Somatic evolution in normal human endometrium

Luiza Moore

Emmanuel College

University of Cambridge

Dissertation is submitted for the degree of Doctor of Philosophy

October 2019

Summary

For decades, the primary focus of cancer research has been the cancer tissue itself. Advances in next generation sequencing technologies have enabled identification and characterisation of driver mutations, provided insights into the tumour burdens and underlying mutational processes, sub-clonal diversification and tumour heterogeneity.

However, all cancers arise from cells that were once normal. Over time, they acquired certain mutations which increased their fitness, giving them a selective advantage over their neighbours and allowing uncontrolled growth, clonal expansion and malignant transformation. Our understanding of somatic evolution occurring in normal tissues with age and in the early stages of tumourigenesis remains relatively poorly understood.

In this thesis, I aimed to investigate somatic evolution in normal ageing human tissues. Firstly, I helped to establish a robust low DNA input whole genome sequencing workflow for laser-capture micro-dissected cellular material. I then utilised this approach to explore genomic and evolutionary landscapes of the normal human endometrium.

In the first results chapter, I investigate the clonal composition of normal endometrial glands. The majority of glands are clonal cell populations that share a common recent ancestor and the monoclonality is independent of whether they have a driver mutation.

In the second results chapter, I investigate the mutational landscape of normal endometrial glands. We show that somatic mutations (base substitutions, indels and genome rearrangements) accumulate with age in a more-or-less linear manner. A small number of ubiquitous mutational processes accounts for the majority of all mutations. A remarkably high proportion of normal endometrial glands carry at least one driver mutation (of the type that one is used to finding in cancers). Accumulation of drivers is negatively affected by parity. Through phylogenetic tree reconstruction of somatic mutations in endometrial glands, we show that driver mutations often occur early in life and continue to accumulate with age.

This work identifies a distinct mutational landscape in normal endometrium that is in keeping with the presence of early positive selection in this highly regenerative tissue.

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and 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 diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee. This dissertation does not exceed the prescribed limit of 60,000 words.

Acknowledgements

First and foremost, I would like to thank my first doctoral supervisor, Mike Stratton for giving me an opportunity to work on these challenging, yet absolutely incredible projects. I am forever indebted to him for his continuous mentorship, trust, patience and supervision of the work presented. I would also like to express my most sincere gratitude and appreciation to my second supervisor, Peter Campbell for his guidance, support and encouragement throughout this PhD. I have learnt so much from working with both of you!

The experimental work in this thesis would not have been possible without Peter Ellis, whose enthusiasm and huge efforts have been instrumental in the development of the low input LCM pipeline that is now used by so many in the Cancer, Ageing and Somatic Mutation (CASM) Programme. I would also like to say special thanks to Mathijs Sanders, Daniel Leongamornlert and Tim Coorens for not only their massive help with the data analysis, but also for all those hours of fruitful variant filtering, tree building and signature extraction discussions!

I am indebted to so many members of the CASM Team. I thank Patrick Tarpey for his support and guidance for when I first started at the department; Inigo Martincorena for endless conversations about selection in normal tissues and cancers; Jyoti Nangalia for sharing their wisdom on science and life in general; Yvette Hooks for help with histology; Andrew Lawson for his unbelievable proofreading skills, which I sadly only discovered at the very end of my PhD; Henry Lee-Six for sharing the pain of late night laser-capture microdissection sessions; Grace Collord for her endless supply of baking goods; Tim Butler, Clare Pacini and Sarah Moody for reminding me to have life outside of work; and to so many others.

Finally, I would like to express my biggest gratitude to my friends and family, especially my very understanding and supportive parents and my son, Sebastien, who have been my rock and kept me smiling.

Table of Contents

Chapter 1 **General introduction**

1.1 Introduction

All cells in the human body are thought to acquire somatic mutations. Most of these mutations are harmless and are termed ‘passengers’. However, some of the mutations confer increased cellular fitness and selective advantage leading to uncontrolled cellular growth, clonal expansion and eventually neoplastic transformation (‘driver mutations’).

1.2 Cancer is a disease of the genome fuelled by somatic mutations

For decades, the primary focus of cancer research has been the cancer tissue itself. Advances in next generation sequencing (NGS) technologies have enabled identification and characterisation of driver mutations, provided insights into the tumour burdens and underlying mutational processes, sub-clonal diversification and tumour heterogeneity. Cancers can now be described in terms of their mutation burden, mutational processes and patterns of selection. These are considered below.

1.1.1 Mutation burden

Large scale next generation sequencing initiatives (Alexandrov, 2018, Cancer Genome Atlas Research et al., 2013, Alexandrov, 2013) have allowed better characterisation of the tumour mutation burden. These analyses have shown a huge variation in the rates of somatic mutations across different types of cancer with the majority of tumours showing 1000-20,000 somatic point mutations and much smaller numbers of insertions, deletions, and rearrangements.

1.1.2 Mutational processes

Cancer genomes carry thousands of somatic mutations, but only a very small proportion of these are “drivers” that are implicated in oncogenesis. The remainder are “passengers”, the bystanders of the mutational processes that have been operative in those tissues throughout life and the development of cancer (Helleday et al., 2014). These mutations occur spontaneously as a result of various processes, termed ‘signatures’ (Alexandrov, 2013). They can be of endogenous source, such as reactive oxygen species, defective DNA repair mechanisms and infidelity in the DNA replication machinery, or of exogenous source, such as ultra-violet light exposure and tobacco smoking (Alexandrov, 2013, Alexandrov et al, 2015).

1.2.1.1 Early work on mutational patterns

Different mutational processes leave specific patterns of mutations on the cancer genomes, which are termed “mutational signatures”. Some of the first efforts to characterise mutational patterns were made back in the 90’s (Hollstein et al., 1991, Hollstein et al., 1999). In a series of studies, multiple samples of the same cancer type were combined to examine patterns of coding mutations in *TP53*. These analyses yielded two key observations. First, ultra-violet light exposure related skin cancers were characterised by frequent C>T transversions occurring primarily at dipyrimidines, which was in keeping with the pattern of mutation observed *in vitro.* Second, a strong C>A pattern was seen in tobacco smoking related lung cancers, which matched the observation made *in vitro* of DNA exposure to benzo(a)pyrene, a known tobacco carcinogen (Nik-Zainal et al., 2015). While these studies provided first insights into mutational patterns, the analyses were primarily focused around processes with strong mutagenic activity that would generate most of mutations detected in individual cancers. However, more than one mutational process may have been operative in a given cancer, but the “signal” from these may not be readily deciphered in the mixture of mutations.

1.2.1.2 Next Generation Sequencing studies

Subsequently, advances in next generation sequencing (NGS) have resulted in large amounts of whole exome and genome sequencing data. This implied that thousands of somatic mutations were identified in individual cancers, which in turn provided sufficient power to apply mathematical algorithms to extract individual mutational signatures.

Large-scale cancer genome sequencing initiatives not only generated comprehensive lists of somatic mutations, but also provided an opportunity to decipher mutational signatures from thousands of cancers (Alexandrov et al., 2013, Cancer Genome Atlas Research et al., 2013, Alexandrov et al., 2018). Some of these signatures are present in most cancer types, for example a signature associated with the APOBEC family of cytidine deaminases, while others are unique to specific tumours. It is now also known that while certain mutational processes operate continuously, leading to accumulation of somatic mutations at a constant rate over decades, in a 'clock-like' fashion, others generate these more intermittently (Petljak et al., 2019). These mutational processes determine the mutation burdens that result in the first "driver" mutations leading to neoplastic change and may contribute to other normal and diseased biological states including ageing. Furthermore, these mutational processes may change in non-cancer disease states in which the metabolic state of the cell is chronically altered and thus may provide us with a record in DNA of these metabolic changes.

As more whole genome sequencing data have become available, a more comprehensive characterisation of the signatures has been possible of not only single base substitutions, but also of dinucleotide substitutions, small insertions and deletions (indels) and structural variants (Alexandrov et al., 2018).

In addition, in an attempt to better our understanding of the underlying mechanisms of the mutational processes, Kucab and colleagues tested 79 known or suspected environmental agents and their effect on single base substitutions (Kucab et al., 2019). The study found that approximately 50% of the tested mutagens were associated with specific mutational processes, several of which matched those previously observed in tumours, including UV-light and tobacco-related carcinogens.

Finally, work by Alexandrov and colleagues made a first attempt at estimating the 'clock-like' mutation rates in normal cells by interrogating thousands of cancer genomes (Alexandrov et al., 2015). The study identified two mutational signatures that were seen in most cancer types and accumulated mutations at a constant rate over time, thus confirming the existence of mutational molecular clocks.

1.1.3 Patterns of selection and driver mutations

The above mentioned large sequencing initiatives have also allowed identification and characterisation of cancer-associated mutations. As a result, there are now more than 600 genes that are thought to be implicated in oncogenesis (COSMIC). Statistical models (*dN/dS*) were subsequently applied to identify genes that are under selection across cancer types (Martincorena et al., 2017). These analyses have also highlighted driver burden differences between cancers with some types, such as chromophobe renal cell carcinomas and ovarian carcinomas, characterised by only a handful of driver genes, and others, such as urothelial and endometrial carcinomas showing a much broader range of genes under selection.

1.2.1.3 Multi-step clonal tumour evolution and heterogeneity

The multistep process of tumourigenesis was first proposed in 1958 (Foulds, 1958). Molecular events that drive cancer development and progression were further characterised over the following 30 years (Farber and Cameron, 1980; Weinberg, 1989). Some of the key analyses included work by Fearon and Vogelstein in colon in which they showed the complexity of the genetic path in colorectal cancer development (Fearon and Vogelstein, 1990). They examined different histopathological states in the colone, from normal epithelium to invasive colorectal adenocarcinoma. The work showed that the great majority of early adenomatous polyps carried inactivating mutations of the tumour-suppressor gene *APC*. Approximately half of the intermediate-sized lesions carried activating mutations of *ras* oncogenes and about half of the advanced colorectal carcinomas had mutations in the tumour-suppressor gene *TP53* (Kinzler and Vogelstein, 1996).

1.3 Current knowledge of somatic evolution in normal tissues

1.3.1.1 Driver mutations and clonal expansion

Some of the first studies reporting somatic mutations in normal tissues were carried out in blood. Gene fusion events that are typically seen in leukaemias and lymphomas, were detected in nearly 30% of clinically normal individuals studied (Biernaux et al., 1995, Bose et al., 1998). Furthermore, work on cord blood showed that *TEL-AML1* and *AML1-ETO* gene fusions associated with leukaemia can occur early in life with such events identified in around 1% of healthy neonates (Mori et al., 2002).

In 2014, seminal publications based on whole exome sequencing of large cohorts of patients showed that driver mutations, including *DNMT3A*, *TET2* and *JAK2* that are implicated in myeloid neoplasms, are frequently found in the blood of older but otherwise healthy individuals (Jaiswal et al., 2014, Genovese et al., 2014). The observation was termed clonal haematopoiesis. Work by Jaiswal and colleagues later showed that the presence of those clonal expansions conferred a small but significant risk of leukaemia (0.5%-1% per year) and that these clones represent early steps of tumourigenesis (Jaiswal et al., 2014). It was later shown that clonal haematopoiesis with cancer-associated mutations can occur at all ages (3% in 20-29-year olds; 20% in 60-69-year olds).

Subsequently, driver mutations identified in blood were also shown to be associated with non-malignant diseases: in addition to an increased risk of haematological neoplasms, the rates of coronary heart disease and ischaemic stroke were also increased (Jaiswal et al., 2017).

Detection of somatic mutations in normal solid tissues has been more challenging due to biological limitations, including slower proliferation, clonally restrictive tissue architecture, more difficult tissue access, and technical issues. A series of studies assessing clonal expansions in normal tissues, such as colon, prostate and liver, were carried out using mutations in mitochondrial DNA (Fellous et al., 2009a, Fellous et al., 2009b, Blackwood et al., 2011, Greaves, 2003, Greaves et al., 2006). However, while these analyses provided some insights into clonal composition of those tissues, the role of mitochondrial mutations in clonal expansion and tumourigenesis is poorly understood.

The first ground breaking analysis of somatic mutations in normal solid tissues was carried out by Martincorena and colleagues, in which extensive clonal patches bearing mutations in cancer genes, including *TP53, NOTCH1, NOTCH2, NOTCH3* and *FAT1*, were identified in normal sun-exposed skin of middle-aged to elderly individuals (Martincorena et al., 2015). Later, accumulation of somatic mutations, including those in cancer genes, and associated tissue remodelling have been shown in normal oesophagus (Martincorena et al., 2018, Yokoyama et al., 2019).

Finally, accumulation of cancer-associated mutations with age is not limited to somatic cells. Targeted studies on testicular tissue from healthy men have shown that mutations conferring predisposition to cancer could also confer a selective advantage to spermatogonia stem cells leading to clonal expansion similar to the process of oncogenesis (Maher et al., 2016). Over time, this clonal expansion leads to the relative enrichment of mutant sperm and in some cases, to large clones with driver mutations, such as *FGFR3* and *HRAS*, expanding within the testes, and can be associated with spermatocytic seminoma in older men (Goriely et al., 2009).

1.3.1.2 Mutational processes and burden

DNA mutations are inevitable, but it is the alterations that occur in the genomes of adult stem cells (ASC) that have the greatest impact on the tissue mutational burden and are thought to be most significant in terms of cancer risk (Tomasetti and Vogelstein, 2015). Tissues with high ASC turnovers show higher cancer incidence in comparison to those with lower ASC turnover rates. It is therefore important to assess somatic mutation accumulation in ASCs of different tissues. Previous work on clonal organoid cultures derived from liver, small intestine and colon has shown that despite significant variation in the cancer incidence in these tissues, somatic mutations accumulate at a similar rate of around 40 single base substitutions per year (Blokzijl et al., 2016). Although age-associated signatures (Signature 1 and 5) were observed in all three tissues, their contribution in the liver was markedly different from that observed in the small intestine and colon with the majority of substitutions attributed to signature 5, a signature of an unknown underlying mechanism. Interestingly, there was little intra-tissue inter-individual variation in the mutational spectra across ages.

As mentioned earlier, age-associated accumulation of somatic mutations is not unique to the soma, but has also been reported in the germline. Studies on trios have shown that *de novo* mutations accumulate with age in the paternal germline, and that there is a degree of variability across individuals. Surprisingly, the underlying mutational processes (mostly attributed to signature 5 and to a lesser extend to signature 1) are similar between paternal and maternal germlines as well as across individuals from a range of ages (Rahbari et al., 2016, Jonsson et al., 2017) .

1.1.4 Methods for studying somatic mutations in normal tissues

Normal tissues are complex systems comprising different populations of cells with distinct morphological and functional properties and specific spatial arrangements. However, this cellular heterogeneity implies that normal tissues are composed of many clones that are usually too small to provide sufficient amount of DNA that is necessary for standard sequencing protocols. In recent years, a number of approaches have been developed with the aim to study normal tissues (Table 1.1). Some of these are considered below.

1.3.1.3 Single cell genomics

Ideally, one would like to explore tissue heterogeneity targeting one cell at a time, and single cell technologies have the potential to provide new insights into the genomic landscapes of tumour and normal tissues. Recently, Casasent and colleagues applied this approach to laser-capture micro-dissected cells to assess genomic changes, particularly copy number variants, and to delineate clonal evolution in early-stage breast cancer (Casasent et al., 2018). However, the majority of such work has been performed on single cells in suspension and not laser-captured material. Overall, these technologies are still under development and are frequently associated with a whole myriad of issues, including incomplete genome coverage, whole genome amplification-induced errors and suboptimal variant calling sensitivity (Gawad et al., 2016, Navin, 2015) (Table 1.1).

1.3.1.4 **Single stem cell derived organoids**

An alternative way to study genomic landscapes of individual cells is through the use of *in-vitro* clonal organoid experimental models derived from single adult stem cells (Roerink et al., 2018, Blokzijl et al., 2016, Fatehullah et al., 2016). These provide sufficient amounts of DNA for standard 'bulk' sequencing methods while circumventing whole genome amplification and associated issues. However, while this approach has substantial utility, these are often challenging to derive, are highly laborious to generate in large numbers, may show bias towards certain subtypes of cell in a tissue, lack spatial information, may favour cells with driver mutations and will introduce additional mutations during cell culture that often include additional mutational signatures.

1.3.1.5 **Error-corrected next generation sequencing (ecNGS)**

Another way to study genomic changes in normal tissues at a cellular level is through removal of sequencing errors and identification of variants that are present at very low frequencies (Hoang et al., 2016, Kennedy et al., 2014, Schmitt et al., 2012). One of these approaches is Duplex sequencing, in which both strands of DNA are tagged and mutations are only considered *bona fide* if they are present in both strands of DNA and are complimentary (Schmitt et al., 2012). Subsequently, this approach was applied to detect somatic mutations, including those in *TP53*, at frequency <0.01% in peritoneal fluid samples from women without cancer (Krimmel et al., 2016).

Another example of ecNGS method is the bottleneck sequencing system (BotSeqS), which aims to reduce the error rate of NGS by utilising the consensus of reads from individual template molecules to discriminate *bona fide* variants from PCR artefacts. This has been achieved by circularisation of the DNA template, the addition of unique molecular identifiers (UMIs) to asymmetric (Y-shaped) adapters and utilising the mapping coordinates of reads as endogenous barcodes. The theoretical error rate for these approaches is reported to be <1 artefact per 10^9 nucleotides sequenced, which is calculated by assuming two independent mutational events (one on each strand of the original template molecule) occurring at the average substitution rate for high-fidelity DNA polymerases.

Method	Advantages	Disadvantages
Single cell sequencing	Allows to examine genomic changes in individual cells	Usually requires prior WGA
		WGA can be associated with poor genome coverage, allele/locus drop out and artifacts
Organoids	Provides sufficient DNA for standard library preparation and sequencing protocols	Not available for all tissue and cell types
	Does not require prior WGA	Additional mutations introduced during cell culturing
	Provides information on individual adult stem cells	Takes time to grow and is laborious
	Clonal samples, therefore more confident variant calling	Loss of spatial information
Error-corrected methods on bulk sequencing	Allows to detect mutations at a single molecule level	Incomplete genome coverage
		Can only be used for calling single base substitutions and indels but not copy number and structural variants
		Final variants represent an 'average' from a mixture of molecules from a relatively large population of cells and burden can be affected (increased) by cells with higher mutation burdens

Table 1.1 | Methods for studying somatic mutations in normal tissues. WGA, whole genome amplification.

1.4 Thesis aims

In this thesis, I aimed to investigate somatic evolution in normal ageing human tissues. Firstly, I helped to establish a robust low DNA input whole genome sequencing workflow for laser-capture micro-dissected cellular material. I then utilised this approach to explore genomic and evolutionary landscapes of the normal human endometrium.

In the first results chapter, I describe the clonal composition of laser-capture micro-dissected normal endometrial glands with multiple samples derived from 28 pre- and post-menopausal women. I also correlate the effect of menstrual phase, menopause status and presence or absence of driver mutations on clonality.

In the second results chapter, I investigate the mutational landscape of normal endometrial epithelium, including mutation burdens, signatures and prevalence of driver mutations and how these are modulated by age and parity. In addition, through phylogenetic tree reconstruction of somatic mutations in endometrial glands, we estimate the age at which the identified driver mutations occurred. Finally, I compare mutation burdens and patterns of selection of the normal endometrial epithelium and endometrial cancer.

Chapter 2 **Materials and methods**

2.1 Samples

2.1.1 Endometrium

Anonymized snap-frozen endometrial tissue samples were obtained from five different cohorts.

Cohort 1: Samples from individuals PD37605, PD37601, PD37607, PD37613, PD37594, PD37595, PD41871, PD41860, PD41857, PD41865, PD41868, PD41859, PD41861 and PD41869 (age 29 to 46) were provided by Professor Jan Brosens; these were collected from women undergoing hysteroscopy examination at the Tommy's National Early Miscarriage Centre, University Hospitals Coventry and Warwickshire NHS Trust. Informed consent was obtained and biopsies collected and stored at the Arden Tissue Bank, University Hospitals Coventry and Warwickshire NHS Trust in line with the protocols approved by the NRES Committee South Central Southampton B (REC reference 12/SC/0526, 19/04/2013).

Cohort 2: Samples from individuals PD40535, PD39444, PD39953, PD39952, PD39954, PD40107, PD42746 and PD42475 (age 24 to 74) were collected by Mr Kourosh Saeb-Parsy from non-uterine transplant organ donors with an informed consent obtained from the donor's family (REC reference: 15/EE/0152 NRES Committee East of England – Cambridge South).

Cohort 3: Individuals PD36804 and PD36805 (age 47 and 49), underwent total abdominal hysterectomy for benign non-endometrial pathologies and uterine biopsies were collected, snap frozen and stored at the Human Research Tissue Bank, Cambridge University Hospitals NHS Foundation Trust by Dr Mercedes Jimenez-Linan. The samples were collected in line with the protocols approved by the NRES Committee East of England (REC reference 11/EE/0011, 11/03/2011).

Cohorts 4 and 5: Samples from individuals PD37506, PD38812, PD37507 and PD40659 (age 19 to 81) were collected at autopsy following death from non-gynaecological causes. The use of this material was approved by the London, Surrey Research Ethics Committee (REC reference 17/LO/1801, 26/10/2017) and East of Scotland Research Ethics Service (REC reference: 17/ES/0102, 27/07/2017).

2.1.2 Pan-body survey

2.1.2.1 Donor 1

In collaboration with Professor Rebecca Fitzgerald and her research team led by Miss Ayesha Noorani, I collected 252 samples from a variety of macroscopically normal tissues during a rapid ('warm') autopsy. The samples were collected in line with the protocols approved by the NRES Committee East of England (NHS National Research Ethics Service reference 13/EE/0043). The post-mortem sample collection was performed on a 78-year-old male, non-smoker who died of a metastatic oesophageal carcinoma; he had no other co-morbidities. The collection was completed within six hours of the patient's death to ensure tissue integrity for morphology preservation and whole genome sequencing (WGS). Every sampled tissue was photographed and biopsy sites carefully documented. As there was an extensive lower oesophageal tumour that invaded into the pancreas, I was not able to obtain any normal tissue samples from the stomach and pancreas. Once collected, all biopsies were snap frozen in liquid nitrogen and subsequently stored at -80^{0}C. Summary of all sampled tissues is provided in Appendix 1.

2.1.2.2 Donors 2 and 3

Multiple biopsies from twenty-six different tissues were collected from a 54-year-old female and a 47 year old male; both individuals died of non-cancer causes (acute coronary syndrome and traumatic injuries respectively). All samples were obtained within less than five hours of death. The use of these tissues was approved by the London, Surrey Research Ethics Committee (REC reference 17/LO/1801, 26/10/2017). Summary of all obtained tissues is provided in Appendix 2.

2.1.2.3 **Additional limited samples from other donors**

To obtain the most comprehensive catalogue of somatic mutations across as many female and male tissues as possible and to further validate some of our observations, we acquired additional samples, mostly from one or two organs from additional donors. These included, breast, stomach, endometrium, cervix, fallopian tubes, pancreas, testis, colon and others. These samples were obtained at autopsy following death from non-cancer causes. The use of this material was approved by the London, Surrey Research Ethics Committee (REC reference 17/LO/1801, 26/10/2017) and East of Scotland Research Ethics Service (REC reference 17/ES/0102, 27/07/2017).

2.2 Laser-capture microscopy

In this work, we aimed to study somatic mutations in relatively small populations of cells from specific morphological or functional units, such as endometrial glands or colonic crypts. These units typically contain 200-2000 cells, which would equate to approximately 1.2-12 ng of DNA. When I first started my PhD (April 2016), a minimum of 200 ng of input DNA (equivalent to around 33,300 cells) was required for a successful library preparation by the standard sequencing methods.

Fortunately, Peter Ellis, who at the time was a Principle Staff Scientist in the Research and Development Department, was testing different approaches to decrease the amount of input DNA for efficient library construction. I have therefore spent the first 10 months of my PhD working together with Peter to build a workflow that would enable robust processing of low input LCM derived cellular material. The experimental side of this process involved three major components: (a) effective tissue preparation (fixation and morphology), (b) cell lysis and (c) DNA isolation and library construction.

2.2.1 Tissue preparation

Tissue fixation is an essential step in histology as it preserves morphology for accurate microscopic assessment. However, routine histology fixatives, specifically formalin, are known to have a detrimental impact on both the quality and quantity of extracted DNA (Howat and Wilson, 2014). It was therefore essential to optimize this step and to find an alternative fixative. Three non-cross-linking fixatives were tested: acetone (100%), ethanol (70%) and methanol (100%). Out of these three, ethanol fixation provided the most optimal morphology preservation, followed by methanol and acetone.

In general, two types of tissue preparation are used for histology assessment: frozen and paraffin sections. Protocol for the first method usually involves cutting sections from a frozen block, followed by a brief (2-5 minutes) immersion in a fixative (70% ethanol in our protocol), followed by staining with haematoxylin and eosin (H&E) or haematoxylin only (H) (Figure 2.1). The second approach can take up to two days and includes several hours of fixation (to allow fixative to penetrate through the entire tissue block) and embedding in paraffin, followed by sectioning, xylene-based deparaffinisation and staining (Figure 2.1).

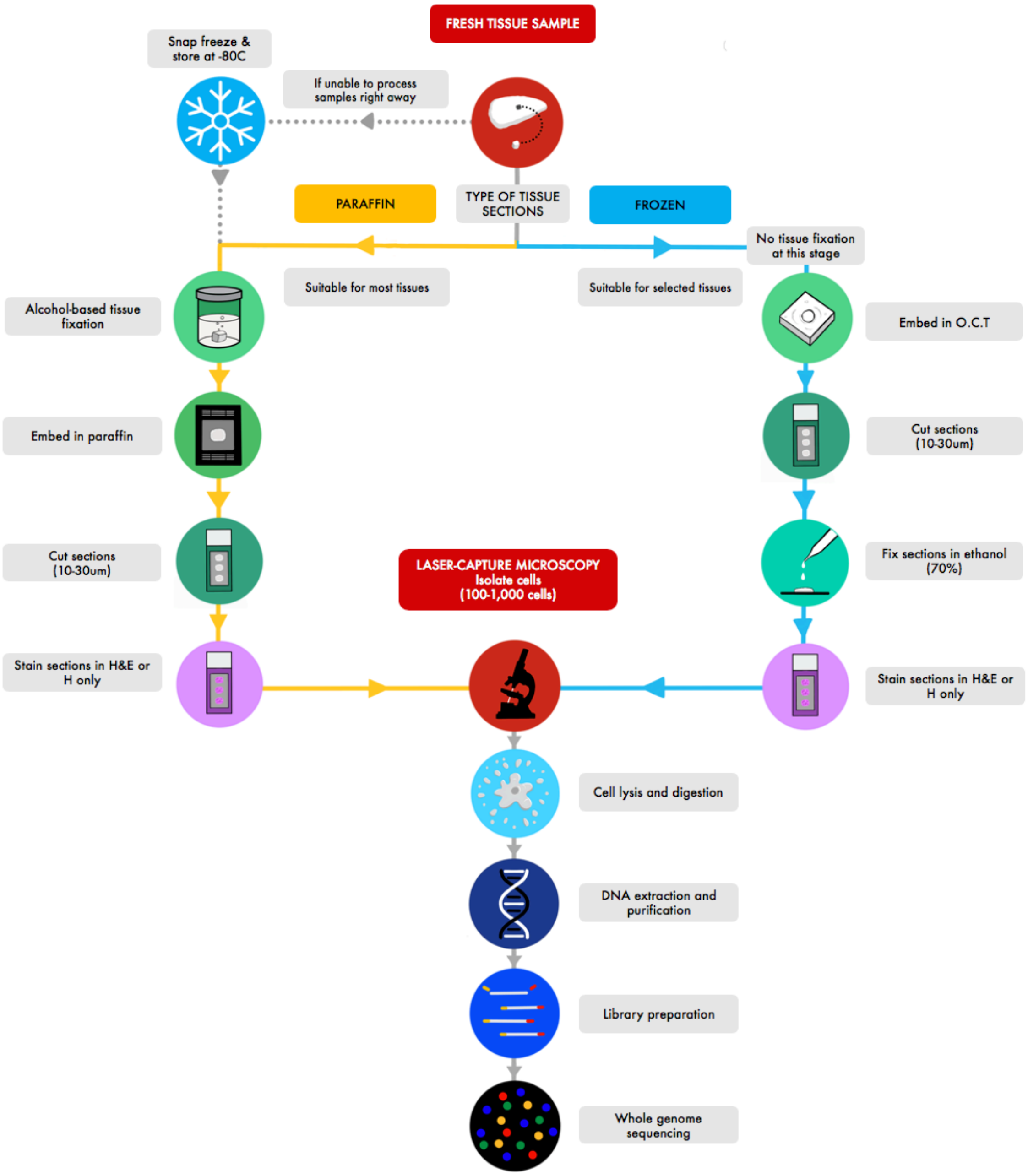

FRESH TISSUE SAMPLE
Snap freeze & store at -80C
If unable to process samples right away
PARAFFIN
TYPE OF TISSUE SECTIONS
FROZEN
No tissue fixation at this stage
Suitable for most tissues
Suitable for selected tissues
Alcohol-based tissue fixation
Embed in O.C.T
Embed in paraffin
Cut sections (10-30um)
Cut sections (10-30um)
Fix sections in ethanol (70%)
LASER-CAPTURE MICROSCOPY Isolate cells (100-1,000 cells)
Stain sections in H&E or H only
Stain sections in H&E or H only
Cell lysis and digestion
DNA extraction and purification
Library preparation
Whole genome sequencing

Figure 2.1 | Summary of the LCM workflow. Tissue morphology can be assessed using frozen and paraffin sections. This figure outlines individual steps in both approaches. Sections can be stained using haematoxylin and eosin (H&E) or haematoxylin only (H). To aid sectioning of frozen tissue blocks, biopsies are embedded in rapidly solidifying optimal cutting temperature compound (O.C.T.). Specific morphological structures or tissue-specific functional units, such as colonic crypts or endometrial glands (typically containing 200-2,000 cells), are laser-capture micro-dissected into individual wells. The cellular material is subjected to our modified protocols for cell-lysis, DNA extraction and library preparation for whole genome sequencing.

Given the fact that we were working with relatively small amounts of input DNA, we wanted to minimise tissue handling and potential DNA degradation. Therefore, we first focused on optimisation of our workflow for frozen sections (Appendix 3). However, while this method is suitable for some tissues, such as colon and endometrium, for many other tissue types, for instance, brain and testis, it results in poor preservation of morphology and inability to accurately type cells and structures (Figure 2.2). We therefore also optimized tissue fixation and preparation protocols for experiments performed on paraffin embedded material (Appendix 4).

Routine clinical histology sections are around 4-5 micron thick. However, to increase the amount of input DNA, while also allowing accurate morphology assessment, the section thickness for most tissue types was chosen to be 10 microns.

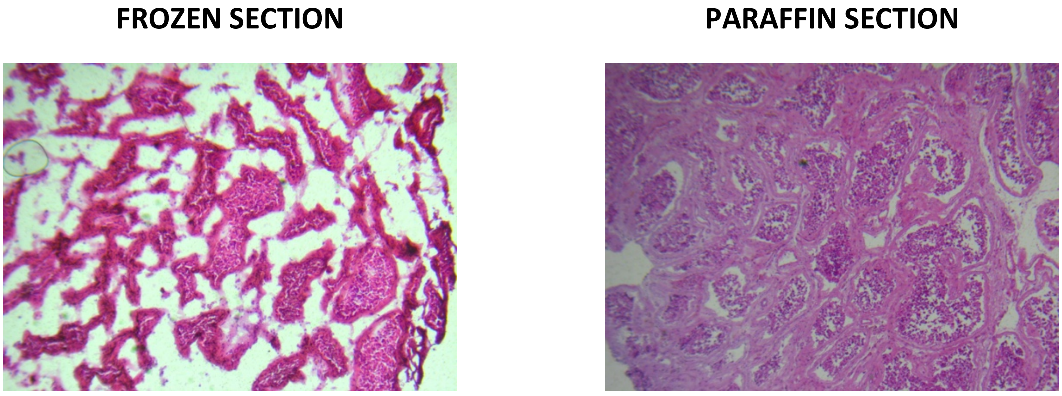


Figure 2.2 |Comparison of testicular histology using frozen and paraffin sections (H&E, 5x magnification, 10 micron thickness). The figure shows an example of the two different tissue preparation methods and their effect on preservation of morphology.

2.2.2 Cell lysis

To maximise DNA recovery from micro-dissected cellular material, three different types of lysis buffers were tested: alkaline lysis, protease lysis (an in-house version, Appendix 5 or a commercially available Arcturus™ PicoPure™ DNA extraction kit) and chaotropic lysis (RLT). Fixatives and lysis buffers were tested jointly. Below are the results of some of these tests. From these and further experiments on other tissue types, a combination of ethanol (70%) and protease lysis buffer was selected (Figure 2.3).

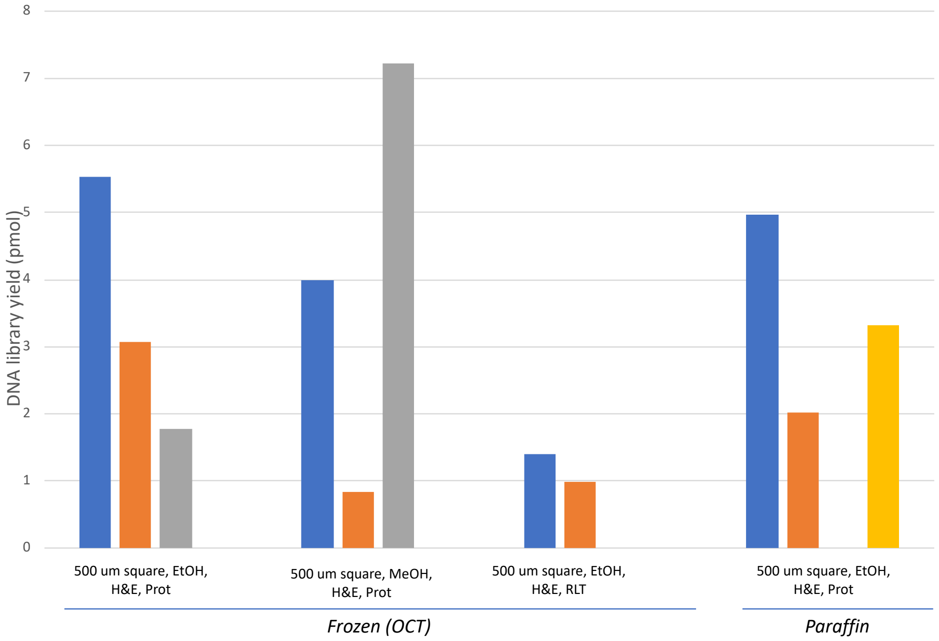


Figure 2.3 | Quantification of libraries for assessment of fixation and lysis conditions. This figure shows DNA library yeilds obtained when testing different types of fixatives (70% ethanol (EtOH) and methanol (MeOH)) and lysis buffers (proteased based buffer (Prot) and chaotropic lysis buffer (RLT). H&E, haematoxylin and eosin; OCT, optimal cutting temperature compound. Adapted from Peter Ellis. Different fixation and lysis conditions were tested on frozen and paraffin tissue sections.

2.2.3 DNA isolation and library construction

Traditionally, DNA purification and quantification are separate steps. In our protocol, to maximize DNA recovery from the low input samples, we introduced a modified solid phase reversible immobilization (SPRI) bead purification step within the library construction workflow and omitted DNA quantification altogether.

Early tests indicated that genomic DNA recovery at the DNA purification step could be as low as 50%, which led us to believe that a large proportion of high molecular weight genomic DNA was refractory to elution from the SPRI beads. The entire post-elution sample (including beads) was therefore integrated into the library construction workflow to minimize these losses. It is likely that a combination of buffer detergent, heat and the action of the fragmentation enzymes in the next step promotes the release of all available DNA into solution.

Standard NGS workflows for whole genome sequencing typically use around 200 ng input DNA material, often fragmented by acoustic shearing. Fragmented DNA is repaired, dA-tailed, ligated to adapter sequences and indexed by PCR amplification for 6 cycles. Additional PCR cycles are introduced to ameliorate lower DNA inputs; however, this approach is useful only when the predefined minimum number of unique DNA templates are present in the final DNA library. For instance, sufficient material can be generated from <1 ng human genomic DNA to perform whole exome sequencing. However, our ability to produce sequencing data with a meaningful library complexity drops dramatically below 10 ng input DNA material. In contrast, we discovered that DNA fragmentation reagents that utilize enzymatic rather than acoustic fragmentation, yielded a >10-fold improvement in DNA library yield. This increase in efficiency led to a dramatic reduction in PCR duplicate rates that enables the generation of whole exome sequencing data from DNA inputs as low as 0.75 ng (Figure 2.4). Comparison to the standard DNA pipelines showed that our approach performed consistently better when reducing the input DNA (Figures 2.4 and 2.5). Duplicate fractions negatively correlated with the number of input cells as well as post-library DNA concentration (Figures 2.6 and 2.7).

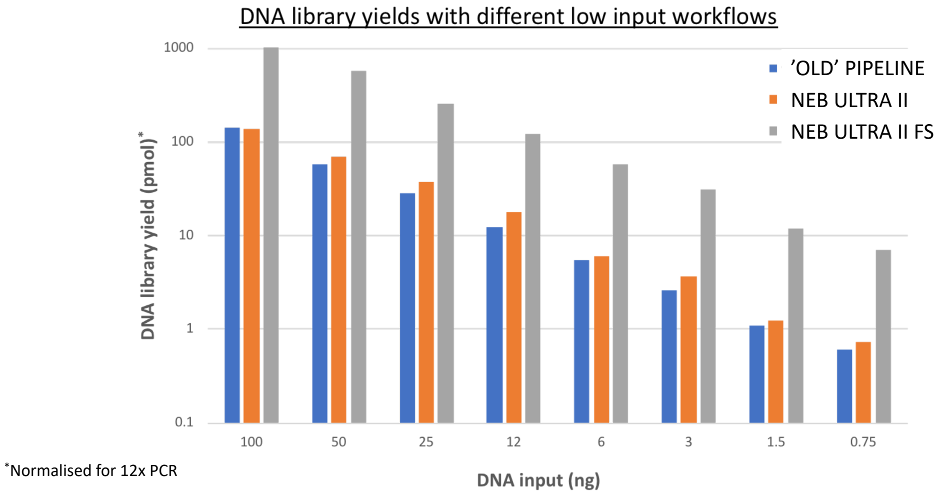


Figure 2.4 | Comparison of standard and our DNA library preparation methods. This figure shows comparison between different low input DNA workflows. Although the decrease in the input DNA inevitably leads to the decrease in the DNA library yields, our new protocol (NEB Ultra II FS) was consistently superior to the standard DNA library preparation protocols ('Old' pipeline utilises sonication in the DNA fragmentation step of library preparation; NEB Ultra II utilises the original version of the enzymatic DNA fragmentation NEB kit)**.** Adapted from Peter Ellis.

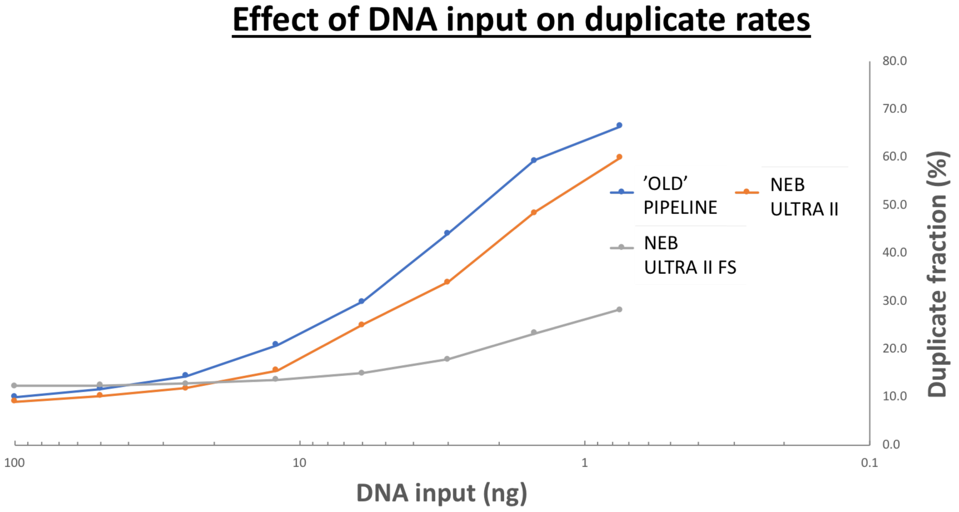


Figure 2.5 | Comparison of standard ('OLD Pipeline' and 'NEB ULTRA II') and our new approach ('NEB ULTRA II FS') for sequencing library preparation. Duplicate fractions increase with the decrease in the amount of input DNA. Although the general trend is the same with all three approaches, our new library preparation approach was superior to the previously available protocols. Adapted from Peter Ellis.

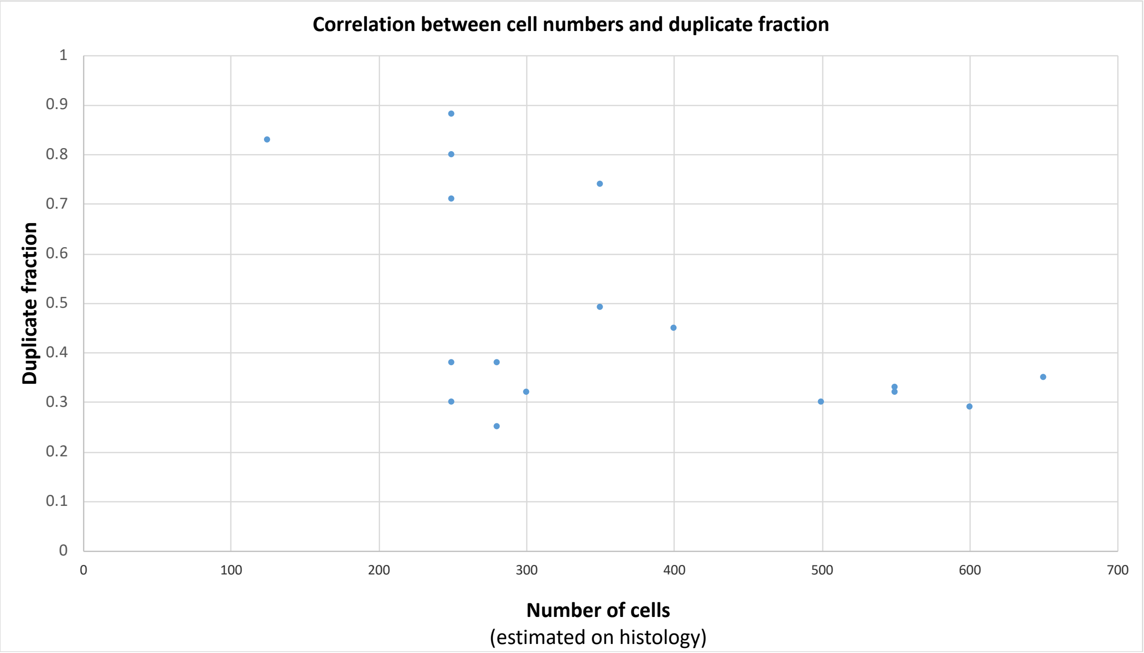


Figure 2.6 | Correlation between cell numbers and duplicate fractions. Duplicate fraction increases with the decrease in the amount of input DNA (in this case the number of laser-capture microdissected cells).

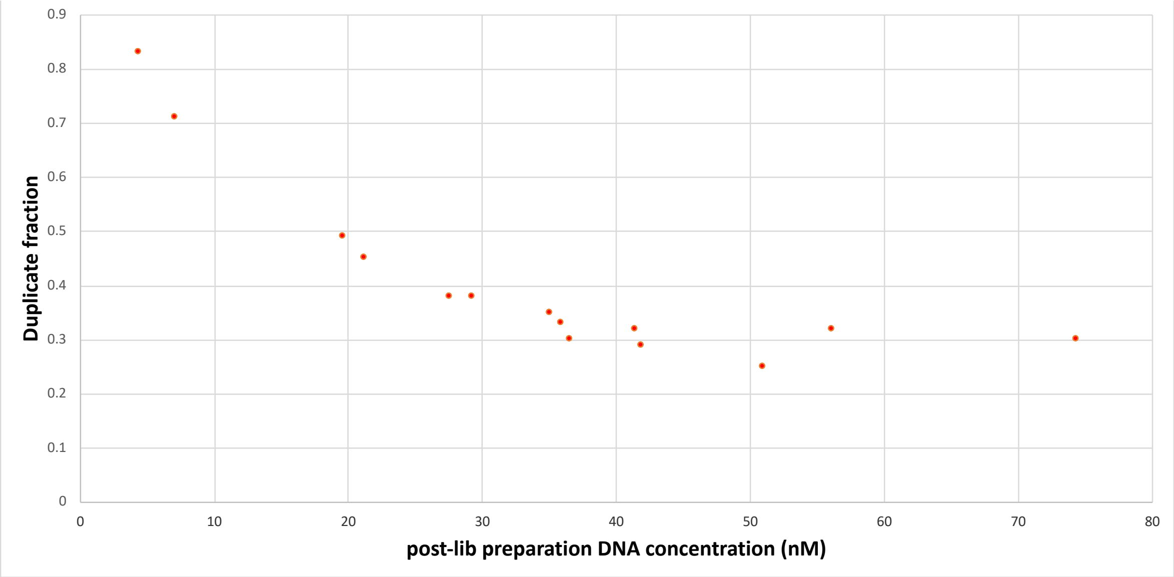


Figure 2.7 | Correlation between post-library DNA concentration and duplicate fractions. Duplicate fractions negatively correlated with post-library preparation DNA concentration.

All samples in my PhD were processed using the low-input enzymatic fragmentation-based library preparation method(Lee-Six et al., 2019). Briefly, each 20 ul LCM lysate was mixed with 50 ul Ampure XP beads (Beckman Coulter) and 50 µl TE buffer (Ambion; 10 mM Tris-HCl, 1 mM EDTA) at room temperature. Following a 5-minute binding reaction and magnetic bead separation, genomic DNA was washed twice with 75% ethanol. Beads were resuspended in 26 µl TE buffer and the bead/genomic DNA slurry was processed immediately for DNA library construction. Each sample (26 µl) was mixed with 7 µl of 5X Ultra II FS buffer, 2 µl of Ultra II FS enzyme (New England BioLabs) and incubated on a thermal cycler for 12 minutes at 37°C then 30 minutes at 65°C. Following DNA fragmentation and A-tailing, each sample was incubated for 20 minutes at 20°C with a mixture of 30 µl ligation mix and 1 µl ligation enhancer (New England BioLabs), 0.9 µl nuclease-free water (Ambion) and 0.1 µl duplexed adapters (100 uM; 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3', 5'-phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'). Adapter-ligated libraries were purified using Ampure XP beads by addition of 65 µl Ampure XP solution (Beckman Coulter) and 65 µl TE buffer (Ambion). Following elution and bead separation, DNA libraries (21.5 µl) were amplified by PCR by addition of 25 µl KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 1 µl PE1.0 primer (100 µM; 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC

TTCCGATC*T-3') and 2.5 µl iPCR-Tag (40 µM; 5'-CAAGCAGAAGACGGCATACGAGATXGAGATCG GTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC-3') where 'X' represents one of 96 unique 8-base indexes. The samples were then mixed and thermal cycled as follows: 98 °C for 5 minutes, then 12 cycles of 98 °C for 30 s, 65°C for 30 s, 72 °C for 1 minute and finally 72 °C for 5 minutes. Amplified libraries were purified using a 0.7:1 volumetric ratio of Ampure Beads (Beckman Coulter) to PCR product and eluted into 25 µl of nuclease-free water (Ambion). DNA libraries were adjusted to 2.4 nM and sequenced on the HiSeq X platform (illumina) according to the manufacturer's instructions with the exception that we used iPCRtagseq (5'-AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC-3') to read the library index.

2.3 Variant calling

2.3.1 Single nucleotide variants (SNVs)

Sequencing data were first aligned to the reference human genome (NCBI build 37) using Burrow-Wheeler Aligner (BWA-MEM) (Li and Durbin, 2009). Duplicates were marked and removed and mapping quality thresholds were set at 30. Single base somatic substitutions were called using Cancer Variants through Expectation Maximization (CaVEMan) algorithm (major copy number 5, minor copy number 2) (Nik-Zainal et al., 2012). These settings were used as they provided the most optimal balance between removing genuine variants and allowing artefacts through.

To exclude germline variants, matched normal samples were collected for each donor and used when running variant calling algorithms. For the endometrial study, we collected either cervix, myometrium, Fallopian tube or endometrial stroma; the type of tissue depended on sample source and availability. For the pan-body study, cerebellum was used as a matched normal in all three donors.

A set of previously described post-processing filters was subsequently applied:

- to remove common single nucleotide polymorphisms, variants were filtered against a panel of 75 unmatched normal samples (Nik-Zainal et al., 2012);
- to remove mapping artefacts associated with BWA-MEM, median alignment score of reads supporting a mutation should be greater than or equal to 140 (Alignment Score ‘ASMD’>=140) and fewer than half of the reads should be clipped (Clipping Score ‘CLPM’=0)(Lee-Six et al., 2019);
- to remove artefacts that are specific to the library preparation for laser capture (LCM) samples, two additional filters were used. A fragment-based filter, which is designed to remove overlapping reads resulting from relatively shorter insert sizes allowed in this protocol that can lead to double counting of variants, and a cruciform filter, which removes erroneous variants that can be introduced due to the incorrect processing of cruciform DNA. For each variant, the standard deviation (SD) and median absolute deviation (MAD) of the variant position within the read was calculated separately for

positive and negative strand reads. If a variant was supported by a low number of reads for one strand, the filtering was based on the statistics calculated from the reads derived from the other strand and it was required that either: (a) ≤ 90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, or (b) that the MAD >0 and SD>4. Where both strands were supported by sufficient reads, it was required for both strands separately to either: (a) ≤90% of supporting reads report the variant within the first 15% of the read as determined from the alignment start, (b) that the MAD>2 and SD>2, or (c) that at least one strand has fulfilled the criteria MAD>1 and SD>10.

2.3.2 Indels

Insertions and deletions were called using cgpPindel (Raine et al., 2015, Ye et al., 2009). To remove germline variants the algorithm was run with the same matched normal samples that were used for calling substitutions. Post-processing filters were applied as previously described (Nik-Zainal et al., 2012). In addition, a 'Qual' filter (the sum of the mapping qualities of the supporting reads) of at least 300 and an average sequencing depth cut-off of ≥ 15 reads were used.

2.3.3 Copy number and structural variants

Allele-specific copy number profiles were reconstructed for the endometrial gland samples by ASCAT (Van Loo et al., 2010, Raine et al., 2016) using matched samples as described above, with a ploidy of 2 and contamination with other cell types of 10%. Only samples with a minimum coverage of 15X and above were used. All putative copy number changes were visually inspected for copy number profiles on Jbrowse (Buels et al., 2016).

Structural variants (SVs) in endometrial glands were called using matched samples (as described above) with the Breakpoints Via Assembly (BRASS) algorithm and further annotated by GRASS (https://github.com/cancerit/BRASS). Potential SVs are detected for the sample of interest and read-pairs clusters supporting the SV are used for breakpoint sequence *de novo* assembly. Absence of supporting evidence in the matched control indicates that the SV was acquired in the sample of interest. The isolation of minute amounts of DNA for sequencing in combination with the LCM enzymatic fragmentation-based library preparation procedure introduces additional artefacts and additonal post-processing filtering was performed in two phases.

2.3.3.1 Further annotation of SVs with statistics that detect LCM specific artefacts

All SVs detected by BRASS were further annotated by AnnotateBRASS. Each SV is defined by two breakpoints and their genomic coordinates.

(**A**) The following statistics were determined for each breakpoint separately:

- The total number of reads supporting the SV.
- The total number of unique reads supporting the SV, based on alignment position and read orientation.
- The standard deviation of the alignment positions of reads supporting the SV.
- The number of chromosomes, based on read-pairs not supporting the SV, to which one read mapped while the mate-read aligned to the SV breakpoint.
- The number of reads supporting the SV that had an alternative alignment (XA-tag).
- The number of reads supporting the SV that had an alternative alignment score (XS-tag) similar to the current alignment score.
- The percentage of read-pairs not supporting the SV with a discordant inferred insert size (default: ≥ 1000bp).

(**B**) A wider search for read-pairs supporting the SV is initiated and the following statistics were calculated for each breakpoint separately:

- The total number of reads supporting the SV.

- The total number of unique reads supporting the SV, based on alignment position and read orientation.
- The standard deviation of the alignment positions of reads supporting the SV.
- The number of reads supporting the SV that had an alternative alignment.
- The number of reads supporting the SV that had an alternative alignment score similar to the current alignment score.

(**C**) Reads spanning the SV breakpoints are often clipped. Clipped sequences of sufficient length can be aligned to other positions on the genome (i.e., supplementary alignment) and it is expected that these align to the proximity of the other SV breakpoint. Based on the clipping positions and supplementary alignments the following was determined for each SV:

- Whether the clipped sequences of read-pairs spanning a SV breakpoint align in the proximity of the other SV breakpoint.
- Whether the clipping within read-pairs supporting the SV occurred at roughly the same genomic position (default: all clipping positions occurred within 10 bp of each other).

(**D**) BRASS uses a single matched control and a panel of normals (PoN, bulk WGS) to determine whether a SV is somatic. SVs observed in the sample of interest but not in the matched control or PoN are considered somatic. However, due to the difference in library preparation and the variance of spatial genomic coverage observed it is not always possible to accurately assess the validity of the SV. Two different approaches were implemented to determine whether the SV is somatic:

1. A wider search in the matched control sample was performed to search for read-pairs that could support the SV. The SV was still considered to be detected in cases where the discovered read-pairs were insufficient for breakpoint sequence *de novo* assembly.

2. Additional controls can be defined in case multiple samples have been isolated for the same individual. Samples from the same individual with little genetic relationship, as determined from the SNVs and indels, can be used as controls to determine whether te detected SV is germline or a recurrent artifact.

2.3.3.2 Post-hoc filtering of SVs based on a combination of the above statistics.

SVs were further filtered based on the described statistics. The optimal set of statistics and their most practical thresholds depends on the achieved coverage and stringency of filtering desired. At default the following criteria were used for detecting somatic SVs:

- For each breakpoint there must be ≥ 4 unique reads supporting the SV (**A.2**).
- The alignment position standard deviation must be > 0 (**A.3**).
- At each breakpoint there are read-pairs not supporting the SV that map to < 5 other chromosomes (**A.4**).
- The total number of chromosomes mapped to by read-pairs not supporting the SV for both breakpoints should be < 7 (**A.4**).
- The percentage of reads supporting the SV with alternative alignments or alternative alignments with similar alignment scores should be ≤ 50% for both SV breakpoints separately (**A.5-A.6**).
- The percentage of discordant read-pairs not supporting the SV should be ≤ 7.5% of total read-pairs for both SV breakpoints separately (**A.7**).
- For the wider search of SV-supporting read-pairs the same thresholds apply as under criteria 1-6 (**B.1-B.5**).
- There are no read-pairs in the matched control that support the SV (**C.1**).

- The SV is not detected in any of the other control samples, or there were ≤ 2 samples carrying the same SV and the proportion of control samples carrying the SV was < 1/3 of the defined control set (**C.2**).
- It was not allowed for read-pairs supporting the SV to have widely divergent clipping positions in terms of genomic location for both SV breakpoints separately (**D.2**).

2.3.4 Validation experiments and sensitivity

To test our approach, we performed a set of validation experiments in different tissue types. First, the reproducibility of the workflow was assessed by generating pairs of biological 'near-replicate' samples and processing them independently using the new library construction methodology. In these experiments, two separate samples were generated from the same tissue structure, such as an appendiceal crypt, and subjected to independent DNA extraction, cell lysis, library preparation and WGS (Figure 2.8a-d). Subsequent analysis of the sequencing data showed similar variant allele frequency (VAF) distributions (Figure 2.8b), a high degree of overlap for single nucleotide variants (SNVs) (Figure 2.8c), and similar single base substitution mutational spectra (Figure 2.8d).

We then compared WGS data generated by our new workflow to LCM lysates processed via traditional acoustic shearing methods. Similarly, pairs of biological 'near-replicate' samples were derived from the same histological structure; this time, one sample was processed with our new workflow and the other with acoustic shearing. Again, comparison of the WGS data between the two, differently processed, samples showed similar VAF distributions, SNVs and mutational spectra (Figure 2.8e-h).

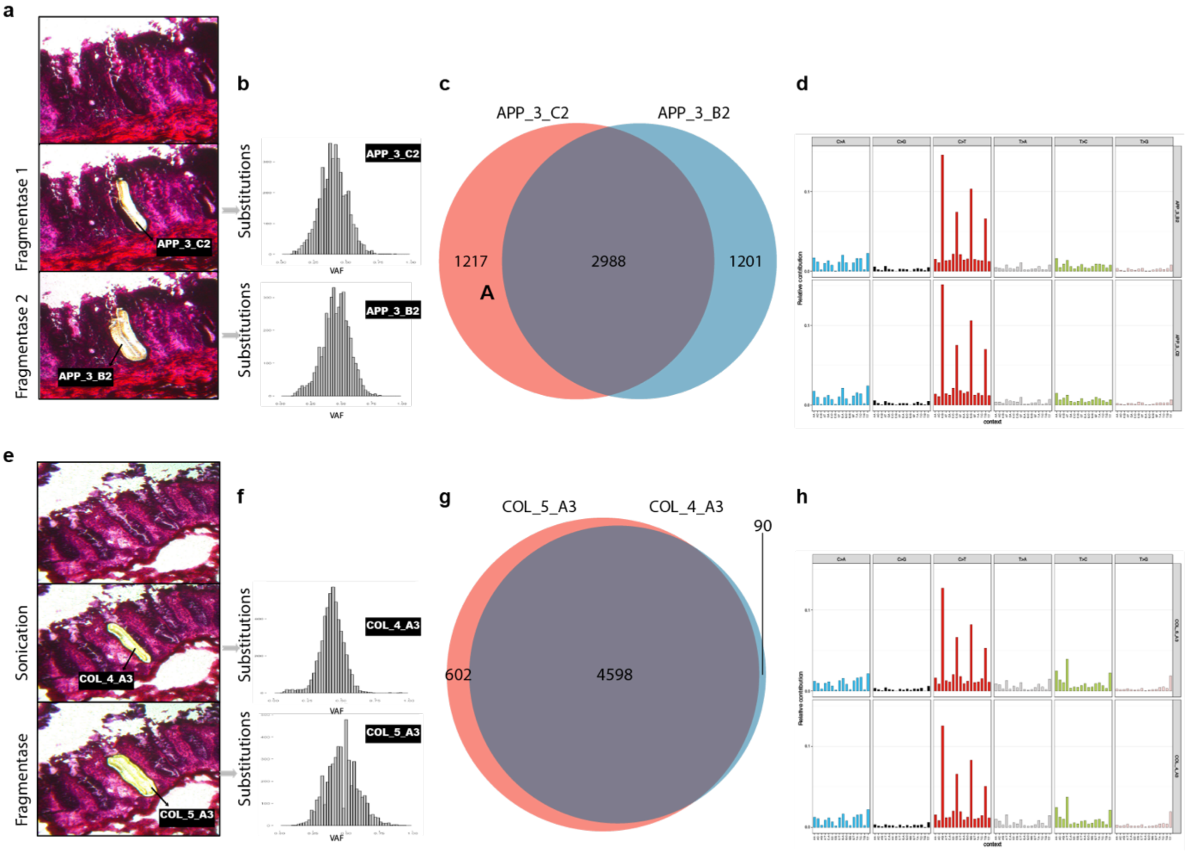


Figure 2.8 | Validation experiments sequencing 'near-replicate' samples. a-d, 'Near-replicate' samples were generated by splitting an appendiceal crypt into two halves, which were then processed and sequenced independently. **b,** VAF of all substitutions in both halves show similar clonal distribution with a median VAF around 0.5. **c,** Venn diagram demonstrating SNV identity between both samples. **d,** Mutational spectra of all substitutions are also similar. **e-h,** 'Near-replicate' samples were generated by splitting a colonic crypt into two halves, which were subsequently processed with our fragmentase-based method (COL_5_A3) and sonication-based method (COL_4_A3). Similar clonal VAF distributions **(f)** SNV calls **(g)** and mutational spectra **(h)** are observed from the two samples.

To calculate sensitivity of our somatic variant calling, for selected endometrial tissue donors, pairs of biological 'near-replicates' were obtained. For these experiments, we collected two samples from the same endometrial gland using a z-stacking approach, in which a structure is 'traced' on consecutive levels (Figure 2.9).

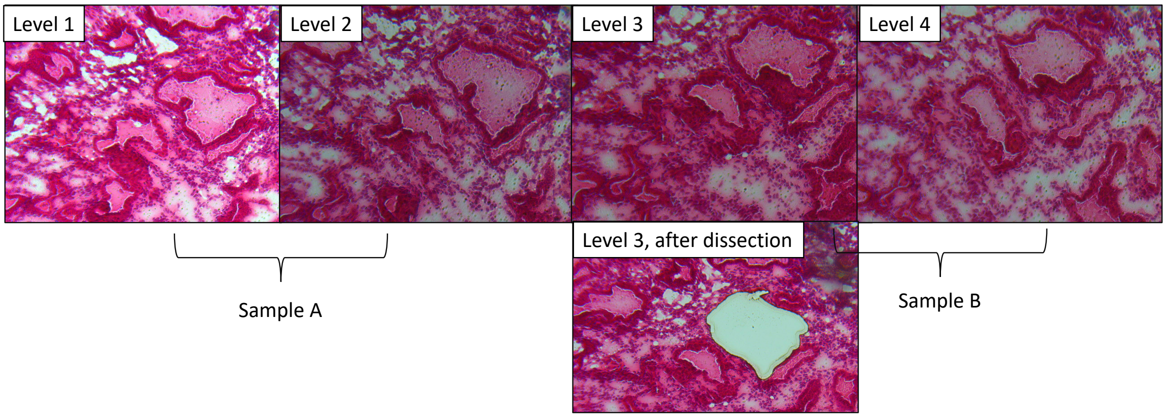


Figure 2.9| An example of a z-stacking approach to 'tracing' and micro-dissecting a specific structure. Sincer the majority of endometrial glands are clonal cell populations, i.e. share the most common recent ancestor, cells derived from the same glands should share most of the somatic mutations. Z-stacking and splitting individual glands into two separate samples allows to generate biological 'near-replicates' that can be used to generate biological 'near-replicates' to calculate sensitivity.

Each sample was then processed separately with independent DNA extraction, library preparation and whole genome sequencing. As these were obtained from the same glands, they should represent derivatives of the same single stem cell and therefore the same sensitivity would be expected in both samples of each pair. The maximum likelihood estimate for sensitivity (*s*) was then calculated as follows:

$$\boxed{S = \frac{2n_2}{n_1 + 2n_2}}$$

where n_1 is the number of variants called only in one of the two LCM samples and n_2 is the number of variants called in both LCM samples in each pair.

2.4 Initial application of the low DNA input LCM workflow

The first part of my PhD was dedicated to exploring somatic mutations across multiple tissues from the same individuals. This study is still ongoing with sequencing data pending from two additional donors (donor 2 and 3). However, the endometrial study stemmed from those initial experiments, and a brief summary is therefore provided below.

2.4.1 Samples

By November 2017, I micro-dissected over 2,900 individual samples from 13 individuals, although the majority of the samples were from one individual (Donor 1) (Figure 2.10). Based on the post-library preparation DNA concentration (a cut off of minimum 3-5 ng/ul was applied), a total of 421 samples were subjected to whole genome sequencing. Only samples with ≥15-fold coverage were processed through the variant calling pipeline (n=225, Appendix 6).

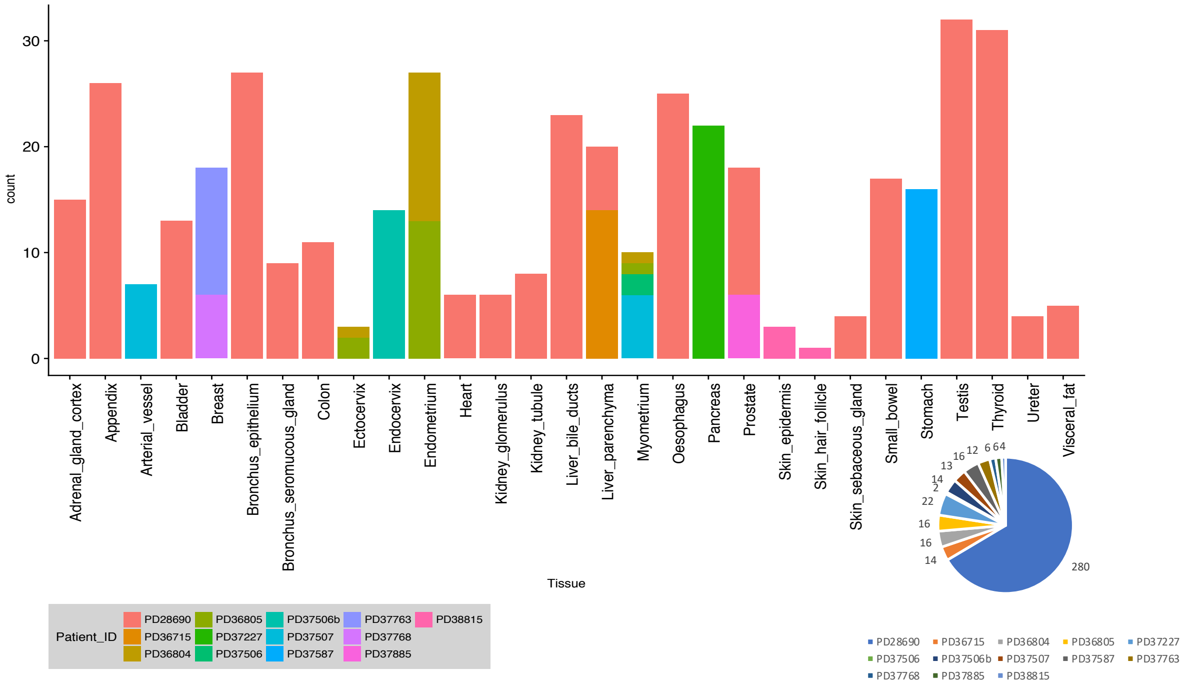


Figure 2.10 | Summary of samples sequenced as part of the initial pan-body survey. A total of 421 samples were micro-dissected from 13 individuals.

2.4.2 Clonality

The clonal architectures of human tissues have been investigated previously by other approaches, in particular there has been a series of studies that utilised mitochondrial DNA (mtDNA) mutations (Fellous et al., 2009a, Blackwood et al., 2011). These have provided evidence for the clonal expansion in colon, small intestine, kidney, pancreas and others. The analyses presented here illustrate the potential of DNA sequence-based approaches to further elucidate tissue architecture and cell lineages providing systematic comparisons of the different clonal architectures of normal human tissues and their microanatomical structures.

Micro-dissected units of cells from different tissues showed markedly different VAF distributions (Figures 2.10 and Table 2.1). 34% (77/224) of all the sampled units, including individual colorectal, appendiceal, small intestinal, prostatic, endometrial crypts or glands showed distributions with peaks between 0.3-0.5 (Figures 2.11). Thus, these cell populations are predominantly constituted of the descendants of a single progenitor stem cell (the most recent common ancestor cell, MRCA) which existed at some point in the past.

Similar VAF distributions of 0.4-0.5 were observed in subsets of microdissected patches from seminiferous tubules, bile ductules, thyroid follicles and segments of bronchial epithelium indicating that these were also predominantly derived from single MRCA cells (Figure 2.10 and Table 2.1). However, other samples from these tissues showed lower VAF peaks indicating the presence of clones derived from multiple MRCA cells. All microdissected patches from oesophagus, bladder, adrenal and adipose tissue showed low median VAFs. Micro-dissections from cardiac muscle and an arterial vessel yielded very few somatic mutations, consistent with these tissues being non-renewing in the adult and/or being composed of so many clones that none achieve the level of clonal dominance required for calling of somatic mutations.

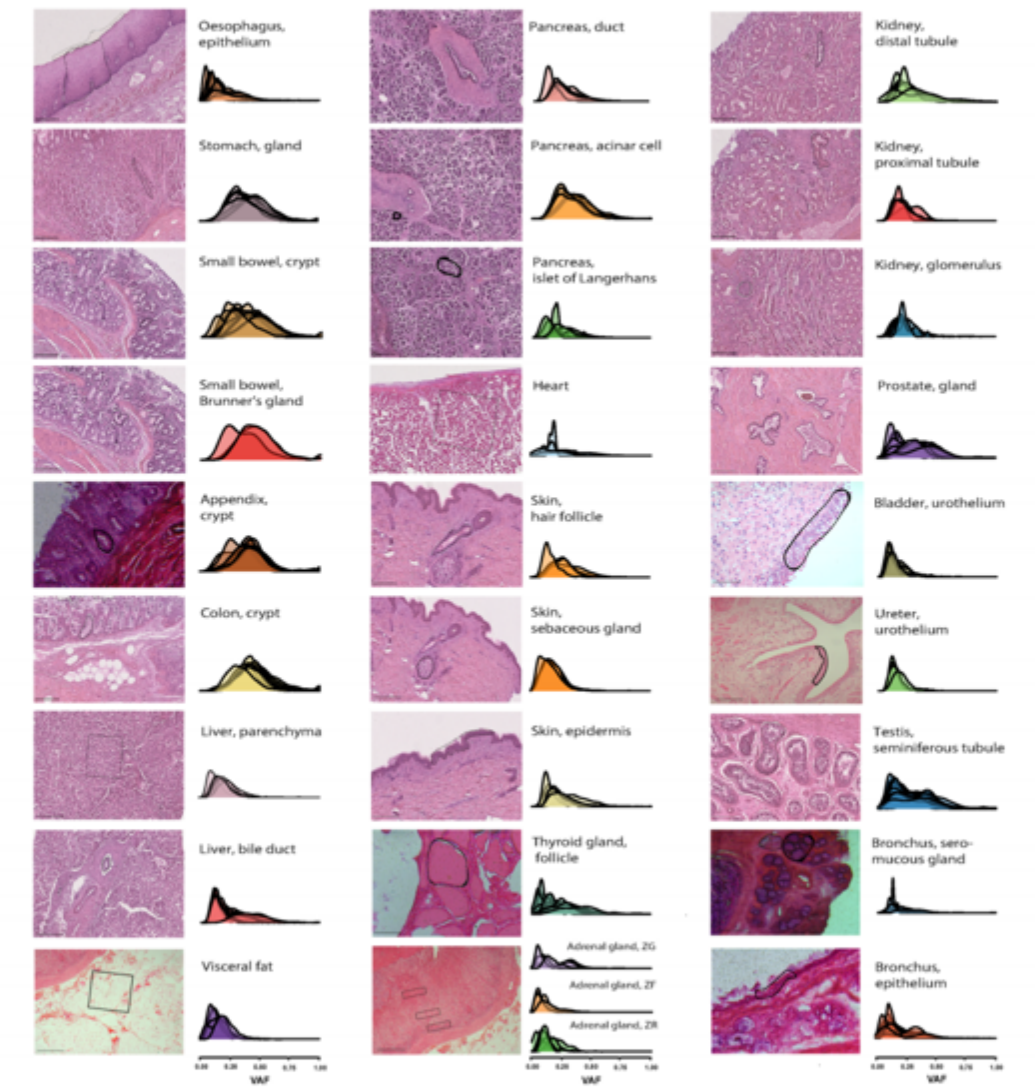


Figure 2.11 | Clonality of some of the sampled microscopic units. To study clonal composition across various normal tissues, laser-capture microdissection and whole genome sequencing were applied in two ways. In some tissues, previously described or putative clonal units such as crypts in the colon and small intestine were targeted. In other cases, including the ectocervix and adrenal gland cortex, variably sized strips or patches of cells were microdissected. Clonal composition of the sampled microscopic units can be studied using variant allele fractions (VAF) of all single base substitutions(Keller et al., 2008, Blokzijl et al., 2016). Each density line represents an individual sample; individual samples are grouped and coloured by tissue type. Samples derived from a clonal population, i.e. sharing the most common recent ancestor, will have VAF peaks around 0.5 as the majority of somatic mutations are heterozygous (e.g. colonic crypts). However, even with LCM approach, there might a contamination with other cells types, such as stromal or inflammatory cells, which would result in a left-sided shift in the density plots. If a sample was oligoclonal, i.e. derived from a few ancestral clones, this would result in an additional VAF peak (e.g. seminiferous tubules in testis). Polyclonal samples are those derived from many different ancestral clones (including different cell types); their VAF distribution will be generally < 0.25.

Tissue/structure	Fraction of clonal samples (%)
Appendix, crypt	100 (20/20)
Colon, crypt	100 (20/20)
Small intestine, jejunum, crypt	89 (8/9)
Small intestine, ileum, crypt	100 (7/7)
Prostate, acini	83 (10/12)
Testis, seminiferous tubules	50 (7/14)
Liver, bile ductules	26 (5/19)
Thyroid, follicle	19 (6/31)
Adrenal gland, cortex	<1 (1/15)
Lung, respiratory epithelium	<1 (1/13)
Oesophagus, squamous epithelium	<1 (1/15)
Bladder, urothelium	0 (0/7)
Kidney, glomerulus	0 (0/4)
Kidney, proximal tubule	0 (0/4)
Kidney, distal tubule	0 (0/6)
Liver, parenchyma	0 (0/3)
Main bronchus, seromucous glands	0 (0/6)
Ureter	0 (0/4)
Visceral fat	0 (0/5)
Skin, sebaceous glands	0 (0/3)
Heart	0 (0/6)
Artery	0 (0/1)

Table 2.1 | Clonality of some of the sampled microscopic units. Different microscopic units were dissected out in different tissues, including individual crypts in the small and large intestines or acini in the prostate. Samples were considered clonal if the median variant allele fraction (VAF) was >=0.3 as previously described (Blokzijl et al, 2016).

2.4.3 Burden

Although estimation of mutation burden from the data in the initial pan-body experiments is complicated by differences in clonality and sequencing coverage between samples, two key observations were made. First, the results showed inter-tissue heterogeneity in the mutation burden within the same individual (Figure 2.12a); tissues of the same chronological age demonstrated different mutation burdens. The findings are likely to be reflective of the differences in physiology, function and exposures as well as stem cell dynamics and turnover rates. Second, although at this stage we only had sequencing data from a very limited number of individuals, there was an age-associated accumulation of somatic mutations in prostate and endometrium (Figure 2.12b and c).

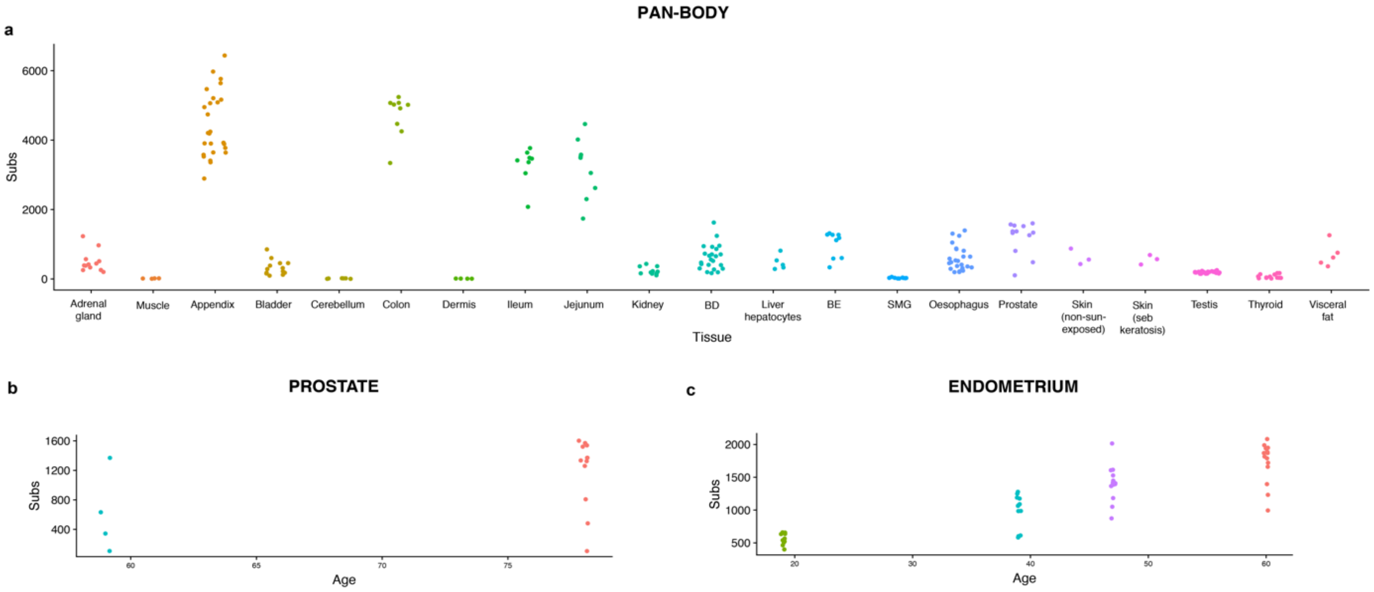


Figure 2.12 | Somatic mutation burden (SNVs). (a) This figure shows mutaiton burden (SNVs) across different tissues derived from one indvidual (Donor 1). BD, bile ductules, SMG, sero-mucous glands, BE, bronchial epithelium. Initial experiments in prostate **(b)** and endometrium **(c)** showed age-associated accumulation of somatic mutations.

2.4.4 Drivers

Filtered CaVEMan and Pindel variants were intersected against a previously published list of 369 genes that are under selection in human cancers (Martincorena et al., 2017). All non-synonymous mutations were annotated to indicate mode of action using the Cancer Gene Census (719 genes) and a catalogue of 764 genes (https://www.cancergenomeinterpreter.org). Variants were triaged against a curated list of 5601 validated cancer driver variants (https://www.cancergenomeinterpreter.org/mutations). Any variant in the sample data which co-presented in this reference list was declared a likely driver. The initial results showed that endometrium had the highest prevalence of driver mutations compare to other tissues (Figure 2.13).

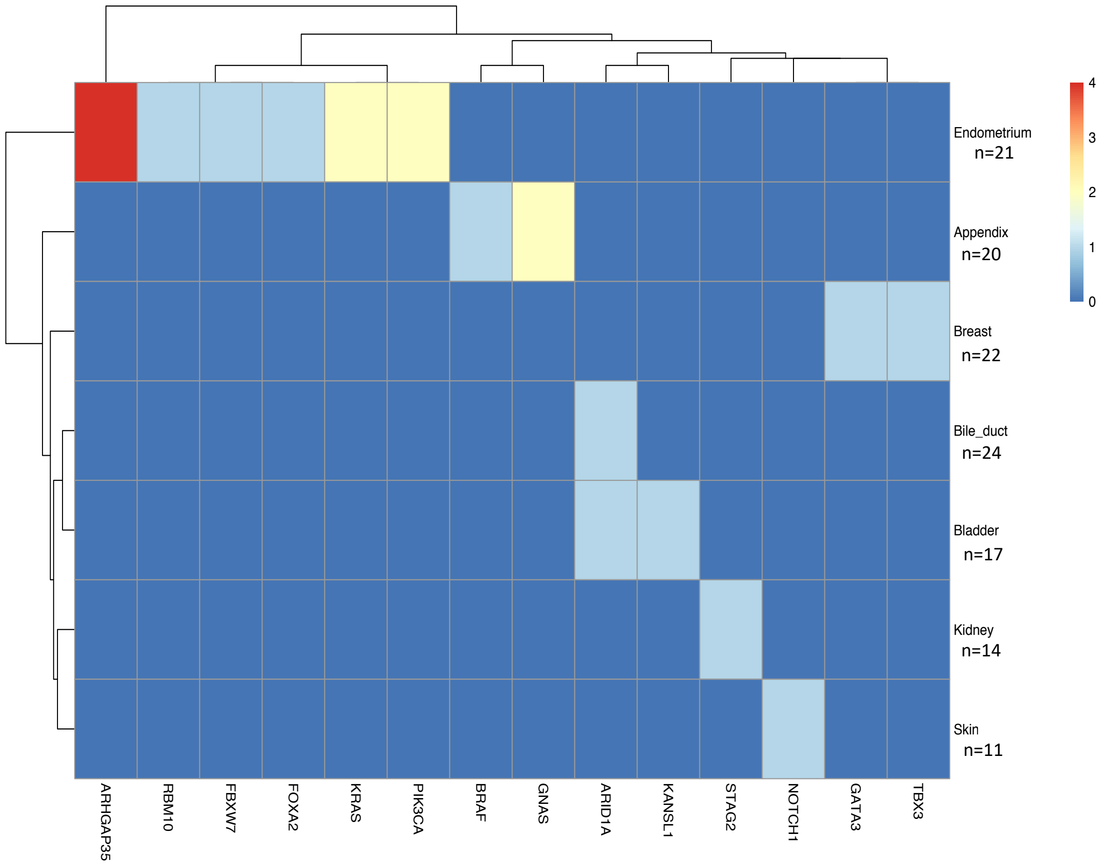


Figure 2.13 | Driver mutation burden across tissues. Although the number of samples studied in the pilot experiments varied between tissues, our first impression was that the endometrium had the highest number of driver variants. This was an unexpected finding which led to the more comprehensive study on the normal endometrium.

2.4.5 Summary of the initial experiments

Preliminary results from the very first pan-body experiments have provided first insights into inter-tissue heterogeneity in terms of somatic mutation burden and clonal expansion. One of the most striking observations was that the majority of the sampled endometrial glands were clonal cell populations and had the highest frequency of driver variants. The latter was particularly surprising given that the these events occur at a much lower frequency in other normal tissues with gland-like structures, such as colon and prostate, yet the documented cancer incidence is greater than reported in the endometrium (CRUK, 2019) that are associated with a higher cancer incidence rates. We therefore decided to carry out a more in-depth analysis of the genomic landscape of normal endometrium to find out how age as well as other known endometrial cancer risk factors affect the rate of (driver) mutation acquisition. The results of this work are discussed in Chapters 3 and 4.

2.5 Construction of phylogenies

As we obtained multiple samples from the same individuals, we needed to differentiate between shared and unique variants to avoid double counting. Phylogenetic trees were therefore reconstructed for individual patients.

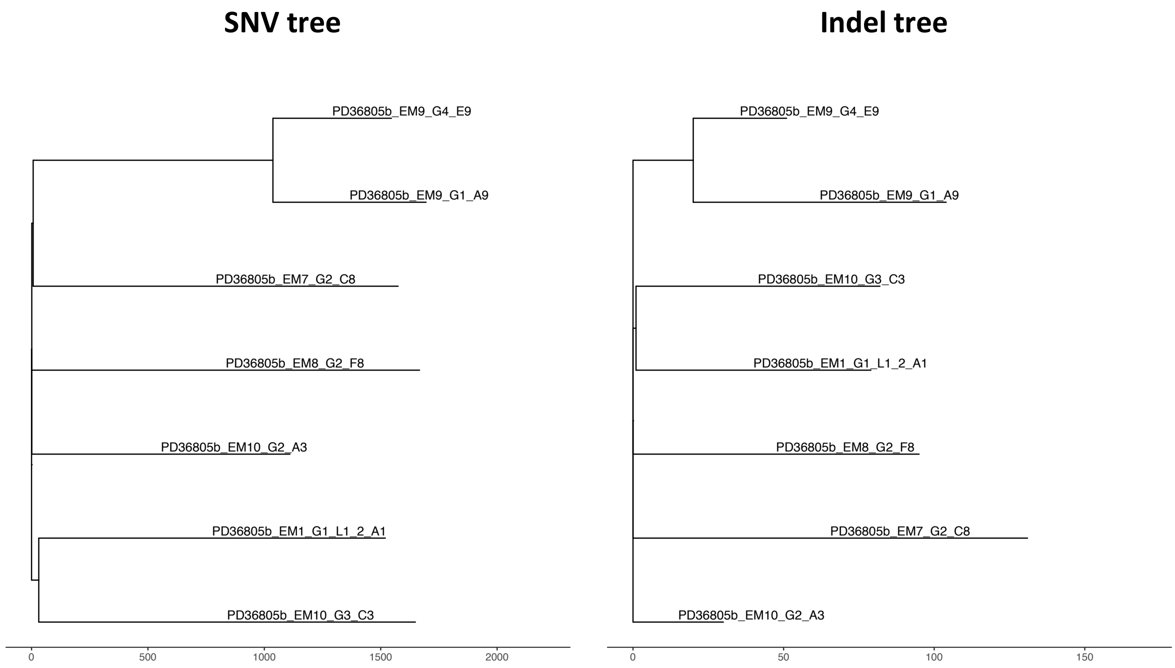
2.5.1 Single nucleotide variants (SNVs)

Phylogenies for endometrial glands were reconstructed for twenty five donors. Due to the low number of available samples, donor PD38812 was not included in this analysis. We first generated trees using substitutions called by CaVEMan; matched normal samples were used to exclude germline variants and post-processing filters were applied as above. Final variants were recalled in all samples from each donor using an in-house re-genotyping algorithm (cgpVAF). Variants with a VAF>0.3 were noted to be present ('1'), VAF<0.1 absent ('0') and between 0.1 and 0.3 as ambiguous ('?'). This approach excludes private sub-clonal variants from the tree building. The tree was reconstructed using a maximum parsimony approach (Hoang et al., 2018) and branch support was calculated using 1000 bootstrap replicates. Nodes with a confidence lower than 50 were collapsed into polytomies and branch lengths of the collapsed tree were determined by the number of assigned substitutions.

2.5.2 Small insertions and deletions (indels)

The constructed phylogenies were validated using indels called by Pindel and filtered as above. The same approach was applied for the final indel matrices. Although the lower number of indels resulted in more polytomous tree, the overall tree topologies were reconcilable with those generated using substitutions (Figure 2.14).

Cancer driver mutations, copy number and structural variants were annotated manually in the trees.



SNV tree

Indel tree

Figure 2.14| Comparison of phylogenetic tree structure using SNVs and indels of endometrial glands obtained from the same donor (PD36805).

2.6 Assessment of clonality

2.6.1 dpClust

To formally assess clonal composition of individual endometrial glands, we applied a previously described method dpclust v2.2.7 (Nik-Zainal et al., 2012) (analysis was performed by Stefan Dentro). This sub-clonal reconstruction caller with default parameters to the SNVs in each endometrial gland to assess the clonality of each gland. SNVs that fell within a detected copy number alteration were excluded from this analysis. The purity of each gland was set to 1, the resulting mutation clusters therefore represent proportions of the overall sequenced cells. Analysis yields, for every sample, the number of mutation clusters and assigned mutations, and the proportion of overall cells that each cluster represents.

2.6.2 PyClone

PyClone is a clustering method that is based on a hierarchical Bayes statistical model (Roth et al., 2014). It was developed for deep (1,000x) targeted sequencing data from one or more samples from the same tumour. The method assigns mutations to putative clonal clusters while also estimating their cellular prevalence and correcting for allelic imbalances, which can result from segmental copy number aberrations as well as contamination with normal cells. We attempted to use this method as an alternative way to infer clonal composition of endometrial glands.

2.6.3 Lichee

Another computational method that utilises single nucleotide variants to infer sub-clonal composition of samples while allowing simultaneous reconstruction of multi-sample cell lineage trees (in our study, per donor lineage) is LICHeE (Lineage Inference for Cancer Heterogeneity and Evolution) (Popic et al., 2015). This approach relies on VAFs of deep-sequenced somatic SNVs. The algorithm was run with default settings: distance between clusters was 0.15, minimum VAF for a mutation to be present was 0.15, maximum VAF for a cluster was 0.65, and a VAF measurement error of 0.10.

2.7 Extraction of mutational signatures

Mutational signature extraction was performed using mutations assigned to every branch of the reconstructed phylogenetic trees and each branch was treated as an individual sample. Such approach allows characterisation and differentiation of specific mutational processes that were operative at various times in individual glands. Substitutions were first categorised into 96 classes following the method used by the Mutational Signature working group of the Pan Cancer Analysis of Whole Genomes (PCAWG) (Alexandrov, 2018). SBS signature analysis was performed in 3 steps: extraction, deconvolution and re-attribution. SBS signatures were extracted using 3 approaches: (i) using the HDP package (https://github.com/nicolaroberts/hdp) that utilises hierarchical Bayesian Dirichlet process either *de novo* or (ii) with reference signatures ('priors') identified by the Mutational Signatures working group of the Pan Cancer Analysis of Whole Genomes (PCAWG) (Alexandrov, 2018), and (iii) non-negative matrix factorization (NMF) (Alexandrov, 2018). Such extensive mutational signature analysis was performed for two reasons: (1) to validate signatures as NMF was originally developed for cancer tissues which usually provide many more mutations than normal/non-cancer tissue samples; (2) to ensure we do not miss any new mutational signatures that are unique to normal tissues. We chose to perform de novo signature extraction as mutational signatures had not been previously described in the normal endometrium. Furthermore, the so-called 'known' signatures or priors were derived using cancer sequencing data. Simply fitting mutations to a cancer derived catalogue of signatures could potentially 'over-fit' certain signatures.

2.7.1 HDP

(i) HDP *de novo* signature extraction revealed 3 components (Components 1, 2 and 0, Figure 2.14); similarity of the components to the 65 reference signatures was assessed; Component 2 had a high Cosine Similarity (>0.95) to SBS 18. (ii) HDP signature extraction with all 65 PCAWG priors yielded the following components: 'priors'/reference SBS signatures (P1 = SBS1, P5 = SBS5, P18 = SBS18, P23 = SBS23, P40 = SBS40); 'new' component that did not match any of the provided 65 reference signatures/priors (N1) and 'Component 0' (Comp 0); all of the components from this extraction were taken to further analysis and deconvolution (Figure 2.16). Because P1, P5, P18, P23 and P40 showed high Cosine Similarity (>0.95) to the respective signatures (SBS1, SBS5, SBS18, SBS23 and SBS40), no further deconvolution of these components was required. As component N1 did not show high Cosine Similarity to any of the reference signatures, deconvolution was performed using a 'deconvolution' catalogue comprising all of the extracted signatures (SBS1, SBS5, SBS18, SBS23, SBS40). Final exposures were derived and signatures re-attributed to the individual samples (branches). As SBS5 and SBS40 are relatively featureless and present particular challenges in estimating their separate contributions (as previously outlined (Alexandrov, 2018)), these have therefore been combined (but are shown separately in Appendix 7). SBS23 was previously found in a small number of liver cancers with high mutation burdens. Given its low mutation burden and small contribution in our cohort it is unclear whether this is really. Therefore, this signature and the associated mutations were placed in the "unattributed" category (Figure 2.16).

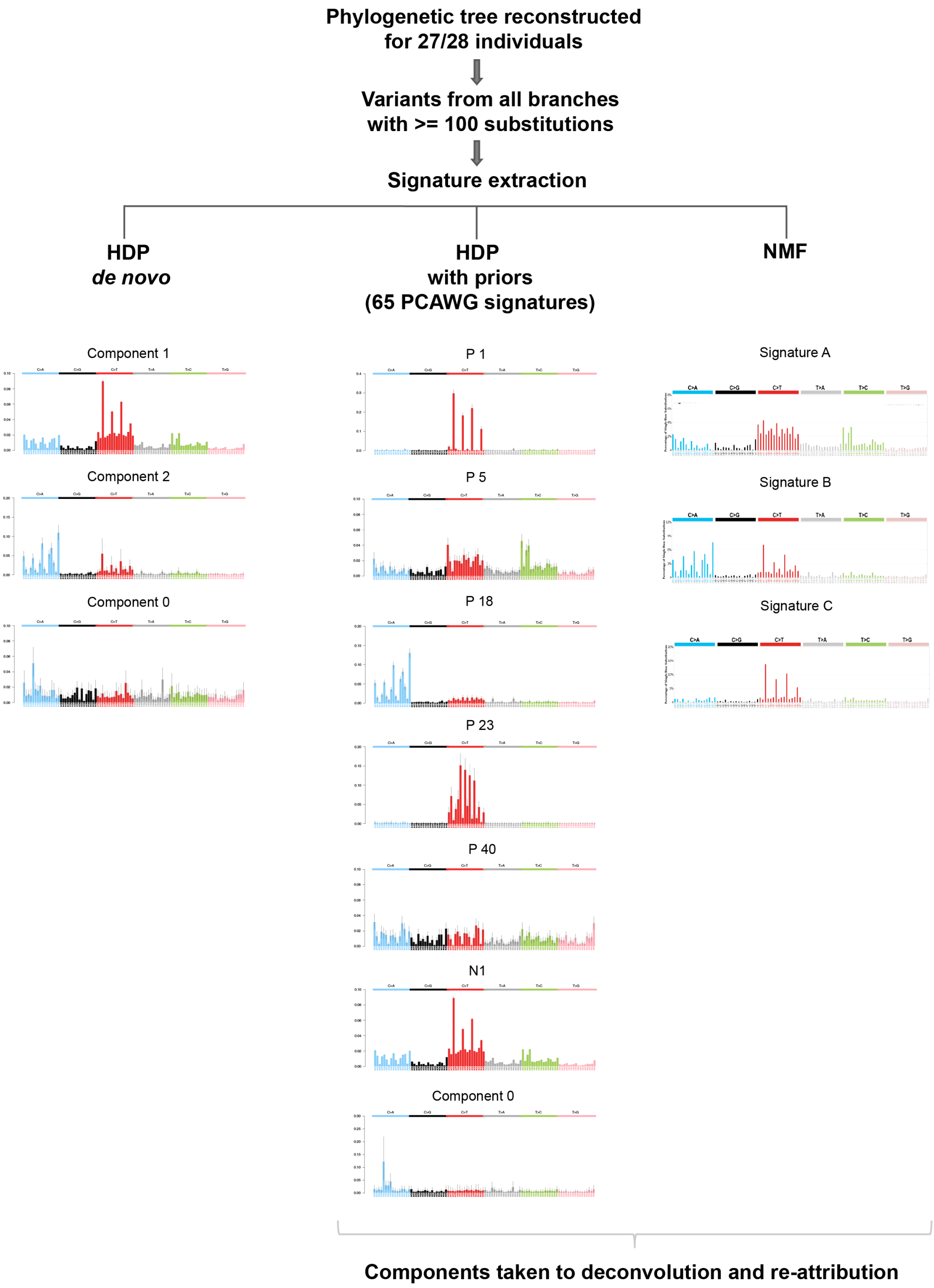


Figure 2.15 | Extraction of Single Base Substitution (SBS) mutational signatures. Final catalogue of single base substitutions were used to re-construct phylogenetic trees for 27 donors. SBS signatures were extracted on a per branch basis first using Hierarchical Dirichlet Process (HDP) *de novo*. HDP *de novo* signature extraction revealed 3 components; similarity of the extracted components to the 65 reference signatures was assessed; only Component 2 had a high Cosine Similarity (>0.95) to a reference signature (SBS 18). Signature extraction methods are continuously being developed and modified. We therefore applied different approaches. If a tissue is relatively homogenous in terms of type mutational processes, this can lead to a weaker signal and some of the components (signatures) not separating. HDP conditioning with priors (or the known signatures) can be used to aid the extraction. However, such approach can also result in a small number of variants falsly attributed to signatures that are not really there ('over-splitting').

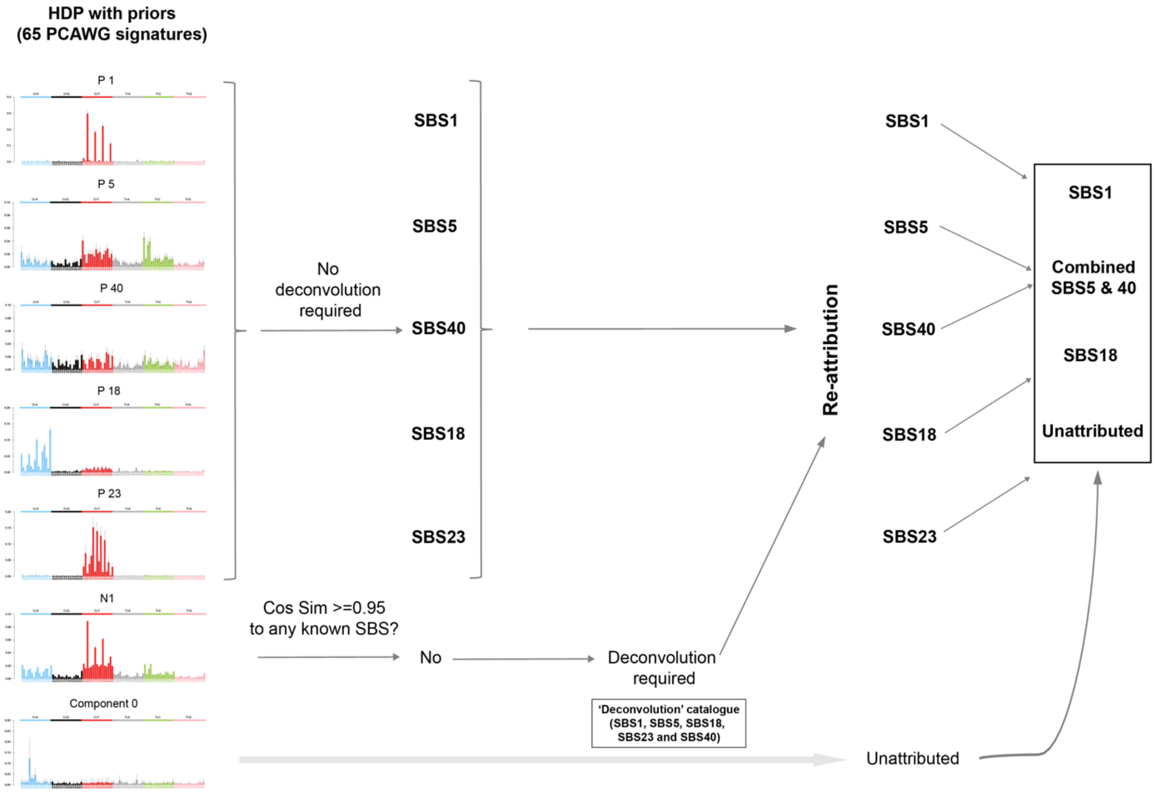


Figure 2.16 | Extraction of Single Base Substitution (SBS) mutational signatures. As P1, P5, P18, P23 and P40 showed high Cosine Similarity (>0.95) to the respective signatures (SBS1, SBS5, SBS18, SBS23 and SBS40), no further deconvolution of these components was required. Because component N1 did not show high Cosine Similarity to any of the reference signatures, deconvolution was performed using a 'deconvolution' catalogue comprising all of the extracted signatures (SBS1, SBS5, SBS18, SBS23, SBS40). Final exposures were derived and signatures re-attributed to the individual samples (branches). As SBS5 and SBS40 are relatively featureless and present particular challenges in estimating their separate contributions, these have therefore been combined (but are shown separately in Supplementary Fig 5). SBS23 was previously found in a small number of liver cancers with high mutation burdens. Given its low mutation burden and small contribution in our cohort it is unclear whether this is really. Therefore, this signature and the associated mutations were placed in the "unattributed" category

2.7.2 NMF

(iii) NMF signature extraction was performed using SigprofilerExtractor Version 0.0.5.51 (https://pypi.org/project/sigproextractor/#history), SigprofilerMatrixGenerator Version 1.0.2 (https://pypi.org/project/SigProfilerMatrixGenerator/#history) and SigprofilerPlotting Version 1.0.3 (https://pypi.org/project/sigProfilerPlotting/) on solutions between 1 and 20 signatures with 3 signatures chosen as the optimal solution running 1000 iterations. The extraction yielded 3 signatures, which were further deconvoluted as following: Signature A into SBS1 (8.16%), SBS5 (79.88%) and SBS23 (11.96%); Signature B into SBS1 (16.18%), SBS5 (22.6%) and SBS18 (61.22%); Signature C into SBS1(42.1%) and SBS5(57.9%) (Figure 2.17).

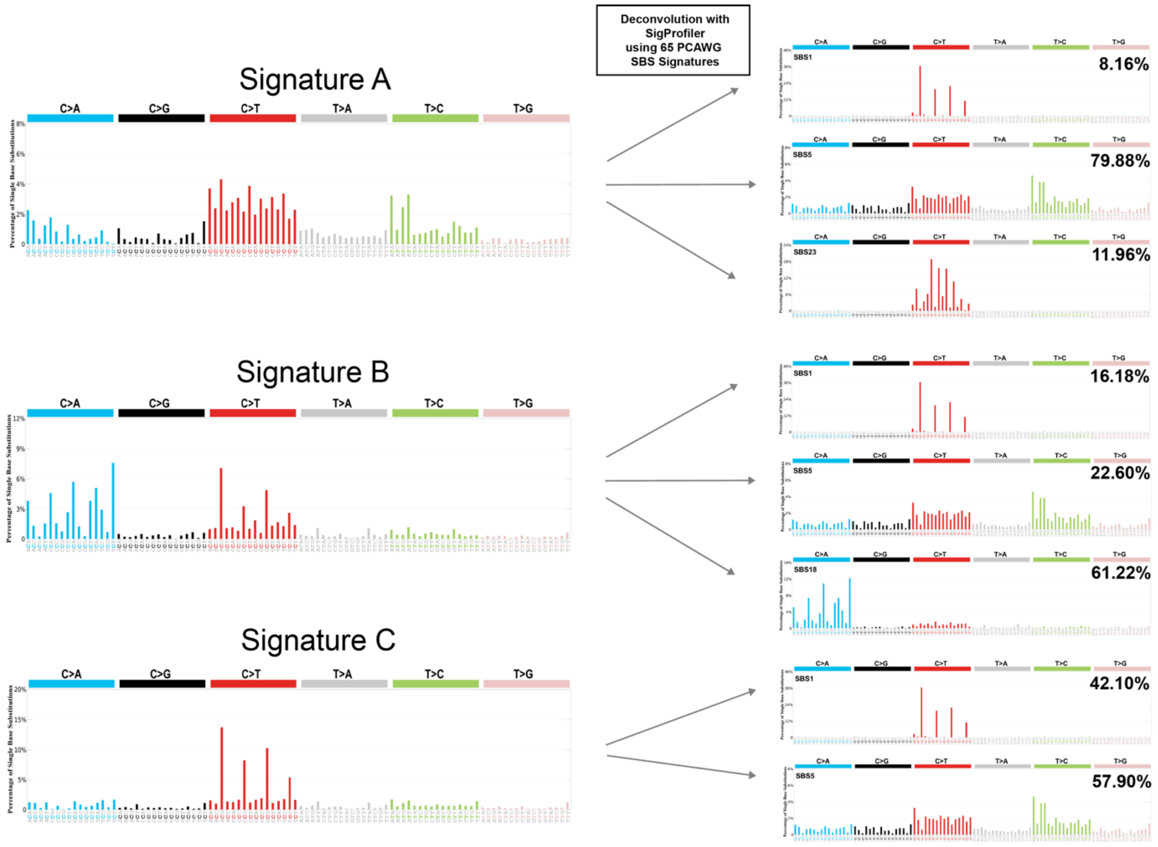


Figure 2.17 | Extraction of Single Base Substitution (SBS) mutational signatures. NMF extraction yielded 3 signatures, which were also taken to further analysis and deconvolution (**c**). Using Sigprofiler Version 1.8 (ref), Signature A was deconvoluted into SBS1 (8.16%), SBS5 (79.88%) and SBS23 (11.96%); Signature B into SBS1 (16.18%), SBS5 (22.6%) and SBS18 (61.22%); Signature C into SBS1 (42.1%) and SBS5 (57.9%).

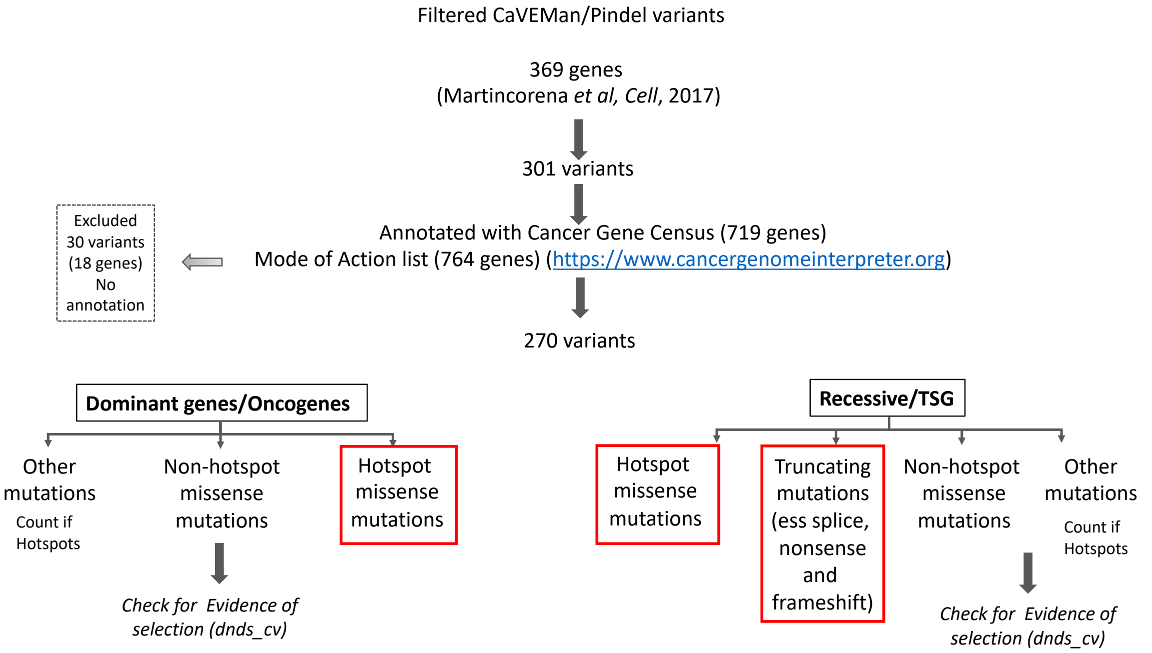
Indels were classified using PCAWG method (Alexandrov, 2018) and composite mutational spectra were generated for each donor (Appendix 8). However, given the relatively low numbers of indels, no formal signature extraction was performed.

2.8 Driver mutations

Analysis of driver variants in the normal endometrial glands was performed in two parts. First, filtered CaVEMan and Pindel variants were intersected against a previously published list of 369 genes that are under selection in human cancers (Martincorena et al., 2017). All non-synonymous mutations were annotated to indicate mode of action using the Cancer Gene Census (719 genes) and a catalogue of 764 genes (https://www.cancergenomeinterpreter.org). Truncating variants (nonsense, frameshift and essential splice), which resided in recessive/tumour-suppressor genes (TSG) were declared likely drivers. Missense mutations in recessive/TSG and dominant/oncogenes were triaged against a database of validated hotspot mutations (http://www.cbioportal.org/mutation_mapper). All mutations that were shown to be known mutational hotspots or 'likely oncogenic' were declared drivers. In addition, identified activating mutations in mutational hotspots in *RRAS2*, involving the RAS/MAPK pathway were declared as likely drivers.

2.8.1 dN/dS

Second, to identify genes that are under positive selection in normal endometrium we used the dN/dS (Martincorena et al., 2017) method that is based on the observed:expected ratios of non-synonymous:synonymous mutations. The analysis was carried out for the whole genome ($q<0.01$ and $q<0.001$) and for 369 known cancer genes (Martincorena et al., 2017) (RHT, restricted hypothesis testing, $q<0.05$). Twelve genes were found to be under positive selection in normal endometrial glands. The output of this analysis was also used to assess whether missense mutations in genes that are under positive selection in normal and/or malignant endometrium (*PIK3CA, ERBB2, ERBB3, FBXW7* and *CHD4*) but are not known mutational hotspots, are likely to be drivers. We calculated the fraction of the mutations tested that are likely to be drivers (f) using the following equation: $f = (w-1)/w$, where w is the observed missense count (52) divided by the expected count (0.14). If f was ≥ 0.95, then all missense mutations in that gene were declared likely drivers.



Filtered CaVEMan/Pindel variants
369 genes
(Martincorena et al, Cell, 2017)
301 variants
Annotated with Cancer Gene Census (719 genes)
Mode of Action list (764 genes) (https://www.cancergenomeinterpreter.org)
Excluded 30 variants (18 genes) No annotation
270 variants
Dominant genes/Oncogenes
Other mutations
Count if Hotspots
Non-hotspot missense mutations
Check for Evidence of selection (dnds_cv)
Hotspot missense mutations
Recessive/TSG
Hotspot missense mutations
Truncating mutations (ess splice, nonsense and frameshift)
Non-hotspot missense mutations
Check for Evidence of selection (dnds_cv)
Other mutations
Count if Hotspots

To compare patterns of selection in normal endometrial epithelium and cancer, we performed *dN/dS* analysis on previously published data from the The Cancer Genome Atlas, TCGA (Martincorena, 2018).

2.8.2 Timing of cancer driver mutations

To estimate the time interval in which specific driver mutations occurred, we applied two approaches: (a) ‘patient-based’, in which we calculated a patient-specific mutation rate by taking the ratio of the patient’s mean mutation burden per endometrial gland and the patient’s age; (b) ‘cohort-based’, in which mutation rate for each patient was derived from the linear mixed-effect model for total mutation rate that included data from the entire cohort (Supplementary Results 5). The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. Both approaches rely on the assumption of a constant mutation rate for endometrial glands throughout a patient’s life.

2.9 Mixed-effect model and estimation of mutation burdens

2.9.1 Estimation of the total somatic mutation rate

Assuming a constant mutation rate, a linear model can be fitted to estimate the number of mutations that occurred due to normal mutational processes, which should correlate with patient age. To estimate the mutation rate, we could use a simple linear model or a mixed effects linear model. Given the fact that we have multiple samples from the same individuals, we chose to apply a linear mixed effects model, which takes into consideration both: (a) variation explained by the independent variables of interest (fixed effects), such as age, parity and others; and (b) variation not explained by the independent variables of interest (random effects), which would give a structure to the error term ϵ (ref).

We used a random slope with fixed intercept as most women will start menarche at a similar age (~13 years), but to account for the potential differences in the rates at which mutations were acquired each year in different individuals due to variation in parity, contraception and other factors.

We tested features with a known effect on mutation burden or endometrial cancer risks:

- Age
- Read depth & VAF ('Vafdepth')
- Driver mutations
- BMI
- Parity
- Cohort

For these analyses, we excluded the following cases: (a) samples from donors with missing meta-data, such as BMI and parity; (b) samples with an adjusted coverage (VAF depth) of <7.5 (adjusted coverage defined as VAF x sequencing depth). To account for the non-independent sampling per patient, we used mixed effects models. In these analyses, we tested features either with a known effect on mutation burden or endometrial cancer risk; age, read depth &

VAF, BMI, Parity. In addition, we tested whether there was any significant difference between different patient cohorts. Finally, we tested whether menstrual phase has an effect on the clonality and mutation burdens. All statistical analyses were performed in R and are summarised in Appendix 9.

2.9.2 Estimation of the driver mutation burden

To our best knowledge, there has been no previous work on estimating driver mutation rates in normal tissues.

Similar to the above, in order to describe estimates of the total mutation rates, we applied a mixed-effects model. However, given the fact that the data (driver variants) are not normally distributed and sparse, we used a generalized linear mixed effects model with Poisson distribution. As above, we also use a random slope with fixed intercept as most women will start menarche at a similar age (~13 years), but to account for the potential differences in the rates at which mutations were acquired in different individuals due to variation in parity, contraception and other factors.

Chapter 3 Clonal composition of normal endometrial epithelium

It is unknown whether endometrial epithelial glands are clones of cells deriving from a recent single common ancestor or whether they are constituted from multiple clones of cells from multiple ancestors. In this chapter I have aimed to use the whole genome sequences to characterise the clonality of normal endometrial glands.

3.1 Introduction to the chapter

Human endometrium is the mucosal lining of the corpus (the body) of the uterus. It is a unique highly dynamic tissue that undergoes over 400 cycles of breakdown, rapid repair, growth and remodelling in response to the oscillating levels of oestrogen and progesterone over a woman's lifetime (Jabbour et al., 2006, Gargett et al., 2007). Histologically, it is composed of two major components: the epithelial compartment in the form of tubular glands that produce glycogen-rich secretion and open up on to the luminal surface, and the mesenchymal compartment comprising cellular endometrial stroma and specialised hormone-sensitive blood vessels (spiral arterioles) (Mills et al., 2012, p.1071). Functionally, it is divided into two layers: the *functionalis*, the superficial layer that is sensitive to hormones and is shed during menses, and the *basalis*, the deep layer which is retained during menstruation or following gestation. The latter represents the germinal compartment of the endometrium containing adult progenitor stem cells from which the *functionalis* regenerates during menstrual cycles or after gestation (Chan et al., 2004, Gargett et al., 2008).

3.1.1 Endometrial adenogenesis

Human uterus differentiates from the Mullerian ducts and doubles in size from the twenty-eighth week of foetal development to birth. During this time, the initial endometrial adenogenesis, the development of glands, occurs. At this stage, the tissue is composed primarily of simple columnar epithelium lining the endometrial surface and from which small invaginations, primordial endometrial glands, are formed (Mills et al., 2012, p.1071). At birth, although the endometrial tissue architecture resembles that of an adult, it is still significantly less developed with only occasional endometrial glands present (Valdes-Dapen, 1973). A considerable amount of growth and adenogenesis occurs postnatally and in early childhood; at puberty, the tissue architecture reaches maturity with coiled, tubular glands radiating through to the myometrium (the underlying smooth muscle layer of the uterus) (Valdes-Dapen, 1973). Importantly, this pattern of gland development is distinct from the one observed in the adult endometrium through menstrual cycles when the glands develop adluminally from the basal layer (Okulicz et al., 1997, Huang et al., 2012).

A number of key genes are thought to be involved in the process of endometrial gland development. Amongst these are members of the *WNT* gene family (*WNT4*, *WNT5a* and *WNT7a*), which regulate essential cell behaviours including movement, adhesion, differentiation and proliferation, that are pivotal to endometrial adenogenesis (Cunha, 1976, Sharpe and Ferguson, 1988). Knock out of beta-catenin (*CTNNB1*), a critical intracellular mediator of Wnt signalling (Jeong et al., 2010), or its downstream target gene, transcription factor *Lef1* (Shelton et al., 2012), has been shown to disrupt gland formation in neonatal uterus. Forkhead box A2 (*FOXA2*) is a key transcription factor for adenogenesis (Jeong et al., 2010) with studies in mice showing that its ablation leads to significant reduction in the number of endometrial glands. Another important gene is *CDH1*, with its loss also leading to a reduction in the number of endometrial glands (Reardon et al., 2012). Notably, both *FOXA1* and *CDH1* genes are also thought to be involved in the Wnt signalling pathway.

3.1.2 Endometrium in reproductive years

During the female reproductive years, from menarche (the first occurrence of menstruation, usually at around 13 years) through to menopause (the cessation of menstruation, usually at around 51 years), the endometrium undergoes cyclical changes in response to oscillating levels of female hormones.

Endometrial function and the menstrual cycle are regulated primarily by steroid hormones secreted by the ovary. Following ovulation, the corpus luteum, a structure in the ovary that develops after an ovum has been discharged, secretes high levels of progesterone to maintain endometrial receptivity to the blastocyst, should fertilisation occur ('Secretory phase'). If pregnancy is not achieved, the *corpus luteum* regresses leading to a rapid decline in progesterone and oestrogen levels. The progesterone withdrawal causes tissue breakdown, local inflammatory response and shedding of the endometrium (Jouager et al., 2007). Following loss of almost the entire endometrial surface, re-epithelization is completed within 48 hours after the start of menses (Salamonsen et al., 1999, Ludwig et al., 1991) and the tissue undergoes further rapid proliferation and growth reaching a thickness of around 5-10 mm, a process that is driven by rising levels of unopposed oestrogen secreted by the ovary ('Proliferative phase'). The cycle then re-starts with the next round of ovulation (Day 14-15) (Figure 3.1).

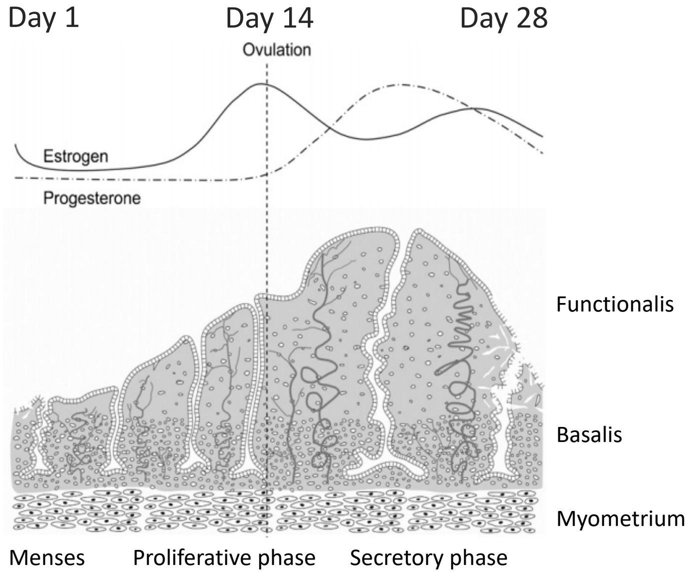


Figure 3.1 | Schematic of the human menstrual cycle. The human menstrual cycle is regulated by the ovary which secretes oestrogen and progesterone; the cycle is divided into three phases: menses, proliferative phase and secretory phase. Adapted from Gargett et al., 2008.

3.1.3 Endometrial adult stem cells

Adult stem cells are rare undifferentiated cells that are retained throughout the body after the completion of embryonic development (Li and Clevers, 2010, Weissman, 2000). They are characterised by the ability to self-renew as well as to produce more differentiated daughter cells (Gargett, 2007, Bongso and Richards, 2004) and play a critical role in maintenance of organs and tissues, and regeneration after damage. The existence of endometrial stem cells was first shown by Chan and colleagues (2004); using purified single cell suspensions obtained from hysterectomy tissues, they showed that 0.22±0.07% of endometrial epithelial cells and 1.25±0.18% of stromal cells formed colonies within 15 days. Both the epithelial and stromal cells generated two types of colonies: large and small colonies. Large putative stem cell colonies were rare (0.08% of single cell suspensions for epithelial cells and 0.02% for stromal cells); they displayed much greater self-renewal capability in comparison to the small colonies that showed a limited proliferation potential. The authors suggested that the large colonies were derivatives of the putative progenitor stem cells while the small colonies were thought to have been derived from transient amplifying (TA) cells.

Subsequently, Schwab and Gargett performed further functional clonogenicity experiments this time including samples not only from two phases of the menstrual cycle (proliferative and secretory) but also from inactive (post-menopausal) endometrium (Schwab and Gargett, 2007). The results showed that there is no variation in the frequency of clonogenic epithelial and stromal cells in two phases of the menstrual cycle or in the post-menopausal endometrium. Importantly, as inactive endometrium comprises the *basalis* layer only, the findings suggested that the endometrial progenitor stem cells reside in the basalis layer and persist beyond the menopause. This suggestion that the endometrial epithelial adult stem cells (eeASCs) reside in the basalis is further supported by the ability to induce proliferation in post-menopausal women who are treated with hormone replacement therapy as well as tissue regeneration and regrowth in patients who undergo extensive endometrial ablation for heavy bleeding (Tresserra et al., 1999).

Since the majority of endometrial cancers are of epithelial origin, the remainder of this section will be focused on the eeASCs. Despite the fact that their existence was first shown over a decade ago, eeASCs have remained poorly characterised in comparison to their counterparts in other tissues, such as colon and stomach. One of the main reasons for this is the lack of a specific marker that would reliably distinguish those cells from their mature progeny. Some of the general stem cell markers, for instance bcl-2, c-kit (CD117) and CD34 have been identified in the normal endometrium (Cho et al., 2004). However, the number of cells that expressed these markers were significantly higher than the number of clonogenic cells that had been previously shown in the functional studies (Chan et al., 2004, Schwab and Gargett, 2007).

Chan and Gargett carried out further experiments to locate label retaining cells (LRCs) to identify somatic progenitor stem cells and characterise their location in the stem cell niche in the absence of specific markers (Chan and Gargett, 2006). They studied mouse endometrium in which the tissue was pulse labelled with Bromodeoxyuridine (BrdU) and examined after an 8-week chase to identify endometrial LRCs. The results showed that 3% of the epithelial nuclei were BrdU+ and were located in the luminal epithelium. The cells did not express Oestrogen Receptor alpha (ERα) through dual labelling immunofluorescence, confirming that luminal epithelial progenitor stem cells are responsible for the growth of glands during development and in cycling mice. With the use of a mouse model with menstrual breakdown and repair, ERα negative glandular epithelial LRCs contributed to the repair of the luminal epithelium following menstruation. Endometrial repair occurred in the absence of oestrogen. BrdU^{+} epithelia were rapidly lost in the chase period, leading to the notion that the epithelial regeneration may depend on self-duplication of a mature epithelial cell type, or that the LRC technique is not sensitive enough to label rare endometrial epithelial cells with an ASC phenotype.

3.1.4 Clonal composition of endometrial glands

Cancers are caused by the accumulation of somatic mutations in normal cells. These mutations allow cells to proliferate uncontrollably, escaping homeostatic controls and providing survival advantage over their neighbours with subsequent clonal expansion. To better understand ageing and early neoplastic transformation, it is essential to expand our knowledge on somatic evolution, selective pressures and remodelling in normal tissues.

Colon is one of the most studied highly proliferative tissues, in which individual crypts, its functional gland-like units, are known to eventually become clonal cell populations that share most common recent ancestors (MCRAs). The monoclonal conversion of individual crypts is thought to occur through neutral drift (Snippert et al., 2010); in mice, it has been shown that the initially multi-coloured colonic crypts became monochrome over a period of around 1-6 months in a pattern that is consistent with neutral dynamics. In humans, the monoclonality of colonic crypts was shown using naturally occurring somatic mitochondrial DNA (mtDNA) mutations through enzyme-histochemical staining for loss of cytochrome C oxidase (CCO) protein (Baker et al., 2014). More recently, Nicholson and colleagues used staining for loss of mild Periodic Acid-Schiff (mPAS), which detects loss of O-acetylation of sialomucins and is a marker of clonality, and showed that in humans, the process of monoclonal conversion of colonic crypts takes several years (13 years for 90% conversion, median 6.3 years) (Nicholson et al., 2018).

However, the above mentioned lack of definitive markers for eeASCs has meant that little is known about the stem cell dynamics and clonal composition of normal endometrial glands. To my best knowledge, work by Tanaka and colleagues from more than 15 years ago is the only study inferring clonal composition of human endometrial glands (Tanaka et al., 2003). Using a collagenase-based digestion approach, they isolated individual human and mouse glands and assessed their clonality using a polymerase chain reaction-based assay for non-random X-chromosome inactivation with an X-linked androgen receptor gene. They found that most of the studied glands were monoclonal populations and that in some of the clonal patches expanded over several adjacent glands. Although this study provided first insights into stem cell dynamics in the tissue, the clonal patches and their distributions were defined by the events that would have occurred in early embryogenesis, whereas the aim of our study was to infer adult stem cell dynamics and associated clonal expansion that occurs throughout life.

3.1.5 Study design and sample selection

Normal endometrium was one of the first tissues that we included in our pan-body survey of somatic mutations (Methods). In July 2017, the results of the initial experiments showed two striking observations: the majority of the sampled endometrial glands were clonal cell populations and mutations in cancer genes, such as *PIK3CA*, *KRAS* and *FBXW7*, were frequent. Yet, there was no morphological evidence of neoplastic transformation in any of those samples. Shortly before we made this observation, a whole exome and targeted sequencing study on endometriosis was published (Anglesio et al., 2017). In this analysis, Anglesio and colleagues investigated genomic changes in deep infiltrating endometriosis (DIE), a disorder in which histologically normal endometrium is found deep in abnormal (ectopic) locations, such as the urinary bladder or the bowel. Although ectopic, the endometrium in these lesions is 'functional' and undergoes cyclical changes similar to those in the eutopic (uterine) tissue; the associated repetitive breakdown and regeneration causes local bleeding, inflammatory reaction and pain. The study showed that driver mutations can be found in these lesions without morphological evidence of cancer. This finding was particularly interesting as unlike ovarian endometriosis, DIE is not known to undergo malignant transformation (Wei et al., 2011). Given their results and our observations in normal endometrial glands from the initial experiments, we decided to carry out a larger study.

As human endometrium is a highly dynamic tissue that adopts various physiological states, to obtain a representative view of its somatic mutagenesis and consequences throughout life, we collected samples from as wide an age range of women as possible. These included biopsies taken from women under investigation for reproductive problems (14), hysterectomies for benign non-endometrial pathologies (2), residual tissues from transplant organ donors (8) and autopsies after death from non-gynaecological causes (4) (Meta-data is summarised in Table 3.2). We also aimed to assess how these are modulated by some of the known endometrial cancer risk factors such as BMI and parity. Finally, to confirm normal histology, all endometrial biopsies were examined by two histopathologists (Dr Mercedes Jimenez-Linan and myself).

Patient ID	Reason for sampling	Age	BMI	Parity	No.of high coverage samples	Menopause status	Menstrual phase
PD37506	Post-mortem (traumatic injury)	19	U	U	10	Pre-menopausal	Undetermined
PD40535	Transplant donor	24	24	3	7	Pre-menopausal	Proliferative
PD41871	Infertility clinic	27	30	0	17	Pre-menopausal	Secretory
PD37605	Infertility clinic	29	27	2	9	Pre-menopausal	Secretory
PD37601	Infertility clinic	31	28	0	10	Pre-menopausal	Secretory
PD41860	Infertility clinic	31	23	0	4	Pre-menopausal	Secretory
PD37607	Infertility clinic	34	24	1	19	Pre-menopausal	Secretory
PD41857	Infertility clinic	34	22	1	14	Pre-menopausal	Secretory
PD39444	Transplant donor	35	24	1	10	Pre-menopausal	Proliferative
PD41865	Infertility clinic	36	31	0	2	Pre-menopausal	Secretory
PD41868	Infertility clinic	36	23	0	6	Pre-menopausal	Secretory
PD39953	Transplant donor	37	18	2	8	Pre-menopausal	Secretory
PD41859	Infertility clinic	38	21	0	1	Pre-menopausal	Secretory
PD37613	Infertility clinic	39	22	0	11	Pre-menopausal	Secretory
PD41861	Infertility clinic	39	21	0	8	Pre-menopausal	Secretory
PD41869	Infertility clinic	40	37	0	13	Pre-menopausal	Secretory
PD37594	Infertility clinic	42	20	1	17	Pre-menopausal	Secretory
PD39952	Transplant donor	44	36	0	11	Pre-menopausal	Proliferative
PD39954	Transplant donor	44	24	1	10	Pre-menopausal	Secretory
PD37595	Infertility clinic	46	19.5	5	9	Pre-menopausal	Secretory
PD36804	TAH for leiomyomata	47	30	3	13	Pre-menopausal	Secretory
PD36805	TAH for benign ovarian tumour	49	27	0	7	Pre-menopausal	Secretory
PD38812	Post-mortem (traumatic injury)	54	U	U	2	Post-menopausal	Proliferative
PD37507	Post-mortem (peritonitis)	60	U	U	14	Post-menopausal	Inactive
PD42746	Transplant donor	67	34	2	2	Post-menopausal	Inactive
PD40107	Transplant donor	69	24	2	10	Post-menopausal	Inactive
PD42475	Transplant donor	74	27	2	8	Post-menopausal	Inactive
PD40659	Post-mortem	81	22	4	5	Post-menopausal	Inactive

Table 3.2 | Summary of clinico-pathological data for all donors.

U, unknown; TAH, total abdominal hysterectomy.

3.2 Results

3.2.1 Sample Collection

I laser-capture microdissected >800 individual endometrial glands. DNA from each gland was subjected to our LCM library-making protocol modified to accommodate small amounts of input DNA (methods). Wherever possible, biopsies from other tissues, including Fallopian tube, cervix and myometrium, were also collected.

3.2.1.1 Paired normal selection

To exclude germline mutations, somatic mutations in each gland were determined by comparison with whole genome sequences from pieces of uterus, cervix or Fallopian tube from the same individuals. The type of sample that was used as a normal was determined by the nature of the procedure during which the endometrial sample was taken: samples from the infertility clinic were taken from live donors during hysteroscopy, which is usually limited to the endometrium layer of the uterus and therefore we had to use endometrial stroma as a paired normal sample; in the case of the hysterectomy resections, post-mortem and transplant donor samples, other tissues were available such as cervix and myometrium. On a selection of samples, we re-ran mutation calling algorithms (CaVEMan and Pindel) using matched normal samples from two different tissues; no significant difference was observed between the two runs (Table 3.3 and Figure 3.2).

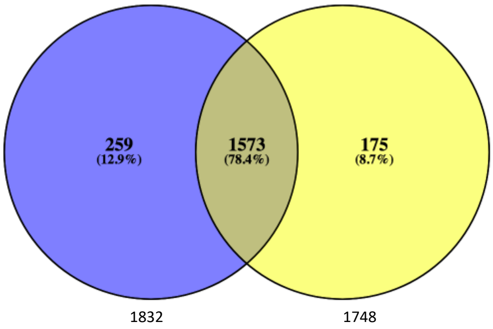


Figure 3.2 | An example of the overlap of variants called in the same sample using two different paired normal samples (cervix and myometrium). In this case (sample PD36805b_EM9_G4_B3) variants were called using cervix and myometrium bulk samples.

Sample ID	Subs called against cervix	Subs called against myometrium
PD36804b_EM4_G3_E5	896	875
PD36804b_EM5_G2_B6	1176	1183
PD36804b_EM5_G3_C6	1059	1052
PD36804b_EMD_7_A1	1743	2016
PD36804b_EMD_7_A5	1420	1418
PD36804b_EMD_7_A6	1408	1399
PD36804b_EMD_7_C2	1522	1529
PD36804b_EMD_7_C3	1621	1615
PD36804b_EMD_7_C6	1396	1398
PD36804b_EMD_7_E3	1618	1608
PD36804b_EMD_7_E4	1390	1387
PD36804b_EMD_7_G4	1403	1397
PD36804b_EMD_7_G5	1362	1367
PD36805b_EM10_G2_A3	1092	1104
PD36805b_EM10_G3_C3	1644	1647
PD36805b_EM1_G1_L1_2_A1	1525	1518
PD36805b_EM7_G2_C8	1575	1577
PD36805b_EM8_G2_F8	1680	1676
PD36805b_EM9_G1_A9	1662	1659
PD36805b_EM9_G4_B3	1832	1748

Table 3.3 | Comparison of substitution mutation burdens in selected samples using different types of tissue as the matched normal.

3.2.2 Sample QC

Based on post-library preparation DNA concentration (a cut-off of ≥5 ng/ul was applied), a total of 292 glands were selected for whole genome sequencing (30x). The mean sequencing coverage was 28-fold (Figure 3.3). Only samples with ≥15-fold coverage were processed through the variant calling pipeline (n=257, Appendix 10).

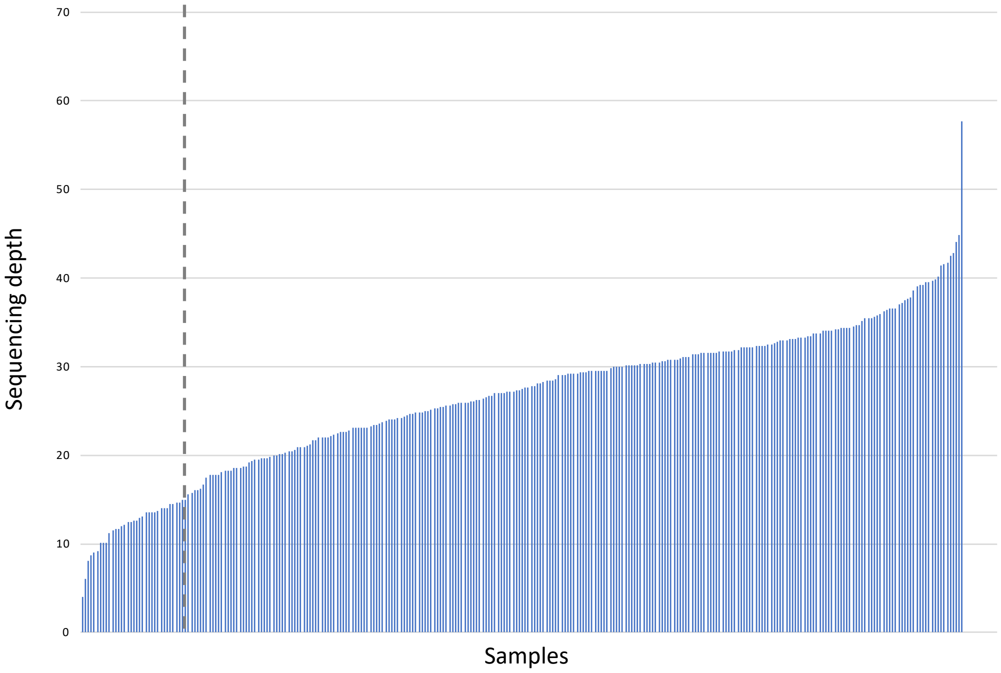


Figure 3.3| Sequencing coverage across all endometrial gland samples. A total of 292 normal endometrial glands were subjected to whole genome sequencing. Only samples with a ≥15-fold coverage (indicated by the dotted line) were used for subsequent analyses.

3.2.3 Variant calling

Using 18 pairs of biological 'near-replicates' (details in methods) we calculated the mean sensitivity of our somatic mutation variant calling at >0.86% (range 0.70-0.95%).

A total of 338,376 single nucleotide variants (SNVs) was found with a median of 1521 (range 209-2833) per sample.

3.2.4 Clonality of endometrial glands

There are a number of ways in which we could infer clonal composition of endometrial glands. Some of these are discussed below, all analyses used filtered caveman input as outlines in methods.

3.2.4.1 Distribution of variant allele fractions (VAFs) of all mutations

In the simplest approach, clonality can be explored through variant allele fractions (VAFs). As most somatic mutations are heterozygous, those mutations present in all cells of a population derived from a single ancestor will have VAFs of 0.5 whereas mutations in cell populations derived from multiple ancestors will have lower VAFs or be undetectable by standard mutation calling approaches. Therefore, to assess whether endometrial glands are clonal cell populations, the VAFs of all called somatic mutations can be used; 91% (234/257) of microdissected endometrial glands showed distributions of base substitution VAFs with peaks between 0.3 and 0.5 (Figure 3.4) indicating that each gland consists predominantly of a cell population descended from a single epithelial progenitor stem cell. Mutations that are present at a lower VAF may represent contamination by other cell types; these potentially include endometrial stromal cells, inflammatory cells, epithelial cells from neighbouring glands or subclonal diversification within the same gland.

Assessment of small insertions and deletions (indels) showed similar VAF distributions confirming the results from base substitutions (Figure 3.5).

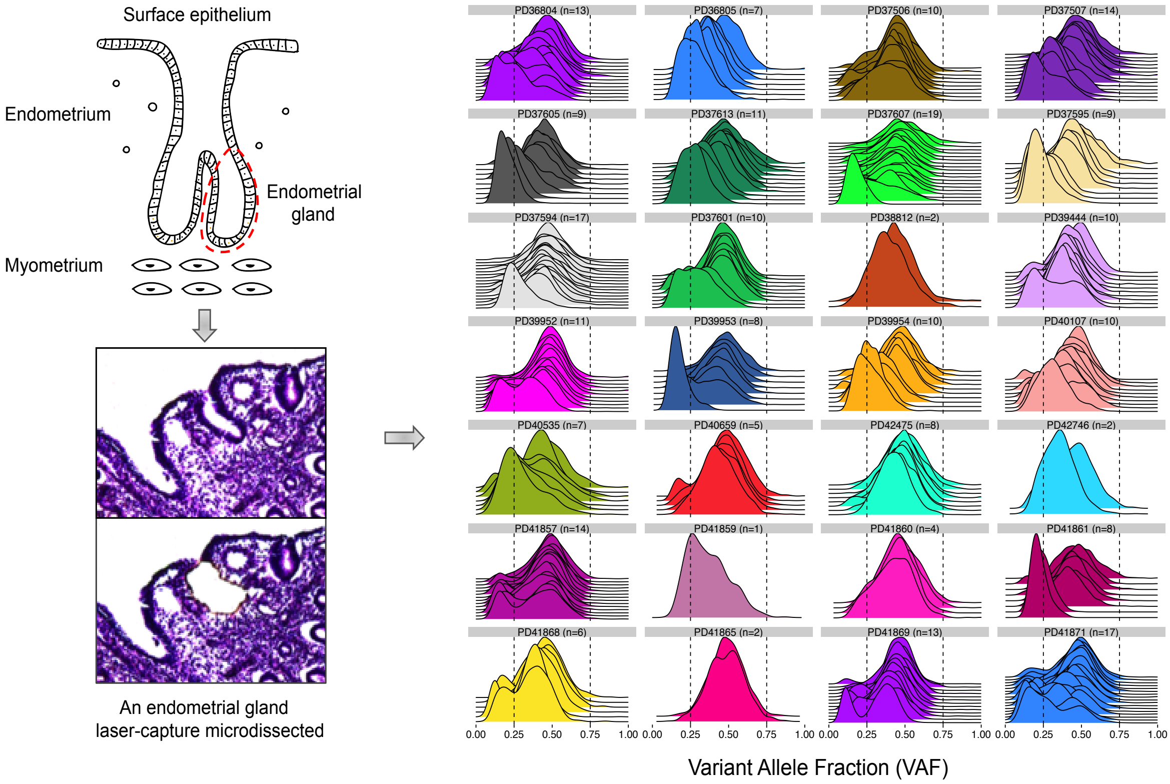


Figure 3.4| Clonality of normal endometrial glands. Individual normal endometrial glands were laser-capture microdissected and whole genome sequenced. The majority (91%) of the sampled glands were clonal cell populations, sharing the most recent common ancestor, with a median variant allele fraction (VAF) between 0.3 and 0.5 for all identified substitutions across individuals. Each density line represents an individual endometrial gland sample; individual samples are grouped and coloured by patient.

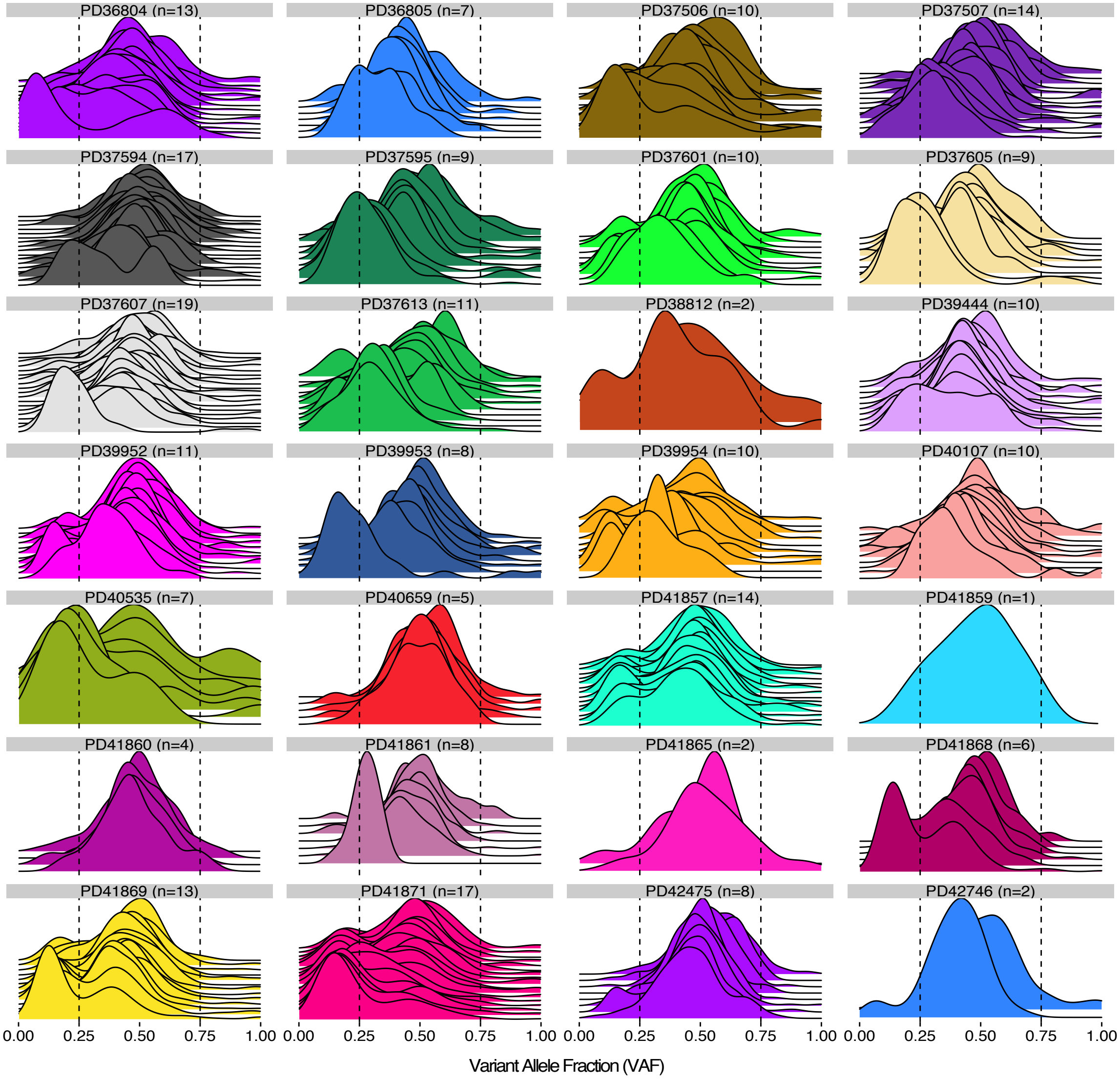


Figure 3.5| Clonality of endometrial glands based on VAFs distributions for indels. The majority of sampled normal endometrial glands were clonal with a median variant allele fraction (VAF) for all identified indels of 0.3 or above.

3.2.4.2 **Two-dimensional clustering algorithms: dpclust**

To formally assess clonal composition of each sampled endometrial gland, we applied a previously described sub-clonal reconstruction caller (dpclust v2.2.7) (Nik-Zainal et al., 2012) with default parameters to the single nucleotide variants (SNVs) in each endometrial gland to assess the clonality of each gland (this work was carried out by Stefan Dentro). The analysis yields, for every sample, the number of mutation clusters and assigned mutations, and the proportion of overall cells that each cluster represents.

A gland was determined to be the result of a single progenitor cell if a single mutation cluster was obtained or when the proportions of cells in which multiple mutation clusters were detected. Akin to the so-called "pigeon-hole" principle (Yates et al., 2015), in such a scenario the sum of the estimated proportions of cells of a pair of cellular populations exceeds 1 (100% of cells), which means at least some cells must contain the mutations in both clusters. Alternatively, if the sum of the estimated proportions does not exceed 1 the populations could be the result of a single or of separate ancestors.

The results of the dpclust analysis concurred with our observations based on the distribution of VAFs; 89.9% (231/257) of all endometrial glands had a major clone (defined as those with ≥ 75% of sequenced cells) with clusters containing on average 79.5% of all substitutions (sd = 24.9%) (Figures 3.6 and 3.7, Appendix 11). 83% (214/257) of glands showed evidence of a further, subclonal cell population which, based on the "pigeon-hole" principle (Yates et al., 2015), is a descendant of the main clonal population. The majority of glands also showed minor contamination by cells that do not share somatic mutations with the observed clonal expansions, potentially including endometrial stromal cells, inflammatory cells and epithelial cells from other glands.

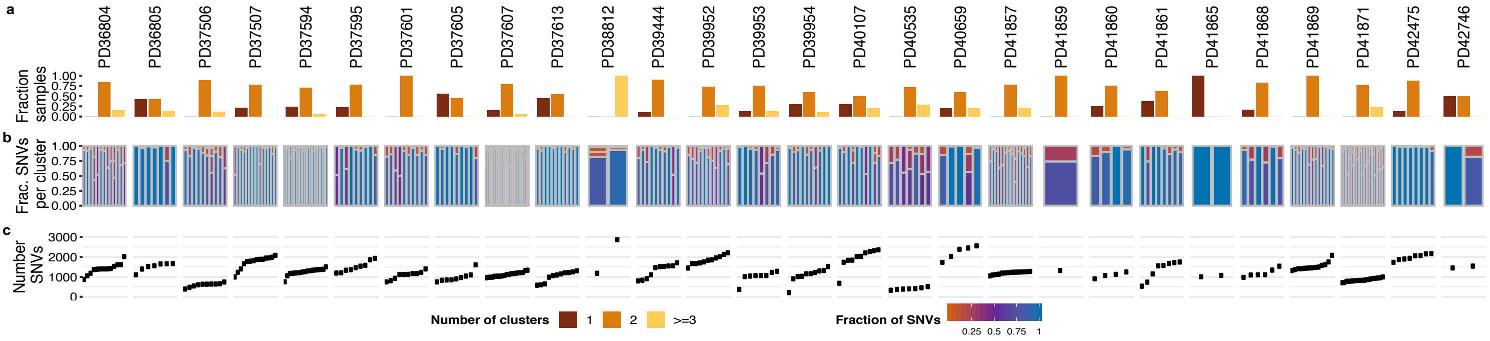


Figure 3.6 | Assessment of clonal composition of individual endometrial glands using mutation clustering method dpclust. Each column contains summary of the clonality analysis for individual donors, showing the fraction of samples in which 1, 2 or 3 or more mutation clusters were found (a), the fraction of mutations assigned per cluster for each sample (b) and at the total number of single nucleotide variants (SNVs) per sample (c).

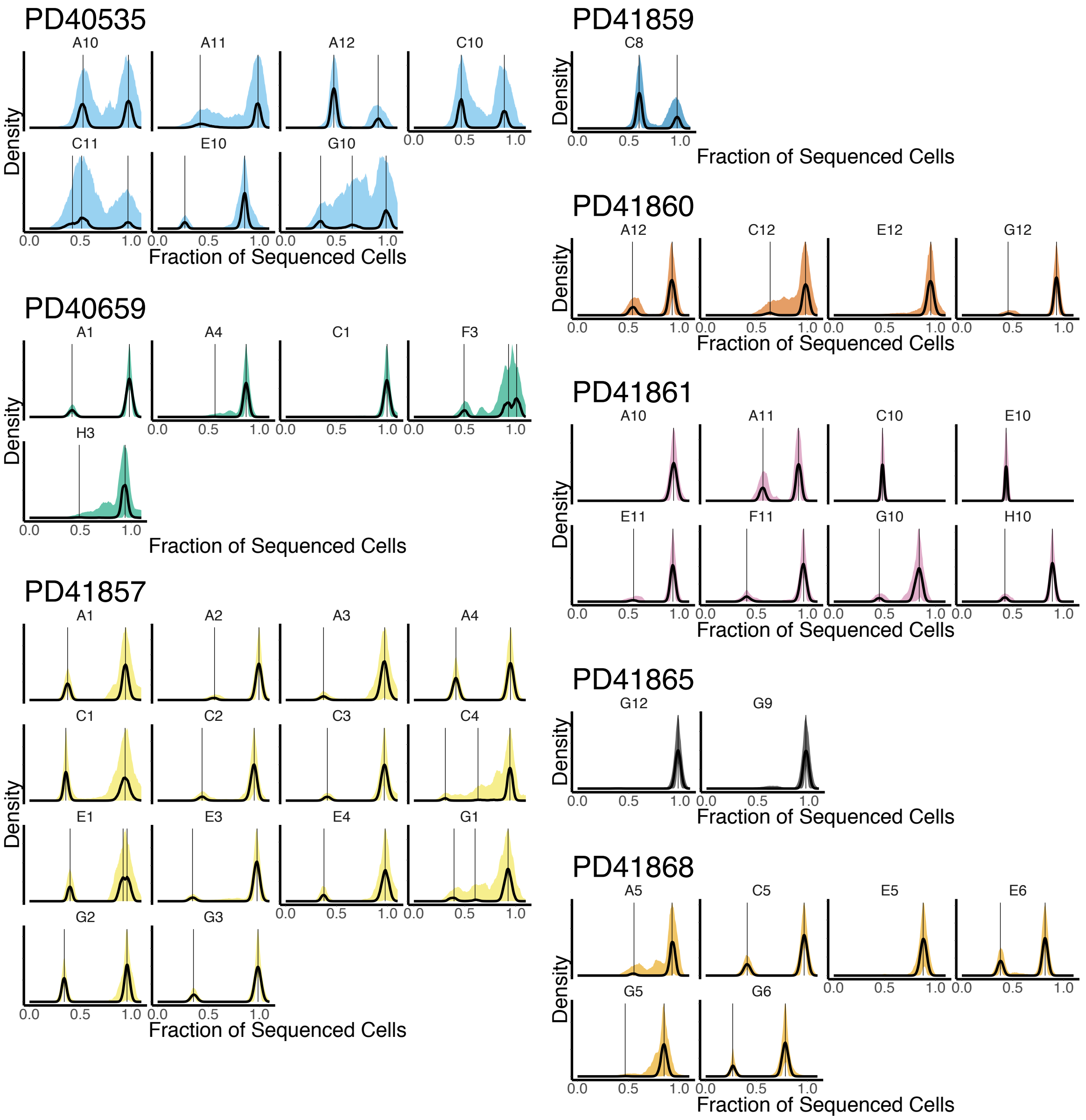


Figure 3.7 | Examples of clusters identified in individual endometrial glands with dpclust. The clonality analysis yields a posterior density estimate of what proportion of sequenced cells likely represents a mutation cluster. Each plot shows the posterior density in black and its corresponding 95% confidence interval coloured. Called clusters are marked with a vertical black line. 89.9% of all sampled glands had a major clone which is defined as a cluster containing >=75% of all base substitutions. The identified subclonal populations can represent either late subclonal diversification occurring in an individual endometrial gland, incomplete monoclonal conversion of a gland and contribution of more tham one adult stem cell or contamination with another clone from an adjacent gland or even stroma. The complete clonal decomposition analysis for all glands and donors is provided in Appendix 18.

3.2.4.3 **Two-dimensional clustering algorithms: LICHeE**

Another computational method that utilises single nucleotide variants to infer sub-clonal composition of samples while allowing simultaneous reconstruction of multi-sample cell lineage trees (in our study, per donor lineage) is LICHeE (Lineage Inference for Cancer Heterogeneity and Evolution) (Popic et al., 2015). This approach relies on VAFs of deep-sequenced somatic SNVs. The algorithm was run with default settings: distance between clusters was 0.15, minimum VAF for a mutation to be present was 0.15, maximum VAF for a cluster was 0.65, and a VAF measurement error of 0.10.

The results showed that 67/257 (26%) glands had one major clone, 189/257 (74%) glands had two clusters and 1/257(<1%) had 3 clusters (Figure 3.8).

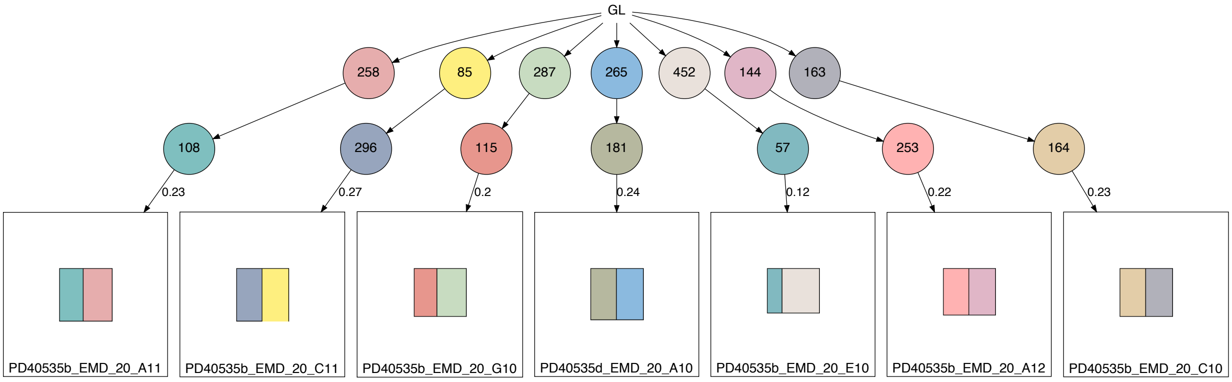


Figure 3.8 | An example of clusters identified with LICHeE. This tool relies on variant allele fractions (VAFs) of deep sequenced SNV's and default pre-defined distances between clusters. It is a rather simplistic but a quick approach to defining number of clusters in an individual sample**.**

3.2.4.4 **Methods comparison**

Comparison with clonal clusters identified using dpclust and LICHeE showed a correlation of 0.40 (Figure 3.9). This method generally proved ineffective due primarily to 2 factors. First, using VAF instead of cancer-cell fraction yielded VAF clusters which broke the pigeonhole principle due to varying contamination by other cell types, LICHeE is unable to handle such trees. Second, the distance between clusters would ideally be dynamically chosen rather than a single fixed value across all donors and samples.

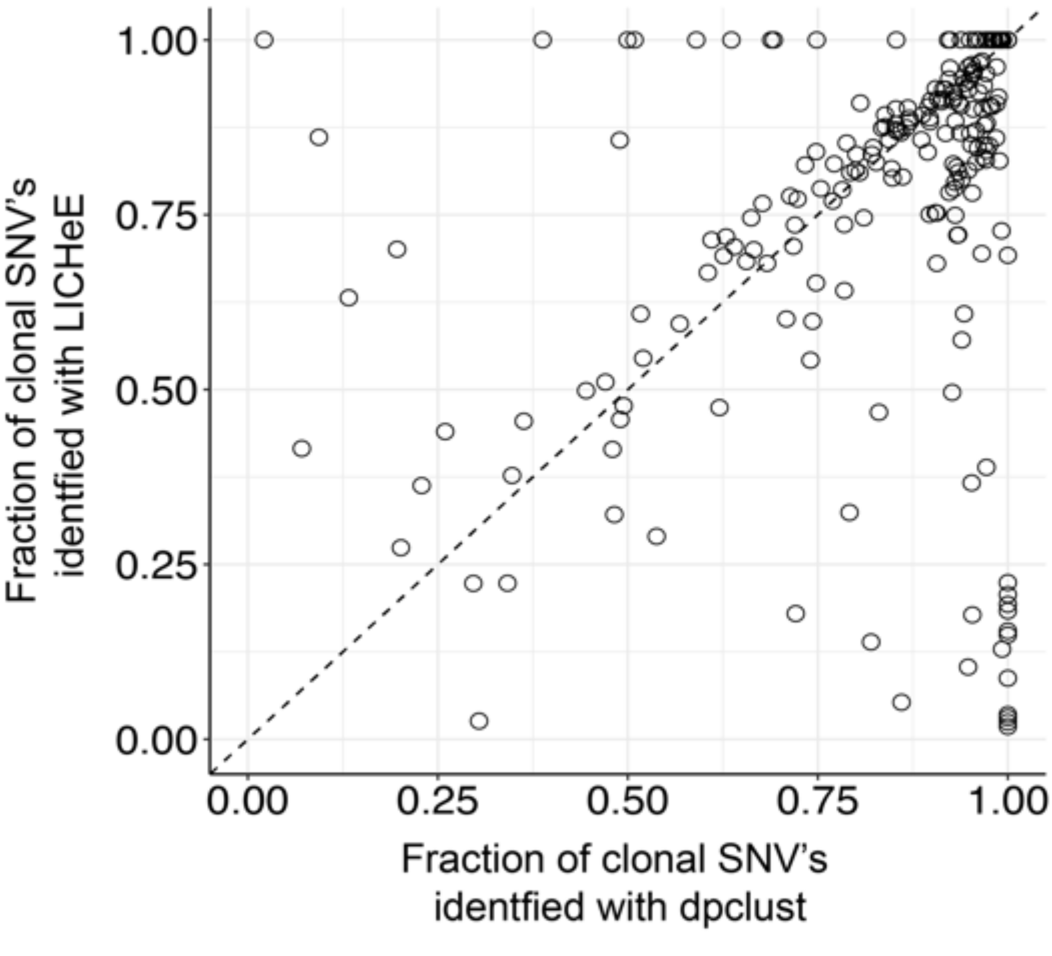


Figure 3.9 | Comparison of clonal mutations identified by dpclust and LICHeE algorithms.

3.2.4.5 **Two-dimensional clustering algorithms: PyClone**

We also attempted to use the PyClone algorithm, which has been previously applied on whole exome sequencing data (Roth et al., 2014) (this analysis was performed with help from Raheleh Rahbari). However, the algorithm was built for whole exome sequencing data with fewer variants, we were unable to run it on our whole genome sequencing data from all glands. Instead, we selected 1000 random substitutions per individual with genotype priors. The result of the PyClone analysis with beta-binomial emission densities with total copy-number priors showed that the majority of mutations were clonal across all individuals. Figure 3.10 illustrates an example of a PyClone density plot for donor PD39952 with 99% of the 1000 selected mutations clustered together at a variant allele fraction (VAF) of 0.5.

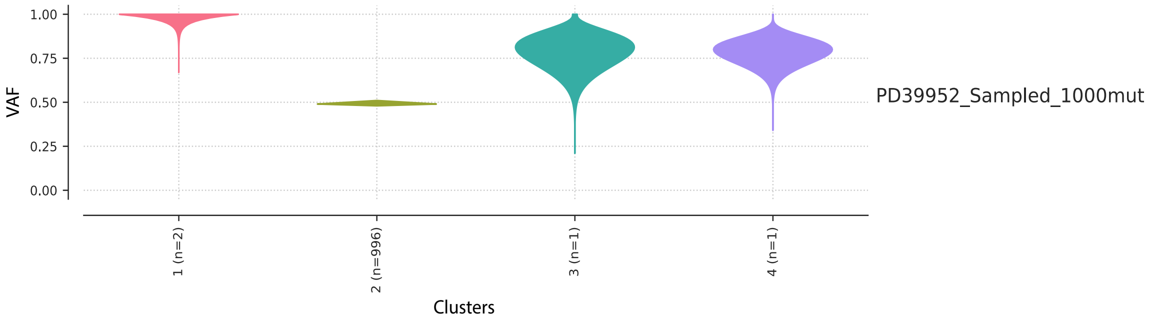


Figure 3.10 | Density plot of identified clusters and substitutions assigned in donor PD39952. For each donor, 1000 random substitutions were selected from different samples. PyClone analysis showed that the majority of the mutations were clonal with VAF = 0.5.

3.2.4.6 **Clonality and presence of driver mutations**

Subsequent analyses demonstrated that many endometrial glands carry "driver" mutations in known cancer genes. Such mutations are known to be advantageous to stem cells – these allow uncontrolled proliferation and provide selective advantage over their neighbours (Stratton et al., 2009). We therefore examined the effect of the presence or absence of a driver mutation on clonality of endometrial glands. The analysis showed that endometrial glands exhibit clonality irrespective of the presence of known driver mutations (Figure 3.11a) with, for example, somatic mutations in all 10 glands from a 19-year-old individual (PD37506) having a median VAF >0.3 but no driver mutations identified (Figure 3.11b). Thus, colonisation of endometrial glands by descendants of single endometrial epithelial stem cells is not contingent on a growth selective advantage provided by driver mutations and may occur by a process analogous to genetic drift, as proposed for other tissues (Lopez-Garcia et al., 2010, Snippert et al., 2010).

Given the highly dynamic nature of the endometrium with cycles of tissue loss, rapid regeneration (proliferative phase) and further growth and expansion (secretory phase) during reproductive years and the lack of these in post-menopausal women, we examined the correlation between the menstrual phase, menopause and clonality of glands. The results showed that the observed monoclonality was also independent of the menstrual phase and menopause status (Appendix 12).

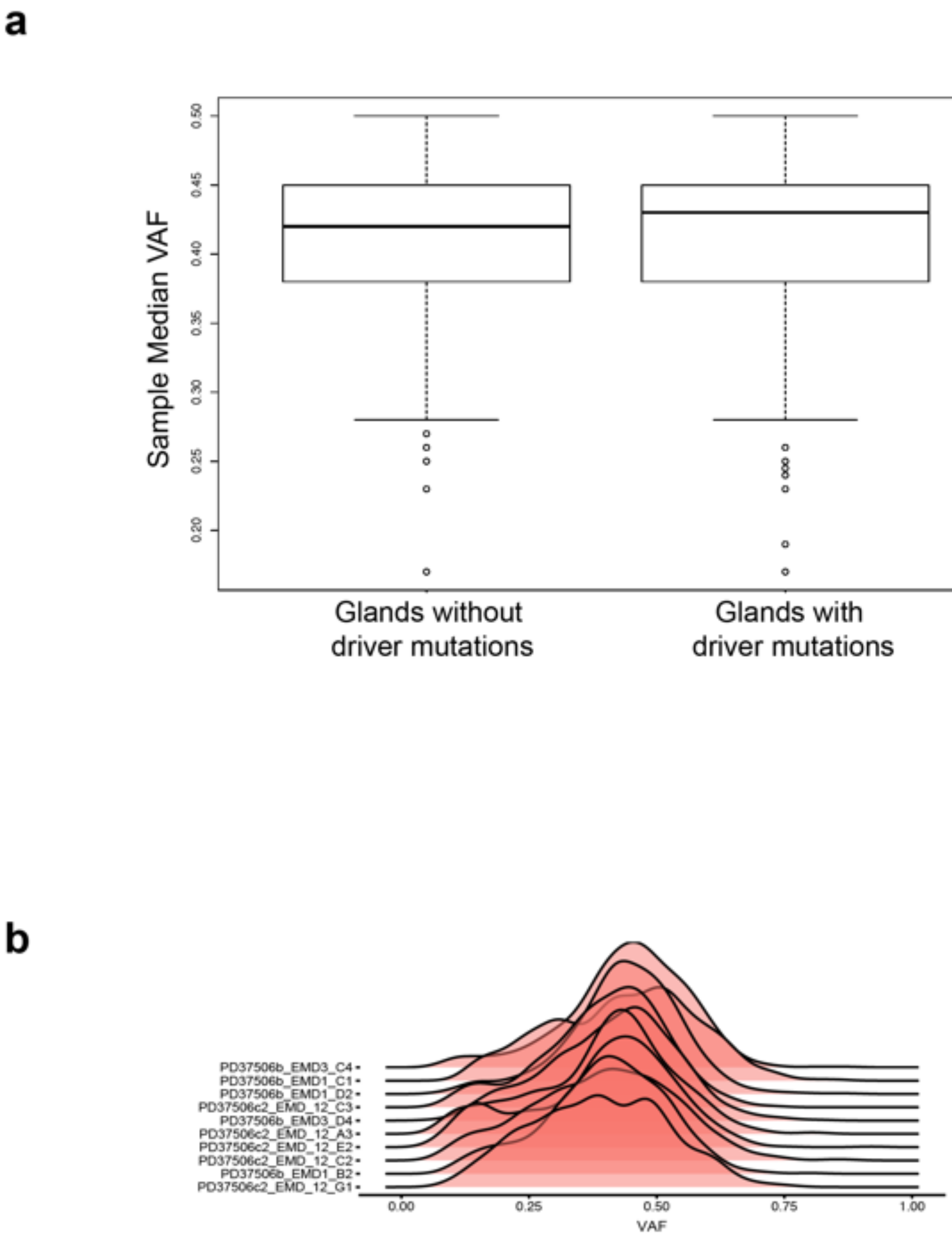


Figure 3.11 | Clonality of endometrial glands and driver mutations. The presence of a driver mutation did not have a significant effect on the observed monoclonality of the glands (Mann-Whitney two-sided test, p = 0.1) (**a**). All glands from the 19-year-old donor (PD37506) were clonal with a median VAF >=0.3, but there were no detectable driver mutations (**b**).

3.3 Summary of results for this chapter

Endometrium is a relatively less studied tissue in comparison to other glandular type tissues, such as colon and stomach. Although endometrial epithelial adult stem cells (eeASCs) were first described over a decade ago, they remain relatively poorly characterised in comparison to their counterparts in other tissues, such as the small and large intestine. In particular, the number of stem cells in individual endometrial glands, their dynamics and clonal expansion remain poorly understood, which at least in part, is due to the lack of robust biomarkers (Tempest et al., 2018) and animal models given that only a limited number of species undergo menstrual cycle with tissue loss and regeneration. Here, we show that irrespective of the 'starting' number of eeASCs, the majority of normal endometrial glands are clonal cell populations that share common recent ancestors. The monoclonal conversion occurs early (all glands from a 19-year old individual were clonal) and is independent of the presence of driver mutations and menstrual phase.

Chapter 4 The mutational landscape of normal endometrial epithelium

4.1 Introduction to the chapter

4.1.1 Somatic mutations in normal endometrium

4.1.1.1 Mutation rates

In recent years, there has been growing interest in somatic mutations in normal ageing tissues. A number of studies, including several from our group, have characterised these changes for different epithelial tissues, such as the small and large intestine(Lee-Six et al., 2019, Blokzijl et al., 2016), liver (Blokzijl et al., 2016, Zhu et al., 2019) and oesophagus (Martincorena, 2018, Martincorena et al., 2018, Yokoyama et al., 2019); similar work has also been carried out on non-epithelial tissues, for instance, skeletal muscle (Franco et al., 2018) and blood (Osorio et al., 2018, Lee-Six et al., 2018b). Recent pan-cancer analyses have examined somatic mutation rates across various tumours, including those originating in the endometrium; the results have given us first estimates of the 'clock-like' mutation rates in normal cells based on the fact that cancers arise from cells that were once normal (Alexandrov et al., 2015). However, such views are likely to be distorted as the estimates were derived from cancer tissues rather than from normal tissues directly. To the best of my knowledge, our study was the first to estimate mutation burden and rates in the normal human endometrium.

4.1.1.2 Genomic changes in normal endometrium

The first insights into the genomic changes in non-neoplastic endometrium were provided in a study by Nair and colleagues, in which they applied ultra-deep, targeted sequencing to screen for cancer driver mutations in uterine lavage fluid from women undergoing hysteroscopy for molecular screening and diagnosis of endometrial cancer (Nair et al., 2016). They showed that a deep targeted sequencing approach can be used to detect early microscopic lesions. In addition, they also found cancer associated mutations in ~49% of all examined women without histological evidence of endometrial pathology. Importantly, the presence or absence of a neoplasm in this study was based on histological assessment made only on a small tissue biopsy that was taken during the hysteroscopy and, undoubtedly, some of the negative cases could represent missed lesions rather than truly non-neoplastic endometrium. In addition, uterine lavage fluid contains a mixture of endometrial and non-endometrial cells, including those shed from the epithelial lining of the Fallopian tubes, cervix and ovary. It is therefore plausible that some of the detected driver mutations were actually representative of genomic changes that occurred in these tissues rather than the endometrium. Nevertheless, this study was the first to suggest that cancer driver mutations may potentially be found in non-neoplastic endometrium.

As described earlier in Chapter 3, shortly before our initial experiments on normal endometrium, a study by Anglesio and colleagues showed that cancer associated mutations can be identified in endometriosis (Anglesio et al., 2017). Known cancer driver mutations in genes such as *PIK3CA*, *KRAS* and *ARID1A* were found in 5/24 patients some of whom were in their 20s. Subsequently, the same group investigated genomic changes in another type of the disorder, iatrogenic endometriosis, which is thought to be associated with previous surgical procedures (Lac et al., 2018). Similarly, driver mutations could be detected in these samples, yet these lesions virtually never undergo malignant transformation (Wei et al., 2011).

The aim of our study was: to use whole genome sequencing to provide a comprehensive characterisation of the mutational landscape of the normal endometrial epithelium; to explore how this landscape is influenced by age, BMI and parity, to estimate the age of driver mutations and to investigate the relationship of clonal evolution to glandular architecture.

4.1.2 Current understanding of endometrial cancer

Endometrial cancer is the most common gynaecological tumour in the developed world with 9,314 new cases and 2,360 deaths a year in the UK (CRUK, 2019). While it is not the ‘deadliest’ malignancy, its incidence has increased by 57% in the UK between 1993-1995 and 2014-2016

(CRUK, 2019). Moreover, the incidence is predicted to rise further, which is at least partially related to the worldwide obesity epidemic (Morice et al., 2016, Onstad et al., 2016), thus making it an important health care issue and burden in the future. Approximately 75% of patients with the disease are diagnosed in the early stages (International Federation of Gynaecology and Obstetrics (FIGO) stages I and II) with a 5-year overall survival of 74-91% (Siegel et al., 2013, Creasman et al., 2006). For patients with advanced disease (stage III and IV), 5-year overall survival is 57-66% and 20-26% respectively (Creasman et al., 2006).

The majority of endometrial cancers are sporadic, but a small proportion of cancers (2-5%) are familial (Le Gallo and Bell, 2014). These include tumours associated with Lynch Syndrome (hereditary nonpolyposis colorectal cancer) with underlying germline mutations in mismatch repair genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) as well certain germline deletions in *EPCAM,* and cancers in patients with Cowden Syndrome who carry germline mutations in *PTEN* (Le Gallo and Bell, 2014).

4.1.2.1 **Classification of endometrial cancer**

Historically, endometrial cancers have been classified into two broad groups based primarily on their clinical, metabolic and endocrine features (Bokhman, 1983). Type I tumours are thought to be linked to unopposed oestrogen exposure and obesity, are hormone-receptor positive and are usually well to moderately differentiated neoplasms that carry a relatively favourable prognosis (Cancer Genome Atlas Research et al., 2013, Murali et al., 2014). Type II cancers are less common, tend to present in older, post-menopausal, non-obese women, arise in the absence of endocrine and metabolic disturbances, are poorly differentiated and have a less favourable outcome.

Since this original classification by Bokhman in 1983, endometrial cancers have been further characterised and subtyped using histological and more recently, molecular features. These are considered below.

4.1.2.2 **Histological classification of endometrial cancer**

According to the World Health Organisation (WHO) classification, neoplasms of the uterine corpus comprise several distinct histological types: epithelial carcinomas (endometrioid, serous, clear cell, mucinous, squamous cell, transitional cell, small cell and undifferentiated), mixed epithelial and mesenchymal tumours (e.g. carcinosarcomas), or mesenchymal tumours (e.g. endometrial stromal sarcomas) and others (Silverberg et al., 2003). However, epithelial carcinomas account for the majority of all endometrial neoplasms, including endometrioid (87-90%) and serous (5-10%) (Liang et al., 2012), and therefore the rest of the discussion will be focused on these tumours.

Endometrioid carcinoma

Endometrioid carcinomas (ECs) are associated with excess exposure to unopposed oestrogen with risk factors including, obesity, early age at menarche (the first occurrence of menstruation), late age at menopause and nulliparity (never having completed a pregnancy beyond 20 weeks). The tumours are typically preceded by hyperplasia (simple or atypical), and endometrial intra-epithelial neoplasia (O'Hara and Bell, 2012). The majority of these neoplasms are diagnosed at an early stage and are associated with a favourable prognosis (Lewin et al., 2010).

On a molecular level, ECs are characterized by frequent mutations in *PIK3CA*, *PTEN* and *PIK3R1*, which result in inappropriate activation of the PI3K pathway (Risinger et al., 1997, Rudd et al., 2011). Other signal transduction pathways that are frequently disrupted in these tumours include the RAS-RAF-MEK-ERK pathway with mutations in *KRAS* seen in 18% of cases. Somatic mutations in the *FGFR2* receptor tyrosine kinase occur in 12% of cases with mutations in *FGFR2* and *KRAS* being mutually exclusive (Byron et al., 2012). ECs also frequently show disruption of the canonical WNT signalling pathway with mutations in *CTNNB1* gene (19-45%) (Byron et al., 2012, Machin et al., 2002). It has been suggested that the mutual exclusivity of *CTNNB1* and *KRAS* mutations and functional cross talk between the RAS-RAF-MEK-ERK and WNT/TCF signalling pathways may occur in this cell type or that functional redundancy exists in the biological consequences of altered RAS-RAF-MEK-ERK and WNT/TCF signalling (Byron et al., 2012). Finally, 34- 40% of all ECs show microsatellite instability (MSI) (Byron et al., 2012, Cancer Genome Atlas Research et al., 2013), which is attributed to defective mismatch repair, primarily due to hypermethylation of the MLH1 promoters; somatic mutations in MSH6 and loss of MSH2 expression have also been observed (Esteller et al., 1999, Simpkins et al., 1999, Goodfellow et al., 2003).

Serous endometrial carcinoma

Serous endometrial carcinomas are high grade neoplasms that are relatively rare accounting for only 5-10% of ECs, but are clinically aggressive and contribute substantially to the mortality from endometrial cancer accounting for 39% of deaths from endometrial cancer (Hamilton et al., 2006). Older age and smoking are thought to be the main risk factors. Serous carcinomas arise from surface endometrial intraepithelial carcinoma (Sherman, 2000) on the background of atrophic endometrium in older post-menopausal women. The tumours are characterised by a high frequency of mutations in *TP53*, which are believed to be the initiating events in the development of these cancers (Fadare and Zheng, 2012).

4.1.2.3 **Molecular classification of endometrial cancer**

Advances in next generation sequencing technologies have allowed better characterisation of many types of cancers. In 2013, The Cancer Genome Atlas Research Network (TCGA) published a comprehensive genomic and transcriptomic analysis of endometrial cancers (endometrioid, serous and mixed endometrioid/serous carcinomas) (Cancer Genome Atlas Research et al., 2013). Based on mutation spectra, copy-number alterations (CNAs) and microsatellite instability status, endometrial cancers were classified into four groups (Figure 4.1):

POLE **(ultra-mutated)** cancers that are characterised by extremely high mutation burdens, hotspot mutations in the exonuclease domain of *POLE,* frequent C>A substitutions, few CNAs and recurrent mutations in *PTEN*, *PIK3R1*, *PIK3CA*, *FBXWA* and *KRAS*. These were also found to be associated with favourable outcome;

Microsatellite-instable (MSI) (hypermutated) cancers that are characterised by microsatellite instability due to *MLH1* promoter methylation, relatively high mutation burdens, few CNAs and frequent mutations in *PTEN*, *KRAS* and *RPL22*;

Copy-number low (endometrioid) cancers which comprise microsatellite stable low grade endometrioid cancers with low mutation burden and frequent mutations in *PTEN* and *CTNNB1*;

Copy-number high (serous-like) cancers that are characterised by extensive CNAs, low mutation burdens and recurrent mutations in TP53 as well as FBXW7 and PPP2R1A, but infrequent alterations in PTEN and KRAS.

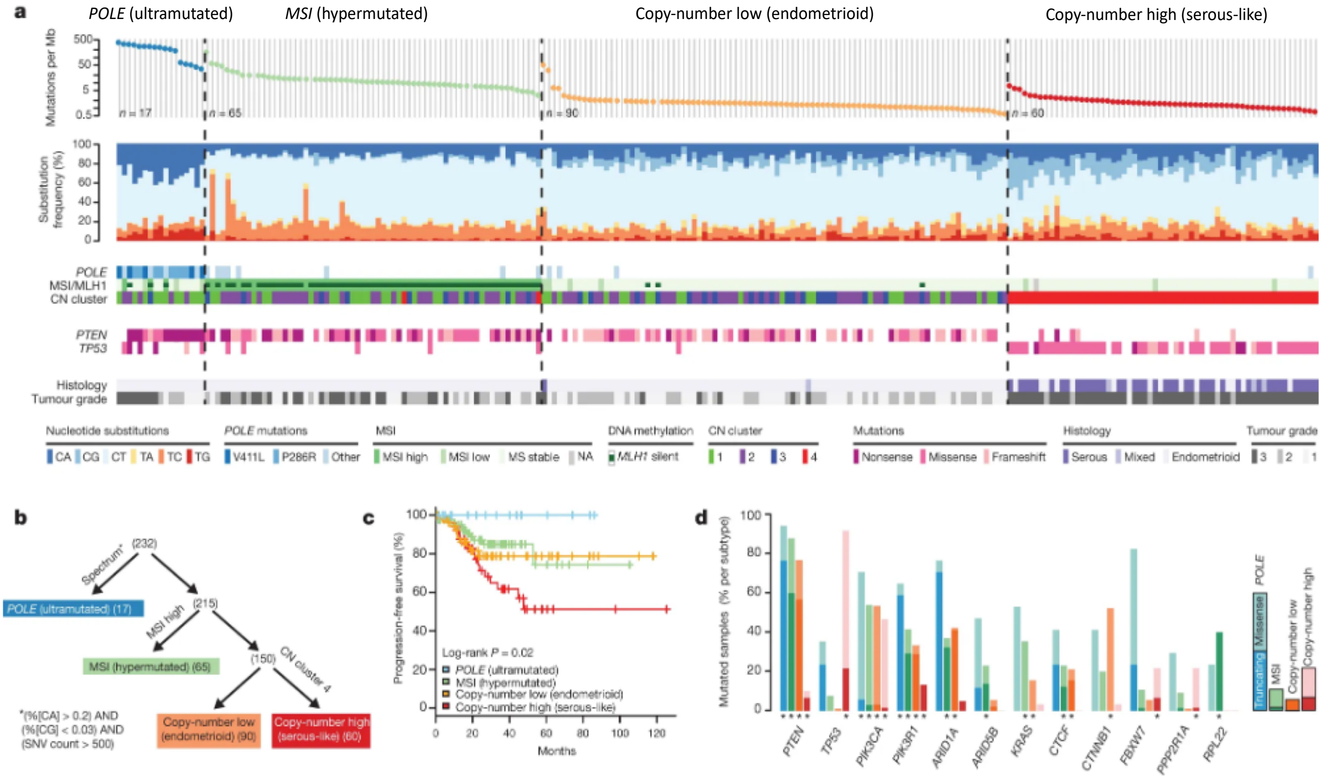


Figure 4.1 | Molecular classification of endometrial cancer. Adapted from G Getz *et al. Nature* 497, 67-73 (2013).

4.1.2.4 Endometrial cancer risk factors

The estimated lifetime risk of being diagnosed with endometrial cancer is 1 in 36 (3%) (CRUK, 2019). The main risk factor for endometrioid carcinomas is exposure to endogenous and exogenous oestrogens in association with obesity, early age of menarche, late-onset menopause, nulliparity, hormone therapy (e.g. tamoxifen) and diabetes. For serous carcinomas tumours, older age (>55 years) and smoking are thought to be the main risk factors although work by McCullough and colleagues have also shown that the incidence increases with elevation in body mass index (BMI) (McCullough et al., 2008). Since the majority of endometrial cancers are endometrioid, some of the key risk factors for the disease are considered in more detail below.

Obesity

Obesity is the second biggest preventable cause of cancer in the UK (CRUK, 2019); it is a well-recognised risk factor for thirteen different types of malignancies, including those arising in the breast, colon, liver, ovary and stomach. In women, obesity has a stronger association with the development of endometrial cancer than with any other cancer type (Reeves et al., 2007) with 34% in the UK and 57% in the US of all such cases attributable to being overweight and obese (Renehan et al., 2008, Calle and Kaaks, 2004). This association is well-documented and shows a dose-response relationship with the cancer incidence increasing with an elevation in the BMI; for every five units of BMI, there is an increase in the risk of developing the disease (relative risk, RR=1.50; $CI_{95\%}$ 1.42-1.59) (World Cancer Research). Furthermore, being obese has an effect on the endometrial cancer prognosis: the RR of disease-specific mortality is 2.53 for mildly obese (BMI 30-34.9) and 6.25 for severely obese (BMI >40) patients (Calle et al., 2003), compared to women with a normal BMI. The underlying mechanistic pathways that link obesity and endometrial cancer are briefly discussed below.

Visceral fat is composed of mature fat cells (adipocytes), less differentiated fat cells (preadipocytes), endothelial, stromal and nerve cells along with mesenchymal stem cells (MSCs) (Mills et al., 2012, p.1071); it serves as an endocrine organ that is responsible for the synthesis and secretion of several hormones amongst multiple other functions (Coelho et al., 2013). During reproductive years, ovaries are the primary source of oestrogens; whereas in post-menopausal women other tissues, in particular visceral fat, become key sites of production and secretion of these hormones (Davis et al., 2015). Adipocytes, preadipocytes and MSCs secrete aromatase, an enzyme that is necessary for the conversion of androgens to oestrogens (Blakemore and Naftolin, 2016, O'Connor et al., 2009). In addition, an increase in adiposity (the state of being fat), leads to a decrease in sex hormone-binding globulin (SHBG) levels which in turn results in an increase in the pool of available, bioactive oestrogen (Simo et al., 2015). When oestrogen is bound to oestrogen alpha and/or beta-receptor, it directly modulates the transcription of a variety of pro-proliferative genes including *IGF1R* and *IGF1* (Westin et al., 2009).

Visceral fat is also a rich source of adipokines, which regulate metabolism and modulate chronic inflammatory states associated with adiposity. Obesity-associated proinflammatory adipokines, including leptin, interleukin-6 and tumour necrosis factor alpha, not only suppress normal insulin signalling and contribute to insulin resistance (Onstad et al., 2016, Renehan et al., 2015, Mu et al., 2012, Kwon and Pessin, 2013), but also promote endometrial proliferation (Onstad et al., 2016).

Finally, Type 2 diabetes which is closely linked with obesity, is characterized by elevated levels of insulin and insulin-like growth factor 1 (*IGF1*) and hyperglycaemia, both of which have been shown to play a role in the pathogenesis of endometrial cancer (Nead et al., 2015, Poloz and Stambolic, 2015). In premenopausal women, oestrogen-induced cyclical changes in *IGF1* expression and signaling modulate endometrial proliferation during normal menstrual cycle (McCampbell et al., 2006). The positive association of endometrial cancer with hyperinsulinaemia and type 2 diabetes is well documented (Nead et al., 2015, Calle and Kaaks, 2004, Lees and Leath, 2015). Increased expression of insulin and *IGF1* receptors is observed in endometrial hyperplasia, which heightens the responsiveness of these cells to insulin and *IGF1* (McCampbell et al., 2006) and promotes hyperactivity of MAPK and PI3K/AKT/mTOR signaling frequently observed in endometrial cancer. Proliferative signaling is further amplified by the loss of the *PTEN* tumor suppressor gene, which acts in opposition to the PI3K/AKT/mTOR pathway and is an early event in the pathogenesis of endometrial cancer. Finally, hyperglycaemia, which occurs as a consequence of insulin insensitivity, serves to further fuel the growth of metabolically active tissue (Masur et al., 2011), including endometrial hyperplasia and cancer.

Parity

A meta-analysis by Wu and colleagues (Wu et al., 2015) has revealed that there is a significant inverse association between parity and risk of endometrial cancer with a RR for parous versus nulliparous women of 0.69 ($CI_{95\%}$ 0.65-0.74). In addition, parity number of 1, 2 or 3 versus nulliparous showed a significant negative association with the relative risk, RR=0.73 ($CI_{95\%}$ 0.64-0.84), RR = 0.62 ($CI_{95\%}$ 0.53–0.74); and RR = 0.68 ($CI_{95\%}$ 0.65–0.70) respectively). Oestrogens are known to stimulate endometrial proliferation and increase mitotic activity, which can lead to tumour development (Henderson and Feigelson, 2000, Akhmedkhanov et al., 2001) while progestins reduce cell proliferation and promote differentiation and can therefore decrease risk of endometrial cancer (Akhmedkhanov et al., 2001). Wu and colleagues suggested that the observed negative correlation between parity and risk of endometrial cancers is due to slightly higher levels of progesterone relative to oestrogen during pregnancy (Wu et al., 2015). They also proposed that the dose-response relationship observed between parity and endometrial cancer risk may be attributable to repeated long-term anti-oestrogenic endometrial effects of progesterone occurring during pregnancies (Preston-Martin et al., 1990), or alternatively, due to 'mechanical shedding of malignant/premalignant endometrial cells' at parturition (Wu et al., 2015, Lambe et al., 1999).

4.1.2.5 **Microbiome**

In recent years, there has been an increasing amount of interest in the uterine microbiota with several studies reporting its association with various disease states including infertility and cancer (Walther-Antonio et al., 2016, Chen et al., 2017, Baker, 2018). Bacterial organisms that were previously found to be enriched in endometrial cancer cases are summarised in Table 4.1. However, there are limitations associated with such investigations, of which contamination is probably one of the most significant (Baker, 2018). The majority of uterine sampling in the published work and in some of the cases in our study, would have been performed transcervically, which makes it difficult to avoid cross-contamination by microbiota from the lower genital tract. In cases where samples were collected in different circumstances, such as the transplant and autopsy donors in our study, contamination with organisms from the lower abdominal and pelvic cavities may also occur. The use of uterine manipulators, cervical dilators and surgical tools as well as histology tissue processing equipment may further contribute to cross-contamination.

Phylum	Genus
Actinobacteria	Atopobium
Bacteroidetes	Porphyromonas
Bacteroidetes	Bacteroides
Firmicutes	Anaerostipes
Firmicutes	ph2
Firmicutes	Peptoniphilus
Firmicutes	1-68
Firmicutes	Anaerotruncus
Firmicutes	Dialister
Firmicutes	Ruminococcus
Proteobacteria	Arthrospira
Spirochaetes	Treponema

Table 4.1 | List of bacterial organisms previously shown to be associated with endometrial cancer. Data extracted from Walther-Antonio et al, Genome Medicine, 2016.

4.2 Results

The results presented in this chapter are based on the final variants that were called in normal endometrial glands from 28 women aged 19 to 81 years. Only samples with ≥15-fold coverage were processed through the variant calling pipeline (n=257, Appendix 10). Our mutation burden and signature analyses are based on the same variants that were used for reconstruction of phylogenetic trees, which mitigates double counting and differentiates between shared and unique variants, which is crucial for timing driver events.

4.2.1 Mutation burden

The somatic mutation burden in normal endometrial glands from the 28 individuals ranged from 209 to 2833 base substitutions (median 1,521) and 1 to 358 indels (median 180) (Figure 3.2a, b). In large part this variation was attributable to the ages of the individuals with a linear accumulation of ~29 base substitutions per gland per year during adult life (linear mixed-effect model, $CI_{95\%}$ 23-34 , p = 3.02 x 10^{-11}) (Appendix 9). However, the possibilities of lower mutation rates pre-menarche and post-menopause cannot be excluded. Positive driver mutation status conferred an additional ~110 substitutions ($CI_{95\%}$ 43-177, p = 1.34 x 10^{-3}). The basis for this correlation is unclear. It is conceivable that an elevated total mutation load increases the chances of including, by chance, a driver. It is also plausible, however, that drivers engender biological changes, for example elevated cell division rates, that result in higher overall mutation loads. There was no obvious correlation between parity and total somatic mutation burden.

In addition, to formally test the effect of “sample cohort” on our observations, we applied a mixed-effect model; the analysis showed that “cohort” (i.e. whether the sample was from a transplant donor/autopsy or from the infertility clinic) had no significant effect on the clonality and mutation burdens of normal endometrial glands (Appendix 9).

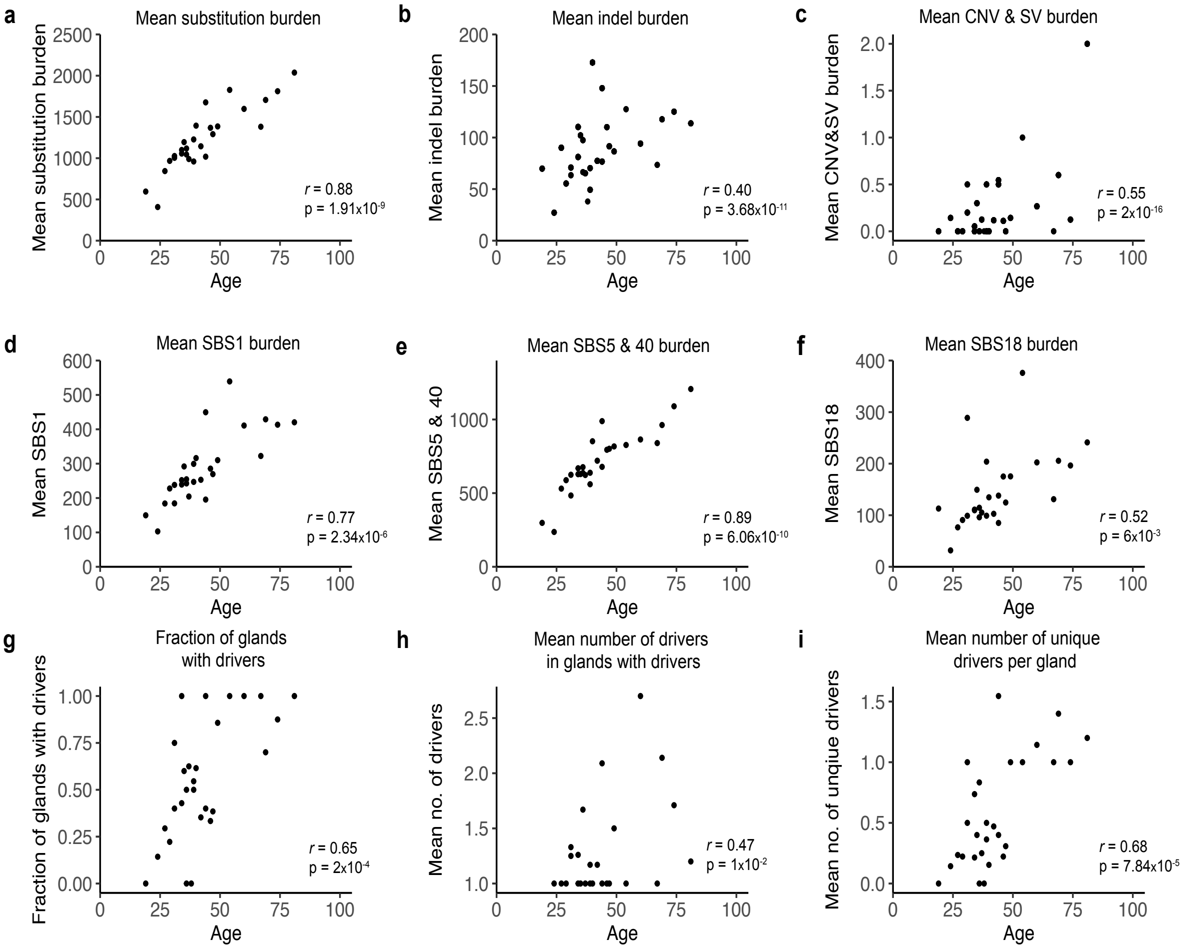


Figure 4.2 | Mutation burden in normal endometrial glands. (a) Substitutions accumulate in the endometrium in a relatively linear fashion. A positive correlation between age and accumulation of indels **(b)**, copy number alterations (CNA) and structural variants (SV) **(c)** and mutations attributed to single base substitution (SBS) mutational signatures SBS1 **(d)**, SBS5/40 **(e)** and SBS18 **(f)** was also observed. The fraction of glands with driver mutations **(g)**, mean number of driver mutations in glands with drivers **(h)** and mean number of unique (different) driver mutations per gland **(i)** all show positive correlation with age.

4.2.1.1 **Coding/non-coding mutation burden**

To explore whether coding and non-coding mutations accumulate at different rates all substitutions were split into the two groups. Our analysis shows that there is an age-associated accumulation of somatic mutations for both types of mutations (linear regression, p = 1.22 x 10^{-6} for coding mutations and p = 4.73 x 10^{-10} for non-coding mutations (Figure 4.3a,b). The median ratio of coding to non-coding mutations was 0.011 (range 0.007 - 0.015) (Figure 4.3c); there was no association with age (r = -0.024; linear regression p = 0.904).

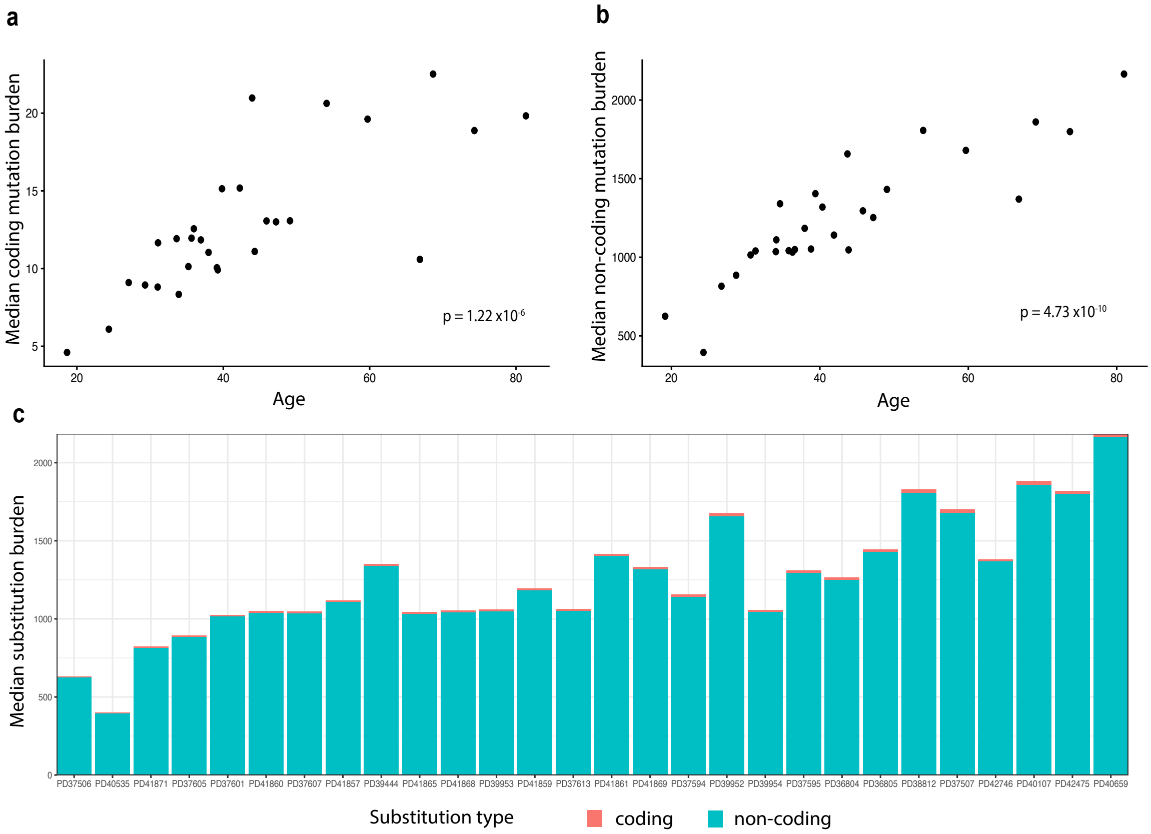


Figure 4.3 | Age-associated accumulation of coding and non-coding mutations in normal endometrial epithelium. For each sample, substitutions were divided into coding and non-coding. **(a)**, **(b)** Both types of mutations accumulated with age. **(c)** The median ratio of coding/non-coding mutations was 0.011 (range 0.007-0.015); there was no correlation with age.

4.2.2 Endometrium and other normal tissues

Tissues across the body differ in their physiology, turnover, exposure to mutagens and architecture with specific stem cell arrangements and dynamics. All of these are likely to be reflected in their mutational landscapes, including mutation burdens. We therefore compared mutation burden of endometrial epithelium to that of other normal tissues.

4.2.2.1 In-house LCM experiments: endocervix

In addition to endometrial glands, nearby normal endocervical glands were micro-dissected from one individual (PD37506). In this analysis, for each cell type, only the samples with a median VAF of ≥0.4 were used. There was a ~2-fold lower somatic mutation burden in endocervical than endometrial glands (Figure 4.4). The finding may reflect the absence, in endocervical glands, of the cyclical process of loss and regeneration that occurs in endometrial glands.

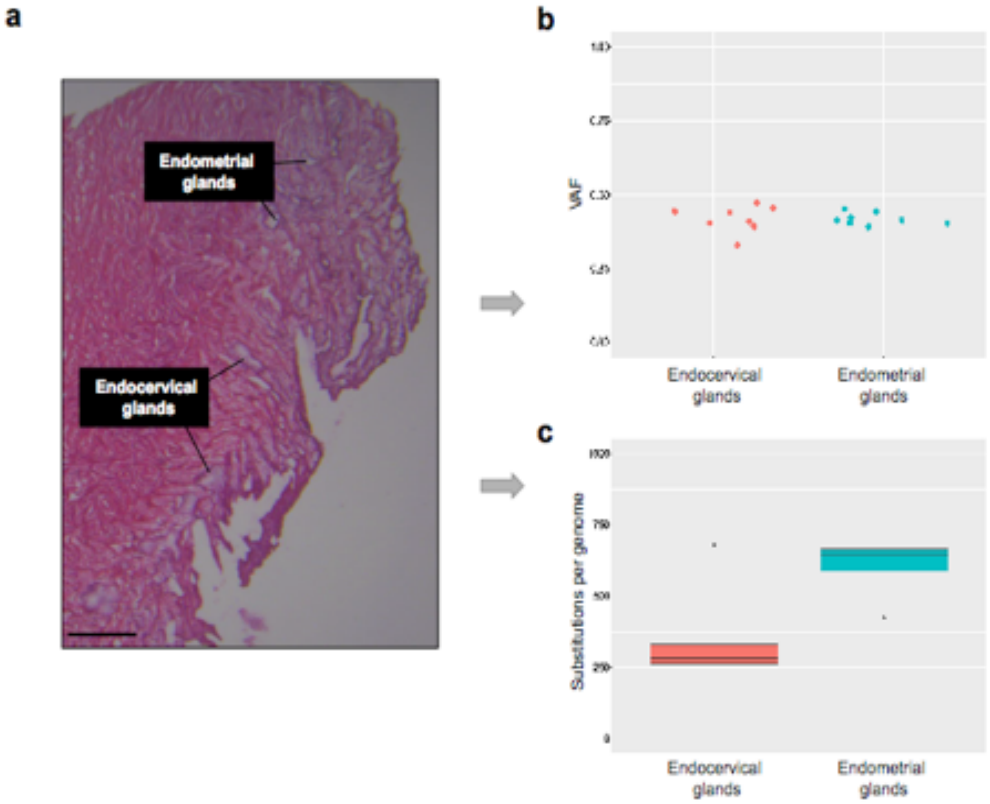


Figure 4.4 | Comparison between normal endometrial and endocervical glands. (a) An overview histology image of an ~2cm^3 tissue biopsy sample from a 19-year-old donor (PD37506). The image shows normal endometrial and adjacent endocervical glands, which were subsequently micro-dissected. **(b)** Endometrial and endocervical glands with a similar median variant allele frequency (VAF) of substitutions were compared. **(c)** There was a ~2-fold difference in the mutation burden between the two types of glands.

4.2.2.2 **Previously published data (non-LCM experiments)**

Using previously published results, we compared mutation rates between normal endometrial epithelial and other types of cells. The results showed that endometrial cells exhibit lower mutation rates than normal skin epidermal (Martincorena et al., 2015), colorectal (Lee-Six et al., 2019, Blokzijl et al., 2016), small intestinal (Lee-Six et al., 2019, Blokzijl et al., 2016) and liver cells (Blokzijl et al., 2016), similar burdens to oesophageal cells (Martincorena, 2018) and higher rates than skeletal muscle cells (Franco et al., 2018) (Figure 4.5). Of the mutational signatures found in endometrial cells, SBS1 and SBS5 are found in all other cell types (Alexandrov et al., 2015). However, the SBS1 mutation rate is higher in colorectal and small intestinal epithelial cells whereas the SBS5 mutation rate is higher in liver cells (Blokzijl et al., 2016). SBS18 has also been found ubiquitously in colonic crypts(Lee-Six et al., 2019).

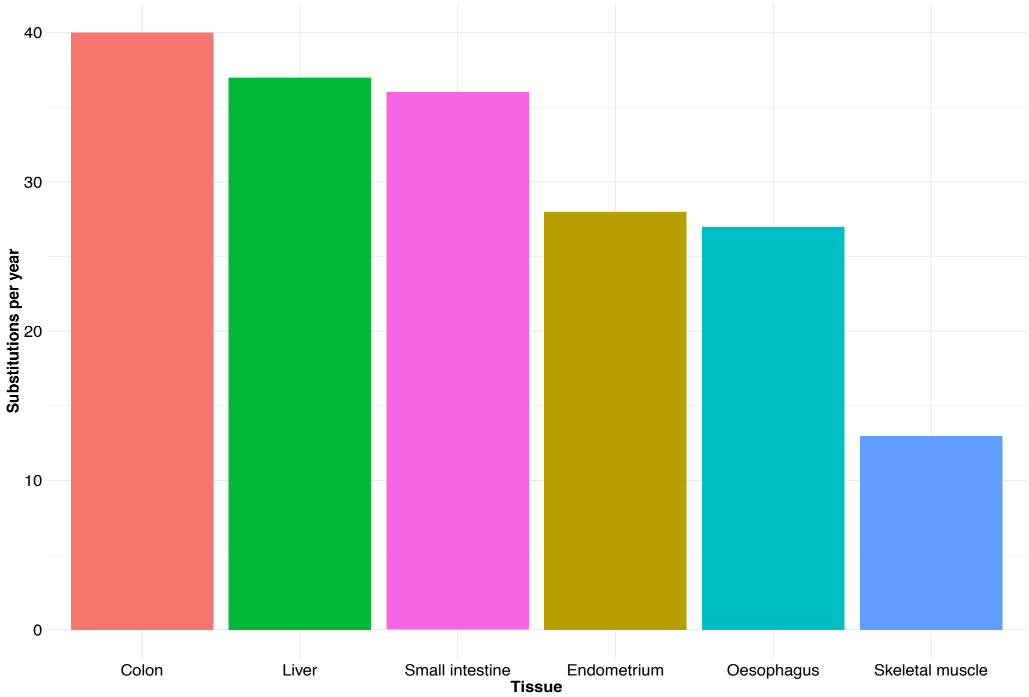


Figure 4.5 | Comparison of mutation rates between endometrial epithelium and other cell types. The barplot shows a comparison of estimated mutilation rates (substitutions) for normal endometrial epithelial and other cell types from previously published studies (liver, colon and small intestine (Blokzijl et al., 2016), oesophagus (Martincorena et al., 2018) and skeletal muscle (Franco et al., 2018).

4.2.3 Normal endometrium and cancer

Acquisition of driver mutations enables uncontrolled proliferation, cancer clone expansion and increased accumulation of somatic mutations. We therefore compared mutation burdens between normal endometrial glands and endometrial cancer. We performed the following analyses:

1. **Raw mutation burden comparison between normal endometrial glands and tumour using endometrial cancer samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) set.** We compared the mutation loads of normal cells observed here with those recently released by the Pan Cancer Analysis of Whole Genomes Project[2]. Endometrial cancers exhibited higher mutation loads than normal endometrial cells, for base substitutions (~5-fold, medians of 1346 and 7330 substitutions observed in normal endometrium and endometrial cancer respectively (Mann-Whitney test, $P = 7.63 \times 10^{-6}$) (Fig. 4.6a) and indels (Figure 4.6b) and these differences also pertain to normal endometrial cells with driver mutations. In most endometrial cancers these differences are attributable to higher mutation burdens of the ubiquitous base substitution and indel mutational signatures. In addition, however, the very high mutation loads of the subsets of endometrial cancers with DNA mismatch repair deficiency and polymerase epsilon/delta mutations were not seen in normal endometrial cells. Differences between endometrial cancers and normal cells were even more marked for structural variants and copy number changes (median number zero in normal endometrial cells and ~23 in endometrial cancers) and this again pertained to normal endometrial cells with drivers.

2. **Tumour (Pan Cancer Analysis of Whole Genomes (PCAWG)) and normal mutation burden comparison using subsampled sequencing data:** These analyses were performed with the help of Tim Coorens and Stefan Dentro. We selected five tumour (PCAWG) and five normal endometrial gland samples with a similar clonal composition (clonal composition was inferred with dpclust (Nik-Zainal et al., 2012) and only samples that had a clonal fraction of mutations of ≥0.8 were included in this analysis). Binary Alignment Map (BAM) files were subsampled at regular

fractions (0.1) of the original coverage to assess the sensitivity of mutation calling across comparable levels of coverage in both cancer and normal samples; when a mutation called in the original BAM file was present in a subsampled BAM file in four or more reads, it was taken to be present in the subsampled BAM file. The results showed that ≥90% of substitutions detected at the original coverage were recovered at a median coverage of 22.1x for tumour (range 21.4-43.4x) and 20.1x for normal endometrial glands (range 18.6-24.2x) (Figure 4.7a). Comparison of the mutation burden between normal and tumour samples at the sequencing coverage of 25-30x, showed an ~4-fold difference (Mann-Whitney test, p = 0.00794, Figure 4.7b), therefore it is highly unlikely that such a marked difference is due to the depth of coverage alone.

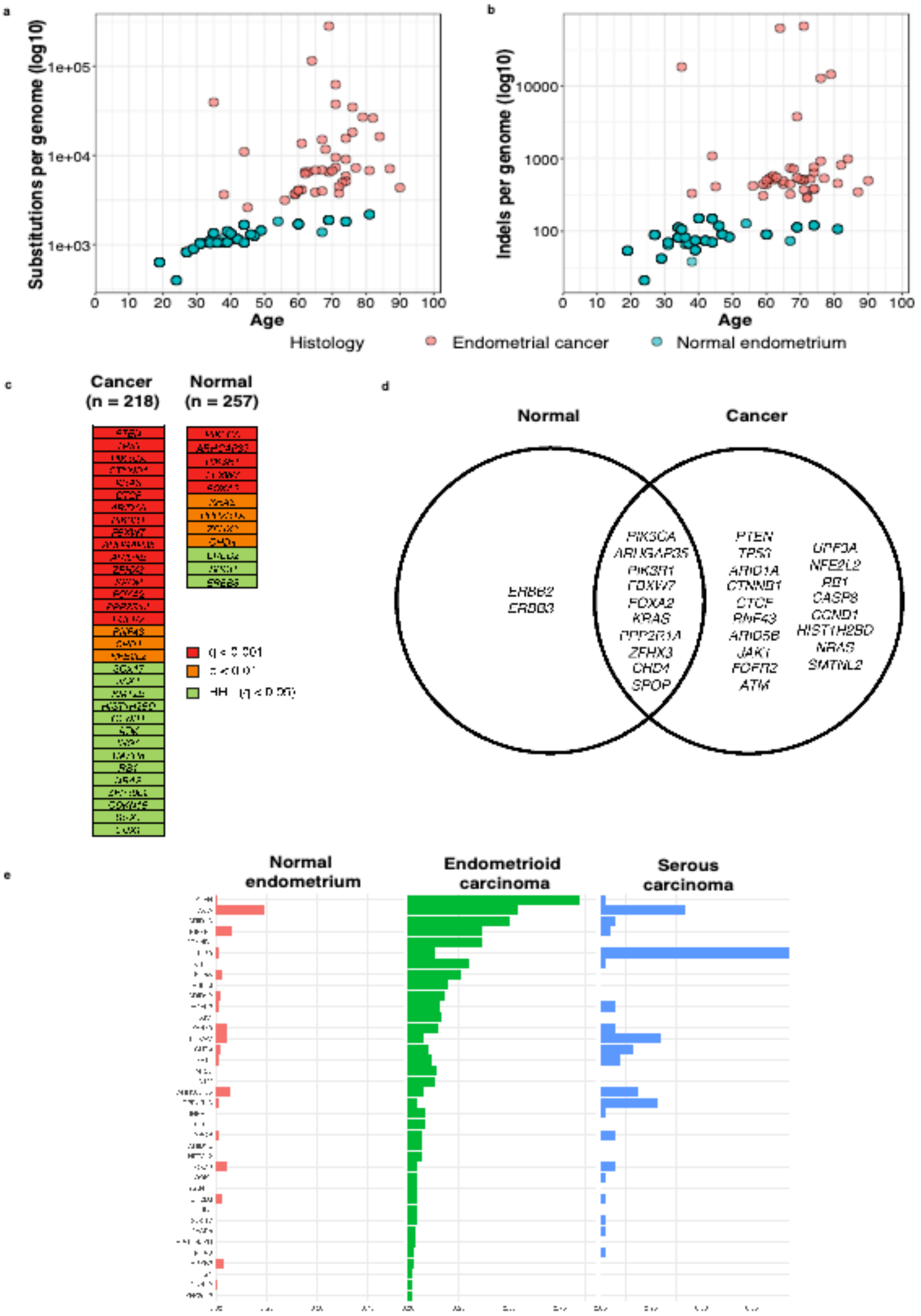


Figure 4.6 | Comparison between normal endometrial epithelium and endometrial cancer. (**a**,**b**) Normal endometrial glands show lower total mutation burden in comparison to endometrial cancer (Pan Cancer Analysis of Whole Genomes Project[2]). (**d**,**e**) Genes that are under significant positive selection (*dN/dS* > 1) in normal endometrial epithelium and endometrial cancer. RHT, restricted hypothesis testing of known cancer genes. *ERBB2* and *ERBB3* are under selection in normal endometrial epithelium, but are not in endometrial cancer. (**f**) Identified driver mutations and their distribution in normal endometrial glands and the two major types of endometrial cancer (endometroid and serous carcinomas).

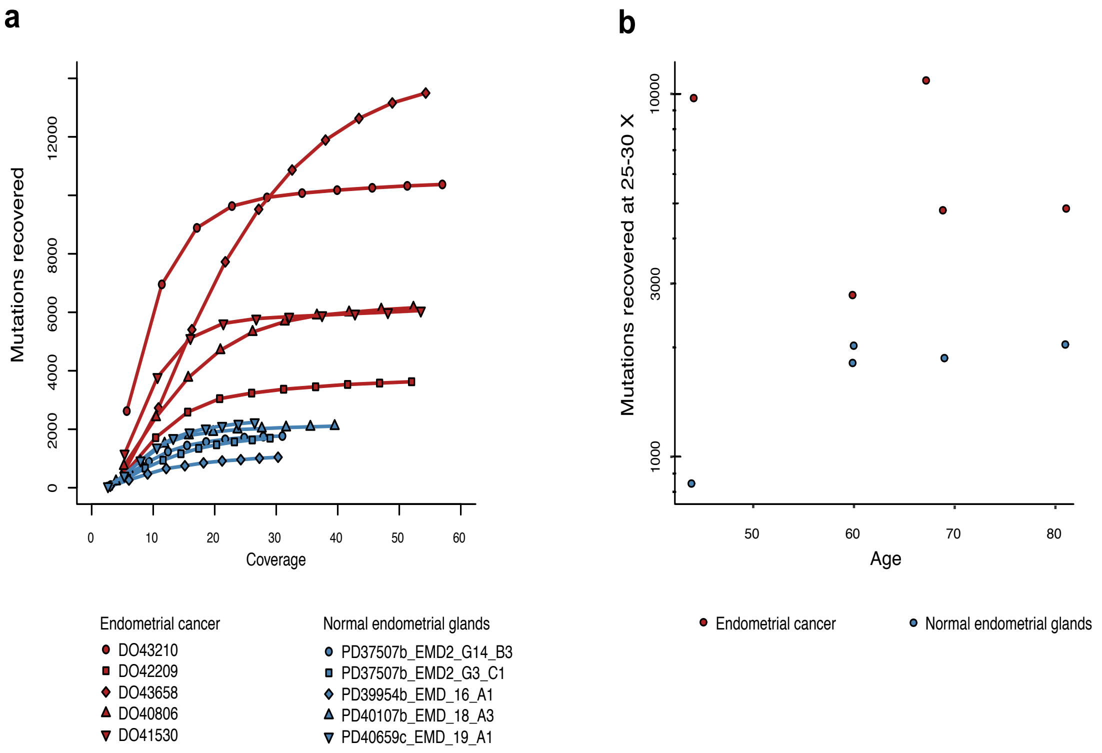


Figure 4.7 | Comparison of mutation burden in subsampled tumour and normal endometrial samples. Five tumour and five normal samples with a clonal fraction of mutations of ≥0.8 were selected, bam files subsampled at a regular interval of the original coverage. **(a)** ≥90% of the mutations called at the original coverage were recovered at a median coverage of 22.1x for tumour (range 21.4-43.4x) and 20.1x for normal endometrial gland samples (range 18.6-24.2x). **(b)** Comparison of the mutation burden between normal and tumour samples at the sequencing coverage of 25-30x, showed an ~4-fold difference.

3. **Tumour (PCAWG) and normal comparison using clonal mutations only:** In addition to the original analysis of the total substitution burden in the normal and tumour (PCAWG) cases (Figure 4.8a), we made this comparison using 'clonal' mutations only (clonal composition of each sample was inferred with dpclust (Nik-Zainal et al., 2012) by Stefan Dentro) (Figure 4.8b). Given that the majority of the endometrial cancer samples are from women aged 60 to 80 years, we also performed an age-restricted comparison using cases from the two decades (Figure 4.8c); the results show a significant difference in the clonal substitution burden between normal and cancer samples (Wilcoxon rank sum test, $p = 4.02 \times 10^{-14}$).

4. **Tumour (TCGA) and normal mutation burden comparison:** Given the above-mentioned differences in pathogenesis, molecular changes and clinical outcomes between the two types of endometrial cancer, we also compared mutation burden of normal endometrial glands and the two classes cancer. Ideally, we would have liked to have performed both comparisons using whole genome sequencing (WGS) data from the same PCAWG data set. Unfortunately, no histology information was available for this cohort and so the cancer cases could not be subtyped in the total mutation burden comparison. Instead, we compared coding mutations in normal endometrial glands and endometrial cancer samples from TCGA. There was a 6-fold and 5-fold difference in the mutation burden comparing to endometrioid and serous carcinoma respectively (Figure 4.9).

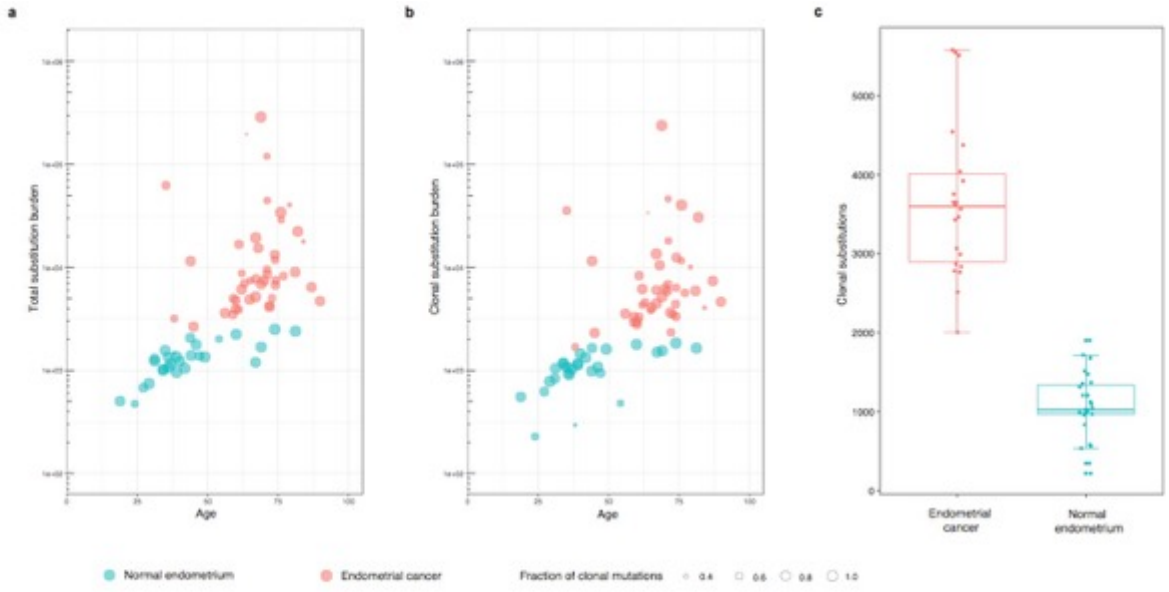


Figure 4.8 | Mutation burden comparison in normal endometrium and endometrial cancer. (a) Scatter plot showing all substitutions identified in normal endometrial glands and endometrial cancer cases (Pan Cancer Analysis of Whole Genomes (Alexandrov, 2018)); **(b)** comparison of clonal mutations (clonal substitutions are those that were assigned to the major clone using dpclust method (Nik-Zainal et al., 2012)); **(c)** boxplot showing clonal substitution burden in tumour and normal endometrium restricted to donors aged 60 to 80 years (Wilcoxon rank test, p = 4 x 10-14). In **(a)** and **(b)**, median mutational burden is calculated for each donor; in **(c)**, all samples are included for each of normal tissue donor and ‘hypermutator’ cancer cases (defined as those above the 75 percentile (>5,631 substitutions)) were excluded in this analysis.

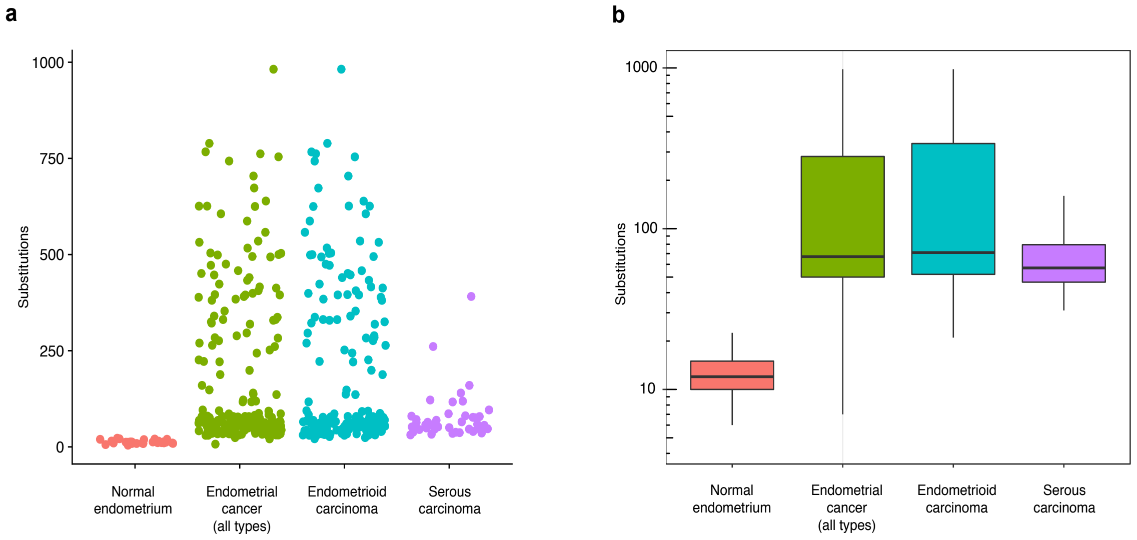


Figure 4.9 | Coding mutation burden comparison in normal endometrium and endometrial cancer (TCGA cohort). (a) Scatter plot showing mutation burden for tumour and normal samples. For cancer cases, each data point represents an individual donor; for normal endometrial samples, each data point represents a median burden for an individual donor. **(b)** Somatic mutation burden in normal endometrium is 6-fold and 5-fold lower than that of endometrioid and serous carcinoma respectively.

4.2.4 Mutational signatures

To explore the underlying processes of somatic mutagenesis operative in normal endometrial epithelial cells mutational signatures were analysed. Five previously described single base substitution (SBS) mutational signatures were identified in endometrial glands (Figure 4.10 and Appendices 8 and 9): SBS1, predominantly characterised by NCG>NTG mutations and likely due to spontaneous deamination of 5-methylcytosine; SBS5 and SBS40, two relatively featureless, 'flat' signatures of uncertain cause; SBS18, predominantly characterised by C>A substitutions and possibly due to reactive oxygen species (Rouhani et al., 2016); and SBS23, a signature predominantly composed of C>T mutations and of unknown aetiology. Because SBS5 and SBS40 are relatively featureless they present particular challenges in estimating their separate contributions (as previously outlined (Alexandrov, 2018)) and have therefore been combined (but shown separately in Appendices 8 and 9). SBS23 was previously found in a small number of liver cancers at high mutation burdens. Given its low mutation burden and small contribution here it is unclear whether this is really the same signature and we have therefore included it in the "unattributable" category. The mean signature exposures were 0.23 for SBS1, 0.58 for SBS5/40 and 0.12 for SBS18. A positive linear correlation with age for the mutation burdens attributable to SBS1, combined SBS5/40 and SBS18 signatures was observed (Figure 4.2).

Interestingly, glands from one donor with a history of recurrent missed miscarriage (RMM) showed much higher mean SBS18 exposure (0.35) compared to the rest of the cohort. As SBS18 has been shown to be associated with base excision repair (BER) deficiency we searched for truncating (somatic and germline) mutations in all samples from the 31-year-old donor (PD37601). In this analysis we used a panel of twenty five genes associated with BER (Table 4.2), but no such variants were identified.

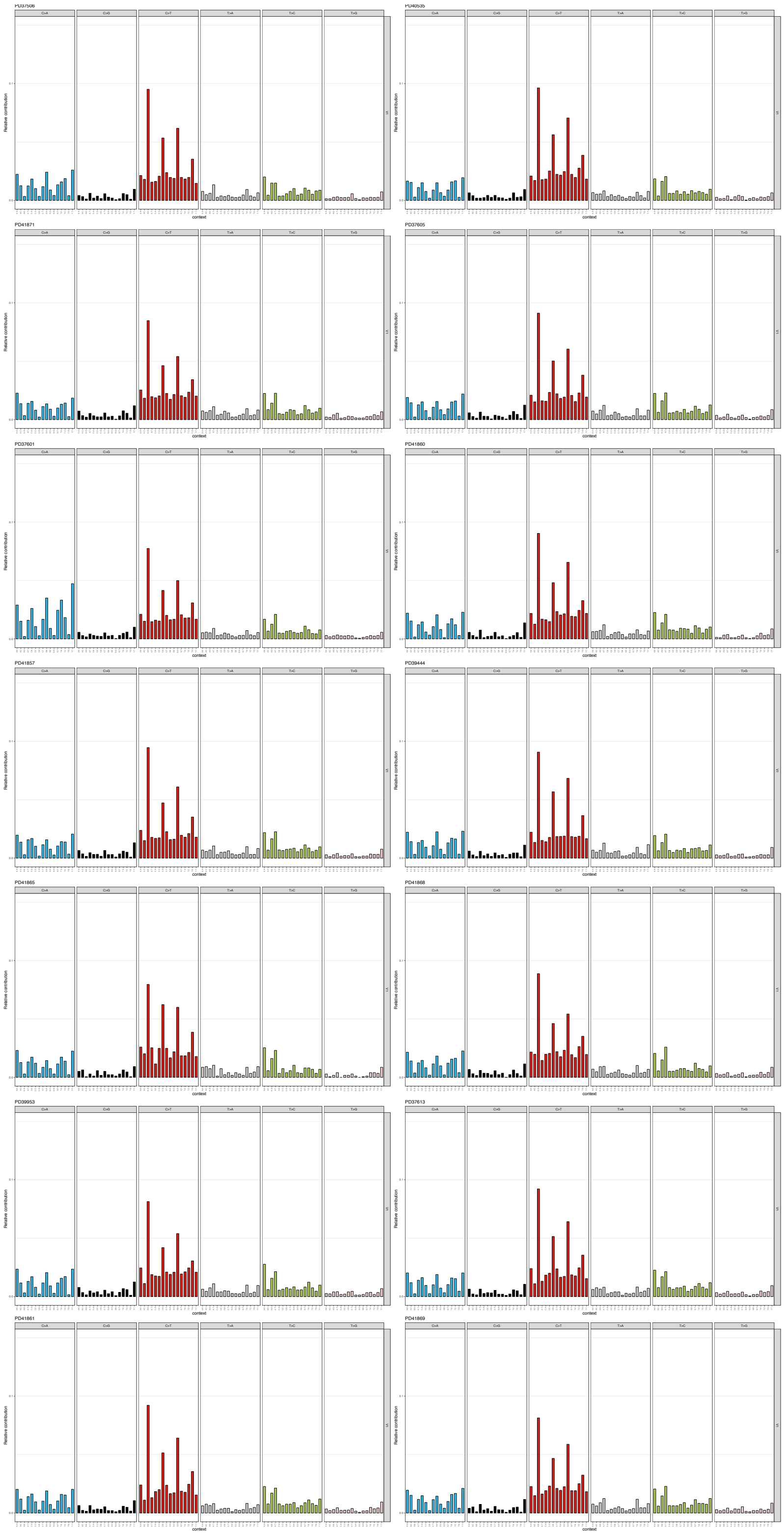


Figure 4.10 | Composite mutational spectra for selected fourteen donors. Composite mutational spectra for twenty seven donors were first generated using single base substitutions identified in all glands from each individual.

APEX1	OGG1
APEX2	PARP1
APLF	PARP2
DUT	PNKP
LIG3	POLB
MBD4	RECQL4
MPG	SMUG1
MUTYH	TDG
NEIL1	TDP1
NEIL2	UNG
NEIL3	WRN
NTHL1	XRCC1
NUDT1	

Table 4.2 | Panel of twenty five genes that were used to screen for base excision repair deficiency.

We currently do not know whether different signatures operate at different times of life. Therefore, to ascertain the periods during which different mutational processes operate, phylogenetic trees of endometrial glands were constructed for each individual using somatic mutations (Figures 4.11, 4.12 and 4.13). These revealed that the mutational processes underlying the three signatures are active throughout life.

With respect to small indels, composite mutational spectra for each donor were generated. These were similar across ages and showed that single T insertions at runs of T bases were the most common mutation type observed. However, due to the relative sparsity of indels in normal endometrial glands, formal signature extraction was not performed (Appendix 8).

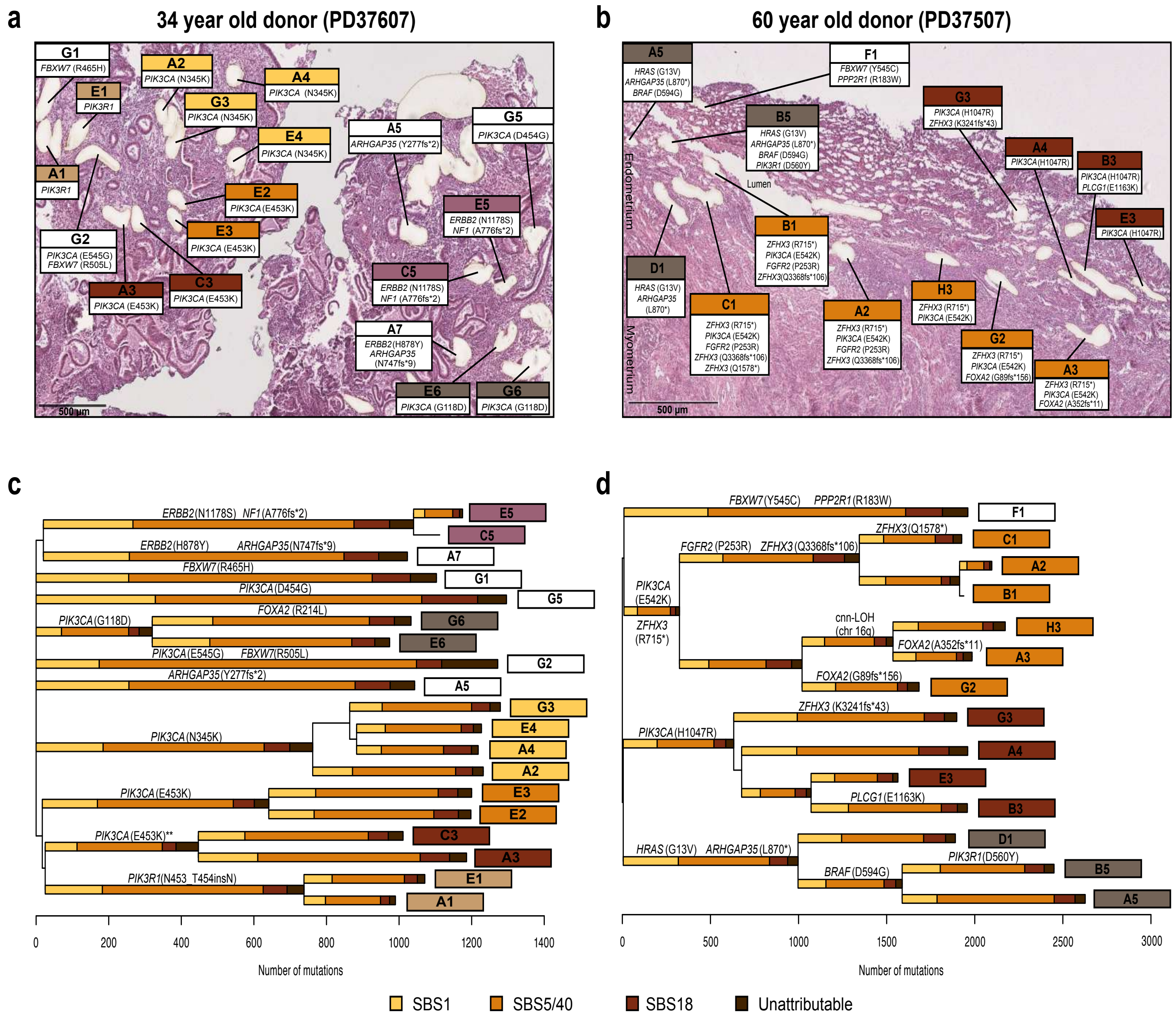


Figure 4.11 | Histology images and reconstructed phylogenetic trees for two individuals in whom every normal endometrial gland contained at least one driver mutation: 34 year old (a,b) and 60 year old (c,d). (**a,b**) Haematoxylin and eosin (H&E) images of endometrial glands were taken after laser-capture microdissection (20x magnification). (**c,d**) Phylogenetic trees were reconstructed using single base substitutions; the length of each branch is proportional to the number of variants; a stacked barplot of attributed single base substitution (SBS) mutational signatures that contributed to each branch is then superimposed onto every branch; signature extraction was not performed on branches with less than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only as it is not possible to time different signatures within individual branches. Glands sharing over 100 variants were considered part of the same clade (indicated by the colour of the sample ID label). Glands that did not belong to any clades are in white. SBS signatures are colour-coded; substitutions that were not attributed to the reference signatures and those attributed to SBS23 are shown as ‘Unattributable’.

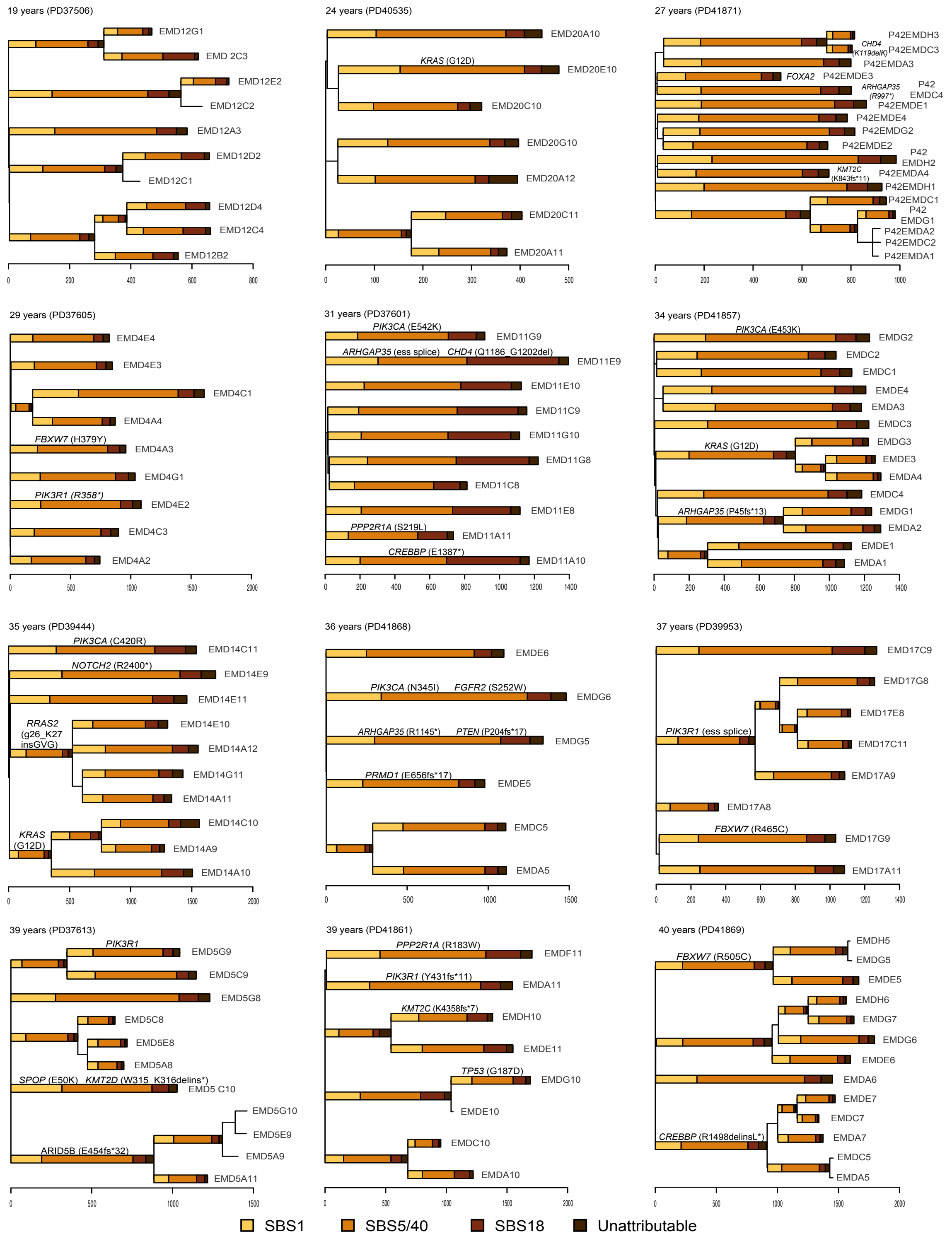


Figure 4.12 | Phylogenetic trees of endometrial glands for donors aged 19-40 years. Phylogenetic trees for the other twelve donors were reconstructed also using single base substitutions with branch length proportional to the number of variants; the stacked bar plots represent attributed SBS mutational signatures that contributed to each branch. Signature extraction was not performed on branches with less than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only as it is not possible to time different signatures within individual branches.

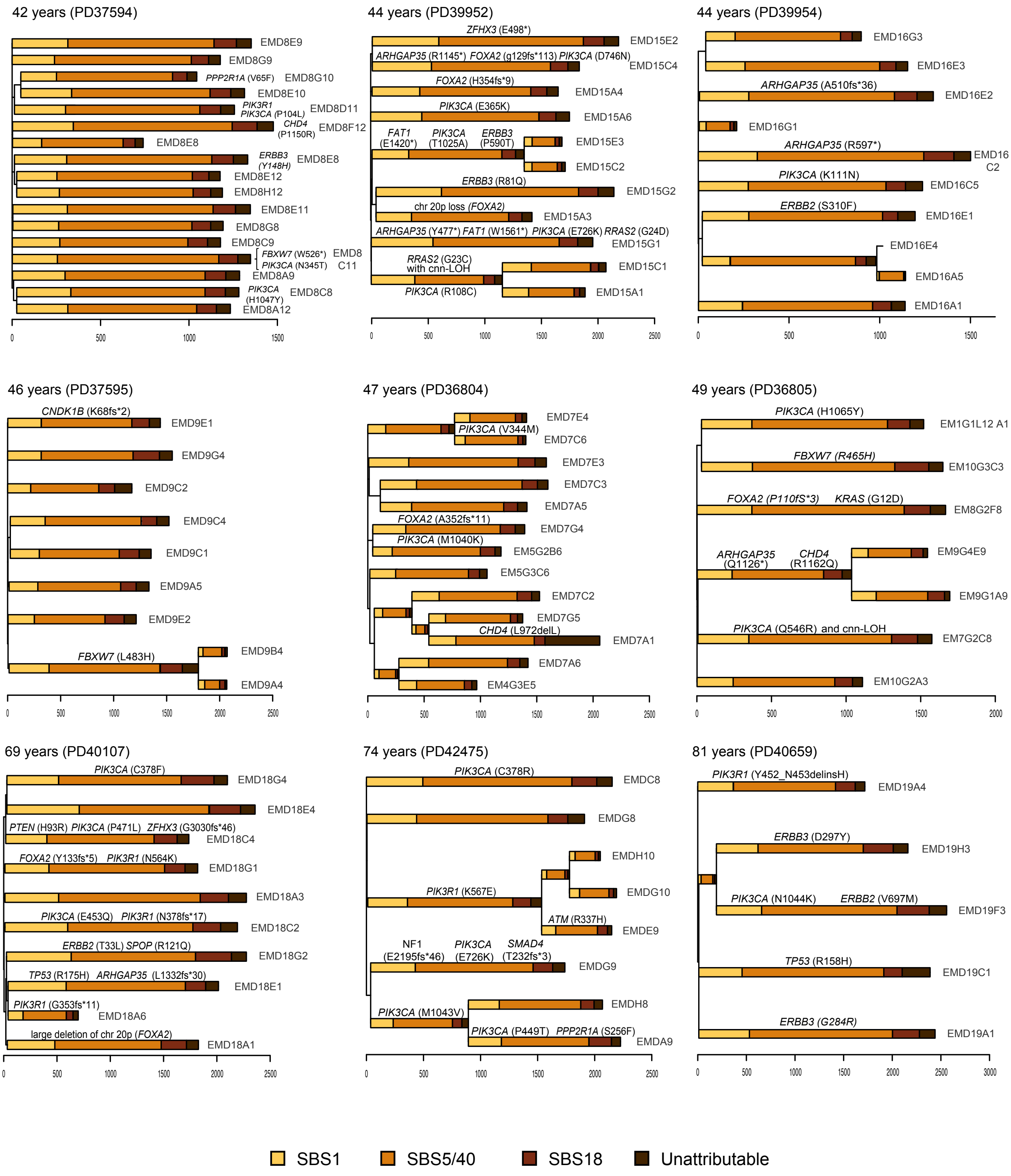


Figure 4.13 | Phylogenetic trees of endometrial glands for donors aged 42 to 81 years. Phylogenetic trees for twelve donors aged 42 to 81 years were also reconstructed as described above. Every single gland from donors PD39952 (44 year old) and PD40659 (81 year old) had at least one driver mutation.

4.2.5 Copy Number and Structural Variants

Serous endometrial carcinomas are characterised by relatively low mutation rates, but extensive CNAs (Cancer Genome Atlas Research et al., 2013). In our cohort, somatic CNAs and structural variants (genome rearrangements) were found in only 27 out of 182 (15%) normal endometrial glands (Figure 4.2, 4.14 and Appendix 13). These included copy number neutral loss of heterozygosity (cnn-LOH) in six glands, whole chromosome copy number increases in one and structural variants in eighteen (12 large deletions, six tandem duplications and nine translocations). The rates are similar to those observed in normal colon with CNAs and/or structural variants seen in ~18% of normal colonic crypts(Lee-Six et al., 2019).

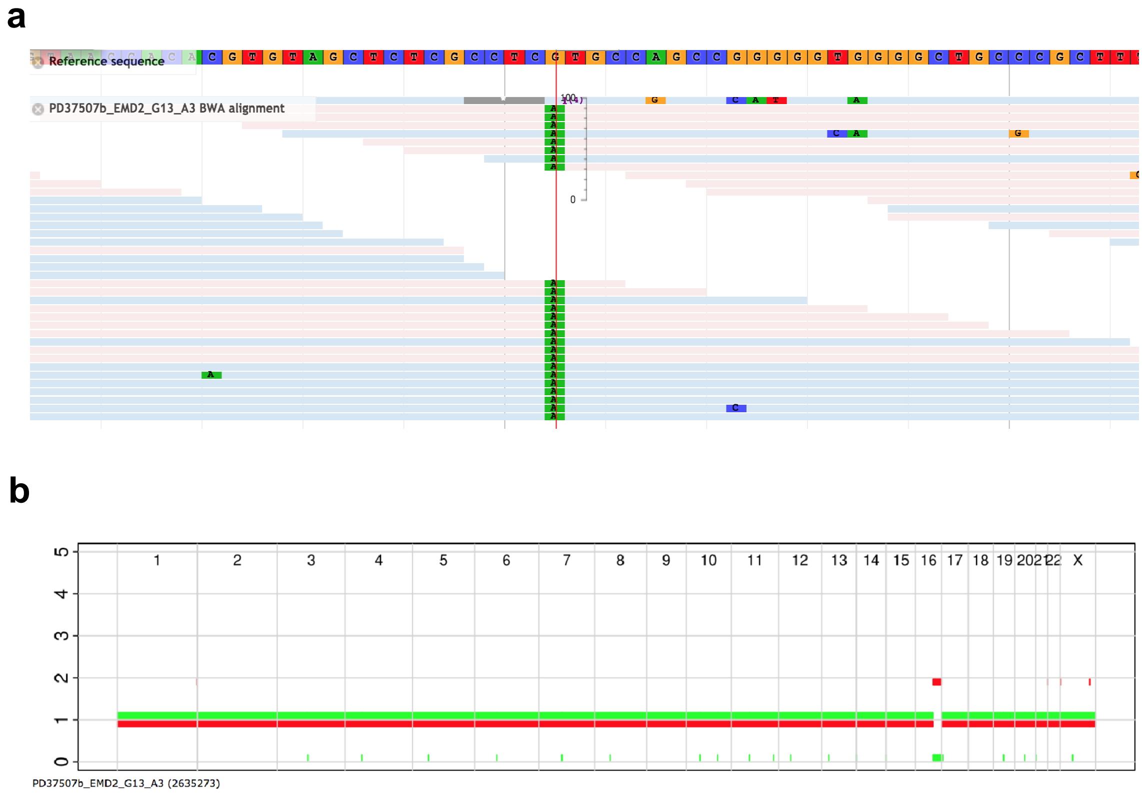


Figure 4.14 | An example of copy-number neutral loss of heterozygosity (cnn-LOH) in a normal endometrial gland. (**a**) biallelic truncating mutation is seen in *ZFHX3* (p.R715*) with every read carrying the variant. (**b**) an associated cnn-LOH is observed on chromosome 16.

The majority of glands showed no change; of those with CNAs/SVs, showed a single change. However, one of two glands carrying a *TP53* mutation (see below) exhibited nine structural variants, indicating that genomic instability caused by defective DNA maintenance occurs in normal cells. Although the observation is only seen in one donor, there are two reasons why we believe this notion:

1. **R175H mutation in *TP53* has a dominant negative effect:** We have identified three missense mutations in *TP53*: R175H (81-year-old donor, VAF = 0.52), R158H (69 year old donor, VAF = 0.5) and G187D (39 year old donor, VAF = 0.29). One of these mutations, R175H, is known to have a dominant negative effect through inactivation of the function of wild-type p53 and is implicated in tumour development (Willis et al., 2004, Aubrey et al., 2018, Boettcher et al., 2019). It is this very mutation that is present in the sample containing 9 structural variants were identified; no structural variants were seen in the two other samples with the other two *TP53* mutations.
2. **Many endometrial cancer cases have heterozygous *TP53* mutations and structural variants:** 25 out of the 44 endometrial cancer cases (PCAWG) had at least one *TP53* mutation of which 21 had no evidence of loss of heterozygosity, LOH (LOH was assumed if VAF was above 0.6 or if there was more than one mutation in *TP53*). Structural variants were detected in all of the studied endometrial cancer cases, however the burden was higher in samples with *TP53* mutations (median = 251, range 8-450) than those without (median = 77, range 8-316) (Wilcoxon rank sum test, p = 0.019) (Figure 4.15).

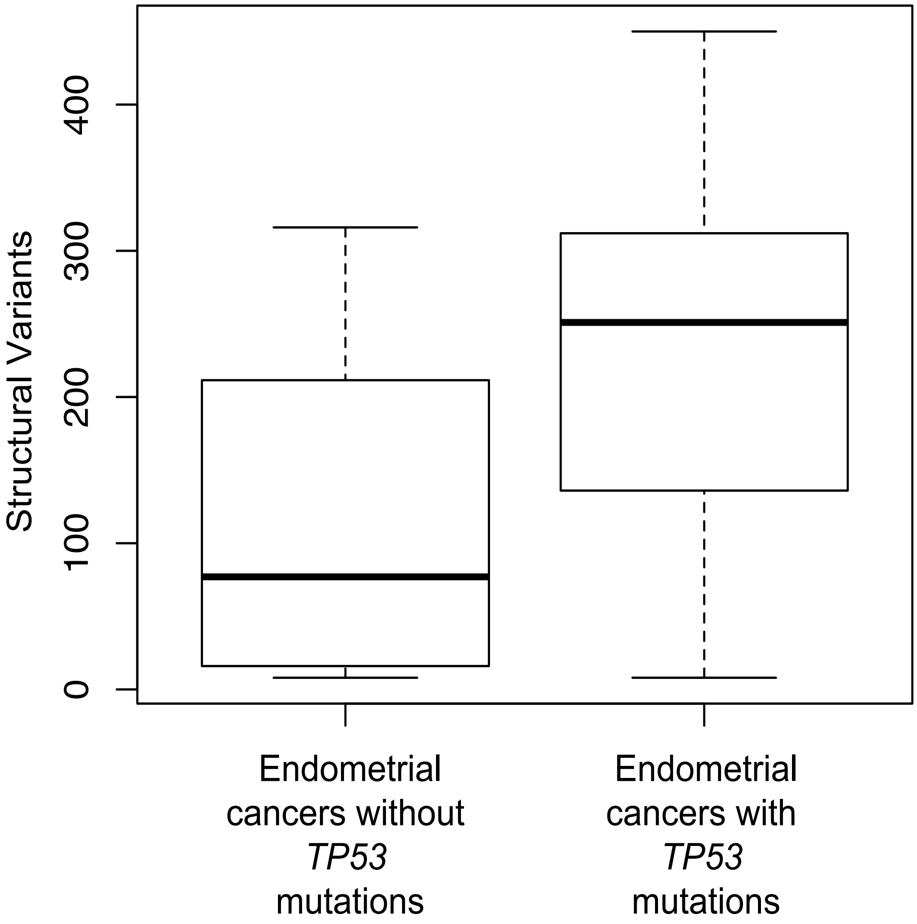


Figure 4.15 | Structural variant burden of endometrial cancer samples with and without *TP53* mutations.

4.2.6 Cancer driver mutations in normal endometrial glands

To identify genes under positive selection a statistical method based on the observed:expected ratios of non-synonymous:synonymous mutations was used (Martincorena et al., 2017). Twelve genes showed evidence of positive selection in the 257 normal endometrial glands; *PIK3CA*, *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP*, *PPP2R1A* and *ERBB3* (Figure 4.6c-e, Appendix 14). All were present in a set of 369 genes previously shown to be under positive selection in human cancer (Martincorena et al., 2017). In addition, four different truncating mutations (and no other mutations) were observed in the progesterone receptor gene (PGR). Although these did not attain standard significance levels the biological role that progesterone plays in normal endometrium as an antagonist of oestrogen driven proliferation raises the possibility that these inactivating mutations confer growth advantage.

To comprehensively identify drivers in the 257 endometrial glands, mutations with the characteristics of drivers in each of the 369 genes were sought (Methods). 209 driver mutations were found in normal endometrial glands from 25/28 women (Appendix 15). The youngest carrier was a 24-year-old (PD40535) with a *KRAS* G12D mutation in 1/7 glands sampled. 57% (147/257) of endometrial glands carried at least one driver mutation, 16% (42/257) carried at least two and 2% (5/257) carried at least four drivers. Remarkably, in four women, aged 34 (19 glands), 44 (11 glands), 60 (14 glands) and 81 (5 glands), all glands analysed carried driver mutations suggesting that the whole endometrium had been colonised by genomically microneoplastic clones (Figure 4.11 and 4.12). The fraction of endometrial glands carrying a driver (Figure 4.2g), the mean number of drivers per gland (Figure 4.2h) and the number of different drivers in each individual (corrected for number of glands sampled) (Figure 4.2i) all positively correlated with age of the individual. However, there were sufficient outliers from this age correlation to suggest that other factors influence colonisation of the endometrium by driver carrying clones.

Driver mutations in both recessive (tumour suppressor genes) and dominant cancer genes were found, similar to recent publications (Suda et al., 2018, Lac et al., 2019, Anglesio et al., 2017). *PIK3CA* was the most frequently mutated cancer gene, with at least one missense mutation in 54% (15/28) of women and five different mutations found in two women (Figure 4.11 and Figure 4.13). Most truncating driver mutations in recessive cancer genes (including in *ZFHX3*, *ARGHAP35* and *FOXA2* which showed formal evidence of selection in normal endometrial glands, see above) were heterozygous without evidence of a mutation inactivating the second, wild type allele. Therefore, haploinsufficiency of these genes appears sufficient to confer growth advantage in normal cells. Nevertheless, further inactivating mutations, including copy number neutral LOH of the wild type allele and truncating mutations, in the same genes in other glands indicate that additional advantage is conferred by complete abolition of their activity (notably for *ZFHX3* in the 60 year old, Figure 4.12 and 4.15). Driver mutations were found in genes encoding growth factor receptors (*ERBB2*, *ERBB3* and *FGFR2*), components of signal transduction pathways (*HRAS*, *KRAS*, *BRAF*, *PIK3CA*, *PIK3R1*, *ARHGAP35*, *RRAS2*, *NF1*, *PP2R1A* and *PTEN*), pathways mediating steroid hormone responses

(*ZFHX3*, *FOXA2*, *ARHGAP35*), proteins involved in chromatin function (*KMT2D* and *ARID5B*) and protein-mediated degradation pathways (*FBXW7*) that target oncoproteins such as mTOR and c-MYC. Many different combinations of mutated cancer genes were found in individual glands.

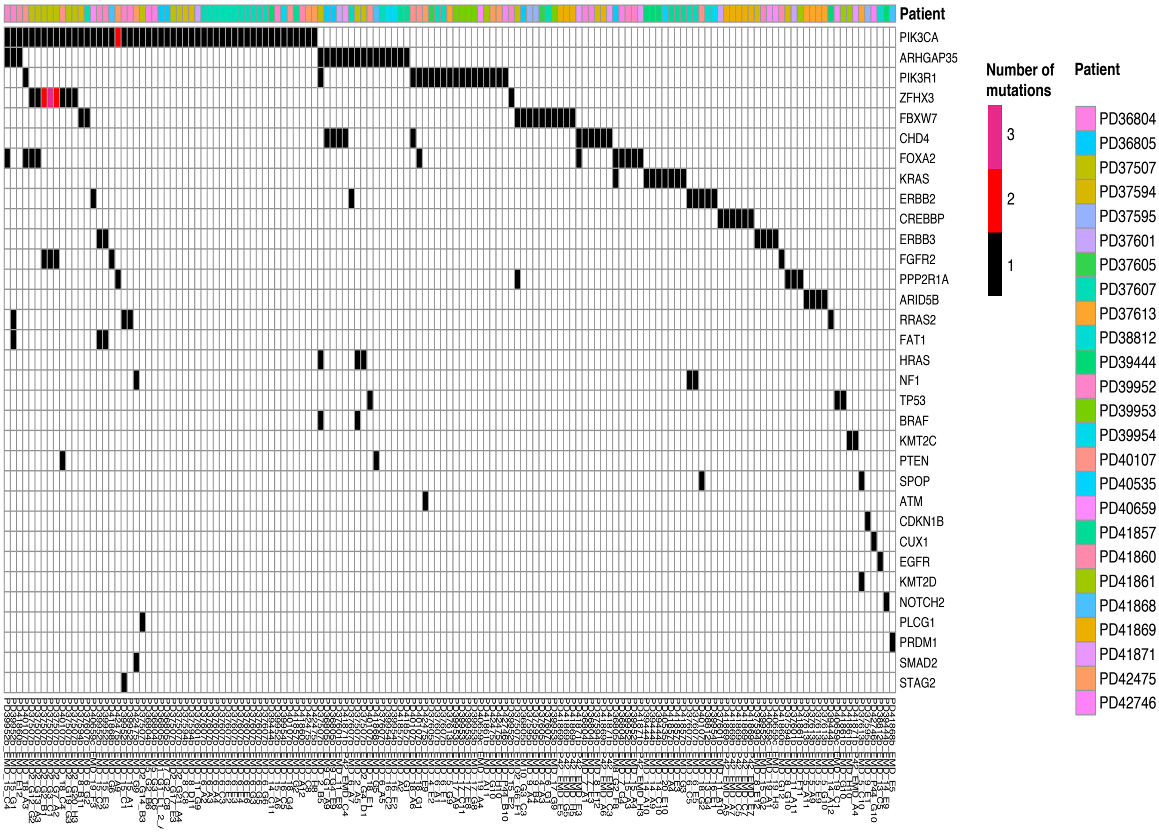


Figure 4.16 | Oncoplot of all driver mutations and their distribution across individual endometrial gland samples and donors. Each cell represents an individual endometrial gland sample and is colour-coded to represent the total number of detected driver mutations (0-3). *PIK3CA* was the most frequently mutated gene with at least one mutation detected in 54% (15/28) of women. In some glands, these co-occurred with mutations in *ZFHX3*, *ARHGAP35*, *FGFR2*, *FOXA2* and other genes that are also selected for in endometrial cancer.

4.2.6.1 Rate of driver mutations and mixed-effect model

The fraction of endometrial glands carrying a driver (Figure 4.2g), the mean number of drivers per gland (Figure 4.2h) and the number of different drivers in each individual (corrected for number of glands sampled) Figure 4.2i) all positively correlated with age of the individual. However, there were sufficient outliers from this age correlation to suggest that other factors influence colonisation of the endometrium by driver carrying clones. Indeed, use of a generalised linear mixed effect model showed that age has a positive association with accumulation of driver mutations (0.035 driver mutation per year, $CI_{95\%}$ 0.01-0.06, p = 3.31 x 10^{-4}) while parity has a negative association (-0.253 per life birth, $CI_{95\%}$ -0.46 to -0.05, p = 1.33 x 10^{-2}) (Appendix 16); no correlation was observed with menstrual phase (Appendix 17).

4.2.6.2 Timing of driver mutations

Constructing phylogenetic trees based on whole genome sequences of individual endometrial glands enabled characterisation of the mode of expansion of normal cell clones with drivers and timing of their initiation. Phylogenetically closely related glands were often in close physical proximity within the endometrium (Figure 4.11). In phylogenetic clusters for which the mutation catalogues were almost identical, this may simply reflect multiple sampling of a single tortuous gland weaving in and out of the plane of section, rather than distinct glands with their own stem cell populations (e.g. glands C5 and E5, Figure 4.11a, c). For other phylogenetic clusters, the different branches within the clade have diverged substantially, sometimes acquiring different driver mutations, and therefore are likely derived from different stem cell populations. In such instances phylogenetically related glands can range over distances of hundreds of microns suggesting that their clonal evolution has entailed capture and colonisation of extensive zones of endometrial lining (e.g. glands C1, A2, B1, H2, A3, B3, Figure 4.11b, d). Conversely, many glands in close physical proximity are phylogenetically distant (e.g. glands E1 and G2, Figure 4.11a, c), indicating that their cell populations have remained isolated from each other.

Driver mutations were positioned on the phylogenetic trees of somatic mutations constructed for each individual and their times of occurrence were estimated by assuming constant somatic mutation rates during life (Figure 4.18, 4.18and 4.20). This assumption is unlikely to be completely correct. However, the results indicate that mutations in normal endometrial cells (and particularly those attributable to SBS1 and SBS5/40) are acquired in more-or-less linear fashions throughout life and that potential modifying factors, including acquisition of a driver, make only modest differences to mutation rates. Furthermore, our approach is, overall, likely to overestimate the ages before which driver mutations have occurred because it does not account for the time taken by a single endometrial stem cell to colonise an individual gland, which in colorectal crypts is estimated at several years (Nicholson et al., 2018). The results indicate, therefore, that at least some driver mutations occur early in life. These included a *KRAS* G12D mutation in three glands from a 35 year old and a *PIK3CA* mutation in two glands from a 34 year old, which are both likely to have arisen during the first decade (Figures 4.11, 4.12, 4.13 and 4.18). A pair of drivers in *ZFHX3* and *PIK3CA*, co-occurring in six glands from a 60 year old, were also acquired during the first decade indicating that driver associated clonal evolution also begins early in life (Figures 4.11 and 4.19). Indeed, it is possible that many more clones with drivers were initiated during the first decade, but their phylogenetic trees are not informative in this regard (Figures 4.19 and 4.20). Three normal cell clones (from 3 individuals) with a driver mutation were demonstrably initiated after age 20 (Figure 4.19). There was evidence, however, for continuing acquisition and clonal expansion of driver mutations into the third and fourth decades and further accumulation beyond this period is not excluded (Figures 4.18, 4.19 and 4.20).

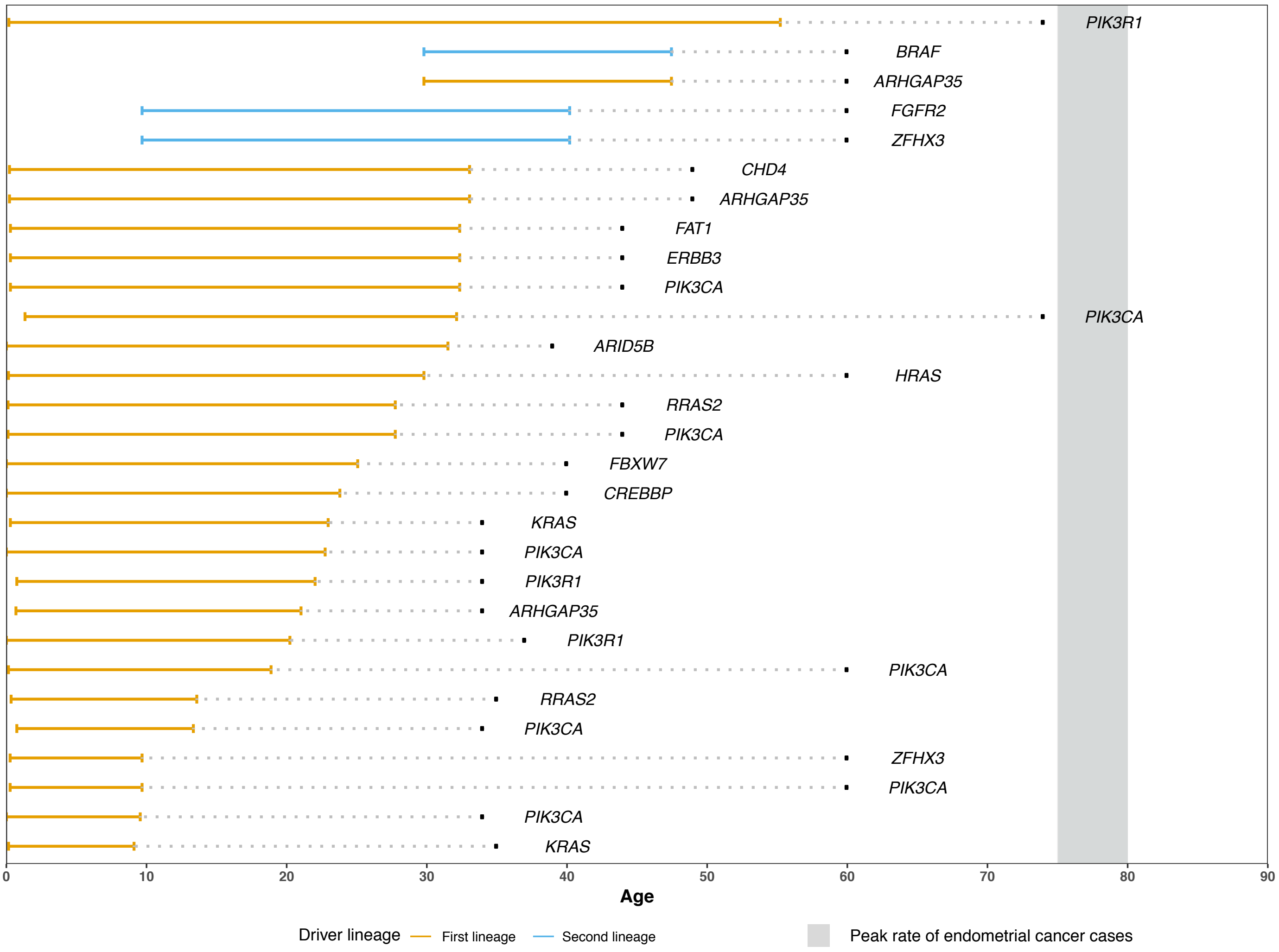


Figure 4.18 | Timing of driver mutations in normal endometrial glands. To time driver mutations, phylogenetic trees were reconstructed for 25 out of the 28 donors using single nucleotide variants (SNVs). To estimate the time interval in which specific mutations occurred, we calculated a patient-specific mutation rate by taking the ratio of the patient's mean mutation burden per endometrial gland and the patient's age. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. This approach relies on the assumption of a constant mutation rate for endometrial glands throughout a patient's life. The same approach was used for timing indels. We timed driver mutations that occurred in the 'trunks' and branches. Here, we display driver variants that occurred in the 'trunks' of the individual trees only. We show that many driver variants occur decades before the reported peak incidence of endometrial cancer (variants with an interval of <1 year between the upper age and the age at sampling were excluded from this plot for illustration purposes). Based on our calculations, four driver variants (*KRAS* G12D, *PIK3CA* G118D, *PIK3CA* E542K and *ZFHX3* R715*) from three different women occurred before the age of 10.

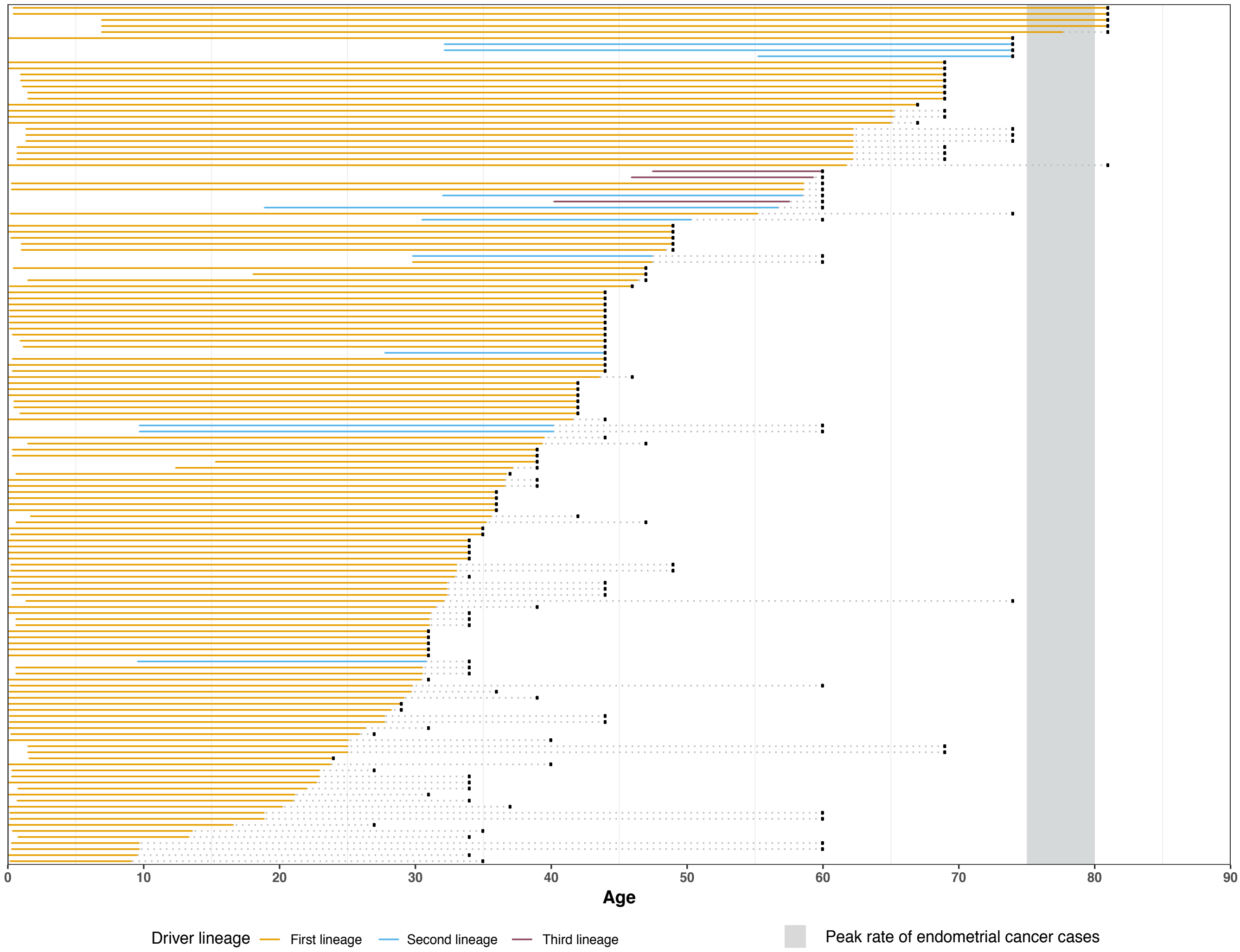


Figure 4.19 | Timing of all driver mutations. To time driver mutations, we used the reconstructed SNV based phylogenetic trees for 25 out of the 28 donors. Here, to estimate the time interval in which specific mutations occurred, we calculated a patient-specific mutation rate by taking the ratio of the patient's mean mutation burden per endometrial gland and the patient's age. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. This approach relies on the assumption of a constant mutation rate for endometrial glands throughout a patient's life. The same approach was used for timing indels. We timed driver mutations that occurred in the 'trunks' and branches.

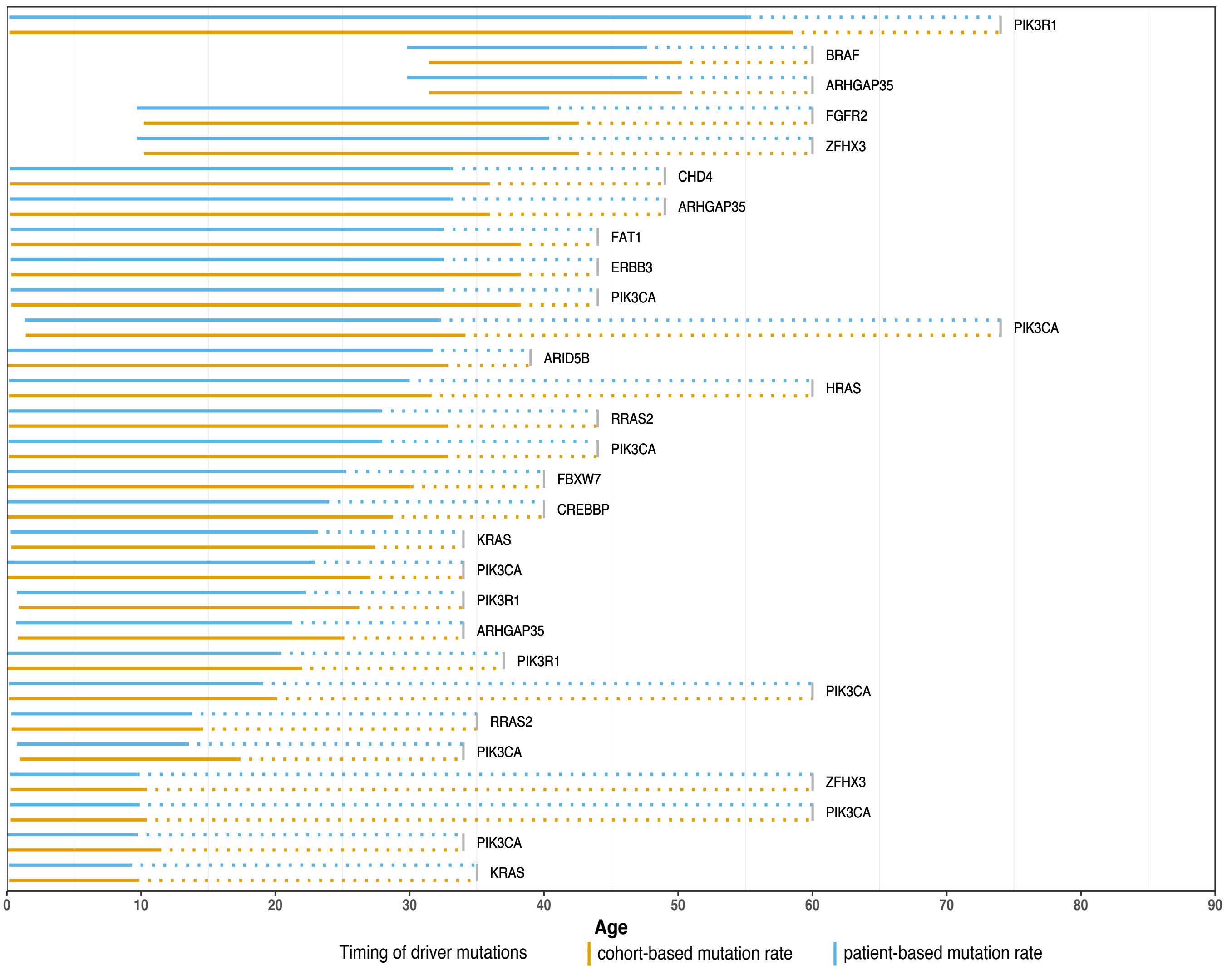


Figure 4.20 | Timing of driver mutations using patient-based and cohort-based estimates of mutation rates. To estimate the time interval in which specific mutations occurred, we applied two approaches: (a) ‘patient-based’, in which we calculated a patient-specific mutation rate by taking the ratio of the patient’s mean mutation burden per endometrial gland and the patient’s age; (b) ‘cohort-based’, in which mutation rate for each patient was derived from the linear mixed-effect model for total mutation rate that included data from the entire cohort. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. Both approaches rely on the assumption of a constant mutation rate for endometrial glands throughout a patient’s life.

4.2.6.3 **Microbiome**

The microbiome content of the endometrium has become a hot topic in recent years. We examined whether there were any correlations of the microbiome and patient age, BMI and somatic mutations. To detect bacterial DNA sequences in the available whole-genome sequencing data from normal endometrial glands, read-pairs which had one or both reads unmapped were identified and bases with Phred quality score < 10 were removed. The remaining sequence was split into non-overlapping 30 bp fragments. Terminal fragments were processed without further splitting (30-59 bp). The obtained fragments were aligned to the viral GOTTCHA database (Freitas et al., 2015) at the taxonomic levels of phylum, class, order, family, genus, species and strain using BWA (Li and Durbin, 2010). For each endometrial gland sample, we also calculated unmapped and mapped read ratios which were included in the mixed-effect model.

First, we looked for the presence of bacterial organisms that have been previously associated with endometrial cancer (Walther-Antonio et al., 2016) (Table 4.1) at a species level. *Porphyromonas asaccharolytica* was identified in two glands from two patients (PD37507b_EMD2_G7_A2 and PD39952b_EMD_15_G1) at a relative abundance of 0.0357 and 0.0229 respectively. Although the species has been previously associated with endometrial cancer, given the fact that the two calls are only identified in one sample from each donor and at a relative abundance <0.05, we are hesitant to make firm conclusions based on these limited observations. No other endometrial cancer associated bacterial organisms were identified in the WGS data from the normal endometrial glands.

Second, we examined the relative abundance of all identified bacterial genomes at the phylum and order levels (Figure 4.21) for each donor. Interestingly, the top three phyla detected in the normal endometrial glands were Proteobacteria, Actinobacteria and Firmicutes, all of which are known to be the most prevalent phyla in normal/”healthy” uterine microbiota (Baker, 2018).

Next, to test whether there is any correlation between the relative abundance of the identified bacteria and the total somatic mutation burden in normal endometrium, we applied a linear mixed-effect model. At the phylum level, relative abundance of Firmicutes has a negative effect on the acquisition of somatic mutations in normal endometrium (-172 substitutions, p = 2 x 10^{-2}). At the order level, there is a negative correlation between the relative abundance of Lactobacillales and the rate of total mutation burden (-309 substitutions, p = 2.1 x 10^{-2}). This is an interesting observation and it is not yet clear what the underlying mechanism might be between the somatic mutation acquisition and the relative abundance of Lactobacillales. It may well be that this association is related to other factors such BMI, age and parity. Further work to explore the endometrial microbiome and its association with somatic mutation burden fully on a larger study with a microbiome-specific hypothesis and methodology, in particularly in relation to the sample collection, and strict control for multiple testing in the statistical analyses, is necessary to draw robust conclusions.

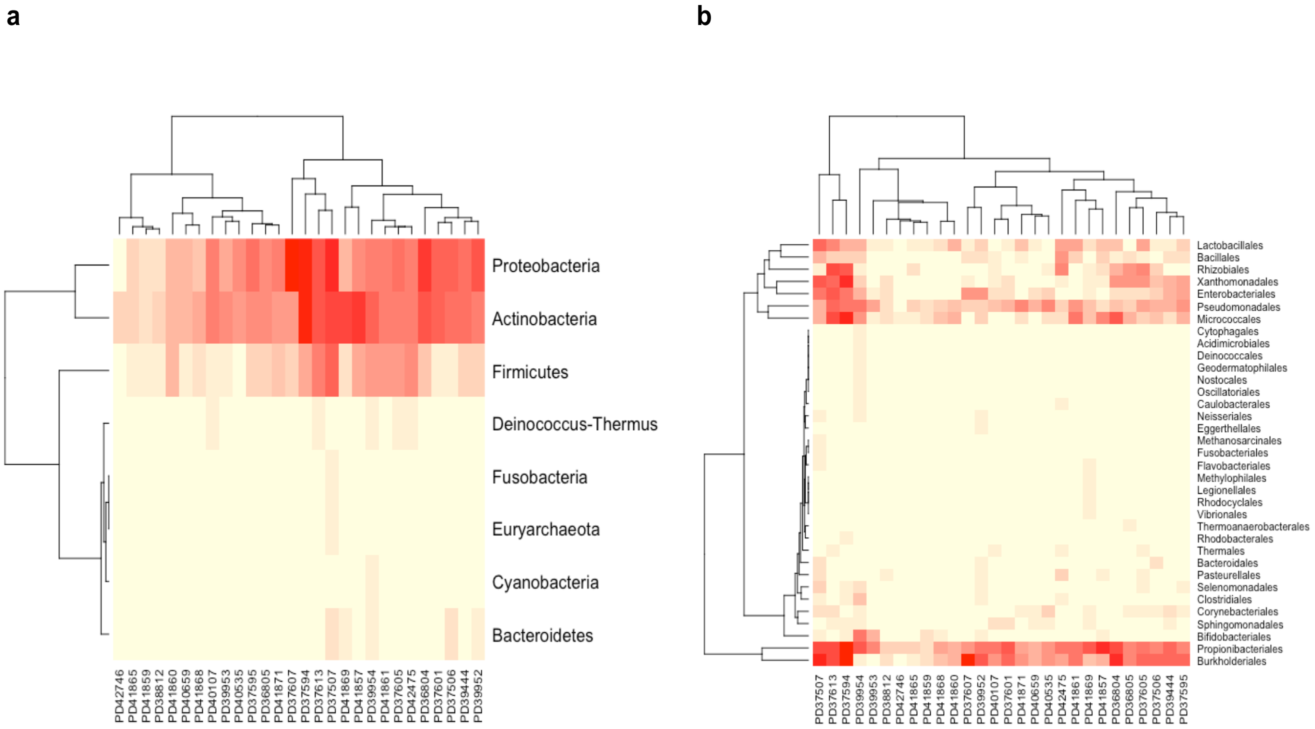


Figure 4.21| Heatmap of bacterial organisms identified in normal endometrial glands. The figure shows summary of the identified bacterial genomes and their relative abundance in normal endometrial glands in each donor on a phylum **(a)** and order **(b)** levels.

4.2.7 Summary of results in this chapter

Using a combination of laser-capture microdissection and whole genome sequencing of individual endometrial glands, we show that the ‘driver’ mutations in normal cell clones are not only abundant in this tissue, but occur in the early decades of life, accumulate with age and in some women appear to colonise the entire endometrium without morphological evidence of neoplastic transformation. We show that parity has a ‘protective’ effect on the rate at which driver events occur in this tissue. Importantly, although we report a high prevalence of driver mutations in this tissue, genomic changes in key cancer genes, such as *PTEN* and *TP53*, that are usually seen in both types of endometrial cancer, are relatively infrequent in the normal endometrium with only five such mutations identified in the entire cohort. Interestingly, other types of genomic alterations (CNAs and structural variants) were also uncommon. Furthermore, unlike cancer, normal endometrial glands are characterised by relatively homogenous mutational processes with the majority of the samples showing primarily SBS1, SBS5 and SBS18 signatures. Together, these observations support the notion that cancer is a complex multi-step process and that single events, such as single base substitutions in cancer genes, alone do not necessarily lead to neoplastic transformation.

A series of studies are being conducted in our group, and elsewhere, in multiple different normal tissues, and we are already seeing that the observed mutation patterns across sites are not the same. Here, we showed that the landscape of somatic mutations is different between endometrium and other normal tissues, such as colon. The epithelial component in both colon and endometrium comprises glandular structures, each containing a pool of stem cells within the basal compartments. Although the incidence of cancer and the rate at which somatic mutations occur is higher in the colon, surprisingly driver mutations have been found in only ~1% of crypts. *PIK3CA*, the second most commonly mutated gene in endometrial cancer, and is also the most frequently mutated cancer gene in normal endometrium and yet, no detectable morphological changes were seen. These findings also highlight that other factors, such as cell context and microenvironment, play role in the development of cancer.

Chapter 5 **General discussion**

5.1 Summary of findings

This, and other, studies of normal endometrial epithelium, together with recent studies of other normal cell types (Blokzijl et al., 2016, Martincorena, 2018, Martincorena et al., 2015, Lee-Six et al., 2018a, Lee-Six et al., 2018b, Genovese et al., 2014, Jaiswal et al., 2014, Suda et al., 2018, Lee-Six et al., 2019), is revealing the landscape of somatic mutations in normal human cells. Somatic mutations are predominantly generated by a limited repertoire of ubiquitous mutational processes generating base substitutions, small indels, genome rearrangements and whole chromosome copy number changes which exhibit more-or-less constant mutation rates during the course of a lifetime. Additional mutational signatures which are present only in some cells, only in some cell types and/or are intermittent also operate in some normal cells, although apparently not the endometrial epithelium, supplementing the mutation load contributed by ubiquitous signatures. The latter include exposures such as ultraviolet light in skin (Martincorena et al., 2015), APOBEC mutagenesis in occasional colon crypts and other signatures of unknown cause in normal colon epithelium(Lee-Six et al., 2019).

A small subset of mutations generated by these mutational processes have the properties of driver mutations. Numerous cell clones with one or more drivers colonise much of the normal endometrial epithelium (Suda et al., 2018, Lac et al., 2019), in contrast to the colon where just 1% of normal crypts in middle-aged individuals carry a driver(Lee-Six et al., 2019, Suda et al., 2018). This marked difference in driver mutation landscape seems unlikely to be due to any relatively modest difference in total somatic mutation rate between endometrial and colonic epithelial stem cells (Blokzijl et al., 2016, Lee-Six et al., 2019, Roerink et al., 2018). However, it may be attributable to intrinsic differences in physiology between endometrium and colon. In the endometrium, the cyclical process of tissue breakdown, shedding and remodelling iteratively opens up denuded terrains for pioneering clones of endometrial epithelial cells with drivers to preferentially colonise compared to wild type cells. By contrast, in the colon the selective advantage of a clone with a driver is usually confined to the small siloed population of a single crypt, with only occasional opportunities for further expansion. Thus, the endometrium, in some respects, resembles more the squamous epithelia of skin and oesophagus in which cell clones derived from basal cells (with or without driver mutations) directly compete against each other for occupancy of the squamous sheet and in which substantial proportions of such sheets become colonised over a lifetime by normal cell clones carrying driver mutations (Martincorena et al., 2015, Martincorena et al., 2018). Although this rampant colonisation by driver clones in endometrium progresses with age, it is already well advanced in some young women, and parity apparently has an inhibitory effect on it, indicating that multiple factors influence its progression. The effect of parity is of particular interest since increased parity also reduces endometrial cancer risk and it is plausible that this is mediated by a suppressive effect of parity on driver clone expansion (Wu et al., 2015). More extensive studies of the mutational landscape in normal endometrium are required to better assess how pregnancy, the premenarchical and postmenopausal states, hormonal contraceptive use and hormone replacement therapies influence it and also the potential impact it has on pregnancy and fertility.

The burdens of all mutation classes are lower in normal endometrial cells, including those with drivers, than in endometrial cancers. However, these differences are most marked for structural variants/copy number changes and for the extreme base substitution/indel hypermutator phenotypes due to DNA mismatch repair deficiency and polymerase delta/epsilon mutations which were not found in normal endometrium. The results therefore indicate that in endometrial epithelium, and in other tissues thus far studied including colon, oesophagus and skin, normal mutation rates are sufficient to generate large numbers of clones with driver mutations behaving as normal cells, but that acquisition of an elevated mutation rate and burden is associated with further evolution to invasive cancer (Lee-Six et al., 2019, Martincorena et al., 2015, Martincorena, 2018). Given that the endometrial epithelium is extensively colonised by clones of normal cells with driver mutations in middle-aged and older women and that the lifetime risk of endometrial cancer is only 3% (CRUK), this conversion from normal cell clone with drivers to symptomatic malignancy appears to be extremely rare.

The frequent colonisation of normal endometrial epithelium by normal cell clones with driver mutations provides a particular opportunity to time the onset of drivers during the lifetime of an individual by construction of phylogenetic trees of cell lineages based on whole genome sequences. The results show that the first drivers in these clones often arise relatively early in life, indicate that some occur within the first decade and do not exclude many more doing so. The modal period of diagnosis of endometrial cancer is 75-80 years. Therefore, if normal cell clones with drivers are progenitors of endometrial cancers, which is plausible given the similar repertoires of cancer genes in which the driver mutations are found, then it is conceivable that some neoplastic clones ultimately manifesting as cancer were initiated during childhood and that evolution to malignancy has taken place over much of the individual's lifetime. This perspective on the long duration of neoplastic evolution of invasive endometrial cancer has resonance with previous observations on leukaemia (Greaves, 2005, Greaves, 2003) and, more recently, other solid malignancies (Mitchell et al., 2018, Anderson et al., 2018, Maura et al., 2018, Gerstung et al., 2018) and may therefore be a common feature of human cancer development.

5.2 Limitations

5.2.1 Method limitations

The low DNA input LCM workflow has been particularly impactful for when we are able to identify and capture clonal units, such as colonic crypts or endometrial glands in mitotically active tissue. Conversely, in mitotically relatively in-active tissues (brain, heart and skeletal muscle) or highly polyclonal tissues (liver and lung), this approach is less informative and requires greater read depth. These tissues would benefit from error-corrected WGS techniques which are currently under development and have the potential to differentiate between sequencing artefacts and genuine variants residing within small clones within a polyclonal sample.

5.2.2 Study limitations

Within the endometrial study, the main issue is the fact that we were restricted by the availability of eligible samples, which impacted our case ascertainment, specifically the age spectrum. In addition, the availability of the associated metadata, such as BMI and parity, was suboptimal reducing the power of our analyses when accounting for these variables. In our comparison of the mutation rate in the endometrium to other tissues, some of the possible limitations include differences in experimental approaches (organoid cultures and LCM-derived material) and additional mutations that could've been acquired during the cell culture, sequencing depth and clonality and purity of the samples.

5.3 Work in context

5.3.1 Relevant work published during my PhD

During the course of our work, a study by Anglesio and colleagues showed that cancer associated mutations can be identified in morphologically normal, but in abnormally located (ectopic) endometrium. Specifically, they studied deep infiltrating endometriosis, a condition that almost never undergoes malignant transformation (Anglesio et al., 2017). Known cancer driver mutations in genes such as *PIK3CA*, *KRAS* and *ARID1A* were found in 5/24 patients, including those in their late 20s. Later, the same group studied another type of endometriosis, iatrogenic endometriosis, which is thought to be associated with previous surgical procedures (Lac et al., 2018). The results showed driver mutations in 11/40 such cases and yet these lesions virtually never undergo malignant transformation.

Finally, Suda and colleagues applied targeted and whole exome sequencing approach to study ovarian endometriosis and concurrent normal endometrium from the same patients; they showed that cancer driver mutations are not only abundant in the endometriotic lesions, but can also be detected in the eutopic (uterine) normal endometrium without morphological evidence of malignancy (Suda et al., 2018).

5.3.2 Early detection

In recent years, significant efforts have been made to improve early cancer diagnosis through the development of techniques to screen blood and other bodily fluids for early cancer driver events. In this work, I show that the 'driver' mutations in normal endometrial epithelium are not only abundant, but occur in the early decades of life, accumulate with age and in some women appear to colonise the entire endometrium without morphological evidence of malignant transformation. These observations along with the recent work in other normal tissues, such as skin and oesophagus, have implications on our understanding of ageing and what constitutes 'normal' and force us to reconsider the rather simplistic binary distinction between 'drivers' and 'passengers'. The findings also highlight that caution should be taken in the development and utilization of mutation-based early detection tools in endometrial and other cancer types and that a multi-dimensional (‘multi-omics’) approach, which would also incorporate methylation and transcriptomics data, should be considered to avoid false positive results and unnecessary diagnostic tests, overtreatment and distress.

5.4 Future work

5.4.1 Endometrium expansion

Based on our initial observation in relation to somatic mutation accumulation, I plan to study more endometrial glands from healthy women expanding across the age range, particularly at the extremes and around perimenopause.

The expanded dataset will allow us to:

1. Model more accurately mutational burden as a function of age and to determine whether the accumulation of mutations is truly linear or whether there are oestrogen-related rate changes, for instance at puberty and menopause.
2. Use better characterise driver landscape of peri- and post-menopausal women to better understand what constitutes ‘normal’ ageing and endometrial tumourigenesis.
3. Model with greater power the effect of known epidemiological cancer risk factors, such as BMI, parity and hormonal therapy.

5.4.2 Panbody completion

The preliminary pan-body analyses on a single male donor (78 year old), which included 224 samples across twenty five tissues have already provided first insights into the clonal architecture, mutational signatures and mutation burden. We have expanded this work to two further donors: one male (47 year old) and one female (54 year old). The additional data will not only validate some of our initial observations in terms of burden and signatures, but it will also make the pan-body survey more comprehensive by including tissues from both genders.

5.5 Conclusions

These preliminary normal tissue analyses have already provided an initial survey of clonal architecture, mutational signatures and mutation burden. More extensive studies of each tissue are required to investigate whether additional mutational signatures occur sporadically, to characterise the accumulation of mutations from each signature with age, to provide more comprehensive estimates of mutation burden and to extend to post-mitotic cell types, such as myocytes and neurones, which are not easily studied our low DNA input LCM approach here. The survey also indicates that small clones of cells carrying driver mutations are present and, given the relatively modest number of samples analysed, relatively common in many normal tissues. This phenomenon similarly requires more in-depth characterisation of the differences between tissues in the proportions of normal cells carrying drivers, the accumulation of driver clones in each tissue with age, and the extent to which driver mutations alter the parameters of clonal expansion. The results of such studies will collectively establish a basis for subsequent exploration of how mutational processes *in vivo* are influenced by inherited genetic background, by lifestyle, occupational and environmental exposures, and by inflammatory, metabolic and degenerative human diseases.

Bibliography

AKHMEDKHANOV, A., ZELENIUCH-JACQUOTTE, A. & TONIOLO, P. 2001. Role of exogenous and endogenous hormones in endometrial cancer: review of the evidence and research perspectives. *Ann N Y Acad Sci,* 943**,** 296-315.

ALEXANDROV, L. 2018. The Repertoire of Mutational Signatures in Human Cancer.

ALEXANDROV, L. B. 2013. Signatures of mutational processes in
human cancer. *Nature,* 500**,** 415–421.

ALEXANDROV, L. B., JONES, P. H., WEDGE, D. C., SALE, J. E., CAMPBELL, P. J., NIK-ZAINAL, S. & STRATTON, M. R. 2015. Clock-like mutational processes in human somatic cells. *Nat Genet,* 47**,** 1402-7.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., CAMPBELL, P. J. & STRATTON, M. R. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep,* 3**,** 246-59.

ANDERSON, N. D., DE BORJA, R., YOUNG, M. D., FULIGNI, F., ROSIC, A., ROBERTS, N. D., HAJJAR, S., LAYEGHIFARD, M., NOVOKMET, A., KOWALSKI, P. E., ANAKA, M., DAVIDSON, S., ZARREI, M., ID SAID, B., SCHREINER, L. C., MARCHAND, R., SITTER, J., GOKGOZ, N., BRUNGA, L., GRAHAM, G. T., FULLAM, A., PILLAY, N., TORETSKY, J. A., YOSHIDA, A., SHIBATA, T., METZLER, M., SOMERS, G. R., SCHERER, S. W., FLANAGAN, A. M., CAMPBELL, P. J., SCHIFFMAN, J. D., SHAGO, M., ALEXANDROV, L. B., WUNDER, J. S., ANDRULIS, I. L., MALKIN, D., BEHJATI, S. & SHLIEN, A. 2018. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science,* 361.

ANGLESIO, M. S., PAPADOPOULOS, N., AYHAN, A., NAZERAN, T. M., NOE, M., HORLINGS, H. M., LUM, A., JONES, S., SENZ, J., SECKIN, T., HO, J., WU, R. C., LAC, V., OGAWA, H., TESSIER-CLOUTIER, B., ALHASSAN, R., WANG, A., WANG, Y., COHEN, J. D., WONG, F., HASANOVIC, A., ORR, N., ZHANG, M., POPOLI, M., MCMAHON, W., WOOD, L. D., MATTOX, A., ALLAIRE, C., SEGARS, J., WILLIAMS, C., TOMASETTI, C., BOYD, N., KINZLER, K. W., GILKS, C. B., DIAZ, L., WANG, T. L., VOGELSTEIN, B., YONG, P. J., HUNTSMAN, D. G. & SHIH, I. M. 2017. Cancer-Associated Mutations in Endometriosis without Cancer. *N Engl J Med,* 376**,** 1835-1848.

AUBREY, B. J., JANIC, A., CHEN, Y., CHANG, C., LIESCHKE, E. C., DIEPSTRATEN, S. T., KUEH, A. J., BERNARDINI, J. P., DEWSON, G., O'REILLY, L. A., WHITEHEAD, L., VOSS, A. K., SMYTH, G. K., STRASSER, A. & KELLY, G. L. 2018. Mutant TRP53 exerts a target gene-selective dominant-negative effect to drive tumor development. *Genes Dev,* 32**,** 1420-1429.

BAKER, A. M., CERESER, B., MELTON, S., FLETCHER, A. G., RODRIGUEZ-JUSTO, M., TADROUS, P. J., HUMPHRIES, A., ELIA, G., MCDONALD, S. A., WRIGHT, N. A., SIMONS, B. D., JANSEN, M. & GRAHAM, T. A. 2014. Quantification of crypt and stem cell evolution in the normal and neoplastic human colon. *Cell Rep,* 8**,** 940-7.

BAKER, J. M. 2018. Uterine Microbiota: Residents, Tourists, or Invaders? *Frontiers in Immunology,* 9.

BIERNAUX, C., LOOS, M., SELS, A., HUEZ, G. & STRYCKMANS, P. 1995. Detection of major bcr-abl gene expression at a very low level in blood cells of some healthy individuals. *Blood,* 86**,** 3118-22.

BLACKWOOD, J. K., WILLIAMSON, S. C., GREAVES, L. C., WILSON, L., RIGAS, A. C., SANDHER, R., PICKARD, R. S., ROBSON, C. N., TURNBULL, D. M., TAYLOR, R. W. & HEER, R. 2011. In situ lineage tracking of human prostatic epithelial stem cell fate reveals a common clonal origin for basal and luminal cells. *The Journal of Pathology,* 225**,** 181-188.

BLAKEMORE, J. & NAFTOLIN, F. 2016. Aromatase: Contributions to Physiology and Disease in Women and Men. *Physiology (Bethesda),* 31**,** 258-69.

BLOKZIJL, F., DE LIGT, J., JAGER, M., SASSELLI, V., ROERINK, S., SASAKI, N., HUCH, M., BOYMANS, S., KUIJK, E., PRINS, P., NIJMAN, I. J., MARTINCORENA, I., MOKRY, M., WIEGERINCK, C. L., MIDDENDORP, S., SATO, T., SCHWANK, G., NIEUWENHUIS, E. E., VERSTEGEN, M. M., VAN DER LAAN, L. J., DE JONGE, J., JN, I. J., VRIES, R. G., VAN DE WETERING, M., STRATTON, M. R., CLEVERS, H., CUPPEN, E. & VAN BOXTEL, R. 2016. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature,* 538**,** 260-264.

BOETTCHER, S., MILLER, P. G., SHARMA, R., MCCONKEY, M., LEVENTHAL, M., KRIVTSOV, A. V., GIACOMELLI, A. O., WONG, W., KIM, J., CHAO, S., KURPPA, K. J., YANG, X., MILENKOWIC, K., PICCIONI, F., ROOT, D. E., RUCKER, F. G., FLAMAND, Y., NEUBERG, D., LINDSLEY, R. C., JANNE, P. A., HAHN, W. C., JACKS, T., DOHNER, H., ARMSTRONG, S. A. & EBERT, B. L. 2019. A dominant-negative effect drives selection of TP53 missense mutations in myeloid malignancies. *Science,* 365**,** 599-604.

BOKHMAN, J. V. 1983. Two pathogenetic types of endometrial carcinoma. *Gynecol Oncol,* 15**,** 10-7.

BONGSO, A. & RICHARDS, M. 2004. History and perspective of stem cell research. *Best Pract Res Clin Obstet Gynaecol,* 18**,** 827-42.

BOSE, S., DEININGER, M., GORA-TYBOR, J., GOLDMAN, J. M. & MELO, J. V. 1998. The presence of typical and atypical BCR-ABL fusion genes in leukocytes of normal individuals: biologic significance and implications for the assessment of minimal residual disease. *Blood,* 92**,** 3362-7.

BUELS, R., YAO, E., DIESH, C. M., HAYES, R. D., MUNOZ-TORRES, M., HELT, G., GOODSTEIN, D. M., ELSIK, C. G., LEWIS, S. E., STEIN, L. & HOLMES, I. H. 2016. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol,* 17**,** 66.

BYRON, S. A., GARTSIDE, M., POWELL, M. A., WELLENS, C. L., GAO, F., MUTCH, D. G., GOODFELLOW, P. J. & POLLOCK, P. M. 2012. FGFR2 point mutations in 466 endometrioid endometrial tumors: relationship with MSI, KRAS, PIK3CA, CTNNB1 mutations and clinicopathological features. *PLoS One,* 7**,** e30801.

CALLE, E. E. & KAAKS, R. 2004. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nat Rev Cancer,* 4**,** 579-91.

CALLE, E. E., RODRIGUEZ, C., WALKER-THURMOND, K. & THUN, M. J. 2003. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med,* 348**,** 1625-38.

CANCER GENOME ATLAS RESEARCH, N., KANDOTH, C., SCHULTZ, N., CHERNIACK, A. D., AKBANI, R., LIU, Y., SHEN, H., ROBERTSON, A. G., PASHTAN, I., SHEN, R., BENZ, C. C., YAU, C., LAIRD, P. W., DING, L., ZHANG, W., MILLS, G. B., KUCHERLAPATI, R., MARDIS, E. R. & LEVINE, D. A. 2013. Integrated genomic characterization of endometrial carcinoma. *Nature,* 497**,** 67-73.

CASASENT, A. K., SCHALCK, A., GAO, R., SEI, E., LONG, A., PANGBURN, W., CASASENT, T., MERIC-BERNSTAM, F., EDGERTON, M. E. & NAVIN, N. E. 2018. Multiclonal Invasion in Breast Tumors Identified by Topographic Single Cell Sequencing. *Cell,* 172**,** 205-217 e12.

CHAN, R. W. & GARGETT, C. E. 2006. Identification of label-retaining cells in mouse endometrium. *Stem Cells,* 24**,** 1529-38.

CHAN, R. W., SCHWAB, K. E. & GARGETT, C. E. 2004. Clonogenicity of human endometrial epithelial and stromal cells. *Biol Reprod,* 70**,** 1738-50.

CHEN, C., SONG, X., WEI, W., ZHONG, H., DAI, J., LAN, Z., LI, F., YU, X., FENG, Q., WANG, Z., XIE, H., CHEN, X., ZENG, C., WEN, B., ZENG, L., DU, H., TANG, H., XU, C., XIA, Y., XIA, H., YANG, H., WANG, J., WANG, J., MADSEN, L., BRIX, S., KRISTIANSEN, K., XU, X., LI, J., WU, R. & JIA, H. 2017. The microbiota continuum along the female reproductive tract and its relation to uterine-related diseases. *Nat Commun,* 8**,** 875.

CHO, N. H., PARK, Y. K., KIM, Y. T., YANG, H. & KIM, S. K. 2004. Lifetime expression of stem cell markers in the uterine endometrium. *Fertil Steril,* 81**,** 403-7.

COELHO, M., OLIVEIRA, T. & FERNANDES, R. 2013. Biochemistry of adipose tissue: an endocrine organ. *Arch Med Sci,* 9**,** 191-200.

CREASMAN, W. T., ODICINO, F., MAISONNEUVE, P., QUINN, M. A., BELLER, U., BENEDET, J. L., HEINTZ, A. P., NGAN, H. Y. & PECORELLI, S. 2006. Carcinoma of the corpus uteri. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer. *Int J Gynaecol Obstet,* 95 Suppl 1**,** S105-43.

CRUK. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/risk-factors#heading-Zero [Accessed 19/12
2018].

CRUK. 2019. *Cancer Research UK Uterine Cancer Statistics* [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer#heading-Zero [Accessed 05/10/2019].

CUNHA, G. R. 1976. Stromal induction and specification of morphogenesis and cytodifferentiation of the epithelia of the Mullerian ducts and urogenital sinus during development of the uterus and vagina in mice. *J Exp Zool,* 196**,** 361-70.

DAVIS, S. R., LAMBRINOUDAKI, I., LUMSDEN, M., MISHRA, G. D., PAL, L., REES, M., SANTORO, N. & SIMONCINI, T. 2015. Menopause. *Nat Rev Dis Primers,* 1**,** 15004.

ESTELLER, M., CATASUS, L., MATIAS-GUIU, X., MUTTER, G. L., PRAT, J., BAYLIN, S. B. & HERMAN, J. G. 1999. hMLH1 promoter hypermethylation is an early event in human endometrial tumorigenesis. *Am J Pathol,* 155**,** 1767-72.

FADARE, O. & ZHENG, W. 2012. Endometrial serous carcinoma (uterine papillary serous carcinoma): precancerous lesions and the theoretical promise of a preventive approach. *Am J Cancer Res,* 2**,** 335-9.

FATEHULLAH, A., TAN, S. H. & BARKER, N. 2016. Organoids as an in vitro model of human development and disease. *Nat Cell Biol,* 18**,** 246-54.

FELLOUS, T. G., ISLAM, S., TADROUS, P. J., ELIA, G., KOCHER, H. M., BHATTACHARYA, S., MEARS, L., TURNBULL, D. M., TAYLOR, R. W., GREAVES, L. C., CHINNERY, P. F., TAYLOR, G., MCDONALD, S. A., WRIGHT, N. A. & ALISON, M. R. 2009a. Locating the stem cell niche and tracing hepatocyte lineages in human liver. *Hepatology,* 49**,** 1655-63.

FELLOUS, T. G., MCDONALD, S. A., BURKERT, J., HUMPHRIES, A., ISLAM, S., DE-ALWIS, N. M., GUTIERREZ-GONZALEZ, L., TADROUS, P. J., ELIA, G., KOCHER, H. M., BHATTACHARYA, S., MEARS, L., EL-BAHRAWY, M., TURNBULL, D. M., TAYLOR, R. W., GREAVES, L. C., CHINNERY, P. F., DAY, C. P., WRIGHT, N. A. & ALISON, M. R. 2009b. A methodological approach to tracing cell lineage in human epithelial tissues. *Stem Cells,* 27**,** 1410-20.

FRANCO, I., JOHANSSON, A., OLSSON, K., VRTACNIK, P., LUNDIN, P., HELGADOTTIR, H. T., LARSSON, M., REVECHON, G., BOSIA, C., PAGNANI, A., PROVERO, P., GUSTAFSSON, T., FISCHER, H. & ERIKSSON, M. 2018. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. *Nat Commun,* 9**,** 800.

FREITAS, T. A., LI, P. E., SCHOLZ, M. B. & CHAIN, P. S. 2015. Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Res,* 43**,** e69.

GARGETT, C. E. 2007. Uterine stem cells: what is the evidence? *Hum Reprod Update,* 13**,** 87-101.

GARGETT, C. E., CHAN, R. W. & SCHWAB, K. E. 2007. Endometrial stem cells. *Curr Opin Obstet Gynecol,* 19**,** 377-83.

GARGETT, C. E., CHAN, R. W. & SCHWAB, K. E. 2008. Hormone and growth factor signaling in endometrial renewal: role of stem/progenitor cells. *Mol Cell Endocrinol,* 288**,** 22-9.

GAWAD, C., KOH, W. & QUAKE, S. R. 2016. Single-cell genome sequencing: current state of the science. *Nat Rev Genet,* 17**,** 175-88.

GENOVESE, G., KAHLER, A. K., HANDSAKER, R. E., LINDBERG, J., ROSE, S. A., BAKHOUM, S. F., CHAMBERT, K., MICK, E., NEALE, B. M., FROMER, M., PURCELL, S. M., SVANTESSON, O., LANDEN, M., HOGLUND, M., LEHMANN, S., GABRIEL, S. B., MORAN, J. L., LANDER, E. S., SULLIVAN, P. F., SKLAR, P., GRONBERG, H., HULTMAN, C. M. & MCCARROLL, S. A. 2014. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med,* 371**,** 2477-87.

GERSTUNG, M., JOLLY, C., LESHCHINER, I., DENTRO, S. C., GONZALEZ ROSADO, S., ROSEBROCK, D., MITCHELL, T. J., RUBANOVA, Y., ANUR, P., YU, K., TARABICHI, M., DESHWAR, A., WINTERSINGER, J., KLEINHEINZ, K., VAZQUEZ-GARCIA, I., HAASE, K., JERMAN, L., SENGUPTA, S., MACINTYRE, G., MALIKIC, S., DONMEZ, N., LIVITZ, D. G., CMERO, M., DEMEULEMEESTER, J., SCHUMACHER, S., FAN, Y., YAO, X., LEE, J., SCHLESNER, M., BOUTROS, P. C., BOWTELL, D. D., ZHU, H., GETZ, G., IMIELINSKI, M., BEROUKHIM, R., SAHINALP, S. C. C., JI, Y., PEIFER, M., MARKOWETZ, F., MUSTONEN, V., YUAN, K., WANG, W., MORRIS, Q. D., SPELLMAN, P. T., WEDGE, D. C. & VAN LOO, P. 2018.

GOODFELLOW, P. J., BUTTIN, B. M., HERZOG, T. J., RADER, J. S., GIBB, R. K., SWISHER, E., LOOK, K., WALLS, K. C., FAN, M. Y. & MUTCH, D. G. 2003. Prevalence of defective DNA mismatch repair and MSH6 mutation in an unselected series of endometrial cancers. *Proc Natl Acad Sci U S A,* 100**,** 5908-13.

GORIELY, A., HANSEN, R. M., TAYLOR, I. B., OLESEN, I. A., JACOBSEN, G. K., MCGOWAN, S. J., PFEIFER, S. P., MCVEAN, G. A., RAJPERT-DE MEYTS, E. & WILKIE, A. O. 2009. Activating mutations in FGFR3 and HRAS reveal a shared genetic origin for congenital disorders and testicular tumors. *Nat Genet,* 41**,** 1247-52.

GREAVES, L. C., PRESTON, S. L., TADROUS, P. J., TAYLOR, R. W., BARRON, M. J., OUKRIF, D., LEEDHAM, S. J., DEHERAGODA, M., SASIENI, P., NOVELLI, M. R., JANKOWSKI, J. A., TURNBULL, D. M., WRIGHT, N. A. & MCDONALD, S. A. 2006. Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. *Proc Natl Acad Sci U S A,* 103**,** 714-9.

GREAVES, M. 2003. Pre-natal origins of childhood leukemia. *Rev Clin Exp Hematol,* 7**,** 233-45.

GREAVES, M. 2005. In utero origins of childhood leukaemia. *Early Hum Dev,* 81**,** 123-9.

HAMILTON, C. A., CHEUNG, M. K., OSANN, K., CHEN, L., TENG, N. N., LONGACRE, T. A., POWELL, M. A., HENDRICKSON, M. R., KAPP, D. S. & CHAN, J. K. 2006. Uterine papillary serous and clear cell carcinomas predict for poorer survival compared to grade 3 endometrioid corpus cancers. *Br J Cancer,* 94**,** 642-6.

HELLEDAY, T., ESHTAD, S. & NIK-ZAINAL, S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet,* 15**,** 585-98.

HENDERSON, B. E. & FEIGELSON, H. S. 2000. Hormonal carcinogenesis. *Carcinogenesis,* 21**,** 427-33.

HOANG, D. T., VINH, L. S., FLOURI, T., STAMATAKIS, A., VON HAESELER, A. & MINH, B. Q. 2018. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol Biol,* 18**,** 11.

HOANG, M. L., KINDE, I., TOMASETTI, C., MCMAHON, K. W., ROSENQUIST, T. A., GROLLMAN, A. P., KINZLER, K. W., VOGELSTEIN, B. & PAPADOPOULOS, N. 2016. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A,* 113**,** 9846-51.

HOLLSTEIN, M., HERGENHAHN, M., YANG, Q., BARTSCH, H., WANG, Z. Q. & HAINAUT, P. 1999. New approaches to understanding p53 gene tumor mutation spectra. *Mutat Res,* 431**,** 199-209.

HOLLSTEIN, M., SIDRANSKY, D., VOGELSTEIN, B. & HARRIS, C. C. 1991. p53 mutations in human cancers. *Science,* 253**,** 49-53.

HOWAT, W. J. & WILSON, B. A. 2014. Tissue fixation and the effect of molecular fixatives on downstream staining procedures. *Methods,* 70**,** 12-9.

HUANG, C. C., ORVIS, G. D., WANG, Y. & BEHRINGER, R. R. 2012. Stromal-to-epithelial transition during postpartum endometrial regeneration. *PLoS One,* 7**,** e44285.

JABBOUR, H. N., KELLY, R. W., FRASER, H. M. & CRITCHLEY, H. O. 2006. Endocrine regulation of menstruation. *Endocr Rev,* 27**,** 17-46.

JAISWAL, S., FONTANILLAS, P., FLANNICK, J., MANNING, A., GRAUMAN, P. V., MAR, B. G., LINDSLEY, R. C., MERMEL, C. H., BURTT, N., CHAVEZ, A., HIGGINS, J. M., MOLTCHANOV, V., KUO, F. C., KLUK, M. J., HENDERSON, B., KINNUNEN, L., KOISTINEN, H. A., LADENVALL, C., GETZ, G., CORREA, A., BANAHAN, B. F., GABRIEL, S., KATHIRESAN, S., STRINGHAM, H. M., MCCARTHY, M. I., BOEHNKE, M., TUOMILEHTO, J., HAIMAN, C., GROOP, L., ATZMON, G., WILSON, J. G., NEUBERG, D., ALTSHULER, D. & EBERT, B. L. 2014. Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med,* 371**,** 2488-98.

JAISWAL, S., NATARAJAN, P., SILVER, A. J., GIBSON, C. J., BICK, A. G., SHVARTZ, E., MCCONKEY, M., GUPTA, N., GABRIEL, S., ARDISSINO, D., BABER, U., MEHRAN, R., FUSTER, V., DANESH, J., FROSSARD, P., SALEHEEN, D., MELANDER, O., SUKHOVA, G. K., NEUBERG, D., LIBBY, P., KATHIRESAN, S. & EBERT, B. L. 2017. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *N Engl J Med,* 377**,** 111-121.

JEONG, J. W., KWAK, I., LEE, K. Y., KIM, T. H., LARGE, M. J., STEWART, C. L., KAESTNER, K. H., LYDON, J. P. & DEMAYO, F. J. 2010. Foxa2 is essential for mouse endometrial gland development and fertility. *Biol Reprod,* 83**,** 396-403.

JONSSON, H., SULEM, P., KEHR, B., KRISTMUNDSDOTTIR, S., ZINK, F., HJARTARSON, E., HARDARSON, M. T., HJORLEIFSSON, K. E., EGGERTSSON, H. P., GUDJONSSON, S. A., WARD, L. D., ARNADOTTIR, G. A., HELGASON, E. A., HELGASON, H., GYLFASON, A., JONASDOTTIR, A., JONASDOTTIR, A., RAFNAR, T., FRIGGE, M., STACEY, S. N., TH MAGNUSSON, O., THORSTEINSDOTTIR, U., MASSON, G., KONG, A., HALLDORSSON, B.

V., HELGASON, A., GUDBJARTSSON, D. F. & STEFANSSON, K. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature,* 549**,** 519-522.

KELLER, P. J., SCHMIDT, A. D., WITTBRODT, J. & STELZER, E. H. 2008. Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science,* 322**,** 1065-9.

KENNEDY, S. R., SCHMITT, M. W., FOX, E. J., KOHRN, B. F., SALK, J. J., AHN, E. H., PRINDLE, M. J., KUONG, K. J., SHEN, J. C., RISQUES, R. A. & LOEB, L. A. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc,* 9**,** 2586-606.

KRIMMEL, J. D., SCHMITT, M. W., HARRELL, M. I., AGNEW, K. J., KENNEDY, S. R., EMOND, M. J., LOEB, L. A., SWISHER, E. M. & RISQUES, R. A. 2016. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A,* 113**,** 6005-10.

KUCAB, J. E., ZOU, X., MORGANELLA, S., JOEL, M., NANDA, A. S., NAGY, E., GOMEZ, C., DEGASPERI, A., HARRIS, R., JACKSON, S. P., ARLT, V. M., PHILLIPS, D. H. & NIK-ZAINAL, S. 2019. A Compendium of Mutational Signatures of Environmental Agents. *Cell,* 177**,** 821-836 e16.

KWON, H. & PESSIN, J. E. 2013. Adipokines mediate inflammation and insulin resistance. *Front Endocrinol (Lausanne),* 4**,** 71.

LAC, V., NAZERAN, T. M., TESSIER-CLOUTIER, B., AGUIRRE-HERNANDEZ, R., ALBERT, A., LUM, A., KHATTRA, J., PRAETORIUS, T., MASON, M., CHIU, D., KOBEL, M., YONG, P. J., GILKS, B. C., ANGLESIO, M. S. & HUNTSMAN, D. G. 2019. Oncogenic mutations in histologically normal endometrium: the new normal? *J Pathol*.

LAC, V., VERHOEF, L., AGUIRRE-HERNANDEZ, R., NAZERAN, T. M., TESSIER-CLOUTIER, B., PRAETORIUS, T., ORR, N. L., NOGA, H., LUM, A., KHATTRA, J., PRENTICE, L. M., CO, D., KOBEL, M., MIJATOVIC, V., LEE, A. F., PASTERNAK, J., BLEEKER, M. C., KRAMER, B., BRUCKER, S. Y., KOMMOSS, F., KOMMOSS, S., HORLINGS, H. M., YONG, P. J., HUNTSMAN, D. G. & ANGLESIO, M. S. 2018. Iatrogenic endometriosis harbors somatic cancer-driver mutations. *Hum Reprod*.

LAMBE, M., WUU, J., WEIDERPASS, E. & HSIEH, C. C. 1999. Childbearing at older age and endometrial cancer risk (Sweden). *Cancer Causes Control,* 10**,** 43-9.

LE GALLO, M. & BELL, D. W. 2014. The emerging genomic landscape of endometrial cancer. *Clin Chem,* 60**,** 98-110.

LEE-SIX, H., ELLIS, P., OSBORNE, R. J., SANDERS, M. A. & MOORE, L. 2018a. The&#x9;landscape&#x9;of somatic&#x9; mutation in normal colorectal epithelial cells.**,**

.

LEE-SIX, H., OBRO, N. F., SHEPHERD, M. S., GROSSMANN, S., DAWSON, K., BELMONTE, M., OSBORNE, R. J., HUNTLY, B. J. P., MARTINCORENA, I., ANDERSON, E., O'NEILL, L., STRATTON, M. R., LAURENTI, E., GREEN, A. R., KENT, D. G. & CAMPBELL, P. J. 2018b. Population dynamics of normal human blood inferred from somatic mutations. *Nature,* 561**,** 473-478.

LEE-SIX, H., OLAFSSON, S., ELLIS, P., OSBORNE, R. J., SANDERS, M. A., MOORE, L., GEORGAKOPOULOS, N., TORRENTE, F., NOORANI, A., GODDARD, M., ROBINSON, P., COORENS, T. H. H., O'NEILL, L., ALDER, C., WANG, J., FITZGERALD, R. C., ZILBAUER, M., COLEMAN, N., SAEB-PARSY, K., MARTINCORENA, I., CAMPBELL, P. J. & STRATTON, M.

R. 2019. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature,* 574**,** 532-537.

LEES, B. & LEATH, C. A., 3RD 2015. The Impact of Diabetes on Gynecologic Cancer: Current Status and Future Directions. *Curr Obstet Gynecol Rep,* 4**,** 234-239.

LEWIN, S. N., HERZOG, T. J., BARRENA MEDEL, N. I., DEUTSCH, I., BURKE, W. M., SUN, X. & WRIGHT, J. D. 2010. Comparative performance of the 2009 international Federation of gynecology and obstetrics' staging system for uterine corpus cancer. *Obstet Gynecol,* 116**,** 1141-9.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H. & DURBIN, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics,* 26**,** 589-95.

LI, L. & CLEVERS, H. 2010. Coexistence of quiescent and active adult stem cells in mammals. *Science,* 327**,** 542-5.

LIANG, H., CHEUNG, L. W., LI, J., JU, Z., YU, S., STEMKE-HALE, K., DOGRULUK, T., LU, Y., LIU, X., GU, C., GUO, W., SCHERER, S. E., CARTER, H., WESTIN, S. N., DYER, M. D., VERHAAK, R. G., ZHANG, F., KARCHIN, R., LIU, C. G., LU, K. H., BROADDUS, R. R., SCOTT, K. L., HENNESSY, B. T. & MILLS, G. B. 2012. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. *Genome Res,* 22**,** 2120-9.

LOPEZ-GARCIA, C., KLEIN, A. M., SIMONS, B. D. & WINTON, D. J. 2010. Intestinal stem cell replacement follows a pattern of neutral drift. *Science,* 330**,** 822-5.

MACHIN, P., CATASUS, L., PONS, C., MUNOZ, J., MATIAS-GUIU, X. & PRAT, J. 2002. CTNNB1 mutations and beta-catenin expression in endometrial carcinomas. *Hum Pathol,* 33**,** 206-12.

MAHER, G. J., MCGOWAN, S. J., GIANNOULATOU, E., VERRILL, C., GORIELY, A. & WILKIE, A. O. 2016. Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proc Natl Acad Sci U S A,* 113**,** 2454-9.

MARTINCORENA, I. 2018. Somatic mutant clones colonize the human esophagus with age. *Science,* 362**,** 911-917.

MARTINCORENA, I., FOWLER, J. C., WABIK, A., LAWSON, A. R. J., ABASCAL, F., HALL, M. W. J., CAGAN, A., MURAI, K., MAHBUBANI, K., STRATTON, M. R., FITZGERALD, R. C., HANDFORD, P. A., CAMPBELL, P. J., SAEB-PARSY, K. & JONES, P. H. 2018. Somatic mutant clones colonize the human esophagus with age. *Science,* 362**,** 911-917.

MARTINCORENA, I., RAINE, K. M., GERSTUNG, M., DAWSON, K. J., HAASE, K., VAN LOO, P., DAVIES, H., STRATTON, M. R. & CAMPBELL, P. J. 2017. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell,* 171**,** 1029-1041 e21.

MARTINCORENA, I., ROSHAN, A., GERSTUNG, M., ELLIS, P., VAN LOO, P., MCLAREN, S., WEDGE, D. C., FULLAM, A., ALEXANDROV, L. B., TUBIO, J. M., STEBBINGS, L., MENZIES, A., WIDAA, S., STRATTON, M. R., JONES, P. H. & CAMPBELL, P. J. 2015. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science,* 348**,** 880-6.

MASUR, K., VETTER, C., HINZ, A., TOMAS, N., HENRICH, H., NIGGEMANN, B. & ZANKER, K. S. 2011. Diabetogenic glucose and insulin concentrations modulate transcriptome and protein levels involved in tumour cell migration, adhesion and proliferation. *Br J Cancer,* 104**,** 345-52.

MAURA, F., BOLLI, N., ANGELOPOULOS, N., DAWSON, K. J., LEONGAMORNLERT, D., MARTINCORENA, I., MITCHELL, T. J., FULLAM, A., GONZALEZ, S., SZALAT, R., RODRIGUEZ-MARTIN, B., SAMUR, M. K., GLODZIK, D., RONCADOR, M., FULCINITI, M., TAI, Y. T., MINVIELLE, S., MAGRANGEAS, F., MOREAU, P., CORRADINI, P., ANDERSON, K. C., TUBIO, J. M. C., WEDGE, D. C., GERSTUNG, M., AVET-LOISEAU, H., MUNSHI, N. & CAMPBELL, P. J. 2018.

MCCAMPBELL, A. S., BROADDUS, R. R., LOOSE, D. S. & DAVIES, P. J. 2006. Overexpression of the insulin-like growth factor I receptor and activation of the AKT pathway in hyperplastic endometrium. *Clin Cancer Res,* 12**,** 6373-8.

MCCULLOUGH, M. L., PATEL, A. V., PATEL, R., RODRIGUEZ, C., FEIGELSON, H. S., BANDERA, E. V., GANSLER, T., THUN, M. J. & CALLE, E. E. 2008. Body mass and endometrial cancer risk by hormone replacement therapy and cancer subtype. *Cancer Epidemiol Biomarkers Prev,* 17**,** 73-9.

MITCHELL, T. J., TURAJLIC, S., ROWAN, A., NICOL, D., FARMERY, J. H. R., O'BRIEN, T., MARTINCORENA, I., TARPEY, P., ANGELOPOULOS, N., YATES, L. R., BUTLER, A. P., RAINE, K., STEWART, G. D., CHALLACOMBE, B., FERNANDO, A., LOPEZ, J. I., HAZELL, S., CHANDRA, A., CHOWDHURY, S., RUDMAN, S., SOULTATI, A., STAMP, G., FOTIADIS, N., PICKERING, L., AU, L., SPAIN, L., LYNCH, J., STARES, M., TEAGUE, J., MAURA, F., WEDGE, D. C., HORSWELL, S., CHAMBERS, T., LITCHFIELD, K., XU, H., STEWART, A., ELAIDI, R., OUDARD, S., MCGRANAHAN, N., CSABAI, I., GORE, M., FUTREAL, P. A., LARKIN, J., LYNCH, A. G., SZALLASI, Z., SWANTON, C., CAMPBELL, P. J. & CONSORTIUM, T. R. R. 2018. Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell,* 173**,** 611-623 e17.

MORI, H., COLMAN, S. M., XIAO, Z., FORD, A. M., HEALY, L. E., DONALDSON, C., HOWS, J. M., NAVARRETE, C. & GREAVES, M. 2002. Chromosome translocations and covert leukemic clones are generated during normal fetal development. *Proc Natl Acad Sci U S A,* 99**,** 8242-7.

MORICE, P., LEARY, A., CREUTZBERG, C., ABU-RUSTUM, N. & DARAI, E. 2016. Endometrial cancer. *Lancet,* 387**,** 1094-1108.

MU, N., ZHU, Y., WANG, Y., ZHANG, H. & XUE, F. 2012. Insulin resistance: a significant risk factor of endometrial cancer. *Gynecol Oncol,* 125**,** 751-7.

MURALI, R., SOSLOW, R. A. & WEIGELT, B. 2014. Classification of endometrial carcinoma: more than two types. *Lancet Oncol,* 15**,** e268-78.

NAIR, N., CAMACHO-VANEGAS, O., RYKUNOV, D., DASHKOFF, M., CAMACHO, S. C., SCHUMACHER, C. A., IRISH, J. C., HARKINS, T. T., FREEMAN, E., GARCIA, I., PEREIRA, E., KENDALL, S., BELFER, R., KALIR, T., SEBRA, R., REVA, B., DOTTINO, P. & MARTIGNETTI, J. A. 2016. Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Mutations in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study. *PLoS Med,* 13**,** e1002206.

NAVIN, N. E. 2015. The first five years of single-cell cancer genomics and beyond. *Genome Res,* 25**,** 1499-507.

NEAD, K. T., SHARP, S. J., THOMPSON, D. J., PAINTER, J. N., SAVAGE, D. B., SEMPLE, R. K., BARKER, A., AUSTRALIAN NATIONAL ENDOMETRIAL CANCER STUDY, G., PERRY, J. R., ATTIA, J., DUNNING, A. M., EASTON, D. F., HOLLIDAY, E., LOTTA, L. A., O'MARA, T., MCEVOY, M., PHAROAH, P. D., SCOTT, R. J., SPURDLE, A. B., LANGENBERG, C., WAREHAM, N. J. & SCOTT, R. A. 2015. Evidence of a Causal Association Between

Insulinemia and Endometrial Cancer: A Mendelian Randomization Analysis. *J Natl Cancer Inst,* 107.

NICHOLSON, A. M., OLPE, C., HOYLE, A., THORSEN, A. S., RUS, T., COLOMBE, M., BRUNTON-SIM, R., KEMP, R., MARKS, K., QUIRKE, P., MALHOTRA, S., TEN HOOPEN, R., IBRAHIM, A., LINDSKOG, C., MYERS, M. B., PARSONS, B., TAVARE, S., WILKINSON, M., MORRISSEY, E. & WINTON, D. J. 2018. Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell,* 22**,** 909-918 e8.

NIK-ZAINAL, S., ALEXANDROV, L. B., WEDGE, D. C., VAN LOO, P., GREENMAN, C. D., RAINE, K., JONES, D., HINTON, J., MARSHALL, J., STEBBINGS, L. A., MENZIES, A., MARTIN, S., LEUNG, K., CHEN, L., LEROY, C., RAMAKRISHNA, M., RANCE, R., LAU, K. W., MUDIE, L. J., VARELA, I., MCBRIDE, D. J., BIGNELL, G. R., COOKE, S. L., SHLIEN, A., GAMBLE, J., WHITMORE, I., MADDISON, M., TARPEY, P. S., DAVIES, H. R., PAPAEMMANUIL, E., STEPHENS, P. J., MCLAREN, S., BUTLER, A. P., TEAGUE, J. W., JONSSON, G., GARBER, J. E., SILVER, D., MIRON, P., FATIMA, A., BOYAULT, S., LANGEROD, A., TUTT, A., MARTENS, J. W., APARICIO, S. A., BORG, A., SALOMON, A. V., THOMAS, G., BORRESEN-DALE, A. L., RICHARDSON, A. L., NEUBERGER, M. S., FUTREAL, P. A., CAMPBELL, P. J., STRATTON, M. R. & BREAST CANCER WORKING GROUP OF THE INTERNATIONAL CANCER GENOME, C. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell,* 149**,** 979-93.

NIK-ZAINAL, S., KUCAB, J. E., MORGANELLA, S., GLODZIK, D., ALEXANDROV, L. B., ARLT, V. M., WENINGER, A., HOLLSTEIN, M., STRATTON, M. R. & PHILLIPS, D. H. 2015. The genome as a record of environmental exposure. *Mutagenesis,* 30**,** 763-70.

O'CONNOR, K. A., FERRELL, R. J., BRINDLE, E., SHOFER, J., HOLMAN, D. J., MILLER, R. C., SCHECHTER, D. E., SINGER, B. & WEINSTEIN, M. 2009. Total and unopposed estrogen exposure across stages of the transition to menopause. *Cancer Epidemiol Biomarkers Prev,* 18**,** 828-36.

O'HARA, A. J. & BELL, D. W. 2012. The genomics and genetics of endometrial cancer. *Adv Genomics Genet,* 2012**,** 33-47.

OKULICZ, W. C., ACE, C. I. & SCARRELL, R. 1997. Zonal changes in proliferation in the rhesus endometrium during the late secretory phase and menses. *Proc Soc Exp Biol Med,* 214**,** 132-8.

ONSTAD, M. A., SCHMANDT, R. E. & LU, K. H. 2016. Addressing the Role of Obesity in Endometrial Cancer Risk, Prevention, and Treatment. *J Clin Oncol,* 34**,** 4225-4230.

OSORIO, F. G., ROSENDAHL HUBER, A., OKA, R., VERHEUL, M., PATEL, S. H., HASAART, K., DE LA FONTEIJNE, L., VARELA, I., CAMARGO, F. D. & VAN BOXTEL, R. 2018. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep,* 25**,** 2308-2316 e4.

PETLJAK, M., ALEXANDROV, L. B., BRAMMELD, J. S., PRICE, S., WEDGE, D. C., GROSSMANN, S., DAWSON, K. J., JU, Y. S., IORIO, F., TUBIO, J. M. C., KOH, C. C., GEORGAKOPOULOS-SOARES, I., RODRIGUEZ-MARTIN, B., OTLU, B., O'MEARA, S., BUTLER, A. P., MENZIES, A., BHOSLE, S. G., RAINE, K., JONES, D. R., TEAGUE, J. W., BEAL, K., LATIMER, C., O'NEILL, L., ZAMORA, J., ANDERSON, E., PATEL, N., MADDISON, M., NG, B. L., GRAHAM, J., GARNETT, M. J., MCDERMOTT, U., NIK-ZAINAL, S., CAMPBELL, P. J. & STRATTON, M. R. 2019. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell,* 176**,** 1282-1294 e20.

POLOZ, Y. & STAMBOLIC, V. 2015. Obesity and cancer, a case for insulin signaling. *Cell Death Dis,* 6**,** e2037.

POPIC, V., SALARI, R., HAJIRASOULIHA, I., KASHEF-HAGHIGHI, D., WEST, R. B. & BATZOGLOU, S. 2015. Fast and scalable inference of multi-sample cancer lineages. *Genome Biol,* 16**,** 91.

PRESTON-MARTIN, S., PIKE, M. C., ROSS, R. K., JONES, P. A. & HENDERSON, B. E. 1990. Increased cell division as a cause of human cancer. *Cancer Res,* 50**,** 7415-21.

RAHBARI, R., WUSTER, A., LINDSAY, S. J., HARDWICK, R. J., ALEXANDROV, L. B., TURKI, S. A., DOMINICZAK, A., MORRIS, A., PORTEOUS, D., SMITH, B., STRATTON, M. R., CONSORTIUM, U. K. & HURLES, M. E. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet,* 48**,** 126-133.

RAINE, K. M., HINTON, J., BUTLER, A. P., TEAGUE, J. W., DAVIES, H., TARPEY, P., NIK-ZAINAL, S. & CAMPBELL, P. J. 2015. cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics,* 52**,** 15 7 1-12.

RAINE, K. M., VAN LOO, P., WEDGE, D. C., JONES, D., MENZIES, A., BUTLER, A. P., TEAGUE, J. W., TARPEY, P., NIK-ZAINAL, S. & CAMPBELL, P. J. 2016. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics,* 56**,** 15 9 1-15 9 17.

REARDON, S. N., KING, M. L., MACLEAN, J. A., 2ND, MANN, J. L., DEMAYO, F. J., LYDON, J. P. & HAYASHI, K. 2012. CDH1 is essential for endometrial differentiation, gland development, and adult function in the mouse uterus. *Biol Reprod,* 86**,** 141, 1-10.

REEVES, G. K., PIRIE, K., BERAL, V., GREEN, J., SPENCER, E., BULL, D. & MILLION WOMEN STUDY, C. 2007. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ,* 335**,** 1134.

RENEHAN, A. G., TYSON, M., EGGER, M., HELLER, R. F. & ZWAHLEN, M. 2008. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet,* 371**,** 569-78.

RENEHAN, A. G., ZWAHLEN, M. & EGGER, M. 2015. Adiposity and cancer risk: new mechanistic insights from epidemiology. *Nat Rev Cancer,* 15**,** 484-98.

RISINGER, J. I., HAYES, A. K., BERCHUCK, A. & BARRETT, J. C. 1997. PTEN/MMAC1 mutations in endometrial cancers. *Cancer Res,* 57**,** 4736-8.

ROERINK, S. F., SASAKI, N., LEE-SIX, H., YOUNG, M. D., ALEXANDROV, L. B., BEHJATI, S., MITCHELL, T. J., GROSSMANN, S., LIGHTFOOT, H., EGAN, D. A., PRONK, A., SMAKMAN, N., VAN GORP, J., ANDERSON, E., GAMBLE, S. J., ALDER, C., VAN DE WETERING, M., CAMPBELL, P. J., STRATTON, M. R. & CLEVERS, H. 2018. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature,* 556**,** 457-462.

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-COTE, A. & SHAH, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nat Methods,* 11**,** 396-8.

ROUHANI, F. J., NIK-ZAINAL, S., WUSTER, A., LI, Y., CONTE, N., KOIKE-YUSA, H., KUMASAKA, N., VALLIER, L., YUSA, K. & BRADLEY, A. 2016. Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. *PLoS Genet,* 12**,** e1005932.

RUDD, M. L., PRICE, J. C., FOGOROS, S., GODWIN, A. K., SGROI, D. C., MERINO, M. J. & BELL, D. W. 2011. A unique spectrum of somatic PIK3CA (p110alpha) mutations within primary endometrial carcinomas. *Clin Cancer Res,* 17**,** 1331-40.

SCHMITT, M. W., KENNEDY, S. R., SALK, J. J., FOX, E. J., HIATT, J. B. & LOEB, L. A. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A,* 109**,** 14508-13.

SCHWAB, K. E. & GARGETT, C. E. 2007. Co-expression of two perivascular cell markers isolates mesenchymal stem-like cells from human endometrium. *Hum Reprod,* 22**,** 2903-11.

SHARPE, P. M. & FERGUSON, M. W. 1988. Mesenchymal influences on epithelial differentiation in developing systems. *J Cell Sci Suppl,* 10**,** 195-230.

SHELTON, D. N., FORNALIK, H., NEFF, T., PARK, S. Y., BENDER, D., DEGEEST, K., LIU, X., XIE, W., MEYERHOLZ, D. K., ENGELHARDT, J. F. & GOODHEART, M. J. 2012. The role of LEF1 in endometrial gland formation and carcinogenesis. *PLoS One,* 7**,** e40312.

SHERMAN, M. E. 2000. Theories of endometrial carcinogenesis: a multidisciplinary approach. *Mod Pathol,* 13**,** 295-308.

SIEGEL, R., NAISHADHAM, D. & JEMAL, A. 2013. Cancer statistics, 2013. *CA Cancer J Clin,* 63**,** 11-30.

SIMO, R., SAEZ-LOPEZ, C., BARBOSA-DESONGLES, A., HERNANDEZ, C. & SELVA, D. M. 2015. Novel insights in SHBG regulation and clinical implications. *Trends Endocrinol Metab,* 26**,** 376-83.

SIMPKINS, S. B., BOCKER, T., SWISHER, E. M., MUTCH, D. G., GERSELL, D. J., KOVATICH, A. J., PALAZZO, J. P., FISHEL, R. & GOODFELLOW, P. J. 1999. MLH1 promoter methylation and gene silencing is the primary cause of microsatellite instability in sporadic endometrial cancers. *Hum Mol Genet,* 8**,** 661-6.

SNIPPERT, H. J., VAN DER FLIER, L. G., SATO, T., VAN ES, J. H., VAN DEN BORN, M., KROON-VEENBOER, C., BARKER, N., KLEIN, A. M., VAN RHEENEN, J., SIMONS, B. D. & CLEVERS, H. 2010. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell,* 143**,** 134-44.

STRATTON, M. R., CAMPBELL, P. J. & FUTREAL, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-724.

SUDA, K., NAKAOKA, H., YOSHIHARA, K., ISHIGURO, T., TAMURA, R., MORI, Y., YAMAWAKI, K., ADACHI, S., TAKAHASHI, T., KASE, H., TANAKA, K., YAMAMOTO, T., MOTOYAMA, T., INOUE, I. & ENOMOTO, T. 2018. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep,* 24**,** 1777-1789.

TANAKA, M., KYO, S., KANAYA, T., YATABE, N., NAKAMURA, M., MAIDA, Y., OKABE, M. & INOUE, M. 2003. Evidence of the Monoclonal Composition of Human Endometrial Epithelial Glands and Mosaic Pattern of Clonal Distribution in Luminal Epithelium. *The American Journal of Pathology,* 163**,** 295-301.

TEMPEST, N., MACLEAN, A. & HAPANGAMA, D. K. 2018. Endometrial Stem Cell Markers: Current Concepts and Unresolved Questions. *Int J Mol Sci,* 19.

TOMASETTI, C. & VOGELSTEIN, B. 2015. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science,* 347**,** 78-81.

TRESSERRA, F., GRASES, P., UBEDA, A., PASCUAL, M. A., GRASES, P. J. & LABASTIDA, R. 1999. Morphological changes in hysterectomies after endometrial ablation. *Hum Reprod,* 14**,** 1473-7.

VAN LOO, P., NORDGARD, S. H., LINGJAERDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B., PEROU, C. M., BORRESEN-DALE, A. L. & KRISTENSEN, V. N. 2010. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A,* 107**,** 16910-5.

WALTHER-ANTONIO, M. R., CHEN, J., MULTINU, F., HOKENSTAD, A., DISTAD, T. J., CHEEK, E. H., KEENEY, G. L., CREEDON, D. J., NELSON, H., MARIANI, A. & CHIA, N. 2016. Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Med,* 8**,** 122.

WEI, J. J., WILLIAM, J. & BULUN, S. 2011. Endometriosis and ovarian cancer: a review of clinical, pathologic, and molecular aspects. *Int J Gynecol Pathol,* 30**,** 553-68.

WEISSMAN, I. L. 2000. Stem cells: units of development, units of regeneration, and units in evolution. *Cell,* 100**,** 157-68.

WESTIN, S. N., BROADDUS, R. R., DENG, L., MCCAMPBELL, A., LU, K. H., LACOUR, R. A., MILAM, M. R., URBAUER, D. L., MUELLER, P., PICKAR, J. H. & LOOSE, D. S. 2009. Molecular clustering of endometrial carcinoma based on estrogen-induced gene expression. *Cancer Biol Ther,* 8**,** 2126-35.

WILLIS, A., JUNG, E. J., WAKEFIELD, T. & CHEN, X. 2004. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene,* 23**,** 2330-8.

WU, Q. J., LI, Y. Y., TU, C., ZHU, J., QIAN, K. Q., FENG, T. B., LI, C., WU, L. & MA, X. X. 2015. Parity and endometrial cancer risk: a meta-analysis of epidemiological studies. *Sci Rep,* 5**,** 14243.

YATES, L. R., GERSTUNG, M., KNAPPSKOG, S., DESMEDT, C., GUNDEM, G., VAN LOO, P., AAS, T., ALEXANDROV, L. B., LARSIMONT, D., DAVIES, H., LI, Y., JU, Y. S., RAMAKRISHNA, M., HAUGLAND, H. K., LILLENG, P. K., NIK-ZAINAL, S., MCLAREN, S., BUTLER, A., MARTIN, S., GLODZIK, D., MENZIES, A., RAINE, K., HINTON, J., JONES, D., MUDIE, L. J., JIANG, B., VINCENT, D., GREENE-COLOZZI, A., ADNET, P. Y., FATIMA, A., MAETENS, M., IGNATIADIS, M., STRATTON, M. R., SOTIRIOU, C., RICHARDSON, A. L., LONNING, P. E., WEDGE, D. C. & CAMPBELL, P. J. 2015. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med,* 21**,** 751-9.

YE, K., SCHULZ, M. H., LONG, Q., APWEILER, R. & NING, Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics,* 25**,** 2865-71.

YOKOYAMA, A., KAKIUCHI, N., YOSHIZATO, T., NANNYA, Y., SUZUKI, H., TAKEUCHI, Y., SHIOZAWA, Y., SATO, Y., AOKI, K., KIM, S. K., FUJII, Y., YOSHIDA, K., KATAOKA, K., NAKAGAWA, M. M., INOUE, Y., HIRANO, T., SHIRAISHI, Y., CHIBA, K., TANAKA, H., SANADA, M., NISHIKAWA, Y., AMANUMA, Y., OHASHI, S., AOYAMA, I., HORIMATSU, T., MIYAMOTO, S., TSUNODA, S., SAKAI, Y., NARAHARA, M., BROWN, J. B., SATO, Y., SAWADA, G., MIMORI, K., MINAMIGUCHI, S., HAGA, H., SENO, H., MIYANO, S., MAKISHIMA, H., MUTO, M. & OGAWA, S. 2019. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature,* 565**,** 312-317.

ZHU, M., LU, T., JIA, Y., LUO, X., GOPAL, P., LI, L., ODEWOLE, M., RENTERIA, V., SINGAL, A. G., JANG, Y., GE, K., WANG, S. C., SOROURI, M., PAREKH, J. R., MACCONMARA, M. P., YOPP, A. C., WANG, T. & ZHU, H. 2019. Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. *Cell,* 177**,** 608-621 e12.

List of abbreviations and acronyms

ASC	**adult stem cells**
CCO	**cytochrome C oxydase**
eeASCs	**endometrial epithelial adult stem cells**
HDP	**Hierarchical Dirichlet Process**
LCM	**Laser-capture microscopy**
LRCs	**Label retaining cells**
NGS	**next generation sequencing**
SNV	**single nucleotide variant**
SV	**structural variant**
TA	**Transient amplifying**
TAH	**total abdominal hysterectomy**
VAF	**variant allele fraction**
WGS	**Whole genome sequencing**

Appendix 1

Tissue_type	**Biopsy_site**	**Structure**	**SampleID**
Appendix	Appendix_tip	Crypt	PD28690bv_APP1_F2
Appendix	Appendix_tip	Crypt	PD28690bv_APP1_G3
Appendix	Appendix_tip	Crypt	PD28690bv_APP_4_A7
Appendix	Appendix_tip	Crypt	PD28690bv_APP_4_A8
Appendix	Appendix_tip	Crypt	PD28690bv_APP_4_C7
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_A1
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_B2
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_C4
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_D1
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_D2
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_D4
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_D5
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_F2
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_F3
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_F4
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_G3
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_G4
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_H3
Appendix	Appendix_mid	Crypt	PD28690bw_APP_3_H4
Colon	Colon_transverse	Crypt	PD28690cc_COL_1_B11
Colon	Colon_transverse	Crypt	PD28690cc_COL_1_B12
Colon	Colon_transverse	Crypt	PD28690cc_COL_1_C11
Colon	Colon_transverse	Crypt	PD28690cc_COL_1_C12
Colon	Colon_transverse	Crypt	PD28690cc_COL_2_B8
Colon	Colon_transverse	Crypt	PD28690cc_COL_2_F9
Colon	Colon_transverse	Crypt	PD28690cc_COL_2_G8
Colon	Colon_transverse	Crypt	PD28690cc_COL_2_G9
Colon	Colon_transverse	Crypt	PD28690cc_COL_2_H8
Colon	Colon_transverse	Crypt	PD28690cc_COL_5_A3
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_A9
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_B8
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_B9
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_D8
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_E9
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_G8
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_G9
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_H8
Small_intestine	Jejunum	Crypt	PD28690bp_SB1_H9
Small_intestine	Ileum	Crypt	PD28690bt_SB2_A11
Small_intestine	Ileum	Crypt	PD28690bt_SB2_F10
Small_intestine	Ileum	Crypt	PD28690bt_SB2_F11
Small_intestine	Ileum	Crypt	PD28690bt_SB2_G10
Small_intestine	Ileum	Crypt	PD28690bt_SB2_H10
Small_intestine	Ileum	Crypt	PD28690bt_SB3_B5
Small_intestine	Ileum	Crypt	PD28690bt_SB3_F5
Liver_bile_duct	Liver_left_lobe	Bile_duct	PD28690cr_BD_3_A8
Liver_bile_duct	Liver_left_lobe	Bile_duct	PD28690cr_BD_3_A9
Liver_bile_duct	Liver_left_lobe	Bile_duct	PD28690cr_BD_3_C7
Liver_bile_duct	Liver_left_lobe	Bile_duct	PD28690cr_BD_3_C8

Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690cx_BD_2_C1
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690da_BD_5_A1
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690da_BD_5_C1
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690da_BD_5_E1
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690db_BD_6_A2
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690db_BD_7_A3
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690db_BD_7_C3
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690di_BD_1_B2
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dj_BD_8_C4
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dr_BD_4_A10
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dr_BD_4_C10
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dr_BD_4_E10
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dw_BD_10_A6
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dw_BD_9_A5
Liver_bile_duct	Liver_right_lobe	Bile_duct	PD28690dw_BD_9_G5
Liver_parenchyma	Liver_right_lobe	Liver_parenchyma	PD28690di_HEP1_Z2
Liver_parenchyma	Liver_right_lobe	Liver_parenchyma	PD28690di_HEP2_Z1
Liver_parenchyma	Liver_right_lobe	Liver_parenchyma	PD28690di_HEP2_Z2
Ureter	Urothelium	Urothelium	PD28690ip_U_1_C5
Ureter	Urothelium	Urothelium	PD28690ip_U_1_A5
Ureter	Urothelium	Urothelium	PD28690ip_U_1_B5
Ureter	Urothelium	Urothelium	PD28690ip_U_1_D5
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_B1
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_B3
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_C2
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_D2
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_F2
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L1_G2
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L2_C3
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L2_C4
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L2_F3
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L4_B6
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L4_C6
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L4_E6
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L4_F5
Testis	Left_testis	Seminiferous_tubule	PD28690id_T3_L4_H5
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES1_CU1
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES1_CU2
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES2_CU1
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES2_CU2
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU1
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU2
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU3
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU4
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU5
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU6
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU7
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU8
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU1
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU2
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU3
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU5
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU6
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU3
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU4
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU5
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU7
Oesophagus	Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU8

Prostate	Prostate_right_lobe	Acinus	PD28690fd_PA_1_A1
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_A10
Prostate	Prostate_right_lobe	Acinus	PD28690fd_PA_1_A2
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_A3
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_A6
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_A8
Prostate	Prostate_right_lobe	Acinus	PD28690fd_PA_1_C2
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_C3
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_C5
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_C6
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_E10
Prostate	Prostate_left_lobe	Acinus	PD28690fd_PA_1_E12
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_A2
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_C2
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_E1
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_E2
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_G1
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L2_A3
Bronchus_epithelium	Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L2_C3
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_A7
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_A8
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_B8
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_D9
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_G9
Bronchus_epithelium	Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_H7
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1A
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG1D
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1D
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG2B
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG1B
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG2A
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG2A
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG2B
Bronchus_sero_mucous_glands	Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1C
Adrenal_gland_cortex	Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L1
Adrenal_gland_cortex	Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L2
Adrenal_gland_cortex	Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L3
Adrenal_gland_cortex	Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L4
Adrenal_gland_cortex	Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L5
Adrenal_gland_cortex	Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L1
Adrenal_gland_cortex	Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L2
Adrenal_gland_cortex	Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L3
Adrenal_gland_cortex	Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L4
Adrenal_gland_cortex	Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L5
Adrenal_gland_cortex	Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L1
Adrenal_gland_cortex	Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L2
Adrenal_gland_cortex	Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L3
Adrenal_gland_cortex	Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L4
Adrenal_gland_cortex	Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L5
Periadrenal_visceral_fat	Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L1
Periadrenal_visceral_fat	Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L2
Periadrenal_visceral_fat	Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L3
Periadrenal_visceral_fat	Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L4
Periadrenal_visceral_fat	Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L5
Skin_sebaceous_gland	Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_C2
Skin_sebaceous_gland	Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_E1
Skin_sebaceous_gland	Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_H1
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_3_E3

Kidney	Right_kidney_superior	Glomerulus	PD28690hk_KD_3_A3
Kidney	Right_kidney_superior	Glomerulus	PD28690hk_KD_5_G2
Kidney	Right_kidney_superior	Glomerulus	PD28690hk_KD_1_D1
Kidney	Right_kidney_superior	Proximal_tubule	PD28690hk_KD_6_A2
Kidney	Right_kidney_superior	Glomerulus	PD28690hk_KD_6_A4
Kidney	Right_kidney_superior	Proximal_tubule	PD28690hk_KD_5_H2
Kidney	Right_kidney_superior	Proximal_tubule	PD28690hk_KD_4_D4
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_4_A4
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_5_E2
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_4_C4
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_1_A1
Kidney	Right_kidney_superior	Proximal_tubule	PD28690hk_KD_1_E1
Kidney	Right_kidney_superior	Distal_tubule	PD28690hk_KD_6_G3
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F1_2_A12
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F2_2_B12
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F3_2_C12
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F4_2_D12
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F5_2_E12
Thyroid	Thyroid_left_inferior_lobe	Follicle	PD28690fl_F6_2_F12
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_A1
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_A11
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_B1
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F2_1_B11
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F2_2_B2
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F3_1_C11
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F4_1_D11
Thyroid	Thyroid_left_superior_lobe	Follicle	PD28690fm_F5_1_E11
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_EW_CT_A2
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_EW_CT_D3
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_1_A1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_3_E1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_4_G1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_6_G2
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F2_3_F1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F2_6_H2
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F3_1_C1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F3_5_F2
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F4_1_E1
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_F5_1_A3
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_L1_CL2_C3
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_L1_CL4_G3
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_L2_CL2_C7
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_L5_CL2_G5
Thyroid	Thyroid_right_superior_lobe	Follicle	PD28690fq_L5_CL3_A7
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_C10
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_C9
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_E10
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_E9
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_G10
Heart	Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_G9
Bladder	Bladder_left_wall	Urothelium	PD28690ch_BL2_CU1_L3_4_D
Bladder	Bladder_left_wall	Urothelium	PD28690ch_BL2_CU2_L3_4_E
Bladder	Bladder_left_wall	Urothelium	PD28690ch_BL2_CU3_L3_4_F
Bladder	Bladder_right_wall	Urothelium	PD28690cm_BL1_CU1_L1_2_A
Bladder	Bladder_right_wall	Urothelium	PD28690cm_BL1_CU2_L1_2_B
Bladder	Bladder_right_wall	Urothelium	PD28690cm_BL1_CU3_L3_4_C
Bladder	Bladder_right_wall	Urothelium	PD28690cm_BL1_CU4_L3_4_H
Artery	Right_kidney_superior	Renal arteriole	PD28690hk_RA_1_F5

Appendix 2

Sex	**Female**	**Sex**	**Male**
Age	**54**	**Age**	**47**
Donor ID	**11-S11**	**Donor ID**	**11-S7**
Tissue	Adrenal gland	Tissue	Adrenal gland
Tissue	Bladder (urinary)	Tissue	Bladder (urinary)
Tissue	Brain, Cerebellum	Tissue	Brain, Cerebellum
Tissue	Breast	Tissue	Cecum
Tissue	Cecum	Tissue	Colon, ascending
Tissue	Colon	Tissue	Colon, descending
Tissue	Duodenum	Tissue	Colon, sigmoid
Tissue	Fallopian tube	Tissue	Colon,transversal
Tissue	Gallbladder	Tissue	Duodenum
Tissue	GI Tract	Tissue	Esophagus
Tissue	Ileum	Tissue	Gallbladder
Tissue	Jejunum	Tissue	Ileum
Tissue	Kidney	Tissue	Jejunum
Tissue	Kidney, cortex	Tissue	Kidney
Tissue	Kidney, medulla	Tissue	Kidney, medulla
Tissue	Liver	Tissue	Liver
Tissue	Lung	Tissue	Lung
Tissue	Ovary	Tissue	Pancreas
Tissue	Pancreas	Tissue	Prostate
Tissue	Rectum	Tissue	Rectum
Tissue	Skin	Tissue	Salivary gland
Tissue	Stomach (fundus)	Tissue	Skin
Tissue	Thyroid	Tissue	Stomach (fundus)
Tissue	Uterus, cervix	Tissue	Testis
Tissue	Uterus, endometrium	Tissue	Thyroid

Appendix 3

Fixation of Frozen Tissue Sections for LCM

Ethanol

Add 100 ul of 70% ethanol to a single slide with unfixed frozen sections for 2-3 minutes

Wash 2-3x with PBS (10 sec)

Place slides into a petri dish/coplin jar with PBS until ready for staining

Phosphate-buffered paraformaldehyde 4%

Add 100 ul 4% phosphate-buffered paraformaldehyde (PFA) to a single slide with unfixed frozen sections for 5 minutes

Wash 3x with PBS

Place slides into a petri dish/coplin jar with PBS before staining

Phosphate-buffered paraformaldehyde 1%

Add 100 ul 1% phosphate-buffered paraformaldehyde (PFA) to a single slide with unfixed frozen sections for 5 minutes

Wash 3x with PBS

Place slides into a petri dish/ coplin jar with PBS before staining

Methanol

Add 100 ul of ice-cold methanol to a single slide with unfixed frozen sections for 2-3 minutes

Wash 3x with PBS

Place slides into a petri dish/coplin jar with PBS before staining

Acetone

Add 100 ul of ice-cold acetone to a single slide with unfixed frozen sections for 2 minutes

Wash 3x with PBS

Place slides into a petri dish/coplin jar with PBS before staining

Staining frozen sections with haematoxylin

Staining should be done in a fume hood (CGP Containment level 1 lab)

Ensure stains and alcohols have been recently changed

Place fixed unstained tissue slides into haematoxylin for 10 seconds

Rinse 2x with tap water

Place the slides into 70% ethanol 2x for approximately 5 seconds

Place the slides into 100% ethanol 2x for approximately 5 seconds

Place the slides into xylene 1x for 5 seconds

Staining frozen sections with haematoxylin and eosin

Staining should be done in a fume hood (CGP Containment level 1 lab)

Ensure stains and alcohols have been recently changed

Place fixed unstained tissue slides into haematoxylin for 10 seconds

Rinse with tap water 2x

Place slides into eosin for 5 seconds

Rinse with tap water 1x

Place the slides into 70% ethanol for 5-10 seconds

Place the slides into 100% ethanol 2x for 5-10 seconds

Place the slides into xylene (or Neo-clear xylene substitute) for 5 seconds

Appendix 4

H&E staining for LCM paraffin sections

Staining should be done in a fume hood (CGP Containment level 1 lab)

Remove paraffin/dewax by sequential immersion in the following:

Xylene – 2 min

Xylene – 2 min

Ethanol 100% – 1 min

Ethanol 100% – 1 min

Ethanol 70% - 1 min

Deion – 1 min

Stain with Haematoxylin (Gills) and eosin

Haematoxylin – 10-20 sec

Tap water – 20 sec wash

Tap water – 20 sec wash

Eosin – 5-10 sec

Tap water – 10-20 sec wash

Ethanol 70% - 10-20 sec

Ethanol 70% - 10-20 sec

Ethanol 100% -10-20 sec

Ethanol 100% -10-20 sec

Xylene – 10-20 sec

Appendix 5

PAXGENE PROTOCOL

General Information on PAXgene

PAXgene Tissue FIX rapidly penetrates and fixes tissue, with a fixation rate of approximately 1 mm/30 minutes. The reagent preserves morphology and biomolecules without the destructive cross-linking and degradation associated with formalin fixation.

The process includes two steps:

1. Tissue fixation – Immersion of tissue in PAXgene Tissue FIX
2. Tissue stabilisation and storage - PAXgene Tissue STABILIZER. Tissue samples can be stored in PAXgene Tissue STABILIZER for 7 days at room temperature, up to 4 weeks at 2-8°C and indefinitely at -20°C or -80°C.

Equipment needed

PAXgene Tissue Fix Container

PAXgene Tissue STABILIZER

Tissue Cassettes (for smaller biopsies)

Use one of the following protocols:

Protocol A: for storing multiple small biopsies in a Single PAXgene Tissue FIX Container.

Protocol B: for storing a single biopsy (20 x 20 x 20 mm) in a PAXgene Tissue FIX Container.

Protocol A - For multiple small samples

1 - Resect and cut tissue into max. 4 x 15 x 15 mm sections.

2 - Place each section into a tissue cassette.

3 - Place up to 4 tissue cassettes into a single PAXgene Tissue FIX Container.

4 – Fixation at room temperature for 2 – 24 hours, depending on tissue type and size, assuming a fixation rate of approximately 1mm in 30 minutes. **Recommended standard fixation time of 24 hours**.

5 – After fixation step is complete pour off the PAXgene Tissue FIX solution from the Tissue FIX Container and fill the container with PAXgene Tissue STABILIZER.

7 – Transfer to -20°C or -80°C for long-term storage.

Protocol B – For a single, larger tissue sample

1 – Tissue sample can have max. dimensions 20 x 20 x 20 mm.

2 - Place tissue directly into a PAXgene Tissue FIX Container.

3 – Fixation at room temperature for 6 – 48 hours, depending on tissue type and size, assuming a fixation rate of approximately 1mm in 30 minutes. **Recommended standard fixation time of 48 hours**.

4 – After fixation step is complete pour off the PAXgene Tissue FIX solution from the Tissue FIX Container and fill the container with PAXgene Tissue STABILIZER.

5 – Transfer to -20°C or -80°C for long-term storage.

Appendix 6

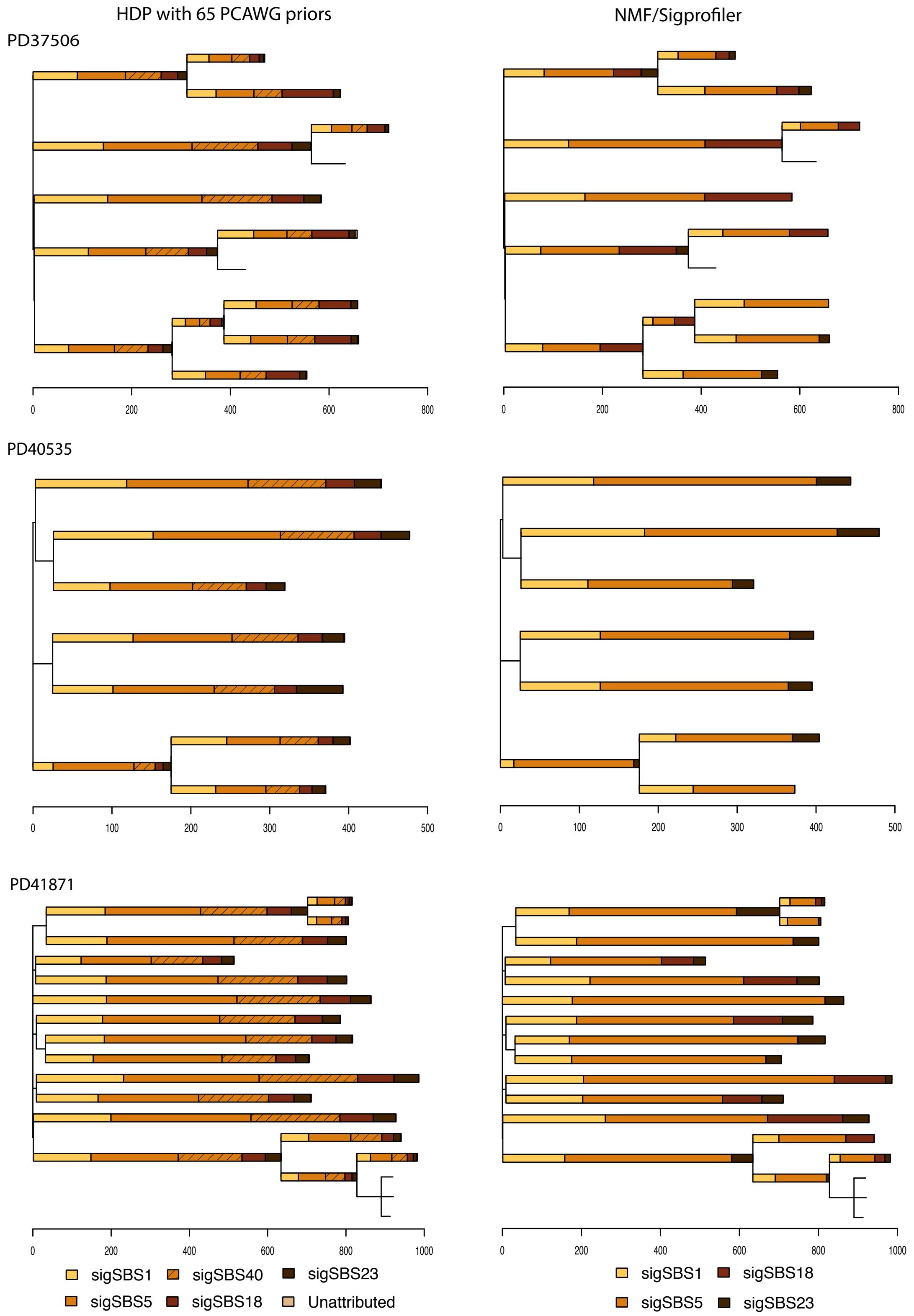
Biopsy_site	Structure	SampleID	Seq_depth
Appendix_tip	Crypt	PD28690bv_APP1_C3	54.0
Appendix_tip	Crypt	PD28690bv_APP1_F2	55.6
Appendix_tip	Crypt	PD28690bv_APP1_G3	40.5
Appendix_tip	Crypt	PD28690bv_APP_4_A7	35.5
Appendix_tip	Crypt	PD28690bv_APP_4_A8	33.9
Appendix_tip	Crypt	PD28690bv_APP_4_C7	25.2
Appendix_mid	Crypt	PD28690bw_APP_3_A1	26.7
Appendix_mid	Crypt	PD28690bw_APP_3_B2	25.2
Appendix_mid	Crypt	PD28690bw_APP_3_C4	31.8
Appendix_mid	Crypt	PD28690bw_APP_3_D1	27.6
Appendix_mid	Crypt	PD28690bw_APP_3_D2	28.1
Appendix_mid	Crypt	PD28690bw_APP_3_D4	32.1
Appendix_mid	Crypt	PD28690bw_APP_3_D5	27.1
Appendix_mid	Crypt	PD28690bw_APP_3_F2	35.2
Appendix_mid	Crypt	PD28690bw_APP_3_F3	23.8
Appendix_mid	Crypt	PD28690bw_APP_3_F4	26.8
Appendix_mid	Crypt	PD28690bw_APP_3_G3	23.7
Appendix_mid	Crypt	PD28690bw_APP_3_G4	30.3
Appendix_mid	Crypt	PD28690bw_APP_3_H3	24.8
Appendix_mid	Crypt	PD28690bw_APP_3_H4	27.2
Colon_transverse	Crypt	PD28690cc_COL_1_B11	20.9
Colon_transverse	Crypt	PD28690cc_COL_1_B12	15.4
Colon_transverse	Crypt	PD28690cc_COL_1_C11	18.9
Colon_transverse	Crypt	PD28690cc_COL_1_C12	15.3
Colon_transverse	Crypt	PD28690cc_COL_2_B8	29.9
Colon_transverse	Crypt	PD28690cc_COL_2_F9	24.2
Colon_transverse	Crypt	PD28690cc_COL_2_G8	33.1
Colon_transverse	Crypt	PD28690cc_COL_2_G9	25.9
Colon_transverse	Crypt	PD28690cc_COL_2_H8	27.5
Colon_transverse	Crypt	PD28690cc_COL_5_A3	51.9
Jejunum	Crypt	PD28690bp_SB1_A9	19.3
Jejunum	Crypt	PD28690bp_SB1_B8	38.4
Jejunum	Crypt	PD28690bp_SB1_B9	26.6
Jejunum	Crypt	PD28690bp_SB1_D8	27.2
Jejunum	Crypt	PD28690bp_SB1_E9	26.9
Jejunum	Crypt	PD28690bp_SB1_G8	17.8
Jejunum	Crypt	PD28690bp_SB1_G9	29.0
Jejunum	Crypt	PD28690bp_SB1_H8	15.5
Jejunum	Crypt	PD28690bp_SB1_H9	50.6
Ileum	Crypt	PD28690bt_SB2_A11	21.4
Ileum	Crypt	PD28690bt_SB2_F10	23.8
Ileum	Crypt	PD28690bt_SB2_F11	24.0
Ileum	Crypt	PD28690bt_SB2_G10	17.5
Ileum	Crypt	PD28690bt_SB2_H10	17.2
Ileum	Crypt	PD28690bt_SB3_B5	19.2
Ileum	Crypt	PD28690bt_SB3_F5	22.3
Liver_left_lobe	Bile_duct	PD28690cr_BD_3_A8	21.8
Liver_left_lobe	Bile_duct	PD28690cr_BD_3_A9	15.3
Liver_left_lobe	Bile_duct	PD28690cr_BD_3_C7	16.9
Liver_left_lobe	Bile_duct	PD28690cr_BD_3_C8	25.2
Liver_right_lobe	Bile_duct	PD28690cx_BD_2_C1	24.9
Liver_right_lobe	Bile_duct	PD28690da_BD_5_A1	30.3

Liver_right_lobe	Bile_duct	PD28690da_BD_5_C1	30.2
Liver_right_lobe	Bile_duct	PD28690da_BD_5_E1	31.1
Liver_right_lobe	Bile_duct	PD28690db_BD_6_A2	26.2
Liver_right_lobe	Bile_duct	PD28690db_BD_7_A3	26.1
Liver_right_lobe	Bile_duct	PD28690db_BD_7_C3	30.3
Liver_right_lobe	Bile_duct	PD28690di_BD_1_B2	16.1
Liver_right_lobe	Bile_duct	PD28690dj_BD_8_C4	27.7
Liver_right_lobe	Bile_duct	PD28690dr_BD_4_A10	31.9
Liver_right_lobe	Bile_duct	PD28690dr_BD_4_C10	30.1
Liver_right_lobe	Bile_duct	PD28690dr_BD_4_E10	32.7
Liver_right_lobe	Bile_duct	PD28690dw_BD_10_A6	30.0
Liver_right_lobe	Bile_duct	PD28690dw_BD_9_A5	33.5
Liver_right_lobe	Bile_duct	PD28690dw_BD_9_G5	30.6
Liver_right_lobe	Liver_parenchyma	PD28690di_HEP1_Z2	23.9
Liver_right_lobe	Liver_parenchyma	PD28690di_HEP2_Z1	26.2
Liver_right_lobe	Liver_parenchyma	PD28690di_HEP2_Z2	30.1
Urothelium	Urothelium	PD28690ip_U_1_C5	35.0
Urothelium	Urothelium	PD28690ip_U_1_A5	31.6
Urothelium	Urothelium	PD28690ip_U_1_B5	30.3
Urothelium	Urothelium	PD28690ip_U_1_D5	29.9
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_B1	26.4
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_B3	28.1
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_C2	26.8
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_D2	26.9
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_F2	28.7
Left_testis	Seminiferous_tubule	PD28690id_T3_L1_G2	28.7
Left_testis	Seminiferous_tubule	PD28690id_T3_L2_C3	30.6
Left_testis	Seminiferous_tubule	PD28690id_T3_L2_C4	29.0
Left_testis	Seminiferous_tubule	PD28690id_T3_L2_F3	31.7
Left_testis	Seminiferous_tubule	PD28690id_T3_L4_B6	24.4
Left_testis	Seminiferous_tubule	PD28690id_T3_L4_C6	26.5
Left_testis	Seminiferous_tubule	PD28690id_T3_L4_E6	15.9
Left_testis	Seminiferous_tubule	PD28690id_T3_L4_F5	28.6
Left_testis	Seminiferous_tubule	PD28690id_T3_L4_H5	25.2
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES1_CU1	22.9
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES1_CU2	22.7
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES2_CU1	43.1
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES2_CU2	45.6
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU1	33.9
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU2	24.2
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU3	28.0
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU4	32.7
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU5	20.5
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU6	31.7
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU7	32.9
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES3_CU8	35.4
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU1	40.4
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU2	30.1
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU3	35.9
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU5	45.4
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES4_CU6	41.3
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU3	33.1
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU4	31.8
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU5	30.1
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU7	27.5
Oesophagus_upper_third	Squamous_epithelium	PD28690bl_OES5_CU8	34.2
Prostate_right_lobe	Acinus	PD28690fd_PA_1_A1	26.4
Prostate_left_lobe	Acinus	PD28690fd_PA_1_A10	26.0

Prostate_right_lobe	Acinus	PD28690fd_PA_1_A2	26.8
Prostate_left_lobe	Acinus	PD28690fd_PA_1_A3	27.2
Prostate_left_lobe	Acinus	PD28690fd_PA_1_A6	29.8
Prostate_left_lobe	Acinus	PD28690fd_PA_1_A8	21.8
Prostate_right_lobe	Acinus	PD28690fd_PA_1_C2	19.8
Prostate_left_lobe	Acinus	PD28690fd_PA_1_C3	25.3
Prostate_left_lobe	Acinus	PD28690fd_PA_1_C5	26.1
Prostate_left_lobe	Acinus	PD28690fd_PA_1_C6	26.4
Prostate_left_lobe	Acinus	PD28690fd_PA_1_E10	28.0
Prostate_left_lobe	Acinus	PD28690fd_PA_1_E12	29.8
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_A2	35.6
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_C2	34.0
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_E1	44.2
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_E2	39.0
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L1_G1	31.0
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L2_A3	36.7
Left_distal_bronchus	Bronchial_epithelium	PD28690ef_BR4_L2_C3	42.7
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_A7	31.7
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_A8	36.6
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_B8	34.9
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_D9	35.0
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_G9	37.8
Right_distal_bronchus	Bronchial_epithelium	PD28690eh_BR5_L2_H7	41.4
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1A	37.3
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG1D	32.9
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1D	31.7
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG2B	30.3
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG1B	29.1
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG2A	29.0
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG2A	28.0
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L2_SMG2B	17.0
Left_distal_bronchus	Sero_mucous_gland	PD28690ef_BR4_L1_SMG1C	15.5
Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L1	40.1
Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L2	40.4
Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L3	39.5
Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L4	36.1
Zona_fasciculata	Zona_fasciculata	PD28690gu_AG1_ZF_L5	39.1
Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L1	25.2
Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L2	40.4
Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L3	30.8
Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L4	36.4
Zona_glomerulosa	Zona_glomerulosa	PD28690gu_AG1_ZG_L5	27.7
Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L1	20.7
Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L2	34.0
Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L3	35.1
Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L4	34.6
Zona_reticularis	Zona_reticularis	PD28690gu_AG1_ZR_L5	34.6
Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L1	23.5
Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L2	27.5
Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L3	24.4
Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L4	43.4
Visceral_fat	Visceral_fat	PD28690gu_AG1_AT_L5	24.6
Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_C2	24.2
Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_E1	36.2
Skin_lower_abdomen	Skin_sebaceous_gland	PD28690bf_SKN2_H1	33.1
Right_kidney_superior	Distal_tubule	PD28690hk_KD_3_E3	26.9
Right_kidney_superior	Glomerulus	PD28690hk_KD_3_A3	26.4
Right_kidney_superior	Glomerulus	PD28690hk_KD_5_G2	24.9

Right_kidney_superior	Glomerulus	PD28690hk_KD_1_D1	23.8
Right_kidney_superior	Proximal_tubule	PD28690hk_KD_6_A2	22.7
Right_kidney_superior	Glomerulus	PD28690hk_KD_6_A4	22.7
Right_kidney_superior	Proximal_tubule	PD28690hk_KD_5_H2	21.5
Right_kidney_superior	Proximal_tubule	PD28690hk_KD_4_D4	21.0
Right_kidney_superior	Distal_tubule	PD28690hk_KD_4_A4	18.9
Right_kidney_superior	Distal_tubule	PD28690hk_KD_5_E2	18.4
Right_kidney_superior	Distal_tubule	PD28690hk_KD_4_C4	18.2
Right_kidney_superior	Distal_tubule	PD28690hk_KD_1_A1	18.1
Right_kidney_superior	Proximal_tubule	PD28690hk_KD_1_E1	15.9
Right_kidney_superior	Distal_tubule	PD28690hk_KD_6_G3	15.2
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F1_2_A12	31.1
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F2_2_B12	27.5
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F3_2_C12	28.8
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F4_2_D12	31.7
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F5_2_E12	20.2
Thyroid_left_inferior_lobe	Follicle	PD28690fl_F6_2_F12	19.8
Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_A1	61.2
Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_A11	26.3
Thyroid_left_superior_lobe	Follicle	PD28690fm_F1_1_B1	50.7
Thyroid_left_superior_lobe	Follicle	PD28690fm_F2_1_B11	34.0
Thyroid_left_superior_lobe	Follicle	PD28690fm_F2_2_B2	32.9
Thyroid_left_superior_lobe	Follicle	PD28690fm_F3_1_C11	27.1
Thyroid_left_superior_lobe	Follicle	PD28690fm_F4_1_D11	28.7
Thyroid_left_superior_lobe	Follicle	PD28690fm_F5_1_E11	29.8
Thyroid_right_superior_lobe	Follicle	PD28690fq_EW_CT_A2	31.9
Thyroid_right_superior_lobe	Follicle	PD28690fq_EW_CT_D3	36.3
Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_1_A1	57.0
Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_3_E1	56.2
Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_4_G1	46.7
Thyroid_right_superior_lobe	Follicle	PD28690fq_F1_6_G2	60.8
Thyroid_right_superior_lobe	Follicle	PD28690fq_F2_3_F1	42.4
Thyroid_right_superior_lobe	Follicle	PD28690fq_F2_6_H2	57.1
Thyroid_right_superior_lobe	Follicle	PD28690fq_F3_1_C1	29.5
Thyroid_right_superior_lobe	Follicle	PD28690fq_F3_5_F2	54.0
Thyroid_right_superior_lobe	Follicle	PD28690fq_F4_1_E1	30.1
Thyroid_right_superior_lobe	Follicle	PD28690fq_F5_1_A3	45.8
Thyroid_right_superior_lobe	Follicle	PD28690fq_L1_CL2_C3	15.3
Thyroid_right_superior_lobe	Follicle	PD28690fq_L1_CL4_G3	25.0
Thyroid_right_superior_lobe	Follicle	PD28690fq_L2_CL2_C7	24.0
Thyroid_right_superior_lobe	Follicle	PD28690fq_L5_CL2_G5	25.8
Thyroid_right_superior_lobe	Follicle	PD28690fq_L5_CL3_A7	19.7
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_C10	22.0
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_C9	27.5
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_E10	16.6
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_E9	29.1
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_G10	19.1
Heart_left_ventricle	Cardiac_myocytes	PD28690gd_HEART_2_G9	29.8
Bladder_left_wall	Urothelium	PD28690ch_BL2_CU1_L3_4_D11	34.7
Bladder_left_wall	Urothelium	PD28690ch_BL2_CU2_L3_4_E11	34.4
Bladder_left_wall	Urothelium	PD28690ch_BL2_CU3_L3_4_F11	28.9
Bladder_right_wall	Urothelium	PD28690cm_BL1_CU1_L1_2_A10	39.6
Bladder_right_wall	Urothelium	PD28690cm_BL1_CU2_L1_2_B10	42.4
Bladder_right_wall	Urothelium	PD28690cm_BL1_CU3_L3_4_G10	31.0
Bladder_right_wall	Urothelium	PD28690cm_BL1_CU4_L3_4_H10	37.2
Right_kidney_superior	Renal arteriole	PD28690hk_RA_1_F5	21.7

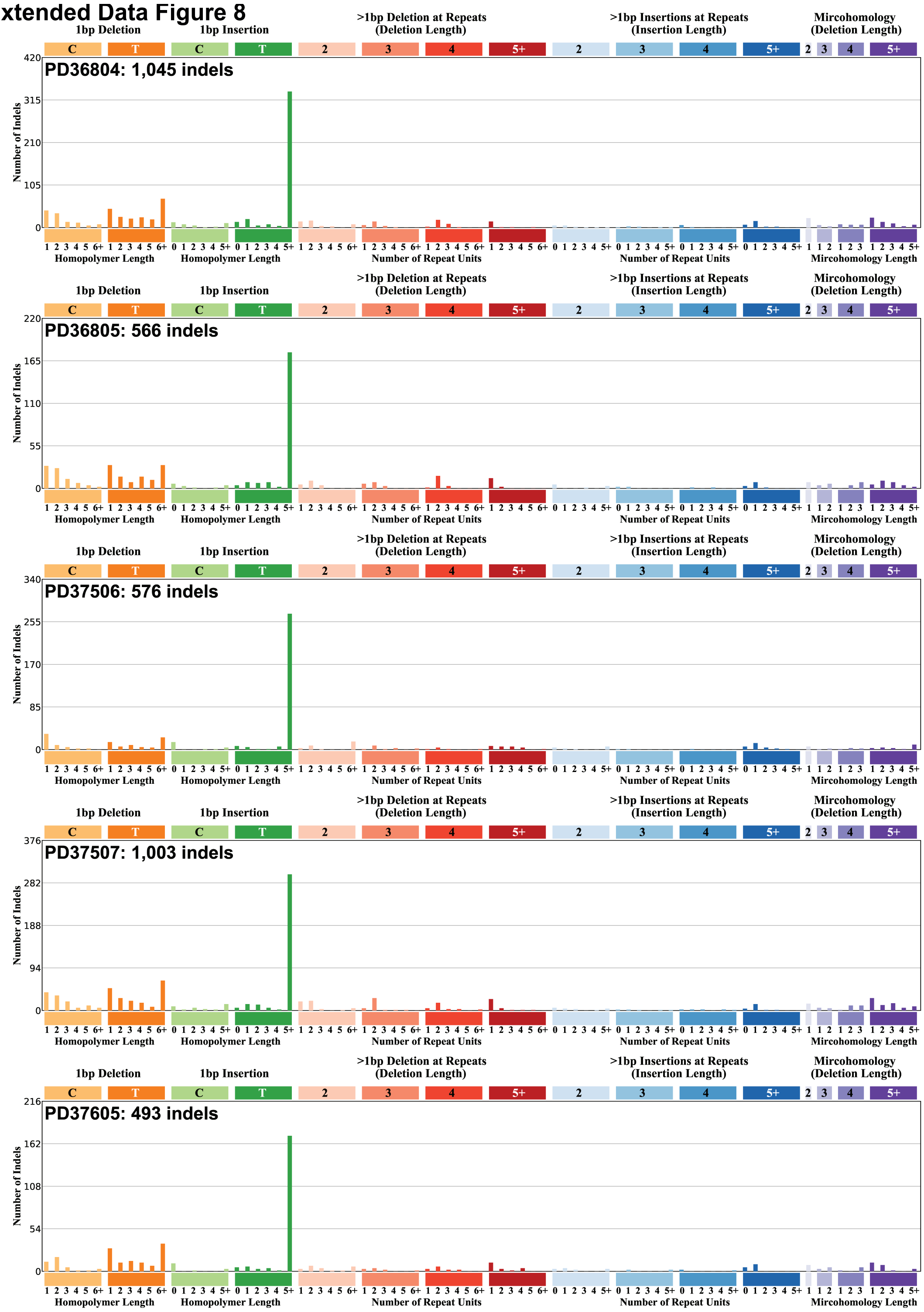
Appendix 7



Comparison of SBS signatures using two different approaches: HDP with 65 PCAWG priors and NMF with Sigprofiler attribution. Final signatures from HDP with 65 priors and NMF extraction and attribution for selected individuals.

Appendix 8

Extended Data Figure 8



Composite mutational spectra of all small insertions and deletions (indels) for each donor. Indels were classified and composite mutational spectra for each individual were generated; due to the relative sparsity of indels detected, no formal signature extraction was performed.

Appendix 9

R Notebook

Luiza Moore

26062019

Modelling total mutation burden in normal endometrium

Markdown file to document methods used in the analysis of the total mutation burden in normal endometrium.

Load Libraries

```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(sjPlot)
library(sjmisc)
```

Load in data

Load in sample level data for 28 donors with associated meta-data on age, body mass index (BMI) and parity.

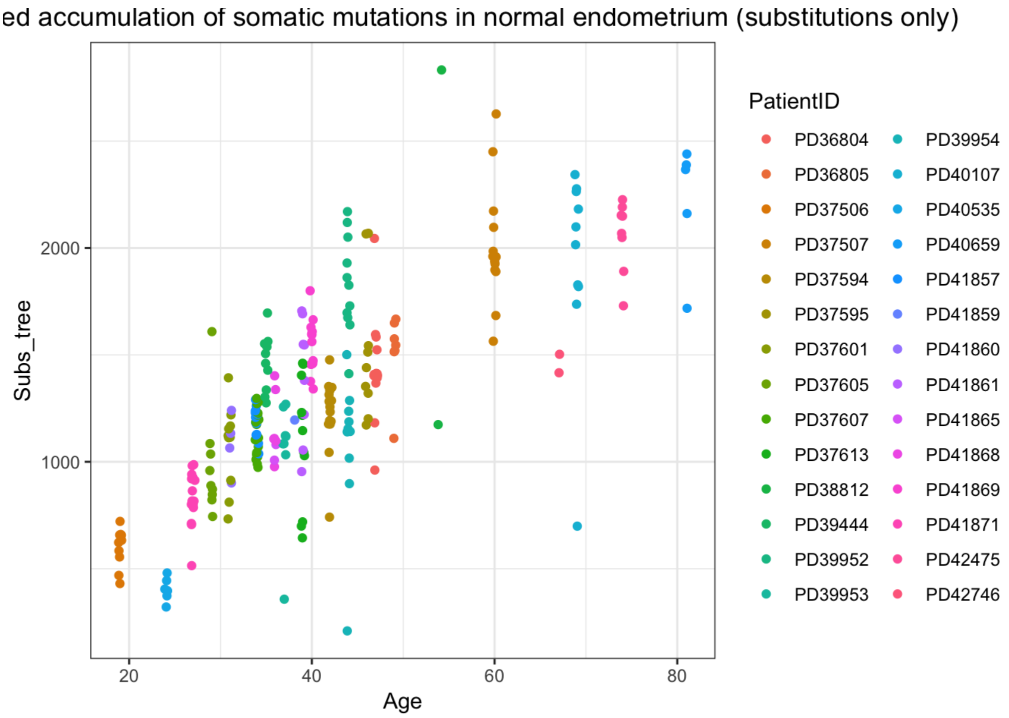
```
endom_burden <- read.csv("~/Desktop/Endometrium_for_model_26062019.csv")

# Samples per patient
endom_burden %>% group_by(PatientID) %>%  count(PatientID) %>%  rename(`Sample count` = n)
 %>% arrange(desc(`Sample count`)) %>%  kable() %>%  kable_styling(bootstrap_options = c("s
triped", "condensed"), full_width = F, position = "left")
```

PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD37507	14
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11

PatientID	Sample count
PD37506	10
PD37601	10
PD39444	10
PD39954	10
PD40107	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD42475	8
PD36805	7
PD40535	7
PD41868	6
PD40659	5
PD41860	4
PD38812	2
PD41865	2
PD42746	2
PD41859	1

```
# Look at raw data
endom_burden %>% ggplot(aes(Age, Subs_tree, colour = PatientID)) +
  geom_jitter(width = 0.2) +
  theme(plot.title = element_text(size = 8)) +
  ggtitle("Age-associated accumulation of somatic mutations in normal endometrium (substitu
tions only)") +
  theme(plot.title = element_text(size = 14)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5))
```


ed accumulation of somatic mutations in normal endometrium (substitutions only)

Fit linear mixed effects models and estimate mutation rate per year

To account for the non-independent sampling per patient we use a linear mixed-effects model as the observed frequencies of all substitutions approximates a normal distribution. We also use a random slope with fixed intercept as most women will start menarche at a similar age (~13 years), but to account for the potential differences in the rates at which mutations were acquired in different individuals due to variation in parity, contraception and other factors.

We test features with a known affect on mutation burden or endometrial cancer risks:

- Age
- Read depth & VAF ('Vafdepth')
- Driver mutations
- BMI
- Parity
- Cohort

We use backwards elimination to define the final model

Make the full model and drop each fixed effect in turn

```
# Combine read depth and median sample depth as Vafdepth
  endom_burden %<>%  mutate(Vafdepth = Seq_X*SampleMedianVAF)

# Make BMI and Parity numeric
  endom_burden %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden %<>%  mutate(Parity.QC = as.numeric(Parity))

# Exclude cases without Parity data
  endom_burden.qc <- endom_burden %>% filter(!is.na(Parity.QC))

# Build the full model

  full_lmer_model = lmer(Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
 Cohort + (Age - 1|PatientID),  data=endom_burden, REML=F)

  print(full_lmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
##     Cohort + (Age - 1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3566.797  3605.836 -1772.398  3544.797       246
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.651
##  Residual       219.661
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##           (Intercept)                     Age                Vafdepth
##              -280.880                  29.666                  27.855
##         Driver_status                  BMI.QC               Parity.QC
##               110.348                   7.572                 -16.138
##     CohortPost-mortem               CohortTAH  CohortTransplant donor
##                30.250                 -56.199                 -97.972
```

```
# Drop each fixed effect
  lme4:::drop1.merMod(full_lmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
##     Cohort + (Age - 1 | PatientID)
##               Df    AIC    LRT   Pr(Chi)
## <none>           3566.8
## Age            1 3611.0 46.170 1.084e-11 ***
## Vafdepth       1 3590.9 26.116 3.215e-07 ***
## Driver_status  1 3575.2 10.362  0.001286 **
## BMI.QC         1 3565.2  0.436  0.509086
## Parity.QC      1 3565.1  0.299  0.584717
## Cohort         3 3562.8  1.979  0.576675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove feature with largest P > 0.05 to make reduced model 1

```
# Remove Parity from full model
  reduced1_glmer_model <- update(full_lmer_model, ~ . -Parity.QC )
  anova(full_lmer_model,reduced1_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced1_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
## reduced1_glmer_model:     (Age - 1 | PatientID)
## full_lmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
## full_lmer_model:     Cohort + (Age - 1 | PatientID)
##                      Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced1_glmer_model 10 3565.1 3600.6 -1772.5   3545.1
## full_lmer_model      11 3566.8 3605.8 -1772.4   3544.8 0.2987      1
##                      Pr(>Chisq)
## reduced1_glmer_model
## full_lmer_model          0.5847
```

```
  print(reduced1_glmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
##     (Age - 1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3565.095  3600.586 -1772.548  3545.095       247
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.654
##  Residual       219.783
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##           (Intercept)                    Age               Vafdepth
##              -327.209                 29.847                 28.011
##         Driver_status                 BMI.QC       CohortPost-mortem
##               111.647                  9.277                -64.864
##             CohortTAH  CohortTransplant donor
##               -77.080               -115.590
```

```
  lme4:::drop1.merMod(reduced1_glmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
##     (Age - 1 | PatientID)
##                Df    AIC    LRT   Pr(Chi)
## <none>            3565.1
## Age            1 3610.2 47.111 6.707e-12 ***
## Vafdepth       1 3589.5 26.442 2.716e-07 ***
## Driver_status  1 3573.7 10.629  0.001113 **
## BMI.QC         1 3563.8  0.705  0.401140
## Cohort         3 3561.5  2.387  0.496036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove next feature with largest P > 0.05 to make reduced model 2

```
# Remove Cohort from reduced model 1
reduced2_glmer_model <- update(reduced1_glmer_model, ~ . -Cohort)
anova(reduced1_glmer_model,reduced2_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced2_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
## reduced2_glmer_model:     1 | PatientID)
## reduced1_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
## reduced1_glmer_model:     (Age - 1 | PatientID)
##                      Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced2_glmer_model  7 3561.5 3586.3 -1773.7   3547.5
## reduced1_glmer_model 10 3565.1 3600.6 -1772.5   3545.1 2.3871      3
##                      Pr(>Chisq)
## reduced2_glmer_model
## reduced1_glmer_model      0.496
```

```
print(reduced2_glmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
##     1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3561.482  3586.326 -1773.741  3547.482       250
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.771
##  Residual       220.280
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##   (Intercept)            Age       Vafdepth  Driver_status         BMI.QC
##      -323.464         28.952         28.681        110.772          6.553
```

```
lme4:::drop1.merMod(reduced2_glmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
##     1 | PatientID)
##               Df    AIC    LRT   Pr(Chi)
## <none>           3561.5
## Age            1 3605.6 46.093 1.128e-11 ***
## Vafdepth       1 3587.3 27.855 1.308e-07 ***
## Driver_status  1 3569.9 10.413  0.001251 **
## BMI.QC         1 3560.1  0.593  0.441211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove next feature with largest P > 0.05 to make reduced model 3

```
# Remove BMI information from reduced model 2
reduced3_glmer_model <- update(reduced2_glmer_model, ~ . -BMI.QC)
anova(reduced2_glmer_model,reduced3_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced3_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
## reduced2_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
## reduced2_glmer_model:     1 | PatientID)
##                      Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced3_glmer_model  6 3560.1 3581.4 -1774.0   3548.1
## reduced2_glmer_model  7 3561.5 3586.3 -1773.7   3547.5 0.5931      1
##                      Pr(>Chisq)
## reduced3_glmer_model
## reduced2_glmer_model     0.4412
```

Define the final model

```
# Define final model keeping all features that are significant with P < 0.05
  final_glmer_model <- reduced3_glmer_model

# Print the final model summary
  print(summary(final_glmer_model))
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
##    Data: endom_burden
##
##      AIC      BIC   logLik deviance df.resid
##   3560.1   3581.4  -1774.0   3548.1      251
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0371 -0.4099  0.0067  0.4361  3.9936
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     14.78   3.845
##  Residual       48474.42 220.169
## Number of obs: 257, groups:  PatientID, 28
##
## Fixed effects:
##               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   -267.398    120.757   57.039  -2.214  0.03082 *
## Age             28.620      2.732   28.290  10.477 3.02e-11 ***
## Vafdepth        29.028      5.266  255.958   5.513 8.61e-08 ***
## Driver_status  109.881     33.881  249.039   3.243  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.829
## Vafdepth    -0.543  0.081
## Driver_stts  0.131 -0.220 -0.161
```

```
# Estimate confidence intervals using "likelihood profile" method
  # confint.merMod(final_glmer_model, method = "profile")
  confint.merMod(final_glmer_model, method = "Wald")
```

```
##                    2.5 %    97.5 %
## .sig01                NA        NA
## .sigma                NA        NA
## (Intercept)   -504.07833 -30.71845
## Age             23.26647  33.97419
## Vafdepth        18.70793  39.34852
## Driver_status   43.47519 176.28725
```

```
# Calculate mutation rates for each donor from this model
# # randomEffects.df <- as.data.frame(ranef(final_glmer_model))
# write_csv(randomEffects.df, "model_rates.csv")
```

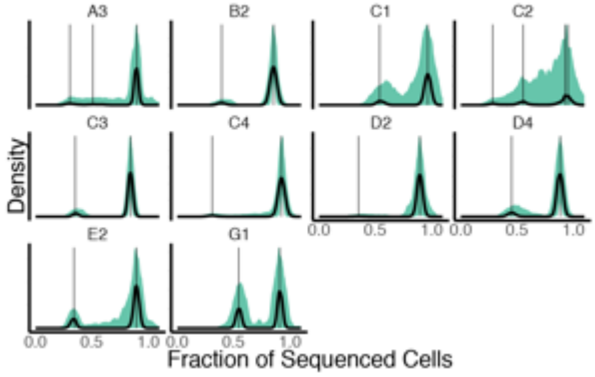
Appendix 10

Patient ID	Reason for sampling	Age	BMI	Parity	Number of high coverage samples	Menopause status	Menstrual phase
PD37506	Post-mortem (traumatic injury)	19	U	U	10	Pre-menopausal	Undetermined
PD40535	Transplant donor	24	24	3	7	Pre-menopausal	Proliferative
PD41871	Infertility clinic	27	30	0	17	Pre-menopausal	Secretory
PD37605	Infertility clinic	29	27	2	9	Pre-menopausal	Secretory
PD37601	Infertility clinic	31	28	0	10	Pre-menopausal	Secretory
PD41860	Infertility clinic	31	23	0	4	Pre-menopausal	Secretory
PD37607	Infertility clinic	34	24	1	19	Pre-menopausal	Secretory
PD41857	Infertility clinic	34	22	1	14	Pre-menopausal	Secretory
PD39444	Transplant donor	35	24	1	10	Pre-menopausal	Proliferative
PD41865	Infertility clinic	36	31	0	2	Pre-menopausal	Secretory
PD41868	Infertility clinic	36	23	0	6	Pre-menopausal	Secretory
PD39953	Transplant donor	37	18	2	8	Pre-menopausal	Secretory
PD41859	Infertility clinic	38	21	0	1	Pre-menopausal	Secretory
PD37613	Infertility clinic	39	22	0	11	Pre-menopausal	Secretory
PD41861	Infertility clinic	39	21	0	8	Pre-menopausal	Secretory
PD41869	Infertility clinic	40	37	0	13	Pre-menopausal	Secretory
PD37594	Infertility clinic	42	20	1	17	Pre-menopausal	Secretory
PD39952	Transplant donor	44	36	0	11	Pre-menopausal	Proliferative
PD39954	Transplant donor	44	24	1	10	Pre-menopausal	Secretory
PD37595	Infertility clinic	46	19.5	5	9	Pre-menopausal	Secretory
PD36804	Hysterectomy for leiomyomata	47	30	3	13	Pre-menopausal	Secretory
PD36805	Hysterectomy for benign ovarian tumour	49	27	0	7	Pre-menopausal	Secretory
PD38812	Post-mortem (traumatic injury)	54	U	U	2	Post-menopausal	Proliferative
PD37507	Post-mortem (peritonitis)	60	U	U	14	Post-menopausal	Inactive
PD42746	Transplant donor	67	34	2	2	Post-menopausal	Inactive
PD40107	Transplant donor	69	24	2	10	Post-menopausal	Inactive
PD42475	Transplant donor	74	27	2	8	Post-menopausal	Inactive
PD40659	Post-mortem	81	22	4	5	Post-menopausal	Inactive

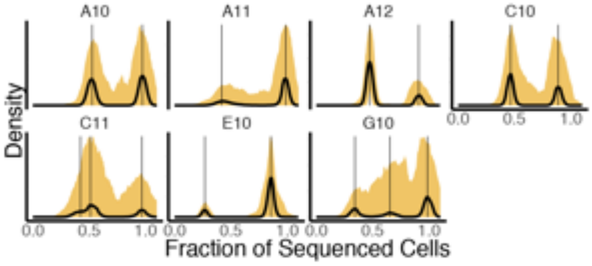
U = unknown

Appendix 11

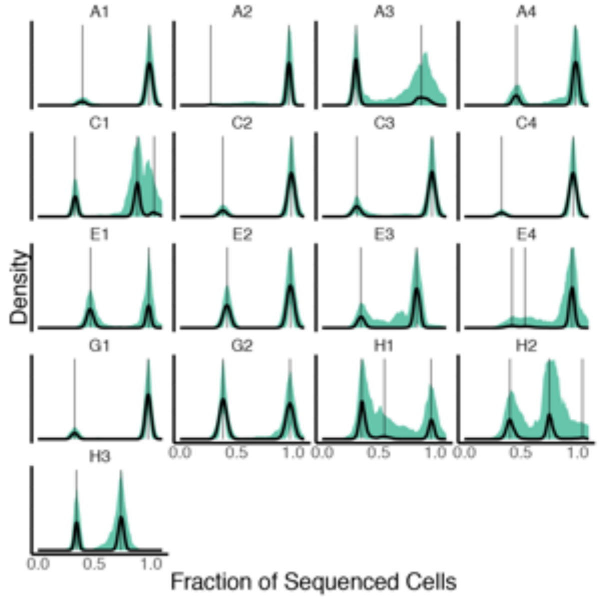
PD37506



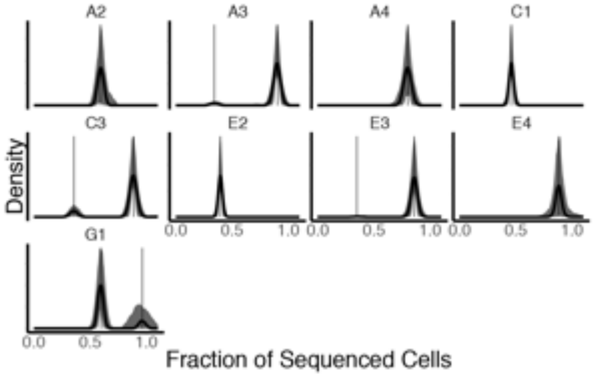
PD40535



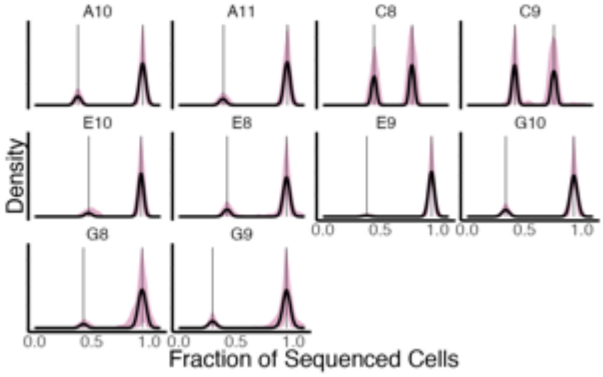
PD41871



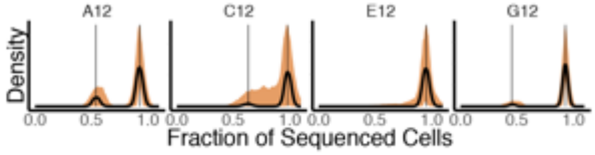
PD37605

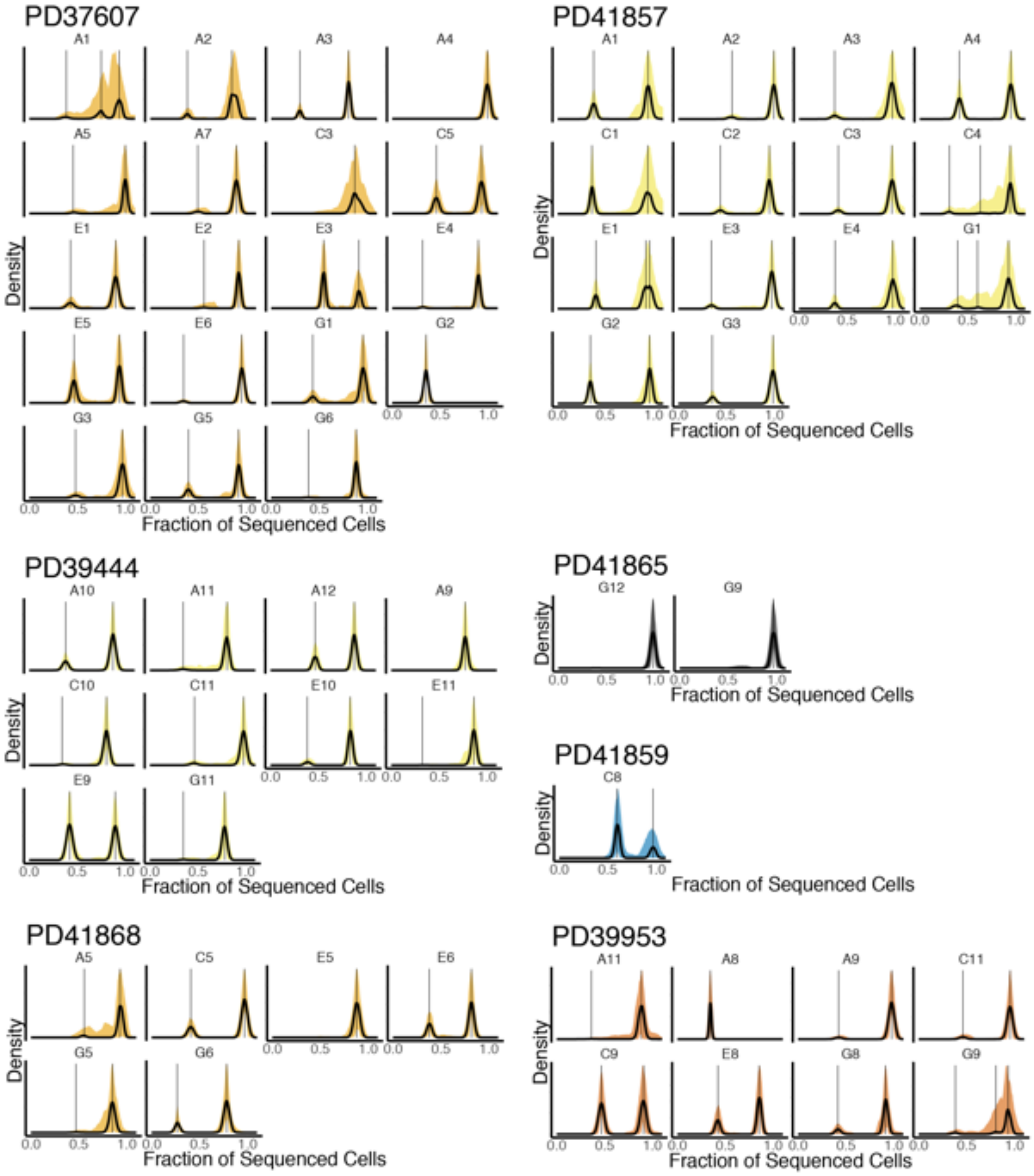


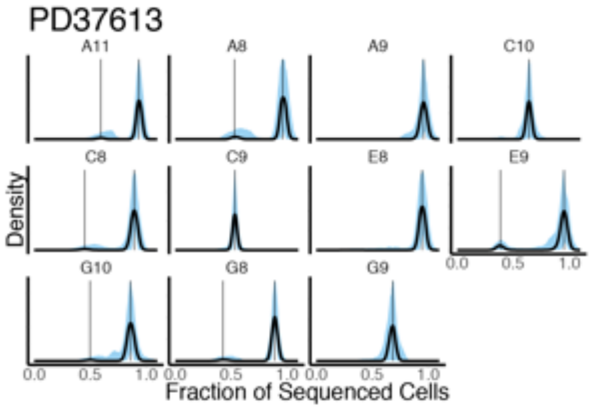
PD37601

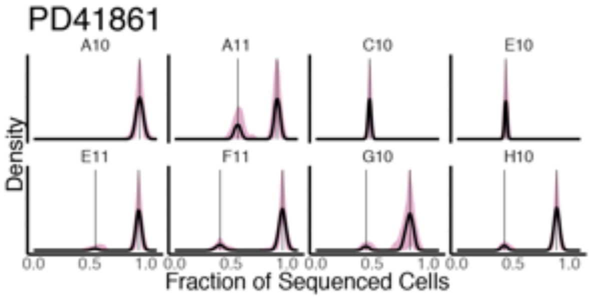


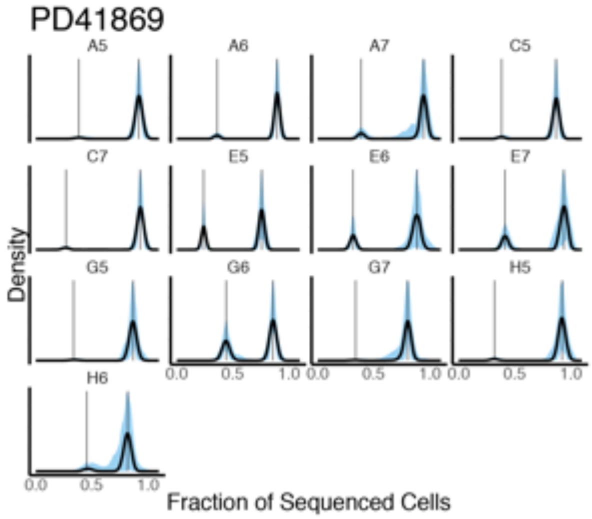
PD41860

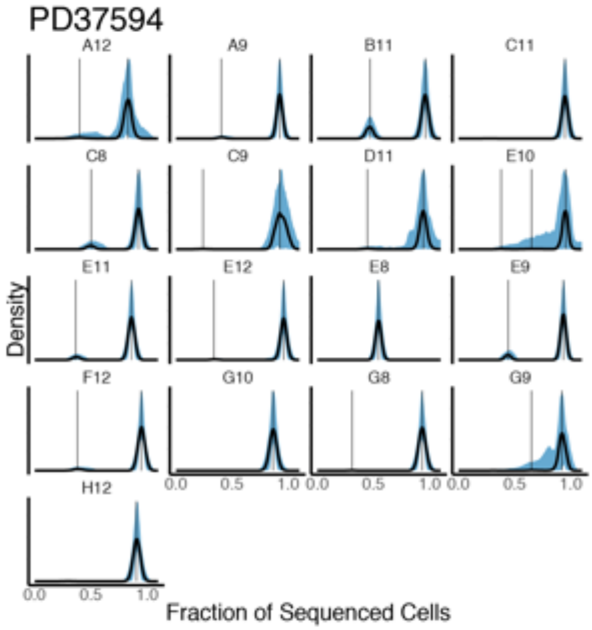


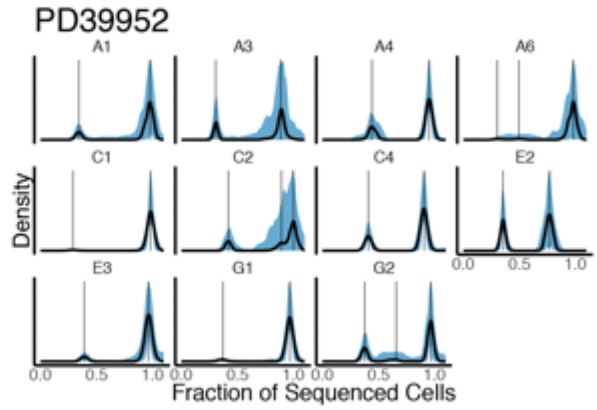

PD37607
A1
A2
A3
A4
A5
A7
C3
C5
E1
E2
E3
E4
E5
E6
G1
G2
G3
G5
G6
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41857
A1
A2
A3
A4
C1
C2
C3
C4
E1
E3
E4
G1
G2
G3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD39444
A10
A11
A12
A9
C10
C11
E10
E11
E9
G11
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41865
G12
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41859
C8
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41868
A5
C5
E5
E6
G5
G6
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD39953
A11
A8
A9
C11
C9
E8
G8
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

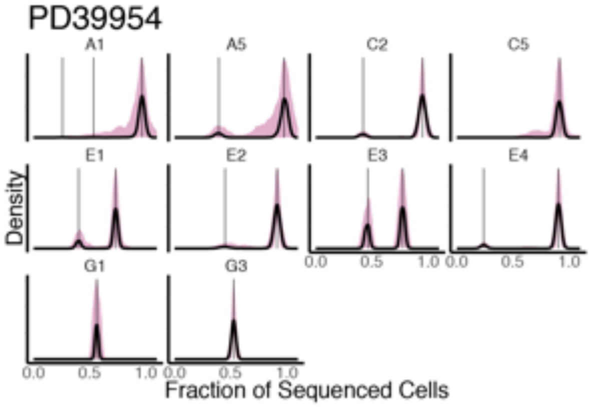

PD37613
A11
A8
A9
C10
C8
C9
E8
E9
G10
G8
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

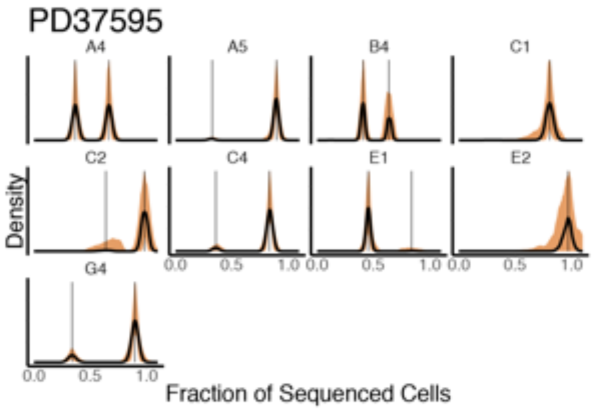

PD41861
A10
A11
C10
E10
E11
F11
G10
H10
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

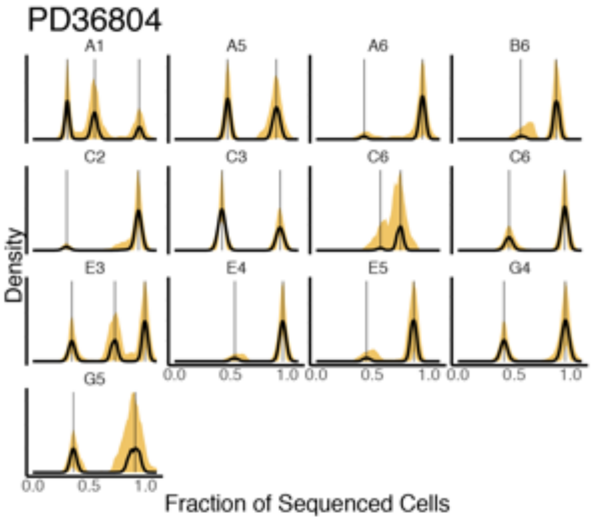

PD41869
A5
A6
A7
C5
C7
E5
E6
E7
G5
G6
G7
H5
H6
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

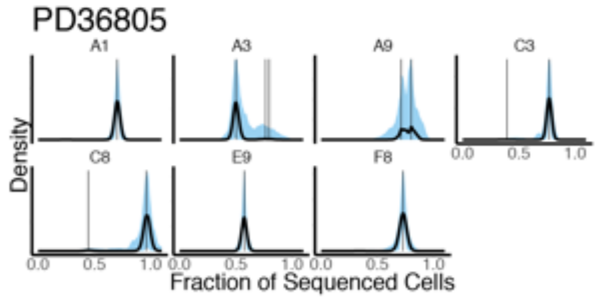

PD37594
A12
A9
B11
C11
C8
C9
D11
E10
E11
E12
E8
E9
F12
G10
G8
G9
H12
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

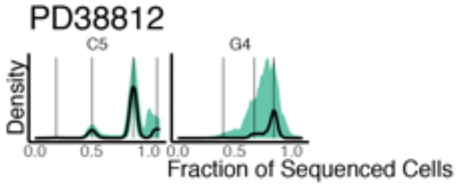

PD39952
A1
A3
A4
A6
C1
C2
C4
E2
E3
G1
G2
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

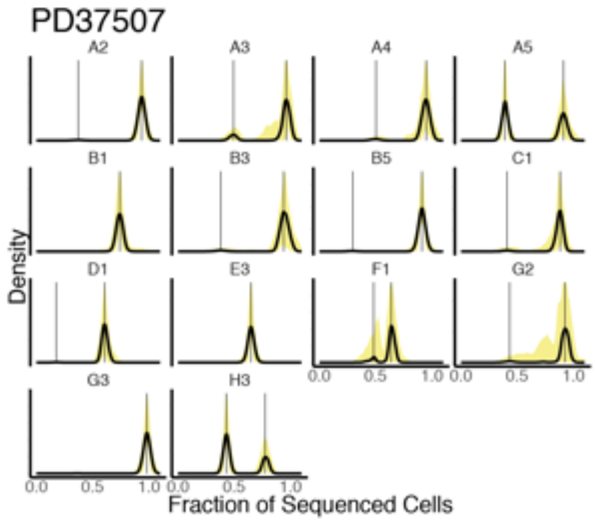

PD39954
A1
A5
C2
C5
E1
E2
E3
E4
G1
G3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

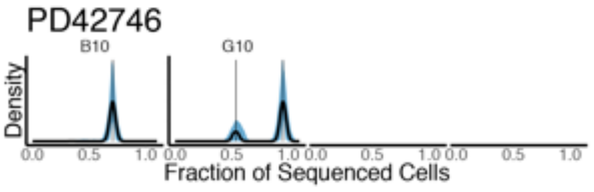

PD37595
A4
A5
B4
C1
C2
C4
E1
E2
G4
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

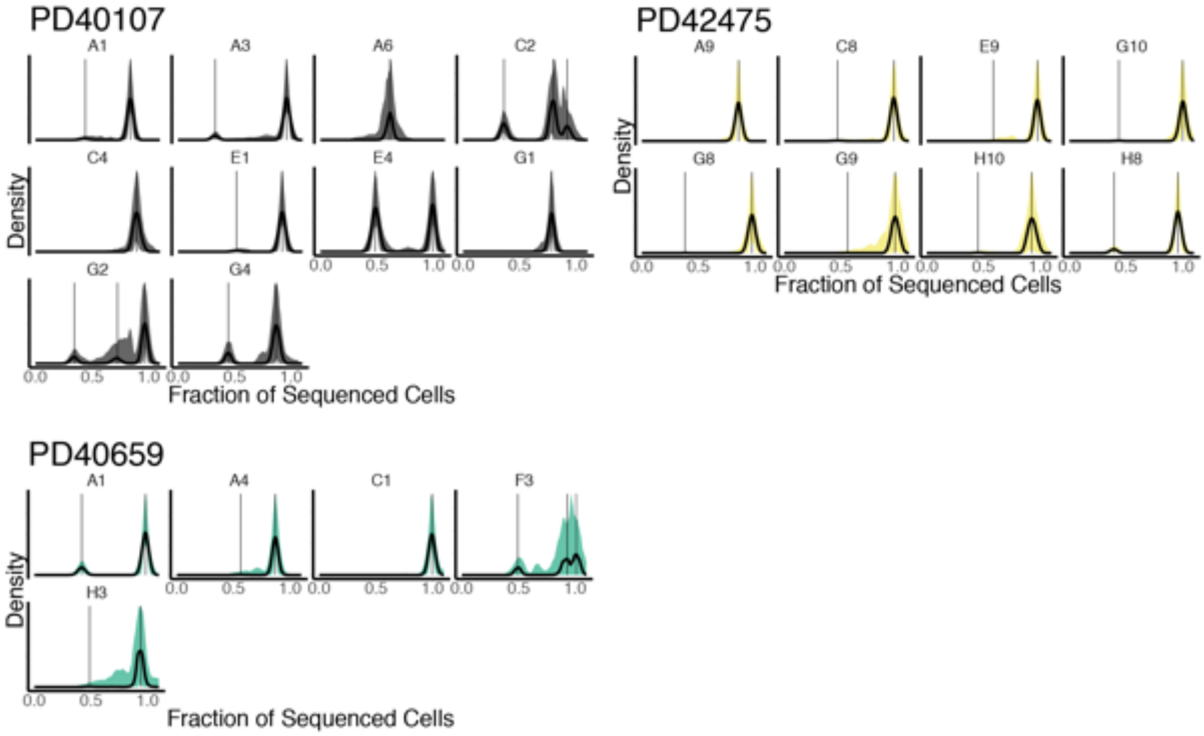

PD36804
A1
A5
A6
B6
C2
C3
C6
C6
E3
E4
E5
G4
G5
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD36805
A1
A3
A9
C3
C8
E9
F8
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD38812
C5
G4
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD37507
A2
A3
A4
A5
B1
B3
B5
C1
D1
E3
F1
G2
G3
H3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD42746
B10
G10
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD40107
A1
A3
A6
C2
C4
E1
E4
G1
G2
G4
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD42475
A9
C8
E9
G10
G8
G9
H10
H8
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD40659
A1
A4
C1
F3
H3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

Appendix 12

R Notebook

Luiza Moore

26062019

Modelling the effect of menstrual phase on total mutation burden and clonality

Markdown file to document methods used in the analysis of the menstrual phase and its effect on the total mutation burden and clonality

Load Libraries

```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(pbkrtest)
```

Load in data

Load in sample level data for all 28 donors, but exclude post-menopausal women and women with undetermined menstrual phase.

```
  endom_burden <- read.csv("Endometrium_for_model_26062019.csv", stringsAsFactors = F, na.s
trings = c("", "NA", "Unknown", "Uncertain"))
  dim(endom_burden)
```

```
## [1] 257  25
```

```
# Make BMI and Parity numeric
  endom_burden %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden %<>%  mutate(Parity.QC = as.numeric(Parity))

# Exclude post-menopausal women
  endom_burden.qc <- endom_burden %>% filter(Menopause_status_num == 0)
  dim(endom_burden.qc)
```

```
## [1] 218  27
```

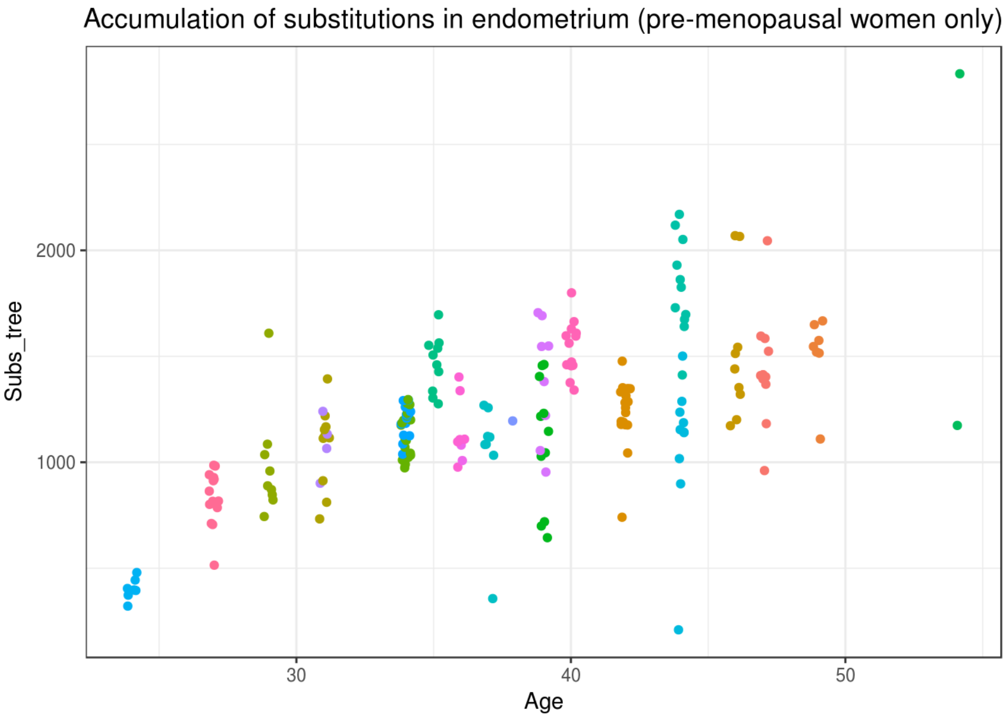
```
# Exclude cases with undetermined menstrual phase
  endom_burden.qc <- endom_burden.qc %>% filter(Menstrual_phase_num >0)
  dim(endom_burden.qc)
```

```
## [1] 208  27
```

```
 # Samples per patient
endom_burden.qc %>% group_by(PatientID) %>%  count(PatientID) %>%  rename(`Sample count` =
 n) %>% arrange(desc(`Sample count`)) %>%  kable() %>%  kable_styling(bootstrap_options = c
("striped", "condensed"), full_width = F, position = "left")
```

PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11
PD37601	10
PD39444	10
PD39954	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD36805	7
PD40535	7
PD41868	6
PD41860	4
PD38812	2
PD41865	2
PD41859	1

```
# Plot data
endom_burden.qc %>% ggplot(aes(Age, Subs_tree, colour = PatientID)) +
  geom_jitter(width = 0.2) +
  theme(plot.title = element_text(size = 3)) +
  ggtitle("Accumulation of substitutions in endometrium (pre-menopausal women only)") +
  theme(plot.title = element_text(size = 3)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5)) +
  theme(legend.position="none")
```


Accumulation of substitutions in endometrium (pre-menopausal women only)

Does menstrual phase have an effect on the total mutation burden?

To test the effect of menstrual phase on the total mutation burden we apply the final mixed-effect model with features that have been shown to be significant in the full cohort of patients.

These significant features are:

- Age
- Read depth & VAF ('Vafdepth')
- Driver mutations

```
# Combine read depth and median sample depth as 'Vafdepth'
  endom_burden.qc %<>%  mutate(Vafdepth = Seq_X*SampleMedianVAF)

# Total mutation burden
  full_lmer_model1 = lmer(Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num
 + (Age - 1|PatientID),  data=endom_burden.qc, REML=F)
  summary(full_lmer_model1)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   2853.6   2873.6  -1420.8   2841.6      202
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5372 -0.4404  0.0263  0.4820  4.0069
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     14.5    3.807
##  Residual       42357.8  205.810
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   -474.721    184.103   30.774  -2.579   0.0149 *
## Age             36.876      4.798   23.455   7.685 7.43e-08 ***
## Vafdepth        21.747      5.419  207.876   4.013 8.36e-05 ***
## Driver_status  132.336     32.969  201.308   4.014 8.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.925
## Vafdepth    -0.338  0.018
## Driver_stts  0.083 -0.113 -0.190
```

```
anova(full_lmer_model1,reduced_lmer_model1)
```

```
## Data: endom_burden.qc
## Models:
## reduced_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
## full_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
## full_lmer_model1:     (Age - 1 | PatientID)
##                     Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced_lmer_model1  6 2853.6 2873.6 -1420.8   2841.6
## full_lmer_model1     7 2854.9 2878.2 -1420.4   2840.9 0.7026      1
##                     Pr(>Chisq)
## reduced_lmer_model1
## full_lmer_model1        0.4019
```

Does menstrual phase have an effect on clonality?

To test the effect of menstrual phase on clonality, we used a linear mixed-effect model with SampleMedianVAF as a proxy for clonality

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   2853.6   2873.6  -1420.8   2841.6      202
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5372 -0.4404  0.0263  0.4820  4.0069
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     14.5    3.807
##  Residual       42357.8  205.810
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   -474.721    184.103   30.774  -2.579   0.0149 *
## Age             36.876      4.798   23.455   7.685 7.43e-08 ***
## Vafdepth        21.747      5.419  207.876   4.013 8.36e-05 ***
## Driver_status  132.336     32.969  201.308   4.014 8.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.925
## Vafdepth    -0.338  0.018
## Driver_stts  0.083 -0.113 -0.190
```

```
anova(full_lmer_model1,reduced_lmer_model1)
```

```
## Data: endom_burden.qc
## Models:
## reduced_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
## full_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
## full_lmer_model1:     (Age - 1 | PatientID)
##                     Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced_lmer_model1  6 2853.6 2873.6 -1420.8   2841.6
## full_lmer_model1     7 2854.9 2878.2 -1420.4   2840.9 0.7026      1
##                     Pr(>Chisq)
## reduced_lmer_model1
## full_lmer_model1        0.4019
```

Does menstrual phase have an effect on clonality?

To test the effect of menstrual phase on clonality, we used a linear mixed-effect model with SampleMedianVAF as a proxy for clonality

```
  full_lmer_model2 = lmer(SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phas
e_num + (Age - 1|PatientID),  data=endom_burden.qc, REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0184371 (tol =
## 0.002, component 1)
```

```
  summary(full_lmer_model2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
##     (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   -584.8   -561.5    299.4   -598.8      201
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.97712 -0.48971  0.05725  0.56190  2.74962
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  2.486e-07 0.0004986
##  Residual       3.055e-03 0.0552702
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##                      Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)          2.236e-01  4.256e-02  3.449e+01   5.253 7.75e-06 ***
## Age                  5.753e-04  8.292e-04  1.954e+01   0.694    0.496
## Vafdepth             1.390e-02  1.365e-03  1.827e+02  10.185  < 2e-16 ***
## Driver_status       -4.209e-03  8.558e-03  2.072e+02  -0.492    0.623
## Menstrual_phase_num  2.068e-04  1.540e-02  2.217e+01   0.013    0.989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth Drvr_s
## Age         -0.659
## Vafdepth    -0.348  0.058
## Driver_stts  0.037 -0.162 -0.198
## Mnstrl_phs_ -0.586 -0.083 -0.071  0.077
## convergence code: 0
## Model failed to converge with max|grad| = 0.0184371 (tol = 0.002, component 1)
```

```
  reduced_lmer_model2 = lmer(SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1|Pa
tientID),  data=endom_burden.qc, REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0180755 (tol =
## 0.002, component 1)
```

```
summary(reduced_lmer_model2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula: SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1 |
##     PatientID)
##    Data: endom_burden.qc
## 
##      AIC      BIC   logLik deviance df.resid
##   -586.8   -566.8    299.4   -598.8      202
## 
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.97672 -0.49076  0.05746  0.56106  2.74980
## 
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  2.486e-07 0.0004986
##  Residual       3.055e-03 0.0552703
## Number of obs: 208, groups:  PatientID, 22
## 
## Fixed effects:
##                Estimate Std. Error         df t value Pr(>|t|)    
## (Intercept)    2.239e-01  3.448e-02  3.567e+01   6.495 1.59e-07 ***
## Age            5.762e-04  8.264e-04  1.987e+01   0.697    0.494    
## Vafdepth       1.390e-02  1.361e-03  1.836e+02  10.212  < 2e-16 ***
## Driver_status -4.218e-03  8.532e-03  2.063e+02  -0.494    0.622    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.876              
## Vafdepth    -0.483  0.053       
## Driver_stts  0.102 -0.157 -0.194
## convergence code: 0
## Model failed to converge with max|grad| = 0.0180755 (tol = 0.002, component 1)
```

```
anova(full_lmer_model2,reduced_lmer_model2)
```

```
## Data: endom_burden.qc
## Models:
## reduced_lmer_model2: SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1 |
## reduced_lmer_model2:     PatientID)
## full_lmer_model2: SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phase_num
+
## full_lmer_model2:     (Age - 1 | PatientID)
##                     Df     AIC     BIC logLik deviance Chisq Chi Df
## reduced_lmer_model2  6 -586.84 -566.82 299.42  -598.84             
## full_lmer_model2     7 -584.84 -561.48 299.42  -598.84 2e-04      1
##                     Pr(>Chisq)
## reduced_lmer_model2           
## full_lmer_model2        0.9893
```

Appendix 13

Structural variants

SampleID	Chr1	start1	end1	Chr2	start2	end2	strand1	strand2	svclass
PD40535b_EMD_20_A11	12	120886465	120886466	12	123019772	123019773	+	+	deletion
PD37601b_EMD_11_E9	18	22857100	22857101	18	22859098	22859099	-	-	tandem-duplication
PD37601b_EMD_11_G10	5	113338567	113338568	5	113488147	113488148	+	+	deletion
PD37607b_EMD_6_E2	16	78780536	78780537	16	78824915	78824916	+	+	deletion
PD39444b_EMD_14_E9	19	47148553	47148555	19	47241742	47241744	+	+	deletion
PD39444b_EMD_14_E9	19	47148554	47148556	20	2795831	2795833	-	-	translocation
PD39444b_EMD_14_E9	19	47241742	47241743	20	2795831	2795832	+	+	translocation
PD39953b_EMD_17_C9	1	207866091	207866094	1	208150175	208150178	-	-	tandem-duplication
PD41861b_EMD_F11	22	29443121	29443122	X	12249093	12249094	-	-	translocation
PD37594b_EMD_8_A9	3	153811859	153811860	3	153818239	153818240	-	-	tandem-duplication
PD37594b_EMD_8_F12	6	90123273	90123274	6	90124479	90124480	+	+	deletion
PD39952b_EMD_15_C2	10	76122556	76122557	10	76164984	76164985	+	+	deletion
PD39952b_EMD_15_E3	10	76122556	76122557	10	76164984	76164985	+	+	deletion
PD39954b_EMD_16_E3	X	110302620	110302621	X	110304074	110304075	+	+	deletion
PD39954b_EMD_16_C2	X	66635162	66635163	X	66738873	66738874	+	-	inversion
PD39954b_EMD_16_E2	X	66635165	66635166	X	66738873	66738874	+	-	inversion
PD39954b_EMD_16_E3	X	66635165	66635166	X	66738873	66738874	+	-	inversion
PD39954b_EMD_16_G3	X	66635164	66635165	X	66738873	66738874	+	-	inversion
PD37595b_EMD_9_C1	12	60041293	60041294	12	60046767	60046768	-	-	tandem-duplication
PD38812b_EMD_13_C5	14	69063692	69063693	14	69129713	69129714	+	+	deletion
PD38812b_EMD_13_C5	7	154208554	154208555	7	154221315	154221316	+	+	deletion
PD37507b_EMD_2_B5	14	87635387	87635388	14	87649060	87649061	-	-	tandem-duplication
PD37507b_EMD2_G7_A2	4	110760126	110760127	4	110761792	110761793	-	-	tandem-duplication
PD40107b_EMD_18_A1	20	22312601	22312602	20	23066262	23066263	+	+	deletion
PD40107b_EMD_18_A3	9	11230240	11230242	9	11233177	11233179	+	+	deletion
PD40107b_EMD_18_A3	9	11231379	11231380	9	11234244	11234245	-	-	tandem-duplication
PD42475b_EMD_A9	5	41940779	41940780	5	41943347	41943348	-	-	tandem-duplication
PD40659c_EMD_19_A1	1	15342186	15342187	12	49278653	49278654	+	-	translocation
PD40659c_EMD_19_C1	1	109642333	109642334	3	37034595	37034596	+	+	translocation
PD40659c_EMD_19_C1	1	109642338	109642339	3	37034579	37034580	-	-	translocation
PD40659c_EMD_19_C1	3	41330863	41330865	3	56348950	56348952	+	+	deletion
PD40659c_EMD_19_C1	4	24731077	24731078	5	133211928	133211929	-	+	translocation
PD40659c_EMD_19_C1	4	24731078	24731079	5	133211919	133211920	+	-	translocation
PD40659c_EMD_19_C1	4	54948131	54948132	4	61183815	61183816	+	-	inversion
PD40659c_EMD_19_C1	4	54948132	54948133	4	61183813	61183814	-	+	inversion
PD40659c_EMD_19_C1	4	135500411	135500412	5	63470404	63470405	-	-	translocation
PD40659c_EMD_19_C1	7	73366851	73366852	7	73665165	73665166	+	+	deletion
PD40659c_EMD_19_F3	6	111447964	111447965	7	77557167	77557168	-	-	translocation

Copy number variants

Age	SampleID	Chrom	Start	End	Total copy number	Minor allele copy number
49	PD36805b_EM7_G2_C8	3	151924874	197908615	2	0
60	PD37507b_EMD2_G13_A3	16	67451927	90292766	2	0
60	PD37507b_EMD2_G20_H3	16	67347740	90292766	2	0
44	PD39952b_EMD_15_A1	11	87268	38612664	2	0
44	PD39952b_EMD_15_A3	20	61098	29650825	1	0
44	PD39952b_EMD_15_C1	11	87268	38511931	2	0
69	PD40107b_EMD_18_G2	17	49346457	81185372	2	0
69	PD40107b_EMD_18_G4	13	62420270	115108598	4	1
81	PD40659c_EMD_19_C1	3	41336053	56347925	1	0
31	PD41860b_EMD_G12	11	85897571	114112013	1	0
31	PD41860b_EMD_G12	13	19020095	115108598	1	0
39	PD41861b_EMD_E10	20	61098	21998953	2	0
39	PD41861b_EMD_G10	20	61098	26054883	2	0

Appendix 14

		Normal endometrium						Endometrial cancer (TCGA)			
Whole exome, q<0.01		Whole exome, q<0.001		RHT, q<0.05		Whole exome, q<0.01		Whole exome, q<0.001		RHT, q<0.05	
gene_name	qglobal_cv	gene_name	qglobal_cv	gene_name	qglobal_RHT	gene_name	qglobal_cv	gene_name	qglobal_cv	gene_name	qglobal_RHT
PIK3CA	0	PIK3CA	0	PIK3CA	0	PTEN	0	PTEN	0	PTEN	0
ARHGAP35	0	ARHGAP35	0	ARHGAP35	0	TP53	0	TP53	0	TP53	0
PIK3R1	3.64E-07	PIK3R1	3.64E-07	PIK3R1	6.69E-09	PIK3CA	0	PIK3CA	0	PIK3CA	0
FBXW7	3.90E-06	FBXW7	3.90E-06	FBXW7	7.17E-08	CTNNB1	0	CTNNB1	0	CTNNB1	0
FOXA2	0.0002395	FOXA2	0.00023946	FOXA2	4.40E-06	KRAS	0	KRAS	0	KRAS	0
KRAS	0.0013681			KRAS	2.51E-05	CTCF	0	CTCF	0	CTCF	0
PPP2R1A	0.005791			PPP2R1A	0.00010637	ARID1A	0	ARID1A	0	ARID1A	0
ZFHX3	0.0064149			ZFHX3	0.00011782	PIK3R1	0	PIK3R1	0	PIK3R1	0
CHD4	0.0091925			CHD4	0.00016884	FBXW7	4.46E-06	FBXW7	4.46E-06	FBXW7	8.19E-08
				ERBB2	0.00584202	ARHGAP35	6.29E-06	ARHGAP35	6.29E-06	ARHGAP35	1.16E-07
				SPOP	0.00657231	ARID5B	8.81E-06	ARID5B	8.81E-06	ARID5B	1.62E-07
				ERBB3	0.01518232	ZFHX3	9.43E-06	ZFHX3	9.43E-06	ZFHX3	1.73E-07
						SPOP	1.07E-05	SPOP	1.07E-05	SPOP	1.97E-07
						FOXA2	0.00011264	FOXA2	0.00011264	FOXA2	2.07E-06
						PPP2R1A	0.00012485	PPP2R1A	0.00012485	PPP2R1A	2.29E-06
						FGFR2	0.0001309	FGFR2	0.0001309	FGFR2	2.40E-06
						RNF43	0.00202553			RNF43	3.72E-05
						CHD4	0.00326925			CHD4	6.00E-05
						NFE2L2	0.00388559			NFE2L2	7.14E-05
										FAT1	0.0004304
										ARID1B	0.0007817
										SOX17	0.0012432
										JAK1	0.0016882
										KMT2B	0.0019627
										HIST1H2BD	0.0027622
										CCND1	0.003446
										ATM	0.0040176
										ING1	0.0092608
										CASP8	0.0109251
										RB1	0.0114702
										NRAS	0.0243062
										ZFP36L2	0.0357105
										CDKN1B	0.0432904
										SGK1	0.0433312
										CUX1	0.0482893

Appendix 15

Sample	Chrom	Pos	Ref	Alt	Gene	Protein	Type	Effect	VAF
PD36804b_EM5_G2_B6	3	178952064	T	A	PIK3CA	p.M1040K	Sub	missense	0.45
PD36804b_EM5_G3_C6	3	178921548	G	A	PIK3CA	p.V344M	Sub	missense	0.5
PD36804b_EMD_7_A1	12	6701590	TTAG	T	CHD4	p.L972delL	Del	inframe	0.21
PD36804b_EMD_7_E3	12	6701191	A	G	CHD4	p.F994S	Sub	missense	0.27
PD36804b_EMD_7_G4	20	22562813	AGCAGGTGGGCCGCG	A	FOXA2	p.A352fs*11	Del	frameshift	0.65
PD36805b_EM1_G1_L1_2_A1	3	178952138	C	T	PIK3CA	p.H1065Y	Sub	missense	0.38
PD36805b_EM10_G3_C3	4	153249384	C	T	FBXW7	p.R465H	Sub	missense	0.44
PD36805b_EM7_G2_C8	3	178936095	A	G	PIK3CA	p.Q546R	Sub	missense	1
PD36805b_EM8_G2_F8	20	22563552	GC	G	FOXA2	p.P110fs*3	Del	frameshift	0.24
PD36805b_EM8_G2_F8	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.5
PD36805b_EM9_G1_A9	19	47425308	C	T	ARHGAP35	p.Q1126*	Sub	nonsense	0.35
PD36805b_EM9_G1_A9	12	6697096	C	T	CHD4	p.R1162Q	Sub	missense	0.5
PD36805b_EM9_G4_E9	19	47425308	C	T	ARHGAP35	p.Q1126*	Sub	nonsense	0.32
PD36805b_EM9_G4_E9	12	6697096	C	T	CHD4	p.R1162Q	Sub	missense	0.24
PD37507b_EMD_2_A5	19	47424541	T	G	ARHGAP35	p.L870*	Sub	nonsense	0.44
PD37507b_EMD_2_A5	7	140453154	T	C	BRAF	p.D594G	Sub	missense	0.43
PD37507b_EMD_2_A5	11	534285	C	A	HRAS	p.G13V	Sub	missense	0.5
PD37507b_EMD_2_B5	19	47424541	T	G	ARHGAP35	p.L870*	Sub	nonsense	0.45
PD37507b_EMD_2_B5	7	140453154	T	C	BRAF	p.D594G	Sub	missense	0.71
PD37507b_EMD_2_B5	11	534285	C	A	HRAS	p.G13V	Sub	missense	0.5
PD37507b_EMD_2_B5	5	67591085	G	T	PIK3R1	p.D560Y	Sub	missense	0.29
PD37507b_EMD2_G12_G2	20	22563615	C	CG	FOXA2	p.G89fs*156	Ins	frameshift	0.36
PD37507b_EMD2_G12_G2	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.52
PD37507b_EMD2_G12_G2	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.6
PD37507b_EMD2_G13_A3	20	22562813	AGCAGGTGGGCCGCG	A	FOXA2	p.A352fs*11	Del	frameshift	0.39
PD37507b_EMD2_G13_A3	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.96
PD37507b_EMD2_G13_A3	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.4
PD37507b_EMD2_G14_B3	20	39802384	G	A	PLCG1	p.E1163K	Sub	missense	0.59
PD37507b_EMD2_G14_B3	3	178952085	A	G	PIK3CA	p.H1047R	Sub	missense	0.39
PD37507b_EMD2_G17_E3	3	178952085	A	G	PIK3CA	p.H1047R	Sub	missense	0.5
PD37507b_EMD2_G19_G3	16	72822452	CT	C	ZFHX3	p.K3241fs*43	Del	frameshift	0.48
PD37507b_EMD2_G19_G3	3	178952085	A	G	PIK3CA	p.H1047R	Sub	missense	0.3

PD37507b_EMD2_G2_B1	10	123279674	G	C	FGFR2	p.P253R	Sub	missense	0
PD37507b_EMD2_G2_B1	16	72822038	CTGCTGCTGCTGAATTGCCTCCTGCAGACTCTGCT	C	ZFHX3	p.Q3368fs*106	Del	frameshift	0.2
PD37507b_EMD2_G2_B1	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.38
PD37507b_EMD2_G2_B1	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.2
PD37507b_EMD2_G20_H3	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.67
PD37507b_EMD2_G20_H3	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.36
PD37507b_EMD2_G21_A4	3	178952085	A	G	PIK3CA	p.H1047R	Sub	missense	0.5
PD37507b_EMD2_G3_C1	10	123279674	G	C	FGFR2	p.P253R	Sub	missense	0.41
PD37507b_EMD2_G3_C1	16	72831849	G	A	ZFHX3	p.Q1578*	Sub	nonsense	0.26
PD37507b_EMD2_G3_C1	16	72822038	CTGCTGCTGCTGAATTGCCTCCTGCAGACTCTGCT	C	ZFHX3	p.Q3368fs*106	Del	frameshift	0.35
PD37507b_EMD2_G3_C1	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.43
PD37507b_EMD2_G3_C1	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.4
PD37507b_EMD2_G4_D1	19	47424541	T	G	ARHGAP35	p.L870*	Sub	nonsense	0.24
PD37507b_EMD2_G4_D1	11	534285	C	A	HRAS	p.G13V	Sub	missense	0.39
PD37507b_EMD2_G6_F1	4	153247168	T	C	FBXW7	p.Y545C	Sub	missense	0.44
PD37507b_EMD2_G6_F1	19	52715982	C	T	PPP2R1A	p.R183W	Sub	missense	0.37
PD37507b_EMD2_G7_A2	10	123279674	G	C	FGFR2	p.P253R	Sub	missense	0.46
PD37507b_EMD2_G7_A2	16	72822038	CTGCTGCTGCTGAATTGCCTCCTGCAGACTCTGCT	C	ZFHX3	p.Q3368fs*106	Del	frameshift	0.36
PD37507b_EMD2_G7_A2	16	72991902	G	A	ZFHX3	p.R715*	Sub	nonsense	0.52
PD37507b_EMD2_G7_A2	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.3
PD37594b_EMD_8_B11	4	153247224	C	T	FBXW7	p.W526*	Sub	nonsense	0.52
PD37594b_EMD_8_B11	3	178921552	A	C	PIK3CA	p.N345T	Sub	missense	0.55
PD37594b_EMD_8_C8	3	178952084	C	T	PIK3CA	p.H1047Y	Sub	missense	0.45
PD37594b_EMD_8_D11	3	178916924	C	T	PIK3CA	p.P104L	Sub	missense	0.44
PD37594b_EMD_8_E12	12	56480335	T	C	ERBB3	p.Y148H	Sub	missense	0.39
PD37594b_EMD_8_F12	12	6697480	G	C	CHD4	p.P1150R	Sub	missense	0.6
PD37594b_EMD_8_G10	19	52709239	G	T	PPP2R1A	p.V65F	Sub	missense	0.57
PD37594b_EMD_8_G10	5	67589650			PIK3R1	p.S474_Q475insHRTS	inframe		0.17
PD37595b_EMD_9_A4	4	153247354	A	T	FBXW7	p.L483H	Sub	missense	0.38
PD37595b_EMD_9_B4	4	153247354	A	T	FBXW7	p.L483H	Sub	missense	0.28
PD37595b_EMD_9_E1	12	12870972	CACAA	C	CDKN1B	p.K68fs*2	Del	frameshift	0.53
PD37601b_EMD_11_A10	16	3789700	C	A	CREBBP	p.E1387*	Sub	nonsense	0.49
PD37601b_EMD_11_A11	19	52716212	C	T	PPP2R1A	p.S219L	Sub	missense	0.67
PD37601b_EMD_11_E9	19	47440667	T	C	ARHGAP35	p.?	Sub	ess_splice	0.43
PD37601b_EMD_11_E9	12	6696973	AGCCCAGGCCGCACCACTAGATGCGTCAGCATCATTTTCTTCTTTGCCACCT	A	CHD4	p.Q1186_G1202delQVAKKKMMLTHLVVRPG	Del	inframe	0.41

PD37601b_EMD_11_G9	3	178936082	G	A	PIK3CA	p.E542K	Sub	missense	0.55
PD37605b_EMD_4_A3	4	153250925	G	A	FBXW7	p.H379Y	Sub	missense	0.44
PD37605b_EMD_4_E2	5	67588981	C	T	PIK3R1	p.R358*	Sub	nonsense	0.16
PD37607b_EMD_6_A1	5	67589591	T	TATA	PIK3R1	p.N453_T454insN	Ins	inframe	0.6
PD37607b_EMD_6_A2	3	178921553	T	A	PIK3CA	p.N345K	Sub	missense	0.43
PD37607b_EMD_6_A3	3	178928079	G	A	PIK3CA	p.E453K	Sub	missense	0.36
PD37607b_EMD_6_A4	3	178921553	T	A	PIK3CA	p.N345K	Sub	missense	0.33
PD37607b_EMD_6_A5	19	47422760	G	GT	ARHGAP35	p.Y277fs*2	Ins	frameshift	0.6286
PD37607b_EMD_6_A7	19	47424170	TA	T	ARHGAP35	p.N747fs*9	Del	frameshift	0.4082
PD37607b_EMD_6_A7	17	37881440	C	T	ERBB2	p.H878Y	Sub	missense	0.33
PD37607b_EMD_6_C3	3	178928079	G	A	PIK3CA	p.E453K	Sub	missense	0.47
PD37607b_EMD_6_C5	17	37884062	A	G	ERBB2	p.N1178S	Sub	missense	0.18
PD37607b_EMD_6_C5	17	29554541	GC	G	NF1	p.A776fs*15	Del	frameshift	0.4722
PD37607b_EMD_6_E1	5	67589591	T	TATA	PIK3R1	p.N453_T454insN	Ins	inframe	0.3871
PD37607b_EMD_6_E2	3	178928079	G	A	PIK3CA	p.E453K	Sub	missense	0.35
PD37607b_EMD_6_E3	3	178928079	G	A	PIK3CA	p.E453K	Sub	missense	0.5
PD37607b_EMD_6_E4	3	178921553	T	A	PIK3CA	p.N345K	Sub	missense	0.41
PD37607b_EMD_6_E5	17	37884062	A	G	ERBB2	p.N1178S	Sub	missense	0.28
PD37607b_EMD_6_E5	17	29554541	GC	G	NF1	p.A776fs*15	Del	frameshift	0.4872
PD37607b_EMD_6_E6	3	178917478	G	A	PIK3CA	p.G118D	Sub	missense	0.39
PD37607b_EMD_6_G1	4	153249384	C	T	FBXW7	p.R465H	Sub	missense	0.5
PD37607b_EMD_6_G2	4	153247288	C	A	FBXW7	p.R505L	Sub	missense	0.19
PD37607b_EMD_6_G2	3	178936092	A	G	PIK3CA	p.E545G	Sub	missense	0.15
PD37607b_EMD_6_G3	3	178921553	T	A	PIK3CA	p.N345K	Sub	missense	0.5
PD37607b_EMD_6_G5	3	178928083	A	G	PIK3CA	p.D454G	Sub	missense	0.36
PD37607b_EMD_6_G6	3	178917478	G	A	PIK3CA	p.G118D	Sub	missense	0.29
PD37607b_EMD_6_G6	20	22563239	C	A	FOXA2	p.R214L	Sub	missense	0.47
PD37613b_EMD_5_A11	10	63845619	AAG	A	ARID5B	p.E454fs*32	Del	frameshift	0.61
PD37613b_EMD_5_A9	10	63845619	AAG	A	ARID5B	p.E454fs*32	Del	frameshift	0.2
PD37613b_EMD_5_C10	12	49446997	TTCC	T	KMT2D	p.W315_K316delins*	Del	nonsense	0.4
PD37613b_EMD_5_C10	17	47699360	C	T	SPOP	p.E50K	Sub	missense	0.2
PD37613b_EMD_5_E9	10	63845619	AAG	A	ARID5B	p.E454fs*32	Del	frameshift	0.39
PD37613b_EMD_5_G10	10	63845619	AAG	A	ARID5B	p.E454fs*32	Del	frameshift	0.4
PD37613b_EMD_5_G9	5	67589556			PIK3R1	p.N441_I442del	Del	frameshift	0.1
PD38812b_EMD_13_C5	7	55249017	C	CCCA	EGFR	p.H773_V774insH	Ins	inframe	0.48

PD38812b_EMD_13_G4	17	37868208	C	T	ERBB2	p.S310F	Sub	missense	0.24
PD39444b_EMD_14_A10	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.38
PD39444b_EMD_14_A12	11	14380338	T	TGCCCACGCC	RRAS2	p.G26_K27insGVG	Ins	inframe	0.23
PD39444b_EMD_14_A9	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.41
PD39444b_EMD_14_C10	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.32
PD39444b_EMD_14_C11	3	178927980	T	C	PIK3CA	p.C420R	Sub	missense	0.5
PD39444b_EMD_14_E9	1	120458147	G	A	NOTCH2	p.R2400*	Sub	nonsense	0.22
PD39952b_EMD_15_A1	3	178916935	C	T	PIK3CA	p.R108C	Sub	missense	0.31
PD39952b_EMD_15_A1	11	14380350	C	A	RRAS2	p.G23C	Sub	missense	0.97
PD39952b_EMD_15_A3	20				FOXA2		20p loss		
PD39952b_EMD_15_A4	20	22562806	CGGGCCCAGCAGGTG	C	FOXA2	p.H354fs*9	Del	frameshift	0.45
PD39952b_EMD_15_A6	3	178922324	G	A	PIK3CA	p.E365K	Sub	missense	0.38
PD39952b_EMD_15_C1	§	178916935	C	T	PIK3CA	p.R108C	Sub	missense	0.47
PD39952b_EMD_15_C1	X	123220599	G	T	STAG2	p.E1086*	Sub	nonsense	0.5
PD39952b_EMD_15_C1	11	14380350	C	A	RRAS2	p.G23C	Sub	missense	0.97
PD39952b_EMD_15_C2	12	56488249	C	A	ERBB3	p.P590T	Sub	missense	0.62
PD39952b_EMD_15_C2	4	187554903	C	A	FAT1	p.E1420*	Sub	nonsense	0.54
PD39952b_EMD_15_C2	3	178952018	A	G	PIK3CA	p.T1025A	Sub	missense	0.54
PD39952b_EMD_15_C4	19	47425365	C	T	ARHGAP35	p.R1145*	Sub	nonsense	0.41
PD39952b_EMD_15_C4	20	22563487	GGCCAGGCC	G	FOXA2	p.G129fs*113	Del	frameshift	0.5
PD39952b_EMD_15_C4	3	178941917	G	A	PIK3CA	p.D746N	Sub	missense	0.36
PD39952b_EMD_15_E2	16	72992553	C	A	ZFHX3	p.E498*	Sub	nonsense	0.17
PD39952b_EMD_15_E3	12	56488249	C	A	ERBB3	p.P590T	Sub	missense	0.42
PD39952b_EMD_15_E3	4	187554903	C	A	FAT1	p.E1420*	Sub	nonsense	0.38
PD39952b_EMD_15_E3	3	178952018	A	G	PIK3CA	p.T1025A	Sub	missense	0.4
PD39952b_EMD_15_G1	19	47423363	T	G	ARHGAP35	p.Y477*	Sub	nonsense	0.42
PD39952b_EMD_15_G1	4	187549436	C	T	FAT1	p.W1561*	Sub	nonsense	0.26
PD39952b_EMD_15_G1	3	178938934	G	A	PIK3CA	p.E726K	Sub	missense	0.53
PD39952b_EMD_15_G1	11	14380346	C	T	RRAS2	p.G24D	Sub	missense	0.41
PD39952b_EMD_15_G2	12	56478786	G	A	ERBB3	p.R81Q	Sub	missense	0.52
PD39953b_EMD_17_A9	5	67591147	CTTGATGT	C	PIK3R1	p.?	Del	ess_splice	0.64
PD39953b_EMD_17_C11	5	67591147	CTTGATGT	C	PIK3R1	p.?	Del	ess_splice	0.62
PD39953b_EMD_17_E8	5	67591147	CTTGATGT	C	PIK3R1	p.?	Del	ess_splice	0.56
PD39953b_EMD_17_G8	5	67591147	CTTGATGT	C	PIK3R1	p.?	Del	ess_splice	0.6
PD39953b_EMD_17_G9	4	153249385	G	A	FBXW7	p.R465C	Sub	missense	0.67

PD39954b_EMD_16_C2	19	47423721	A	T	ARHGAP35	p.R597*	Sub	nonsense	0.61
PD39954b_EMD_16_C5	3	178916946	G	T	PIK3CA	p.K111N	Sub	missense	0.5
PD39954b_EMD_16_E1	17	37868208	C	T	ERBB2	p.S310F	Sub	missense	0.42
PD39954b_EMD_16_E2	19	47423460	GC	G	ARHGAP35	p.A510fs*36	Del	frameshift	0.65
PD40107b_EMD_18_A3	20				FOXA2		large deletion		
PD40107b_EMD_18_A3	3	178928079	G	C	PIK3CA	p.E453Q	Sub	missense	0.54
PD40107b_EMD_18_A3	5	67589143	T	TA	PIK3R1	p.N378fs*17	Ins	frameshift	0.71
PD40107b_EMD_18_A6	12	6697058	C	T	CHD4	p.V1175M	Sub	missense	0.24
PD40107b_EMD_18_A6	5	67588964	A	AC	PIK3R1	p.G353fs*11	Ins	frameshift	0.34
PD40107b_EMD_18_C4	3	178928226	C	T	PIK3CA	p.P471L	Sub	missense	0.5
PD40107b_EMD_18_C4	10	89692794	A	G	PTEN	p.H93R	Sub	missense	0.67
PD40107b_EMD_18_C4	16	72827491	GC	G	ZFHX3	p.G3030fs*46	Del	frameshift	0.62
PD40107b_EMD_18_E1	19	47492885	AC	A	ARHGAP35	p.L1332fs*30	Del	frameshift	0.27
PD40107b_EMD_18_E1	17	7578406	C	T	TP53	p.R175H	Sub	missense	0.52
PD40107b_EMD_18_G1	20	22563483	A	AGG	FOXA2	p.Y133fs*5	Ins	frameshift	0.38
PD40107b_EMD_18_G1	5	67591099	C	A	PIK3R1	p.N564K	Sub	missense	0.33
PD40107b_EMD_18_G2	17	37879903	C	T	ERBB2	p.T733I	Sub	missense	0.62
PD40107b_EMD_18_G2	17	47696461	C	T	SPOP	p.R121Q	Sub	missense	0.54
PD40107b_EMD_18_G4	3	178922364	G	T	PIK3CA	p.C378F	Sub	missense	0.3
PD40535b_EMD_20_E10	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.32
PD40659c_EMD_19_A1	12	56481922	G	A	ERBB3	p.G284R	Sub	missense	0.57
PD40659c_EMD_19_A4	5	67589590			PIK3R1	p.Y452_N453delinsH	Del	inframe	0.37
PD40659c_EMD_19_C1	17	7578457	C	T	TP53	p.R158H	Sub	missense	0.5
PD40659c_EMD_19_F3	17	37879794	G	A	ERBB2	p.V697M	Sub	missense	0.39
PD40659c_EMD_19_F3	3	178952077	T	A	PIK3CA	p.N1044K	Sub	missense	0.59
PD40659c_EMD_19_H3	12	56482341	G	T	ERBB3	p.D297Y	Sub	missense	0.57
PD41857b_EMD_A2	19	47422061	GCGCCCGAGTGCTGA	G	ARHGAP35	p.P45fs*13	Del	frameshift	0.5
PD41857b_EMD_A4	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.38
PD41857b_EMD_E3	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.52
PD41857b_EMD_G1	19	47422061	GCGCCCGAGTGCTGA	G	ARHGAP35	p.P45fs*13	Del	frameshift	0.4776
PD41857b_EMD_G2	3	178928079	G	A	PIK3CA	p.E453K	Sub	missense	0.51
PD41857b_EMD_G3	12	25398284	C	T	KRAS	p.G12D	Sub	missense	0.55
PD41860b_EMD_A12	3	178936091	G	C	PIK3CA	p.E545Q	Sub	missense	0.5
PD41860b_EMD_E12	19	47422399	TGGTTGATGGTTTTCTTCTTGGTATTGATGTTAGCA	T	ARHGAP35	p.V157fs*4	Del	frameshift	0.3793
PD41860b_EMD_E12	3	178951957	G	A	PIK3CA	p.M1004I	Sub	missense	0.65

PD41860b_EMD_G12	10	123258034	A	T	FGFR2	p.N550K	Sub	missense	0.55
PD41861b_EMD_A11	5	67589299	C	CA	PIK3R1	p.Y431fs*11	Ins	frameshift	0.5789
PD41861b_EMD_F11	19	52715982	C	T	PPP2R1A	p.R183W	Sub	missense	0.55
PD41861b_EMD_G10	17	7578289	C	T	TP53	p.G187D	Sub	missense	0.29
PD41861b_EMD_H10	7	151845935	CTTCT	C	KMT2C	p.K4358fs*7	Del	frameshift	0.6296
PD41868b_EMD_E5	6	106554849	GAGAAACC	G	PRDM1	p.E656fs*17	Del	frameshift	0.6296
PD41868b_EMD_G5	19	47425365	C	T	ARHGAP35	p.R1145*	Sub	nonsense	0.45
PD41868b_EMD_G5	10	89711991	TC	T	PTEN	p.P204fs*17	Del	frameshift	0.4516
PD41868b_EMD_G6	10	123279677	G	C	FGFR2	p.S252W	Sub	missense	0.41
PD41868b_EMD_G6	3	178921552	A	T	PIK3CA	p.N345I	Sub	missense	0.37
PD41869b_P42_EMD_A5	16	3786718	C	CAAA	CREBBP	p.R1498delinsL*	Ins	nonsense	0.5263
PD41869b_P42_EMD_A7	16	3786718	C	CAAA	CREBBP	p.R1498delinsL*	Ins	nonsense	0.5625
PD41869b_P42_EMD_C5	16	3786718	C	CAAA	CREBBP	p.R1498delinsL*	Ins	nonsense	0.4375
PD41869b_P42_EMD_C7	16	3786718	C	CAAA	CREBBP	p.R1498delinsL*	Ins	nonsense	0.3654
PD41869b_P42_EMD_E5	4	153247289	G	A	FBXW7	p.R505C	Sub	missense	0.37
PD41869b_P42_EMD_E7	16	3786718	C	CAAA	CREBBP	p.R1498delinsL*	Ins	nonsense	0.4167
PD41869b_P42_EMD_G5	4	153247289	G	A	FBXW7	p.R505C	Sub	missense	0.59
PD41869b_P42_EMD_H5	4	153247289	G	A	FBXW7	p.R505C	Sub	missense	0.52
PD41871b_P42_EMD_A4	7	151944989	GT	G	KMT2C	p.K843fs*11	Del	frameshift	0.1513
PD41871b_P42_EMD_C3	12	6711206	GCTT	G	CHD4	p.K119delK	Del	inframe	0.1509
PD41871b_P42_EMD_C4	19	47424921	C	T	ARHGAP35	p.R997*	Sub	nonsense	0.61
PD41871b_P42_EMD_E3	20	22563158	AGGGTCCAGAAGGAGCCCTTGCCGGGCTTGTCGGGCGAGC	A	FOXA2	p.R228_T240delRSPDKPGKGSFWT	Del	inframe	0.3016
PD41871b_P42_EMD_H3	20	22563158	AGGGTCCAGAAGGAGCCCTTGCCGGGCTTGTCGGGCGAGC	A	FOXA2	p.R228_T240delRSPDKPGKGSFWT	Del	inframe	0.05
PD42475b_EMD_A9	3	178928067	C	A	PIK3CA	p.P449T	Sub	missense	0.65
PD42475b_EMD_A9	3	178952072	A	G	PIK3CA	p.M1043V	Sub	missense	0.5
PD42475b_EMD_A9	19	52716323	C	T	PPP2R1A	p.S256F	Sub	missense	0.42
PD42475b_EMD_C8	3	178922363	T	C	PIK3CA	p.C378R	Sub	missense	0.52
PD42475b_EMD_E9	11	108117799	G	A	ATM	p.R337H	Sub	missense	0.43
PD42475b_EMD_E9	5	67591106	A	G	PIK3R1	p.K567E	Sub	missense	0.69
PD42475b_EMD_G10	5	67591106	A	G	PIK3R1	p.K567E	Sub	missense	0.56
PD42475b_EMD_G9	18	45391465	G	GT	SMAD2	p.T232fs*3	Ins	frameshift	0.5
PD42475b_EMD_G9	17	29664534	TGA	T	NF1	p.E2195fs*46	Del	frameshift	0.6452
PD42475b_EMD_G9	3	178938934	G	A	PIK3CA	p.E726K	Sub	missense	0.44
PD42475b_EMD_H10	5	67591106	A	G	PIK3R1	p.K567E	Sub	missense	0.65
PD42475b_EMD_H8	3	178952072	A	G	PIK3CA	p.M1043V	Sub	missense	0.38

PD42746b_EMD_P44_B10	5	67591117	TATCCAGCTG	T	PIK3R1	p.I571_L573delIQL	Del	inframe	0.4444
PD42746b_EMD_P44_G10	7	101844683	T	TCTCTCTCTCG	CUX1	p.D714fs*5	Ins	frameshift	0.5789

Appendix 16

R Notebook

LM

26062019

Modelling driver mutation burden in normal endometrium

Markdown file to document methods used in the analysis of the driver mutation burden in normal endometrium.

Load Libraries

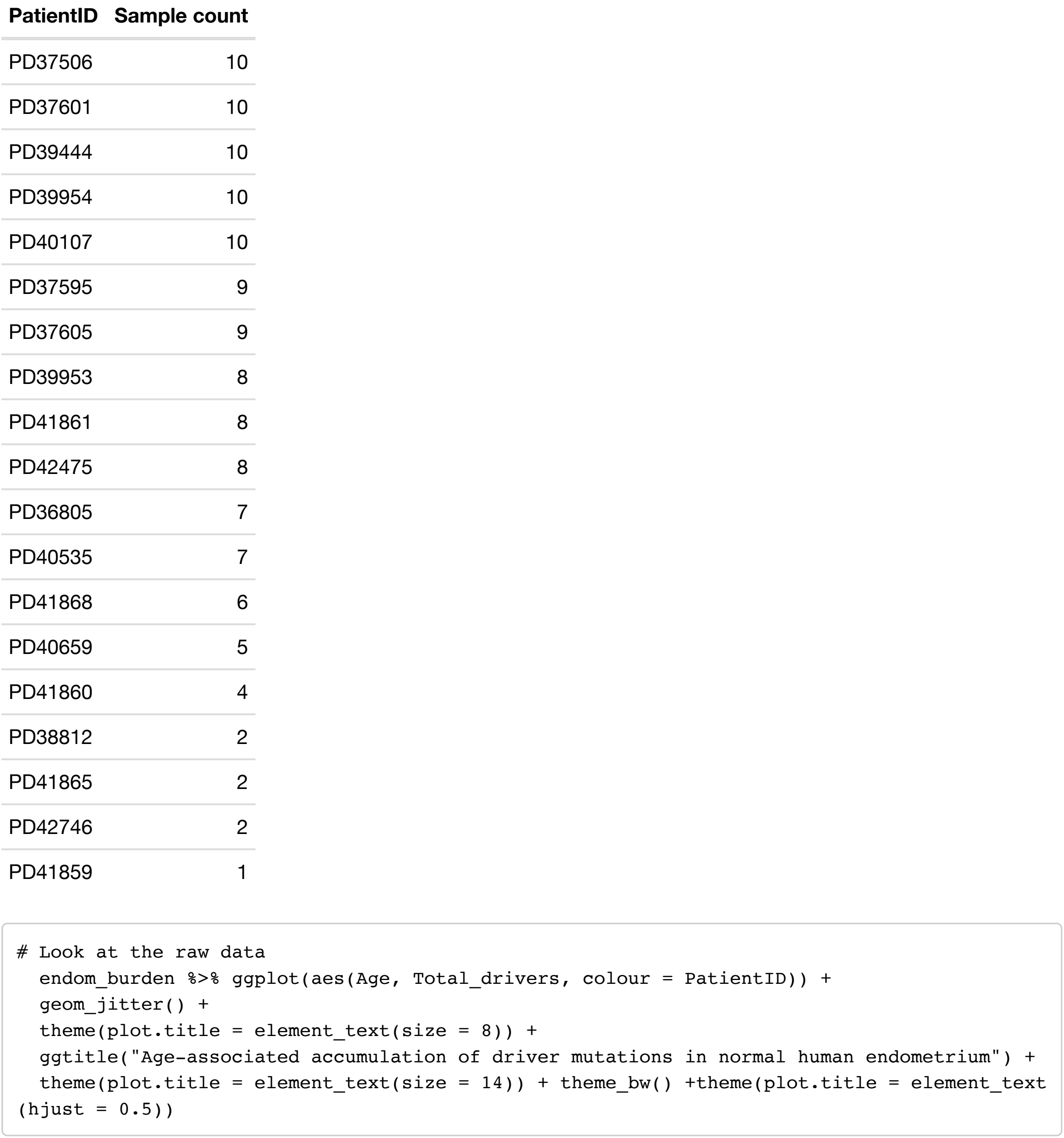
```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(pbkrtest)
```

Load in data files

Load in sample level data for the 28 donors with associated meta-data, including Body Mass Index (BMI), Parity and Cohort (sample source).

```
endom_burden <- read.csv("Endometrium_for_model_26062019.csv", stringsAsFactors = F, na.str
ings = c("", "NA", "Unknown", "Uncertain"))
# Samples per patient
endom_burden %>% group_by(PatientID) %>%  count(PatientID) %>%  rename(`Sample count` = n)
 %>% arrange(desc(`Sample count`)) %>%  kable() %>%  kable_styling(bootstrap_options = c("s
triped", "condensed"), full_width = F, position = "left")
```

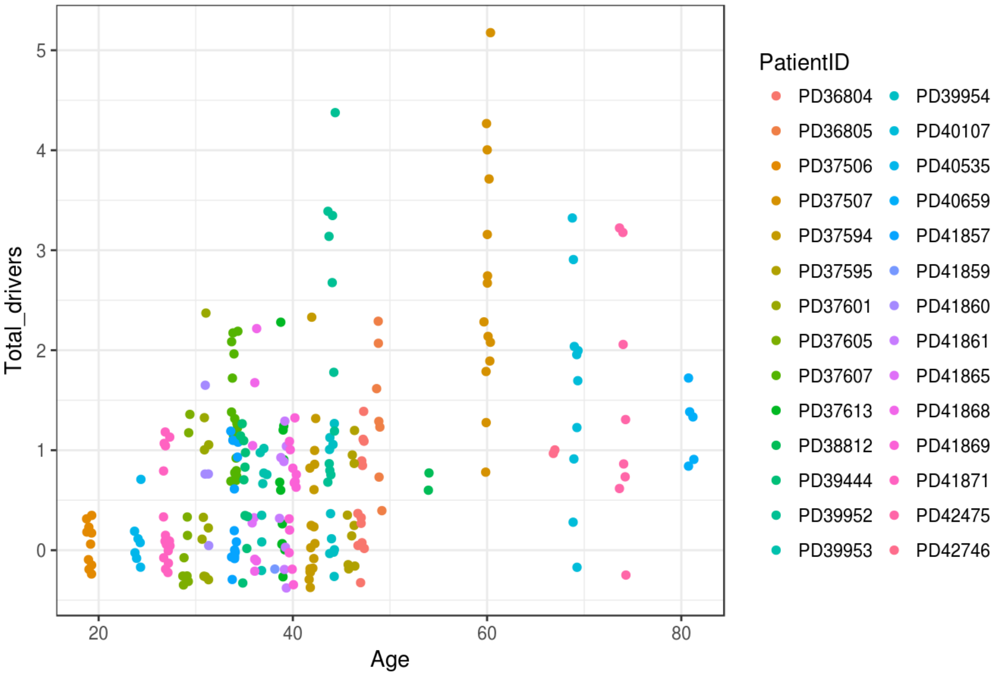
PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD37507	14
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11



PatientID	Sample count
PD37506	10
PD37601	10
PD39444	10
PD39954	10
PD40107	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD42475	8
PD36805	7
PD40535	7
PD41868	6
PD40659	5
PD41860	4
PD38812	2
PD41865	2
PD42746	2
PD41859	1

```
# Look at the raw data
  endom_burden %>% ggplot(aes(Age, Total_drivers, colour = PatientID)) +
  geom_jitter() +
  theme(plot.title = element_text(size = 8)) +
  ggtitle("Age-associated accumulation of driver mutations in normal human endometrium") +
  theme(plot.title = element_text(size = 14)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5))
```

sociated accumulation of driver mutations in normal human endometrium



Fit a mixed-effect model to estimate driver mutation rates

To account for the non-independent sampling per patient we use a generalized linear mixed effects model with Poisson distribution. We also use a random slope with fixed intercept as most women will start menarche at a similar age (~13 years), but to account for the potential differences in the rates at which mutations were acquired in different individuals due to variation in parity, contraception and other factors.

We test features that can have an effect on mutation burden or are modulate endometrial cancer risk:

- Age
- Read depth & VAF (‘Vafdepth’)
- BMI
- Parity
- Cohort

We use backwards elimination to define the final model

Define full model and drop each fixed effect in turn

```
# Combine read depth and median sample depth (Seq_X) as 'Vafdepth'
  endom_burden %<>%  mutate(Vafdepth = Seq_X*SampleMedianVAF)

# Make BMI and Parity numeric
  endom_burden %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden %<>%  mutate(Parity.QC = as.numeric(Parity))

# Exclude cases without Parity data
  endom_burden.qc <- endom_burden %>% filter(!is.na(Parity.QC))

# Define the full model containing all features
  full_glmer_model = glmer(Total_drivers ~ Age + Vafdepth + BMI.QC + Parity.QC + Cohort +(A
ge - 1|PatientID), data=endom_burden.qc, family = poisson(link = "log"), control =   glmerC
ontrol(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))

  print(summary(full_glmer_model))
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Vafdepth + BMI.QC + Parity.QC + Cohort +
##     (Age - 1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    483.6    514.6   -232.8    465.6      222
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.2757 -0.7002 -0.1361  0.5323  2.0615
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  4.832e-05 0.006951
## Number of obs: 231, groups:  PatientID, 25
##
## Fixed effects:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -1.937221   0.728279  -2.660  0.00781 **
## Age                     0.031603   0.011826   2.672  0.00753 **
## Vafdepth                0.044643   0.028273   1.579  0.11434
## BMI.QC                 -0.006626   0.023231  -0.285  0.77547
## Parity.QC              -0.259493   0.113226  -2.292  0.02192 *
## CohortPost-mortem       0.242012   0.917639   0.264  0.79199
## CohortTAH               0.153797   0.424937   0.362  0.71741
## CohortTransplant donor  0.304985   0.280186   1.089  0.27637
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth BMI.QC Prt.QC ChrtP- ChrTAH
## Age         -0.493
## Vafdepth    -0.311  0.087
## BMI.QC      -0.626 -0.136 -0.275
## Parity.QC   -0.271 -0.211 -0.003  0.371
## ChrtPst-mrt  0.300 -0.502  0.000 -0.013 -0.264
## CohortTAH    0.243 -0.275  0.115 -0.225 -0.197  0.281
## ChrtTrnspld  0.305 -0.450  0.045 -0.167 -0.210  0.412  0.388
```

```
# "user" parametric boot function as defined in drop1.merMod help example
  PBSumFun <- function(object, objectDrop, ...) {
    pbnames <- c("stat", "p.value")
    r <- if (missing(objectDrop)) {
      setNames(rep(NA, length(pbnames)), pbnames)
    } else {
      pbtest <- PBmodcomp(object, objectDrop, nsim = nsim, ref = NULL, seed=12345, details
 = 0)
      unlist(pbtest$test[2, pbnames])
    }
    attr(r, "method") <- c("Parametric bootstrap via pbkrtest package")
    r
  }
# Drop each fixed effect from model and test significance
# Use 1000 samples to form the reference distribution
nsim <- 1000
drop1(full_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Vafdepth + BMI.QC + Parity.QC + Cohort +
##     (Age - 1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##             stat p.value
## <none>
## Age       6.7178 0.05277
## Vafdepth  2.4586 0.14317
## BMI.QC    0.0821 0.83577
## Parity.QC 5.3143 0.08761
## Cohort    1.1445 0.85466
```

Remove feature with the largest P > 0.05 to make reduced model 1

```
# Remove Cohort from the full model
  reduced1_glmer_model <- update(full_glmer_model, ~ . -Cohort, control=glmerControl(optimi
zer="bobyqa", optCtrl = list(maxfun = 100000)))
# Drop each fixed effect from the model and test significance
  drop1(reduced1_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Vafdepth + BMI.QC + Parity.QC + (Age -
##     1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##              stat p.value
## <none>
## Age       10.8137 0.00326
## Vafdepth   2.3500 0.13436
## BMI.QC     0.0160 0.91478
## Parity.QC  4.7712 0.06361
```

Remove next feature with the largest P > 0.05 to make reduced model 2

```
# Remove BMI from the above model
  reduced2_glmer_model <- update(reduced1_glmer_model, ~ . -BMI.QC, control=glmerControl(op
timizer="bobyqa", optCtrl = list(maxfun = 100000)))
# Drop each fixed effect from the model and test significance
  drop1(reduced2_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Vafdepth + Parity.QC + (Age - 1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##              stat  p.value
## <none>
## Age       10.8621 0.002105
## Vafdepth   2.4033 0.137539
## Parity.QC  5.0721 0.037190
```

Remove next feature with the largest P > 0.05 to make reduced model 3

```
# Remove Vafdepth from the above model
  reduced3_glmer_model <- update(reduced2_glmer_model, ~ . -Vafdepth, control=glmerControl
(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))
# Drop each fixed effect from model and test significance
  drop1(reduced3_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##              stat  p.value
## <none>
## Age       10.3793 0.003125
## Parity.QC  5.8943 0.019348
```

Define the final model

```
# Define the final model keeping only the significant features  (P < 0.05)

  final_glmer_model <- reduced3_glmer_model

# Print summary for the final model
  print(summary(final_glmer_model))
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    477.1    490.9   -234.6    469.1      227
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.2451 -0.6912 -0.1927  0.6225  2.0057
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  5.987e-05 0.007738
## Number of obs: 231, groups:  PatientID, 25
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.643601   0.391387  -4.199 2.68e-05 ***
## Age          0.035460   0.009878   3.590 0.000331 ***
## Parity.QC   -0.253115   0.102227  -2.476 0.013285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Age
## Age       -0.930
## Parity.QC  0.204 -0.440
```

```
# Estimate confidence intervals using "likelihood profile" method
  confint.merMod(final_glmer_model, method = "profile")
```

```
## Computing profile confidence intervals ...
```

```
##                    2.5 %      97.5 %
## .sig01       0.002577037  0.01361534
## (Intercept) -2.493282376 -0.87980304
## Age          0.015388799  0.05650318
## Parity.QC   -0.463678195 -0.05087779
```

Appendix 17

R Notebook

Luiza Moore

26062019

Modelling the effect of menstrual phase on driver mutation burden

Markdown file to document methods used in the analysis of the driver mutation burden in normal endometrium.

Load Libraries

```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(pbkrtest)
```

Load in data

Load in sample level data for all 28 donors, but exclude post-menopausal women and women with undetermined menstrual phase.

```
  endom_burden <- read.csv("Endometrium_for_model_26062019.csv", stringsAsFactors = F, na.s
trings = c("", "NA", "Unknown", "Uncertain"))
  dim(endom_burden)
```

```
## [1] 257  25
```

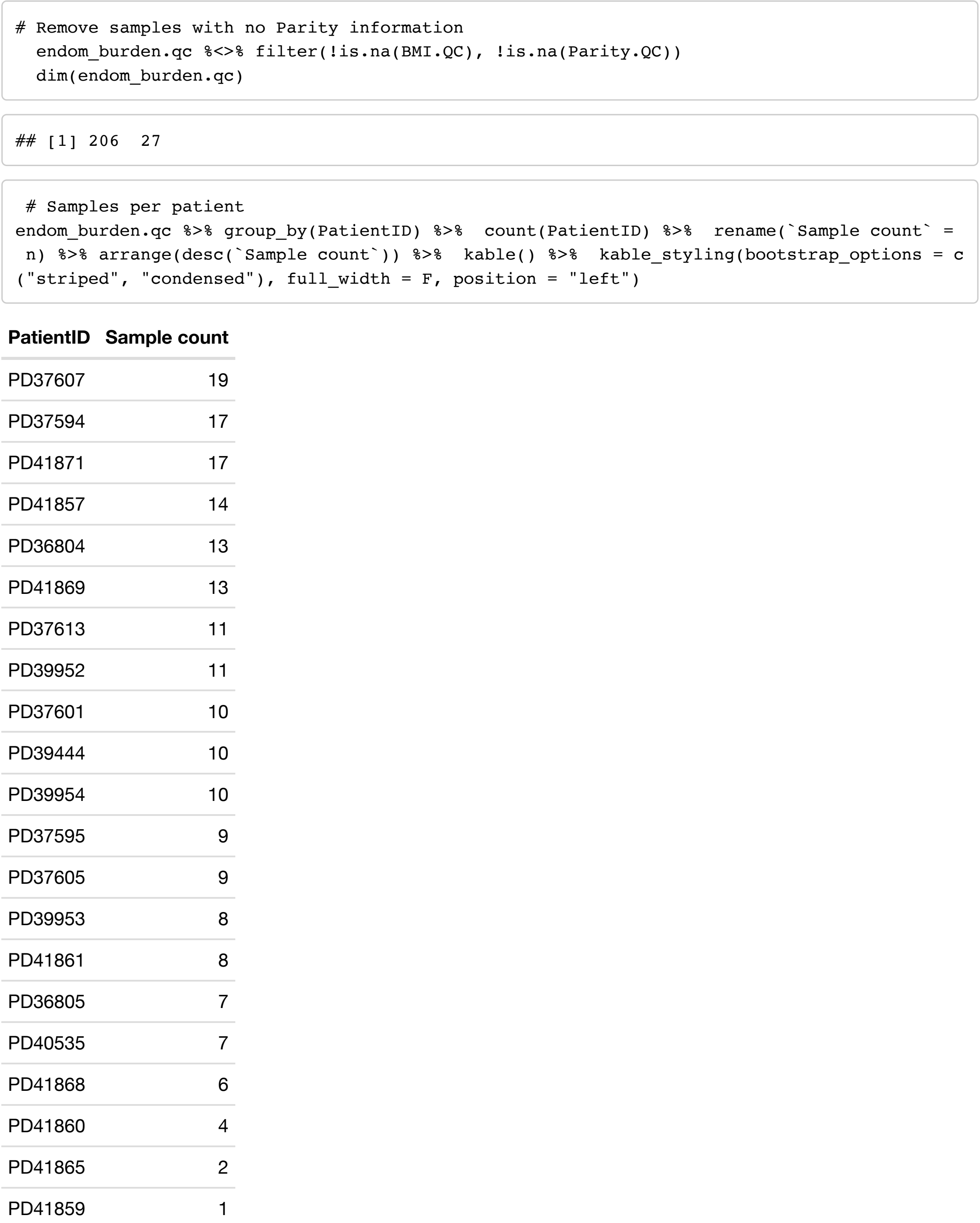
```
# Make BMI and Parity numeric
  endom_burden %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden %<>%  mutate(Parity.QC = as.numeric(Parity))

# Exclude post-menopausal women
  endom_burden.qc <- endom_burden %>% filter(Menopause_status_num == 0)
  dim(endom_burden.qc)
```

```
## [1] 218  27
```

```
# Exclude cases with undetermined menstrual phase
  endom_burden.qc <- endom_burden.qc %>% filter(Menstrual_phase_num >0)
  dim(endom_burden.qc)
```

```
## [1] 208  27
```



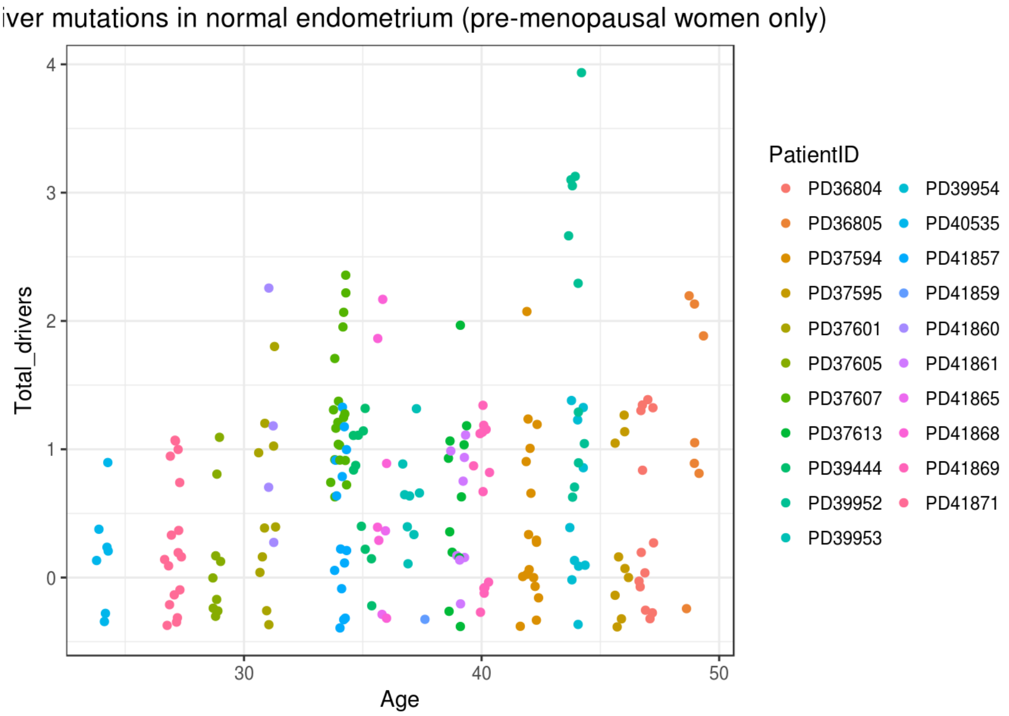
```
# Remove samples with no Parity information
  endom_burden.qc %<>% filter(!is.na(BMI.QC), !is.na(Parity.QC))
  dim(endom_burden.qc)
```

```
## [1] 206  27
```

```
 # Samples per patient
endom_burden.qc %>% group_by(PatientID) %>%  count(PatientID) %>%  rename(`Sample count` =
 n) %>% arrange(desc(`Sample count`)) %>%  kable() %>%  kable_styling(bootstrap_options = c
("striped", "condensed"), full_width = F, position = "left")
```

PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11
PD37601	10
PD39444	10
PD39954	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD36805	7
PD40535	7
PD41868	6
PD41860	4
PD41865	2
PD41859	1

```
# Look at the raw data
  endom_burden.qc %>% ggplot(aes(Age, Total_drivers, colour = PatientID)) +
  geom_jitter() +
  theme(plot.title = element_text(size = 8)) +
  ggtitle("Driver mutations in normal endometrium (pre-menopausal women only)") +
  theme(plot.title = element_text(size = 14)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5))
```



Does menstrual phase have an effect on the driver mutation burden?

To test the effect of menstrual phase on the driver mutation burden we add Menstrual phase to the final generalized linear mixed-effects model with Poisson distribution with features that have been shown to be significant in the full cohort of patients.

The significant features are:

- Age
- Read depth & VAF ('Vafdepth')
- Parity

We use backwards elimination to define the final model

Define full model and drop each fixed effect in turn

```
# Combine read depth and median sample depth (Seq_X) as 'Vafdepth'
  endom_burden.qc %<>%  mutate(Vafdepth = Seq_X*SampleMedianVAF)

# Make BMI and Parity numeric
  endom_burden.qc %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden.qc %<>%  mutate(Parity.QC = as.numeric(Parity))

# Define the full model containing all features
  full_glmer_model = glmer(Total_drivers ~ Age + Parity.QC + Menstrual_phase_num +(Age - 1|
PatientID), data=endom_burden.qc, family = poisson(link = "log"), control =   glmerControl
(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))

  print(summary(full_glmer_model))
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Parity.QC + Menstrual_phase_num + (Age -
##     1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    403.8    420.4   -196.9    393.8      201
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0933 -0.6763 -0.5314  0.6787  2.0963
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  7.543e-05 0.008685
## Number of obs: 206, groups:  PatientID, 21
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.97608    0.95243  -1.025   0.3054
## Age                  0.04002    0.01914   2.091   0.0366 *
## Parity.QC           -0.24689    0.10749  -2.297   0.0216 *
## Menstrual_phase_num -0.46049    0.32828  -1.403   0.1607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Prt.QC
## Age         -0.751
## Parity.QC   -0.031 -0.101
## Mnstrl_phs_ -0.649  0.011  0.024
## convergence code: 0
## Model failed to converge with max|grad| = 0.0018568 (tol = 0.001, component 1)
```

```
# "user" parametric boot function as defined in drop1.merMod help example
  PBSumFun <- function(object, objectDrop, ...) {
    pbnames <- c("stat", "p.value")
    r <- if (missing(objectDrop)) {
      setNames(rep(NA, length(pbnames)), pbnames)
    } else {
      pbtest <- PBmodcomp(object, objectDrop, nsim = nsim, ref = NULL, seed=12345, details
 = 0)
      unlist(pbtest$test[2, pbnames])
    }
    attr(r, "method") <- c("Parametric bootstrap via pbkrtest package")
    r
  }
# Drop each fixed effect from model and test significance
# Use 1000 samples to form the reference distribution
nsim <- 1000
drop1(full_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Parity.QC + Menstrual_phase_num + (Age -
##     1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##                       stat  p.value
## <none>
## Age                 4.2999 0.056701
## Parity.QC           5.1460 0.048857
## Menstrual_phase_num 1.7141 0.260549
```

Remove feature with the largest P > 0.05 to make reduced model

```
# Remove Menstrual phase from the full model
  reduced_glmer_model <- update(full_glmer_model, ~ . -Menstrual_phase_num, control=glmerCo
ntrol(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))
# Drop each fixed effect from the model and test significance
  drop1(reduced_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##             stat  p.value
## <none>
## Age       3.8150 0.067708
## Parity.QC 4.3927 0.063017
```

Define the final model

```
# Define the final model keeping only the significant features  (P < 0.05)
  final_glmer_model <- reduced_glmer_model

# Print summary for the final model
  print(summary(final_glmer_model))
```

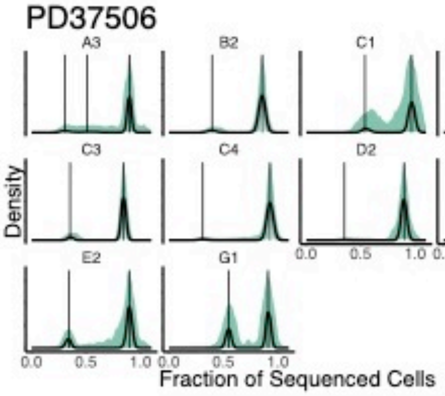
```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    403.5    416.8   -197.8    395.5      202
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1113 -0.6955 -0.4384  0.6488  2.1040
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  9.975e-05 0.009987
## Number of obs: 206, groups:  PatientID, 21
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.87925    0.77163  -2.435   0.0149 *
## Age          0.04092    0.02062   1.985   0.0471 *
## Parity.QC   -0.24412    0.11431  -2.136   0.0327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Age
## Age       -0.978
## Parity.QC  0.024 -0.145
```

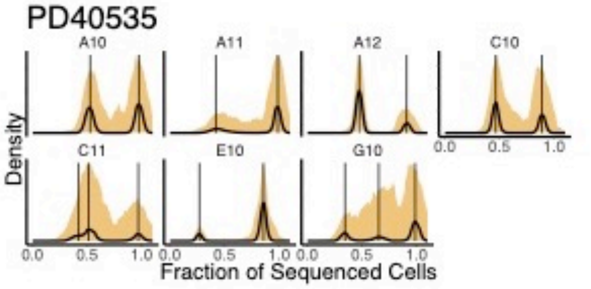
```
# Estimate confidence intervals using "likelihood profile" method
  confint.merMod(final_glmer_model, method = "profile")
```

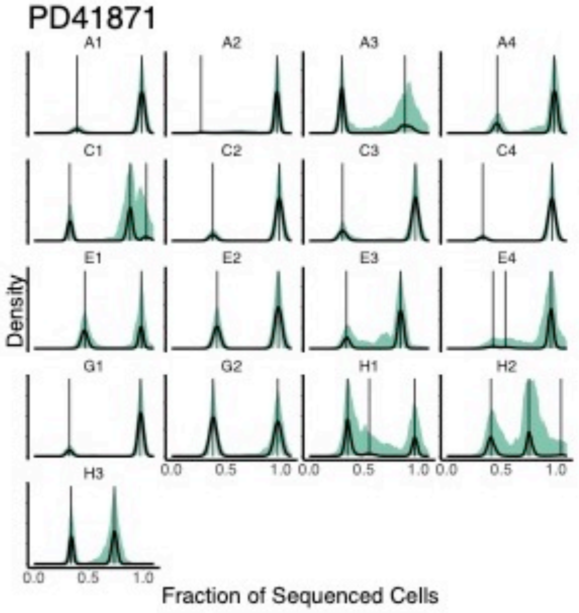
```
## Computing profile confidence intervals ...
```

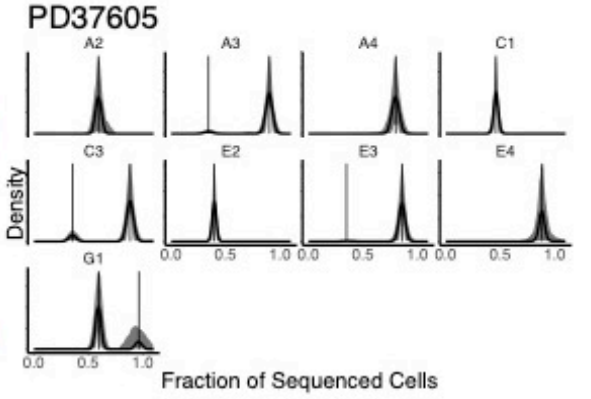
```
##                     2.5 %      97.5 %
## .sig01       0.0049871530  0.01716773
## (Intercept) -3.5142119925 -0.37232360
## Age         -0.0001535045  0.08423708
## Parity.QC   -0.4821811109 -0.01665213
```

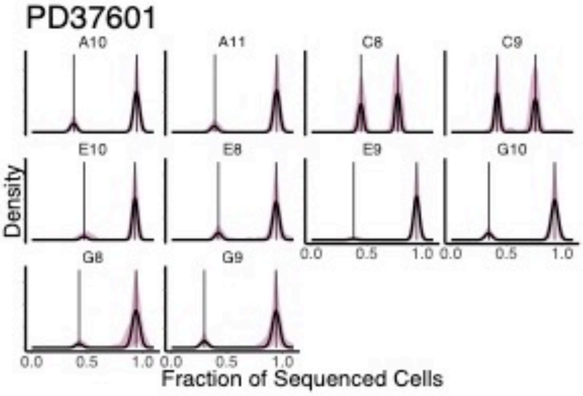
Appendix 18

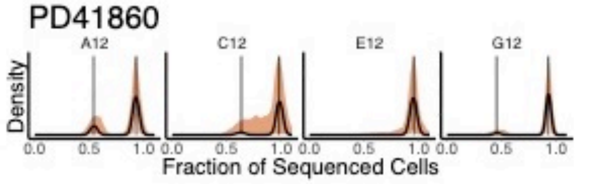

PD37506
A3
B2
C1
C2
C3
C4
D2
D4
E2
G1
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

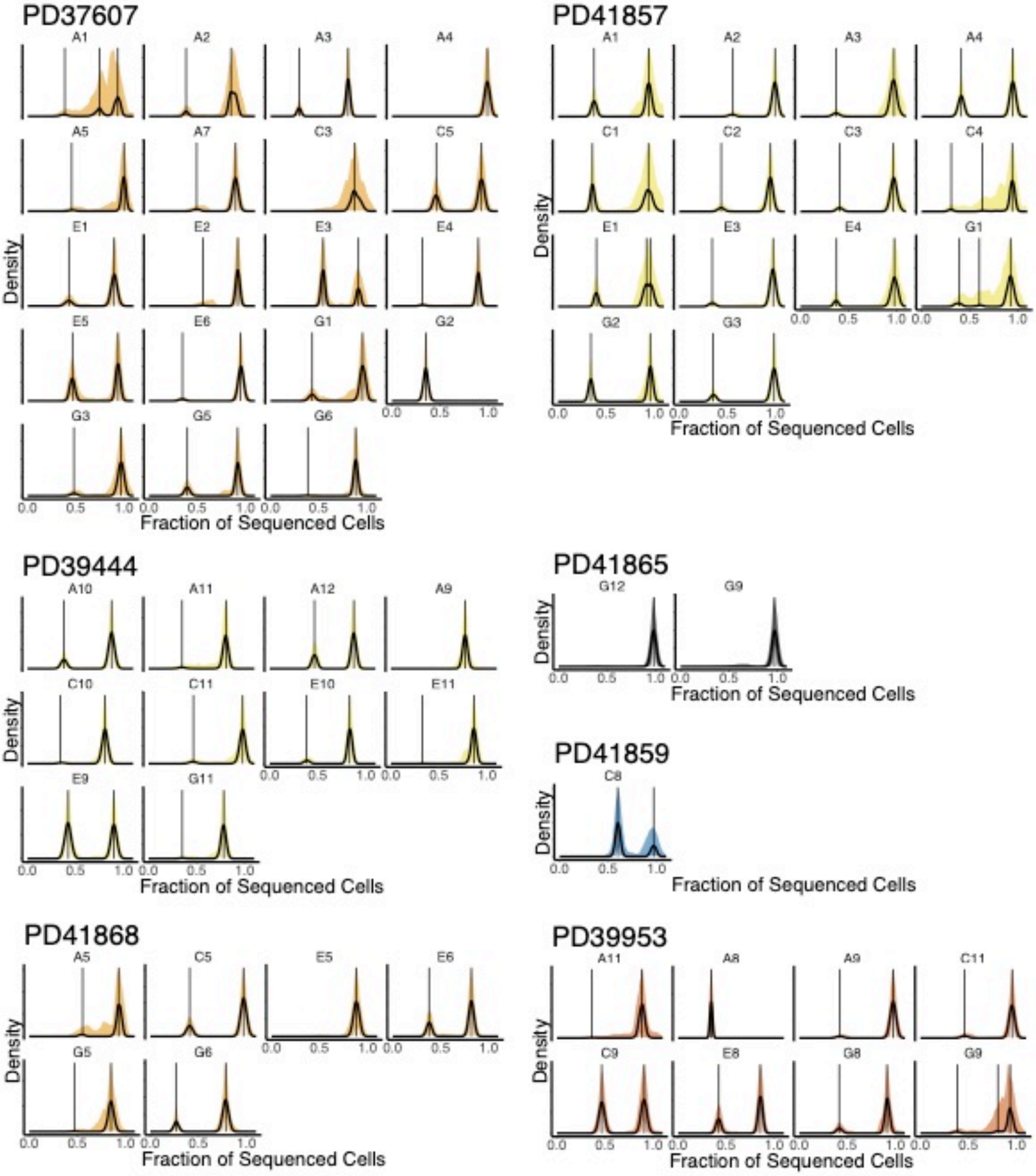

PD40535
A10
A11
A12
C10
C11
E10
G10
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

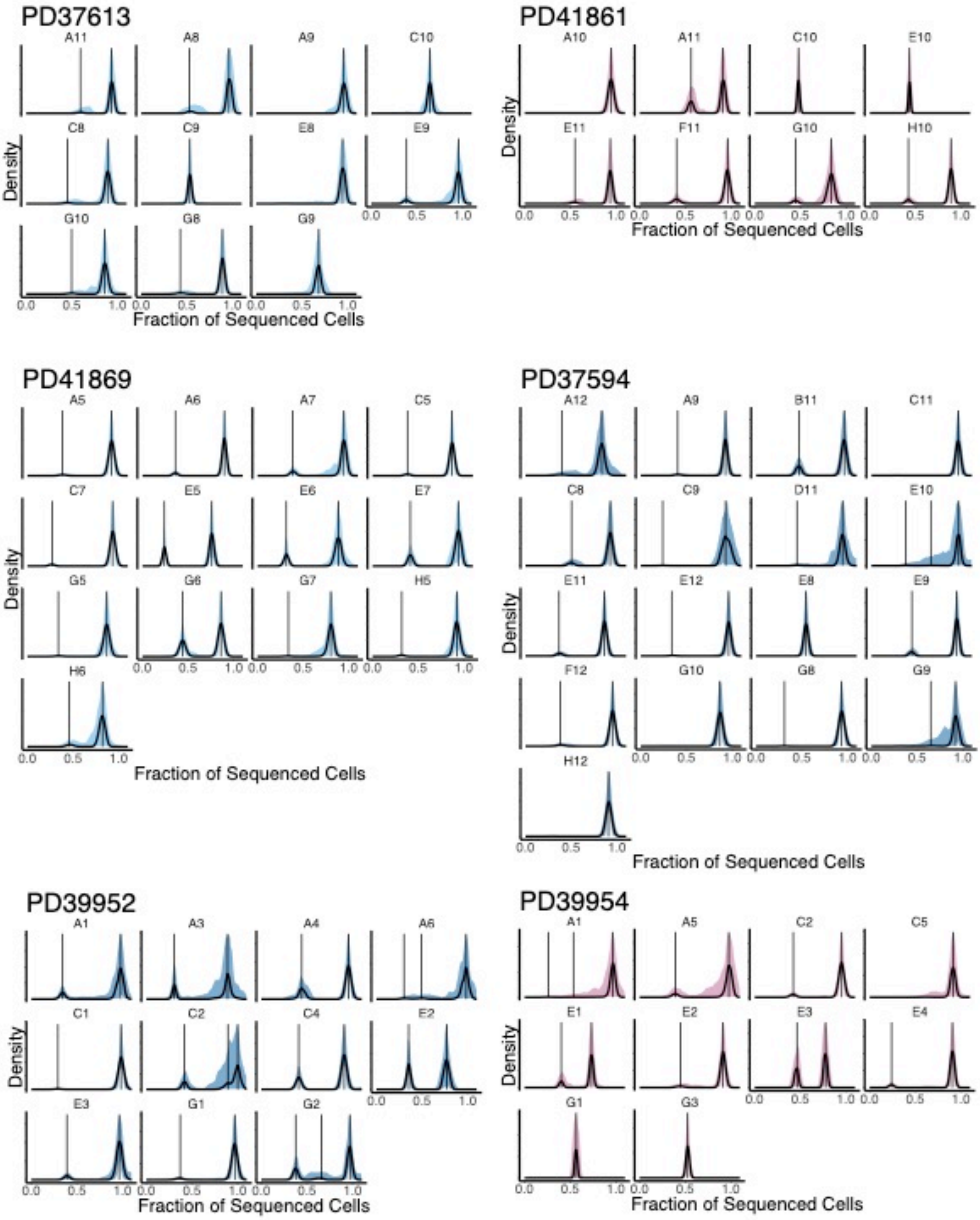

PD41871
A1
A2
A3
A4
C1
C2
C3
C4
E1
E2
E3
E4
G1
G2
H1
H2
H3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

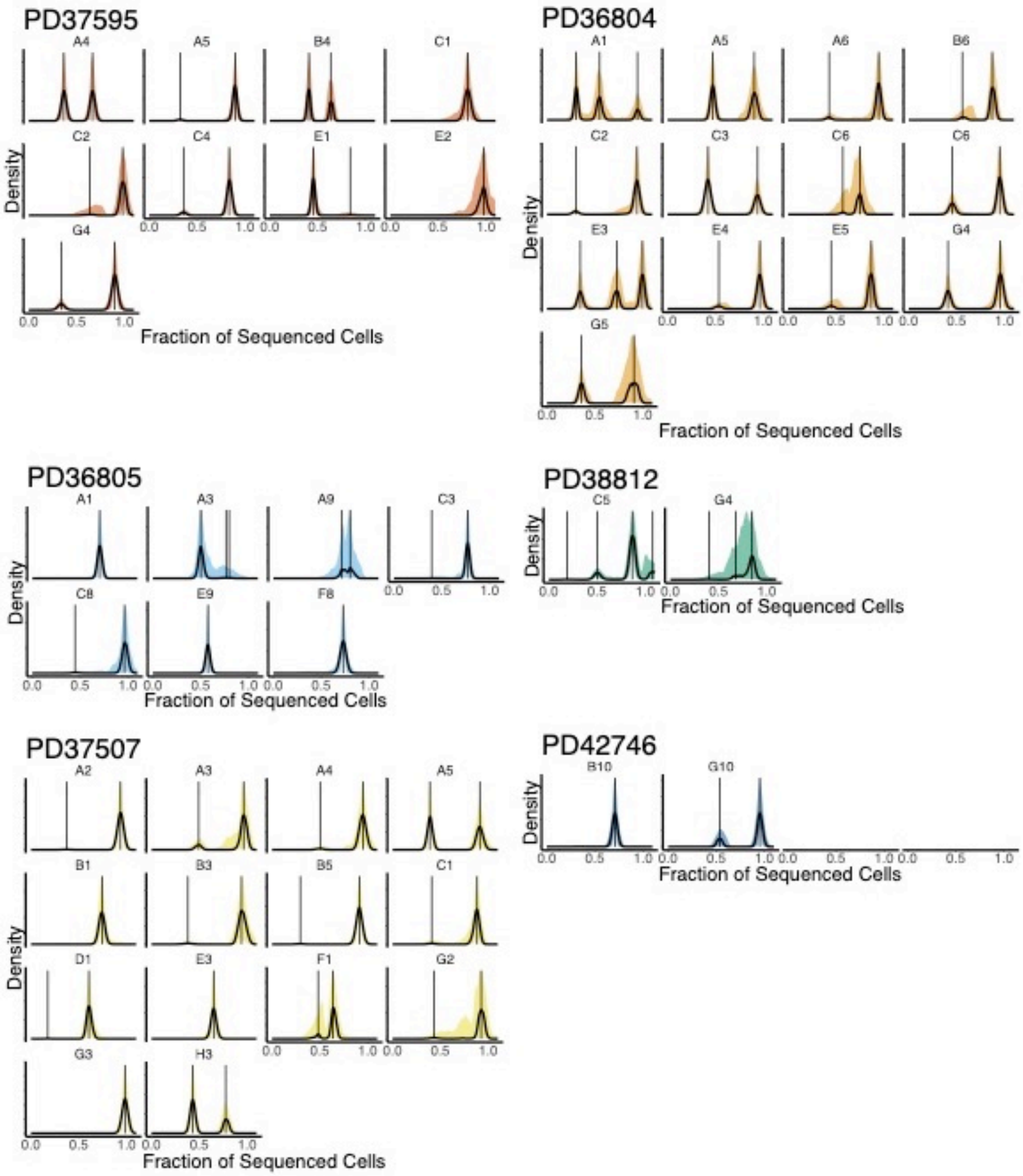

PD37605
A2
A3
A4
C1
C3
E2
E3
E4
G1
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

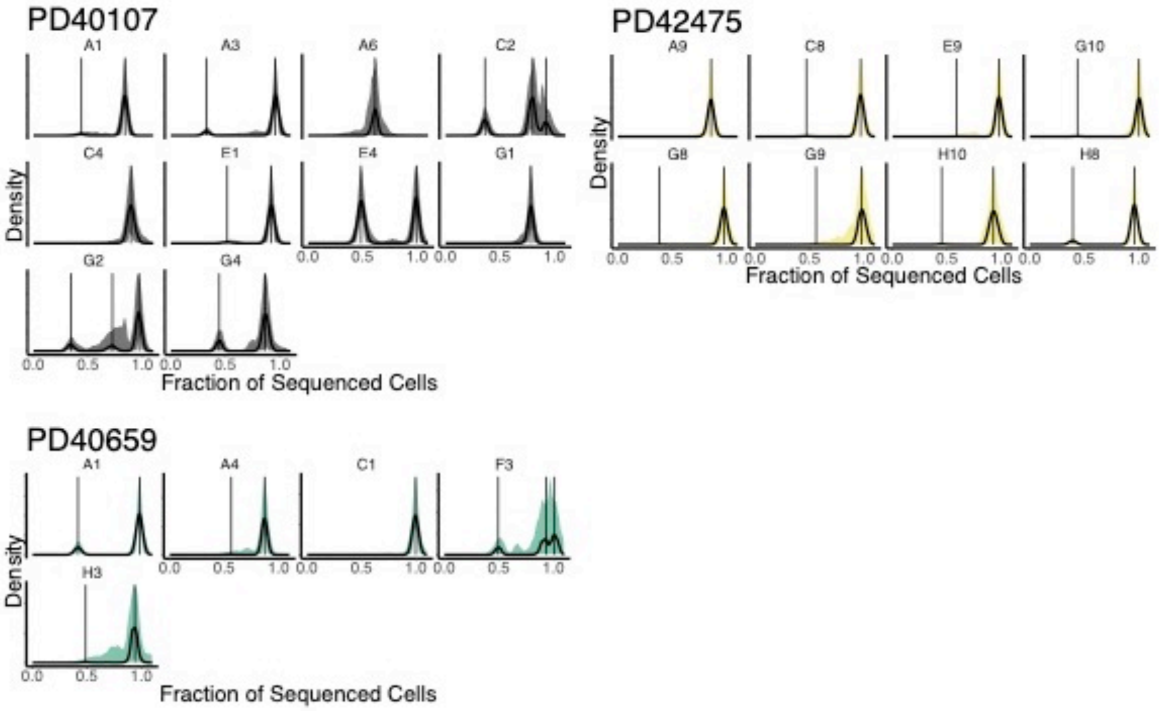

PD37601
A10
A11
C8
C9
E10
E8
E9
G10
G8
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD41860
A12
C12
E12
G12
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD37607
A1
A2
A3
A4
A5
A7
C3
C5
E1
E2
E3
E4
E5
E6
G1
G2
G3
G5
G6
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41857
A1
A2
A3
A4
C1
C2
C3
C4
E1
E3
E4
G1
G2
G3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD39444
A10
A11
A12
A9
C10
C11
E10
E11
E9
G11
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41865
G12
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41859
C8
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41868
A5
C5
E5
E6
G5
G6
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD39953
A11
A8
A9
C11
C9
E8
G8
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells


PD37613
A11
A8
A9
C10
C8
C9
E8
E9
G10
G8
G9
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD41861
A10
A11
C10
E10
E11
F11
G10
H10
Density
Fraction of Sequenced Cells
PD41869
A5
A6
A7
C5
C7
E5
E6
E7
G5
G6
G7
H5
H6
Density
Fraction of Sequenced Cells
PD37594
A12
A9
B11
C11
C8
C9
D11
E10
E11
E12
E8
E9
F12
G10
G8
G9
H12
Density
Fraction of Sequenced Cells
PD39952
A1
A3
A4
A6
C1
C2
C4
E2
E3
G1
G2
Density
Fraction of Sequenced Cells
PD39954
A1
A5
C2
C5
E1
E2
E3
E4
G1
G3
Density
Fraction of Sequenced Cells


PD37595
A4
A5
B4
C1
C2
C4
E1
E2
G4
Density
0.0 0.5 1.0
Fraction of Sequenced Cells
PD36804
A1
A5
A6
B6
C2
C3
C6
C6
E3
E4
E5
G4
G5
Density
0.0 0.5 1.0
Fraction of Sequenced Cells
PD36805
A1
A3
A9
C3
C8
E9
F8
Density
0.0 0.5 1.0
Fraction of Sequenced Cells
PD38812
C5
G4
Density
0.0 0.5 1.0
Fraction of Sequenced Cells
PD37507
A2
A3
A4
A5
B1
B3
B5
C1
D1
E3
F1
G2
G3
H3
Density
0.0 0.5 1.0
Fraction of Sequenced Cells
PD42746
B10
G10
Density
0.0 0.5 1.0
Fraction of Sequenced Cells


PD40107
A1
A3
A6
C2
C4
E1
E4
G1
G2
G4
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD42475
A9
C8
E9
G10
G8
G9
H10
H8
Density
0.0
0.5
1.0
Fraction of Sequenced Cells
PD40659
A1
A4
C1
F3
H3
Density
0.0
0.5
1.0
Fraction of Sequenced Cells

Appendix 19

Article

The mutational landscape of normal human endometrial epithelium

https://doi.org/10.1038/s41586-020-2214-z

Received: 19 December 2018

Accepted: 20 March 2020

[redacted]

Check for updates

Luiza Moore[1,2], Daniel Leongamornlert[1], Tim H. H. Coorens[1], Mathijs A. Sanders[1,3], Peter Ellis[1,4], Stefan C. Dentro[1,5], Kevin J. Dawson[1], Tim Butler[1], Raheleh Rahbari[1], Thomas J. Mitchell[1], Francesco Maura[1,6], Jyoti Nangalia[1], Patrick S. Tarpey[1], Simon F. Brunner[1], Henry Lee-Six[1], Yvette Hooks[1], Sarah Moody[1], Krishnaa T. Mahbubani[7,8,9], Mercedes Jimenez-Linan[2], Jan J. Brosens[10], Christine A. Iacobuzio-Donahue[11,12], Inigo Martincorena[1], Kourosh Saeb-Parsy[7,8], Peter J. Campbell[1] & Michael R. Stratton[1]✉

All normal somatic cells are thought to acquire mutations, but understanding of the rates, patterns, causes and consequences of somatic mutations in normal cells is limited. The uterine endometrium adopts multiple physiological states over a lifetime and is lined by a gland-forming epithelium[1,2]. Here, using whole-genome sequencing, we show that normal human endometrial glands are clonal cell populations with total mutation burdens that increase at about 29 base substitutions per year and that are many-fold lower than those of endometrial cancers. Normal endometrial glands frequently carry 'driver' mutations in cancer genes, the burden of which increases with age and decreases with parity. Cell clones with drivers often originate during the first decades of life and subsequently progressively colonize the epithelial lining of the endometrium. Our results show that mutational landscapes differ markedly between normal tissues—perhaps shaped by differences in their structure and physiology—and indicate that the procession of neoplastic change that leads to endometrial cancer is initiated early in life.

Acquisition of mutations is a ubiquitous feature of cells in living organisms. Although there has been comprehensive characterization of the somatic mutation landscape of human cancer[3–5], knowledge of the patterns of somatic mutation in normal cells is limited. This has mainly been due to the challenge of detecting somatic mutations in normal tissues. Several strategies have recently been developed to address this, including the sequencing of in vitro-derived cell clones from normal tissues[6–8], the sequencing of small biopsies that contain limited numbers of microscopic clones[9–12], the sequencing of microscopically distinguishable structural elements that are clonal units[13–15], highly error-corrected sequencing[16,17] and the sequencing of single cells[18]. Together, these approaches have begun to reveal differing mutation burdens between cell types, the patterns of acquisition of mutation burdens over time and the underlying mutational processes. These strategies have also shown that clones of normal cells with driver mutations in cancer genes are present in normal tissues. In the glandular epithelium of the colon, these mutations are relatively uncommon[14]—but in the squamous epithelia of the skin[9] and oesophagus[10], and in the blood[19–21], clones that carry drivers can constitute substantial proportions of the normal cells present after middle age.

The factors that determine differences in the mutation landscape between normal cell types are incompletely understood. However, these factors plausibly include the intrinsic structural and physiological features of each tissue. The endometrium is a uniquely dynamic tissue composed of a stromal cell layer invaginated by a contiguous glandular epithelial sheet that covers the luminal surface. Endometrium adopts multiple different physiological states during life, including in premenarche, menstrual cycling, pregnancy and postmenopause. During reproductive years, the endometrium undergoes cyclical breakdown, shedding, repair and remodelling in response to oscillating levels of oestrogen and progesterone, which together entail the iterative restoration of the contiguity of the interrupted glandular epithelial sheet that is effected by stem cells within basal glands retained after menstruation[1,2,22].

The characterization of the mutational landscapes of normal tissues is advancing our understanding of the succession of intermediate neoplastic stages between normal cells and the cancers that originate from them. Endometrial cancer is the most common gynaecological tumour in high-income countries, with a peak incidence at 75–80 years of age[23]. There are two major histological classes[24,25]. Type I, endometrioid carcinoma, is the more common of the two; the main known risk factor is oestrogen exposure, influenced by ages of menarche and menopause, and body mass index[24,26]. Type II, which includes serous and clear cell carcinomas, occurs in older women, with smoking and

[1]Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Cambridge, UK. [2]Department of Pathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [3]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. [4]Inivata Ltd, Cambridge, UK. [5]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK. [6]Myeloma Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Department of Surgery, University of Cambridge, Cambridge, UK. [8]Cambridge NIHR Biomedical Research Centre, Cambridge, UK. [9]Department of Haematology, University of Cambridge, Cambridge, UK. [10]Tommy's National Miscarriage Research Centre, Warwick Medical School, University of Warwick, Coventry, UK. [11]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [12]Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ✉e-mail: mrs@sanger.ac.uk

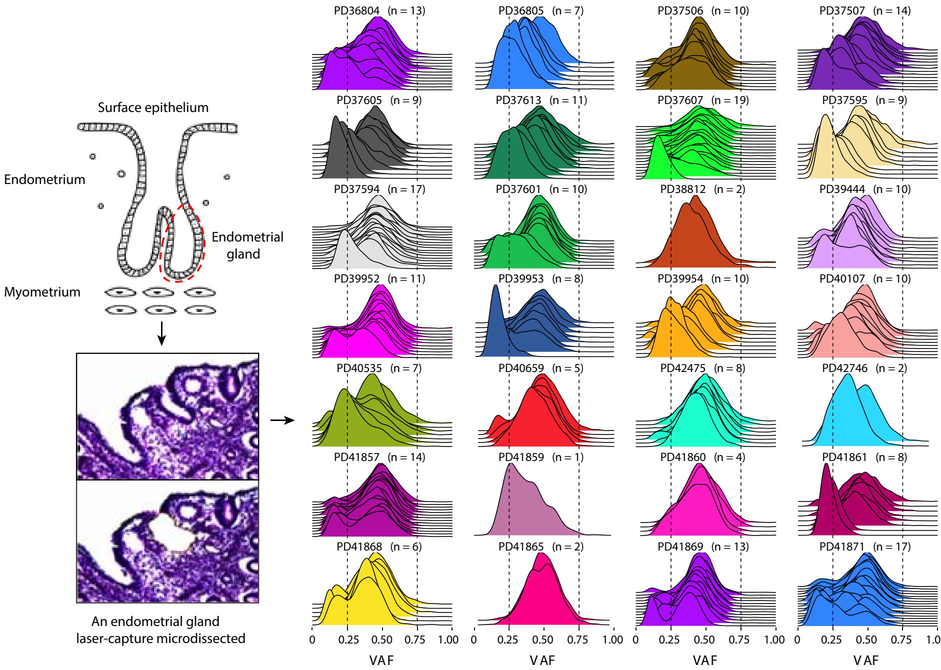


Fig. 1 | Clonality of normal endometrial glands. Individual normal endometrial glands were laser-capture microdissected and whole-genome sequenced. Most (91%, 234 out of 257) of the glands were clonal cell populations with a median VAF between 0.3 and 0.5 for base substitutions. Each density line represents an endometrial gland sample; individual samples are grouped and coloured by patient (n = 28).

body mass index as risk factors[27]. Commonly mutated cancer genes include *PTEN*, *TP53*, *PIK3CA*, *KRAS*, *ARID1A*, *FBXW7* and *PIK3R1*[28], and subsets of endometrial cancer carry many base substitution and/or small insertion and deletion (indel) mutations due to defective DNA mismatch repair or polymerase proof-reading mutations, or many copy number changes and genome rearrangement[29].

Recent studies using targeted sequencing have revealed driver mutations in known cancer genes in a high proportion of endometrial glands in endometriosis[13,30,31] and eutopic normal endometrial epithelium[13,32]. Here, by whole-genome sequencing of individual glands, we comprehensively characterize the mutational landscape of normal endometrial epithelium, explore the influences of age and parity, and estimate the timing of driver mutations.

Samples and sequencing

We used laser-capture microdissection to isolate 292 histologically normal endometrial glands from 28 women aged between 19 and 81 years. Samples were obtained from biopsies taken for the investigation of reproductive problems (14 women), hysterectomies for benign non-endometrial pathologies (2 women), residual tissues from transplant organ donors (8 women) and autopsies after death from nongynaecological causes (4 women). DNA from each gland was whole-genome sequenced using a protocol that accommodates small amounts of input DNA[14]. The mean sequencing coverage was 28-fold; only samples with >15-fold coverage were included in subsequent analyses (*n* = 257) (Supplementary Results 1, 2). Somatic mutations in each gland were determined by comparison with whole-genome sequences from other tissues from the same individuals.

Clonality of endometrial glands

To assess whether endometrial glands comprise clonal cell populations, we examined the variant allele fractions (VAFs) of somatic mutations. Ninety-one per cent (234 out of 257) of microdissected endometrial glands showed distributions of VAFs with peaks between 0.3 and 0.5 (Fig. 1, Extended Data Fig. 1a), indicating that each gland consists predominantly of a cell population that is descended from a single epithelial progenitor stem cell (a formal clonality analysis is described in Methods, Extended Data Fig. 2, Supplementary Results 3). Subsequent analyses (described in 'Driver mutations') revealed that many endometrial glands carry driver mutations in known cancer genes. However, endometrial glands exhibited clonality irrespective of the presence of driver mutations (Extended Data Fig. 1b, Supplementary Results 4). Thus, colonization of endometrial glands by descendants of single endometrial epithelial stem cells is not contingent on a selective growth advantage provided by driver mutations, and may occur by a process analogous to genetic drift (as previously proposed for other tissues[33,34]).

Mutation burdens and signatures

Somatic mutation burdens in normal endometrial glands from the 28 women ranged from 209 to 2,833 base substitutions (median of 1,521) and 1 to 358 indels (median of 180) (Fig. 2a, b). This variation was predominantly attributable to age, with about 29 base substitutions per gland per year being acquired during adult life (linear mixed-effect model, 95% confidence interval 23–34, $P = 3.02 \times 10^{-11}$) (Supplementary Results 5, 6). The presence of a driver mutation was also associated with

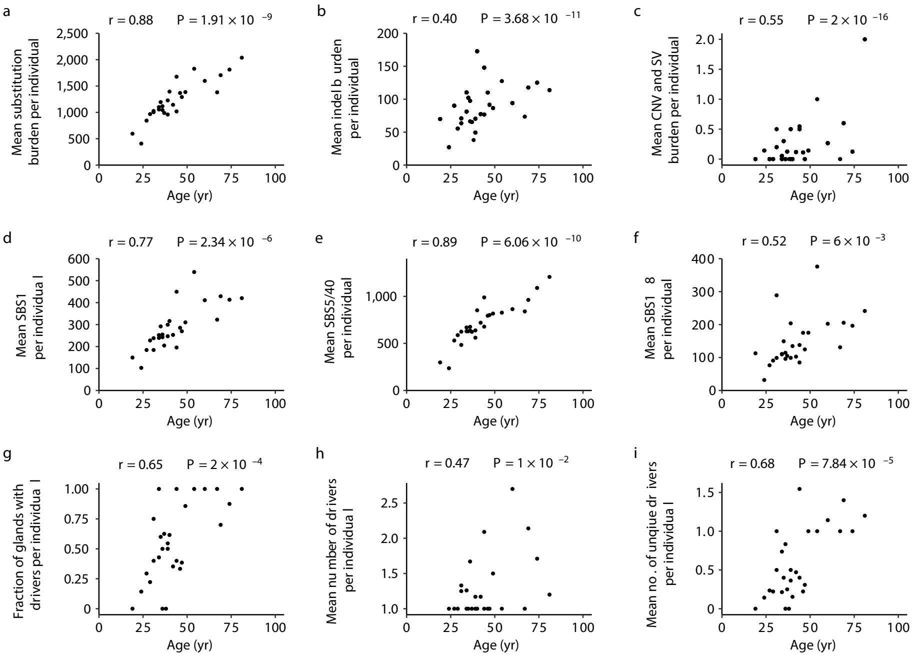


Fig. 2 | Mutation burden correlates with age in normal endometrial glands. Mutation burdens shown as mean for each donor (n = 28 donors), with Pearson correlation (r) with age and P values (P) from linear regression (burden–age). a–c, Variant burdens. a, Substitution burden. b, Indel burden. c, Copy-number variant (CNV) and structural variant (SV) burden. d–f, SBS burdens. d, SBS1 burden. e, SBS5/40 burden. f, SBS18. g–i, Driver mutation burden per gland. g, Fraction of glands with drivers, per individual. h, Mean number of driver mutations in glands with drivers. i, Mean number of unique (different) driver mutations per gland.

an additional approximately 110 substitutions (95% confidence interval 43–177, $P = 1.34 \times 10^{-3}$). There was no obvious correlation between parity and total somatic mutation burden.

We identified five previously described single-base-substitution (SBS) mutational signatures (Supplementary Results 7–9): SBS1, which is predominantly characterized by NCG > NTG mutations and is probably due to spontaneous deamination of 5-methylcytosine; SBS5 and SBS40, two relatively featureless 'flat' signatures of uncertain cause; SBS18, predominantly characterized by C > A substitutions and possibly due to reactive oxygen species [35]; and SBS23, a signature predominantly composed of C > T mutations and of unknown aetiology. Because SBS5 and SBS40 are relatively featureless, it is challenging to estimate their separate contributions [4] and they have therefore been combined (designated SBS5/40) (but shown separately in Supplementary Results 8, 9). SBS23 has previously occasionally been found in liver cancers with high mutation burdens. Given the low mutation burden and small contribution of SBS23 in the data reported here, it is unclear whether this is the same signature and so SBS23 was included in the 'unattributable' category. The mean signature exposures were 0.23 for SBS1, 0.58 for SBS5/40 and 0.12 for SBS18. There were positive linear correlations with age for the mutation burdens attributable to each of these three signatures (Fig. 2d–f). To ascertain the periods during which different mutational processes operate, we constructed phylogenetic trees of endometrial glands for each individual, which indicated that the mutational processes that underlie these three signatures are active throughout life (Figs. 3, 4, Extended Data Fig. 3). In regard to small indels, single T insertions at runs of T bases were the most common type of mutation that we observed (Supplementary Results 10).

Somatic copy-number changes and structural variants were found in 36 out of 257 (14%) normal endometrial glands, almost all of which carried just a single change (Extended Data Fig. 4, Supplementary Results 4). These changes included copy-number neutral loss of heterozygosity in 8 glands, whole chromosome copy-number increase in 1 gland and structural variants in 18 glands (12 large deletions, 6 tandem duplications and 9 translocations). One of three glands carrying a TP53 mutation exhibited nine structural variants, indicating that genomic instability caused by defective DNA maintenance occurs in normal cells.

Driver mutations

To identify genes under positive selection, we used a statistical method based on the observed:expected ratios of nonsynonymous:synonymous mutations [28]. Twelve genes showed evidence of positive selection in the 257 normal endometrial glands: PIK3CA , PIK3R1 , ARHGAP35 , FBXW7 , ZFHX3 , FOXA2 , ERBB2 , CHD4 , KRAS , SPOP , PPP2R1A and ERBB3 (Supplementary Results 11). All were listed among 369 genes that have previously been shown to be under positive selection in human cancer [28]. To identify additional drivers in the 257 endometrial glands, we sought mutations with the characteristics of drivers in those 369 genes (Methods). In total, we found 209 driver mutations in normal endometrial glands from 25 out of 28 women (Supplementary Results 4). The youngest carrier was a 24-year-old woman (patient PD40535) with a KRAS ^{G12D} mutation in 1 out of 7 glands that we sampled. We found that 147 out of 257 endometrial glands carried at least 1 driver mutation; 42 out of 257 glands carried at least 2 drivers; and 5 out 257 glands carried at least 4 drivers. In 4 women (aged 34 (19 glands), 44 (11 glands),

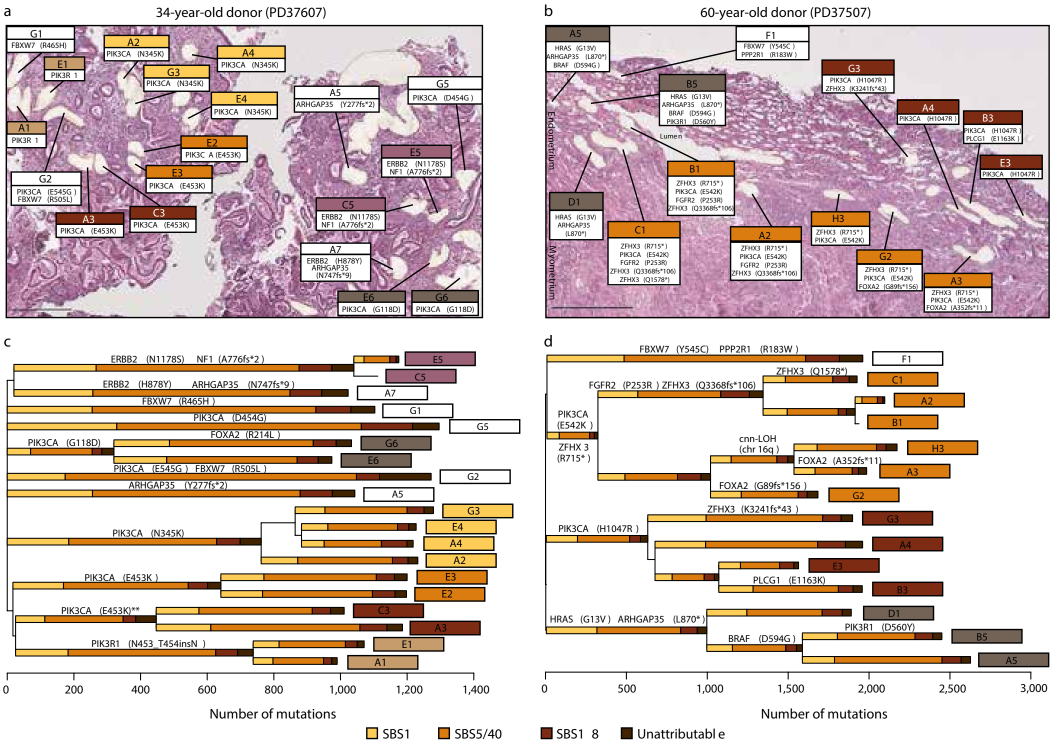


Fig. 3 | Histology images and reconstructed phylogenetic trees for two individuals in whom every normal endometrial gland contained at least one driver mutation. a, b, Haematoxylin and eosin images of endometrial glands from a 34-year-old woman (a) and a 60-year-old woman (b) were taken after laser-capture microdissection (20× magnification). c, d, Phylogenetic trees were reconstructed for the 34-year-old woman (c) and 60-year-old woman (d) using SBSs; the length of each branch is proportional to the number of variants. A stacked bar plot of the attributed SBS mutational signatures that contributed to each branch is then superimposed onto every branch; signature extraction was not performed on branches with fewer than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only, as it is not possible to time the different signatures within individual branches. Glands that shared over 100 variants were considered part of the same clade (indicated by the colour of the sample identifier label). Glands that did not belong to any clades are in white. SBS signatures are colour-coded; substitutions that were not attributed to the reference signatures, and those attributed to SBS23, are shown as 'unattributable'. Scale bars, 500 μm.

60 (14 glands) and 81 (5 glands)), all of the glands that we analysed carried driver mutations, which suggests that the whole endometrium had been colonized by microneoplastic clones (Fig. 3, Extended Data Fig. 3). The fraction of endometrial glands carrying a driver (Fig. 2g), the mean number of drivers per gland (Fig. 2h) and the number of different drivers in each individual (corrected for the number of glands sampled) (Fig. 2i) all positively correlated with age of the individual. However, there were sufficient outliers to suggest that other factors influence the colonization of the endometrium by driver-carrying clones. Indeed, our generalized linear mixed-effect model showed that in addition to the positive association of age with the accumulation of driver mutations (0.035 driver mutations per year, 95% confidence interval 0.01–0.06, $P = 3.31 \times 10^{-4}$), parity had a negative association (−0.253 driver mutations per life birth, 95% confidence interval −0.46 to −0.05, $P = 1.33 \times 10^{-2}$) (Supplementary Results 12, 13).

We found driver mutations in recessive (tumour-suppressor genes) and dominant cancer genes, similar to recent publications[13,30,32]. PIK3CA was the most frequently mutated cancer gene (Fig. 3, Extended Data Figs. 3, 5, Supplementary Results 14). Most truncating drivers in recessive cancer genes were heterozygous, indicating that haploinsufficiency confers a growth advantage in normal cells. Nevertheless, further inactivating mutations in the same genes in other glands show that an additional advantage is conferred by complete abolition of their activity (notably for ZFHX3 in the 60-year-old woman) (Fig. 3). Driver mutations were found in genes that encode growth factor receptors (ERBB2 , ERBB3 and FGFR2), components of signal transduction pathways (HRAS , KRAS , BRAF , PIK3CA , PIK3R1 , ARHGAP35 , RRAS2 , NF1, PPP2R1A and PTEN), pathways that mediate responses to steroid hormones (ZFHX3 , FOXA2 and ARHGAP35), proteins involved in chromatin function (KMT2D and ARID5B) and protein-mediated degradation pathways (FBXW7) that target oncoproteins, such as mTOR and MYC. Many different combinations of mutated cancer genes were found in individual glands.

Timing of driver mutations

Constructing phylogenetic trees of individual endometrial glands enabled the characterization of the mode of expansion of normal cell clones with drivers and the timing of their initiation. Glands with a phylogenetically close relationship were often in close physical

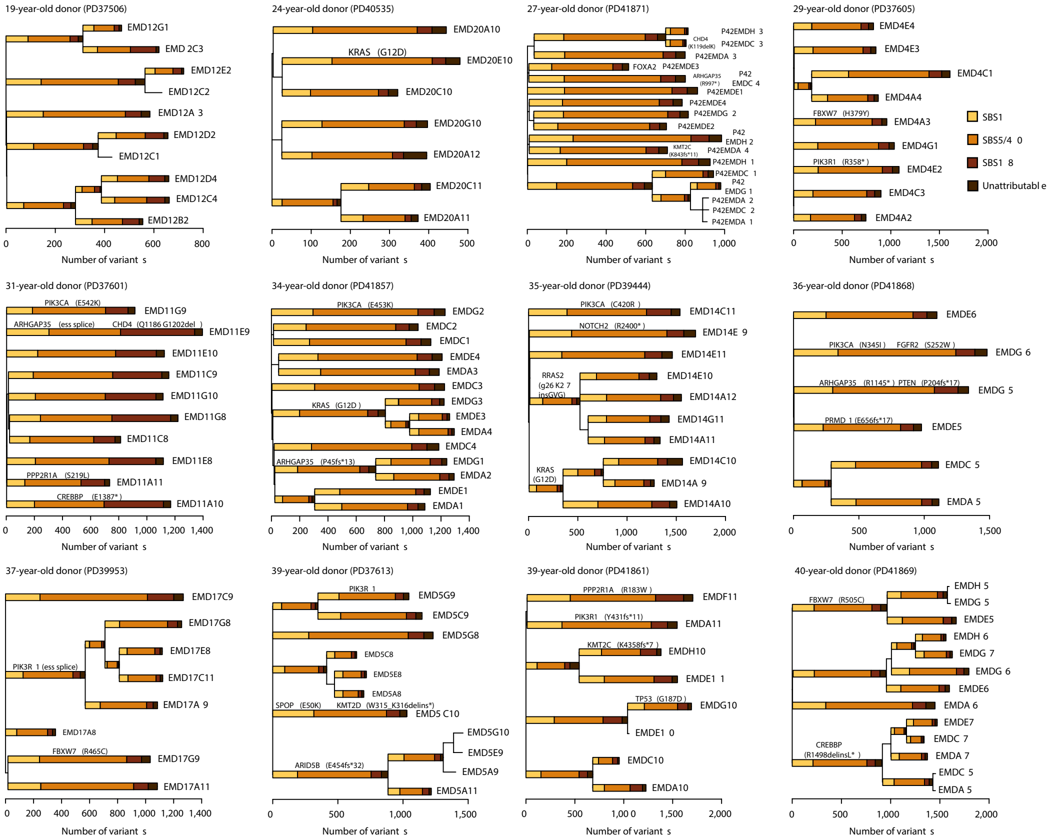


Fig. 4 | Phylogenetic trees of endometrial glands for donors aged 19 to **40 years.** Phylogenetic trees for individuals aged 19 to 40 years were reconstructed using SBSs with branch length proportional to the number of variants; the stacked bar plots represent the attributed SBS mutational signatures that contributed to each branch. Signature extraction was not performed on branches with fewer than 100 substitutions. The ordering of signatures within each branch is for visualization purposes only, as it is not possible to time different signatures within individual branches. SBS signatures are colour-coded; substitutions that were not attributed to the reference signatures, and those attributed to SBS23, are shown as 'unattributable'. EMD codes refer to individual endometrial glands.

proximity within the endometrium (Fig. 3). In phylogenetic clusters for which the mutation catalogues were almost identical, this may simply reflect multiple sampling of a single tortuous gland that weaves in and out of the plane of section, rather than distinct glands with their own stem cell populations (for example, glands C5 and E5 in Fig. 3a, c). For other phylogenetic clusters, the different branches within the clade have diverged substantially, sometimes acquiring different driver mutations, and therefore are probably derived from different stem cell populations. In such instances, phylogenetically related glands can range over distances of hundreds of micrometres, which suggests that their clonal evolution has entailed the capture and colonization of extensive zones of the endometrial lining (for example, glands C1, A2, B1, H2, A3 and B3 in Fig. 3b, d). Conversely, some glands in close physical proximity are phylogenetically distant (for example, glands E1 and G2 in Fig. 3a, c), indicating that their cell populations have remained isolated from each other.

Driver mutations were positioned on the phylogenetic trees for each individual, and times of occurrence were estimated by assuming constant somatic mutation rates during life (Fig. 5, Extended Data Figs. 6, 7, Methods). Although this assumption is unlikely to be completely correct, the results show that mutations in normal endometrial cells are acquired in a more-or-less linear fashion throughout life and potential modifying factors, including acquisition of a driver, make only modest differences to mutation rates. Furthermore, overall our approach is likely to overestimate the ages before which driver mutations have occurred, because it does not account for the time taken for a single endometrial stem cell to colonize an individual gland, which—in colorectal crypts—has been estimated to take several years [36]. Therefore, our results indicate that at least some driver mutations occur early in life. These included a KRAS^{G12D} mutation in 3 glands from a 35-year-old woman, and a PIK3CA mutation in 2 glands from a 34-year-old woman, both of which are likely to have arisen during the first decade of life (Figs. 3, 4, Extended Data Figs. 6, 7). A pair of drivers in ZFHX3 and PIK3CA , which co-occur in 6 glands from a 60-year-old woman, was also acquired during the first decade of life, indicating that driver-associated clonal evolution also begins early in life (Figs. 3, 5). It is possible that

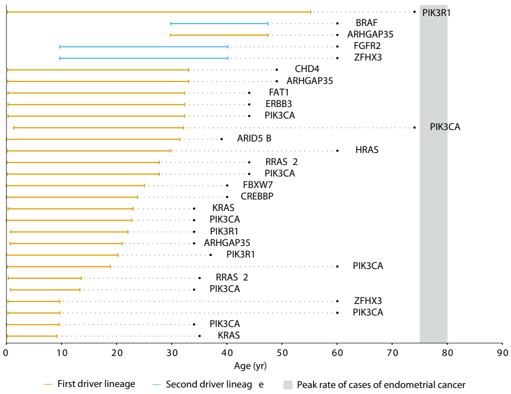


Fig. 5 | Timing of driver mutations in normal endometrial glands. To time the driver mutations, we reconstructed phylogenetic trees using SBSs. To estimate the time interval in which specific mutations occurred, we used two approaches (Methods). We calculated a patient-specific mutation rate by taking the ratio of the mean mutation burden per endometrial gland of the patient and age of the patient. The mutation number at the start and end of a branch in the phylogenetic tree was then converted to a lower and upper age by dividing these numbers by the estimated mutation rate. A similar approach was used for timing indels. We timed the driver mutations that occurred in the 'trunks' and branches. We display only those driver variants that occurred in the 'trunks' of the individual trees. We show that many such events occur decades before the reported peak incidence of endometrial cancer (variants with an interval of <1 year between the upper age and the age at sampling were excluded from this plot for illustration purposes). On the basis of our calculations, four driver variants (KRAS$^{\mathrm{G12D}}$, PIK3CA$^{\mathrm{G118D}}$, PIK3CA$^{\mathrm{E542K}}$ and ZFHX3$^{\mathrm{R715*}}$) from three different women occurred before the age of ten.

many more clones with drivers were initiated during the first decade of life, but their phylogenetic trees are not informative in this regard (Extended Data Figs. 6, 7). However, there was also evidence for the continued accumulation and clonal expansion of driver mutations into the later decades of life (Fig. 5, Extended Data Figs. 6, 7).

Comparison between normal tissue and cancer

Endometrial cancers (from the recent Pan Cancer Analysis of Whole Genomes (PCAWG) dataset[4]) exhibited higher mutation loads than normal endometrial cells for base substitutions (about 5-fold higher, medians of 1,346 and 7,330 in normal endometrium and endometrial cancer, respectively (Mann–Whitney U-test, $P = 7.63 \times 10^{-6}$)) and indels (Extended Data Fig. 8a, b). These differences also pertained to normal endometrial cells with driver mutations. In most endometrial cancers, the differences are attributable to higher mutation burdens of the ubiquitous base substitution and indel mutational signatures[3,4]. In addition, however, the very high mutation loads of the subsets of endometrial cancer with deficiencies in DNA mismatch repair and proof-reading mutations in polymerase-ε or polymerase-δ were not seen in normal endometrial cells. Differences between endometrial cancers and normal cells were even more marked for structural variants and copy-number changes (median number zero in normal endometrial cells and about 23 in endometrial cancers[37]), and this difference again pertained to normal endometrial cells with drivers.

There were also differences in the repertoire of cancer genes in which driver mutations were found (Extended Data Fig. 8c–e, Supplementary Results 4, 11). Notably, mutations in PTEN, CTCF, CTNNB1 and ARID1A in endometrioid, and in TP53 in serous carcinoma of the endometrium accounted for higher proportions of driver mutations than in normal endometrial cells. It is possible that PTEN, ARID1A, TP53 and CTCF require biallelic mutation to confer a growth advantage and this may account for their lower prevalence in normal cells. However, heterozygous mutations in PTEN and TP53 were found, albeit only in around 2% (5 out of 257) of all sampled glands, and this explanation would not account for the relative deficit of CTNNB1 mutations. Overall, the results suggest that driver mutations in some cancer genes are relatively effective at enabling the colonization of normal tissues, but confer a limited risk of conversion to invasive cancers. Conversely, other drivers may require biallelic mutation and/or confer limited advantage in colonizing normal tissues, but are relatively effective at the conversion to malignancy.

Discussion

Studies of normal endometrial epithelium and other types of normal cell[6,7,9,10,13–15,19,20] are revealing the landscape of somatic mutations in normal human cells. Somatic mutations are predominantly generated by a limited repertoire of ubiquitous mutational processes that generate base substitutions, small indels, genome rearrangements and whole chromosome copy-number changes, which exhibit more-or-less constant mutation rates during life. Additional mutational processes present only in some cells, some cell types and/or that are intermittent also contribute to the mutation burden—albeit apparently not in the endometrial epithelium.

The prevalence of clones with driver mutations is substantially different in different types of normal cell. Numerous cell clones with one or more driver mutations colonize much of the normal endometrial epithelium (as discussed in this Article, and in previous studies[13,32]), in contrast to another glandular epithelium, the colon, in which about 1% of normal crypts in middle-aged individuals carry a driver[13,14]. This is unlikely to be due to differences in the somatic mutation rate between endometrial and colonic epithelial cells, which are relatively modest; in any case, the somatic mutation rate is higher in the colon[6,14,38]. However, it may be attributable to intrinsic differences in structure and physiology between the endometrium and colon. In the endometrium, the cyclical process of tissue breakdown, shedding and remodelling iteratively opens up denuded terrain for pioneering clones of endometrial epithelial cells with drivers to preferentially colonize, compared to wild-type cells. In the colon, however, the selective advantage of a clone with a driver is usually confined to the small, siloed population of a single crypt, with only occasional opportunities for further expansion. Although the colonization of endometrium by driver clones progresses with age, it is already well-advanced in some young women—and parity has an inhibitory effect on it. The effect of parity is of particular interest as increased parity reduces the risk of endometrial cancer and it is conceivable that this is mediated by its effect on the expansion of driver clones[39]. Further studies of normal endometrium are required to assess how premenarchical and postmenopausal states, hormone contraceptive use and hormone replacement therapies influence the mutational landscape and its potential effect on pregnancy and fertility.

The burdens of all mutation classes are lower in normal endometrial cells (including those with drivers) than in endometrial cancers. Therefore, in endometrial epithelial stem cells, and in all other tissues studied thus far (including colon, oesophagus and skin), normal mutation rates are sufficient to generate large numbers of clones with driver mutations that behave as normal cells, but acquisition of an elevated mutation rate and burden is associated with further evolution to invasive cancer[9,10,14]. Because the endometrial epithelium is extensively colonized by clones of normal cells with driver mutations in middle-aged women, and the lifetime risk of endometrial cancer is only 3% (ref.[23]), this conversion from a normal cell clone with drivers to symptomatic malignancy appears to be extremely rare.

The first driver mutations in normal endometrial clones with drivers can arise within the first decade of life, and our results are compatible with of many doing so. The modal period of diagnosis of endometrial cancer is 75–80 years of age. Therefore, if normal cell clones with drivers are progenitors of endometrial cancers (which is plausible given the similar driver mutations found), our results suggest that many cancers are initiated during childhood and evolution to malignancy takes place over the lifetime of an individual. This perspective on the long duration of neoplastic evolution of invasive endometrial cancer has resonance with previous observations on leukaemia[40,41] and, more recently, other solid malignancies[42–45], and may be a common feature of the development of human cancers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-020-2214-z.

1. Gargett, C. E., Schwab, K. E. & Deane, J. A. Endometrial stem/progenitor cells: the first 10 years. Hum. Reprod. Update 22, 137–163 (2016).
2. Kaitu'u-Lino, T. J., Ye, L. & Gargett, C. E. Reepithelialization of the uterine surface arises from endometrial glands: evidence from a functional mouse model of breakdown and repair. Endocrinology 151, 3386–3395 (2010).
3. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
4. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. Nature 578, 94–101 (2020).
5. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. Nature 458, 719–724 (2009).
6. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264 (2016).
7. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. Nature 561, 473–478 (2018).
8. Franco, I. et al. Somatic mutagenesis in satellite cells associates with human skeletal muscle aging. Nat. Commun. 9, 800 (2018).
9. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 348, 880–886 (2015).
10. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. Science 362, 911–917 (2018).
11. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature 565, 312–317 (2019).
12. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. Cell 177, 608–621 (2019).
13. Suda, K. et al. Clonal expansion and diversification of cancer-associated mutations in endometriosis and normal endometrium. Cell Rep. 24, 1777–1789 (2018).
14. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature 574, 532–537 (2019).
15. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 574, 538–542 (2019).
16. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl Acad. Sci. USA 109, 14508–14513 (2012).
17. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. Proc. Natl Acad. Sci. USA 113, 9846–9851 (2016).
18. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. Science 359, 555–559 (2018).
19. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. 371, 2488–2498 (2014).
20. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487 (2014).
21. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. 20, 1472–1478 (2014).
22. Tempest, N., Maclean, A. & Hapangama, D. K. Endometrial stem cell markers: current concepts and unresolved questions. Int. J. Mol. Sci. 19, E3240 (2018).
23. CRUK. Uterine cancer risk. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/risk-factors#heading-Zero (accessed 28 March 2020).
24. Morice, P., Leary, A., Creutzberg, C., Abu-Rustum, N. & Darai, E. Endometrial cancer. Lancet 387, 1094–1108 (2016).
25. Le Gallo, M. & Bell, D. W. The emerging genomic landscape of endometrial cancer. Clin. Chem. 60, 98–110 (2014).
26. Onstad, M. A., Schmandt, R. E. & Lu, K. H. Addressing the role of obesity in endometrial cancer risk, prevention, and treatment. J. Clin. Oncol. 34, 4225–4230 (2016).
27. Setiawan, V. W. et al. Type I and II endometrial cancers: have they different risk factors? J. Clin. Oncol. 31, 2607–2618 (2013).
28. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. Cell 171, 1029–1041 (2017).
29. The Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. Nature 497, 67–73 (2013).
30. Anglesio, M. S. et al. Cancer-associated mutations in endometriosis without cancer. N. Engl. J. Med. 376, 1835–1848 (2017).
31. Lac, V. et al. Iatrogenic endometriosis harbors somatic cancer-driver mutations. Hum. Reprod. 34, 69–78 (2019).
32. Lac, V. et al. Oncogenic mutations in histologically normal endometrium: the new normal? J. Pathol. 249, 173–181 (2019).
33. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. Science 330, 822–825 (2010).
34. Barker, N. et al. Lgr5^{+ve} stem cells drive self-renewal in the stomach and build long-lived gastric units in vitro. Cell Stem Cell 6, 25–36 (2010).
35. Rouhani, F. J. et al. Mutational history of a human cell lineage from somatic to induced pluripotent stem cells. PLoS Genet. 12, e1005932 (2016).
36. Nicholson, A. M. et al. Fixation and spread of somatic mutations in adult human colonic epithelium. Cell Stem Cell 22, 909–918 (2018).
37. Zhang, Y. et al. A pan-cancer compendium of genes deregulated by somatic genomic rearrangement across more than 1,400 cases. Cell Rep. 24, 515–527 (2018).
38. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. Nature 556, 457–462 (2018).
39. Wu, Q. J. et al. Parity and endometrial cancer risk: a meta-analysis of epidemiological studies. Sci. Rep. 5, 14243 (2015).
40. Greaves, M. In utero origins of childhood leukaemia. Early Hum. Dev. 81, 123–129 (2005).
41. Greaves, M. Pre-natal origins of childhood leukemia. Rev. Clin. Exp. Hematol. 7, 233–245 (2003).
42. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. Cell 173, 611–623 (2018).
43. Anderson, N. D. et al. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. Science 361, eaam8419 (2018).
44. Maura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. Nat. Commun. 10, 3835 (2019).
45. Gerstung, M. et al. The evolutionary history of 2,658 cancers. Nature 578, 122–128 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix 20

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-020-00437-6

Check for updates

Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing

Peter Ellis[1,3,4], Luiza Moore[1,4], Mathijs A. Sanders[1,2,4], Timothy M. Butler[1,4], Simon F. Brunner[1], Henry Lee-Six[1], Robert Osborne[1,3], Ben Farr[1], Tim H. H. Coorens[1], Andrew R. J. Lawson[1], Alex Cagan[1], Mike R. Stratton[1], Inigo Martincorena[1] and Peter J. Campbell[1]✉

Somatic mutations accumulate in healthy tissues as we age, giving rise to cancer and potentially contributing to ageing. To study somatic mutations in non-neoplastic tissues, we developed a series of protocols to sequence the genomes of small populations of cells isolated from histological sections. Here, we describe a complete workflow that combines laser-capture microdissection (LCM) with low-input genome sequencing, while circumventing the use of whole-genome amplification (WGA). The protocol is subdivided broadly into four steps: tissue processing, LCM, low-input library generation and mutation calling and filtering. The tissue processing and LCM steps are provided as general guidelines that might require tailoring based on the specific requirements of the study at hand. Our protocol for low-input library generation uses enzymatic rather than acoustic fragmentation to generate WGA-free whole-genome libraries. Finally, the mutation calling and filtering strategy has been adapted from previously published protocols to account for artifacts introduced via library creation. To date, we have used this workflow to perform targeted and whole-genome sequencing of small populations of cells (typically 100–1,000 cells) in thousands of microbiopsies from a wide range of human tissues. The low-input DNA protocol is designed to be compatible with liquid handling platforms and make use of equipment and expertise standard to any core sequencing facility. However, obtaining low-input DNA material via LCM requires specialized equipment and expertise. The entire protocol from tissue reception through whole-genome library generation can be accomplished in as little as 1 week, although 2–3 weeks would be a more typical turnaround time.

Introduction

Normal and cancerous tissues are complex ecosystems comprising different cell populations with distinct morphologies, functional properties and spatial arrangements. Single-cell DNA sequencing technologies provide insights into the genomic landscapes of tumor and normal tissues[1]. However, these approaches remain suboptimal for identifying somatic mutations in normal cells as a consequence of incomplete genome coverage, allelic dropout and amplification-induced errors[1]. An alternative to single-cell sequencing is to expand single cells into colonies or clonal organoids in vitro, providing sufficient material to use standard genome sequencing approaches to investigate the genomes of individual cancer and normal stem cells[2,3]. Although these models have the advantage of providing high-quality genome data derived from a single cell, they are challenging to derive for certain tissues, might show biases toward particular cell types or toward cells with or without driver mutations and can be afflicted by additional mutational processes activated during cell culture[4]. Both single-cell sequencing and colony/organoid sequencing as approaches lose information about the spatial distribution and histopathological features of mutant cells within a tissue.

We have established a method based on LCM that allows whole-genome sequencing (WGS) of small, often clonal, cell populations for which precise phenotypic and spatial information is preserved[5–8]. Identifying somatic mutations in normal tissues is more challenging than sequencing tumors, as normal tissues are typically polyclonal, whereas cancers are monoclonally derived. Nonetheless, the limited mobility and steady cell turnover in many tissues, particularly epithelial cells,

[1]Cancer, Ageing and Somatic Mutation (CASM), Wellcome Sanger Institute, Hinxton, UK. [2]Department of Hematology, Erasmus University Medical Center, Rotterdam, the Netherlands. [3]Present address: Invivata Limited, The Glenn Berge Building, Babraham Research Campus, Babraham, UK. [4]These authors contributed equally: Peter Ellis, Luiza Moore, Mathijs A. Sanders, Timothy M. Butler. ✉e-mail: pc8@sanger.ac.uk

means that localized clonal patches do develop over time. The challenge for identifying mutations in normal somatic cells is to isolate and sequence DNA from microbiopsies not much larger than the size of these clonal patches—hence, the need for a robust, high-fidelity library production protocol, effective using only a few hundred cells.

Comparison with other methods

Our primary goal was to generate a DNA library construction workflow capable of processing low-input DNA to enable effective identification, isolation and lysis of small cell populations or tissue structures of interest (e.g., colonic crypts or gastric glands). Several library preparation techniques have been previously developed that enable interrogation of DNA from single cells, including degenerate oligonucleotide-primed polymerase chain reaction (DOP-PCR)[9], multiple displacement amplification (MDA)[10] and multiple annealing and looping-based amplification cycles (MALBAC)[11]. Although the protocol described here is not suitable for analysis of single cells, as a diploid human cell contains only 6.6 pg of DNA, our approach has some advantages over current single-cell sequencing techniques.

MDA and DOP-PCR both involve exponential amplification steps, and so small differences in amplification efficiency during early cycles can result in substantial over- and under-representation of loci in the final library, leading to allelic dropout rates on the order of 20%[1]. Our protocol also involves an exponential amplification step (PCR), but the increased amount of input material (100–1,000 cell equivalents rather than 1–2 copies of each locus) allows for fewer cycles of amplification, thus reducing the effect of variable amplification efficiency. MALBAC attempts to overcome this issue by introducing complementary sequences at the ends of mature amplicons, which protects them from further amplification by the formation of loop structures at intermediate temperatures, resulting in quasi-linear amplification. However, despite the increased uniformity of coverage afforded by MALBAC compared to other single-cell sequencing techniques, the false-positive rate for single-nucleotide variant (SNV) calls is very high compared to bulk data, with ~1.1×10^5 false-positive calls per genome[11]. As many of the samples we have sequenced have only 10^3–10^4 somatic SNVs per genome, the number of false-positive calls associated with MALBAC might be considerably larger than the number of true calls, making MALBAC best suited for detection of large-scale structural variants in single cells rather than accurate SNV calling.

Phasing of putative somatic SNVs with germline single-nucleotide polymorphisms (SNPs) has been used to improve quality of SNV calls from single-cell data[3]. However, this approach is useful only for the ~20% of SNVs that lie sufficiently close to germline SNPs. Therefore, no extant single-cell technique is able to achieve both genome-wide coverage and false-positive rates similar to bulk sequencing. By contrast, the technique described here achieves high, even coverage across the vast majority of the genome and has false-positive rates for SNVs below the somatic mutation burden in many tissues. Therefore, we recommend using this protocol for studies that prioritize these features. However, we note that single-cell approaches will still be necessary for highly polyclonal samples for which colonies from single cells cannot be derived.

Development and overview of the protocol

To overcome limitations associated with the aforementioned methods[9–11], we developed a robust, streamlined and high-throughput approach to generate whole-genome or targeted sequencing data from just a few hundred cells isolated from tissue sections[5–8], while avoiding the errors and biases introduced by WGA protocols, such as MDA or PicoPLEX[12]. The procedure consists of three major components: (i) effective tissue fixation, histology, LCM and cell lysis; (ii) genomic DNA isolation and library construction from limited DNA amounts; and (iii) variant discovery and filtering. Where possible, laboratory methods were developed to accommodate automation on robotic liquid handling platforms (e.g., the Agilent Bravo and Beckman NX platforms) and to align with next-generation sequencing (NGS) pipelines currently in operation. However, the workflow accommodates other automation platforms or could be performed manually with, for instance, multi-channel pipettes. The workflow from tissue preparation to DNA sequencing data is outlined in Fig. 1.

Tissue fixation is an essential step in histology to preserve tissue morphology for accurate microscopic assessment. However, standard histology fixatives, such as formalin, have a detrimental effect on both the quality and quantity of extracted DNA[13]. We, therefore, tested several non-crosslinking fixatives and discovered that alcohol-based preparations (ethanol, methanol or commercial alternatives, e.g., PAXgene Tissue FIX) were suitable for the proposed workflow. We routinely

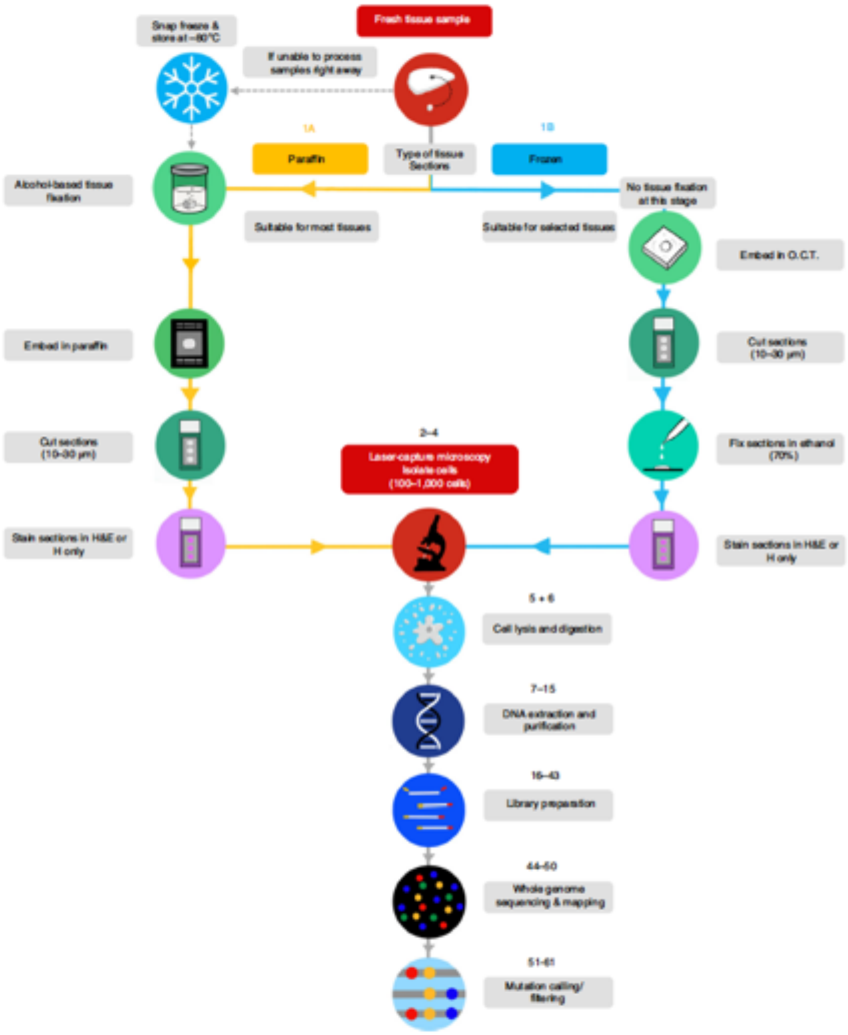

Fresh tissue sample
Snap freeze & store at −80°C
If unable to process samples right away
1A
Paraffin
Type of tissue Sections
1B
Frozen
Alcohol-based tissue fixation
Suitable for most tissues
Suitable for selected tissues
No tissue fixation at this stage
Embed in O.C.T.
Embed in paraffin
Cut sections (10–30 µm)
Cut sections (10–30 µm)
Fix sections in ethanol (70%)
2–4
Laser-capture microscopy Isolate cells (100–1,000 cells)
Stain sections in H&E or H only
Stain sections in H&E or H only
5 + 6
Cell lysis and digestion
7–15
DNA extraction and purification
16–43
Library preparation
44–50
Whole genome sequencing & mapping
51-61
Mutation calling/ filtering

Fig. 1 | Tissue processing and library preparation workflow. H, hematoxylin; H&E, hematoxylin and eosin; OCT, optimal cutting temperature.

use alcohol-based fixatives to prepare paraffin and frozen tissue sections and prefer paraffin embedding as it is compatible with most tissue types and results in high-quality morphology preservation. It should be noted that significant optimization has gone into the process of generating microbiopsies via LCM, extraction of DNA and construction of WGS libraries. Additional optimization of these steps could improve the success rate of the protocol. The tissue preparation and staining steps are included to document our approach, but it is by no means the only approach possible.

We tested several methods to maximize DNA recovery from microdissected cellular material. We discovered that proteinase K-based buffers work best within the proposed workflow and use either an

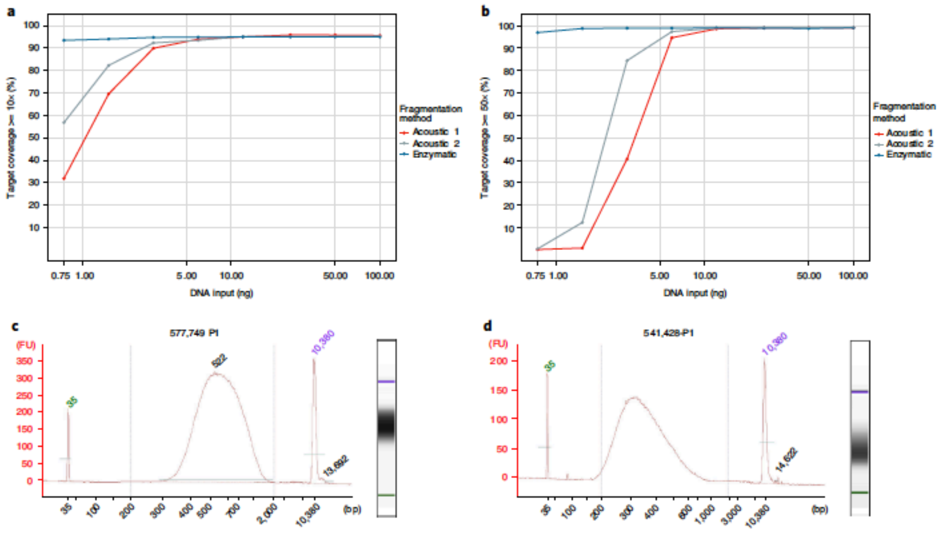


Fig. 2 | Comparison of DNA library performance in targeted sequencing workflows (Agilent SureSelect). We prepared human DNA libraries (mean insert length, -175 bp) using reagents supporting enzymatic or acoustic fragmentation methods. Equal amounts of each library were enriched for exome (**a**) or 2-Mb custom cancer panel (**b**) targets. Enriched libraries were subjected to 75-bp paired-end sequencing using the Illumina HiSeq 2500 platform. Data were normalized to -50 million (exome) or 30 million (custom cancer panel) reads per sample before analysis. **c**, Typical results from Agilent 2100 Bioanalyzer after low-input WGS library construction (Step 47). **d**, Typical results from Agilent 2100 Bioanalyzer after WGS library construction (Step 47) and subsequent targeted hybrid capture.

in-house version (described in 'Reagent setup') or the commercially available Arcturus PicoPure DNA Extraction Kit. The ability to bypass traditional DNA purification and quantification steps is a distinct feature of our proposed method. We propose a modified SPRI bead purification within the library construction workflow and omit DNA quantification altogether. Early tests indicated that genomic DNA recovery at the DNA purification step could be as low as 50% (unpublished results, P.E.), which led us to think that a large proportion of high-molecular-weight genomic DNA was refractory to elution from the SPRI beads. The entire post-elution sample (including beads) was integrated into the library construction workflow to avoid these losses. It is likely that a combination of buffer detergent, heat and action of the fragmentation enzymes in the next step promotes the release of bound DNA into solution.

Standard NGS workflows operational in our institute typically use 200 ng of input DNA material, often fragmented by acoustic shearing. Fragmented DNA is repaired, dA-tailed, ligated to adapter sequences and indexed by PCR amplification for six cycles. With the standard pipeline, our ability to produce sequencing data with meaningful library complexity drops dramatically when using less than 10 ng of input DNA. In developing the new protocol, we discovered that DNA fragmentation approaches that use enzymatic, rather than acoustic, fragmentation, yielded a >ten-fold improvement in DNA library yield. This increase in efficiency led to a dramatic reduction in PCR duplicate rates, enabling the generation of whole-exome or custom-targeted sequencing data from DNA inputs as low as 0.75 ng (Fig. 2). Importantly, we could readily control the mean fragment length for different applications independent of DNA input. This approach has, therefore, been implemented to generate DNA sequencing libraries from LCM microbiopsies.

A series of previously described post-processing filters were used to remove erroneous somatic variants[5–7,14]. However, we discovered that our low-input, enzymatic fragmentation-based LCM workflow also generated erroneous variants within inverted repeats capable of forming cruciform DNA (Fig. 3). Reads containing these erroneous variants had similar, but not identical, alignment start positions and could, therefore, not be marked as PCR duplicates. The erroneous variants often coincided with other erroneous variants in close proximity (1–30 bp) within the same read (Fig. 3a).

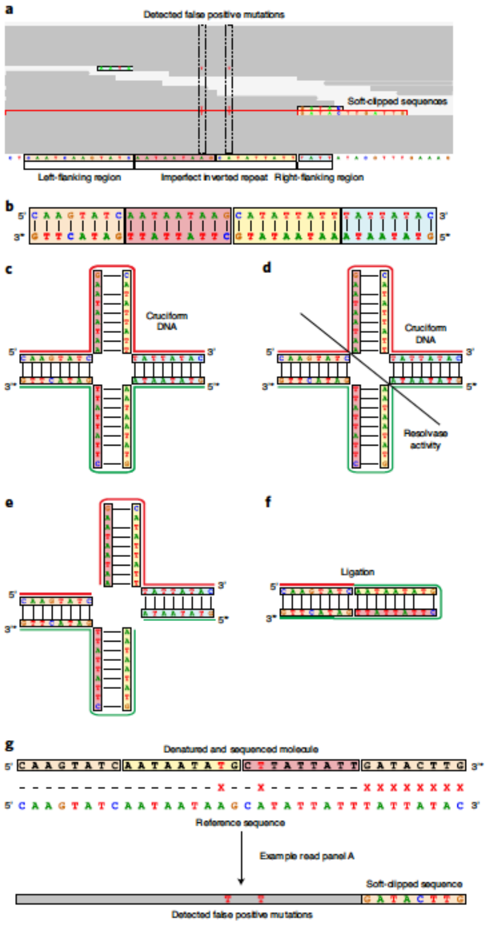

a
Detected false positive mutations
Soft-clipped sequences
Left-flanking region
Imperfect inverted repeat
Right-flanking region
b
5' C A A G T A T C A A T A A T A A G C A T A T T A T T T A T T A T A C 3'
3'' G T T C A T A G T T A T T A T T C G T A T A A T A A A T A A T A T G 5''
c
Cruciform DNA
5'
3'
3''
5''
d
Cruciform DNA
Resolvase activity
e
f
Ligation
g
Denatured and sequenced molecule
5' C A A G T A T C A A T A A T A T G C T T A T T A T T G A T A C T T G 3''
5' C A A G T A T C A A T A A T A A G C A T A T T A T T T A T T A T A C 3'
Reference sequence
Example read panel A
Soft-clipped sequence
T T
G A T A C T T G
Detected false positive mutations

Fig. 3 | Genesis of false-positive cruciform DNA-induced variants. a, Integrative Genomics Viewer (IGV) screenshot of reads containing false-positive variants calls (red Ts bracketed by dashed lines) typically introduced by the erroneous processing of cruciform DNA. (Red and Yellow) Two sides of the semi-inverted repeats with mismatches upon the formation of a hairpin. (Orange) Left flanking site of the imperfect inverted repeats. (Blue) Right flanking site of the imperfect inverted repeats. **b**, A DNA fragment containing the imperfect inverted repeats. **c**, Before or during DNA fragmentation, cruciform DNA is formed from two inverted repeats present in both strands. **d**, **e**, Resolvase activity cuts across cruciform structure. **f**, Ligation of the resolved hairpin. **g**, The hairpin is effectively transferred from one strand to the other. The processed DNA fragment bears a similar sequence to the original DNA fragment (**b**) with a few mismatches and, depending on the location of the single-strand nick, a small piece of ectopic sequence that derives from the opposite strand (reverse complement left flanking region, **a** and **g**).

Erroneous processing of cruciform DNA, either existing before DNA isolation or formed during library construction, is the most likely explanation for these artifactual variants (Fig. 3b–d). Reads containing these false-positive variants tended to align in close proximity to one another, which served as a hallmark for their detection. We used the variant position in the read with respect to the alignment start, the standard deviation (s.d.) and the median absolute deviation (MAD) of the variant

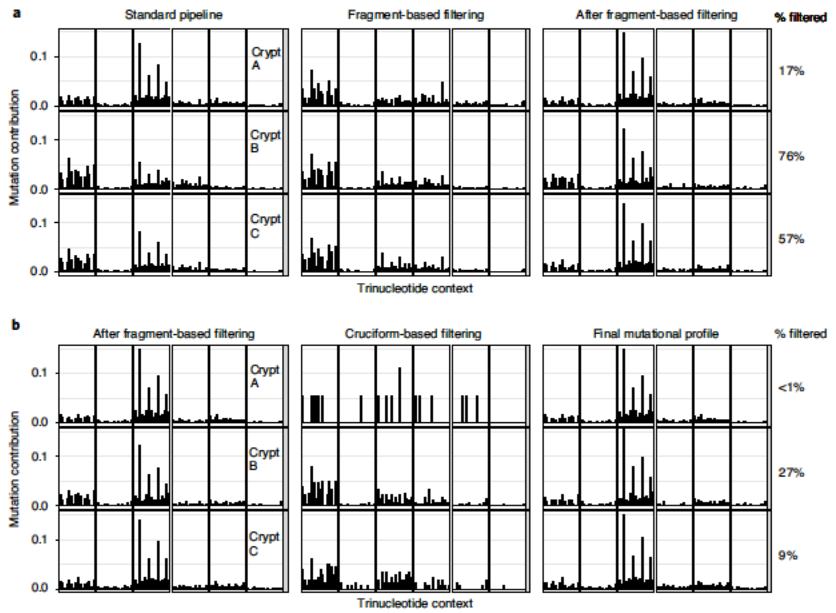


Fig. 4 | Effects of the different filtering steps. a, Fragment-based filtering is applied on three LCM small bowel microbiopsy samples. Left panel, Mutational spectra after CaVEMan variant detection and application of the standard filtering practices (Protocol Steps 51–55). Middle panel, Mutational spectra of the variants that get discarded by the fragment-based filter (Steps 56–58). Right panel, Mutational spectra of the variants retained after fragment-based filtering (Steps 56–58). Right, The percentage of variants filtered by fragment-based filtering. **b**, Cruciform-based filtering is applied to the variants retained after fragment-based filtering for the same three LCM small bowel microbiopsy samples. Right panel, Mutational spectra of variants retained after fragment-based filtering (Steps 56–58). Middle panel, Mutational spectra of variants that get discarded by the cruciform-based filter (Steps 59 and 60). Right panel, The final mutational spectra after fragment-based and cruciform-based filtering (final spectrum after Step 61). Right, The percentage of variants that get filtered by the cruciform-based filtering after applying the fragment-based filter first. Color indicates mutation type: blue - C>A, black - C>G, red - C>T, gray - T>A, green - T>C, pink - T>G. Trinucleotide context indicates the context in which the mutation occurred identified by the base preceding the mutation, the mutated base and the base succeeding the mutation.

position within the read as features for filtering. Application of these filters to bulk tissue WGS data showed that our new filtering approach removed only a few true somatic mutations per genome, while specifically removing erroneous variants associated with low-input, enzymatic fragmentation-based LCM experiments (Fig. 4).

We performed a set of validation experiments to test the developed workflow. First, the reproducibility of the workflow was assessed by generating pairs of biological 'near-replicate' samples and processing them independently using our new library construction methodology. In these experiments, two separate samples were generated from the same tissue structure, such as an appendiceal crypt, and subjected to independent DNA extraction, cell lysis, library preparation and WGS (Fig. 5a–d). Comparisons of somatic SNVs identified in each 'near-replicate' showed similar variant allele frequency (VAF) distributions (Fig. 5b), a high degree of overlap for SNVs (Fig. 5c) and similar substitution mutational spectra (Fig. 5d).

We next compared WGS data generated by our new end-to-end workflow to LCM lysates processed via traditional acoustic shearing methods. Similarly, pairs of biological 'near-replicate' samples were derived from the same histological structure; this time, one sample was processed with our new workflow and the other with acoustic shearing. Again, comparison of the WGS data between the two differently processed samples showed similar VAF distributions, SNVs and mutational spectra (Fig. 5e–h).

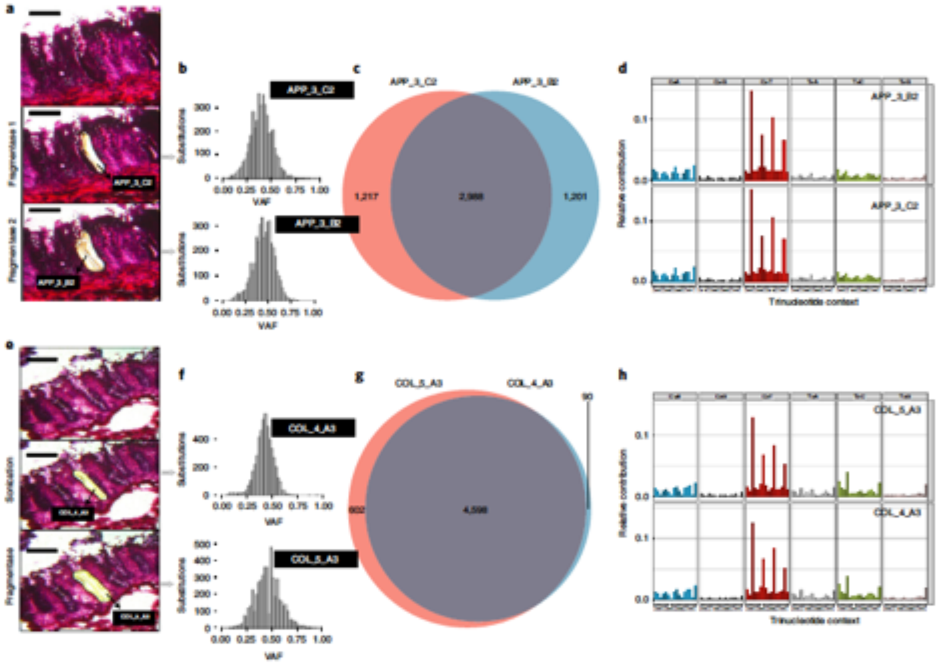


Fig. 5 | Validation experiments sequencing 'near-replicate' samples. a-d, 'Near-replicate' samples were generated by splitting an appendiceal crypt into two halves, which were then processed and sequenced independently. **a**, Images acquired on the LCM of an appendiceal crypt before (upper), after one half was microdissected (middle) and after the other half was microdissected (lower). Scale bar, 200 μm. **b**, VAF of all substitutions in both halves show similar clonal distribution with a median VAF ~0.5. **c**, Venn diagram demonstrating SNV identity between both samples. **d**, Trinucleotide context of all substitutions are also similar. **e–h**, 'Near-replicate' samples were generated by splitting a colonic crypt into two halves, which were subsequently processed with our fragmentase-based method (COL_5_A3) and sonication-based method (COL_4_A3). **e**, Images acquired on the LCM of a colonic crypt before (upper), after one half was microdissected (middle) and after the other half was microdissected (lower). Scale bar, 200 μm. Similar clonal VAF distributions (**f**), SNV calls (**g**) and trinucleotide context (**h**) are observed from the two samples.

Advantages and limitations of the protocol

Using our protocol, we obtained sufficient DNA from small cell populations for accurate WGS data, while circumventing the artifacts typically observed with single-cell WGA. Sequencing data from these experiments have already provided important insights into somatic mutations present in adult stem cells and their consequent clonal expansions[5–7].

Where possible, protocols were developed to support high-throughput, automated pipelines, and we aimed to minimize the number of steps from tissue preparation to variant detection. To date, we have successfully processed more than 40,000 LCM microbiopsies across ~550 96-well plates. Successful whole-genome libraries (library concentration >5 ng/ul) have been generated from 80% of microbiopsies (Fig. 6a,b). Failures are generally attributable to unsuccessful placement of LCM material in the collection vessel before lysis, resulting in a negative well or an insufficient microbiopsy size/cell count. We think that a success rate of ~80% is an acceptable level for this protocol. Our tradeoff here is between microdissecting a sufficient number of cells for which we can successfully make a sequencing library versus microdissecting sufficiently few cells as to minimize the number of clonal structures sampled. We can assess the quality of a given library before sequencing it, which means that we can cost-effectively tolerate a 20% library failure rate.

The minimum number of cells required for generating sufficient genome coverage is an obvious limitation to our protocol. Reference genomic DNA (gDNA) was used to formulate pass/fail criteria,

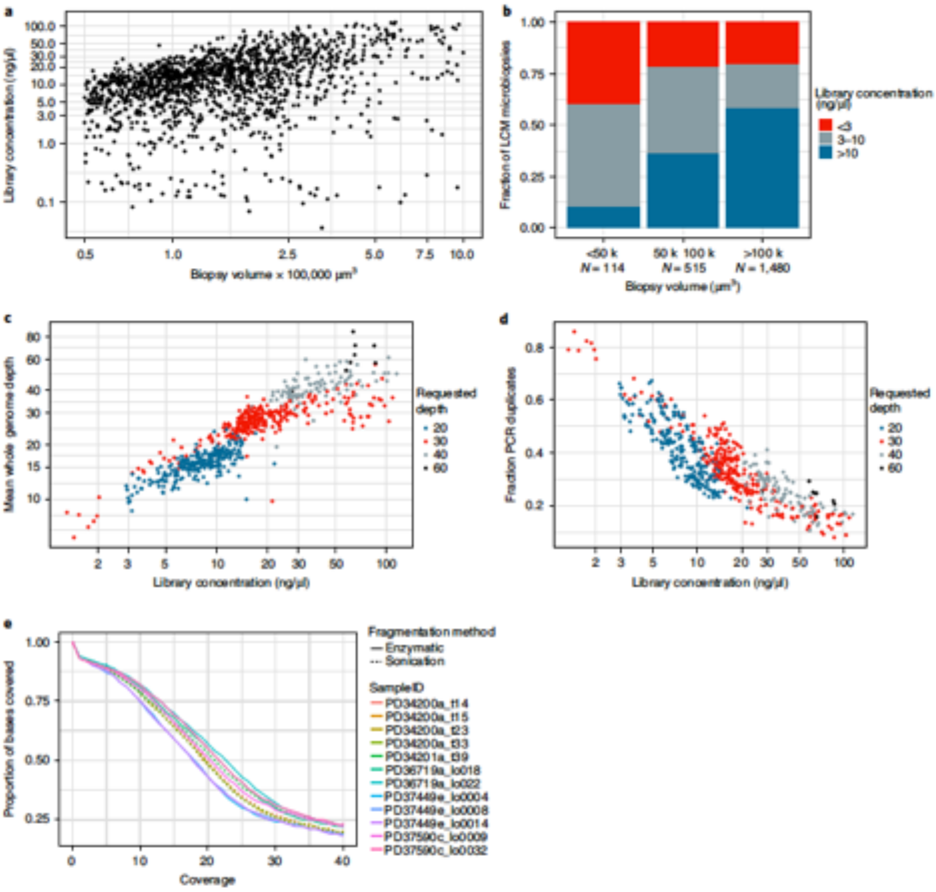


Fig. 6 | Library concentrations of epithelial LCM samples. a, Post-PCR library concentration of 1,721 LCM microbiopises of breast epithelial tissue. **b**, Fraction of breast microbiopsies meeting varying library concentration thresholds. **c**, Mean WGS depth achieved for 526 breast and lung epithelial microbiopsies across a range of library concentrations. Requested sequencing depth was varied to avoid unnecessary sequencing of PCR duplicates. **d**, Fraction of reads identified as PCR duplicates. **e**, Proportion of GRCh37 covered at a given depth, comparing colonic crypts that underwent low-input library generation via either sonication or enzymatic fragmentation.

and the DNA library yield from samples passing these criteria was subsequently used to predict WGS data quality based on previously sequenced standard WGS libraries (Table 1). We estimate a minimum requirement of 30–100 cell equivalents of DNA (180–600 pg) to obtain a coverage of 10–20× reads for WGS (Figs. 2, 6). This translates to an efficiency of 5–33% of all DNA molecules being successfully converted to WGS data. We routinely captured 100–1,000 cells, depending on the tissue histology of the sample and were, therefore, unaffected by this constraint in the setting of WGS. Tissue microanatomic structures or clonal expansions can permeate multiple tissue sections. In this context, depending on the specifics of the research question, it could be beneficial to sample the same clonal structure from serial tissue sections to increase cell numbers. Processing new tissue types usually involves cutting LCM microbiopsies from a wide range of sizes to identify the threshold that yields a desired final library concentration. Typically, microbiopsy volume more often governs LCM cutting than cell count, as the cutting area is quickly and automatically calculated by the LCM

Table 1 | Identifying pass/fail criteria for WGS based on DNA library yield

Cell equivalents	PCR cycles	Library yield (fmol)	Theoretical yield (fmol)	q20 data (Gb)	SNP coverage	Reads mapped (%)	Duplicate reads (%)
4,000	8	2,670	42,720	96.0	31	99.6	5.8
1,000	10	2,435	9,740	100.7	34	99.6	7.7
250	12	2,362	2,362	95.2	29	99.6	17.2
60	14	2,348	587	93.5	21	99.2	39.3
30	14	1,240	310	85.3	12	98.2	54.8

Human genomic DNA (Horizon Discovery Tru-Q 6 Refernce Standard DNA) was subjected to DNA library construction using our LCM-based NGS workflow. DNA libraries were sequenced using the Illumina HiSeq 4000 platform. Depth of coverage is indicated by averaging call depths across 26 common SNPs. Theoretical yield indicates the amount of library produced if our standard 12 cycles of PCR were used and assumes linear response and 100% PCR efficiency.

software. The typical minimum input volume for epithelial tissues is 100,000 µm^3, with library concentrations >3 ng/µl in over 75% of samples (Fig. 6a,b). Alcohol-based tissue fixation is another protocol limitation, with an option for formalin-fixed tissues currently lacking.

Our workflow is not limited to LCM-based experiments, and low-input material from multiple sources has been successfully processed by our workflow. The two primary reasons to use this workflow on other sources are: (i) we have demonstrated the ability to produce high-quality DNA sequencing data from very limited amounts of DNA, and (ii) the workflow bypasses unnecessary DNA isolation and quantification steps. We have successfully produced whole-genome or targeted sequencing data from esophageal and skin biopsies, organoids, fluorescence-activated cell-sorted lymphocytes and parasitic blood flukes (*Schistosoma*) (unpublished results, L.M. and P.E.). Although our workflow offers a powerful approach for sequencing these sample types, DNA quantity must be considered to avoid overwhelming the limited enzymatic reagents. We recommend an input of 100–1,000 cells for our workflow, which is relatively simple to control during capture. Working within this range enables the running of a standardized workflow that processes all samples in a 96-well plate under identical conditions. For new tissue types, we recommend running a pilot experiment, measuring the DNA library yields and using these values to infer a DNA amount typical for the starting material. Sample cohorts should be adjusted to achieve library concentrations greater than 10 ng/µl, or, for some cases, we reduce the number of library amplification cycles. A threshold of 10 ng/µl allows for WGS read depth of ~20×, whereas a lower threshold of 3 ng/µl can be used to generate 10× read depth, for situations where that depth is sufficient (Fig. 6c,d).

Several potential concerns were investigated and found not to be an issue. Using an enzymatic rather than acoustic fragmentation could potentially lead to uneven genomic coverage. However, by comparing sequencing data from either sonicated or enzymatic fragmented DNA from colonic crypts, we could not find any difference in genome coverage (Fig. 6e and Supplementary Data 1). Another concern relates to microbiopsy sizes, which might result in the differential incidence of any artifacts. We took two microbiopsies from the same breast tumor sample that differed in size tenfold (Fig. 7a). Unsurprisingly, the larger cut was able to be sequenced to a greater median whole-genome depth (33× versus 21×), at the expense of a lower median VAF (0.36 versus 0.44) (Fig. 7b,c). Despite these differences, the two samples had a large proportion of shared mutations and no evidence of different mutational patterns caused by a size-related artifact (Fig. 7d–f).

Experimental design

There are four major experimental design factors that require consideration before applying this method: input material, germline filtering strategy, expected tissue clonality and sequencing requirements.

Input material

Our protocol was specifically designed to be used with LCM-derived low-input material, but the low-input library preparation protocol (Steps 7–49) has been successfully applied in non-LCM settings. Studies investigating blood colonies[14,15] and single-cell-derived organoids[16] have successfully been used in the low-input pipeline to generate high-quality whole-genome libraries. However, optimization of the required DNA input amount followed by the necessary number of PCR cycles is

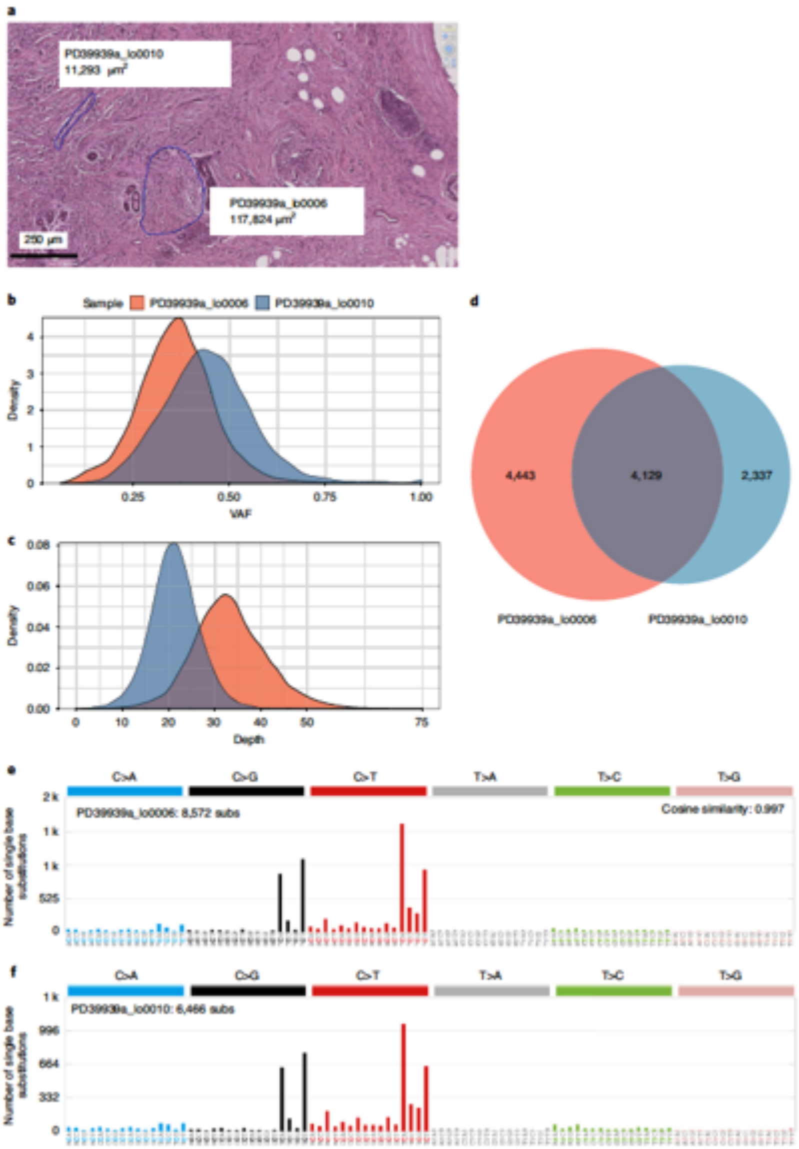


Fig. 7 | Variation of microbiopsy size in samples from the same breast tumor. a, Annotated slide-scanner image showing two regions of the same lobular carcinoma breast tumor that were used to assess the effect of varying microbiopsy size (blue) and sampled regions not included in the analysis (white). **b**, Density plot of VAF of SNVs. **c**, Density plot of sequencing depth of SNVs. **d**, Venn diagram showing overlap of SNV calls between the two samples. **e**, **f**, Mutational spectra of SNV calls.

advised when DNA is extracted through alternative approaches. Two factors are considered pivotal for LCM experiments and require further tailoring to achieve the optimal outcome: the thickness of tissue sections (thinner sections provide better histology but yield less DNA) and the volume of tissue to cut (tissues have varying cellular densities; therefore, larger cuts increase the risk of capturing more unrelated clones, making mutation detection more difficult). Increasing the DNA input amount can yield a more complex sequencing library, enabling deeper sequencing coverage, which might be necessary for certain applications.

Germline filtering strategy

Germline filtering is important for identifying true somatic over germline variants. We used three different strategies, depending on study design and tissue availability:

The 'matched bulk normal' approach. A bulk DNA sample derived from a different tissue than what is being microdissected (typically a blood sample) undergoes traditional WGS. This approach provides a high-quality normal sequence from which to identify germline variants but requires the collection of multiple tissues from a single patient. Furthermore, the input DNA amount and library preparation methods are fundamentally different, potentially leading to a mismatch in artifacts present in the matched-normal and LCM-derived samples.

The 'matched LCM normal' approach. An LCM microbiopsy is cut from the same tissue but comprises unrelated cell types, ideally from polyclonal tissues, to avoid clonal mutations or copy number changes in the matched normal. Examples include lymph nodes or areas of stroma or smooth muscle. Such microbiopsies are typically large (~1,000 cells) to ensure a high-depth library. The matched LCM normal is subsequently used in mutation calling similar to the bulk normal, with the added benefit of being subjected to the exact same library preparation conditions. This approach will not always be possible depending on the cell types present in a tissue section and the architecture of the tissue of interest.

The 'unmatched' approach. Variant calling can be performed against an artificial genome generated from the reference sequence, and all variant sites are aggregated across all samples from the same individual. When analyzing multiple largely unrelated microbiopsies from the same individual, germline variants can be conservatively removed by calculating the global allele frequency (VAF) of a mutant site across all available samples from the same individual. Germline mutations will be present at global allele frequencies ~0.5 (heterozygous mutations) or 1 (homozygous mutations). Removal of germline mutations can be done using a global VAF cutoff or a binomial test (testing for global VAF <0.5). Finally, a beta-binomial test is run, filtering out low-frequency artifacts that, unlike genuine somatic mutations, are often present at similarly low frequencies across multiple libraries from a given individual. An advantage of the 'unmatched' approach when multiple unrelated samples are available from a patient is that no second source of tissue is required and that early embryonic variants will not be filtered out by being present in a matched normal. This approach improves as the aggregate coverage increases and requires at least a total of 50-fold coverage per individual.

Tissue clonality

Levels of clonality are largely dependent on tissue histology and the microenvironment, which should be considered in the experimental design. Clonal entities, where all cells are derived from a single progenitor, will have somatic mutations shared across all cells within the histological structure and predicted VAFs at 0.5. Examples of these are colonic crypts, endometrial glands and prostate glands. For these examples, high-confidence somatic mutations can be identified from samples sequenced to a moderate depth (15×). Other tissues show a more disordered histological structure (sheets of epithelial cells in the bronchus, esophagus and bladder, for example), or the histological structure is derived from multiple progenitors or cell types and, therefore, is not clonal (breast acini and pancreatic islets). Mutations in these microbiopsies are typically shared by a subset of cells; VAFs in a range of 0.2–0.4 are often seen in microbiopsies comprising two or three distinct cell populations ('oligoclonal' samples). Truly polyclonal samples can show lower VAFs or even yield no detectable somatic mutations. In tissues with no clearly defined structure, it is not uncommon to see a wide range VAFs depending on the number of clones present in each microbiopsy and the degree of contamination by other cell types. One approach that can be useful to obtain high-quality whole genomes from tissues with oligoclonal microbiopsies is to screen 2–3 times as many microbiopsies using targeted or whole-exome sequencing as you intend to perform WGS on. Use the mutations detected to identify libraries with sufficiently high VAFs (e.g. >0.25), or interesting driver mutations, for subsequent WGS. We do this by generating WGS libraries on all microbiopsies and performing targeted or exome sequencing on a fraction of the available amount of library. A small pilot experiment to determine the optimal sampling strategy is advised when the clonal architecture of a tissue is unknown a priori before initiating a larger study. Using this protocol, ~80% of samples pass our predefined quality controls, but this percentage could be further increased by taking larger microbiopsies, with the added risk of sampling more individual clones.

Sequencing considerations

The protocol as outlined here primarily focusses on the use of LCM in combination with WGS to identify structural variants (SVs), copy number alterations, SNVs and small insertions and deletions (indels). The large number of somatic mutations identified allows for building of robust phylogenetic trees describing the relationships between the samples and enables the identification of mutational signatures present in the samples. We have successfully sequenced low-input LCM libraries to >100× depth using several small (<5 megabase) targeted panels[17,18] or a whole-exome panel, although the maximum achievable coverage can be lower in libraries from small numbers of cells. For these targeted samples, we first generated whole-genome libraries and subjected a subset of that library to targeted pulldown, saving the remainder for potential follow-up WGS. Due to increased median insert size of the whole-genome libraries, this approach can yield more off-target reads in targeted sequencing than is typical (>30%).

Future development

We have developed a robust set of protocols for sequencing the genomes of discrete cell populations. In addition, we are interested in integrating this workflow with multi-omic sequencing approaches, including DNA, RNA and methylation analyses performed on the same microbiopsies. To date, we have successfully generated Smart-Seq2 (ref. [19]) RNA sequencing libraries from PAXgene-fixed tissues; however, additional optimization steps are currently under development to ensure consistent RNA yields and high-quality sequencing data.

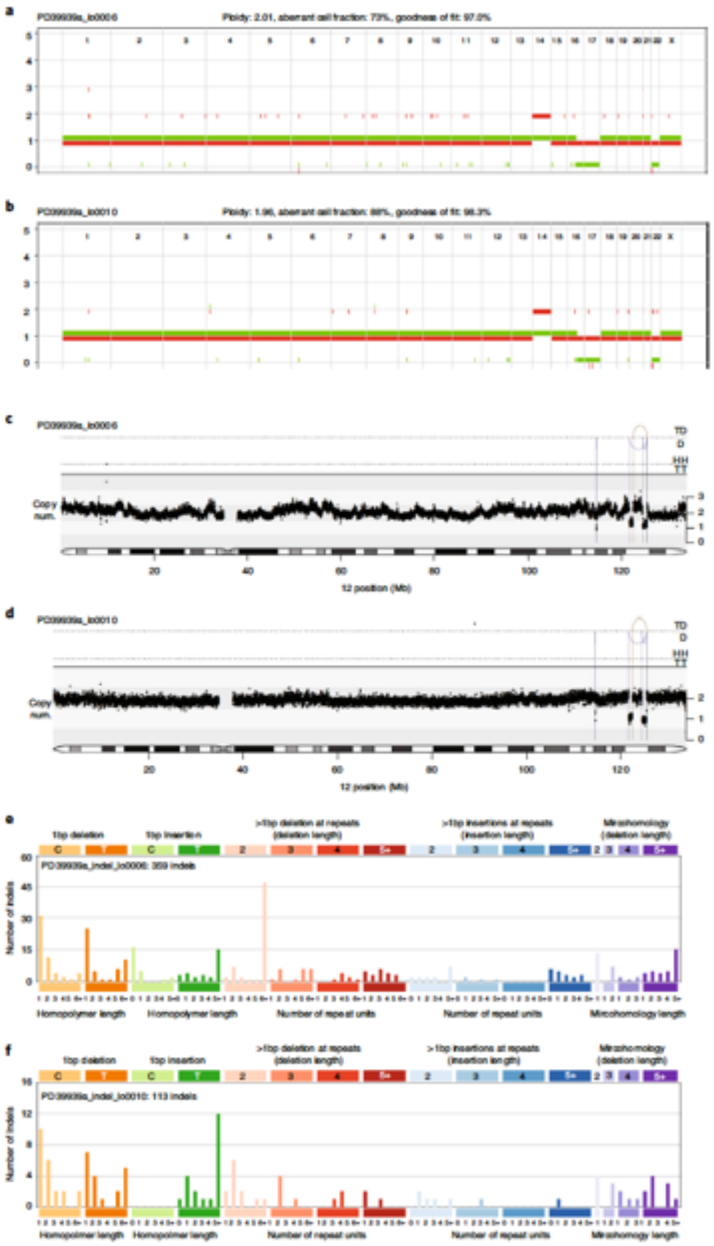
In the context of oligoclonal or polyclonal tissues, low-frequency variants could be reliably detected by using an error-corrected sequencing approach, such as 'duplex sequencing' or other unique molecular identifier approaches[20,21]. The additional sequencing cost would make this cost-prohibitive for WGS but would be highly effective for targeted gene panels. It is also possible that some of the observed artifacts would be distinguished by these error-corrected sequencing approaches.

Copy number variants, SVs and small indels were successfully identified from our WGS libraries using our standard WGS analysis tools (ASCAT, BRASS and Pindel, respectively; available at https://github.com/cancerit/dockstore-cgpwgs) (Fig. 8). The estimated purity from the copy number calls robustly correlates with the median VAF identified from the SNVs, and the large chromosomal gains and losses are identical between two samples taken from the same tumor (Figs. 7a,b and 8a,b). Similarly, the tandem duplication and deletion calls made on chromosome 12 are the same between both samples, despite a noisier copy number profile for sample PD39939a_lo0006 (Fig. 8c,d). Although this is encouraging, the noisiness of the data, particularly the indels, will require some additional filtering—an approach we are currently pursuing. Indels appear particularly error-prone on account of the additional PCR cycles introducing higher numbers of single-base indels, which can complicate separating true from artifactual mutations.

Materials

Biological materials

- This approach was successfully used on a variety of human tissues (including bladder, breast, colon, endometrium, lung and stomach) and colon and small intestine from mice.
- Sequencing data from colonic and appendiceal tissue used in Fig. 5 were obtained from a warm-body autopsy of a 78-year-old man. Tissue was processed by ethanol fixation and frozen sectioning (Protocol Step 1B). The samples were collected in line with the protocols approved by the NRES Committee East of England (NHS National Research Ethics Service reference 13/EE/0043).
- Sequencing data from breast tissue used in Figs. 7 and 8 were obtained from a mastectomy sample of an estrogen receptor-positive lobular carcinoma from a 70-year-old woman. Tissue was processed using PAXgene fixation and paraffin embedding (Protocol Step 1A). Ethics approval was obtained from the Wellcome Sanger Institute (UK) and the University of Queensland (Australia).


a
PD39939a_lo0006
Ploidy: 2.01, aberrant cell fraction: 73%, goodness of fit: 97.0%
b
PD39939a_lo0010
Ploidy: 1.96, aberrant cell fraction: 88%, goodness of fit: 98.3%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
c
PD39939a_lo0006
TD
D
HH
TT
Copy num.
12 position (Mb)
d
PD39939a_lo0010
e
1bp deletion
1bp insertion
>1bp deletion at repeats (deletion length)
>1bp insertions at repeats (insertion length)
Microhomology (deletion length)
PD39939a_indel_lo0006: 359 indels
Number of indels
Homopolymer length
Homopolymer length
Number of repeat units
Number of repeat units
Microhomology length
f
PD39939a_indel_lo0010: 113 indels

◀ **Fig. 8 | Copy-number variant, SV and small indel calling. a**, **b**, Allele-specific copy number analysis of tumors (ASCAT)[29]; copy number segments for two microbiopsies taken from the same breast tumor (Fig. 7a). **c**, **d**, SV identified on chromosome 12 using BReakpoin AnalySiS BRASS (https://github.com/cancerit/BRASS). D, deletion; HH, head-to-head inversion; TD, tandem duplication; TT, tail-to-tail inversion. **e**, **f**, Mutational spectrum for small indel calls made using Pindel[30].

! CAUTION All experiments on human and animal tissues must have ethics approval in accordance with governmental and institutional regulations. Informed consent must be obtained for all human samples. Human samples were processed to destruction.

Reagents

Tissue fixation, sectioning and staining

- Absolute ethanol (VWR International, cat. no. 1.08543.0250)
- PAXgene FIX Kit (PreAnalytiX, cat. no. 765312)
- PAXgene Tissue STABILIZER Concentrate (PreAnalytiX, cat. no. 765512)
- Water (VWR International, cat. no. 22934.K7)
- Phosphate-buffered saline (PBS; 1×)
- Gill's hematoxylin II (Leica, cat. no. 3801501)
- Aqueous eosin 1% (Leica, cat. no. 3801591)
- Xylene (VWR International, cat. no. 28975.325) **! CAUTION** Flammable and harmful; wear protective clothes and gloves; perform all procedures in a fume hood.
- Neo-Clear xylene substitute for microscopy (Sigma-Aldrich, cat. no. 1.09843) **! CAUTION** Flammable and harmful; wear protective clothes and gloves; perform all procedures in a fume hood.
- Poly-ethylene napthalate (PEN)-membrane slides (Arcturus, cat. no. LCM0522 or Leica, cat. no. 11600288) **▲CRITICAL** To minimize DNA cross-contamination, it is important to place all slides in a UV crosslinker (recommended time, 30 min at maximum power) before mounting of tissue sections.
- Optimal cutting temperature (OCT) compound (Thermo Fisher Scientific, cat. no. 23-730-625)
- Paraffin (congealing point, 55–58 °C; VWR International, cat. no. 361077E)

Cell lysis and digestion

▲CRITICAL This step can be performed using either a commercially available kit (Arcturus PicoPure DNA Extraction Kit; Thermo Fisher Scientific, cat. no. KIT0103) or our in-house protease buffer composed of the reagents below (for preparation, see 'Reagent setup').

- Proteinase-K (Qiagen, Protease, 7.5 AU, cat. no. 19155)
- Tris-HCl, pH 8.0 (Sigma-Aldrich, cat. no. RES3098T-B701X)
- Tween-20 (Sigma-Aldrich, cat. no. P1379-500ML)
- IGEPAL CA-630 (Sigma-Aldrich, cat. no. I8896-100ML)
- Nuclease-free water (Ambion, cat. no. AM9937)

Purification of gDNA from LCM lysates/purification of amplified libraries

- Eppendorf TwinTec PCR plates (Eppendorf, cat. no. 0030128648)
- TE buffer, pH 8.0 (Ambion, cat. no. AM9858)
- Nuclease-free water (Ambion, cat. no. AM9937)
- Agencourt AMPure XP Beads (Beckman Coulter, cat. no. A63881)
- Ethanol, 95–97% (vol/vol), AnalaR NORMAPUR analytical reagent (VWR International, cat. no. 20823.327).

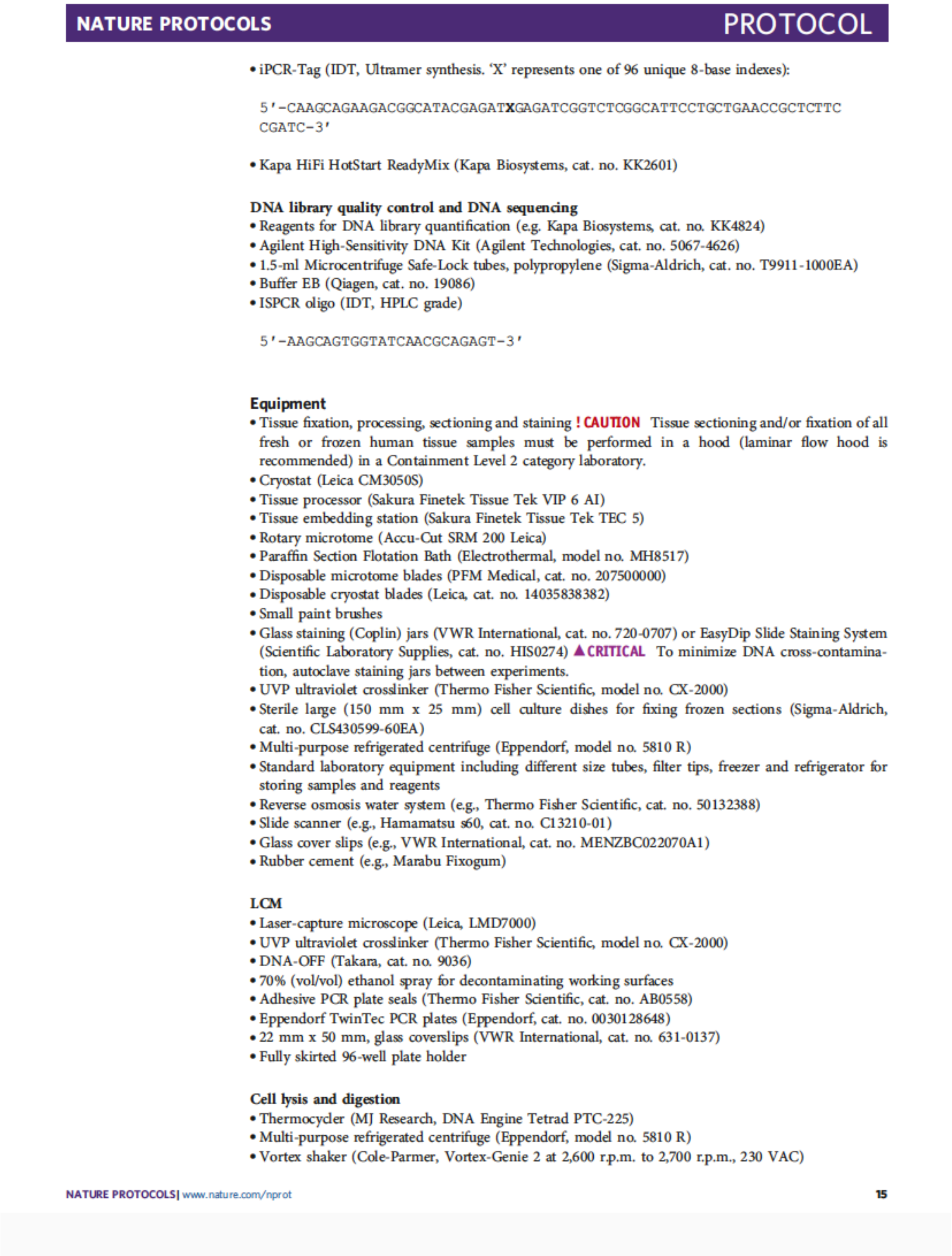
DNA library construction and amplification of adapter-ligated libraries

- NEBNext Ultra II FS DNA Library Prep Kit for Illumina (New England Biolabs, cat. no. E7805L)
- Duplexed adapters (IDT, HPLC grade, '*' represents phosphorothioate modification)

```
5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'
5'-phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'
```

- PE1.0 primer (IDT, Ultramer synthesis, standard desalt, '*' represents phosphorothioate modification):

```
5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3'
```



- iPCR-Tag (IDT, Ultramer synthesis. 'X' represents one of 96 unique 8-base indexes):

  5′-CAAGCAGAAGACGGCATACGAGAT**X**GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC CGATC-3′

- Kapa HiFi HotStart ReadyMix (Kapa Biosystems, cat. no. KK2601)

DNA library quality control and DNA sequencing

- Reagents for DNA library quantification (e.g. Kapa Biosystems, cat. no. KK4824)
- Agilent High-Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626)
- 1.5-ml Microcentrifuge Safe-Lock tubes, polypropylene (Sigma-Aldrich, cat. no. T9911-1000EA)
- Buffer EB (Qiagen, cat. no. 19086)
- ISPCR oligo (IDT, HPLC grade)

  5′-AAGCAGTGGTATCAACGCAGAGT-3′

Equipment

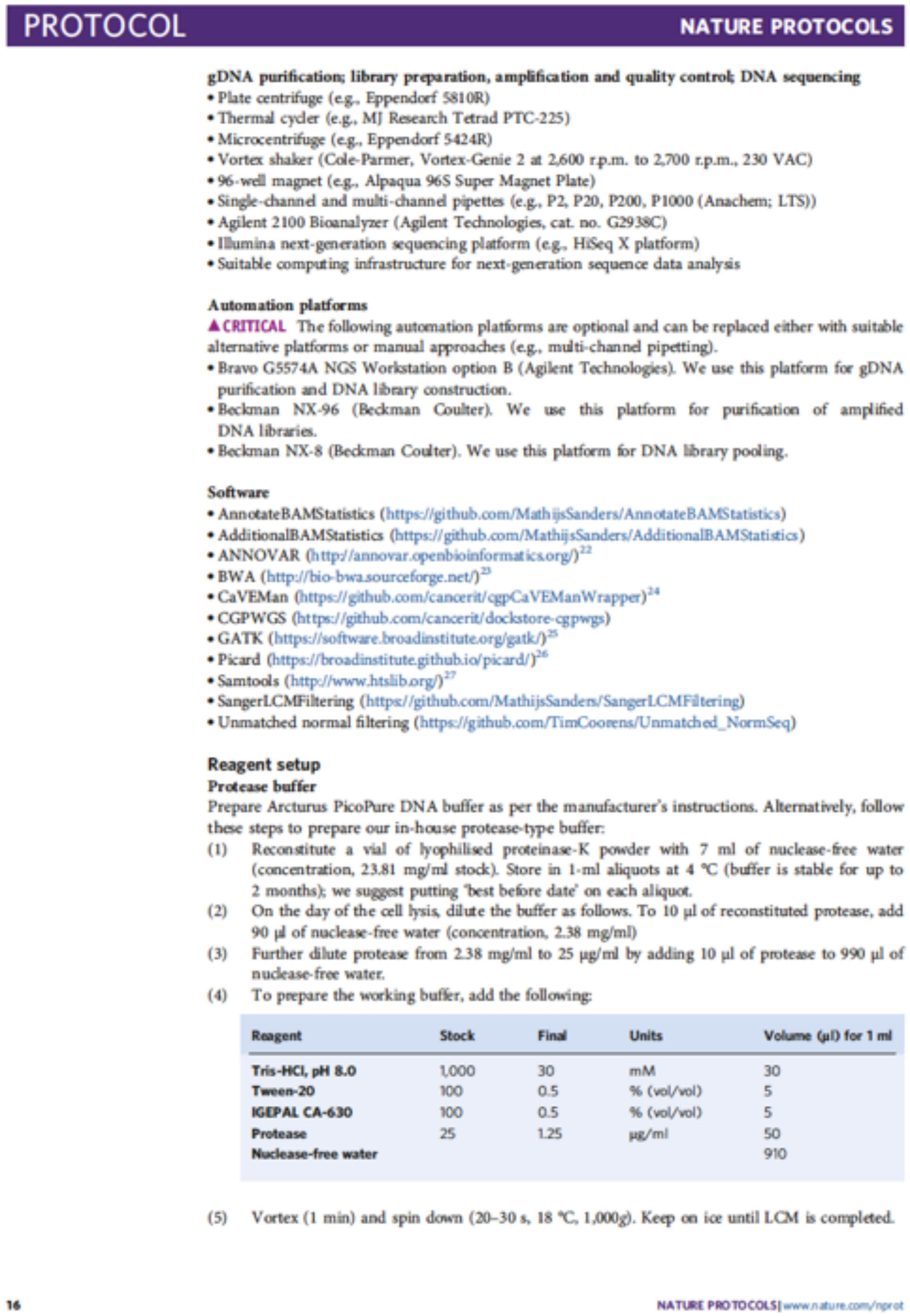
- Tissue fixation, processing, sectioning and staining **! CAUTION** Tissue sectioning and/or fixation of all fresh or frozen human tissue samples must be performed in a hood (laminar flow hood is recommended) in a Containment Level 2 category laboratory.
- Cryostat (Leica CM3050S)
- Tissue processor (Sakura Finetek Tissue Tek VIP 6 AI)
- Tissue embedding station (Sakura Finetek Tissue Tek TEC 5)
- Rotary microtome (Accu-Cut SRM 200 Leica)
- Paraffin Section Flotation Bath (Electrothermal, model no. MH8517)
- Disposable microtome blades (PFM Medical, cat. no. 207500000)
- Disposable cryostat blades (Leica, cat. no. 14035838382)
- Small paint brushes
- Glass staining (Coplin) jars (VWR International, cat. no. 720-0707) or EasyDip Slide Staining System (Scientific Laboratory Supplies, cat. no. HIS0274) **▲CRITICAL** To minimize DNA cross-contamination, autoclave staining jars between experiments.
- UVP ultraviolet crosslinker (Thermo Fisher Scientific, model no. CX-2000)
- Sterile large (150 mm x 25 mm) cell culture dishes for fixing frozen sections (Sigma-Aldrich, cat. no. CLS430599-60EA)
- Multi-purpose refrigerated centrifuge (Eppendorf, model no. 5810 R)
- Standard laboratory equipment including different size tubes, filter tips, freezer and refrigerator for storing samples and reagents
- Reverse osmosis water system (e.g., Thermo Fisher Scientific, cat. no. 50132388)
- Slide scanner (e.g., Hamamatsu s60, cat. no. C13210-01)
- Glass cover slips (e.g., VWR International, cat. no. MENZBC022070A1)
- Rubber cement (e.g., Marabu Fixogum)

LCM

- Laser-capture microscope (Leica, LMD7000)
- UVP ultraviolet crosslinker (Thermo Fisher Scientific, model no. CX-2000)
- DNA-OFF (Takara, cat. no. 9036)
- 70% (vol/vol) ethanol spray for decontaminating working surfaces
- Adhesive PCR plate seals (Thermo Fisher Scientific, cat. no. AB0558)
- Eppendorf TwinTec PCR plates (Eppendorf, cat. no. 0030128648)
- 22 mm x 50 mm, glass coverslips (VWR International, cat. no. 631-0137)
- Fully skirted 96-well plate holder

Cell lysis and digestion

- Thermocycler (MJ Research, DNA Engine Tetrad PTC-225)
- Multi-purpose refrigerated centrifuge (Eppendorf, model no. 5810 R)
- Vortex shaker (Cole-Parmer, Vortex-Genie 2 at 2,600 r.p.m. to 2,700 r.p.m., 230 VAC)



gDNA purification; library preparation, amplification and quality control; DNA sequencing

- Plate centrifuge (e.g., Eppendorf 5810R)
- Thermal cycler (e.g., MJ Research Tetrad PTC-225)
- Microcentrifuge (e.g., Eppendorf 5424R)
- Vortex shaker (Cole-Parmer, Vortex-Genie 2 at 2,600 r.p.m. to 2,700 r.p.m., 230 VAC)
- 96-well magnet (e.g., Alpaqua 96S Super Magnet Plate)
- Single-channel and multi-channel pipettes (e.g., P2, P20, P200, P1000 (Anachem; LTS))
- Agilent 2100 Bioanalyzer (Agilent Technologies, cat. no. G2938C)
- Illumina next-generation sequencing platform (e.g., HiSeq X platform)
- Suitable computing infrastructure for next-generation sequence data analysis

Automation platforms

▲CRITICAL The following automation platforms are optional and can be replaced either with suitable alternative platforms or manual approaches (e.g., multi-channel pipetting).

- Bravo G5574A NGS Workstation option B (Agilent Technologies). We use this platform for gDNA purification and DNA library construction.
- Beckman NX-96 (Beckman Coulter). We use this platform for purification of amplified DNA libraries.
- Beckman NX-8 (Beckman Coulter). We use this platform for DNA library pooling.

Software

- AnnotateBAMStatistics (https://github.com/MathijsSanders/AnnotateBAMStatistics)
- AdditionalBAMStatistics (https://github.com/MathijsSanders/AdditionalBAMStatistics)
- ANNOVAR (http://annovar.openbioinformatics.org/)[22]
- BWA (http://bio-bwa.sourceforge.net/)[23]
- CaVEMan (https://github.com/cancerit/cgpCaVEManWrapper)[24]
- CGPWGS (https://github.com/cancerit/dockstore-cgpwgs)
- GATK (https://software.broadinstitute.org/gatk/)[25]
- Picard (https://broadinstitute.github.io/picard/)[26]
- Samtools (http://www.htslib.org/)[27]
- SangerLCMFiltering (https://github.com/MathijsSanders/SangerLCMFiltering)
- Unmatched normal filtering (https://github.com/TimCoorens/Unmatched_NormSeq)

Reagent setup

Protease buffer

Prepare Arcturus PicoPure DNA buffer as per the manufacturer's instructions. Alternatively, follow these steps to prepare our in-house protease-type buffer:

(1) Reconstitute a vial of lyophilised proteinase-K powder with 7 ml of nuclease-free water (concentration, 23.81 mg/ml stock). Store in 1-ml aliquots at 4 °C (buffer is stable for up to 2 months); we suggest putting 'best before date' on each aliquot.

(2) On the day of the cell lysis, dilute the buffer as follows. To 10 µl of reconstituted protease, add 90 µl of nuclease-free water (concentration, 2.38 mg/ml)

(3) Further dilute protease from 2.38 mg/ml to 25 µg/ml by adding 10 µl of protease to 990 µl of nuclease-free water.

(4) To prepare the working buffer, add the following:

Reagent	Stock	Final	Units	Volume (µl) for 1 ml
Tris-HCl, pH 8.0	1,000	30	mM	30
Tween-20	100	0.5	% (vol/vol)	5
IGEPAL CA-630	100	0.5	% (vol/vol)	5
Protease	25	1.25	µg/ml	50
Nuclease-free water				910

(5) Vortex (1 min) and spin down (20–30 s, 18 °C, 1,000*g*). Keep on ice until LCM is completed.

Equipment setup

Microtome

! CAUTION Care must be taken when handling microtome blades. Always use a brush to clean wax residuals from the blade. If interrupted during tissue cutting, ensure that the blade guard is in place and the hand wheel is locked **▲ CRITICAL** Only required when following the paraffin tissue sectioning, fixation and staining procedure (Fig. 1; Procedure Step 1A).

1 Pre-cool paraffin blocks on ice (~2 h to ~4 °C).

2 Place the blade in the holder on the microtome and clamp in place. Check that there is no movement of the blade. **! CAUTION** The clearance angle should be set at 5°.

3 With the hand wheel locked, clamp the wax block in the chuck, tissue outermost, and with the bevelled edge of the cassette pointing to the right of the microtome. Check that there is no movement of the block in the chuck.

4 Release the hand wheel lock on the rotary wheel. Select tissue section thickness (can vary from 5 to 30 μm).

Cryostat

! CAUTION Tissue section preparation of all unfixed human samples must be performed in a Containment Level 2 laboratory. All users must be screened for immunizations in accordance with local guidelines. Care must be taken when handling cryostat blades; always use a brush to clean tissue residuals from a blade **▲ CRITICAL** Only required when following the frozen tissue sectioning, fixation and staining procedure (Fig. 1, right; Procedure Step 1B)

1 Ensure that all trimmings and solid material are removed from the instrument. Disinfect the chamber using 70% (vol/vol) ethanol spray.

2 Pre-cool the cryostat by changing temperature for the chamber (CT) and specimen holder (OT). Refer to the manufacturer's instructions for specific temperature settings, as these will vary depending on the tissue type. We found that many tissues can be cut with the following temperature settings: CT: −18°C to −25°C and OT: −18°C to −25°C.

3 Slide a disposable blade into the holder to pre-cool, making sure that it is securely held in place using the locking mechanism.

4 Place the knife guard over the blade until it is required. Place brushes and chucks in the cryostat to pre-cool (15–20 min).

5 Select tissue section thickness; use the '+' and '–' buttons on the left-hand side. We usually cut 10–30-μm thick sections.

LCM

! CAUTION To minimize DNA cross-contamination, wear a lab coat and gloves during the LCM experiments. Decontaminate all working surfaces with DNA-OFF, paying particular attention to the microscope stage, slide and plate holders. Next, repeat the decontamination procedure with 70% (vol/vol) ethanol spray.

Set up the LCM through the following steps:

1 Switch on the microscope and laser power control and open the LMD software. Select the correct camera type and set up image brightness, exposure and white balance.

2 Laser cutting settings vary depending on the tissue type and section thickness; however, the following setting can be used on many tissues: aperture 2, power 17–25, speed 5–10. **▲ CRITICAL** Make sure to calibrate laser settings and align plate position before the start of dissection. To maximize cell capture, avoid draft from air conditioning and open windows.

3 Select correct slide and sample holders. We usually dissect directly into 96-well plates, but dissection can also be performed into strips of caps or PCR-type tubes, depending on the experiment setup. Open MIC control and laser control and calibrate the laser. Regularly calibrate the plate holder to ensure that tissue microbiopsies are properly landing in wells.

Procedure

Tissue sectioning, fixation and staining

1 For instructions on how to section, fix and stain tissues embedded in paraffin, follow Option A. To section frozen material, follow Option B.

(A) **Paraffin sections ● Timing 1–2 d**

(i) To fix a fresh tissue sample, place it either into 70% (vol/vol) ethanol or in PAXgene FIX Kit and follow the manufacturer's instructions. Fixation time will vary depending on the tissue dimensions, but a rate of 1 mm/h penetration can be used for general guidance for both

Table 2 | Tissue processor program for small tissue (<0.5 cm maximum dimension)

Solution	Time	Temperature	P/V (pump/vacuum)	Mix
90% ethanol (vol/vol)	10 min	35 °C	On	Slow
100% ethanol (vol/vol)	10 min	35 °C	On	Slow
100% ethanol (vol/vol)	10 min	35 °C	On	Slow
100% ethanol (vol/vol)	10 min	35 °C	On	Slow
Xylene	10 min	35 °C	On	Slow
Xylene	10 min	35 °C	On	Slow
Xylene	10 min	35 °C	On	Slow
Wax	20 min	63 °C	On	Slow
Wax	10 min	63 °C	On	Slow
Wax	10 min	63 °C	On	Slow

Table 3 | Tissue processor program for large tissue (>0.5 cm maximum dimension)

Solution	Time	Temperature	P/V	Mix
50% (vol/vol) ethanol	45 min	Ambient (18–22 °C)	On	Slow
70% (vol/vol) ethanol	45 min	Ambient	On	Slow
90% (vol/vol) ethanol	45 min	Ambient	On	Slow
100% (vol/vol) ethanol	60 min	Ambient	On	Slow
100% (vol/vol) ethanol	60 min	Ambient	On	Slow
100% (vol/vol) ethanol	60 min	Ambient	On	Slow
Xylene	60 min	Ambient	On	Slow
Xylene	60 min	Ambient	On	Slow
Xylene	60 min	Ambient	On	Slow
Wax	75 min	60 °C	On	Slow
Wax	75 min	60 °C	On	Slow
Wax	75 min	60 °C	On	Slow
Wax	75 min	60 °C	On	Slow

PAXgene FIX Kit and ethanol approaches. To fix frozen tissue, first briefly thaw it at 4 °C (thawing time will vary based on the tissue size; we typically thaw a 1-cm^3 sample for 15 min).

(ii) Place fixed tissue samples into standard histology cassettes and process using the conditions outlined in either Table 2 or Table 3, depending on the tissue size. For small tissue samples (<0.5 cm in maximum dimension), using the tissue processor, execute the program outlined in Table 2. For larger samples (>0.5 cm in maximum dimension), using the tissue processor, execute the program outlined in Table 3.

(iii) Once the tissue processing is complete, transfer tissue samples onto the tissue embedding station for embedding and trimming.

(iv) To cut sections, cool down the paraffin tissue block by placing it on ice (~2–3 h for a standard histology block). Fill the water bath with reverse osmosis water, set temperature to 37 °C and then cut sections to the desired thickness. We have successfully generated libraries from 10–30-µm thick sections.

▲CRITICAL STEP To minimize contamination with external DNA, use new microtome blades and fresh water for each batch of samples and ensure pre-labeled PEN-slides have been crosslinked in a UV crosslinker instrument (recommended time, 30 min).

(v) Before staining tissue sections, remove xylene by placing slides in Coplin jars or staining pots and sequentially immersing in the following: xylene (2 min, twice), 100% (vol/vol) ethanol (1 min, twice), 70% (vol/vol) ethanol (1 min) and deionised water (1 min).

(vi) Stain tissue sections by sequential immersing in the following: hematoxylin (10–20 s), tap water (10–15 s, twice), eosin (5–10 s), tap water (10–15 s), 70% (vol/vol) ethanol (5–10 s, twice), 100% (vol/vol) ethanol (5–10 s, twice) and xylene (or Neo-Clear, 5 s, once). Proceed to LCM cell isolation or store slides at 4 °C.

▲CRITICAL STEP To minimize DNA cross-contamination, change Coplin jars/staining pots between samples from different patients.

■PAUSE POINT To preserve DNA quality, store paraffin tissue blocks and slides at 4 °C until ready to proceed with LCM. We have successfully generated libraries from slides stored at 4 °C for up to 1.5 years.

(B) **Frozen sections** ● Timing **2 h**

(i) To prepare frozen sections, place tissue sample in OCT compound in a moulding block and leave to solidify on dry ice (15–30 min).

(ii) Once the block is set, use cryostat to cut sections of desired thickness (typically, 10–30 μm) and mount them onto PEN-slides.

▲CRITICAL STEP To minimize contamination with external DNA, crosslink pre-labeled PEN-slides in a UV-crosslinker instrument for 30 min before mounting tissue onto slides. Once tissue is mounted to slides, proceed to fixation.

■PAUSE POINT Alternatively, unfixed sections can be stored at 80 °C for up to 1 week until ready to proceed with the next steps.

(iii) Immerse tissue slides in 0.5–1 ml of 70% (vol/vol) ethanol in a 6" Petri dish. Leave to fix for 2 min and then aspirate residual ethanol with a pipette. Wash with 0.5–1 ml of PBS (15–30 s, twice).

▲CRITICAL STEP Ensure that the entire tissue section is covered in ethanol during fixation.

! CAUTION Take care not to wash tissue sections off the slides.

(iv) Leave tissue slides in PBS at 4 °C until ready for staining, 1 min–1 h (ideally, proceed to staining and LCM experiments within 24 h).

(v) Place fixed tissue slides in Coplin jars and stain by sequential immersing in the following: hematoxylin (10–20 s), tap water (10–15 s, twice), eosin (5–10 s), 70% (vol/vol) ethanol (5–10 s, twice), 100% (vol/vol) ethanol (5–10 s, twice) and xylene (or Neo-Clear, 5 s). Proceed to LCM cell isolation within 24 h.

■PAUSE POINT To preserve DNA quality, store frozen tissue blocks and slides at −80 °C until ready to proceed with LCM. We have successfully generated libraries from slides stored at −80 °C for up to 1.5 years.

LCM ● Timing Variable

▲CRITICAL The duration of this part of the workflow is highly variable and depends on the experimental design. However, when working on frozen sections, we recommend to complete LCM experiments within 24 h to minimize DNA degradation. For PAXgene-fixed, paraffin-embedded samples (from Step 1A), we have successfully constructed libraries from membrane slides stored at 4 °C for over 1 year.

▲CRITICAL We recommend taking a whole-slide overview image before and after LCM cutting for all tissue slides. To speed up imaging and increase image quality, ×20 or ×40 before images of slides with temporarily mounted glass coverslips are taken using the NanoZoomer S60 digital slide scanner. As the tissue cut on the LCM is typically dried out, and without a coverslip, using the higher-resolution slide scanner image to guide the LCM cutting is recommended.

2 Using the LCM software, acquire images immediately before and after taking each LCM microbiopsy.

3 Cut microbiopsies into a 96-well plate.

▲CRITICAL STEP Occasionally, tissue might not release from the PEN membrane. In that case, adjust the LCM laser settings or manually pulse with stronger power settings to detach the tissue into a well.

▲CRITICAL STEP To minimize DNA cross-contamination, crosslink all plates before the start of LCM for 30 min. We recommend leaving several empty wells to serve as negative controls to test for potential cross-contamination.

4 Once dissection is completed, proceed immediately to cell lysis and digestion.

■PAUSE POINT If necessary, the plate can be covered and stored at 4 °C for up to 48 h. This is not recommended, as any additional plate seal application and removal can lead to unnecessary loss of microbiopsies.

Cell lysis and digestion ● Timing 2 h

▲CRITICAL No DNA quantification is done after cell lysis, but successful library generation rates in excess of 80% are routine (Fig. 6b).

5 Visually inspect the plate to ensure that tissue did not miss the wells. Tissue in between wells can be an indication of potential cross-contamination.

6 To use the in-house protease buffer, follow Option A. If using the Arcturus PicoPure buffer, follow Option B.

(A) **In-house protease buffer**

(i) Add 20 µl of in-house protease buffer to each well.

(ii) Lightly vortex the plate for 5–10 s and centrifuge at 1,500*g* for 1 min at 4 °C.

▲CRITICAL STEP Visually inspect the wells and make sure that LCM samples are at the bottom of each well, but bear in mind that, for microbiopsies smaller than 100,000 µm^3, tissue might not be visible to the human eye.

(iii) Run the following program on the thermocycler:

Step	Temperature (°C)	Time (min)
1	50	12
2	75	30
3	4	Hold

■PAUSE POINT LCM lysates can be stored at −20 °C for up to 1 month.

(B) **Arcturus PicoPure protease buffer**

(i) Add 155 µl of Arcturus PicoPure DNA reconstitution buffer to the Arcturus PicoPure Proteinase K. Pulse vortex and briefly spin down (5–10 s).

(ii) Add 20 µl of solution to each well.

(iii) Lightly vortex the plate for 5–10 s and centrifuge at 1,500*g* for 1 min at 4 °C.

▲CRITICAL STEP Visually inspect the wells and make sure that LCM samples are at the bottom of each well. For microbiopsies smaller than 100,000 µm^3, tissue might not be visible to the human eye.

(iv) Seal plate and place on thermocycler and run the following program: 60 °C for 3 h, 75 °C for 30 min, hold at 4 °C.

■PAUSE POINT LCM lysates can be stored at −20°C for up to 1 month.

Purification of gDNA from LCM lysates ● Timing 45 min

▲CRITICAL Perform all steps in a dedicated pre-PCR amplification space. We use an Agilent Bravo NGS Workstation to perform Steps 9–34, but these steps can be performed on other platforms or manually. We refer to sample processing in 96-well plates throughout the remaining steps in the workflow, but the protocol may also be performed in individual tubes.

7 Allow Agencourt AMPure beads to reach room temperature for 15 min before use. Vortex AMPure beads to ensure that the beads are resuspended.

8 Thaw LCM lysate plates and centrifuge for 1 min at 1,000*g* at 4 °C. Each sample well should contain 20 µl of liquid in total.

9 To each LCM lysate, add 100 µl of a 1:1 mixture of AMPure beads and TE buffer (pH 8.0).

10 Mix thoroughly by pipetting up and down and allow the mixture to stand for 5 min at room temperature. We use diluted AMPure beads to avoid dispensing small volumes. The final ratio of AMPure beads to sample is ~0.7:1.

11 Transfer the plate to the magnet, allow the beads to settle for 5 min at room temperature and then carefully remove and discard the supernatant.

12 With the plate still on the magnet, wash the beads with 150 µl of 80% (vol/vol) ethanol for 1 min. Remove and discard the ethanol wash.

13 Repeat Step 12 one more time, removing all traces of the ethanol wash.

14 Air dry the beads for 5 min at room temperature.

15 Resuspend the beads in 26 µl of TE buffer by repeated pipetting up and down. Proceed directly to Step 16.

DNA library construction and amplification of adapter-ligated libraries ● Timing 2.5 h

▲CRITICAL Continue the protocol in the same pre-PCR amplification laboratory until DNA library amplification (Step 35) and then move to the post-PCR amplification laboratory.

16 Prepare the fragmentation/end repair/dA-tailing mix as described in the table below, which will provide sufficient mastermix for one 96-well plate. Mix thoroughly by pipetting up and down and place on ice.

Component	Volume (µl)
NEBNext Ultra II FS Reaction Buffer	770
NEBNext Ultra II FS Reaction Enzyme	220

17 Place a fresh 96-well plate onto a chilled position of the robot deck or on ice and add 9 µl of fragmentation/end repair/dA-tailing mix from Step 16 to each well.

18 Transfer the entire volume, including the beads (26 µl) from Step 15, to the reaction plate. It might be necessary to resuspend the beads if they have settled.

19 Place the plate in a thermal cycler with a heated lid and perform the following steps:

Step	Temperature (°C)	Time (min)
1	37	12
2	65	30
3	4	Hold

20 Prepare the ligation mix as described in the table below, which will provide sufficient mastermix for one 96-well plate. Mix thoroughly by pipetting up and down and place on ice.

Component	Volume (µl)
NEBNext Ultra II Ligation Master Mix	3,300
NEBNext Ultra II Ligation Enhancer	110
Nuclease-free water	247.5
Duplexed adapter (100 µM)	27.5

21 Place a fresh 96-well plate onto a chilled position of the robot deck or on ice and add 33.5 µl of ligation mix from Step 20 to each well.

22 Transfer the entire volume, including the beads (35 µl) from Step 19, to the reaction plate from Step 21 and mix thoroughly by pipetting up and down.

23 Incubate for 15 min at 20 °C and proceed immediately to Step 24.

24 To each ligation reaction, add 35 µl of a 1:1 mixture of AMPure beads and TE and mix thoroughly by pipetting up and down. Allow the mixture to stand for 5 min at room temperature.

25 Transfer the plate to the magnet, allow the beads to settle for 5 min at room temperature and then carefully remove and discard the supernatant.

26 With the plate still on the magnet, wash the beads with 150 µl of 80% (vol/vol) ethanol for 1 min. Remove and discard the ethanol wash.

27 Repeat Step 26 one more time, removing all traces of the ethanol wash.

28 Air dry the beads for 5 min at room temperature.

29 Resuspend the beads in 23 µl of nuclease-free water by repeated pipetting up and down and incubate for 2 min at room temperature.

30 Place the plate on the magnet and prepare the PCR reaction.

31 Prepare the mix as described in the table below, which will provide sufficient PCR mastermix for one 96-well plate. Mix thoroughly by pipetting up and down and place on ice.

Component	Volume (µl)
Kapa HiFi HotStart ReadyMix (2×)	2,750
PE1.0 primer (100 µM)	110

32 Place a new 96-well plate containing indexed iPCR-Tag primers (2.5 µl of each primer at 40 µM) onto a chilled position of the robot deck or on ice if performed manually. This plate will become the PCR reaction plate.

33 Using a separate pipette tip for each well, transfer 26 µl of PCR mix to each well containing indexed primers.

34 Keeping the plate on the magnet, transfer 21.5 µl of adapter-ligated library from Step 30 to a well of the PCR plate (Step 33) and mix thoroughly by pipetting up and down 4–6 times.

35 Seal the plate, move to the post-PCR amplification laboratory and perform library amplification as detailed in the table below.

Cycle number	Denature	Anneal	Extend	Hold
1	95 °C, 5 min			
2–12	98 °C, 30 s	65 °C, 30 s	72 °C, 1 min	
13			72°C, 5 min	
14				4 °C, ∞

▲CRITICAL STEP We use 12 cycles of PCR for LCM lysates from 100–1,000 cells, which is equivalent to 5–50 cells per µl of lysis buffer. If lysates contain a higher concentration of cells and, therefore, available DNA, then either dilute the lysate or reduce the number of PCR cycles. More than 12 amplification cycles can be used for lysates containing fewer than 100 cells. However, we estimate from tests using purified human gDNA that 60–100 cells are required to achieve a sequenced human genome with an average read depth of 20. Below this input, the level of PCR duplicates in the sequence data can exceed 50% (Table 1).

■ PAUSE POINT Amplified libraries can be stored at −20°C for several months before purification.

Purification of amplified libraries ● Timing 45 min

▲CRITICAL We use a Beckman NX-96 to perform Steps 36–43, but these steps can be performed on other liquid handling platforms or manually.

36 To each PCR reaction from Step 35, add 35 µl of AMPure beads and mix thoroughly by pipetting up and down. Allow the mixture to stand for 5 min at room temperature.

▲CRITICAL STEP We use a bead-to-sample ratio of 0.7:1 to target a median insert length of 350 bp. Lower ratios (e.g., 0.6:1) will lead to a reduced fraction of overlapping reads (e.g., on 150-bp paired-end reads) but a lower concentration of DNA library.

37 Transfer the plate to the magnet, allow the beads to settle for 5 min at room temperature and then carefully remove and discard the supernatant.

38 With the plate still on the magnet, wash the beads with 150 µl of 80% (vol/vol) ethanol for 1 min. Remove and discard the ethanol wash.

39 Repeat Step 38 one more time, removing all traces of the ethanol wash.

40 Air dry the beads for 5 min at room temperature.

41 Resuspend the beads in 25 µl of nuclease-free water by repeated pipetting up and down and incubate for 2 min at room temperature.

42 Place the plate on the magnet for 2 min at room temperature.

43 Transfer the amplified and purified libraries to a fresh 96-well plate and proceed to DNA library quality control.

DNA library quality control and DNA sequencing ● Timing Approximately 3 d for HiSeqX platform

▲CRITICAL To ensure even representation of libraries within a multiplex sequencing reaction, we quantify each library using a fluorescence-based DNA quantitation assay (Accuclear; Biotium). We have successfully used Picogreen (Invitrogen), UV absorption or q-PCR (Kapa Library Quantification Kit for Illumina platforms; Kapa Biosystems) for high-throughput approaches and Qubit (Invitrogen) or Bioanalyzer (Agilent Technologies) for low-throughput applications. For DNA library pooling, we use a Beckman NX with eight independent pipetting channels, but these steps can be performed on other platforms or manually.

44 Quantify DNA libraries and transfer an equimolar amount of each DNA library into a 1.5-ml microcentrifuge tube.

▲CRITICAL STEP This quantification step is the first opportunity to assess the success and quality of the library.

? TROUBLESHOOTING

45 Mix thoroughly and briefly centrifuge.

46 Dilute the library pool to ~5 nM. We use a conversion of 2.45 ng/µl to 1 nM to estimate the pool concentration.

47 Check the quality and quantity of the library pool using a high-sensitivity chip on an Agilent Bioanalyzer. A typical library pool will consist of DNA fragments from 300 to 1,000 bp (Fig. 2c).

48 Record the concentration of the library pool (nM). We dilute each pool or individual sample to 2.8 nM such that 5 µl of the library pool solution can enter the recommended workflow for Illumina HiSeqX sequencing.

49 Perform paired-end sequencing in accordance with the manufacturer's protocols.

▲**CRITICAL STEP** We use a custom ISPCR primer to read the 8-base index sequence introduced with our ISPCR-Tags. For many 'off-the-shelf' adapters and sequences, this will not be necessary.

Mapping data and marking duplicates ● Timing 2–3 d

50 Align processed sequencing data (in fastQ format) to the human reference genome (data presented here used NCBI build 37; build 38 has also been used without issue) using the Burrows–Wheeler aligner (BWA-MEM)[23]. Mark PCR and optical duplicates with the MarkDuplicates command within the GATK package[25]. This will output a binary alignment map (BAM) file.

? TROUBLESHOOTING

SNV discovery ● Timing 1 d

51 Compare the BAM file of the LCM sample of interest to the BAM file of a control sample to determine the somatic variants. We highlight three potential control approaches: (A) matched bulk normal, (B) matched LCM normal and (C) unmatched normal. This comparison will output a variant call format (VCF) file.

(A) **Matched bulk normal**

(i) Generate a WGS sequencing library from a bulk tissue sample obtained from the same donor using a more traditional amount of DNA input (>500 ng) and a standard library preparation protocol[28].

(ii) Sequence this sample to at least 30× depth.

(iii) Proceed with paired somatic mutation calling (Step 52) using the LCM sample of interest as 'tumor' and the bulk normal sample as 'normal'.

(B) **Matched LCM normal**

(i) Generate a low-input LCM-derived sequencing library taken from the same donor but from a different tissue type (lymph node, stroma, muscle, etc.). Follow the procedure outlined in Steps 2–50.

(ii) Sequence this sample to at least 30× depth.

(iii) Proceed with paired somatic mutation calling (Step 52) using the LCM sample of interest as 'tumor' and the nominated normal LCM sample as 'normal'.

(C) **Unmatched normal**

▲**CRITICAL** For more detail on filtering out germline SNPs, and for the appropriate scripts, see https://github.com/TimCoorens/Unmatched_NormSeq. The approach outlined below is different from conventional approaches, as germline SNPs are removed when consistently detected at VAFs of 0.5 across all microbiopsies from the same donor.

The following steps are required to filter germline SNPs and artifactual variants when using unmatched controls:

(i) Identify sites different from the reference genome for all samples taken from the same donor.

(ii) Recount all variant sites across all samples and aggregate total coverage and variant-supporting reads (using 'AnnotateBAMStatistics' in the SangerLCMFiltering Singularity container).

(iii) Filter out germline variants using the exact binomial test— null hypothesis $P = 0.5$; alternative hypothesis $P < 0.5$. A germline variant should be omnipresent across all samples, and the total number of variant-supporting reads is binomially distributed with $P = 0.5$ (use 'germline_exact_binom.R' in Unmatched_NormSeq).

(iv) Filter out artifactual variants using a beta-binomial test (using 'beta_binom_filter.R' in Unmatched_NormSeq).

(v) Proceed to Step 56.

52 Detect variants using the Cancer Variants Through Expectation Maximization (CaVEMan)[24] algorithm, available on GitHub (https://github.com/cancerit/dockstore-cgpwgs), using a copy number state of 10/2, which we found maximizes sensitivity.

SNV filtering ● Timing 1 d

▲CRITICAL Several post-processing filters can be applied to maximize the variant detection specificity. We first apply three filtering steps commonly used in our filtering strategies to remove a substantial fraction of erroneous variants. The CGPWGS container provides software that calculates a set of statistics that inform which variants are likely false. In addition, detected variants are further filtered based on a list of previously detected variants in a panel of 75 control samples that have been sequenced using the same platform and at the same sequencing center[24]. Examples of the number of variants filtered by these steps can be found in Fig. 9.

Filtering out erroneous variants

▲CRITICAL In the following steps, CaVEMan calculates a set of values, some of which (listed in Step 53) are 'hard filters' that would cause a variant to automatically fail, whereas the filtering based on the ASRD and CLPM score described in Steps 54 and 55 must be done manually.

53 Remove variants that fail one or more of the post-processing filters. The current post hoc filters are described below and are added to the VCF file upon using CaVEMan from the CGPWGS container. The indicated values are flags for variants that fail filtering:

- CR, centromeric repeat: Position falls within centromeric repeat using a predefined BED file.
- DTH, depth: Fewer than 1/3 of mutant alleles were ≥25 base quality.
- GI, germline indel: Position falls within a germline indel for this sample.
- MN, matched normal: Matched normal has VAF ≥0.03 for this mutation with base quality ≥15.
- MNP, matched normal proportion: Tumor VAF—Normal VAF <0.2.
- MQ. mapping quality: Mean mapping quality of the mutant allele reads <21.
- SE, single end: Coverage is ≥10 on each strand, but mutant allele is present only on one strand.
- SR, simple repeat: Position falls within simple repeat region using a predefined BED file.
- VUM, variant unmatched normal panel: Variant is present at VAF ≥0.03 in 1% of the unmatched normal panel. Our panel comprises 75 whole-genome sequencing samples generated from whole blood, sequenced to at least 30× depth.

  ▲CRITICAL STEP We recommend generating your own VUM panel, as platform- and institution-specific sequencing artifacts have been observed, and an unmatched normal panel is a reasonable approach to addressing those.

54 For each variant, the 'Read-adjusted Median Alignment Score' (ASRD) is calculated by CGPWGS based on the variant-supporting reads and adjusted by read length. Discard variants with a score lower than 0.93 (corresponding to a 'Median Alignment Score' of 140 for 150-bp paired-end reads).

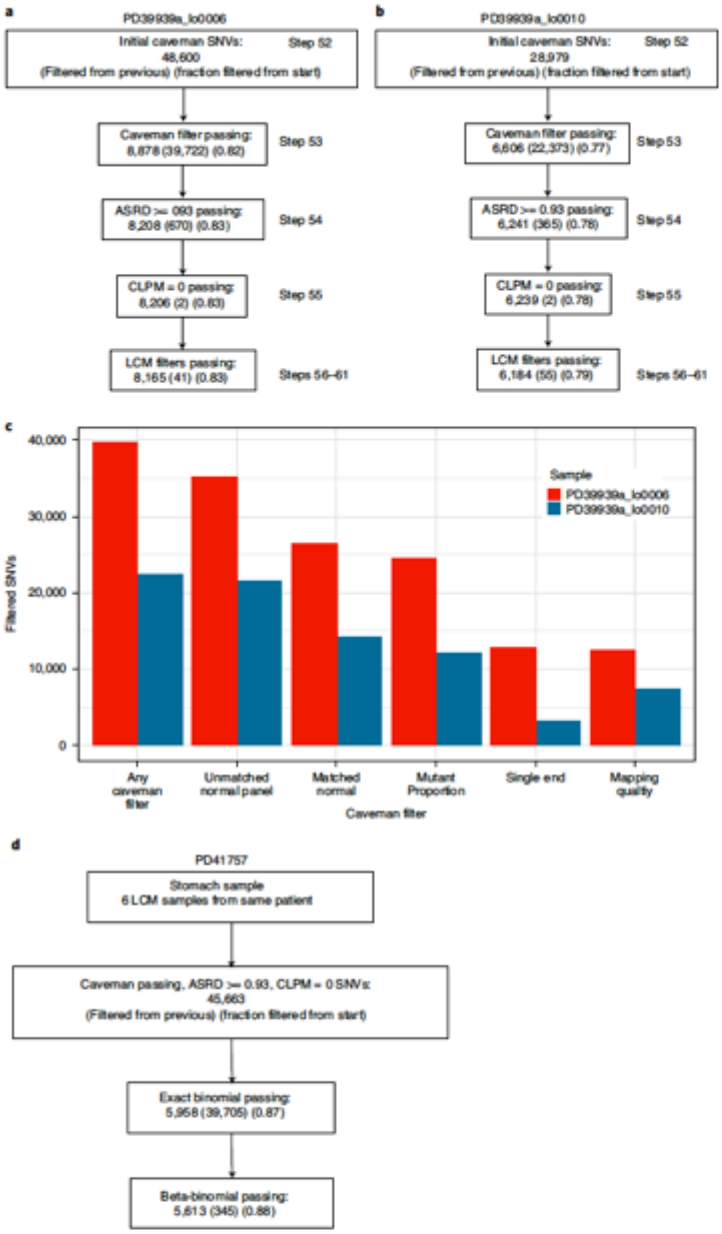
55 For each variant, the 'Median Clipping Length' (CLPM) is calculated by CGPWGS based on variant-supporting reads. Discard variants with a median clipping length score greater than 0.

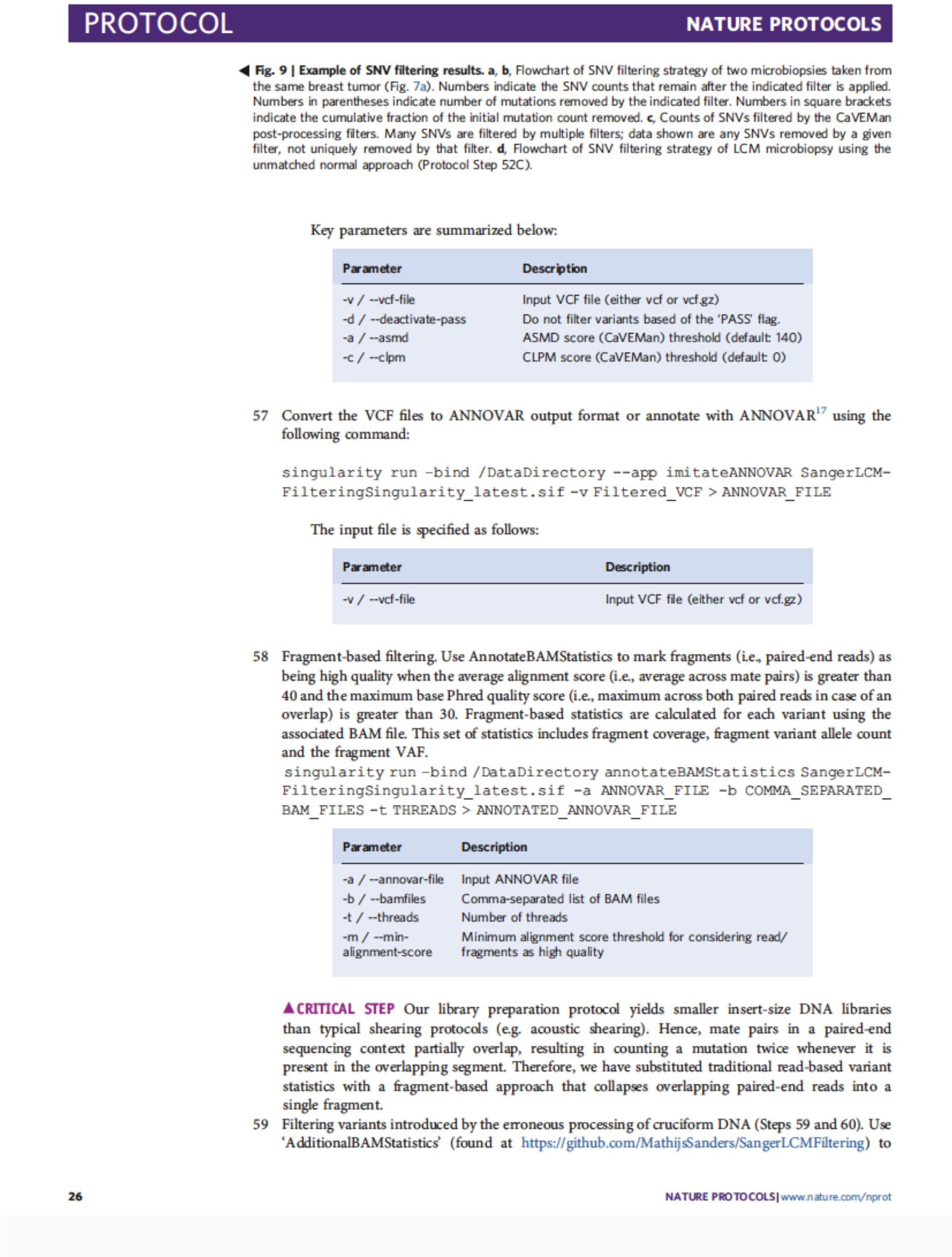
Filtering specific to low-input samples

▲CRITICAL The following filtering steps were developed specifically to improve variant detection for this low-input pipeline. Variants that pass the filters described in Steps 53–55 are run through the filters in Steps 56–61.

56 Select variants passing filtering flags and predefined criteria specified in Steps 53–55 by running the following command. The algorithm automatically detects whether the VCF is produced by CaVEMan or a different variant caller (substitute DataDirectory with the directory where data are stored).

```
singularity run -bind /DataDirectory --app preselect SangerLCMFilter-
ingSingularity_latest.sif -v Input_VCF > Filtered_VCF
```


a
PD39939a_lo0006
Initial caveman SNVs: Step 52
48,600
(Filtered from previous) (fraction filtered from start)
Caveman filter passing:
8,878 (39,722) (0.82)
Step 53
ASRD >= 093 passing:
8,208 (670) (0.83)
Step 54
CLPM = 0 passing:
8,206 (2) (0.83)
Step 55
LCM filters passing:
8,165 (41) (0.83)
Steps 56–61
b
PD39939a_lo0010
Initial caveman SNVs: Step 52
28,979
(Filtered from previous) (fraction filtered from start)
Caveman filter passing:
6,606 (22,373) (0.77)
Step 53
ASRD >= 0.93 passing:
6,241 (365) (0.78)
Step 54
CLPM = 0 passing:
6,239 (2) (0.78)
Step 55
LCM filters passing:
6,184 (55) (0.79)
Steps 56–61
c
Filtered SNVs
40,000
30,000
20,000
10,000
0
Sample
PD39939a_lo0006
PD39939a_lo0010
Any caveman filter
Unmatched normal panel
Matched normal
Mutant Proportion
Single end
Mapping quality
Caveman filter
d
PD41757
Stomach sample
6 LCM samples from same patient
Caveman passing, ASRD >= 0.93, CLPM = 0 SNVs:
45,663
(Filtered from previous) (fraction filtered from start)
Exact binomial passing:
5,958 (39,705) (0.87)
Beta-binomial passing:
5,613 (345) (0.88)



◀ **Fig. 9 | Example of SNV filtering results. a**, **b**, Flowchart of SNV filtering strategy of two microbiopsies taken from the same breast tumor (Fig. 7a). Numbers indicate the SNV counts that remain after the indicated filter is applied. Numbers in parentheses indicate number of mutations removed by the indicated filter. Numbers in square brackets indicate the cumulative fraction of the initial mutation count removed. **c**, Counts of SNVs filtered by the CaVEMan post-processing filters. Many SNVs are filtered by multiple filters; data shown are any SNVs removed by a given filter, not uniquely removed by that filter. **d**, Flowchart of SNV filtering strategy of LCM microbiopsy using the unmatched normal approach (Protocol Step 52C).

Key parameters are summarized below:

Parameter	Description
-v / --vcf-file	Input VCF file (either vcf or vcf.gz)
-d / --deactivate-pass	Do not filter variants based of the 'PASS' flag.
-a / --asmd	ASMD score (CaVEMan) threshold (default: 140)
-c / --clpm	CLPM score (CaVEMan) threshold (default: 0)

57 Convert the VCF files to ANNOVAR output format or annotate with ANNOVAR[17] using the following command:

```
singularity run -bind /DataDirectory --app imitateANNOVAR SangerLCM-
FilteringSingularity_latest.sif -v Filtered_VCF > ANNOVAR_FILE
```

The input file is specified as follows:

Parameter	Description
-v / --vcf-file	Input VCF file (either vcf or vcf.gz)

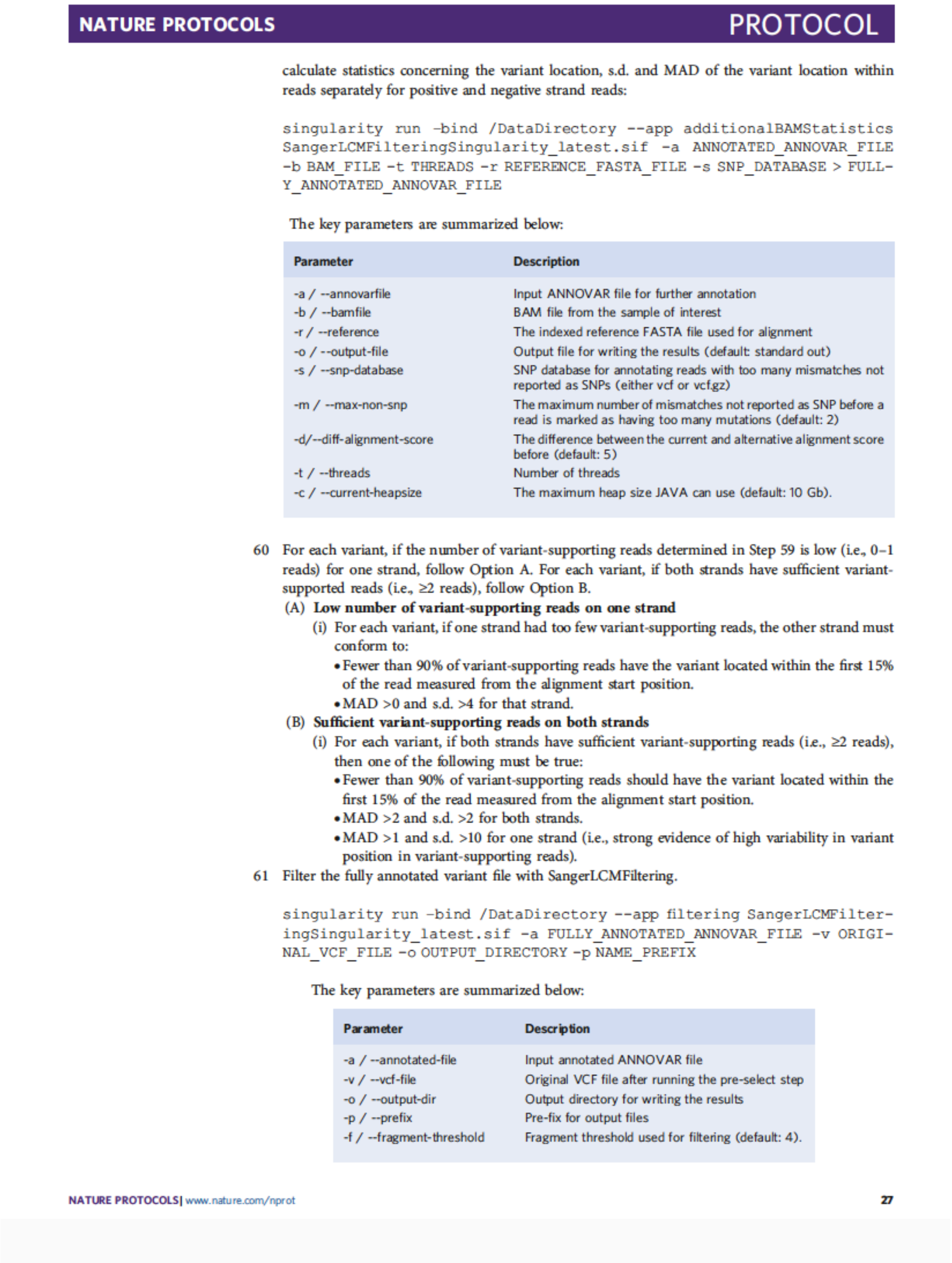
58 Fragment-based filtering. Use AnnotateBAMStatistics to mark fragments (i.e., paired-end reads) as being high quality when the average alignment score (i.e., average across mate pairs) is greater than 40 and the maximum base Phred quality score (i.e., maximum across both paired reads in case of an overlap) is greater than 30. Fragment-based statistics are calculated for each variant using the associated BAM file. This set of statistics includes fragment coverage, fragment variant allele count and the fragment VAF.

```
singularity run -bind /DataDirectory annotateBAMStatistics SangerLCM-
FilteringSingularity_latest.sif -a ANNOVAR_FILE -b COMMA_SEPARATED_
BAM_FILES -t THREADS > ANNOTATED_ANNOVAR_FILE
```

Parameter	Description
-a / --annovar-file	Input ANNOVAR file
-b / --bamfiles	Comma-separated list of BAM files
-t / --threads	Number of threads
-m / --min-alignment-score	Minimum alignment score threshold for considering read/ fragments as high quality

▲**CRITICAL STEP** Our library preparation protocol yields smaller insert-size DNA libraries than typical shearing protocols (e.g. acoustic shearing). Hence, mate pairs in a paired-end sequencing context partially overlap, resulting in counting a mutation twice whenever it is present in the overlapping segment. Therefore, we have substituted traditional read-based variant statistics with a fragment-based approach that collapses overlapping paired-end reads into a single fragment.

59 Filtering variants introduced by the erroneous processing of cruciform DNA (Steps 59 and 60). Use 'AdditionalBAMStatistics' (found at https://github.com/MathijsSanders/SangerLCMFiltering) to



calculate statistics concerning the variant location, s.d. and MAD of the variant location within reads separately for positive and negative strand reads:

```
singularity run -bind /DataDirectory --app additionalBAMStatistics
SangerLCMFilteringSingularity_latest.sif -a ANNOTATED_ANNOVAR_FILE
-b BAM_FILE -t THREADS -r REFERENCE_FASTA_FILE -s SNP_DATABASE > FULL-
Y_ANNOTATED_ANNOVAR_FILE
```

The key parameters are summarized below:

Parameter	Description
-a / --annovarfile	Input ANNOVAR file for further annotation
-b / --bamfile	BAM file from the sample of interest
-r / --reference	The indexed reference FASTA file used for alignment
-o / --output-file	Output file for writing the results (default: standard out)
-s / --snp-database	SNP database for annotating reads with too many mismatches not reported as SNPs (either vcf or vcf.gz)
-m / --max-non-snp	The maximum number of mismatches not reported as SNP before a read is marked as having too many mutations (default: 2)
-d/--diff-alignment-score	The difference between the current and alternative alignment score before (default: 5)
-t / --threads	Number of threads
-c / --current-heapsize	The maximum heap size JAVA can use (default: 10 Gb).

60 For each variant, if the number of variant-supporting reads determined in Step 59 is low (i.e., 0–1 reads) for one strand, follow Option A. For each variant, if both strands have sufficient variant-supported reads (i.e., ≥2 reads), follow Option B.

(A) **Low number of variant-supporting reads on one strand**

(i) For each variant, if one strand had too few variant-supporting reads, the other strand must conform to:

- Fewer than 90% of variant-supporting reads have the variant located within the first 15% of the read measured from the alignment start position.
- MAD >0 and s.d. >4 for that strand.

(B) **Sufficient variant-supporting reads on both strands**

(i) For each variant, if both strands have sufficient variant-supporting reads (i.e., ≥2 reads), then one of the following must be true:

- Fewer than 90% of variant-supporting reads should have the variant located within the first 15% of the read measured from the alignment start position.
- MAD >2 and s.d. >2 for both strands.
- MAD >1 and s.d. >10 for one strand (i.e., strong evidence of high variability in variant position in variant-supporting reads).

61 Filter the fully annotated variant file with SangerLCMFiltering.

```
singularity run -bind /DataDirectory --app filtering SangerLCMFilter-
ingSingularity_latest.sif -a FULLY_ANNOTATED_ANNOVAR_FILE -v ORIGI-
NAL_VCF_FILE -o OUTPUT_DIRECTORY -p NAME_PREFIX
```

The key parameters are summarized below:

Parameter	Description
-a / --annotated-file	Input annotated ANNOVAR file
-v / --vcf-file	Original VCF file after running the pre-select step
-o / --output-dir	Output directory for writing the results
-p / --prefix	Pre-fix for output files
-f / --fragment-threshold	Fragment threshold used for filtering (default: 4).

In our experience, the proposed filtering strategies remove erroneous variants due to (i) poor overall quality of the called variant or the presence of the variant in a panel of normal cases; (ii) a low number of variant-supporting fragments; and (iii) the likelihood that the variant is introduced due to erroneous processing of cruciform DNA. Application of the described pipeline to bulk tissue WGS data revealed that the final filtering steps (i.e., filtering variants introduced by erroneous cruciform DNA processing) removes virtually none of the detected variants while specifically removing erroneous variants commonly observed in LCM experiments (Fig. 4).

Troubleshooting

Troubleshooting advice can be found in Table 4. Most, if not all, problems that we observe can be explained by too little or too much starting DNA. Possible causes and actions are summarized.

Table 4 | Troubleshooting table

Step	Problem	Possible reason	Solution
44	DNA library yield too low	Failed DNA library construction	Validate own protocol and reagents, such as oligonucleotides, using isolated DNA
		Failed collection of LCM material	In cases where LCM samples are large enough to be visible to the eye, the capture rate can be improved as follows: (i) Check the rim of each well for the presence of LCM sample. Slide microbiopsy to the bottom of the well using the micropipette tip (ii) After the addition of protease buffer, make sure that each LCM sample is at the bottom of the well, immersed in buffer. If the sample is found sticking on the wall of the well, wash down using protease buffer
			Ensure plate holder is properly calibrated to prevent wells from being missed
			Minimize air flow near LCM to prevent microbiopsies from missing wells
		Not enough cells captured	Increase area or cell count of microbiopsy; consider adding multiple adjacent *z*-sections of the same histological feature to increase DNA input
			Increase section thickness
	High adapter contamination	DNA input too low	See 'DNA library yield too low' above
		Adapter input too high	Use amounts recommended in the Procedure. If necessary, batch test and titrate adapters using isolated DNA tests
	Overamplification of DNA libraries	DNA input too high	For human DNA, adjust input to 100–1,000 cells. For known higher inputs, reduce PCR cycling
50	PCR duplicates >50%	Typically caused by low DNA inputs	See 'DNA library yield too low' above

Timing

Tissue sectioning, fixation and staining
Step 1Ai–vi, paraffin sections: 1–2 d
Step 1Bi–v, frozen sections: 2 h
Steps 2–4, LCM: variable
Steps 5–6, cell lysis and digestion: 2 h
Steps 7–15, purification of gDNA from LCM lysates: 45 min
Steps 16–35, DNA library construction and amplification of adapter-ligated libraries: 2.5 h
Steps 36–43, purification of amplified libraries: 45 min
Steps 44–49, DNA library quality control and DNA sequencing (Steps 44–49): 3 d for HiSeqX platform
Steps 50–61, data analysis: 4–5 d

Table 5 | Performance metrics based on HiSeqX lane aiming for 30X genome (human) WGS

Metric	Typical value
DNA library yield[a]	250–2,500 fmol; 70–700 ng
% adapters	<1%
% mapped to reference genome	>99%
GC content	40–42
% PCR duplicates	10–50[b]
No. of variants (SNV)	Variable[c]
Median insert length (bp)	300–400

[a] 12 cycles of PCR library amplification [b] Includes duplicates derived from Illumina Flow Cell clustering process [c] Number of variants is tissue specific and age dependent. As an example, in hepatocytes, we observe ~1,200 SNVs per genome of a healthy 60 year old.

Anticipated results

We demonstrate that our protocol reliably generates WGS libraries from LCM-derived microbiopsies of fewer than 1,000 cells. The protocol is compatible with frozen and alcohol-fixed tissue and was tested over a wide variety of human tissues (including bladder, breast, colon, endometrium, lung and stomach). Typical DNA library concentrations prepared from LCM material are shown in Fig. 6. Typical performance metrics are summarized in Table 5. Typical SNV results can be seen in Figs. 4 and 5. The effect of the SNV filtering strategy on SNV counts is shown in Fig. 9.

Data availability

Sequencing data referred to in this study have been deposited in the European Genome-phenome Archive with accession code EGAD00001006088.

References

1. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
2. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
3. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
4. Fatehullah, A., Tan, S. H. & Barker, N. Organoids as an in vitro model of human development and disease. *Nat. Cell Biol.* **18**, 246–254 (2016).
5. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
6. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
7. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
8. Olafsson, S. et al. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* **182**, 672–684 (2020).
9. Telenius, H. et al. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
10. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095–1099 (2001).
11. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
12. Deleye, L. et al. Performance of four modern whole genome amplification methods for copy number variant detection in single cells. *Sci. Rep.* **7**, 1–9 (2017).
13. Mathieson, W. et al. A critical evaluation of the PAXgene tissue fixation system morphology, immunohistochemistry, molecular biology, and proteomics. *Am. J. Clin. Pathol.* **146**, 25–40 (2016).
14. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

15. Chapman, M. S. et al. Lineage tracing of human embryonic development and foetal haematopoiesis through somatic mutations. Preprint at https://www.biorxiv.org/content/10.1101/2020.05.29.088765v1 (2020).
16. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
17. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
18. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
19. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
20. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
21. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
22. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164–e164 (2010).
23. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
24. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
25. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
26. Broad Institute. Picard Tools. http://broadinstitute.github.io/picard/.
27. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
28. Kozarewa, I. et al. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nat. Methods* **6**, 291–295 (2009).
29. Loo, P. V. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
30. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).

Acknowledgements

We acknowledge Y. Hooks and K. Roberts (Wellcome Sanger Institute) for assistance in tissue preparation and sectioning, P. Scott (Wellcome Sanger Institute) for advice and training for LCM, D. Rasil (Royal Papworth Hospital) for pathology support and R. Hoogenboezem (Erasmus University Medical Center) for assistance in software development. Finally, we would like to thank L. Apone, K. McKay, V. Panchapakesa and E. Dimalanta (NEB) for assistance in optimizing the library preparation process. This work was supported by the Wellcome Trust. S.F.B. was supported by the Swiss National Science Foundation (P2SKP3-171753 and P400PB-180790). T.M.B. was supported by a Cancer Research UK Grand Challenge Award (C98/A24032). M.A.S. was supported by a Rubicon Fellowship from the Netherlands Organisation for Scientific Research (019.153LW.038). L.M. is a recipient of a CRUK Clinical PhD Fellowship (C20/A20917) and a Pathological Society of Great Britain and Ireland Trainee Small Grant (grant reference no. 1175). I.M. is funded by Cancer Research UK (C57387/A21777). P.J.C. is a Wellcome Trust Senior Clinical Fellow (WT088340MA).

Author contributions

P.E., L.M., M.A.S., T.M.B., A.R.J.L. and A.C. wrote the manuscript with contributions from all authors. P.E., L.M., R.O., B.F., S.F.B. and H.L.S. devised the protocol for laser-capture microscopy, DNA extraction and sequencing of microbiopsies. M.A.S. developed filters to remove fragmentase-associated artifacts. L.M. and T.M.B. performed LCM experiments, data curation and analysis. T.C. developed the 'unmatched normal' filtering strategy. A.R.J.L., M.R.S., I.M. and P.J.C. assisted with data analysis. P.J.C. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41596-020-00437-6.

Correspondence and requests for materials should be addressed to P.J.C.

Peer review information *Nature Protocols* thanks Edwin Cuppen, Subhajyoti De and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 8 October 2019; Accepted: 8 October 2020;
Published online: 14 December 2020

Related links

Key references using this protocol

Brunner, S. et al. *Nature* **574**, 538-542 (2019): https://doi.org/10.1038/s41586-019-1670-9

Lee-Six, H. et al. *Nature* **574**, 532-537 (2019): https://doi.org/10.1038/s41586-019-1672-7

Moore, L. et al. *Nature* **580**, 640-646 (2020): https://doi.org/10.1038/s41586-020-2214-z