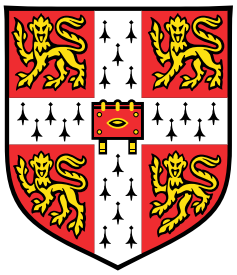
# Somatic evolution in healthy and chronically inflamed colon and skin



**Sigurgeir Olafsson**

Supervisors: Dr. Carl Anderson
Dr. Peter J. Campbell

Wellcome Sanger Institute
University of Cambridge

This dissertation is submitted for the degree of
*Doctor of Philosophy*

Hughes Hall College

February 2022

# Declaration

I hereby declare that this thesis is my own original work. Some of the figures and parts of the text have been published in the pages of scientific journals as described in the beginning of each chapter. No part of this thesis has been submitted for consideration for any other degree or qualification in this, or any other university.

Modern scientific work is usually collaborative in nature and I have been fortunate to enjoy the help and support of many skilled people during the course of my PhD. I highlight their contributions in relevant places in the text. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Sigurgeir Olafsson
February 2022

# Acknowledgements

With sincere gratitude, I thank my two doctoral supervisors, Drs. Carl Anderson and Peter Campbell, for their support and advise during the last four years. Many thanks as well to my colleagues in CASM and department of Human Genetics who helped and encouraged me along the way. I also want to thank the Wellcome Trust for their financial support and Hughes Hall college for creating a social structure to which I enjoyed belonging.

Special thanks go out to the IBD and psoriasis patients who donated samples for research. I sincerely hope the work presented herein will meaningfully add to our understanding of these diseases. In the writing of this thesis I used a Latex template created by Krishna Kumar in 2013. I used version 2.3.1 of the template, released 24 May 2017. I am grateful to Krishna for publishing and maintaining this resource.

# Abstract

The human body is made up of trillions of cells which cooperate to reproduce their genetic material. While all the cells are a part of a whole, each is also an individual and will selfishly give rise to a clonal expansion of cells within a tissue given the chance, even to the detriment of the organism. This thesis discusses the evolutionary forces acting on cells within the body, specifically on epithelial cells in the colon and skin.

After a general introduction of the evolutionary forces acting on normal cells and the methods used to study them, Chapter 2 focuses specifically on genetic drift within the colon, where clones expand through the process of crypt fission. I apply a statistical framework called Approximate Bayesian Computation to estimate the crypt fission rate in the normal colon and in individuals with Familial adenomatous polyposis (FAP). I estimate the rate of crypt fission to be one every 27 years in the normal colon and one every 13 years in (FAP).

In Chapter 3, I describe somatic evolution in the colon under conditions of chronic inflammation. I used whole-genome sequencing of individual colonic crypts from patients with inflammatory bowel disease (IBD) to show that the IBD-colon is characterized by a higher mutation burden and larger clonal expansions than the healthy colon. I also show that mutations in immune-related genes, including *PIGR*, *ZC3H12A* and genes in the interleuking 17 and toll-like receptor pathways, are under positive selection in the colons of IBD patients and may contribute to the disease pathogenesis.

In Chapter 4, I focus on the skin. I performed whole-exome sequencing of microbiopsies of epidermis from patients with psoriasis, a second chronic inflammatory disease. In contrast to IBD, I did not find increased mutation burden and clonal spread in psoriasis, except when the skin had been treated with psoralens + UVA (PUVA) phototreatment. The selection landscape of psoriatic skin resembles that of normal skin, and mutations in *NOTCH1*, *FAT1*, *TP53*, *PPM1D* and *NOTCH2* are positively selected. *ZFP36L2* was the only gene found to be enriched in mutations that has not been previously reported in normal skin, but it is as yet uncertain if selection of *ZFP36L2* mutant cells is a feature specific to psoriatic skin or not.

Finally, Chapter 5 discusses my findings in the broader context of cancer and complex-trait genomics. I discuss how a causal relationship between somatic evolution and non-neoplastic diseases may be established and the different ways somatic evolution may affect disease

progression for good or ill. I further discuss how to design a study to search for germline determinants of somatic evolution and the need for developing methods to enable such studies to be conducted at scale.

# Table of contents

# List of figures

# List of tables

# Chapter 1

# Introduction

Figure 1.7 and parts of the text in this chapter (section 1.3 in particular) have been previously published in Trends In Genetics in an article titled “Somatic mutations provide important and unique insights into the biology of complex diseases” by myself and Carl Anderson.

## 1.1 The evolving body

In 1859, Charles Darwin published his theory of evolution and origin of species by means of natural selection. Darwin presented three postulates from which evolution by natural selection is a logical outcome. The postulates are (1) Individuals in a population are variable; (2) This variation is at least partially heritable; (3) Individuals who carry favorable variations are more likely to raise viable offspring than those who don’t (Darwin, 1876). Darwin sought to explain species evolution but as I will discuss, his work also applies on a smaller scale, to the evolution of cells within a body.

A human is a partnership of about 37 trillion cells (Bianconi et al., 2013), all derived from one, which come together to cooperatively reproduce their genetic material. Although originating from a single progenitor, the individual cells that make up this system are not genetically identical. As a cell divides to produce two daughter cells, each daughter accrues a small number of mutations which distinguish it both from her mother and sister. These variations are passed on to subsequent generations as more mutations are introduced in each round of replication. A group of cells that have separated late from their mutual ancestor and have many more mutations in common than are unique to each are often referred to as a clone.

An organism can sustain a finite number of cells, which introduces a struggle for space and existence within the body, homologous to that between individuals of a species living

in a finite environment. Competition between cells is a zero-sum game and (with the exception of neoplastic growth) expansion of a clone derived from one cell results in the reduction or elimination of competitor clones. Darwin's postulates for evolution of species in an environment by natural selection therefore also apply to cell types in a body since (1) Individual cells are variable; (2) much of this variation is the result of heritable genetic variation; and (3) cells that reproduce the most are those carrying favourable mutations.

## 1.2 Evolutionary forces

In classic evolutionary theory, species evolution is shaped by four evolutionary forces which affect allele frequencies in populations of individuals. These are mutations, natural selection, genetic drift and gene flow. As the cells of the body don't exchange genetic material, gene flow does not influence the evolution of somatic cells (ignoring the edge cases of gene therapy, viral insertions and organ donation). What follows is a brief description of how each of the remaining three evolutionary forces affects somatic cells.

### 1.2.1 Mutagenesis

Inherited changes to the base composition of a cell are the substrate of evolution. This section will describe somatic mutagenesis, processes by which mutations accumulate in somatic cells. I will use the term 'somatic mutations' to describe everything from a single-base substitution change to whole-genome amplification, although different types of somatic mutations have different causal mechanisms which need to be considered in turn. While a comprehensive description of all mutagenic mechanisms affecting somatic cells is beyond the scope of this thesis, I discuss some examples of endogenous and exogenous mutational mechanisms of particular relevance to my own work. I also highlight the important interplay between the damaging agents themselves and the DNA repair and replication pathways that are called upon by the cell to resolve the damage.

Mutagens are often very base- and sequence- specific and this specificity means that each mutational process leaves a characteristic imprint on the genome of the affected cell. This pattern is termed a mutational signature, and can serve as a physiological documentation of the biological history of the cell, capturing both the mutagens that have affected the genome and quantifying the exposure to each (Alexandrov et al., 2015, 2020, 2013a,b; Nik-Zainal et al., 2012).

To define mutational signatures, mutations are assigned to mutually exclusive classes. For example, single base substitutions are commonly classified according to the base substitution C>A, C>G, C>T, T>A, T>C, T>G (all substitutions are referred to by the pyrimidine of the mutated base pair) and the bases 5' and 3' of the mutated base, yielding a total of 96 classes. The standard way of plotting a signature is with a bar-plot which has the mutation classes on the x-axis and the frequency of each mutation type on the y-axis. A number of mutational signatures have been described as part of the Catalogue of Somatic Mutations in Cancer (COSMIC) and the Pan-Cancer Analyses of Whole Genomes (PCAWG) efforts to sequence and characterize the genomes of different cancer types (Alexandrov et al., 2020, 2013a). As I describe some important mutagenic processes, I will also often mention their corresponding signatures. Methods for mutation calling are described in section 1.4.2 and methods for signature extraction in section 1.4.3.

### Substitutions and indels

#### Endogenous mutagenic processes

Base substitutions can result from mutational processes that are either endogenous or exogenous to the cell. Let us start by discussing endogenous factors. The DNA sequence may be altered as a result of direct chemical changes. For example, an important endogenous mutation process is deamination, which occurs spontaneously across all DNA bases that contain primary amine groups. In particular, 5-methylcytosines at CpG sites are prone to hydrolytic deamination, which gives rise to C>T mutations at a rate which correlates closely with the rate of cell division (Alexandrov et al., 2015). This is the process which produces COSMIC and PCAWG single base substitution signature 1 (SBS1 or signature 1 for simplicity, Figure 1.1).

Chemical changes may also be catalyzed by enzymatic activity in the cell. The best characterized example of this is perhaps the catalyzation of cytosine deamination to uracil by apolipoprotein B mRNA editing enzyme (APOBEC). This family of enzymes, which fulfills diverse physiological functions in the cell, including restriction of retroviruses and mobile retroelements, have been shown *in vitro* and *in vivo* to have a preference for TpCpN motifs in stretches of single-stranded DNA (Nik-Zainal et al., 2012; Suspène et al., 2011). This gives rise to mostly C>T and C>G mutations which make up COSMIC signatures 2 and 13 (Figure 1.2). What determines which type of mutation occurs is unknown, but one might speculate that subtly different repair mechanisms may play a role.

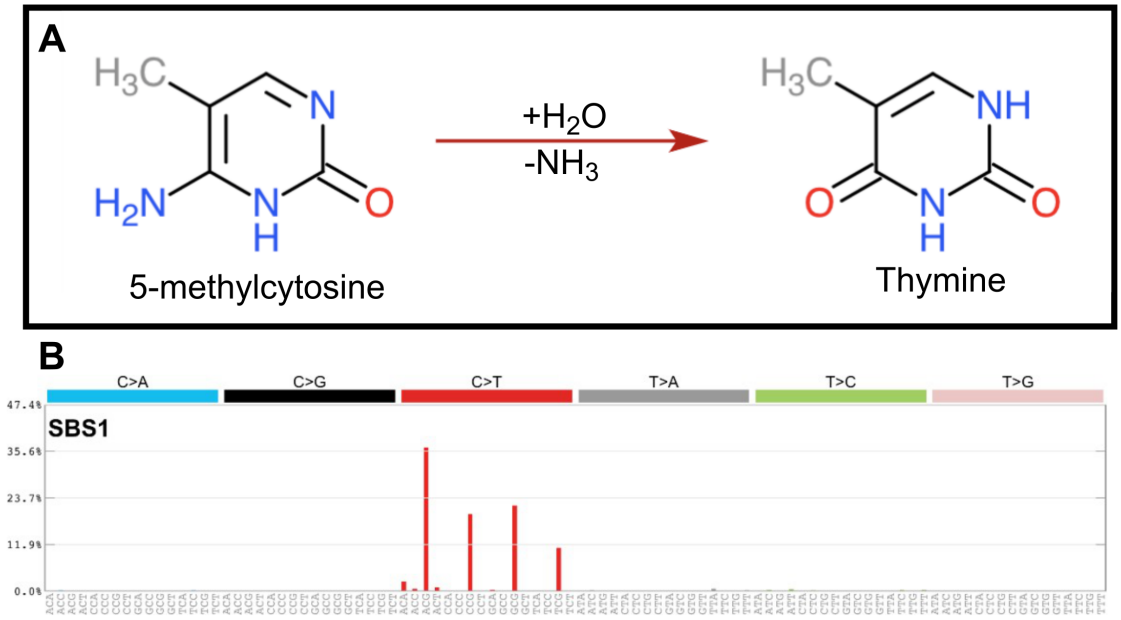



Fig. 1.1 **Deamination at CpG sites.** A) 5-methylcytosine undergoes spontaneous deamination of the primary amine group to yield thymine. B) Single base substitution signature 1 (SBS1) which is thought to be the result of this process. Figure from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

Endogenous mutational processes also include DNA replication errors. The two primary polymerases responsible for replication of the genome are polymerases $\varepsilon$ and $\theta$ which synthesise the leading and the lagging strand, respectively. Although the error rate of these enzymes is low, estimated at 1 error every 10 million nucleotides after accounting for intrinsic proofreading capacity (Shevelev and Hübscher, 2002), the size of the genome is such that some mistakes are inevitably made. Additionally, somatic and germline mutations in these proteins can increase the rate of mutagenesis by orders of magnitude. A well established example is the mutation profile seen in some hypermutated colorectal carcinomas that carry polymerase $\varepsilon$ mutations. These are characterized by a very large number of C>A and C>T mutations at TpCpG and TpCpT sites and which constitute mutational signatures 10a and 10b (Alexandrov et al., 2013a) (Figure 1.3).

Polymerase errors during normal DNA synthesis can cause somatic indels as well as substitutions. Areas consisting of long stretches of homopolymers are particularly prone to errors caused by polymerase slippage and COSMIC indel signatures 1 and 2 are attributed to slippage of the nascent and the template strands respectively (Figure 1.4). They exhibit clock-like properties and high correlations with substitution signature 1 (Alexandrov et al.,

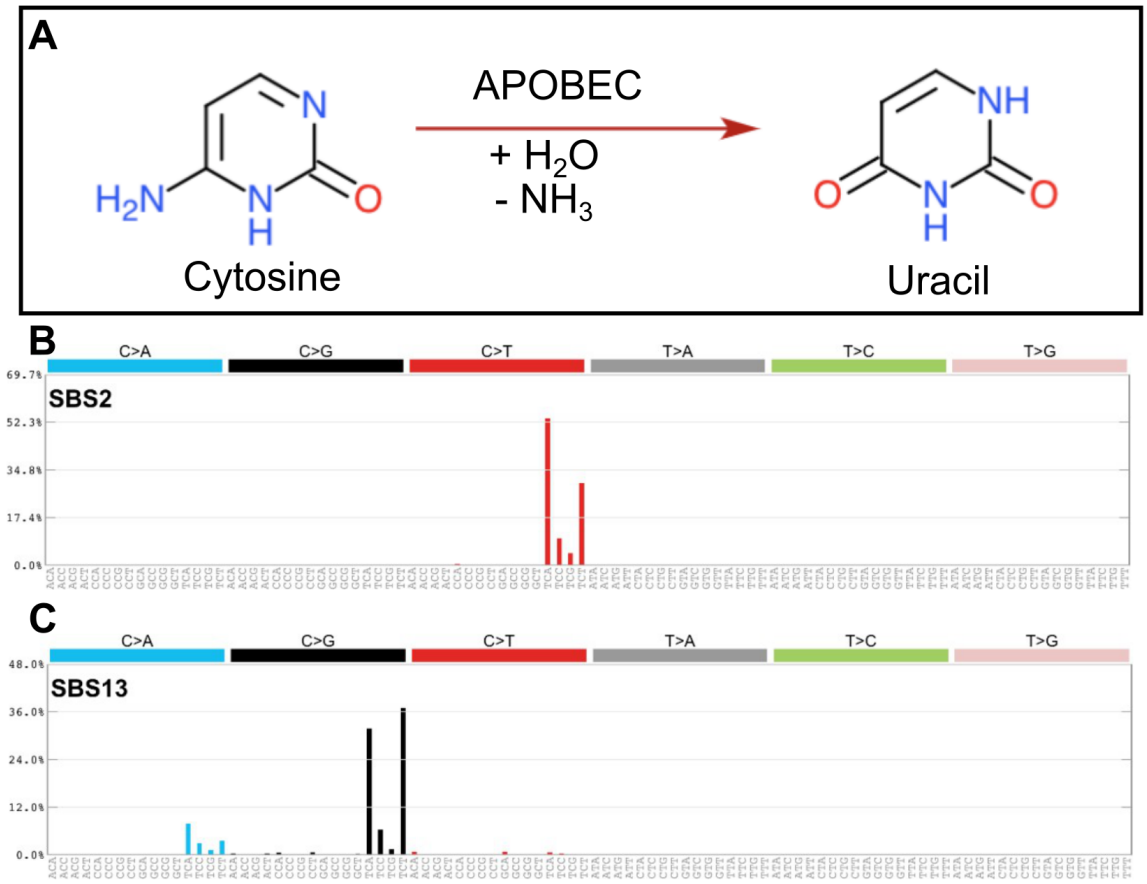



Fig. 1.2 **APOBEC catalyzed mutagenesis.** A) APOBEC catalyzes the deamination of cytosine to uracil. B and C) Single-base substitution signatures 2 and 13. Characterized by cytosine deamination at TpC sites and attributed to the apolipoprotein B mRNA editing enzyme (APOBEC) family of proteins. Figures from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

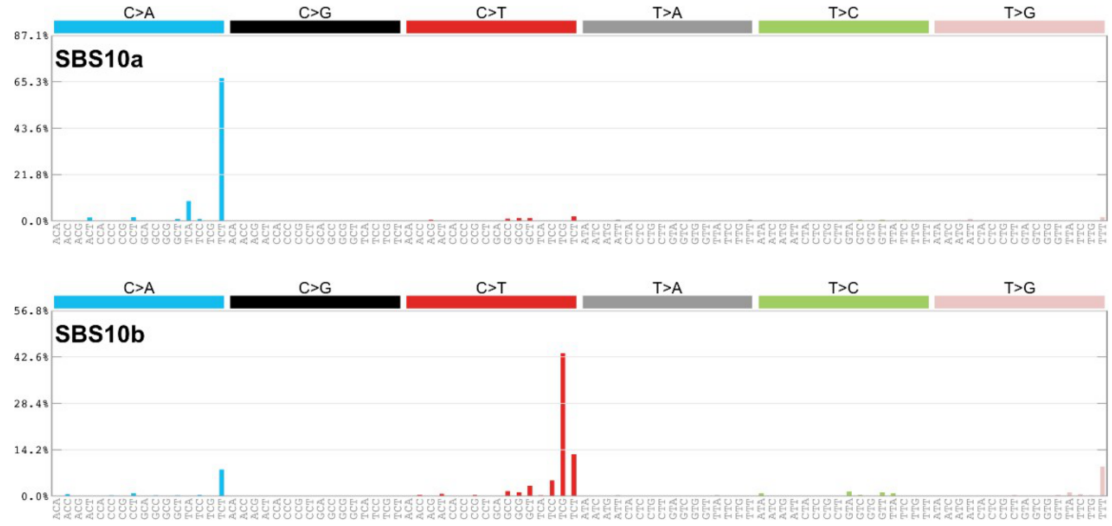



Fig. 1.3 **Single base substitution signatures 10a and 10b.** Both are associated with mutations in DNA polymerase $\varepsilon$. Figures from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

2020).

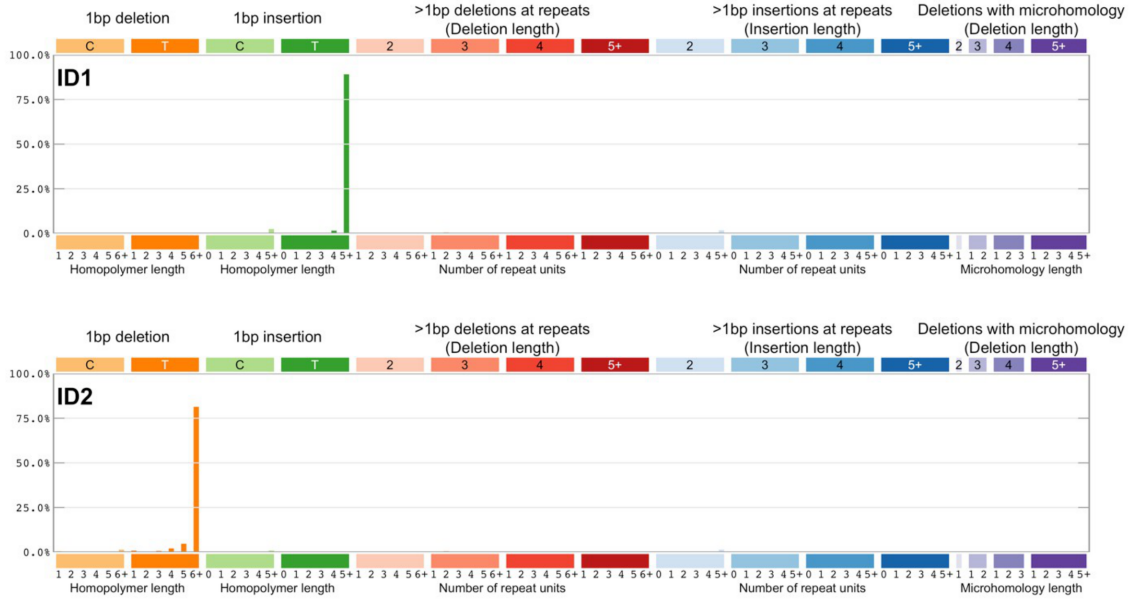



Fig. 1.4 **Indel signatures 1 and 2.** Both are attributed to DNA polymerase slippage during replication of homopolymer regions of the nascent and template strands, respectively. Figures from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

### Exogenous mutational processes

Many external physical and chemical factors are mutagenic to the cell. Among the physical factors, the one most important to this work is doubtless ultraviolet (UV) light. UV rays cause the formation of covalent bonds between neighboring pyrimidines resulting in pyrimidine dimers, pyrimidine photoproducts or cyclobutane pyrimidine dimers, which halt the replication machinery until repaired by transcription-coupled nucleotide excision repair (TC-NER) (see below). The reliance on TC-NER to correct damaged bases results in considerably more mutations occurring on the non-transcribed strand than on the transcribed (Alexandrov et al., 2020). Mechanistic studies in cell lines, bacteria and rodents have shown that different UV wavelengths are associated with different types of mutations (Pfeifer et al., 2005). UVA (320-400 nm) is thought to cause mainly C>T and C>A substitutions and small tandem base deletions following oxidative damage. UVA is not absorbed by the DNA itself but the damage is done by other molecules which absorb the energy and may form radicals, including reactive oxygen species. The higher energy wavelengths, UVB (280–320 nm) and UVC (200–280 nm), are absorbed by the DNA itself and frequently cause the formation of photoproducts like cyclobutane pyrimidine dimers and (6-4) pyrimidone photoproducts (6-4 PPs). Other photoproducts like purine dimers and pyrimidine mono-adducts can also be formed in small quantities (Pfeifer et al., 2005).

A handful of mutational signatures of UV-light exposure have been extracted, reflecting the complexity of this mutagen. The signatures capture different types of single base substitutions, deletions at tandem pyrimidines and double-base substitutions at pyrimidine sites (Figure 1.5).

Exogenous mutagens of chemical origin include such well known agents as tobacco smoke and aflatoxin. However, drugs taken for treatment of cancer and complex diseases can affect the mutation burdens in recipients (Pich et al., 2019). This thesis focuses on non-neoplastic tissue and the patient cohorts used include only the rare cancer survivor. However, some of the treatments prescribed for the two chronic inflammatory diseases which will feature in this work, inflammatory bowel disease and psoriasis, have mutagenic effects. For example, COSMIC single base substitution signature 32 has been attributed to azathioprine treatment (Inman et al., 2018) for immunosuppression but azathioprine and other thiopurines are also used as immunosuppressants for the treatment of IBD. Psoralens are phototherapeutic agents used in the treatment of psoriasis and although a psoralen signature is not described in COSMIC, they are known to lead to pyrimidine mutations at TpA motifs (Esposito et al., 1988; Papadopoulo et al., 1993; Yang et al., 1994; Zhen et al., 1986).

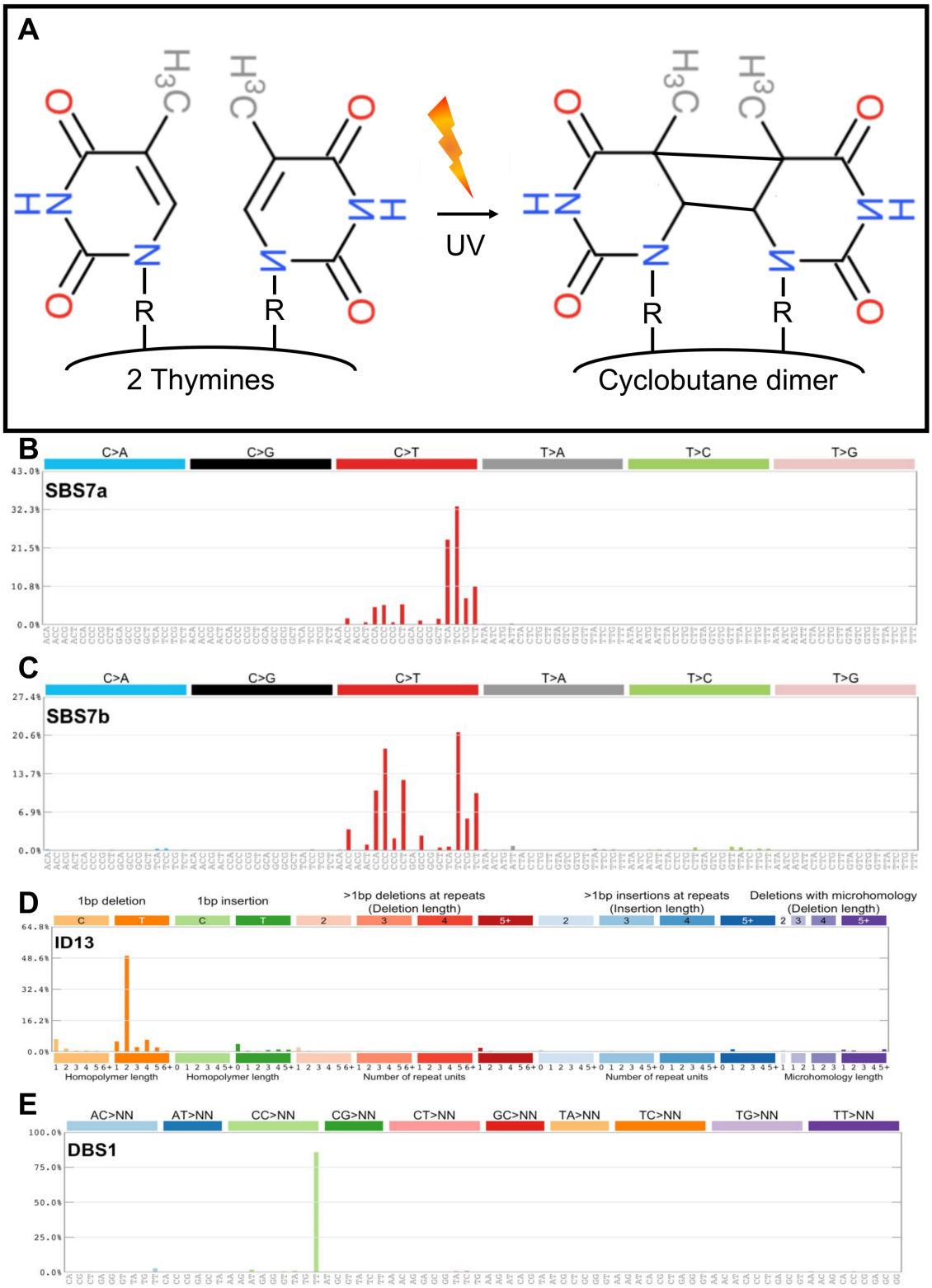



Fig. 1.5 **Mutations associated with ultraviolet light.** A) UV light causes the formation of bonds between carbons 5 and 6 of adjacent thymines to form a cyclobutane pyrimidine dimer. B) and C) The two most prominent single base substitution signatures associated with UV-light exposure. D) An indel signature characterized by single base deletions at tandem thymine sites and associated with UV light exposure. E) A double base substitution signature characterized by substitution of tandem cytosines and associated with UV light exposure. Figures B) - E) from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

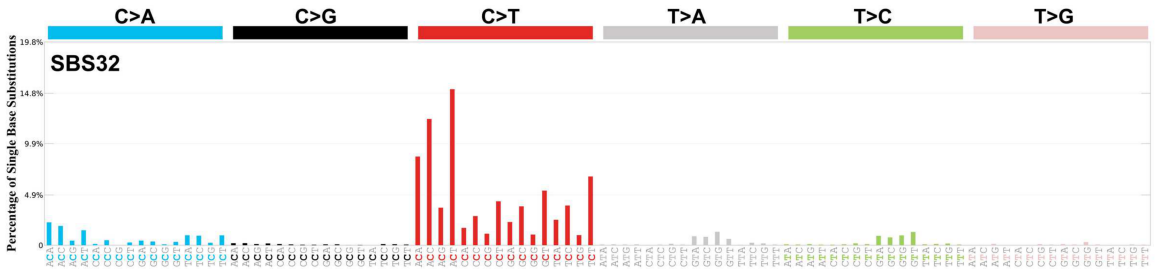



Fig. 1.6 **Single base substitution signature 32.** The proposed aetiology of this signature is azathioprine treatment for immunosuppression. Figure from https://cancer.sanger.ac.uk/signatures/, accessed in March 2020.

### DNA repair pathways

An observed mutational pattern is the combined effect of a mutagen and the DNA repair mechanism the cell employs to fix the damage (Volkova et al., 2020). Defects in the repair machinery itself lead to distinct mutation patterns. Germline variants in *BRCA1/2* for example, leave the cell unable to repair double strand breaks via homologous repair. Instead, breaks are fixed by non-homologous end joining which results in numerous substitutions and indels (Powell and Kachnic, 2003), and increased cancer risk in breast and ovarian tissue in particular. More relevant to the tissues studied herein, mutations in the mismatch repair pathway are observed in a fraction of colorectal carcinomas and are often associated with hypermutator phenotypes and as many as seven distinct mutational signatures (Alexandrov et al., 2020). The mutation spectrum of mismatch repair deficiency in all likelihood depends on an interaction between the specific repair defect and the mutagens the cell is exposed to. Finally, one mechanism of lesion detection is when RNA polymerase II is arrested during transcription of a gene. This invokes transcription-coupled repair (TCR), a sub pathway of nucleotide excision repair (NER) and results in the transcriptional strand biases observed for several of the COSMIC and PCAWG signatures like single base substitution signatures 4, 7, 16 and 24 (Alexandrov et al., 2020, 2013a).

### Genome wide patterns of mutagenesis

In addition to sequence context, more “higher level” variables also affect the chances of a mutation occurring. Mutation rate negatively correlates with the expression level of genes (Lawrence et al., 2013) and with chromatin accessibility and modification (Polak et al., 2015). Furthermore, regions of the chromosomes that are replicated late in S-phase show a marked increase of somatic mutations, possibly due to these regions lagging in single-stranded form as the dNTP pool is exhausted (Polak et al., 2015; Stamatoyannopoulos et al., 2009).

Together, expression levels, chromatin accessibility and replication timing have been shown to explain up to 86% of the variance in substitution rates along cancer genomes (Polak et al., 2015), which has important implications for driver identification, as discussed in section 1.4.5.

Finally, cancer mutations due to exogenous mutagens tend to show pronounced strand asymmetries. In large regions of the genome, sometimes entire chromosomes, mutagens have only affected one strand. This phenomenon, termed lesion segregation, results from a combination of segregation of unrepaired lesions during mitosis and selection (Aitken et al., 2020). Mutagens generate DNA lesions on both strands, but these are not immediately repaired and may be passed on through several cell divisions. Each strand segregates to a different daughter cell so that each has its own set of lesions derived from the mutagen. If the different sets of mutations confer on each daughter cell different "fitness" (for example if one of the cells now has a driver mutation) the fitter daughter dominates in subsequent expansions, which gives rise to the observed strand asymmetry (Aitken et al., 2020).

### Structural variants

I will use the term structural variation to describe large scale deletions, translocations, inversions and amplifications affecting parts of chromosomes, whole chromosomes or multiple chromosomes. I will also use it to describe somatic retrotranspositions and viral insertion events. These events occur through a multitude of mechanisms, including chromosome segregation errors following erroneous repair of double strand breaks, breakage-fusion-bridge cycles and genome doubling (Ghezraoui et al., 2014; Ly et al., 2019). While mutations often accumulate gradually over time, sometimes catastrophic mutational events, termed chromothripsis, are observed that cause a huge number of rearrangements and copy number variations in a single replication cycle. These events are pervasive across cancers (Cortés-Ciriano et al., 2020) but rare in non-neoplastic tissue, although they have been reported in the liver (Brunner et al., 2019).

In the PCAWG analysis of 2,658 cancers across 38 tumour types, the authors divided the structural variants into mutually exclusive categories by variant type, size, replication timing and whether or not they were found within fragile sites of the genome. Although this relies on arbitrary cut-offs, 16 mutational signatures of structural variants could be identified (Li et al., 2020b). As structural variants are rare in non-neoplastic tissues, I do not consider these signatures further in the work presented in this thesis.

### 1.2.2 Natural selection

Most adult tissues are maintained by a population of stem cells which undergo stochastic divisions in such a way that on average, the number of active stem cells remains constant (Klein and Simons, 2011). The fixed size of an adult tissue shapes differences in evolutionary dynamics between cancer and non-neoplastic tissue. In the latter, clones derived from single stem cells compete for space in a zero-sum game where the expansion and proliferation of a particular clone is compensated for by the loss of others. In the following sections, I describe how genetic changes affect the evolutionary dynamics of somatic cells.

#### Positive and negative selection

Species evolution, including human evolution, has been dominated by negative selection. Evolution of somatic cells within the body however, is dominated by positive selection, suggesting that a majority of the genes in the genome are dispensable for any given cell type (Martincorena et al., 2017). A commonly used method to detect selection in sequencing datasets is to look at the ratio of the mutation rate of non-synonymous (dN) and synonymous mutations (dS). The synonymous mutations are assumed to be selectively neutral to the cell but if some of the non-synonymous mutations found in a gene are under positive selection, this will result in dN/dS >1 for that gene. Similarly, dN/dS <1 for a gene indicates negative selection of mutations in the gene. dN/dS ratios and a software implementation for their study in somatic cells are further described in section 1.4.5.

While dN/dS ratios are useful for identifying genes and pathways in which mutations confer advantage to the carrier, they are not properties of individual mutations and do not directly translate into the selection coefficients often used in evolutionary genetics. Williams et al showed how dN/dS ratios can be combined with clone size information to derive a selection coefficient for each mutation (and thus estimate the distribution of fitness effects for the gene) (Williams et al., 2020).

#### Drivers and passengers

Somatic mutations may be classified as driver mutations or passenger mutations. A driver mutation, or driver, may be defined as any mutation that confers upon a cell a selective advantage within a population of cells. In contrast, passenger mutations are those which are selectively neutral to the cell. In the classical model of cancer development, normal cells gradually accumulate driver mutations over a person's lifetime, ultimately resulting in cancer formation (Fearon and Vogelstein, 1990). Drivers are thus often acquired years or

decades before cancer diagnosis (Gerstung et al., 2020). However, the vast majority of the somatic mutations found in a cell are generally passengers. A recent large scale analysis of multiple solid tumour types found 4.6 drivers per tumour on average, among thousands of passengers (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). The driver landscape of normal tissues varies markedly by tissues. In some, for example, oesophagus and endometrium, 70-100% of cells in middle aged individuals carry one or more driver mutations (Martincorena et al., 2018; Moore et al., 2020), while in other tissues, like colon, liver and prostate, this fraction is <5% (Brunner et al., 2019; Grossmann et al., 2021; Lee-Six et al., 2019). The fitness of a mutant clone is only defined relative to its neighbours in the tissue and there is mounting evidence that in tissues where a very large fraction of cells carry drivers, like the oesophagus, initial exponential growth of a driver-carrying clone is followed by reversion to near neutral drift as cells collide with others of similar 'fitness' (Colom et al., 2020; Martincorena et al., 2015). The mutational and selection landscape of normal tissues is further discussed in section 1.3 below.

Driver mutations are commonly loss of function (LoF) mutations in tumour suppressor genes or gain of function mutation in oncogenes. Gain of function (GoF) mutations include structural variants giving rise to fusion genes, enhancer hijacking and oncogene amplification (Li et al., 2020b; Rodriguez-Martin et al., 2020), but also point mutations. Single nucleotide changes that are GoF often disrupt ubiquitin sites and prevent or reduce ubiquitin-mediated proteolysis (Martínez-Jiménez et al., 2020), but they can also directly change the function of the protein. A classic example is the *KRAS* G12 site, which is frequently mutated in multiple cancer types, including cancers of the pancreas, colon and small intestine (Forbes et al., 2017). The KRAS protein switches between inactive and active states via binding to guanosine diphosphate (GDP) and guanosine triphosphate (GTP), respectively. Mutations of the G12 site disrupt the hydrolysis of GTP, causing the protein to be perpetually "stuck" in the active state, signalling to the cell to continue dividing (Liu et al., 2019).

Large-scale analyses of whole genomes have failed to identify many driver mutations outside coding regions (Rheinbay et al., 2020). The most notable class of non-coding drivers are mutations in the promoter of the *TERT* gene, which encodes the catalytic subunit of telomerase. Such mutations are frequently found across a range of cancers and are thought to recruit transcription factors to the promoter that normally don't regulate *TERT* expression, but once recruited increase the expression of the gene (Bell et al., 2016). Rheinbay et al argue the paucity of other non-coding drivers is a result of differential fitness effects of coding and non-coding mutations and suggest that point mutations rarely assert sufficient effect on the

function of regulatory elements to be selected for. However, the Genotype-Tissue expression project identified a large number of germline variants which alter the expression of the gene by a standard deviation or more (Chiang et al., 2017) and I personally remain uncertain as to whether I believe the paucity of non-coding drivers isn't simply due to them being more difficult to detect than coding drivers.

Some drivers start off as passengers and only become drivers once the selection landscape is changed. This can, for example, happen as a result of cancer treatment, where a mutation which otherwise is neutral may allow the cell to escape the effects of therapy and cause a relapse of the cancer (Pich et al., 2021; Wong et al., 2014). One can also imagine an alteration of the selective landscape of a tissue by disease as well as therapy, as will be discussed in later chapters.

### 1.2.3 Genetic drift

The third evolutionary force operating on normal cells is genetic drift. The impact of genetic drift depends largely on the cellular structure of the tissue. In the skin, where many stem cells line a two dimensional basal layer (see Chapter 4.1.2), some clones will grow with age as others are lost simply through neutral drift. However, the chances that a single cell takes over a larger patch of tissue by drift alone are small.

In the colon, a small number of stem cells reside at the bottom of structures called colonic crypts. Drift influences the spread of these stem cells in two ways: Firstly, cycles of neutral sweeps repeatedly occur with the progeny of a single stem cell taking over the entire crypt, thus creating a clonal unit of cells (Lopez-Garcia et al., 2010; Snippert et al., 2010). Secondly, the crypts undergo a process called crypt fission and "divide" to populate the gut. Both of these processes are further described in the Introduction of Chapter 2, where I discuss the crypt structure and dynamics of the normal colon.

## 1.3 Somatic evolution in normal tissues

The major challenge to the study of somatic evolution in solid tissues in the non-neoplastic state is usually the highly-polyclonal structure of the tissue. This means that when whole tissue biopsies are sequenced, nearly all somatic mutations are present in such a small fraction of cells that they are indistinguishable from sequencing errors. Early studies would seek to overcome this by sequencing to very high depth and such studies are still being published.

However, during my PhD I have witnessed (and contributed to) a large increase in studies that use more sophisticated methods that allow more insights to be gained. Three methods in particular stand out (Figure 1.7). The first involves the expansion of single cells in culture followed by sequencing. This method has for example been used to study somatic mutations in healthy and diseased colon (Blokzijl et al., 2016; Nanki et al., 2020), in fibroblasts and melanocytes from skin (Abyzov et al., 2017; Tang et al., 2020), hematopoietic stem cells (Lee-Six et al., 2018), skeletal muscle (Franco et al., 2018) and in the bronchial epithelium of the lung (Yoshida et al., 2020). This method has the advantages that samples have high clonality and cells found to carry mutations of interest are not all destroyed but can be subjected to experimental functional assays. Disadvantages include that culturing cells can be labour intensive and difficult and many cell types cannot be expanded in culture using existing methods. Furthermore, the spatial information between stem cells is lost as the tissue is disassociated. Finally, culturing can affect the mutational landscape of the cells, both because selective forces may operate in the culture that favour the expansion of cells carrying specific mutations and because cells accrue mutations during culturing (Kucab et al., 2019).

The second method, laser capture microdissection (LCM), is the one I have used in the work presented in this thesis. This involves using a combination of a laser and a microscope to dissect small populations (often 100-2000) cells, usually comprising some distinct morphological features. It has the advantage that the spatial relationships between the groups are known, which can inform about clonal spread over millimeters or centimeters as well as on locally active mutational processes. The method relies on the ability to identify at least semi-clonal populations of a few hundred cells in the tissue and is not feasible for highly polyclonal tissues like muscle or brain (Ellis et al., 2021). LCM has been employed at the Sanger to establish the mutational landscape of tissues like colon (Lee-Six et al., 2019), endometrium (Moore et al., 2020), liver (Brunner et al., 2019), urothelium (Lawson et al., 2020), prostate (Grossmann et al., 2021) and more. The small number of cells dissected means that sequencing must be performed on very little input material. The strategy used at the Sanger institute to obtain good sequencing libraries from low DNA inputs is discussed in section 1.4.1.

The third method is single cell sequencing. Mutation calling from single cells is attractive because it would allow mutation calling even in highly polyclonal tissues like the brain and would enable us to study the differences between the mutation profiles of stem cells and their differentiated progeny. Methods have been developed to call mutations both from scDNA and scRNA datasets (Dong et al., 2017; Luquette et al., 2019; Vu et al., 2019) but

all suffer from significant limitations, the most significant of which are associated with the need for exponential amplification like multiple displacement amplification (MDA) in single cell sequencing. This process is associated with a high rate of artificial chimeric DNA molecules which result in a very high fraction of false positive mutation calls (Dong et al., 2017; Luquette et al., 2019). Early studies estimated that the number of false positive calls was an order of magnitude greater than the number of true positives (Zong et al., 2012). Methods have since been developed to improve the quality of the data by various means but the false positive rate remains high. Like when cells are expanded in culture, single cell sequencing also loses information about spatial distribution and histopathological features of the cells.

During my PhD, the somatic mutation landscapes of many diverse tissues have been established and the effects of some environmental exposures and diseases have been characterized. I describe the findings for normal colon and skin in some detail in the introductions to Chapters 3 and 4 but otherwise, general insights from these studies can be summarized as follows:

- Mutation burden varies substantially by tissue but generally increases linearly with age across tissues. Mutation burden is not necessarily linked with the mitotic activity of the tissue. For example, hematopoietic stem cells accumulate about 14 mutations per year (Lee-Six et al., 2018) while non-dividing cortical neurons accrue 17 mutations per year (Abascal et al., 2021).

- There is within patient and between patient variation in mutation burden which is as yet unexplained. Adjacent cells can be differentially affected by external mutagens like tobacco smoke (Yoshida et al., 2020) but also endogenous mutagens like APOBEC activation (Lawson et al., 2020).

- Excluding hypermutators, many tissues show a similar burden of substitutions to cancers of those tissues. However, structural variants and chromosomal abnormalities are rare in normal tissues compared with cancers, but can increase in frequency in diseased states like cirrhosis (Brunner et al., 2019).

- The cancer driver frequency varies markedly by tissue type, from less than 1% of cells in the colon and prostate (Grossmann et al., 2021; Lee-Six et al., 2019) to essentially every cell in the oesophagus and endometrium of individuals over 50 years of age (Martincorena et al., 2018; Moore et al., 2020).

- Some genes are found to be mutated more often in normal tissue than in cancers and vice versa. For example, *NOTCH1* mutations seem to be more common in normal

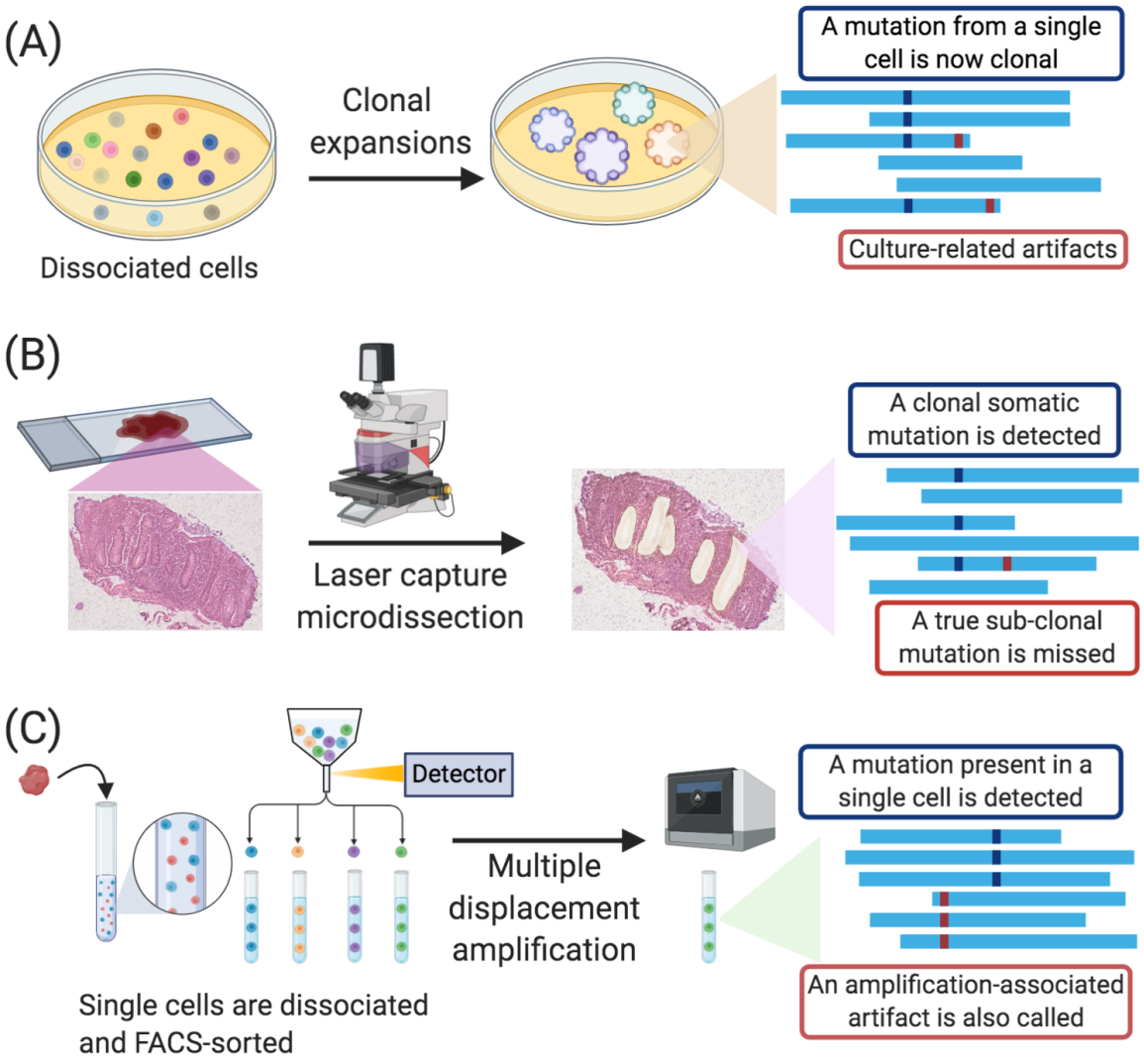



Fig. 1.7 **Methods for the study of somatic mutations.** (A) Expansion of single cells in culture followed by sequencing. (B) Laser capture microdissection of tissue sections can isolate clonal or semi-clonal populations of cells that can be sequenced. (C) Single cell DNA sequencing after dissociation and sorting. Figure created with BioRender.com.

oesophagus than in oesophageal cancers (Martincorena et al., 2018; Yokoyama et al., 2019) while only the subset of urothelial cancer genes that are classified as chromatin remodelers are found to be mutated in the normal urothelium (Lawson et al., 2020), with mutations in other genes likely responsible for malignant transformation.

- Although clones carrying cancer-driver mutations are common in some tissues, the average number of driver mutations per cell in normal tissues is typically lower than that in cancer cells. Normal cells typically carry 0-2 drivers while the median number of cancer drivers in cancers is 4.6 (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020).
- Parallel evolution of clones carrying distinct mutations in the same genes is common in normal tissues. Selection pressures seem to vary, with mutations in different genes favored in different individuals (Lawson et al., 2020).

## 1.4 Methods background

### 1.4.1 Laser capture microdissection and low-input DNA sequencing

The ability to create sequencing libraries from microdissections of a few hundred cells is a cornerstone of all the work presented in this thesis. The LCM workflow starts with tissue fixation. Some histology fixatives, like formalin, cause the formation of cross-links and DNA adducts that have a detrimental effect on the quantity of extracted DNA and cause sequencing errors. The LCM workflow therefore employs alcohol-based fixatives, with ethanol, methanol and commercial preparations like PAXGene Tissue FIX and RNA-Later buffer all having shown good results.

Tissues can either be embedded in paraffin or frozen before sectioning. Histological sections made for pathological assessments are often sectioned at 4 micrometers, but to increase the DNA yield from each section, I have used 10-20 micrometer thick sections in the work presented in this thesis. Furthermore, I often repeatedly cut the same crypt or histological feature, visible in serial tissue sections, in the hopes of increasing DNA yield. This practice is referred to as z-stacking.

After dissection, DNA is extracted using the commercially available Arcturus PicoPure DNA Extraction Kit. The DNA purification is modified in such a way that quantification is omitted (as this would result in loss of DNA) and SPRI beads are integrated into the library

construction workflow to avoid losing DNA to imperfect elution. Rather than using acoustic shearing for fragmentation, the low-input DNA pipeline uses enzymatic fragmentation, as this has been found to result in >10-fold improvement in library yield (Ellis et al., 2021). As in "standard" library preparation, the fragmented DNA is end-repaired, dA-tailed and ligated to adapter sequences. It is then indexed by six cycles of PCR amplification. While whole genome amplification is associated with high error rates in single cell experiments, as described above in section 1.3, the increased input material of the LCM approach compared with single cell sequencing allows for fewer cycles of PCR, which reduces the effect of variable amplification.

### 1.4.2 Mutation calling

#### CaveMan

To call somatic substitutions in samples of colon and skin, I used the CaveMan algorithm. This is an expectation-maximisation algorithm for genotype probabilities. The process can be divided into four steps (not counting post hoc filtering at the end). The steps are split, mstep, merge, and estep (Jones et al., 2016).

The split step involves dividing the genome up into regions to optimize memory and job scheduling time. The precise number of regions varies with coverage of the sample. The second step, mstep, iterates through the positions in a region, pre-calculating parameters that will be used in the Bayesian estimation of genotype probabilities in later steps. These include base quality of the reads, documenting the reference base, called base, read position, mapping quality of reads and the mapping strand. These individual region profiles are merged in the merge step. Finally, the estep uses these metrics, as well as the sequence data itself, to assign each putative mutation site a genotype probability using an expectation maximization algorithm.

The distribution of the observed reads is determined by the unobserved genotype. The likelihood of the data can be estimated in a mixture model as

$$Pr(X_{i,n}) = \sum_{\gamma}(Pr(X_{i,n}|\tau_n = \gamma) \times Pr(\tau_n = \gamma))$$

Where $X_{i,n}$ is the genotype i at position n and $\gamma$ is an element from the set of possible genotypes for the nth base (defined from the reference base and the copy number estimate of

the sample and the normal and depending on whether or not the site is a known germline polymorphic site (in dbSNP)). The E-step of the algorithm estimates the genotype probabilities for all sites sequenced using the parameters calculated in the mstep described above and first assuming that all bases are the reference genotype. The M-step of the EM algorithm in turn updates the mutation rate and SNP rate estimates to iteratively update the base call probabilities. In this way, somatic probabilities are assigned to each base position and a mutation is reported if the base probability exceeds 0.8 for somatic genotypes or 0.95 for germline.

Finally, sites are filtered to remove false positives while retaining true somatic mutations. The mutation list is compared against a panel of 75 unrelated normal samples. This is to remove both common polymorphisms and recurrent errors resulting from misalignment and sequencing artifacts. Two further filters are typically applied to remove mapping artefacts. The median alignment score of reads supporting a mutation is required to be at least 140 and fewer than half of the reads supporting a mutation should be clipped.

### Pindel

To call deletions and insertions in samples of colon and skin, I used a slightly modified version of the PIndel algorithm (Raine et al., 2015; Ye et al., 2009). Pindel can detect large deletion events and small-to-moderate size insertions (size of insertions depends on the length of the reads). It requires paired-end libraries and it works only with those read pairs for which one member has mapped to a unique position in the reference and the other hasn't mapped at all, or maps with the inclusion of an indel.

An anchor point at the 3' end of the mapped read is defined, which restricts the search space for both deletions and insertions to a small area 3' of the read. To identify deletions, Pindel splits the un-mapped mate in two and uses a pattern growth algorithm to search for the maximum unique substring of that read (from the 3' end). The deletion is presumed to occur at the end of this substring and Pindel next tries to map the remainder of the read within the range of read length + Max_deletion_size (parameter given by the user). If the read can be completely reconstructed by the two substrings and more than one read supports the event, a deletion is called (Ye et al., 2009).

Deletions can thus in principle be quite long but the length of insertions that can be detected is capped by the length of the reads. The algorithm for insertions works much the same as the approach used for deletions except now the un-mapped read is split into

three with the middle part representing the insertion. Pindel searches the area <2*insert size (average distance between the mates) from the anchor point for the maximum substring matching the 3' end of the read, then does the same for the 5' end assuming that the middle section corresponds to the insertion (Ye et al., 2009).

cgpPindel is an adaptation of Pindel used at the Sanger institute that has been optimized for detection of somatic variants and that has a higher tolerance of indels in mapped reads (as reads have grown longer since the initial release of Pindel). It includes a filtering step where calls are compared against a normal panel to capture recurrent sequencing and mapping artifacts and implements a local realignment of reads overlapping indel calls to correct mapping errors that may be caused by the indel (Raine et al., 2015).

### 1.4.3 Mutational signature extraction

The mutational profile of a sample is the result of exposure to different mutagens, replication error and defective repair which each leave a characteristic fingerprint, a signature, on the genome (see section 1.2.1). Signatures can be defined as discrete probability distributions over a set of categorical mutation classes which are mixed in sample-specific ratios to create the observed profile. Formally, a mutation can be represented as a letter from a K-letter alphabet, $\Xi$ , and a signature as a K-element vector, $P_1 = [p_1^1, p_1^2, ..., p_1^k]$, where $p_1^k$ is the probability that the signature $P_1$ causes a mutation of type $\Xi[k]$ (Alexandrov et al., 2013b). For example, most studies define $\Xi$ as the trinucleotide context, yielding K=96 (section 1.2.1).

The original approach for signature extraction is to use non-negative matrix factorization (NMF) (Alexandrov et al., 2013b; Nik-Zainal et al., 2012). The method seeks to solve $M \approx R \times E$, where M is a $K \times N$ matrix containing non-negative mutation counts for K mutational classes and N samples, R is a $K \times P$ matrix, where K is the number of mutational classes and P is the number of signatures, and E is a $P \times N$ exposure matrix containing non-negative values between 0 and 1. A problem arises when we wish to allow for the possibility of signatures not present in the reference set of signatures to be extracted and so P is not known. Alexandrov et al apply Monte Carlo bootstrap resampling of the mutation matrix and calculate the NMF solutions at a range of values for P. They return the consensus solution with stability across bootstraps and minimal reconstruction error by the Frobenius norm (Alexandrov et al., 2013a,b). This is the method implemented in the commonly used SigProfiler software (Alexandrov et al., 2020).

A second approach for signature extraction that has become widely used is based on Hierarchical Dirichlet Processes (HDP) implemented in the R-package hdp (Roberts, 2018). This is a non-parametric Bayesian approach to carry out mutational signature extraction. Dirichlet processes are probability distributions used in Bayesian inference methods to determine how likely it is that a set of random variables follow a particular probability distribution (in this case, how likely it is that a set of somatic mutations is drawn from a particular set of mutational signatures). A Dirichlet process (DP) is defined by a base distribution, H, and a concentration parameter, $\alpha$. In a similar way a normal distribution draws numbers around its mean, a DP draws an infinite sample of distributions around its base distribution with increasingly small weights (Teh et al., 2006).

In hdp, the data is organized into a tree structure. The top node of the tree is the base distribution, which in this case is the uniform probability over the infinite set of all possible signatures. Its daughter node represents the distribution of signatures in the whole dataset and lower level nodes represent the signature distribution in some grouping of the samples, for example by disease and patient, if there are multiple samples from the same patient. A DP is associated with each node and the draw from the DP at a given node serves as a base measure for its children so at each level of the tree, a more discrete probability density is outputted (Roberts, 2018).

To estimate the signature identity and exposure at each level of the tree, hdp implements a Markov chain Monte Carlo (MCMC) algorithm for inferring the posterior distribution of signatures under the DP mixtures. All mutations are initially assigned to random clusters and then a Gibbs sampler cycles through each mutation in turn, probabilistically moving the mutation to a cluster with one of the following: 1) a cluster with a high proportion of that same mutation class across all samples. 2) a cluster with a high proportion of mutations in that sample or its parent node. 3) the mutation gets assigned to a new cluster all by itself. Posterior samples are taken at regular intervals in the MCMC chain that give snapshots of possible cluster allocations (Roberts, 2018; Teh et al., 2006).

HDP has the advantage over NMF based methods that it can simultaneously quantify exposure from known signatures and discover any new signatures that may be present. It can model relationships between samples and sample groups to identify shared signatures, while also providing a quantification of the differences between them (Roberts, 2018).

Regardless of the choice of method, mutational signatures are typically extracted for single base substitutions, double base substitutions and indels separately, even if the same mutational process may generate all three types in different mixtures (for example, a defective homologous recombination due to a *BRCA1/2* mutation will result in the accumulation of both indels and substitutions). There is in principle no reason why mutational signatures couldn't be defined using multiple mutation classes which would make them more reflective of the mutagenic process they represent. In practice however, there are usually so many more single base substitutions than there are mutations of the other classes that these come to dominate in the signature extraction by existing methods. Other challenges to signature extraction include resolving the difficulty posed by relatively flat or uniform signatures. These are often more difficult to accurately extract and distinguish from each other than are signatures defined by distinct mutational classes. Finally, each signature in COSMIC or PCAWG is represented as a single reference. However, the effects of different mutational processes likely vary slightly between tissues and individuals.

### 1.4.4 Phylogenetic tree building

Phylogenetic trees display evolutionary relationships between biological entities. Several methods exist for building phylogenetic trees, one of which is the method of maximum parsimony. This method assumes the most plausible tree is the one requiring the fewest mutation events to explain. A complication of maximum parsimony is that finding the most parsimonious tree in the space of all possible trees is an NP-hard problem (it can't be solved in polynomial time) so for a large number of taxa, computing the parsimony score for every possible tree becomes computationally prohibitive. Furthermore, the tree space may contain a number of equally most-parsimonious trees.

In the work presented in this thesis, I used the MPBoot software for phylogenetic inference (Hoang et al., 2018b). MPBoot uses a method called parsimony ratchet to carry out a heuristic search of the tree space and progressively approach the best tree (Nixon, 1999). While methods for maximum parsimony often employ bootstrap to assess the robustness of branches, MPBoot implements ultrafast bootstrap approximation, greatly reducing the computational time needed to derive a consensus tree (Hoang et al., 2018a).

It is worth concluding this section with a comment on the great robustness of phylogenetic trees constructed from somatic cells, which are free from some of the limitations that may affect, for example, studies of species comparison. Firstly, in contrast to species evolution, the ancestral state of somatic cells is known. This is the germline, and knowing it makes

rooting the tree trivial. Secondly, the mutation rate of somatic cells is typically so low that the chances of recurrent mutations at the same site or of a mutation reverting back to the ancestral state are miniscule.

### 1.4.5 Selection analyses

The selection analyses in this thesis were carried out using dNdScv, a software for quantifying selection in cancer and somatic evolution (Martincorena et al., 2017). dNdScv builds on methods with a long history of use in species evolution but introduces some modifications tailored to the study of somatic cells.

dN/dS is the ratio between the rate of non-synonymous substitutions per non-synonymous site and the rate of synonymous substitutions per synonymous site. dNdScv implements a group of maximum likelihood methods for quantifying this ratio after accounting for sequence composition, trinucleotide mutation rates and variable mutation rates across the genome. Synonymous mutations are modelled as a Poisson process, for example:

$$n_{C>T,s} \sim Poisson(\lambda = t \times r_{C>T} \times L_{C>T,s})$$

Represents the number of synonymous C>T mutations in the dataset, where t is the mutation rate per site, $r_{C>T}$ is the relative mutation rate of C>T mutations and $L_{C>T,s}$ accounts for the sequence composition, it is the number of C-sites where a C>T mutation would result in synonymous amino acid substitution of the protein. Non-synonymous mutations are modelled in the same way, but with an additional parameter, $\omega$, reflecting the effect of selection on the mutation count. For example, missense mutations at C>T sites are modelled as:

$$n_{C>T,m} \sim Poisson(\lambda = t \times r_{C>T} \times L_{C>T,m}, \omega_m)$$

Where $\omega_m$ represents the dN/dS ratio after correcting for mutation rate and sequence composition and a maximum likelihood estimate for it can be derived by Poisson regression.

The parameters above may be further refined. For example, there is not one site-wise mutation rate parameter, r, for each of the 6 mutation classes, but rather the model uses 192 rates. Genomes of somatic cells show a strong context dependence, particularly for the bases immediately 3' and 5' of the mutated base (see section 1.2.1 on mutations and section 1.4.3 on mutational signatures). The substitution rate model incorporates the 96 mutation

classes of a trinucleotide model and also accounting for transcriptional strand asymmetry, this number becomes 192.

The parameter t, the mutation rate per site, can also be further refined. Mutation rates are known to vary depending on the expression levels of genes, replication time and chromatin state of the region. dNdScv models t as following a Gamma distribution. The number of synonymous substitutions is modelled as a negative binomial distribution which allows the background mutation rate of each gene to be modelled combining local information like the size and sequence composition of the gene as well as more global information on mutation rates across genes. dNdScv uses as covariates the first 20 principal components of 169 chromatin marks from the ROADMAP project (Roadmap Epigenomics Consortium et al., 2015).

A similar strategy is used to incorporate indels in the selection model. Since there are no synonymous indels, the expected rate of indels per gene is modeled using a negative binomial regression model accounting for the length of the gene and the same epigenomic covariates as described above. As a default, a list of known cancer genes are excluded to avoid them inflating the background model. The P-value for the indel regression is then combined with that from the substitution model using Fisher's method.

# Chapter 2

# Estimating the crypt fission rate of the normal colon

In this chapter, I describe my work to use approximate Bayesian computation to estimate the crypt fission rate in the normal human colon and in the colons of patients with Familial adenomatous polyposis. The methods described herein and Figure 2.5 have been previously published as part of the manuscript "The landscape of somatic mutations in normal colorectal epithelial cells", Lee-Six et al. 2019. Nature. My contribution to that project was limited to estimating the crypt fission rate as described below.

## 2.1 Chapter Introduction

### 2.1.1 Colonic crypts

#### Intestinal stem cells

The epithelial sheet lining the human colon is made up of a single layer of columnar epithelial cells and organized into millions of colonic crypts, finger-like invaginations into the lamina propria below. A small number of stem cells (commonly known as crypt base columnar cells (CBCs)) reside at the bottom of the crypt. The CBCs are defined by expression of *LGR5* Barker et al. (2007) and an average of 5-10 LGR5+ stem cells are thought to exist at the bottom of every crypt (Nicholson et al., 2018; Stamp et al., 2018).

Stem cell identity (stemness, the ability to self renew while generating differentiated daughter cells) is not fixed but is a state that can be lost and gained by cell removal from the stem cell niche and re-entry by partially differentiated cells (Gehart and Clevers, 2019). The

CBCs are interspersed with deep crypt secretory (DCS) cells and stemness depends on direct contact between a CBC and a DCS, which provides the CBC with WNT ligands, epidermal growth factors and Notch stimuli required for its maintenance, as Paneth cells do in the small intestine (Gehart and Clevers, 2019; Sasaki et al., 2016). In addition to the CBCs, so-called +4 cells have also been reported to contribute to the stem cell dynamics of the crypt. The term +4 refers to the cell position just above the uppermost DCS cell, one cell up from the edge of the stem cell zone of the crypt, and four cells up from the crypt base. These partially differentiated cells may in some cases descend back into contact with DCS cells and re-gain stem-cell properties, although it remains debated whether this primarily happens following epithelial regeneration after injury or if the +4 cells also contribute to the stem cell pool under crypt homeostasis (Gehart and Clevers, 2019).

### Stem cell competition in the niche

Dividing stem cells at the base of crypts stochastically have one of three fates: They can produce two daughter stem cells, one stem cell and one differentiated cell or two differentiated cells. On average each division results in one stem cell and one differentiated cell. When a cell divides symmetrically to produce two differentiated cells, that clone becomes effectively extinct and is replaced by another cell symmetrically dividing to produce two stem cells. Over time, the crypt drifts towards clonality as the progeny of a single stem cell take over the crypt (Lopez-Garcia et al., 2010; Snippert et al., 2010), such that all somatic mutations found in the ancestor cell become fixed in the crypt. These neutral sweeps of the stem cell niche occur many times over a typical lifetime and have been estimated to occur at a rate of one sweep every 6.3 years, on average (Nicholson et al., 2018).

Although stem cell division is stochastic it may become biased if cells acquire driver mutations. Vermeulen et al quantified the selective advantage of cells carrying known key drivers in the stem cell population and estimated that in the mouse colon, $Kras^{G12D}$ mutants have an 80% chance of replacing a wild-type neighbour while $APC^{+/-}$ cells have a 62% chance (Vermeulen et al., 2013). This type of biasing of stem cell fate is an important mechanism through which drivers rise to high frequency in many tissues. The study also highlights that although cells carrying drivers have an advantage over wild type cells, they can nevertheless be lost by random drift. Vermeulen et al also described how the advantage of mutant cells may be altered in disease. They showed that the probability that a $TP53^{R172H}$ mutant cell takes over a crypt is increased in a mouse model of colitis (Vermeulen et al., 2013).

### Crypt fission and fusion

Stem cell clones may further expand beyond individual crypts, through a process known as crypt fission. Here, bifurcation of the crypt starts at the base and proceeds in a zipper-like manner towards the crypt opening at the intestinal lumen. Crypt fission occurs rapidly neonatally as the colon elongates, and continues to be observed at low rates during adulthood (Greaves et al., 2006; Nicholson et al., 2018). The opposite process, that of crypt fusion, has also been reported, first in mice (Bruens et al., 2017) and subsequently in humans (Baker et al., 2019).

Colorectal tumorigenesis occurs as a consequence of changes that disrupt normal crypt dynamics. The crypt fission rate of the normal mucosa (CFR) is an important parameter for our understanding of stem cell dynamics of the colon and, by extension, the origins of colorectal cancers. Most efforts to date to estimate the crypt fission rate have used histochemical staining for particular somatic mutations, often those causing loss of cytochrome c oxidase activity (Baker et al., 2014, 2019; Greaves et al., 2006). Estimates of the fission rate vary substantially, from one crypt fission every 13.5 years (Totafurno et al., 1987), to one every 36 years (Baker et al., 2014), to one every 91 years (Baker et al., 2019) and one every 139 years (Nicholson et al., 2018). My gut feeling is that this variation is mostly due to technical reasons. Early estimates of the crypt fission rate were based on simply counting the number of crypts within histological sections that were bifurcating and comparing that number with the total number of crypts. This depends on the bifurcation occurring "in plane" of the histological section and more importantly, does not take into account the more recently discovered phenomenon of crypt fusion (Baker et al., 2019). If all bifurcating crypts are assumed to be undergoing fission then the fission rate estimate would be inflated in such an analysis. Later studies used histological staining of tissue sections to track mutations in particular genes, for example the mitochondrial gene cytochrome c oxidase (Baker et al., 2019) or genes subject to X-inactivation (Nicholson et al., 2018). Unpublished work done in the Campbell lab in the Sanger Institute indicates that mitochondrial mutations are unreliable markers for lineage tracing. Somatic mutations in mitochondrial genomes are likely heteroplasmic and are not necessarily passed on during cell division, which would deflate crypt fission estimates based on this marker. The work described herein is to my knowledge the only estimate of the crypt fission rate based on whole genome sequencing data.

### 2.1.2 Familial adenomatous polyposis

Familial adenomatous polyposis (FAP) is an autosomal dominant syndrome caused by a germline loss-of-function variant in the *APC* gene. Afflicted individuals develop tens to thousands of colorectal adenomatous polyps, some of which inevitably develop into adenocarcinomas if left untreated. FAP adenomas grow by rapid crypt fission, most often driven by a somatic loss of the remaining wild-type *APC* allele (Li et al., 2020a). Overall, the crypt fission rate has been shown to be increased in the colons of FAP patients (Wasan et al., 1998), but as previous studies have not genotyped the crypts, it is not clear whether this increase is driven by crypts in the process of transformation or if *APC* heterozygous crypts also undergo fission at a different rate than wild-type.

### 2.1.3 Approximate Bayesian computation

Approximate Bayesian computation (ABC) is a statistical framework used to estimate the posterior distribution of model parameters when the likelihood function cannot be inferred analytically. Its use is well established in complex models in population genetics (Beaumont et al., 2002) and cancer biology. For example, it has recently been used for the estimation of the population size of the blood stem-cell pool (Lee-Six et al., 2018), to propose an introgression of archaic humans into Asia and Oceania (Mondal et al., 2019) and the inference of parameters of colorectal cancer evolution (Hu et al., 2019; Sottoriva et al., 2015).

ABC aims to approximate the posterior distributions of the parameters in question by simulating a large number of datasets with different parameter values. Informative summary statistics are computed for the simulated data and compared with the same summary statistics calculated for the observed data. Simulations yielding summary statistics close to those calculated for the observed data are retained and used to estimate the parameter values. A regression step is sometimes performed to give additional weight to values that minimize the distance between simulated and observed data (Bertorelle et al., 2010).

Formally, let M be a model used to create data, D, which is determined by a vector of parameters, $\theta$. Denote the prior density by $p(\theta)$. We wish to estimate the posterior distribution of the parameters, calculated by Bayes rule as:

$$\pi(\theta|D) = c \times f_m(D|\theta) \times \pi(\theta)$$

where $f_m(D|\theta)$ is the likelihood of the data and c is a normalizing constant. The problem facing us is that the likelihood function cannot be calculated analytically and so we seek to

empirically reconstruct the posterior distribution through simulation.

Let s be a vector of summary statistics and let $\varepsilon$ be a distance cut-off. We carry out a large number of simulations and each time $dist(s, s_{obs}) < \varepsilon$, we write the parameters of that simulation to a list $P = \{\theta_1, ..., \theta_N\}$.The distance function dist() can for example be the Euclidean distance between the two vectors. In the limit $\varepsilon -> 0$, then $\pi(\theta|dist(s, s_{obs})) = \pi(\theta|D)$. Note however, that the smaller $\varepsilon$ is set, the larger number of simulations will be required since the condition $dist(s, s_{obs}) < \varepsilon$ is rarely satisfied, especially when the number of summary statistics is large (Bertorelle et al., 2010).

One way to speed up the calculations and reduce the number of simulations needed is to incorporate a regression step. $\varepsilon$ is relaxed so a larger fraction of simulations is retained in the first step. The most intuitive method is then to perform a local linear weighted regression between the vector of summary statistics and the retained parameters, assigning to each point a weight inversely proportional to the distance from the observed statistics. The intercept of the line is the best estimate of the parameter. Alternative regression approaches are the general linear model proposed by Leuenberger and Wegmann (Leuenberger and Wegmann, 2010) and non-linear machine learning approaches as suggested by Blum and Francois (Blum and François, 2010).

## 2.2 Chapter aims

In this chapter I describe my efforts to use approximate Bayesian computation to estimate the crypt fission rate in normal colon. I then use the same framework to estimate the crypt fission rate in an independent dataset consisting of *APC* heterozygous crypts dissected from FAP patients and compare the crypt fission rate between the two cohorts.

## 2.3 Methods

### 2.3.1 Input data

As a part of a study of the mutation landscape of the normal colon (Lee-Six et al., 2019), I used ABC to estimate the crypt fission rate of the normal colon. In this study, 571 microdissected crypts from 42 individuals were whole genome sequenced, of which 449 had >10X coverage and were considered in this analysis. The median sequencing depth of the crypts was 16.3X. The lead author of the study, Henry Lee-Six, called and filtered the

mutation calls, constructed phylogenetic trees for each individual and extracted mutational signatures for each branch of each tree with contributions from other authors as described in the 'Author contributions' statement of the original publication. Henry Lee-Six also reviewed microscopic images of the tissue sections to establish spatial relationship matrices for each biopsy. The physical distance between any pair of crypts was established in the unit of crypts separating the pair. Due to varying quality of the tissue sections and the microscopic images, not all sequenced crypts could be included and some biopsies contained cliques of crypts where the distances between crypts in a clique could be established but not the distance between the different cliques. In such instances, the cliques of crypts were treated as independent biopsies and any potential coalescent events linking crypts from different cliques were ignored (this is unlikely to bias the analysis, since crypts from different cliques are usually quite distant in the tissue and so are unlikely to be linked by a recent coalescent event). The final input data for the normal colon cohort included pairwise distances from 324 crypts dissected from 102 independent "biopsies" from 36 donors (Figure 2.1).

To estimate the crypt fission rate in patients with FAP, I used a dataset generated by Dr. Philip Robinson, a clinical PhD student at the Sanger institute. This dataset consisted of 87 crypts dissected from 28 biopsies from 14 donors. Of these 12 biopsies were collected from a single donor (Figure 2.2). Philip Robinson called and filtered the mutation calls, constructed phylogenetic trees and extracted mutational signatures for this data. He also provided a spatial relationship matrix for each biopsy. At the time of writing, Philip Robinson is working on a study of the somatic evolution landscape in FAP that includes samples from normal segments of colon, polyps and carcinomas. I used only the crypts dissected from normal colon and all crypts were confirmed to be *APC* heterozygous (that is, they had one defective copy due to a germline variant and had not acquired a mutation of the second allele).

### 2.3.2 Simulating the colon

To estimate the crypt fission rate, I simulated clonal spread in each biopsy assuming different values of the crypt fission parameter. The epithelial sheet is essentially a two-dimensional structure, bent in space to form the crypts, and I simulated each biopsy independently as a two dimensional $n \times n$ grid of cells, with each cell of the grid representing a crypt (Figure 2.3). Clones could not expand beyond the edges of the grid, and to allow clones sufficient space to spread in every direction while also keeping the simulation from getting prohibitively large, n was set to equal three times the largest distance seen between any two crypts from that biopsy (Figure 2.1D).

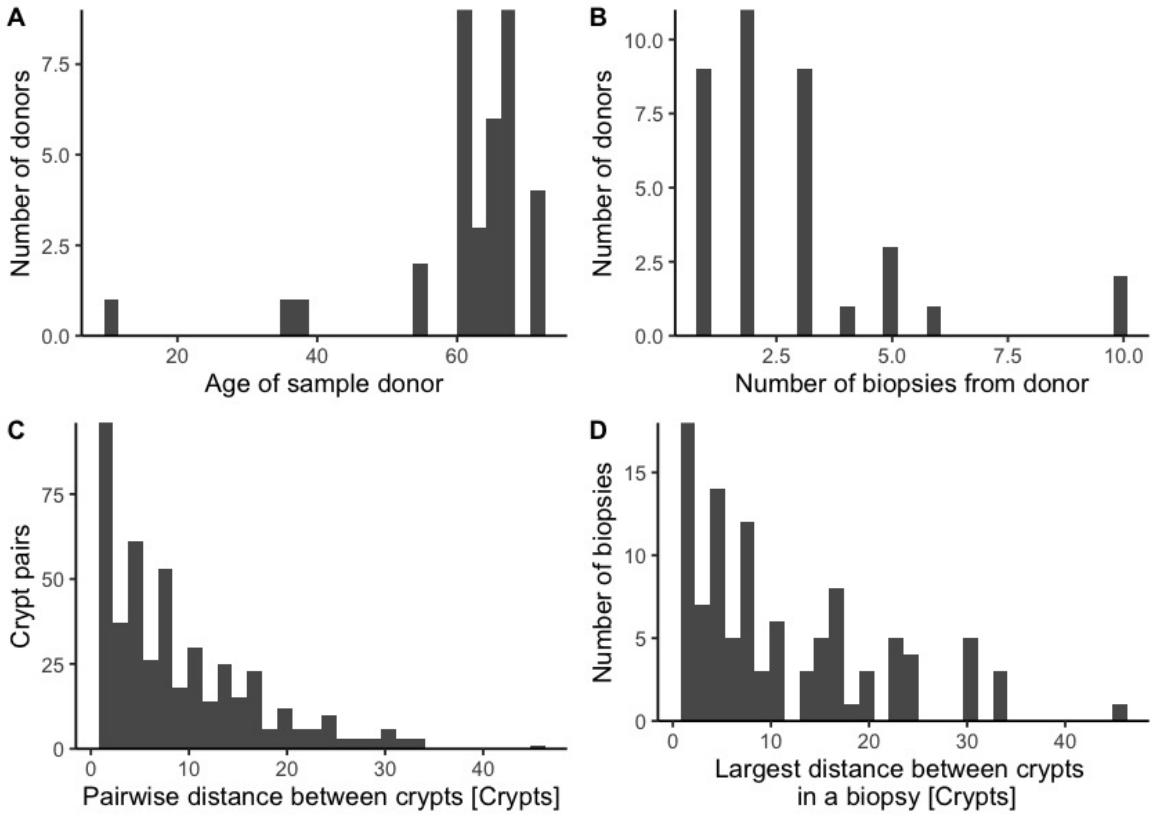



Fig. 2.1 **An overview of the input data used in the simulations of the normal colon.** A) The age distribution of the sample donors. B) The number of biopsies (or independent cliques of crypts, see the main text) from each sample donor. C) The distribution of pairwise distances between crypts within the same biopsy. Most coalescent events are observed between crypts close together in space. D) The largest distance between crypts in a biopsy.

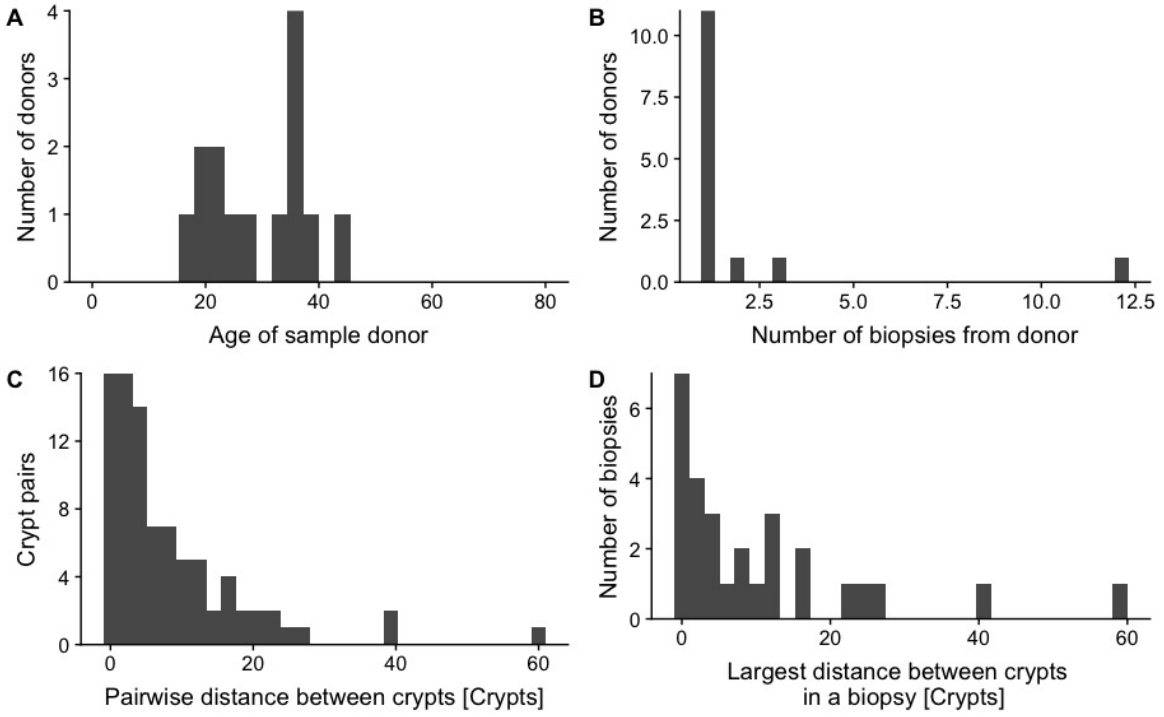



Fig. 2.2 **An overview of the input data used in the simulations of the FAP cohort.** A) The age distribution of the sample donors. B) The number of biopsies from each sample donor. C) The distribution of pairwise distances between crypts within the same biopsy. D) The largest distance between crypts in a biopsy.

I simulated each biopsy thousands of times assuming different values of the crypt fission rate parameter. I drew a crypt fission rate for each simulation from a uniform prior between zero and 0.25 fissions per crypt per year (or 1 fission per crypt every 4 years, this being considerably higher than any previous estimate of the crypt fission rate). The term 'simulation' here refers to a simulation of all the biopsies in the observed dataset using the same value for the crypt fission rate parameter. Starting from time zero, I drew the time until the next crypt fission event from an exponential distribution with the rate parameter determined by the fission rate for that simulation. At each event, a cell in the grid (but not at the edge) was randomly chosen to die and be replaced by one of its eight neighbours, chosen at random (Figure 2.3). The latter cell is considered having undergone fission. This process was repeated until the total time passed exceeded the age of the patient from which the biopsy was drawn, allowing clones to spread in the grid (Figure 2.4A). I next sampled the grid in a way that preserved the spatial relationships between the crypts in the observed data and identified the timing of the coalescent events linking the sampled cells (Figure 2.4B and C).

As described above, ABC uses vectors of summary statistics to compare simulations with observed data. In this case, the summary statistics were the numbers of coalescent events linking the sampled crypts in the simulated data vs the observed data. The coalescent events were grouped into 10-year intervals from birth to 80 years of age (Figure 2.4D) and counted for each biopsy and summed across all biopsies to derive a summary statistics vector for that simulation. To time the coalescent events in the observed data, I used the number of mutations on each edge of the phylogenetic trees that were assigned to substitution signature 1. As stated in Chapter 1, signature 1 represents a clock-like mutational process (Alexandrov et al., 2015) and the timing of the coalescent events could be estimated given the location of the biopsy and the mutation rate of signature 1 in different sectors of the colon (16.8, 16.1, 12.8 and 12.7 mutations per year in the right, transverse and left side of the colon and the ileum, respectively, as estimated in Lee-Six et al).

One cannot simply count the coalescent events of each tree, since a full binary tree with n leafs always has 2n-1 nodes. Instead, I ignored events occurring earlier than 4 years in molecular time in both the observed data and the simulated data to avoid counting events occurring as part of embryogenesis or the neonatal expansion of the colon. The choice of a 4 year cutoff was empirically chosen. The goal was to remove the many coalescent events clustered right at the base of the trees in the observed data and setting the cutoff at 4 years (or 51-67 SBS1 mutations in molecular time, depending on the location of the biopsy in the colon) seemed to accomplish this.

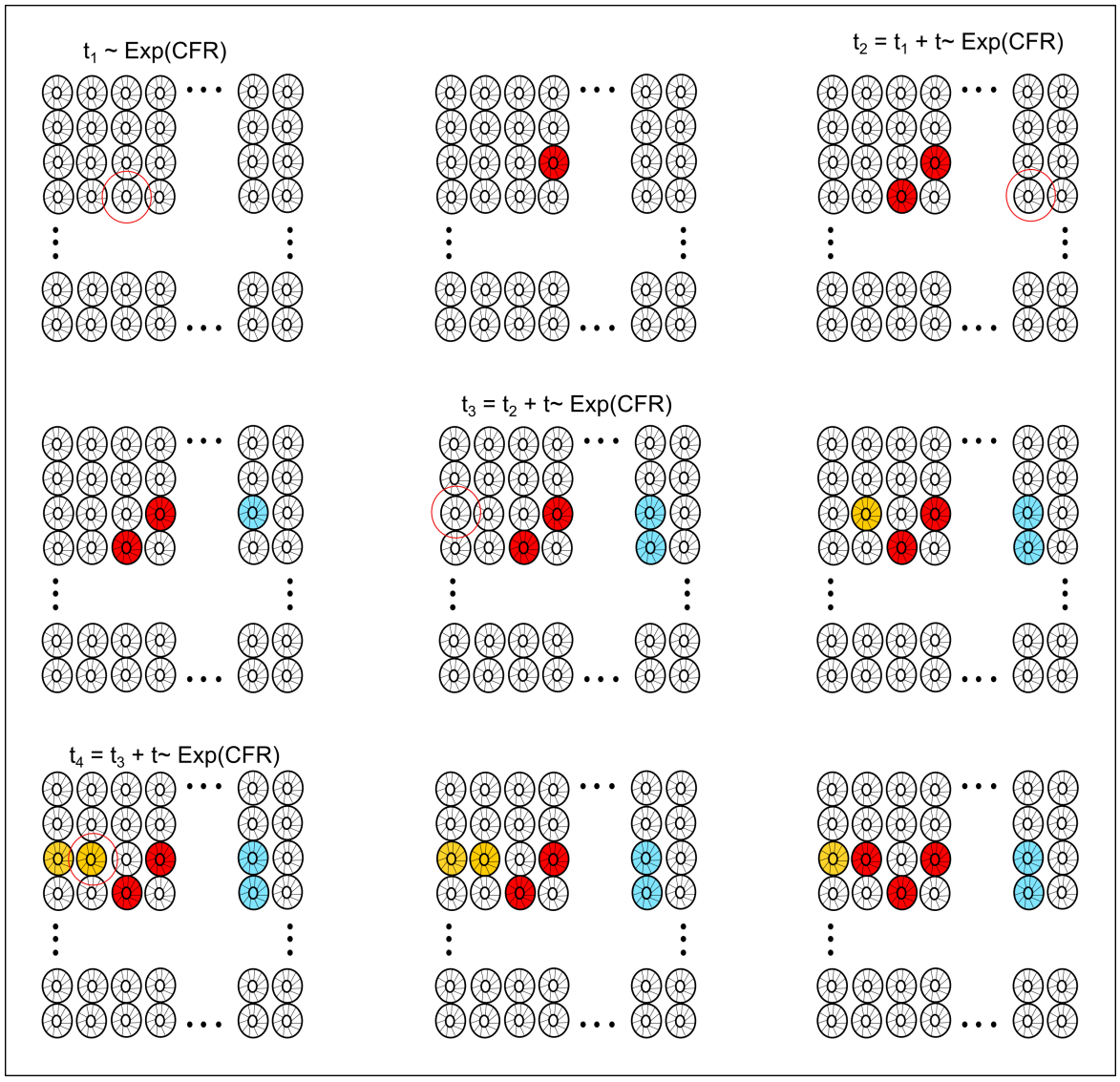



Fig. 2.3 **Simulation of clonal dynamics of the colon.** The time until the next crypt fission is drawn from an exponential distribution with the rate parameter determined by the crypt fission rate (CFR) for that simulation. A random crypt (red circle) is chosen to die and be replaced by one of its neighbours (red crypt). This process is repeated and clones, represented in red, blue and yellow, emerge in the grid. In one instance, a crypt of the yellow clone is replaced by its neighbour.

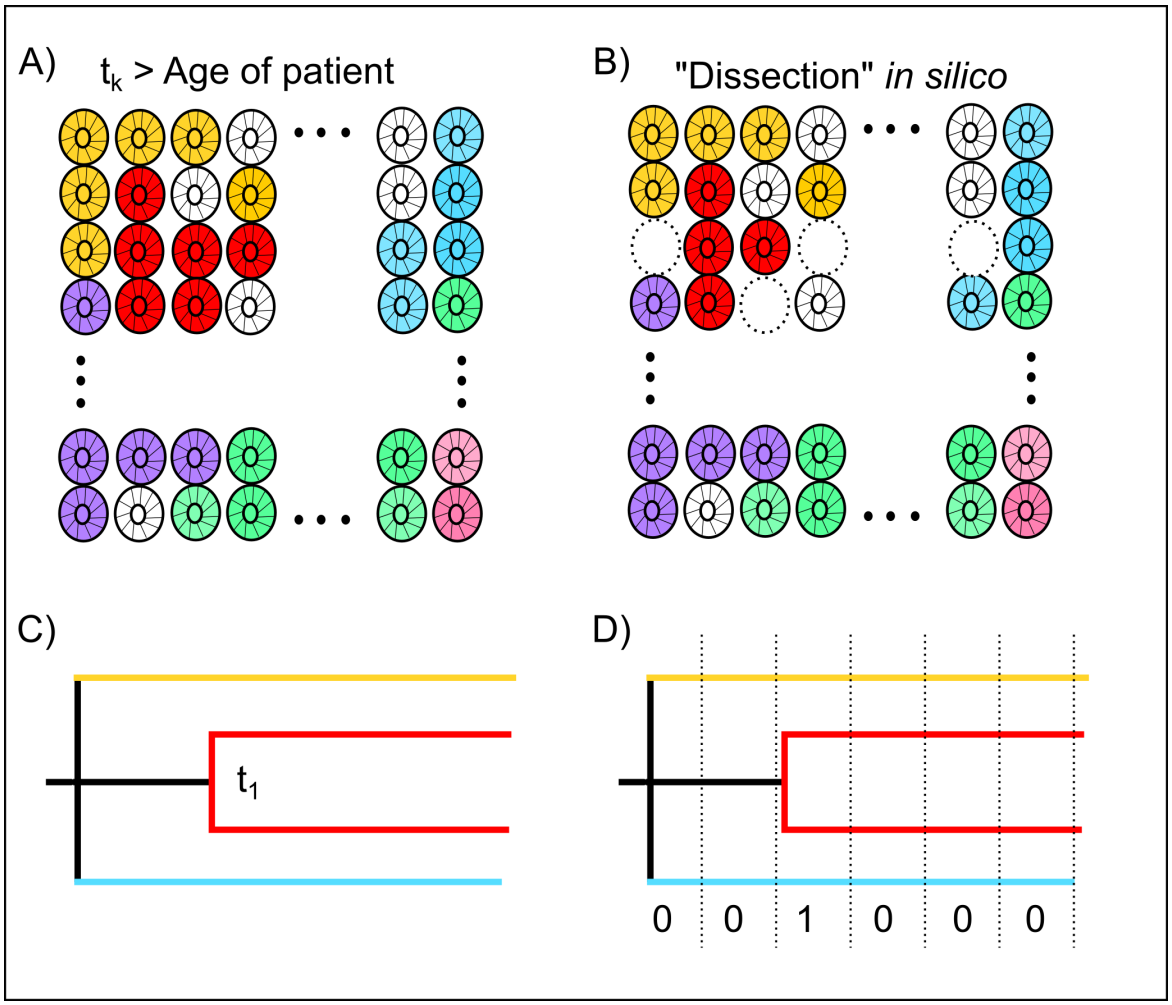



Fig. 2.4 **Sampling the grid at the end of an ABC simulation.** A) When the sampled time exceeds the age of the patient, the simulation is halted. B) The grid is sampled in a way that preserves the distances between the crypts in the original biopsy (in silico dissection). The dashed circles in the figure denote dissected crypts. C) A phylogenetic tree for the "dissected" crypts is constructed and the time of coalescent events is noted. D) A vector containing counts of coalescent events in 10 year intervals is constructed by aggregating counts across all 102 biopsies.

### 2.3.3 Estimating the posterior distribution

To estimate the posterior distribution of the crypt fission rate, I calculated the Euclidean distance between the summary statistics vectors generated for the observed data on one hand and simulations on the other. Rather than using a simple rejection algorithm, I employed the regression-based correction method proposed by Blum and François (Blum and François, 2010) and implemented in the 'abc' package in R (Csilléry et al., 2012). This is a neural network approach designed to reduce the dimensionality of large summary statistic vectors. It fits a nonlinear conditional heteroscedastic regression of the crypt fission rate on the summary statistics and then uses importance sampling to improve the estimation further (Blum and François, 2010). I chose the neural network regression over, say, a general linear model (Leuenberger and Wegmann, 2010), to account for potential non-linearity and uneven variance across the summary statistics dimensions. The summary statistic vectors count coalescent events occurring in 10-year intervals from birth to 80 years of age. As many patients were younger than that at the time of sampling, they don't contribute any counts to the late time intervals, resulting in unequal variance across summary statistic dimensions.

I first set a rejection threshold such that the 5% of simulations with Euclidean distances closest to the observed data were retained (Figure 2.5B). The crypt fission rates of the accepted simulations give an estimate of the posterior distribution of the crypt fission rate parameter, but the estimate can be improved by applying a neural net regression which uses the following equation in the vicinity of the observed data $(S(y_0))$:

$$\theta_i = m(S(y_i)) + \varepsilon_i$$

where $\theta_i$ is the crypt fission rate of simulation i, $S(y_i)$ is the summary statistics for simulation, m() is the regression function and $\varepsilon_i$ represents centred random variables with equal variance. The adjusted crypt fission rate, $\theta_i'$, is obtained as follows:

$$\theta_i' = m'(S(y_0)) + \frac{\sigma(S(y_0))}{\sigma(S(y_i))} \times \varepsilon_i'$$

where m'() is the estimated conditional mean and $\varepsilon'$ is the residual of the regression (which is adjusted for heteroscedasticity in the equation).

## 2.4 Results

Under the model described above, I estimated the crypt fission rate in the normal human colon to be 0.037 (0.021 – 0.063, 95% credibility interval) crypt fissions per crypt per year. This corresponds to one fission per crypt every 27 years on average (Figure 2.5C).

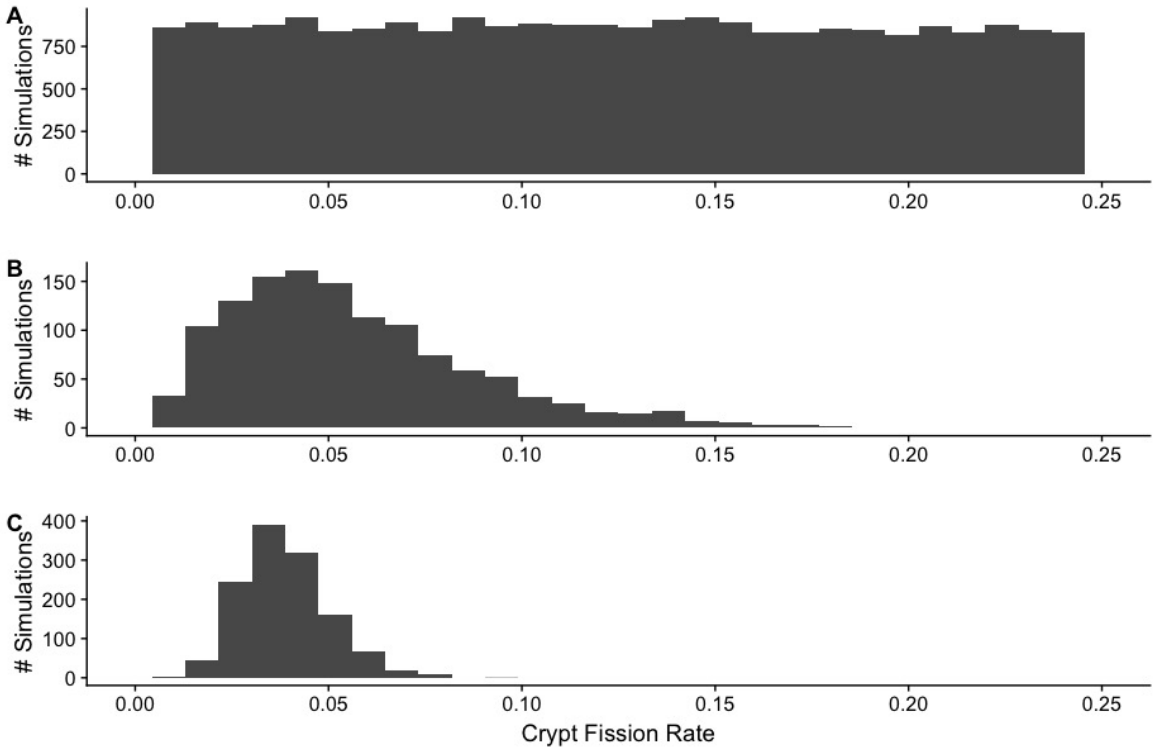



Fig. 2.5 **Approximate Bayesian computation of the crypt fission rate in the human colon.** A) The prior distribution of the crypt fission rate used to simulate many biopsies of the colon. The unit for the crypt fission rate is fissions per crypt per year. B) The crypt fission rates of the 5% of simulations that produced summary statistics most similar to those calculated for the observed data. C) The posterior distribution of the crypt fission rate parameter estimated by neural network regression on the simulations in B. The 95% credibility interval is 0.021-0.063 fissions per crypt per year.

In contrast, I estimated the crypt fission rate in the FAP sample to be 0.076 (0.034-0.146) fissions per crypt per year, corresponding to one fission every 13.2 years (Figure 2.6C). This estimate falls outside the 95% credibility interval for the normal colon (that is to say, 97.5% of the posterior distribution for the normal colon is smaller than the estimate for FAP), indicating that the fission rate is likely higher in FAP than in the control cohort. I compare the two posterior distributions in Figure 2.7. The larger variance of the posterior distribution

for the FAP cohort compared with the normal cohort is a consequence of the smaller sample size of the input data.

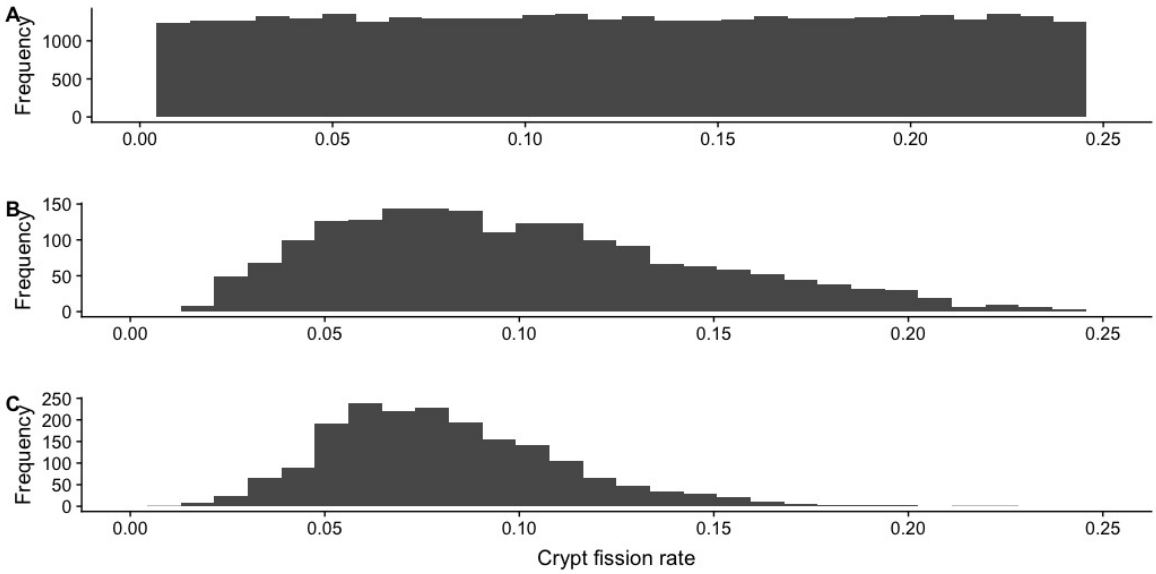



Fig. 2.6 **Approximate Bayesian computation of the crypt fission rate in the FAP cohort.** A) The prior distribution of the crypt fission rate used to simulate 87 colonic biopsies. The unit for the crypt fission rate is fissions per crypt per year. B) The crypt fission rates of the 5% of simulations that produced summary statistics most similar to those calculated for the observed data. C) The posterior distribution of the crypt fission rate parameter estimated by neural network regression on the simulations in B. The 95% credibility interval is 0.034-0.146 fissions per crypt per year.

## 2.5 Discussion

In this chapter, I have described my work to use approximate Bayesian computation to estimate the crypt fission rate in the normal human colon and in colons of patients with FAP. Knowing the fission rate in the normal colon establishes a baseline to which other conditions may be compared. For example, I have presented evidence that the crypt fission rate is accelerated in FAP, even in the absence of a second *APC* hit.

The use of whole genome sequencing frees this analysis from some of the limitations that may have affected previous estimates (as described above) but there are still some limitations and assumptions this analysis makes that need to be considered. One of the main difficulties is the low number of coalescent events in both of the observed data sets, which reduces the

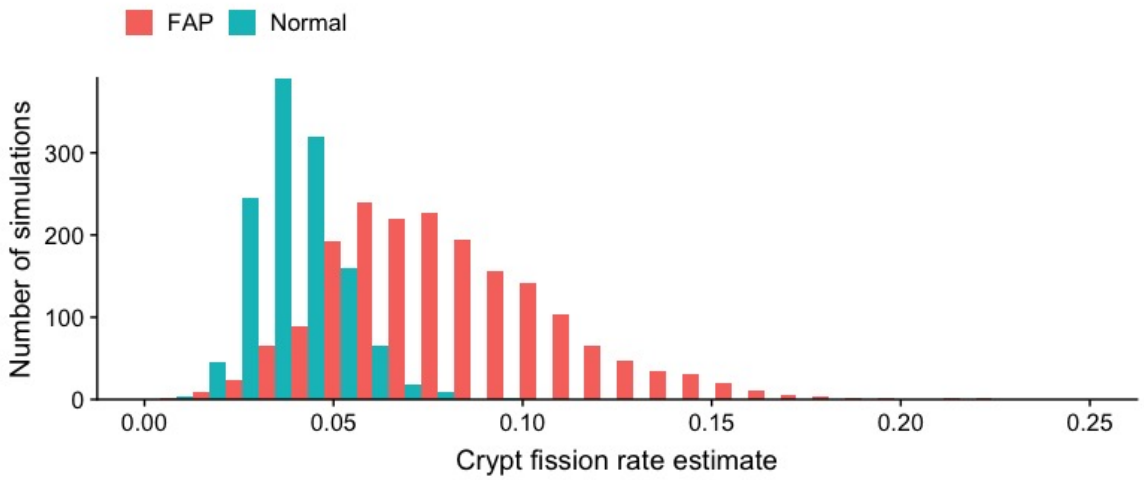



Fig. 2.7 **A comparison of the ABC posterior distributions for the normal and the FAP cohort.**

power to accurately estimate the crypt fission rates.

The greatest limitation of the work presented in this chapter is likely that I did not carry out any formal comparison of my model with alternative models, for example one where crypts have six neighbours rather than eight or one where a crypt death causes a distant crypt to fission, rather than a neighbour of the dead crypt, or where the probability of a fission is affected by the crypt having recently undergone a previous fission. This was due to the massive computational resources needed to carry out the simulations, which prohibited multiple executions. Although I did not model crypt fusion specially, the simulation does allow for fusion.

The model used makes the following four assumptions:

1. The crypt fission rate is constant after four years of life and is the same in all sectors of the colon and the ileum and does not vary between individuals.
2. The number of crypts is constant after four years of life.
3. The effect of selection is negligible in the colon.
4. The mutation burden of substitution signature 1 in each sector of the colon is fixed through life.

Some elongation of the colon may occur between the age of four and adulthood, causing violations of assumptions 1 and 2. If assumption 1 is violated, the crypt fission rate estimate

should be interpreted as the average rate over the period, colon locations and individuals. The effect of violating assumption 2 will be most pronounced over large distances, as extra crypts are spread along the length of the colon. Since the biopsies used in this study are only a few millimeters across, the effect of violating assumption 2 should be modest.

The model does allow for crypt fusion. Following fusion of crypts A and B, the progeny of one of the stem cells from A and B takes over the new crypt. If that cell originates from A, the fusion event corresponds to the death of crypt B in the model.

The number of crypts is not constant in the colons of FAP patients but grows with time and this manifests as a large number of polyps. It is possible that prior to polyp formation, the density of crypts is locally increased. However, a second driver mutation, especially a mutation of the remaining wild-type *APC* allele appears to be the driving force for polyp formation. By excluding all such crypts from this analysis, I hope the effects of violating assumption 2 is minimal.

In Lee-Six et al, the paper in which the crypt fission rate estimate for the normal colon was reported, we also report that driver mutations are very rare in the colon compared with other normal tissues (Lee-Six et al., 2019; Martincorena et al., 2018; Moore et al., 2020). Fewer than 5% of crypts carry a putative driver mutation and even then, the effect of a lone driver on the crypt fission rate is uncertain. I therefore think assumption 3 is reasonable in this setting. In the FAP dataset, all crypts carry a germline *APC* variant. This does not violate assumption 4 as the fitness of all crypts should be equal. Crypts carrying other drivers, and a second *APC* mutation in particular, have an obviously accelerated crypt fission rate which results in the formation of polyps but these were excluded as described previously.

Assumption number 4 should also hold true. Signature 1 has been shown to have clock-like properties across a range of cancers and normal tissues (Alexandrov et al., 2015), including colon (Blokzijl et al., 2016; Lee-Six et al., 2019). Signature 1 mutation rate is used to place the coalescent events in 10-year bins. Even if assumption 4 were violated under some conditions, many coalescent events in the observed data would likely still be placed in the right 10-year bin, unless the violation was very severe.

Since the work on the crypt fission rate in normal colon was published, Kakiuchi et al have published their own estimate of the crypt fission rate from whole-exome sequencing groups of crypts isolated from three individuals (Kakiuchi et al., 2020) (This study is further

discussed in the discussion of Chapter 3). They suggest one fission occurs every 7.4 years up until the age of 20 and one fission occurs every 242 years thereafter. This estimate of crypt fission rate in adulthood is much lower than any previously published estimate. It is derived from the simple equation:

$$Fissionrate = (N_{branch_{p}oint}/N_{crypt})/Y_{after20}$$

Where $N_{branch_{p}oint}$ is the number of coalescence events observed after 20 years of age (estimated from the mutation rate), $N_{crypt}$ is the total number of crypts in the tree and $Y_{after20}$ is the age of the patient minus twenty years.

This formula for the crypt fission rate is in my opinion flawed because it makes the implicit assumption that all pairwise comparisons of crypts sampled from a biopsy are equally likely to uncover a coalescent event linking the two crypts. This assumption biases the estimate of the crypt fission rate downwards. In reality, late-occurring (after 4 years or 20, it makes no matter) coalescent events are most likely to link two adjacent crypts and the probability of observing an event decreases with distance between two crypts of a pair at a rate which is proportional to the crypt fission rate itself. My approach of using ABC controls for this without making any assumptions about the rate at which the probability of observing a crypt fission event drops with distance. The simulations should mirror the observed data in this regard.

The above assumptions may be more severely violated in diseases such as IBD. The cycles of crypt death and healing that characterize the disease may drastically affect the microenvironment of crypts in affected regions. The decrease in crypt density associated with a disease flare is followed by rapid fission as the damage is repaired. These cycles may also be associated with a distinct selection landscape and may for instance allow crypts carrying driver mutations to rapidly expand in the colon. Finally, increased cell proliferation may be associated with increased burden of mutational signature 1 after disease onset. While I describe in the next chapter the somatic evolution landscape of IBD-affected colon, I did not formally estimate the crypt fission rate.

# Chapter 3

# Somatic evolution in the non-neoplastic IBD affected colon

The work presented in this chapter was published in the manuscript "Somatic evolution in non-neoplastic IBD-affected colon", Olafsson et al. 2020. Cell. I contributed to the design of the project, carried out all histological processing of the samples (apart from sectioning of the tissue, which was done by Yvette Hooks) and carried out all the laser capture microdissectioning. I further called the mutations and carried out all bioinformatic analyses except for those where I have explicitly highlighted the contributions of my colleagues in the main text. I interpreted the results together with my supervisors, Drs Carl Anderson and Peter Campbell, and wrote the paper. While all authors contributed to the final text of the manuscript on which this chapter is based, the text of the chapter is mine.

## 3.1 Chapter introduction

Many human diseases are associated with increased risk of cancer. Chronic diseases often have profound consequences on the cellular constitution of affected tissues and this can affect the evolution of cells in many different ways. For example, mutagen exposure and mutation rate may be altered as a result of inflammation, different cell turnover and/or medication and lifestyle factors associated with the disease. Genetic drift may be accelerated by faster cell division and the selection forces operating within a tissue may be changed, making mutations that were neutral under normal conditions advantageous in disease conditions.

Methodological difficulties have until recently limited the study of the changes to the somatic evolutionary landscape that accompany a disease to studies of cancers or premalig-

nant structures like polyps or precancerous fields, from which a comparatively clonal sample may be obtained. In most cases, the early changes to the somatic evolutionary landscape that accompany the disease and pre-date these structures remain poorly understood. Furthermore, to detect changes to the evolutionary landscape that don't increase cancer risk, or that may even be protective against cancer, we need to study non-neoplastic tissues.

The study of somatic evolution in complex diseases is not only motivated by a need to understand cancer risk. Somatic mutations may contribute to complex disease pathogenesis, affect the disease progression and/or drug response. In this chapter, I will describe somatic evolution in inflammatory bowel disease (IBD) affected colonic mucosal tissue and compare the IBD affected colon to normal colon. While much of the focus of the project was to understand the differences in somatic evolution that lead to increased cancer risk among IBD patients, I also found some exciting evidence that somatic mutations may directly contribute to disease pathogenesis.

### 3.1.1 Inflammatory bowel disease

Inflammatory bowel disease is a chronic inflammatory disease of the gastrointestinal tract that has two main subtypes, Crohn's disease (CD) and Ulcerative colitis (UC). The disease is thought to arise as a result of an inappropriate immune response against the resident microbiota and other incompletely understood environmental triggers in genetically predisposed individuals.

Together, CD and UC affect over 2.5 million people of European ancestry and their incidences in developing nations seem to be on the rise (Molodecky et al., 2012; Ng et al., 2017). Both UC and CD are characterized by abdominal cramps, diarrhea and rectal bleeding. Both occur in flares and both are most often diagnosed in early adolescence. They are distinguished by disease location and continuity. UC affects only the large intestine, usually spreading continuously from the left side of the colon, with no healthy regions separating inflamed areas. In contrast, CD may affect any part of the gastrointestinal tract, from mouth to anus, and causes patches of inflamed regions with healthy, un-inflamed regions in between. Importantly, flares tend to re-occur in the same region of the colon both in CD and UC, suggesting that some permanent alterations of the gut biology may play a causal role.

The causes of IBD have not been fully deciphered. The rising incidence in developing nations suggests that various components of modern 'Western' lifestyle contribute to disease pathogenesis. There is evidence linking early life exposure to antibiotics, diet, smoking

and availability of vitamin D with IBD onset (reviewed in (Ananthakrishnan et al., 2017)). The microbiome is altered in IBD, especially in CD patients who have increased abundance of Bacteroidetes and Proteobacteria and decreased abundance of Firmicutes, but a causal relationship has not yet been established.

Germline genetic factors also play an important role in IBD pathogenesis. Genome wide association studies (GWAS) have identified over 240 statistically independent associations between single nucleotide polymorphisms (SNPs) and disease risk (de Lange et al., 2017; Goyette et al., 2015). GWAS has implicated multiple biological pathways in the pathogenesis of IBD. These include innate and adaptive immune regulation, microbial defense and autophagy, but also intestinal permeability. Sequence variants in or near *GNA12*, *CDH1*, *MUC19* and *PTPN2*, for example, are hypothesized to contribute to variation in epithelial barrier integrity and intestinal permeability.

Although generally considered a complex disease with a complex genetic architecture, IBD-like phenotypes can also develop as a result of a single or a few very rare, highly penetrant mutations (Uhlig, 2013; Uhlig et al., 2014). Loss of function variants in 67 genes are thought to cause monogenic disease, with varying levels of evidence.

### 3.1.2 Colitis-associated colorectal cancers

IBD patients are at increased risk of developing colorectal cancers (CRCs). Cancer risk is associated with the duration, extent and severity of disease, but on average CRC risk of IBD patients has been estimated to be 1.7 fold that of the general population (Adami et al., 2016; Beaugerie and Itzkowitz, 2015; Lutgens et al., 2013). The overall cancer risk is higher in UC than in CD. However, this may be due to CD commonly affecting the small bowel, where the cancer rate is much lower than in the large bowel in the general population. The CRC risk is likely similar after correction for the extent of colonic involvement (Gillen et al., 1994). In the work presented in this chapter I tested for differences between CD and UC at every level of analysis but found no significant differences in somatic evolution between the two types of IBD. Most of the analysis is therefore presented as comparison between the combined cohort of IBD patients and controls.

As a result of the increased cancer risk, IBD patients require regular endoscopic screening and may undergo prophylactic colectomy to mitigate this risk (Adami et al., 2016; Beaugerie and Itzkowitz, 2015). The clinical presentation of colitis-associated cancers differs from that of sporadic cancers. Patients with colitis develop cancers at a younger age, the lesions are

more likely to be synchronous and to have mucinous or signet ring cell histology (Choi et al., 2017). Colitis-associated CRCs frequently grow from a precancerous field, where a mutant clone (often carrying a *TP53* mutation) has taken over a large section of the colon (Galandiuk et al., 2012; Leedham et al., 2009) (reviewed in (Choi et al. 2017)).

Colitis associated cancers also differ from sporadic cancers on the molecular level. Whole-exome sequencing studies have suggested that there may be differences in the frequencies with which key genes are mutated. In particular *TP53* may be more often mutated and *KRAS* and *APC* more seldom mutated (Baker et al., 2018; Din et al., 2018; Robles et al., 2016; Yaeger et al., 2016).

### 3.1.3 Somatic evolution in the normal colon

Before describing somatic evolution in the IBD-affected colon, I will in this section give a brief overview of the somatic evolution landscape of the normal colon. This is mostly based on 'The landscape of somatic mutation in normal colorectal epithelial cells' by Lee-six et al mentioned in Chapter 2. I used part of this data as control cohort in my work on IBD, as described in the methods section.

#### Mutation burden and mutational signatures

Lee-Six et al estimated the mutation burden of the colon to be 43.6 mutations per crypt per year of life (Lee-Six et al., 2019). This is comparable with the earlier work of Blokzijl et al, who estimated a mutation rate of 40 mutations per crypt per year (Blokzijl et al., 2016).

Three single base substitution signatures, SBS1, SBS5 and SBS18 were found in over 85% of crypts, as were the indel signatures ID1, ID2 and ID5. The mutation burden of all of these signatures showed a linear relationship with age and all showed the same pattern of the highest mutation burden being found in the right side of the colon, then the transverse and lowest in the left side of the colon. Rarer mutational signatures were found more sporadically and included the previously described SBS2 and SBS13, which are attributed to active APOBEC (as described in section 1.2.1) and were found in only two crypts from different individuals. Additionally, Lee-Six et al described four substitution base signatures and two indel signatures that had not been previously identified in studies of cancer genomes. These are termed SBSA, SBSB, SBSC, SBSD and IDA and IDB. Since being reported by Lee-Six et al, SBSA and IDA (which are highly correlated) have been shown to be caused by Escherichia coli bacteria carrying the pathogenicity island pks (Pleguezuelos-Manzano et al.,

2020). This island encodes enzymes that synthesize the genotoxic compound colibactin. SBSA is now referred to as COSMIC signature 88.

SBSB and IDB are also highly correlated and likely result from the same underlying mutational process. The etiologies of SBSB, SBSC and IDB are unknown but SBSD was only found in one patient with a history of treatment with multiple chemotherapeutic agents for the treatment of lymphoma and likely represents the effect of the treatment.

#### Clonal structure and driver mutation landscape

The analysis of Lee-Six et al showed that crypt fissions occur rarely in the normal colon (see Chapter 2). The most recent common ancestor of most crypts dissected from an individual therefore exists very early in molecular time, and has often existed during the neonatal expansion of the colon.

When compared with other epithelial tissues like the skin, endometrium or oesophagus, the colon is comparatively devoid of drivers. Using an admittedly conservative driver definition, Lee-Six et al estimated that only about 1% of crypts carry driver mutations. Two recessive tumour suppressors, *AXIN2* and *STAG2*, were found to be recurrently affected by truncating mutations and nine further mutations occurring in canonical driver hotspots of other known cancer genes were identified.

## 3.2 Chapter aims

In this chapter, I explore the changes to the somatic evolution landscape of the colon that are associated with IBD. I describe the microdissection and sequencing of colonic crypts from IBD patients and compare those with crypts sequenced as part of the project on the normal colon described above (Lee-Six et al., 2019). I compare IBD affected mucosa to normal colon in terms of the mutation burden, mutational signature exposure, clonal structure of the tissue and driver mutation landscape.

## 3.3 Methods

### 3.3.1 Human tissue attainment and processing

Colonic pinch-biopsies were donated by 49 IBD patients undergoing regular surveillance of their disease at Addenbrooke's hospital, Cambridge (Figure 3.2 and Table 3.1). All samples

were obtained with informed consent of the donor and the study was approved by the National Health Service (NHS) Research Ethics Committee (Cambridge South, REC ID 17/EE/0338) and by the Wellcome Sanger Institute Human Materials and Data Management Committee (approval number 17/113).

All donors are of white-European ancestries. The time between clinical diagnosis and date of biopsy was used to define the disease duration of a given individual. I further added six months to this number for all patients because symptoms often precede diagnosis by several months, and to avoid setting the disease duration to zero for patients who donated samples at the time of diagnosis. Dr. Tim Raine estimated the time of purine treatment by consulting electronic health records from NHS databases. He further annotated the biopsies as never, previously or actively inflamed using all available clinical data and NHS histopathology archives. The biopsy images (or an image of a second biopsy from the same site of the colon) were reviewed by Monika Tripathy, histopathologist. Despite these efforts, there remained some uncertainty about the past-inflammation status of biopsies annotated as never inflamed. Gaps may inevitably exist when patients have suffered from a disease for decades and samples appearing healthy at the time of sampling may have been affected in the past. In particular, there was uncertainty regarding the past inflammation histories of biopsies P29B1, P35B1 and P41B1. None of the patients had colorectal cancer, adenoma or dysplasia.

Biopsies from patients 1-26 were embedded in optimal cutting temperature (OCT) compound and sectioned, stained and fixed as previously described (Lee-Six et al., 2019). None of the samples were fixed in formalin. Subsequent biopsies were embedded in paraffin because this better preserved the morphology of the tissue. Yvette Hooks sectioned the biopsies (10-20 $\mu$m) and fixed the sections to 4 $\mu$m PEN membrane slides (11600288, Leica). I stained the sections with hematoxylin and eosin and dissected individual crypts using laser capture microdissection microscopy (LMD7000, Leica). I lyzed the cells using ARCTURUS PicoPure DNA extraction kit (Applied Biosystems) according to the manufacturer's instructions. DNA libraries were prepared by the Sanger Institute Core Pipelines team using a previously optimized method for obtaining DNA from low input material (Ellis et al., 2021).

The control cohort was obtained from our previous publication on somatic mutations in the normal colon described in Chapter 2 and in the introduction to this chapter (Lee-Six et al., 2019). It consists of seven deceased organ donors, 31 individuals who underwent colonoscopy following a positive faecal occult blood test in a screening programme (16 of which were not found to have an adenoma or a carcinoma and 15 of which had colorectal carcinoma,

Table 3.1 Clinical characteristics of the IBD patients

| Patient-ID | Sex | Age | Disease | Disease Duration | Years on purine | Years of smoking |
|---|---|---|---|---|---|---|
| patient1 | F | 59 | CD | 13 | 4 | NA |
| patient2 | M | 38 | CD | 11 | 5 | NA |
| patient3 | F | 27 | UC | 0.5 | 0 | NA |
| patient4 | F | 26 | CD | 3 | 0.5 | NA |
| patient8 | F | 25 | UC | 1.5 | 0 | NA |
| patient9 | F | 49 | UC | 5 | 4 | NA |
| patient10 | F | 51 | UC | 15 | 0 | 0 |
| patient11 | M | 80 | CD | 5 | 0 | 0 |
| patient12 | M | 31 | UC | 13 | 10 | 0 |
| patient13 | M | 31 | UC | 3 | 4 | NA |
| patient14 | F | 35 | CD | 18 | 1 | 0 |
| patient15 | M | 58 | UC | 21 | 0 | 0 |
| patient16 | M | 62 | UC | 4 | 2 | 0 |
| patient17 | M | 44 | UC | 18 | 1 | 12 |
| patient18 | F | 45 | CD | 14 | 8 | 8 |
| patient19 | M | 42 | CD | 13 | 2 | 0 |
| patient22 | F | 58 | UC | 28 | 0 | 0 |
| patient24 | M | 38 | CD | 20 | 0 | 0 |
| patient25 | F | 77 | UC | 19 | 0 | 0 |
| patient26 | M | 70 | UC | 7 | 7 | 10 |
| patient28 | F | 40 | UC | 3 | 0 | 0 |
| patient29 | M | 48 | UC | 5 | 0 | 0 |
| patient30 | F | 40 | CD | 10 | 0 | 0 |
| patient31 | M | 37 | UC | 2 | 1 | 0 |
| patient33 | F | 31 | CD | 20 | 2 | 0 |
| patient34 | M | 22 | UC | 4 | 3 | 0 |
| patient35 | M | 66 | UC | 25 | 7 | 45 |
| patient36 | F | 61 | UC | 27 | 20 | 30 |
| patient37 | F | 37 | CD | 25 | 0 | 0 |
| patient38 | F | 40 | UC | 4 | 5 | 20 |
| patient40 | M | 58 | CD | 40 | 7 | 0 |
| patient41 | M | 42 | CD | 19 | NA | 0 |
| patient42 | F | 31 | UC | New diagnosis | 0 | NA |
| patient43 | F | 36 | CD | 0.67 | 0 | NA |
| patient44 | F | 29 | CD | 3 | 1 | 6 |
| patient45 | F | 52 | CD | 24 | 10 | 0 |
| patient46 | M | 38 | UC | 21 | 0 | 0 |
| patient47 | M | 46 | CD | 0.16 | NA | NA |
| patient48 | F | 40 | CD | 14 | 1.5 | 0 |
| patient49 | M | 33 | UC | 5 | 4 | 9 |
| patient50 | F | 36 | UC | 4 | 5 | 3 |
| patient51 | M | 61 | UC | 16 | 16 | 30 |
| patient52 | F | 69 | UC | 30 | 0 | NA |
| patient53 | M | 61 | UC | 6 | 0 | 30 |
| patient54 | F | 50 | UC | 8 | 0 | 20 |
| patient59 | F | 50 | UC | 5 | 0 | 0 |

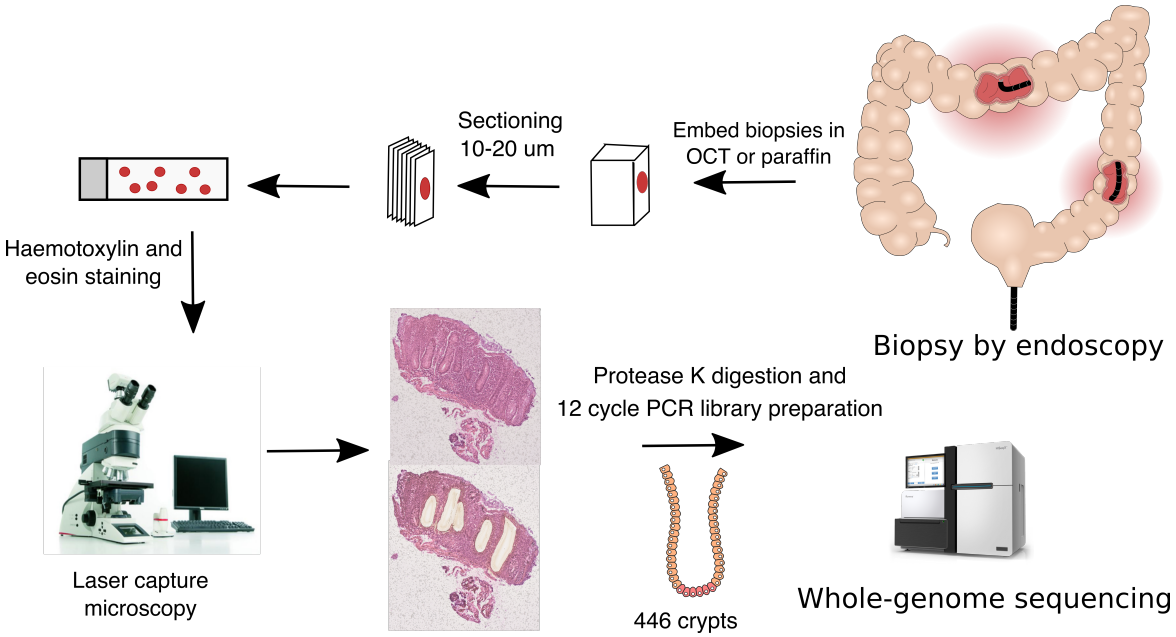



Fig. 3.1 **Overview of the experimental procedure.** Pinch biopsies were taken at the time of colonoscopy. These were embedded, either in OCT or paraffin blocks and the blocks sectioned. Histological sections were fixed to membrane slides and stained using hematoxylin and eosin. Laser capture microscopy was used to isolate crypts for whole genome sequencing.

although the biopsies used were distant from these lesions) and three paediatric patients who underwent colonoscopy to exclude IBD and who were found to have a histologically and macroscopically normal mucosa (Figure 3.2). I excluded one subject from the control cohort who had undergone chemotherapy and was a clear outlier in terms of mutation burden and showed an abnormal mutation profile.

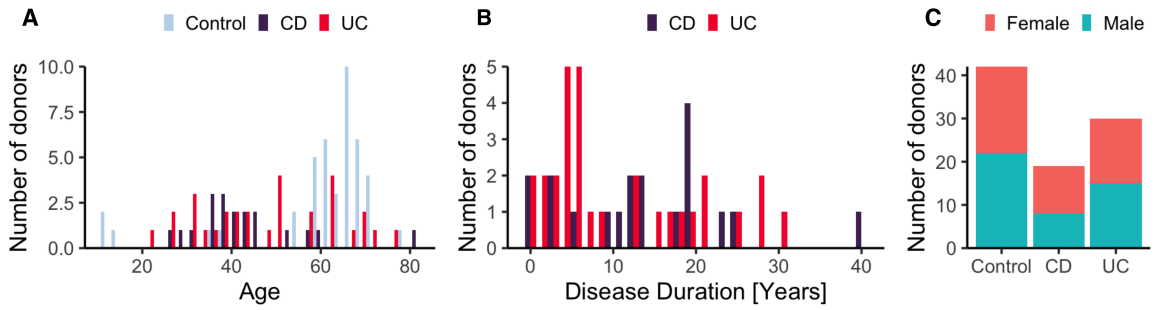



Fig. 3.2 **Cohort characteristics.** A) Age distribution by disease status. B) Disease duration distribution by IBD subtype. C) The sex distribution of the participants by disease status. CD: Crohn's disease. UC: Ulcerative colitis.

### 3.3.2 DNA sequencing

Samples from patient 1 through 19 were whole genome sequenced on Illumina XTEN machines using 150bp long paired-end reads by the Sanger Institute Core Sequencing team. Samples from other patients were whole genome sequenced on Illumina Htp NovaSeq 6000 machines using 150bp, paired end reads except for patients 60-62, which were whole exome sequenced on the same platform using the Human All Exon V5 bait set. Reads were aligned to the human reference genome (NCBI build37) using BWA-MEM by the Sanger Institute Core Informatics team. There was no difference in mutation burden between samples sequenced on different platforms (P>0.05, likelihood ratio test of linear mixed-effect models, as described below).

We multiplexed samples aiming to achieve a coverage of 15X. In reality, the median-median coverage was 18.2X as described below. We opted for 15X coverage after observing that the sensitivity of calling germline singletons at this depth exceeded 90% in other projects taking place in the lab. We hypothesized that the sensitivity when calling heterozygous somatic mutations in clonal colonic crypts would be similar. As described below, this proved to be the case.
Given the choice of sequencing one crypt at 30X or two crypts at 15X each, the latter strategy is preferable. Imagine the crypts have 2000 true somatic mutations on average and that the sensitivity is 90% at 15X but 99% at 30X. Splitting the sequencing resources between two crypts will result in the identification of 3600 somatic mutations compared with 1980 identified by sequencing one crypt. This increases the power for mutation signature extraction and driver discovery. Additionally, by sequencing more crypts we observe more coalescent events which gives a more complete picture of the clonal composition of the tissue.

### 3.3.3 Mutation calling and filtering

#### Substitutions

Base substitution calling was carried out in four steps: Discovery, filtering of the discovery set, genotyping and filtering of the genotypes. Mutations were first called using the Cancer Variants through Expectation Maximisation (CaVEMan) algorithm (Jones et al., 2016). CaVEMan is a Bayesian variant caller described in section 1.4.2. CaVEMan copy number options were set to major copy number 10 and minor copy number 2 for normal clones. Out of concern for field cancerization effect, patients 1 through 26, and patients from which only a few crypts were sequenced, were analysed using a matched normal sample dissected from non-epithelial tissue from one of the biopsies. As it became apparent that clones did not

stretch between biopsies, I stopped sequencing non-epithelial tissue control samples from patients if crypts were dissected from multiple biopsies.

The substitution calls were next filtered to remove mapping artefacts, common single nucleotide polymorphisms and calls associated with the formation of cruciform DNA structures during library preparation. The samples were compared against a 'normal panel' consisting of 75 unrelated normal samples to remove common SNPs and recurrent sequencing errors. I further removed mutations if the median alignment score of supporting reads was lower than 140 or if >50% of the reads supporting the mutations were clipped. The enzymatic fragmentation-based LCM workflow commonly results in errors within inverted repeats of the genome that are capable of forming cruciform DNA (Ellis et al., 2021). These erroneous variants tend to be in close proximity with other erroneous variants within the same read and the reads containing them have very similar alignment start positions. I applied a script written by Mathijs Sanders designed to remove these calls. It uses as features for filtering the variant position within the read, the standard deviation of the position of the variant relative to the alignment start site, and the median absolute deviation of the same (Ellis et al., 2021).

All sites where a somatic mutation was called in any crypt from a given patient were subsequently genotyped in all other samples from that patient by constructing read pileups and counting the number of mutant and wild-type reads. Only reads with a mapping quality of 30 or higher, and bases with a base quality of 30 or higher, were counted.

When matched normal samples were unavailable for the calling (see above), a large number of (rare) germline variants remained post filtering. I removed those by applying a script written by Tim Coorens to carry out an exact binomial test on the variant allele frequency (VAF) of each mutation across samples. True heterozygous germline variants should be present at a VAF of 0.5 in all samples from an individual. Across all samples from a given individual, I aggregated variant and read counts at sites where a single nucleotide variant was called in at least one sample. I then used a one-sided exact binomial test to distinguish germline variants from somatic variants. The null hypothesis was that germline variants were drawn from a binomial distribution with a probability of success of 0.5, or 0.95 for the sex chromosomes in men. The alternative hypothesis was that these variants were drawn from distributions with a lower probability of success. The resulting p-values were corrected for multiple testing using the Benjamini-Hochberg method. A variant was classified as somatic if $q < 10^{-3}$ , or $q < 10^{-2}$ if fewer than five crypts had been dissected for the patient. For variants classified as somatic, I fitted a beta-binomial distribution to the

number of variant supporting reads and total number of reads across crypts from the same patient. For every mutation, I determined the maximum likelihood overdispersion parameter ($\rho$) in a grid-based way (ranging the value of $\rho$ from $10^{-6}$ to $10^{-0.05}$). A low overdispersion captures artefactual variants because they appear to be randomly distributed across samples and can be modelled as being drawn from a binomial distribution. In contrast, true somatic variants will be present at a VAF close to 0.5 in some, but not in all crypt genomes, and are thus best represented by a beta-binomial with a high overdispersion. To distinguish artefacts from true variants, I used $\rho = 0.1$ as a threshold, below which variants were considered artefacts. The code for this filtering approach was similarly provided by Tim Coorens and is an adaptation of the Shearwater variant caller (Gerstung et al., 2014). Finally, I filtered out variants that were supported by fewer than three reads or where the sequencing depth was less than five

### Indels

Short deletions and insertions were called using the Pindel algorithm (Ye et al., 2009). Further description of Pindel can be found in section 1.4.2. I applied the same restrictions on median VAF and read counts as for substitutions, and germline indel calls were filtered using the same binomial filters as described above.

### Structural variants

Copy number variants were called using the BRASS algorithm. When a matched normal sample was not available for a patient, I used a clonally unrelated sample from the same individual to filter germline variants. All variants passing filters were manually reviewed in a genome browser. For discovery of deletions at fragile sites of the genome, I manually reviewed the three regions in all the genomes.

Somatic retrotranspostions were called by Hyunchul Jung using the TraFic algorithm (Rodriguez-Martin et al., 2020). Somatic events supported by read clusters without exact breakpoints were also included. To further identify somatic transduction events, translocation calls (i.e., read clusters) related with known L1 germline sources (Rodriguez-Martin et al., 2020) from the BRASS algorithm were manually examined by Hyunchul Jung, as were all somatic retrotransposition events. Chromosome aneuploidies and deletions or duplications affecting large areas of chromosomes or whole chromosome arms were called using the ASCAT algorithm (Raine et al., 2016; Van Loo et al., 2010).

### 3.3.4 Sensitivity analysis

To estimate sensitivity I dissected and sequenced five crypts twice. Assuming the same sensitivity in both samples, a maximum likelihood estimate for the sensitivity when mutations not present in either sample go unobserved is:

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

Where $n_2$ is the number of mutations called in both samples and $n_1$ is the sum of mutations called in only one sample. As sensitivity depends on coverage, which is uneven for the members of a pair, this estimate should be considered to be a lower bound.

I compared the sensitivity estimates for the five biological duplicates with internal sensitivity estimation for CaveMan carried out by Tim Coorens (Figure 3.3). This used 170 samples from the same individual sequenced to varying depths and, to remove the effect of clonality of the sample, estimated the sensitivity for calling heterozygous germline variants in these samples. The colonic crypts are expected to have slightly lower sensitivity than this estimate for the following reasons:

1. The curve assumes perfect clonality (median VAF of 0.5), but the median-median VAF in the IBD and control cohorts is 0.44.
2. The curve doesn't capture indels, for which sensitivity is expected to be slightly lower than for substitutions.
3. To increase specificity, I had required a coverage of 5 and at least 3 reads supporting the mutation, while standard for CaVEMan is coverage of 4 and 2 mutant reads.

### 3.3.5 Constructing phylogenetic trees

I used the MPBoot software (Hoang et al., 2018a,b) to create a phylogenetic tree for each patient. MPBoot is further described in section 1.4.4. I assigned mutations to branches using a maximum likelihood approach implemented in a script originally written by Nick Williams. I removed mutations which didn't adhere to the tree structure ($P<0.01$, maximum likelihood estimation).

### 3.3.6 Mutation rate comparisons between IBD patients and controls.

Any test for a difference in mutation burden between cohorts must take into account all factors, biological and technical, which correlate with disease and/or affect mutation calling sensitivity. For our comparison of IBD and normal, I fitted linear mixed effects models taking the following factors into account:

1. Age is the most important predictor of mutation burden and the age distribution of the two cohorts is different. I included a fixed effect for age in the models to account for this.

2. Mutation burden differs for different sectors of the colon (Lee-Six et al., 2019). The IBD cohort is enriched with samples from the left side, as this is the area predominantly affected in UC patients. I included a fixed effect for location within the colon to account for this.

3. Mutation counts are non-independent. I included in the models random effects for patient and for biopsy, with the random effect for biopsy nested within that for the patient.

4. Most embryonic mutations will be filtered as germline so at birth the mutation count is near zero. Therefore, I did not include a random intercept in the models but constrained the intercept to zero. The biological interpretation of this is that there are no somatic mutations present at time zero (birth).

5. The between-patient variance is likely greater in the IBD cohort as patients vary in the duration, extent and severity of their disease. The within-patient variance is also likely greater in the IBD cohort as biopsies taken from different sites of the colon vary in their disease exposure, number and duration of flares etc. To model this, I constructed a general positive-definite variance-covariance matrix for the random effects of patient and biopsy by cohort.

6. Any difference in the clonality of the colon between IBD patients and controls will affect the relative sensitivity to detect somatic mutations. To account for this, I adjusted the branch lengths of the phylogenetic trees and used the adjusted mutation counts as the response variable in the models. Mutations with low variant allele frequencies (VAFs) will be missed at low coverage. Therefore, for each crypt, I first fitted a truncated binomial distribution to the VAF distribution of the crypt to estimate the true underlying median VAF (this is different from 0.5 because recent mutations may not yet have been fixed in the stem cell niche, and because of contamination of lymphocytes

and other cells from the lamina propria, which do not carry the same somatic mutations as the epithelial cells). I next simulated 100,000 mutation call attempts by drawing the coverage of each call from a Poisson distribution, with the lambda set as the median coverage of the sample, and multiplying that with the median VAF estimate from the truncated binomial. The resulting value represents the number of reads that carry the mutated allele. I calculated sensitivity for the sample, $S_s$, as the fraction of draws that resulted in four or more mutant reads, which is the number required by CaVEMan to call a mutation. The sensitivity of a branch with n daughter crypts, $S_b$, was then calculated as:

$$S_b = 1 - (1 - S_{s_1}) \times ... * (1 - S_{s_i}) \times ... \times (1 - S_{s_n})$$

The adjusted mutation count is thus the observed mutation count divided by the sensitivity of the branch. In this way, the mutation count of clones formed of stand-alone crypts is augmented more than that of branches with multiple daughter crypts. Even after these steps, a significant effect of coverage remained (38 mutations per 1X increase in coverage, $P = 2.8 \times 10^{-13}$) and a fixed effect for coverage was included in the models.

I compared the fit of these models with and without disease duration as a fixed effect using likelihood ratio tests. The disease durations for never inflamed regions of the colons of IBD patients were set to zero.

As comparatively few structural variants are found in the dataset, I used Poisson regression within a generalized linear mixed effects framework to test for differences in structural variant number between cases and controls. I included the same random and fixed effects described above for base substitutions and indels and compared models with and without disease duration using likelihood ratio tests.

### 3.3.7 Mutational signature extraction and analyses

Refer to section 1.2.1 for a definition of mutational signatures and a discussion on the mutational processes underlying them. Methods for mutational signature extraction are described in section 1.4.3. I extracted mutational signatures using the 'hdp' package in R which implements a hierarchical Dirichlet process based method for signature extraction (Roberts, 2018). This has the advantage of allowing simultaneous fitting to existing signatures and discovery of new signatures. I pooled the control and the IBD data and extracted signatures from the

combined dataset for indels and single base substitutions separately. I mapped mutations to branches of a phylogenetic tree and treated each branch with more than 50 mutations as a sample. To set known signatures as priors, I initialized the HDP process with pseudo-count nodes where mutations were distributed by known signatures. I used signatures reported in colorectal cancer and also included signature 32, which is attributed to azathioprine therapy (Inman et al., 2018), and signature 35, attributed to platinum-based chemotherapy, as there are patients in our cohort with a history of using these drugs. Using the PCAWG terminology (Alexandrov et al., 2020), the prior signatures used were SBS1, SBS2, SBS3, SBS5, SBS13, SBS16, SBS17a, SBS17b, SBS18, SBS25, SBS28, SBS30, SBS32, SBS35,SBS37, SBS40, SBS41, SBS43, SBS45 and SBS49 for substitutions and ID1, ID2, ID3, ID4, ID5, ID6, ID7, ID8, ID10 and ID14 for indels.

I used expectation maximization to deconvolute the HDP components into known PCAWG signatures. In particular, the cosine similarity between the HDP component corresponding to SBS1 was <0.95 and visual inspection of the component suggested it was contaminated by SBS5 and SBS18, which are highly correlated with SBS1. I used expectation maximization to break the component down into PCAWG signatures and then reconstituted the components using only those PCAWG signatures that accounted for >10% of the mutations (this was done to avoid overfitting). This helped resolve the correlation between SBS1, SBS5 and SBS18. I merged components corresponding to SBS5 and SBS40 under the name of SBS5, as both are flat, with the mutation probability distributed near uniformly across mutation classes, and so are difficult to distinguish. No other components had cosine similarity <0.95 with their corresponding signatures and other PCAWG signatures accounting for >10% of the mutations.

### 3.3.8 Selection analyses

To search for mutations under positive selection, I used the dNdScv method (Martincorena et al., 2017), see section 1.4.5. I included never inflamed samples from the IBD cohort in the analysis as some uncertainty existed regarding the annotation of a handful of never-inflamed biopsies and I estimated that since driver mutations are quite rare in the colon, the analysis would suffer more from potential exclusion of drivers than from inclusion of more neutral mutations. I used the Benjamini-Hochberg method to correct for multiple testing.

To look for enrichment of mutations in pathways I defined a priori, 15 gene-sets, with input from Drs. Carl Anderson and Tim Raine. We included all genes found to be under selection in colorectal cancer (Priestley et al., 2019) as well as a list of genes significant

in a pan-cancer analysis of solid tumours (Priestley et al., 2019). We also chose a set of cellular pathways known to be important in IBD pathogenesis and epithelial homeostasis. The Reactome database was used to define the pathways (Fabregat et al., 2018). We chose the cytokine pathways TNF-Signaling, TNFR2, IL6, TGFb and IL17 for testing. We also defined a combined list of cytokines which included all of the above as well as IFNg, IL10, IL20, IL23, IL28, and IL36. We also decided to test other pathways shown by the Anderson group and others through genome-wide association studies to be important in IBD pathogenesis (de Lange et al., 2017). These were Toll-like receptor cascades, NOD-signaling, autophagy, unfolded protein response and epithelial cell-cell junctions. We included the PIP3/AKT signaling pathways as it is downstream of many of the pathways defined above and I had discovered two large scale deletions affecting genes in this pathway before performing the analysis. Finally, we defined a list of genes known to cause early-onset, monogenic forms of IBD. Many of the genes defined in the literature affect myeloid cell development and cause severe immunodeficiencies (Uhlig, 2013; Uhlig et al., 2014). We restricted our analysis to the union of monogenic-IBD genes which either are specifically thought to affect epithelial cells or were members of any of the pathways above.

I extracted global dN/dS values for missense and truncating variants separately and used the Benjamini-Hochberg method to correct for multiple testing.

## 3.4 Results

The final dataset comprised whole genome sequence data from 446 crypts, microdissected from endoscopic biopsies taken from 28 UC patients and 18 CD patients, as well as whole exome sequence data from 187 crypts dissected from 2 UC patients and 1 CD patient. This was combined with whole-genome sequence data from 412 crypts sequenced as part of the study of the normal colon described in Chapter 2 and in the introduction to this chapter (hereafter referred to as the control data).

The median sequencing coverage of the dissected crypts was 18.2X for the whole genome sequenced crypts and 30.2X for the whole exome sequenced crypts from the IBD patients. The median coverage of the crypts in the control cohort was 16.3X. The clonality of crypts was comparable in both cohorts and the median-median VAF was identical at 0.44 (Figure 3.3 A and B). Five pairs of crypts were sequenced twice and could be used to estimate the sensitivity of the mutation calling (Table 3.2). From this material I estimated the average sensitivity to be 79%. However, in four out of five duplicate pairs coverage of both crypts is

Table 3.2 Sensitivity analysis of technical duplicates.

| Original crypt | Coverage | Duplicate | Coverage | Sensitivity |
|---|---|---|---|---|
| P12B1_16 | 12.2 | P12B1_18 | 11.5 | 0.88 |
| P12B2_7 | 11.2 | P12B2_13 | 11.0 | 0.64 |
| P12B2_8 | 13.7 | P12B2_3 | 13.1 | 0.71 |
| P17B2_7 | 13.4 | P17B2_17 | 10.7 | 0.75 |
| P44B1a_1 | 20.5 | P44B1_1_repl | 17.6 | 0.97 |

<14X, which is well below the median of the study. In the one pair where crypts had 17.6X and 20.5X coverage, I estimated sensitivity of 97% (Table 3.2). These values were compared against an internal validation set (see methods) and interpolation from the curve suggests that at the median coverage for cases (18.2X) and controls (16.3X), sensitivity is 97% and 95%, respectively (Figure 3.3).

### 3.4.1 IBD increases the substitution and indel rate of normal colonic epithelium.

To assess if IBD is associated with a difference in the mutation burden of the colonic epithelium, I focused only on the WGS crypts from the IBD and control cohorts (only three IBD patients underwent whole-exome sequencing). I fitted linear mixed-effects models (LMMs) to estimate the independent effects of age, disease duration and biopsy location on mutation burden, while controlling for the within-patient and within-biopsy correlations inherent in the sampling strategy, as described in the methods section. I estimated the effect of IBD to be 55 substitutions per crypt per year of disease duration (35-75 95% CI, $P = 3.1 \times 10^{-7}$, LMMs and likelihood ratio test - Figure 3.4). These mutations are in addition to the 40 (31-50, 95% CI) substitutions I estimated are accumulated on average per year of life under normal conditions, suggesting that mutation rates are increased, on average, 2.3-fold in regions of the IBD-affected colon. Compared to controls, patients with IBD had greater between-patient variance in mutation burden (SD=776 versus 383 substitutions and SD=80 versus 34 indels for cases and controls, $P = 4.2 \times 10^{-8}$ and $P = 1.1 \times 10^{-16}$ respectively - LMMs and likelihood ratio test) and greater within-patient variance (SD = 955 versus 407 substitutions and SD = 81 and 18 indels for cases and controls, P=0.032 and P=0.0011, respectively). The increased between-patient variance likely reflects differences in inflammation exposure not captured by disease duration, as it doesn't account for variable disease severity, response to treatment etc. among patients. The increased within-patient variance probably reflects region-to-region differences in disease severity along the colon. I similarly estimated an

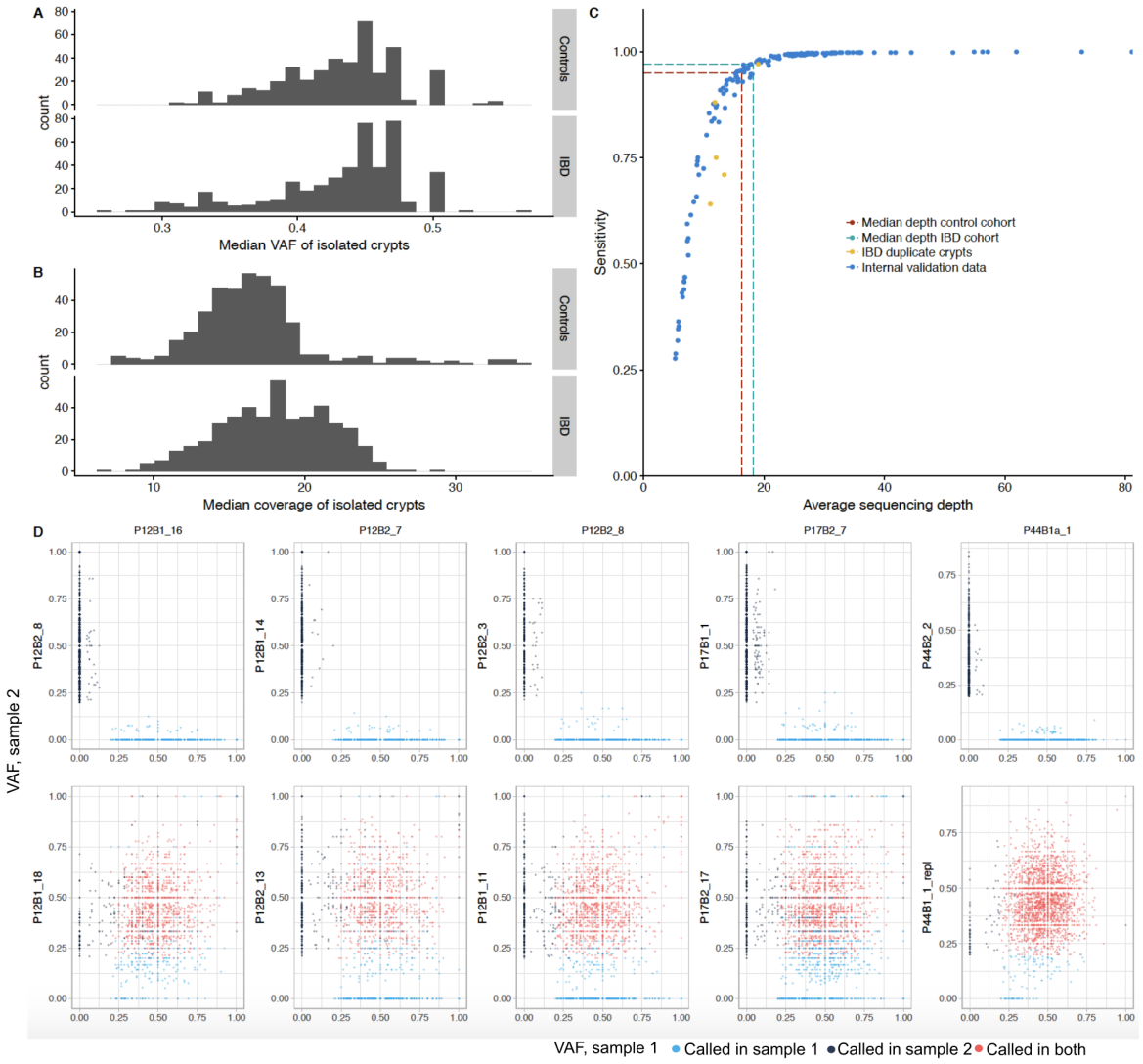



Fig. 3.3 **Clonality, coverage and sensitivity of crypts and mutation calls.** A) The distribution of the median variant allele fraction (VAF) of mutations called in each crypt. The median-median VAF of the two cohorts is identical (0.44). B) The distribution of the median coverage of sequenced crypts. C) Internal analysis of CaVEMan sensitivity. The dashed lines show interpolation of the sensitivity given the median coverage of cases (18.2X - 97% sensitivity) and controls (16.3X - 95% sensitivity). The yellow dots represent biological duplicates where sensitivity was estimated by dissecting and sequencing the same crypts twice (Table 3.2). D) VAFs of variants called in crypts that were sequenced twice (referred to as sample 1). Each dot represents a variant. The VAFs are compared against variants called in unrelated crypts (top) and in biological duplicates (bottom, referred to as sample 2). The high concordance between biological duplicates but not between unrelated samples suggests high specificity.

increase in the indel burden in IBD, with an excess of 6.8 indels per crypt per year of IBD (5.0 - 8.7 95% CI, $P = 5.7 \times 10^{-11}$ - Figure 3.4) in addition to the estimated 1.0 (0.3-1.7 95% CI) indel that is accumulated per crypt per year of life. As shown in Figure 3.4, a handful of clones and patients had a much higher mutation burden than expected given their age. This is partially driven by the effect of smoking and cancer driver status, as discussed below. The effect of IBD on the mutation burden remains significant if crypt carrying driver mutations or the five IBD patients with the highest mutation burdens are excluded (P=0.0014 and 0.0099 for substitutions and $P = 6.8 \times 10^{-6}$ and $1.1 \times 10^{-5}$ for indels, respectively). There was no significant difference in the mutation burden between UC and CD patients.

Smoking history was available for a subset of the IBD cohort (362 crypts from 35 patients), encoded as the number of years of active smoking (no estimation of pack-years was available). In this restricted dataset, there is a significant effect of smoking duration of 49 (18 - 81 95% CI, P=0.0024) substitutions and 5.3 (2.3 - 8.2 95% CI, $P = 6.5 \times 10^{-4}$) indels per crypt per year of smoking. The effect of disease duration is unchanged, suggesting the estimated effect of smoking in the model is not driven by differences in smoking habits between cases and controls. Smoking has been reported to increase the risk of CD and be protective for UC (Mahid et al., 2006) but I found no interaction effect between smoking and disease type (P=0.68). Smoking status was not available for the control cohort.

### 3.4.2 Mutational signatures in IBD affected epithelium

The somatic mutations found in the cells of a colonic crypt reflect the mutational processes that have acted on the stem cells and their progenitors since conception. Distinct mutational processes each leave a characteristic pattern, a mutational signature, within the genome, distinguished by the specific base changes and their local sequence context (Alexandrov et al., 2020, 2013a), as discussed in section 1.2.1.

I extracted mutational signatures jointly for IBD and control crypts and discovered 12 substitution signatures (SBS) and five indel signatures (ID), all of which have been previously observed in tissues from individuals without IBD (Figure 3.5).

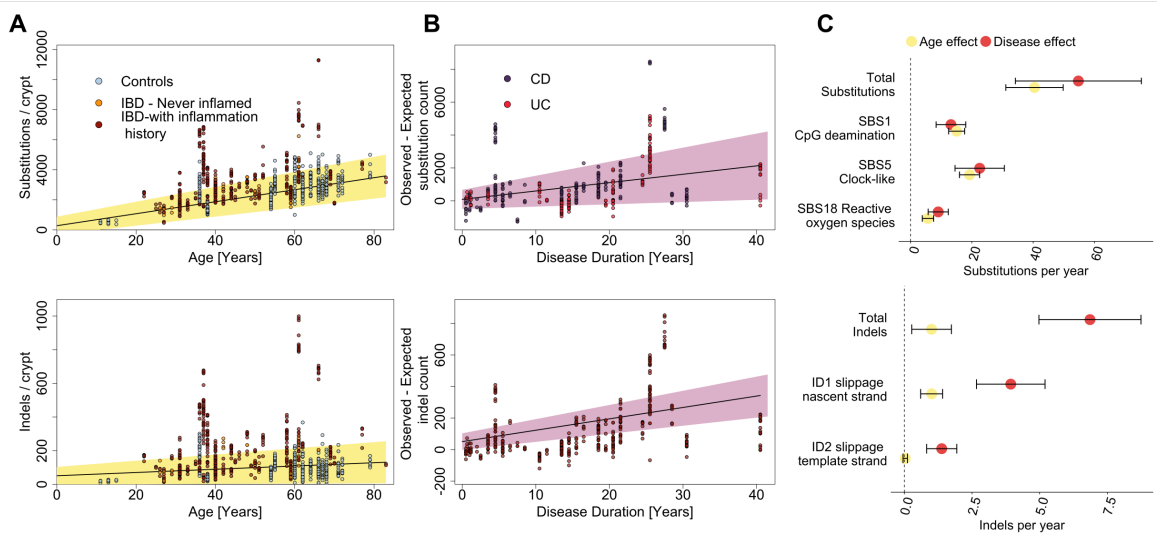



Fig. 3.4 **Mutation burden in the IBD colon.** Substitution (top) and indel (bottom) burden as a function of age. Each point represents a colonic crypt and is coloured by disease status. The line shows the effect of age on mutation burden as estimated by fitting a linear mixed effects model, correcting for sampling location, sequencing coverage and the within-biopsy and within-patient correlation structure, considering both IBD cases and controls. The yellow shaded area represents the 95% confidence interval of the age effect estimate. B) Estimated excess of substitutions (top) and indels (bottom) in crypts from IBD patients as function of disease duration. The pink shaded area represents the 95% confidence interval of the disease duration effect estimate. C) A comparison of the effects of age and disease duration on the total mutation burden and on the burden of mutational signatures that associate with IBD duration. Error bars represent the 95% confidence intervals of the estimates. IBD: Inflammatory bowel disease. CD: Crohn's disease. UC: Ulcerative colitis. SBS: Single base substitution signature. ID: Indel signature.

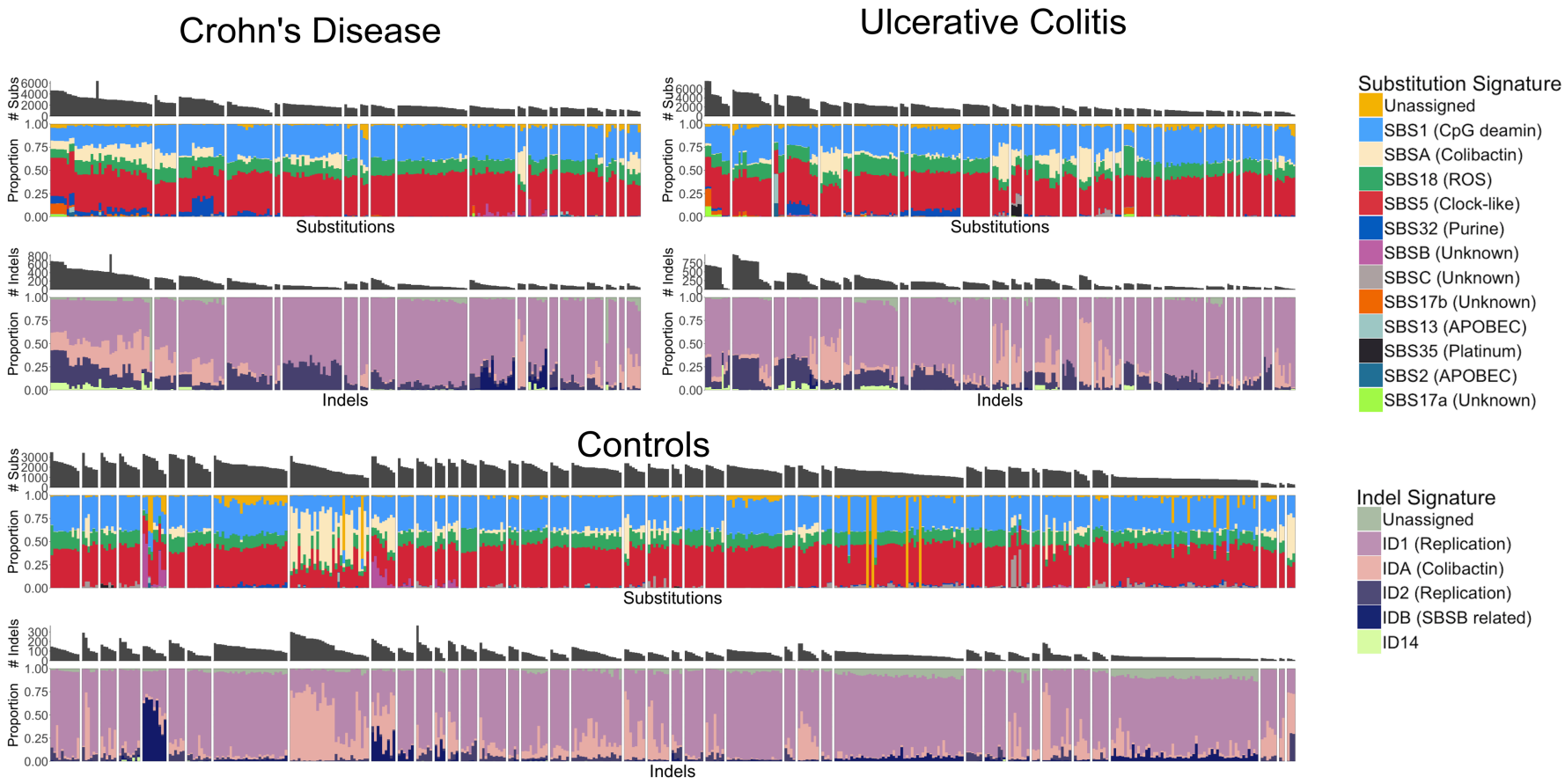



Fig. 3.5 **Mutational signatures in colonic crypts.** stacked barplot showing the proportional contribution of single-base-substitution (SBS) signatures (Top) and Indel (ID) signatures (Bottom) to the mutation burden of each crypt. Crypts are grouped by patient and crypts from CD, UC and controls are shown separately. Signature nomenclature is the same as in COSMIC. The 'Unassigned' component represents uncertainty of the signature extraction.

As expected, the HDP components showed high cosine similarities with their corresponding PCAWG signatures (Figure 3.6). The lowest cosine similarity was between the component corresponding to substitution signature 5, which is a flat and featureless signature (meaning the probability is roughly uniformly distributed across all mutation classes) of the type which is hardest to extract.

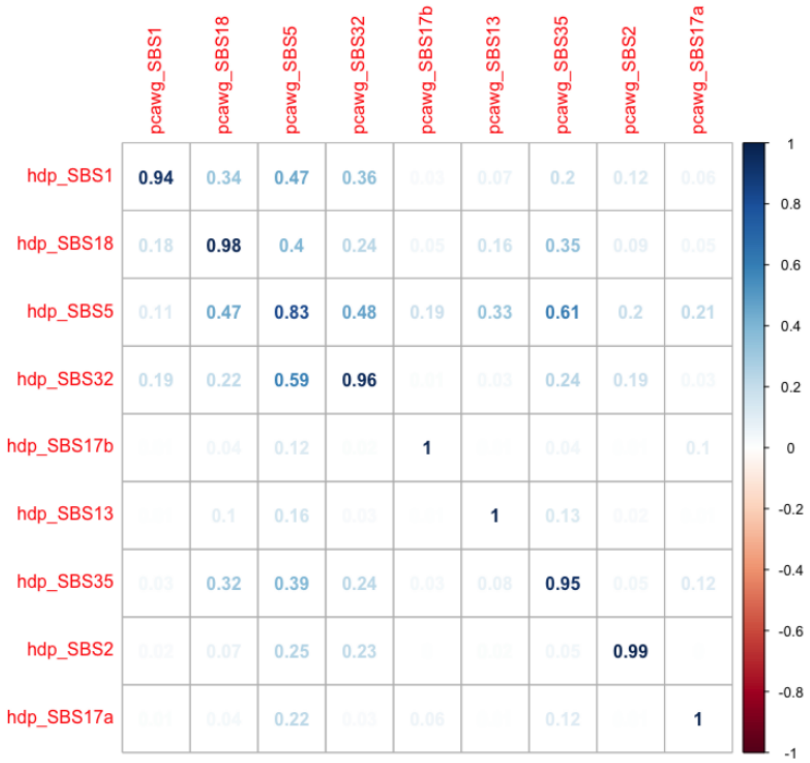



Fig. 3.6 **Cosine similarity between HDP components and corresponding PCAWG signatures.**

Comparing the IBD cases and controls, I found that approximately 80% of the increase in mutation burden in cases is explained by signatures that are also found ubiquitously in normal colon (Blokzijl et al., 2016; Lee-Six et al., 2019) (Figure 3.4C). These are substitution signatures 1, 5 and 18 and indel signatures 1 and 2, as defined in COSMIC, which cause an increase of 13 (8-18 95% CI), 23 (15-30 95% CI) and 9 (6-12 95% CI) substitutions per crypt per year of disease, respectively ($P = 2.4 \times 10^{-7}$, $1.0 \times 10^{-7}$ and $3.2 \times 10^{-7}$), and 4.3 (3.3-5.4 95% CI) and 1.7 (1.1-2.3, 95% CI) indels per crypt per year, respectively ($P = 4.0 \times 10^{-12}$ and $P = 9.5 \times 10^{-8}$ , LMMs and likelihood ratio tests). Substitution signatures 1 and 5 are clock-like and thought to be associated with cell proliferation, while signature 18 has been linked with reactive oxygen species (Alexandrov et al., 2020). The

indel signatures ID1 and ID2 are both thought to be the result of polymerase slippage during DNA replication (Alexandrov et al., 2020).

The remaining 20% of the increase in substitution burden is a consequence of rarer mutational processes and treatment. For example, 96 crypts had over 150 mutations attributed to purine treatment in a subset of seven IBD patients, five of whom have a documented history of such treatment. However, the number of mutations attributed to purine was not associated with purine therapy duration, and some patients showed large mutation burdens despite brief, or indeed no, documented exposure (Figure 3.7). Crypts dissected from the same patient sometimes showed vastly different mutation burdens (Figure 3.7). The largest range was observed for patient 40, who has a 7 year history of purine treatment. The estimated burden of the purine signature in crypts from this patient ranged from 69 to 1005 mutations. I speculate that it may be rapidly dividing cells at the time of treatment that are most affected by purine since the burden of the purine signature (SBS32) associates with the number of driver mutations (see below). This interpretation is supported by recent work on the mutational signatures of cancer therapy. Pich et al showed how 5-fluorouracil only leaves a mutational signature on cancer cells actively dividing at the time of treatment (Pich et al., 2021). Thiopurines, which mimic the structure of metabolic purines, would similarly preferentially be incorporated into the DNA in cells that are actively dividing at the time of treatment. Thiopurine use has been associated with higher overall cancer risk in epidemiological studies, but this is mostly driven by an effect on lymphoid cancers and possibly on urinary tract cancers, but not colorectal cancers (Adami et al., 2016; Pasternak et al., 2013). The relationship between purines and colon cancer is complicated and requires further study. On one hand, these results show purine-related mutations accumulating in the crypts of a subset of patients but on the other, effective purine treatment may prevent disease related mutagenesis.

I also observed great inter-patient variation in the mutation burden attributed to purine treatment. For example, Figure 3.8A shows a phylogenetic tree of a patient with long term exposure to azathioprine. The tree is overlaid with the signature exposure of each branch and shows that this patient did not accrue any purine-related mutations. In contrast, Figure 3.8B shows a second patient who received azathioprine for only two weeks and mercaptopurine for two weeks and had significant adverse reactions to both drugs. This brief treatment resulted in a median of 204 mutations (range: 120-374) attributed to purine treatment in the crypts from this individual.

Five signatures previously discovered in the normal colon (Lee-Six et al., 2019), SBSA, SBSB and SBSC, IDA and IDB were also present in the context of IBD. SBSA and IDA and

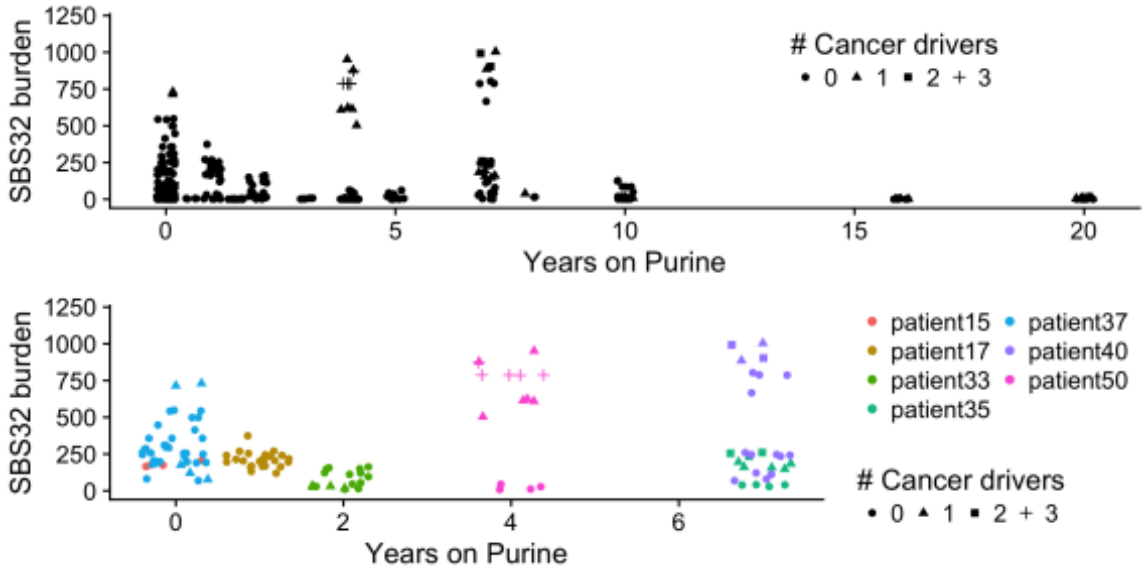



Fig. 3.7 **Burden of substitution signature 32 as a function of purine treatment duration.** (Upper) Purine signature burden for all patients with known duration of purine treatment. (Lower) Patients for which any crypt carries more than 150 mutations attributed to SBS32.

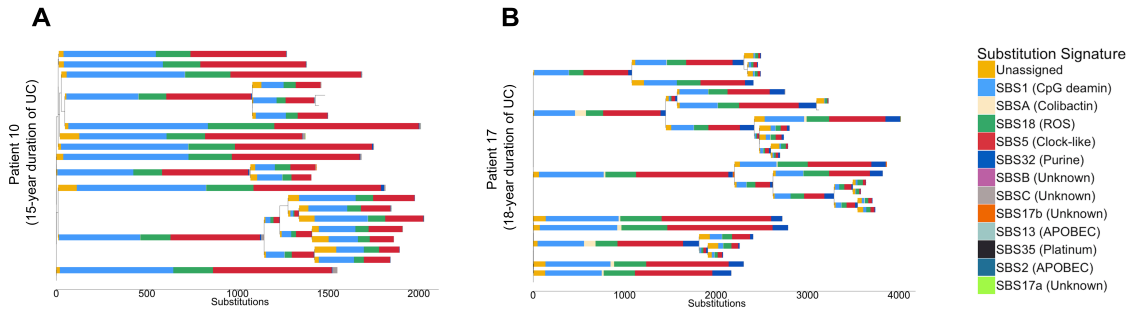



Fig. 3.8 **Phylogenetic trees of two patients with widespread ulcerative colitis who received purine treatment** The colours of the branches reflect the relative contribution of each mutational signature extracted for those branches. A) The patient received azathioprine treatment for 10 years but shows no SBS32 burden (dark blue). In contrast, the patient on the right received azathioprine for 2 weeks and mercaptopurine for 2 weeks and had significant adverse reactions to both drugs. SBS32 is found in most crypts from this patient. All crypts are from biopsies of actively inflamed regions.

SBSB and IDB are highly correlated (Figure 3.9) and likely represent the same underlying mutational processes.

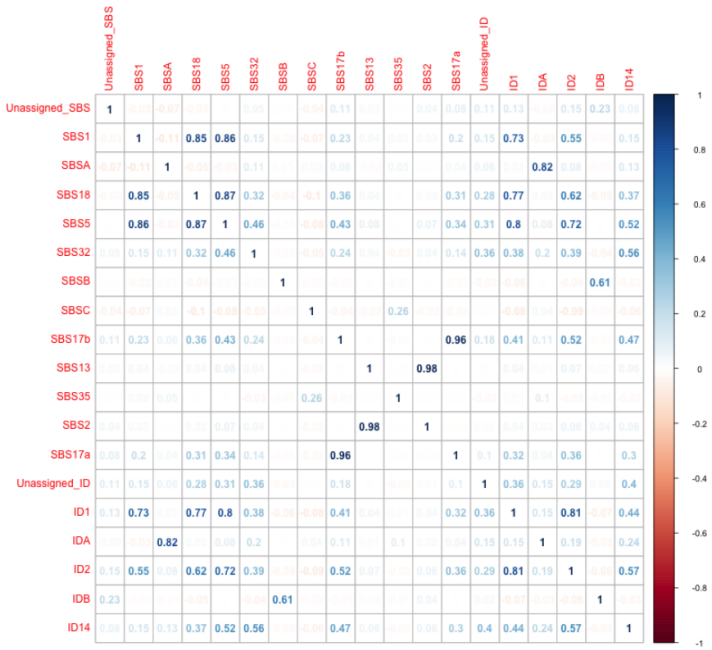



Fig. 3.9 **A correlation matrix between mutational signatures identified in the IBD and healthy colon.**

SBSA and IDA are of particular interest since they have recently been shown to be caused by the genotoxin colibactin, which is produced by bacteria harboring a polyketide synthases (pks) pathogenicity island (Pleguezuelos-Manzano et al., 2020). *pks*+ E. coli have been reported at increased frequency in IBD (Arthur et al., 2012), but I found no relationship between SBSA or IDA burden and disease status or disease duration after correcting for higher burden of both in the left-side of the colon (the site primarily affected in UC) (Figure 3.10).

Signatures SBSB, SBSC and SBS32 have not been reported in studies of sporadic colorectal cancers (Alexandrov et al., 2020), perhaps due to the comparative complexity and diversity of cancer mutation profiles. However, SBS32 would only be expected in patients receiving purine therapy and so would not be present in sporadic colorectal cancers. These signatures have also not been reported in studies of colitis-associated colorectal cancers but this is likely due to a relative lack of power due to the small number of sequenced exomes (Baker et al., 2018; Din et al., 2018; Robles et al., 2016).

Signatures 2 and 13, which are associated with APOBEC activity, and signatures 17a and 17b, which are of unknown aetiology, were active in a small number of crypts with high

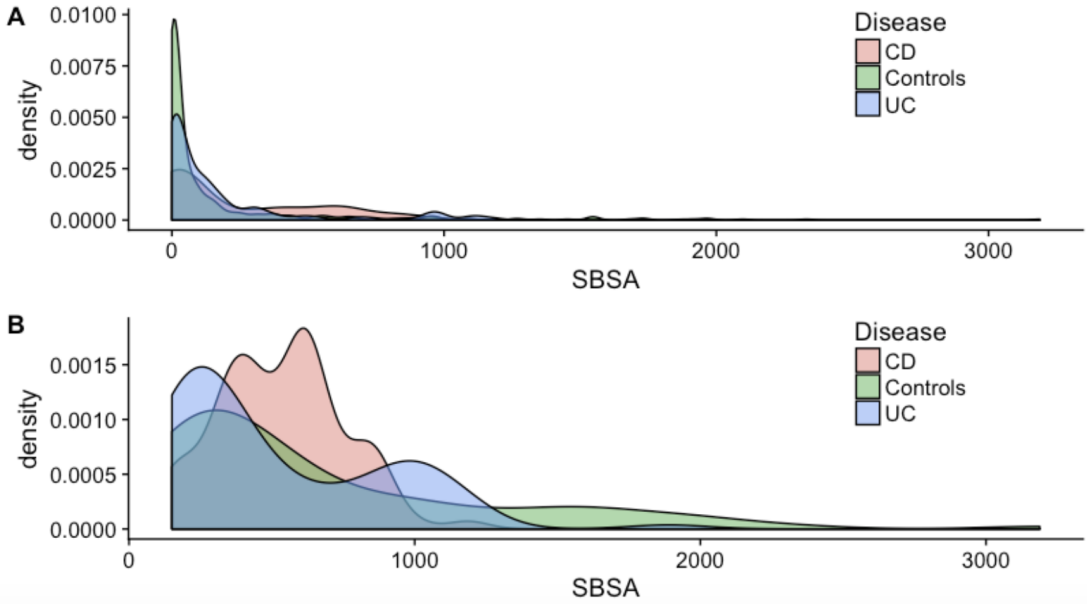



Fig. 3.10 **Colibactin signature exposure by disease.**A) A density plot showing the distribution of SBSA by disease type across all crypts. B) A density plot of crypts with more than 150 mutations attributed to SBSA.

mutation burdens. SBSB, SBSC, SBS17a/b and SBS2/SBS13 are too rare for this study to be well-powered to detect any difference between IBD and controls or to associate these with any clinical feature documented in our metadata. Finally, I found signature 35, associated with platinum compound therapy, in one patient with a history of platinum treatment for squamous cell carcinoma of the tongue. The patient received $40mg/m^2$ of cisplatin therapy on a weekly basis. He completed three of six planned treatment cycles with therapy termination due to toxicity. This relatively brief treatment resulted in a medium of 430 mutations (range 350-461) per crypt that were attributed to signature 35, equivalent to about 10 years of normal mutagenesis.

### 3.4.3 IBD associates with the burden of structural variants

The burden of structural variants is modest in both datasets (Figure 3.11) but for IBD, the occasional clone carried a large number of CNVs and retrotranspositions (Figure 3.11 A and B). The numbers of CNVs and retrotranspositions are associated with IBD duration. I estimated the CNV mutation rate to be 0.067 CNVs per crypt per year of disease (0.027 - 0.11 95% CI, $P = 1.1 \times 10^{-3}$, Likelihood ratio test of mixed-effects Poisson regressions) and the retrotransposition mutation rate to be 0.065 (0.018 - 0.11, 95% CI, $P = 6.9 \times 10^{-3}$). This corresponds to one CNV per crypt every 14.9 years of disease duration and one retrotranspo-

sition event every 15.4 years of disease duration on average. However, a handful of clones accumulated many structural variants (SVs), while the majority had none, suggesting that the processes driving their acquisition may be episodic rather than continuous. This would be in line with findings from other reports linking rapid accrual of SVs with the transition from normal to dysplastic mucosa (Baker et al., 2018) and cancers accruing copy number gains in a punctuated manner (Gerstung et al., 2020).

I found a higher fraction of IBD crypts carrying aneuploidies than in controls (43/419 compared with 13/412, Figure 2c). However, this was driven by large clones carrying aneuploidies and the number of events was not significantly associated with disease duration (P=0.38). The numbers of CNVs, retrotranspositions and aneuploidies are associated with higher substitution burden (112 (49-175 95% CI, $P = 6.4 \times 10^{-4}$), 59 (38-81 95% CI, $P = 1.5 \times 10^{-7}$) and 199 (65-331 95% CI, $P = 3.7 \times 10^{-3}$), respectively) and retrotranspositions and CNVs are associated with higher indel burden (11 (8-14 95% CI, $P = 2.6 \times 10^{-12}$) ,and 17 (10-24 95% CI, $P = 6.7 \times 10^{-6}$), respectively).

### 3.4.4 IBD creates a patchwork of millimeter-scale clones

As described in Chapter 2, colonic crypts divide by a process called crypt fission, whereby a crypt bifurcates at the base and branching elongates in a zip-like manner towards the lumen. I estimated the crypt fission rate as part of a study profiling the somatic mutation landscape of the normal colon (Lee-Six et al., 2019). This is described in Chapter 2 of this thesis. I estimated that each crypt fissions on average only once every 27 years and other sources have estimated even lower crypt fission rates (Baker et al., 2018; Nicholson et al., 2018). I did not apply Approximate Bayesian Computation to estimate the crypt fission rate in IBD as I did for the normal colon in Chapter 2. This is because, in contrast to the normal colon, the crypt fission rate is unlikely constant in IBD. Presumably, it is rapidly accelerated during a flare up and then gradually slows down to normal levels as the patient goes into remission. Rather than estimating the average increase over a period of temporarily increased crypt fission, I will describe the clonal expansions which represent the permanent consequence of the increased crypt fission in IBD.

Compared to normal colon, I found much larger clonal expansions in IBD patients, evident of numerous crypt fission events occurring late in molecular time. I observed several examples of individual clones spanning entire 2-3 mm endoscopic biopsies (Figure 3.12, Figure 3.13, Figure 3.14 and Figure 3.15a). The ability to estimate clone sizes is restricted by the small size of the biopsies, which are pinch biopsies a few millimeters across (Methods).

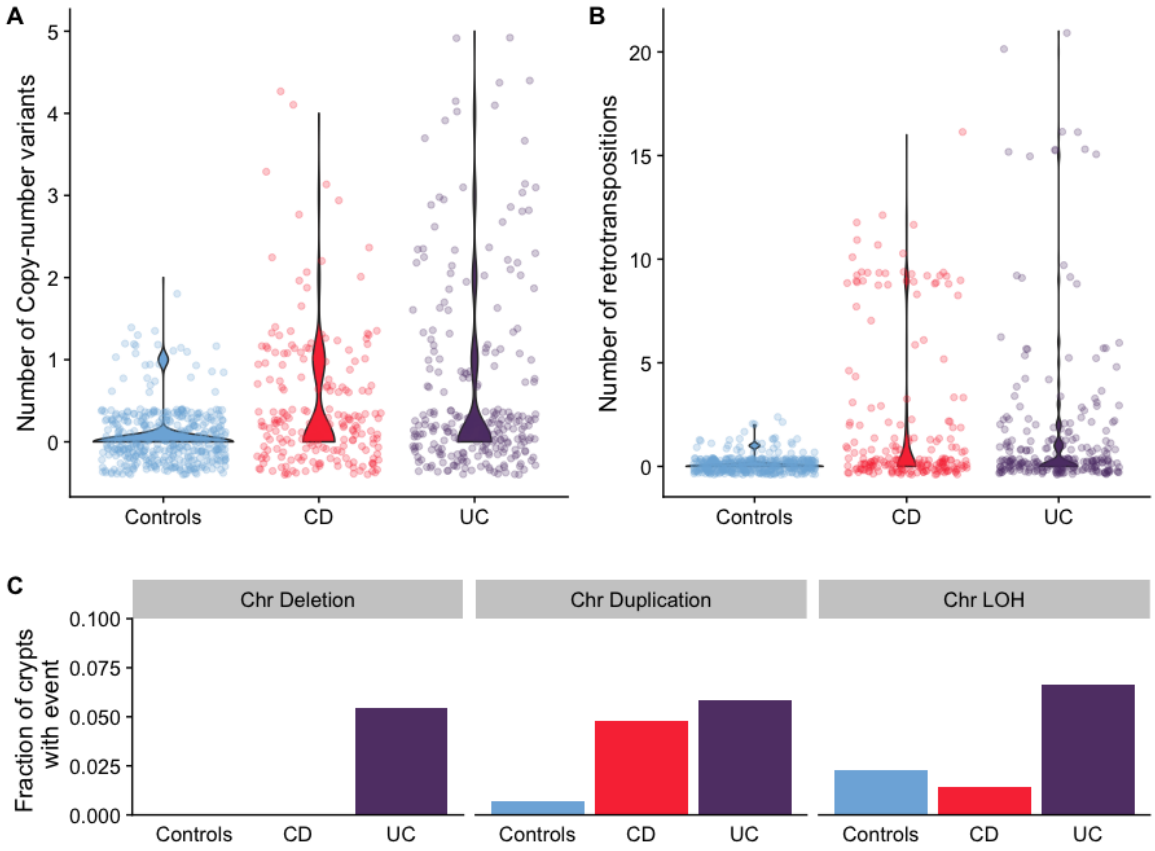



Fig. 3.11 **Burden of structural variants in inflammatory bowel disease affected colon compared with IBD-unaffected colon.** A) Number of copy number variants in IBD subtypes compared with controls. B) Number of somatic retrotranspositions in IBD subtypes compared with controls. C) Fraction of crypts with inflammation history that carry chromosomal aneuploidies.

However, when the same inflamed or previously inflamed region of the colon was biopsied more than once, on only one occasion out of 19 such biopsy pairs did I observe a clone stretching between biopsies that were taken several millimeters apart (Figures 3.14 and 3.13), while most biopsies contained more than one clone. To improve our ability to detect larger clones, three patients were sampled more broadly. From each patient, nine biopsies were taken forming a 3x3 grid with 1 cm separating biopsies. I dissected 187 crypts from these biopsies and performed whole exome sequencing on individual crypts. Phylogenetic trees were reconstructed based on somatic mutations identified (Figure 3.15B-D). While clonal expansions within biopsies were common, I found clones extending between neighboring biopsies in only one of these patients, who showed a very high degree of clonality (Figure 3.15D).

A substantial body of evidence exists documenting widespread clonal expansions giving rise to dysplasia and ultimately to colorectal cancer in IBD (reviewed in (Choi et al., 2017)). Colitis-associated colorectal cancers, which are enriched with synchronous lesions (Choi et al., 2015; Lam et al., 2014), commonly grow from a background of a pre-cancerous field which has expanded many centimeters or even the whole length of the colon (Galandiuk et al., 2012; Leedham et al., 2009). Mutations in *TP53* are thought to be especially prominent in the growth of these clones but aneuploidies and *KRAS* mutations are also commonly observed (Galandiuk et al., 2012; Holzmann et al., 1998; Leedham et al., 2009). In this material of non-dysplastic tissue from individuals without colorectal neoplasia, I find smaller clones and only a total of five mutations in *TP53*, *KRAS* or *APC*. In summary, IBD-affected regions are generally not dominated by a single major clone, but are more accurately viewed as an oligoclonal patchwork of clones that often grow considerably larger than in healthy colon.

Clonal spread within the tissue presumably doesn't occur linearly with time but rather takes place rapidly during or shortly following a flare up of the disease. Nevertheless, it would be interesting to study clonal spread as a function of disease duration similar to how the mutation rate is described above. Unfortunately, our sampling was too sparse to enable such an analysis.

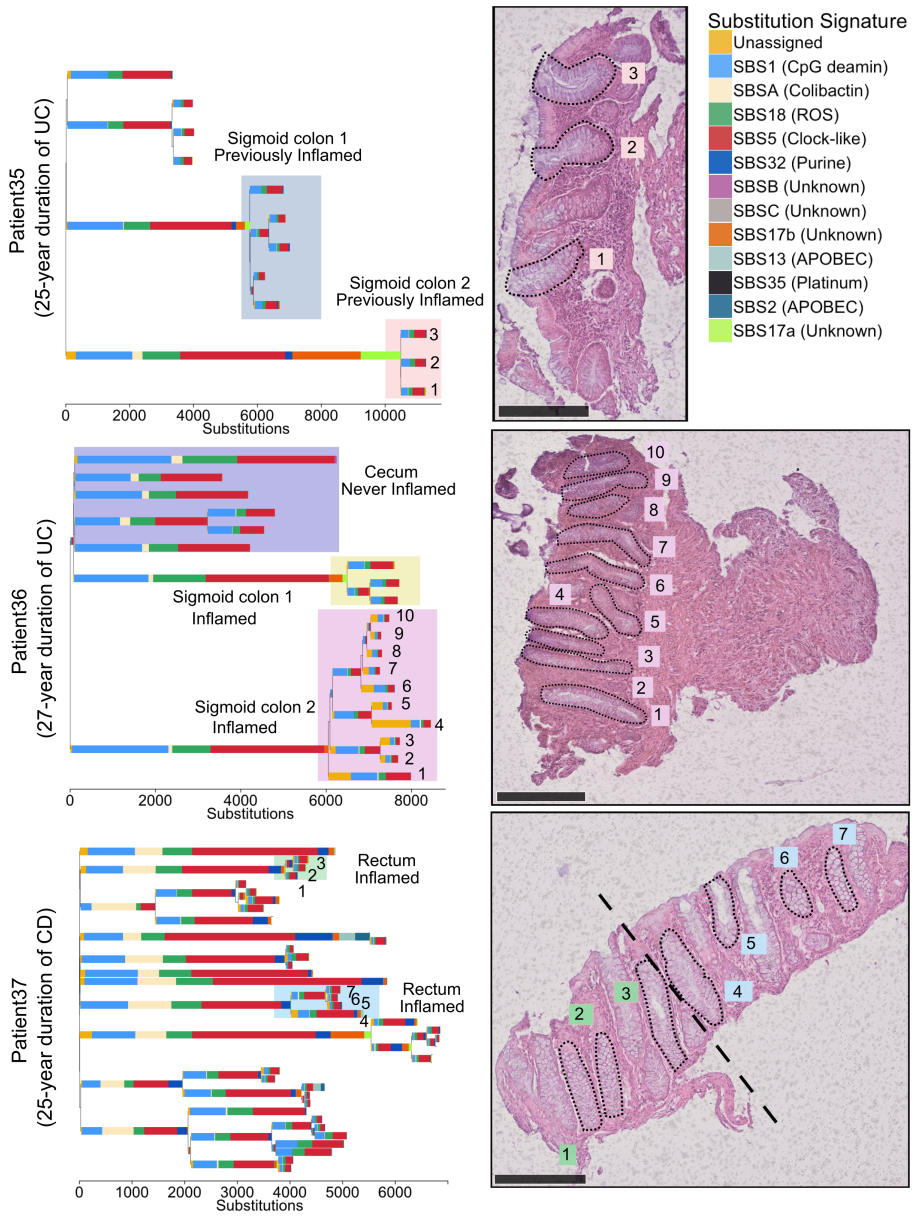



Fig. 3.12 **Examples of clonal expansions in three IBD patients.** (Top) A phylogenetic tree of crypts sampled from a 66 year old patient with a 25 year history of ulcerative colitis. The accompanying biopsy image shows the crypts from the orange shaded area. The clones highlighted in blue and orange come from the same previously inflamed site and were millimeters apart. A large difference in the mutation burden of these clones is driven by a local activation of signatures 17a and 17b in the orange shaded clone. (Middle) A phylogenetic tree of crypts sampled from a 61 year old patient with a 27 year history of ulcerative colitis. The clones highlighted in purple and yellow come from biopsies taken millimeters apart. The accompanying biopsy image shows the crypts from the purple clone. (Bottom) A phylogenetic tree of crypts sampled from a 37 year old patient with a 25 year history of Crohn's disease affecting the colon. A biopsy overlaps two clones (in blue and green).

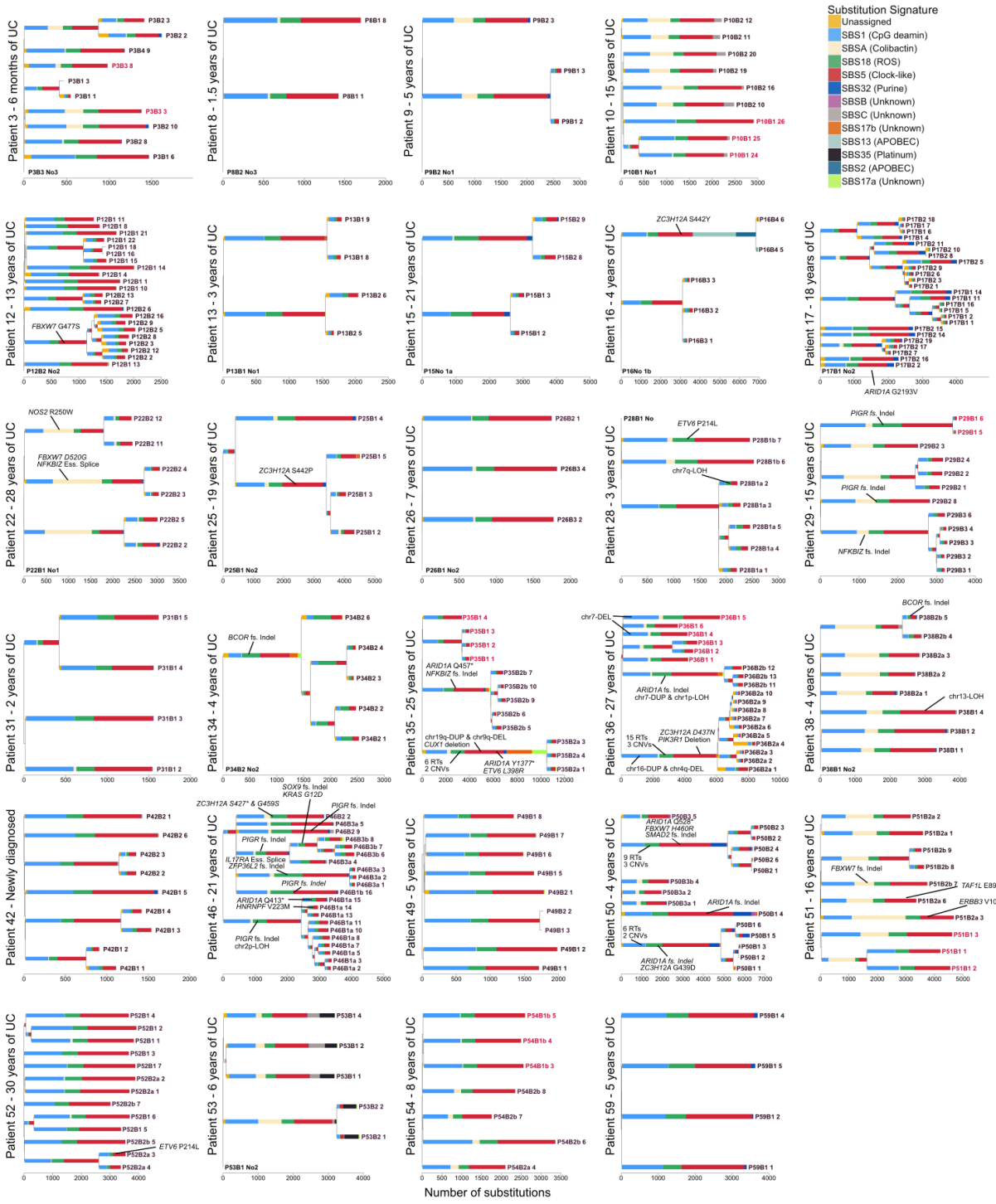



Fig. 3.13 **Phylogenetic trees for all ulcerative colitis patients.** Mutational signatures are overlaid on the trees and likely driver mutations are mapped to the branch in which they occur. Crypts are labelled on the form *PXBY_Z* where PX is the patient number, BY the biopsy number (with a,b and c denoting biopsies taken a few millimeters apart from the same site) and Z is the crypt number. The colour of the labels indicates whether a crypt comes from an inflamed, previously inflamed or never inflamed site.

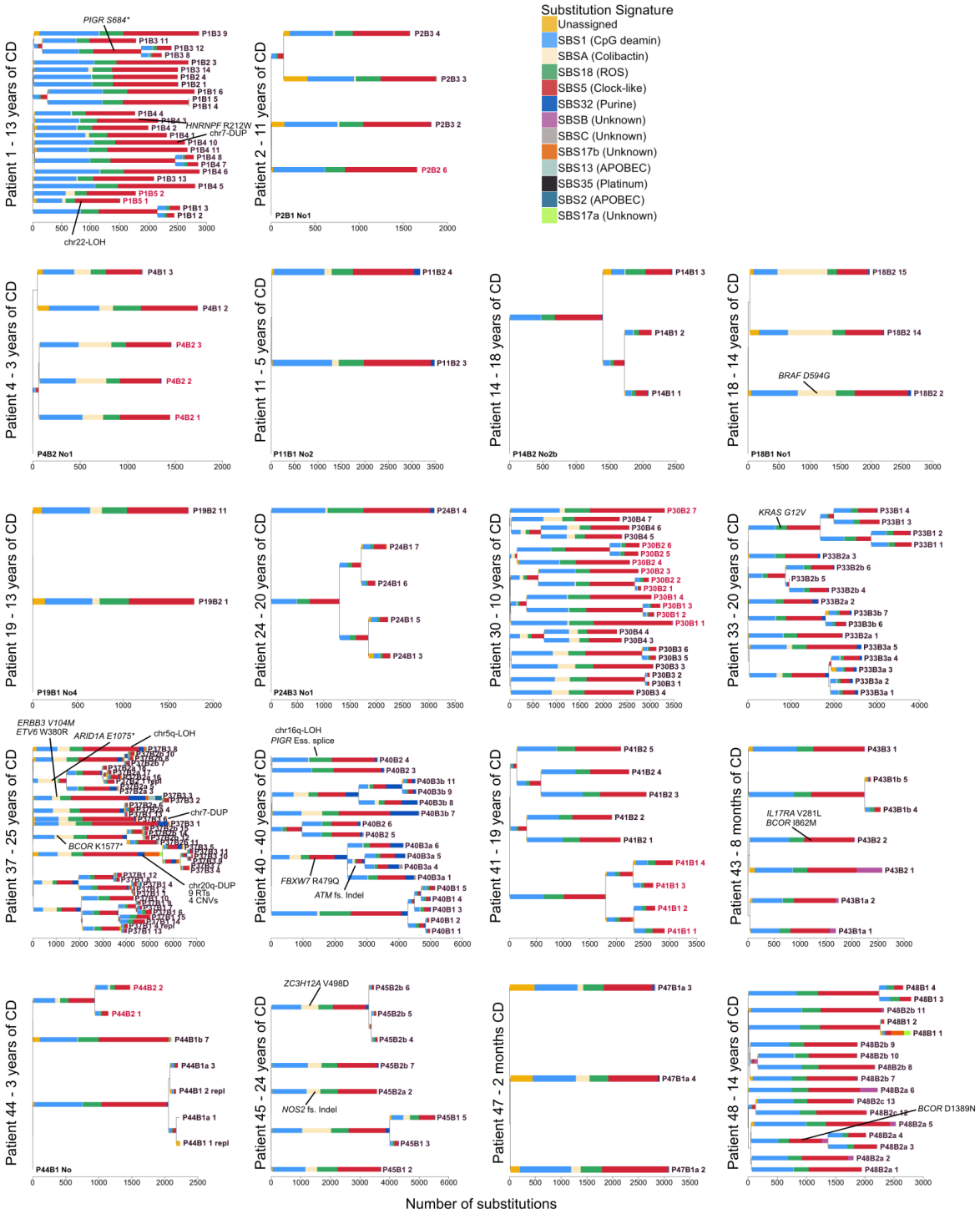



Fig. 3.14 **Phylogenetic trees for all Crohn's disease patients.** Mutational signatures are overlaid on the trees and likely driver mutations are mapped to the branch in which they occur. Crypts are labelled on the form *PXBY_Z* where PX is the patient number, BY the biopsy number (with a,b and c denoting biopsies taken a few millimeters apart from the same site) and Z is the crypt number. The colour of the labels indicates whether a crypt comes from an inflamed, previously inflamed or never inflamed site.

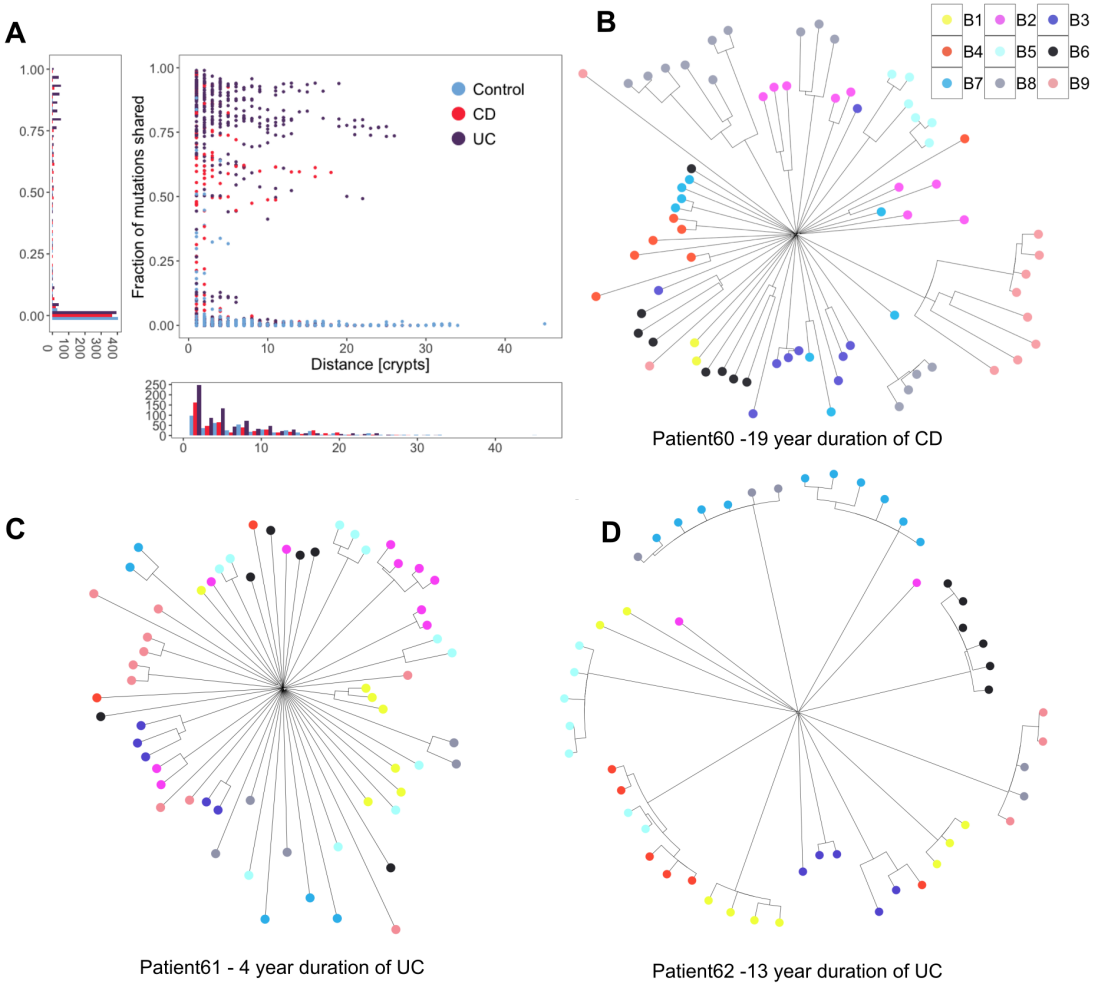



Fig. 3.15 **Clonal structure of the IBD colon.** a) For pairs of crypts from the same biopsy, the figure shows the number of mutations that are shared between a pair as a fraction of the average mutation burden of the two crypts and this is plotted as a function of the distance between the pair. b) A phylogenetic tree showing crypts sampled from 9 biopsies from the sigmoid colon of a 36 year old male diagnosed with CD 19 years prior to sampling. c) A phylogenetic tree showing crypts sampled from 9 biopsies from the rectum of a 71 year old male diagnosed with UC 4 years prior to sampling. d) A phylogenetic tree showing crypts sampled from 9 biopsies from the rectum of a 42 year old female diagnosed with UC 13 years prior to sampling.

Table 3.3 Mutations occurring in canonical cancer hotspots of genes that don't show a significant enrichment of mutations in the IBD mucosa.

| Branch-ID | Gene | Chr | Pos | Ref | Alt | aaChange |
|---|---|---|---|---|---|---|
| Patient18_2 | *BRAF* | 7 | 140453154 | T | C | D594G |
| Patient62_93 | *ERBB2* | 17 | 37879658 | G | A | R678Q |
| Patient37_70 | *ERBB3* | 12 | 56478854 | G | A | V104M |
| Patient51_4 | *ERBB3* | 12 | 56478854 | G | A | V104M |
| Patient33_31 | *KRAS* | 12 | 25398284 | G | A | G12V |
| Patient46_40 | *KRAS* | 12 | 25398284 | G | A | G12V |
| Patient61_33 | *TP53* | 17 | 7577548 | C | T | G245S |
| Patient61_45 | *TP53* | 17 | 7577120 | C | T | R273H |

### 3.4.5 Distinct patterns of selection in IBD compared with normal epithelium

The recurrent cycles of inflammation and remission which characterise IBD could create an environment in which clones containing advantageous mutations may selectively spread in the mucosa. This advantage may manifest either through faster cell division and elevated crypt fission rate or through increased resistance to the cytotoxic effects of inflammation. To identify mutations which likely confer selective advantage on the cell, I searched for mutations occurring in canonical mutation hotspots from the Cancer Genome Atlas. This revealed a total of 10 missense mutations in *KRAS*, *BRAF*, *TP53*, *ERBB2*, *ERBB3* and *FBXW7* occurring at canonical hotspots (Table 3.3). Additionally, I found a heterozygous nonsense mutation in *APC* and frameshift indels in known colorectal tumour suppressors; *ATM*, *SOX9*, *RNF43*, *SMAD2*, *TAF1L* and *ZFP36L2*, of likely driver status (Table 3.4). Furthermore, two large-scale deletions in the dataset overlap known tumour suppressors, *PIK3R1* and *CUX1*, and are likely drivers (Figure 3.16). These mutations, hereafter referred to as putative drivers, are mapped to the phylogenetic trees in figures 3.13 and 3.14.

The number of putative cancer drivers found in a crypt is associated with increased burden of both substitutions (269 substitutions per driver, 90-447 95% CI, $P = 5.6 \times 10^{-3}$) and indels (40 indels per driver, 20-60 95% CI, $P = 1.7 \times 10^{-4}$), as well as with each of the replication-related signatures (SBS1, SBS5, SBS18, ID1 and ID2, Table 3.5). There was also a significant association with the purine signature (SBS32). I estimated the burden of purine signature to be increased by 30 (14-47, 95% CI, $P = 3.7 \times 10^{-4}$, Figure 3.7) substitutions per driver, suggesting that rapidly dividing cells may be particularly susceptible to the mutagenic effect of purine treatment.

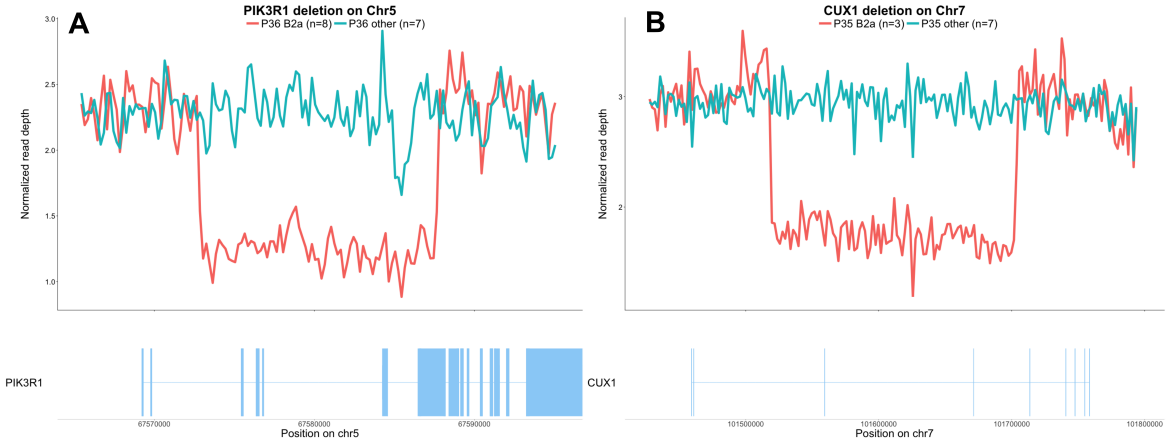



Fig. 3.16 **Structural variants of probable driver status.** The figure compares normalized read depths of crypts called as carriers and non carriers. A) A deletion covering five exons of *PIK3R1* found to precede a clonal expansion in biopsy 2a of patient 36 (Figure 3 of the main text, middle panel, purple clone). B) A deletion covering three exons of *CUX1* and found to precede a clonal expansion in biopsy 2a of patient 35.

Table 3.4 Loss of function mutations in known colorectal tumour suppressors that don't show a significant enrichment of mutations in the IBD mucosa.

| Branch-ID | Gene | Chr | Pos | Ref | Alt | aaChange |
|---|---|---|---|---|---|---|
| Patient37_5 | *ATM* | 11 | 108165721 | AG | A | fs-Indel |
| Patient40_28 | *ATM* | 11 | 108214044 | TA | T | fs-Indel |
| Patient37_78 | *RNF43* | 17 | 56435947 | GC | C | fs-Indel |
| Patient60_39 | *RNF43* | 17 | 56435982 | A | AGGGCCCAT | fs-Indel |
| Patient46_37 | *ZFP36L2* | 2 | 43452828 | G | GCGTCC | fs-Indel |
| Patient61_21 | *ZFP36L2* | 2 | 43452186 | T | TG | fs-Indel |
| Patient50_26 | *SOX9* | 17 | 70117717 | A | AG | fs-Indel |
| Patient61_23 | *SOX9* | 17 | 70117841 | GCACGTCAA | GGGACGT | fs-Indel |
| Patient51_3 | *TAF1L* | 9 | 32632904 | C | A | E892* |
| Patient50_26 | *SMAD2* | 18 | 45374928 | TG | T | fs-Indel |
| Patient60_6 | *APC* | 5 | 112174249 | T | A | Y986* |

Table 3.5 Association between the number of putative drivers found in the crypts and the mutation burden. P-values are calculated with a likelihood ratio test of models with and without the driver count variable.

| Response variable | Effect per driver | 95% CI | P-value |
|---|---|---|---|
| Total substitution burden | 269 | (90-447) | 0.0056 |
| SBS1 | 107 | (51-164) | 0.00038 |
| SBS5 | 149 | (77-221) | 0.00014 |
| SBS18 | 56 | (20-93) | 0.0047 |
| SBSA | -7 | (-47-33) | 0.73 |
| SBS32 | 30 | (14-47) | $3.7 \times 10^{-4}$ |
| | | | |
| Total Indel burden | 40 | (20-60) | 0.00017 |
| ID1 | 33 | (21-45) | $3.7 \times 10^{-7}$ |
| ID2 | 0.81 | (-2.2-3.8) | 0.6 |
| IDA | -1 | (-7-5) | 0.75 |
| ID14 | 0.03 | (-0.23-0.3) | 0.8 |

To search for genes under positive selection, I used the dNdScv software, described in section 1.4.5. dNdScv estimates the ratio of non-synonymous to synonymous mutations (dN/dS) across all genes while correcting for regional and context-dependent variation in mutation rates (Martincorena et al., 2017). Genes with dN/dS ratios significantly different from 1 are considered to be under selective pressure. This analysis revealed four genes, *ARID1A*, *FBXW7*, *PIGR* and *ZC3H12A*, to be under significant positive selection in the IBD colon after Benjamini-Hochberg correction for multiple testing (Figure 3.17 and 3.18).

*ARID1A* and *FBXW7* are well-established tumor suppressors and are found mutated at similar frequencies in sporadic- and colitis-associated colorectal cancers (Baker et al., 2018; Martincorena et al., 2017). I included mutations in *ARID1A* and *FBXW7* in the regression of driver count against mutation burden described above.

In several instances, distinct heterozygous mutations in the same gene were found in different crypts from the same patient (Figures 3.13 and 3.14). For example, in one patient suffering from pan-colitis I found a different *PIGR* mutation in four biopsies from the right, transverse and left side of the colon (Figure 3.18B). As I generally sampled rather few crypts per patient however, these few examples of parallel evolution are difficult to interpret.

I did not detect a significant signal of selection of mutation in the two genes, *AXIN2* or *STAG2*, reported to be under positive selection in the normal colon (Lee-Six et al., 2019) (P

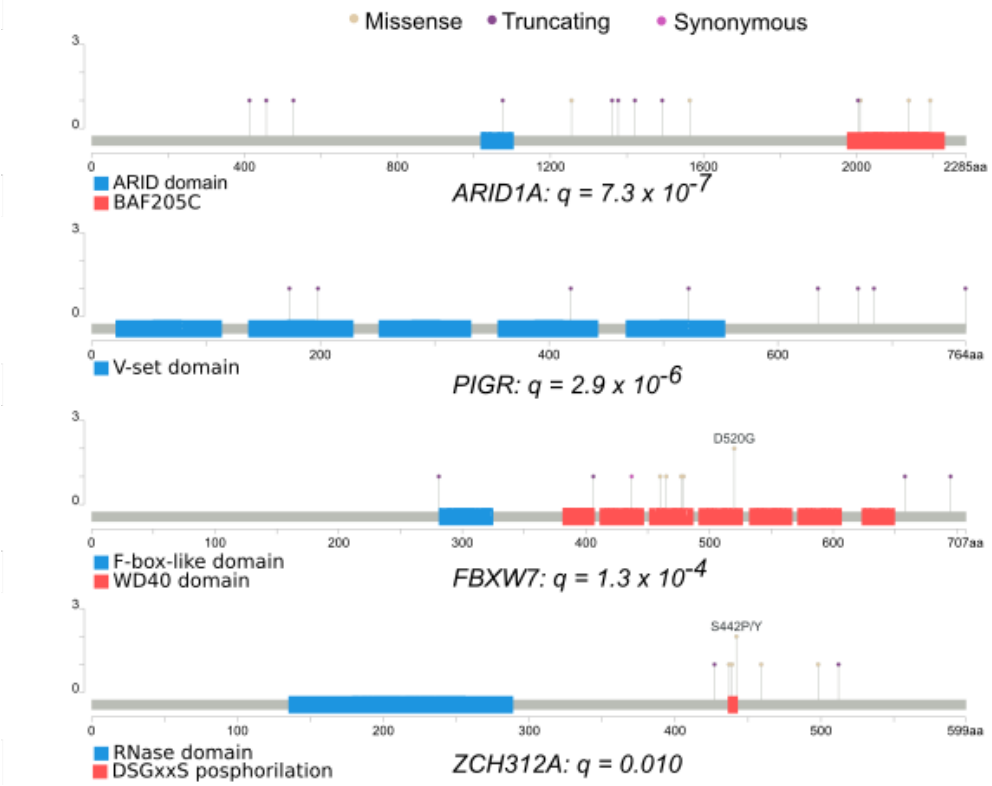



Fig. 3.17 **Mutations under positive selection** The plot shows the location of mutations found in genes that are enriched for non-synonymous coding mutations.

= 0.98 and 0.74, respectively) nor was there any evidence of selection of *PIGR* or *ZC3H12A* mutants in the normal colon. I did not find a significant difference in the mutation burden of any of these genes between UC and CD, although it should be noted that the power of the Poisson analysis is lower when comparing two datasets than when one dataset is compared against the expected distribution.

Recurrent mutations in *PIGR* and *ZC3H12A* are of particular interest since these have not been described in cancer but have roles in immunoregulation and reflect distinct mechanisms of positive selection in the IBD colon. *PIGR* encodes the poly-immunoglobulin (Ig) receptor, which transfers polymeric immunoglobulins produced by plasma cells in the mucosal wall across the epithelium to be secreted into the intestinal lumen (Johansen and Kaetzel, 2011). *Pigr* knock-out mice exhibit decreased epithelial barrier integrity and increased susceptibility to mucosal infections and penetration of commensal bacteria into tissues (Johansen et al., 1999). *ZC3H12A* encodes an RNAse, Regnase-1 (also known as MCPIP1). It is activated in response to TLR stimulation and degrades mRNA of many downstream immune signaling genes (Matsushita et al., 2009), including *PIGR* (Nakatsuka et al., 2018), *NFKBIZ* (Mino et al., 2015) and members of the IL17 pathway (Garg et al., 2015). Four of the mutations in *ZC3H12A* occur in a DSGxxS motif which when phosphorylated marks the protein for ubiquitin-mediated degradation. Mutations of the corresponding residues in mice attenuate the phosphorylation (Iwasaki et al., 2011) and stabilize the protein so these are likely gain of function.

I next carried out a pathway-level dN/dS analysis, searching for enrichment of missense and truncating variants across 15 gene sets that were defined a priori because of their relevance in either colorectal carcinogenesis or IBD pathology (Figure 3.18C, Supplementary Tables 12, 13 and 14, Methods). There was a 6.5-fold (1.8 - 23.6, 95% CI) enrichment of truncating mutations in genes associated with colorectal cancer (q=0.011) as well as a 1.9-fold (1.3-2.8, 95% CI) enrichment in genes significant in a pan-cancer analysis of selection (Priestley et al., 2019) (q=0.011). Interestingly, the pathway-level dNdS also revealed a 4.0-fold (1.7-9.4, 95% CI) enrichment of truncating mutations in the interleukin-17 (IL17) signaling pathway (q=0.011) and a 3.3-fold (1.6-6.7, 95% CI) enrichment in Toll-like receptor (TLR) cascades (q=0.011) with mutations from both UC and CD derived crypts contributing to the enrichment (Figure 3.19).

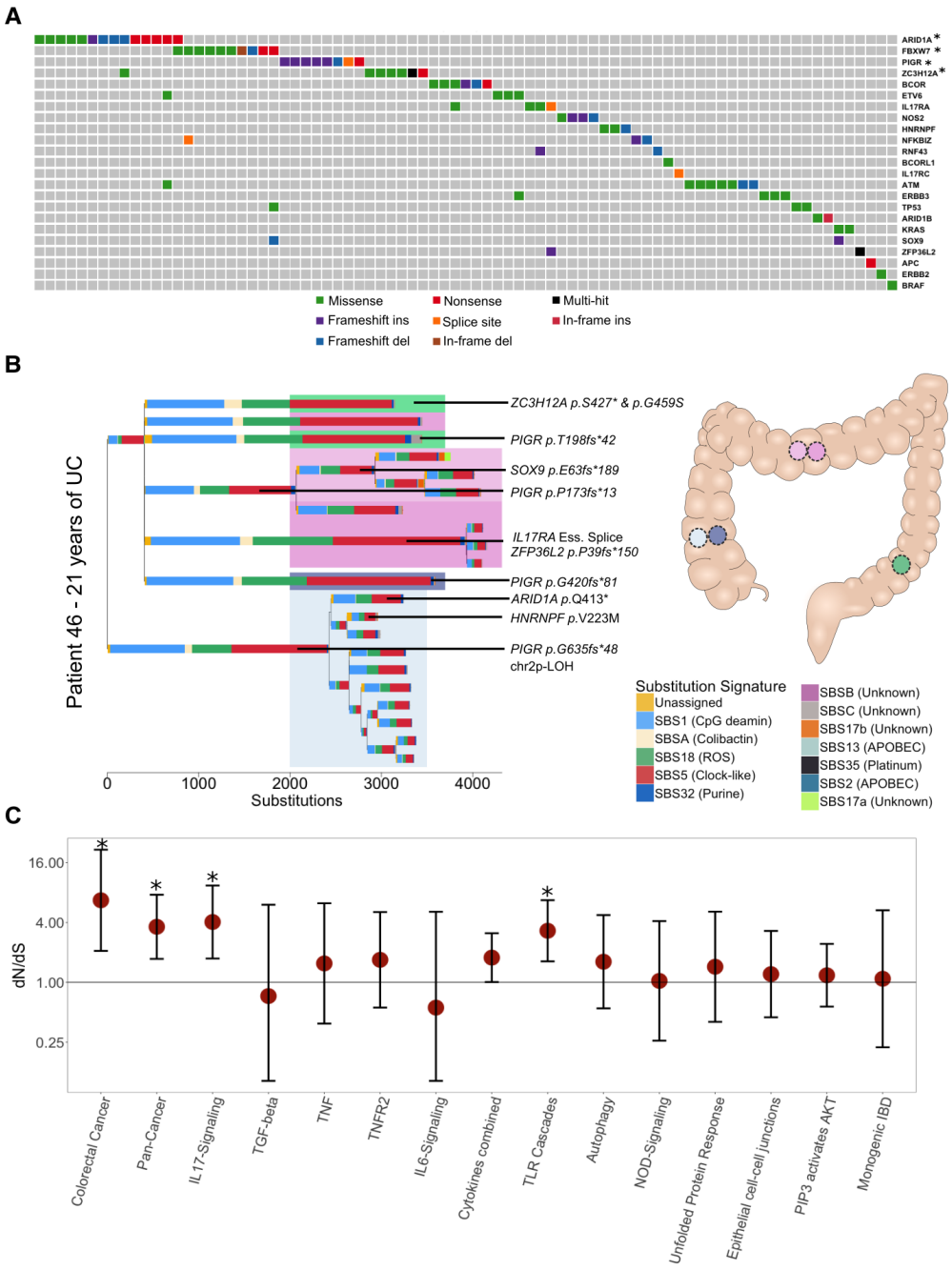



Fig. 3.18 **Driver mutations and positive selection in IBD.** A) An oncoplot showing the distribution of potential driver mutations mapped to branches of phylogenetic trees. Each column represents a branch of a phylogenetic tree and a mutation may be found in multiple crypts if the branch precedes a clonal expansion. Branches without potential drivers are not shown for simplicity. *Genes significantly enriched in non-synonymous coding mutations. B) A phylogenetic tree of the crypts dissected from a 38 year old male suffering from UC for 21 years. Crypts are dissected from five biopsies from three previously inflamed sites of the colon. Crypts carrying distinct *PIGR* truncating mutations are found in four of the biopsies and in all three colonic sites. C) Pathway-level dN/dS ratios for truncating mutations in known cancer genes and cellular pathways important in IBD pathogenesis. Error bars represent 95% confidence intervals. *Significant enrichment of mutations after Benjamini-Hochberg correction for multiple testing ($q<0.05$).

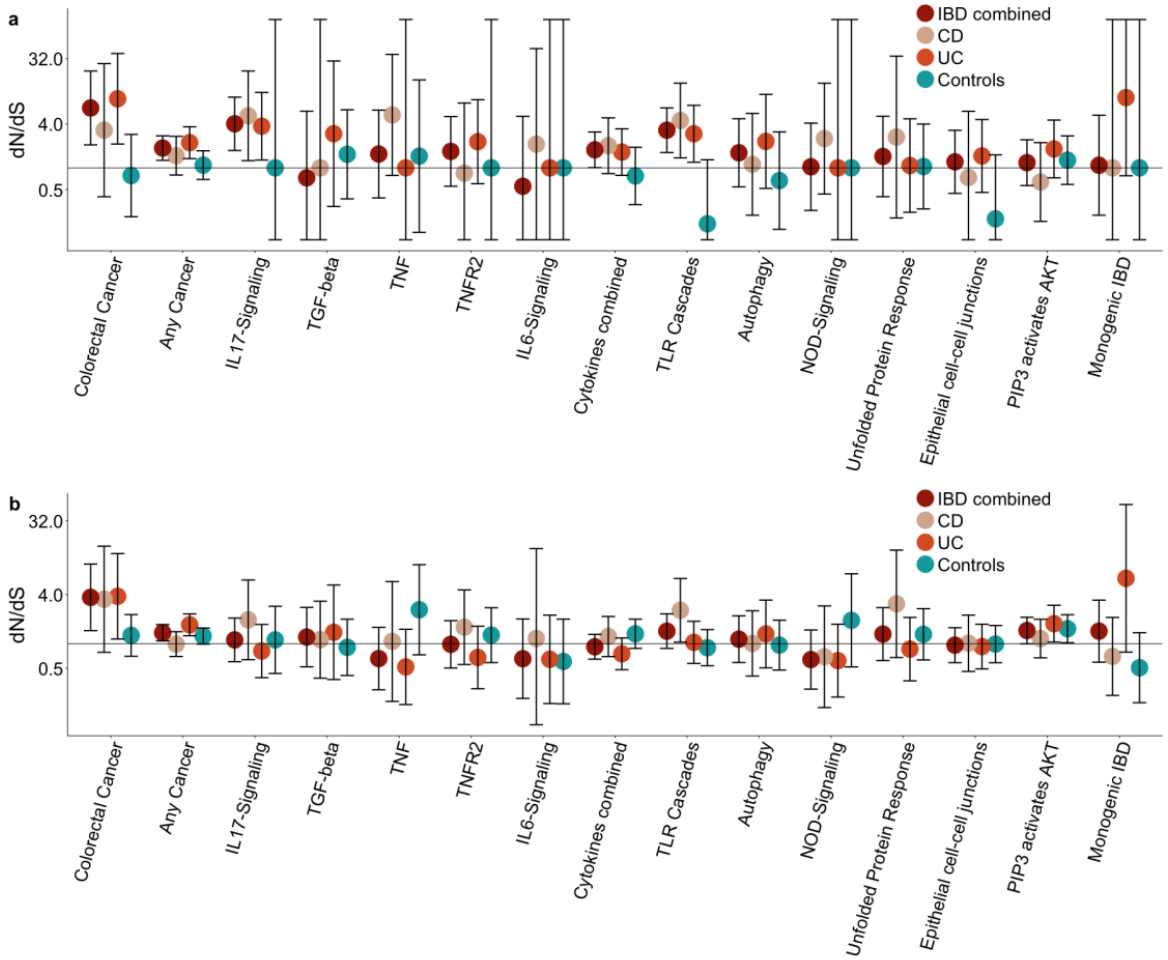



Fig. 3.19 **Pathway-level dN/dS ratios for mutations in known cancer genes and cellular pathways important in IBD pathogenesis.** a) Pathway-level dN/dS for truncating mutations. Same as Figure 3.18C but also showing the ratios when analysis is restricted to CD, UC or control crypts. b) Pathway-level dN/dS for missense mutations.

## 3.5 Discussion

I have used whole genome sequencing of individual colonic crypts to provide the most accurate characterization of the somatic mutation landscape of the IBD affected colon to date. The results suggest that somatic mutagenesis of the mucosa is accelerated 2.4-fold in disease and that this increase is mostly driven by acceleration of common mutational processes which are associated with cell division and metabolic stress and are ubiquitous in IBD-unaffected colon. Metabolic stress also results in an increased burden of somatic structural variants, which nevertheless remain rare in the IBD-affected mucosa. Structural variants are common in colorectal cancers and thus rapid increase in structural variation may be a hallmark of neoplastic transition, in line with previous reports (Baker et al., 2018). Increase in structural variation from healthy tissue to non-neoplastic disease has also been observed in liver disease (Brunner et al., 2019).

Colitis-associated colorectal cancers commonly arise from a background of large clonal fields (Choi et al., 2017). In this sample of non-dysplastic tissue I find millimeter scale clonal expansions, although I note that for many inflamed regions only a single small biopsy is available which limits my ability to detect large clones. *TP53* and *KRAS* mutations are thought to be key events in clonal spread in the IBD mucosa but while I do observe a number of canonical cancer driver mutations in genes including *TP53* and *KRAS*, only *ARID1A* and *FBXW7* show significant evidence of positive selection.

While there is substantial overlap in the driver landscape of IBD and non-IBD colon, important differences also exist. The findings of enrichment of mutations in *PIGR*, *ZC3H12A* and in the IL17 and TLR pathways suggest there are distinct selection mechanisms in the colitis-affected colon and that somatic mutations potentially play a causal role in the pathogenesis of IBD. While this work was under peer review proceeding publication, two studies of somatic mutations in UC patients from the Japanese population were published which confirm the positive selection of mutations in *ARID1A*, *FBXW7*, *PIGR*, *ZC3H12A* and in the IL17-pathway in UC mucosa (Kakiuchi et al., 2020; Nanki et al., 2020). Importantly, this study shows that the same selective pressures are operative in mucosal tissue in both ulcerative colitis and Crohn's disease.

The two papers also report mutations in additional genes including *NFKBIZ*, *IL17RA*, *TRAF3IP2* and *NOS2*. I performed restricted-hypothesis testing of a set of 13 genes reported in these other two papers and replicated six at $q<0.05$ (Table 3.6).

Table 3.6 Restricted hypothesis testing of genes reported to be under positive selection in the UC mucosa in Kakiuchi et al or Nanki et al. $n_{syn}$, $n_{mis}$, $n_{non}$, $n_{spl}$, $n_{ind}$: Number of mutations annotated as synonymous, missense, nonsense, splice site and indels, respectively. $q_{rht}$: Restricted-hypothesis testing q-value (after Benjamini-Hochberg correction of P-values for 13 tests).

| Gene | $n_{syn}$ | $n_{mis}$ | $n_{non}$ | $n_{spl}$ | $n_{ind}$ | $q_{rht}$ |
|---|---|---|---|---|---|---|
| *NOS2* | 0 | 1 | 0 | 0 | 3 | 0.0069 |
| *NFKBIZ* | 1 | 0 | 0 | 1 | 2 | 0.0069 |
| *BCOR* | 1 | 3 | 1 | 0 | 2 | 0.0077 |
| *RNF43* | 0 | 0 | 0 | 0 | 2 | 0.0200 |
| *HNRNPF* | 0 | 2 | 0 | 0 | 1 | 0.0210 |
| *ETV6* | 0 | 4 | 0 | 0 | 0 | 0.0407 |
| *IL17RA* | 1 | 3 | 0 | 1 | 0 | 0.0958 |
| *KRAS* | 0 | 2 | 0 | 0 | 0 | 0.0958 |
| *TP53* | 0 | 3 | 0 | 0 | 0 | 0.160 |
| *ARID1B* | 0 | 1 | 0 | 0 | 1 | 0.259 |
| *IL17RC* | 0 | 0 | 0 | 1 | 0 | 0.261 |
| *TRAF3IP2* | 0 | 0 | 0 | 0 | 0 | 0.9845 |
| *BCORL1* | 1 | 1 | 0 | 0 | 0 | 0.9845 |

Importantly, the enrichments of truncating mutations observed in the IL17 and TLR pathways, which share many genes in common, are not driven by the genes discussed above because *PIGR*, *ZC3H12A*, *NFKBIZ* and *NOS2* are not part of these pathways (according to Reactome), and no mutations were found in *TRAF3IP2*. This suggests that additional positively selected genes related to IL17 and TLR signaling may be discovered in the IBD colon as sample size is increased. The difference in the number of *NFKBIZ* mutant crypts between the studies is noticeable. Only 3 truncating mutations in *NFKBIZ*, which is the most commonly mutated gene in Kakiuchi et al (Kakiuchi et al., 2020), were found in my dataset. This is reminiscent of a previous report of different selection of *NOTCH2* mutants in normal skin between individuals of European and South Asian ancestry (Martincorena et al., 2015). Together with the observation that distinct mutations in the same gene are often found in crypts from the same individual, this leads us to speculate that differences in local environment or a person's genetic background affects the strength of selective advantage posed by somatic variants and studies with larger sample sizes may be able to detect those interactions.

In their study, Nanki et al show how IL17A may be cytotoxic to epithelial cells and argue that clones carrying IL17 pathway mutations are able to avert this cytotoxicity and thereby

selectively expand in the inflamed environment (Nanki et al., 2020). This has implications for the direction of effect of these mutations on IBD pathogenesis since selective pressure would only be asserted following disease onset as Th17 cells infiltrate the tissue and secrete IL17A in the vicinity of the epithelium. However, it could also be hypothesized that these mutations play a causal role in the pathogenesis of IBD through an effect on dysbiosis. Indeed, the discovery by Nanki et at (Nanki et al., 2020) that *PIGR* mutations do not confer upon cells survival advantage in the presence of IL17A may add weight to this hypothesis. While *ZC3H12A* and *NFKBIZ* are involved in IL17 signaling, both are also induced downstream of TLRs (Matsushita et al., 2009; Yamamoto et al., 2004) where they regulate the transcriptional changes that follow TLR signaling. Disruption of the IL17 pathway itself may also play a causal role in the disease as intestinal epithelial-cell specific knock-out of components of the IL17 pathway in mice results in commensal dysbiosis through down-regulation of *Pigr* and other genes (Kumar et al., 2016). Thus, a positive feedback loop may be established, leading to ever greater spread of a pathogenic clone. It is worth noting that clinical trials of anti-IL17A and anti-IL17RA antibodies for the treatment of Crohn's disease have been carried out but either show no efficacy over placebo or worsen the disease (Hueber et al., 2012; Targan et al., 2016).

Our understanding of somatic evolution in normal tissues has improved greatly over the last few years but how and if somatic evolution contributes to the pathogenesis of complex traits other than cancer remains poorly understood. Clonal hematopoiesis has been associated with coronary heart disease (Jaiswal et al., 2014) and our work suggests that somatic evolution in the colonic mucosa may initiate, maintain or perpetuate IBD. Large scale analyses of cancers have started to reveal common themes of cancer evolution across tissues (Gerstung et al., 2020) and extending this work to other tissues exposed to chronic inflammation may similarly reveal patterns of remodeling of the selection landscapes associated with disease, but which need not drive neoplastic growth. Comparing the evolutionary forces in the IBD mucosa with those operating in psoriasis, celiac disease, asthma and other diseases affecting epithelial cells is an area of special interest.

The cohort in this study is small and sampling was biased in the sense that more crypts were often dissected from patients with long-standing disease. It would be interesting to expand the study of the driver mutations in particular to a larger cohort. This would enable us to associate the presence of particular drivers with clinical variables. A disadvantage of the LCM method is that it is time consuming and doesn't scale particularly well to hundreds or thousands of samples. Kakiuchi et al used a clever method for isolating clusters of crypts together. They applied an adhesive to the crypts and then loosened the epithelial cells from

the underlying tissue by treating the tissue with EDTA. Groups of crypts would stick to the adhesive and could be sequenced together. This is a promising method for scaling up the sample size, especially when many crypts in the region belong to the same clone.

# Chapter 4

# Somatic evolution in normal and psoriatic human skin

## 4.1 Chapter introduction

The work in this chapter remains unpublished. I conceived of the project, applied to the faculty of Human Gentics for funding for the project, processed all samples as described in the introduction to chapter 3 and performed out all statistical and bioinformatic analyses.

### 4.1.1 Psoriasis

Excited by the differences we observed in the somatic evolution landscape of the IBD-affected colon compared with normal colon, I was interested in studying somatic evolution in psoriasis. This is a second chronic inflammatory disease affecting an epithelial tissue, this time the skin. Psoriasis is the most common autoimmune disease in the Western world, affecting about 2-3% of adults of white European descent (Parisi et al., 2020). It has several subtypes, the most common of which is psoriasis vulgaris, which accounts for about 90% of cases and manifests as well-defined plaques of thickened skin with an overlying silvery scale, most often on the knees, elbows and scalp (Greb et al., 2016; Griffiths and Barker, 2007). All participants in the study described herein were diagnosed with this most common type of psoriasis and I will hereafter refer to psoriasis vulgaris simply as psoriasis.

The causes of psoriasis are not fully known. Like IBD, it occurs in cycles of flares and remission where the same anatomical location tends to be recurrently affected, suggesting some permanent alteration of the affected region. Many risk factors, both environmental and genetic have been identified and associated with both disease onset and severity. The genetic

risk factor which mediates the largest risk is HLA-Cw*0602, which confers an odds-ratio of 3-3.5 for developing the disease (Okada et al., 2014) and has been linked with earlier onset and more severe disease (Gudjonsson et al., 2003, 2006). Close to 70 additional loci with smaller effect sizes have been identified through GWAS (Tsoi et al., 2017). As for IBD, some genes point towards T-cell biology and immune signaling while others, for example *SERPINB8*, *KLF4*, *KLF13* and *TP63*, are thought to affect epidermal differentiation (Tsoi et al., 2017). Lifestyle and environmental factors associated with the disease most notably include obesity, infections and trauma to the skin, such as burns or cuts (called Köbner phenomenon) (Greb et al., 2016; Griffiths and Barker, 2007).

### 4.1.2 Cellular structure of the epidermis

The two main structures of the skin are the dermis and the epidermis. The lower dermis consists mainly of fibroblasts and immune cells like macrophages and mast cells, as well as matrix components like collagen, elastin and various extrafibrillar matrix components. The upper epidermis, which will be the focus of this chapter, consists mainly of keratinocytes derived from a population of stem cells that reside at the bottom of the epidermis along a basement membrane that separates the epidermis from the dermis. As cells differentiate they stratify vertically up to the outer layers of the epidermis. They first enter the granule layer, so named because in this layer cells start to lose their nuclei and cytoplasmic organelles and a large number of granules appear under the microscope. Finally, the outermost layer of the skin is the cornified layer which consists of several layers of dead cells that have lost their nuclei and are ultimately shed.

The epidermis is punctuated by hair follicles and sweat ducts rising from the dermis below. Both hair follicles and sweat ducts contain their own stem cells and form distinct proliferative compartments that do not contribute to the normal maintenance of the epidermis, although lineage tracing experiments suggest that they can acquire an epidermal phenotype and be called upon for wound healing (Ito et al., 2005; Lu et al., 2012).

Histologically, psoriatic skin has a distinct appearance characterized by epidermal hyperplasia, which drives thickening of the epidermis and elongation of epidermal rete into the dermis below. The differentiation of keratinocytes is also altered such that the granular layer is largely absent and the cornified layer is, as a result, formed from incompletely differentiated keratinocytes, some of which retain their nuclei. This is known as parakeratosis. The scales observed at psoriatic lesions result from the failure of the differentiated keratinocytes to stack and adhere to one another (Lowes et al., 2007). Psoriatic skin is also characterized

by extensive immune cell infiltration, especially into the dermis but also to a lesser extent into the epidermis (Boehncke and Schön, 2015; Greb et al., 2016).

### 4.1.3 Keratinocyte cancers

Keratinocyte cancers are the most common malignant neoplasms affecting humans. They are especially common among fair-skinned individuals who are exposed to large doses of UV-light. Broadly, two subtypes of keratinocyte cancers exist. Basal cell carcinoma (BCC), which originates either in basal stem cells in the interfollicular epidermis or within hair follicles, (Grachtchouk et al., 2011; Peterson et al., 2015) represents three-quarters of all cases. BCC is generally thought to have limited metastatic potential and although it represents a large fraction of skin cancers, it results in a much lower fraction of deaths (Verkouteren et al., 2017). BCCs are characterized by one of the highest mutation burdens seen in any cancer, with mutation burdens estimated at 65-76 mutations per megabase, of which over 75% are the result of UV exposure (Bonilla et al., 2016; Jayaraman et al., 2014). Uncontrolled Hedgehog signalling, usually driven by loss of *PTCH1*, activation of *SMO* or loss-of-function mutations in *SUFU*, is a well established driving mechanism of these cancers (Bonilla et al., 2016; Jayaraman et al., 2014; Peterson et al., 2015).

The second class of keratinocyte cancers are cutaneous squamous cell carcinomas (cSCC), which arise within the more superficially placed squamous cell layers of the epidermis. Like BCCs, cSCCs have very high mutation burdens, 50-61 mutations per megabase, with the majority of mutations resulting from UV exposure (Inman et al., 2018; Pickering et al., 2014). Mutations in *NOTCH1*, *NOTCH2*, *FAT1*, *TP53*, *HRAS*, *CDKN2A*, *PIK3CA* and other genes are common driver events, as are copy number changes affecting *NOTCH1* (Agrawal et al., 2011; Inman et al., 2018; Pickering et al., 2014; South et al., 2014). cSCCs most often occur in the head and neck region, presumably because this is the area most exposed to the mutagenic effects of UV-light. They are histologically and pathologically related to head-and-neck squamous cell carcinomas, with both showing extremely high frequency of mutations affecting the NOTCH signalling pathway (Loganathan et al., 2020).

Most cSCCs respond well to first-line therapy like electrodessication, cryo-surgery or radiotherapy. Only 3-5% of tumours metastasize or recur (Veness, 2007). This is likely due to the lesions being visible from the outset, which leads to earlier detection and treatment, rather than intrinsic tumour factors as such. Once metastasized however, few treatment options are available and survival is poor (Veness, 2007).

Interestingly, immunosuppression carries a major risk for the development of keratinocyte cancers. The incidence of BCC is tenfold in organ transplant recipients compared with the general population and the incidence of cSCC is more than a hundredfold (Euvrard et al., 2003; Harwood et al., 2013). These findings suggest an important role for immune surveillance in skin cancer prevention and that the less extreme immunosuppression achieved in the treatment of autoimmune diseases like psoriasis might also affect skin cancer frequencies. Indeed, many drugs, for example mycophenolate mofetil, cyclophosphamide, cyclosporine A and azathioprine are used to treat both transplant recipients and autoimmune diseases, and some of these drugs are themselves mutagenic. Azathioprine, for example, leaves a distinct mutational signature in cSCCs (Inman et al., 2018) and in the colonic mucosa of IBD patients (as discussed in Chapter 3). This may also explain a part of the increased skin cancer risk associated with the use of these drugs.

### 4.1.4 Psoriasis related cancers

Cancer is one of the serious comorbidities of psoriasis. However, whether cancer risk is related to mechanisms involved in psoriasis itself, such as the chronic inflammation, or is a consequence of the immunosuppressive treatment or other treatments undertaken is not completely clear (Boehncke and Schön, 2015). Using a primary care medical records database of patients in the UK, Chiesa-Fuxench et al estimated risk of cancers in mild and moderate-to-severe psoriasis after correcting for comorbid, well-established risk factors for cancer, such as smoking, drinking and obesity (Chiesa Fuxench et al., 2016). They found that the risk of keratinocyte cancer is modestly increased in psoriasis patients compared with the general population (adjusted hazard ratio 1.12 overall and 1.61 in patients with moderate-to-severe disease).

Chiesa-Fuxench et al also reported that psoriasis patients are at an increased risk of lung cancer and lymphomas. The association with lung cancer however was not apparent when the analysis was restricted to people who had never smoked, indicating that this association may be driven by different smoking habits between cases and controls. The largest increase in cancer risk was seen for cutaneous T-cell lymphoma and was 3.82 overall and 9.25 in patients with moderate-to-severe psoriasis. There was also increased risk of lymphomas if cutaneous T-cell lymphomas were excluded (hazard ratio 1.25 overall) (Chiesa Fuxench et al., 2016). Similar hazard ratios have been reported in other studies (Gelfand et al., 2006; Vaengebjerg et al., 2020).

### 4.1.5 Somatic evolution in normal epidermis

#### Mutations under selection in normal epidermis

The skin was the first normal tissue to be extensively characterized in terms of its somatic mutation landscape. In a landmark study, Martincorena et al performed deep targeted sequencing of 74 cancer genes in epidermis isolated from the eyelids of four donors and showed that over a quarter of normal skin cells harbored at least one driver mutation (Martincorena et al., 2015). More recently, Fowler et al sequenced isolated epidermis from a range of body sites from 35 donors (Fowler et al., 2021). Together, these studies revealed evidence of positive selection of mutations in 14 genes, almost all of which are well recognized drivers of cutaneous SCC. Much less overlap was observed with BCC or melanoma. The prevalence of mutations in *NOTCH1* and *FAT1* matches that seen in cSCCs, suggesting that these genes may appear recurrently mutated in these cancers simply by virtue of their colonization of normal skin rather than them playing a direct role in malignant transformation (Fowler et al., 2021). *NOTCH1* was the most commonly mutated gene in both studies with other mutations in genes involved in NOTCH signalling also being prevalent. These include two of the other members of this gene family, *NOTCH2* and *NOTCH3*, and also regulators of NOTCH signalling like *ARID2* and *AJUBA*. Although negative selection is rarely seen in somatic tissues (Martincorena et al., 2017), Fowler et al found evidence of negative selection of mutations in five genes in the skin. For one of these, *PIK3CA*, there was positive selection of mutations predicted to be activating and negative selection of inactivating mutations.

#### Effect of UV-light on selection

The skin is the largest organ of the body and different areas are subjected to different environmental exposures, most notably different levels of UV-radiation. Fowler et al reported differential selection of *TP53*, *NOTCH1* and *FAT1* across different body sites with mutations in the latter two genes being overrepresented in the leg compared with all other sites. In contrast, mutations in *TP53* and *FAT1* were depleted in the head, the most sun-exposed site of the body. Wei et al explored differences between sun-exposed (dorsal forearm) and non-sun-exposed (buttock) skin in more detail, sequencing 100bp segments of highly mutated genes. They found an enrichment of mutations in segments corresponding to p53 p.227-261 and NOTCH1 p.449-481 in the sun-exposed tissue compared with the non-sun-exposed tissue and an enrichment on the gene level for mutation in *TP53* in the same (Wei et al., 2021). Wei et al also reported that six mutational hotspots were more often mutated in sun-exposed compared with non-exposed tissue. These include TP53 R248W and G245D and NOTCH1

P460L, P460S, S385F and E424K, but the mechanisms driving these enrichments are not known.

### Clonal dynamics in the epidermis

As described above, a single layer of basal stem cells maintains the epidermis, with differentiating cells mostly stratifying vertically from their progenitor through the suprabasal layers until they are ultimately shed. On division, a progenitor can produce two differentiating cells, two progenitor cells or one of each. Clones expand, both through drift and selection, when a cell produces two progenitor cells and takes over the space of another cell that has produced two differentiating cells. Notch signalling induces differentiation via multiple mechanisms (Nowell and Radtke, 2017) and it is likely that LoF mutations in *NOTCH1*, and many other drivers, have the effect of biasing the cell fate towards more progenitor cells being produced.

In epithelial tissues with high rates of drivers, such as oesophagus and skin, clones are initially free to rapidly expand, as the flat architecture of the tissue places little structural limits on spread (as opposed to the organization of the colon into crypts). However, the high density of drivers means that expanding clones soon encounter neighbouring clones that also carry distinct mutations that render their fitness higher than the germline "wildtype". This reduces the comparative advantage of both clones and slows down their spread (Colom et al., 2020; Murai et al., 2018). Remarkably, there is budding evidence that *NOTCH1* mutations in the oesophagus can impair tumour growth and that mutant clones in the normal epithelium often outcompete and eliminate tumours in their early stages (Abby et al., 2021; Colom et al., 2021).

## 4.2 Chapter aims

In this chapter I aim to compare the somatic evolution of keratinocytes in lesional and non-lesional skin from patients with psoriasis vulgaris. I will compare the two in terms of the mutation burden, the clonal structure of the tissue, the mutagen exposure and driver landscape of the tissue. This study may also offer opportunities to expand our understanding of evolution of keratinocytes in general, as it is the first to apply whole-exome sequencing to normal cells, as opposed to deep targeted sequencing of cancer genes.

## 4.3 Methods

### 4.3.1 Human tissue attainment and processing

Round punch biopsies, 4 mm in diameter, from lesional and non-lesional skin were donated by psoriasis patients presenting to the Kiel University Skin Clinic between 2017 and 2019. Biopsies were taken under sterile conditions under local anesthesia and the emerging holes were closed with seam stitching and bandaged. All donors gave informed consent for genetic research of the material and the study was approved by the research ethics committee of Christian-Albrechts University in Kiel (A100/12), the National Health Service (NHS) Research Ethics Committee (Yorkshire & The Humber - South Yorkshire Research Ethics Committee, REC ID 20/YH/0244, IRAS ID 286843) and by the Wellcome Trust Sanger Institute Human Materials and Data Management Committee (approval number 20/0085).

Biopsies were fixed in RNAlater (AM7021, ThermoFisher) upon collection following the manufacturer's instructions. One half of each biopsy was used in this study and one half retained by Dr. Weidinger for use in future projects. Yvette Hooks embedded the biopsies in paraffin, sectioned them and fixed the sections to 4 $\mu$m PEN membrane slides (11600288, Leica). I stained the sections with hematoxylin and eosin and dissected samples from this material using laser capture microdissection microscopy (LMD7000, Leica) (Figure 4.1B). I lyzed the cells using ARCTURUS PicoPure DNA extraction kit (Applied Biosystems) according to the manufacturer's instructions.

The volume of the microbiopsies (referred to as "samples") was determined by adding together the size estimates of the cuts from the LCM software and multiplying this by the thickness of the sections (10 micrometers). I commonly dissected the same histological features from serial sections into the same well to increase the DNA-yield of the samples. The surface area of the samples was determined by measuring the width of the samples along the basal membrane and multiplying this with the section thickness and the number of sections (z-stacks) separating the first and last sections dissected into the same well. When LCM-ing, occasionally, individual dissections do not drop to the bottom of wells but either fall outside a well or get stuck to the side of a well. Those dissections do not contribute any DNA to the sample but are nevertheless part of the volume and surface area estimation. The size estimates in Figure 4.1C and D should therefore be considered as upper bound estimates.

### 4.3.2 Genome sequencing

403 samples from 38 individuals were whole-exome sequenced on Illumina Htp NovaSeq 6000® machines using 150bp, paired end reads and the Human All Exon V5 bait set. The median-median sequencing depth was 66X (Figure 4.1E). Reads were aligned to the human reference genome (build hg38) and PCR duplicates were marked by the Sanger core informatics team.

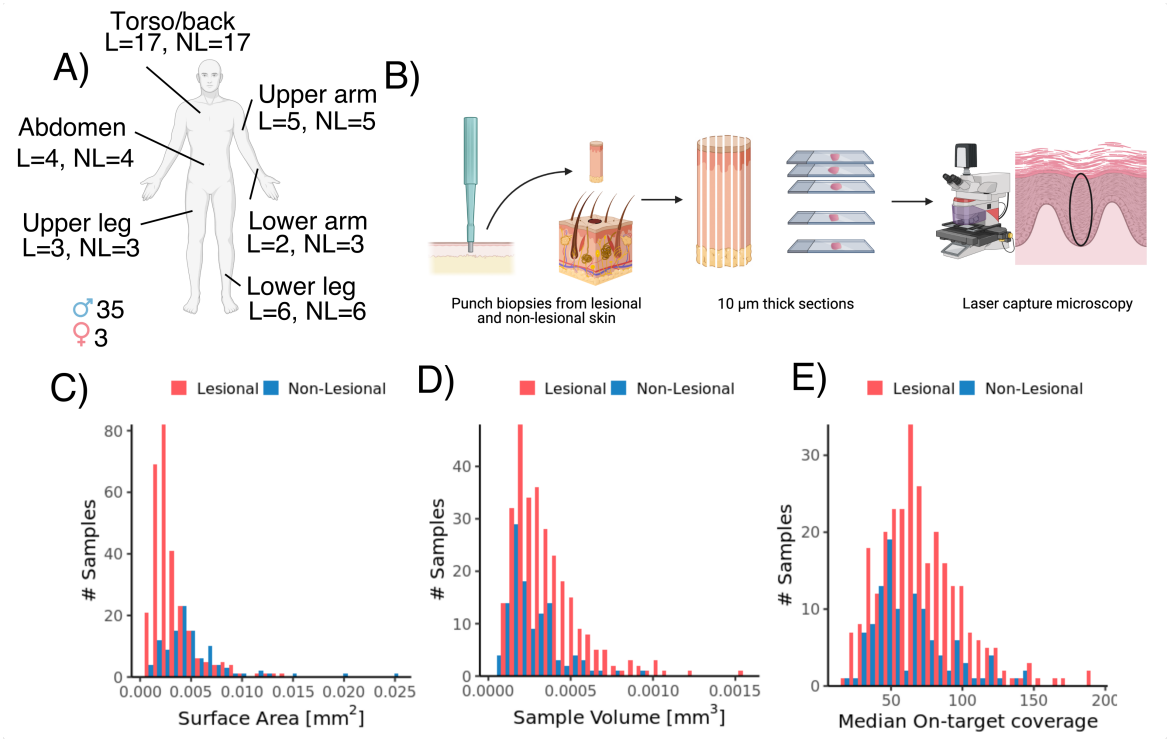



Fig. 4.1 **An overview of the samples of epidermis used in the study.** A) Locations of skin biopsies donated for this study and sex of the donors. L: Lesional; NL: Non-lesional. B) The sample processing pipeline. 4mm punch biopsies were histologically sectioned. I used laser capture microscopy to isolate small parts of epidermis from the histological sections. C) The size distribution in two-dimensions if looking on the sample top-down. D) The size distribution in three-dimensions, after taking into account the thickness of the epidermis. E) Distribution of the median on-target coverage across all samples.

### 4.3.3 Mutation calling and filtering

Substitution calling was performed in much the same way as described for the IBD samples in Chapter 3. Mutations were called by running CaVEMan (Jones et al., 2016), see section 1.4.2, against an unmatched normal with the copy number options set to 10 and 2 for the

major and minor copy numbers, respectively. The samples were compared against a normal panel consisting of 75 unrelated normal samples to remove common SNPs. I also removed mutations if the reads reporting the mutations had a median alignment score lower than 140 or if >50% of the reads were clipped. I did not apply the filters described in Chapter 3 for removing errors associated with the formation of cruciform DNA as I found they were not needed. There are fewer inverted repeats within the exome compared with the whole genome, alignment may be improved to hg38 compared with hg37, and the beta-binomial filters described below do a good job capturing recurrent sequencing errors. Typically, unfiltered "norm-seq" samples would have a signature 8-like abundance of C>A mutations but no trace of this was seen in this exome dataset, indicating that the hairpin filters were redundant. Indels were called using a modified version of the Pindel algorithm (Raine et al., 2015; Ye et al., 2009), described in section 1.4.2.

I next grouped samples by patient and used the bam2R() function of the deepSNV package (Gerstung et al., 2014) to construct read pileups for all sites at which a mutation was called in any sample from that patient. Only reads with a mapping quality of 30 or greater and bases with a base quality of 30 or greater, were counted. To merge adjacent substitutions into double-base-substitutions, I compared the coverage and the number of reads reporting the alternative allele of the two adjacent sites with a Fisher's exact test. I adjusted the P-values for multiple testing on a per-patient basis using the Benjamini-Hochberg method and merged substitutions with q>0.05.

I filtered germline variants not removed by the comparison with the normal panel by applying an exact binomial test of the number of reads reporting each mutation, as described in Chapter 3. Heterozygous germline variants are expected to be present at a VAF of 0.5 in every sample from a patient. For each mutation, I compared the number of reads reporting the reference and alternate alleles across all samples from that patient. I tested the hypothesis that the read counts for the variants were drawn from a binomial distribution with a probability of success of 0.5, or 0.95 for mutations on the sex chromosomes in men. I applied Benjamini-Hochberg correction for multiple testing and excluded mutations with $q > 10^{-3}$. I also used binomial filtering to remove erroneous mutation calls. Recurrent sequencing artefacts will be randomly distributed across samples and can be modelled as being drawn from a binomial distribution. In contrast, true somatic mutations will have a high VAF in some samples whilst being completely absent from others. The latter will be best represented by a beta-binomial with a high overdispersion. For every mutation call, I calculated the maximum likelihood overdispersion parameter ($\rho$) in a grid-based way (ranging the value

of $\rho$ from $10^{-6}$ to $10^{-0.05}$), like described in Chapter 3. Calls with $\rho < 0.1$ were filtered as likely artifactual.

After these filtering steps, I compared the mutational spectra of mutations with low VAF (<0.05 and 0.05<VAF<0.1) with those of high VAF (>0.2). I also compared mutations with >20 supporting reads with those with <6 supporting reads. The mutational spectra of these classes of mutations were near-identical, indicating that further filtering based on VAF or read depth wasn't necessary.

### 4.3.4 Mutation rate estimation and comparisons between lesional and non-lesional skin

I estimated the mutation rate per megabase by summing the numbers of single-base substitutions, double-base substitutions and indels called in each sample and dividing it by the number of bases that had a coverage of 4X or greater. To compare the mutation rates between lesional and non-lesional skin I used linear mixed effects models. I included fixed effects for age, anatomical location of the biopsy, coverage, and the median VAF of each sample and random effects for patients and biopsies. I compared the fit of those models with that of models that additionally included a fixed effect for disease duration using a likelihood ratio test.

### 4.3.5 Mutational signature extraction

As in the work described in Chapter 3, mutational signatures of single base substitutions were extracted using the hdp package in R (section 1.4.3). I used the probability distributions for single-base-substitution signatures 1, 5, 2, 13, 7a, 7b, 7c, 7d, 17a, 17b, 18 and 38 as priors in the hdp process. Signatures of Indels were not extracted due to low number of indels observed and signatures of double-base substitutions were not extracted as a manual inspection of the mutation profiles showed that essentially all mutations were CC>NN substitutions, indicative of double-base-signature 1, attributed to UV-light.

### 4.3.6 Selection analyses

I used the dNdScv software (Martincorena et al., 2017) to identify genes enriched in non-synonymous mutations, indicative of positive selection.
To estimate the fraction of cells in each individual that carry a mutation in a particular gene,

I used the formula:

$$\frac{\sum_i VAF_{G,j} \times 2 \times V_i}{\sum_i V_i}$$

where $VAF_{G,i}$ is the variant-allele fraction of mutations in gene G in sample i and $V_i$ is the volume of sample i. This makes the incorrect but simplifying assumption that there is only one mutation per gene per sample. I also estimated the total fraction of cells in each individual as the sum of the mutation fractions across all the genes, again making the incorrect but simplifying assumption that no clones carry more than one driver. In the future, I intend to apply the pigeonhole principle to determine which mutations co-occur in the same clone and which do not. In the meantime, these estimates should be considered to be upper limits. I used a two-sided Wilcoxon signed rank test for paired data to compare the fraction of mutated cells between lesional and non-lesional biopsies.

## 4.4 Results

I used laser capture microscopy to dissect 403 samples of epidermis from lesional (N=288) and non-lesional (N=115) skin of 38 psoriasis vulgaris patients (Figure 4.1). Most samples covered a surface area of < 0.01 mm2 of the skin (Figure 4.1C). I called somatic mutations as described above.

### 4.4.1 Psoriatic skin shows a similar clonal structure and mutation burden as non-lesional skin

I found that even these small samples of skin rarely comprised fully clonal populations of cells, with most samples containing a mixture of clones (Figures 4.2 and 4.3). Psoriasis is characterized by hyperproliferation of keratinocytes and under such conditions of shortened generation time, evolutionary theory would predict accelerated clonal spread driven both by drift and selection. It was therefore surprising that the variant allele frequency did not differ between samples taken from lesional and non-lesional skin (Figure 4.3). While adjacent samples occasionally shared a fraction of their mutations in common, most mutations were private to individual samples and I generally did not observe clones spreading over large distances within biopsies either from lesional or non-lesional skin. The exception to that was when biopsies had a history of phototreatment (see below).

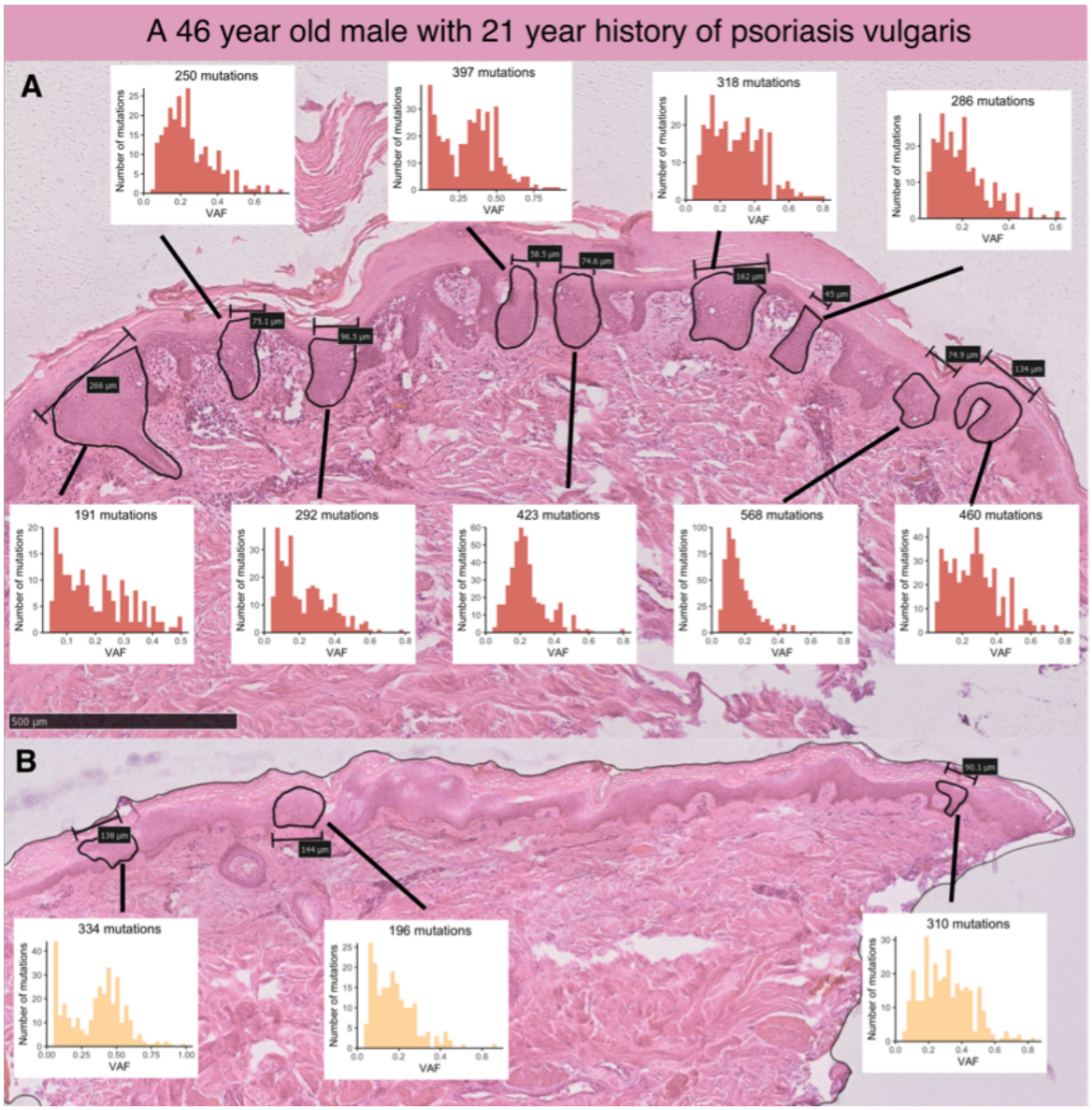



Fig. 4.2 **An example showing the samples sequenced and their corresponding VAF distributions from one of the donors, a 46 year old male with a long history of psoriasis.** A) The lesional biopsy shows a thickening of both the cornified layer and the epidermis, with rete of keratinocytes extending into the dermis below. B) The samples sequenced from the non-lesional biopsy of this patient.

As many of the samples are polyclonal, accurately estimating the mutation burden per cell is challenging. However, the mutation burden per sample increased linearly with age (0.58 mutations per megabase per year (0.23-0.94 95% CI) (Figure 4.3 middle and right panels). A few outlier samples were observed that showed extremely high mutation burden (Figure 4.3 middle). These were found to be from one of the patients with a history of phototreatment and will be discussed in the next section. To identify factors associated with mutation burden, I fit a linear mixed effects model that included a random effect for patient and biopsy and fixed effects for age, anatomical site of the biopsy, coverage and median VAF. I compared this with a model that additionally included a fixed effect for a disease duration but found that including disease duration did not improve the fit of the model (P=0.16).

I observed a striking level of heterogeneity in mutation burden of samples dissected from the same biopsy. The mutation rate of samples with similar coverage and median VAFs and that were sometimes separated by less than a millimeter of tissue, could vary 2-3 fold. This difference is driven by variation in UV-associated mutagenesis.

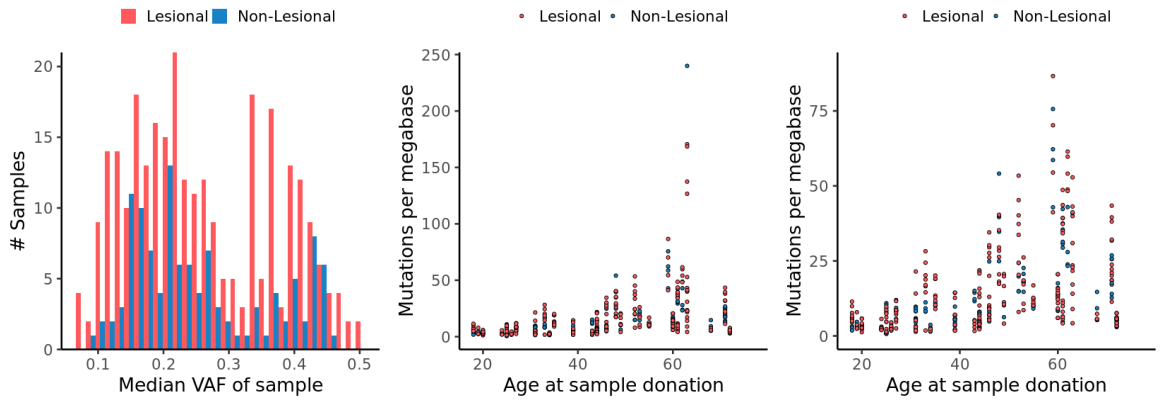



Fig. 4.3 **Clonal composition and mutation burden of epidermal samples.** Left: A histogram showing the VAF distribution of lesional and non-lesional samples. Middle: The raw mutation burden as a function of age, including outliers. Right: Same as the middle panel but the y-axis is capped at 90 mutations per megabase to better highlight the relationship between mutation burden and age.

To determine which mutagens are active in psoriatic skin, I extracted mutational signatures using the hdp package, as described in the Methods. This resulted in ten signature-components being extracted (Figure 4.4), which were compared with the COSMIC reference signatures (Alexandrov et al., 2020). Unsurprisingly, the component that explains by far the greatest number of mutations corresponds to reference signature SBS7b (Figure 4.4B), which has been attributed to UV-light (Alexandrov et al., 2013b). A component correspond-

ing to SBS7d was also extracted (Figure 4.4H), as were two components I have termed UV-component N1 and UV-component N2 (Figure 4.4E and F). The samples from which these components were extracted clustered by patients and, upon inspection of the mutational profiles of affected samples, their effects were clearly visible (Figure 4.5), indicating they represent true mutational processes. I hypothesize that these two components may reflect individual variation in UV-response, although what might cause this variation is not clear at this time.

The COSMIC database lists four mutational signatures as likely resulting directly from UV-exposure, SBS7a, SBS7b, SBS7c and SBS7d. Additionally, SBS38 is only found in melanoma samples and has been hypothesized to be the result of indirect effect of sun-exposure (Alexandrov et al., 2020). In line with a previous study of somatic mutations in normal epidermis (Fowler et al., 2021), I did not find evidence of SBS7a, SBS7c or SBS38 in this dataset. These signatures may represent extraction-artifacts or processes that are specific to melanocytes, as they are all extracted from melanomas while the PCAWG project did not include any non-melanoma skin cancers (Alexandrov et al., 2020).

The component to which second-most mutations were attributed did not correspond to any COSMIC signature but was characterized by a large number of T>A, T>C and T>G mutations at TpA sites (Figure 4.4C). This is consistent with the known mutagenic effects of treatment with psoralens and high-dose UV-A (PUVA treatment) (Esposito et al., 1988; Zhen et al., 1986). The signature was observed in both lesional and the adjacent non-lesional skin of 7 out of the 38 donors (Figure 4.7). Only 2 out of the 7 had documented history of PUVA treatment in the metadata but as this is a common treatment for psoriasis and psoralens are well-known to affect TpA sites, I nevertheless feel confident in ascribing this signature to PUVA treatment.

Additional characterization of the PUVA signature identified a large transcriptional strand bias indicative of an important role for transcription coupled repair in repairing psoralen-associated lesions. The untranscribed strand displayed an average of 2.6, 2.4 and 2.2-fold enrichment of mutations over the transcribed strand for T>A, T>C and T>G mutations, respectively (Figure 4.6A). Furthermore, I found an effect of sequence context that extends beyond the trinucleotide spectrum, with sites that have ApT 3' of the mutated base being preferentially mutated (Figure 4.6B). Psoralens have chemotherapeutic properties similar to cisplatin and mitomycin C. They function by inducing interstrand cross-links between thymines on opposite strands at TpA sites. Interstrand cross-links are highly toxic to cells as they prevent the separation of the two strands of DNA that is necessary both for transcription

and replication. Their repair usually involves a double-strand break, followed by activation of the Fanconi anemia pathway and homologous recombination. However, interstrand cross-links associated with psoralens have been found to be preferentially removed and repaired by a NEIL3 DNA-glycosylase-dependent unhooking mechanism that is independent of the Fanconi anemia pathway and avoids double strand breaks (Semlow et al., 2016). In line with this observation, PUVA treatment did not appear to affect the mutation burden of indels or double-base-substitutions.

Samples from one patient, patient34, were clear outliers in terms of mutation burden and mutational signature profile (Figures 4.3 and 4.7). In the most extreme sample, over 10,000 substitutions were observed in the exome (>250 mutations/Mb, 3-4 times the average mutation burden of keratinocyte cancers), with 90% being attributable to the PUVA signature. Clonal expansions were also associated with PUVA exposure, with the largest clones in the dataset being found within exposed biopsies. The largest clone I observed spanned 1.2 mm and was found in patient34 (Figure 4.6C).

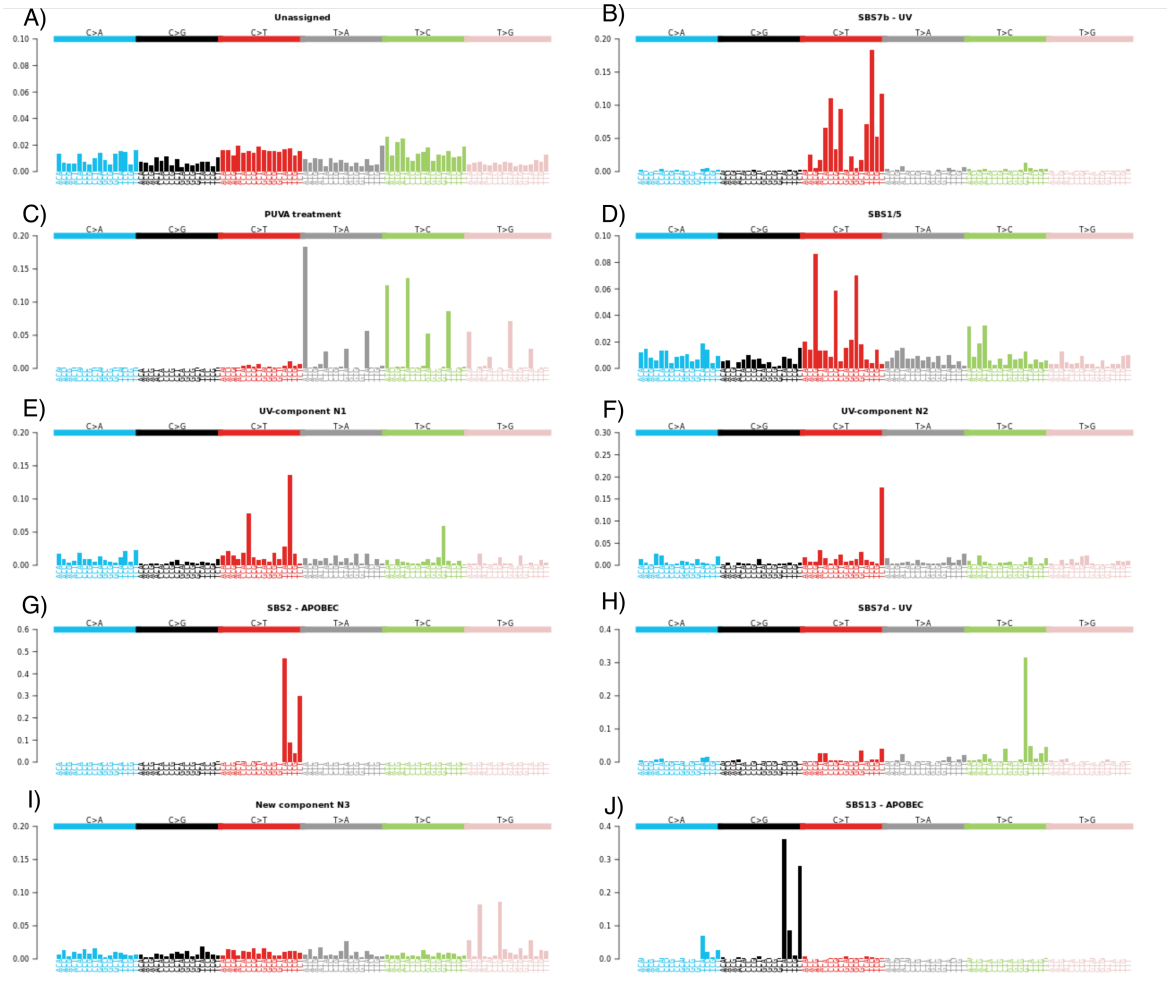



Fig. 4.4 **Mutational signature components extracted by the hierarchical Dirichlet-process algorithm.** Excluding the Unassigned component in A), the components are ordered by the number of mutations attributed to them in the dataset.

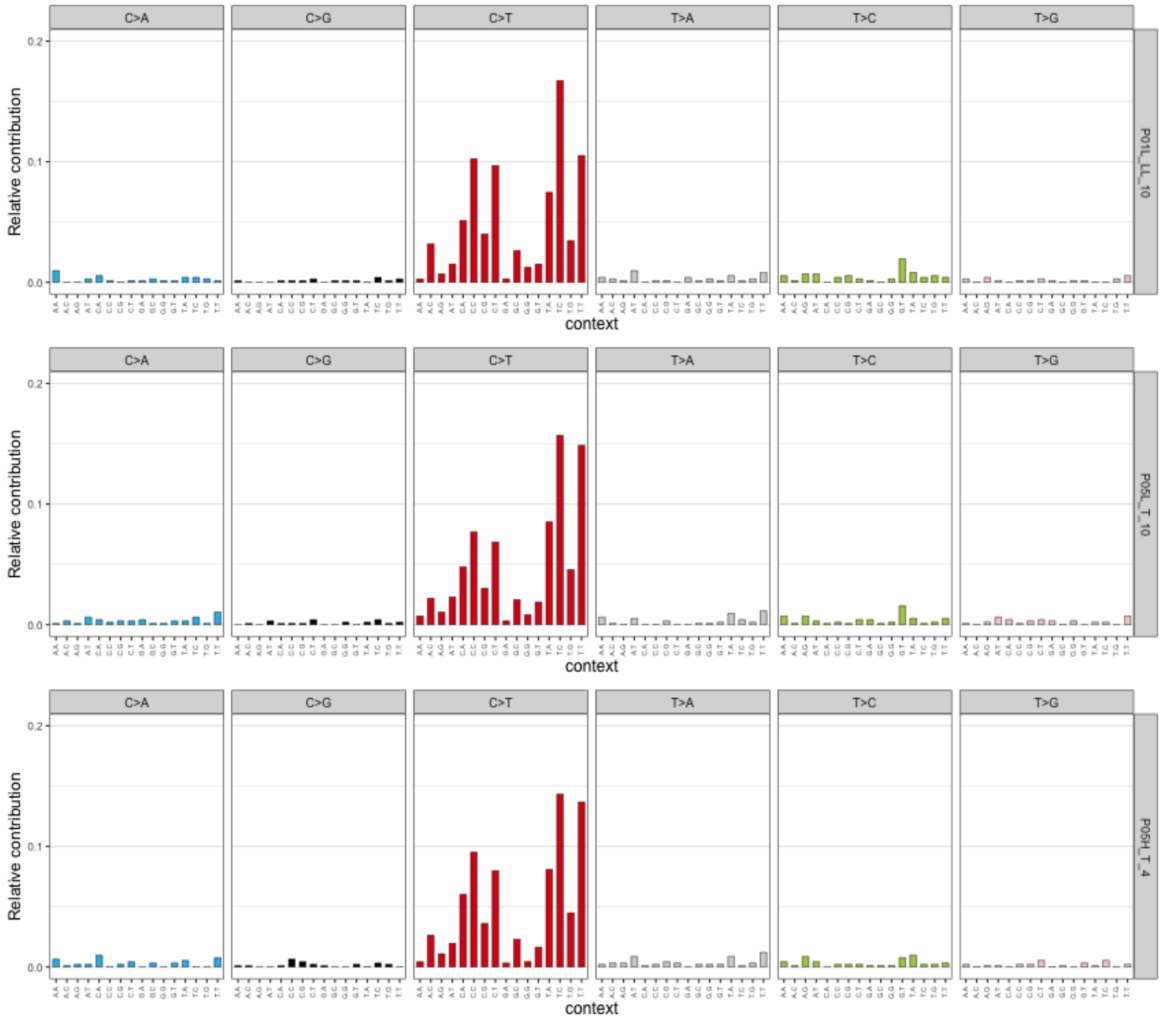



Fig. 4.5 **Individual variation in UV exposure.** The top panel shows the mutational profile of a typical sample from patient01 with a high exposure of SBS7b. The two lower panels show mutational profiles of two samples from patient05, which is one of the patients that shows a high burden of UV-component N2 (Figure 4.4F). A larger fraction of the mutations are Tp[C>T]pT mutations in the two lower panels. All samples have >1200 mutations, so this difference would be unlikely to occur due to Poisson variation alone.

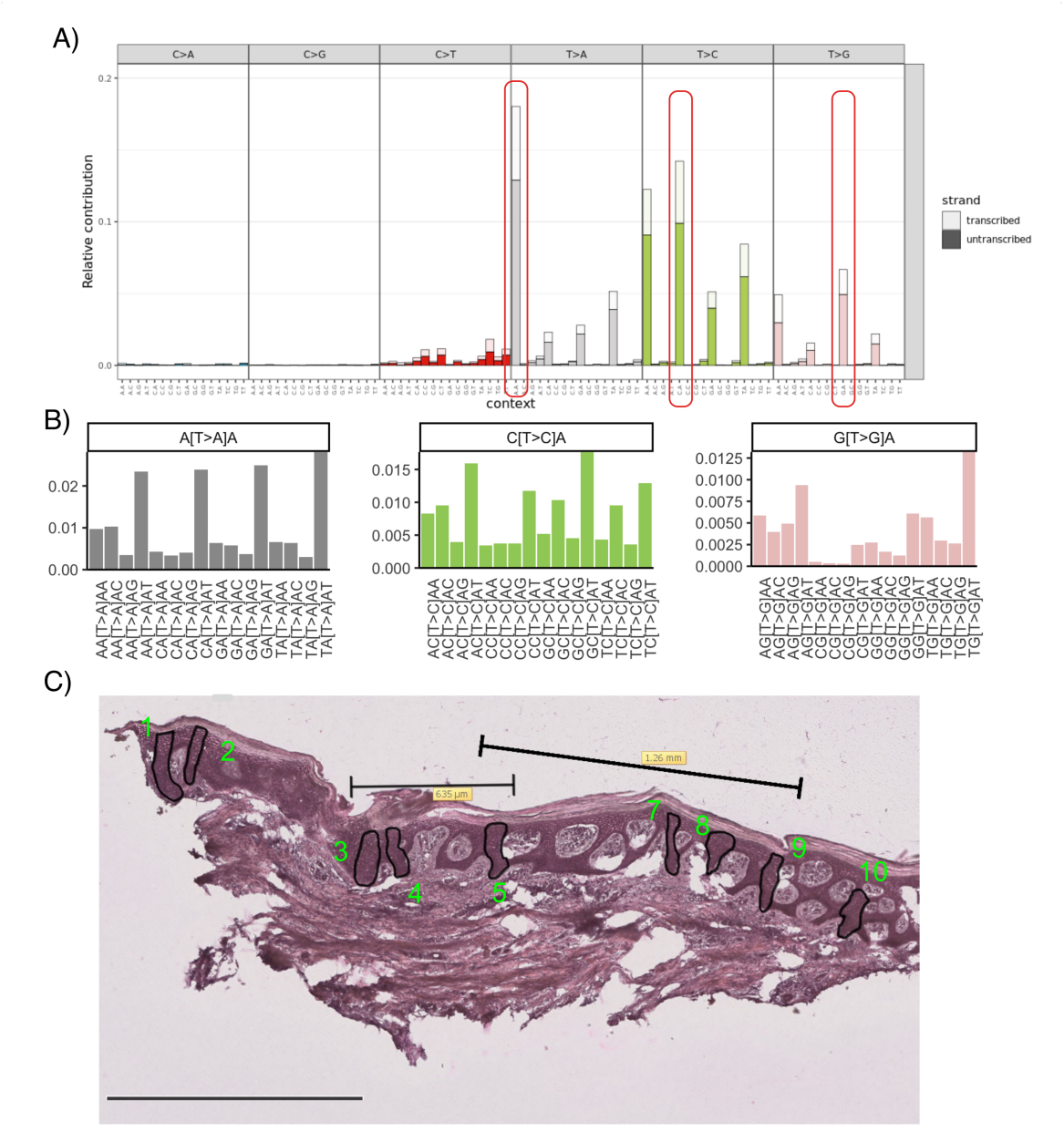



Fig. 4.6 **The effects of treatment with psoralens + UV-A (PUVA) on the mutation landscape of the skin.** A trinucleotide mutational signature of PUVA exposure showing the large transcriptional strand bias characterizing the signature. B) Pentanucleotide mutation frequencies of the peaks from (A) that are highlighted in red. NpNpTpApT are most commonly mutated. C) The lesional biopsy from patient 34 harbours two large clones. The first includes samples 3-5 and the second samples 5-9 from this biopsy (sample 5 is a mixture of both clones). The second clone is the largest clone observed in the study.

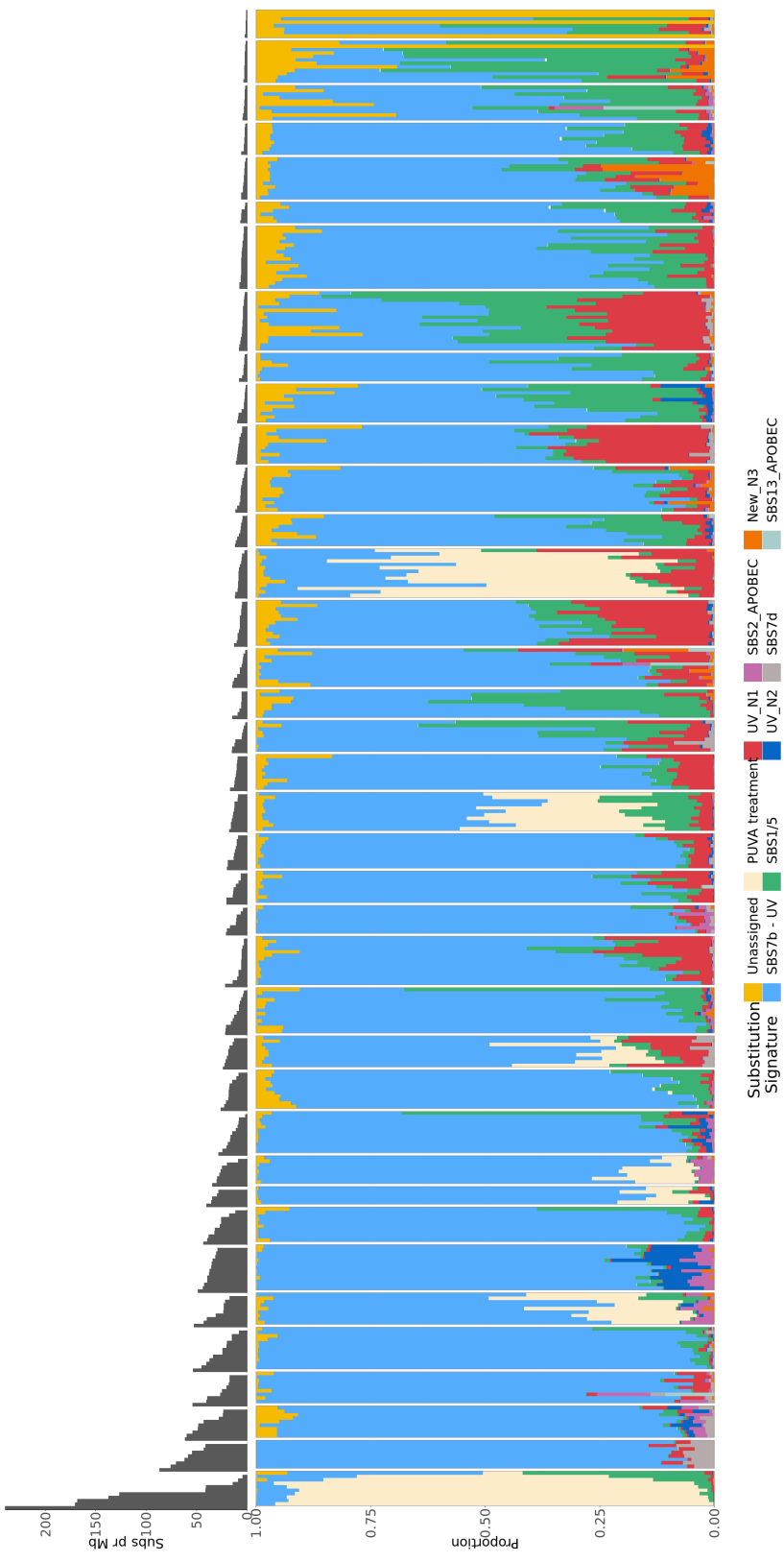



Fig. 4.7 **Signature exposure barplot.** Samples are grouped by patient and patients are ordered by descending mutation burden. The colours show the relative contribution of each signature to the mutation profile of each sample.

Signatures of cell-intrinsic mutational processes were observed in the data. One of the HDP components clearly corresponds to a mixture of mutational signatures SBS1 and SBS5, which are highly correlated and are merged into a single component by the HDP process. Both SBS1 and SBS5 are universally found in normal tissues (Moore et al., 2021) where they correlate closely with the age of the donor. Signatures 2 and 13 were observed in a handful of samples (Figure 4.7), indicating that APOBEC is occasionally activated in normal skin, as it is occasionally activated in normal colon (Lee-Six et al., 2019) and urothelium (Lawson et al., 2020). Finally, one of the HDP components was characterized by T>G mutations (Figure 4.4I). Only 617 mutations (0.3%) were attributed to this component across the entire dataset and although affected samples clustered by patient, as would be expected for a genuine mutation process, the affected samples mostly have low mutation burden and relatively flat mutational spectra. I remain uncertain as to whether this is a genuine signature or not.

In Chapter 3, I described how the burden of SBS1 and SBS5 is increased in the IBD-affected colon. The mutational spectrum of colonic epithelium is dominated by cell-intrinsic mutational processes. In contrast, the mutation spectrum of the skin is dominated by the cell-extrinsic effects of UV-light, which also causes a much larger variation in the mutation burden of the skin than in the colon. This large variation reduces the power to detect an effect of the disease on the mutation burden. To test for an effect of disease duration on cell-intrinsic mutational processes, I subtracted from the total substitution burden of each sample the number of mutations attributed either to UV-light exposure or PUVA-treatment. This reduced the median number of substitutions of each sample from 8.7 to just 0.7 mutations per megabase. I tested for an effect of disease duration on the number of mutations attributed to cell-intrinsic processes using the same LMM framework as described above but did not find a significant effect of disease duration (P=0.27).

I did not formally extract double-base-substitution signatures. Manual inspection of the raw mutation profiles showed that essentially all double-base-substitutions were CC>TT substitutions, corresponding to the reference signature DBS1, which has been attributed to UV-light (Alexandrov et al., 2020). As UV-exposure is not as strongly associated with the formation of indels as substitutions, they are a smaller fraction of mutations than in the colon. The median number of indels per sample was just 3 and I did not extract indel signatures for this thesis chapter.

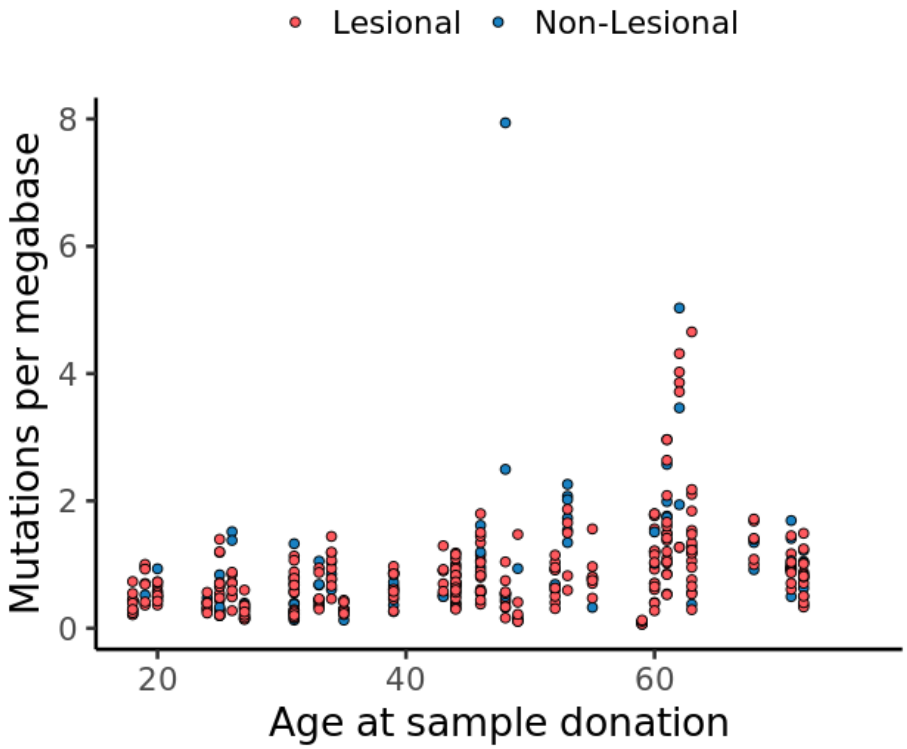



Fig. 4.8 **Mutation burden attributed to cell-intrinsic processes as a function of age.**

Table 4.1 Recurrently mutated genes in lesional and non-lesional skin from psoriasis patients. Shown are the number of mutations in each annotation class: Synonymous (syn), missense (mis), nonsense (non), splice site (splice) and indels or double-base-substitutions.

| Gene | $n_{syn}$ | $n_{mis}$ | $n_{non}$ | $n_{splice}$ | $n_{indels/DBS}$ | q |
|---|---|---|---|---|---|---|
| *NOTCH1* | 4 | 41 | 10 | 14 | 18 | $< 2.2 \times 10^{-16}$ |
| *FAT1* | 2 | 15 | 14 | 7 | 13 | $< 2.2 \times 10^{-16}$ |
| *PPM1D* | 0 | 0 | 6 | 0 | 2 | $1.1 \times 10^{-4}$ |
| *TP53* | 0 | 6 | 1 | 0 | 5 | $5.4 \times 10^{-4}$ |
| *ZFP36L2* | 2 | 8 | 0 | 0 | 5 | $2.3 \times 10^{-3}$ |
| *NOTCH2* | 2 | 18 | 2 | 0 | 9 | $4.0 \times 10^{-3}$ |

### 4.4.2 Positive selection in psoriatic skin resembles that in normal skin

I used the dNdScv package to test if any genes were enriched or depleted of mutations, which would be indicative of selection of those mutations. I found evidence of positive selection of mutations in six genes, *NOTCH1*, *FAT1*, *PPM1D*, *TP53*, *ZFP36L2* and *NOTCH2* (Table 4.1).

All genes but one, *ZFP36L2*, have been previously described in studies of normal skin (Fowler et al., 2021; Martincorena et al., 2015). The mutations in those genes followed the same non-random distribution as has been previously described (Martincorena et al., 2015) (Figure 4.9). *NOTCH1* and *NOTCH2* had a large number of missense mutations affecting the extracellular epidermal growth factor–like domains, as well as many truncating mutations scattered throughout the genes. Mutations in *FAT1* did not obviously cluster but showed a similar pattern as in a previous report (Martincorena et al., 2015). Several mutatins in known mutation hotspots in *TP53* were found, including R282W, R175H, R248Q and R248W. There were also two nonsense mutations in *TP53*. Lastly, the mutations in *PPM1D* were all truncating mutations in exon 6 of this gene. Nonsense mutations in this region have been shown to result in PPM1D overexpression due to loss of a C-terminal degradation signal and reduced proteasomal degradation (Kahn et al., 2018). This in turn can result in impaired p53 function (Kleiblova et al., 2013).

*ZFP36L2* belongs to a family of zinc-finger proteins that bind to the 3' untranslated regions of particular mRNAs and promote their decay. While this gene is not a part of the targeted panels used in previous studies of the skin, mutations in *ZFP36L2* have been previously reported to be under positive selection in the normal oesophagus, and to be more frequently mutated in the normal tissue than in oesophageal squamous cell carcinomas (Yokoyama et al., 2019). Together with another family member, ZFP36L1, ZFP36L2 has been reported to down-regulate Notch1 during lymphocyte development (Hodson et al., 2010). It may

be that it affects the evolution of epithelial cells in skin and oesophagus through an effect on the Notch-pathway and that mutations in this gene are not specific to psoriatic skin. In support of that hypothesis, mutations in the gene are found both in samples from lesional and non-lesional biopsies at a similar rate (Figure 4.10).

However, the possibility that mutations in *ZFP36L2* play a role in the pathogenesis of psoriasis cannot be discarded. Mice that are full *Zfp36l2* knock-outs die soon after birth (Stumpo et al., 2009) and I am not aware that the gene has ever been conditionally knocked out specifically in keratinocytes. However, this has been done for another member of the family, *Zfp36* itself, which encodes the endogenous antiinflammatory protein tristetraprolin. Mice that are full *Zfp36* knock-outs develop a systemic inflammatory syndrome characterized by cachexia, myeloid hyperplasia, arthritis and progressive dermatitis (Taylor et al., 1996). This phenotype can be rescued by administering anti-TNFa antibodies (Taylor et al., 1996), a common treatment for psoriasis. Conditional knock-out of *Zfp36* in mouse keratinocytes leads to the spontaneous formation of psoriatic-like skin lesions and dactylitis (Andrianne et al., 2017) while enhanced stability of the protein is protective against imiquimod-induced dermatitis, a common experimental model of psoriasis (Patial et al., 2016). Several cytokines key to the pathogenesis of psoriasis have been reported to be targets of ZFP36, including TNF (Carballo et al., 1998; Taylor et al., 1996), IL-12 (Jalonen et al., 2006), IL-17 (Lee et al., 2012), IL-23 (Qian et al., 2011), IFNg (Ogilvie et al., 2009) and more (Brooks and Blackshear, 2013). While ZFP36L2 has been less extensively characterized than ZFP36, the zinc-finger mRNA binding domains of the ZFP36 family members are highly conserved and all have been shown to behave similarly in terms of RNA binding in cell-free systems. For example, all family members can destabilize TNFa mRNA (Lai et al., 2003) and likely the same is true of other targets. Since three of the mutations in *ZFP36L2* are out-of-frame indels and the missense variants do not cluster in any particular area of the gene (Figure 4.9), the mutations are likely loss-of-function, meaning they would be predicted to result in increased stability of mRNAs of a wide range of pro-inflammatory cytokines within the epidermis and thus potentially contribute to psoriasis pathogenesis. The non-lesional biopsies are directly adjacent to lesional skin and it cannot be ruled out that those areas have been previously affected by psoriasis. The *ZFP36L2* mutations would also be predicted to increase the stability of *NOTCH1* mRNA, resulting in gain-of-function of *NOTCH1*. This is the opposite direction of effect compared with mutations in *NOTCH1* itself, which are probably loss-of-function. It will be necessary to assess the frequency of *ZFP36L2* mutations in skin from healthy donors to decide which is the more likely interpretation: That mutations in *ZFP36L2* are selected because they affect Notch1 signaling in the skin in general or if they

are selected only in psoriatic skin, where they potentially contribute to disease pathogenesis.

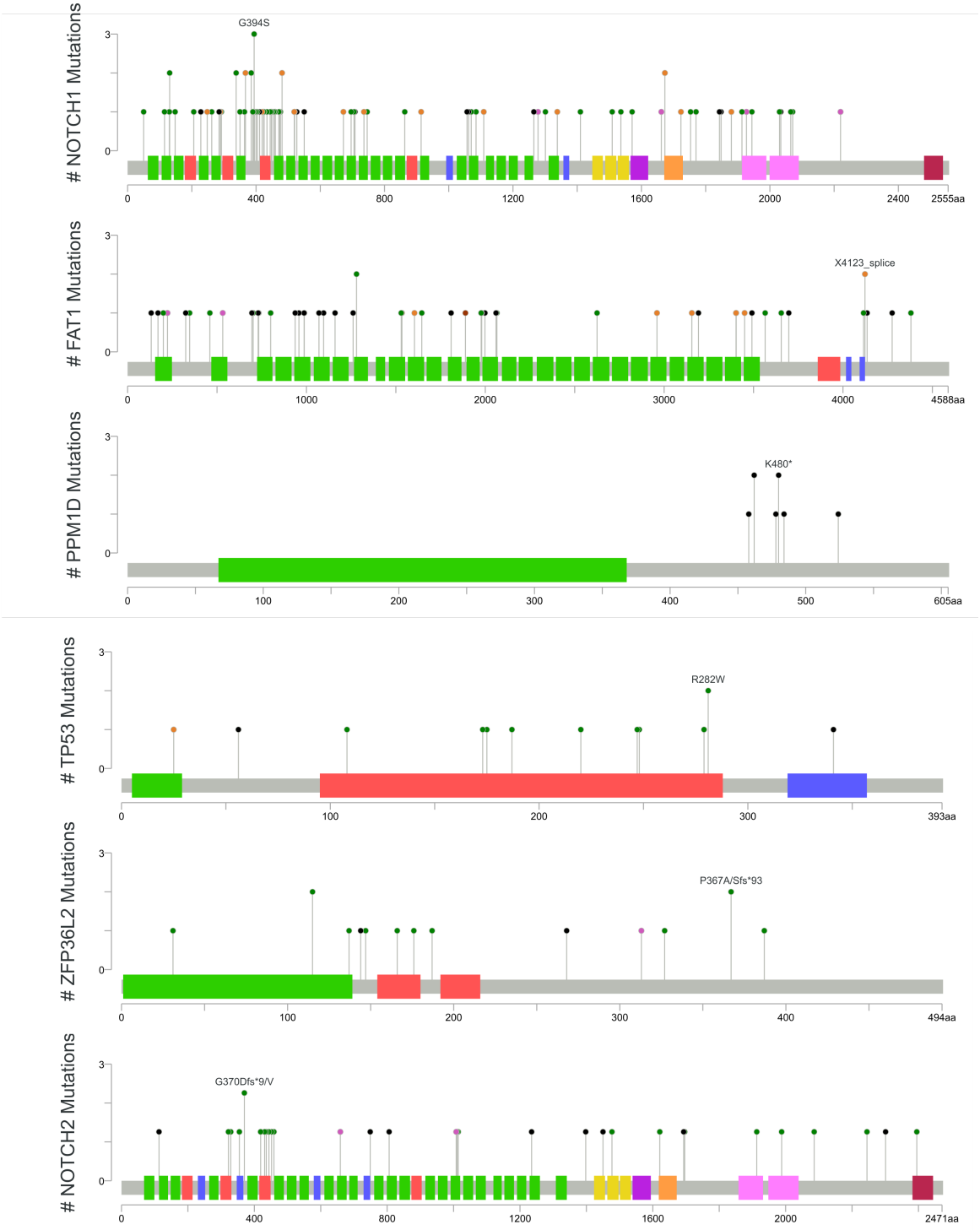



Fig. 4.9 **Distribution of mutations in positively selected genes.**

I next estimated the fraction of cells that carry a mutation in a particular gene across all samples from each individual (Figure 4.10A and B). As expected, the fraction of mutated cells increases with age but no significant differences were observed between lesional and non-lesional biopsies (Figure 4.10C, $P > 0.3$ for all genes. Wilcoxon signed rank test).

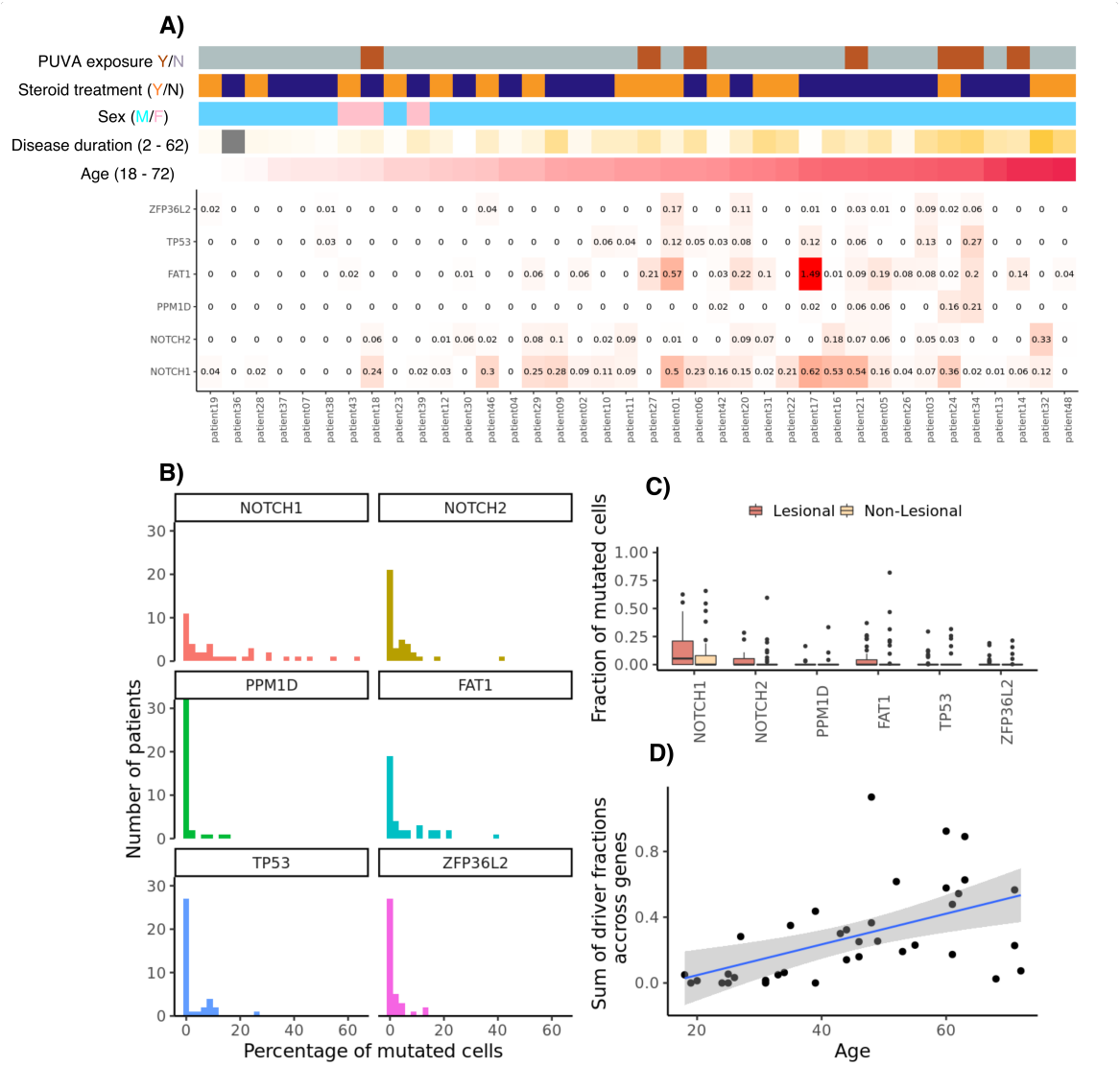



Fig. 4.10 **Fraction of cells that carry mutations likely to be under positive selection.** A) The fraction of cells that carry non-synonymous mutations in any of the genes found to be under positive selection. Patients are ordered by ascending age. B) Histograms of the fractions presented in (A). C) Boxplots showing the fraction of mutated cells in lesional vs non-lesional biopsies. D) Fraction of cells carrying a mutation in any of the six genes as a function of age.

## 4.5 Discussion and future direction

At the time of writing this thesis chapter, this project is still on-going. In addition to the 400 exomes from 38 patients presented here, a further 800 samples from 71 more patients are in various stages of the exome-sequencing pipeline. The addition of those samples will add greatly to our power to detect mutations under selection in particular but also to detect additional factors that influence the mutation burden of the skin and add to our power to extract mutational signatures.

In contrast to chapter 3, where we used whole-genome sequencing, I opted for using whole-exome sequencing for the work presented in this chapter. The skin lacks a clear and clonal histological feature like the crypt and more polyclonal samples require higher depth of sequencing for calling mutations. The lower cost of exome sequencing enables us to sequence many more samples than would have been possible had we opted for whole genome sequencing. This choice was made to maximize our power to identify driver mutations but comes at a cost of lower power to identify mutational signatures and greater uncertainty of the signature exposure estimates in each sample.

That samples from lesional skin are no more clonal than samples from non-lesional skin was surprising. Under conditions of hyperproliferation, one would expect clones to grow larger, even in the absence of selection. One possible explanation may be that there is greater cell movement in the lesional skin than in the non-lesional. Under normal conditions, differentiating cells stratify vertically from the basal layer, through the suprabasal layers until they are shed. I speculate that psoriasis may be associated with increased lateral displacement and "mixing" of squamous cells derived from a larger population of stem cells. The fact that putative driver mutations were not found at increased frequency in lesional skin compared with non-lesional would also suggest that psoriasis has minimal effect on clonal spread.

I observed a high variation in mutation burden, with samples separated by 1mm differing up to 2-3 fold in their mutation burden. This variance could not be explained by differences in coverage or clonal composition of the samples. Fowler et al similarly found large variation in the mutation burden of punch-biopsies 250 micrometers in diameter (Fowler et al., 2021), indicating that this variation is likely biological, rather than technical and that similar UV exposure can affect adjacent cells differently.

That a majority of the mutations were attributed to the mutagenic effects of UV-light is unsurprising. Two components of the HDP extraction process did not correspond to

any COSMIC reference signatures but likely reflect individual variation in the mutational spectrum of UV-light. One of the limitations of both the NMF and the HDP methods for signature extraction is that they both treat signatures of active mutational processes as static probability distributions when in reality the same mutagen can likely result in slightly different mutational spectrums depending on the tissue/cell type and on the germline background of the individual. The HDP component in Figure 4.4F is characterized by a peak in T[C>T]T, but smaller peaks also exist at G[C>T]T, C[C>T]T and A[C>T]T, indicating that the variation is due to processing of C>T mutations when T is the 3' base. A direction of future work is to carry out the signature extraction with strand information. If this component shows a strong transcription strand bias, it may be an indication that the variation in UV-related mutagenesis has to do with repair of UV-associated lesions.

Psoralens + UVA (PUVA) treatment has been known to cause mutations at TpA sites for decades and patients receiving this treatment are already monitored for skin cancers at the exposed sites. Nevertheless, I believe that my characterization of the mutagenic effect of PUVA treatment will add significantly to our understanding of this mutational process. For example, I have demonstrated the effect of extended sequence context and the presence of a large transcriptional strand bias. The transcriptional strand bias will likely translate into a negative relationship between the gene-mutation burden and expression levels, although this analysis is still pending. Unfortunately, I lack the data to establish a dose-response curve for the relationship between PUVA treatment and mutation burden, as treatment quantity is not recorded in the clinical metadata. To explore this signature in even greater detail, I have submitted 16 clonally unrelated samples, which showed evidence of PUVA exposure in the exome mutation profiles, for subsequent whole-genome sequencing. This will enable me to assess the effect of PUVA treatment on the burden of structural variants and the effects of genomic features like replication timing and chromatin state on the mutation burden.

The largest clones in the dataset were found in biopsies that showed clear evidence of PUVA exposure. There are two possible explanations for this. The first is that the high mutation burden has resulted in the accumulation of multiple driver mutations in a single clone, rendering it much "fitter" than its neighbours and enabling its expansion on a large scale. The presence of the PUVA signature did not seem to be associated with a larger fraction of cells carrying mutations in any of the genes as yet found to be under positive selection (Figure 4.10A) but I cannot rule out the presence of other mutations that may be selected for specifically under the conditions of PUVA treatment. The second explanation is that it is the cytotoxicity of the phototreatment that enables clonal expansion, similar to the

effects of inflammation in IBD. Of course, these two explanations are not mutually exclusive.

I detected positive selection of mutations in six genes. These include just 5 of the 14 genes previously found to be under positive selection in the skin (Fowler et al., 2021), suggesting that more genes are likely to reach significance as sample size is increased. The identification of *ZFP36L2* even at this early stage highlights the utility of WES over targeted panels. As discussed above, it is not clear whether mutations in this gene are preferentially selected in the skin of psoriatic individuals and determining the frequency with which it is mutated in the skin of donors without psoriasis is a direction of future research.

# Chapter 5

# Discussion

In this thesis, I have described my contributions to furthering our understanding of somatic evolution in non-neoplastic colon and skin, and the changes to the somatic evolution landscape of these tissues associated with IBD and psoriasis. In this final chapter, I will describe how my results fit into the broader context of somatic evolution of cells within the entire body, and the importance of understanding somatic evolution in common non-neoplastic diseases in general. I shall also discuss the difficulty of assigning a causal direction for genes found to be recurrently mutated in non-neoplastic diseases. Finally, I will give my perspective for the future and offer thoughts on how to carry out joint study of somatic mutations and germline variation and to scale up studies of somatic evolution in solid tissues.

Figures 5.1 and 5.2 in this chapter and parts of the text have been previously published in Trends In Genetics in an article titled "Somatic mutations provide important and unique insights into the biology of complex diseases" by myself and Carl Anderson.

## 5.1 Somatic evolution during normal aging

### 5.1.1 The relationship between mutagenesis and cancer risk is unclear

In this thesis I have cited multiple studies of normal tissues that have been published in recent years. These have confirmed what had been suggested in studies of cancers, that SBS1, SBS5 and, to a lesser extent, SBS18 represent cell intrinsic mutational processes operating near universally across all cells of the body, even those which rarely divide post-mitotically (Abascal et al., 2021; Franco et al., 2018). The APOBEC-associated SBS2 and SBS13 are cell-intrinsic signatures seen more sporadically but these have still been observed across a range of normal tissues. Signatures of exogenous exposures like alcohol, smoking, UV-light

and mutagenic drugs have also been reported mostly in the expected tissues and there haven't been too many surprises from studies of normal tissues. The discovery of SBS88 in the colon (Lee-Six et al., 2019) and its attribution to colibactin produced by *pks*+ E.Coli (Pleguezuelos-Manzano et al., 2020) serves as an example of how mutational signature analyses can reveal mutagenic mechanisms of potential importance to public health. On the whole however, I think mutational signature analyses of normal tissues have mostly revealed what one might have predicted from studying cancers of those tissues.

The utility of measuring the mutation burden in normal cells and attributing it to the different mutagens that have acted on the cell is limited by our poor understanding of the relationship between mutation burden and cancer risk. The view that cancer risk is proportional to mutagenesis is overly simplistic. On one hand, many known carcinogens are clearly mutagenic and germline variants associated with high lifetime cancer risk frequently drive increased mutagenesis (for example germline variants in *BRCA*, *POLE*, *POLD* and genes of the MMR pathway). On the other hand, it is becoming clear that cells in normal tissues can have extremely high mutation burdens without undergoing neoplastic transformation (although the number of structural variants tends to be higher in cancers). Early results from the Mutograph project show that neither differences in mutation burden nor mutational signature compositions can explain the varying incidence of oesophageal cancers across different parts of the world (Moody et al., 2021). Finally, a recent study of 20 human carcinogens found that most did not generate distinct mutational signatures or increase mutation burden (Riva et al., 2020), challenging the classical view that carcinogens cause cancer simply through effect on mutagenesis.

In Chapter 3, I describe how in IBD the substitution rate is increased 2.3 fold and the indel rate is increased 7 fold compared with normal colon. The increased risk of colorectal cancer among IBD patients is potentially due to increased mutation burden, greater opportunity for clonal expansions following widespread cell death, changes to the selection landscape or, as I think is most probable, some mixture of all three. It would aid in the interpretation of these results if it were possible to deconstruct the cancer risk into its individual components. Which is more dangerous, a single crypt with a high mutation burden or a widely expanded clone with a lower mutation burden? The ability to deconstruct the risk would also be helpful when trade-offs present themselves during treatment. For example, I found a mutational signature of purine treatment in some of the IBD patients with a history of this treatment. Even if a dose-response curve between purines and mutagenesis could be established, it

would not be possible to determine the point at which the mutagenic effect of the treatment offsets the reduction in cancer risk gained by bringing the disease under control.

### 5.1.2 Drivers do not always lead to cancer

Genes are typically linked to cancer by observing that they are mutated more often than expected by chance (Martincorena et al., 2017). However, this in itself is not proof the gene plays a role in malignant transformation but merely shows that the mutated clone has been positively selected for at some point in the life of the individual. The finding that *NOTCH1*, a gene previously thought to be an oesophageal cancer gene, is mutated more often in normal oesophagus than in cancer (Martincorena et al., 2018; Yokoyama et al., 2019) serves as the best illustration of this principle. The oesophagus undergoes extensive remodelling with age to the point that nearly every wild-type cell in the tissue is replaced by a *NOTCH1* mutant cell but this does not seem to drive cancer development and indeed, *NOTCH1* mutations may even protect the tissue from cancer (Abby et al., 2021; Colom et al., 2021). The observed enrichment of *NOTCH1* mutations in oesophageal cancers appears to be merely a consequence of an even higher enrichment of mutations in this gene in normal cells and the interpretation of *NOTCH1* as an oesophageal cancer gene is the result of the field of cancer genomics historically lacking normal control samples.

A second observation worth noting is the difference in mutation frequency of known "cancer genes" in some normal tissues. The study of normal urothelium by Lawson et al serves as a great example of this point. Lawson et al discovered that among established urothelial cancer genes, only a subset, primarily those involved in chromatin remodelling, are commonly mutated in normal urothelium while genes in the RTK-Ras-PI3K and p53-Rb pathways are rarely mutated (Lawson et al., 2020). This suggests that the chromatin remodelling genes dominate evolution in the urothelium under normal conditions but that mutations in other genes may be necessary for malignant transformation. In my opinion, the growing understanding of somatic evolution within normal tissues should bring about a change in the way we think about driver mutations and the biological consequences of clonal expansions within a tissue. Not all "cancer genes" may actually be cancer genes, or at least they may drive different levels of pre-malignancy and a further sub-classification may be required for this term to remain useful. This has implications for drug development. I speculate that systemic targeting of mutations which are found in cancer but are also prevalent in normal tissues is likely to be associated with more severe side effects than targeting of the mutations which have driven the neoplastic transformation itself.

The precise definition of a cancer gene is of course only of indirect relevance to this work, which focuses on evolution in non-neoplastic diseases, except to highlight that rich clonal evolution can and does take place within normal tissues and that clonal expansions are not always associated with cancer development. This leads one to ask, could somatic mutations and clonal expansions play a role in other, non-malignant, phenotypes?

## 5.2 Somatic evolution and non-neoplastic disease

The relationship between somatic evolution and non-neoplastic disease presents a chicken-or-the-egg problem. Chronic diseases often have profound consequences on the cellular constitution and the environment of affected tissues. The selection forces operating within a tissue are likely changed by disease and/or the disease treatment, and mutations that were neutral under normal conditions may become advantageous in disease conditions. Disease can also be associated with accelerated mutagenesis, exposure to novel mutagens and affect genetic drift by altering cell proliferation. On the other hand, it is possible that somatic mutations may directly contribute to a disease process either by initiating, maintaining or even potentially resolving a disease. The work presented in Chapter 3 of this thesis and other published work suggests that further testing the hypothesis that somatic mutations play a causal role in diseases is a reasonable thing to do. The case for cause or consequence will be discussed in subsequent sections. It is clear, that to affect the organismal level phenotype, mutations in specific genes must reach sufficient frequency within a tissue. This minimal prevalence within the tissue is likely to vary between mutations, tissues and phenotypes and possibly between individuals, who may differ in their ability to tolerate specific mutations.

### 5.2.1 Somatic evolution as a consequence of disease

#### Disease and mutagenesis

Complex diseases are sometimes associated with changes in mutation burden of the affected tissues. I showed in Chapter 3 of this thesis how mutagenesis is accelerated in the IBD colon compared with normal, 2.3 fold for substitutions and 7 fold for indels. An increase in the mutation burden has also been reported in liver cirrhosis compared with normal liver (Brunner et al., 2019) as well as a pronounced increase in the number of structural variants and copy-number alterations in the diseased tissue. However, the same effect of accelerated mutagenesis is not seen across all diseased conditions. In Chapter 4, I could not detect an effect of disease duration on the mutation burden of psoriatic skin. This may indicate that the hyperproliferation of keratinocytes, which characterizes the disease, does not translate into

higher mutation burden.

When a disease is associated with increased mutation burden, I think it is most likely the disease which accelerates the mutation rate, rather than the other way around. The alternative hypothesis, which assumes that somatic mutations contribute to disease pathogenesis, is that individuals with naturally higher rates of mutations would be at an increased risk of developing a disease and the mutation rate would therefore appear higher in affected individuals and have nothing to do with the disease per se. In this scenario, the same mutations would have to be selected in both normal and diseased conditions. This may happen in some diseases but I don't think it is the case for IBD, where the selection forces appear to be altered compared with normal colonic mucosa.

**Disease-driven clonal expansions and changes in selection landscapes of affected tissues**

Diseases can be associated with differences in clonal expansions within affected tissues. I would expect this effect to be especially pronounced for diseases that involve significant cell death as these create the conditions for bottleneck-expansion cycles where surviving clones rapidly expand to replace their perished neighbours within the tissue. Accelerated clonal expansions are evident in IBD (Kakiuchi et al., 2020; Olafsson et al., 2020) and in liver cirrhosis (Brunner et al., 2019), where clones in affected regions far exceed those from healthy regions in size. In general, mutated cells may reach high frequency in a tissue either through the large-scale expansion of a single clone or through parallel evolution of multiple clones each carrying a distinct mutation in the same gene or pathway. Both modes of expansion have been observed in IBD. Although the small size of our biopsies prevented me from detecting very large clones in the European IBD cohort (Chapter 3), Kakiuchi et al reported many instances of massive clones in their cohort of Japanese IBD patients, with the largest clone covering 19 $cm^2$ of colonic epithelium (Kakiuchi et al., 2020). Evidence of parallel evolution was found both in my British cohort and the Japanese cohort. For example, in one patient in the British cohort, I detected four clones from three distant sites of the colon that each carried a distinct LoF mutation in *PIGR* (Figure 3.18). All three sites had a history of inflammation.

The skin biopsies available in the psoriasis study were similarly small but it was nevertheless evident that clonal expansions on the scale of millimeters or centimeters are rare and clones spanning entire psoriatic lesions, which can cover many square centimeters of skin, likely do not exist or are negligibly rare. Parallel evolution however, was evident and I found

up to 7 distinct *NOTCH1* mutations within the same biopsy.

Psoriasis does not seem to drive clonal expansions in two dimensions as there are no obvious differences in VAFs of samples of similar surface area. Clones may however expand in the third dimension, as rete extending into the dermis have similar clonal structure as "flat" epidermis of comparable width. Thus, we may have samples of comparable clonal composition and covering equal surface area but the clones from the lesional biopsy have a greater volume. It is unclear however, if this has any lasting effect on the clonal structure of the epidermis once the disease is in remission. I could not find any evidence that a greater fraction of cells carry putative driver mutations in lesional compared with non-lesional skin.

In contrast to IBD, the selection landscape of psoriatic skin does not seem to be characterized by mutations in immune-related genes. The only gene found to be under selection that hasn't previously been described in studies of normal skin is *ZFP36L2*, and there are reasons to believe that may be under selection in normal skin as well, as listed in Chapter 4. In contrast to psoriasis, IBD is associated with extensive cell death. Crypt density is reduced and ulcers may form in affected areas of the colon. Under these conditions, clones that carry mutations that enable them to withstand the local cytotoxic forces may be able to rapidly expand. In contrast, psoriasis is associated with hyperproliferation rather than cell death and the affected area of the tissue becomes more, not less crowded with dividing cells. Clones carrying potential disease specific mutations have to outcompete the clones already present in the normal skin, which may already have a high fitness due to accumulation of mutations in *NOTCH1*, *FAT1* etc before the disease onset or during periods of remission. At this stage in the analysis, psoriasis does not seem to affect the fitness landscape of the skin.

### 5.2.2 Somatic evolution as a pathogenic force in disease

One of the best characterized examples of somatic evolution leading to complex disease is the relationship between clonal hematopoiesis and cardiovascular disease. Between 10 and 20% of individuals over the age of 70 harbor a mutant clone that accounts for >4% of their blood cells, typically carrying a mutation in *DNMT3A*, *TET2* or *ASXL1* (Jaiswal et al., 2014; Watson et al., 2020), and copy number alterations and clonal expansions without driver mutations are also common (Loh et al., 2018; Zink et al., 2017). In addition to increased risk of developing a hematological malignancy, clonal hematopoiesis is associated with many cardiovascular outcomes, including ischemic stroke, atherosclerosis, myocardial infarction and more (Jaiswal and Ebert, 2019).

Evidence from mouse studies suggests that the relationship is a causal one. When mice are xenografted with *Tet2* mutant bone marrow and fed on a high-fat diet they are more likely to develop atherosclerosis than control mice (Fuster et al., 2017; Jaiswal et al., 2017). Mutations in *Tet2*, which encodes a chromatin modifier that represses the transcription of pro-inflammatory molecules, result in increased expression in monocytes and macrophages of various cytokines and chemokines including interleukin (IL) 6, IL-1b and members of the CXC family. This drives a higher expression of endothelial adhesion molecules, leading to increased leukocyte recruitment to the aortic site. Macrophage uptake of lipids and cholesterol crystals ultimately leads to plaque formation and atherosclerosis (Fuster et al., 2017; Jaiswal et al., 2017).

I described in Chapter 3 how IBD patients frequently carry somatic mutations in *PIGR* and genes in toll-like receptor and interleukin-17 pathways, most notably *NFKBIZ* and *ZC3H12A* (Kakiuchi et al., 2020; Nanki et al., 2020; Olafsson et al., 2020). These genes are not known drivers of colorectal cancer and may in fact be negatively selected during tumour development (Kakiuchi et al., 2020). They do, however, play a key role in maintaining microbe-epithelial homeostasis. Knock-out studies in mice suggest that mutations in *PIGR* may contribute to the deterioration of the epithelial barrier and allow microbes to penetrate the underlying tissue (Johansen et al., 1999; Sait et al., 2007; **?**). Conditional knockout of the IL-17 pathway has similarly been shown to cause dysbiosis and promote autoimmunity in a mouse model (Kumar et al., 2016). The mouse data is in my opinion convincing evidence that the mutations identified may play a causal role in IBD (Figure 5.1). However, if this is the case, many questions remain unanswered. While Nanki et al showed that mutations in the IL-17 pathway may protect cells against cytotoxic effects mediated by IL-17A (Nanki et al., 2020), it remains unclear what selective advantage *PIGR* mutations confer upon cells and if those clones are able to expand within the colons of "IBD susceptible individuals" even before disease onset.

The IBD cohort described in Chapter 3 is small and is biased for patients with long-standing disease. A larger, random sample would be required to determine what fraction of IBD patients carry mutant clones of pathogenic potential. A larger study would also likely discover additional genes in the IL-17 pathway under positive selection, as the enrichment I detected in this pathway was not driven by any of the genes that individually reached significance, like *NFKBIZ* or *ZC3H12A*. Finally, a larger study would be able to look for germline variants that affect clonal evolution in the IBD colon. Germline variants near *NFKBIZ* and components of the TLR/IL-17 pathways have been associated with IBD (de Lange et al., 2017) and variants in *ZC3H12A* with related autoimmune diseases (Tsoi et al., 2017) but it is

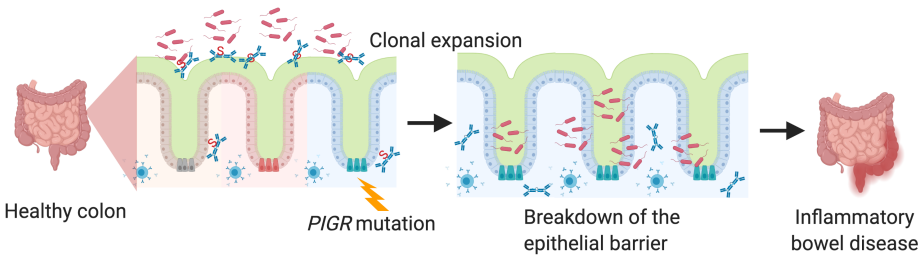



Fig. 5.1 **A causal theory for *PIGR* mutations in IBD.**Somatic mutations in *PIGR* may contribute to the pathogenesis of inflammatory bowel disease, although this has not been confirmed. A clonal expansion of a *PIGR* mutant cell results in locally reduced transfer of IgA across the epithelial membrane from its site of production in the lamina propria. This may facilitate the breakdown of the epithelial barrier and enable resident microbes of the colon to cross, raising an immune response. Figure created with BioRender.com.

unclear if those variants affect clonal spread or contribute to the disease process in a different manner.

In blood, hematopoietic stem cells are able to freely mix within the tissue. In this setting, a large fraction of the tissue frequently comes to be dominated by a single clone of cells. In hematology, clonal hematopoiesis of indeterminate potential is clinically defined as the state where a clonal mutation is found in at least 4% of the nucleated blood cells (Jaiswal and Ebert, 2019). The threshold of 4% for a clinical diagnosis of clonal hematopoiesis is arbitrarily chosen and we do not know what frequency of cells must be mutated to affect different phenotypes. Clones in solid tissues are more spatially constrained and the idea that a single clone could grow to cover 4% or more of a solid tissue like the colon or the skin is neigh unthinkable. It seems much more probable that in solid tissues, mutant cells could reach the frequency needed to affect a phenotype through parallel evolution of multiple clones carrying distinct mutations in the same gene or set of genes, rather than through a large-scale expansion of a single pathogenic clone.

Finally, I want to make the point that if somatic mutations play a role in common complex diseases, they should not be seen as deterministic of the outcome. Rather, they should be regarded as risk factors which may be orthogonal or correlated with other established risk factors for the disease. For example, clonal hematopoiesis (i.e at least 4% of the blood being clonally derived), is a risk factor for cardiovascular disease with a similar hazard ratio to common clinical risk factors like high blood pressure and smoking (reviewed in (Jaiswal and Ebert, 2019)). The same may be true for IBD and psoriasis, where the fraction of IL-17 or

*ZFP36L2* mutant cells may be associated with risk of these diseases but environmental factors and germline background also play a role. That somatic mutations cannot be deterministic can be demonstrated with a thought exercise. Consider the human colon, which is composed of about 15 million colonic crypts. In Chapter 3, I estimated that each crypt accumulates 40 substitutions and 1 indel per year of life (Lee-Six et al., 2019; Olafsson et al., 2020), meaning that by a person's 25th birthday, $1.5 \times 10^{10}$ substitutions and $1.5 \times 10^{7}$ indels will have been fixed in at least one crypt in their colon. Acknowledging that somatic mutations are not uniformly distributed across the genome, every gene should still be hit with a truncating mutation in at least one crypt in every individual. This means that everyone carries a truncating mutation in *APC* and *PIGR* and yet colorectal cancer and IBD are, thankfully, rare outcomes.

### 5.2.3 Disease-expansion feedback loops

Until now, my discussion of the causal relationship between diseases and mutations has assumed a single causal direction. However, it is also possible that a clone that contributes to the disease process also has a survival advantage under the disease condition. The disease then drives the expansion of the pathogenic clone which in turns facilitates the continuation of symptoms. Heyde et al showed how atherosclerosis-associated factors increase the proliferation of hematopoietic stem cells, driving clonal expansions both through drift and selection (Heyde et al., 2021). The pro-inflammatory effects of some of the drivers and elevated levels of myeloid cells then beget more atherosclerosis and so on.

A disease-expansion feedback loop may also exist in IBD. I have described above how conditional *il-17* knockout mice develop dysbiosis and autoimmunity. Nanki et al showed how IL-17A elicits a pro-apoptotic response in organoids derived from normal colonic mucosa but organoids that were knock-outs in any of *NFKBIZ*, *IL17RA* or *TRAF3IP2*, or carried a gain of function mutation in *ZC3H12A*, were protected against the effect. This suggests a model where cells carrying mutations in these or other IL-17-related genes drive dysbiosis and inflammation. Th17 cells are recruited to the site which locally secrete IL-17A, killing off wild-type crypts and enabling the mutant crypts, which are resistant to IL-17A, to take their place within the tissue.

### 5.2.4 Somatic mutations as Nature's gene therapy

In some cases, somatic mutations may confer upon cells resistance to the effects of disease. The expansion of disease-resistant cells within a tissue could potentially restore some of the

tissue function. The existence of mutations restorative to tissue function has been proposed in chronic liver disease (Zhu et al., 2019) but remains a theoretical prediction for most other complex diseases. However, they have been reported in several germline conditions in which the molecular effects of pathogenic germline variants are rescued by somatic variants conferring selective advantage on the revertant cells (Lai-Cheong et al., 2011). Those observations also further underline the importance of parallel evolution. Multiple revertant clones evolving in parallel have, for example, been reported in patients with Wiskott–Aldrich syndrome (Boztug et al., 2008), ichthyosis (Choate et al., 2010; Gudmundsson et al., 2017) and in various types of genodermatoses (Jonkman et al., 1997; Pasmooij et al., 2005; Suzuki et al., 2019).

Identifying and pharmacologically mimicking mutations driving the positive selection of disease-resistant clones could restore tissue homeostasis and alleviate disease symptoms. For such efforts to be successful it will be important to establish the causal relationship between variant and disease, and rule out disease-expansion feedback loops where the mutations that facilitate cell survival also contribute to the disease process.

## 5.3 Integrating germline and somatic variants for the study of complex traits

Many of the studies of somatic evolution in normal tissues published to date have uncovered a marked heterogeneity in somatic evolution between individuals. Lawson et al formally showed that in the urothelium, individuals exhibit "driver preference" where parallel evolution leads to a high fraction of cells carrying mutations in some driver genes but not others. For example, Lawson et al found one patient with 35 distinct mutations in *KDM6A* and only two in *ARID1A* while a second patient carried 20 different *ARID1A* mutations but only 4 *KDM6A* mutations (Lawson et al., 2020). Similar, but less pronounced, examples of parallel evolution have been described in other tissues, but the sample sizes of the respective studies have been insufficient to identify the factors driving this heterogeneity. The observed driver preference is likely driven in part by the environment and in part by the germline background of the individual. The joint study of germline variants and somatic mutations might reveal germline variants that affect the selection coefficients of particular mutations in normal tissues, which would have implications for cancer risk prediction and prevention and help us understand the potential role of somatic mutations in non-neoplastic diseases.

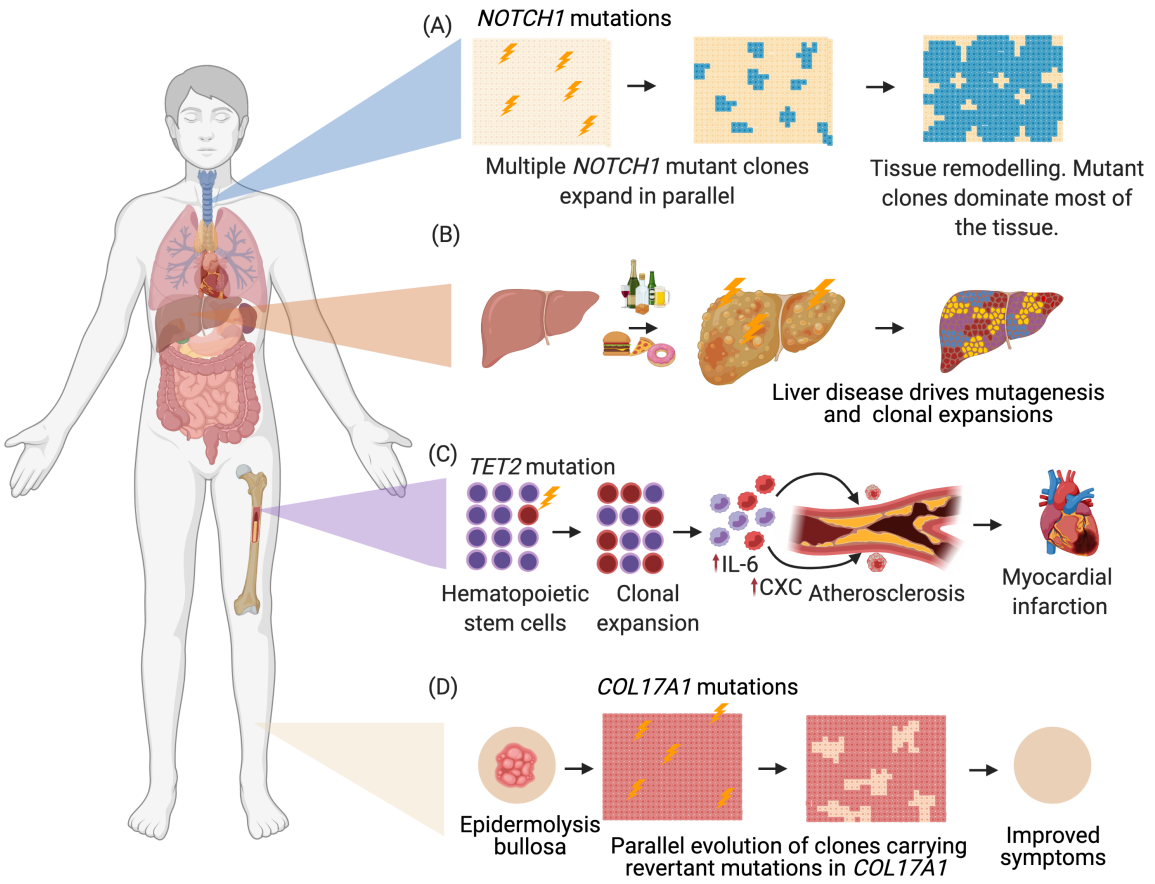



Fig. 5.2 **Somatic evolution in health and disease.** (A) A tissue can be completely remodeled by somatic evolution and remain healthy. In this example, independent clones carrying distinct *NOTCH1* mutations replace nearly every wild type cell of the oesophagus. The tissue retains normal physiological function and cancer growth is not promoted. (B) Disease can change the somatic evolution landscape of a tissue. In this example, the liver, which rarely harbors large clones, becomes diseased as a result of environmental exposures. After disease onset, mutagenesis is increased and clonal growth promoted. (C) A single pathogenic clone may expand to cause a disease. In this example, a mutation in *TET2* in a single blood stem cell leads to clonal hematopoiesis. The initial mutated cell gives rise to a large number of monocytes which infiltrate the artery and differentiate into macrophages. These express high amounts of inflammatory cytokines including interleukin 6, interleukin 1 and chemokines from the CXC family. Macrophages ingest cholesterol and lipids and form foam cells at the lesion, facilitating the formation of atherosclerosis and myocardial infarction. (D) Somatic mutations can be restorative to tissue function. In this example the patient suffers from an autosomal recessive disease of the skin characterized by germline variants in *COL17A1*. Somatic mutations in this gene restore the function of the tissue and mutant clones outcompete the germline "wildtype", resulting in symptom improvement. Figure created with BioRender.com.

### 5.3.1 Genome-wide association studies of somatic evolution

Genome wide association studies (GWAS) have identified thousands of associations between common germline variants and complex diseases, including cancers. The identified variants tend to have small effect sizes and due to the high burden of multiple testing in GWASs, tens of thousands of participants are often needed to reach statistical significance. Part of the reason for the small effect sizes is that diseases are complex, composite phenotypes and multiple causal factors can lead to the same outcome. The closer one gets to a cellular phenotype however, the larger effect sizes tend to become, which is why studies of expression-quantitative trait loci (eQTLs) for example, can be well powered at much smaller sample sizes. It is unclear whether we should expect small effect sizes for somatic evolution variables, similar to those observed for complex traits, or larger effect sizes, similar to those observed for cellular phenotypes. There may also be some variability between tissues, as discussed below.

The availability of hundreds of thousands of blood samples from biobanks has enabled the discovery of 156 germline variants associated with somatic loss of chromosome Y in men (Thompson et al., 2019; Wright et al., 2017) and of 10-20 variants associated with the likelihood of clonal hematopoiesis (Bick et al., 2020; Hinds et al., 2016; Loh et al., 2018; Zink et al., 2017). The variants associated with ChrY loss have odds ratios (OR) ranging from 1.03 to 2.02 (Thompson et al., 2019) and similar effects are observed for the common variants associated with clonal hematopoiesis (Bick et al., 2020; Hinds et al., 2016; Loh et al., 2018; Zink et al., 2017). Most fall within non-coding regions of the genome. These observations suggest that clonal evolution phenotypes are complex in nature and have a genetic architecture similar to that of complex quantitative traits. Rare chromosomal alterations and loss-of-heterozygosity events (frequency <0.05%) have also been reported that have much larger effects on clonal hematopoiesis (ORs=18-698) (Loh et al., 2018, 2020; Terao et al., 2020).

Interactions between germline variants and known somatic drivers in solid tissues have been most systematically studied in the context of cancer. A study of The Cancer Genome Atlas (TCGA) dataset (Carter et al., 2017), which was only powered to detect large effects (1.8-3 fold increase in mutation burden, depending on the frequency with which genes are mutated in cancers), reported 17 associations between common germline variants and the frequency of somatic mutations in known cancer genes. The effect sizes reported range from 1.8 to 14.8 fold increase in mutation frequency. These are much larger effect sizes than those reported for clonal hematopoiesis but the reasons for the discrepancy are not clear.

It is possible that germline effects on selection pressure are higher in solid tissues than in blood, where the admixture of cells is greatest. Tissue architecture and driver prevalence can influence clonal spread. For example, the glandular structure of the colon seems to curb clonal spread, (Lee-Six et al., 2019) while in the flat structure of the oesophagus, where the density of driver mutations is high, the spread of any one clone is constrained by its collision with neighbours of similar fitness (Colom et al., 2020). How such variables influence germline effect sizes is unknown. It is also possible the effects of germline variants are larger in cancers than in evolution of non-neoplastic tissues.

### 5.3.2 Germline modulators of selection in psoriatic skin

Once all samples have been sequenced, the study of psoriatic skin described in Chapter 4 will comprise samples from 109 tissue donors. This large sample size may enable the identification of factors that influence the somatic evolution of cells in the skin. These include some lifestyle and environmental factors like smoking, alcohol intake and BMI, drugs or treatments the patients have received, and also germline variants.

To enable us to look for germline modulators of somatic evolution (fraction of mutated cells in particular), all tissue donors will be genotyped on an Illumina GSA-MD v3.0 genotyping array. With 109 tissue donors, it is unlikely that a GWAS would be powered to detect associations between germline variants and driver prevalence. Instead, I mean to carry out two analyses which focus on variation in or near genes found to be under selection in the skin: The first uses a two-tiered approach similar to an eQTL study (Figure 5.3). I will first carry out an unbiased dN/dS analysis, exome wide, to identify genes that are under selection in psoriatic skin. For each gene under selection, I will search for germline variants that associate with the fraction of cells carrying a somatic mutation in that gene by testing only germline variants within a 1 Mb window of the transcriptional start site. Many variants within that window will be in tight linkage disequilibrium and in order to determine at what threshold associations should be considered significant, I will permutate the phenotypes and covariates.

The second analysis uses a design previously employed to show that non-coding regulatory variants modify the penetrance of rare coding variants in diseases like cancer and autism (Castel et al., 2018). The idea is that haplotypes that contain germline variants which cause a gene to be upregulated will be responsible for more than one half of the mRNA produced from that gene. In other words, the germline variants create an allelic imbalance when present in a heterozygous state. A mutation of the highly expressed allele should confer

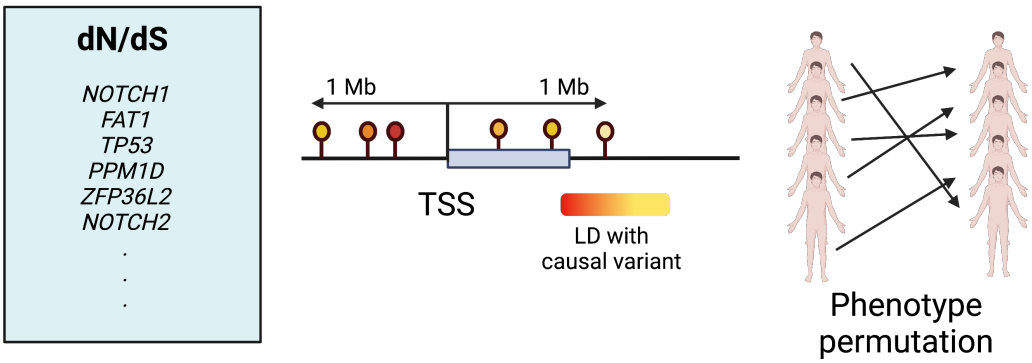



Fig. 5.3 **An eQTL-like study design to identify germline variants that influence the selection of somatic mutations in psoriatic skin.** Figure created with Biorender.com.

a greater selective effect than a mutation of the lower expressed allele, as it is present in a greater fraction of the protein product (Figure 5.4). This would manifest as a departure from the expected 1:1 haplotype ratio in individuals who are heterozygous for the germline variants and information can be pooled across individuals and different germline variants to boost power.

This analysis makes two assumptions. The first is that there are known germline regulatory variants for the gene that have large enough effect sizes in the tissue of interest to cause consequential allelic imbalance. Castel et al limit their analysis to those eQTLs in the GTEx database that have absolute effect sizes in the top 25% of all the eQTLs in GTEx. It remains to be determined if genes enriched in mutations in the skin have eQTLs that fulfil these or similar conditions.

The second assumption is that somatic mutations and germline variants can be phased to determine which variants are on the same haplotype. While germline variants can be phased by comparing them against large external reference panels of haplotypes, no such references exist for somatic mutations and these must be phased based on reads that overlap both somatic mutations and one or more germline polymorphic sites. Unpublished work from the Campbell lab at the Sanger institute suggests that 20-30% of somatic mutations can be phased in this way.

Both of the approaches described above are only able to identify cis-effects on clonal evolution. It is unclear what fraction of variants that modify selection are likely to act in cis vs trans. Cis variants likely play some role in shaping positive selection. For example, there is evidence that germline variants in or near *JAK2* and *TET2* are associated with clonal

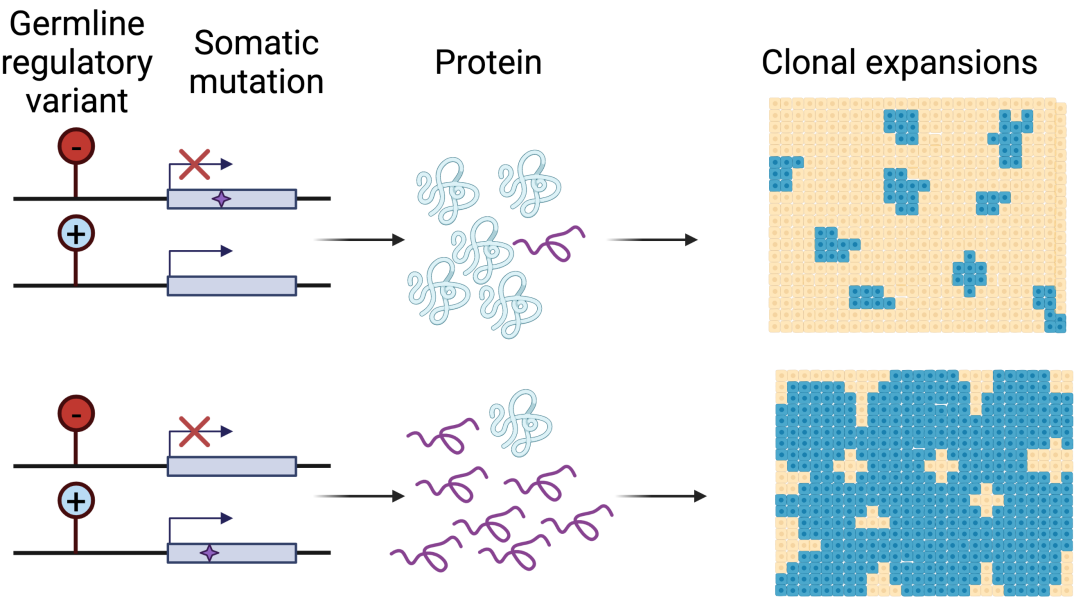



Fig. 5.4 **Regulatory germline variants that affect the selection coefficients of somatic mutations.** An individual is heterozygous for a germline regulatory variant which causes allelic imbalance in the protein product. If the somatic mutation occurs on the "lowly expressed" haplotype then it's effect is small. In contrast, if the somatic mutation occurs on the highly expressed haplotype we expect the selection coefficient to be larger, manifesting as more widespread clonal expansion of mutant cells.

hematopoiesis (Bick et al., 2020). In IBD, *NFKBIZ* is both positively selected in inflamed epithelial tissue and a likely causal gene in a GWAS locus for the disease (de Lange et al., 2017), although it is not clear if the IBD associated germline variant affects selection of *NFKBIZ* mutant cells. However, most variants associated with clonal hematopoiesis lie outside of known driver genes and variants associated with somatic ChrY loss are scattered all across the genome. Furthermore, of the 17 associations discovered in the TCGA dataset mentioned above, the leading variant was never within 1Mb of the gene whose mutation burden it was associated with (Carter et al., 2017). It seems likely that only a fraction of germline variants that influence selection will act in cis. To identify variants acting in trans, larger sample sizes will be needed.

### 5.3.3 Scaling studies of somatic evolution in solid tissues

To build more complete genetic maps of somatic evolution that include trans-acting associations, studies must be scaled up to include hundreds or thousands of tissue donors. The methods discussed in Chapter 1 and summarized in Figure 1.7, organoid culture, LCM and single cell sequencing are labour intensive and expensive and will be difficult to scale to the sample sizes needed unless a much greater level of automation can be achieved.

Studies of normal tissues have relied on samples donated post-mortem, surgical resections or on the willingness of living patients to undergo, or extend, invasive procedures in order to donate biopsies for research. High-throughput and less invasive sampling methods may need to be developed and/or methods that avoid artifacts associated with formalin fixation, as these would enable the repurposing of existing clinical biobanks for research. One exciting prospect is the recently published NanoSeq method (Abascal et al., 2021), which enables mutation calling from a single molecule of DNA and can therefore be applied to a polyclonal sample of cells. NanoSeq is a duplex consensus sequencing method where individual molecules of DNA are barcoded and copies of both strands of the molecule are sequenced multiple times. This makes it possible to distinguish between true mutations and sequencing errors, which are present only on individual reads, and PCR errors and formalin-associated adducts, which are present on only one of the two strands of DNA. NanoSeq might for example enable the study of somatic mutations from swabs of buccal, colon or vaginal swabs, or from biofluids like urine or even cerebrospinal fluid. The limitations of the method include that it requires flow sorting of cells and, because it uses restriction enzyme fragmentation, it only covers 29% of the genome, making it unsuitable for driver mutation identification (Abascal et al., 2021). However, targeted NanoSeq is currently under development at the Sanger Institute and will give exome-wide coverage and enable driver mutation identification.

### 5.3.4 Mendelian randomization

If somatic mutations are to inform complex disease drug target identification then the causal relationship between variant and disease must be established. Mouse models have been useful in the studies of clonal hematopoiesis, epilepsy ichthyosis and IBD (Fuster et al., 2017; Jaiswal et al., 2017; Johansen et al., 1999; Kumar et al., 2016; Lim et al., 2015; Sano et al., 2018; Zhao et al., 2019), but evidence of causation in humans is still outstanding.

Mendelian randomization (MR) has become widely used in the human genetics field for causal inference in epidemiological studies (Davey Smith and Hemani, 2014; Lawlor et al., 2008) and could potentially be used for causal inference of somatic mutations if germline variants that affect their selection can be identified. Assuming that mutations in a given gene drive clonal expansions in a given tissue causing disease, then any variable that influences the propensity of clonal expansion in that tissue should also be associated with the disease. A causal relationship between the clonal expansion and the disease can be inferred if an 'instrument', a variable reliably associated with clonal expansion in a known direction, can be identified. Genetic variants are excellent choices for such instruments because associations between germline variants and human traits represent causal relationships since alleles segregate randomly during meiosis and reverse causation is not possible (disease cannot 'cause' a germline variant). MR, and specifically bidirectional MR, could theoretically be applied to distinguish somatic variants with a causal effect on complex disease from those that are a consequence of disease. If the expansion of clones carrying mutations in a specific gene causes a disease, then genetic variants associated with the propensity for clonal growth will also be associated with the disease (assuming that this association occurs only through the effect on clonal growth).

In addition to the conventional assumptions and limitations of MR, which I will not discuss in detail here (Davey Smith and Hemani, 2014; Lawlor et al., 2008), the assumption of a single causal direction merits special consideration when studying somatic evolution. Disease-expansion feedback loops, where disease drives clonal growth, which in turn perpetuates the disease and so on, have been proposed for clonal hematopoiesis and atherosclerosis (Heyde et al., 2021), as mentioned above, and may also exist between IBD and IL-17 mutant clones. It has been suggested that these can be modeled in a structural equation modeling framework (Evans and Davey Smith, 2015) but, as far as I am aware, these methods await development. Furthermore, care must be taken when interpreting associations between germline variants and somatic mutation frequencies. Cis-associations need not represent

causal relationships because germline haplotypes can be in linkage disequilibrium with fragile sites of the genome and other mutational hotspots.

## 5.4 Population differences in somatic evolution

There is evidence that somatic evolution varies across different ancestries. In a study of normal skin from just four individuals, the one donor of South Asian ancestry seemed to exhibit a different evolutionary landscape from the other three (Martincorena et al., 2015). Similarly, the selection of *NFKBIZ*, *ZC3H12A* and *PIGR* mutant clones seems more pronounced in IBD patients of Japanese ancestry than Europeans (Kakiuchi et al., 2020; Olafsson et al., 2020), as discussed in Chapter 3. Furthermore, white Europeans are nearly twice as likely as Hispanics and East Asians to develop clonal hematopoiesis after adjusting for age (Bick et al., 2020). European individuals also have a 2-6 fold increase in the mutation rate of specific sites linked with clonal hematopoiesis compared to Japanese individuals, a finding which helps explain the different rates of B and T-cell cancers in these populations (Terao et al., 2020). While it is unclear to what extent these population-specific differences are due to environmental or germline differences, they highlight the need to study somatic evolution in diverse populations.

## 5.5 Final remarks

Darwin didn't know about the molecular mechanisms behind natural selection and may never have thought about evolution of cells within the body of an individual. It is tempting to speculate that he would have been fascinated by the parallels between individuals of a species and individual cells that cooperate to make up a body. That he would have approved of this view of life as seen through the lens of somatic evolution. Of a single cell giving rise to a mosaic of billions of microscopic clones, each different from the next, that make up a body. Of the continuous struggle for existence every cell must face, be it a unicellular organism or a cell that is a miniscule part of a much larger whole. Of the ruthlessness of these natural forces which promote the expansion of an individual cell even to the detriment, disease or death of the organism. There is grandeur in this view, even of the smallest units of life.

# References


Abascal, F., Harvey, L. M. R., Mitchell, E., Lawson, A. R. J., Lensing, S. V., Ellis, P., Russell, A. J. C., Alcantara, R. E., Baez-Ortega, A., Wang, Y., Kwa, E. J., Lee-Six, H., Cagan, A., Coorens, T. H. H., Chapman, M. S., Olafsson, S., Leonard, S., Jones, D., Machado, H. E., Davies, M., Øbro, N. F., Mahubani, K. T., Allinson, K., Gerstung, M., Saeb-Parsy, K., Kent, D. G., Laurenti, E., Stratton, M. R., Rahbari, R., Campbell, P. J., Osborne, R. J., Martincorena, I., 2021. Somatic mutation landscapes at single-molecule resolution. Nature 593(7859), 405–410.

Abby, E., Dentro, S. C., Hall, M. W. J., Fowler, J. C., Ong, S. H., Sood, R., Siebel, C. W., Gerstung, M., Hall, B. A., Jones, P. H., 2021. Notch1 mutation drives clonal expansion in normal esophageal epithelium but impairs tumor growth.

Abyzov, A., Tomasini, L., Zhou, B., Vasmatzis, N., Coppola, G., Amenduni, M., Pattni, R., Wilson, M., Gerstein, M., Weissman, S., Urban, A. E., Vaccarino, F. M., 2017. One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. Genome Res. 27(4), 512–523.

Adami, H.-O., Bretthauer, M., Emilsson, L., Hernán, M. A., Kalager, M., Ludvigsson, J. F., Ekbom, A., 2016. The continuing uncertainty about cancer risk in inflammatory bowel disease. Gut 65(6), 889–893.

Agrawal, N., Frederick, M. J., Pickering, C. R., Bettegowda, C., Chang, K., Li, R. J., Fakhry, C., Xie, T.-X., Zhang, J., Wang, J., Zhang, N., El-Naggar, A. K., Jasser, S. A., Weinstein, J. N., Treviño, L., Drummond, J. A., Muzny, D. M., Wu, Y., Wood, L. D., Hruban, R. H., Westra, W. H., Koch, W. M., Califano, J. A., Gibbs, R. A., Sidransky, D., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., Wheeler, D. A., Kinzler, K. W., Myers, J. N., 2011. Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. Science 333(6046), 1154–1157.

Aitken, S. J., Anderson, C. J., Connor, F., Pich, O., Sundaram, V., Feig, C., Rayner, T. F., Lukk, M., Aitken, S., Luft, J., Kentepozidou, E., Arnedo-Pac, C., Beentjes, S. V., Davies, S. E., Drews, R. M., Ewing, A., Kaiser, V. B., Khamseh, A., López-Arribillaga, E., Redmond, A. M., Santoyo-Lopez, J., Sentís, I., Talmane, L., Yates, A. D., Liver Cancer Evolution Consortium, Semple, C. A., López-Bigas, N., Flicek, P., Odom, D. T., Taylor, M. S., 2020. Pervasive lesion segregation shapes cancer genome evolution. Nature 583(7815), 265–270.

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(12), 1402–1407.

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, Getz, G., Rozen, S. G., Stratton, M. R., PCAWG Consortium, 2020. The repertoire of mutational signatures in human cancer. Nature 578(7793), 94–101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Australian Pancreatic Cancer Genome Initiative, ICGC Breast Cancer Consortium, ICGC MMML-Seq Consortium, ICGC PedBrain, Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., Stratton, M. R., 2013a. Signatures of mutational processes in human cancer. Nature 500(7463), 415–421.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., Stratton, M. R., 2013b. Deciphering signatures of mutational processes operative in human cancer. Cell Rep. 3(1), 246–259.

Ananthakrishnan, A. N., Bernstein, C. N., Iliopoulos, D., Macpherson, A., Neurath, M. F., Ali, R. A. R., Vavricka, S. R., Fiocchi, C., 2017. Environmental triggers in IBD: a review of progress and evidence. Nat. Rev. Gastroenterol. Hepatol. 15(1), 39–49.

Andrianne, M., Assabban, A., La, C., Mogilenko, D., Salle, D. S., Fleury, S., Doumont, G., Van Simaeys, G., Nedospasov, S. A., Blackshear, P. J., Dombrowicz, D., Goriely, S., Van Maele, L., 2017. Tristetraprolin expression by keratinocytes controls local and systemic inflammation. JCI Insight 2(11).

Arthur, J. C., Perez-Chanona, E., Mühlbauer, M., Tomkovich, S., Uronis, J. M., Fan, T.-J., Campbell, B. J., Abujamel, T., Dogan, B., Rogers, A. B., Rhodes, J. M., Stintzi, A., Simpson, K. W., Hansen, J. J., Keku, T. O., Fodor, A. A., Jobin, C., 2012. Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 338(6103), 120–123.

Baker, A.-M., Cereser, B., Melton, S., Fletcher, A. G., Rodriguez-Justo, M., Tadrous, P. J., Humphries, A., Elia, G., McDonald, S. A. C., 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(4), 940–947.

Baker, A.-M., Cross, W., Curtius, K., Bakir, I. A., Choi, C.-H. R., Davis, H. L., Temko, D., Biswas, S., Martinez, P., Williams, M. J., Lindsay, J. O., Feakins, R., Vega, R., Hayes, S. J., Tomlinson, I. P. M., McDonald, S. A. C., Moorghen, M., Silver, A., East, J. E., Wright, N. A., Wang, L. M., Rodriguez-Justo, M., Jansen, M., Hart, A. L., Leedham, S. J., Graham,

T. A., 2018. Evolutionary history of human colitis-associated colorectal cancer. Gut pp. gutjnl–2018–316191.

Baker, A.-M., Gabbutt, C., Williams, M. J., Cereser, B., Jawad, N., Rodriguez-Justo, M., Jansen, M., Barnes, C. P., Simons, B. D., McDonald, S. A., Graham, T. A., Wright, N. A., 2019. Crypt fusion as a homeostatic mechanism in the human colon. Gut 68(11), 1986–1993.

Barker, N., van Es, J. H., Kuipers, J., Kujala, P., van den Born, M., Cozijnsen, M., Haegebarth, A., Korving, J., Begthel, H., Peters, P. J., Clevers, H., 2007. Identification of stem cells in small intestine and colon by marker gene lgr5. Nature 449(7165), 1003–1007.

Beaugerie, L., Itzkowitz, S. H., 2015. Cancers complicating inflammatory bowel disease. N. Engl. J. Med. 372(15), 1441–1452.

Beaumont, M. A., Zhang, W., Balding, D. J., 2002. Approximate bayesian computation in population genetics. Genetics 162(4), 2025–2035.

Bell, R. J. A., Rube, H. T., Xavier-Magalhães, A., Costa, B. M., Mancini, A., Song, J. S., Costello, J. F., 2016. Understanding TERT promoter mutations: A common path to immortality. Mol. Cancer Res. 14(4), 315–323.

Bertorelle, G., Benazzo, A., Mona, S., 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol. Ecol. 19(13), 2609–2625.

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., Canaider, S., 2013. An estimation of the number of cells in the human body. Ann. Hum. Biol. 40(6), 463–471.

Bick, A. G., Weinstock, J. S., Nandakumar, S. K., Fulco, C. P., Bao, E. L., Zekavat, S. M., Szeto, M. D., Liao, X., Leventhal, M. J., Nasser, J., Chang, K., Laurie, C., Burugula, B. B., Gibson, C. J., Lin, A. E., Taub, M. A., Aguet, F., Ardlie, K., Mitchell, B. D., Barnes, K. C., Moscati, A., Fornage, M., Redline, S., Psaty, B. M., Silverman, E. K., Weiss, S. T., Palmer, N. D., Vasan, R. S., Burchard, E. G., Kardia, S. L. R., He, J., Kaplan, R. C., Smith, N. L., Arnett, D. K., Schwartz, D. A., Correa, A., de Andrade, M., Guo, X., Konkle, B. A., Custer, B., Peralta, J. M., Gui, H., Meyers, D. A., McGarvey, S. T., Chen, I. Y.-D., Shoemaker, M. B., Peyser, P. A., Broome, J. G., Gogarten, S. M., Wang, F. F., Wong, Q., Montasser, M. E., Daya, M., Kenny, E. E., North, K. E., Launer, L. J., Cade, B. E., Bis, J. C., Cho, M. H., Lasky-Su, J., Bowden, D. W., Cupples, L. A., Mak, A. C. Y., Becker, L. C., Smith, J. A., Kelly, T. N., Aslibekyan, S., Heckbert, S. R., Tiwari, H. K., Yang, I. V., Heit, J. A., Lubitz, S. A., Johnsen, J. M., Curran, J. E., Wenzel, S. E., Weeks, D. E., Rao, D. C., Darbar, D., Moon, J.-Y., Tracy, R. P., Buth, E. J., Rafaels, N., Loos, R. J. F., Durda, P., Liu, Y., Hou, L., Lee, J., Kachroo, P., Freedman, B. I., Levy, D., Bielak, L. F., Hixson, J. E., Floyd, J. S., Whitsel, E. A., Ellinor, P. T., Irvin, M. R., Fingerlin, T. E., Raffield, L. M., Armasu, S. M., Wheeler, M. M., Sabino, E. C., Blangero, J., Williams, L. K., Levy, B. D., Sheu, W. H.-H., Roden, D. M., Boerwinkle, E., Manson, J. E., Mathias, R. A., Desai, P., Taylor, K. D., Johnson, A. D., NHLBI Trans-Omics for Precision Medicine Consortium, Auer, P. L., Kooperberg, C., Laurie, C. C., Blackwell, T. W., Smith, A. V., Zhao, H., Lange, E., Lange, L., Rich, S. S., Rotter, J. I., Wilson, J. G., Scheet, P., Kitzman, J. O., Lander, E. S., Engreitz, J. M., Ebert, B. L., Reiner, A. P., Jaiswal, S., Abecasis, G., Sankaran, V. G.,

Kathiresan, S., Natarajan, P., 2020. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. Nature 586(7831), 763–768.

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. S., Verstegen, M. M. A., van der Laan, L. J. W., de Jonge, J., IJzermans, J. N. M., 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(7624), 260–264.

Blum, M. G. B., François, O., 2010. Non-linear regression models for approximate bayesian computation. Stat. Comput. 20(1), 63–73.

Boehncke, W.-H., Schön, M. P., 2015. Psoriasis. Lancet 386(9997), 983–994.

Bonilla, X., Parmentier, L., King, B., Bezrukov, F., Kaya, G., Zoete, V., Seplyarskiy, V. B., Sharpe, H. J., McKee, T., Letourneau, A., Ribaux, P. G., Popadin, K., Basset-Seguin, N., Ben Chaabene, R., Santoni, F. A., Andrianova, M. A., Guipponi, M., Garieri, M., Verdan, C., Grosdemange, K., Sumara, O., Eilers, M., Aifantis, I., Michielin, O., de Sauvage, F. J., Antonarakis, S. E., Nikolaev, S. I., 2016. Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. Nat. Genet. 48(4), 398–406.

Boztug, K., Germeshausen, M., Avedillo Díez, I., Gulacsy, V., Diestelhorst, J., Ballmaier, M., Welte, K., Maródi, L., Chernyshova, L., Klein, C., 2008. Multiple independent second-site mutations in two siblings with somatic mosaicism for Wiskott-Aldrich syndrome. Clin. Genet. 74(1), 68–74.

Brooks, S. A., Blackshear, P. J., 2013. Tristetraprolin (TTP): interactions with mRNA and proteins, and current thoughts on mechanisms of action. Biochim. Biophys. Acta 1829(6-7), 666–679.

Bruens, L., Ellenbroek, S. I. J., van Rheenen, J., Snippert, H. J., 2017. In vivo imaging reveals existence of crypt fission and fusion in adult mouse intestine. Gastroenterology 153(3), 674–677.e3.

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., Sanders, M. A., Ellis, P., Alder, C., Hooks, Y., Abascal, F., Stratton, M. R., Martincorena, I., Hoare, M., Campbell, P. J., 2019. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature 574(7779), 538–542.

Carballo, E., Lai, W. S., Blackshear, P. J., 1998. Feedback inhibition of macrophage tumor necrosis factor-alpha production by tristetraprolin. Science 281(5379), 1001–1005.

Carter, H., Marty, R., Hofree, M., Gross, A. M., Jensen, J., Fisch, K. M., Wu, X., DeBoever, C., Van Nostrand, E. L., Song, Y., Wheeler, E., Kreisberg, J. F., Lippman, S. M., Yeo, G. W., Gutkind, J. S., Ideker, T., 2017. Interaction landscape of inherited polymorphisms with somatic events in cancer. Cancer Discov. 7(4), 410–423.

Castel, S. E., Cervera, A., Mohammadi, P., Aguet, F., Reverter, F., Wolman, A., Guigo, R., Iossifov, I., Vasileva, A., Lappalainen, T., 2018. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat. Genet. 50(9), 1327–1334.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., GTEx Consortium, Montgomery, S. B., Battle, A., Conrad, D. F., Hall, I. M., 2017. The impact of structural variation on human gene expression. Nat. Genet. 49(5), 692–699.

Chiesa Fuxench, Z. C., Shin, D. B., Ogdie Beatty, A., Gelfand, J. M., 2016. The risk of cancer in patients with psoriasis: A Population-Based cohort study in the health improvement network. JAMA Dermatol. 152(3), 282–290.

Choate, K. A., Lu, Y., Zhou, J., Choi, M., Elias, P. M., Farhi, A., Nelson-Williams, C., Crumrine, D., Williams, M. L., Nopper, A. J., Bree, A., Milstone, L. M., Lifton, R. P., 2010. Mitotic recombination in patients with ichthyosis causes reversion of dominant mutations in KRT10. Science 330(6000), 94–97.

Choi, C.-H. R., Bakir, I. A., Hart, A. L., Graham, T. A., 2017. Clonal evolution of colorectal cancer in IBD. Nat. Rev. Gastroenterol. Hepatol. 14(4), 218–229.

Choi, C.-H. R., Rutter, M. D., Askari, A., Lee, G. H., Warusavitarne, J., Moorghen, M., Thomas-Gibson, S., Saunders, B. P., Graham, T. A., Hart, A. L., 2015. Forty-Year analysis of colonoscopic surveillance program for neoplasia in ulcerative colitis: An updated overview. Am. J. Gastroenterol. 110(7), 1022–1034.

Colom, B., Alcolea, M. P., Piedrafita, G., Hall, M. W. J., Wabik, A., Dentro, S. C., Fowler, J. C., Herms, A., King, C., Ong, S. H., Sood, R. K., Gerstung, M., Martincorena, I., Hall, B. A., Jones, P. H., 2020. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. Nat. Genet. .

Colom, B., Herms, A., Hall, M. W. J., Dentro, S. C., King, C., Sood, R. K., Alcolea, M. P., Piedrafita, G., Fernandez-Antoran, D., Ong, S. H., Fowler, J. C., Mahbubani, K. T., Saeb-Parsy, K., Gerstung, M., Hall, B. A., Jones, P. H., 2021. Precancer: Mutant clones in normal epithelium outcompete and eliminate esophageal micro-tumors.

Cortés-Ciriano, I., Lee, J. J.-K., Xi, R., Jain, D., Jung, Y. L., Yang, L., Gordenin, D., Klimczak, L. J., Zhang, C.-Z., Pellman, D. S., PCAWG Structural Variation Working Group, Park, P. J., PCAWG Consortium, 2020. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. Nat. Genet. 52(3), 331–341.

Csilléry, K., François, O., Blum, M. G. B., 2012. abc: an R package for approximate bayesian computation (ABC): R package: abc. Methods Ecol. Evol. 3(3), 475–479.

Darwin, C., 1876. The Origin of Species: By Means of Natural Selection, Or the Preservation of Favoured Races in the Struggle for Life. Cambridge University Press.

Davey Smith, G., Hemani, G., 2014. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum. Mol. Genet. 23(R1), R89–98.

de Lange, K. M., Moutsianas, L., Lee, J. C., Lamb, C. A., Luo, Y., Kennedy, N. A., Jostins, L., Rice, D. L., Gutierrez-Achury, J., Ji, S.-G., Heap, G., Nimmo, E. R., Edwards, C., Henderson, P., Mowat, C., Sanderson, J., Satsangi, J., Simmons, A., Wilson, D. C., Tremelling, M., Hart, A., Mathew, C. G., Newman, W. G., Parkes, M., Lees, C. W., Uhlig, H., Hawkey, C., Prescott, N. J., Ahmad, T., Mansfield, J. C., Anderson, C. A., Barrett,

J. C., 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. Nat. Genet. 49(2), 256–261.

Din, S., Wong, K., Mueller, M. F., Oniscu, A., Hewinson, J., Black, C. J., Miller, M. L., Jiménez-Sánchez, A., Rabbie, R., Rashid, M., Satsangi, J., Adams, D. J., Arends, M. J., 2018. Mutational analysis identifies therapeutic biomarkers in inflammatory bowel Disease-Associated colorectal cancers. Clin. Cancer Res. 24(20), 5133–5142.

Dong, X., Zhang, L., Milholland, B., Lee, M., Maslov, A. Y., Wang, T., Vijg, J., 2017. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat. Methods 14(5), 491–493.

Ellis, P., Moore, L., Sanders, M. A., Butler, T. M., Brunner, S. F., Lee-Six, H., Osborne, R., Farr, B., Coorens, T. H. H., Lawson, A. R. J., Cagan, A., Stratton, M. R., Martincorena, I., Campbell, P. J., 2021. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. Nat. Protoc. 16(2), 841–871.

Esposito, F., Brankamp, R. G., Sinden, R. R., 1988. DNA sequence specificity of 4,5',8-trimethylpsoralen cross-linking. effect of neighboring bases on cross-linking the 5'-TA dinucleotide. J. Biol. Chem. 263(23), 11466–11472.

Euvrard, S., Kanitakis, J., Claudy, A., 2003. Skin cancers after organ transplantation. N. Engl. J. Med. 348(17), 1681–1691.

Evans, D. M., Davey Smith, G., 2015. Mendelian randomization: New applications in the coming age of Hypothesis-Free causality. Annu. Rev. Genomics Hum. Genet. 16, 327–350.

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C. D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H., D'Eustachio, P., 2018. The reactome pathway knowledgebase. Nucleic Acids Res. 46(D1), D649–D655.

Fearon, E. R., Vogelstein, B., 1990. A genetic model for colorectal tumorigenesis. Cell 61(5), 759–767.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S., De, T., Campbell, P. J., 2017. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. 45(D1), D777–D783.

Fowler, J. C., King, C., Bryant, C., Hall, M. W. J., Sood, R., Ong, S. H., Earp, E., Fernandez-Antoran, D., Koeppel, J., Dentro, S. C., Shorthouse, D., Durrani, A., Fife, K., Rytina, E., Milne, D., Roshan, A., Mahububani, K., Saeb-Parsy, K., Hall, B. A., Gerstung, M., Jones, P. H., 2021. Selection of oncogenic mutant clones in normal human skin varies with body site. Cancer Discov. 11(2), 340–361.

Franco, I., Johansson, A., Olsson, K., Vrtačnik, P., Lundin, P., Helgadottir, H. T., Larsson, M., Revêchon, 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(1), 800.

Fuster, J. J., MacLauchlan, S., Zuriaga, M. A., Polackal, M. N., Ostriker, A. C., Chakraborty, R., Wu, C.-L., Sano, S., Muralidharan, S., Rius, C., Vuong, J., Jacob, S., Muralidhar, V., Robertson, A. A. B., Cooper, M. A., Andrés, V., Hirschi, K. K., Martin, K. A., Walsh, K., 2017. Clonal hematopoiesis associated with TET2 deficiency accelerates atherosclerosis development in mice. Science 355(6327), 842–847.

Galandiuk, S., Rodriguez-Justo, M., Jeffery, R., Nicholson, A. M., Cheng, Y., Oukrif, D., Elia, G., Leedham, S. J., McDonald, S. A. C., Wright, N. A., Graham, T. A., 2012. Field cancerization in the intestinal epithelium of patients with crohn's ileocolitis. Gastroenterology 142(4), 855–864.e8.

Garg, A. V., Amatya, N., Chen, K., Cruz, J. A., Grover, P., Whibley, N., Conti, H. R., Hernandez Mir, G., Sirakova, T., Childs, E. C., Smithgall, T. E., Biswas, P. S., Kolls, J. K., McGeachy, M. J., Kolattukudy, P. E., Gaffen, S. L., 2015. MCPIP1 endoribonuclease activity negatively regulates Interleukin-17-Mediated signaling and inflammation. Immunity 43(3), 475–487.

Gehart, H., Clevers, H., 2019. Tales from the crypt: new insights into intestinal stem cells. Nat. Rev. Gastroenterol. Hepatol. 16(1), 19–34.

Gelfand, J. M., Shin, D. B., Neimann, A. L., Wang, X., Margolis, D. J., Troxel, A. B., 2006. The risk of lymphoma in patients with psoriasis. J. Invest. Dermatol. 126(10), 2194–2201.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., Mitchell, T. J., Rubanova, Y., Anur, P., Yu, K., Tarabichi, M., Deshwar, A., Wintersinger, J., Kleinheinz, K., Vázquez-García, 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., Ji, Y., Peifer, M., Markowetz, F., Mustonen, V., Yuan, K., Wang, W., Morris, Q. D., PCAWG Evolution & Heterogeneity Working Group, Spellman, P. T., Wedge, D. C., Van Loo, P., PCAWG Consortium, 2020. The evolutionary history of 2,658 cancers. Nature 578(7793), 122–128.

Gerstung, M., Papaemmanuil, E., Campbell, P. J., 2014. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics 30(9), 1198–1204.

Ghezraoui, H., Piganeau, M., Renouf, B., Renaud, J.-B., Sallmyr, A., Ruis, B., Oh, S., Tomkinson, A. E., Hendrickson, E. A., Giovannangeli, C., Jasin, M., Brunet, E., 2014. Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. Mol. Cell 55(6), 829–842.

Gillen, C. D., Walmsley, R. S., Prior, P., Andrews, H. A., Allan, R. N., 1994. Ulcerative colitis and crohn's disease: a comparison of the colorectal cancer risk in extensive colitis. Gut 35(11), 1590–1592.

Goyette, P., Boucher, G., Mallon, D., Ellinghaus, E., Jostins, L., Huang, H., Ripke, S., Gusareva, E. S., Annese, V., Hauser, S. L., Oksenberg, J. R., Thomsen, I., Leslie, S., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Anderson, C. A., Andrews, J. M., Annese, V., Aumais, G., Baidoo, L., Baldassano, R. N., Balschun, T., Bampton, P. A., Barclay, M., Barrett, J. C., Bayless, T. M., Bethge, J., Bis, J. C., Bitton, A., Boucher, G., Brand, S., Brant, S. R., Büning, C., Chew, A., Cho,

J. H., Cleynen, I., Cohain, A., Croft, A., Daly, M. J., D'Amato, M., Danese, S., De Jong, D., De Vos, M., Denapiene, G., Denson, L. A., Devaney, K. L., Dewit, O., D'Inca, R., Dubinsky, M., Duerr, R. H., Edwards, C., Ellinghaus, D., Essers, J., Ferguson, L. R., Festen, E. A., Fleshner, P., Florin, T., Franchimont, D., Franke, A., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Goyette, P., Green, T., Griffiths, A. M., Guthery, S. L., Hakonarson, H., Halfvarson, J., Hanigan, K., Haritunians, T., Hart, A., Hawkey, C., Hayward, N. K., Hedl, M., Henderson, P., Hu, X., Huang, H., Hui, K. Y., Imielinski, M., Ippoliti, A., Jonaitis, L., Jostins, L., Karlsen, T. H., Kennedy, N. A., Khan, M. A., Kiudelis, G., Kugathasan, S., Kupcinskas, L., Latiano, A., Laukens, D., Lawrance, I. C., Lee, J. C., Lees, C. W., Leja, M., Van Limbergen, J., Lionetti, P., Liu, J. Z., Louis, E., Mahy, G., Mansfield, J., Massey, D., Mathew, C. G., McGovern, D. P. B., Milgrom, R., Mitrovic, M., Montgomery, G. W., Mowat, C., Newman, W., Ng, A., Ng, S. C., Ng, S. M. E., Nikolaus, S., Ning, K., Nöthen, M., Oikonomou, I., Palmieri, O., Parkes, M., Phillips, A., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Proctor, D. D., Radford-Smith, G., Rahier, J.-F., Raychaudhuri, S., Regueiro, M., Rieder, F., Rioux, J. D., Ripke, S., Roberts, R., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schadt, E. E., Schreiber, S., Schumm, L. P., Scott, R., Seielstad, M., Sharma, Y., Silverberg, M. S., Simms, L. A., Skieceviciene, J., Spain, S. L., Steinhart, A. H., Stempak, J. M., Stronati, L., Sventoraityte, J., Targan, S. R., Taylor, K. M., Velde, A. T., Theatre, E., Torkvist, L., Tremelling, M., van der Meulen, A., van Sommeren, S., Vasiliauskas, E., Vermeire, S., Verspaget, H. W., Walters, T., Wang, K., Wang, M.-H., Weersma, R. K., Wei, Z., Whiteman, D., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhang, H., Zhang, W., Zhao, H., Zhao, Z. Z., Daly, M. J., Van Steen, K., Duerr, R. H., Barrett, J. C., McGovern, D. P. B., Schumm, L. P., Traherne, J. A., Carrington, M. N., Kosmoliaptsis, V., Karlsen, T. H., Franke, A., Rioux, J. D., Rioux, J. D., 2015. High-density mapping of the MHC identifies a shared role for HLA-DRB1*01:03 in inflammatory bowel diseases and heterozygous advantage in ulcerative colitis. Nat. Genet. 47(2), 172–179.

Grachtchouk, M., Pero, J., Yang, S. H., Ermilov, A. N., Michael, L. E., Wang, A., Wilbert, D., Patel, R. M., Ferris, J., Diener, J., Allen, M., Lim, S., Syu, L.-J., Verhaegen, M., Dlugosz, A. A., 2011. Basal cell carcinomas in mice arise from hair follicle stem cells and multiple epithelial progenitor populations. J. Clin. Invest. 121(5), 1768–1781.

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. Z., Turnbull, D. M., Wright, N. A., McDonald, S. A. C., 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(3), 714–719.

Greb, J. E., Goldminz, A. M., Elder, J. T., Lebwohl, M. G., Gladman, D. D., Wu, J. J., Mehta, N. N., Finlay, A. Y., Gottlieb, A. B., 2016. Psoriasis. Nat Rev Dis Primers 2, 16082.

Griffiths, C. E., Barker, J. N., 2007. Pathogenesis and clinical features of psoriasis. Lancet 370(9583), 263–271.

Grossmann, S., Hooks, Y., Wilson, L., Moore, L., O'Neill, L., Martincorena, I., Voet, T., Stratton, M. R., Heer, R., Campbell, P. J., 2021. Development, maturation, and maintenance of human prostate inferred from somatic mutations. Cell Stem Cell 28(7), 1262–1274.e5.

Gudjonsson, J. E., Karason, A., Antonsdottir, A., Runarsdottir, E. H., Hauksson, V. B., Upmanyu, R., Gulcher, J., Stefansson, K., Valdimarsson, H., 2003. Psoriasis patients who are homozygous for the HLA-Cw*0602 allele have a 2.5-fold increased risk of developing psoriasis compared with cw6 heterozygotes. Br. J. Dermatol. 148(2), 233–235.

Gudjonsson, J. E., Karason, A., Runarsdottir, E. H., Antonsdottir, A. A., Hauksson, V. B., Jónsson, H. H., Gulcher, J., Stefansson, K., Valdimarsson, H., 2006. Distinct clinical differences between HLA-Cw*0602 positive and negative psoriasis patients–an analysis of 1019 HLA-C- and HLA-B-typed patients. J. Invest. Dermatol. 126(4), 740–745.

Gudmundsson, S., Wilbe, M., Ekvall, S., Ameur, A., Cahill, N., Alexandrov, L. B., Virtanen, M., Hellström Pigg, M., Vahlquist, A., Törmä, H., Bondeson, M.-L., 2017. Revertant mosaicism repairs skin lesions in a patient with keratitis-ichthyosis-deafness syndrome by second-site mutations in connexin 26. Hum. Mol. Genet. 26(6), 1070–1077.

Harwood, C. A., Mesher, D., McGregor, J. M., Mitchell, L., Leedham-Green, M., Raftery, M., Cerio, R., Leigh, I. M., Sasieni, P., Proby, C. M., 2013. A surveillance model for skin cancer in organ transplant recipients: a 22-year prospective study in an ethnically diverse population. Am. J. Transplant 13(1), 119–129.

Heyde, A., Rohde, D., McAlpine, C. S., Zhang, S., Hoyer, F. F., Gerold, J. M., Cheek, D., Iwamoto, Y., Schloss, M. J., Vandoorne, K., Iborra-Egea, O., Muñoz-Guijosa, C., Bayes-Genis, A., Reiter, J. G., Craig, M., Swirski, F. K., Nahrendorf, M., Nowak, M. A., Naxerova, K., 2021. Increased stem cell proliferation in atherosclerosis accelerates clonal hematopoiesis. Cell 184(5), 1348–1361.e22.

Hinds, D. A., Barnholt, K. E., Mesa, R. A., Kiefer, A. K., Do, C. B., Eriksson, N., Mountain, J. L., Francke, U., Tung, J. Y., Nguyen, H. M., Zhang, H., Gojenola, L., Zehnder, J. L., Gotlib, J., 2016. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood 128(8), 1121–1128.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., Vinh, L. S., 2018a. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35(2), 518–522.

Hoang, D. T., Vinh, L. S., Flouri, T., Stamatakis, A., von Haeseler, A., Minh, B. Q., 2018b. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. BMC Evol. Biol. 18(1), 11.

Hodson, D. J., Janas, M. L., Galloway, A., Bell, S. E., Andrews, S., Li, C. M., Pannell, R., Siebel, C. W., MacDonald, H. R., De Keersmaecker, K., Ferrando, A. A., Grutz, G., Turner, M., 2010. Deletion of the RNA-binding proteins ZFP36L1 and ZFP36L2 leads to perturbed thymic development and T lymphoblastic leukemia. Nat. Immunol. 11(8), 717–724.

Holzmann, K., Klump, B., Borchard, F., Hsieh, C. J., Kühn, A., Gaco, V., Gregor, M., Porschen, R., 1998. Comparative analysis of histology, DNA content, p53 and ki-ras mutations in colectomy specimens with long-standing ulcerative colitis. Int. J. Cancer 76(1), 1–6.

Hu, Z., Ding, J., Ma, Z., Sun, R., Seoane, J. A., Scott Shaffer, J., Suarez, C. J., Berghoff, A. S., Cremolini, C., Falcone, A., Loupakis, F., Birner, P., Preusser, M., Lenz, H.-J., Curtis, C., 2019. Quantitative evidence for early metastatic seeding in colorectal cancer. Nat. Genet. 51(7), 1113–1122.

Hueber, W., Sands, B. E., Lewitzky, S., Vandemeulebroecke, M., Reinisch, W., Higgins, P. D. R., Wehkamp, J., Feagan, B. G., Yao, M. D., Karczewski, M., Karczewski, J., Pezous, N., Bek, S., Bruin, G., Mellgard, B., Berger, C., Londei, M., Bertolino, A. P., Tougas, G., Travis, S. P. L., Secukinumab in Crohn's Disease Study Group, 2012. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. Gut 61(12), 1693–1700.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020. Pan-cancer analysis of whole genomes. Nature 578(7793), 82–93.

Inman, G. J., Wang, J., Nagano, A., Alexandrov, L. B., Purdie, K. J., Taylor, R. G., Sherwood, V., Thomson, J., Hogan, S., Spender, L. C., South, A. P., Stratton, M., Chelala, C., Harwood, C. A., Proby, C. M., Leigh, I. M., 2018. The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. Nat. Commun. 9(1), 3667.

Ito, M., Liu, Y., Yang, Z., Nguyen, J., Liang, F., Morris, R. J., Cotsarelis, G., 2005. Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. Nat. Med. 11(12), 1351–1354.

Iwasaki, H., Takeuchi, O., Teraguchi, S., Matsushita, K., Uehata, T., Kuniyoshi, K., Satoh, T., Saitoh, T., Matsushita, M., Standley, D. M., Akira, S., 2011. The IκB kinase complex regulates the stability of cytokine-encoding mRNA induced by TLR-IL-1R by controlling degradation of regnase-1. Nat. Immunol. 12(12), 1167–1175.

Jaiswal, S., Ebert, B. L., 2019. Clonal hematopoiesis in human aging and disease. Science 366(6465).

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(26), 2488–2498.

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(2), 111–121.

Jalonen, U., Nieminen, R., Vuolteenaho, K., Kankaanranta, H., Moilanen, E., 2006. Down-regulation of tristetraprolin expression results in enhanced IL-12 and MIP-2 production and reduced MIP-3alpha synthesis in activated macrophages. Mediators Inflamm. 2006(6), 40691.

Jayaraman, S. S., Rayhan, D. J., Hazany, S., Kolodney, M. S., 2014. Mutational landscape of basal cell carcinomas by whole-exome sequencing. J. Invest. Dermatol. 134(1), 213–220.

Johansen, F.-E., Kaetzel, C. S., 2011. Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. Mucosal Immunol. 4(6), 598–602.

Johansen, F. E., Pekna, M., Norderhaug, I. N., Haneberg, B., Hietala, M. A., Krajci, P., Betsholtz, C., Brandtzaeg, P., 1999. Absence of epithelial immunoglobulin a transport, with increased mucosal leakiness, in polymeric immunoglobulin receptor/secretory component-deficient mice. J. Exp. Med. 190(7), 915–922.

Jones, D., Raine, K. M., Davies, H., Tarpey, P. S., Butler, A. P., Teague, J. W., Nik-Zainal, S., Campbell, P. J., 2016. cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. Curr. Protoc. Bioinformatics 56(1), 15.10.1–15.10.18.

Jonkman, M. F., Scheffer, H., Stulp, R., Pas, H. H., Nijenhuis, M., Heeres, K., Owaribe, K., Pulkkinen, L., Uitto, J., 1997. Revertant mosaicism in epidermolysis bullosa caused by mitotic gene conversion. Cell 88(4), 543–551.

Kahn, J. D., Miller, P. G., Silver, A. J., Sellar, R. S., Bhatt, S., Gibson, C., McConkey, M., Adams, D., Mar, B., Mertins, P., Fereshetian, S., Krug, K., Zhu, H., Letai, A., Carr, S. A., Doench, J., Jaiswal, S., Ebert, B. L., 2018. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. Blood 132(11), 1095–1105.

Kakiuchi, N., Yoshida, K., Uchino, M., Kihara, T., Akaki, K., Inoue, Y., Kawada, K., Nagayama, S., Yokoyama, A., Yamamoto, S., Matsuura, M., Horimatsu, T., Hirano, T., Goto, N., Takeuchi, Y., Ochi, Y., Shiozawa, Y., Kogure, Y., Watatani, Y., Fujii, Y., Kim, S. K., Kon, A., Kataoka, K., Yoshizato, T., Nakagawa, M. M., Yoda, A., Nanya, Y., Makishima, H., Shiraishi, Y., Chiba, K., Tanaka, H., Sanada, M., Sugihara, E., Sato, T.-A., Maruyama, T., Miyoshi, H., Taketo, M. M., Oishi, J., Inagaki, R., Ueda, Y., Okamoto, S., Okajima, H., Sakai, Y., Sakurai, T., Haga, H., Hirota, S., Ikeuchi, H., Nakase, H., Marusawa, H., Chiba, T., Takeuchi, O., Miyano, S., Seno, H., Ogawa, S., 2020. Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. Nature 577(7789), 260–265.

Kleiblova, P., Shaltiel, I. A., Benada, J., Ševčík, J., Pecháčková, S., Pohlreich, P., Voest, E. E., Dundr, P., Bartek, J., Kleibl, Z., Medema, R. H., Macurek, L., 2013. Gain-of-function mutations of PPM1D/Wip1 impair the p53-dependent G1 checkpoint. J. Cell Biol. 201(4), 511–521.

Klein, A. M., Simons, B. D., 2011. Universal patterns of stem cell fate in cycling adult tissues. Development 138(15), 3103–3111.

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(4), 821–836.e16.

Kumar, P., Monin, L., Castillo, P., Elsegeiny, W., Horne, W., Eddens, T., Vikram, A., Good, M., Schoenborn, A. A., Bibby, K., Montelaro, R. C., Metzger, D. W., Gulati, A. S., Kolls, J. K., 2016. Intestinal interleukin-17 receptor signaling mediates reciprocal control of the gut microbiota and autoimmune inflammation. Immunity 44(3), 659–671.

Lai, W. S., Kennington, E. A., Blackshear, P. J., 2003. Tristetraprolin and its family members can promote the cell-free deadenylation of AU-rich element-containing mRNAs by poly(a) ribonuclease. Mol. Cell. Biol. 23(11), 3798–3812.

Lai-Cheong, J. E., McGrath, J. A., Uitto, J., 2011. Revertant mosaicism in skin: natural gene therapy. Trends Mol. Med. 17(3), 140–148.

Lam, A. K.-Y., Chan, S. S.-Y., Leung, M., 2014. Synchronous colorectal cancer: clinical, pathological and molecular implications. World J. Gastroenterol. 20(22), 6815–6820.

Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N., Davey Smith, G., 2008. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. Stat. Med. 27(8), 1133–1163.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., DiCara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., Getz, G., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499(7457), 214–218.

Lawson, A. R. J., Abascal, F., Coorens, T. H. H., Hooks, Y., O'Neill, L., Latimer, C., Raine, K., Sanders, M. A., Warren, A. Y., Mahbubani, K. T. A., Bareham, B., Butler, T. M., Harvey, L. M. R., Cagan, A., Menzies, A., Moore, L., Colquhoun, A. J., Turner, W., Thomas, B., Gnanapragasam, V., Williams, N., Rassl, D. M., Vöhringer, H., Zumalave, S., Nangalia, J., Tubío, J. M. C., Gerstung, M., Saeb-Parsy, K., Stratton, M. R., Campbell, P. J., Mitchell, T. J., Martincorena, I., 2020. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science 370(6512), 75–82.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., 3rd, Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R., Lee, C., Kharchenko, P. V., Park, P. J., Cancer Genome Atlas Research Network, 2012. Landscape of somatic retrotransposition in human cancers. Science 337(6097), 967–971.

Lee-Six, H., Øbro, 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., 2018. Population dynamics of normal human blood inferred from somatic mutations. Nature 561(7724), 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(7779), 532–537.

Leedham, S. J., Graham, T. A., Oukrif, D., McDonald, S. A. C., Rodriguez-Justo, M., Harrison, R. F., Shepherd, N. A., Novelli, M. R., Jankowski, J. A. Z., Wright, N. A., 2009. Clonality, founder mutations, and field cancerization in human ulcerative colitis-associated neoplasia. Gastroenterology 136(2), 542–50.e6.

Leuenberger, C., Wegmann, D., 2010. Bayesian computation and model selection without likelihoods. Genetics 184(1), 243–252.

Li, J., Wang, R., Zhou, X., Wang, W., Gao, S., Mao, Y., Wu, X., Guo, L., Liu, H., Wen, L., Fu, W., Tang, F., 2020a. Genomic and transcriptomic profiling of carcinogenesis in patients with familial adenomatous polyposis. Gut 69(7), 1283–1293.

Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., Kumar, K., Khurana, E., Waszak, S., Korbel, J. O., Haber, J. E., Imielinski, M., PCAWG Structural Variation Working Group, Weischenfeldt, J., Beroukhim, R., Campbell, P. J., PCAWG Consortium, 2020b. Patterns of somatic structural variation in human cancer genomes. Nature 578(7793), 112–121.

Lim, J. S., Kim, W.-I., Kang, H.-C., Kim, S. H., Park, A. H., Park, E. K., Cho, Y.-W., Kim, S., Kim, H. M., Kim, J. A., Kim, J., Rhee, H., Kang, S.-G., Kim, H. D., Kim, D., Kim, D.-S., Lee, J. H., 2015. Brain somatic mutations in MTOR cause focal cortical dysplasia type II leading to intractable epilepsy. Nat. Med. 21(4), 395–400.

Liu, P., Wang, Y., Li, X., 2019. Targeting the untargetable KRAS in cancer therapy. Acta Pharm Sin B 9(5), 871–879.

Loganathan, S. K., Schleicher, K., Malik, A., Quevedo, R., Langille, E., Teng, K., Oh, R. H., Rathod, B., Tsai, R., Samavarchi-Tehrani, P., Pugh, T. J., Gingras, A.-C., Schramek, D., 2020. Rare driver mutations in head and neck squamous cell carcinomas converge on NOTCH signaling. Science 367(6483), 1264–1269.

Loh, P.-R., Genovese, G., Handsaker, R. E., Finucane, H. K., Reshef, Y. A., Palamara, P. F., Birmann, B. M., Talkowski, M. E., Bakhoum, S. F., McCarroll, S. A., Price, A. L., 2018. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature 559(7714), 350–355.

Loh, P.-R., Genovese, G., McCarroll, S. A., 2020. Monogenic and polygenic inheritance become instruments for clonal selection. Nature 584(7819), 136–141.

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(6005), 822–825.

Lowes, M. A., Bowcock, A. M., Krueger, J. G., 2007. Pathogenesis and therapy of psoriasis. Nature 445(7130), 866–873.

Lu, C. P., Polak, L., Rocha, A. S., Pasolli, H. A., Chen, S.-C., Sharma, N., Blanpain, C., Fuchs, E., 2012. Identification of stem cell populations in sweat glands and ducts reveals roles in homeostasis and wound repair. Cell 150(1), 136–150.

Luquette, L. J., Bohrson, C. L., Sherman, M. A., Park, P. J., 2019. Identification of somatic mutations in single cell DNA-seq using a spatial model of allelic imbalance. Nat. Commun. 10(1), 3908.

Lutgens, M. W. M. D., van Oijen, M. G. H., van der Heijden, G. J. M. G., Vleggaar, F. P., Siersema, P. D., Oldenburg, B., 2013. Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies. Inflamm. Bowel Dis. 19(4), 789–799.

Ly, P., Brunner, S. F., Shoshani, O., Kim, D. H., Lan, W., Pyntikova, T., Flanagan, A. M., Behjati, S., Page, D. C., Campbell, P. J., Cleveland, D. W., 2019. Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. Nat. Genet. 51(4), 705–715.

Mahid, S. S., Minor, K. S., Soto, R. E., Hornung, C. A., Galandiuk, S., 2006. Smoking and inflammatory bowel disease: a meta-analysis. Mayo Clin. Proc. 81(11), 1462–1471.

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(6417), 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(5), 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(6237), 880–886.

Martínez-Jiménez, F., Muiños, F., López-Arribillaga, E., Lopez-Bigas, N., Gonzalez-Perez, A., 2020. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. Nature Cancer 1(1), 122–135.

Matsushita, K., Takeuchi, O., Standley, D. M., Kumagai, Y., Kawagoe, T., Miyake, T., Satoh, T., Kato, H., Tsujimura, T., Nakamura, H., Akira, S., 2009. Zc3h12a is an RNase essential for controlling immune responses by regulating mRNA decay. Nature 458(7242), 1185–1190.

Mino, T., Murakawa, Y., Fukao, A., Vandenbon, A., Wessels, H.-H., Ori, D., Uehata, T., Tartey, S., Akira, S., Suzuki, Y., Vinuesa, C. G., Ohler, U., Standley, D. M., Landthaler, M., Fujiwara, T., Takeuchi, O., 2015. Regnase-1 and roquin regulate a common element in inflammatory mRNAs by spatiotemporally distinct mechanisms. Cell 161(5), 1058–1073.

Molodecky, N. A., Soon, I. S., Rabi, D. M., Ghali, W. A., Ferris, M., Chernoff, G., Benchimol, E. I., Panaccione, R., Ghosh, S., Barkema, H. W., Kaplan, G. G., 2012. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 142(1), 46–54.e42.

Mondal, M., Bertranpetit, J., Lao, O., 2019. Approximate bayesian computation with deep learning supports a third archaic introgression in asia and oceania. Nat. Commun. 10(1), 246.

Moody, S., Senkin, S., Islam, S. M. A., Wang, J., Nasrollahzadeh, D., Penha, R. C. C., Fitzgerald, S., Bergstrom, E. N., Atkins, J., He, Y., Khandekar, A., Smith-Byrne, K., Carreira, C., Gaborieau, V., Latimer, C., Thomas, E., Abnizova, I., Bucciarelli, P. E., Jones, D., Teague, J. W., Abedi-Ardekani, B., Serra, S., Scoazec, J.-Y., Saffar, H., Azmoudeh-Ardelan, F., Sotoudeh, M., Nikmanesh, A., Eden, M., Richman, P., Campos, L. S., Fitzgerald, R. C., Ribeiro, L. F., Dzamalala, C., Mmbaga, B. T., Shibata, T., Menya, D., Goldstein, A. M., Hu, N., Malekzadeh, R., Fazel, A., McCormack, V., McKay, J., Perdomo, S., Scelo, G., Chanudet, E., Humphreys, L., Alexandrov, L. B., Brennan, P., Stratton, M. R., 2021. Mutational signatures in esophageal squamous cell carcinoma from eight countries of varying incidence.

Moore, L., Cagan, A., Coorens, T. H. H., Neville, M. D. C., Sanghvi, R., Sanders, M. A., Oliver, T. R. W., Leongamornlert, D., Ellis, P., Noorani, A., Mitchell, T. J., Butler, T. M., Hooks, Y., Warren, A. Y., Jorgensen, M., Dawson, K. J., Menzies, A., O'Neill, L., Latimer, C., Teng, M., van Boxtel, R., Iacobuzio-Donahue, C. A., Martincorena, I., Heer, R., Campbell, P. J., Fitzgerald, R. C., Stratton, M. R., Rahbari, R., 2021. The mutational landscape of human somatic and germline cells. Nature .

Moore, L., Leongamornlert, D., Coorens, T. H. H., Sanders, M. A., Ellis, P., Dentro, S. C., Dawson, K. J., Butler, T., Rahbari, R., Mitchell, T. J., Maura, F., Nangalia, J., Tarpey, P. S., Brunner, S. F., Lee-Six, H., Hooks, Y., Moody, S., Mahbubani, K. T., Jimenez-Linan, M., Brosens, J. J., Iacobuzio-Donahue, C. A., Martincorena, I., Saeb-Parsy, K., Campbell, P. J., Stratton, M. R., 2020. The mutational landscape of normal human endometrial epithelium. Nature 580(7805), 640–646.

Murai, K., Skrupskelyte, G., Piedrafita, G., Hall, M., Kostiou, V., Ong, S. H., Nagy, T., Cagan, A., Goulding, D., Klein, A. M., Hall, B. A., Jones, P. H., 2018. Epidermal tissue adapts to restrain progenitors carrying clonal p53 mutations. Cell Stem Cell 23(5), 687–699.e8.

Nakatsuka, Y., Vandenbon, A., Mino, T., Yoshinaga, M., Uehata, T., Cui, X., Sato, A., Tsujimura, T., Suzuki, Y., Sato, A., Handa, T., Chin, K., Sawa, T., Hirai, T., Takeuchi, O., 2018. Pulmonary regnase-1 orchestrates the interplay of epithelium and adaptive immune systems to protect against pneumonia. Mucosal Immunol. 11(4), 1203–1218.

Nanki, K., Fujii, M., Shimokawa, M., Matano, M., Nishikori, S., Date, S., Takano, A., Toshimitsu, K., Ohta, Y., Takahashi, S., Sugimoto, S., Ishimaru, K., Kawasaki, K., Nagai, Y., Ishii, R., Yoshida, K., Sasaki, N., Hibi, T., Ishihara, S., Kanai, T., Sato, T., 2020. Somatic inflammatory gene mutations in human ulcerative colitis epithelium. Nature 577(7789), 254–259.

Ng, S. C., Shi, H. Y., Hamidi, N., Underwood, F. E., Tang, W., Benchimol, E. I., Panaccione, R., Ghosh, S., Wu, J. C. Y., Chan, F. K. L., Sung, J. J. Y., Kaplan, G. G., 2017. Worldwide

incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. Lancet 390(10114), 2769–2778.

Nicholson, A. M., Olpe, C., Hoyle, A., Thorsen, A.-S., Rus, T., Colombé, M., Brunton-Sim, R., Kemp, R., Marks, K., Quirke, P., Malhotra, S., Ten Hoopen, R., Ibrahim, A., Lindskog, C., Myers, M. B., Parsons, B., Tavaré, S., Wilkinson, M., Morrissey, E., Winton, D. J., 2018. Fixation and spread of somatic mutations in adult human colonic epithelium. Cell Stem Cell 22(6), 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., Jönsson, G., Garber, J. E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J. W. M., Aparicio, S. A. J. R., Borg, Å., Salomon, A. V., Thomas, G., Børresen-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 Consortium, 2012. Mutational processes molding the genomes of 21 breast cancers. Cell 149(5), 979–993.

Nixon, K. C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15(4), 407–414.

Nowell, C. S., Radtke, F., 2017. Notch as a tumour suppressor. Nat. Rev. Cancer 17(3), 145–159.

Ogilvie, R. L., Sternjohn, J. R., Rattenbacher, B., Vlasova, I. A., Williams, D. A., Hau, H. H., Blackshear, P. J., Bohjanen, P. R., 2009. Tristetraprolin mediates interferon-gamma mRNA decay. J. Biol. Chem. 284(17), 11216–11223.

Okada, Y., Han, B., Tsoi, L. C., Stuart, P. E., Ellinghaus, E., Tejasvi, T., Chandran, V., Pellett, F., Pollock, R., Bowcock, A. M., Krueger, G. G., Weichenthal, M., Voorhees, J. J., Rahman, P., Gregersen, P. K., Franke, A., Nair, R. P., Abecasis, G. R., Gladman, D. D., Elder, J. T., de Bakker, P. I. W., Raychaudhuri, S., 2014. Fine mapping major histocompatibility complex associations in psoriasis and its clinical subtypes. Am. J. Hum. Genet. 95(2), 162–172.

Olafsson, S., McIntyre, R. E., Coorens, T., Butler, T., Jung, H., Robinson, P. S., Lee-Six, H., Sanders, M. A., Arestang, K., Dawson, C., Tripathi, M., Strongili, K., Hooks, Y., Stratton, M. R., Parkes, M., Martincorena, I., Raine, T., Campbell, P. J., Anderson, C. A., 2020. Somatic evolution in non-neoplastic IBD-Affected colon. Cell 182(3), 672–684.e11.

Papadopoulo, D., Laquerbe, A., Guillouf, C., Moustacchi, E., 1993. Molecular spectrum of mutations induced at the HPRT locus by a cross-linking agent in human cell lines with different repair capacities. Mutat. Res. 294(2), 167–177.

Parisi, R., Iskandar, I. Y. K., Kontopantelis, E., Augustin, M., Griffiths, C. E. M., Ashcroft, D. M., Global Psoriasis Atlas, 2020. National, regional, and worldwide epidemiology of psoriasis: systematic analysis and modelling study. BMJ 369, m1590.

Pasmooij, A. M. G., Pas, H. H., Deviaene, F. C. L., Nijenhuis, M., Jonkman, M. F., 2005. Multiple correcting COL17A1 mutations in patients with revertant mosaicism of epidermolysis bullosa. Am. J. Hum. Genet. 77(5), 727–740.

Pasternak, B., Svanström, H., Schmiegelow, K., Jess, T., Hviid, A., 2013. Use of azathioprine and the risk of cancer in inflammatory bowel disease. Am. J. Epidemiol. 177(11), 1296–1305.

Patial, S., Curtis, A. D., 2nd, Lai, W. S., Stumpo, D. J., Hill, G. D., Flake, G. P., Mannie, M. D., Blackshear, P. J., 2016. Enhanced stability of tristetraprolin mRNA protects mice against immune-mediated inflammatory pathologies. Proc. Natl. Acad. Sci. U. S. A. 113(7), 1865–1870.

Peterson, S. C., Eberl, M., Vagnozzi, A. N., Belkadi, A., Veniaminova, N. A., Verhaegen, M. E., Bichakjian, C. K., Ward, N. L., Dlugosz, A. A., Wong, S. Y., 2015. Basal cell carcinoma preferentially arises from stem cells within hair follicle and mechanosensory niches. Cell Stem Cell 16(4), 400–412.

Pfeifer, G. P., You, Y.-H., Besaratinia, A., 2005. Mutations induced by ultraviolet light. Mutat. Res. 571(1-2), 19–31.

Pich, O., Cortes-Bullich, A., Muiños, F., Pratcorona, M., Gonzalez-Perez, A., Lopez-Bigas, N., 2021. The evolution of hematopoietic cells under cancer therapy. Nat. Commun. 12(1), 4803.

Pich, O., Muiños, F., Lolkema, M. P., Steeghs, N., Gonzalez-Perez, A., Lopez-Bigas, N., 2019. The mutational footprints of cancer therapies. Nat. Genet. 51(12), 1732–1740.

Pickering, C. R., Zhou, J. H., Lee, J. J., Drummond, J. A., Peng, S. A., Saade, R. E., Tsai, K. Y., Curry, J. L., Tetzlaff, M. T., Lai, S. Y., Yu, J., Muzny, D. M., Doddapaneni, H., Shinbrot, E., Covington, K. R., Zhang, J., Seth, S., Caulin, C., Clayman, G. L., El-Naggar, A. K., Gibbs, R. A., Weber, R. S., Myers, J. N., Wheeler, D. A., Frederick, M. J., 2014. Mutational landscape of aggressive cutaneous squamous cell carcinoma. Clin. Cancer Res. 20(24), 6582–6592.

Pleguezuelos-Manzano, C., Puschhof, J., Huber, A. R., van Hoeck, A., Wood, H. M., Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P. B., Paganelli, F. L., Geurts, M. H., Beumer, J., Mizutani, T., van der Linden, R., van Elst, S., Ambrose, J. C., Arumugam, P., Baple, E. L., Bleda, M., Boardman-Pretty, F., Boissiere, J. M., Boustred, C. R., Brittain, H., Caulfield, M. J., Chan, G. C., Craig, C. E. H., Daugherty, L. C., de Burca, A., Devereau, A., Elgar, G., Foulger, R. E., Fowler, T., Furió-Tarí, P., Hackett, J. M., Halai, D., Hamblin, A., Henderson, S., Holman, J. E., Hubbard, T. J. P., Ibáñez, K., Jackson, R., Jones, L. J., Kasperaviciute, D., Kayikci, M., Lahnstein, L., Lawson, K., Leigh, S. E. A., Leong, I. U. S., Lopez, F. J., Maleady-Crowe, F., Mason, J., McDonagh, E. M., Moutsianas, L., Mueller, M., Murugaesu, N., Need, A. C., Odhams, C. A., Patch, C., Perez-Gil, D., Polychronopoulos, D., Pullinger, J., Rahim, T., Rendon, A., Riesgo-Ferreiro, P., Rogers, T., Ryten, M., Savage, K., Sawant, K., Scott, R. H., Siddiq, A., Sieghart, A., Smedley, D., Smith, K. R., Sosinsky, A., Spooner, W., Stevens, H. E., Stuckey, A., Sultana, R., Thomas, E. R. A., Thompson, S. R., Tregidgo, C., Tucci, A., Walsh, E., Watters, S. A., Welland, M. J., Williams, E., Witkowska, K., Wood, S. M., Zarowiecki, M., Top,

J., Willems, R. J. L., Giannakis, M., Bonnet, R., Quirke, P., Meyerson, M., Cuppen, E., van Boxtel, R., Clevers, H., Genomics England Research Consortium, 2020. Mutational signature in colorectal cancer caused by genotoxic pks+ e. coli. Nature .

Polak, P., Karlić, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds, A., Rynes, E., Vlahoviček, K., Stamatoyannopoulos, J. A., Sunyaev, S. R., 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518(7539), 360–364.

Powell, S. N., Kachnic, L. A., 2003. Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. Oncogene 22(37), 5784–5791.

Priestley, P., Baber, J., Lolkema, M. P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., Roepman, P., Voda, M., Bloemendal, H. J., Tjan-Heijnen, V. C. G., van Herpen, C. M. L., Labots, M., Witteveen, P. O., Smit, E. F., Sleijfer, S., Voest, E. E., Cuppen, E., 2019. Pan-cancer whole-genome analyses of metastatic solid tumours. Nature 575(7781), 210–216.

Qian, X., Ning, H., Zhang, J., Hoft, D. F., Stumpo, D. J., Blackshear, P. J., Liu, J., 2011. Posttranscriptional regulation of IL-23 expression by IFN-gamma through tristetraprolin. J. Immunol. 186(11), 6454–6464.

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, 15.7.1–15.7.12.

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(1), 15.9.1–15.9.17.

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z., Feuerbach, L., Sabarinathan, R., Madsen, T., Kim, J., Mularoni, L., Shuai, S., Lanzós, A., Herrmann, C., Maruvka, Y. E., Shen, C., Amin, S. B., Bandopadhayay, P., Bertl, J., Boroevich, K. A., Busanovich, J., Carlevaro-Fita, J., Chakravarty, D., Chan, C. W. Y., Craft, D., Dhingra, P., Diamanti, K., Fonseca, N. A., Gonzalez-Perez, A., Guo, Q., Hamilton, M. P., Haradhvala, N. J., Hong, C., Isaev, K., Johnson, T. A., Juul, M., Kahles, A., Kahraman, A., Kim, Y., Komorowski, J., Kumar, K., Kumar, S., Lee, D., Lehmann, K.-V., Li, Y., Liu, E. M., Lochovsky, L., Park, K., Pich, O., Roberts, N. D., Saksena, G., Schumacher, S. E., Sidiropoulos, N., Sieverling, L., Sinnott-Armstrong, N., Stewart, C., Tamborero, D., Tubio, J. M. C., Umer, H. M., Uusküla-Reimand, L., Wadelius, C., Wadi, L., Yao, X., Zhang, C.-Z., Zhang, J., Haber, J. E., Hobolth, A., Imielinski, M., Kellis, M., Lawrence, M. S., von Mering, C., Nakagawa, H., Raphael, B. J., Rubin, M. A., Sander, C., Stein, L. D., Stuart, J. M., Tsunoda, T., Wheeler, D. A., Johnson, R., Reimand, J., Gerstein, M., Khurana, E., Campbell, P. J., López-Bigas, N., PCAWG Drivers and Functional Interpretation Working Group, PCAWG Structural Variation Working Group, Weischenfeldt, J., Beroukhim, R., Martincorena, I., Pedersen, J. S., Getz, G., PCAWG Consortium, 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 578(7793), 102–111.

Riva, L., Pandiri, A. R., Li, Y. R., Droop, A., Hewinson, J., Quail, M. A., Iyer, V., Shepherd, R., Herbert, R. A., Campbell, P. J., Sills, R. C., Alexandrov, L. B., Balmain, A., Adams, D. J., 2020. The mutational signature profile of known and suspected human carcinogens in mice. Nat. Genet. 52(11), 1189–1197.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., Kellis, M., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518(7539), 317–330.

Roberts, N. D., 2018. Patterns of somatic genome rearrangement in human cancer. Ph.D. thesis, University of Cambridge.

Robles, A. I., Traverso, G., Zhang, M., Roberts, N. J., Khan, M. A., Joseph, C., Lauwers, G. Y., Selaru, F. M., Popoli, M., Pittman, M. E., Ke, X., Hruban, R. H., Meltzer, S. J., Kinzler, K. W., Vogelstein, B., Harris, C. C., Papadopoulos, N., 2016. Whole-Exome sequencing analyses of inflammatory bowel Disease-Associated colorectal cancers. Gastroenterology 150(4), 931–943.

Rodriguez-Martin, B., Alvarez, E. G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y. S., Temes, J., Garcia-Souto, D., Detering, H., Li, Y., Rodriguez-Castro, J., Dueso-Barroso, A., Bruzos, A. L., Dentro, S. C., Blanco, M. G., Contino, G., Ardeljan, D., Tojo, M., Roberts, N. D., Zumalave, S., Edwards, P. A. W., Weischenfeldt, J., Puiggròs, M., Chong, Z., Chen, K., Lee, E. A., Wala, J. A., Raine, K., Butler, A., Waszak, S. M., Navarro, F. C. P., Schumacher, S. E., Monlong, J., Maura, F., Bolli, N., Bourque, G., Gerstein, M., Park, P. J., Wedge, D. C., Beroukhim, R., Torrents, D., Korbel, J. O., Martincorena, I., Fitzgerald, R. C., Van Loo, P., Kazazian, H. H., Burns, K. H., PCAWG Structural Variation Working Group, Campbell, P. J., Tubio, J. M. C., PCAWG Consortium, 2020. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. .

Sait, L. C., Galic, M., Price, J. D., Simpfendorfer, K. R., Diavatopoulos, D. A., Uren, T. K., Janssen, P. H., Wijburg, O. L. C., Strugnell, R. A., 2007. Secretory antibodies reduce systemic antibody responses against the gastrointestinal commensal flora. Int. Immunol. 19(3), 257–265.

Sano, S., Oshima, K., Wang, Y., MacLauchlan, S., Katanasaka, Y., Sano, M., Zuriaga, M. A., Yoshiyama, M., Goukassian, D., Cooper, M. A., Fuster, J. J., Walsh, K., 2018. Tet2-

Mediated clonal hematopoiesis accelerates heart failure through a mechanism involving the IL-1$\beta$/NLRP3 inflammasome. J. Am. Coll. Cardiol. 71(8), 875–886.

Sasaki, N., Sachs, N., Wiebrands, K., Ellenbroek, S. I. J., Fumagalli, A., Lyubimova, A., Begthel, H., van den Born, M., van Es, J. H., Karthaus, W. R., Li, V. S. W., López-Iglesias, C., Peters, P. J., van Rheenen, J., van Oudenaarden, A., Clevers, H., 2016. Reg4+ deep crypt secretory cells function as epithelial niche for lgr5+ stem cells in colon. Proc. Natl. Acad. Sci. U. S. A. 113(37), E5399–407.

Semlow, D. R., Zhang, J., Budzowska, M., Drohat, A. C., Walter, J. C., 2016. Replication-Dependent unhooking of DNA interstrand Cross-Links by the NEIL3 glycosylase. Cell 167(2), 498–511.e14.

Shevelev, I. V., Hübscher, U., 2002. The 3' 5' exonucleases. Nat. Rev. Mol. Cell Biol. 3(5), 364–376.

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(1), 134–144.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., Curtis, C., 2015. A big bang model of human colorectal tumor growth. Nat. Genet. 47(3), 209–216.

South, A. P., Purdie, K. J., Watt, S. A., Haldenby, S., den Breems, N., Dimon, M., Arron, S. T., Kluk, M. J., Aster, J. C., McHugh, A., Xue, D. J., Dayal, J. H., Robinson, K. S., Rizvi, S. H., Proby, C. M., Harwood, C. A., Leigh, I. M., 2014. NOTCH1 mutations occur early during cutaneous squamous cell carcinogenesis. J. Invest. Dermatol. 134(10), 2630–2638.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., Sunyaev, S. R., 2009. Human mutation rate associated with DNA replication timing. Nat. Genet. 41(4), 393–395.

Stamp, C., Zupanic, A., Sachdeva, A., Stoll, E. A., Shanley, D. P., Mathers, J. C., Kirkwood, T. B. L., Heer, R., Simons, B. D., Turnbull, D. M., Greaves, L. C., 2018. Predominant asymmetrical stem cell fate outcome limits the rate of niche succession in human colonic crypts. EBioMedicine 31, 166–173.

Stumpo, D. J., Broxmeyer, H. E., Ward, T., Cooper, S., Hangoc, G., Chung, Y. J., Shelley, W. C., Richfield, E. K., Ray, M. K., Yoder, M. C., Aplan, P. D., Blackshear, P. J., 2009. Targeted disruption of zfp36l2, encoding a CCCH tandem zinc finger RNA-binding protein, results in defective hematopoiesis. Blood 114(12), 2401–2410.

Suspène, R., Aynaud, M.-M., Guétard, D., Henry, M., Eckhoff, G., Marchio, A., Pineau, P., Dejean, A., Vartanian, J.-P., Wain-Hobson, S., 2011. Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proc. Natl. Acad. Sci. U. S. A. 108(12), 4858–4863.

Suzuki, S., Nomura, T., Miyauchi, T., Takeda, M., Fujita, Y., Nishie, W., Akiyama, M., Ishida-Yamamoto, A., Shimizu, H., 2019. Somatic recombination underlies frequent revertant mosaicism in loricrin keratoderma. Life Sci Alliance 2(1).

Tang, J., Fewings, E., Chang, D., Zeng, H., Liu, S., Jorapur, A., Belote, R. L., McNeal, A. S., Tan, T. M., Yeh, I., Arron, S. T., Judson-Torres, R. L., Bastian, B. C., Shain, A. H., 2020. The genomic landscapes of individual melanocytes from human skin. Nature 586(7830), 600–605.

Targan, S. R., Feagan, B., Vermeire, S., Panaccione, R., Melmed, G. Y., Landers, C., Li, D., Russell, C., Newmark, R., Zhang, N., Chon, Y., Hsu, Y.-H., Lin, S.-L., Klekotka, P., 2016. A randomized, Double-Blind, Placebo-Controlled phase 2 study of brodalumab in patients with Moderate-to-Severe crohn's disease. Am. J. Gastroenterol. 111(11), 1599–1607.

Taylor, G. A., Carballo, E., Lee, D. M., Lai, W. S., Thompson, M. J., Patel, D. D., Schenkman, D. I., Gilkeson, G. S., Broxmeyer, H. E., Haynes, B. F., Blackshear, P. J., 1996. A pathogenetic role for TNF alpha in the syndrome of cachexia, arthritis, and autoimmunity resulting from tristetraprolin (TTP) deficiency. Immunity 4(5), 445–454.

Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M., 2006. Hierarchical dirichlet processes. J. Am. Stat. Assoc. 101(476), 1566–1581.

Terao, C., Suzuki, A., Momozawa, Y., Akiyama, M., Ishigaki, K., Yamamoto, K., Matsuda, K., Murakami, Y., McCarroll, S. A., Kubo, M., Loh, P.-R., Kamatani, Y., 2020. Chromosomal alterations among age-related haematopoietic clones in japan. Nature 584(7819), 130–135.

Thompson, D. J., Genovese, G., Halvardson, J., Ulirsch, J. C., Wright, D. J., Terao, C., Davidsson, O. B., Day, F. R., Sulem, P., Jiang, Y., Danielsson, M., Davies, H., Dennis, J., Dunlop, M. G., Easton, D. F., Fisher, V. A., Zink, F., Houlston, R. S., Ingelsson, M., Kar, S., Kerrison, N. D., Kinnersley, B., Kristjansson, R. P., Law, P. J., Li, R., Loveday, C., Mattisson, J., McCarroll, S. A., Murakami, Y., Murray, A., Olszewski, P., Rychlicka-Buniowska, E., Scott, R. A., Thorsteinsdottir, U., Tomlinson, I., Moghadam, B. T., Turnbull, C., Wareham, N. J., Gudbjartsson, D. F., International Lung Cancer Consortium (INTEGRAL-ILCCO), Breast Cancer Association Consortium, Consortium of Investigators of Modifiers of BRCA1/2, Endometrial Cancer Association Consortium, Ovarian Cancer Association Consortium, Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) Consortium, Kidney Cancer GWAS Meta-Analysis Project, eQTLGen Consortium, Biobank-based Integrative Omics Study (BIOS) Consortium, 23andMe Research Team, Kamatani, Y., Hoffmann, E. R., Jackson, S. P., Stefansson, K., Auton, A., Ong, K. K., Machiela, M. J., Loh, P.-R., Dumanski, J. P., Chanock, S. J., Forsberg, L. A., Perry, J. R. B., 2019. Genetic predisposition to mosaic Y chromosome loss in blood. Nature 575(7784), 652–657.

Totafurno, J., Bjerknes, M., Cheng, H., 1987. The crypt cycle. crypt and villus production in the adult intestinal epithelium. Biophys. J. 52(2), 279–294.

Tsoi, L. C., Stuart, P. E., Tian, C., Gudjonsson, J. E., Das, S., Zawistowski, M., Ellinghaus, E., Barker, J. N., Chandran, V., Dand, N., Duffin, K. C., Enerbäck, C., Esko, T., Franke, A., Gladman, D. D., Hoffmann, P., Kingo, K., Kõks, S., Krueger, G. G., Lim, H. W., Metspalu,

A., Mrowietz, U., Mucha, S., Rahman, P., Reis, A., Tejasvi, T., Trembath, R., Voorhees, J. J., Weidinger, S., Weichenthal, M., Wen, X., Eriksson, N., Kang, H. M., Hinds, D. A., Nair, R. P., Abecasis, G. R., Elder, J. T., 2017. Large scale meta-analysis characterizes genetic architecture for common psoriasis associated variants. Nat. Commun. 8, 15382.

Uhlig, H. H., 2013. Monogenic diseases associated with intestinal inflammation: implications for the understanding of inflammatory bowel disease. Gut 62(12), 1795–1805.

Uhlig, H. H., Schwerd, T., Koletzko, S., Shah, N., Kammermeier, J., Elkadri, A., Ouahed, J., Wilson, D. C., Travis, S. P., Turner, D., Klein, C., Snapper, S. B., Muise, A. M., COLORS in IBD Study Group and NEOPICS, 2014. The diagnostic approach to monogenic very early onset inflammatory bowel disease. Gastroenterology 147(5), 990–1007.e3.

Vaengebjerg, S., Skov, L., Egeberg, A., Loft, N. D., 2020. Prevalence, incidence, and risk of cancer in patients with psoriasis and psoriatic arthritis: A systematic review and meta-analysis. JAMA Dermatol. .

Van Loo, P., Nordgard, S. H., Lingjærde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., Børresen-Dale, A.-L., Kristensen, V. N., 2010. Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. U. S. A. 107(39), 16910–16915.

Veness, M. J., 2007. High-risk cutaneous squamous cell carcinoma of the head and neck. J. Biomed. Biotechnol. 2007(3), 80572.

Verkouteren, J. A. C., Ramdas, K. H. R., Wakkee, M., Nijsten, T., 2017. Epidemiology of basal cell carcinoma: scholarly review. Br. J. Dermatol. 177(2), 359–372.

Vermeulen, L., Morrissey, E., van der Heijden, M., Nicholson, A. M., Sottoriva, A., Buczacki, S., Kemp, R., Tavare, S., Winton, D. J., 2013. Defining stem cell dynamics in models of intestinal tumor initiation. Science 342(6161), 995–998.

Volkova, N. V., Meier, B., González-Huici, V., Bertolini, S., Gonzalez, S., Vöhringer, H., Abascal, F., Martincorena, I., Campbell, P. J., Gartner, A., Gerstung, M., 2020. Mutational signatures are jointly shaped by DNA damage and repair. Nat. Commun. 11(1), 2169.

Vu, T. N., Nguyen, H.-N., Calza, S., Kalari, K. R., Wang, L., Pawitan, Y., 2019. Cell-level somatic mutation detection from single-cell RNA sequencing. Bioinformatics 35(22), 4679–4687.

Wasan, H. S., Park, H. S., Liu, K. C., Mandir, N. K., Winnett, A., Sasieni, P., Bodmer, W. F., Goodlad, R. A., Wright, N. A., 1998. APC in the regulation of intestinal crypt fission. J. Pathol. 185(3), 246–255.

Watson, C. J., Papula, A. L., Poon, G. Y. P., Wong, W. H., Young, A. L., Druley, T. E., Fisher, D. S., Blundell, J. R., 2020. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science 367(6485), 1449–1454.

Wei, L., Christensen, S. R., Fitzgerald, M. E., Graham, J., Hutson, N. D., Zhang, C., Huang, Z., Hu, Q., Zhan, F., Xie, J., Zhang, J., Liu, S., Remenyik, E., Gellen, E., Colegio, O. R., Bax, M., Xu, J., Lin, H., Huss, W. J., Foster, B. A., Paragh, G., 2021. Ultradeep sequencing

differentiates patterns of skin clonal mutations associated with sun-exposure status and skin cancer burden. Sci Adv 7(1).

Williams, M. J., Zapata, L., Werner, B., Barnes, C. P., Sottoriva, A., Graham, T. A., 2020. Measuring the distribution of fitness effects in somatic evolution by combining clonal dynamics with dN/dS ratios. Elife 9.

Wong, C. C., Martincorena, I., Rust, A. G., Rashid, M., Alifrangis, C., Alexandrov, L. B., Tiffen, J. C., Kober, C., Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium, Green, A. R., Massie, C. E., Nangalia, J., Lempidaki, S., Döhner, H., Döhner, K., Bray, S. J., McDermott, U., Papaemmanuil, E., Campbell, P. J., Adams, D. J., 2014. Inactivating CUX1 mutations promote tumorigenesis. Nat. Genet. 46(1), 33–38.

Wright, D. J., Day, F. R., Kerrison, N. D., Zink, F., Cardona, A., Sulem, P., Thompson, D. J., Sigurjonsdottir, S., Gudbjartsson, D. F., Helgason, A., Chapman, J. R., Jackson, S. P., Langenberg, C., Wareham, N. J., Scott, R. A., Thorsteindottir, U., Ong, K. K., Stefansson, K., Perry, J. R. B., 2017. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. Nat. Genet. 49(5), 674–679.

Yaeger, R., Shah, M. A., Miller, V. A., Kelsen, J. R., Wang, K., Heins, Z. J., Ross, J. S., He, Y., Sanford, E., Yantiss, R. K., Balasubramanian, S., Stephens, P. J., Schultz, N., Oren, M., Tang, L., Kelsen, D., 2016. Genomic alterations observed in Colitis-Associated cancers are distinct from those found in sporadic colorectal cancers and vary by type of inflammatory bowel disease. Gastroenterology 151(2), 278–287.e6.

Yamamoto, M., Yamazaki, S., Uematsu, S., Sato, S., Hemmi, H., Hoshino, K., Kaisho, T., Kuwata, H., Takeuchi, O., Takeshige, K., Saitoh, T., Yamaoka, S., Yamamoto, N., Yamamoto, S., Muta, T., Takeda, K., Akira, S., 2004. Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein IkappaBzeta. Nature 430(6996), 218–222.

Yang, S. C., Lin, J. G., Chiou, C. C., Chen, L. Y., Yang, J. L., 1994. Mutation specificity of 8-methoxypsoralen plus two doses of UVA irradiation in the hprt gene in diploid human fibroblasts. Carcinogenesis 15(2), 201–207.

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(21), 2865–2871.

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(7739), 312–317.

Yoshida, K., Gowers, K. H. C., Lee-Six, H., Chandrasekharan, D. P., Coorens, T., Maughan, E. F., Beal, K., Menzies, A., Millar, F. R., Anderson, E., Clarke, S. E., Pennycuick, A.,

Thakrar, R. M., Butler, C. R., Kakiuchi, N., Hirano, T., Hynds, R. E., Stratton, M. R., Martincorena, I., Janes, S. M., Campbell, P. J., 2020. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature 578(7794), 266–272.

Zhao, S., Li, Z., Zhang, M., Zhang, L., Zheng, H., Ning, J., Wang, Y., Wang, F., Zhang, X., Gan, H., Wang, Y., Zhang, X., Luo, H., Bu, G., Xu, H., Yao, Y., Zhang, Y.-W., 2019. A brain somatic RHEB doublet mutation causes focal cortical dysplasia type II. Exp. Mol. Med. 51(7), 1–11.

Zhen, W. P., Buchardt, O., Nielsen, H., Nielsen, P. E., 1986. Site specificity of psoralen-DNA interstrand cross-linking determined by nuclease bal31 digestion. Biochemistry 25(21), 6598–6603.

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(3), 608–621.e12.

Zink, F., Stacey, S. N., Norddahl, G. L., Frigge, M. L., Magnusson, O. T., Jonsdottir, I., Thorgeirsson, T. E., Sigurdsson, A., Gudjonsson, S. A., Gudmundsson, J., Jonasson, J. G., Tryggvadottir, L., Jonsson, T., Helgason, A., Gylfason, A., Sulem, P., Rafnar, T., Thorsteinsdottir, U., Gudbjartsson, D. F., Masson, G., Kong, A., Stefansson, K., 2017. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. Blood 130(6), 742–752.

Zong, C., Lu, S., Chapman, A. R., Xie, X. S., 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science 338(6114), 1622–1626.