayesian clustering method that does two things simultaneously. Starting with a matrix of counts per mutational category per sample, the algorithm groups mutations into clusters to define a library of signatures. Then, the probabilistic contribution of each signature to each sample is estimated.

Samples are manually arranged into a hierarchy (e.g., grouped first by diagnosis, then donor, then cell type) such that samples within a group are assumed to be more similar to each other. An initial signature library is defined by randomly assigning mutational categories to each of a random number of signatures (clusters). The likelihood of these signature definitions is assessed by comparison to the observed mutational frequencies, taking into account similarities between samples as defined by the hierarchical structure, and updated iteratively through a set number of cycles. Contribution of each signature to each sample is determined by the resulting posterior distribution.

The algorithm can identify *de novo* signatures as well as detect known signatures by incorporating known signatures' mutation distributions as fake data nodes, which makes real data more likely to be drawn into a common cluster with them. For mutational signature analysis of my dataset, lymphocyte and epithelial samples were grouped by tissue, cell type, and diagnosis, and the signature extraction was run using 30 known signatures from the COSMIC database included as priors. The approach identified matches to the known signatures as well as *de novo* signatures.

II.12 Detection of viral elements

To detect viral nucleotide sequences in whole genome epithelial and lymphocyte samples, I used the GOTTCHA metagenome analysis tool[158]. GOTTCHA breaks up input unmapped and split fastq reads into smaller fragments and aligns them to a library of viral genomes. Unique

segments of viral genomes at several taxonomic levels, including strain, genus, and species, were used for classification of input reads. The output was filtered to remove low confidence hits and only keep those in human-tropic viruses, excluding bacteriophage viruses.

II.13 Single cell RNA expression analysis

Analysis of single cell RNA sequencing data, from both Smart-Seq2 and 10X Genomics platforms, was performed using the Seurat suite of tools in R[159,160]. Raheleh Rahbari provided scripts and advice for analysis. The input to the pipeline was a raw matrix of counts per gene per cell, generated by the FeatureCounts[141] tool (as described in the Alignment section of this chapter).

The counts matrix was initially filtered to exclude cells with fewer than 500 detected genes ("features") and to exclude genes expressed by fewer than 5 cells. Meta data was added to the Seurat object, which included patient number, diagnosis, sex, age, and cell type (referring to phenotype data from fluorescence activated cell sorting, performed prior to Smart-Seq2 library preparation). Further filtering was done based on number of total features detected, removing those with more than 10,000 in the full-length cDNA (Smart-Seq2) dataset to get rid of "doublets", where two cells are sequenced instead of one. In the 10X Genomics dataset, which was comprised of shorter reads, fewer features were detected (as expected), and the threshold of maximum number of features was set to 2,500. Cells were then filtered based on mitochondrial content, removing those with more than 10% of reads derived from mitochondrial genes. A high proportion of mitochondrial gene expression indicates low quality, apoptosis, or cellular stress[161].

In the full-length cDNA dataset, the percentage of synthetic spike-in *ERCC* control genes was assessed as an additional cell quality metric. Normally, high *ERCC* content indicates low quality cells, however in this dataset it was discovered that an excess amount of *ERCC* spike-in was added during library preparation, resulting a high proportion of *ERCC* transcripts across many cells. To assess whether the high *ERCC* content obfuscated biological signal in the data, I performed principal component analysis (PCA) to determine the contribution of *ERCC* genes to variability in the dataset. It turned out that *ERCC* genes did not contribute greatly to the

main principal components. Downstream analysis demonstrated that it was possible to correctly identify cell types based on gene expression, as validated by the cell phenotype information from fluorescence-activated cell sorting. Therefore, sufficient biological transcripts were detectable, and it was possible to proceed with analysis despite the high proportion of *ERCC* reads.

After quality control and filtering, read counts were normalized to 10,000 counts per cell and log-transformed. This meant that normalization was carried out by dividing the read counts per gene for each cell by the total number of reads for the cell and multiplying by a scaling factor of 10,000, after which the values were log-transformed. Highly variable genes were then identified by calculating average expression and dispersion of genes across the dataset. Next, counts were scaled and centred, and a linear transformation was applied to reduce the effects of unwanted variation, such as the *ERCC* content (only in the full-length cDNA dataset), with the goal of improving downstream dimensionality reduction and clustering[162].

After normalization and scaling of counts, linear dimensionality reduction was performed by principal component analysis. The statistical significance of principal components as contributors to variation was evaluated to select the number of components used in downstream steps. Non-linear dimensionality reduction, t-stochastic neighbour embedding (tSNE), was then performed using the identified principal components. tSNE projections were used for visualization of dataset variability and cell-to-cell relationships. To identify groups of similar cells, a graph-based clustering approach called shared nearest neighbour analysis (SNN)[163] was used, which utilized the previously identified principal components to calculate distance between cells and define clusters. The number of identified clusters depended on the number of principal components used, as well as a resolution parameter. The resolution parameter was empirically set to 0.8 in the Smart-Seq2 dataset and 0.6 in the 10X dataset.

After defining clusters, differential gene expression analysis was conducted to identify the cell types and phenotypic states of cells that occupy the clusters. A non-parametric Wilcoxon rank sum test was used to compare gene expression between clusters. The differentially expressed genes were then manually annotated to identify cell types by comparing marker genes and expression signatures to those found in the literature.

Chapter 3: DNA sequencing of lymphocytes from minor salivary gland biopsies

With emerging data describing cancer-driver mutations existing in normal tissues and reshaping their landscapes, we hypothesized that somatic mutational processes might underlie chronic diseases as well. We chose to explore this hypothesis in primary Sjögren's syndrome, an autoimmune disease characterized by middle-age onset and a high predisposition to lymphoma. This chapter describes my study of the somatic mutational landscape of tissue-infiltrating lymphocytes in primary Sjögren's syndrome (PSS), in order to understand whether somatic mutations may play a role in this disease.

The ideal experimental scenario for somatic mutation study in autoimmune disease would be to isolate and sequence lymphocyte clones that are known to be autoreactive and mediate the pathogenic immune response. However, this presents a significant technical challenge in PSS, most notably due to the lack of ability to define and isolate autoreactive cells. Such a feat may be possible in certain diseases where there exists a clearly identified antigen to which there is a monoclonal response, and there is a means of selecting the lymphocytes responsible for this response by the affinity of their B-cell or T-cell receptor. For example, in pemphigus vulgaris, antibodies target the cell junction protein desmoglein, which results in an autoantibody-mediated skin condition that causes loss of cell adhesion and painful blistering[164]. The B cells secreting these directly pathogenic autoantibodies have been identified and well characterized, and targeted CAR T-cell therapies are currently being developed for antigen-specific B cell depletion[165]. The situation is not nearly as clear in the case of many complex autoimmune diseases such as PSS, which have elusive molecular targets and multiple associated autoantibodies whose pathogenic roles remain dubious. This makes targeting autoreactive lymphocytes in PSS and related diseases more challenging.

As discussed in Chapter 1, B cells producing the canonical rheumatoid factor autoantibody have recently been successfully extracted from blood samples of patients with cryoglobulinemic vasculitis in the context of Sjögren's syndrome, in which rheumatoid factor

is proven to be directly pathogenic[127]. In less advanced disease, such as PSS without cryoglobulinemia, expanded clones producing pathogenic rheumatoid factor autoantibodies are less likely to be found in blood. Other autoantibodies associated with PSS, such as anti-Ro and anti-La, do not have a known pathogenic role in disease, and their B cell receptor sequences are less well defined [166]. Since there is therefore not a clearly identifiable pathogenic lymphocyte clone that can be targeted and isolated in early PSS, we focused our attention on a disease-affected tissue as a potential reservoir of autoimmune cells. Activated B cells and plasma cells have been detected in minor salivary gland biopsies of PSS patients and associated with auto-antibody production and disease severity[94,167]. Germinal-centre like structures are found in a subset of patient salivary glands and correlate with disease severity as well[81,100]. Disease-affected salivary glands are most often the location of discovery of Sjögren's syndrome-associated lymphomas, which are thought to develop from auto-reactive marginal zone B cells residing in the tissue[94,168]. Salivary glands are also infiltrated by a large number of T cells implicated in pathogenesis[169].

For these reasons, I undertook a genomic study of lymphocytes that invade minor salivary glands in PSS to investigate the presence of somatic mutations and their possible role in pathogenesis. To our knowledge, this is the first study to characterize somatic mutational dynamics and the genomic landscape of tissue-resident lymphocytes in autoimmune disease.

STUDY AIMS

The main aim of this study was employing targeted and whole genome DNA sequencing to survey the mutational dynamics of tissue-infiltrating lymphocytes. The specific questions we hoped to answer include the following:

1. *Are there somatic mutations present in lymphoma-associated genes, cell-cycle regulating genes, and immune-related genes in tissue-infiltrating lymphocytes?*

   I investigated the presence of somatic mutational events that may contribute to the fitness and loss of immune tolerance of a lymphocyte clone, and/or set it on the path to lymphoma transformation. While the association of salivary gland B cell lymphoma

and PSS has been firmly established, the hypothesis of somatic mutations as a primary cause of the autoimmune pathology has not been investigated. Whether potential somatic mutations in lymphocytes are the cause of immune dysfunction or whether they are secondary to it are not mutually exclusive scenarios. Demonstrating the presence of somatic mutations in lymphocytes would be a pivotal first step towards a novel avenue of inquiry of autoimmunity in PSS.

2. *Is there clonal expansion of lymphocytes in affected minor salivary glands?*

   While there are reports in the literature of lymphocyte clonal expansion in PSS salivary glands, this has not been explored through next generation sequencing technologies. The clonality of tissue resident lymphocytes is an important consideration for disease pathogenesis, in order to understand whether potentially autoreactive cells are being produced at the site of disease, or whether affected tissue is merely an inflamed milieu attracting a diverse population of circulating lymphocytes.

3. *Is there an observable trend in the B-cell and T-cell receptor usage?*

   The immunoglobulin and T-cell receptor repertoire sequences of lymphocytes in PSS-affected minor salivary glands have been studied through classical methods such as PCR amplification of V(D)J segments. The tandem extraction of repertoire sequences and somatic mutation information from the same sample by DNA sequencing is a newer approach that permits us to infer correlations between the two findings. The repertoire sequences can be used to assess clonality and potential disease-associated trends in the usage of heavy chain and light chain genes.

To address the aforementioned aims, we devised two ways of extracting lymphocytes from minor salivary glands to maximize the possibility of capturing clonal populations. The initial approach entailed fluorescence activated cell sorting (FACS) of T and B lymphocyte subsets from salivary gland biopsies, paired with an ultra-low input library preparation method to yield sequencing libraries from each population. The secondary approach employed laser capture microdissection (LCM) to extract spatially sequestered lymphocyte clusters from

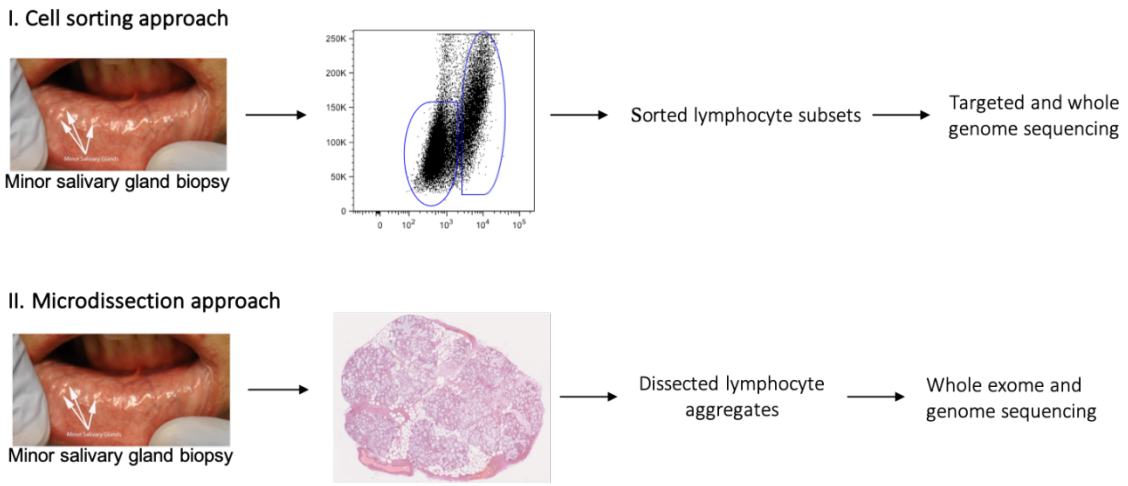


Figure 1. Overview of lymphocyte extraction approach from minor salivary gland biopsies.

histology sections of salivary gland tissue. (**Figure 1**) The results of each approach will be described separately in this chapter and then combined for further interpretation.

I. SEQUENCING OF SORTED LYMPHOCYTE POPULATIONS FROM MINOR SALIVARY GLANDS

Minor salivary gland biopsies are often undertaken to confirm a suspected diagnosis of primary Sjögren's syndrome, where patients suffer from sicca symptoms (dry eye and dry mouth), but serological findings are insufficient for diagnosis. We took advantage of this diagnostic practice to obtain minor salivary gland tissue for our research purposes. Patients were subsequently diagnosed as having confirmed PSS, early or possible PSS, or non-PSS sicca based on the degree of focal lymphocytic infiltration in salivary glands, in conjunction with other diagnostic criteria. It is important to note that while most biopsies from patients deemed negative for PSS appear histologically normal, the patients from whom they were obtained clinically presented with sicca symptoms and therefore may have salivary gland dysfunction of a different aetiology. Occasionally, it is also possible that a patient with a diagnosis of Sjögren's syndrome based on symptoms and serological tests has no immune cell infiltration on histological examination of minor salivary gland tissue.

I.1 Patient cohort

Minor salivary gland biopsies were obtained from 55 patients at the University of Newcastle, in collaboration with Fai Ng and Matthew Collin. Approval by the UK Research Ethics Committee and written informed consent from patients were granted for the use of these biopsies for research purposes. Biopsies were processed immediately and sorted into lymphocyte subpopulations by Paul Milne. The cohort included 35 biopsies with focal lymphocytic infiltration that confirmed a diagnosis of PSS, 11 that were histologically ambiguous and had suspected early PSS, and 9 that were negative for PSS. Of the 55 patients, 46 were female, and ages at the time of biopsy ranged from 23 to 91. (**Figure 2a,b**) This is representative of the broader demographics of PSS patients, ~90% of whom are female and most of whom are 40-50 years of age at the time of diagnosis[78].

I.2 Cell sorting of lymphocyte compartments

Upon enzymatic digestion of biopsy tissue to obtain single cell suspensions, fluorescence-activated cell sorting (FACS) was used by Paul Milne to separate cells into subpopulations by common haematopoietic and lymphoid surface markers (Methods). The main subsets

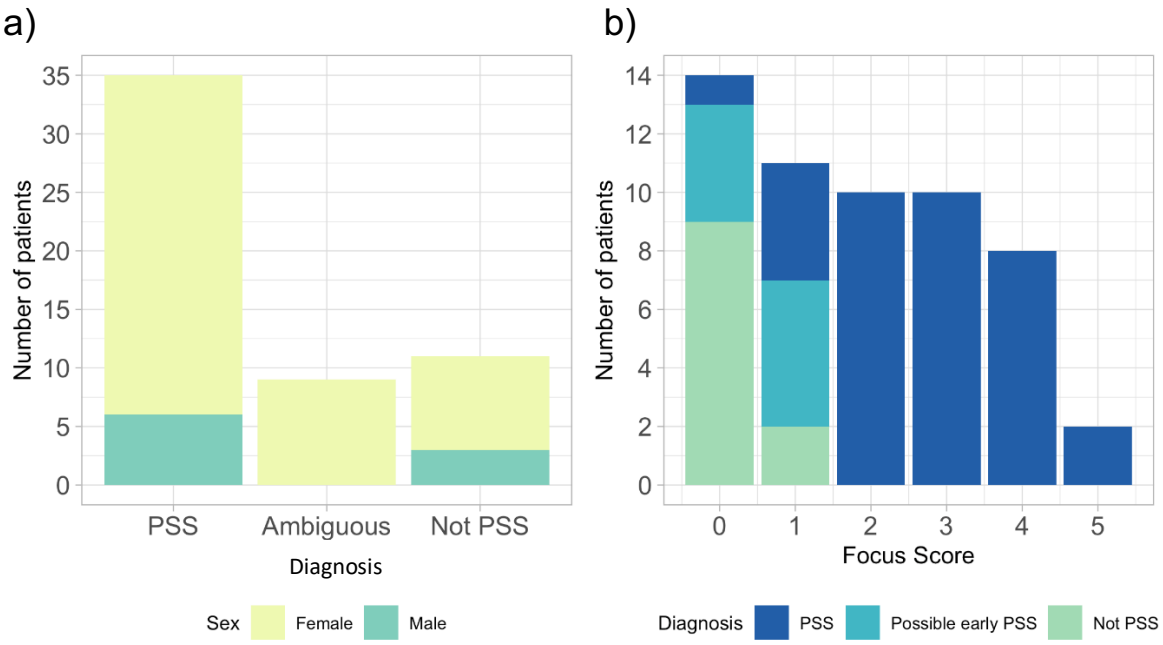


Figure 2. a) Patient cohort by diagnosis and sex. **b)** Patient cohort stratified by diagnosis and focus score.

captured were CD4 T cells, CD8 T cells, B cells, plasma cells, and plasmablasts (**Figure 3**). The sorted subsets were gated by FACS using the following cell surface markers:

B cells: $CD45^{+}CD3^{-}CD19^{+}CD38^{-}$

Plasmablasts: $CD45^{+}CD3^{-}CD19^{+}CD38^{+}$

Plasma cells: $CD45^{+}CD3^{-}CD19^{-}CD38^{hi}$

CD4 T cells: $CD45^{+}CD3^{+}CD19^{-}CD4^{+}$

CD8 T cells: $CD45^{+}CD3^{+}CD19^{-}CD8^{+}$

Antigen-presenting cells: $CD45^{+}CD3^{-}CD19^{-}HLADRA^{+}$

The number of total extracted immune cells varied between biopsies, from a few hundred to several thousand or more cells. Low cell counts precluded the extraction of all desired lymphocyte subsets from some biopsies, commonly biopsies from patients who were ultimately not diagnosed to be PSS. In these cases, for example, if there were not enough lymphocytes to sort both CD4 and CD8 T cells, then bulk $CD3^{+}$ T cells or a bulk $CD45^{+}$ sample was isolated instead.

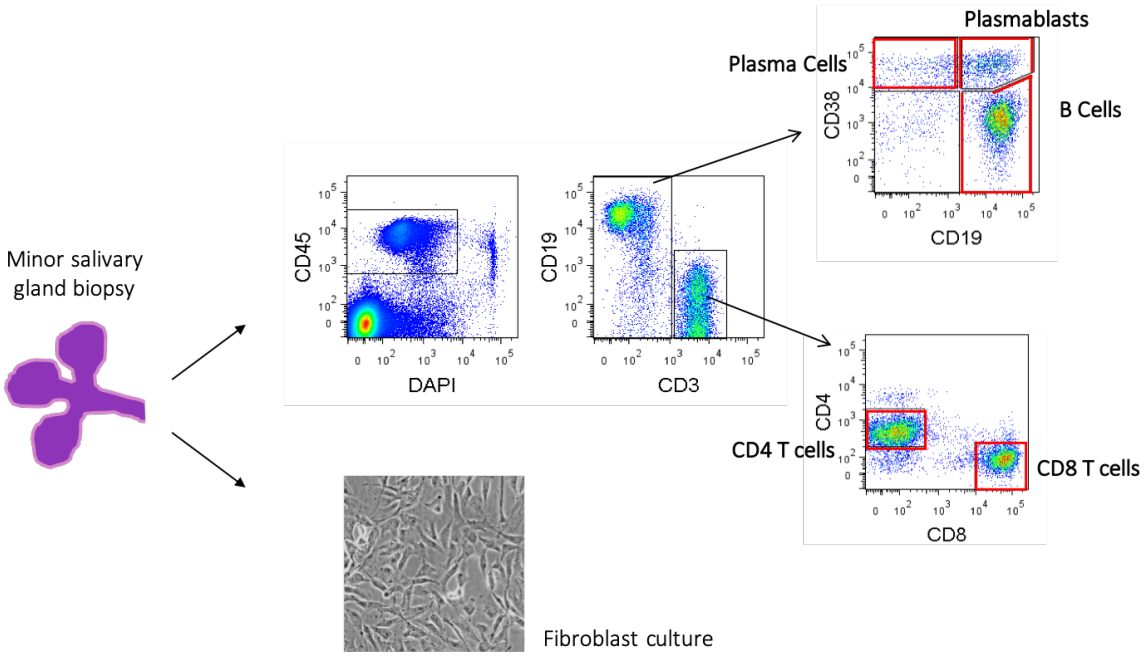


Figure 3. FACS separation of lymphocyte compartments, gating strategy. Figure courtesy of Paul Milne.

Table 1. Total number of sorted samples per category, across patients. Mean sequencing depth per sample type. Number of matching samples from blood (PBMC), per category. Sequenced subsets not referenced in this table include one or more samples from the following groups: CD45+HLADR+ cells, CD45+CD19-CD3-HLADR- cells, stromal cells, CD14+ monocytes from blood, and CD56+ NK cells from blood.

	B Cells	*Plasma Cells*	*Plasma-blasts*	*CD8 T Cells*	*CD4 T Cells*	*Bulk T Cells*	*Bulk B cells*	*Bulk Cell*	*Fibro-blasts*
FACS markers	CD19/CD38	CD19/CD38	CD19/CD38	CD3/CD8	CD3/CD4	CD3	CD19	CD45	NA
Number Samples	31	30	29	31	31	3	3	8	40
Seq Depth	111	124	98	130	109	160	324	269	169
Matched PBMC	22	0	15	20	22	0	0	0	0

I.3 Library preparation

The sorted cells were lysed and DNA libraries prepared through a bespoke in-house pipeline specially created at the Sanger Institute for low-input samples with 100-1000 cells (Methods). Some samples failed library preparation during the early stages of optimizing this pipeline, or they dropped out due to lack of enough genetic material to be sequenced at adequate depth. Samples used for library preparation contained no more than 1,000 cells. After creating adapter-ligated libraries, those with at least 10 ng/µl concentration were recommended for pull-down and enrichment of genes with the custom-designed bait set. However, due to the precious nature of patient samples, all samples with at least 5 ng/µl were submitted for targeted gene pull-down. This resulted in some samples with lower coverage but still adequate signal to detect mutational events with higher clonality. Of the total 55 patient biopsies obtained, library preparation and targeted sequencing yielded data from sorted lymphocytes of 31 patients and bulk B and T cells from an additional 3 patients (**Table 1**). Subsequent whole genome sequencing (described in section II of this chapter) was performed on remaining library material that did not undergo pull-down of target genes.

A subset of patients also had available blood samples, from which lymphocyte populations matching those from the biopsy were isolated. Additionally, fibroblasts were extracted from the stromal part of the gland biopsy and cultured *in vitro* by Paul Milne to produce a bulk population of non-immune control cells (**Table 1**).

I.4 Targeted sequencing (TGS) of bulk sorted lymphocytes

To investigate the presence of somatic mutations in key genes associated with lymphoma and immune dysregulation, we designed a custom gene panel for targeted deep sequencing. The panel covers genes curated from the literature and includes around 350 lymphoma and immune related genes, in addition to HLA loci, SNP sites, ncRNAs associated with PSS, and immunoglobulin and T-cell receptor V(D)J genes. The complete list of targeted genes can be found in the Appendix section (**Table A1**).

All samples underwent library preparation with a custom in-house pipeline for low DNA input and subsequent hybridization-based pull-down of the genes of interest prior to sequencing. Samples were aligned to the NCBI build37 human genome as described in the Methods section, with an extra step to remove contamination that was found to be caused by foetal calf serum used for FACS sorting. After removal of PCR duplicates, the mean depth of coverage across all regions of the targeted gene set was calculated to be 149X, meaning targeted regions were covered by 149 sequencing reads on average (**Figure 4**).

Calling of single nucleotide substitutions and small insertion/deletion events was performed using the Shearwater algorithm[2,170] as described in the Methods section, with fibroblast and matched blood samples used as a combined panel of controls. Variants called by Shearwater were filtered by an FDR-adjusted q-value cut-off of 0.01 and filters that removed reads with soft clipped bases and low alignment scores. An additional filter was used to remove remaining germline variants that were not automatically removed, which consisted of an exact binomial model that discriminates between germline and somatic variants based on sharing of variants between multiple samples from the same individual (Methods).

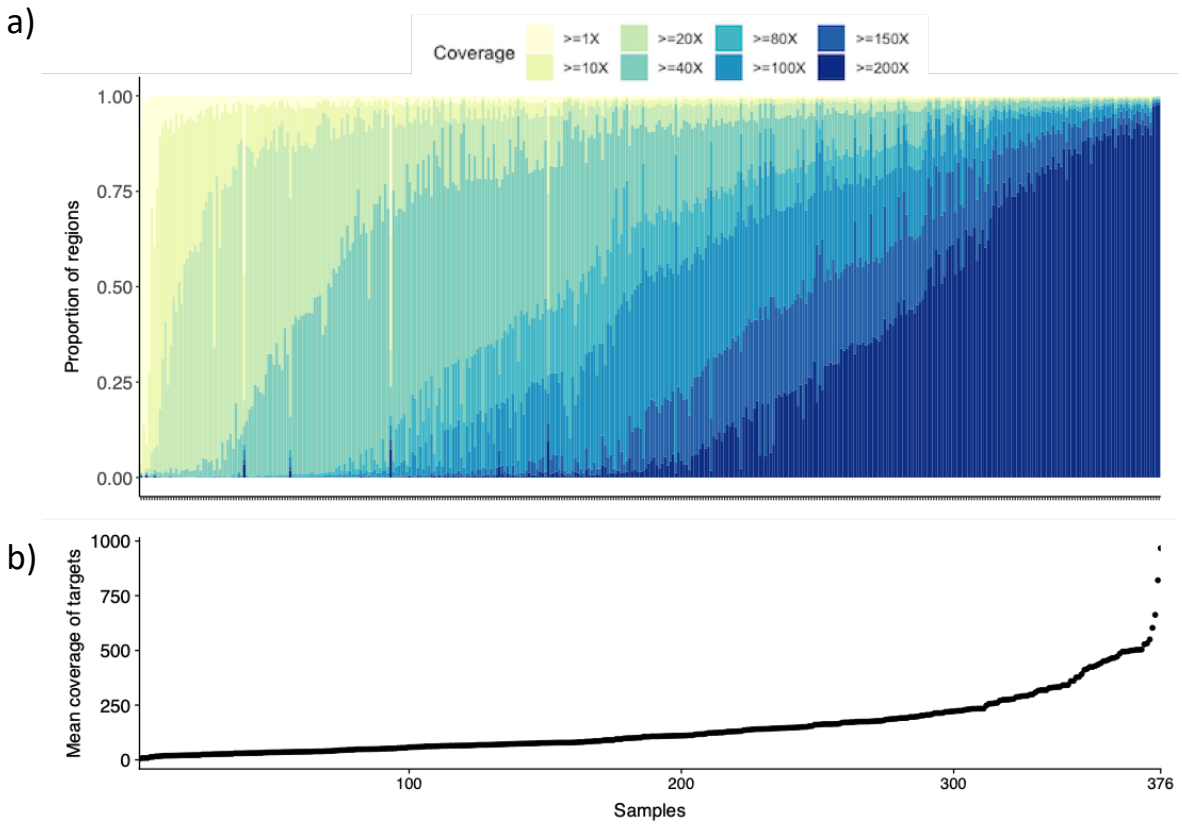


Figure 4. Coverage of targeted regions. **a)** Proportion of regions covered at given depths, per sample. **b)** Mean coverage across all targeted regions, per sample.

I.5 Mutational findings in TGS dataset

The mutational analysis from targeted deep sequencing identified 618 total variants across 199 bulk sorted lymphocyte samples from minor salivary gland biopsies of PSS patients and non-PSS sicca controls. The majority of variants were found in B cell, plasmablast, and plasma cell subsets (**Figure 5b**). These variants were clustered in immunoglobulin gene regions of chromosomes 2, 14, and 22, and had a mutational spectrum indicative of canonical AID (activation induced deaminase) enzyme activity, which is enriched in mutations at the GCT trinucleotide (**Figure 5c**), indicating that somatic hypermutation had taken place in many of the cells. The variant allele frequencies (VAF) of these mutations are mostly small (< 0.1), indicating polyclonal populations of cells (**Figure 5a**), both in T and B cells, as well as in other cell types which include bulk lymphocytes and antigen-presenting cells. Several samples harboured VAFs in the range of 0.15 to 0.30, suggesting a more oligoclonal population where 30% to 60% of cells in the given sample would carry the variant (assuming heterozygosity of variants, the fraction of mutated cells is calculated as VAF*2).

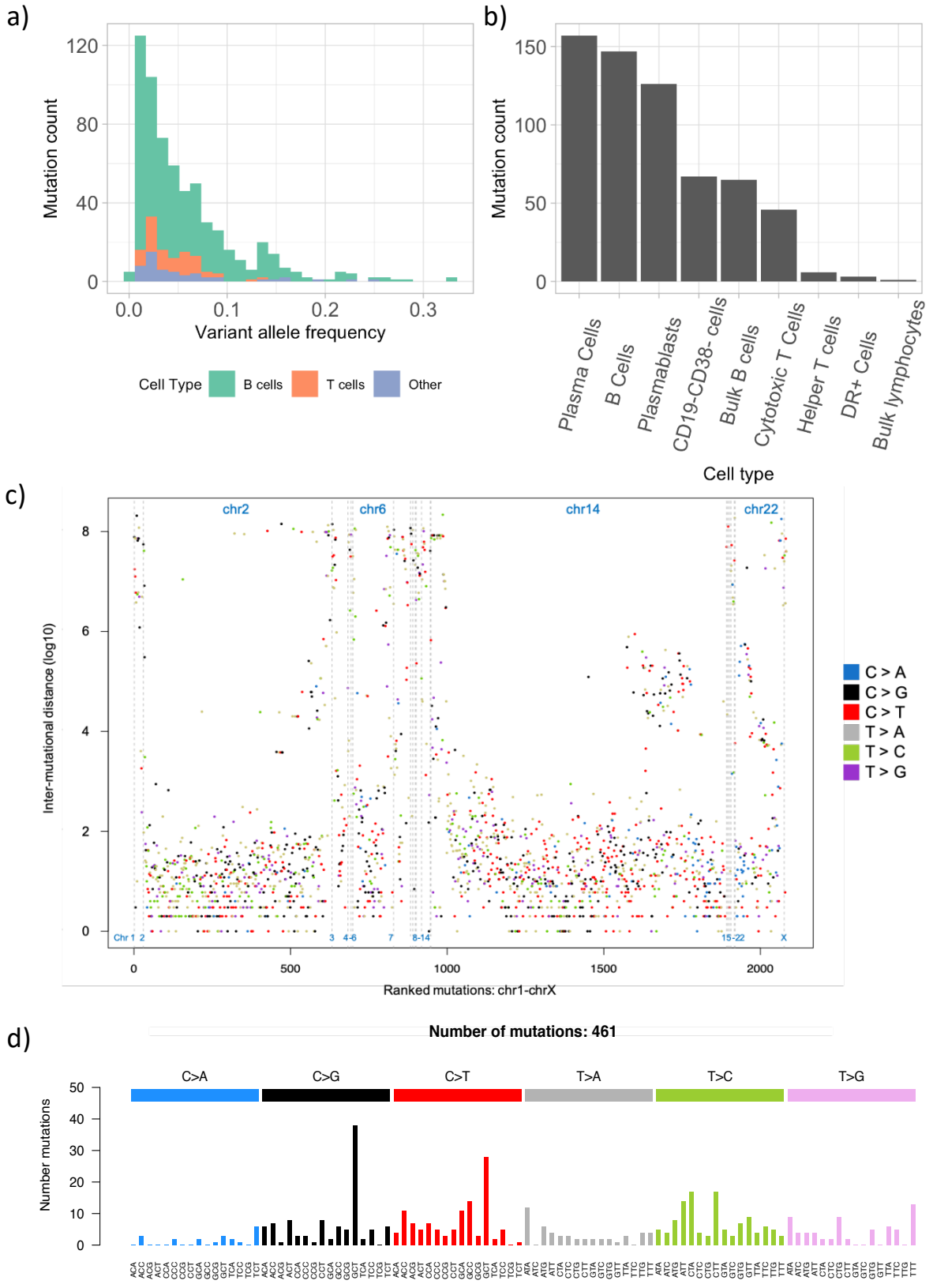


Figure 5. Mutational landscape across target genes. **a)** Histogram of the median variant allele frequency across all cell populations. **b)** Number of variants detected in each cell type subset. **c)** Distance between adjacent variants plotted in chromosomal order for all B cell and plasma cell sample s. **d)** Mutational spectrum of all B cell and plasma cell samples (showing canonical AID signature).

In the context of cancer, "driver" mutations are those that impart a selective advantage to a cell, causing it to proliferate preferentially. As seen in normal tissue studies, driver mutations can exist in cells without causing cancer but still altering features of the tissue[2,4]. To detect putative driver variants which might drive clonal expansion of lymphocytes, I used two approaches: manual annotation of coding variants and application of a dN/dS algorithm to detect gene selection based on relative rates of nonsynonymous and synonymous variation[171] (described in Methods). From these analyses, a prominently mutated gene emerged: *KDM6A*, which encodes the X-linked histone demethylase and tumour suppressor, UTX (Ubiquitously transcribed tetratricopeptide repeat)[172]. Three truncating *KDM6A* mutations were found in CD8 T cell samples, and a fourth was found in a CD4 sample, all in PSS patients. Two of the four *KDM6A* mutations were found in one CD8 sample. Three of the four variants were called by the Shearwater pipeline and passed the filters described above, and a fourth variant was subsequently found by manual inspection of *KDM6A* reads. This variant was filtered out by the Shearwater q-value cut-off due to having only two unique supporting reads, however the mutated reads appeared clean and the mutation was represented on both forward and reverse strands, so it was included in further analysis. dN/dS analysis for gene selection in the targeted gene set across all samples identified *KDM6A* as a gene under significant positive selection (FDR q-value = 0.00068).

The VAF of *KDM6A* mutations ranged from 3% in the rescued sample to 23% in the largest clone (**Figure 6a-c, Table 2**). Sample PD42055c did not carry other detectable mutations in the subset of genes targeted by this analysis, while PD42056c harboured two other variants at comparable VAFs to the *KDM6A* variants, in *TLR9* and *GRID2* genes. This CD8 T cell sample therefore harbours two truncating *KDM6A* variants along with potentially pathogenic variants in two immune-related genes. If co-occurring in the same clone, these mutations may have a proliferative or immune-activating effect. We recently obtained blood from patients with *KDM6A* mutations in order to assess whether the mutations are found in matched blood T cells as well; results are pending.

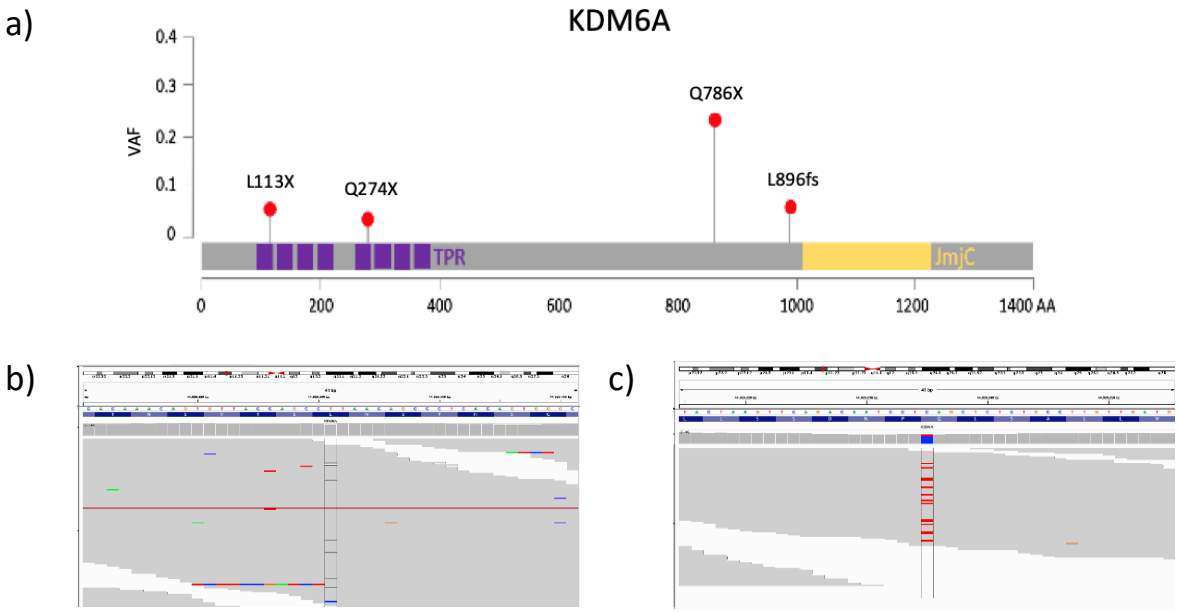


Figure 6. KDM6A mutations: **a)** All truncating (nonsense) KDM6A variants found in T lymphocyte samples, with VAF noted. **b,c)** Snapshot of sequencing reads from Integrated Genomics Viewer software showing KDM6A mutation in patient **b)** PD42056c and **c)** PD42055c, where each mutated read is marked as coloured dash at genomic position highlighted in the centre.

In addition to truncating *KDM6A* mutations identified in T cell samples, a loss-of-function mutation was found in another X-linked tumour suppressor gene, *STAG2.* This mutation was observed in a B cell sample at a VAF of 10% (**Table 2**).

KDM6A is a tumour suppressor gene commonly mutated in many cancers, including haematological malignancies[173,174]. It does not undergo X-inactivation in females, so both copies should remain expressed in normal female cells. Male cells have a Y-chromosome paralogue, UTY, with some overlapping functions with UTX[175]. There is evidence of haploinsufficiency and a dosage-dependent effect of UTX-loss in lymphoma mouse models[174], so it is possible that loss of a single copy, as observed in PSS T cell samples, could have a phenotypic effect. However, if both copies of this gene were to be lost, this would likely have significant effects on proliferation and demethylation of downstream genes, as documented in cancer studies[172,174,176].

I.6 Copy number alterations in TGS dataset

Copy number analysis was carried out to assess possible somatic copy number variations in lymphocyte samples from PSS patients and non-PSS controls. Using the available

Table 2. X-linked truncating mutations and additional mutations co-occurring in the same samples.

sample ID	*cell type*	*chr*	*amino acid*	*vaf*	*depth*	*gene*	*type*	*trinucleotide context*
PD42075a	CD4 T	X	Q274X	0.024	86	KDM6A	stopgain	CCA>CTA
PD42056c	CD8 T	X	L113X	0.058	55	KDM6A	stopgain	TTA>TGA
PD42056c	CD8 T	X	L896fs	0.042	124	KDM6A	frameshift	T>del
PD42056c	CD8 T	3	M58T	0.063	67	TLR9	nonsyn	ATG>ACG
PD42056c	CD8 T	4	V957M	0.042	118	GRID2	nonsyn	CGT>CAT
PD42055c	CD8 T	X	Q786X	0.226	76	KDM6A	stopgain	TCA>TTA
PD42042c	B cell	X	R614X	0.100	44	STAG2	stopgain	GCG>GTG

chromosomal regions covered by the bait set, copy number profiles were inferred by the CNVkit algorithm (Methods), with a control panel of patient fibroblast samples used for normalization. An immediately noticeable trend of X chromosome aneuploidy emerged in female lymphocyte samples. Female T cells, but not B cells or plasma cells, had evidence of subclonal monosomy X, which was especially prevalent in female CD8 T cells (**Figure 7a-c,e-g**). Copy number variation is shown in two ways: as the log-normalized ratio of the sample coverage to the coverage of the control panel (**Figure 7a-c**), and as the total copy number across chromosomes (**Figure 7c-e**).

While the existence of monosomy X was obvious in some samples simply by looking at a log copy ratio plot (**Figure 7a**), other samples had subclonal monosomy X that resulted in subtle visual differences between X chromosome copy number and that of autosomes (**Figure 7b**). To define a threshold for detection of subclonal monosomy X within a sample, I used a p-value cut-off and an effect size cut-off to identify samples in which the copy number of X was significantly lower than the copy number of autosomes. The log copy ratio of autosomes was compared to that of the X chromosome by a one-sided T-test, and effect size of that difference was calculated. The threshold for X loss within a sample was then defined by an FDR-adjusted q-value < 0.01 and an effect size >0.1 (**Figure 7d**); the smallest subclone with monosomy X detected by this threshold comprised 5% of the sample (**Table 5, Figure 8g**).

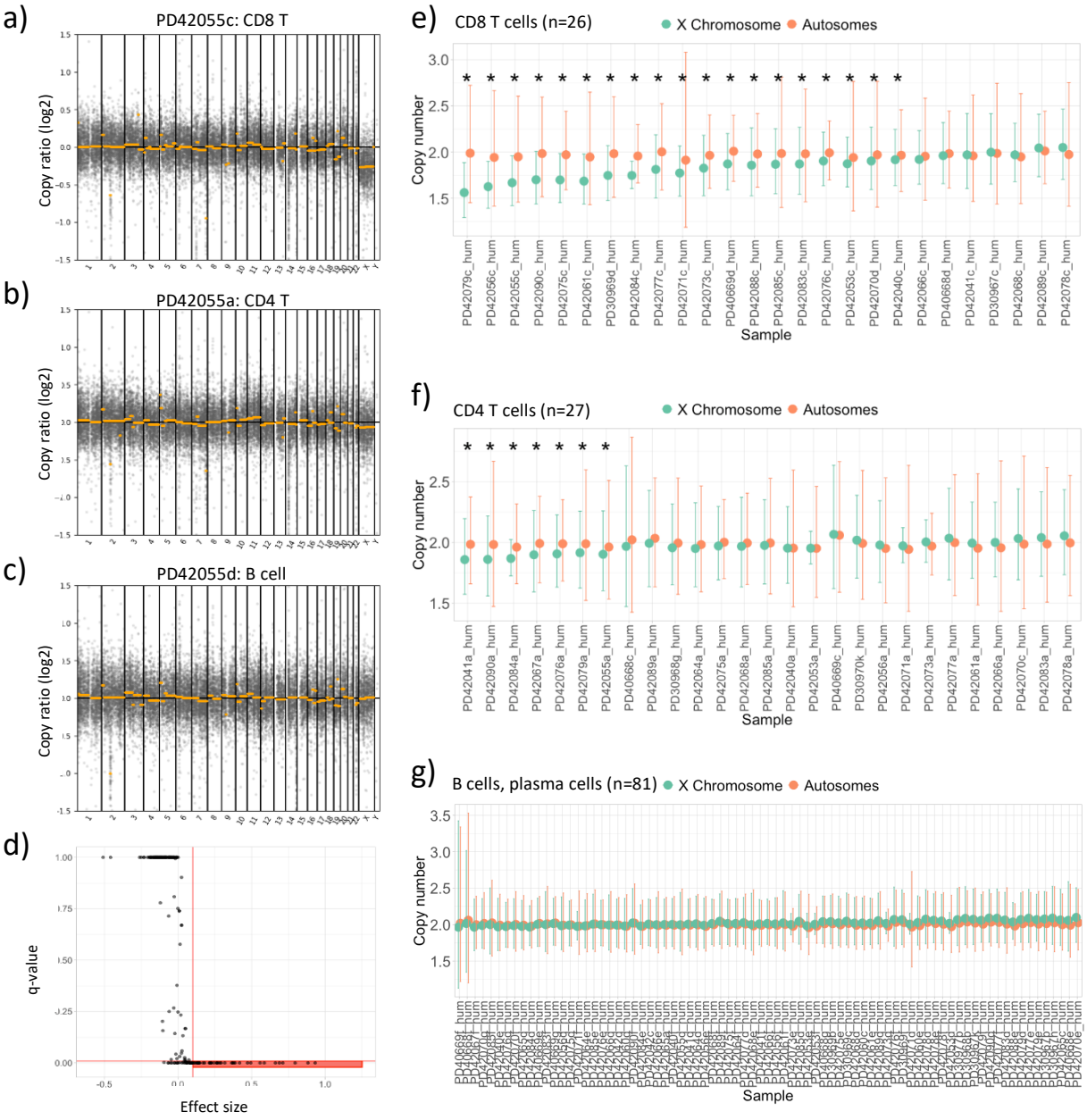


Figure 7. a-c) Log copy ratios across all chromosomes from CD8, CD4, and B cell samples from patient PD42055 biopsy. **d)** FDR q-value (adjustment by number of target regions quantified) < 0.01 and effect size > 0.1 thresholds used to determine significance of X chromosome loss across lymphocyte samples. **e)** CD8 T cell samples: mean X chromosome copy number compared to mean autosomal copy number. **f)** CD4 T samples: mean X chromosome copy number compared to mean autosomal copy number. **g)** B cell, plasmablast and plasma cell samples: mean X chromosome copy number compared to mean autosomal copy number.

Using these metrics, subclonal chromosome X loss was observed in 73% of female CD8 T cell samples (19 out of 26, **Figure 7e**), 26% of female CD4 T cell samples (7 out of 27, **Figure 7f**), and zero out of 81 total B cell, plasmablast and plasma cell female samples (**Figure 7g**). These findings demonstrate a striking prevalence of monosomy X in T cells, especially CD8 T cells, but a complete lack of this phenomenon in B lineage cells.

Table 3. Large copy number alterations in lymphocyte subsets of female patients. "Y" indicates evidence of chr X loss, "N" indicates no detectable chr X loss, as identified by statistical threshold outlined in Fig.7b. Diagnosis of patients is stratified into three categories: "Non-PSS" for patients with sicca symptoms but unconfirmed PSS, "Early PSS" for histologically ambiguous biopsies but other indications suggesting possible early PSS, and "PSS FS" for confirmed diagnosis of PSS with numeric focus score of biopsy sample. The "-" symbol indicates that the cell subset was not recovered in sufficient quantities or failed QC for DNA sequencing. "*" indicates significant q-value but effect size < 0.1; applied only to PBMC-derived samples.

Patient	*Age*	*Diagnosis*	*CD8 T cells*	*CD4 T cells*	*Bulk CD3 T cells*	*CD19 B cells*	*CD38 plasma*	*CD8 T PBMC*
PD42055	56	Early PSS	Y	Y	-	N	N	N
PD42076	-	Early PSS	Y	Y	-	N	N	-
PD42079	-	PSS FS2-3	Y	Y	-	N	N	-
PD42084	64	Non-PSS	Y	Y	-	N	N	Y*
PD42090	52	PSS FS2-3	Y	Y	-	N	N	-
PD42085	48	Non-PSS	Y	N	-	N	N	Y*
PD42073	-	PSS FS2-3	Y	N	-	N	N	Y
PD40669	49	Early PSS	Y	N	-	N	N	N
PD42040	52	Early PSS	Y	N	-	N	N	-
PD42053	49	PSS FS2-3	Y	N	-	N	N	-
PD42056	71	PSS FS4-5	Y	N	-	partial chX loss	partial chX loss	N
PD42061	58	PSS FS4-5	Y	N	-	N	N	-
PD42070	50	PSS FS2-3	Y	N	-	N	N	Y*
PD42071	-	PSS FS2-3	Y	N	-	N	N	-
PD42075	-	PSS FS2-3	Y	N	-	N	N	-
PD42077	-	PSS FS2-3	Y	N	-	N	N	-
PD42083	54	PSS FS2-3	Y	N	-	N	N	N
PD30969	63	Non-PSS	Y	-	-	N	N	-
PD42088	52	PSS FS2-3	Y	-	-	N	N	N
PD42041	52	PSS FS4-5	N	Y	-	N	N	-
PD42067		PSS FS2-3	-	Y	-	-	-	-
PD40668	51	PSS FS4-5	N	N	-	N	N	N
PD42068	57	Early PSS	N	N	-	N	N	N
PD42078	-	PSS FS4-5	N	N	-	N	N	-
PD42089	27	PSS FS4-5	N	N	-	N	N	N
PD42066	53	Early PSS	N	N	-	N	N	N
PD30967	73	PSS FS2-3	N	-	-	N	N	-
PD30968	50	PSS FS2-3	-	N	-	N	N	-
PD30970	64	PSS FS2-3	-	N	-	-	-	-
PD42064	69	PSS FS2-3	-	N	-	N	N	-
PD42042	52	Non-PSS	-	-	Y	N	N	N
PD42060	23	Early PSS	-	-	N	N	N	-
PD42065	39	Non-PSS	-	-	-	N	N	N
PD30974	35	Non-PSS	-	-	-	-	N	-
PD30971	64	PSS FS4-5	-	-	chX gain	N	N	-
PD30973	73	PSS FS0-1	-	-	-	ch7 gain	N	-

Table 3 details the samples in which X chromosome ploidy was evaluated using the statistical thresholds defined above (**Figure 7d**); all samples shown are female. Not all patient biopsies yielded CD4, CD8, and B cell subsets, therefore only the available cell types are shown from each patient. Lymphocyte samples evaluated for monosomy X fell into all three diagnostic groups: confirmed PSS patients, patients with suspected early PSS based on ambiguous biopsy, and non-PSS sicca patients. Out of the 7 biopsies from female patients with non-PSS sicca, only 4 had infiltrating lymphocytes that were successfully isolated. Of those four, all have detectable X chromosome loss in CD8, CD4, or bulk CD3 T cell subsets, and two also have subtle X chromosome loss noted in paired PBMC-derived CD8 samples. The presence of monosomy X in the control group implies that it is not a feature specific to PSS, however it is worth noting that non-PSS sicca patients have salivary gland dysfunction of a different aetiology or possible early PSS that did not meet diagnostic criteria at the time of biopsy. Therefore, the non-PSS cohort is not a true control group for this study, though it is the closest we were able to obtain. Some non-PSS biopsies had a focus score of 1, indicating the presence of a histologically observable focal lymphocyte aggregate. Clinical details of the female PSS-negative patients are shown in Table 3b.

Monosomy X was observed in a majority of CD8 T samples and a quarter of CD4 T samples and was the sole genetic abnormality detected in most of them. The previously described *KDM6A* mutations also occurred in CD8 T cell samples with monosomy X, thereby these samples had two somatic events targeting the X chromosome. In the CD4 T cell sample which harboured the other *KDM6A* variant, the X chromosome copy number fell just short of the statistical threshold set for detection of subclonal monosomy X. However, this sample was

Table 4. Clinical details of PSS-negative patients. Samples in red have detectable X chromosome loss in one or more T cell subsets. Patient IDs highlighted in red indicate those that had a subset of T samples with subclonal loss of X chromosome.

Patient	*Diagnosis*	*Age*	*Focus Score*	*Clinical detail*
PD42084	Non-PSS	64	1	Non-PSS sicca; acellular cyst
PD42085	Non-PSS	48	1	Non-PSS sicca; reflux disease
PD30969	Non-PSS	63	0	Non-PSS sicca; few scattered inflammatory cells
PD42042	Non-PSS	52	0	Non-PSS sicca; no details available
PD42065	Non-PSS	39	0	Non-PSS sicca; chronic fatigue
PD30974	Non-PSS	35	0	Non-PSS sicca; no details available

very polyclonal, with VAF of *KDM6A* mutation being 3%, so a similarly-sized monosomy X subclone may have existed but was too small to detect. The matched CD8 sample from that patient's biopsy, however, had distinct monosomy X. The estimated fraction of cells with X loss in CD8 samples is shown in Table 3c, as calculated by the difference of mean autosomal copy number and mean X chromosome copy number, multiplied by two. The proportion of cells in samples PD42055c and PD42056c that appear to lack the X chromosome is around 30%, while the observed *KDM6A* variant allele frequencies in these samples are 23%, 6%, and 4%. It is plausible that X chromosome loss and *KDM6A* mutations co-occur in the same subclone, however this is difficult to prove from bulk sequencing data.

If monosomy X and a truncating *KDM6A* mutation do exist in the same clone, then both copies of the tumour suppressor UTX would be lost, which would have significant downstream epigenetic effects on many genes and would provide the clone with a selective advantage, as has been observed in cancer studies[174].

Table 5. Estimated size (fraction) of subclone with monosomy X in CD8 lymphocyte subsets, fraction of dominant clone identified by V(D)J rearrangement with CDR3 sequences and gene rearrangement.

Sample	*X-loss clone*	*Mutation*	*Largest TRB clone*	*TRB clone CDR3*	*V/D/J usage*
PD42079c	0.43	-	0.33	CASTRGEGTGELFF	TRBV2/TRBD1/TRBJ2
PD42056c	0.32	*KDM6A (6%, 4%)*	0.15	CASSTGQLTNTEAFF	TRBV9/TRBD1/TRBJ1
PD42090c	0.28	-	0.13	CASSAEAGTSTDTQYF	TRBV5-5/TRBD1/TRBJ2
PD42055c	0.28	*KDM6A (23%)*	0.10	CASSLARAAQETQYF	TRBV7-6/TRBD1/TRBJ2
PD42075c	0.27	-	0.31	CASSDKQGNYGYTF	TRBV5-1/TRBD1/TRBJ1
PD42061c	0.26	-	0.33	CAISDVGGGNQPQHF	TRBV10-3/TRBD2/TRBJ1
PD30969d	0.23	-	0.18	CASSLPYGPYGYTF	TRBV28/TRBD2/TRBJ1
PD42084c	0.21	-	0.21	CASSQSPGGTQYF	TRBV14/TRBD1/TRBJ2
PD42077c	0.19	-	0.16	CASSYSLDRGDTEAFF	TRBV6-3/TRBD1/TRBJ1
PD42073c	0.14	-	NA	CASSSEGGNTEAFF	TRBV5-1/TRBD1/TRBJ1
PD42071c	0.14	-	0.18	CASSLAWGADEQFF	TRBV12-3/TRBD2/TRBJ2
PD40669d	0.14	-	0.27	CASNPTGTSYEQYF	TRBV27/TRBD1/TRBJ2
PD42088c	0.12	-	NA	CSVGSQGTNEKLFF	TRBV29-1/TRBD1/TRBJ1
PD42085c	0.12	-	0.15	CASSLEGQGPTGSPLHF	TRBV4-1/TRBD1/TRBJ1
PD42083c	0.11	-	0.14	CASSLEGGAKNGYTF	TRBV7-9/TRBD2/TRBJ1
PD42076c	0.09	-	NA	CASSPSGDARDNEQFF	TRBV11-1//TRBJ2
PD42053c	0.07	-	NA	CASSVSGTRSGHQPQHF	TRBV2/TRBD1/TRBJ1
PD42070d	0.07	-	0.18	CASSLGRAVNEKLFF	TRBV5-1/TRBD1/TRBJ1
PD42040c	0.05	-	0.13	CASSDTDIENTEAFF	TRBV6-1/TRBJ1

Previous studies have noted that loss of sex chromosomes occurs more frequently in lymphocytes than other cell types, with up to a few percent of peripheral blood lymphocytes losing the X or Y chromosome with age[177,178]. However, the rate of monosomy X in cytotoxic T cells seen here is much higher than that observed in healthy females in studies of peripheral blood. The phenotypic effects of losing the X chromosome are not completely clear, but some studies suggest that they might be associated with various morbidities such as cancer and autoimmune disease[179,180].

While copy loss of chromosome X was very common in T cell samples, one patient had a copy gain of chromosome X was in their bulk CD3 T cell sample (patient PD30971). The only whole-chromosome aneuploidy event in autosomes was a gain of chromosome 7 in the bulk CD19 B cells of patient PD30973 (data not shown). Chromosome 7 gain is frequently observed in B cell lymphomas, suggesting its putative role in malignant transformation or progression[181].

I.7 Receptor sequences of tissue-derived lymphocytes

The targeted gene set queried across the set of lymphocyte samples included the numerous V, D, and J genes that recombine to form antigen receptors of B and T cells. These included immunoglobulin IGH, IGK, and IGL, as well T cell receptor TRA and TRB genes. Due to the pulldown gene enrichment method, short-read platform, and overall polyclonality of the samples, reconstructing the receptor sequences from their discontinuous DNA segments proved challenging. I used the Mixcr algorithm (Methods) to extract heavy and light chain V(D)J rearrangements from T and B cell samples and considered only those rearrangements which were productive and yielded fully reconstructed CDR3 sequences. Some samples had many V(D)J clones reconstructed while others had few, presumably due to underlying differences in clonality and sequencing coverage. Overall, the number of total and unique clones detected per sample increased with the sequencing coverage (**Figure 8a**). The utility of reconstructing V(D)J sequences was twofold: assessing the clonal composition of lymphocyte samples and attempting to identify any disease-associated trends in repertoire usage.

To address the question of clonality, we considered only samples in which at least 10 independent V(D)J clones were reconstructed. The findings confirm what was observed by

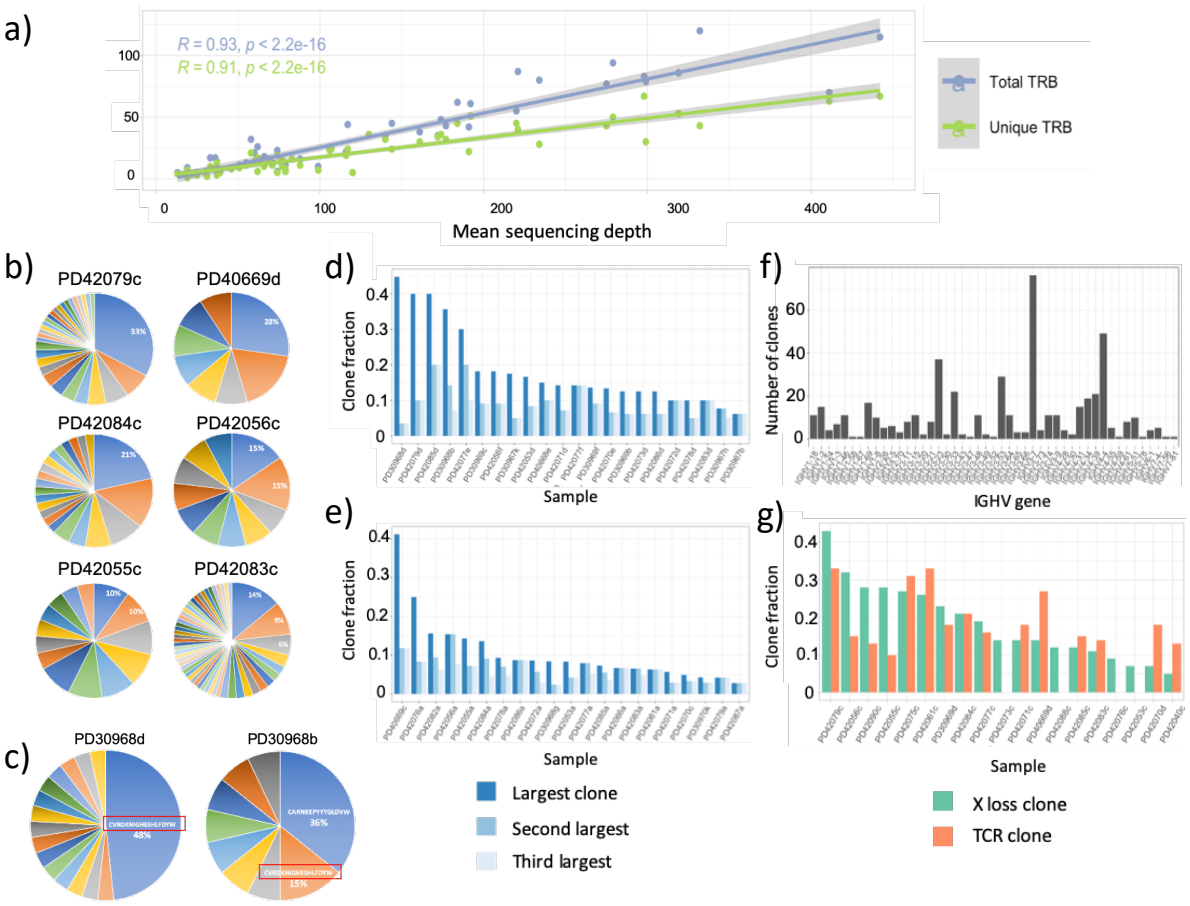


Figure 8. T and B cell receptor sequences. **a)** Number of total and unique T cell beta chain rearrangements detected versus sequencing coverage. **b)** T-cell receptor heavy chain diversity in a subset of CD8 samples (including PD42055c and PD42056c which harbour *KDM6A* mutations) with percentage of largest TCR clone labelled. **c)** B-cell and plasma cell BCR from patient PD30968 showing shared clone. **d)** B cell and plasma cell V(D)J: size of top three largest IGH clones per sample, from samples containing 10 or more identifiable CDR3 sequences. **e)** CD4 T cell clonality per patient, from samples containing 10 or more identifiable CDR3 TRB clones. **f)** Usage of IGHV genes across all detected BCR clones in B cells and plasma cells from biopsies. **g)** Comparison of estimated size of clones harbouring monosomy X and size of largest TCR clone across female CD8 T cell samples with observed X loss; TCR clone unavailable for some samples.

the variant allele frequency trend, that salivary gland T and B cells are not monoclonal lymphocyte populations, but contain some expanded clones. In the CD8 T cell samples, the largest clone comprised 33% of a sample, as suggested by the TCR beta rearrangements identified, 33% of which shared an identical CDR3 sequence (**Figure 8b,d**). The size of the CDR3 clones somewhat correlated with the estimated size of clones harbouring monosomy X (**Table 5, Figure 8g**), suggesting that the largest expanded CDR3 clone may be the one lacking an X chromosome. Of the 15 biopsies which had matched CD8 T cell samples from peripheral blood, the major clone identified in biopsy-derived cells was also found in blood in only one case: CASSQDTSIGSPLHF was present in tissue-derived CD8 at 27% frequency and in PBMC-derived CD8 in 1% of reconstructed CDR3 sequences. The significantly higher proportion of this clone in the tissue than in blood indicates local expansion of a circulating T cell and/or

selective recruitment of that clonotype to the tissue. The lack of similar findings in other blood-derived CD8 cells is likely a technical hindrance, where the polyclonality of blood and limited sampling of the lymphocyte pool make it difficult to reconstruct repertoires from rare clones.

The largest heavy chain CDR3 clone observed in B cells comprised 45% of the repertoire of the sample (**Figure 8c**). Interestingly, the clone identified in that $CD19^+$ B cell sample was also found in the $CD19^-CD38^+$ plasma cell sample from this patient (**Figure 8c**), demonstrating that this clone likely matured from a B cell phenotype to a plasma cell phenotype within the salivary gland tissue. Sharing of clones between B cell and plasma cell compartments of a patient's biopsy was observed in several other samples as well. No immunoglobulin heavy chain (IGH) CDR3 sequences were shared between patients, although several light chain CDR3s were found to be present in multiple patients.

The most frequently used IGHV gene detected across all CDR3 clones observed in biopsy-derived B cells was IGHV3-7 (**Figure 8f**). This finding prompted comparison to the previously described public "Po" idiotype of the rheumatoid factor antibody, which is comprised of a IGHV3-7/IGHJ3 heavy chain paired with an IGKV3-15 light chain[97]. This combination of heavy and light chains was found in 7 B cell samples from different patients, however since multiple heavy and light chain CDR3s were identified in each sample, it is not possible to know if the IGHV3-7/IGHJ3 was expressed by the same clone as IGKV3-15 in those samples. Other public clonotypes, such as the "Wa" idiotype of rheumatoid factor[97], were not detected.

Lastly, the TCR sequences reconstructed from CD4 samples were comprised of multiple small subclones. A few samples were exceptions and harboured dominant clones that comprised up to 40% of the heavy chain repertoires. The clonality of CD4 T cell samples by beta chain CDR3 usage is shown in **Figure 8e**.

Overall, even though polyclonality of samples and variable depth of sequencing made repertoire extraction challenging, some important observations can be made from the findings. The lymphocyte populations were not monoclonal by the coarse grouping into CD4, CD8, B cell, plasmablast, and plasma cell populations, but there was distinctly less clonal

diversity observed in tissue samples than in their matched PBMC-derived counterparts. This observation supports the notion of localized clonal expansion within the tissue. The size of dominant CDR3 clones can be compared to the size of clones harbouring mutations or copy number changes to infer whether these events might co-occur in the same cells. In samples where the proportion of cells with monosomy X is approximately the same as the proportion of the largest TCR clone, we may conclude that X is lost in this particular receptor-bearing clone. In cases where the proportion of cells with monosomy X is larger than the largest TCR clone, it is possible that the chromosomal loss occurred either in a precursor T cell prior to V(D)J diversification or independently in two or more TCR clones. The latter scenario would suggest that loss of X is a common tissue-driven phenomenon that preferentially affects the cytotoxic T cell population and might be more common in disease and inflammation.

To further explore the set of sorted lymphocyte samples, I analysed a subset of them by whole genome sequencing. The aim of this was to detect mutations outside of regions targeted by our gene panel, perform genome-wide copy number analysis, and gain insights into genome-wide mutation patterns.

II. Whole genome sequencing (WGS) of bulk sorted lymphocytes

In order to further characterize a subset of bulk lymphocyte samples which contained intriguing features by targeted sequencing analysis, we sequenced whole genomes from the remaining DNA material. The samples selected for whole genome sequencing included those that carry a *KDM6A* mutation, have evidence of X chromosome loss, or harbour an expanded clone as determined by TCR/BCR repertoire analysis. The number of samples per cell type category that have been whole-genome sequenced is outlined in **Table 6**; additional samples are currently pending sequencing and analysis. This section will briefly discuss the WGS findings specific to bulk lymphocyte samples, while integrated analysis of these and LCM-derived lymphocyte samples will be addressed later in this chapter.

Table 7. Number of samples with WGS data per cell type category.

CD8 T cell	*CD4 T cell*	*B cell*	*Plasmablast*	*Plasma cell*	*Fibroblast*
7	2	2	0	1	6

II.1 WGS findings

Whole genomes of lymphocyte samples were sequenced to an average depth of 31X (**Figure 9a**). Single nucleotide substitutions were called by the Caveman algorithm and small insertion/deletion events by the Pindel algorithm (Methods), using matched fibroblast control samples. The median VAF of single nucleotide substitutions genome-wide was in the range of 0.1-0.2, (**Figure 9b**) in line with the degree of oligoclonality seen in targeted sequencing data. The number of single nucleotide substitutions per sample ranged from 100 to 500, with the lower end of the spectrum likely due to dropout of low VAF variants which were not adequately captured at the given depth of sequencing. WGS confirmed the nonsense *KDM6A* variant identified in sample PD42055c, however the two *KDM6A* variants in PD42056c which occurred at lower VAF were not detected, likely due to low sensitivity. Other potentially pathogenic coding variants were identified in multiple samples; however, none were obviously deleterious or recurrent.

Genome-wide copy number profiles were detected by the Ascat and Battenberg algorithms (Methods) and confirmed the chromosomal findings from the TGS dataset. Copy loss of chromosome X was confirmed in six female CD8 T cell samples, while the seventh sample had normal diploidy of chromosome X in both datasets. The total copy number, logR (log ratio of sample copy number to control copy number), and BAF (B-allele frequency, or minor allele frequency) profiles shown here for sample PD42079c (**Figure 10a-c**) were similar to those of the other samples with X loss. Ascat analysis estimated that up to 60% of cells in this sample lack an X chromosome, suggesting a clonal expansion harbouring monosomy X. However, on closer inspection of the BAF distribution of polymorphic sites on X, we observe that there is

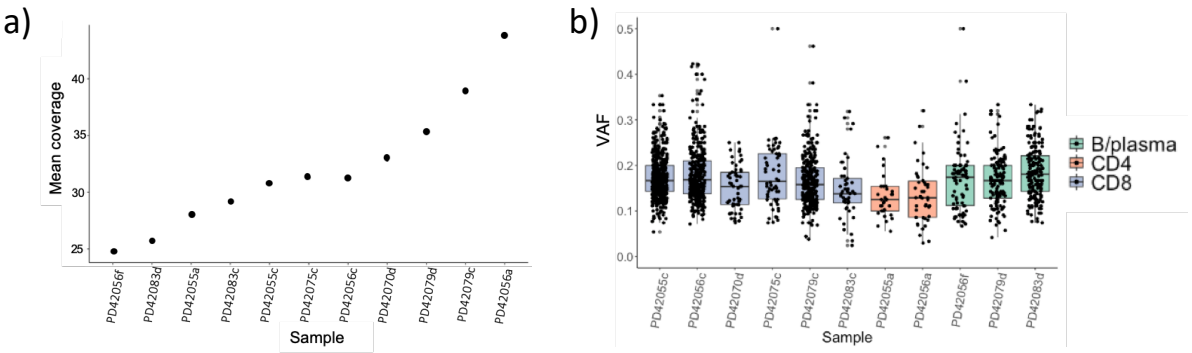


Figure 9. a) Mean coverage of whole genome samples. **b)** Median VAF across whole genomes of lymphocyte samples.

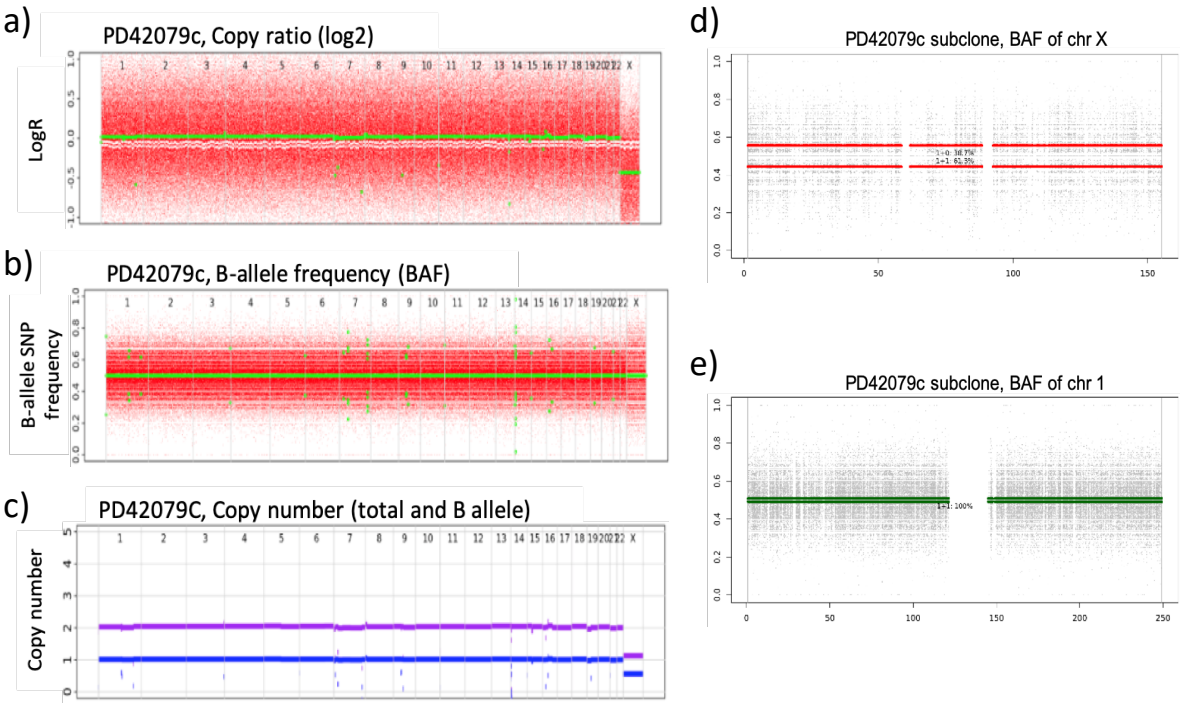


Figure 10. WGS copy number analysis of CD8 T cell sample PD42079c. **a)** Ascat algorithm logR (log ratio of sample versus control copy number across chromosomes), **b)** Ascat BAF (B allele, or minor allele, frequency at heterozygous SNP sites), and **c)** Ascat copy number of B allele (blue) and total copy number (purple), **d-e)** Battenberg BAF of subclone harbouring copy number alteration in **d)** chromosome X and **e)** chromosome 1.

not a clear divergence away from 0.5 as might be expected if around half the reads from either the maternal or the paternal X were missing, in which case we would expect bands around 0.33 and 0.66. Instead, the frequency of the minor allele is centred around 0.5, implying that both the maternal and the paternal X are almost equally represented, even though the overall copy number of X is lower. There is more of a spread of BAF values on X than autosomes indicating a slight allelic disbalance, but there are no concentrated bands diverging from 0.5. This intriguing finding suggests that there might be multiple clones within the sample that have lost the X chromosome, some losing the maternal and others the paternal copy, resulting in a nearly even distribution of heterzyogous alleles.

The Battenberg algorithm is better equipped to infer subclonal somatic copy number changes than Ascat, and it can depict the copy number profile of a subclone rather than the whole sample. For sample PD42079c, Battenberg detected a subclone with a divergence of BAF away from 0.5 on the X chromosome (**Figure 10d-e**), confirming allelic imbalance and a clonal loss of X. This sample, PD42079c, had the most pronounced copy loss of chromosome X among all

samples queried by both TGS and WGS. For most of the other samples, Battenberg failed to extract copy number profiles, likely due to clonality and coverage issues. In samples where Battenberg did generate subclonal copy number profiles, they did show the same allelic imbalance demonstrated by a split in the BAF. Therefore, it remains likely that in some samples, multiple clones might independently be losing the X chromosome.

The implication of multiple monosomy X clones within a cell population is that X chromosome loss is a frequent mitotic event in this context. Whether that context is specific to PSS, related to general inflammation, or a result of tissue-driven expansion requires further investigation. The prevalence of X chromosome loss in T cells, especially CD8 T cells, and the absence of it in B cells suggests a cell-type specific phenomenon.

III. Laser-capture microdissection approach for sequencing lymphocytes from PSS biopsies

The lymphocyte populations sorted by FACS yielded valuable observations about mutational trends in different subsets of T, B, and plasma cells. However, the limitation encountered through this approach was the low degree of clonality across most samples. This presented a challenge for mutation detection of low frequency variants, matching heavy and light chain receptor rearrangements, pairing receptor sequences with mutational events, etc. To improve the probability of capturing a clonal population for sequencing, we turned to the histology of minor salivary glands. In PSS, infiltrating lymphocytes form focal aggregates, which are frequently used as a diagnostic criterion. These aggregates are often spatially isolated from each other in the gland and can be viewed as discrete units. We therefore hypothesised that individual lymphocytic aggregates have a higher likelihood of being clonal populations than do bulk lymphocyte subsets sorted from the entire biopsy, and we addressed this question using laser-capture microdissection (LCM).

Due to the small size of the minor salivary gland biopsies, the entirety of tissue that was obtained in the first cohort of samples was used up for FACS isolation of lymphocytes. We acquired a new set of snap-frozen biopsy samples from the University of Newcastle Biobank

Table 9. Cohort of patient biopsies were used for laser-capture microdissection.

Sample	*Diagnosis*	*Focus score*	*Sex*	*Age*	*Clinical detail*
PD42760	Non-PSS	0	F	51	
PD42764	Non-PSS	0	M	57	
PD42766	Non-PSS	0	M	57	
PD42763	Non-PSS	0	F	59	
PD42765	Non-PSS	0	F	59	
PD42759	Non-PSS	0	F	65	
PD42761	Non-PSS	0	M	67	
PD42767	PSS	3	-	22	
PD42768	PSS	3	F	31	
PD42769	PSS	3	F	42	
PD45528	PSS	4	F	58	
PD45532	PSS	2	-	60	
PD45527	PSS	5	F	68	
PD42773	PSS	3	M	71	Peripheral neuropathy, IgM kappa positive, anti-MAG antibody low-positive
PD42770	PSS	3	M	72	
PD42771	PSS	4	F	-	
PD42772	PSS	5	F	-	
PD42774	PSS	5	F	-	
PD45529	PSS	0	F	61	Small fibre neuropathy, mild small vessel disease
PD45530	PSS	0	F	46	Meibomian gland dysfunction, punctate epithelial erosions in left eye, musculoskeletal pain, bilateral keratoconjunctivitis, sicca, vestibulitis.

with the help of rheumatologist Fai Ng and our collaborators Paul Milne and Matt Collin. The new cohort consisted of 13 PSS-positive and 7 PSS-negative biopsies. The biopsies were ethanol-fixed and paraffin-embedded (Methods) prior to sectioning to a thickness of 10 µm and mounting onto histology slides, which was performed by Yvette Hooks. They were subsequently stained either by haematoxylin and eosin (H&E) or by immunohistochemistry (IHC) for surface markers of interest. LCM was used to dissect individual lymphocyte aggregates from the tissue, as well as glandular epithelium, which will be discussed in a separate chapter. Sequencing libraries were made from microdissected tissue by the bespoke library preparation protocol and submitted for whole genome sequencing (Methods).

III.1 Histology and immunohistochemistry of minor salivary glands in PSS

By H&E staining, we observed the classic histological features of minor salivary glands in PSS and how they compared to biopsies of non-PSS sicca controls. The biopsies from PSS patients ranged from those that were mildly inflamed, with one or few visible lymphocytic aggregates, to those completely inundated with lymphocytes and almost completely lacking in glandular acini (**Figure 11b-c**). While non-PSS biopsies had characteristic glandular architecture characterised by compact mucous acini, serous acini, and ducts enveloped into distinct lobules by connective tissue (**Figure 11a)**, PSS biopsies frequently displayed destruction of these structures (**Figure 11b-c**). In addition to focal lymphocytic infiltration, PSS salivary glands often showed degradation or absence of acini and dilated, fibrotic ducts with thickened walls. These features were present in varying degrees across all but two of the biopsies from patients with confirmed PSS, which appeared histologically normal and were assigned a focus score of zero. These two patients were diagnosed based on sufficient clinical and serological evidence of disease, despite normal biopsies; this scenario occurs in a small subset of PSS patients.

Many of the PSS biopsies had large lymphocytic infiltrates or large areas of the tissue covered by lymphocytes. To better assess the lymphocyte subtype composition and localization, I stained the slides for relevant B, T and plasma cell markers by immunohistochemistry (IHC). The IHC staining was used to guide microdissection, so that clusters of the same cell type from a regional aggregate were dissected as a unique sample. Staining for CD4 and CD8 T cell markers, CD20 B cell marker, and CD38 plasma cell marker helped identify the regional deposition of these cells. Non-PSS samples were stained as well, revealing interspersed plasma cells throughout the tissue (**Figure 11d**), along with occasional T cells. In PSS biopsies, plasma cells were much more prevalent throughout, while discrete clusters were often composed of CD20 B cells loosely surrounded by T cells (**Figure 11e-g**). The tight clustering of B cells made them good targets for microdissection (**Figure 11h**), while T cells proved more difficult to dissect due to their diffuse distribution. Libraries were made from the dissected samples, and those that had a minimum concentration of 5 ng/µl were selected for whole genome sequencing. Targeted sequencing of the previously queried set of genes was not

performed due to the higher starting amount of DNA necessary for hybridization-based selection of targets (10 ng/µl), which most LCM libraries did not meet.

III.2 WGS of LCM-derived lymphocytes

This section will highlight specific findings from whole genome sequencing of LCM-derived lymphocytes samples, and an integrated overview of these and previously discussed bulk sorted lymphocyte samples will follow in the next section of this chapter.

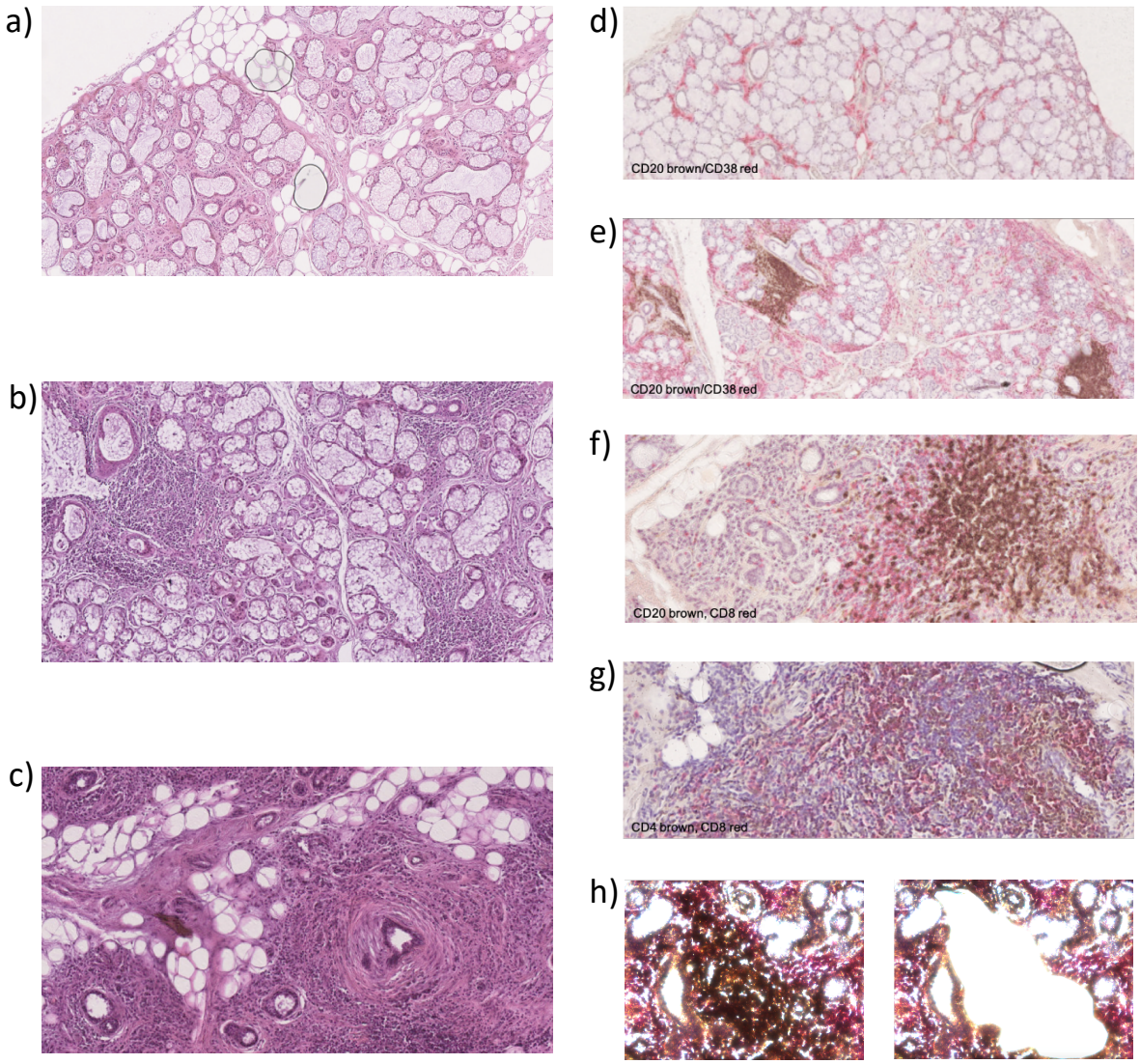


Figure 11. H&E staining of minor salivary gland biopsies. **a)** Non-PSS biopsy, focus score = 0. **b)** PSS biopsy, focus score = 3. **c)** PSS biopsy, focus score ≥ 5. **d)** Non-PSS biopsy IHC, CD20 brown (DAB), CD38 red. **e)** PSS biopsy IHC, CD20 brown, CD38 red. **f)** PSS biopsy IHC, CD20 brown, CD8 red. **g)** PSS biopsy IHC, CD4 brown, CD8 red. **h)** LCM microscope image before and after dissection of DAB (brown) stained lymphocyte aggregate from PSS biopsy.

Table 11. Number of LCM WGS samples with > 10X depth

CD8 T cell	*CD4 T cell*	*B cell*	*Plasma cell*
7	2	8	3

The laser-capture microdissection method yielded variable numbers of B and T cell samples from each biopsy, depending on the localization and ability to dissect enough cells for library preparation. Due to the lower cellularity of samples, the genomic sequencing coverage was lower than that of bulk sorted samples, with average depth ~15X (**Figure 12**a), as compared to ~30X for sorted samples. Samples with a mean coverage below 10X were excluded from analysis due to the limited ability to detect somatic variants (**Table 8**).

The clonality of samples, represented by median VAF of single nucleotide variants is shown in **Figure 12b.** None of the samples were highly clonal (with a median VAF > 0.4), although many were in the range of 0.2 to 0.35, suggesting oligoclonal or near-clonal populations.

Two female PSS biopsies (PD42771 and PD45527) had CD8 T cell clusters sequenced, for a total of seven CD8 T cell whole genomes. All of these seven showed lower copy number of X than of the autosomes, revealing subclonal monosomy X, as observed in the CD8 samples of the bulk sorted lymphocyte dataset (**Figure 12c**). Even though the coverage was lower and the copy number profiles were noisier than that of the bulk sorted CD8 samples, the X loss was discernible. No copy loss of X was observed in corresponding B or plasma cell samples from the same biopsies (**Figure 12c**). No *KDM6A* mutations were found in the CD8 T cell samples, however the sensitivity of detecting subclonal variants was low given that the coverage was only 10-15X (**Figure 12a**), so they may have been missed.

From the pool of LCM-derived B cell samples, the most pronounced finding came from the biopsy of a 71-year old male patient with confirmed PSS and a focus score of 3 (PD42773). CD20 B cell clusters were sampled from three isolated regions of the biopsy and harboured a shared set of variant calls with high VAFs (**Table 9, Figure 12b,d**). These variants included numerous protein-coding mutations, including a known pathogenic variant in the *MYD88*

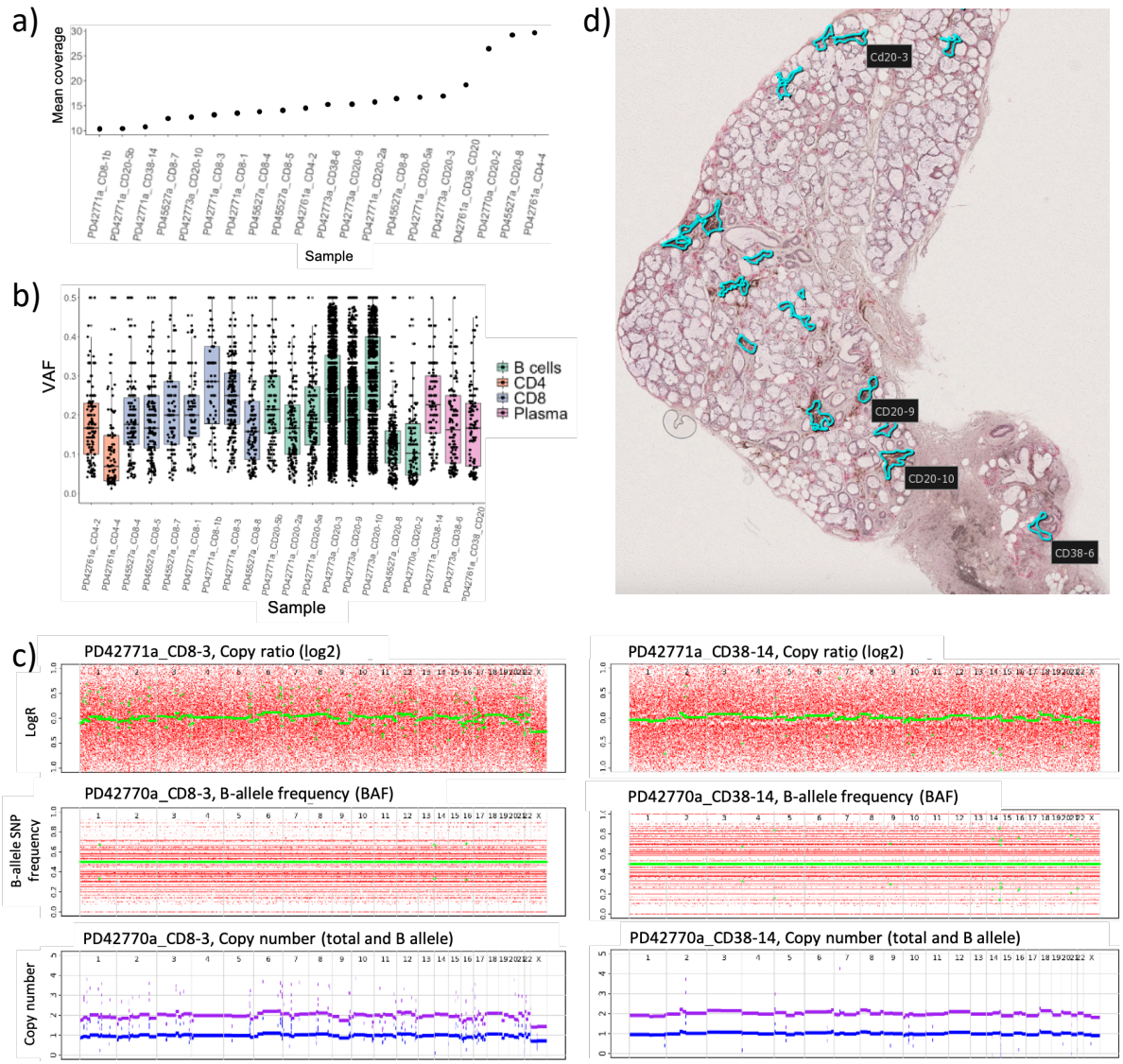


Figure 12. a) Mean genome-wide depth of coverage across LCM lymphocyte samples. **b)** Median VAF of WGS from LCM-derived lymphocyte samples. **c)** Copy number profiles of two female CD8 T cell samples from patients PD42771 and PD45527, plots from top down show: log copy ratio, B-allele frequency, total and B-allele copy number. **d)** Dissections of three CD20 B cell clusters (labelled 3, 6 ,and 9) and one plasma cell cluster (CD38-6) from patient PD42773.

gene, L273P (a.k.a L265P, depending on transcript used for annotation), which is mutated in lymphoproliferative disorders. Additionally, nonsense mutations in tumour suppressor genes (*ERBIN* and *PGC)* and a nonsense mutation in the immuno-inhibitory *LILRB2* gene were also found. The mutations were not shared with a nearby aggregate of plasma cells (sample CD38-6). A common single B cell receptor rearrangement was found in all three CD20 samples as well (with a heavy chain CDR3 sequence CAKAGIGVGSGLNRDLQHW and Ig kappa light chain sequence CHQYNSFPLTF*).* The genome-wide number of mutations in these B cells was higher than that of other samples, with the excess burden of mutations attributable to off-target effects of activation-induced deaminase (polymerase-eta), as defined by signature SBS9 in the

Table 13. Coding mutations found in discrete B cell clusters (3, 9, and 10) and a plasma cell cluster of biopsy from patient PD42773. "Nonsyn" denotes nonsynonymous amino acid change.

Gene	*Amino acid*	*Trinucleotide context*	*Mutation type*	*CD20-3 VAF*	*CD20-9 VAF*	*CD20-10 VAF*	*CD38-6 VAF*
CCT6B	N23S	TTA>TCA	nonsyn	0.32	0.30	0.53	0
LILRB5	R95*	GCG>GTG	stopgain	0.40	0.38	0.44	0
IGLL5	L39P	CTG>CCG	nonsyn	0.22	0.25	0.56	0
MYD88	**L265P**	CTG>CCG	nonsyn	0.36	0.33	0.23	0
ERBIN	Q320*	GCA>GTA	stopgain	0.25	0.41	0.30	0
PGC	E327*	CCT>CAT	stopgain	0.32	0.27	0.41	0
NYAP1	V524D	GTC>GAC	nonsyn	0.46	0.16	0.17	0
KIAA1432	L1063P	TTT>TAT	nonsyn	0.32	0.32	0.44	0
NCS1	E24G	CTC>CCC	nonsyn	0.14	0.23	0.40	0
CXorf38	N95K	ATG>AGG	nonsyn	0.63	0.80	1.00	0
GRIA3	R660G	CTC>CCC	nonsyn	0.40	0.50	0.57	0
CGN	R211W	ACG>ATG	nonsyn	0.22	0.14	0.06	0
IGFN1	A2567S	ACG>AAG	nonsyn	0.12	0.17	0.20	0

COSMIC database. Further discussion of mutational signature analysis across all samples will follow in a subsequent section.

The findings from the B cell sample of patient PD42773 illustrate the ability to extract expanded cell populations by tissue microdissection. This biopsy shows a CD20 B cell clone that has populated the gland and whose expansion was very likely driven by underlying proliferation-associated mutations. This was the only patient with a detectable monoclonal population of B cells harbouring known cancer driver mutations. It could be the case that this patient is on the trajectory to developing lymphoma and the B cell lesions could be considered pre-lymphoma. There are additional features that set this patient apart from the rest of the cohort and suggest more advanced disease, namely the clinical presence of peripheral neuropathy and a low-positive finding of anti-MAG (myelin-associated glycoprotein) antibody and IgM kappa paraprotein in the blood. The L265P mutation has been previously identified in patients with IgM anti-MAG paraprotein-associated peripheral neuropathy[182]. The L265P is also found in >90% of patients with Waldenstrom's macroglobulinemia[183], a low-grade B cell malignancy similarly associated with IgM kappa paraprotein, as well as in IgM monoclonal gammopathy of uncertain significance (MGUS)[184]. *MYD88* L265P has also been found in marginal zone B-cell lymphomas, including MALT lymphoma[185]. Given the serological finding of IgM paraprotein and its association with the MYD88 mutation, it is possible that the B cell

clone harbouring this mutation would be detectable in the blood of this individual; sequencing of blood-derived B cells is pending.

It is unclear whether the B cell findings from patient PD42773 are an isolated case, or if it is representative of more advanced PSS, as suggested by the clinical findings of peripheral neuropathy and a monoclonal component in the blood. Sequencing of additional LCM-derived samples is currently underway, which will determine whether B cell clusters from other patients are also clonally expanded. For this purpose, the LCM approach is more sensitive than bulk sequencing for evaluating B cell aggregates. However, for CD4 and CD8 T cell samples, the same success of isolating clonal populations was not replicated, since their distribution is more diffuse and therefore more difficult to microdissect. Extracting clonal populations for sequencing analysis from minor salivary gland biopsies remains challenging, but by using two approaches I have been able to demonstrate the respective strengths of each.

IV. Integrated analysis of minor salivary gland lymphocytes

With the combined set of FACS-sorted and LCM-derived lymphocyte samples, I performed further analyses to infer mutational patterns across the cell subtypes. These included mutational signature analysis to evaluate the mutagenic processes active in lymphocytes, and analysis of positive selection of genes by implementation of a dN/dS algorithm across samples.

IV.1 Mutational signature analysis

Single nucleotide variant (SNV) calls were used to extract mutational signatures and compare them to signatures of known exposure in the COSMIC database. This was done using a hierarchical Dirichlet process to break down the contribution of signatures in a sample using the distribution of mutations across a spectrum of 96 mutational trinucleotide contexts (Methods). Due to a low number of observed variants, resulting from low coverage and/or polyclonality, the signature extraction could not produce meaningful results for some

samples. Therefore, results from samples with less than 100 observed SNVs were not taken into consideration.

In the group of B cell samples with sufficient SNVs called, we observed the presence of COSMIC Signature 9 (SBS9), associated with the off-target effects of B cell somatic hypermutation (SHM) machinery (**Figure 13a**). The three expanded B cell samples from patient PD42773 had the highest number of SNVs (up to 2,000), and the burden of additional mutations in these samples is attributed to SBS9, suggesting highly active SHM processes have occurred in this B cell clone at some stage of ontogeny. However, the pathogenic coding mutations observed in these samples are not the T>G transversions characteristic of SBS9, suggesting that SHM activity was not the cause of these mutations. Rather, the increased hypermutation activity may be subsequent to the mutations and a result of the activated phenotype of the cell. Endogenous signatures SBS1 and SBS5, attributed to normal "clock-

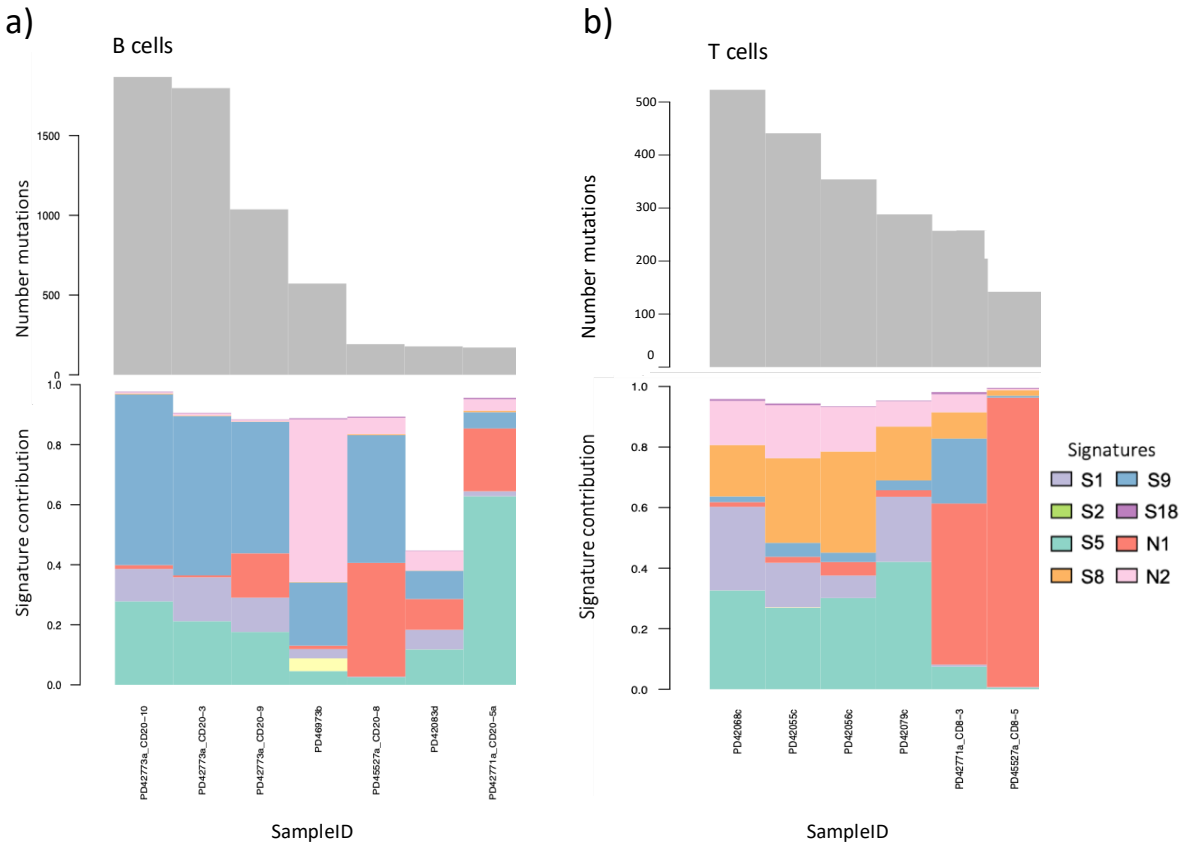


Figure 13. WGS mutational signatures in **a)** B cell samples; top bar chart shows number of variants, bottom chart shows proportion of mutations explained by given signature (white space denotes unattributed mutations); red line separates samples to the right which have too few mutations for extracting meaningful results. **b)** Mutational signatures in T cell samples. "S" signatures are those found in COSMIC database, "N" denotes novel signatures extracted by HDP algorithm

like" processes of cellular ageing and replication, were also found ubiquitously across the B cell samples, as expected.

In T cell samples, we likewise observed signatures SBS1 and SBS5 across all samples with sufficient variant calls. Additionally, signature SBS8 was observed, and to a smaller extent, SBS18 (**Figure 13b**). The aetiology of SBS8 is not known, though it has been observed across a wide array of cancer types, mostly solid tumours[11]. SBS18 is thought to be a result of damage caused by reactive oxygen species and has been observed in varying degrees across normal tissues and cancer. Two of the three *KDM6A* substitutions are T>C transitions, possibly attributable to signature SBS5.

IV.2 Analysis of genes under selection

To assess if any genes in the combined lymphocyte dataset were under positive selection, we applied the dN/dS algorithm (Methods) to infer whether the ratio of nonsynonymous to synonymous mutations in any given gene was higher than expected based on its background mutation rate. This analysis was done using variant calls from whole genome and targeted sequencing of tissue-derived lymphocytes. The first method utilized only variants from WGS, and selection analysis was done in an unbiased manner across all coding regions of the genome. The second method used combined variant calls from TGS and WGS and focused the analysis only on genes targeted by the pull-down panel. In the latter approach, *KDM6A* was under significant positive selection (qval = $9.3 \cdot 10^{-4}$, **Table 10**), which has been previously discussed in the TGS results section, but was not significant in the WGS-only analysis. This is due to the fact that the low VAF variants identified by TGS were not confirmed in the WGS due to lower sequencing depth and sensitivity of detection.

Additionally, mutations in the *IGLL5* gene were found to be significantly recurrent (qval = $3.3 \cdot 10^{-6}$ in targeted analysis, not significant in genome-wide analysis, **Table 10**). *IGLL5* variants

Table 15. Genes under positive selection by dN/dS analysis of genes in targeted gene panel.

Gene	# Synonymous	# Missense	# Nonsense	# Splicing	# Indels	FDR q-value
KDM6A	0	0	3	0	1	$9.3 \cdot 10^{-4}$
IGLL5	1	5	0	0	0	$3.3 \cdot 10^{-6}$

were found in B cell and plasma cell samples, at variable VAFs, up to 0.56 in the clonally expanded PD42773 B cell sample. This gene is in the immunoglobulin lambda locus on chromosome 22 and forms part of the joining and constant regions of B cell receptor sequences, but not variable regions. Little else is known about the function of *IGLL5*, though it is found recurrently mutated in chronic lymphocytic lymphoma (CLL) and diffuse large B cell lymphoma (DLBCL)[186], and in a CRISPR-screen study has been identified as a tumour suppressor in DLBCL[187]. Due to its location in the immunoglobulin locus, it is targeted by canonical AID hypermutation, but also by non-canonical AID effects[186]. We observed a total of six distinct coding variants in *IGLL5* and several noncoding variants, several of them shared between multiple B lineage samples from the same individual.

V. Discussion

This study is the first to perform DNA sequencing for somatic mutation detection on lymphocytes from salivary gland biopsies from primary Sjögren's syndrome. As far as we are aware, it is also the first to do so with tissue-infiltrating lymphocytes in any autoimmune disease. As discussed in the Introduction section, the interest in somatic mutations in normal tissue and chronic diseases has gained traction only in the past few years, driven by the development of technologies that have enabled these studies. The challenges faced centre around sensitivity of detection due to the low clonality of normal, non-cancerous tissues. In this study, we paired an ultra-low input library preparation protocol with techniques for cell type enrichment to investigate the mutational landscapes of tissue infiltrating lymphocyte subpopulations.

We observed recurrent mutations in *KDM6A* in T cell samples, a total of four protein-truncating variants. Apart from one *KDM6A* mutation which had a variant allele frequency (VAF) of 23%, the remaining variants were at much lower VAF, down to 3%. The sequencing coverage at the sites of *KDM6A* mutations was less than 100X in three out of four samples. Therefore, several of the variants found were just above the technical limit of detection. This begs the question: had it been possible to sequence these samples at higher depth, would we have found additional *KDM6A* mutations? To further explore this, more sensitive methods would be required. In particular, single cell methods would be best suited to overcome the

limitation posed by polyclonality of samples. We have made attempts to grow single cell lymphocyte colonies, however the success rate of this protocol is low and yielded no viable cultures from minor salivary gland lymphocytes. Sequencing of genomes from individual cells does not currently lend itself well to mutation discovery and signature analysis in polyclonal samples, as the rate of artefacts introduced by whole genome amplification is high.

Nevertheless, the recurrence of loss-of-function UTX (*KDM6A)* variants in PSS is a novel and intriguing finding. Paired with loss of X chromosome in potentially the same clones as the mutations, this would result in cells with no functional UTX protein. *KDM6A* is a tumour suppressor commonly mutated in haematological malignancies and certain solid tumours, with inactivating mutations often found in lymphomas and T cell acute lymphoblastic leukaemia (T-ALL)[172]. *KDM6A* is one of several genes on the X chromosome that escapes X-inactivation, thereby both copies are expressed in normal female cells. The mechanism of tumour suppression by *KDM6A* is incompletely understood but thought to possibly be independent of its histone demethylase function[188]. While it has been shown that inactivating mutations in UTX lead to increased H3K27 histone methylation of many genes and impart a proliferative advantage, it is unclear how the loss of UTX activates oncogenes[172]. Overexpression of wildtype *KDM6A,* on the other hand, has been shown to cause cell cycle arrest in human fibroblast cell lines[189]. The effect of UTX loss is dosage dependent, as demonstrated by mouse models of B cell lymphoma. Loss of both copies of UTX in female Eμ-Myc transgenic mice (which are predisposed to lymphoma) greatly accelerates lymphomagenesis, while loss of one copy does so to a lesser but still significant extent[174]. Male Eμ-Myc mice with UTX knockout and an intact copy of the Y-chromosome paralog, UTY, developed lymphoma significantly faster than Eμ-Myc $UTX^{+/-}$ female mice, suggesting that UTY does not provide a comparable tumour suppressive function[174]. These and other relevant findings have been used to hypothesize that X-linked escape-from-inactivation tumour suppressor genes might be responsible for the higher rate of cancer in men[190]. If indeed loss of both copies of UTX imparts proliferative disinhibition, it would lead to selective expansion of the T cells that harbour it. To understand if this also has an immune activating effect, further studies such as expression and methylation analysis would be needed. UTX acts epigenetically on many downstream genes, including immune-related genes[191,192], so assessing these downstream effects would be crucial. We were not able to perform additional

studies on the samples in which *KDM6A* mutations were identified since all biopsy material was consumed for the preparation of sequencing libraries.

Much more frequently than *KDM6A* mutations, we observed loss of the X chromosome, in around three quarters of female CD8 T samples and a third of CD4 T samples (but no B lineage samples). It has long been recognized through cytogenetics studies that lymphocytes frequently lose sex chromosomes with age[193,194]. Why this happens preferentially in lymphocytes is not known. Estimates of the rate of Y and X chromosome loss range from one to several percent of blood lymphocytes in ageing individuals[193,194]. Loss of sex chromosomes was for a long time not thought of as a deleterious event, since it preferentially affects the inactive X chromosome in women and Y chromosome in men, both of which were considered dispensable to an ageing cell. More recently however, loss of the Y chromosome in blood cells has been associated with an increased cancer risk and overall higher mortality in men[195]. As such, this phenomenon is similar to clonal haematopoiesis of indeterminate potential (CHIP) in ageing individuals, which carries a similar risk of cancer and morbidity. The implications of age-associated monosomy X are unclear.

It is known that monosomy X preferentially affects the inactive X (Barr body), and in a study of peripheral blood and buccal cells, rates of monosomy X were up to 0.45% in 75-year old apparently healthy women[178]. Due to the female predominance of many autoimmune diseases, some studies have investigated the presence of X chromosome aneuploidy and skewed patterns of X-inactivation in this context, and both events were found to be associated with certain autoimmune diseases including primary biliary cirrhosis, autoimmune thyroid disease, and scleroderma[179]. In one study, monosomy X was found in about 5.5% of peripheral lymphocytes of women affected by these diseases, compared to 1.5% of lymphocytes from healthy age-matched women[179]. Additionally, women with Turner's syndrome (45,XO) are also predisposed to some of these particular autoimmune diseases[196,197]. Primary Sjögren's syndrome has not specifically been associated with Turner's syndrome however. It has instead been observed that men with Klinefelter syndrome (47,XXY) are over-represented in cohorts of male patients with PSS and lupus[198,199]. These observations point to an important but complex and poorly understood role of the X

chromosome in autoimmune diseases, which seemingly revolves around careful dosage control of multiple genes on X.

Additionally, monosomy X has been observed in acute lymphoblastic leukaemia, sometimes as the only cytogenetic abnormality[200]. It has also been found in some cancers, including breast carcinoma[180]. Association with cancer suggests a role of the inactive X in controlling cellular proliferation, as has been previously alluded to by the discovery of tumour suppressor genes which escape X inactivation.

The frequency of monosomy X in our dataset is significantly higher than what has been reported in blood lymphocytes of healthy women. Additionally, the previously mentioned studies did not observe a difference in rates of monosomy X between T and B lymphocytes, while our investigation shows a markedly higher rate in T cells, especially cytotoxic T cells.
These findings suggest that perhaps monosomy X imparts a selective advantage that is specific to T cells in the context of disease. Tissue resident CD8 cells are driven to proliferate by innate immune activation and MHC-I expressing cells marked for cytotoxic deletion. In inflamed PSS salivary glands, CD8s may be aberrantly targeting healthy glandular epithelial cells expressing self-antigen, killing them and releasing more self-antigens into the environment, thereby stimulating a further immune response. In this context, CD8s might be continually primed for activation and selected for based on their survival ability. It is therefore tempting to conjecture that monosomy X, especially when paired with the loss of an X-linked tumour suppressor, may be providing the survival advantage that is selected in an environment of chronic immune stimulation.

However, it remains difficult to draw confident conclusions about the pathogenicity of this phenomenon and its specificity to PSS due to the lack of normal tissue-derived lymphocytes as a true control for PSS biopsies. Several biopsies from patients who were not ultimately diagnosed with PSS showed evidence of monosomy X in cytotoxic T cells from tissue, along with subtle evidence of monosomy X in blood CD8s as well. It is difficult to contextualize this finding when the patients in question were not truly healthy controls, as they had symptoms of salivary gland dysfunction similar to that in PSS, and they had similar numbers of immune cells extracted from their salivary glands as some of the patients with confirmed PSS.

Therefore, to investigate whether monosomy X in tissue-resident T cells is truly a feature of PSS or if it is a phenomenon related to normal tissue-driven expansion, general inflammation, or age-related aneuploidy requires further investigation. This would entail extraction of tissue-resident lymphocytes from healthy uninflamed tissue and examining their karyotypes by cytogenetics or their copy number variations by sequencing. Additionally, a closer examination of the rates of monosomy X in blood T cells versus B cells would be beneficial, as this has not been thoroughly examined in previous studies.

Another intriguing finding related to monosomy X in this study is that the aneuploidy does not in all cases seem clonal in origin, given the B-allele frequency trends and the comparison to size of largest TCR clone (Table 3c, Fig. 8f). In some CD8 samples, the estimated size of a clone with monosomy X is very close to the size of the largest TCR clone identified in that sample, so it is plausible to suspect that loss of X occurred in that particular TCR clone, which expanded in the tissue. In other samples, the clone sizes are not concordant, and the fraction of cells with monosomy X appears two or three times greater than the largest TCR clone. While the sensitivity of V(D)J detection in these polyclonal samples can legitimately be called into question, if we assume for the sake of argument that the TCR clone sizes detected are accurate, then there are two scenarios that can explain the data observed. The first possibility is that T cells with diverse V(D)J repertoires are produced early in life, released into the circulation, and eventually come to inhabit the salivary glands, where they are driven to proliferate by antigen, and in the process lose their X chromosome and subsequently lose the remaining copy of UTX by mutation as well. Thereby, the progression would be: repertoire diversification, X chromosome loss, acquisition of UTX mutation. Conversely, the second possibility is that the X loss and UTX mutation occur in a T cell precursor, prior to V(D)J diversification, which would result in several TCR clones harbouring both monosomy X and the UTX mutation. Given the clonal frequencies in sample PD42055c (XO clone = 28%, UTX clone = 23%, TCR clone = 10%; Table 3c), it seems the latter scenario is more likely for this sample. If this is true, then a significant proportion of T cells in blood of this individual would also harbour X loss and the UTX mutation. We have not yet been able to test this patient's blood to determine if this is the case.

The B cell findings of this study highlighted the recurrently mutated *IGLL5* gene in PSS. While significant by dN/dS analysis, it is difficult to interpret the importance of this finding since the function of this gene is poorly understood. It is not a known cancer driver, but it has been repeatedly mutated in chronic lymphocytic leukaemia[186] and diffuse large B cell lymphoma[187], therefore it could be affecting proliferation of B cells in the context of PSS as well. This study also yielded evidence of varying degrees of clonal expansion of B lineage cells by repertoire analysis, identified non-canonical AID activity as a major contributor to genome-wide mutagenesis in these cells, and identified *IGHV3-7* as the most commonly used immunoglobulin heavy chain V gene in this cohort.

However, the B cell findings overall revealed an unexpected paucity of lymphoma-driver mutations. Given the multiple previous studies that suggest active B cell proliferation in the PSS salivary glands and a propensity for development of B cell lymphoma[88,94,168], we suspected that B cells would be accumulating detectable driver mutations during the course of the disease. While this is still a plausible hypothesis, the findings of this study have not strongly confirmed it, despite using different methods of cell isolation and sequencing to investigate the question. I observed bona fide cancer driver mutations in the clonally expanded B cells of patient PD42773, although this individual may be an isolated case in our cohort. Therefore, the original hypothesis that somatic mutations in "rogue" autoimmune B cells may be driving the pathogenesis of PSS from an early stage has not been proven by this study. Instead, the results suggest that perhaps lymphoma driver events occur at later stages of the disease, after many years of B cell stimulation and oligoclonal expansion in the inflammatory environment. Patient PD42773 seems to have had more advanced disease than the rest of the cohort, as indicated by peripheral neuropathy and a monoclonal component in the blood. If we were to sequence lymphocytes from patients with later stages of disease and paraprotenemia or other monoclonal features, perhaps we would see more mutation-driven clonal expansion in tissue-resident B cells. Our overall cohort represented the earliest clinically validated stages of PSS, at the time of diagnosis, at which point we see little evidence of driver mutations in B cells. It is all the more striking perhaps, that we instead see recurrent somatic mutations and copy number changes in T cells (particularly cytotoxic T cells) at this stage, which may underlie the early pathogenesis of PSS. This study therefore identifies cytotoxic T cells as potential key players in PSS, a disease canonically considered to be a "B

cell disease", highlighting the need for further investigation of their molecular function and effect of the somatic alterations.

Chapter 4: Single-cell expression profiling of lymphocytes infiltrating minor salivary glands

To date, most immunophenotyping and transcriptomic studies in primary Sjögren's syndrome (PSS) have focused on whole blood to detect systemic changes that characterize the autoimmune response[102,107]. This approach has successfully identified cell type trends and gene signatures significantly altered in the disease, such as lymphopenia and a persistent signature of interferon activation[78,109]. Conversely, the lymphocytic infiltrate in affected minor salivary glands has not been well characterized. Classical analysis by immunohistochemistry staining for lymphocyte markers and gene expression studies has identified an abundance of T cells in minor salivary glands of PSS patients[169], along with increased levels of plasma cells and B cells, which sometimes aggregate into germinal centre-like structures[94,201]. Recently, a more comprehensive investigation was carried out by mass cytometry with 34 specific antibodies to characterize various cell types present in the minor salivary glands and blood of PSS patients[112]. The key tissue-related findings from this study were high levels of activated CD8 T cells expressing MHC class II molecules, terminally differentiated plasma cells, and activated epithelial cells in PSS biopsies compared to non-PSS controls, which correlated with disease severity. These important observations point to previously under-appreciated cell types which likely play a role in disease, highlighting the need for further studies to understand their functions and characteristics.

Transcriptomic studies can identify genes and signalling pathways that are differentially enriched in disease states and point to potential mechanisms of disease activity. Single cell transcriptome profiling takes this concept a step further, allowing for high resolution analysis and identification of rare and novel cell types. Single cell RNA sequencing has yielded many exciting discoveries in recent years, however, to our knowledge, it has not yet been used to query tissue infiltrating lymphocytes in PSS minor salivary glands. To characterize the cell types and transcriptomic features of these infiltrating immune cells, I performed a single cell RNA sequencing analysis of biopsy-derived immune cells from donors with PSS and those with non-PSS sicca symptoms as controls. The findings are described in this chapter.

STUDY AIMS

1. *Use single cell RNA (scRNA) sequencing to describe the cell types and gene expression profiles of lymphocytic infiltrates in minor salivary gland biopsies from PSS and non-PSS donors.*

   As mentioned, previous tissue phenotyping studies have described certain key characteristics of lymphocytic infiltrates in PSS minor salivary glands. scRNA sequencing has the potential to build on and expand those findings by analysing a greater number of genes and cells from a given donor in a high throughput manner.

2. *Investigate clonality and expansion trends by surveying the T-cell and B-cell receptor sequences.*

   Certain techniques employed for scRNA library preparation allow for the downstream bioinformatic reconstruction of V(D)J rearrangements in T and B cells. The receptor sequences can inform us of clonal expansion with high resolution, and in conjunction with expression analysis, inform us of the cell type and functionality of clonally expanded cells.

3. *Look for potential correlates between findings of scRNA sequencing and findings of DNA sequencing of a different cohort of PSS and non-PSS biopsy-derived lymphocytes (described in Chapter 3).*

   Our genomic investigation of tissue infiltrating lymphocytes yielded observations about clonality, receptor usage, and somatic alterations (particularly in cytotoxic T cells). Single cell transcriptome profiling, even from a different set of PSS and non-PSS donors, could potentially offer complementary insights.

STUDY DESIGN

To conduct a single cell transcriptomic study, fresh minor salivary gland tissue was obtained from patients undergoing diagnostic biopsies based on suspicion of Sjögren's syndrome. Biopsies were obtained at the Newcastle University clinic under a research protocol approved by the UK Research Ethics Committee. The samples obtained were not from the same individuals as those used for the study described in Chapter 3, which investigated somatic mutations by DNA sequencing. While matching DNA and RNA sequencing from the same donors might have been optimal for research purposes, biopsy size of minor salivary glands was too small to provide enough cellular material for both studies from the same tissue. All patients enrolled in the study experienced oral dryness and other symptoms indicative of Sjögren's syndrome, and histopathological examination of their minor salivary gland biopsies determined whether there was focal lymphocytic infiltration which meets the diagnostic criteria for Sjögren's syndrome. Based on the histological findings in conjunction with other criteria such as serum autoantibody findings, a diagnosis was made of either primary Sjögren's syndrome or non-Sjögren's sicca (non-PSS in further text). For the purposes of the study, both PSS and non-PSS samples were examined, using the latter as a control for the disease. However, it should be kept in mind that the "control" group had salivary gland dysfunction of a different nature, even if the tissue appeared histologically normal.

Biopsies were processed and digested with a collagenase enzyme protocol by Dr Paul Milne at Newcastle University to obtain single cell suspensions. Downstream processing of disaggregated cells involved two protocols for two separate methods of single cell RNA sequencing. One batch of samples was processed for SmartSeq2 full length cDNA sequencing and another batch for 10X 5' cDNA single cell sequencing with V(D)J enrichment. I therefore have two datasets of single cell transcriptomes, one from a full-length cDNA plate-based method and the other from a 5' cDNA enriched droplet-based method (**Figure 14**).

The first protocol involved fluorescence-activated cell sorting (FACS) of the disassociated cell suspensions into single wells of a plate, using lymphocyte surface markers for T and B cells. In this way, single cells were separated into wells and categorized by their lymphocyte compartments, which included the following groups (initially gated on $CD45^{+}$ and $CD3^{+/-}$):

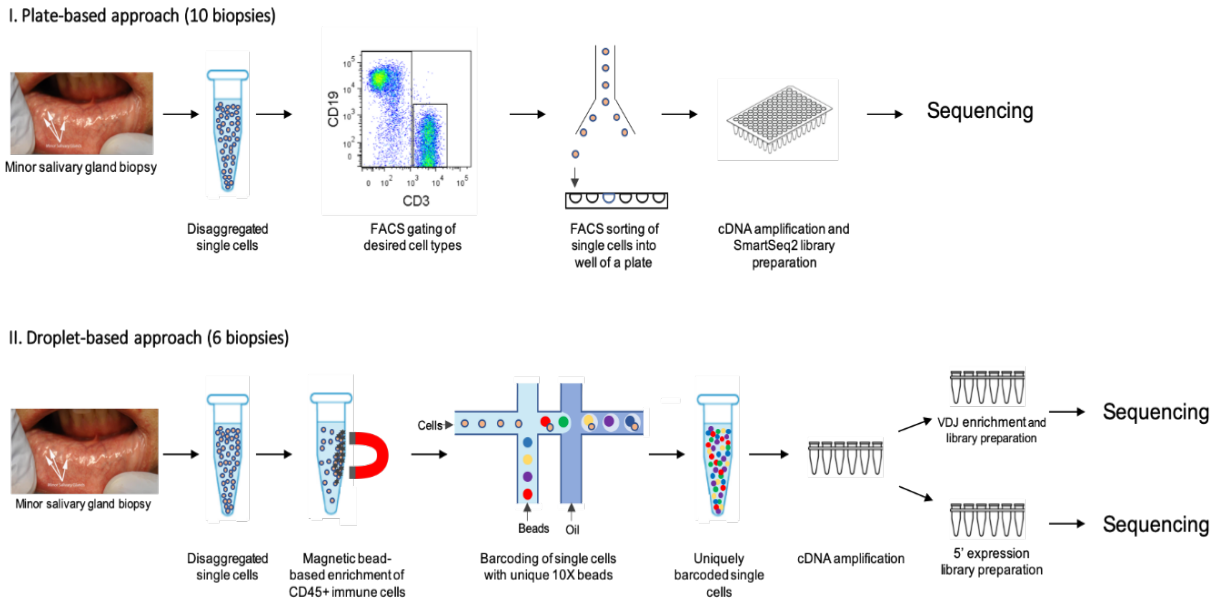


Figure 14. Two approaches to single cell RNA sequencing: plate-based full-length cDNA approach (SmartSeq2) and droplet-based 5' expression with V(D)J enrichment approach (10X Genomics).

$CD4^+$ T cells, $CD8^+$ T cells, $CD3^+$ T cells, $CD19^+CD38^-$ B cells, $CD19^+CD38^+$ plasmablasts, $CD19^-CD38^+$ plasma cells, and $CD19^-CD38^-$ "other" cells. The single cells were made into individual RNA sequencing libraries by the SmartSeq2 protocol. This process yielded full length cDNA sequences which allowed for both expression analysis and reconstruction of T cell and B cell receptor sequences through their expression of V(D)J genes. A total of six confirmed PSS biopsies and four non-PSS biopsies were processed in this way.

The second protocol relied on enrichment of immune cells by selection of those expressing the $CD45^+$ cell surface marker by a bead-based method, also done at Newcastle University. The enriched immune cells were processed according to the 10X Genomics bead capture protocol, which involves tagging individual cells with unique cellular barcodes prior to cDNA retrotranscription. The barcoded cDNA samples from each donor were then split into two, one for constructing a 5' enriched transcriptome library and the other for making a PCR-enriched library of T and B cell V(D)J regions. The expression and V(D)J libraries were then sequenced separately and later combined for bioinformatic analysis based on the unique cellular barcodes. A total of three confirmed PSS biopsies and three non-PSS biopsies were sequenced in this way.

I. Results of plate-based full-length cDNA single cell sequencing (SmartSeq2 protocol)

Upon sequencing both sets of samples, I analysed the results separately due to the different underlying library preparation methods and the significant batch effects they would impart if the datasets were combined. I opted to compare the findings observed across both datasets and use the particular strengths of each approach to extract any additional information. The findings of the SmartSeq2 plate-based method are described first, followed by a validation of the key findings in the 10X droplet-based method.

A total of 50 plates and 4,714 single cells were FACS-sorted from the minor salivary gland biopsies of six patients with confirmed PSS and four patients with symptoms of oral dryness for whom diagnosis of PSS was excluded, referred to as the "non-PSS" cohort (Table 1). Of these cells, 4,292 were sequenced by the full-length cDNA SmartSeq2 method. Alignment of reads to the hg19 genome build was performed using the STAR aligner for RNA sequencing, followed by trimming of adapter sequences and generation of a count matrix using the FeatureCounts tool. The raw matrix of counts per gene per cell was analysed with the

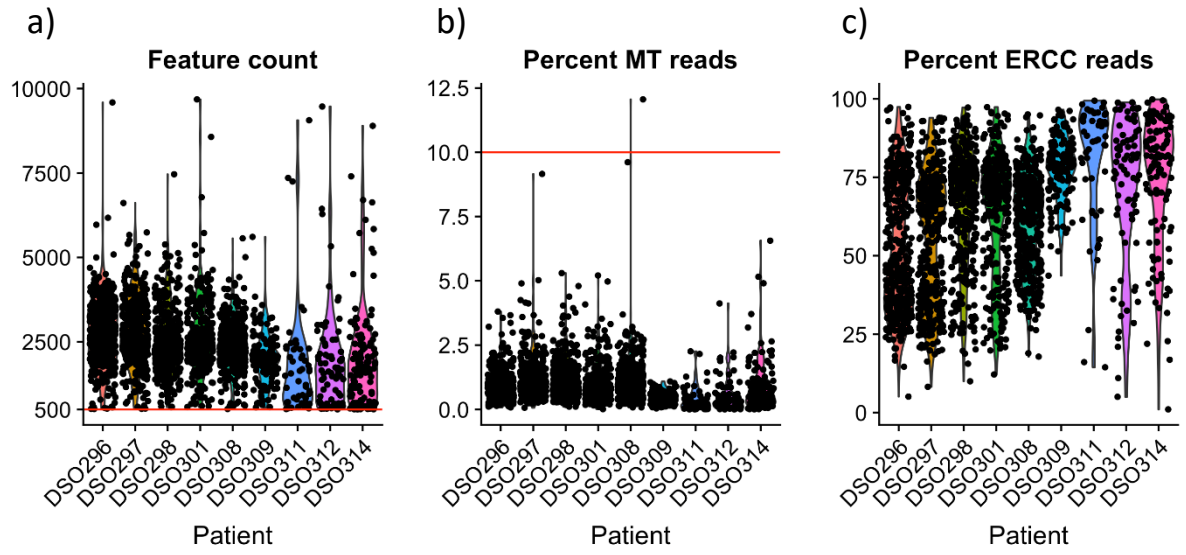


Figure 15. QC of SmartSeq2 cells. **a)** Number of features (genes) detected per cell, grouped by patient; cells with fewer than 500 features removed (denoted by red line). **b)** Percentage of mitochondrial genes per cell, grouped by patient; cells with more than 10% mitochondrial genes subsequently removed (denoted by red line). **c)** Percentage of ERCC spike-in per cell, grouped by patient.

Table 16. Samples sequenced with SmartSeq2 protocol.

Patient	Diagnosis	Focus Score	Age	Sex	Diagnostic notes	# Cells sorted
DSO-297	PSS	2-3	61	F		410
DSO-298	PSS	2-3	74	M		476
DSO-301	PSS	2-3	57	M		412
DSO-308	PSS	2-3	76	F	Gastric MALT lymphoma	556
DSO-312	PSS	2-3	35	F		732
DSO-315	PSS	1	59	F		84
DSO-296	Not PSS	-	29	F	Fibromyalgia	437
DSO-309	Not PSS	-	77	F	Basal cell carcinoma, meningioma	184
DSO-311	Not PSS	-	52	F	Chronic fatigue syndrome	624
DSO-314	Not PSS	-	44	F	Peripheral neuropathy, fibromyalgia	801

SingleCellExperiment and Seurat tools in R (described in Methods). Cells with fewer than 500 genes detected as well as genes present in fewer than four cells were excluded from analysis (**Figure 15a**). Additionally, cells with more than 10,000 genes detected were excluded due to a high likelihood of being "doublets", i.e. two cells instead of one. Cells were also filtered based on mitochondrial content, removing those with greater than 10% mitochondrial content which indicate low quality (**Figure 15b**). Further inspection of gene counts revealed an unexpectedly high proportion of the External RNA Controls Consortium (ERCC) synthetic spike-ins across a majority of cells (**Figure 15c**), which were added as an internal gene control to the RNA sequencing experiment. Despite the observation that ERCC genes frequently accounted for >50% of all reads for a given cell, I was able to proceed with downstream analysis because of sufficient depth of sequencing which enabled detection of biological signal amidst the noise of ERCCs. To minimize the effect of ERCC genes in downstream steps, ERCC percentage was controlled for by linear regression (Methods). Principal component analysis (PCA) confirmed that ERCC content was not a significant source of variability in the dataset, since ERCC genes did not contribute to the main two principal components, and cells with high and low ERCC content did not separate into distinct groups in the PCA projection (**Figure 16a**). Therefore, cells with high ERCC content were not removed from the analysis. After filtering cells by the number of genes detected and the mitochondrial content, the number of cells that passed quality control for further analysis was 2,290.

I.1 Clustering and cell type detection

Downstream analysis of filtered cells required normalization of gene counts, which was done with the Seurat package and entailed normalizing by the total counts in a given cell, multiplying by a scaling factor of 10,000, and log-transforming (Methods). The normalized and scaled gene counts were used in all subsequent steps. To understand and visualize the sources of variation in the dataset, linear and nonlinear dimensionality reduction methods were used: principal component analysis (PCA, **Figure 16a**) and t-distributed stochastic neighbour embedding (tSNE, **Figure 16b-f**), respectively. The t-SNE projection was coloured by cell type identities previously determined by FACS sorting, which showed that cells of the same type (including B cells, T cells, plasma cells, and HLA-DR$^+$ cells) were correctly grouped together by the unsupervised t-SNE algorithm (**Figure 16c**). After labelling the cell type groups which corresponded to given areas of the t-SNE plot, colouring of the t-SNE projection by high and low ERCC expression was informative for determining that cells with a lower proportion of ERCCs were almost exclusively plasma cells and HLA-DR+ antigen presenting cells (**Figure 16b,c**). This finding was biologically plausible, given that both cells types express high levels of transcripts which would lower the relative contribution of ERCCs: plasma cells express very high levels of immunoglobulins and HLA-DR+ cells are large antigen presenting cells with high transcriptional activity.

Clusters of cells with shared characteristics were identified using a graph-based clustering approach, called shared nearest neighbour (SSN) analysis, shown on the t-SNE projection (**Figure 16d**). Many of the identified clusters corresponded to the cell type categories previously determined by FACS sorting (**Figure 16c**), confirming that the SNN analysis clustered cells into groups based on biological function and not by other effects such as inter-patient variation (**Figure 16f**). Colouring the t-SNE projection by patient diagnosis, i.e. whether PSS diagnosis was ultimately confirmed or disproved upon biopsy, showed that some clusters had a higher proportion of cells from biopsies with confirmed PSS (**Figure 16e**). This includes SNN cluster 8, which corresponds to the FACS-sorted B cells (defined by CD45$^+$CD3$^-$ CD19$^+$CD38$^-$ FACS gating). More than 95% of the 120 cells in this cluster originated from

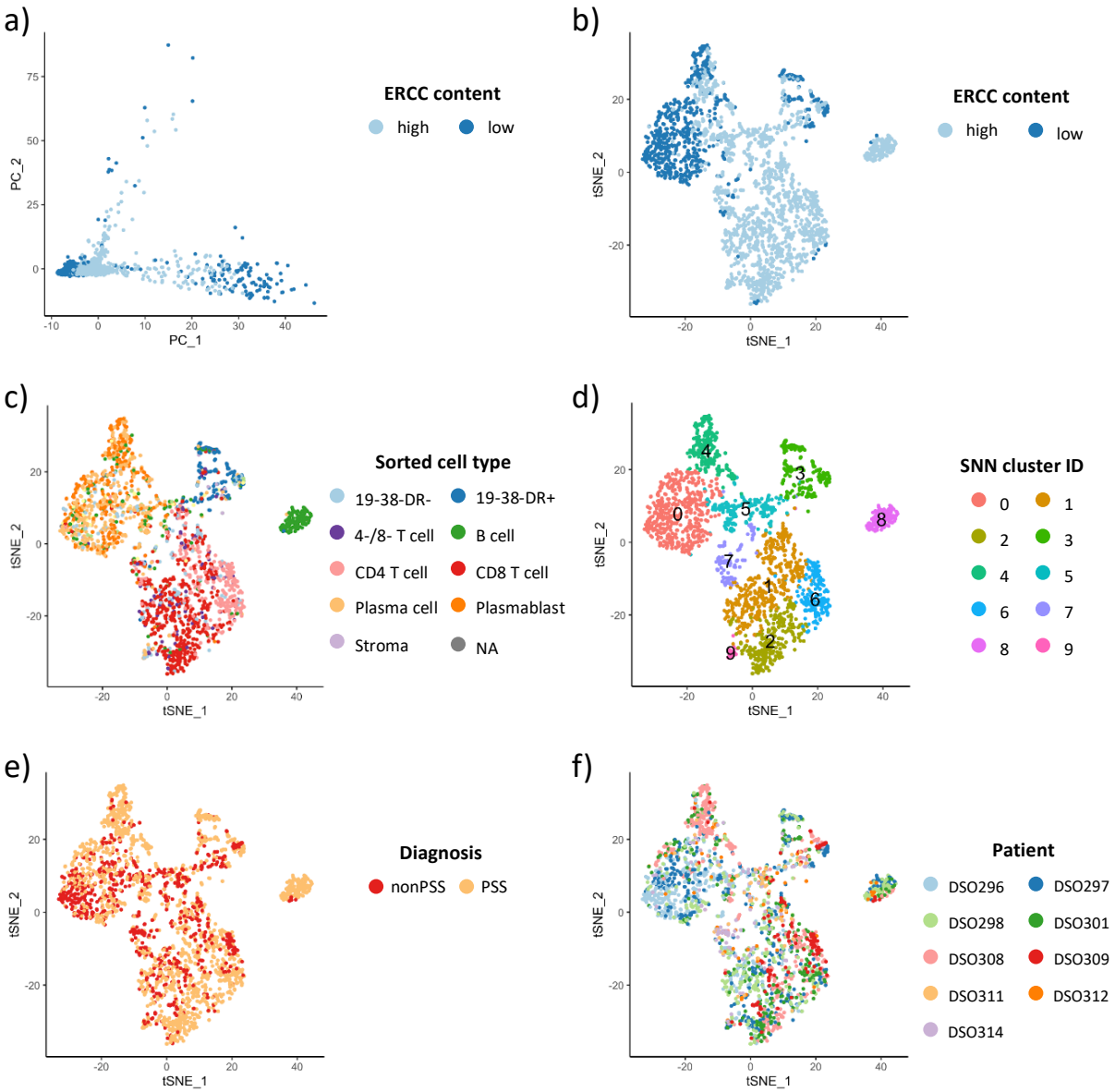


Figure 17. Visualization by dimensionality reduction. **a)** Projection of principal components 1 and 2, coloured by ERCC content of cells (>50% = ”high”, >50% = ”low”) **b)** tSNE projection coloured by ERCC content of cells. **c)** tSNE projection coloured by index sort cell types. **d)** tSNE projection coloured by clusters identified through shared nearest neighbour (SNN) analysis.

confirmed PSS biopsies, suggesting that this cell type may be overrepresented in the disease. Similarly, over 90% of cells in SNN cluster 4, corresponding to FACS-sorted plasmablasts and plasma cells (defined by $CD45^{+}CD3^{-}CD19^{+}CD38^{+}$ and by $CD45^{+}CD3^{-}CD19^{-}CD38^{++}$, respectively), derive from PSS patient biopsies. Conversely, most of the plasmablasts and plasma cells from non-PSS biopsies were grouped into SNN cluster 0. It should be noted that there are about twice as many PSS cells as non-PSS cells in this analysis (1,560 versus 730). To evaluate the relative distribution of PSS and non-PSS cells among the clusters, the percentage of total PSS

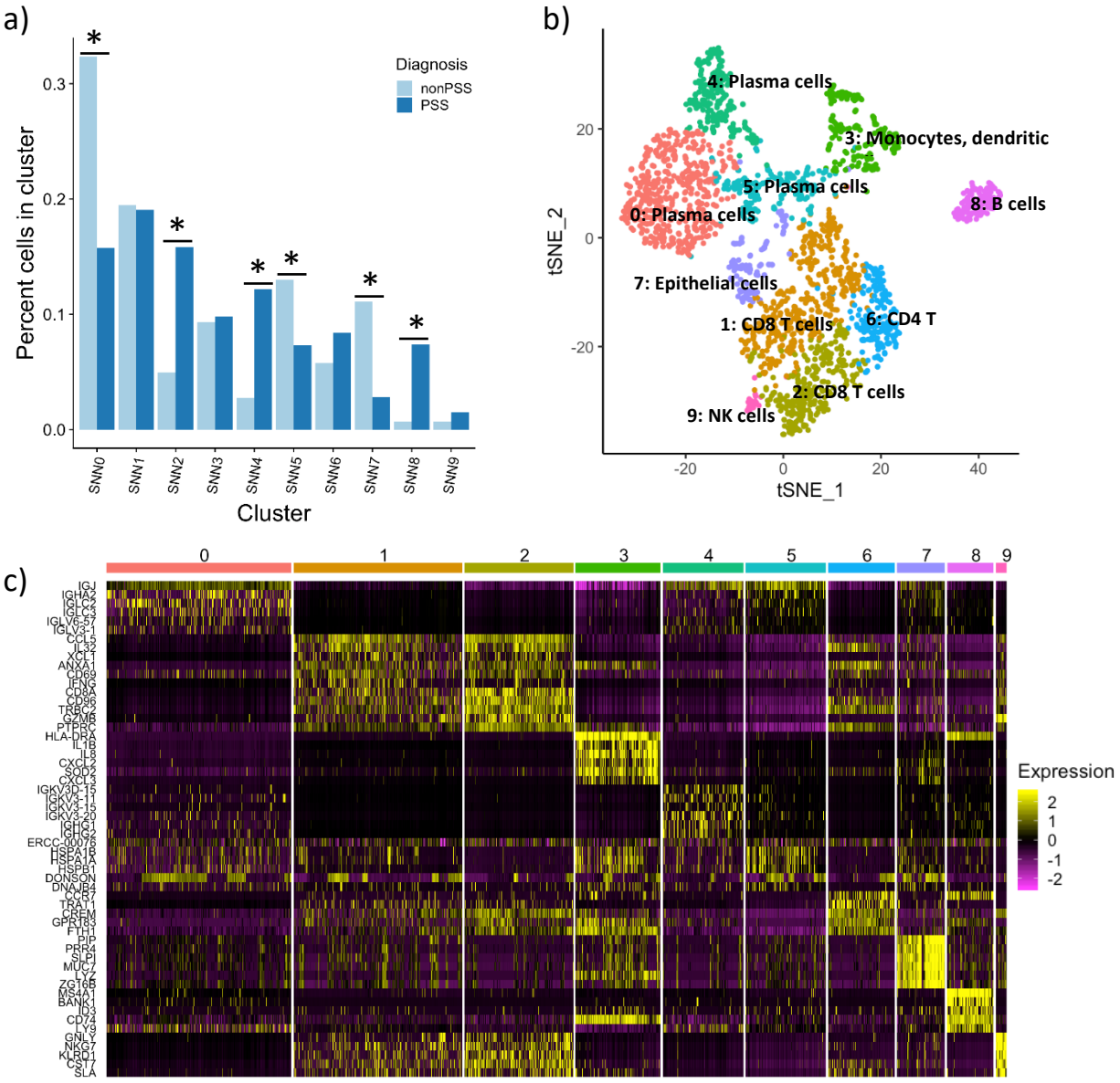


Figure 18. SNN cluster features. **a)** Fraction of total PSS and non-PSS in each cluster, compared by Fisher's exact test (adjusted by multiple testing correction): SNN0 $p=9.9\cdot10^{-18}$; SNN2 $p=3.1\cdot10^{-15}$; SNN4 $p=1.8\cdot10^{-14}$; SNN5 $p=1.7\cdot10^{-4}$; SNN7 $p=9.8\cdot10^{-14}$; SNN8 $p=1.9\cdot10^{-13}$; SNN1,3,6,9 p-value not significant. **b)** Putative cell types annotated on t-SNE projection. **c)**Heatmap showing top six differentially expressed genes from each SNN cluster, coloured by Z-scored expression (where 0 is equal to population mean).

and non-PSS cells forming each cluster was compared by a Fisher's exact test (**Figure 17a**). It became evident that PSS biopsies had a significantly higher proportion of cells belonging to SNN clusters 2,4,5,7, and 8 and a significantly lower proportion of cells in cluster 0 than the non-PSS biopsies. The proportion of total cells contributing to clusters 1,3,6, and 9 was similar between PSS and non-PSS biopsies.

To further interrogate the cell types and functionalities defining the SNN clusters, differentially expressed genes were identified for each cluster by comparing expression levels with a non-parametric test (Wilcox test) in the Seurat suite of tools. The relative expression of the top six biomarkers from each cluster were visualized by a heatmap for comparison (**Figure 17c**). By performing a literature search of the cluster biomarkers in conjunction with known cell type information from the FACS sorting data, I inferred the most likely cell type identity for each cluster (**Figure 17b**). In addition to the expected T cell, B cell, and plasma cell groups that were FACS sorted, the analysis identified the presence of NK cells, dendritic cells, monocytes, and glandular epithelial cells. NK cells co-localized with SNN cluster 9, as suggested by high expression of common NK cell marker genes including granulysin (GNLY), NK cell granule protein 7 (NKG7), and killer cell lectin-like receptor D1 (KLRD1), (**Figure 17c,5c**). Myeloid cells were identified in cluster 3 by the expression of MHC class II, with a subset of cells expressing monocyte markers CD14 and/or CD16 and a subset expressing dendritic cell markers CD1c and CD103 (**Figure 17c, Figure 18c**). Glandular epithelial cells were detected in cluster 7 based on high expression of commonly secreted salivary gland proteins such as salivary mucin 7 (MUC7), lysozyme (LYZ), and prolactin-induced protein (PIP) (**Figure 17c**).

I.2 T cell dynamics

The unsupervised clustering analysis identified one cluster of CD4 T cells (cluster 6) and two clusters of CD8 T cells (clusters 1 and 2). The fraction of total cells from PSS samples in cluster 6 was slightly higher than the fraction of non-PSS (not significant, **Figure 17a**). A majority of these CD4 cells were characterized by a $CCR7^{+}$ profile (**Figure 18a**), suggesting a naïve or central memory CD4 phenotype. However, a comparison of subsets of PSS and non-PSS derived CD4 cells within cluster 6 showed that PSS CD4 cells had lower CCR7 expression ($p<0.001$, Wilcoxon test), slightly lower CD62L/SELL expression (p not significant) and higher STAT4 and CD69 expression ($p<0.05$, **Figure 19a**), indicative of T cell activation, suggesting that in disease the infiltrating CD4 cells might shift towards an effector memory phenotype.

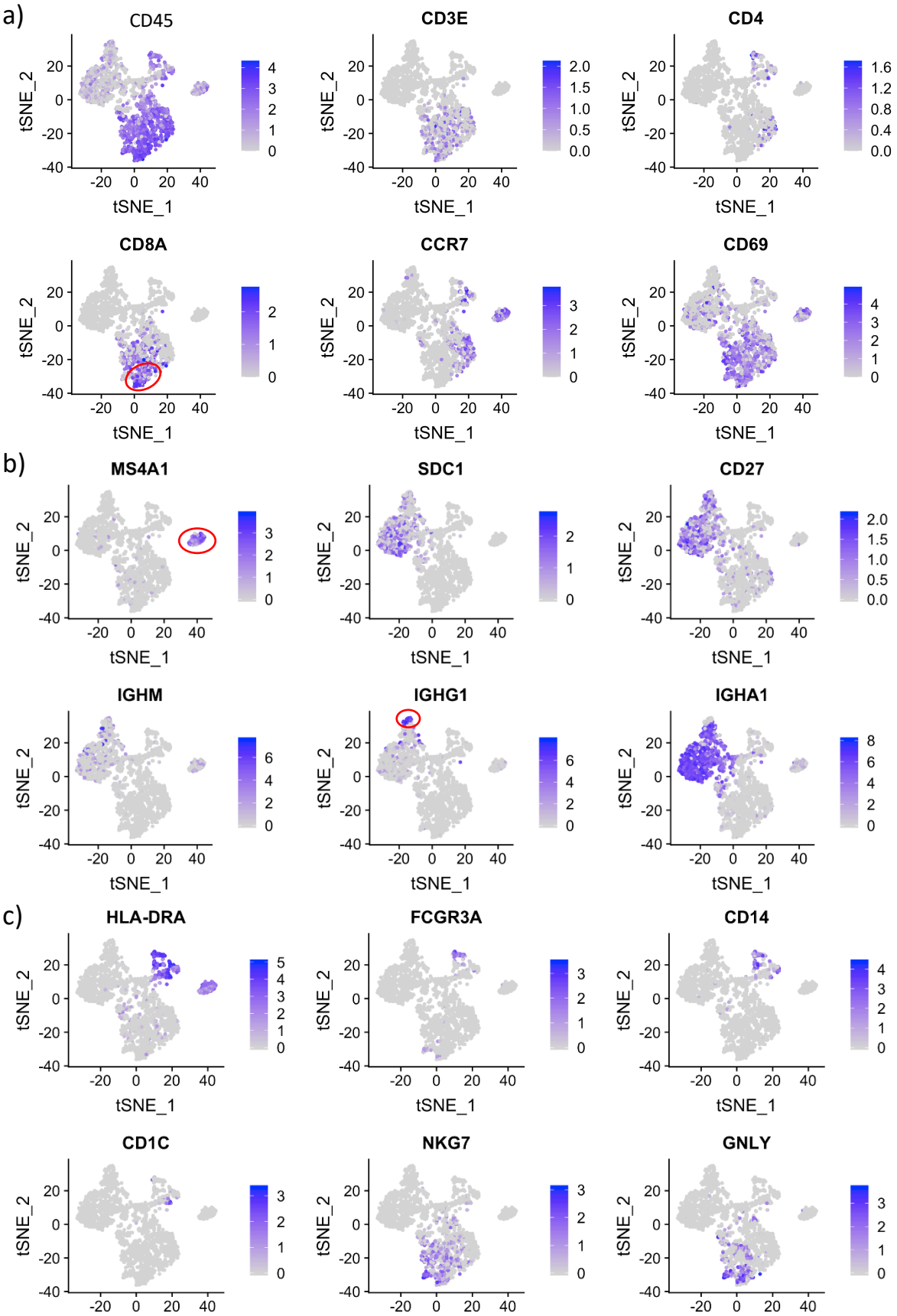


Figure 19. Expression of key markers (log normalized expression). **a)** T cell markers: CD45, CD3, CD4, CD8, CCR7, and CD69. **b)** B cell markers: CD20 (MS4A1), CD38 (SDC1), CD27, IgM, IgG, IgA; groups enriched in PSS cells circled in red. **c)** Myeloid and NK cell markers: HLA-DRA, CD16 (FCGR3A), CD14, CD1C, NK cell granule protein 7 (NKG7), granulysin (GNLY)

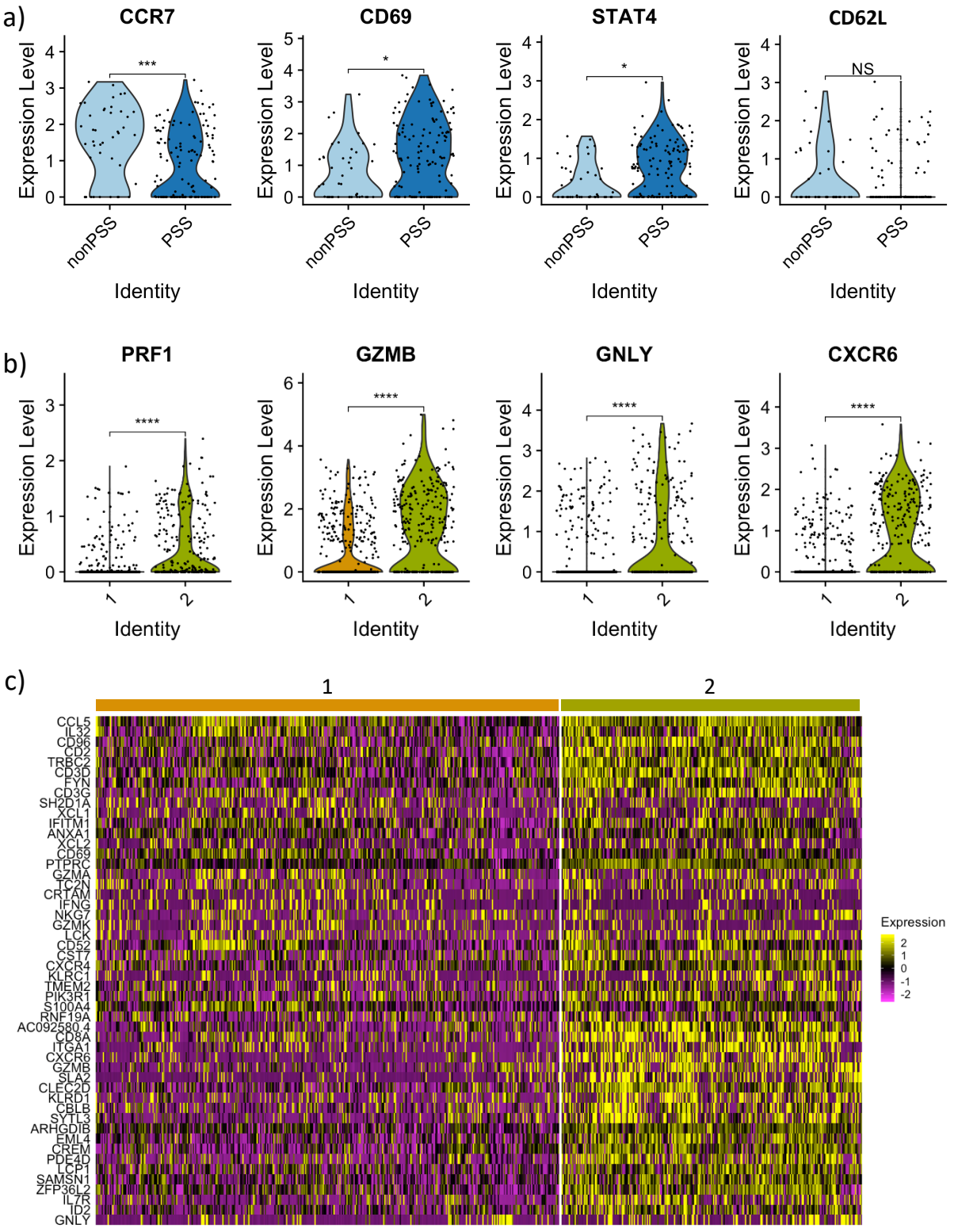


Figure 20. Markers of T cells. **a)** CD4 T cluster 6: comparison of within-cluster PSS and non-PSS cell activation markers CCR7, CD69, STAT4, and CD62L, log normalized expression compared by Wilcoxon rank sum test (* p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001). **b)** CD8 T cell clusters 1 and 2: comparison of perforin, granzyme B, granulysin, and CXCR6 expression. **c)** Heatmap comparing differentially expressed genes between CD8 T cell clusters 1 and 2, Z-scored expression.

Two groups of CD8 T cells were identified in the analysis, defined as clusters 1 and 2. Cluster 2 constituted a significantly higher proportion of total PSS cells than non-PSS cells ($p<10^{-13}$, **Figure 17a**) and was characterized by higher expression of genes related to cytotoxicity and effector function than cluster 1, including perforin, granzyme B, granulysin, and CXCR6 chemokine ($p<0.0001$, **Figure 19b**). CD8 T cells in cluster 2 expressed higher levels of many activation and cytotoxicity-related genes than cluster 1 (**Figure 19c**), suggesting that the excess of cytotoxic T cells present in PSS salivary glands of these patients is made up of cells with a more active effector phenotype than CD8 T cells present in non-PSS biopsies. There was not notably higher expression of MHC class II molecules on these CD8 cells however, as identified by previous phenotyping studies using cytometry by time of flight.

I.3 B cell and plasma cell dynamics

Clustering analysis identified three clusters that are likely comprised of plasma cells and plasmablasts (clusters 0, 4, and 5) and one cluster of cells comprised of the $CD45^{+}CD3^{-}CD19^{+}CD38^{-}$ FACS-sorted population of B cells (cluster 8). Cluster 8 was distinctly separated from the plasma cells clusters on the tSNE projection and is composed of >95% cells derived from PSS biopsies, making this group of cells significantly enriched in the PSS cohort ($p<10^{-12}$, **Figure 17a**). Cluster 8 highly expresses the B cell marker CD20 (MS4A1) and MHC class II molecules (HLA-DRA), but does not highly express CD27, CD138 (SDC1), and CD38 markers of memory B cell and plasma cell differentiation (**Figure 18b**). While a majority of the plasma cells in clusters 0, 4 and 5 expressed high levels of IgA transcripts (**Figure 20a, Figure 18b**), B cell cluster 8 expressed lower levels of IgA, moderate levels of IgM, and a small amount of IgD, suggesting that this group is comprised, at least in part, of IgM^{+} B cells which have not undergone class switch recombination. The high expression of MHC class II molecules implies an antigen presenting function, characteristic of an activated B cell
phenotype. Cluster 8 did not express significant levels of germinal centre markers such as GL7, PD1, CD94, and BCL6 (not shown). The specificity of $CD20^{+}CD38^{-}$ B cells to PSS minor salivary gland biopsies was shown in Chapter 3 with immunohistochemistry staining that demonstrated the presence of $CD20^{-}CD38^{+}$ plasma cells in non-PSS biopsies, while PSS biopsies showed localized aggregates of B cells expressing CD20 in addition to an increased number of plasma cells permeating the tissue throughout (**Chapter 3, Figure 11d,e**).

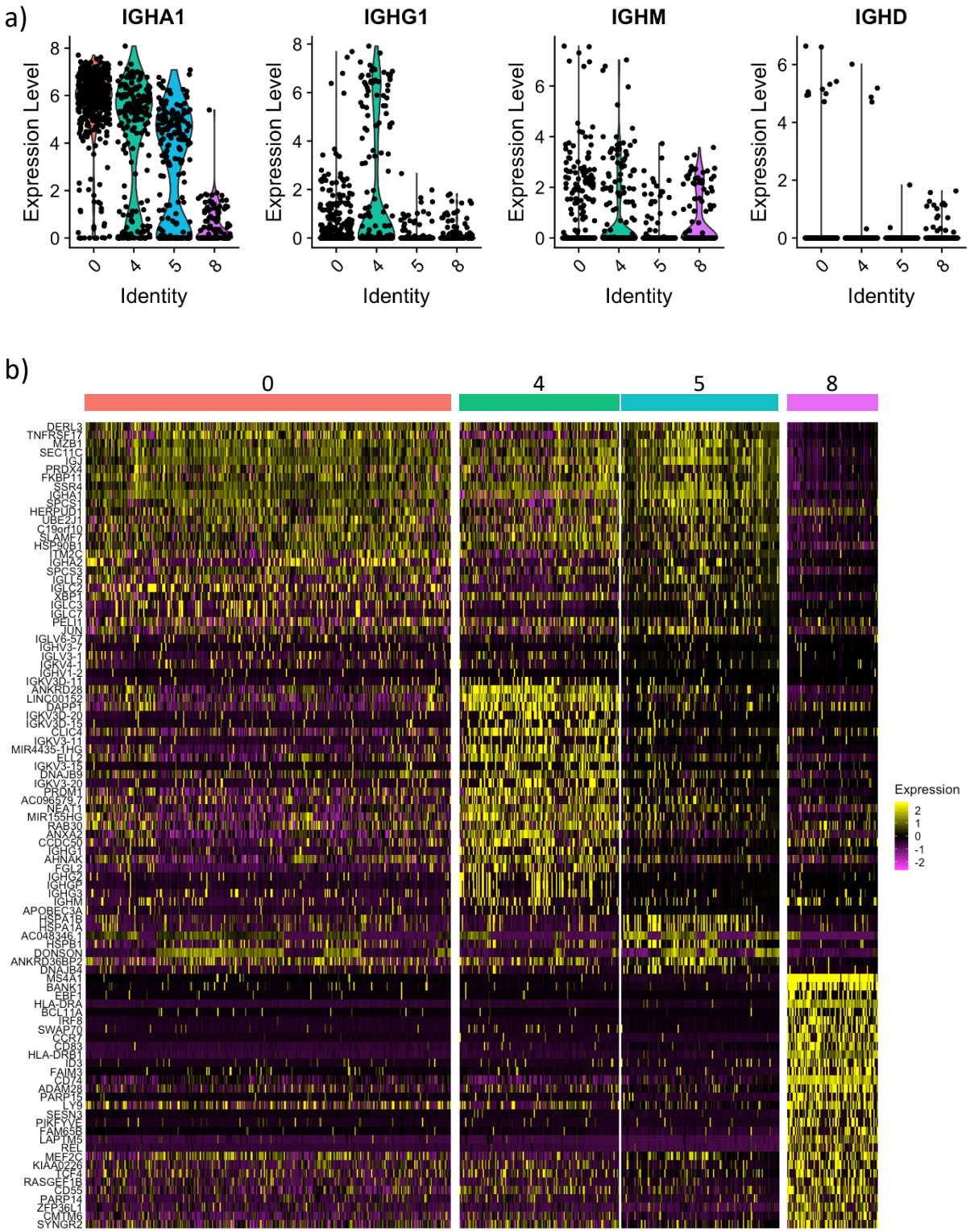


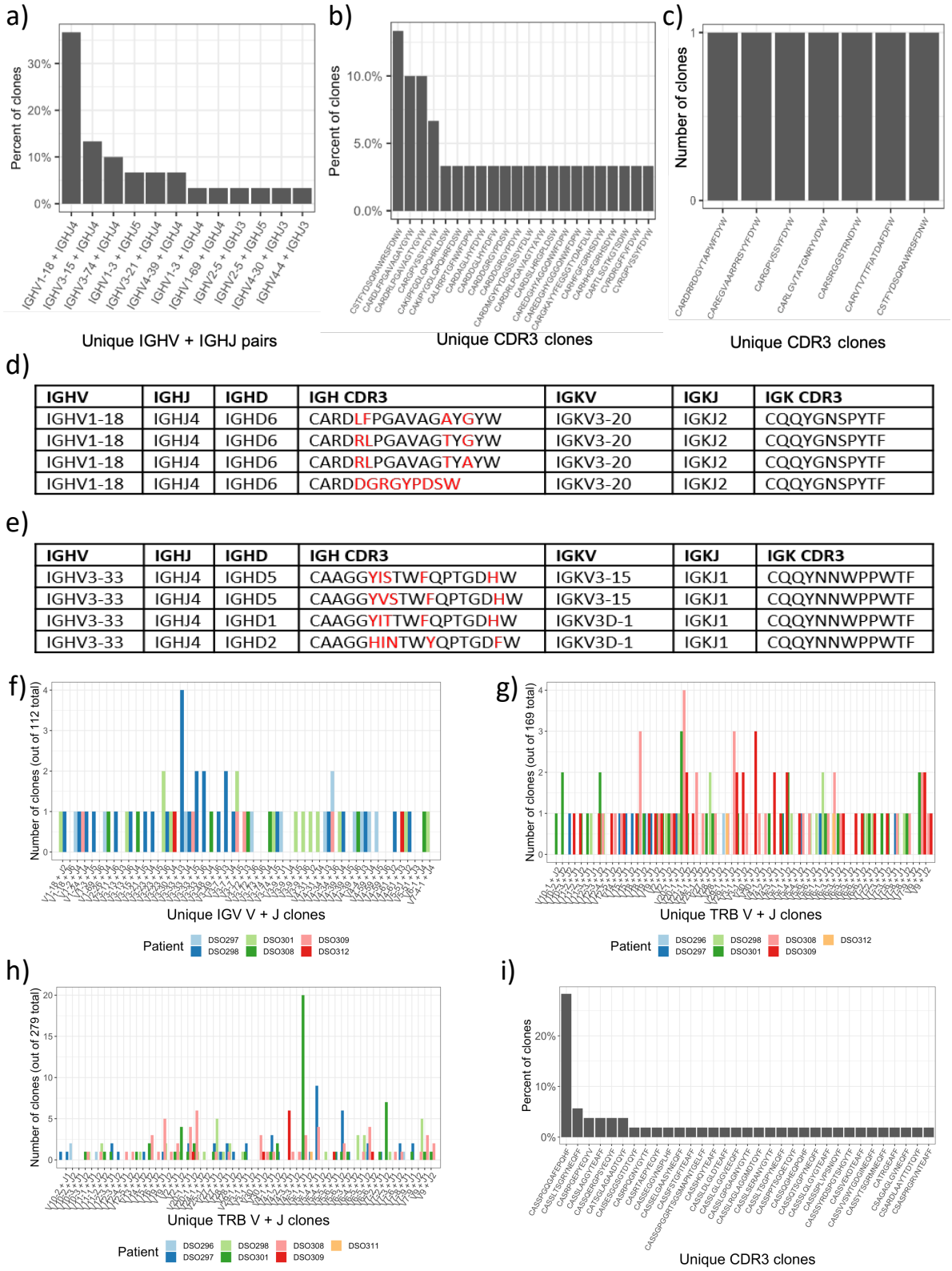
Figure 21. Markers of B cells and plasma cells. **a)** Plasma cell clusters 0, 4, 5 and B cell cluster 8: log normalized expression of immunoglobulins A1, G1, M and D. **b)** Heatmap of differentially expressed genes between plasma cell clusters (0,4,5) and B cell cluster 8 (Z-scored expression).

In addition to B cells, cluster 4 plasma cells were comprised mostly of cells from PSS biopsies (>90%) and contained a significantly higher portion of total PSS than total non-PSS cells ($p=1.8\cdot10^{-14}$, **Figure 17a**). This cluster contained a subset of cells which highly express IgG, setting it apart from the rest of the plasma cell population of clusters 0, 4, and 5 which highly expressed IgA (**Figure 20a**). A group of differentially expressed markers set cluster 4 apart from the other plasma cell clusters, 4 and 5 (**Figure 20b**). These include immunoglobulin genes as well as proliferation-associated genes such as long noncoding RNA LINC00152, which is highly expressed in various cancers[202]. It should be noted that a majority of cells in cluster 4 (64%) derive from patient DSO308, who had a finding of earlier gastric mucosal-associated tissue (MALT) lymphoma noted in their clinical history. While the lymphoma was not found in salivary glands, it is worth keeping in mind the previous history of a B cell disorder in this individual and treating this sample as a potentially special case.

I.4 V(D)J expression in single cells

The full-length cDNA sequences obtained by the Smart-Seq2 library preparation method allowed for extraction of recombined V(D)J sequences that form the B cell and T cell receptors. To extract the expressed V(D)J sequences, the Mixcr algorithm was used, as mentioned in Chapter 3 for extraction of repertoires from bulk DNA sequencing reads (described in Methods). Additionally, the single cell platform allowed matching of heavy and light chain sequences from the same cell. The clonality of lymphocytes by donor and cell type group or cluster was assessed to determine the dynamics of clonal expansion of the infiltrating cells.

As mentioned in the previous section, cluster 4 was populated mostly by plasma cells and plasmablasts from patient DSO308 and contained a subset of cells with high expression of IgG, which stood out from the remainder of cells which highly expressed IgA. To determine if cluster 4 IgG^+ cells from DSO308 were clonal in origin, immunoglobulin heavy chain sequences were inspected. More than a third of these clones (36%) harboured receptors comprised of a heavy chain pairing between IGHV1-18 and IGHJ4 (**Figure 21a**), which formed several distinct but related CDR3 sequences, indicating that hypermutation of the original V(D)J sequence had taken place (**Figure 21b,d**). The overall population of cluster 4 (including



IGHV	IGHJ	IGHD	IGH CDR3	IGKV	IGKJ	IGK CDR3
IGHV1-18	IGHJ4	IGHD6	CARDLFPGAVAGAYGYW	IGKV3-20	IGKJ2	CQQYGNSPYTF
IGHV1-18	IGHJ4	IGHD6	CARDRLPGAVAGTYGYW	IGKV3-20	IGKJ2	CQQYGNSPYTF
IGHV1-18	IGHJ4	IGHD6	CARDRLPGAVAGTYAYW	IGKV3-20	IGKJ2	CQQYGNSPYTF
IGHV1-18	IGHJ4	IGHD6	CARDDGRGYPDSW	IGKV3-20	IGKJ2	CQQYGNSPYTF

IGHV	IGHJ	IGHD	IGH CDR3	IGKV	IGKJ	IGK CDR3
IGHV3-33	IGHJ4	IGHD5	CAAGGYISTWFQPTGDHW	IGKV3-15	IGKJ1	CQQYNNWPPWTF
IGHV3-33	IGHJ4	IGHD5	CAAGGYVSTWFQPTGDHW	IGKV3-15	IGKJ1	CQQYNNWPPWTF
IGHV3-33	IGHJ4	IGHD1	CAAGGYITTWFQPTGDHW	IGKV3D-1	IGKJ1	CQQYNNWPPWTF
IGHV3-33	IGHJ4	IGHD2	CAAGGHINTWYQPTGDFW	IGKV3D-1	IGKJ1	CQQYNNWPPWTF

Figure 22. B and T cell receptor sequence assessment. **a)** IGHV and IGHJ usage of IgG+ cluster 4 plasma cells from patient DSO308. **b)** CDR3 sequences of IgG+ cluster 4 plasma cells from patient DSO308. **c)** Clones found in cluster 8 B cells from patient DSO308. **d)** Receptor sequences of IgG+ cluster 4 plasma cells from patient DSO308. **e)** Receptor sequences of cluster 8 B cells from patient DSO298. **f)** IGHV + IGHJ pairing in cluster 8 B cell clones, coloured by patient. **g)** IGHV + IGHJ pairing in cluster 6 CD4 T cell clones, coloured by patient. **h)** IGHV + IGHJ pairing in cluster 2 CD8 T cell clones, coloured by patient. **i)** T cell heavy chain CDR3 sequences of patient DSO301 cells from CD8 cluster 2.

both IgG^+ and IgA^+ cells) was more oligoclonal, with the largest IGHV/IGHJ clone forming 11% of the cluster (not shown). Cluster 8, the group of B cells highly enriched in the PSS patient population, included only a few cells from individual DSO308. Among them the IGHV1-18 /IGHJ4 clones were not found, however the second most-expanded clone in cluster 4 IgG+ cells, IGHV3-15/IGHJ4 (*CSTFYDSQRAWRSFDNW),* was observed (**Figure 21c**). Therefore, this clone existed both as an activated $CD20^+HLA^-DR^+$ B cell and a class-switched antibody-producing cell, suggesting that maturation into the plasma cell phenotype occurred within the tissue, or alternatively that this clone was selectively recruited to the site.

Receptor analysis of B cell cluster 8, which was comprised of 112 reconstructed sequences, revealed a diverse, largely polyclonal V(D)J landscape with little evidence of clonal expansion. The largest clone was observed in patient DSO298, in which four out of 56 cells harboured the same IGHV3-33/IGHJ4 pairing and the same kappa light chain sequence. Given some variability in the IGHD and IGKV genes identified, along with mismatches in the CDR3 sequence, these four cells may not actually be derived from the same parent cell but instead converged on a similar repertoire sequence. However, this is difficult to parse out definitively due to the similarity of small immunoglobulin genes to each other, which makes exact alignment challenging. Overall, this cluster of B cells, which is highly enriched in the PSS cohort, does not show evidence of significant intra-patient clonal expansion or inter-patient repertoire sharing based on analysis of the 112 reconstructed V(D)J sequences.

As previously discussed, CD8 T cell cluster 2 contained a large proportion of PSS patient cells and displayed a more cytotoxic phenotype than the CD8 T cells of cluster 1. Repertoire analysis found that 20% of this cluster of 279 total cells was made up of an expanded clone of CD8 cells from patient DSO301, the largest clone of CD8 cells observed (**Figure 21g**). This clone was found in 19 out of 68 cells (28%) from patient DSO301 in this cluster (**Figure 21h**), while only two cells with this receptor were found in CD8 cluster 1 (not shown). The remaining clones in cluster 2 were smaller than 10% of the total cells but suggested oligoclonality of receptor usage (**Figure 21g**). CD4 T cells in cluster 6 were notably less clonal, with the largest clone comprising 2% of the total cells (**Figure 21i**). The oligoclonality of effector CD8 T cells observed here aligns with the levels of CD8 clonal expansion observed in bulk DNA sequencing

analysis in Chapter 3 and suggests that activated, clonally expanded cytotoxic T cells may be more prevalent in PSS minor salivary gland than non-PSS controls.

The 2,290 cells analysed in this dataset represent a preliminary set of findings to guide further analysis. For a subsequent batch of samples, we opted to use the 10X Genomics single cell platform to sequence 5' transcripts with enriched V(D)J sequences. The following section will describe how the findings of the droplet-based technique relate to the key findings from the plate-based analysis.

II. Results of droplet-based 5' expression with V(D)J enrichment single cell sequencing (10X Genomics protocol)

In the previous section, the results of the plate-based single cell sequencing of 10 patient biopsies were described. For a further six biopsies, we opted to use a droplet-based platform instead, which has the advantage of higher throughput and obviates the need for FACS sorting of cells, a step which often results in high levels of sample attrition. The batch included three biopsies from patients with confirmed PSS and three from non-PSS controls (**Table 12**). This section will summarize the results of the droplet-based sequencing, focusing on the validation of key findings from the previous dataset rather than a methodical layout of all results, for the sake of brevity.

Table 18. Samples sequenced with 10X Genomics 5' expression with V(D)J expression protocol.

Patient	Diagnosis	Focus Score	Age	Sex	# Cells sequenced
DSO-469	PSS	2	25	F	2,147
DSO-470	PSS	5	32	F	771
DSO-477	PSS	2	61	F	1,398
DSO-478	Non-PSS	-	64	F	1,132
DSO-479	Non-PSS	-	57	F	701
DSO-480	Non-PSS	-	45	F	989

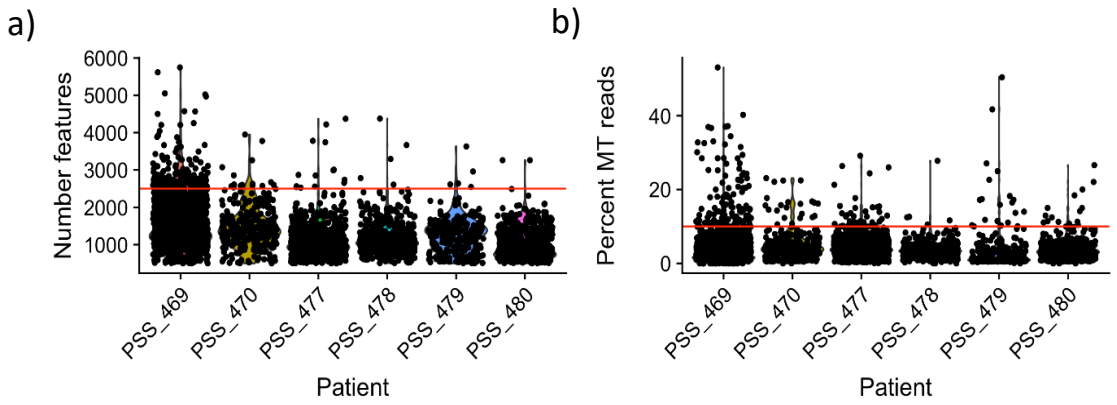
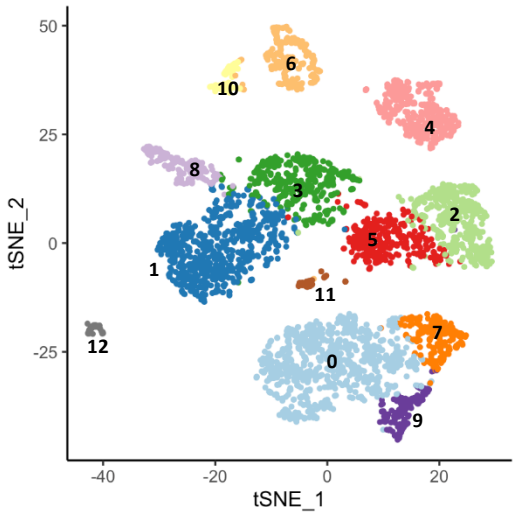


Figure 24. 10X droplet-based single cell sequencing QC, with red lines indicating filtering thresholds. **a)** Total features (genes) detected per cell. **b)** Percent mitochondrial reads per cell.

The droplet-based method allowed for a higher number of cells per biopsy to be sequenced, however the average number of features (genes) detected per cell was lower than with the full-length cDNA platform (**Figure 22a**). For this dataset, cells with greater than 2,500 detected features were excluded to remove doublets. The percentage of mitochondrial DNA per cell was similar between the two platforms, and the same threshold of maximum 10% mitochondrial reads was applied (**Figure 22b**). Once low-quality cells were removed, 3,754 total cells remained for downstream analysis. Normalization, dimensionality reduction, clustering, and differential expression analysis were done in the same way as described previously.

The clustering analysis identified 13 clusters, as shown in the t-SNE projection (**Figure 23a**), and their cell type identities were inferred based on the cluster biomarkers (**Figure 23b**). Cell types identified correlated with those found in the plate-based approach and included B cells, plasma cells, CD4 and CD8 T cells, monocytes, dendritic cells, NK cells, and epithelial cells. Additionally, a cluster of stromal fibroblasts was identified based on high expression of stromal factors such as CXCL12, COL6A2, LUM, FN1, and MMPI4[203]. Like in the previous dataset, the percent contribution of total PSS cells was higher than that of non-PSS cells in B cell, plasma cell, and CD4 T cell clusters; however, unlike in the previous dataset, there was not a significant enrichment in the number of CD8 T cells (Fisher's exact test, **Figure 23d,c**). CD8 clusters 1 and 3 had similar proportions of PSS and non-PSS cells. The differentially expressed genes of these two clusters showed different activation states, with cluster 1 expressing higher levels of MHC class II molecules, interferon-gamma, and granzyme A, demonstrating an activated phenotype, while cluster 3 expressed higher levels of perforin,

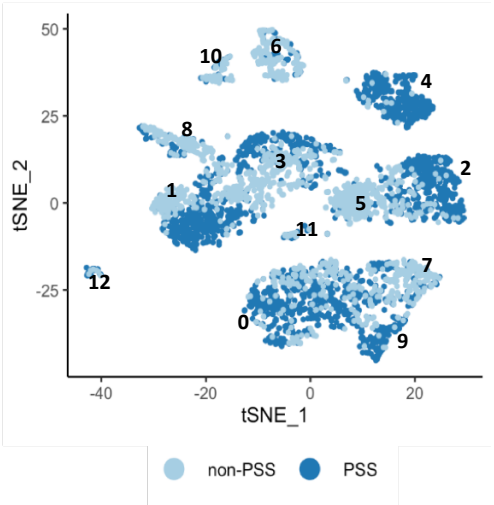
a)



Cluster	Assignment
0	Plasma cells
1	CD8 T cells
2	CD4 T cells
3	CD8 T cells
4	B cells
5	CD4 T cells
6	CD14 monocytes, dendritic cells
7	Plasma cells
8	NK cells
9	Plasma cells
10	CD16 monocytes
11	Stromal fibroblasts
12	Epithelial cells

b)

c)



d)

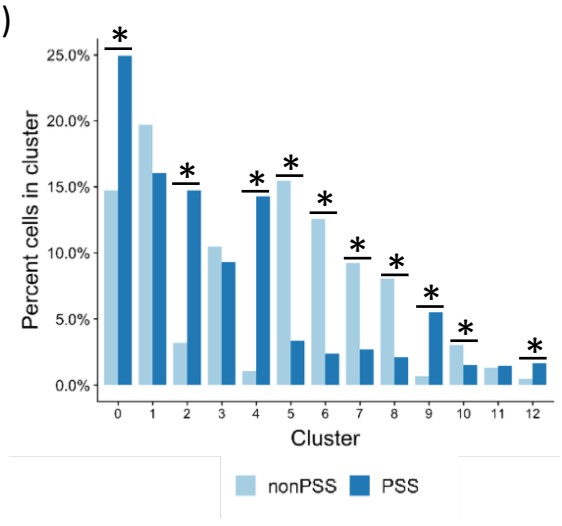


Figure 25. 10X dataset single cell projection and cluster biomarkers. **a)** t-SNE projection with inferred cell type identities of clusters labelled. **b)** Heatmap of top 6 differentially expressed genes per cluster (Z-scored expression). **c)** t-SNE projection labelled by diagnosis. **d)** Percent contribution of total PSS and non-PSS cells to each clusters, compared by Fisher's exact test (Bonferonni adjusted p-values: SNN0=$3.1{\cdot}10^{-13}$, SNN1=$6.3{\cdot}10^{-2}$, SNN2=$3.0{\cdot}10^{-33}$, SNN3=1, SNN4=$4.0{\cdot}10^{-53}$, SNN5=$1.0{\cdot}10^{-38}$, SNN6=$4.7{\cdot}10^{-34}$, SNN7=$1.5{\cdot}10^{-16}$, SNN8=$1.5{\cdot}10^{-16}$, SNN9=$1.9{\cdot}10^{-16}$, SNN10=$3.1{\cdot}10^{-2}$, SNN1=1, SNN12=$1.3{\cdot}10^{-2}$)

granzymes A and K, and CXCR6, suggesting a cytotoxic phenotype. ($p<0.0001$, Wilcox test, **Figure 24a**). Within cluster 3, PSS CD8 cells displayed higher expression levels of perforin, granzymes A and K, and CXCR6 than the non-PSS cell population (**Figure 24a**), suggesting that the disease-associated CD8 T cells are in fact more cytotoxic than in non-PSS sicca, as observed in the previous dataset. Of the CD4 cells, cluster 2 (which is enriched in PSS cells, $p<10^{-32}$), is distinct from CD4 cluster 5 (enriched in non-PSS cells, $p<10^{-37}$) by higher expression of immune checkpoint and other genes associated with T cell activation, including CTLA4 and CD278 (**Figure 23b**), also underscoring the presence of a more active helper T cell population in disease.

The 10X dataset validated the presence of the B cell population highly enriched in the disease setting: the $CD20^{+}HLA^{-}DR^{+}$ B cells of cluster 4 (>95% PSS-derived cells). Likewise, an IgG producing plasma cell population (cluster 9) was highly enriched in PSS, as in the previous dataset (**Figure 24a, Figure 23c**). Both of these disease-specific clusters contained a large number of cells from donor PSS-477, so to examine potential clonal expansion in this patient,

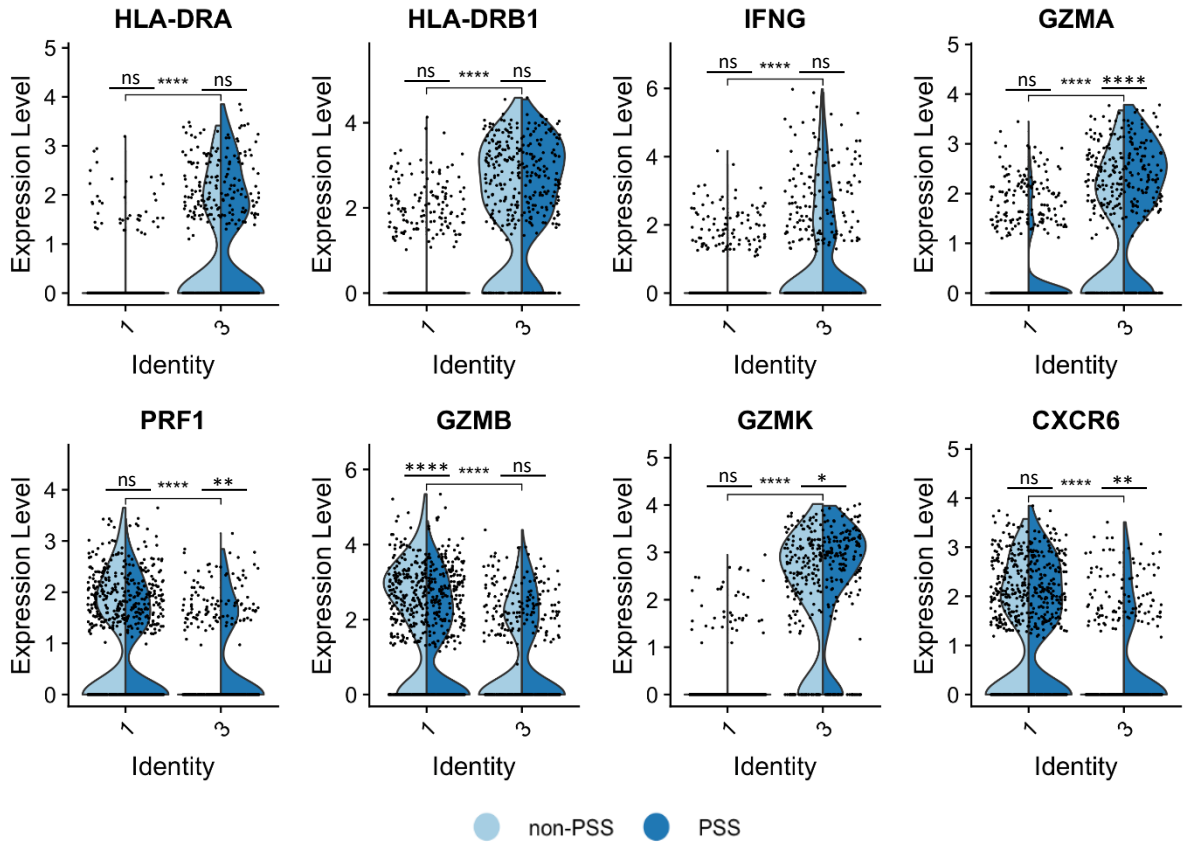


Figure 27. CD8 T cell markers. Comparison of CD8 T cell markers between cluster 1 and 3, as well as between PSS and non-PSS cells within each group (Wilcoxon rank sum test).

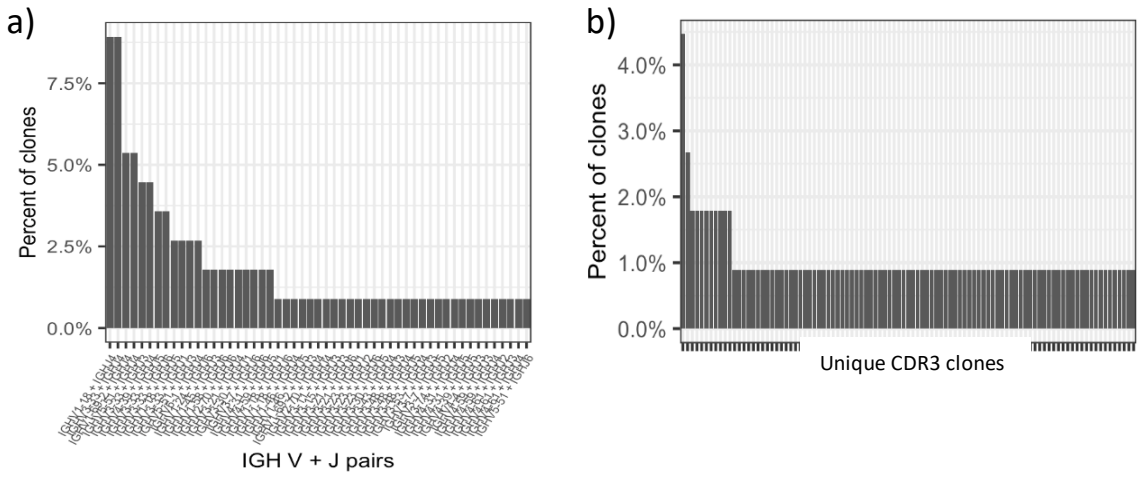


Figure 28. Heavy chain BCR repertoire rearrangements in cluster 9 plasma cells of patient PSS277. **a)** IGHV and IGHJ pairing of clones. **b)** Unique CDR3 clones (sequences not shown).

the V(D)J repertoire was assessed. Repertoire sequences were obtained from the 10X CellRanger algorithm, which assembled the PCR-enriched V(D)J libraries. As before, the receptor sequences were oligoclonal, with the largest clone in this patient comprising 8% of the repertoire of this patient's plasma cells in cluster 9 (**Figure 25a**). Cluster 4 B cells had few successfully assembled repertoires, and of those few, the main clones found in the patient's plasma cells were not identified (not shown). While the 10X platform worked well for V(D)J assembly of plasma cells, it performed less well than the full-length cDNA approach for reconstructing repertoires of B cells in this group of samples. Overall, the B cell and plasma cell profiles were consistent across both datasets and highlighted two disease-specific populations with varying degrees of oligoclonal expansion.

III. Discussion

III.1 Summary of approach

In this chapter, I undertook an analysis of single cell transcriptome dynamics of the infiltrating lymphocytes in PSS and non-PSS control biopsies. To our knowledge, this was the first study to analyse single cell genome-wide expression profiles in minor salivary glands of PSS. The approach was two-fold: sequencing one batch of samples with a full-length cDNA plate-based method (Smart-Seq2) and another with a droplet-based 5' enriched cDNA method (10X Genomics). The two methods displayed distinct strengths and weaknesses for this type of tissue-derived sample, which was limited in cell number and required an enzyme-based tissue

disaggregation step. The full-length cDNA method, which involved fluorescence-activated cell sorting (FACS) prior to library preparation, had lower throughput in terms of number of cells successfully sequenced per donor. However, it had the advantage of more genes detected per cell, full length transcripts, and a more robust ability to detect V(D)J recombination. Conversely, the droplet-based method obviated the need for fluorescence-activated sorting of single cells, a step which results in significant sample attrition, yielding a higher number of successfully sequenced cells per donor. Despite these technical differences, both approaches yielded several similar conclusions.

III.2 B cell findings

Previous studies of tissue-infiltrating lymphocytes in minor salivary glands highlighted the presence of plasma cells and CD4 T cells as likely key cell types involved in the production of pathogenic autoantibodies[88,169]. Numerous studies have also described that, in some patients, infiltrating lymphocytes are organized into germinal centre-like structures that facilitate the production of antigen-specific plasma cells through T-cell help, and this feature is associated with worse clinical outcomes[100,204]. While the expression of germinal centre markers was not detected in plasma cells and B cells in this dataset, I did observe two distinct features of B lineage cells in the PSS patient cohort that differed from those of patients with non-PSS sicca. Firstly, the presence of $CD20^{+}HLADR^{+}CD38^{-}CD27^{-}$ B cells was significantly enriched, with >95% of cells in this category deriving from PSS samples in both datasets. These B cells expressed higher levels of IgM and IgD, and lower levels of IgG and IgA, than plasma cells from the biopsies, suggested a mature activated B cell phenotype before class switching. This population of cells was also observed by immunohistochemistry, described in Chapter 3, wherein PSS biopsies displayed discrete clusters of $CD20^{+}$ cells surrounded by diffuse plasma cell infiltration, which was not seen in non-PSS specimen. These B cell clusters were more likely to be clonal than surrounding plasma cells. The presence of this B cell population is in accordance with previous studies, including a cytometry by time of flight (CYTOF)[112] investigating PSS infiltrating lymphocytes, which also described an increased number of glandular B cells in PSS compared to non-PSS, phenotyped by $CD19^{+}CD38^{-}CD27^{-}$ expression. Given the physiological presence of plasma cells but not B cells in mucosal-associated lymphoid tissue, the finding of this B cell population in PSS is highly relevant. Whether they

are enriched in auto-reactive B cells that actively differentiate into autoantibody-producing plasma cells in the salivary glands is important to investigate in future functional studies.

The number of infiltrating plasma cells in PSS biopsies was not significantly higher than in non-PSS biopsies in this analysis, a finding which also agrees with the CYTOF study[112]. However, within the plasma cell population profiled here, there was an IgG producing subset highly enriched in the PSS samples, which was distinct from the remainder of plasma cells which expressed IgA. This group of IgG^+ plasma cells comprised a small proportion of the overall plasma cell population, but it was more clonal than the IgA^+ plasma cells, with up to one third of sequenced cells in a given patient harbouring the same V(D)J recombination. Within these clonal V(D)J sequences, there was evidence of somatic hypermutation. In some cases, it was possible to identify the same V(D)J clone in IgG^+ plasma cells and in the $CD20^+$ B cells from the same patient, indicating either that B cell maturation was occurring within the gland or that this particular clone was being selectively recruited to the site of disease from circulation. Given the strong association of autoantibodies with disease, the PSS-specific IgG^+ plasma cells are strong candidate for further studies evaluating their pathogenicity.

III.3 T cell findings

As mentioned previously, there has long been interest in the role of CD4 T helper cells as mediators of B cell activation and the autoimmune humoral response which is characteristic of PSS. Indeed, our study shows a higher number of activated CD4 T cells in PSS-derived biopsies as compared to non-PSS. Overall numbers of infiltrating CD4 cells do not appear to be increased in PSS, and they display largely a naïve or central memory phenotype. Conversely, the CD8 infiltrating T cells of PSS and non-PSS display a terminal effector memory phenotype. However, when comparing PSS and non-PSS CD8 T cells, the PSS group has markedly higher expression of cytotoxicity genes, such as those encoding perforin, granzymes, and granulysin. The increased presence of activated CD8 T cells was also noted by the CYTOF study[112], based on high expression of MHC class II molecule HLA-DRA. In this study, I characterized this further by showing a terminal effector CD8 phenotype, as well as evidence of oligoclonal expansion in some patients. This finding underscores the emerging role of cytotoxic T cells in PSS and other autoimmune diseases, which has seen increased scientific

interest in recent years. This has been suggested by several mouse and human studies[126,205,206], including a recent study of T cells in psoriatic arthritis which identified clonal expansion of activated cytotoxic T cells at the site of disease[207].

The findings of bulk DNA sequencing in Chapter 3 also highlighted the potential role of CD8 T cells in PSS, as they frequently showed evidence of monosomy X in female samples, along with truncating mutations in the X-linked tumour suppressor *KDM6A*. This finding was unexpected given the supposition that B cells and CD4 T cell are key players in PSS; however, it seems that cytotoxic T cells might play a more crucial role than previously recognized. This notion is supported by the transcriptomic finding of significant upregulation in genes related to cytotoxicity in CD8 cells. Ideally, my analyses would have been performed on matched DNA and RNA sequencing data, to be able to compare the transcriptomic profile of cells with X chromosome alterations to those without it. However, this was not technically feasible with our protocol. Assessment of X loss in female cells by single cell RNA sequencing is challenging due to X chromosome inactivation, since physiologically only one copy of X is expressed. The exception to this are genes which escape X-inactivation (*KDM6A* being one of them), though many of these genes were lowly expressed in this dataset and did not show significant difference in expression between CD8 T cells and other cell types. Therefore, evaluating the transcriptional effects of monosomy X ultimately requires a paired single cell genome and transcriptome approach or acquiring enough sample to perform both analyses on bulk cells of the same tissue.

III.4 Limitations and future directions

This study used two single-cell methods to profile the transcriptomes of infiltrating lymphocytes. The main biological limitation underlying both approaches was the low number of input cells from the small tissue biopsies obtained. To overcome this, more tissue will be obtained from additional donors, as having more cells would strengthen the conclusions of the analysis thus far. Additionally, validation of the observations by a secondary method (flow cytometry) is currently underway.

CHAPTER 5: DNA sequencing of salivary gland epithelial cells

While the initial focus of this thesis project was to examine the somatic mutational dynamics of tissue-infiltrating lymphocytes in primary Sjögren's syndrome (as described in previous chapters), we also explored a complementary avenue examining the mutational landscape of affected epithelial cells in the minor salivary glands of patients and controls. Lymphocytes are considered the primary effectors of the autoimmune response in PSS, as evidenced by frequent serum autoantibody findings and numerous markers of an altered humoral immune response[78]. Nevertheless, affected exocrine gland tissue has itself been recently shown to have a more active role in promoting disease than previously thought, through a phenotype of "activated" epithelial cells that promote an immune response through elevated expression of Toll-like receptors and other immune-stimulating factors[90,113]. The drivers of this epithelial cell activation are not known.

As discussed in the Introduction section, recent years have seen an increased interest in research of somatic mutations in normal tissue and chronic disease. There are now several examples of somatic evolution in the context of non-malignant disease that point to distinct selective pressures shaping the disease evolution. Profiling of colonic crypts in inflammatory bowel disease (IBD) points to an enrichment of mutations in the interleukin-17 signalling pathway that are thought to impart a selective advantage to clones by evasion of inflammatory damage[56–58]. Thus, the somatic mutations found in IBD appear to be an adaptation to the disease-mediated damage. A similar phenomenon is observed in non-alcoholic fatty liver disease, where clonal selection favours mutations that drive hepatocyte regeneration and evade lipotoxicity[29,54]. In rheumatoid arthritis synovium, recurrently observed *TP53* mutations are similarly thought to be an adaptation to inflammation[131,132,208], since *TP53* mutant clones are known to outcompete wildtype cells in an environment of oxidative stress[27]. The *TP53* mutant clones could subsequently proliferate and invade surrounding bone tissue irrespective of an inflammatory stimulus, and thus potentially propagate joint disease of their own accord. It should be noted that the findings of *TP53* mutations in rheumatoid arthritis synovium date back to the 1990s and early 2000s and have not been validated by next generation sequencing methods. Taken together, these findings

highlight the importance of further investigation of somatic mutations in tissues affected by inflammatory disorders to elucidate patterns of somatic mutation underpinning clonal adaption, disease progression, and possibly disease initiation.

An atlas of somatic mutation trends across all normal tissues is currently in the making, and as of yet, the genomes of normal, non-malignant salivary glands have not been sequenced. In our study, we utilized minor salivary gland biopsies from PSS patients and those determined not to have PSS upon biopsy. As with my study of lymphocytes, I compared the genomic landscape of salivary epithelial cells in PSS to non-PSS controls, bearing in mind that the control biopsies were obtained from individuals with complaints of oral dryness similar to that of PSS but who ultimately did not satisfy histological diagnostic criteria for PSS. This caveat notwithstanding, most of the non-PSS control biopsies appeared histologically normal and could be used to glean new insights about the genomic landscape of normal salivary gland epithelium. Therefore, my study of minor salivary glands is at the same time an investigation into inflamed tissue affected by the autoimmune response in PSS and a characterization of mutational dynamics in "normal" non-malignant salivary gland epithelium.

By comparing PSS patient and non-PSS control biopsies I hoped to detect features of the mutational landscape that might differ between disease and normal state. In the previous examples, the mutational burden in IBD was higher than that of normal colon[58], as was the mutational burden in chronic liver disease compared to healthy liver[29]. It therefore seems that in an inflammatory microenvironment the rate of mutation accumulation is altered, likely due to increased cell turnover and regeneration following inflammatory insult. The goal of this study was to comparatively investigate the genomic features that characterize salivary epithelium in PSS and non-PSS biopsies.

STUDY AIMS

1. *Describe the mutational burden, clonality, driver events, and mutational signatures operative in acinar and ductal epithelium of normal minor salivary glands.*

   As the normal salivary gland has never been genomically profiled, I sought to describe its mutational features in order to infer tissue-specific dynamics. Similar studies of other normal tissues have revealed trends in mutation accumulation rates, selection of driver genes, unique mutational signatures, etc. that are informative of underlying processes in the tissue and can be extrapolated into its tendency for cancerous transformation.

2. *Compare the mutational features of salivary glandular epithelium between PSS and non-PSS donors.*

   Using the non-PSS donor samples as controls, I attempted to identify any features that may be unique to the disease and potentially related to its pathogenesis. As suggested by the sequencing studies of IBD-affected colon, rheumatoid arthritis synovium, and non-alcoholic fatty liver, we might expect to find similar evidence of clonal selection and increased mutation burden in PSS minor salivary glands relative to controls.

STUDY DESIGN

To address the aims described above, I used laser capture microdissection (LCM) to extract sections of individual acini and ducts from histology slides. The biopsy specimen were the same ones as those used for LCM extraction of lymphocytic aggregates, described in Chapter 3. The cohort of biopsy donors includes seven who are negative for PSS and thirteen with confirmed PSS diagnoses (**Table 13**).

As previously discussed, minor salivary gland biopsies of patients with PSS frequently have several hallmark histological features, most notably lymphocytic aggregates of >50 clustered cells, a feature termed focal lymphocytic sialadenitis. Lymphocytic infiltration is accompanied

by degradation of mucous and serous acini, as well as fibrosis and dilation of ducts. The destruction of glandular microarchitecture in PSS biopsies is evident when compared to biopsies from non-PSS donors, which appear histologically normal (**Figure 26a,b**).

LCM was used to dissect sections of individual acini and ducts from the tissue specimen (**Figure 26 d,e**). By extracting these discrete histological structures, we aimed to maximize the probability of selecting clonal cell populations. Because cross sections of acini and small ducts often contain few nuclei, this required dissecting the same feature from 5-6 multiple successive tissue sections to get enough genetic material for making sequencing libraries (**Figure 26 c-e**). As previously described, libraries were made through a bespoke protocol allowing low per-sample input, down to ~100 cells (Methods). The pairing of LCM with the low-input library preparation method allowed dissection of samples with a sufficient number of cells for library preparation but small enough to capture individual clones.

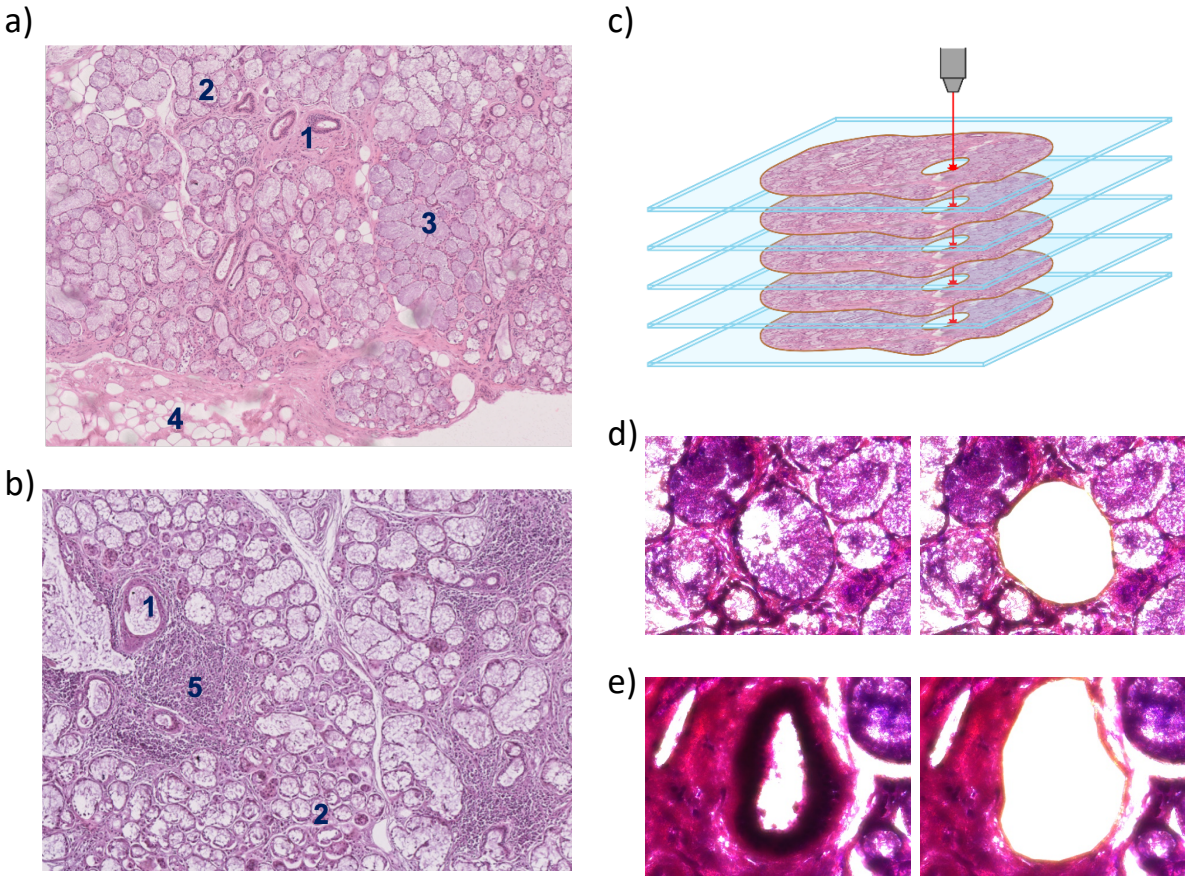


Figure 29. a) H&E of minor salivary gland section from biopsy of an individual negative for PSS and **b)** from an individual with confirmed PSS. Numbers indicate histological features as following, 1 – ducts, 2 – mucous acini, 3 – serous acini, 4 – fibroadipose tissue, 5 – lymphocytic aggregate. **c)** Depiction of LCM dissection of identical feature across multiple Z-stacked tissue sections. **d,e)** LCM microscope image before (*left*) and after (right) dissection of a single acinus and a single duct, respectively.

RESULTS

I.1 Whole genome and exome sequencing of glandular epithelium

Samples of acinar or ductal epithelium with adequate genomic material were sequenced by whole genome (WGS) or whole exome sequencing (WES). A minimum library concentration of 3 ng/ul was required for WGS, and a minimum of 10 ng/ul required for whole exome pulldown and sequencing. Due to low cellularity of the dissected features, many samples did not pass these thresholds. This was in part intentional, since I wanted to dissect samples of minimum required size for library preparation, in order to increase the probability of capturing individual clones. Samples with higher amounts of DNA were largely from microdissections of ducts, and many of these were processed for WES. Samples that underwent WGS have roughly equal contributions from acini and ducts. For downstream analysis of WGS samples, only those with a minimum mean coverage of 10X were considered (**Figure 27a**). Whole genome sequences of 86 samples were generated from four non-PSS and seven PSS donors, and an additional 62 whole exome sequenced were generated from a subset of the same donors plus one additional PSS patient donor (**Table 13**).

For somatic mutation calling, we used the Caveman algorithm for single nucleotide substitutions and Pindel for small insertions and deletions (Methods), bearing in mind the limitation that in some samples low coverage may hinder the ability to detect variants present at lower allele frequency. Variants were called against an unmatched normal, and a filter to remove LCM-specific artefacts was applied. Germline variants were subsequently removed based on consensus between all samples from a given donor, using an exact binomial model

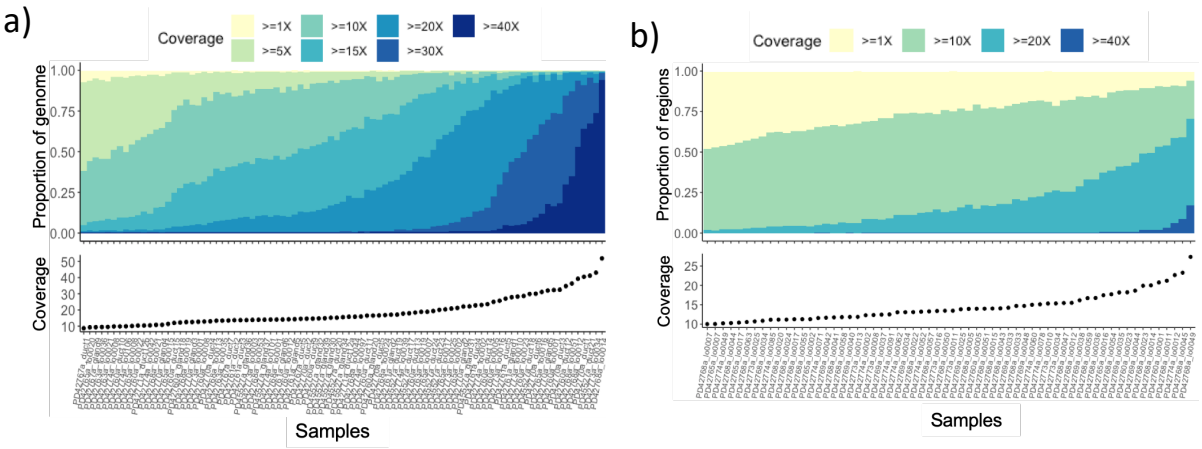


Figure 30. Breadth and depth of coverage of **a)** WGS samples and **b)** WES samples. Coverage ranges in breadth from at least 1x (yellow) to at least 40x (dark blue).

approach. An additional filter based on a beta binomial model was used to remove low quality variants (Methods).

Table 19. Cohort of non-PSS and confirmed PSS biopsy donors.

Sample	*Diagnosis*	*Focus score*	*Sex*	*Age*	*WGS samples*	*WES samples*
PD42760	Non-PSS	0	F	51	14	3
PD42763	Non-PSS	0	F	59	4	-
PD42765	Non-PSS	0	F	59	6	4
PD42761	Non-PSS	0	M	67	8	-
PD42767	PSS	3	-	22	4	-
PD42768	PSS	3	F	31	10	28
PD42769	PSS				-	10
PD45527	PSS	5	F	68	14	-
PD42773	PSS	3	M	71	6	7
PD42770	PSS	3	M	72	10	-
PD42771	PSS	4	F	-	3	-
PD42774	PSS	5	F	-	9	10

I.2 Clonality and phylogenetic relationships

Understanding the clonality of tissue structures not only aids in the discovery of somatic variants, it also informs us of the developmental dynamics underlying its microarchitecture. To assess the clonality of the samples, we considered the median variant allele frequency (VAF) of somatic variants from WGS (**Figure 28a,b**). For reference, a median VAF of 0.5 would suggest a perfectly clonal population where the same somatic heterozygous variants are present in every cell. Practically speaking, this is difficult to achieve and a VAF greater than 0.3 implies a near-clonal population where 60% or more of cells are derived from a single ancestral clone. The median VAF distribution across samples ranged from 0.15 to >0.3, with a trend of higher clonality in acinar samples. The median overall VAF across all acini was

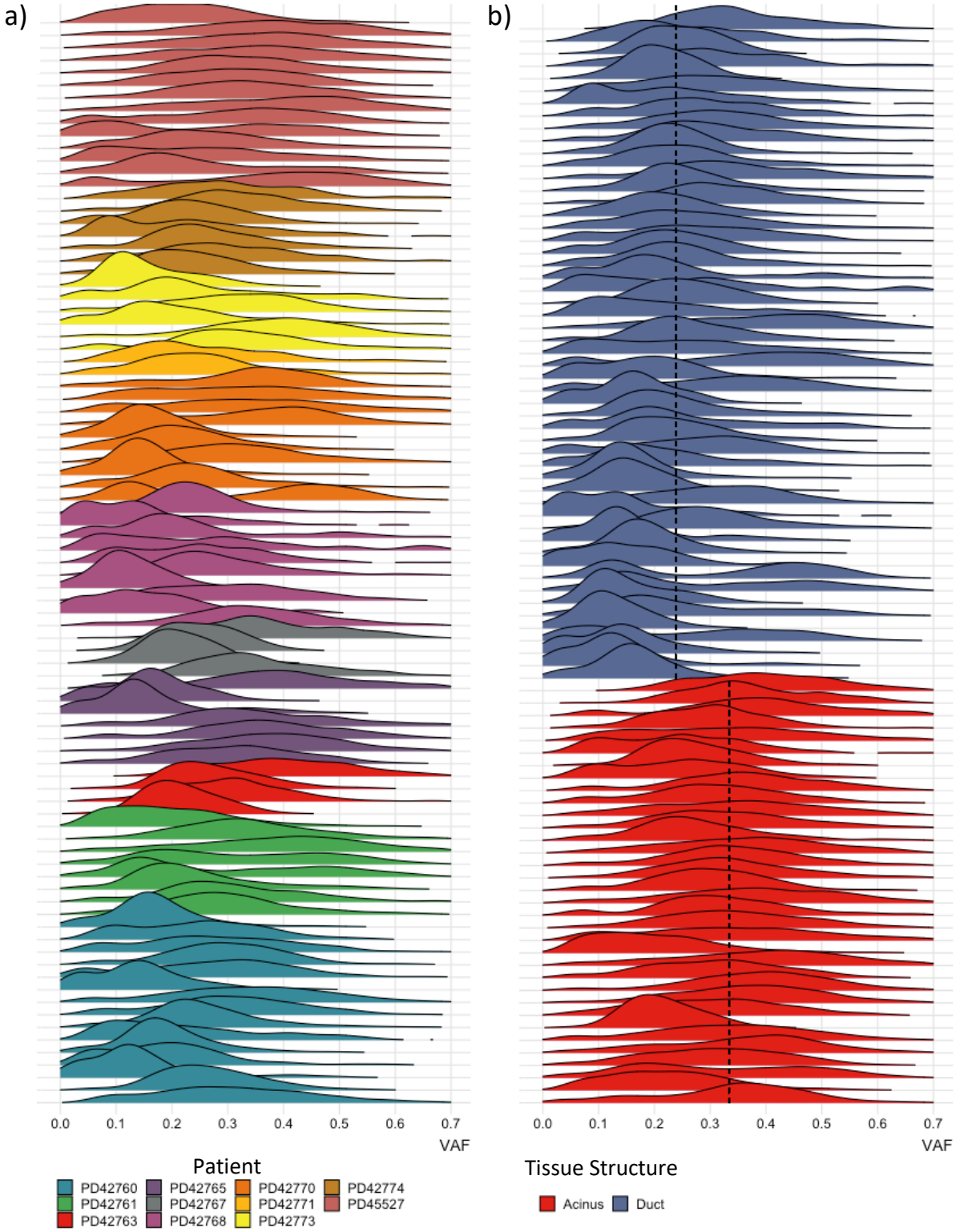


Figure 31. Variant allele frequency distribution of somatic variants genome-wide in epithelial cell samples. Kernel density estimation of VAF in all samples, coloured by **a)** patient of origin and **b)** type of feature (acinus or duct). Dashed line denotes median VAF of all acini and duct samples, respectively.

0.33, while the median VAF of all ductal samples was 0.24 ($p=2.2\cdot10^{16}$, Welch's T test). This suggests that a single acinus is a clonal unit, comprised of daughter cells originating from a single ancestral progenitor cell, while ducts are larger, polyclonal structures that originate from two or more progenitor cells. This finding was consistent across both PSS and non-PSS biopsies and is similar to that of other healthy glandular tissues that have been genomically profiled in our laboratory, such as breast and prostate tissue (unpublished data).

To construct phylogenetic trees of single base substitutions for the purposes of detecting shared common origins of samples, we used a maximum parsimony approach (Methods). Due to low sequencing depth or low clonality of some samples, the reconstructed phylogenies had limited resolution. Despite this, we observed that nearby structures often have shared mutations, while physically distant structures do not. In cases where adjacent acini were sequenced, we found that they are often clonally related but not identical, with branch divergence caused by their respective private mutations (**Figure 29a-d**). Acini and ducts from PSS biopsies did not appear to have more clonal relatedness than those from non-PSS biopsies. As such, there was no obvious evidence of clonal dominance that might occur under the cytotoxic pressure of an inflammatory environment. However, given that there were a limited number of samples sequenced per donor and that low sequencing depth frequently precluded high resolution clustering, a confident conclusion cannot be made.

I.3 Mutational burden and driver detection

The mutational burden of a tissue is an important feature that is often significantly altered in cancer and other disease states, the latter of which is evidenced by findings of elevated mutation number in chronic liver disease[29] and inflammatory bowel disease[5,58] compared to their respective normal tissues. To evaluate the effect of various factors that might impact the mutational burden across minor salivary gland samples, a linear mixed effects (LME) regression model was used. The model evaluated the contribution of age, diagnosis, and sex on the number of mutations per sample, while accounting for inter-patient variation. The number of mutations per sample used in the model was adjusted by a sensitivity coefficient that accounted for the mean depth and VAF of each sample (Methods). The LME model

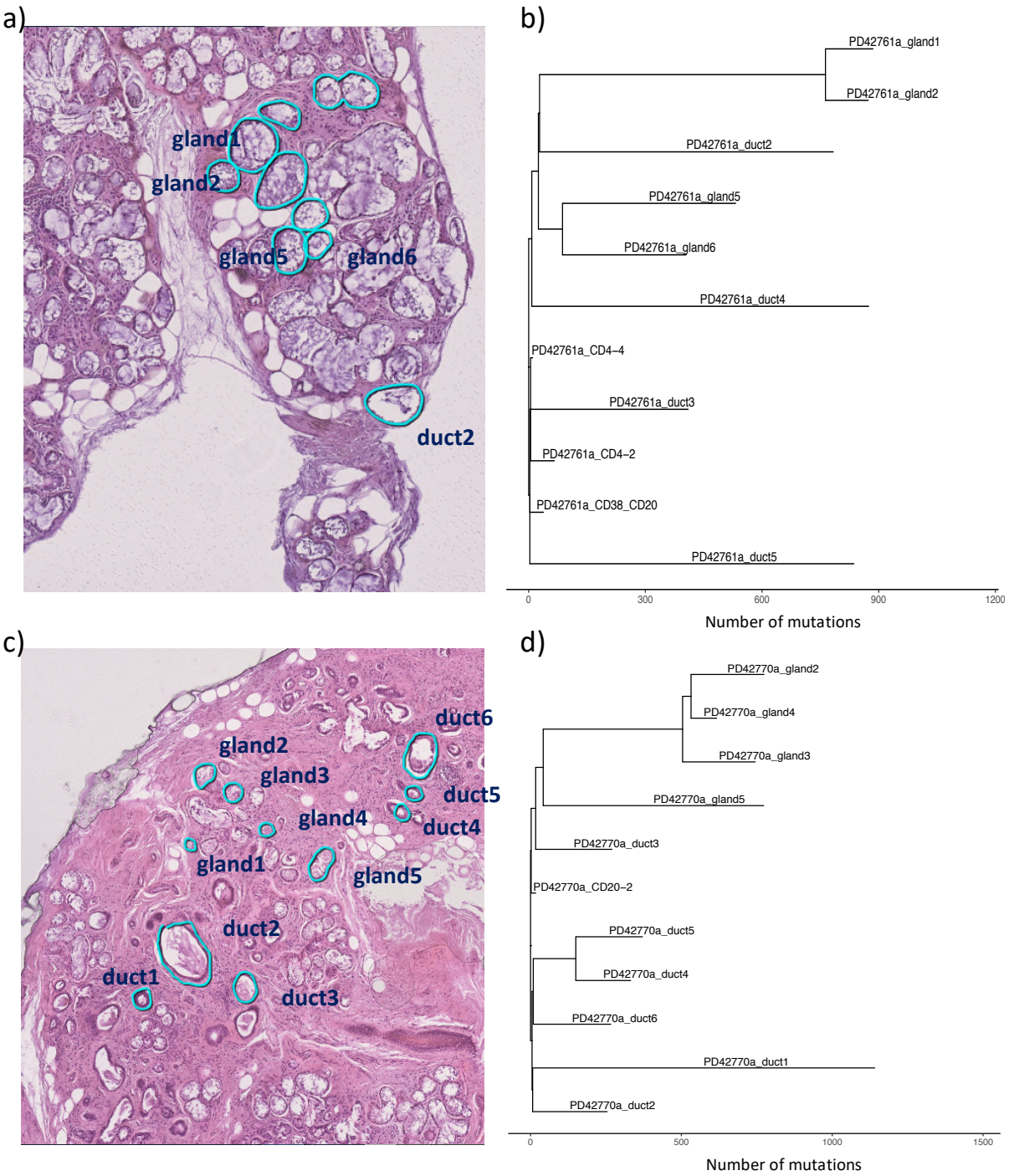


Figure 32. Histology images showing microdissected acinar and ductal samples (circled), along with phylogenetic trees of single nucleotide variants reconstructed from these samples. a) H&E stained section from non-PSS patient biopsy PD42761, b) phylogeny from PD42761 variants, c) H&E stained section from PSS patient biopsy PD42771, d) phylogeny from PD42771 variants.

found that mutational burden was not significantly impacted by diagnosis or sex, however it did vary significantly with age of the donor (p-value=0.0003419, LME regression Chi-squared test, **Figure 30**). The increase in mutational burden with age has been observed consistently in previous sequencing studies of normal tissue[3,5], and this finding confirms the expected

trend. The linearity of this relationship implies that mutations are accumulated throughout life at a constant rate in glandular epithelium. The lack of overt effect of PSS diagnosis on mutational burden suggests that perhaps the inflammatory microenvironment does not contribute a significant excess of mutations to the epithelium, which we might have expected. However, with a small cohort of seven PSS and four non-PSS samples, the analysis is underpowered to detect the true effect of PSS diagnosis on mutational burden.

Next, we sought to determine whether there were putative driver mutations among the somatic variants observed. To do this, we combined the single nucleotide substitution and small indel calls from the WGS and WES datasets and used a dN/dS approach to detect genes under selection (Methods). This was done first in an unbiased way across all genes, and then in a restricted hypothesis manner with a set of known cancer genes curated from TCGA (The Cancer Genome Atlas) database. Neither approach identified any genes under significant positive selection in the epithelium. However, we observed that missense mutations in the *IGFN1* gene occurred in samples from four PSS donors and zero non-PSS donors, though this finding was not statistically significant by dN/dS analysis. The function of *IGFN1* has not been elucidated and it is not thought to a cancer driver gene, however one of the *IGFN1* variants identified has previously been reported in two cancer samples in the COSMIC database. In the GTEx gene expression database[209], salivary glands did not express high levels of *IGFN1,*

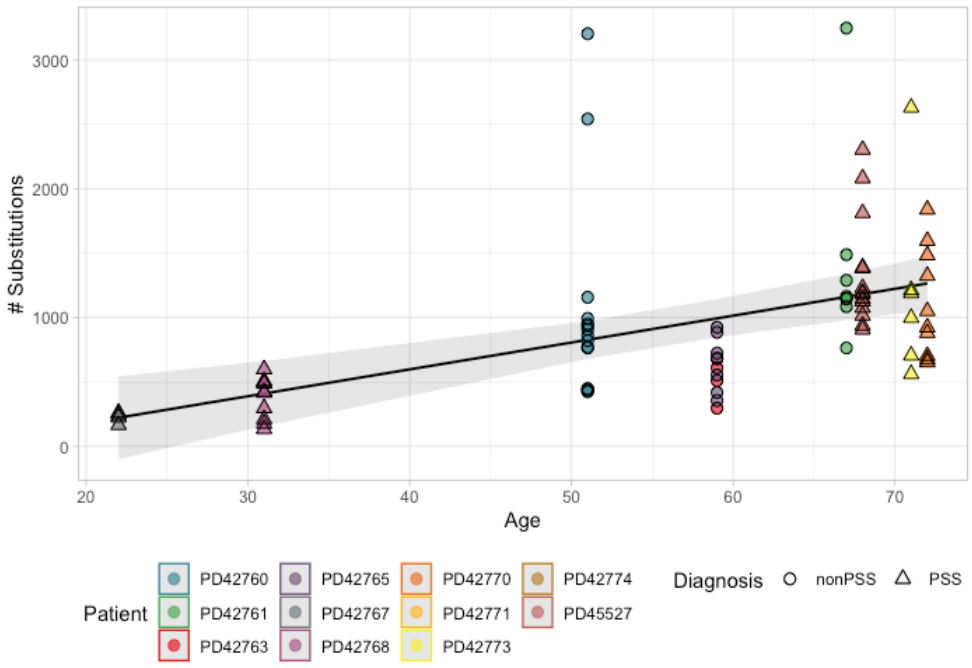


Figure 33. Linear mixed effects model of mutational burden variation per sample (acini or duct) with age, by patient (colour) and by diagnosis (shape).

therefore there is currently no strong evidence of functional relevance of these mutations to PSS.

A few individual samples, from both the PSS and non-PSS cohorts, harboured mutations in cancer-associated genes such as *RB1*, *ARID1B,* and *FBXW10*. However, none of the mutations were in known hotspots identified by previous cancer sequencing studies or were obviously deleterious by variant effect prediction algorithms, so their functional effect is not clear. Genes that were frequently mutated in cancers of the salivary gland (such as *TP53, HRAS, NOTCH1, EGFR,* and genes in the cyclin and PIK3 pathways)[210–212] were not mutated in this cohort. We also did not observe the hallmark *MYB-NFIB* fusion that is commonly found in adenoid cystic carcinoma of salivary gland[210,213]. Overall, with the small dataset analysed we did not have enough statistical power to confidently identify genes under positive selection, therefore we cannot exclude the possibility of their existence without analysing a larger cohort of PSS and non-PSS patients.

I.4 Mutational signature analysis

To identify the mutagens acting on glandular epithelium, we performed mutational signature analysis using a signature extraction approach based on a Bayesian hierarchical Dirichlet process (Methods). This approach enables both the extraction of novel signatures and the matching of identified components to known signatures in the COSMIC database[40], which were provided as priors for the analysis. The algorithm identified the ageing ("clock-like") signatures SBS1 and SBS5 ubiquitously across the samples (**Figure 31f,g**), which was expected based on prior sequencing studies of normal tissues[3,25,31,33]. In addition, several samples exhibited a high contribution of signatures SBS2 and SBS13 to their mutational burden, which are associated with the APOBEC family of DNA cytidine deaminases[214] (**Figure 31a-c,f,g)**. The presence of APOBEC-associated signatures accounts for a higher mutational burden in samples that harbour it (**Figure 31e**). Signatures SBS2 and SBS13 are seen in both PSS (12 out of 53 samples, 23%) and non-PSS biopsies (7 out of 34 samples, 21%). Interestingly, almost all samples with high contributions of these signatures originated from ductal epithelium.

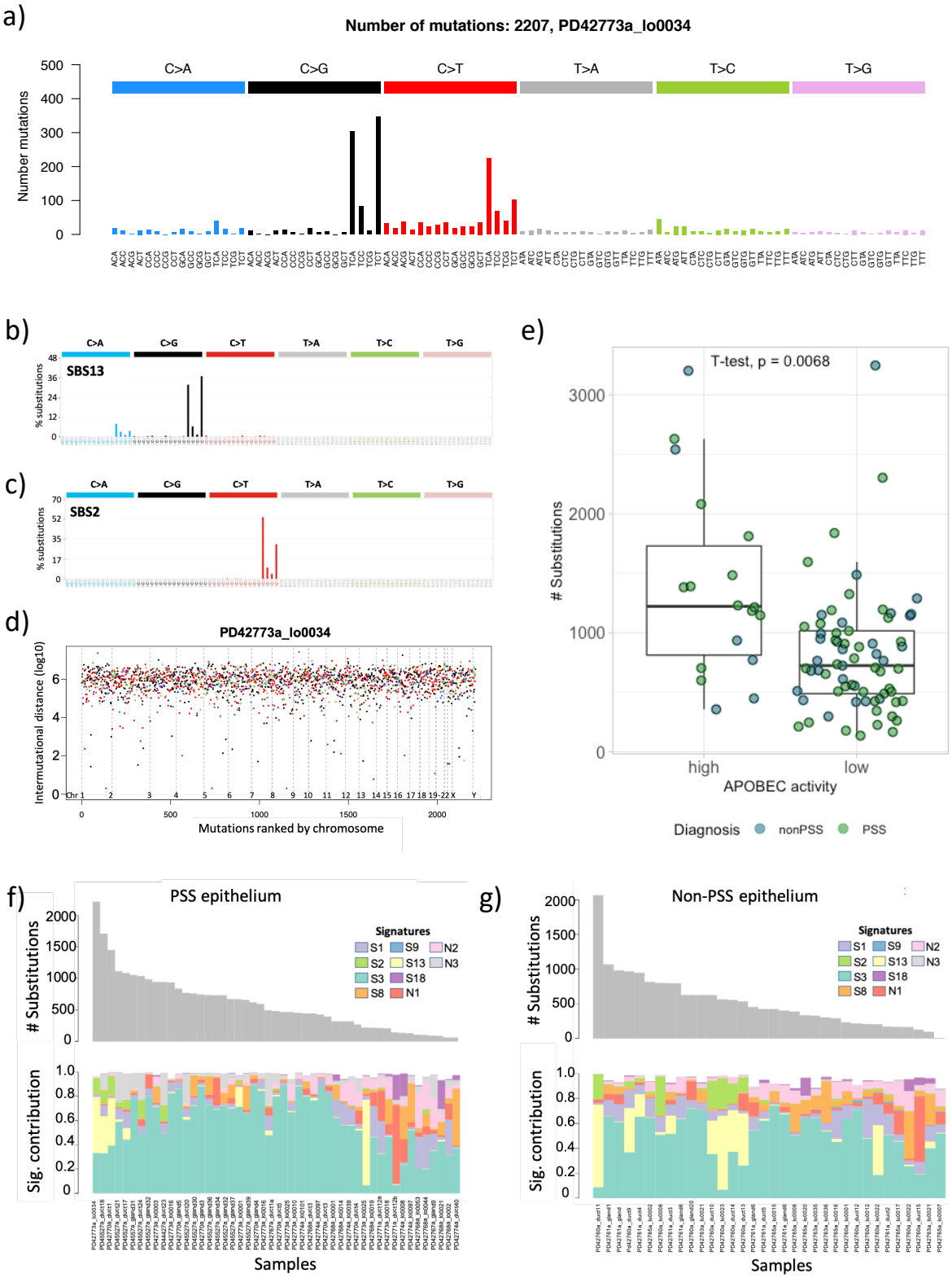


Figure 34. Mutational signature analysis. **a)** The trinucleotide spectrum of mutations in a sample with high APOBEC contribution. **b,c)** The SBS13 and SBS2 APOBEC-associated signatures from the COSMIC database, respectively. **d)** Intermutational distance plot of sample with high APOBEC contribution showing regional clustering of mutations. **e)** Total number of mutations in samples with high and low APOBEC contribution. **f,g)** Mutation count and signature contribution in PSS and non-PSS epithelial samples, respectively; "S" refers to known signatures from COSMIC database, "N" refers to novel signatures identified by HDP algorithm.

APOBEC activity has been observed in various cancers[10] but less commonly in normal tissues[2,4,215]. It is primarily associated with cellular defence against viral integration and retrotransposon jumping[216,217] but thought to also be activated in the context of cancer and tissue inflammation as well[39,214]. The mechanisms of APOBEC activation are not well understood. It is thought to often exert its hypermutation activity in bursts, called kataegis[39], although we did not see evidence of mutational clustering in samples with high APOBEC activity, instead observing an even distribution of mutations across the genome (**Figure 31d)**. Previous sequencing studies of salivary gland tumours have also identified APOBEC as a significant contributor to their mutational landscape[211,218], a finding that paired with the analysis of normal tissue described here, suggests a pervasiveness of this process in salivary glands, both in healthy state and in cancer.

I.5 Metagenomic analysis of viral elements

As mentioned, APOBEC-associated signatures have been seen in various cancers[39,214,219]. In particular, they have been found in cancers with known association to oncogenic viruses, such as cervical cancer[10,220,221]. Given the suspected physiological function of APOBEC in viral defence, we sought to determine the presence of viral DNA in samples with and without APOBEC signatures. We used the GOTTCHA algorithm to align reads from WGS samples to a library of viral reference genomes (Methods). Taking into consideration all hits of human-tropic viruses, we identified numerous instances of lesser known strains of human papillomavirus (HPV 10, 41, 131) as well as several instances of Epstein-Barr virus, in both non-PSS and PSS samples (**Table 14**). For reference, viral elements found in tissue-derived lymphocyte samples are also shown. Viral sequences found in epithelial samples were limited to HPV and EBV, while those found in lymphocytes were mainly from other human herpesvirus strains (6 and 7).

Out of 87 epithelial samples, 19 showed significant contribution of APOBEC mutagenesis (>5% of observed mutations), while 20 samples had viral sequences detected; however only one sample (PD42773a_lo0001) had both viral presence and APOBEC signatures, indicating an overall mutual exclusion of the two observations (p=0.06, Fisher exact test). Since APOBEC is often activated as a defence against the integration of retroviruses[217], it is plausible that cells

which activated APOBEC enzymes have successfully prevented viral integration into the genome during infection, while those that failed to activate APOBEC were more permissive to integration.

These observations suggest that salivary epithelial cells are a frequent target of viral infection, which sometimes results in retroviral integration into the genome. Neither HPV nor EBV sequences were found exclusively in samples from patients with PSS diagnosis, therefore there is not an apparent association of virus with the disease. However, this does not exclude

Table 20. Sequences of human viruses found in genomes of epithelial and lymphocyte samples. Samples excluded from list did not have any findings of human-tropic viruses.

EPITHELIUM						
Sample	**Feature**	**Taxa/strain**	**Rel. abundance**	**Length**	**Bases mapped**	**Hit count**
PD42760a_duct10	duct	HPV type 41	0.1107	665	1079	31
PD42760a_gland18	acinus	HPV type 131	0.2093	285	873	27
PD42760a_lo0001	duct	HPV type 41	0.1442	684	1557	46
PD42760a_lo0008	acinus	HPV type 41	0.0754	234	521	15
PD42760a_lo0022	acinus	HHV-4 (EBV)	0.0236	1235	1287	40
PD42760a_lo0023	duct	HPV type 10	0.1244	801	1071	33
PD42765a_lo0002	acinus	HPV type 128	0.1279	496	1083	33
PD42765a_lo0007	acinus	HPV type 41	0.0713	301	395	12
PD42765a_lo0020	acinus	HPV type 41	0.0308	491	516	16
PD42767a_duct7	duct	HPV type 131	0.2019	439	648	20
PD42768a_lo0001	acinus	HPV type 4	0.0807	675	762	23
PD42768a_lo0012	duct	HPV type 128	0.2014	296	789	24
PD42768a_lo0051	duct	HPV type 41	0.0811	759	791	24
PD42770a_gland4	acinus	HPV type 92	0.1791	151	302	10
PD42771a_gland1	acinus	HPV type 131	0.1574	889	1552	46
PD42773a_lo0001	acinus	HPV type 10	0.0465	181	544	18
		HHV-4 (EBV)	0.0212	1023	1401	42
PD42773a_lo0003	acinus	HPV type 41	0.0977	699	1468	43
		HPV type 109	0.0619	241	321	10
PD42773a_lo0010	acinus	HPV type 41	0.0645	370	370	12
PD42774a_lo0042	duct	HPV type 109	0.1094	274	719	21
		HPV type 41	0.0693	471	783	24
PD42774a_lo0106	duct	HHV-4 (EBV)	0.0309	1222	1607	51
LYMPHOCYTES						
Sample	**Cell type**	**Taxa/Strain**	**Rel. abundance**	**Length**	**Bases mapped**	**Hit count**
PD42070d	CD8 T	HHV-6B	0.0149	536	536	15
PD42079c	CD8 T	HHV-6B	0.0242	48035	252586	7745
		HHV-6A	0.0094	428	871	29
PD42079c	CD8 T	HPV type 50	0.0053	624	713	22
PD42079d	B cells	HHV-6B	0.0227	46838	231347	7063
		HHV-6A	0.0138	300	901	30
PD42089c	CD8 T	HPV type 50	0.001	498	498	15
		HHV-6B	0.001	432	432	13
PD42761a_CD38_CD20	B cells	HHV-7	0.0963	6829	11835	361
PD42761a_CD4-2	CD4 T	HHV-7	0.2272	2083	3795	112
PD42771a_CD20-2a	B cells	HPV type 41	0.187	397	615	18
PD42771a_CD8-1b	CD8 T	HHV-7	0.122	2767	5837	177

the possibility that viral infection may play a role in the pathogenesis of PSS, a long-standing hypothesis which remains difficult to prove[76].

I.6 Copy number variation

To infer copy number states, we employed the Ascat and Battenberg algorithms for genome-wide analysis, and CNVkit for exome-wide analysis. Structural variation analysis was done with BRASS (Methods). The copy number profiles observed were fairly stable across both PSS and non-PSS samples with one particular exception: a duplication of chromosome 7 in several acinar and ductal samples from three PSS donors and none of the non-PSS donors. By both Ascat and Battenberg, the log copy ratios of chromosome 7 are higher while the B-allele frequencies show a divergence away from 0.5, indicating over-representation of one copy of chromosome 7 (**Figure 32 a-f, Table 15**). Samples with chromosome 7 duplication in patient PD45527 are phylogenetically related by single nucleotide variant analysis and are adjacent to each other in tissue sections (**Figure 33a,b**). In the non-PSS samples, the only large copy number change observed was a duplication of chromosome 13 in one sample. No significant or recurrent structural variants were detected in either cohort.

Chromosome 7 contains numerous critical genes for cell cycle regulation and immune system signalling, and duplication of chromosome 7 has been observed in various cancers, notably in salivary ductal carcinoma[222]. Prior studies of other normal tissue have not identified recurrent duplication of chromosome 7, in fact the copy number profiles of normal tissue tended to be quite stable in general[3–5,33]. In light of this, the observation of chromosome 7 duplication in 3 separate biopsies from PSS patients suggests that this event likely imparts a selective advantage to salivary epithelial cells in the context of disease.

Table 21. Summary of chromosome 7 duplication findings.

Donor	Diagnosis	Samples with chr7 gain	Sample type
PD42770	PSS	1	acinus
PD42773	PSS	1	duct
PD45527	PSS	3	acinus

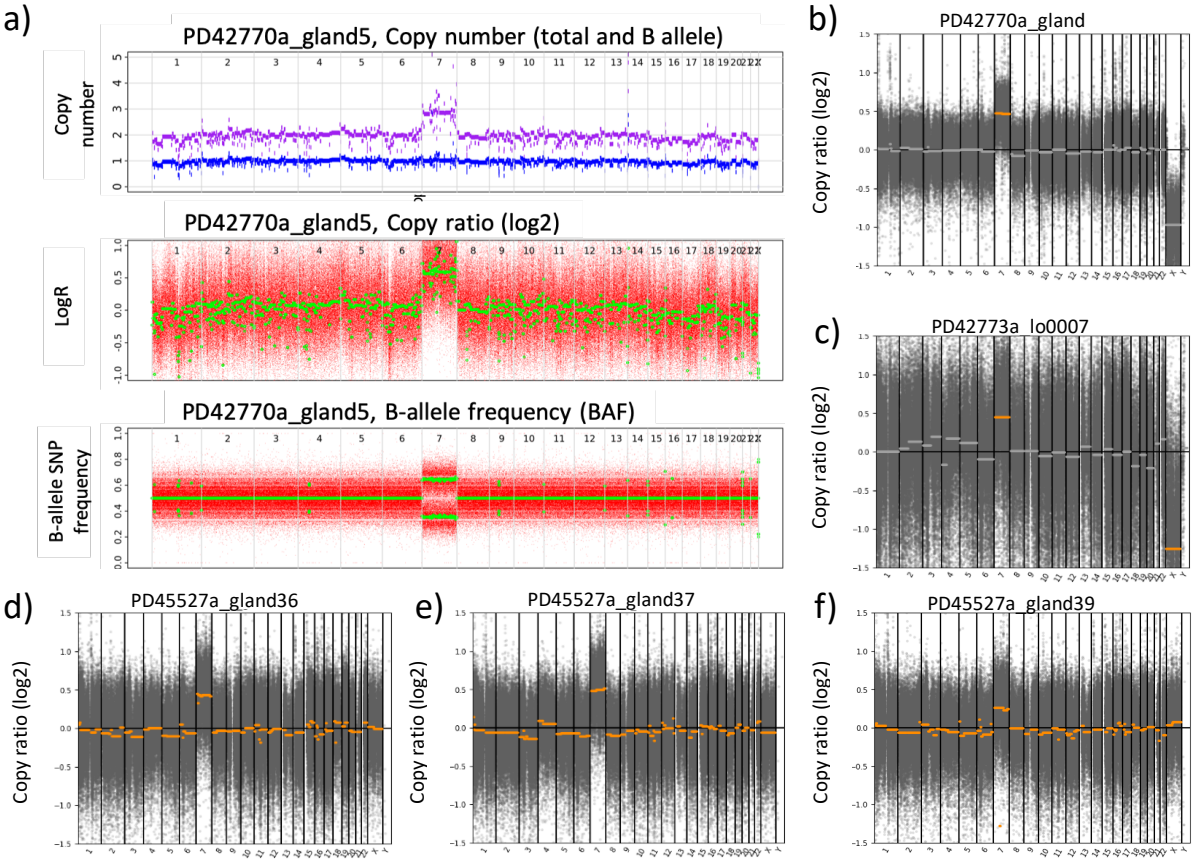


Figure 35. Copy number alterations of chromosome 7. **a)** Copy number, copy log ratio, and B-allele frequency genome-wide profiles for samples PD42770a_gland5. **b-f)** Exome-wide copy log ratio profiles of five samples with duplication of chr 7.

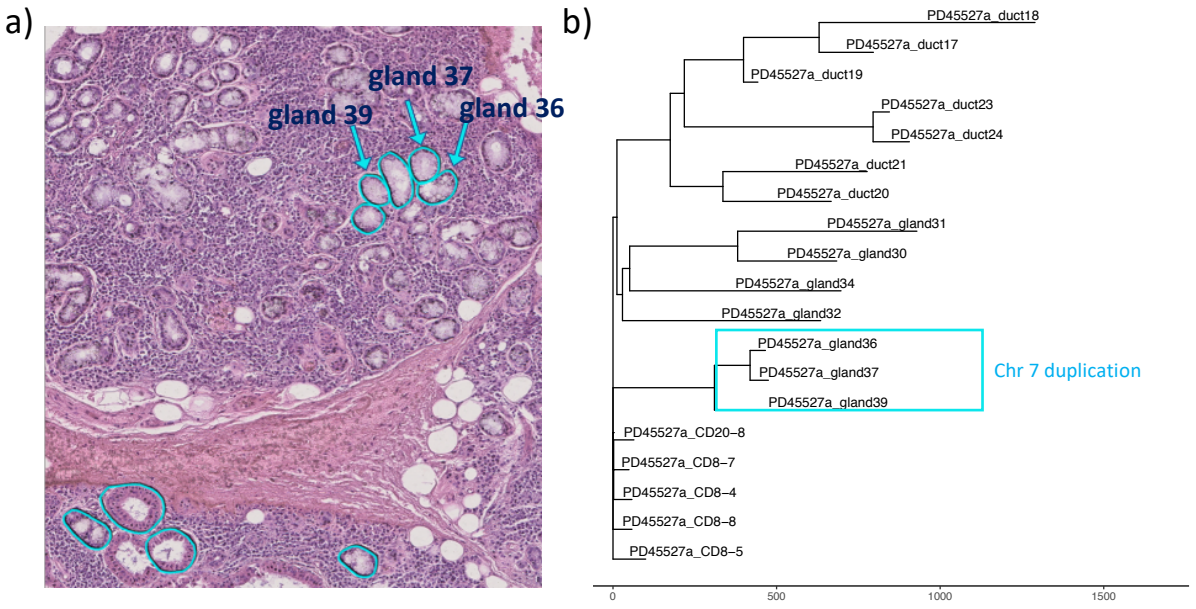


Figure 36. Clustering of samples with chr7 duplication in patient PD45527. **a)** H&E section of PD45527 biopsy showing location of three samples with chr7 duplication. **b)** Phylogenetic relationship of PD45527 by SNP sharing shows recent common origin of three indicated samples.

DISCUSSION

II.1 Findings related to normal minor salivary gland epithelium

Previous studies have profiled the exomes and genomes of salivary gland tumours, which were informative of the oncogenic events and mutational signatures operative in the malignantly transformed tissue[211,213,223,224]. However, they gave little insight into the underlying pre-malignant and normal tissue mutational dynamics. Our study of minor salivary glands has yielded the first glimpse into the genomic landscape of normal salivary epithelium. This investigation has allowed us to determine if any of the processes observed in cancer are also present in the normal tissue. Furthermore, it has contextualized the salivary gland genome within the broader scope of normal tissue genomics, demonstrating comparative trends in clonal structure, mutational signatures, and mutational burden between the minor salivary glands and other normal tissues.

The major goal of this study was to profile the somatic mutational landscape of salivary glands in primary Sjögren's syndrome. Our cohort included biopsies of minor salivary gland tissue from patients with complaints of oral dryness, some of whom were subsequently diagnosed with PSS and others for whom PSS diagnosis was excluded upon biopsy. Therefore, we had biopsies from a PSS cohort histologically characterized by lymphocytic infiltration and acinar destruction, and a non-PSS cohort with histologically normal salivary glandular architecture. The non-PSS biopsies were used as a control to detect changes occurring specifically in the PSS cohort, but they were also used to discover normal genomic features of the salivary gland. I therefore considered features shared between PSS and non-PSS biopsies to be inherent to minor salivary gland tissue.

One such feature of normal salivary gland epithelium is the clonal composition. By profiling discrete histological structures using laser capture microdissection, I discovered that a single acinus is a clonal unit, while a larger feature such as a duct is often comprised of the progeny of multiple stem cells. Physically proximal structures often shared somatic variants, indicating a shared lineage from a localized common progenitor. These findings agree with previous sequencing studies of glandular tissues across the body, contributing to the understanding of

glandular development and turnover and enabling the design of studies that maximise sample clonality.

The occurrence of canonical cancer driver mutations across several normal tissues has been an unexpected and striking finding from previous normal tissue sequencing endeavours[2–4]. The accumulation of cancer drivers is intrinsically linked to tissue-specific characteristics such as cell turnover, mutagen exposure, and selective pressures. Driver discovery analysis of non-PSS minor salivary glands did not reveal any overt oncogenic events, though some variants of uncertain significance were observed in cancer-associated genes. Though there were 39 individual non-PSS samples sequenced, they derived from only three donors, a cohort size too small to detect genes under positive selection using a dN/dS approach. Therefore, future plans include sequencing additional acinar and ductal samples from a larger cohort of non-PSS biopsies. This will illustrate a baseline burden of driver mutations in normal minor salivary glands, which could then be used to evaluate changes in the context of disease such as PSS or cancer.

Mutational signature analysis is a potent tool for discovery of underlying mutagen exposures that shaped the somatic mutational landscape of a tissue. Previous normal tissue sequencing studies have identified tissue-specific mutagen exposures, such as ultraviolet light on skin[2] or activation-induced cytidine deaminase in B cells[44]. Through our analysis, we discovered that pervasive APOBEC-mediated mutagenesis is a prominent feature of normal salivary epithelium, due to its presence in ~20% of both non-PSS and PSS samples. A recent study of salivary ductal carcinoma also identified the APOBEC-associated signatures SBS2 and SBS13 in four out of ten whole-genome sequenced samples and demonstrated that the overall mutational burden was higher in those samples[211]. Taken together, these findings suggest that APOBEC is an important mutagen in the salivary gland that is frequently activated in normal tissue before the onset of malignancy and likely plays a key role in its progression.

APOBEC-associated signatures have been found in numerous cancers, most notably the following: breast, lung, bladder, cervix, and head and neck squamous cell carcinoma[10,214,218]. The APOBEC family of enzymes has 11 members, which include AID (activation-induced cytidine deaminase, involved in lymphocyte receptor diversification), APOBEC1 (involved in

mRNA editing), and the APOBEC3 subfamily thought to be involved in viral and retrotransposon inhibition[219]. APOBEC3 signatures are found in the aforementioned cancers, some of which have well-known involvement of oncogenic viruses. For example, cervical cancers show high rates of APOBEC-driven mutagenesis, which contributes to accumulation of driver mutations in addition to the proliferative effect imposed by HPV infection through expression of viral cell cycle genes E6 and E7[219]. Therefore, in cervical cancer the association of viral infection and APOBEC3 activation is evident. Conversely, in breast cancer, it is more difficult to understand the trigger of APOBEC3 activation when viral infections are not a common occurrence. The mobilization of retroelements in the genome is thought to be a possible trigger for APOBEC, as is replicative stress in the context of cancer; these are ongoing areas of research. In the normal tissues sequenced thus far, there has been little evidence of APOBEC activity in lung[30], oesophagus[4], colon[5], blood[25], and endometrium[3]; however, sequencing of normal bladder tissue demonstrated a significant contribution of these signatures along with patterns of mutational clustering (kataegis)[215]. It is not understood why bladder urothelial cells frequently activate APOBEC, though viral infection is a possibility.

In the context of head and neck cancers, viral infection remains a plausible trigger for APOBEC3 activation. It has long been suspected that salivary gland tumours, of which there are several varieties including adenoid cystic carcinoma and salivary ductal carcinoma, might be caused by oncogenic viruses. However, there has been no replicable evidence of this. Some studies reported a high prevalence of oncogenic viruses such as human beta papillomaviruses and herpesviruses in salivary gland tumours[225], while other studies disagreed, finding very little human papillomavirus material[226]. These studies looked for evidence of directly oncogenic viruses, which induce epithelial cell proliferation and incite tumorigenesis by hijacking cellular replication machinery. What has not been addressed in the context of salivary gland cancers, however, is the potential of viral infections to promote tumorigenesis indirectly, through APOBEC activation.

In addition to finding prevalent APOBEC-associated mutagenesis in our minor salivary gland samples, we also found that benign strains of human papillomavirus as well as Epstein-Barr virus are frequently present in both non-PSS and PSS biopsies. Importantly, APOBEC signatures and viral elements were not found in the same samples, with one sample out of

20 being an exception. It is possible that in the minor salivary glands APOBEC activation is not primarily related to viral infection but to other endogenous causes, similarly to the case of normal bladder urothelium. However, given the findings of viral presence in >20% samples, it is more likely that APOBEC activity is significantly related to viral infection. The lack of co-occurrence of APOBEC activity and viral elements in the same samples can be explained by the fact that successful activation of APOBEC enzymes results in hypermutation of viral sequences, prevention of integration, and clearance of the virus from a cell. This would explain why previous studies have struggled to find evidence associating viral infection and salivary gland tumours. Our findings suggest that viral infections of the salivary glands do increase the risk of tumorigenesis, not through directly oncogenic viral mechanisms as previously hypothesized, but indirectly through APOBEC-mediated mutagenesis which increases the mutational burden of infected cells and thus also the risk of acquiring cancer driver mutations.

II.2 Findings related to minor salivary gland epithelium in PSS

To identify features of minor salivary gland epithelium that might be specific to PSS, we used non-PSS epithelium as a control. As described above, we concluded that APOBEC-mediated mutagenesis and viral infections are not disease-specific but rather ubiquitous, although this does not mean that these phenomena are unrelated to PSS pathogenesis. It has been hypothesized that viral infections may be a trigger for autoimmune activity in PSS, though as with salivary gland carcinomas, the evidence has been elusive. A mouse model study demonstrated that development of Sjögren's syndrome-like sialadenitis in lupus-prone mice occurs three months after infection with cytomegalovirus, but by the time sialadenitis developed, the virus was undetectable in salivary glands[117]. Therefore, it is possible that long-term autoimmune activity can be triggered by transient viral infections that are later successfully cleared from infected cells, likely with the help of APOBEC3 enzyme activity, though this remains difficult to prove.

Despite the small cohort of biopsies analysed, we observed some genomic trends that are possibly specific to PSS. Protein-altering mutations (three missense and one nonsense) were found in the *IGFN1* gene in four out of seven PSS biopsies. Though not statistically significant

by dN/dS analysis, it is a gene we will keep under close consideration in the analysis of future samples. Without a larger cohort of donors, it is not possible to determine with certainty the presence or absence of genes under positive selection.

We identified chromosome 7 duplication in samples from three out of seven PSS biopsies. While there are numerous cell cycle and immune-related genes on chromosome 7 that could be functionally relevant, the observation is particularly interesting because chromosome 7 duplication is frequent in salivary ductal carcinoma and is associated with worse outcomes[222]. Additionally, the *EGFR* (epidermal growth factor receptor) driver gene found on chromosome 7 is also frequently mutated and overexpressed in salivary gland cancers (ductal carcinoma[222], mucoepidermoid carcinoma[212], adenoid cystic carcinoma[227]). These studies suggest that copy number gain of chromosome 7 likely provides a selective advantage to cells that harbour it. In the context of PSS, this may have emerged as an adaptation to the cytotoxic inflammatory environment, favouring clones with copy gain of chromosome 7 for survival and regeneration of damaged glands.

II.3 Limitations and future directions

A major technological limitation of this study was its reliance on microdissection of clonal features from minor salivary gland biopsies that have enough DNA to create sequencing libraries. This required Z-stacking microdissections of the same feature, typically single acini which are 10-15 cells in perimeter, from up to six adjacent tissue sections. For this reason, most dissected samples did not meet the minimum DNA requirement and were not sequenced. However, the successfully sequenced samples demonstrated the validity of dissecting small discrete structures, since the individual acini proved to be largely clonal.

The most important next step for furthering the exploratory findings of this study is obtaining additional biopsies and dissecting more samples through the laser-capture microdissection approach, as described above. A larger, confirmatory cohort may reinforce the significance of the preliminary findings described, particularly regarding the presence of genes under positive selection in PSS. Secondly, the observation of chromosome 7 duplication would benefit from validation by RNA sequencing to determine the transcriptional effect imparted

by the aneuploidy. This could be done by microdissection of the same features from remaining sections of tissue, provided enough cells remain for the preparation of low-input RNA sequencing libraries. The transcriptomic profiles could validate increased expression of chromosome 7 genes and show whether the duplication has an effect on other genes and pathways transcriptome-wide that could promote growth and selection. Thirdly, the finding of viral DNA could be further investigated by searching for the sites of integration of the viral genome into the host genome of salivary epithelial cells. This would provide more definitive proof of viral presence in a given sample. Finally, obtaining samples from major salivary glands would be a useful comparison to determine whether APOBEC activation and viral presence vary across different salivary gland sites.

Chapter 6: Conclusion

In this thesis, I have explored the somatic mutational landscape of lymphocytes and epithelial cells in the minor salivary glands of primary Sjögren's syndrome (PSS) patients and controls, as well as the transcriptomic profiles of tissue-infiltrating lymphocytes. The approaches used here represent novel ways of interrogating the genetic basis and molecular pathophysiology of this disease. The findings highlight enrichment of disease-relevant cell types and somatic alterations in these cell types, which fit into the context of current knowledge and provide novel directions for future research into functional effects and possible therapeutics.

I. B cells

The primary hypothesis of this project was that somatic mutations may exist in autoreactive lymphocytes, driving their chronic activation and causing the autoimmune pathogenesis of PSS. To isolate autoreactive lymphocytes is challenging because we don't quite know what defines them. Therefore, we chose to investigate tissue-infiltrating lymphocytes from minor salivary glands, a common site of disease-mediated damage in PSS, hoping that this tissue is a reservoir enriched in auto-reactive cells. Given current knowledge of PSS pathology, B lymphocytes are considered key players by multiple lines of evidence: activated B cells and plasma cells are found to infiltrate minor salivary glands in PSS, they produce disease-associated auto-antibodies, they can organize into germinal centre-like structures in the salivary glands, and they are associated with transformation to B cell lymphoma[76,94,100,168,201]. For these reasons, our hypothesis was that somatic mutations were most likely to be found in the B cell and plasma cell compartments of lymphocytes, especially given their physiological ability to induce somatic hypermutation for repertoire diversification, which results in an excess number of somatic mutations.

Given this notion of PSS as a "B cell disease", we were surprised to discover a relative paucity of activating or proliferation-inducing mutations in B cells. With the exception of one patient, who had a demonstrable clonal expansion of CD20 B cells with multiple activating mutations, no recurrent trend of driver mutations was found in B cells from minor salivary glands across PSS patients. This may suggest then, that instead of being an early event in disease

development, pathogenic somatic mutations in B cells may arise in later stages of disease and be associated with worse outcomes, as the patient who harboured the mutated clone also had peripheral neuropathy and a monoclonal immunoglobulin in blood, the latter of which is associated with higher risk of lymphoma. To confirm this, further studies would be needed examining samples from patients at later stages of disease, in addition to the diagnostic biopsies used here, which represent a cohort at early stages of disease.

Nevertheless, the significance of B cells and plasma cells in PSS is highlighted by other findings in this project. Single cell RNA sequencing identified a population of $CD20^{+}CD27^{-}CD38^{-}$ B cells that were almost entirely specific to PSS minor salivary glands, compared to non-PSS sicca control biopsies. This finding corresponds to the what was observed by immunohistochemistry, where PSS patients had discrete localized clusters of $CD20^{+}$ cells which were not present in non-PSS controls. A proportion of these cells had an $IgM^{+}IgD^{+}$ unswitched phenotype by single cell RNA analysis, making them distinctly different from the differentiated plasma cells and plasmablasts which make up the majority of B lineage cells, in both PSS and non-PSS tissue. While relative numbers of plasma cells in the PSS and control biopsies were not significantly different, a subset of plasma cells was highly specific to the patient cohort. These were plasma cells which expressed IgG, as opposed to the remaining majority of plasma cells which expressed IgA transcripts. Since serum levels of immunoglobulin are commonly elevated in PSS, the finding of tissue-infiltrating IgG^{+} cells is significant and warrants further investigation of their function and antigen specificity. The IgG^{+} plasma cell population and the $CD20^{+}$ B cell populations identified by this analysis are important targets for further functional studies and underscore the crucial role of B cells in PSS.

II. T cells

As mentioned, PSS is considered to be a disease heavily mediated by aberrant B cell activity, and correspondingly by T helper cells as well[169,228], which are required for antigen-dependent B cell activation. However, it was not in B cells nor in CD4 helper T cells that we observed the most striking somatic alterations, but in the often-overlooked subset of cytotoxic T cells. By

bulk DNA sequencing, we identified subclonal monosomy X in CD8 T cell samples from up to 75% of female patient biopsies. In two of those samples, we also identified three different truncating mutations in the X-linked tumour suppressor gene *KDM6A*, which encodes the protein UTX (Ubiquitously transcribed X chromosome tetratricopeptide repeat protein). A similar but less pronounced phenomenon was seen in CD4 T cells, where 33% of samples had evidence of subclonal monosomy X and one sample had a truncating *KDM6A* mutation. In comparison, none of the bulk sorted B cell, plasma cell, or plasmablast samples had evidence of either monosomy X or *KDM6A* mutations. In total, we observed four truncating *KDM6A* mutations in T cells in this cohort of 30 patients, a finding unlikely by chance, as assessed by dN/dS selection analysis (q-value = 0.00068).

The *KDM6A* gene encode UTX, which is a tumour suppressor and histone demethylase that escapes X-inactivation, and it is commonly mutated in various cancers, including leukaemia and lymphoma[172,173]. If a clone has lost both copies of this gene, as would be the case if the mutations and monosomy X co-existed in the same clone, then these T cells would lose a potent tumour suppressor and would be lacking its histone demethylating activity, which modulates the expression of many downstream genes. Whether this leads to constitutive activation of T cells and promotes autoimmunity in PSS requires further studies. A caveat of this finding is that monosomy X (but not *KDM6A* mutation) was also found in non-PSS sicca control patients, making it difficult to define the specificity of this phenomenon to PSS. As discussed in previous chapters, these non-PSS patients have similar sicca symptoms as PSS patients, but do not fulfil the necessary diagnostic criteria, allowing for the possibility that they might have an early stage or 'forme fruste' of PSS.

The fact that both of the somatic alterations observed in T cells are linked to the X chromosome is of particular interest given that PSS is a disease with a 90% female predominance[78]. If losing one X chromosome is the first event in a two-hit model of T cell activation, it is curious why loss of the Y chromosome in males does not happen with similar frequency and have a similar outcome. Potential differences in sex chromosome segregation during mitosis may be at play, as well as complex gene dosage effects of X chromosome genes and their Y chromosome paralogues.

The results of single cell RNA sequencing also underscore the role of cytotoxic T cells in PSS. Both CD4 and CD8 T cells were found to express significantly higher levels of activation markers in PSS minor salivary glands than in non-PSS controls. While CD4 T cells often displayed a naïve phenotype, CD8 cells had an effector-memory phenotype and expressed significantly higher levels of cytotoxicity-related genes in PSS patients than controls. This finding is complementary to other studies in the past few years which have implicated CD8 T cell activation in PSS. For example, a cytometry-by-time-of-flight study identified activated HLA-DR+ CD8 cells in minor salivary glands of PSS patients and found them to be associated with worse outcomes[112]. Additionally, a multi-omic study of immune cells in the blood of PSS patients identified a signature of CD8 T cell activation[229]. Supporting evidence also came from a mouse model of Sjögren's syndrome, where depletion of CD8 T cells abrogated the development of Sjögren's syndrome-like sialadenitis and other pathologic manifestations of the disease[205]. The findings of this thesis project and other recent studies underscore the previously underappreciated importance of cytotoxic T cells in the pathogenesis of PSS and point towards an immediate need for further research examining their functional role in this disease and their potential as a therapeutic target.

III. Epithelial cells

Recently, a potential role of salivary epithelial cells in the pathogenesis of PSS has emerged through evidence of epithelial cell activation by expression of immune-stimulating factors[90,112,113]. To examine whether somatic mutations might underlie the reprogramming of epithelial cells to become chronically immunostimulatory, we sequenced DNA from minor salivary gland acini and ducts from PSS patients and non-PSS controls to perform a somatic mutation analysis. The data showed a copy number alteration, chromosome 7 gain, specific to PSS samples which may be related to this activation[222]. Currently ongoing sequencing of additional samples and future plans for RNA sequencing of epithelial cell structures are in place to further this finding.

We identified several features of salivary epithelium that were shared between PSS patients and controls, indicating inherent features of the tissue that have not been previously

described. These include identification of single acini as clonal units, demonstrating that the genome-wide burden of somatic mutations in salivary epithelial cells increases linearly with age, prevalence of APOBEC-associated mutational signatures (in ~20% of samples), and evidence of retroviral DNA integration (in ~20% samples). APOBEC activation is associated with cellular defence against viral integration[216,217], complementing the observation of viral DNA. The finding of viral elements and APOBEC mutagenesis was not specific to PSS samples, instead it highlights the general susceptibility of salivary epithelial cells to viral infection and the prevalence of APOBEC activation as a defence mechanism.

The role of viruses in PSS initiation has long been suspected but never proven, which may be because the virus is undetectable in tissue by the time autoimmune disease has developed[76,117]. Here we've found evidence of past viral activity (directly by detecting viral DNA and indirectly through APOBEC activity) in a significant fraction of normal and PSS salivary gland samples, which then also implies past activation of cytotoxic T cells as an immune defence against viral infection. Whether this CD8 T cell activation by viruses is an inciting step in PSS pathogenesis remains to be proven, however the activated phenotype and somatic alterations in CD8 T cells described in this thesis lend support to this idea. It is plausible that the CD8 T cells activated by viral infection then acquire somatic mutations and copy number changes during antigen-driven proliferation and expansion, and that these changes impart a constitutively activated phenotype that perpetuates inflammation and autoimmune disease. Further investigation of CD8 T cells and viral infection of salivary glands is needed to evaluate this novel hypothesis.

Original publications

Smita Jha*, **Aleksandra Ivovic*** et al. Distribution and functional consequences of somatic MAP2K1 mutations in affected skin associated with bone lesions in melorheostosis. Journal of Investigative Dermatology. 2020. (*co-first authors)

Heeseog Kang*, Smita Jha*, **Aleksandra Ivovic*** et al. SMAD3 Somatic Activating Mutations Cause Melorheostosis with an Endosteal Radiographic Pattern by Upregulating the TGF-β/SMAD Pathway. J Exp Med. 2020. (*co-first authors)

Laura T. Donlin, Sung-Ho Park, Eugenia Giannopoulou, **Aleksandra Ivovic**, et al. Insights into rheumatic diseases from next-generation sequencing. Nature Reviews Rheumatology. 2019.

Fratzl-Zelman N, Roschger P, Kang H, Jha S, Roschger A, Blouin S, Deng Z, Cabral WA, **Ivovic A**, Katz J, Siegel RM, Klaushofer K, Fratzl P, Bhattacharyya T, Marini JC. Melorheostotic Bone Lesions Caused by Somatic Mutations in MAP2K1 Have Deteriorated Microarchitecture and Periosteal Reaction. J Bone Miner Res. 2019 Jan 22.

Heeseog Kang, Smita Jha, Zuoming Deng, Nadja Fratzl-Zelman, Wayne A Cabral, **Aleksandra Ivovic**, Françoise Meylan, et al. 2018. Somatic Activating Mutations in MAP2K1 Cause Melorheostosis. Nature Communications. 2018.

Smita Jha, M.D., Nadja Fratzl-Zelman, Ph.D., Paul Roschger, Ph.D., Georgios Z. Papadakis, M.D., Ph.D., Edward W. Cowen, M.D., M.H. Sc., Heeseog Kang, Ph.D., Tanya J. Lehky, M.D., Katharine Alter, M.D., Zuoming Deng, Ph.D., **Aleksandra Ivovic**, et al. Distinct clinical and pathological features of melorheostosis associated with somatic MAP2K1 mutations. Journal of Bone and Mineral Research. 2018.

References

1. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science (80-.).* **349**, 1483 LP-1489 (2015).
2. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
3. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
4. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
5. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
6. Rose, N. R. Prediction and Prevention of Autoimmune Disease in the 21st Century: A Review and Preview. *Am. J. Epidemiol.* **183**, 403–406 (2016).
7. BURNET, M. SOMATIC MUTATION AND CHRONIC DISEASE. *Br. Med. J.* **1**, 338–342 (1965).
8. Goodnow, C. C. Multistep pathogenesis of autoimmune disease. *Cell* **130**, 25–35 (2007).
9. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
10. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
11. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
12. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821–836.e16 (2019).
13. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N Engl J Med* **371**, 2477–87 (2014).
14. Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
15. McKerrell, T. *et al.* Leukemia-Associated Somatic Mutations Drive Distinct Patterns of Age-Related Clonal Hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
16. Jaiswal, S. *et al.* Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N. Engl. J. Med.* **377**, 111–121 (2017).
17. Fuster, J. J. *et al.* Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. *Science* **355**, 842–847 (2017).
18. Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).

19. Busque, L. *et al.* Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* **88**, 59–65 (1996).

20. Yoshizato, T. *et al.* Somatic Mutations and Clonal Hematopoiesis in Aplastic Anemia. *N. Engl. J. Med.* **373**, 35–47 (2015).

21. Atkin, C., Richter, A. & Sapey, E. What is the significance of monoclonal gammopathy of undetermined significance? *Clin. Med.* **18**, 391–396 (2018).

22. Mikulasova, A. *et al.* The spectrum of somatic mutations in monoclonal gammopathy of undetermined significance indicates a less complex genomic landscape than that in multiple myeloma. *Haematologica* **102**, 1617–1625 (2017).

23. Strati, P. & Shanafelt, T. D. Monoclonal B-cell lymphocytosis and early-stage chronic lymphocytic leukemia: diagnosis, natural history, and risk stratification. *Blood* **126**, 454–462 (2015).

24. van de Donk, N. W. C. J. *et al.* The clinical relevance and management of monoclonal gammopathy of undetermined significance and related disorders: recommendations from the European Myeloma Network. *Haematologica* **99**, 984–996 (2014).

25. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

26. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).

27. Fernandez-Antoran, D. *et al.* Outcompeting p53-Mutant Cells in the Normal Esophagus by Redox Manipulation. *Cell Stem Cell* **25**, 329–341.e6 (2019).

28. Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777–1789 (2018).

29. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).

30. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

31. Lodato, M. A. *et al.* Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science (80-.).* **359**, 555 LP-559 (2018).

32. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science (80-.).* **364**, eaaw0726 (2019).

33. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).

34. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science (80-.).* **253**, 49 LP-53 (1991).

35. Olivier, M., Hussain, S. P., Caron de Fromentel, C., Hainaut, P. & Harris, C. C. TP53 mutation spectra and load: a tool for generating hypotheses on the etiology of cancer. *IARC Sci. Publ.* 247–270 (2004).

36. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

37. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

38. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

39. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).

40. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* **Chapter 10**, Unit-10.11 (2008).

41. David, S. S., O'Shea, V. L. & Kundu, S. Base-excision repair of oxidative DNA damage. *Nature* **447**, 941–950 (2007).

42. Ng, A. W. T. *et al.* Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci. Transl. Med.* **9**, (2017).

43. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

44. Zhang, L. *et al.* Single-cell whole-genome sequencing reveals the functional landscape of somatic mutations in B lymphocytes across the human lifespan. *Proc. Natl. Acad. Sci.* **116**, 9014 LP-9019 (2019).

45. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).

46. Kanwal, F. *et al.* Risk of Hepatocellular Cancer in Patients With Non-Alcoholic Fatty Liver Disease. *Gastroenterology* **155**, 1828–1837.e2 (2018).

47. Dias, C. & Isenberg, D. A. Susceptibility of patients with rheumatic diseases to B-cell non-Hodgkin lymphoma. *Nat. Rev. Rheumatol.* **7**, 360–368 (2011).

48. Kang, H. *et al.* Somatic activating mutations in MAP2K1 cause melorheostosis. *Nat. Commun.* **9**, 1390 (2018).

49. Kang, H. *et al.* Somatic SMAD3-activating mutations cause melorheostosis by up-regulating the TGF-β/SMAD pathway. *J. Exp. Med.* **217**, (2020).

50. Lindhurst, M. J. *et al.* A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N. Engl. J. Med.* **365**, 611–619 (2011).

51. Dowdell, K. C. *et al.* Somatic FAS mutations are common in patients with genetically undefined autoimmune lymphoproliferative syndrome. *Blood* **115**, 5164–5169 (2010).

52. Tanaka, N. *et al.* High Incidence of NLRP3 Somatic Mosaicism in Patients With Chronic Infantile Neurologic, Cutaneous, Articular Syndrome: Results of an International Multicenter Collaborative Study. *Arthritis Rheum.* **63**, 3625–3632 (2011).

53. Kawasaki, Y. *et al.* Identification of a High-Frequency Somatic NLRC4 Mutation as a Cause of Autoinflammation by Pluripotent Cell–Based Phenotype Dissection. *Arthritis Rheumatol.* **69**, 447–459 (2016).

54. Zhu, M. *et al.* Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. *Cell* **177**, 608–621.e12 (2019).

55. Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of

NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).

56. Kakiuchi, N. *et al.* Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**, 260–265 (2020).

57. Nanki, K. *et al.* Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *Nature* **577**, 254–259 (2020).

58. Olafsson, S. *et al.* Somatic evolution in non-neoplastic IBD-affected colon. *bioRxiv* 832014 (2020) doi:10.1101/832014.

59. McGovern, D. P. B., Kugathasan, S. & Cho, J. H. Genetics of Inflammatory Bowel Diseases. *Gastroenterology* **149**, 1163–1176.e2 (2015).

60. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

61. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

62. Rossi, D. J., Jamieson, C. H. M. & Weissman, I. L. Stems Cells and the Pathways to Aging and Cancer. *Cell* **132**, 681–696 (2008).

63. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science (80-.).* **347**, 78 LP-81 (2015).

64. Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909–918.e8 (2018).

65. Wang, Y. & Navin, N. E. Advances and applications of single-cell sequencing technologies. *Mol. Cell* **58**, 598–609 (2015).

66. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).

67. Schmitt, M. W. *et al.* Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods* **12**, 423–425 (2015).

68. Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. doi:10.1073/pnas.1607794113.

69. Fairweather, D., Frisancho-Kiss, S. & Rose, N. R. Sex differences in autoimmune disease from a pathological perspective. *Am. J. Pathol.* **173**, 600–609 (2008).

70. Hu, X. & Daly, M. What have we learned from six years of GWAS in autoimmune diseases, and what is next? *Curr. Opin. Immunol.* **24**, 571–575 (2012).

71. Ye, J., Gillespie, K. M. & Rodriguez, S. Unravelling the Roles of Susceptibility Loci for Autoimmune Diseases in the Post-GWAS Era. *Genes (Basel).* **9**, 377 (2018).

72. Alperin, J. M., Ortiz-Fernández, L. & Sawalha, A. H. Monogenic Lupus: A Developing Paradigm of Disease. *Front. Immunol.* **9**, 2496 (2018).

73. Wang, J. *et al.* Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proc. Natl. Acad. Sci.* **113**, E2029 LP-E2038 (2016).

74. Nocturne, G. & Mariette, X. Advances in understanding the pathogenesis of primary Sjögren's syndrome. *Nat. Rev. Rheumatol.* **9**, 544–556 (2013).

75. Nordmark, G., Alm, G. V & Rönnblom, L. Mechanisms of Disease : primary Sjögren ' s syndrome and the type I interferon system. **2**, 262–269 (2006).

76. Pontarini, E., Lucchesi, D. & Bombardieri, M. Current views on the pathogenesis of Sjögren's syndrome. *Curr. Opin. Rheumatol.* **30**, 215–221 (2018).

77. Saraux, A., Pers, J.-O. & Devauchelle-Pensec, V. Treatment of primary Sjögren syndrome. *Nat. Rev. Rheumatol.* **12**, 456–471 (2016).

78. Mariette, X. & Criswell, L. A. Primary Sjögren's Syndrome. *N. Engl. J. Med.* **378**, 931–939 (2018).

79. Stefanski, A.-L. *et al.* The Diagnosis and Treatment of Sjögren's Syndrome. *Dtsch. Arztebl. Int.* **114**, 354–361 (2017).

80. Nordmark, G., Alm, G. V. & Rönnblom, L. Mechanisms of disease: Primary Sjögren's syndrome and the type I interferon system. *Nat. Clin. Pract. Rheumatol.* **2**, 262–269 (2006).

81. Theander, E. *et al.* Lymphoid organisation in labial salivary gland biopsies is a possible predictor for the development of malignant lymphoma in primary Sjögren's syndrome. *Ann. Rheum. Dis.* **70**, 1363–1368 (2011).

82. Shiboski, C. H. *et al.* 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: A consensus and data-driven methodology involving three international patient cohorts. *Ann. Rheum. Dis.* **76**, 9–16 (2017).

83. Lessard, C. J. *et al.* Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome. *Nat. Genet.* **45**, 1284–1292 (2013).

84. Nordmark, G. *et al.* Additive effects of the major risk alleles of IRF5 and STAT4 in primary Sjögren's syndrome. *Genes Immun.* **10**, 68–76 (2009).

85. Miceli-Richard, C. *et al.* Association of an IRF5 gene functional polymorphism with Sjögren's syndrome. *Arthritis Rheum.* **56**, 3989–3994 (2007).

86. Lu, Q. The critical importance of epigenetics in autoimmunity. *J. Autoimmun.* **41**, 1–5 (2013).

87. Alevizos, I., Alexander, S., Turner, R. J. & Illei, G. G. MicroRNA expression profiles as biomarkers of minor salivary gland inflammation and dysfunction in Sjögren's syndrome. *Arthritis Rheum.* **63**, 535–544 (2011).

88. Youinou, P., Devauchelle-Pensec, V. & Pers, J.-O. Significance of B cells and B cell clonality in Sjögren's syndrome. *Arthritis Rheum.* **62**, 2605–2610 (2010).

89. Nocturne, G. & Mariette, X. Sjögren Syndrome-associated lymphomas: An update on pathogenesis and management. *British Journal of Haematology* vol. 168 317–327 (2015).

90. Amft, N. *et al.* Ectopic expression of the B cell-attracting chemokine BCA-1 (CXCL13) on endothelial cells and within lymphoid follicles contributes to the establishment of

germinal center-like structures in Sjögren's syndrome. *Arthritis Rheum.* **44**, 2633–2641 (2001).

91. Lindop, R. *et al.* Molecular signature of a public clonotypic autoantibody in primary Sjögren's syndrome: a 'forbidden' clone in systemic autoimmunity. *Arthritis Rheum.* **63**, 3477–3486 (2011).

92. Lindop, R. *et al.* Long-term Ro60 humoral autoimmunity in primary Sjögren's syndrome is maintained by rapid clonal turnover. *Clin. Immunol.* **148**, 27–34 (2013).

93. Waaler, E. ON THE OCCURRENCE OF A FACTOR IN HUMAN SERUM ACTIVATING THE SPECIFIC AGGLUTINATION OF SHEEP BLOOD CORPUSCLES. *Acta Pathol. Microbiol. Scand.* **17**, 172–188 (1940).

94. Nocturne, G. & Mariette, X. B cells in the pathogenesis of primary Sjögren syndrome. *Nat. Rev. Rheumatol.* **14**, 133–145 (2018).

95. Meltzer, M., Franklin, E. C., Elias, K., McCluskey, R. T. & Cooper, N. Cryoglobulinemia-A clinical and laboratory study. II. Cryoglobulins with rheumatoid factor activity. *Am. J. Med.* **40**, 837–856 (1966).

96. Kunkel, H. G., Agnello, V., Joslin, F. G., Winchester, R. J. & Capra, J. D. Cross-idiotypic specificity among monoclonal IGM proteins with anti-γ-globulin activity. *J. Exp. Med.* **137**, 331–342 (1973).

97. Tzioufas, A. G., Boumba, D. S., Skopouli, F. N. & Moutsopoulos, H. M. Mixed monoclonal cryoglobulinemia and monoclonal rheumatoid factor cross-reactive idiotypes as predictive factors for the development of lymphoma in primary Sjögren's syndrome. *Arthritis Rheum.* **39**, 767–772 (1996).

98. Nocturne, G. *et al.* Rheumatoid Factor and Disease Activity Are Independent Predictors of Lymphoma in Primary Sjögren's Syndrome. *Arthritis Rheumatol. (Hoboken, N.J.)* **68**, 977–985 (2016).

99. Bende, R. J. *et al.* Among B cell non-Hodgkin's lymphomas, MALT lymphomas express a unique antibody repertoire with frequent rheumatoid factor reactivity. *J. Exp. Med.* **201**, 1229–1241 (2005).

100. Risselada, A. P., Looije, M. F., Kruize, A. A., Bijlsma, J. W. J. & van Roon, J. A. G. The role of ectopic germinal centers in the immunopathology of primary Sjögren's syndrome: a systematic review. *Semin. Arthritis Rheum.* **42**, 368–376 (2013).

101. Daridon, C. *et al.* Aberrant expression of BAFF by B lymphocytes infiltrating the salivary glands of patients with primary Sjögren's syndrome. *Arthritis Rheum.* **56**, 1134–1144 (2007).

102. Mariette, X. *et al.* The level of BLyS (BAFF) correlates with the titre of autoantibodies in human Sjögren's syndrome. *Ann. Rheum. Dis.* **62**, 168–171 (2003).

103. Mackay, F. *et al.* Mice transgenic for BAFF develop lymphocytic disorders along with autoimmune manifestations. *J. Exp. Med.* **190**, 1697–1710 (1999).

104. Bowman, S. J. *et al.* Randomized Controlled Trial of Rituximab and Cost-Effectiveness Analysis in Treating Fatigue and Oral Dryness in Primary Sjögren's Syndrome. *Arthritis Rheumatol. (Hoboken, N.J.)* **69**, 1440–1450 (2017).

105. Devauchelle-Pensec, V. *et al.* Treatment of primary Sjögren syndrome with rituximab: a randomized trial. *Ann. Intern. Med.* **160**, 233–242 (2014).

106. Nordmark, G., Eloranta, M.-L. & Ronnblom, L. Primary Sjögren's syndrome and the type I interferon system. *Curr. Pharm. Biotechnol.* **13**, 2054–2062 (2012).

107. Emamian, E. S. *et al.* Peripheral blood gene expression profiling in Sjögren's syndrome. *Genes Immun.* **10**, 285–296 (2009).

108. Hjelmervik, T. O. R., Petersen, K., Jonassen, I., Jonsson, R. & Bolstad, A. I. Gene expression profiling of minor salivary glands clearly distinguishes primary Sjögren's syndrome patients from healthy control subjects. *Arthritis Rheum.* **52**, 1534–1544 (2005).

109. Hall, J. C. *et al.* Precise probes of type II interferon activity define the origin of interferon signatures in target tissues in rheumatic diseases. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17609–17614 (2012).

110. Gottenberg, J.-E. *et al.* Activation of IFN pathways and plasmacytoid dendritic cell recruitment in target organs of primary Sjögren's syndrome. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2770 LP-2775 (2006).

111. Gota, C. & Calabrese, L. Induction of clinical autoimmune disease by therapeutic interferon-alpha. *Autoimmunity* **36**, 511–518 (2003).

112. Mingueneau, M. *et al.* Cytometry by time-of-flight immunophenotyping identifies a blood Sjögren's signature correlating with disease activity and glandular inflammation. *J. Allergy Clin. Immunol.* **137**, 1809–1821.e12 (2016).

113. Manoussakis, M. N. & Kapsogeorgou, E. K. The role of intrinsic epithelial activation in the pathogenesis of Sjögren's syndrome. *J. Autoimmun.* **35**, 219–224 (2010).

114. Spachidou, M. P. *et al.* Expression of functional Toll-like receptors by salivary gland epithelial cells: increased mRNA expression in cells derived from patients with primary Sjögren's syndrome. *Clin. Exp. Immunol.* **147**, 497–503 (2007).

115. Deshmukh, U. S., Nandula, S. R., Thimmalapura, P.-R., Scindia, Y. M. & Bagavant, H. Activation of innate immune responses through Toll-like receptor 3 causes a rapid loss of salivary gland function. *J. oral Pathol. Med. Off. Publ. Int. Assoc. Oral Pathol. Am. Acad. Oral Pathol.* **38**, 42–47 (2009).

116. Iwakiri, D. *et al.* Epstein-Barr virus (EBV)-encoded small RNA is released from EBV-infected cells and activates signaling from Toll-like receptor 3. *J. Exp. Med.* **206**, 2091–2099 (2009).

117. Fleck, M., Kern, E. R., Zhou, T., Lang, B. & Mountz, J. D. Murine cytomegalovirus induces a Sjögren's syndrome-like disease in C57Bl/6-lpr/lpr mice. *Arthritis Rheum.* **41**, 2175–2184 (1998).

118. Balada, E., Vilardell-Tarrés, M. & Ordi-Ros, J. Implication of human endogenous retroviruses in the development of autoimmune diseases. *Int. Rev. Immunol.* **29**, 351–370 (2010).

119. Manoussakis, M. N., Tsinti, M., Kapsogeorgou, E. K. & Moutsopoulos, H. M. The salivary gland epithelial cells of patients with primary Sjögren's syndrome manifest

significantly reduced responsiveness to 17β-estradiol. *J. Autoimmun.* **39**, 64–68 (2012).

120. Ishimaru, N. *et al.* Development of autoimmune exocrinopathy resembling Sjögren's syndrome in estrogen-deficient mice of healthy background. *Am. J. Pathol.* **163**, 1481–1490 (2003).

121. Chu, Y. *et al.* B cells lacking the tumor suppressor TNFAIP3/A20 display impaired differentiation and hyperactivation and cause inflammation and autoimmunity in aged mice. *Blood* **117**, 2227–2236 (2011).

122. Duncan, C. J. A. *et al.* Early-onset autoimmune disease due to a heterozygous loss-of-function mutation in TNFAIP3 (A20). *Ann. Rheum. Dis.* **77**, 783–786 (2018).

123. Zhou, Q. *et al.* Loss-of-function mutations in TNFAIP3 leading to A20 haploinsufficiency cause an early onset autoinflammatory syndrome. *Nat. Genet.* **48**, 67–73 (2016).

124. Sisto, M. *et al.* A failure of TNFAIP3 negative regulation maintains sustained NF-κB activation in Sjögren's syndrome. *Histochem. Cell Biol.* **135**, 615–625 (2011).

125. Nocturne, G., Boudaoud, S. & Mariette, X. Germline and somatic genetic variations of TNFAIP3 in lymphoma complicating primary Sjögren's syndrome.

126. Savola, P. *et al.* Somatic mutations in clonally expanded cytotoxic T lymphocytes in patients with newly diagnosed rheumatoid arthritis. *Nat. Commun.* **8**, (2017).

127. Singh, M. *et al.* Lymphoma Driver Mutations in the Pathogenic Evolution of an Iconic Human Autoantibody. *Cell* **180**, 878–894.e19 (2020).

128. Jeelall, Y. S. *et al.* Human lymphoma mutations reveal CARD11 as the switch between self-antigen-induced B cell death or proliferation and autoantibody production. *J. Exp. Med.* **209**, 1907–1917 (2012).

129. Casellas, R. *et al.* Mutations, kataegis, and translocations in B lymphocytes: towards a mechanistic understanding of AID promiscuous activity. *Nat. Rev. Immunol.* **16**, 164–176 (2016).

130. Pettersen, H. S. *et al.* AID expression in B-cell lymphomas causes accumulation of genomic uracil and a distinct AID mutational signature. *DNA Repair (Amst).* **25**, 60–71 (2015).

131. Firestein, G. S., Echeverri, F., Yeo, M., Zvaifler, N. J. & Green, D. R. Somatic mutations in the p53 tumor suppressor gene in rheumatoid arthritis synovium. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 10895–10900 (1997).

132. Inazuka, M. *et al.* Analysis of p53 tumour suppressor gene somatic mutations in rheumatoid arthritis synovium. *Rheumatology* **39**, 262–266 (2000).

133. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).

134. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).

135. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

136. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

137. Conway, T. *et al.* Xenome—a tool for classifying reads from xenograft samples. *Bioinformatics* **28**, i172–i178 (2012).

138. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

139. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1 Next Gener. Seq. Data Anal.* (2011) doi:10.14806/ej.17.1.200.

140. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

141. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

142. Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1-15.10.18 (2016).

143. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

144. Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr. Protoc. Bioinforma.* **52**, 15.7.1-15.7.12 (2015).

145. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).

146. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).

147. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

148. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

149. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

150. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).

151. Network, C. G. A. R. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

152. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910 LP-16915 (2010).

153. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

154. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput. Biol.* **12**, e1004873–e1004873 (2016).

155. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).

156. Hoang, D. T. *et al.* MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).

157. Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical Dirichlet Processes. *J. Am. Stat. Assoc.* **101**, 1566–1581 (2006).

158. Freitas, T. A. K., Li, P.-E., Scholz, M. B. & Chain, P. S. G. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res.* **43**, e69 (2015).

159. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).

160. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

161. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).

162. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).

163. Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31**, 1974–1980 (2015).

164. Gazit, E. & Loewenthal, R. The immunogenetics of pemphigus vulgaris. *Autoimmun. Rev.* **4**, 16–20 (2005).

165. Ellebrecht, C. T. & Payne, A. S. Setting the target for pemphigus vulgaris therapy. *JCI insight* **2**, e92021–e92021 (2017).

166. Youinou, P., Devauchelle-Pensec, V. & Pers, J.-O. Significance of B Cells and B Cell Clonality in Sjögren's Syndrome. *ARTHRITIS Rheum.* **62**, 2605–2610 (2010).

167. Hansen, A. *et al.* Diminished peripheral blood memory B cells and accumulation of memory B cells in the salivary glands of patients with Sjögren's syndrome. *Arthritis Rheum.* **46**, 2160–2171 (2002).

168. Nocturne, G., Pontarini, E., Bombardieri, M. & Mariette, X. Lymphomas complicating primary Sjögren's syndrome: from autoimmunity to lymphoma. *Rheumatology* (2019) doi:10.1093/rheumatology/kez052.

169. Maehara, T. *et al.* Selective localization of T helper subsets in labial salivary glands from primary Sjögren's syndrome patients. *Clin. Exp. Immunol.* **169**, 89–99 (2012).

170. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btt750.

171. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **173**, 1823 (2018).

172. Wang, L. & Shilatifard, A. UTX Mutations in Human Cancer. *Cancer Cell* **35**, 168–176 (2019).

173. Stief, S. M. *et al.* Loss of KDM6A confers drug resistance in acute myeloid leukemia. *Leukemia* **34**, 50–62 (2020).

174. Li, X. *et al.* UTX is an escape from X-inactivation tumor-suppressor in B cell lymphoma. *Nat. Commun.* **9**, 2720 (2018).

175. Walport, L. J. *et al.* Human UTY(KDM6C) is a male-specific Nϵ-methyl lysyl demethylase. *J. Biol. Chem.* **289**, 18302–18313 (2014).

176. Chang, S., Yim, S. & Park, H. The cancer driver genes IDH1/2, JARID1C/ KDM5C, and UTX/ KDM6A: crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism. *Exp. Mol. Med.* **51**, 66 (2019).

177. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in leukocytes matters. *Nat. Genet.* **51**, 4–7 (2019).

178. MacHiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7**, (2016).

179. Invernizzi, P. *et al.* X Chromosome Monosomy: A Common Mechanism for Autoimmune Diseases. *J. Immunol.* **175**, 575 LP-578 (2005).

180. Richardson, A. L. *et al.* X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell* **9**, 121–132 (2006).

181. Bernell *et al.* Gain of chromosome 7 marks the progression from indolent to aggressive follicle centre lymphoma and is a common finding in patients with diffuse large B-cell lymphoma: a study by FISH. *Br. J. Haematol.* **101**, 487–491 (1998).

182. Vos, J. M. *et al.* High prevalence of the MYD88 L265P mutation in IgM anti-MAG paraprotein-associated peripheral neuropathy. *Journal of neurology, neurosurgery, and psychiatry* vol. 89 1007–1009 (2018).

183. Treon, S. P. *et al.* MYD88 L265P somatic mutation in Waldenström's macroglobulinemia. *N. Engl. J. Med.* **367**, 826–833 (2012).

184. Jiménez, C. *et al.* MYD88 L265P is a marker highly characteristic of, but not restricted to, Waldenström's macroglobulinemia. *Leukemia* **27**, 1722–1728 (2013).

185. Martinez-Lopez, A. *et al.* MYD88 (L265P) somatic mutation in marginal zone B-cell lymphoma. *Am. J. Surg. Pathol.* **39**, 644–651 (2015).

186. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).

187. Reddy, A. *et al.* Genetic and Functional Drivers of Diffuse Large B Cell Lymphoma. *Cell* **171**, 481–494.e15 (2017).

188. Andricovich, J. *et al.* Loss of KDM6A Activates Super-Enhancers to Induce Gender-Specific Squamous-like Pancreatic Cancer and Confers Sensitivity to BET Inhibitors.

Cancer Cell **33**, 512–526.e8 (2018).

189. Wang, J. K. *et al.* The histone demethylase UTX enables RB-dependent cell fate control. *Genes Dev.* **24**, 327–332 (2010).

190. Dunford, A. *et al.* Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nat. Genet.* **49**, 10–16 (2017).

191. Li, X. *et al.* Demethylase Kdm6a epigenetically promotes IL-6 and IFN-β production in macrophages. *J. Autoimmun.* **80**, 85–94 (2017).

192. Itoh, Y. *et al.* The X-linked histone demethylase Kdm6a in CD4+ T lymphocytes modulates autoimmunity. *J. Clin. Invest.* **129**, 3852–3863 (2019).

193. Fitzgerald, P. H. & McEwan, C. M. Total aneuploidy and age-related sex chromosome aneuploidy in cultured lymphocytes of normal men and women. *Hum. Genet.* **39**, 329–337 (1977).

194. Guttenbach, M., Koschorz, B., Bernthaler, U., Grimm, T. & Schmid, M. Sex chromosome loss and aging: in situ hybridization studies on human interphase nuclei. *Am. J. Hum. Genet.* **57**, 1143–1150 (1995).

195. Forsberg, L. A. *et al.* Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat. Genet.* **46**, 624–628 (2014).

196. Jørgensen, K. T. *et al.* Autoimmune diseases in women with Turner's syndrome. *Arthritis Rheum.* **62**, 658–666 (2010).

197. Lleo, A., Moroni, L., Caliari, L. & Invernizzi, P. Autoimmunity and Turner's syndrome. *Autoimmun. Rev.* **11**, A538-43 (2012).

198. Harris, V. M. *et al.* Klinefelter's syndrome (47,XXY) is in excess among men with Sjögren's syndrome. *Clin. Immunol.* **168**, 25–29 (2016).

199. Scofield, R. H. *et al.* Klinefelter's syndrome (47,XXY) in male systemic lupus erythematosus patients: support for the notion of a gene-dose effect from the X chromosome. *Arthritis Rheum.* **58**, 2511–2517 (2008).

200. Bueno, J. L. *et al.* Monosomy X as the sole cytogenetic abnormality in acute lymphoblastic leukemia: a report of two new patients. *Leuk. Lymphoma* **32**, 381–384 (1999).

201. Dörner, T. & Lipsky, P. E. Abnormalities of B cell phenotype, immunoglobulin gene expression and the emergence of autoimmunity in Sjögren's syndrome. *Arthritis Res.* **4**, 360–371 (2002).

202. Bian, Z. *et al.* Long non-coding RNA LINC00152 promotes cell proliferation, metastasis, and confers 5-FU resistance in colorectal cancer by inhibiting miR-139-5p. *Oncogenesis* **6**, 395 (2017).

203. Li, J. *et al.* Stromal microenvironment promoted infiltration in esophageal adenocarcinoma and squamous cell carcinoma: a multi-cohort gene-based analysis. *Sci. Rep.* **10**, 18589 (2020).

204. Pontarini, E., Lucchesi, D. & Bombardieri, M. Current views on the pathogenesis of ¨ gren ' s syndrome Sjo. 215–221 (2018) doi:10.1097/BOR.0000000000000473.

205. Gao, C.-Y. *et al.* Tissue-Resident Memory CD8+ T Cells Acting as Mediators of Salivary Gland Damage in a Murine Model of Sjögren's Syndrome. *Arthritis Rheumatol.* **71**, 121–132 (2019).

206. Deng, Q. *et al.* The Emerging Epigenetic Role of CD8+T Cells in Autoimmune Diseases: A Systematic Review . *Frontiers in Immunology* vol. 10 856 (2019).

207. Penkava, F. *et al.* Single-cell sequencing reveals clonal expansions of pro-inflammatory synovial CD8 T cells expressing tissue-homing receptors in psoriatic arthritis. *Nat. Commun.* **11**, 4767 (2020).

208. Yamanishi, Y. *et al.* Regional analysis of p53 mutations in rheumatoid arthritis synovium. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10025–10030 (2002).

209. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

210. Stephens, P. J. *et al.* Whole exome sequencing of adenoid cystic carcinoma. *J. Clin. Invest.* **123**, 2965–2968 (2013).

211. Kim, Y. *et al.* Integrative genomic analysis of salivary duct carcinoma. *Sci. Rep.* **10**, 14995 (2020).

212. Yan, K., Yesensky, J., Hasina, R. & Agrawal, N. Genomics of mucoepidermoid and adenoid cystic carcinomas. *Laryngoscope Investig. Otolaryngol.* **3**, 56–61 (2018).

213. Rettig, E. M. *et al.* Whole-Genome Sequencing of Salivary Gland Adenoid Cystic Carcinoma. *Cancer Prev. Res. (Phila).* **9**, 265–274 (2016).

214. Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* **45**, 977–983 (2013).

215. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science (80-.).* **370**, 75 LP-82 (2020).

216. Milewska, A. *et al.* APOBEC3-mediated restriction of RNA virus replication. *Sci. Rep.* **8**, 5960 (2018).

217. Harris, R. S. & Dudley, J. P. APOBECs and virus restriction. *Virology* **479–480**, 131–145 (2015).

218. Cannataro, V. L. *et al.* APOBEC-induced mutations and their cancer effect size in head and neck squamous cell carcinoma. *Oncogene* **38**, 3475–3487 (2019).

219. Smith, N. J. & Fenton, T. R. The APOBEC3 genes and their role in cancer: insights from human papillomavirus. *J. Mol. Endocrinol.* **62**, R269–R287.

220. Revathidevi, S., Murugan, A. K., Nakaoka, H., Inoue, I. & Munirajan, A. K. APOBEC: A molecular driver in cervical cancer pathogenesis. *Cancer Lett.* **496**, 104–116 (2021).

221. Zhu, B. *et al.* Mutations in the HPV16 genome induced by APOBEC3 are associated with viral clearance. *Nat. Commun.* **11**, 886 (2020).

222. Williams, M. D. *et al.* Genetic and Expression Analysis of HER-2 and EGFR Genes in Salivary Duct Carcinoma: Empirical and Therapeutic Significance. *Clin. Cancer Res.* **16**, 2266 LP-2274 (2010).

223. Yin, L. X. & Ha, P. K. Genetic alterations in salivary gland cancers. *Cancer* **122**, 1822–

1831 (2016).

224. Ho, A. S. *et al.* The mutational landscape of adenoid cystic carcinoma. *Nat. Genet.* **45**, 791–798 (2013).

225. Chen, A. A. *et al.* Oncogenic DNA viruses found in salivary gland tumors. *Oral Oncol.* **75**, 106–110 (2017).

226. Haeggblom, L. *et al.* No evidence for human papillomavirus having a causal role in salivary gland tumors. *Diagn. Pathol.* **13**, 44 (2018).

227. Saida, K. *et al.* Mutation analysis of the EGFR pathway genes, EGFR, RAS, PIK3CA, BRAF, and AKT1, in salivary gland adenoid cystic carcinoma. *Oncotarget* **9**, 17043–17055 (2018).

228. Blokland, S. L. M. *et al.* Epigenetically quantified immune cells in salivary glands of Sjögren's syndrome patients: a novel tool that detects robust correlations of T follicular helper cells with immunopathology. *Rheumatology* (2019) doi:10.1093/rheumatology/kez268.

229. Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjögren's syndrome. *Ann. Rheum. Dis.* **76**, 1458 LP-1466 (2017).

Appendix Table A1

Lymphoma/immune	IG_genes	TR_genes	HLA_genes	ncRNA_genes	SNP_sites
ACP5	IGHE	TRAJ:3-57	HLA-A	DAOA-AS1	rs10168266
ADA	IGHJ1	TRAV1-1	HLA-C	HSA-MIR-4537	rs10553577
ADRBK1	IGHJ2	TRAV1-2	HLA-DPB1	KIAA0125	rs10954213
AFF3	IGHJ3	TRAV10	HLA-DQA1	LINC01036	rs117026326
AGTPBP1	IGHJ4	TRAV12-1	HLA-DQB1	LINC02405-201	rs13426947
AHR	IGHJ5	TRAV12-2	HLA-DRA	Lnc-EIF2AK3-4	rs17338998
AICDA	IGHJ6	TRAV12-3	HLA-DRB1	Lnc-LRRTM1-3	rs17339836
AIRE	IGHV1-18	TRAV13-1	HLA-DRB5	Lnc-NEGR1-1	rs2409781
AKAP2	IGHV1-2	TRAV13-2	HLA-DRB6	Lnc-PLA2G4A-1	rs2729935
ALK	IGHV1-24	TRAV14DV4	HLA-G	LNC13	rs2736345
ARHGAP25	IGHV1-3	TRAV16	HLA-H	MIR142	rs2856674
ARID1A	IGHV1-45	TRAV17	HLA-C	MIR5011	rs3128917
ARID1B	IGHV1-46	TRAV18	HLA-DRB5	MIR875	rs3129716
ARID2	IGHV1-58	TRAV19		RP11-44H4.1	rs3135394
ARID5B	IGHV1-69	TRAV2			rs3757387
ASB4	IGHV1-8	TRAV20			rs4282438
ASXL1	IGHV2-26	TRAV21			rs4680536
ATM	IGHV2-5	TRAV22			rs4728142
ATRX	IGHV2-70	TRAV23DV6			rs4840568
B2M	IGHV3-11	TRAV24			rs485497
BACH2	IGHV3-13	TRAV25			rs4936443
BAK1	IGHV3-15	TRAV26-1			rs5029939
BAP1	IGHV3-20	TRAV26-2			rs583911
BCL10	IGHV3-21	TRAV27			rs6579837
BCL2	IGHV3-23	TRAV29DV5			rs6998387
BCL2L11	IGHV3-30	TRAV3			rs7119038
BCL6	IGHV3-33	TRAV30			rs73272842
BCL7A	IGHV3-43	TRAV34			rs7732451
BCOR	IGHV3-48	TRAV35			rs9260107
BIRC2	IGHV3-49	TRAV36DV7			rs9264672
BIRC3	IGHV3-53	TRAV38-1			rs9264672
BLK	IGHV3-64	TRAV38-2DV8			rs9271558
BLM	IGHV3-66	TRAV39			rs9271588
BRAF	IGHV3-7	TRAV4			rs9271588
BRCA1	IGHV3-72	TRAV40			rs9271591
BTG1	IGHV3-73	TRAV41			rs9271591
BTG2	IGHV3-74	TRAV5			
BTK	IGHV3-9	TRAV6			
BTLA	IGHV4-28	TRAV7			
C7ORF50	IGHV4-31	TRAV8-1			

CACNA2D1
CADM2
CARD11
CARF
CASP10
CASP8
CBFB
CCL11
CCND1
CCND3
CCR3
CD19
CD226
CD24
CD27
CD28
CD38
CD40
CD47
CD58
CD70
CD79A
CD79B
CDK4
CDKN2A
CDKN2B
CDKN2C
CHM
CHUK
CIITA
COL11A2
CORO1A
CR1
CR2
CREBBP
CTLA4
CTPS1
CUX1
CXCL13
CXCR4
CXCR5
DDX3X
DDX58
IGHV4-34
IGHV4-39
IGHV4-4
IGHV4-59
IGHV4-61
IGHV5-51
IGHV6-1
IGHV7-34-1
IGJ
IGKJ_1-5
IGKV1-12
IGKV1-16
IGKV1-17
IGKV1-27
IGKV1-33
IGKV1-39
IGKV1-5
IGKV1-6
IGKV1-8
IGKV1-9
IGKV1D-12
IGKV1D-13
IGKV1D-16
IGKV1D-17
IGKV1D-33
IGKV1D-39
IGKV1D-43
IGKV1D-8
IGKV2-24
IGKV2-28
IGKV2-30
IGKV2-40
IGKV2D-26
IGKV2D-28
IGKV2D-29
IGKV2D-30
IGKV2D-40
IGKV3-11
IGKV3-15
IGKV3-20
IGKV3D-11
IGKV3D-15
IGKV3D-20
TRAV8-2
TRAV8-3
TRAV8-4
TRAV8-6
TRAV9-1
TRAV9-2
TRBJ2_1-7
TRBV10-1
TRBV10-2
TRBV11-1
TRBV19
TRBV2
TRBV20-1
TRBV24-1
TRBV25-1
TRBV27
TRBV28
TRBV29-1
TRBV3-1
TRBV30
TRBV4-1
TRBV4-2
TRBV5-1
TRBV5-4
TRBV5-5
TRBV5-6
TRBV6-1
TRBV6-4
TRBV6-5
TRBV6-6
TRBV6-8
TRBV6-9
TRBV7-3
TRBV7-4
TRBV7-6
TRBV7-7
TRBV7-8
TRBV9
TRDJ1
TRDJ2
TRDJ3
TRDJ4
TRDV1

DIS3
DNMT3A
DTX1
EBF1
EED
EEF1A1
EGR1
EIF4A2
ELF1
EP300
ERBB2
ETS1
ETV1
ETV6
EZH2
FADD
FAM167A
FAM46C
FAM47C
FAM5C
FAS
FASLG
FBXW7
FLT3LG
FNBP1
FOXO1
FOXO3
FOXP1
FOXP3
GAPDH
GATA1
GATA2
GATA3
GIMAP5
GNAS
GRID2
GTF2I
GUCY1A2
H3F3A
H3F3B
HIST1H1B
HIST1H2AG
HIST1H2BD
IGKV3D-7
IGKV4-1
IGKV5-2
IGKV6-21
IGKV6D-21
IGLC3
IGLJ1
IGLJ2
IGLJ3
IGLJ4
IGLJ5
IGLJ6
IGLJ7
IGLL5
IGLV1-36
IGLV1-40
IGLV1-44
IGLV1-47
IGLV1-51
IGLV10-54
IGLV2-11
IGLV2-14
IGLV2-18
IGLV2-23
IGLV2-8
IGLV3-1
IGLV3-10
IGLV3-12
IGLV3-16
IGLV3-19
IGLV3-21
IGLV3-22
IGLV3-25
IGLV3-27
IGLV3-9
IGLV4-3
IGLV4-60
IGLV4-69
IGLV5-37
IGLV5-45
IGLV5-52
IGLV6-57
IGLV7-43
TRDV2
TRDV3
TRGJ1
TRGJ2
TRGV2
TRGV3
TRGV4
TRGV5
TRGV8
TRGV9

HIST1H2BK
HIST1H2BO
HIST1H3I
HIST1H4I
HIST1H4K
HIST2H2AB
HNRNPU
HRAS
ICOS
ICOSLG
ID3
IDH1
IDH2
IDO1
IFITM1
IKBKB
IKBKG
IKZF1
IKZF3
IL10
IL10RA
IL10RB
IL12A
IL18
IL18RAP
IL21
IL21R
IL22
IL2RA
IL7R
INPP5D
IRF1
IRF4
IRF5
IRF8
ITCH
ITK
ITPKB
JAK1
JAK2
JAK3
KDM5C
KDM6A
IGLV7-46
IGLV8-61
IGLV9-49

KIAA0226L
KLF2
KLHL1
KMT2A
KMT2C
KMT2D
KRAS
LAT2
LCK
LILRA3
LILRB1
LIMA1
LPHN3
LRBA
LTA
LTB
LYN
M3
MAGT1
MALT1
MAP2K1
MAP2K4
MAP3K14
MASP2
MEF2B
MEF2C
MET
MGA
MLL3
MLL5
MNDA
MPEG1
MSH6
MTOR
MYB
MYC
MYD88
NBEAL2
NBN
NCR3
NEAT1
NFE2L2
NFKB1

NFKB2
NFKBIA
NME1
NOTCH1
NOTCH2
NRAS
ORAI1
P2RX7
PAX5
PDGFRA
PDGFRB
PDL1
PELI1
PGM3
PHF19
PIK3CA
PIK3CD
PIK3R1
PIM1
PKN1
PLCG1
PLCG2
PMS2
PNN
PNP
POT1
POU2AF1
POU2F2
PPM1D
PRAMEL
PRDM1
PRKCD
PRKDC
PTEN
PTMA
PTPN1
PTPN11
PTPN22
PYCR1
RAD21
RAP1A
RASSF6
RB1

RCC1
RCSD1
RELA
RET
RHOA
RHOH
RNF34
ROBO1
ROS1
RUNX1
S1PR1
SEL1L3
SERF2
SETD2
SF3B1
SGK1
SH2B3
SH2D1A
SOCS1
SORL1
SP140
SPEN
SRSF2
SSB
SSNA1
SSSCA1
STAG2
STAT1
STAT3
STAT4
STAT5B
STAT6
STIM1
STK11
STK4
SUZ12
SYK
TAS1R1
TBL1XR1
TCF3
TCL1A
TERT
TET1

TET2
TET3
THRAP3
TLR7
TLR9
TMSB4X
TNFAIP3
TNFRSF13B
TNFRSF13C
TNFSF13B
TNFSF4
TNIP1
TNPO3
TP53
TP53INP1
TRAC
TRAF2
TRAF3
TRAF3IP2
TRAF5
TRAF6
TRIM21
TRIP11
TROVE2
TXLNA
TYK2
U2AF1
USP49
VPREB1
WAS
WASL
WEE1
WHSC1
WNK1
WT1
XBP1
XIAP
XPO1
ZAP70
ZBTB37
ZFP36L1
ZNF385B
ZNF608