Somatic mutagenesis in humans with deficient DNA repair

Philip Robinson
Emmanuel College
University of Cambridge

February 2022

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. It does not exceed the prescribed word limit of 60,000 for the Degree Committee.

Summary

Somatic mutagenesis in humans with deficient DNA repair

Philip Robinson

The accumulation of mutations in normal cells causes the development of cancer and is implicated as a potential mechanism in the physiological process of ageing. In recent years our ability to interrogate the genome of human cancers and the normal tissues from which they arise has expanded greatly. These studies have shown that mutations accumulate in normal tissues throughout life and that mutation rates are remarkably similar across individuals. However, the potential impact of increased somatic mutation rates on the risk of developing cancer and the process of ageing is not known.

In this thesis, two inherited syndromes associated with intestinal cancer predisposition were selected to investigate the mutation burdens and mutational processes across different normal tissue types. Individuals with these syndromes have a known elevated risk of cancer which is thought to be underpinned by an increased somatic mutation rate. Chapter 3 summarises experiments that investigate somatic mutagenesis in a selection of normal tissue types from individuals with germline heterozygous mutations in the DNA polymerase genes *POLE* and *POLD1*. Chapter 4 summarises the investigation of somatic mutagenesis in normal tissues from individuals with germline *MUTYH* mutations. In Chapter 5 findings from the two cancer predisposition syndromes are compared and the results are placed in the broader context of intestinal cancer predisposition syndromes. Lastly, the observations from this thesis are interpreted with regards our current understanding of the somatic mutation theory of ageing.

In summary, this thesis presents insight into somatic mutagenesis in normal tissues from individuals with known cancer predisposition. The findings may have potential implications for our understanding of cancer risk in predisposed and non-predisposed individuals. The observation of increased somatic mutation rates in normal healthy tissues also has pertinence to our understanding of the somatic mutation theory of ageing. The data presented in the

thesis serve as a potential proof-of-concept for the measurement of somatic mutagenesis in normal tissues to improve the care of individuals with inherited DNA repair defects.

Table of Contents

Description of contributions

The work described in this Thesis involved the direct and indirect contributions of several colleagues and collaborators. The nature of these contributions are outlined below.

Chapter 3

Experimental design was undertaken with Mike Stratton, Claire Palles (University of Birmingham) and Ian Tomlinson (University of Edinburgh). Patient recruitment, sample collection and extraction of DNA from blood and sperm samples was undertaken by members of the Tomlinson Lab. Tissue processing and sectioning for a small number of samples was undertaken by Yvette Hooks. The low-input DNA sequencing protocol was designed by Pete Ellis and Luiza Moore. Low-input library preparation was conducted by the Wellcome Sanger Institute Core Sequencing Pipelines. NanoSeq library preparation and sequencing was performed by Wellcome Sanger Institute Core Sequencing Pipelines.

Genome alignment and QC was performed by Wellcome Sanger Institute Core informatics. Mutation calling algorithms were developed and maintained by the CASM IT team. Filters to remove low-input DNA library related mutations were developed by Mathjis Sanders. The pipeline for filtering unmatched mutation SBS and ID mutations was developed by Tim Coorens. Data for normal controls was contributed by Henry Lee Six and re-filtered by Sigurgeir Olaffson. Algorithms to assign mutations to phylogenetic trees were written by Nick Williams. Mutational signature attribution code was written by Andrew Lawson. Code for running HDP and SigProfiler mutational signature analysis was provided by Tim Butler. Filtering and calling of NanoSeq (duplex sequencing) data was undertaken by Federico Abascal. Emily Mitchell undertook analysis of ARCH variants in deep targeted sequencing of blood using bespoke pipelines. Tim Coorens and I jointly created figures for this chapter. Tim Coorens performed analysis of embryonic variants.

My contributions included:

- Governance matters relating to samples and their use
- Sample management
- Generation of frozen tissue sections, histological staining and slide scanning

- Laser capture microdissection of tissues
- Somatic mutation calling and filtering of SBS and ID mutations
- Somatic mutation calling, filtering and curation of structural rearrangements and copy-number alterations
- Statistical modelling of mutation rate estimates
- Mutational signature analysis for SBS and ID mutation types.
- Characterisation and validation of mutational signatures
- Mutational signature assignment and characterisation of mutational signatures e.g. replication strand biases etc.
- Driver mutation analysis
- Telomere length analysis and statistical modelling with guidance from Peter Campbell
- Analysis of mutational ages

Chapter 4

The experimental design was undertaken jointly with Mike Stratton with input from Laura Thomas (Swansea University) and Julian Sampson (Cardiff University). Patient recruitment, storage of samples and extraction of blood samples was undertaken by members of the Julian Sampson's Lab. Tissue processing and sectioning for some samples was undertaken by Yvette Hooks. The laser capture microdissection and low-input DNA sequencing protocol was designed by Pete Ellis and Luiza Moore. Low-input library preparation was conducted by the Wellcome Sanger Institute Core Sequencing Pipelines.

Mutation calling algorithms were developed and maintained by the CASM IT team. Filters to remove low-input DNA library related mutations were developed by Mathjis Sanders. The pipeline for filtering unmatched mutation SBS and ID mutations was developed by Tim Coorens. Data for normal controls was contributed by Henry Lee Six and re-filtered by Sigurgeir Olafsson. Algorithms to assign mutations to phylogenetic trees were written by Nick Williams. Hyunchul Jung undertook calling and filtering of CNV and SV mutation types. Code for running HDP and SigProfiler mutational signature analysis was provided by Tim Butler. NanoSeq libraries and sequencing was performed by Wellcome Sanger Institute Core Sequencing Pipelines. Filtering and calling of NanoSeq (duplex sequencing) data was undertaken by Federico Abascal.

My contributions included:

- Governance matters relating to samples and their use
- Sample management
- Generation of frozen tissue sections, histological staining and slide scanning
- Laser capture microdissection of tissues
- Somatic SBS and ID mutation calling
- Filtering and QC of filtering pipeline
- Analysis of mutation burdens
- Telomere length estimation and analysis
- Statistical analysis including modelling of mutation rates and telomere length with guidance from Peter Campbell
- Mutational signature analysis
- Validation of mutational signatures
- Characterisation of mutational signatures e.g. replication strand biases
- Driver mutation analysis
- Creation of all figures

Acknowledgements

With many sincere and profound thanks to my supervisors for their unwavering support throughout the PhD. To my primary supervisor Mike Stratton, thank you for introducing me to the world of scientific discovery. Thank you for mentoring me and helping me to develop as a scientist. It has been a thrilling adventure and tremendous privilege. Many thanks to my secondary supervisor Peter Campbell who has welcomed me so warmly and has opened my eyes to many of the joys of science and bioinformatic analysis.

It has been an honour to share this journey with the talented colleagues at the Sanger Institute whose passion, dedication and generosity has been a pleasure to behold. I am grateful to you all. I would like to make particular mention of several colleagues: Tim Coorens for your mentorship and for sharing the DNA polymerase journey with me. Pete Ellis and Luiza Moore who developed the laboratory protocols that enabled this project and produced the first generation of studies. Henry for your friendly and thoughtful feedback and the foundational role in the development of the low-input pipeline and the invaluable data you generated. Inigo, Fede and Andrew for your support over the years and for all of the time you have generously spent with me. Thank you for the remarkable technologies that you have developed that are so central to the work conducted in this thesis. Tim Butler for all of your help with mutational signature analysis and help getting it all going. Tom Mitchell for being a patient and calm source of encouragement and support during my first experiences of coding. Robert Osbourne who was a go-to for sage advice and guidance. Joe Lee for the great 'chats' and moral support throughout the PhD. Sigurgeir Olafsson for your generous advice, support and feedback. Luke Harvey for all of your help and support with NanoSeq and exome sequencing. Raheleh Rahbari for your calm and wise advice and support. Mathijs Sanders for all of your invaluable work on artefact filtering and for introducing me to the inner workings of mismatch repair. Emily for your infectious positivity and for all of your help. Bernard Lee for your friendship and support during your time at the Sanger. Yvette for all of your help with histology and for your companionship during the many hours spent together in the lab. Joanne Doleman, Carol Smee and George Martin for all of your support with research governance and for teaching me lots along the way. Laura O'Neill, Claire Hardy, Kirsty Roberts, Calli Latimer and Liz Anderson for your outstanding support with laboratory work and sample management. Jo, Wendy, Debbie

and the teams for your outstanding administrative support. Wellcome Sanger Institute Core Sequencing Pipelines and Wellcome Sanger Institute Core informatics for your invaluable role in the generation of these data. CASM IT who create and maintain the analytical pipelines I relied upon for analysis and have patiently helped me navigate many informatic hurdles. Matthew Astley for introducing me to the world of informatic systems.

Thanks to all who have generously given of their time to provide feedback for this Thesis including: Aleks, Andrew, Joe, Emily and Mike Spencer-Chapman.

To all of the colleagues and friends at collaborating Institutions. Thank you for your generosity and openness, without your collaboration these projects would not have been possible. To the many patients and their families who so kindly contribute their time and samples to enable all of this work to occur.

Lastly, to my family and friends who have supported me throughout the PhD fellowship and allowed me to commit wholeheartedly to this endeavour, I thank you for your steadfast support.

Abbreviations and Acronyms

DNA	Deoxyribose nucleic acid
TCGA	The Cancer Genome Atlas
COSMIC	Catalogue of Somatic Mutations In Cancer
ICGC	International Cancer Genome Consortium
SBS	Single base substitution
ID	Insertion or deletion
CNV	Copy number variant
SV	Structural variant
PPAP	Polymerase proofreading associated polyposis
MAP	MUTYH associated polyposis
FAP	Familial adenomatous polyposis
AFAP	Attenuated familial adenomatous polyposis
MCR	Mutation cluster region
CIMP	CpG island methylator phenotype
ACF	Aberrant crypt focus
NSAID	Non-steroidal anti-inflammatory drug
MMR	Mismatch repair
BER	Base excision repair
8-OG	8-oxo-guanine
NNMF	Non-negative matrix factorisation
HDP	Hierarchical Dirichlet process
IQR	Interquartile range

Chapter 1 - Introduction

The cancer genome

Mutations accumulate over time in normal cells causing the development of cancer and potentially other diseases associated with ageing[1]. The transformation of a single cell into a cancer arises from the accumulation of mutations and is moulded by positive evolutionary selection[1-4]. Over the past 40 years since the discovery of the first cancer gene, understanding of the cancer genome has expanded greatly. Through the study of cancer mutations, the processes that lead to their acquisition have also become apparent[5,6]. The cancer genome contains a record of the mutations acquired throughout its lifetime from the fertilised embryo until shortly before the cancer was sequenced. Thus, interrogation of the cancer genome can reveal insights into the processes present in the normal cells that preceded its development. Investigation of the mutations and mutational processes in normal cells may inform our understanding of the cellular mechanisms that operate prior to and during cancer formation. In addition to informing our understanding of the mechanisms of carcinogenesis, these studies may reveal the determinants of the variable cancer risk observed in different individuals and across the different normal tissue types.

Twenty years ago the first draft and initial analysis of the human genome was published[7]. This achievement marked an important milestone and served as a catalyst for the development of several scientific fields including cancer genomics. In 2001, approximately 300 cancer genes had been discovered[8]. At that time, large scale investigation of cancer genomes was technologically and financially unviable. The study of cancer genomics was restricted to the coding exons, panels of a small number of genes or specific mutations. However, within a decade of the publication of the human genome, proof-of-principle studies emerged, demonstrating the application of massively parallel paired-end sequencing to the characterisation of the genome of cancer cells[9,10]. These studies illustrated, on a small scale, that it was possible to survey the entire genome and to identify somatically-acquired genomic alterations that had accumulated during the lifespan of the cancer and its normal precursors. They reported the identification of several different mutation types from genome sequencing data, including base substitutions, small insertion and deletions, copy-number alterations and large-scale structural alterations. In addition, they illustrated the potential of genomic sequencing to infer clonal dynamics and somatic evolution in cancer.

Over the course of the past decade, falling costs and an increased availability of genome sequencing have enabled the study of larger numbers of cancer samples. An increasing emphasis on data sharing has led to the establishment of publicly available catalogues and open source resources such as the Cataloguing the Somatic Mutations In Cancer (COSMIC) resource[3]. In recognition of the value generated through the comparison of large cancer genomics data sets containing a spectrum of cancer types, two large international consortia were established. The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) coordinated the in-depth characterisation of cancer genomes across the most common cancer types. The data generated from these consortia has catalysed research in the field of cancer genomics and serves as a valuable ongoing resource for the field. The scale of cancer genome sequencing is likely to increase further in coming years with large scientific and healthcare organisations aiming to introduce genome sequencing into routine medical care.

Cancer genomes typically have elevated burdens of somatic mutations compared to normal cells in healthy tissues[5]. The number and type of mutations varies between cancer tissue-types and is influenced by various endogenous and exogenous mutational processes[2,5,6]. Mutation types in the genomes of cancer cells include alterations of a single base, known as single base substitutions (SBS), and small insertions or deletions of bases (IDs). Larger scale events such as copy number variants (CNVs) and structural rearrangements / variants (SVs)[1] are also commonly observed in cancer genomes. Copy-number variants involve changes in the number of copies of a sequence of DNA ranging in size, from hundreds or thousands of bases up to entire chromosomes. Structural rearrangements / variants (SVs) are typically large scale events and include deletions, inversions, tandem duplications, translocations and complex rearrangements[1]. Extrachromosomal DNAs are circular DNA fragments containing regions of amplified oncogenes observed in several cancer types and are associated with increased expression of cancer oncogenes[11].

Somatic mutation types vary between cancers arising from different tissues[2,5]. For example skin melanoma has high levels of SBS whereas oesophageal adenocarcinomas commonly have high levels of SVs, although there is a correlation between the mutation burden of each mutation type across different cancers (Figure 1.1). It seems plausible that both normal and neoplastic tissues have different intrinsic mutation rates and repair proficiencies of different mutation types[2,12].

There is substantial variation in the burden of mutations seen in cancers from different individuals (Figure 1.1). A full understanding of the reasons for these differences eludes us. However, two important factors are known. First, differences in the number of age-related mutations that accumulated in the normal cells that precede the formation of the cancer, determined by the age of the individual at diagnosis explain some variation[12]. Second, sporadic mutational processes that are additional to ageing related mutational processes have heterogeneous contributions to the mutation burden across cancers from different individuals. A small proportion of samples may show exceptionally higher mutation burdens. In these samples, somatic mutations in DNA repair genes such as those involved in mismatch repair (MMR) or in the DNA polymerases may contribute to exceptionally high mutation burdens

(100,000s of SBS and ID mutations)[5] and are a potential source of increased inter-individual heterogeneity observed in certain cancer types.

Of the mutations that accumulate during the course of a cancer's development, only a very small proportion confer a selective advantage. These aberrations, which may potentially be of any mutation type, are called 'driver' mutations.

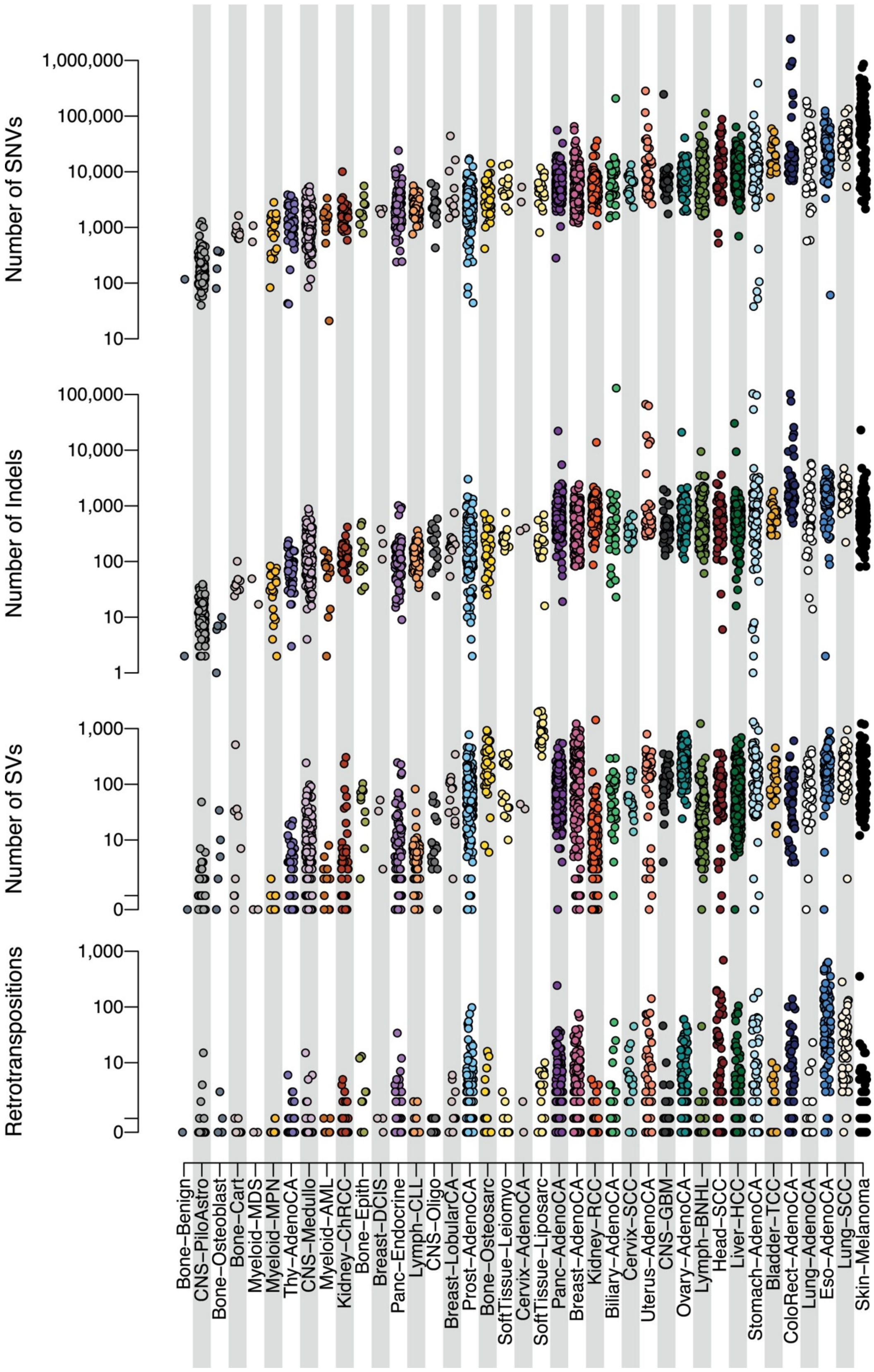

Figure 1.1 | Somatic mutation burdens in human cancer

Somatic mutation burdens organised by cancer type (x-axis) and mutation class (y-axis) and shown on a logarithmic scale. Reproduced from Campbell et al, Nature, 2020[2]

Drivers of cancer development and growth

The molecular events that confer a selective advantage to cancer cells by improving their competitive advantage or fitness are called cancer driver mutations, or just ‘drivers’. An estimate of the number of driver events, which genes they affect and how they alter cellular function is central to developing an understanding of the factors contributing to the growth and development of human cancer. Cancer drivers may be any type of mutation that affects cell function contributing to the cell’s malignant programme.

Characteristics of cancer driver genes

Cancer genes have been identified in most human cancer types. They can be classified according to whether they act in a dominant or recessive manner.

Dominantly acting cancer genes are activated by the acquisition of a driver mutation on one allele. These genes typically have a small number of mutational hotspots at which activating missense mutations are recurrently observed and may also be activated by in-frame insertion and deletion mutations[13]. Examples include the heterozygous missense mutations *KRAS*$^{\mathrm{G12D}}$ in bowel, pancreas and lung cancer and *BRAF*$^{\mathrm{V600E}}$ in thyroid, skin and bowel cancers.

Recessively acting cancer genes, also called tumour suppressor genes, are inactivated by mutations on both alleles. Tumour suppressor genes are usually inactivated by nonsense and frameshift mutations which generate premature stop codons. RNA transcripts carrying premature stop codons may be eliminated by nonsense-mediated decay or, if translated, produce proteins with partial or complete loss of function. Additionally, inactivation of a tumour suppressor gene may be caused by 1) independent truncating mutations on both alleles of a recessive gene or 2) loss-of heterozygosity (LOH), which is a chromosomal event caused by deletion of the wild-type allele in a tumour suppressor gene with a pre-existing germline or somatic heterozygous loss of function mutation. Tumour suppressor gene loss was first described by Knudson as part of his two-hit hypothesis in the context of retinoblastoma which is caused by biallelic inactivating mutations of the *RB1* gene[14]. Mutations in TSGs are found in many cancer types. An example of a canonical TSG is the *APC* gene which encodes the adenomatous polyposis coli protein and is mutated in the majority of colorectal cancers and in several other cancer types[15,16].

TSGs are typically thought of as requiring inactivation of both alleles to have an effect. Mutations affecting one allele of certain TSGs may result in haploinsufficiency, which is a reduction or alteration in protein function as a consequence of inactivation of one allele of a gene that is not sufficiently compensated by the remaining wild-type allele. Examples of cancer genes in which haploinsufficiency is thought to occur include: *NF1*, *TP53*, *PTEN* and *SMAD4*[17]. It is noteworthy that certain genes demonstrate aspects of both dominant and recessively acting oncogenes e.g. *NOTCH1* and *NOTCH2* (reviewed in Lobry et al 2011[18]).

Non-coding point mutations can also contribute to cancer development. Twenty-five percent of tumours carry non-coding driver mutations – mutations in regulatory regions including promoters and enhancers - the most common example of which are mutations in the *TERT* gene promoter, which is involved in telomere maintenance[2]. Cancer drivers need not be due to DNA mutations. Specific epigenetic alterations are also thought to be *bona fide* driver events e.g. *MLH1* promoter methylation in sporadic colorectal cancer[16].

Burden and frequency of cancer driver genes and mutations in cancer

Inference of the number of driver mutations using age-incidence curves led to the observation that 2-5 rate-limiting driver events are required for the development of many cancer types[19,20]. Identification of cancer driver mutations from genome sequencing data has corroborated these estimates, showing that cancers have an average of 4-5 driver mutations[2,21]. All types of mutation can act as drivers and the extent to which each does so depends on the cancer type, cancer gene and presence of specific mutational processes[2,15].

Over 700 genes are implicated in the development of cancer as listed in the COSMIC gene census[15]. These genes are classified into two levels: tier 1 and tier 2. Tier 1 genes (n=~570) have documented evidence of mutation in cancer and strong evidence for their role in the oncogenic transformation. Tier 2 genes (~170) have a documented role in cancer development albeit with less strong supporting evidence[15]. In a complementary effort, a similar number of cancer genes (n=568) were identified from a large assembly of human cancer genome data using the Integrative OncoGenomics (IntOGen) informatic pipeline[22].

Identification of driver mutations in cancer samples

The principal method used to identify driver mutations is the observation of recurrence of mutations at specific sites or in specific genes from different individuals, compared to expectation by chance. The observation of recurrence implies that the particular mutation-gene combination confers a beneficial trait in the affected cell which is positively selected for during the cancer's development, although the alternative interpretation of hypermutability cannot always be excluded. The identification of cancer driver mutations requires DNA sequencing data from a sizeable number of cancer samples of the same histological cancer type and identifies genes and mutations that are overrepresented in a tissue / cancer type. In theory, this approach can be applied to any mutation type. To date, it has been most extensively applied to the study of single base substitution driver mutations, which constitute the most common form of driver mutation[2].

The dNdS approach assesses the ratio of non-synonymous (protein-altering) mutations (dN) compared to the number of synonymous mutations (dS), which do not alter the protein product and hence are selectively neutral. The ratio is used to establish the strength of evolutionary selection[21]. Cancer genomes are principally subject to positive selection. Negative selection is rarely observed in the genome of cancer cells. However, it is prevalent in the human germline.

Several factors can bias the accumulation of mutations across the genome. Therefore, when evaluating the dN/dS ratio integration of genomic covariates and other factors is required to accurately define the expected mutation rate of a gene and thus reduce the possibility of incorrectly identifying or missing potential drivers[21,23].

Examples of factors that can bias the identification of genes that are subject to positive selection include:

1. Large genes that have high mutation rates by virtue of their size alone
2. Recurrent sequencing artefact
3. Biased accumulation of mutations across the genome e.g. euchromatin vs. heterochromatin and exonic vs. intergenic regions
4. Nucleotide context biases of the underlying mutational process

Cancer genes and specific tissue types

The number of driver mutations in cancer genes differs between cancer types. Mutations in some cancer genes are present in a high proportion of some cancers but are largely absent from others e.g. *IDH1* in glioma. By contrast, certain cancer genes are observed in multiple different cancer types e.g. *BRAF*, *KRAS* and *TP53*. *BRAF*^{V600E} is observed in a substantial proportion of papillary thyroid cancer, large intestine adenocarcinoma, hairy-cell leukaemia and skin melanoma. *KRAS*^{G12D} is present in lung, large intestine, biliary and pancreatic adenocarcinoma as well as endometrial, ovarian and peritoneal cancers[3]. Lastly, *TP53* mutations are prevalent in many cancer types, including oesophageal, colorectal, lung small cell, uterine, head and neck, and pancreatic carcinoma. The underlying factors contributing to the recurrent mutation of specific genes are not entirely clear. Several factors may explain some differences in the frequency of cancer gene mutations across different tissues: 1) gene expression 2) chromatin organization 3) endocrine and paracrine signalling 4) tissue microenvironment and 5) immune surveillance[15,24-26]. It is noteworthy that a substantial number of cancer genes demonstrate opposing roles in different tissues. Over 70 genes have tumour promoting roles in some tissues and tumour suppressing roles in others[15,21] suggesting that factors specific to cell and tissue types have an influence on the driver landscape of cancers[27,28]. For example *DNM2* which encodes dynamin 2 - a protein involved in endocytosis and forms part of the cytoskeleton - is thought to be an tumour suppressor in T-cell-acute lymphoblastic leukaemia and has a potential role as an oncogene in prostate cancer (Reviewed in Trochet et al 2021[29-31]).

Cancers often acquire driver mutations in characteristic sets of cancer genes e.g. *APC*, *KRAS* and *TP53* in colorectal adenocarcinoma and *EGFR*, *TP53* and *KRAS* in lung adenocarcinoma. In certain cancer types, the sets of genes may represent mutations sequentially acquired during the early phases of cancer development e.g. *APC*, *KRAS*, *TP53* and *PIK3CA* in intestinal neoplasia[32,33]. While the effect of certain driver mutations is in some cases well understood, the interplay between combinations of driver mutations in specific tissues is largely unknown.

Timing of cancer driver mutations

The timing of cancer driver mutations varies between individuals and tissues. These mutations may occur at any stage: during development[34,35], in early life[36] and in the middle and late decades of life[32,37]. Identification of the timing of the first driver mutations through the study of cancers and normal cells, may generate an improved understanding of the mutational processes that are operative and potentially enable the development of improved early diagnostics and improved prevention strategies.

Driver mutations may also be caused by somatic copy number alterations (sCNAs) and structural rearrangements[2]. sCNAs occur in a substantial proportion of human cancer[38,39]. The proportion of these that are drivers is not entirely clear[40]. However, it is estimated that sCNA driver mutations occur in approximately three quarters of human cancer[2,41]. Fusion genes are potential cancer driver mutations and are present in a substantial proportion of individuals with certain cancer types e.g. *BCR-ABL*; in 90-95% of chronic myeloid leukaemia, *TMPRSS2-ERG*; in 36% of prostate cancers and *EWSR1-FLI1*; in 70% of Ewing sarcoma.

Mutational signatures

Overview and introduction

Mutations accumulate in the cells and tissues of living organisms throughout life. Understanding the processes responsible for mutagenesis in healthy and diseased tissue can reveal important insights into the processes of DNA damage, DNA repair and other aspects of cellular physiology and help to explain how alterations in these processes contribute to the development of disease.

The objective of mutational signature analysis is to identify the mutational processes present in the genomes of the cells being studied. Each mutational signature may correspond to one or more mutational process. Investigation of the mutational processes can enable the identification of specific mechanisms underlying mutational signatures. At the time of writing, over 60 mutational signatures have been described and catalogued[3,6]. The current set of mutational signatures were principally discovered from mutations identified in human cancers[5,6].

Mutational signatures are represented as probability distributions summarising the relative frequency of a base-change categorised by the type of mutation. The COSMIC catalogue of mutational signatures has, to date, primarily focussed on categorisation of mutational signatures using a 96-context profile comprising 6 nucleotide changes (C>A, C>G, C>T, T>A, T>C and T>G) based on 16 permutations of trinucleotide contexts (5' and 3' bases). By convention this annotation is performed based on the pyrimidine annotation i.e. C:G>A:T mutations are represented as C>A.

Mutational signature analysis involves two steps: 1) signature extraction and 2) signature fitting. Signature extraction involves the identification of mutational signatures from the mutations present in the genomes to be analysed, without *a priori* knowledge of the mutational signatures that are present. This process can be employed to discover novel mutational signatures and identify the presence of known mutational signatures in an unbiased manner. Signature extraction requires data from multiple independent samples and works optimally when there is a degree of heterogeneity in the mutational processes present between samples. Signature fitting is the process of assessing the abundance of a set of known mutational signatures in the samples being analysed. This process seeks to fit the observed trinucleotide profile to that of a nominated set of mutational signatures.

Key stages in mutational signature analysis

Mutational signature analysis starts with a count matrix summarising the abundance of mutation types in samples i.e. each row represents a sample, and each column the mutational subtype. When analysing the 96-context mutational signature, each column will correspond to one mutation type/trinucleotide context combination e.g. C>A mutations at T$\underline{\text{C}}$T context (mutated base underlined). Counts of each mutation type in each sample are tallied. Next, mutational signature extraction is performed, in which a set of mutational signatures are identified *a priori* based on the relative proportions of each mutational subtype in each sample. The prevalent method used in signature extraction is non-negative matrix factorisation (NNMF)[42]. NNMF decomposes a matrix (F) into its composites (W) and (H). Matrix F contains the mutational spectra for each sample. Matrix W contains the mutational signatures i.e. the probability distribution of the trinucleotides for each mutational signature and matrix H contains the attributions, i.e. the proportion of each signature in each sample. By iterating over the problem many times, an optimised estimate of the number and abundance of mutational signatures and their activities can be established.

The implementation of NNMF differs between software packages. The following summarises the approach used in the SigProfiler package. First, SigProfiler applies NNMF to a matrix, F. To optimise the approximation of the number of matrices (W and H), and to minimise divergence, the Kullback-Liebler divergence is assessed[6,42]. One of the major challenges of NNMF is that the number of mutational signatures N is not known a priori. Therefore, to approximate the optimal solution, the NNMF process is repeated for multiple values of N. For each value of N, solutions are clustered and ranked. The optimal solution can then be selected by the operator based on its stability and associated accuracy metrics. Lastly, a hierarchical consensus clustering step is applied which aids the identification of less abundant, sporadic mutational signatures.

The next step is signature attribution / fitting which is the process of assessing the relative abundance of each identified mutational signature in each sample in the data set. This can be undertaken simultaneously with the signature extraction process or independently using a specific signature fitting packages such as the 'sigfit' algorithm which uses a Bayesian approach to approximate the abundance of a nominated set of signatures in a sample[43].

An alternative approach to signature extraction, which is also implemented in this thesis, is the Hierarchical Dirichlet Process (HDP)[44]. Using the same mutational count matrix (F) referred to above, this method infers the mutational signatures present in the whole data set based on knowledge of the hierarchical organisation of samples using a Dirichlet process. A Dirichlet process is a distribution of distributions i.e. each value in a distribution is itself approximated by a distribution. The Dirichlet process is used to model the distribution of mutational signatures in each sample based on the distribution of signatures in the parental nodes of the hierarchy. The HDP package then uses a Bayesian method (Gibbs Sampler) to assign mutations to individual clusters / signatures. This method has been used in studies in both cancer genomics[45] as well as normal tissue genomics[46-50].

Mutational signatures in cells and tissues

To date over 60 SBS mutational signatures have been identified from the analysis of sizeable cohorts of cancer genomes[5,6]. The mutational processes underlying many of these mutational signatures have been established; however, the aetiology of many signatures are yet to be identified[51] (Figure 1.2).

Two Single Base Substitution (SBS) mutational signatures; SBS1 and SBS5 are found ubiquitously in human cancers and in normal cells. The burden of these mutational signatures correlates with the age at cancer diagnosis[12] leading to the suggestion that these mutational processes operate in normal cells and accumulate in a continuous manner throughout life. SBS1 is due to deamination of 5-methylcytosine (5-mc) at NCG trinucleotides (mutated base underlined), previously termed CpG dinucleotides, causing C>T mutations. Approximately 70-80% of CG dinucleotides are methylated, deamination of these bases generates two mismatches U:G and T:G[52]. During subsequent replication U and T bases bind with adenines. Whilst U is recognisable as an abnormal base in DNA and typically repaired, T bases are normal constituents of DNA and hence may be unrepaired resulting in a C>T / G>A transition mutation. SBS5 is a ‘flat’ signature affecting all six SBS mutation types (C>A, C>G, C>T, T>A, T>C and T>G). Its aetiology is unknown and it is thought to possibly represent a composite of multiple mutational processes.

SBS18 is characterised by C>A transversion mutations predominantly at ACA, CCA, GCA and TCT trinucleotides and is associated with DNA damage due to reactive oxygen species[53]. SBS18 is observed in abundance in paediatric neuroblastoma and is also observed in several other cancer types. SBS18 is associated with *MYCN* amplification and 17q gain in neuroblastoma, which are thought to alter mitochondrial gene expression, resulting in the generation of reactive oxygen species[54]. SBS18 is also induced *in vitro* following exposure to free-radical producing agents such as hydroxyl radical producing chemicals[53].

Defective DNA repair is a known cause of several mutational signatures: SBS3, SBS6, SBS10a, SBS10b, SBS14, SBS15, SBS20, SBS21, SBS26, SBS30, SBS36 and SBS44. SBS3 is due to defective homologous repair of double-strand DNA breaks and is associated with germline and somatic mutations in BRCA1 and BRCA2. Seven signatures are associated with defective DNA mismatch repair: SBS6, SBS14, SBS15, SBS20, SBS21, SBS26 and SBS44. Of these, two signatures, SBS14 and SBS20, are associated with defective DNA mismatch repair and concurrent mutations in the DNA polymerases[55]. DNA mismatch repair mutational signatures are responsible for some of the highest mutation burdens observed in human cancers including substantial increases in the number of small insertion and deletion mutations. Microsatellite instability is a common feature in cancers with DNA mismatch repair deficiency and is usually present in samples with these mutational signatures.

SBS30 is caused by defective NTHL1 which is a DNA glycosylase involved in the base excision repair of oxidative damage with a particular preference for the repair at cytosine bases. This signature is characterised by C>T transversion mutations. Inherited *NTHL1* mutations cause a hereditary cancer syndrome with an elevated risk of polyposis and colorectal cancer.

SBS36 is caused by defective MUTYH DNA glycosylase which is involved in the removal of adenine bases mispaired opposite 8-oxoguanine as part of the base excision repair pathway. SBS36 is characterised by C>A transversion mutations with distinctive peaks at TCA and TCT trinucleotides. It has been identified in colorectal polyps, colorectal cancers and adrenocortical cancers with biallelic germline and somatic mutations in the *MUTYH* gene. SBS36 bears similarity to SBS18 and these two signatures are often observed together in cancers with *MUTYH* mutations[56-58].

Environmental mutagens cause a variety of mutational signatures[53]. In cancer, the most commonly observed mutational signatures associated with an environmental mutagen are SBS7a, SBS7b, SBS7c and SBS7d which are attributable to ultraviolet (UV) light exposure. These mutational signatures are frequently observed in skin cancer and result from the repair of UV photoproducts. They show an excess of mutations on the un-transcribed strand which reflects the role of transcription-coupled nucleotide excision repair in the resolution of UV light damage. Other less common environmental and occupational mutagens that are observed in cancer include aristolochic acid, aflatoxin and haloalkanes, which cause SBS22, SBS24 and SBS42 respectively.

Inhaled and chewing tobacco cause SBS4 and SBS29 respectively. SBS4 is seen in lung, head and neck and biliary cancer from individuals with a history of tobacco smoking[59] and is attributed to the effects of tobacco smoke carcinogens including benzo[a]pyrene. Interestingly, SBS4 is not observed in several cancer types in which smoking is a recognised risk factor including colorectal and renal cancer, implying that other carcinogenic mechanisms may underlie the increased cancer risk in these cancer types.

Chemotherapeutics used in the treatment of cancer and non-malignant conditions cause several mutational signatures including: SBS11 (temozolomide), SBS31 & SBS35 (platinum therapies) and SBS32 (azathioprine).

SBS17a and SBS17b are characterised by T>C and T>G mutations respectively. The mechanism of SBS17a and SBS17b is not known. However, in a proportion of breast and colon cancers SBS17b is associated with exposure to the chemotherapy agent 5-flourouracil / capecitabine[60].

SBS88 is caused by colibactin producing *E.coli*[61]. SBS88 was first identified in normal intestinal epithelial cells[46] and head and neck cancer[62]. It has subsequently been implicated as a possible mutagen in colorectal cancer[61]. SBS88 is the first mutational signature to be attributed to the effects of a bacterial mutagen.

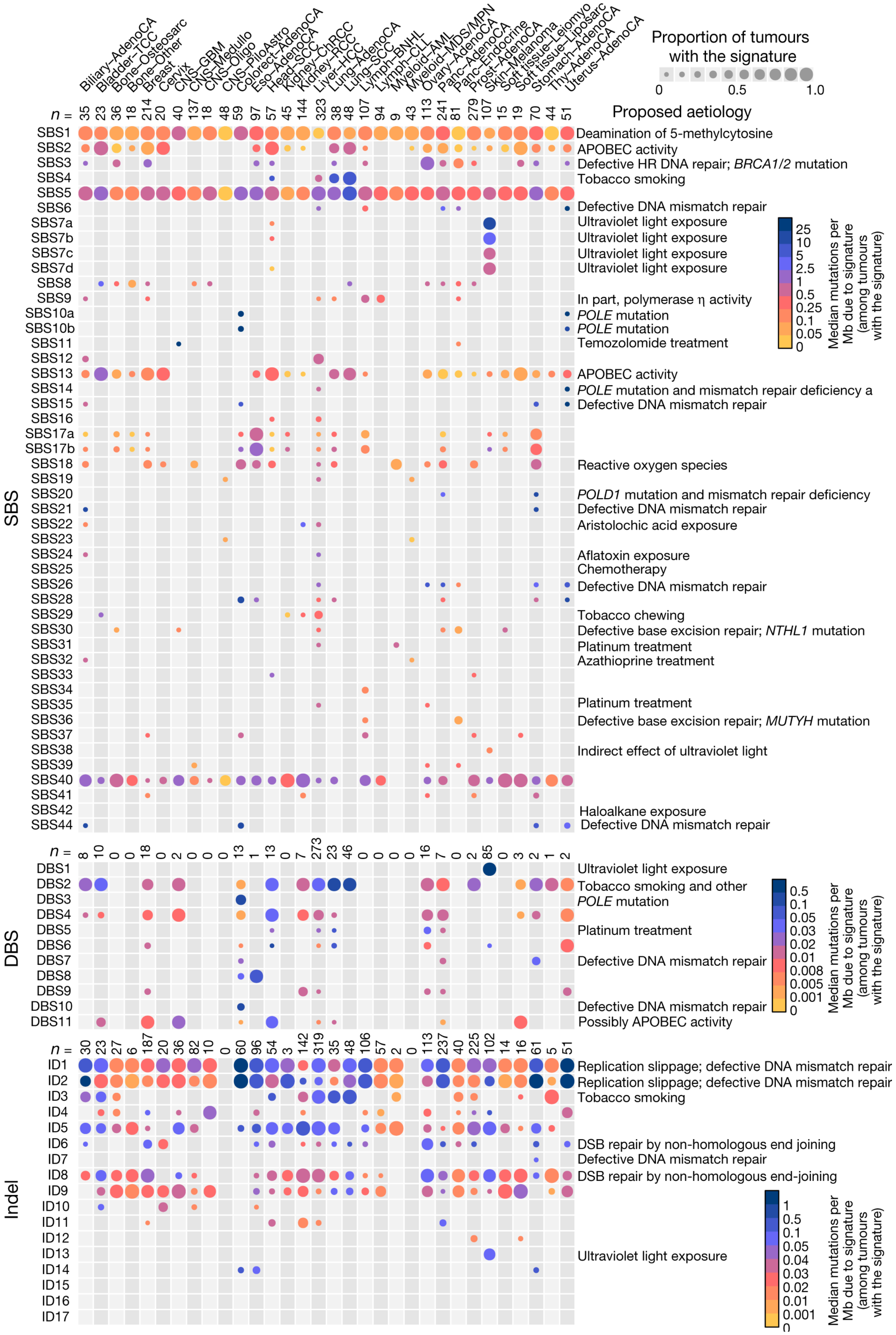
Biliary–AdenoCA
Bladder–TCC
Bone–Osteosarc
Bone–Other
Breast
Cervix
CNS–GBM
CNS–Medullo
CNS–Oligo
CNS–PiloAstro
Colorect–AdenoCA
Eso–AdenoCA
Head–SCC
Kidney–ChRCC
Kidney–RCC
Liver–HCC
Lung–AdenoCA
Lung–SCC
Lymph–BNHL
Lymph–CLL
Myeloid–AML
Myeloid–MDS/MPN
Ovary–AdenoCA
Panc–AdenoCA
Panc–Endocrine
Prost–AdenoCA
Skin–Melanoma
Soft tissue–Leiomyo
Soft tissue–Liposarc
Stomach–AdenoCA
Thy–AdenoCA
Uterus–AdenoCA
Proportion of tumours with the signature
0
0.5
1.0
Proposed aetiology
n = 35 23 36 18 214 20 40 137 18 48 59 97 57 45 144 323 38 48 107 94 9 43 113 241 81 279 107 15 19 70 44 51
SBS1 Deamination of 5-methylcytosine
SBS2 APOBEC activity
SBS3 Defective HR DNA repair; BRCA1/2 mutation
SBS4 Tobacco smoking
SBS5
SBS6 Defective DNA mismatch repair
SBS7a Ultraviolet light exposure
SBS7b Ultraviolet light exposure
SBS7c Ultraviolet light exposure
SBS7d Ultraviolet light exposure
SBS8
SBS9 In part, polymerase η activity
SBS10a POLE mutation
SBS10b POLE mutation
SBS11 Temozolomide treatment
SBS12
SBS13 APOBEC activity
SBS14 POLE mutation and mismatch repair deficiency a
SBS15 Defective DNA mismatch repair
SBS16
SBS17a
SBS17b
SBS18 Reactive oxygen species
SBS19
SBS20 POLD1 mutation and mismatch repair deficiency
SBS21 Defective DNA mismatch repair
SBS22 Aristolochic acid exposure
SBS23
SBS24 Aflatoxin exposure
SBS25 Chemotherapy
SBS26 Defective DNA mismatch repair
SBS28
SBS29 Tobacco chewing
SBS30 Defective base excision repair; NTHL1 mutation
SBS31 Platinum treatment
SBS32 Azathioprine treatment
SBS33
SBS34
SBS35 Platinum treatment
SBS36 Defective base excision repair; MUTYH mutation
SBS37
SBS38 Indirect effect of ultraviolet light
SBS39
SBS40
SBS41
SBS42 Haloalkane exposure
SBS44 Defective DNA mismatch repair
SBS
25
10
5
2.5
1
0.5
0.25
0.1
0.05
0
Median mutations per Mb due to signature (among tumours with the signature)
n = 8 10 0 0 18 0 2 0 0 0 13 1 13 0 7 273 23 46 0 0 0 0 16 7 0 2 85 0 3 2 1 2
DBS1 Ultraviolet light exposure
DBS2 Tobacco smoking and other
DBS3 POLE mutation
DBS4
DBS5 Platinum treatment
DBS6
DBS7 Defective DNA mismatch repair
DBS8
DBS9
DBS10 Defective DNA mismatch repair
DBS11 Possibly APOBEC activity
DBS
0.5
0.1
0.05
0.03
0.02
0.01
0.008
0.005
0.001
0
Median mutations per Mb due to signature (among tumours with the signature)
n = 30 23 27 6 187 20 36 82 10 0 60 96 54 3 142 319 35 48 106 57 2 0 113 237 40 225 102 14 16 61 5 51
ID1 Replication slippage; defective DNA mismatch repair
ID2 Replication slippage; defective DNA mismatch repair
ID3 Tobacco smoking
ID4
ID5
ID6 DSB repair by non-homologous end joining
ID7 Defective DNA mismatch repair
ID8 DSB repair by non-homologous end-joining
ID9
ID10
ID11
ID12
ID13 Ultraviolet light exposure
ID14
ID15
ID16
ID17
Indel
1
0.5
0.1
0.05
0.04
0.03
0.02
0.01
0.001
0
Median mutations per Mb due to signature (among tumours with the signature)

Figure 1.2 | Mutational processes in human cancers.

Plot shows the frequency and abundance of mutational processes in human cancer types. The size of the coloured dots indicates the proportion of tumour samples with the mutational process and the colour indicates the mutation rate attributable to each signature. Processes with a known aetiology are annotated in the right hand margin of the table. Reproduced from Alexandrov et al, *Nature*, 2020[6].

Eighteen ID mutational signatures have been identified to date. The most common processes are ID1 and ID2, which are caused by strand slippage during DNA replication. ID1 is predominant in many normal tissues, whereas ID2 is the principal ID signature in cancer. Eleven doublet base substitution (DBS) signatures have been identified to date, with several of known aetiology. Associations are observed between specific mutational signatures of different mutational classes e.g. ageing / clock-like signatures: SBS1 & SBS5 & ID1, platinum therapy exposure: SBS35 & DBS5, colibactin exposure: SBS88 & ID18 and UV light: SBS7a & DBS1.

Temporal dynamics of mutational processes

Of the known SBS mutational processes two, SBS1 and SBS5, are thought to act throughout life in all normal cells and also accrue in neoplastic cells. A further mutational signature, SBS18 also appears to accumulate in a linear manner with age but is only present in some normal cell and cancer types[6,50]. Mutational processes may also be continuous in cells with mutations in DNA repair genes, such as the DNA polymerases and *MUTYH*, which are characterised in this thesis. In contrast, other mutational processes act on the genome in an episodic manner. Mutational signatures associated with environmental exposures e.g. SBS88 due to mutagenic *E.coli*, or chemotherapy e.g. SBS35 due to platinum therapy, occur around the time of mutagen exposure. Mutagenesis due to APOBEC cytosine deaminases generates mutational signatures SBS2 and SBS13 which are thought to occur in an episodic manner[63]. They also can produce a characteristic mutational phenomenon called *kataegis*, which is a spatially and temporally localized pattern of hypermutation[2,64,65].

Extension of the mutational signature analysis frame work

While the most familiar form of mutational signature analysis uses the 96-context classification, several analysis tools permit the analysis of the extended sequence context incorporating other DNA bases that are beyond the immediate trinucleotide context. These include 1536-category analysis which accounts for the pentanucleotide extended sequence context (-2 to +2 bases). Other genomic data can also be investigated with mutational signature analysis. For example, signatures incorporating the transcriptional strand bias (288-context) or replication strand bias (192-context) can be analysed using the mutational signature framework. At present, reference signatures incorporating these and other covariates are at a relatively early stage in their development.

In addition to SBS, ID and DBS, mutational signature analysis has also been applied to other mutation types. Signature analysis of genomic rearrangements / structural variants (SV)[66] and copy number changes[67] have been conducted. However, a ratified set of reference signatures for these mutation types is yet to be published.

Lastly, novel mutational signature analysis tools have been developed that permit the analysis of data sets with heterogeneous mutational signatures incorporating their genomic covariates. One example is the TensorSignatures package which extracts mutational signatures incorporating their genomic features[68]. Tensors are mathematical objects that can contain multidimensional data. Just as vectors represent data in 1-dimension, and matrices in 2-dimensions (x & y), tensors represent data in n-dimensions where n>2. Therefore, in the context of signature analysis they can be employed to simultaneously incorporate several covariates into the mutational signature extraction process. This approach offers particular benefits when signatures are not easily differentiated on the basis of the 96-context annotation, and may permit the identification of characteristics that can discriminate between similar looking signatures. This analytical approach is also helpful when analysing cohorts with multiple heterogeneous mutational processes and is particularly adept at identifying mutational signatures with specific genomic characteristics e.g. those with unusual genomic distribution or distinctive strand biases[68,69]. This multi-dimensional approach represents a valuable method to characterise different and diverse mutational signatures in an unbiased manner, and may signal a ‘direction of travel’ for analysis tools in this scientific field.

Normal colonic epithelium

The colon is lined by a simple columnar epithelium, a thin layer of cells that serves as a barrier between the contents of the bowel lumen and the underlying bowel wall. The colonic epithelium is organised into microscopic crypts, each of which contains approximately 2000 cells. The cell population of the intestinal crypt is hierarchically organised, with pluripotent intestinal stem cells residing at its base, and transit amplifying cells and enterocytes towards the lumen[70-72]. Intestinal stem cells divide every two to three days to ensure that the population of differentiated cells in the crypt are maintained[73]. Renewal of the entire differentiated enterocyte cell population of the crypt is thought to occur every seven days[73].

It is estimated that there are ~15 million crypts in the human colon. The number of crypts is believed to stay constant throughout life; however, it may vary between individuals owing to factors such as height[74]. Intestinal crypts periodically divide to form two daughter crypts in a process termed crypt fission. This process occurs physiologically during foetal development to populate the colonic epithelium and continues throughout life to maintain the integrity of the intestinal epithelium[75-79]. The rate of crypt fission in healthy adult colon is very low, occurring once every ~27 years per crypt[46,79]. An additional process, crypt fusion, has recently been described in humans[79]. It is the joining of two adjacent intestinal crypts. In healthy epithelium the rate of crypt fission and fusion are similar thus maintaining a relatively constant number of crypts throughout life[78]. Heterozygous germline mutations in the tumour suppressor gene, *APC*, cause the cancer predisposition syndrome familial adenomatous polyposis (FAP), and are thought to increase the rate of both crypt fission and fusion[79]. The rate of crypt fission is also increased as a consequence of inflammation in colonic epithelium from individuals with inflammatory bowel disease (IBD)[80]. It is proposed that the increased crypt fission rate in FAP and IBD may contribute to the elevated risk of colorectal cancer associated with these conditions.

Colorectal cancer

General introduction

Colorectal cancer is the third commonest malignancy in the UK and is responsible for 10% of all cancer-related deaths[81]. The incidence of colorectal cancer has remained largely unchanged over the past 20 years in the UK. The risk of colorectal cancer climbs steeply after the fifth decade of life and peaks between 85 and 89 years of age.

Risk factors for colorectal cancer

Most colorectal cancers (>95%) are sporadic i.e. have no identifiable mendelian genetic cause. The main risk factor for colorectal cancer is age. It is thought that increasing burdens of mutations accumulate in intestinal stem cells as we age, contributing to the risk of cancer development[12,46,82,83]. Other risk factors for colorectal cancer include: obesity, smoking tobacco, increased alcohol consumption, inflammatory bowel disease, a history of radiation exposure, immunosuppression, low levels of physical activity, a diet low in fibre and high in processed meat and a first-degree family history. The factors that contribute to colorectal cancer are therefore many and varied. Many of the colorectal cancer risk factors are modifiable: hence a substantial burden of cases are considered to be potentially preventable[84].

Diagnosis and staging of colorectal cancer

Each year, in excess of 42,000 individuals are diagnosed with colorectal cancer in the UK[85]. The majority of suspected cases are identified in primary care and approximately 15% are identified as part of the Bowel Cancer Screening Programme[86]. Early diagnosis is associated with improved survival[87]. Detection and removal of pre-cancerous adenomas via the Bowel Cancer Screening Programme (BCSP) is attributed with identifying over 25,000 cases of colorectal cancer since its inception in 2006[88].

Suspected colorectal cancer is investigated and diagnosed using endoscopy. Colorectal cancer is staged according to TNM (Tumour, Node and Metastasis) and Duke's systems which include assessment of the extent of local growth, involvement of lymph nodes and the presence of spread / metastatic disease. Prognosis and response to treatment are strongly predicted by the tumour stage.

Prognostic and predictive factors in colorectal cancer

Additional factors may inform the prognosis and management of individuals with colorectal cancer including the microsatellite status and presence of mutations in specific genes i.e. *BRAF* and *KRAS*. Approximately 15% of colorectal cancers are microsatellite unstable (MSI-H / MSI-high). Microsatellites are repetitive DNA sequences that are present throughout the genome. Their highly repetitive sequences make them prone to the accumulation of errors when they are replicated by the DNA polymerases. In normal and cancer cells, the DNA mismatch repair machinery effectively identifies replication-related errors including those affecting microsatellites. However, in a small proportion of cancers, acquired mutations in the mismatch repair machinery reduce the ability of the mismatch repair machinery to identify and repair mutations in microsatellite sequences. In affected cancers, high mutation burdens may result, which may be detected through the accumulation of mutations in microsatellites termed microsatellite instability. Microsatellite instability (MSI) is pathognomonic for DNA mismatch repair deficiency and is usually caused by somatic mutations or epigenetic inactivation of four genes that encode components of the DNA mismatch repair pathway: *MLH1*, *MSH2*, *MSH6* and *PMS2*. Cancers with these mutations tend to have substantially elevated tumour mutational burden[16]. MSI colorectal cancers have a favourable prognosis in early stage disease and respond to cancer immunotherapy which is now indicated in certain clinical circumstances[89].

The RAS family of genes are key effectors of intracellular signalling transduction responsible for cell growth and cell fate. Mutations in the *KRAS* and *BRAF* oncogenes are observed in 40% and 10% of colorectal cancers respectively[3,90]. *KRAS* mutations are typically an early event in colorectal cancer development: they are observed in 24% of adenomas[3]. *BRAF* mutations are more common in right-sided colorectal cancers often contributing to an alternative molecular pathogenesis of colorectal cancer known as the serrated pathway. *KRAS* and *BRAF* mutations predict outcome following anti-EGFR therapy with either cetuximab and panitumumab[91]. Several additional targeted systemic therapies are approved or under evaluation for specific clinical indications in the management of colorectal cancers with specific gene mutations.

The adenoma-carcinoma sequence in colorectal cancer development

Colorectal cancer arises from normal intestinal stem cells[92]. The progression from normal intestinal mucosa to adenocarcinoma is known as the adenoma-carcinoma sequence and describes the pathway by which most colorectal cancers are thought to develop[93]. It describes the development from a normal crypt to an early adenoma, growth and development of the adenoma and subsequently, the development of adenocarcinoma[93]. The landmark stages in the adenoma-carcinoma sequence are accompanied by the acquisition of mutations in specific cancer genes: *APC*, *KRAS*, *SMAD4* and *TP53*. *APC*, is mutated in the majority of early adenomas. *KRAS* mutations are found in 10% of early adenomas and 40% of large adenomas[93]. Lastly, *TP53* mutations which are observed in a small proportion of late adenomas and ~40-50% of colorectal adenocarcinoma. A substantial proportion of colorectal adenocarcinomas demonstrate other molecular defects and trajectories. Alternative molecular pathways are demonstrated by *BRAF*-mutant and CpG Island Methylator Phenotype (CIMP) colorectal cancers which are more commonly observed in right-sided colorectal cancers[94].

The events that precede the development of adenomas in colorectal epithelium are not well understood. The earliest histologically visible microneoplasm in the colon is the aberrant crypt focus (ACF). ACFs are observed in normal colonic epithelium in healthy individuals and individuals with colorectal cancer and may harbour oncogenic mutations in genes including *APC*, *KRAS* and *BRAF*[95-97]. Whether these lesions are a precursor to adenoma development is highly debated[95].

APC gene mutations in colorectal cancer

The Adenomatous Polyposis Coli (APC) gene is a tumour suppressor with an essential role in the WNT signalling pathway which regulates: 1) cell proliferation, 2) stem cell maintenance / cell fate specification and 3) cell migration. Mutations in APC are thought to initiate the development of colorectal neoplasia[93,98-100]. Somatic mutations in APC are identified in 80-85% of human colorectal cancer[16] and are also observed at lower frequencies in other cancer types[3]. Consistent with its role as a tumour suppressor gene, >95% of truncating mutations in *APC* are estimated to be drivers whereas missense mutations in *APC* are not thought to be deleterious[21]. Copy number loss of the *APC* locus on chromosome 5q is observed in approximately 20% of colorectal cancers[16]. The most common mechanism of *APC* inactivation involves two point mutations or a point mutation and deletion[16]. Large scale structural rearrangements are not a common cause of *APC* inactivation in colorectal cancer.

Copy number changes and structural rearrangement in colorectal cancer

Aneuploidy is a common feature in colorectal cancer. Chromosomal changes include gains of 1q, 7p and q, 8p and q, 12q,13q,19q, and 20p, as well as losses of 1p, 4q, 8p,14q,15q, 17, 20p and 22q[16,38,101]. Several of these changes are also observed in adenomas prior to the formation of colorectal cancer[90,102]. P53 is a key regulator of genome integrity and acquired mutations in this gene confer tolerance to chromosomal instability, aneuploidy and the generation of copy number changes[33,103]. *TP53* mutations typically arise in advanced adenomas and early stage colorectal carcinomas where they hasten the accumulation of somatic copy number changes[90]. Structural rearrangements in colorectal cancer are dominated by the presence of breaks at fragile sites and tandem duplications: other rearrangement types are rare[66]. Fusion genes are observed in colorectal cancer. However, they are not thought to play a role in the early development of colorectal neoplasia[16].

Hypermutation in colorectal cancer

Hypermutation, typically defined as >12 mutations/Mb, is observed in approximately 10% of colorectal cancers[16]. Hypermutation is commonly caused by acquired biallelic inactivation of components of the DNA MMR machinery or heterozygous mutations in the DNA polymerase mutations, *POLE* and *POLD1*. The most frequent MMR defect is the silencing of the mismatch repair gene, *MLH1*, which is associated with the development of microsatellite instability[16].

Somatic mutations in MMR genes and the DNA polymerases *POLE* and *POLD1* may also cause hypermutation. Hypermutation is associated with a different set of cancer driver gene mutations when compared to non-hypermutator colorectal cancer[16]. This may be caused by the particular SBS and ID mutational processes that are present in cells with MMR deficiency and DNA proofreading defects.

Colorectal cancer predisposition

Introduction

Inherited causes account for less than 5% of colorectal cancer cases[104]. By far the most common inherited cause of colorectal cancer is Lynch syndrome, which is responsible for up to 4% of all colorectal cancer cases[104-106]. Familial adenomatous polyposis (FAP) causes less than 1% of colorectal cancer[104]. Other important yet rare causes of colorectal predisposition include: *MUTYH*-associated Polyposis (MAP), juvenile polyposis syndrome (JPS), proofreading polymerase associated polyposis (PPAP) and *NTHL1*-associated polyposis. Colorectal cancer risk is also modestly increased to various extents in Cowden's syndrome, Bannayan-Riley-Ruvalcaba syndrome and Peutz-Jeghers syndrome (PJS)[107].

Multiple genes are associated with inherited predisposition to polyposis and intestinal cancer (Table 1.1.) and can be categorised into three groups:

1. Regulators and effectors of WNT signalling e.g. *APC, AXIN2* and *CTNNB1*
2. Effectors of DNA repair and replication e.g. MMR genes, *POLE*, *POLD1*, *NTHL1* and *MBD4*
3. Other regulators of cell signalling e.g. *PTEN, SMAD4 and STK11*

A substantial proportion of individuals with intestinal polyposis have no identifiable causative germline mutation[108,109]. There is substantial ongoing interest in identifying novel genes responsible for inherited causes of intestinal polyposis. An example of a recently discovered, albeit rare intestinal polyposis causing gene is *MBD4*, which is associated with colorectal polyposis, an increased risk of colorectal cancer and acute myeloid leukaemia[69,110].

The development of intestinal polyps is a feature that is common to many inherited intestinal neoplasia syndromes. Unusually among intestinal predisposition syndromes, individuals with Lynch syndrome do not develop polyps at an increased rate compared with healthy individuals, and this observation is alluded to in the condition's alternative name, hereditary non-polyposis colorectal cancer (HNPCC). By contrast, the next most common intestinal cancer syndromes, FAP and MAP, are characterised by the development of tens, hundreds or thousands of intestinal adenomatous polyps.

Syndrome	Gene	Pathway
Lynch Syndrome	*MLH1 / PMS2 / MSH2 / MSH6 / EPCAM*	MMR DNA repair
FAP	*APC*	WNT signalling
MAP	*MUTYH*	BER DNA repair
PPAP	*POLE / POLD1*	DNA proof reading
NTHL1 associated polyposis	*NTHL1*	BER DNA repair
Juvenile polyposis syndrome	*SMAD4 / BMPR1A*	TGF-β signalling
MBD4-deficiency	*MBD4*	BER DNA repair
Cowden’s syndrome	*PTEN*	multiple
Peutz-Jeghers syndrome	*STK11*	multiple

Table 1.1 | Summary of intestinal cancer predisposition syndromes and associated inherited gene defects

MMR and Lynch syndrome

DNA damage arising from errors of DNA replication is predominantly repaired by the MMR pathway. MMR is a highly conserved DNA repair pathway. In eukaryotes it is principally effected by four genes: *MLH1*, *MSH2*, *MSH6* and *PMS2*. MMR proteins form two complexes which associate PCNA, the motor of the replication machinery and the DNA polymerases, which are the effectors of DNA synthesis. The MutL homologue complex, comprising MLH1 and PMS2, undertakes DNA mismatch sensing and recruits the MutS homologue complex, comprising MSH6 and MSH2, to initiate repair.

Lynch syndrome (LS) is caused by inherited heterozygous inactivation of one of the MMR genes. It is an inherited cancer predisposition characterised by an increased risk of early onset colorectal, endometrial and ovarian cancer. Lynch syndrome has a strong penetrance with a 40-80% lifetime risk of developing colorectal cancer and similar risks of endometrial cancer[111-115]. The rate of generation of adenomas in LS is comparable to unaffected individuals, leading to the suggestion that cancer risk in LS may act through an accelerated progression of adenomas into colorectal cancer[116,117]. The majority of adenomas (77%) in individuals with LS are MMR deficient and colorectal driver mutations acquired in adenoma development typically demonstrate an MMR deficient mutational signature suggesting that the mutational processes associated with deficient MMR are often present during adenoma development[117].

Mutations in the four main MMR genes: *MLH1*, *MSH2*, *PMS2* and *MSH6*, account for the majority of LS cases. Inherited *MLH1* mutations account for 60% of LS cases, *MSH2* 30% and *MSH6* and *PMS2* 10%[118,119]. LS is also described in individuals with mutations in the epithelial cell adhesion molecule gene, *EPCAM*, which is adjacent to *MSH2*. Deletions in *EPCAM* can affect MSH2 expression and thus cause LS. Second-hit mutations in LS are typically caused by loss of the unmutated allele and may also, less commonly, be caused by small insertion or deletion mutations or single-base substitutions[120,121].

Diagnosis of individuals with LS is made using molecular testing[122]. Endoscopy is a mainstay in the surveillance and management of colorectal cancer risk in LS. However, compared with screening in low-risk populations, individuals with LS have a more modest benefit from endoscopic screening and polypectomy[123,124]. Screening is also indicated for surveillance in other cancer-prone organs in individuals with LS[125].

Important differences are observed between sporadic and LS colorectal adenocarcinoma. LS cancers have high tumour mutational burdens, with elevated levels of both SBS and ID mutations[126]. LS colorectal cancers and sporadic colorectal cancers with mutations in MMR gene mutations have higher levels of tumour lymphocytes[127] and many have increased expression of immune checkpoint markers[128,129]. This has led to the introduction of immune checkpoint inhibitors in the treatment of metastatic cancer from individuals with LS / MSI-H colorectal cancer[89].

Lastly, inherited biallelic mutation in MMR genes cause a very rare syndrome associated with very early-onset of multiple cancer types with a high risk of cancer-associated mortality in the early decades of life[130].

Adenomatous Polyposis Coli (*APC*) and Familial Adenomatous Polyposis (FAP)

FAP is an autosomal dominant condition caused by inherited mutations in the tumour suppressor gene *APC*. The syndrome is characterised by the development of hundreds to thousands of colonic adenomas starting in the second and third decades of life. In the absence of appropriate surveillance and surgical intervention, FAP carries a 100% lifetime risk of developing colorectal cancer[131]. FAP affects an estimated 1:7000 to 1:20000 individuals in the UK. Whilst most cases of FAP are associated with inherited mutations, 20-30% of individuals with FAP result from *de novo APC* mutations[131-133]. Individuals with FAP inherit a mutation in one allele of the *APC* gene. During life, some cells in at-risk tissues acquire a second-hit mutation in *APC*, resulting in biallelic inactivation and the initiation of neoplasia. The *APC* gene and development of FAP follows the pattern of the classic description of a tumour suppressor gene as formulated in Knudson's two-hit hypothesis.

FAP is associated with two main clinical phenotypes; classic FAP (FAP) and attenuated FAP (AFAP). Classic FAP is defined by a colonic polyp burden in excess of 100 polyps whereas individuals with attenuated FAP typically have fewer than 100 adenomas, often 30-50. The anatomical distribution of polyps in FAP is greater on the left / descending colon than the right/ascending colon. Conversely, individuals with AFAP tend to have more right-sided adenomas. The cause of the anatomical distribution observed in these two conditions is not clearly understood. Intestinal polyps in FAP are typically adenomas; either tubular or villous in architecture. Other histological types are uncommon. The most common histological type of cancer in FAP is adenocarcinoma.

The location of the germline mutation within the *APC* gene affects the clinical phenotype including the polyp burden of individuals with FAP. Mutations in certain parts of the N-terminus, affecting the armadillo repeats and the mutation cluster region (MCR) are associated with the classic FAP phenotype whereas mutations in the C-terminus and parts of the N-terminus are associated with an attenuated FAP phenotype (AFAP).

Polyposis in FAP also affects the duodenum and stomach. Duodenal polyposis occurs in approximately 90% of individuals with FAP and has a high risk of progression to duodenal adenocarcinoma[134]. The risk factors for the interindividual variation in the development of duodenal disease in individuals with FAP are not fully understood. In the stomach, fundic gland polyps predominate affecting 30-90% of individuals.

In addition to intestinal neoplasia, FAP is associated with a range of extra-intestinal disease manifestations including congenital hypertrophy of the retinal pigment epithelium (CHRPE); desmoid tumours; osteomas; as well as certain other cancers including adrenal adenoma; childhood hepatoblastoma; thyroid cancer; and central nervous system tumours[135]. The prevalence of extra-intestinal manifestations is more common in FAP than AFAP[135].

Endoscopic surveillance is a cornerstone in the management of FAP [122]. In classic FAP, the definitive treatment is prophylactic colectomy, with or without removal of the rectum. Colectomy in early adulthood mitigates most or all of the risk of colorectal cancer in individuals with FAP, depending on the type of resection[136]. In AFAP, where the polyp burdens are substantially lower, routine surveillance combined with polypectomy is often sufficient. Other anatomical sites also require regular surveillance, including the duodenum, which is monitored using regular endoscopy, and the frequency of surveillance is guided by the Spiegelman scoring system[122,137].

Medical management of FAP is at present limited. Medications including non-steroidal anti-inflammatory (NSAID) agents and targeted small molecule inhibitors have been the subject of clinical trials in patients with FAP[138,139]. The NSAID Sulindac reduces the overall polyp burden but does not alter the risk of progression to colorectal cancer[140]. Use of Sulindac as a chemopreventative agent in the paediatric population, prior to the onset of endoscopically visible polyposis, does not prevent the development of polyps[141]. Use of cyclooxygenase-2 (COX-2) inhibitors has shown modest reductions in polyp burden when administered to individuals with FAP but did not prevent progression to colorectal cancer[142]. Hence, the current recommendation is that monotherapy chemoprevention does not have a role in FAP[143]. Combination therapies show some promise in reducing adenoma burden both in the duodenum and colorectum[138] and the results of further studies are awaited.

MUTYH and MUTYH-Associated Polyposis (MAP)

Base-excision repair and DNA glycosylases

Reactive oxygen species cause DNA damage and are a threat to the integrity of the cellular genome. Oxidative species are generated as a result of multiple cellular processes and cell-extrinsic processes Oxidative damage causes a wide variety of DNA lesions[144]. The most common alteration is 8-oxoguanine (8-OG)[145]. Repair of 8-oxoguanine is conducted by the base excision repair pathway (BER) and is effected by the DNA glycosylases oxoguanine glycosylase[146] (OGG1) and MutY DNA glycosylase[147] (MUTYH). Under normal circumstances OGG1 excises 8-OG and MUTYH removes adenine mispaired opposite 8-OG. If unrepaired, 8-OG is misrecognised as thymine by DNA polymerases during DNA replication. Thus, it is paired opposite an adenine base. Following subsequent replication the 8-OG base is replaced with a thymine base and thus a G:C>T:A (C>A) mutation results[148] (Figure 1.3).

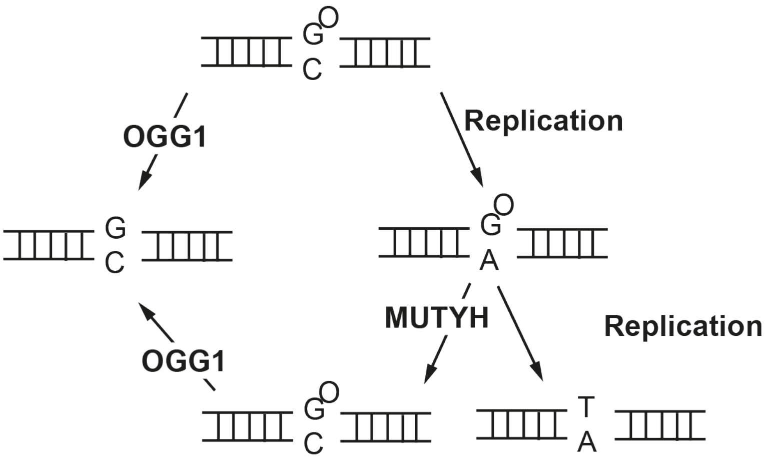

Figure 1.3 | Schematic overview of the role of DNA glycosylases, MUTYH and OGG1 in the repair of 8-OG.

8-OG is usually repaired by OGG1. If unrepaired, 8-OG is misrecognised as a thymine base (T) and mispaired with adenine. MUTYH excises adenine mispaired opposite 8-OG. Subsequently, OGG1 acts to excise 8-OG and thus restore the correct G:C pairing.

Clinical syndrome associated with mutations in *MUTYH*

Mutations in *MUTYH* impair its glycosylase activity and cause an increased rate of C>A mutations[149-153]. Inherited biallelic mutations in *MUTYH* cause the autosomal recessive cancer predisposition syndrome: ***M***_UTYH_-**A**ssociated **P**olyposis (MAP), which is characterised by intestinal polyposis and a substantially elevated risk of colorectal and duodenal cancer[154-157].

A spectrum of germline *MUTYH* mutations and genotypes are associated with MAP[158]. The allele frequency of MAP alleles and genotypes varies between different continents and population groups[159]. The most common European pathogenic *MUTYH* alleles are *MUTYH*$^{\text{Y179C}}$ and *MUTYH*$^{\text{G396D}}$ [158,159]. Other, less common germline missense mutations are also described[159]. Additionally, a small number of families with homozygous truncating *MUTYH* mutations are described in association with MAP[158-160]. The *MUTYH*$^{\text{Y179C}}$ mutation affects the helix-loop-helix (HLH) domain which is involved in DNA binding and is associated with a severe reduction in glycosylase activity[149,151,152,161]. The *MUTYH*$^{\text{G396D}}$ mutation affects the C-terminal domain and is thought to impair 8-OG recognition and glycosylase activity[151,152,161]. Truncating *MUTYH* mutations such as *MUTYH*$^{\text{E480*}}$ also severely impair the protein's glycosylase activity *in vitro*[162]. Differences in glycosylase activity associated with the common *MUTYH* mutations *in vitro* are reflected in the severity of clinical phenotype in individuals with these germline mutations. Individuals with *MUTYH*$^{\text{Y179C+/+}}$ and homozygous truncating *MUTYH* germline mutations have higher polyp burdens and earlier age of onset of colorectal cancer compared to those with the *MUTYH*$^{\text{Y179C+/- G396D+/-}}$ and *MUTYH*$^{\text{G396D+/+}}$ genotypes[158].

Inherited heterozygous *MUTYH* mutations occur at relatively higher frequency in healthy individuals and do not confer a polyposis phenotype. Whether these individuals carry an increased risk of cancer is debated[163-170].

Clinical features and management of MAP

MAP is associated with colorectal polyposis and an elevated risk of early-onset colorectal cancer. Colorectal polyp burdens are typically between ~30-50 adenomas; 20-30% of individuals have >100 polyps. Polyps in MAP are usually adenomatous. The median age of onset of colorectal cancer is 43-51 years of age. There is substantial heterogeneity in both the colorectal polyp burden and age of onset of colorectal cancer: some is due to differences in the *MUTYH* germline genotype, but much is unexplained[158]. The lifetime risk of colorectal cancer in MAP is markedly elevated – potentially as high as 90% - but estimates vary substantially between studies[169-171].

Up to a third of individuals (17-34%) with MAP develop duodenal polyposis[172-174]. At a median age of 51yrs, 15% of individuals with MAP have duodenal polyposis with a median burden of 3 polyps. Lifetime duodenal cancer risk is ~1.4% to 4%[174] which is substantially higher than the ~0.05% lifetime risk of sporadic small intestinal cancer in non-predisposed individuals. Individuals with *MUTYH*$^{Y179C/Y179C}$ and *MUTYH*$^{E480*/E480*}$ genotypes are most likely to have duodenal polyps at first endoscopy and carry the greatest cumulative risk of developing duodenal adenomas[174]. This is broadly consistent with the observation that individuals homozygous for *MUTYH*^{Y179C} and those with truncating genotypes have an earlier onset of colorectal cancer and more severe colonic polyposis[158].

Risks of other cancers including ovarian and bladder cancer are significantly increased in MAP[170,172]. Endometrial, breast, thyroid, gastric and skin cancer all occur in small numbers of individuals with MAP, albeit not at a level that reaches epidemiological significance[170,172,175].

Routine endoscopic surveillance and polypectomy is the mainstay of the management in MAP. In the UK, colonic endoscopy is advised annually from 18-20 years of age and upper GI endoscopy from 35 years of age[122]. Unlike FAP, prophylactic colectomy is not indicated in MAP[122]. There are no chemoprevention agents currently approved for use in MAP.

Mutational processes associated with defective MUTYH

Colorectal cancers and adenomas with biallelic inactivating *MUTYH* mutations show a predominance of C>A mutations[56-58]. Two C>A mutational signatures are abundant in neoplasms with *MUTYH* mutations: SBS18 and SBS36. SBS18 is associated with DNA damage due to reactive oxidative species[53] and SBS36 arises through impaired cleavage of adenine bases mispaired opposite 8-OG[56]. SBS36 is exclusively observed in samples with impaired *MUTYH* function, whereas SBS18 is also found in multiple sporadic cancer types and in normal tissues[6,56,57]. Several studies have identified the presence of SBS18 and SBS36 mutational signatures in colorectal polyps and colorectal cancers from individuals with inherited *MUTYH* mutations[6,56,57,176]. In these studies, the presence of SBS18 and SB36 was additive to the ubiquitous mutational signatures typically found in sporadic neoplasms without *MUTYH* mutations, thus implying that neoplasms with *MUTYH* mutations have an increased somatic mutation burden. The mutation burdens in normal tissues from individuals with MAP have not been extensively investigated. However, bulk normal intestinal epithelium from individuals with MAP appears to have a higher mutation burden than control samples[58] implying the potential presence of elevated mutation burdens.

Neoplasms from individuals with MAP are associated with a distinctive spectrum of driver mutations: indeed, this was the first clue that led to the discovery of *MUTYH* as a cancer predisposition gene[156]. Al-Tassan et al identified a predominance of truncating C>A mutations causing stop codons at glutamic acid residues in the *APC* gene in colorectal cancer from individuals with inherited *MUTYH* mutations. Subsequent studies have identified a similar C>A spectra in other canonical colorectal cancer driver genes in colorectal adenomas and adenocarcinomas from individuals with MAP[56-58,177]. These data suggest that the mutational processes associated with defective MUTYH in adenomas and colorectal cancers are responsible for the generation of driver mutations. This implies that defective MUTYH-associated mutational processes may be active in the early stages of neoplasia and possibly also in normal intestinal epithelium.

Polymerase Proofreading-Associated Polyposis (PPAP)

DNA replication and proofreading

Cell division is required for the growth and maintenance of human tissues. With each cell division, the entire genome - comprising over three billion bases - is replicated. Replication of DNA is chiefly undertaken by DNA polymerases ε and δ which are encoded by the genes *POLE* and *POLD1* respectively. These replicative DNA polymerases are highly processive and undertake DNA replication from a reference template strand with a high degree of fidelity. Pol ε and Pol δ undertake leading and lagging strand replication respectively[178,179]. Uniquely among DNA polymerases, Pol ε and Pol δ possess an exonuclease / proofreading domain which is responsible for identifying and removing mispaired bases from the nascent DNA strand. Cells with engineered mutations in the proofreading domain of *POLE* and *POLD1* show elevated mutation rates. The highly accurate replication and separate proofreading mechanisms make Pol ε and Pol δ important contributors to the maintenance of genome integrity.

DNA polymerase mutations and human cancer

DNA polymerase exonuclease domain mutations (EDM) are observed in 3% of colorectal cancers and 7% of endometrial cancers[16,180]. Cancers with somatically-acquired heterozygous *POLE* and *POLD1* proofreading domain mutations have substantially elevated rates of SBS and small ID mutations[16,180]. Cancers with *POLE* mutations generally have higher base substitution burdens whereas those with *POLD1* mutations have increases in both base substitutions and small insertions and deletion mutations[180].

Mutational spectra in DNA polymerase mutant cancers

Defective POLE generates two mutational signatures; SBS10a and SBS10b[6,181]. SBS10a is characterised by C>A mutations at TC̲T and TC̲A trinucleotide contexts (mutated base is underlined). SBS10b is characterised by C>T mutations at TC̲G and TC̲T trinucleotides. Both mutational signatures account for a large number of mutations in cancers with *POLE* mutations. Both SBS10a and SBS10b have a leading strand replication bias in keeping with the role of Pol ε in leading strand DNA synthesis and proofreading. A further signature; SBS28, is commonly seen in samples with high SBS10a/SBS10b exposure. SBS28 is a T>G signature with characteristic peaks at AT̲T, CT̲T and TT̲T trinucleotides[12,182]. The ratio of these three *POLE* signatures varies somewhat between tumours, with some of the variation attributable to the type and location of the *POLE* mutation[183]. A further signature, SB14, is observed in cancers with concurrent *POLE* mutations and MMR deficiency[5]. Defective POLD1 is typically observed alongside microsatellite instability where it generates a C>A mutational signature, SBS20[180]. Mutational spectra arising from *POLD1* mutations in cells with proficient MMR have not been identified in cancer.

Errors that occur during normal DNA replication are identified and repaired by DNA polymerase proofreading. In addition to this important line of defence, MMR identifies and repairs a substantial burden of DNA lesions that are generated during DNA replication and are missed by the DNA polymerase proofreading. Therefore, the mutational spectra observed in cells with faulty DNA polymerases represent mutations that were unrepaired by the faulty DNA polymerases and not identified by the competent MMR machinery[55,183]. Impaired DNA polymerase proofreading and MMR presents a different mutational spectrum that is thought to be more comparable to the spectrum of errors generated by DNA polymerase activity[183].

Clinical features of Proofreading Polymerase Associated Polyposis

DNA polymerase EDM mutations inherited in the human germline cause the cancer predisposition syndrome Proofreading Polymerase Associated Polyposis (PPAP). PPAP is an autosomal dominant condition caused by germline heterozygous germline mutation in *POLE* or *POLD1* gene.

PPAP is characterised by colorectal adenomatous polyposis and an increased risk of early onset colorectal and endometrial cancer[184,185]. Cancers in extra-intestinal tissues are also observed in individuals with PPAP, including ovarian cancer in *POLE* mutation carriers and brain malignancies in individuals with *POLE* and *POLD1* mutations[185]. Almost all individuals with PPAP (95%) develop intestinal adenomas and or colorectal carcinoma[185]. The median age at diagnosis of colorectal adenomas and adenocarcinoma are 36 and 44 years respectively. The median number of colonic polyps observed in PPAP is 12 (range 0-100) which is relatively modest compared to individuals with the more common polyposis syndromes. In FAP, individuals typically have hundreds or thousands of adenomas by their third decade of life and in MAP polyp burdens are typically ~30-50 with a proportion of individuals having 100’s of polyps. Polyposis in PPAP is typically adenomatous: an increased risk of other polyp types is not reported. There is no overt association between the polyp burden and cancer risk in PPAP. Almost a third of individuals develop colorectal cancer in the absence of polyposis.

Much like the more common polyposis syndromes, PPAP is associated with the development of small bowel neoplasia[185]. Duodenal adenomatous polyposis affects ~15% of individuals with *POLE* mutations and is less frequent in individuals with *POLD1* mutations. Similarly, duodenal cancers, which affect 10% of individuals with *POLE* mutations, are not observed in individuals with *POLD1* mutations. Whilst not formally calculated and limited by the small number of individuals with PPAP, the incidence of duodenal / small bowel neoplasia represents a many fold increase when compared to normal healthy individuals. The age of onset of neoplasia is later in the duodenum than the colon.

	POLE	*POLD1*
Large bowel polyposis	Y	Y
Small bowel polyposis	Y	N
Early onset colorectal cancer	Y	Y
Early onset duodenal cancer	Y	N
Endometrial cancer	Y	Y
Ovarian cancer	Y	N
Breast cancer	Y	Y
CNS cancer	Y	Y

Table 1.2 | Summary of cancer types associated with PPAP

Source data: Palles et al, *Familial Cancer*, 2021[185]

Female reproductive tract cancers are also a common feature of PPAP[185]. Endometrial cancer is the most common reproductive cancer in female *POLD1* mutation carriers and is also observed, albeit less commonly, in individuals with *POLE* mutations. Ovarian cancer is observed in *POLE* carriers but is not in *POLD1* carriers. Other cancer types including breast and brain tumours are now also described in association with PPAP[185].

Inherited *POLE* and *POLD1* mutations in PPAP have, to date, only been identified in the heterozygous state. Homozygous, compound heterozygous or co-mutation of both main replicative DNA polymerases - Pol ε and Pol δ - have not been reported. Whether this is a function of the extremely low allele frequency, making the inheritance of two mutant DNA polymerase alleles very unlikely, or a possible indication of the inviability of living with two defective DNA polymerase alleles, is unclear.

Besides the known cancer predisposition, individuals with PPAP do not display overt features of ageing or premature onset of diseases commonly attributed to or associated with ageing such as hypertension, type 2 diabetes or heart disease. This is an observation that is important when contextualising the data presented in this thesis and their implications for our understanding of the somatic mutation theory of ageing.

Clinical management of PPAP

The mainstay of management in PPAP is regular endoscopic surveillance[122]. The treatment of neoplasia in PPAP is largely similar to sporadic cancers. Currently, there are no medical therapies specifically approved for the treatment of cancer in individuals with PPAP. However, there is increasing interest in the use of immune checkpoint inhibitors in sporadic colorectal and endometrial cancers with *POLE* mutations owing to the elevated neoantigen generation and PD-1 expression[186].

Progeria syndrome associated with DNA polymerase domain mutations

Germline mutations in the exonuclease domain of the DNA polymerase mutations cause PPAP. Inherited mutations in the polymerase domain of the DNA polymerase genes have been described in a small number of individuals. The clinical phenotype of these individuals includes dysmorphia[187], multi-systemic features including lipodystrophy; and premature ageing[188-192]. It is proposed that the ageing features of this syndrome result from an increased frequency of stalled DNA replication forks, increased cell apoptosis and stem cell exhaustion. The pathogenesis of this very rare condition may share some overlap with mechanisms seen in other more established progeria syndromes.

NTHL1-associated polyposis

NTHL1-associated polyposis is an intestinal polyposis and colorectal cancer predisposition syndrome associated with inherited biallelic inactivating mutations in the *NTHL1* gene[193,194]. NTHL1 is a DNA glycosylase responsible for the excision of oxidised DNA bases as part of the base excision repair pathway. Polyp burdens in NTHL1-associated polyposis are between 8-50. Early onset and synchronous colorectal carcinomas are reported in NTHL1-associated polyposis. Extra-intestinal malignancies are common in this syndrome, including breast, bladder, skin, uterine and brain cancers. The mutation burden in cancers from individuals with *NTHL1* mutations are thought to be increased compared with sporadic cancers and are caused by the accumulation of C>T mutations due to the characteristic NHTL1-associated mutational signature, SBS30[195].

MBD4-associated neoplasia syndrome

Germline biallelic mutations in the base excision repair gene *MBD4* cause a cancer predisposition syndrome characterised by acute myeloid leukaemia, colonic polyposis and an increased risk of colorectal cancer[110,196]. Biallelic *MBD4* mutations are estimated to occur at a frequency of 1 in 5 million individuals. Thus this syndrome represents an exceptionally rare cause of intestinal cancer predisposition. MBD4-associated neoplasia syndrome (MANS) is associated with colonic polyposis, colorectal cancer and acute myeloid leukaemia (AML). Intestinal polyposis occurs with burdens of 17-90 adenomas. Colorectal cancer was diagnosed in three individuals with MANS at the ages of 31, 37 and 47 years of age. AML was diagnosed in the early 30s in three individuals with MANS.

Deamination of 5-methylcytosine is a cause of age-related mutation accumulation in many normal cell types[46,47,49,50,197]. MBD4 acts to remove thymine mispaired opposite deaminated 5-methycytosine (5-MC). Loss of function variants in *MBD4* cause the accumulation of C>T transversion mutations at sites of deaminated 5-MC. AML cancers bearing biallelic *MBD4* loss have up to ~33 fold higher mutation burden than *MBD4* wild-type AML cancers[196]. At least six-fold increases in the mutation burden are observed in adenomas with biallelic *MBD4* loss[110]. Higher mutation burdens were observed in adenomas from older individuals, implying that *MBD4* mutations may also increase the mutation rate in normal tissues. *MBD4* mutations cause a distinctive mutational signature characterised by mutations at CpG dinucleotides which resembles the reference mutational signature SBS1.

Intestinal polyposis and colorectal cancer predisposition

Intestinal polyps are a relatively common finding in the large intestine of humans with advancing age, typically increasing in prevalence from the sixth decade of life onwards[198]. Intestinal polyps generally refer to an abnormal overgrowth of cells arising from the intestinal epithelium. The most common histological type of polyp are adenomas. Adenomatous polyps are early neoplastic lesions that are strongly associated with progression to colorectal adenocarcinoma. Other histological types including serrated and hyperplastic polyps are less common in healthy individuals[199].

Polyps may also be found at an early age of onset and / or with increased frequency in many intestinal cancer predisposition syndromes, leading to these conditions being called polyposis syndromes. There is no universal definition for the number of polyps that is required to be diagnosed as having a polyposis syndrome. Substantial variation in polyp burdens are often observed between individuals with the same polyposis syndrome.

Examples of syndromes associated with adenomatous polyposis include familial adenomatous polyposis (FAP), MUTYH-associated polyposis (MAP) and polymerase proofreading associated polyposis (PPAP). Certain intestinal cancer predisposition syndromes are associated with the development of non-adenomatous polyps e.g. juvenile polyposis syndrome and Cowden's syndrome which are associated with the development of hamartomatous polyps.

Comparison of intestinal neoplasia in inherited cancer predisposition syndromes

The clinical phenotype of individuals with PPAP and MAP - and several other intestinal cancer predisposition syndromes - shows substantial heterogeneity. For example, in MAP intestinal polyp burdens range from tens to ~1000 polyps and the age of colorectal cancer diagnosis ranges from ~20 to 80 years of age[158]. The underlying cause for the heterogeneity observed in intestinal cancer syndromes is not fully understood. Given the variability in phenotype between individuals, comparison between the phenotype of intestinal disease in the different syndromes should be approached with a degree of caution. Nevertheless, there are some clear differences in the colorectal polyp burden and age of onset of colorectal cancer (Table 1.3). The lowest polyp burdens are seen in PPAP, higher polyp burdens are seen in AFAP and MAP and the highest in classic FAP. The earliest age of onset of colorectal cancer is in FAP followed by PPAP, MAP, LS and AFAP.

	PPAP	MAP	FAP	AFAP	LS
Colorectal polyp burden	0-100	0-1000	100-1000s	0-100	Unaffected
Age of cancer dx.	35-45yrs	45-55yrs	35-40yrs*	55-60yrs**	40-60yrs^

Table 1.3 | Colorectal polyp burden and age at colorectal cancer diagnosis in PPAP, MAP, FAP, AFAP and LS

*Management of FAP involves prophylactic colectomy. Hence, colorectal cancer is unusual in individuals who are known to have FAP and age estimates may therefore be inaccurate.

** Knudsen et al 2003[200]

^Giardiello et al 2014[125]

Normal tissue genomics

Overview

Somatic mutations accumulate in normal cells with age[46,47,49,50,82,197,201-204]. These mutations serve as a record of past mutational processes including mutagen exposure. In addition to information about the genomic changes that drive the development and evolution of cancers, genome sequencing of healthy and diseased human tissues can offer a diversity of insights into cell biology, developmental processes and the clonal dynamics of normal tissues[34,205-209].

Genome sequencing of healthy tissues presents certain technical challenges compared to cancers:

1. The cells in a cancer arise from a single shared ancestral clone. Hence, mutations in the ancestral cell are also present in all of the cells that comprise the cancer. Conversely, normal tissues typically comprise numerous small clones that are ancestrally unrelated. The mutations found in individual cells from normal tissues are unique and not shared by the neighbouring cells. Therefore, clonal mutations in cancers are relatively easily identifiable whereas mutations in normal tissues that occur at low allele frequencies are not easily identifiable because they cannot easily be distinguished from sequencing errors.
2. Sampling of small clones or even single cells from normal tissues may improve the detection of somatic mutations. However, the yield of DNA from small numbers of normal cells is typically insufficient for conventional next-generation DNA sequencing.
3. When sequencing small quantities of DNA, error rates from library preparation and sequencing artefacts may become difficult to distinguish from true somatic mutations.

Several methods have been developed that enable the study of somatic mutations in the DNA of normal cells from healthy tissues. Approaches applied to investigate somatic mutations in normal tissues include deep targeted sequencing of small tissue biopsies; single cell DNA sequencing; *in vitro* expansion of single cells and single-cell derived organoid generation; tissue microdissection and low-input DNA sequencing; and duplex sequencing methods.

- Deep targeted sequencing of tissue biopsies involves DNA sequencing using a panel of prospectively selected genes of interest to high coverage, often ~100-1000-fold. This approach has been applied to the identification of cancer gene mutations in normal healthy skin and oesophageal epithelium[203,204]. It is also the gold-standard manner of identifying mutations in peripheral blood from individuals with suspected clonal haematopoiesis. A key limitation of this method is that it only informs about mutations in the regions of the genome covered by the targeted gene panel: it is unable to inform about other regions. In addition, in most experiments, this approach seeks to identify mutations in the dominant clone or in large sub-clones. Mutations that occur in smaller clones or individual cells will not be identified. Lastly, inference of the genome-wide somatic mutation burdens and mutational processes may be imprecise owing to the small and usually non-representative portion of the genome surveyed.

- Single cell DNA (scDNA) sequencing involves the isolation of individual cells from a tissue of interest. Dissociated cells can be flow-sorted to identify specific cell populations and then DNA libraries are generated from selected individual cells. Whole-genome amplification is typically required to ensure that there is sufficient DNA for sequencing. ScDNA sequencing suffers from certain technological and analytical challenges including recurrent and sporadic artefacts, allelic dropout and high error rates. Furthermore, since scDNA sequencing requires the isolation of cells from a tissue, it loses the original orientation of the cell within the tissue being studied, and hence loses the associated histological information.
- *in vitro* expansion of tissues using culture techniques such as the generation of single-cell derived organoids[82] and single-cell derived colonies[207,210-212] are techniques involving the clonal expansion of single cells to generate a sufficient quantity of DNA for conventional next generation sequencing. This approach has been successfully applied to the study of several tissue types and benefits from the ability to generate high quality DNA libraries originating from a single cell. However, not all cell types are amenable to these techniques and retrospectively collected samples may not be suitably preserved for use with these methods. Furthermore, *in vitro* expansion may induce culture-related artefacts and, much like single-cell sequencing, the isolation of cells from tissues loses the spatial and histological information regarding the cell that is being expanded.
- Tissue microdissection involves the isolation of cells or tissue structures guided by histological visualisation using a microdissecting device, which can be either a needle

or a computer-controlled laser. Laser-capture microdissection (LCM) allows direct visualisation and histological assessment of tissue prior to isolation. Certain factors limit its application. Firstly, although some tissues have a microscopically visible clonal structure which is amenable to microdissection, others do not. Secondly, LCM requires tissues to be histologically prepared to enable visualisation: which may impact the preservation of cell constituents, particularly RNA and to a lesser extent DNA. Lastly, not all tissues are naturally amenable to the histological preparation that is a requirement for laser capture microdissection e.g. osteocytes in bone.

- Duplex sequencing methods involve the identification of a mutation in a single molecule of DNA based on sequencing both the Watson and Crick strands[201,213-215]. In sequencing both strands, mutations can be independently verified on both DNA strands and thus can be differentiated from PCR and sequencing preparation artefacts that would be present on just one DNA strand. Since this method can identify mutations in individual molecules of DNA, it is possible to identify mutations that occur in single cells. Therefore this method does not rely on a mutation being present in a substantial proportion of cells to be detectable. This method has two principal limitations: 1) since it detects mutations in single molecules of DNA, it is highly sensitive to contamination 2) fragmentation of the genome using sonication can lead to additional damage which may result in the introduction of artefactual mutations. An alternative approach is to use restriction enzymes to fragment the DNA. However, this approach means that large parts of the genome that do not contain the restriction sites are not sampled. Duplex sequencing is therefore well suited to assessment of mutation rates and mutational signatures but does not provide reliable coverage of the entire genome.
- Finally, there are several methods that use germline single nucleotide polymorphisms (snp) or artificially introduced barcodes to evaluate the validity of mutations. By phasing a recurrent base change to a nearby snp, the mutation can be differentiated from a sequencing artefact[216]. However, only somatic mutations that occur in close proximity to a snp in an individual read are evaluable using this method and since snp frequency varies across the genome many base changes will be unevaluable.

Mutation rates in normal tissues

The mutation rate in normal tissues is thought to be an important factor in the development of cancer. It is proposed that the somatic mutation rate in normal cells explains the observed variability in cancer risk across different tissues[217,218]. Over the past half-decade many studies using complementary approaches have sought to investigate the mutational burdens in various normal tissue types including skin[203,219], oesophagus[204,220], blood[211,212], lung[48], liver[47], intestine[46], endometrium[49], brain[201,202,221], bladder[197], foetal tissues[208,222], placenta[34] and sperm[50]. In addition, several studies have investigated and compared somatic mutation burdens in multiple different cell types[50,82,201,223,224].

The most abundant mutation type in the genome of cells from normal tissues is single base substitutions. Small insertion and deletion mutations are much less frequently observed, approximately 60-fold lower than SBS mutations[50]. A common finding from the normal tissue studies listed above is that somatic mutation rates vary between different normal tissue types, with the highest mutation burdens observed in intestinal epithelium[46,50] and the lowest mutation burden in sperm cells[50]. The cause of the variable mutation burdens observed across different normal tissue types is not clear. Cell proliferation / turnover is thought to be an important factor contributing to the generation and accumulation of DNA mutations in the genome[218]. It is thought that with each cycle of DNA replication, errors accumulate from mismatched base pairing, a small proportion of which are not identified by the cell's inbuilt surveillance mechanisms including DNA polymerase mediated proofreading and the mismatch repair machinery. Additionally, DNA damage that results from other mutational processes may, if unrepaired, become fixed during DNA replication and form a mutation. Indeed, many of the tissues that are known to have high rates of cell turnover and hence replicate frequently, have high SBS mutation rates e.g. colon, endometrium and skin.

However, there are differences in the mutation burden and signature composition between proliferative tissue types that do not appear to be explained by cell turnover alone. Thus investigation of mutational processes across normal tissue types is an essential endeavour to characterise and explain the cell type specific differences in mutation burden. Characterisation of the mutational processes, their relative activity, temporal dynamics and contribution to the generation of protein-altering mutations may help to develop our understanding of cancer risk as well as the tissue-specific factors that contribute to the development of human disease and ageing.

Having mentioned normal tissues with high mutation burdens, it is worthwhile also briefly discussing the surprise observation of very low mutation burdens in specific tissues. Perhaps the most striking example of this is the very low mutation rate in sperm cells. Sperm cells are produced by the testes in great abundance continuously throughout life, yet appear to maintain a remarkably low mutation rate. Moore et al compared the somatic mutation rates across multiple tissues from three individuals supplemented by in-depth investigation of seminiferous tubules from a larger cohort of men[50]. They found that the mutation rate in testis is 27-fold lower than colon, which has the highest mutation rate. It is postulated that the mutation rate in sperm is kept exceptionally low through additional or enhanced DNA repair processes which minimise the accumulation of mutations. Why this occurs and how the testis has developed to achieve this remarkable feat whilst retaining very high levels of proliferation is an important question that needs addressing. Of course, there is an important biological rationale for maintaining a low mutation rate in sperm as the consequences of mutations in the sperm are considerable. Acquisition of a pathogenic mutation in germ cells may, if transmitted to the next generation, result in a disease phenotype or reduced reproductive capability.

In contrast to SBS mutations, structural rearrangements and copy number changes are rarely observed in normal tissue[50]. In a study of multiple tissues from three individuals, ~480,000 SBS mutations and ~8,400 ID mutations were identified and only 37 copy-number changes (CNA) and 128 structural variants (SV) were identified. The majority of SVs occurred in intestinal crypts and did not affect any known cancer genes. The sparsity of structural variants in the genomes of normal tissues invites potential speculation. Whether this reflects the high competency of DNA repair mechanisms or potentially the elimination of cells with these mutation types is not known. An important caveat is that not all tissues are represented by the studies of normal tissues conducted to date. SVs are driver events in certain cancer types. Normal cells from those tissue types are not represented in normal tissue sequencing studies e.g. bone osteocytes and head and neck squamous epithelium. Whether SV rates are higher in those tissues compared to other normal tissues is not known

Mutational signatures in normal tissues

An understanding of the processes responsible for the accumulation of mutations in cells is a key step in deciphering the causes of human diseases of the genome. Whole genome sequencing studies of normal cells have collectively generated catalogues comprising millions of somatic mutations. Identification of mutational processes in these tissues has led to the observation that two ubiquitous processes account for most of the somatic mutations observed in normal human tissues. SBS1, which is caused by deamination of 5-methylcytosine at CpG dinucleotides and SBS5, which is of unknown aetiology, but is potentially thought to be a composite of multiple underlying processes (discussions with I Martincorena, P Campbell, M Stratton). The rate of accumulation and relative proportion of SBS1 and SBS5 varies substantially between different tissue types[50] suggesting that they are independent processes.

The next most commonly observed mutation process in normal tissues is SBS18 which is thought to be due to the mutagenic effect of reactive oxygen species on DNA. SBS18 is characterised by C>A transversion mutations at specific trinucleotide contexts. Colonic epithelium has the highest burden of SBS18 mutations compared to other normal types, accounting for 13% of all SBS mutations[46]. It is noteworthy that substantially higher proportions are observed in extra-embryonic tissue from the human placenta, where SBS18 mutation burdens account for >40% of the total SBS mutation burden[206]. SBS18 is found in several human cancer types[6] and is abundant in neuroblastoma where it contributes ~20-40% of total SBS mutation burden. In neuroblastoma, SBS18 is postulated to be caused by ROS generated by increased expression of mitochondrial ribosomal genes[54]. SBS18 is thought to be active from the early stages of neuroblastoma development as implied by the presence of driver mutations with an SBS18-like mutational signature. SBS18 is also seen commonly as an artefact of *in vitro* cell culture[63,225].

Additional mutational processes are observed sporadically in human tissues and are summarised below (Table 1.4).

Mutational Signature	Mutation type	Tissue	Mechanism	Studies
SBS2 and SBS13	C>G and C>T respectively	Ileum crypts Oesophageal epithelium Colonic crypts	APOBEC mediated mutagenesis	Moore et al 2021[50] Yokoyama et al 2018[220] Olafsson et al 2020[80]
SBS4	C>A	Liver parenchyma Lung epithelium	Tobacco smoke mutagens e.g. benzo[a]pyrene	Moore et al 2021[50] Yoshida et al 2020[48]
SBS7a and SBS7b	C>T	Skin epithelium	UV light exposure	Martincorena et al 2015[203] Moore et al 2021[50]
SBS9	T>C and T>G	Lymphocytes	Somatic hypermutation in B cells	Machado et al 2021[210]
SBS16	T>C	Oesophageal epithelium	Unknown / possible alcohol	Moore et al 2021[50] Martincorena et al 2018[204] Yokoyama et al 2018[220]
SBS17b	T>G	Lymphocytes Oesophageal epithelium	Unknown	Machado et al 2021[210] Yokoyama et al 2018[220]

SBS88	T>C and T>G	Colonic crypts	Colibactin mediated mutagenesis produces by PKS producing *E.coli*	Lee-Six et al 2019[46] Olafsson et al 2020[80] Moore et al 2021[50]
SBS89	Multiple	Colon crypts	Unknown	Lee-Six et al 2019[46]

Table 1.4 | Mutational processes observed sporadically in normal human tissues

In addition to the sporadic SBS mutational processes listed above, several mutational signatures resulting from the mutagenic effects of drug therapies have been identified in normal tissues (Table 1.5). In affected individuals, therapy-associated signatures contribute 100s and in one case 1000s of additional somatic mutations, equivalent to years or decades of mutation burdens due to normal ageing[46,80]. In one individual who had undergone treatment for lymphoma and caecal adenocarcinoma, the ~3-5-fold additional mutation burden due to chemotherapy would confer an elevated 'mutational age', equivalent to 200-300 years of mutagenesis due to normal ageing processes. The observation of therapy-related mutational signatures in normal intestinal crypts illustrates the off-target / unintended effects of certain types of commonly used drug therapies. The effect of off-target chemotherapy-related mutations on cancer risk and ageing in normal tissues is not entirely clear. However, elevated therapy related mutational signature burdens are observed in secondary cancer development[206,226]. Chemotherapy treatment may also cause remodelling of cell populations in humans[227].

SBS32	C>T	Colonic crypts	Purine treatment	Olafsson et al 2020[80]
SBS35	Multiple	Colonic crypts	Platinum treatment	Olafsson et al 2020[80]
SBSD	T>A	Colonic crypts	Multiple	Lee-Six et al 2019[46]

Table 1.5 | Summary of therapy -associated mutational signatures identified in normal human tissues.

SBSD is of unknown aetiology, but is potentially due to multiple chemotherapy agents.

Driver burdens and frequencies in normal tissues

The spectrum of genes and types of mutations that drive the development of cancer vary between cancer types[21]. The cause of the tissue-specificity of the driver landscapes observed in human cancer and normal tissues is somewhat mysterious. Since cancer driver mutations may originate in normal tissues, prior to the onset of cancer, study of the driver mutations present in normal tissues may inform our understanding of the timing and potential role of these mutations. Cancer drivers are commonplace in several normal tissue types[48,49,203,204]. In normal skin epidermis, expanded clones carry driver mutations in genes including *NOTCH1*, *NOTCH2* and *NOTCH3*; *FAT1*, *TP53* and *RBM10*[203]. In the oesophagus, mutations in *NOTCH1* and *TP53* are commonplace, increase with age and are associated with clonal expansions[220,228]. In lung epithelium, driver mutations are found in up to 14% of cells from non-smokers and up to 25% of current smokers, with *NOTCH1*, *TP53*, and *ARID2* showing signals of positive selection[48]. In endometrium, 25/28 women had driver mutations, with the youngest, a 24 year old with a *KRAS*^{G12D} mutation. In total, almost 60% (147/257) of endometrial glands had 1 or more driver mutation. Twelve genes were found to be under positive selection: *PIK3CA*, *PIK3R1*, *ARHGAP35*, *FBXW7*, *ZFHX3*, *FOXA2*, *ERBB2*, *CHD4*, *KRAS*, *SPOP*, *PPP2R1A* and *ERBB3*[49]. In bladder urothelium positive selection was observed in 12 genes: *KMT2D*, *KDM6A*, *ARID1A*, *RBM10*, *EP300*, *STAG2*, *NOTCH2*, *CDKN1A*, *CREBBP*, *FOXQ1*, *RHOA*, and *ERCC2*[197]. Lastly, Lee-Six et al[46] surveyed the driver landscape of 90 colorectal cancer genes in over 1400 normal intestinal crypts and found that driver mutations were present in 1% of crypts. Signals of positive selection were identified in *AXIN2* and *STAG2*. Hotspot mutations were identified in *PIK3CA*, *ERBB1*, *ERBB3* and *FBXW7* and heterozygous truncating mutations observed in tumour suppressor genes: *ARID2*, *ATM*, *ATR*, *BRCA2*, *CDK12*, *CDKN1B*, *RNF43*, *TB1XR1* and *TP53*.

In normal bladder, skin and oesophagus, substantial burdens of cancer driver mutations are observed. However, there is substantial variation in the number of driver mutations observed in different individuals[197,203,204,220]. In normal bladder urothelium, highly significant differences were observed in the frequency of specific recurrently mutated genes e.g. in two individuals of a similar age, one had 35 *KDM6A* mutations and two *ARID1A* mutations, the second had four *KDM6A* mutations and 20 *ARID1A* mutations[197]. In oesophageal epithelium, *NOTCH1* clones were observed on average in an estimated 25-42% of the cells[204]. Age was a strong determinant of *NOTCH1* mutation frequency; mutant clones were observed in 30-80% of cells in middle aged / elderly individuals and just 1-6% of cells in individuals under 40 years of age. However, much of the interindividual variability was unexplained. Similarly, in normal skin epithelium, *NOTCH2* mutation frequencies show substantial interindividual heterogeneity[203]. Proposed causes of the heterogeneity in the driver mutation frequency in normal bladder urothelium; and oesophageal and skin epithelia, include: differences in environmental mutagen exposure and possibly the influence of germline modifier mutations.

An important paradigm that emerges from these experiments is that histologically normal healthy tissues may tolerate driver mutations and that these are not necessarily synonymous with the development of malignancy. It is noteworthy that in some tissues the landscape of driver mutations in normal tissues does not necessarily reflect those observed in cancers from the same tissue. For example, *APC*, which is recurrently mutated in ~80% of colorectal cancers, was not identified in any of the ~1500 intestinal crypts surveyed[46] and *PTEN* mutations, which are found in 30% of endometrial cancers[3], are present in only 2% of normal endometrium crypts[49]. A further difference is that driver mutations in tumour suppressor genes in normal tissues are commonly heterozygous whereas biallelic mutations are thought to be required to inactivate a gene in most cancers. If biallelic mutations are required to confer a fitness advantage, then heterozygous mutations in recessive cancer genes in normal tissues may not be beneficial. Most cancers are thought to require multiple driver mutations[21]. Accumulation of more than one bona fide driver mutation in normal healthy cells is unusual in most normal tissues, but does occur in endometrium[49]. Together, these factors may explain the apparent toleration of individual cancer driver mutations in otherwise healthy normal tissues.

An accepted criterion that connotes a mutation as being a potential cancer driver is evidence of positive evolutionary selection for the mutation or the gene in which it occurs. Several methods exist to assess the strength of selection. An extensively applied method in cancer genomics is assessment of the dN/dS ratio (discussed previously in the introduction). A principal challenge for the use of this type of method in normal tissue studies is the relatively small sample sizes and low frequency of driver / potential driver mutations. Nevertheless, with extensive sampling, tissues with strong signals of positive selection can be identified. Indeed, strong signals of selection accompany many of the recurrently identified cancer driver genes listed earlier in this section. In those examples, a potential consequence of the accumulation of driver mutations is the development of cancer. Somatic evolution may also operate on molecular pathways in non-neoplastic diseases. Two particularly exciting recent observations of somatic evolution in non-neoplastic disease are: 1) components of the IL-17 and Toll-like receptor pathways which are under positive selection in the normal colonic epithelium in individuals with inflammatory bowel disease (IBD)[80,229,230] and 2) effectors of insulin signalling that are under positive selection in normal liver hepatocytes in individuals with chronic liver disease[231].

Somatic mutation theory of ageing

Ageing is a universal phenomenon characterised by a progressive decline in cellular and organismal function associated with the passage of time. The proposed causes that underlie the process of ageing are wide-ranging and varied[232]. The somatic mutation theory of ageing is one of the forerunner theories of ageing[233,234]. It encompasses the notion that the accumulation of mutations in the genome of normal cells and tissues over the course of the human lifespan are responsible for the observable features and phenotypes associated with ageing[235-238].

The somatic mutation theory of ageing (SMT) was first presented by Leo Szilard in 1958[235]. He proposed that:

1) A proportion of genes are essential to the function of somatic cells and that heterozygous inactivation of these genes is largely tolerated; however, homozygous inactivation of a gene may occur in normal cells and would constitute a "hit".
2) Accumulation of "hits" can inactivate or destroy the chromosome and hence the somatic cell.
3) Ageing "hits" accumulate in the DNA of somatic cells in a linear manner throughout life and that the build-up of hits in chromosomes of healthy cells will, over time, affect the number of somatic cells that survive and that the rate of this decline increases with age.
4) The probability of dying increases with the accumulation of ageing "hits" until it reaches a critical threshold

There have been several iterations of the SMT since it was first proposed[237,238]. Morley (1982) outlined a conception of the somatic mutation theory of ageing that offers additional insight into the mechanism by which mutations may confer an ageing phenotype[236]. He proposed that:

1) Mutations accumulate in normal cells with time.
2) Singular mutations alone are unlikely to cause ageing. Rather, the number of mutations needs to reach a critical threshold.

3) The genes affected by the accumulation of mutations need to be expressed / coding in order for the mutations to have an effect.
4) Ageing is most likely to result from the accumulation of mutations in expressed (coding) genes rather than non-expressed regions of the genome.
    a. (Mutations that do not affect expressed / coding sequences may contribute to an ageing phenotype through the contribution of structural DNA damage.)
5) Upon reaching a critical threshold, cell death or severe impairment of function may result.
6) Death of cells in non-proliferating tissues will result in a decline in cell number, whereas cells that die in proliferative tissues are typically replaced.

In the decades since these and other papers were published, there has been growing interest in, and ability to characterise the somatic mutation burdens in normal healthy tissues. In addition to the broader scientific value of these studies, the observations arising from them may inform and refine the SMT. Progress in the study of somatic mutagenesis has been fuelled by several important technological advances that enable reliable detection of mutations from a small number of cells. These studies have shown that 1) somatic mutations accumulate throughout life in most human cell types and 2) somatic mutation rates in healthy tissues are remarkably similar across different individuals. The second observation implies that a critical threshold of somatic mutation burden exists, suggesting that further increases may be deleterious, causing disease or perhaps death. These observations would be consistent with, and in support of the various formulations of the somatic mutation theory of ageing. However, whether the somatic mutations that accumulate in normal cells are a cause or merely a consequence of ageing remains unclear. Investigation of individuals with an elevated somatic mutation rate may be informative in this regard and help to guide our conception of the somatic mutation theory of ageing. If substantially elevated somatic mutation burdens were observed in healthy tissues this would imply that normal tissues can tolerate higher mutation rates than previously thought and that the rate of acquisition of mutations in normal healthy tissues does not necessarily define the process of ageing. These postulates form one of the key rationales for the work described in this thesis.

Whilst observations regarding the accumulation of mutations in normal tissues and mammalian species are consistent with the principles set out in the SMT, the mechanism by which mutations accumulated in the genome of normal cells cause cellular ageing are not well understood. Although we are now able to investigate somatic mutagenesis in the genome of many normal tissue types, methods to comprehensively assess their cellular effects largely evade us.

In its initial conceptualisation, the SMT did not specify whether the substrate for ageing is the nuclear or mitochondrial genome. In a later formulation Orgel[239] proposed that mitochondrial genome may be more susceptible to the effect of mutations than the nuclear genome owing to the independence of each mitochondrial organelle. The mitochondrial genome is smaller and is ~97% exonic, meaning that a somatic mutation may be more likely to have a deleterious effect on an individual organelle. However, heteroplasmy and a large pool of mitochondria in normal cells may mean that mitochondrial mutations in individual organelles may not impact the function of the entire cell.

Preventing the introduction of errors during the synthesis of DNA requires an extensive system of surveillance and repair, and involves the exertion of considerable cellular resource. The disposable soma theory of the evolution of ageing proposes that higher organisms have evolved mechanisms to prioritise the accuracy of DNA replication in specific cell types that are critical to reproduction of the germ line and human longevity[236,240]. In a recent survey of mutation rates across normal human cells, intestinal stem cells had the highest somatic mutation rate, whereas sperm cells, that convey the genome of the male germline, had the lowest levels[50]. This observation is consistent with the proposal that higher species have developed mechanisms to maintain a low germline mutation rate in order to prevent the development of disease and thus protect the germ line. The evolution of cellular defence mechanisms that defend against the accumulation of mutations in the germline would imply that mutations / somatic mutations have a causal role in the ageing of cells and tissues.

A further observation that would support a potential causal link between the accumulation of somatic mutations and ageing is the evolution of the lifespan of species[240]. There is a ~30-fold difference in lifespan between mice and humans. It is proposed that differences in the lifespan of species are a consequence of evolution processes to develop mechanisms that reduce the somatic mutation rate hence conferring longer lifespans. In a study of somatic mutation rates in 16 mammalian species across a range of different lifespans a strong inverse correlation between lifespan and the somatic mutation rate was observed[241]. That is to say, short-lived mammalian species demonstrated high mutation rates whereas long-lived species had low mutation rates. Thus, at the end of the lifetimes of these species, the absolute mutation burdens are relatively comparable. These data corroborate the proposed theory, that evolution of species acts to minimise somatic mutation rates[236]. These data would also suggest that the somatic mutations contribute to ageing and may dictate the lifespan of species, in support of the somatic mutation theory of ageing.

Summary

In this thesis the somatic mutation burden and mutational processes in two inherited cancer predisposition syndromes are investigated. Common to both syndromes is the inheritance of germline mutations that are thought to confer an elevated cancer risk through an increased somatic mutation rate. These data may contribute to our understanding of the mechanisms of cancer predisposition and cancer risk. They also have potential relevance to our understanding of the somatic mutation theory of ageing.

Chapter 2 - Methods

Sample management and laboratory protocols

Patient recruitment and sample management

Individuals with PPAP were recruited as part of the CORGI-2 study, United Kingdom Research Ethics Committee (REC) 17/SC/0079. Additional sample collection was undertaken under approval from the following committees; London – Westminster, North East-Newcastle and North Tyneside 1 and NRES Committee East of England - Cambridge South (REC references: EC04/015, 16/NE/003 and 07-MRE05-44 respectively).

Individuals with MAP were recruited as part of Wales Research Ethics Committee (REC) 12-WA0071 and 15-WA0075 and samples were approved for use in this project by REC 18/ES/0133. Samples from healthy individuals were recruited as part of the following UK Research Ethics Committee (REC) studies; 15/WA/0131, 15/EE/0152, 18/ES/0133 and 08/h0304/85+5.

Informed consent was obtained from all participants and no monetary compensation was offered for their participation. A complete list of study participants is summarised in Chapters 3 and 4.

DNA extraction from bulk samples

Frozen whole blood underwent DNA extraction using the Gentra Puregene blood kit (Qiagen). Briefly, 1-2ml of frozen blood was thawed, lysed in RBC lysis solution and centrifuged. Cell pellet was resuspended in cell lysis solution and incubated at 37 ℃ for 2 hours. RNA and protein was degraded using RNase A solution and protein precipitation solution. DNA was precipitated with isopropanol.

Tissue Preparation

Frozen tissues were embedded in Optimal Cutting Temperature (OCT) compound. Frozen histological sections were cut at 25-30µm and mounted on polyethylene naphthalate (PEN) slides and fixed in 70% ethanol for 5 minutes followed by two washes with phosphate buffered saline for 1 minute each. Slides were manually stained in haematoxylin and eosin using a conventional staining protocol. A subset of samples were fixed in RNAlater (Sigma Aldrich) or PAXgene (Qiagen) according to manufacturers' instructions. PAXgene / RNAlater fixed tissue samples were embedded in paraffin using a Tissue-Tek tissue processing machine (Sakura). No formalin was used in the preparation, storage, fixation or processing of samples. Tissue blocks were sectioned to 10-16µm thickness and mounted onto PEN slides (Leica). Tissue slides were stained using a standard haematoxylin and eosin (H&E) protocol. Slides were temporarily cover-slipped and scanned on a NanoZoomer S60 Slide Scanner (Hamamatsu), images were viewed with NDP.View2 software (Hamamatsu).

Laser Capture Microdissection

Laser capture microdissection was undertaken using a LMD7000 microscope (Leica) into a skirted 96-well PCR plate. Cell lysis was undertaken using 20µl proteinase-K PicoPure® DNA Extraction kit (Arcturus®), samples were incubated at 65 ℃ for 3 hours followed by proteinase denaturation at 75 ℃ for 30 minutes. Thereafter samples were stored at -20 ℃ prior to DNA library preparation.

Intestinal crypt isolation

Crypts from one tissue block (PD44593e) were isolated using EDTA chelation. In brief, dissected mucosa was incubated in an EDTA solution and gently agitated resulting in dissociation of intestinal crypts from the underlying components of the intestinal epithelium. Crypts were then separated under a light microscope and placed in ATL buffer (Qiagen) containing 10%(v/v) proteinase K and digested overnight at 56℃. DNA extraction was performed using the QiaAMP DNA micro kit (Qiagen) as per manufacturer's instructions. DNA was then stored at -20℃.

Low-input DNA library preparation and sequencing

DNA library preparation of micro-dissected tissue samples was undertaken as previously described using a bespoke low-input enzymatic-fragmentation-based library preparation method[46,47,49,242]. This method was employed as it allows for high quality DNA library preparation from very low starting quantity of material (from 100-500 cells). DNA library concentration was assessed after library preparation and used to guide choice of samples to take forward to DNA sequencing, minimum library concentration was 5ng/µL and libraries with >15ng/µL were preferentially chosen. 150bp paired-end Illumina reads were prepared with Unique Dual Index barcodes (Illumina).

DNA sequencing was undertaken on a NovaSeq 6000 platform using an XP kit (Illumina). Samples were multiplexed in pools of 6-24 samples. Pools were sequenced to achieve a coverage of ~30x.

Mutation calling and post-processing filters

Sequencing reads were aligned to NCBI human genome GRCh37 and aligned using the Burrow-Wheeler Alignment (BWA-MEM). Single Base Substitutions (SBS) were called using the 'Cancer Variants through Expectation Maximization' algorithm (CaVEMan)[243]. Mutations were called using an unmatched normal synthetic bam file to retain early embryonic and somatic mutations. Post-processing filters were applied to remove low-input library preparation specific artefacts and germline mutations using a previously described method[34,48,49,242]. Filters applied were: (1) common single nucleotide polymorphisms were removed by filtering against a panel of 75 unmatched normal samples[64] (2) to remove mapping artefacts, mutations were required to have a minimum median read alignment score of mutant reads (ASMD ≥ 140) and fewer than half of the reads supporting the mutation should be clipped (CLPM =0); (3) a filter to remove overlapping reads that result from the relatively short insert size which could lead to double counting of variant reads; and (4) a filter to remove cruciform DNA structures that can arise during the low-input library preparation method.

Next, multiple filters to remove germline variants and potential artefacts whilst retaining *bona fide* embryonic and somatic variants were applied. This approach has been detailed in previous publications and the code for these filters can be found at https://github.com/TimCoorens/Unmatched_NormSeq. Mutations were aggregated per patient and a read pile-up was performed using an in-house algorithm (cgpVAF) to tabulate the read count of mutant and reference reads per sample for each mutation locus. Germline mutations were filtered out using an exact binomial test. The exact binomial test is used to distinguish germline from somatic variants and uses the aggregate read counts from all samples of the same patient[34,48]. In brief, the read depth across all samples from that individual was calculated (median in this study 496-fold). This high coverage yields a very precise estimate of the true VAF of each mutation. While the VAF estimates of the earliest embryonic SBS mutations and germline variants from samples sequenced at 30x might overlap, the VAFs from the aggregate coverage from that individual will be distinguishable using statistical testing. To achieve this, the beta-binomial test was applied. The overdispersion parameter (rho) threshold for genuine variants of rho>0.1 was used.

Phylogenetic trees were created using MPBoot (version 1.1.0 bootstrapped - 1000) and mutations were mapped to branches using maximum likelihood assignment (https://github.com/NickWilliamsSanger/treemut).

Indels (ID) were called with Pindel[244] using the same synthetic unmatched normal sample employed in SBS mutation calling. ID calls were filtered to remove calls with a quality score of <300 ('Qual'; sum of mapping qualities of the supporting reads) and a read depth of less than 15. Thereafter, ID filtering was performed in a similar manner as SBS to remove germline variants and library preparation / sequencing artefacts.

Copy number and structural rearrangement calling

Copy number variants and structural rearrangements were called separately for the PPAP and MAP cohorts. Analysis of the PPAP cohort was undertaken first and methods were modified and refined for the MAP cohort.

Copy-number variant calling - PPAP

The Somatic copy-number variants (CNVs) were called using the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm[245] (https://github.com/Crick-CancerGenomics/ascat) in the ascatNGS package[246]. Bulk (blood or in one case tissue) samples were used as matched normals. ASCAT was initially run with default parameters. To reduce the number of false-positive calls that arise in normal tissue samples, a segmentation penalty was applied in the ASCAT 'aspcf' step. Optimum performance was observed with a penalty value of 100 which was subsequently applied to all samples. Copy-number calls were further filtered to remove artefacts. Copy-number (CN) calls less than 2MB were excluded. Samples with a goodness-of-fit of less than 95% were excluded. CN calls at specific recurrent breakpoints were removed. Sharing of CNVs between samples from different tissue blocks and across individuals that violated phylogenetic structures implied from SBS and ID phylogenetic trees were treated as artefactual and removed from analysis. Similarly, any recurrent copy-number calls with identical break points observed across different individuals were also removed. CNV calls were manually verified by visualisation of reads in JBrowse[247].

Structural variant calling - PPAP

Whole-genome sequences were analysed for somatic structural variants (SVs) using Breakpoints via assembly (BRASS) algorithm[9], paired blood samples were used as controls. If no blood sample was available, a tissue sample was used that was phylogenetically distant to the sample being analysed. SV calls were filtered using an in-house algorithm in a multi-stage process using bespoke software (https://github.com/MathijsSanders/AnnotateBRASS). Finally, all SV calls were manually inspected to confirm somatic variants. SV calls in L1 transposon donor regions and fragile sites were excluded from the final SV analysis.

Copy-number alteration calling - MAP

Somatic copy-number variants (CNVs) were called using the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm[245] (https://github.com/Crick-CancerGenomics/ascat) in the ascatNGS package[246]. Bulk blood samples or phylogenetically unrelated normal samples were used as matched normals. ASCAT was initially run with default parameters. To reduce the number of false-positive calls that arise when analysing normal tissue samples using ASCAT, a bespoke algorithm ascat-PCA was applied. ascat-PCA extracts a noise profile by aggregating the LogR ratio from across a panel of normal unrelated samples and subtracts this signature from that observed in the sample being analysed using principal component analysis (https://github.com/hj6-sanger/ascatPCA).

Structural variant calling - MAP

Whole-genome sequences were analysed for somatic structural variants (SVs) using the Genomic Rearrangement Identification Software Suite (GRIDSS). In preparation for this analysis, genomes were remapped to Human Genome Version 38 and GRIDSS was run using the same matched normal as used for CNV analysis. Coordinates for SV calls were subsequently converted back to GRCh37. SV calls in L1 transposon donor regions and fragile sites were excluded from the final SV analysis.

Mutational signature analysis

The R package HDP (https://github.com/nicolaroberts/hdp), based on the hierarchical Dirichlet process[248], was used to extract mutational signatures. Analysis of mutational signatures using this package has been applied to normal tissues previously[46-49]. In brief, this nonparametric Bayesian method models categorical count data using the hierarchical Dirichlet process. A hierarchical structure was established using the patient / individual as the first tier (parent nodes) and the samples as the second tier (dependent nodes). Uniform Dirichlet priors were applied across all samples. The algorithm creates a mutation catalogue for each sample and infers the distribution of signatures in any one sample using a Gibbs sampler. Mutational signature analysis was performed per-branch, treating each branch of the phylogenetic tree as a distinct sample to avoid double counting of mutations. Since the MCMC process scales linearly with the number of counts, each branch was randomly subsampled to a maximum of 2500 total SBS. Branches with fewer than 100 mutations were excluded from the mutational signature extraction. No reference signatures were included as priors.

To assess the contribution of each mutational process, mutational signatures were refitted to all mutation counts of branches of phylogenies using the R package sigfit (https://github.com/kgori/sigfit)[43]. To avoid overfitting, a limited subset of reference mutational signatures were included per patient corresponding to the HDP signatures that have been identified in that individual.

The following parameters were used for HDP SBS mutational signature extraction:
Chains: 20 MCMC chains
Iterations: 40,000

Burnin: 20,000

Samples: 100 / chain

Signature components identified: 14

Component Names: HDP0-HDP13

The non-negative matrix factorisation (NMF) based algorithm SigProfiler was used to validate the results generated using HDP. The settings used are outlined below. In general, fewer signature components were identified using SigProfiler than HDP. However, the components SigProfiler extracted had clearly recognisable counterparts in the compendium of HDP components. Additional components that were stably extracted and showed close resemblance to known, reported signatures were identified by HDP but not SigProfiler.

The following parameters were used for SigProfiler SBS mutational signature extraction:

input_type: vcf

startProcess: 1

endProcess: 15

totalIterations: 1000

cpu: 1

hierarchy: True

refgen: GRCh37 genome_build: GRCh37 mtype: [‘default’]

init: random

Mutational signatures analysis - PPAP cohort

Fourteen signature components were extracted using HDP; HDP0-HDP13. An expectation maximisation algorithm based on cosine similarity was employed to deconstruct these extracted signature components into their reference constituents (SBS1, SBS5, SBS7a, SBS7b, SBS10a, SBS10b, SBS17a, SBS17b, SBS25, SBS28, SBS31, SBS35, SBS88 and SBS89). In this way, HDP3 was broken down into SBS1 and SBS5; HDP4 into SBS10a and SBS5; HDP5 into SBS1, SBS5, SBS10a, and SBS28; HDP9 into SBS7a, SBS7b, SBS10a, and SBS28. HDP1 was broken down into SBS10a and SBS10b, but SBS10b poorly reflected the observed C>T component of HDP1 and HDP6, which heavily impacted later fitting of signatures. Hence, a bespoke version of SBS10b was constructed by subtracting the estimated contribution of SBS10a from HDP1. In this vein, HDP1 became SBS10a and SBS10b, and HDP6 was decomposed into SBS10b, SBS1, SBS5, and SBS28.

HDP2 and HDP7, exclusively found in patients with a *POLD1* germline mutation, had not been previously reported. They were renamed SBS10c and SBS10d, respectively, and not subjected to decomposition. HDP13 which was present in multiple samples from an individual with a *POLE* germline mutation was renamed SBS91. HDP0, 8 and 11 only had one major contributor and were replaced by the corresponding reference signatures SBS5, SBS88, and SBS89. HDP10 demonstrated substantial contributions of SBS31 and SBS35 which are attributed to the effects of platinum-based chemotherapy. However, the report of a spectrum of signatures due to these chemotherapy agents[249] and the relatively poor reconstitution using SBS31 and SBS35, favoured the retention of HDP10 without further decomposition hence, it was named SBS35-like. A similar approach was used for the capecitabine-related component HDP12, which resembles SBS17b but has closer similarity to previously reported therapy-related signatures[60]. Hence HDP12 was renamed SBS17b-like. Therefore, in total 14 signatures were identified: SBS1, SBS5, SBS7a-b, SBS10a-d, SBS17b-like, SBS28, SBS35-like, SBS88, SBS89 and SBS91.

SigProfiler reported fewer signature components than HDP (7 vs 14) however the components it extracted had clearly recognisable counterparts in the compendium of HDP components. Additional components that were stably extracted and had close resemblance to known, reported signatures were identified by HDP but not SigProfiler.

Signatures were refitted to the mutation counts for each branch of the phylogenetic tree to establish the absolute contribution of each mutational signature using the R package SigFit (https://github.com/kgori/sigfit). To prevent overfitting, a limited subset of reference signatures was used corresponding to the HDP components identified in that patient. Signatures contributing 10% or less of the total mutation burden were excluded to prevent overfitting. This method was applied similarly in both the PPAP and MAP cohorts.

Mutational signature analysis - MAP cohort

Signature extraction was performed using similar methods to the PPAP analysis. *De novo* signature extraction was performed to extract / identify mutational signatures using HDP yielding 9 signature components (HDP N0-N9). Components showing close similarity to known reference signatures were replaced by their respective reference signature (HDP N0 as SBS5, HDP N1 as SBS36, HDP N5 as SBS5, HDP N6 as SBS38, HDP N8 as SBS17b). To deconvolute the other signature components and equate them to known COSMIC reference signatures, an expectation maximisation algorithm was used. HDP N2 was broken down into SBS18 and SBS36, HDP N3 into SBS1, SBS18 and SBS36 and HDP N7 into SBS18 and SBS88. A further component, HDP N4 was unable to be fully deconvoluted into known mutational signatures so was retained in its original format for the next stage of analysis.

Signature fitting was undertaken as described above. In total seven known mutational signatures were refitted; SBS1, SBS5, SBS17b, SBS18, SBS36, SBS38 and SBS88. Ageing signatures SBS1 and SBS5 were present in all normal intestinal crypts[46]. Lower than expected burdens of SBS1 and SBS5 were observed in most individuals in this study due to: 1) the inherent challenges of accurately estimating mutation burden in hypermutated samples and 2) the appreciable contamination of reference signatures with SBS1 and SBS5. To partially address this the HDP component corresponding to SBS36 was utilised in the refitting stage which has lower SBS1 and SBS5 contamination than the COSMIC reference SBS36 signature. Nevertheless, in individual PD44890 where SBS18 and SBS36 exposures are many tens of times greater than the normal mutation rate, the estimates of SBS1 and SBS5 are substantially lower than would be expected.

Individual PD44890 displayed a substantially elevated mutation burden (~33 fold increase) compared to healthy controls. This increased mutation burden is also many times greater than other individuals with *MUTYH* mutations included in this study. Two germline *OGG1* mutations were observed in PD44890 which may potentially contribute to the increased mutation rate in this individual. In addition, the trinucleotide spectrum of somatic mutations shows distinctive peaks with overrepresentation of C>A mutations at GCA, ACA and to a lesser extent CCA (mutated base underlined). In a recent *in vitro* study, deletion of *OGG1* was shown to induce a distinctive mutational signature that shows similarity to the spectra observed in samples from PD44890[250].

Signature components HDP N2 and SigProfiler A were abundant in PD44890 and were absent when this individual was removed from the signature extraction suggesting that this component was heavily influenced by the mutations from this individual. Initial deconvolution of HDP N2 demonstrated contributions from reference signatures SBS18 and SBS36. The combination of these two signatures recapitulated most of the peaks seen in this individual (cosine similarity of original and reconstituted signature of 0.9). This degree of cosine similarity (>/=0.9) met the prospectively set threshold for acceptability of reconstruction and thus, obviated the need to treat HDP2 as a novel/new signature. Nevertheless, since the HDP N2 signature component bore similarity to the SBSOGG1 signature, quantification of a potential relationship between the SBSOGG1 signature and the extracted signature component HDP N2 was explored. To assess a potential contribution of the SBSOGG1 mutational process, deconvolution of HDP N2 was repeated using SBS18, SBS36 and SBSOGG1. The cosine similarity was assessed comparing the reconstituted component with the original signature component. The reconstituted signature component corresponding to HDP N2 showed improved cosine similarity metrics using SBSOGG1 when compared to deconvolution without SBSOGG1. This would support the proposal that SBSOGG1 may contribute to the spectrum observed in samples from this individual.

Lastly, it is noteworthy that the combined spectra of SBS36 and SBSOGG1 also bore close similarity to the mutational spectra and signature components observed in several cell and tissue experiments ascribed to SBS18 and thought to be due to the effect of reactive oxygen species (ROS) related DNA damage. This may possibly suggest that the known reference mutational signature SBS18 may be caused by the mutational processes, SBS36 and SBSOGG1.

Mutational signatures extracted using the HDP method were validated using SigProfiler. Using the same input data, SigProfiler generated 5 signature components (Sigprofiler A-E), fewer than HDP (5 vs 9). SigProfiler components SigProfiler.A, SigProfiler.B, SigProfiler.C which accounted for the majority of mutations had clear counterparts among the HDP signature components (HDP 1,2,3). Additional signature components were stably extracted by HDP but not by Sig Profiler.

Mutational signature assignment

To investigate the contribution of different mutational processes to the accumulation of specific somatic driver mutations, mutational signature assignment was undertaken. This method assesses the likely causative mutational signature for individual mutations. Mutational signature assignment was performed for all SBS mutations assigned to phylogenetic trees treating each branch / edge as a unique sample, hence ensuring that mutations were not double counted. These assigned probabilities were used to subset mutations for further analysis i.e. replication strand and extended sequence context biases. This method was only applied to the PPAP cohort. Signature assignment was not conducted on the MAP cohort as the predominant mutational signatures - SBS18 and SBS36 - are very similar, this method would not be able to resolve which of these signatures a mutation arose from. Mutational signature assignment was performed using the output of the HDP SBS signature extraction.

Mutational assignment probability was defined as the probability P, that a particular mutation, i, could be assigned to a given signature, j, in genome k and was calculated as follows:

$$P_{i,j,k} \frac{w_{j,k} \cdot f_{i,j}}{\sum_i w_{j,k} \cdot f_{i,j}}$$

where $w_{j,k}$ is the proportion of mutations assigned to signature j in genome k and $f_{i,j}$ is the fraction of mutations in signature j that are the same substitution type and occur at the same trinucleotide context as mutation i.

Additional sequencing protocols

Modified duplex sequencing (NanoSeq)

DNAs from bulk blood samples were extracted as outlined above. Blood samples from normal healthy controls were obtained and processed using the following method. Whole blood was diluted with PBS and mononuclear cells (MNC) were isolated using lymphoprepTM (STEMCELL Technologies) density gradient centrifugation. The red blood cell and granulocyte fraction of the blood was then removed. The MNC fraction was depleted of red blood cells by lysis steps involving three incubations at room temperature for 20 mins/10 mins/10 mins respectively with RBC lysis buffer (BioLegend). Samples from solid tissues were isolated using laser capture microdissection and subjected to protein lysis as outlined above. Cell lysates were processed and whole genome sequenced using the NanoSeq protocol.

This modified duplex sequencing method, called NanoSeq, relies on blunt-end restriction enzymes to fragment the genome in order to avoid errors associated to the filling of 5' overhangs and the extension of internal nicks during end repair after sonication. This modified method has error rates < 5e-9[201].

Given the uneven frequency of trinucleotides in the digested genome, the strong filtering of common SNPs sites (typically occurring at CpG), and the strong dependence of mutation rates on trinucleotide contexts, the estimates of mutation burdens are normalized and projected onto genomic trinucleotide frequencies.

Let *t* denote the count of a given trinucleotide of type *i* = 1...32. The frequency of each trinucleotide is calculated separately for the genome f_i^g and for the NanoSeq experiment f_i^e where:

$$f_i = \frac{t_i}{\sum_{i=1}^{32} t_i}$$

The ratio of genomic to experimental frequencies for a given trinucleotide is:

$$r_i = \frac{f_i^g}{f_i^e}$$

There are j = 1...6 classes of substitution where the mutated base is a pyrimidine. Let s_{ij} denote the count of substitution j in trinucleotide context i, giving a total of 96 substitution classes. Each substitution count is corrected as follows:

$$s'_{ij} = s_{ij} r_i$$

The corrected substitution counts provide a substitution profile projected onto the human genome, and are also used to calculate the corrected mutation burden:

$$\beta' = \frac{\sum_{i=1}^{32} \sum_{j=1}^{6} s'_{ij}}{\sum_{i=1}^{32} t_i}$$

Sequencing for ARCH-related variants in blood - PPAP cohort

Twenty-two blood samples were subjected to deep targeted sequencing (median coverage ~10000x) using a gene panel of known drivers of clonal haematopoiesis[251]. Samples were sequenced on Illumina Hiseq4000 lanes using 75bp paired-end reads. Sequencing reads were aligned to the human reference genome (GRCh37d5) using the Burrows-Wheeler aligner (BWA-aln). ShearwaterML[228] was used to call somatic SBS mutations. This algorithm was developed to detect subclonal mutations in deep-sequencing data by modelling the error rate at each site using information from a panel of normal unrelated samples. The normal panel comprised data from 310 previously sequenced normal individuals (with no identifiable ARCH mutations) aged 42-89 years. Post-processing filtering was performed as previously described, with a requirement for variants to have at least 2 supporting reads in both directions[228]. Germline variants were excluded by removing variants with a variant allele fraction (VAF) > 0.42. Probable false positives were removed by excluding variants with a VAF < 0.005. Further filtering restricted analysis to variants that caused non-synonymous protein coding changes or introduced a stop codon.

Other analyses

Embryonic variant calling

Whole-genome sequencing of bulk blood samples were used to identify early embryonic SBS and ID mutations in the PPAP cohort. Since bulk blood represents a very polyclonal tissue, variants found in blood reflect those generated in the first few cell divisions of life[34]. Variant counts from blood samples were included in the germline and artefact filtering as described above. For SBS, a minimum VAF of 0.15 was required to be included in the embryonic set. Of the remaining SBS, 205 out of a total 385 (53%) were shared with intestinal samples, confirming they must have arisen prior to gastrulation. For ID, the minimum VAF threshold was set to 0.1 to reflect the higher levels of noise accompanying indel calling and variant read counting. For indels, this amounted to 28 out of 30 (93%) mutations.

To investigate the role of *POLE* mutagenesis in the early embryo, given the highly elevated SBS mutation rate, the mutational signature contribution to the observed SBS counts was used. Reference mutational signatures SBS1, SBS5, SBS10a, SBS10b and SBS28 were fitted to patient-specific embryonic counts using SigFit. SBS1 and SBS5 reflect the normal background mutagenesis already present in the embryo[34,252], while the other signatures are caused by defective *POLE*.

For *POLD1* mutagenesis, the number of insertions of T at homopolymers of T, the characteristic peak in ID1 was quantified. These mutations represent the dominant type of somatic indel in *POLD1* patients. Insertions were used rather than SBS mutations because of the relatively modest increase in SBS mutation rate, but much higher relative increase in the rate of insertion acquisition.

Cancer driver mutations

Cancer driver mutations were identified using two methods aiming to identify genes and mutations in this cohort that are subject to positive selection. Firstly, to identify mutations in cancer genes under positive selection in an unbiased manner modified dNdS method[21] was run. To avoid double-counting of mutations, only unique mutations (SBS and ID) which were mapped to branches of the phylogenetic trees were analysed. dNdScv was run using the following parameters; max_coding_muts_per_sample=5000 and max_muts_per_gene_per_sample=20. The extended sequence context bias observed in PPAP violates one of the assumptions of the dNdS model. Hence, to adjust for this potential source of bias, a generalized negative binomial linear model was applied to each mutation subtype accounting for the biased distribution observed. *P*-values were combined using Fisher's method and multiple testing correction was performed with Benjamini-Hochberg method. Genes with a qval of <0.05 were considered to be under positive selection.

A second phase of cancer gene mutation analysis was undertaken identifying mutations in this cohort which are codified in cancer mutation databases and exhibit characteristic traits of cancer driver mutations; an approach previously employed in the study of normal tissues[48,49]. Somatic mutations (SBS and ID) were collated per-sample from all tissues. Analysis was restricted to protein coding regions and mutations were filtered using lists of known cancer genes. Mutations in intestinal epithelium were filtered using a list of 90 genes associated with colorectal cancer and includes variants that are commonly identified in small bowel adenocarcinoma[46]. Samples from all other tissues including blood were filtered using a pan-cancer list of 369 driver genes[21]. Genes were then characterised according to their predominant molecular behaviour; dominant, recessive or intermediate (those demonstrating aspects of both types of behaviour) using the COSMIC Cancer Gene Census[15]. All candidate mutations were annotated using the cBioportal MutationMapper database (https://www.cbioportal.org/mutation_mapper). Mutations meeting the following criteria were considered to be driver mutations: truncating mutations (those that cause a shortened RNA transcript i.e. nonsense, essential splice-site, splice region and frameshift ID) in recessively acting genes, known activating hotspot mutations in dominant (and recessive) genes and lastly mutations that were in neither of the above categories but characterised by the MutationMapper database as being 'likely oncogenic' were also included in the final driver mutation catalogue. The frequency of driver mutations in histologically normal crypts from individuals with *POLE*/*POLD1* and *MUTYH* mutations were compared to 445 normal intestinal crypts[46] from wild-type individuals which was annotated and filtered using the same procedure.

Telomere length estimation

Telomere attrition is a hallmark of cellular ageing and is accelerated in certain disease processes. To assess the length of telomeres in the various tissue samples, estimation of telomere lengths was undertaken. Two established methods of telomere length estimation from next-generation sequencing data were used.

Telomerecat is a ploidy-agnostic method of telomere length estimation (to base pair resolution) from next-generation sequencing data which has been benchmarked across human and animal studies in normal tissues and cancers[253]. This method has been employed in previous studies of somatic mutations in normal mutations[46-48]. Telomere length estimates were generated for all samples in the PPAP cohort. Results for most but not all samples were plausible and showed a positive correlation with those from a second telomere length content algorithm (TelomereHunter). Approximately 30% of samples returned zero values for telomere length. Similar observations have been made in other data sets sequenced on the Illumina NovaSeq platform. Results of the algorithm based on sequencing data generated by the Illumina X10 and other sequencing platforms does not demonstrate this pattern and can be relied upon. Therefore analysis was undertaken using TelomereHunter which is a well-established method used in tumour sequencing analyses[254], shows good concordance with other methods of telomere length estimation[255] and is reliable across all tested samples sequenced on the Illumina NovaSeq platform.

Telomere content measurements were generated by running TelomereHunter using default parameters across all histologically normal crypts in the *POLE/POLD1* and *MUTYH* cohorts, and normal crypts from wild-type healthy individuals from a previous study (n=445)[46]. Modelling was undertaken to assess age-related telomere attrition in normal tissues. A linear mixed-effects model was used to assess the effect of age on telomere length and to test whether telomere attrition is greater in the disease cohort compared to normal controls. Age was fitted as a fixed effect and patient as a random effect, an additional dichotomous genotype variable was added as a fixed effect. A similar telomere length at birth was assumed hence a fixed intercept was fitted. The model fit was compared using ANOVA and the difference between models assessed using a chi-squared test. P-value thresholds of >0.05 were used.

Mutational age calculations

The genome-wide and exon-wide somatic mutation rates were calculated for PPAP and MAP cohorts and for a control dataset of normal intestinal crypts from healthy wild-type individuals (mentioned previously). Mutational ages were calculated from the median mutation rate per individual divided by the expected mutation rate in wild-type healthy controls.

Clinical phenotyping

Clinical phenotype data was collected by collaborators. In addition to details regarding the principle clinical phenotypes, specific enquiry was made regarding rarer cancer types and non-cancer diagnoses. The clinical phenotype of individuals with PPAP and MAP is central to the conclusions drawn regarding the somatic mutation theory of ageing. Thus, to ensure that potentially important clinical features were not being missed, in depth review of published and unpublished cases was undertaken and specific discussions were conducted with fellow experts who routinely care for individuals with these conditions. The outcome of these enquiries was that aside from the recognised cancer predisposition, individuals with PPAP and MAP do not have an overt acceleration or increase in incidence of common age related diseases nor do they have signs or features of accelerated ageing.

In-depth phenotyping of the individuals enrolled in these studies would have been desirable to provide more fine-grained evidence to support the observation that these individuals do not display non-cancer clinical phenotypes. Further investigation could also potentially identify sub-clinical or undeclared ageing phenotypes associated with ageing. Examples of parameters that could be investigated include physiological parameters such as glucose tolerance, blood pressure, cardiac function and investigation of kidney and liver function. In addition cognitive function and memory could potentially be assessed to investigate the possibility of cognitive and memory impairment. Prospective investigation of the type outlined above was unfortunately not possible owing to restrictions on research activities due to the global SARS-CoV-2 pandemic which coincided with this PhD.

Chapter 3 - Mutation burdens in normal and neoplastic cells from individuals with DNA polymerase mutations

Introduction

As outlined in the Introduction, mutations in the proofreading domain of the DNA polymerases are a cause of hypermutation in human cancer and, when inherited through the germline, cause a cancer predisposition syndrome, polymerase proofreading-associated polyposis (PPAP), characterised by intestinal polyposis and an increased risk of colorectal cancer. DNA replication is essential for the growth and maintenance of human tissues. Therefore, defects in the accuracy of DNA replication could conceivably be expected to impact all replicating cells in the human body. Assessment of mutation rates across normal tissues with defective polymerase proofreading may: 1) offer insight into the mutational processes associated with defective DNA proofreading in human tissue, 2) explain the mechanism that underpins cancer development in PPAP, 3) explain the tissue-specific differences in cancer risk observed in PPAP. It may also help to explain why somatic DNA polymerase mutations are observed in some cancer types but not others. Lastly, if increased mutation burdens are observed in normal tissues, this may have implications for our understanding of the somatic mutation theory of ageing. An increased somatic mutation burden in the absence of overt signs of premature ageing would invite us to reframe our understanding of the theory.

Key questions to be addressed in this chapter:

1) Are the increased somatic mutation rates observed in adenomas and cancers with somatic and germline *POLE*/*POLD1* mutations also observed in normal cells with inherited *POLE*/*POLD1* mutations?
2) If the mutation rates are increased, what is the extent of the increase?
3) Do all cells in each tissue have increased mutation burdens or are some cells unaffected?
4) Do all tissues have increased mutation rates?
5) If increased in all cell types, do all tissue have similar increases in their mutation rate or are there differences between tissue types?
6) Which mutational signatures are responsible for any increase in mutation rate in normal cells with *POLE*/*POLD1* mutations?
7) Is transformation from a normal cell to an adenoma accompanied by an increased mutation rate? If so, how great is the increase?
8) Which mutational processes are active in the process of neoplastic transformation?
9) Are the mutational processes associated with defective DNA proofreading activity present in tissues that do not demonstrate an overt cancer risk and tissues that are not known to undergo cell replication in adult life?

These questions were investigated by studying normal healthy tissues samples from the intestinal epithelium, peripheral blood and other normal tissues. Paired adenoma samples were also studied from individuals in whom these tissues were available. See Methods for further detail.

Section 1 – Clinical information and mutation rates

Clinical information and samples

Fourteen individuals aged between 17 and 72 years with confirmed PPAP were recruited to this study. All individuals carried a germline exonuclease domain mutation in *POLE* or *POLD1* (*POLE*^{L424V} (n=8), *POLD1*^{S478N} (n=4), *POLD1*^{L474P} (n=1) and *POLD1*^{D316N} (n=1)). Eleven individuals had a history of five or more colorectal adenomas. The age at first polyp ranged from 15 to 58 years. Five were diagnosed with colorectal cancer, all before age 50. All individuals had a known family history of colorectal adenoma, colorectal cancer and/or other cancers. No other consistent phenotypic abnormalities were reported. Clinical information, including the burden of neoplasia is summarised below (Table 3.1).

current study ID	sex	age	germline mutation	adenoma burden	age at first polyp	colon cancer	age at diagnosis CRC	endometrial cancer	total cancer	other phenotypic features
PD44580	M	71	*POLE*^{L424V}	11	41	0	NA	0	0	none reported
PD44581	F	65	*POLD1*^{S478N}	>10	40	2	29,50	0	2	none reported
PD44582	M	45	*POLD1*^{S478N}	8	29	0	NA	0	0	none reported
PD44584	F	71	*POLD1*^{S478N}	>30	53	0	NA	1	1	none reported
PD44585	M	51	*POLD1*^{S478N}	9	33	0	NA	0	0	none reported
PD44586	M	48	*POLE*^{L424V}	19	29	1	29	0	1	none reported
PD44587	F	72	*POLE*^{L424V}	2	45	0	NA	0	0	none reported
PD44588	M	38	*POLD1*^{L474P}	9	19	0	NA	0	0	none reported
PD44589	F	60	*POLE*^{L424V}	2	47	3	47 (2 CRCs),61	0	5*	none reported
PD44590	F	71	*POLD1*^{D316N}	5	58	0	NA	1	1	none reported
PD44591	F	46	*POLE*^{L424V}	44	36	0	NA	0	1^	none reported
PD44592	M	50	*POLE*^{L424V}	>10	46	1	45	0	2^	none reported
PD44593	M	17	*POLE*^{L424V}	1	15	0	NA	0	0	none reported
PD44594	M	46	*POLE*^{L424V}	6	27	1	26	0	1	none reported

Table 3.1 | Clinical characteristics of individuals included in this study

Colorectal cancer is abbreviated (CRC)

*Total includes duodenal carcinoma

^Total includes basal cell carcinoma

Mutagenesis in normal intestinal stem cells

The intestinal crypt comprises a population of cells arising from a single common ancestor that existed < 10 years prior to sampling[70,72,256]. Thus, somatic mutations arising from the ancestral stem cell are present in all cells in the crypt and occur at high variant allele fractions (VAFs)[46]. To investigate the somatic mutation rates and mutational signatures in stem cells with *POLE* and *POLD1* mutations, intestinal crypts were microdissected and whole genome sequenced using a bespoke low-input library preparation technique that permits the generation of high complexity DNA libraries from very low starting quantities of DNA (Methods)[242].

In total, 109 intestinal crypts (colorectum n=85, ileum n=10 and duodenum n=14) from thirteen individuals were individually sampled using laser-capture microdissection of tissues from surgical resection samples and endoscopic biopsies. Crypts were whole genome sequenced to a median 33.5-fold coverage (Methods). Samples from the fourteenth individual (PD44594, $POLE^{L424V}$) were collected during autopsy; however, preservation of intestinal tissues was poor and library yields were inadequate for sequencing. Analysis of the other tissue types from this individual are discussed later in this chapter.

Somatic base substitution (SBS) mutation rates were increased in all intestinal crypts from all individuals with *POLE* and *POLD1* mutations (Figure 3.1a). SBS burdens increased in a linear manner with age, implying that they accumulate continuously throughout life (Figure 3.1b). Different rates of accumulation were observed in each of the various germline mutation types. Individuals with $POLE^{L424V}$ had an SBS rate of 331/year (linear mixed-effects model 95% confidence interval (C.I.) 259-403, $P=10^{-12}$) (Figure 3.1a and 3.1b, Methods). The $POLD1^{S478N}$ SBS mutation rate was 152/year (linear mixed-effects model 95% C.I. 128-176, $P=10^{-17}$) and $POLD1^{D316N}$ and $POLD1^{L474P}$; 58/year (linear mixed-effects model 95% C.I. 51-65, $P=10^{-22}$).

Normal intestinal crypts from wild-type healthy individuals acquire 49 SBS per year[46] (linear mixed-effects model 95% C.I. 46-52, $P=10^{-36}$). Therefore the increased mutation rates observed in this study correspond to ~7 fold increases in individuals with $POLE^{L424V}$. Mutation rates differed between individuals with POLD1 mutations; individuals with $POLD1^{D316N}$ and $POLD1^{L474P}$ demonstrated relatively modest increases relative to $POLD1^{S478N}$ (~1.2-fold vs ~3-fold). Heterogeneity was also observed between individuals with the $POLE^{L424V}$ mutation, implying the presence of additional genetic and or environmental modifiers of the mutation rate.

Small insertion and deletion mutations rates were also increased in individuals with *POLE* and *POLD1* mutations. Similar to SBS mutations, ID mutation burdens increased in a linear manner with age, accumulating continuously throughout life (Methods). Differences were observed in the mutation rate between the different germline mutations. ID mutation rates were 13/year (*POLE*^{L424V}), 44/year (*POLD1*^{S478N}) and 12/year (*POLD1*^{D316N} and *POLD1*^{L474P}) (linear mixed-effects model 95% C.I., 10-16, 35-53, 9-16, $P=10^{-10}$,$P=10^{-13}$ and $P=10^{-9}$ respectively). By comparison, ID mutation rates in healthy controls accumulate at a rate of ~1/yr (Methods)[46].

Thus, SBS and ID mutation rates were elevated in all intestinal crypts from all individuals studied, consistent with the presence of elevated mutation rates in all epithelial cells of the intestine. Differences were, however, noted in the relative increases in the SBS and ID mutation rates. Individuals with the *POLE*^{L424V} mutation exhibited higher rates of SBS mutations compared to individuals with the *POLD1*^{S478N} mutation whereas individuals with *POLE*^{L424V} had more modest increases in the ID mutation rate than those with *POLD1*^{S478N}. This finding is consistent with observations from human cancers and experimental systems[16,180,257-261]. It is striking that the mutation burdens in normal intestinal crypts from several individuals with *POLE* and *POLD1* mutations were higher than those observed in many human cancers [5,6].

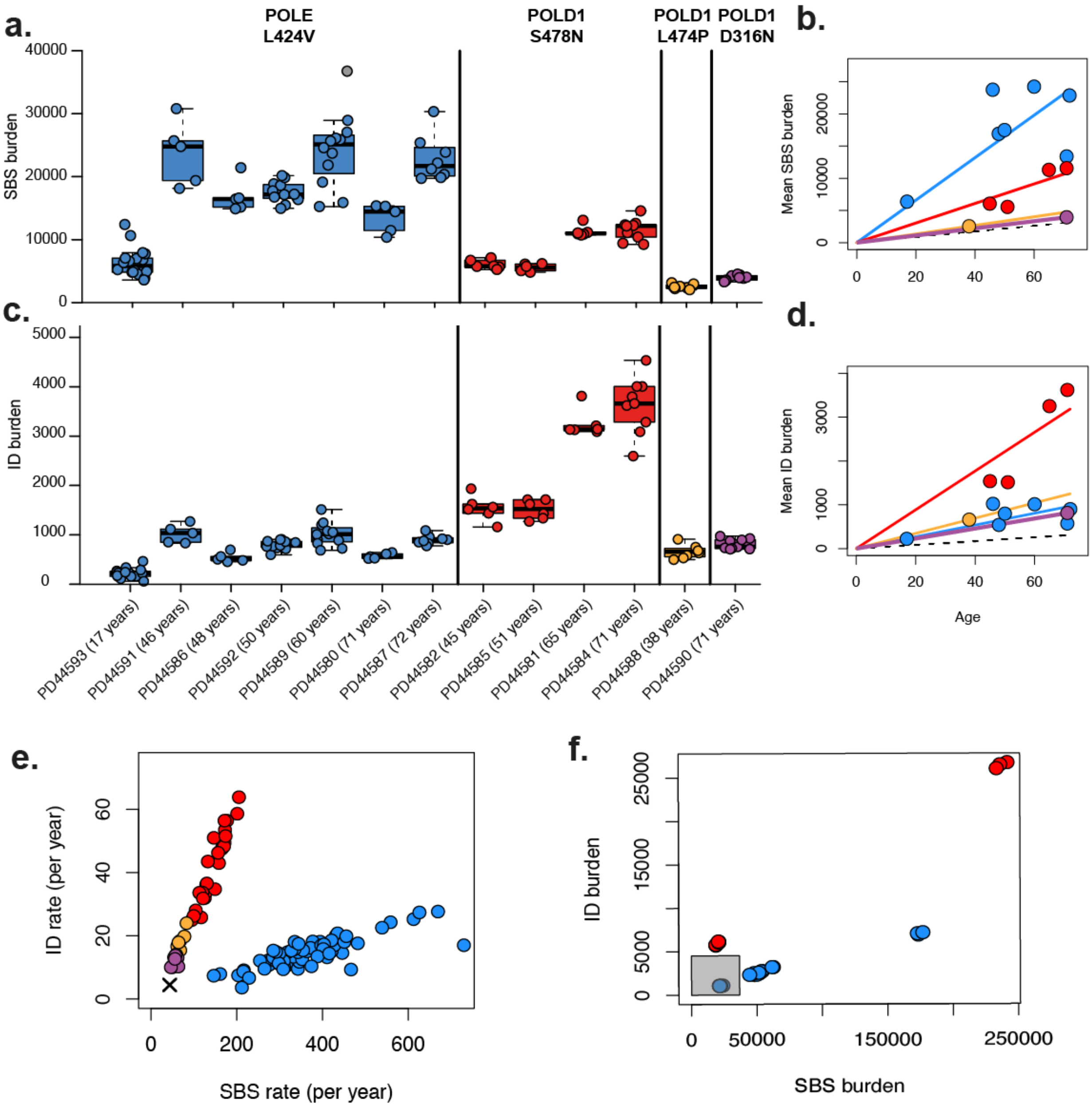

Figure 3.1 | Genome-wide somatic mutation burdens in normal and neoplastic crypts with *POLE* and *POLD1* mutations

(a) Box and whisker plot showing single base substitution (SBS) mutation burdens for normal intestinal crypts organised by individual and coloured by germline *POLE*/*POLD1* mutation. Scatter plots overlaid show genome-wide mutation burden for each crypt. For the box and whisker plots, the central line represents the median, the box represents the inter-quartile range (IQR) from 1st to 3rd quartiles and the whiskers represent the furthest point within 1.5 times the IQR.

(b) Mean genome-wide SBS mutation burden plotted against age (years). Each dot represents the median SBS burden per individual. Dots are coloured according to the germline *POLE*/*POLD1* mutation.

Linear regression lines are coloured according to the germline *POLE*/*POLD1* mutation and the black dashed line indicates the rate of SBS accumulation in normal crypts from wild-type healthy individuals.

(c) Box and whisker plot showing insertion and deletion (ID) mutation burdens for normal intestinal crypts organised by individual and coloured by germline *POLE*/*POLD1* mutation. Scatter plots overlaid show genome-wide mutation burden for each crypt. For the box and whisker plots the central line, box and whiskers represent the median, inter-quartile range (IQR) from 1st to 3rd quartiles and the furthest point within 1.5 times the IQR.

(d) Plot of average genome-wide ID mutation burden against age (years). Each dot represents the median ID burden per individual. Dots are coloured according to the germline *POLE*/*POLD1* mutation. Linear regression lines are coloured according to the germline *POLE*/*POLD1* mutation and the black dashed line indicates the rate of ID accumulation in normal crypts from wild-type healthy individuals.

(e) Plot of genome-wide SBS mutation rate (x-axis) against the genome-wide ID mutation rate (y-axis) for normal intestinal crypts. Each dot represents an individual crypt and is coloured according to the germline mutation.

(f) Plot of genome-wide SBS mutation burden (x-axis) against the genome-wide ID mutation burden (y-axis) for adenoma glands. Each dot represents an individual gland and is coloured according to the germline mutation present. The grey box indicates the upper limit of mutation burden in normal intestinal crypts from individuals with *POLE* and *POLD1* mutations.

An increased burden of protein-coding mutations was also observed in normal tissues with *POLE* and *POLD1* mutations, which comprised increases in nonsense (~7 fold vs. wild-type), missense and synonymous mutations (~4 and ~3 fold respectively). There was also an increase in the number of potential cancer "driver" mutations compared with normal crypts from healthy individuals (20/109 vs 26/445 unique driver mutations, P=0.00005 Chi-squared test). The increase in potential driver mutations was, however, largely in keeping with the general increase in the burden of protein coding mutations, implying that the higher burden of putative driver mutations results from the increased mutation rate rather than positive selection. Furthermore, dNdS ratios, which are indicative of evolutionary selection, were not increased in the DNA polymerase cohort. Thus, the increase in genome-wide mutation rates appears to also cause elevated burdens of protein-coding and potentially driver mutations.

An increased burden of nonsense mutations was observed in normal intestinal crypts from individuals with PPAP compared to wild-type controls (Figure 3.2c), this is likely due to the tendency of the polymerase mutational signatures to generate stop codons at specific trinucleotide contexts[262,263].

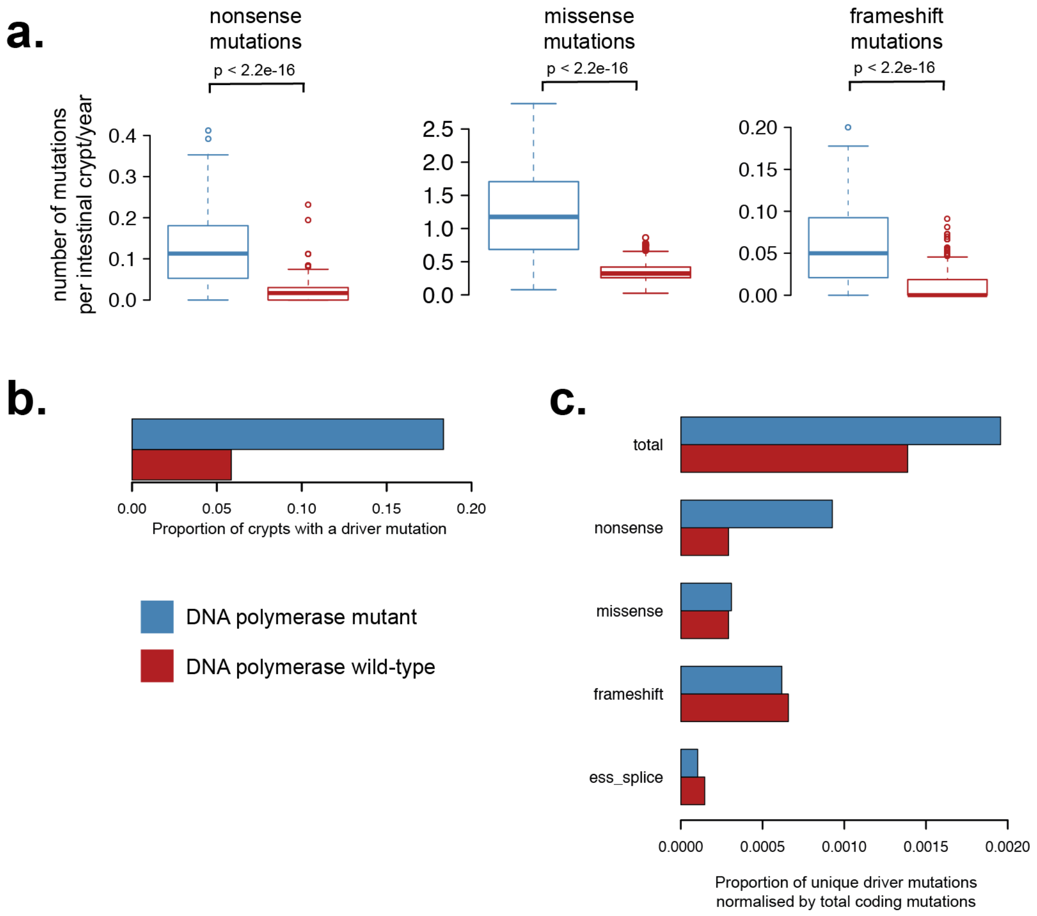

Figure 3.2 | Coding mutation burden in normal intestinal crypts

(a) Box and whisker plots showing the number of mutations per intestinal crypt per year for nonsense, missense and frameshift mutations. *POLE*/*POLD1* crypts shown in blue and wild-type controls crypts in red. For the box and whisker plots the central line, box and whiskers represent the median, inter-quartile range (IQR) from 1st to 3rd quartiles and the furthest point within 1.5 times the IQR.

(b) Proportion of normal intestinal crypts with cancer driver mutations. SBS mutations from n=445 wild-type crypts (red) and n=109 *POLE*/*POLD1* crypts (blue). Wild-type data from Lee-Six et al[46].

(c) Proportion of crypts with a driver mutation normalised for the total coding mutation burden. No statistically significant increase was observed in the total driver mutation burden (Chi-squared test p > 0.05).

The majority of individuals with *POLE* mutations develop colorectal polyposis and almost 70% develop colorectal adenocarcinoma, whereas only ~15% of individuals with *POLE* mutations develop duodenal polyposis and ~10% develop duodenal adenocarcinoma[185]. However, despite marked differences in cancer incidence, the mutation rate in the colon and small bowel of individuals with *POLE*^{L424V} was comparable (346 SBS/yr vs 334 SBS/yr, for large and small bowel respectively). This observation mirrors the similarity in mutation rates observed in normal colon and small intestinal from healthy individuals[46,82].

The burden of other mutation types including CNVs and structural rearrangements, was similar to normal intestinal crypts from healthy individuals[46]. However, in one individual, PD44592, who had been treated with oxaliplatin for colorectal cancer, 83% (10/12) of crypts had one or more somatic structural rearrangement, compared with ~15% of crypts in healthy individuals. Higher rates of accumulation of SVs have previously been observed in neoplastic cells exposed to platinum drugs[264] but have not been reported in cells from normal human tissues exposed to platinum treatments.

Neoplastic glands from 6 colorectal adenomas and one colorectal cancer (total n=3 individuals) were isolated and genome sequenced. The mutation burden in these glands was higher than in normal crypts from the same individuals sampled at the same times. However, the extent of increase varied substantially between neoplasms (Figure 3.1f). Thus, SBS and ID rates appear to increase in the process of transformation from a normal intestinal crypt to a neoplastic gland in PPAP, in keeping with observations from non-predisposed individuals[83,265].

Section 1 summary:

1. SBS and ID mutation burdens were increased in all normal crypts from all individuals with inherited POLE and POLD1 mutations.
2. Mutation accumulation in normal tissues with *POLE* and *POLD1* mutations is linear and continuous throughout life.
3. The extent of increase in the mutation rate varies between individuals according to the *POLE* or *POLD1* mutation.
4. In cells with *POLE* mutations, greater relative increases in SBS than ID are observed. Conversely in cells with *POLD1* mutations, greater relative increases in ID than SBS are observed.
5. Protein-coding and driver mutation rates are also increased in normal crypts with *POLE* and *POLD1* mutations.
6. Mutation burdens are further increased in the process of neoplastic transformation.

Section 2 - Mutational signatures in normal tissues with *POLE* and *POLD1* mutations

Mutational signatures were extracted from the combined catalogue of somatic mutations identified in normal crypts and adenoma glands of individuals with *POLE* and *POLD1* mutations. Eleven mutational signatures were observed in total, nine of which are have been previously reported; SBS1, SBS5, SBS10a, SBS10b, SBS17b, SBS28, SBS35, SBS88 and SBS89 (https://cancer.sanger.ac.uk/cosmic/signatures)[6,46]. SBS1 is characterised by C>T substitutions at NCG trinucleotides (mutated base underlined) and is thought to be caused by the deamination of 5-methylcytosine. SBS5 spans multiple SBS mutation types and is of unknown aetiology. Both SBS1 and SBS5 are found in most normal cell types and cancers where they accumulate in a linear manner with age[5,6,12,42,46,48-50,201]. SBS10a and SBS10b are characterised by C>A and C>T mutations respectively and are associated with the activity of defective POLE proofreading. SBS10a and SBS10b are observed in cancers with somatic POLE mutations and are often accompanied by SBS28[6,180].

Two previously unreported mutational signatures were identified in individuals with *POLD1* mutations; SBS10c, which was abundant in normal crypts with *POLD1* mutations and SBS10d, which was abundant in an adenoma from an individual with a *POLD1* mutation. SBS10c is characterised by C>A mutations at ACC, CCA, CCT, TCA and TCT trinucleotides and SBS10d is characterised by C>A mutations at TCA and TCT trinucleotides (mutated base underlined).

Signatures of defective DNA polymerase proofreading accounted for the increased SBS mutation burden observed in normal tissues. Signatures SBS10a, SBS10b and SBS28 accounted for the additional mutation burdens observed in normal crypts from individuals with *POLE* mutations. Signature SBS10c accounted for the additional mutation burdens in normal crypts from individuals with *POLD1* mutations.

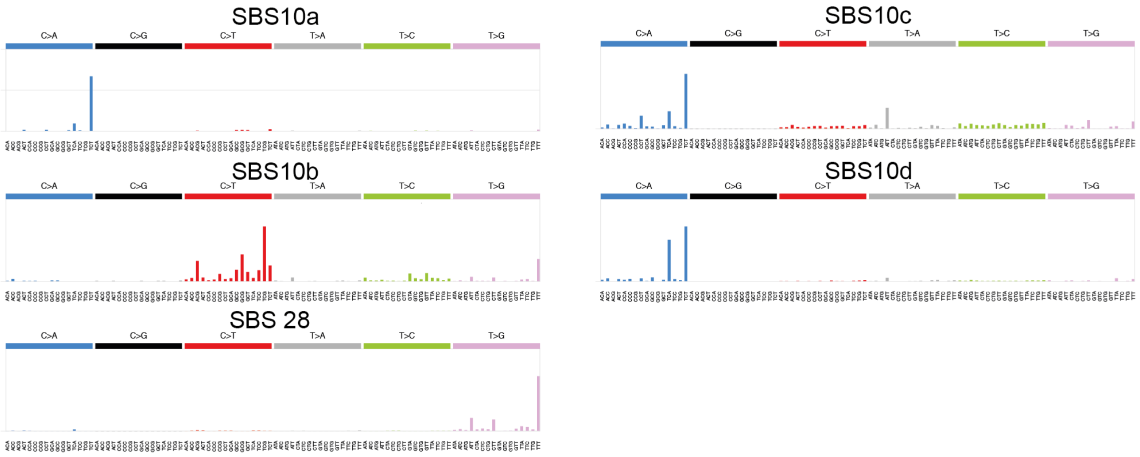

Figure 3.3 | DNA polymerase mutational signatures

Bar plots showing the trinucleotide distribution for COSMIC reference mutational signatures SBS10a, SBS10b and SBS28, associated with *POLE* mutations and novel mutational signatures SBS10c and SBS10d, associated with *POLD1* mutations.

During DNA replication, Polε and Polδ are responsible for leading and lagging strand synthesis respectively[178,179]. Consistent with these roles, defective *POLE*-generated mutational signatures demonstrate a replication strand bias (SBS10a and SBS10b) and *POLD1*-generated signatures show the opposite replication strand bias (SBS10c and SBS10d) (Figure 3.4, Methods). Mirroring this observation, significant leading and lagging replication strand bias was observed in the ID spectra of *POLE* and *POLD1* respectively. Somewhat surprisingly, a T>A leading strand bias was observed in SBS10c, whether this reflects the leading strand activity of Polδ[266] or, potentially, the misidentification of small insertions mutations as T>A SBS mutations, is unclear. An extended sequence context bias (+/-9 bases from the mutated base) was seen in all of the DNA polymerase signatures that were extracted, which was previously reported in *POLE*-related mutational signatures in cancer genomes[21,262] (Figure 3.5).

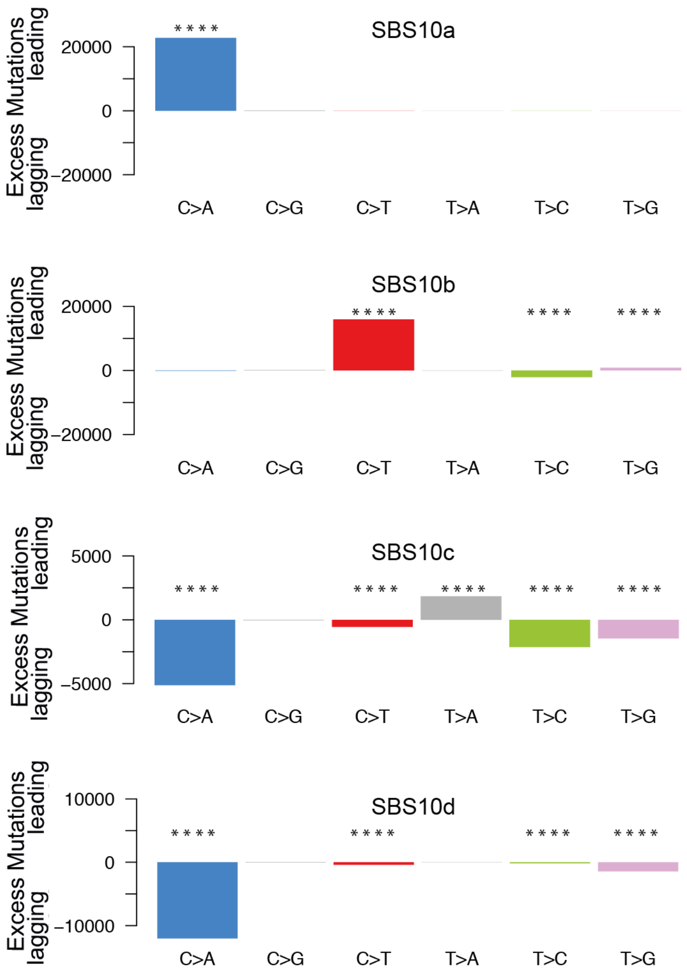

Figure 3.4 | Replication strand bias in DNA polymerase signatures

Bar plots showing replication strand bias of SBS mutations assigned to the DNA polymerase SBS mutational signatures. P-values were calculated using two-sided Poisson tests and corrected for multiple testing using the Benjamini-Hochberg method. Mutation types with statistically significant replication strand bias (adjusted $p < 0.0001$) are annotated with “****”. See Methods for a description of the approach used to assign mutations to mutational signatures.

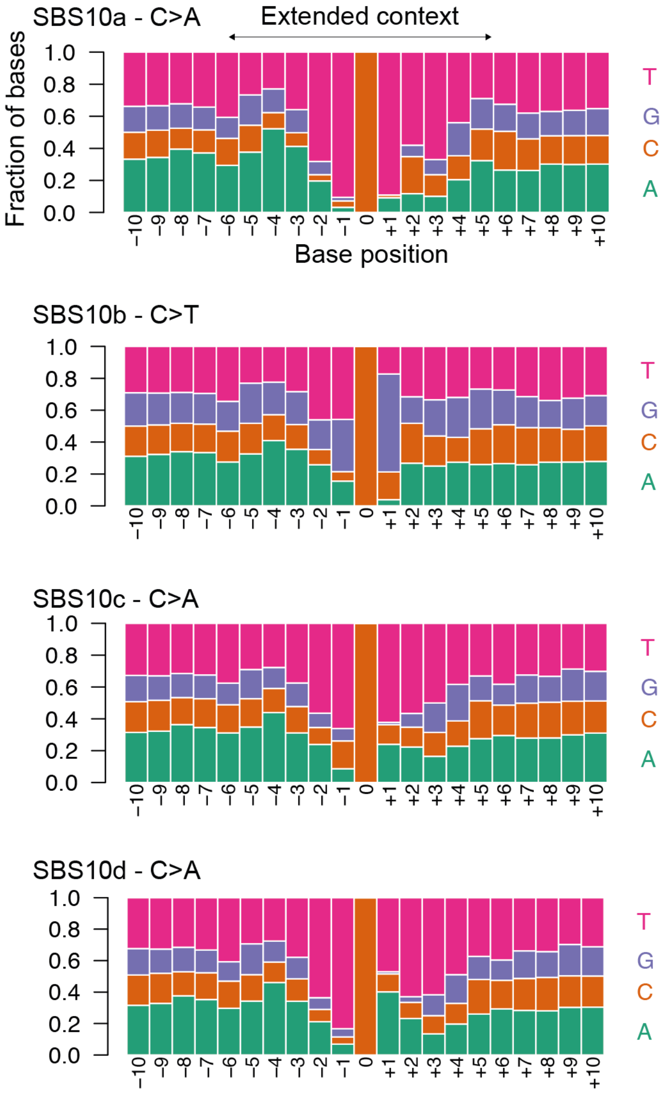

Figure 3.5 | Extended sequence context bias in mutational strand bias in DNA polymerase signatures

Stacked bar plots showing the extended sequence context of mutations assigned to the DNA polymerase SBS mutational signatures. Mutated base is displayed at position ‘0’. Y-axis shows the proportion of bases at each position.

The increased mutation burden in neoplasms was attributable to the contribution of DNA polymerase mutational signatures. In adenomas with *POLE* mutations SBS10a and SBS10b were increased, and in those with *POLD1* mutations SBS10c and SBS10d were increased. Adenoma glands showed greater relative increases in SBS10a than SBS10b compared with normal crypts. The increase in these DNA replication-related mutational signatures would imply the presence of increased cell turn-over in adenoma glands[267]. However, other possibilities, such as reduced fidelity of DNA synthesis, polymerase proofreading or mismatch repair, might also reasonably account for this observation.

Chemotherapy associated mutational signatures were observed in two individuals. SBS17b, which is principally a T>G mutational signature, associated with 5-FU / capecitabine exposure, was observed in normal crypts from an individual treated with capecitabine chemotherapy[6,60]. An SBS35-like signature, associated with exposure to platinum chemotherapies, was observed in an individual who had been treated with oxaliplatin[6,249].

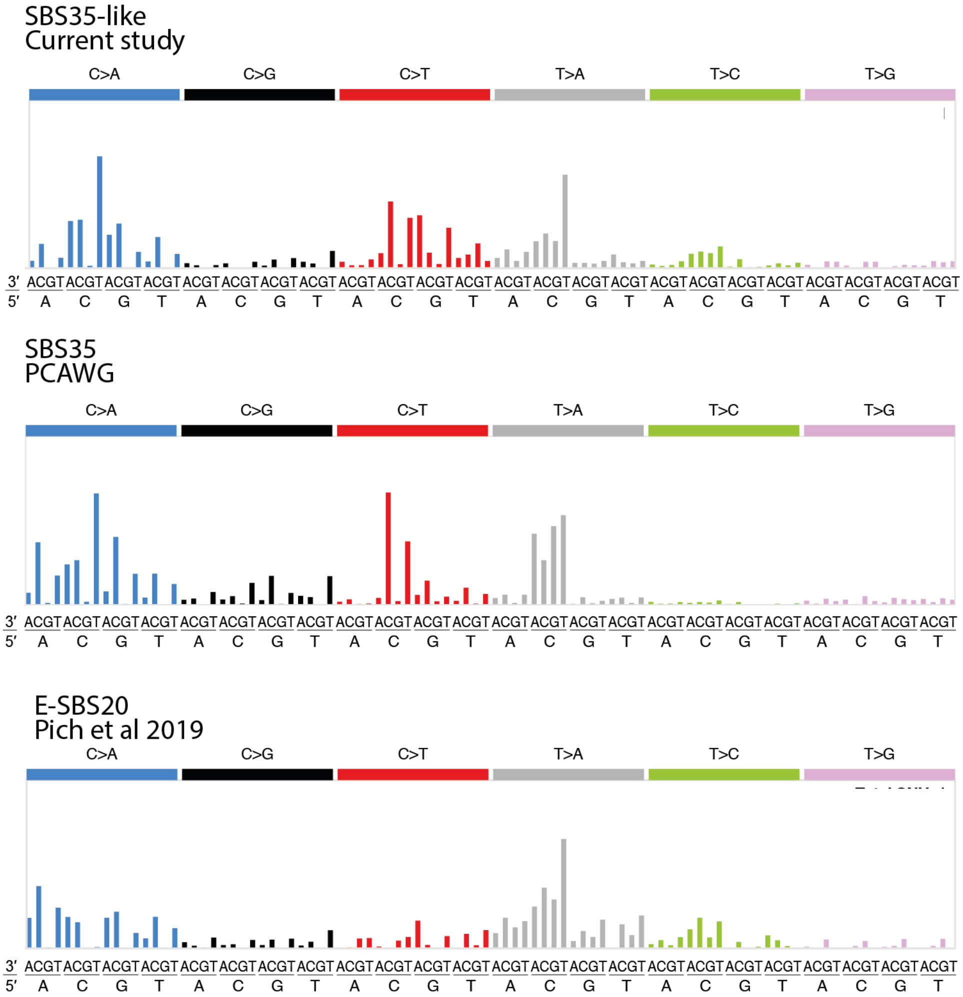

Figure 3.6 | SBS35 signature identified in normal cells exposed to platinum chemotherapy

Bar plot showing the mutational spectrum of the SBS35-like signature identified, COSMIC reference signature, SBS35 (https://cancer.sanger.ac.uk/cosmic/signatures) and the previously reported E-SBS20 signature extracted from cancer samples exposed to platinum chemotherapy[249]. Cosine similarity of SBS35-like to SBS35 was 0.88 and SBS35-like to E-SBS20 was 0.69.

The level of SBS1, SBS5, SBS88 and SBS89 in normal crypts with *POLE* and *POLD1* mutations was comparable to those found in normal crypts from healthy individuals[46], thus indicating that the processes underlying these mutational signatures are not affected by the germline DNA polymerase mutations.

Insertion and deletion mutations in normal crypts were predominantly generated by ID1, which is characterised by 1bp T insertions at T homopolymer tracts. ID1 was responsible for the additional mutation burden observed in normal cells from individuals with *POLE* and *POLD1* mutations. ID1 burdens were further increased in neoplastic cells. In addition, ID18, which is due to the mutagenic effects of colibactin producing *E.coli*, was observed in a small number of crypts that also had evidence of the SBS colibactin signature, SBS88[46].

Lastly, the mutational signature of cancer "driver" mutations from normal intestinal crypts and neoplastic glands of individuals with *POLE* and *POLD1* mutations showed a similar pattern to the genome-wide SBS and ID spectra. Therefore, the mutational processes associated with defective DNA polymerase activity appear to contribute to the accumulation of potential cancer "driver" mutations and thus the development of neoplasia in normal cells from individuals with germline *POLE*/*POLD1* mutations.

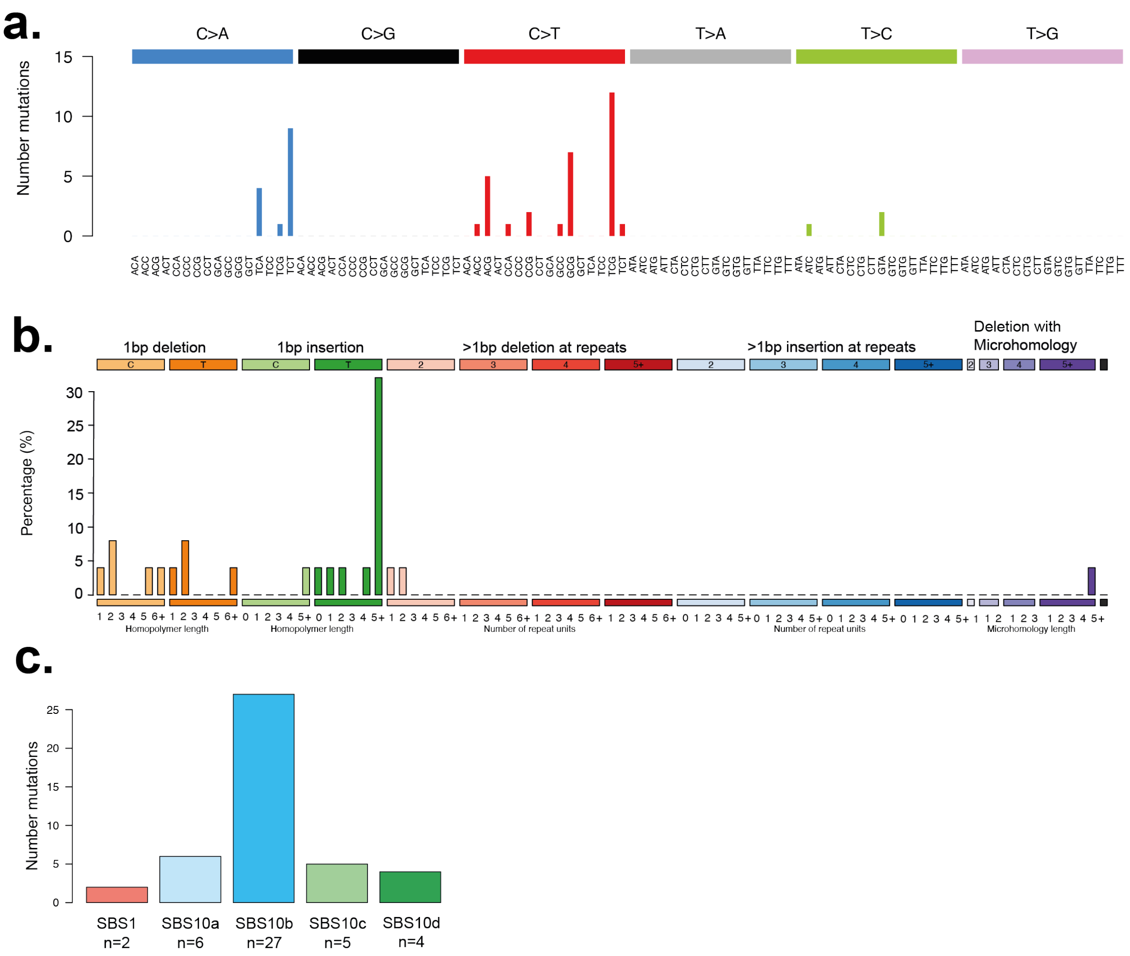

Figure 3.7 | Mutational spectrum of cancer "driver" mutations

(a/b) Spectrum of (a) SBS and (b) ID cancer "driver" mutations identified in normal intestinal crypts and adenoma glands with *POLE* and *POLD1* mutations.

(c) SBS driver mutations assigned to mutational signatures with a confidence metric >0.7 (Methods).

Section 2 summary:

1. SBS10a and SBS10b account for the additional mutation burdens observed in normal crypts from individuals with germline *POLE* mutations.
2. SBS10c accounts for the additional mutation burdens observed in normal crypts from individuals with germline *POLD1* mutations.
3. SBS10a/b show a leading replication strand bias whereas SBS10c shows a lagging replication strand bias.
4. Therapy associated signatures SBS17b and SBS35 were observed in normal crypts from two individuals previously treated with cancer chemotherapies.
5. Increased ID mutation burdens were attributable to signature ID1.
6. The mutational spectrum of driver mutations reflected the genome-wide mutational processes implying that mutational signatures of defective DNA polymerases contribute to the generation of driver mutations and hence neoplasia in PPAP.

Section 3 – Mutagenesis in other tissues with *POLE* and *POLD1* mutations

Mutagenesis in endometrium and other tissues

Colorectal and endometrial neoplasia are the principle cancer types associated with somatic and germline *POLE*/*POLD1* mutations[184,185]. To assess whether the elevated mutation rates observed in intestinal crypts are also observed in normal endometrium cells from individuals with *POLE* mutations, whole genome sequencing of endometrial tissue was undertaken. The endometrial lining is organised into glandular structures similar to the intestinal crypts that line the GI tract. The cell population of the endometrial gland is maintained by a small pool of stem cells with a recent common ancestor[268]. Therefore, somatic mutations in the stem cell will also be present in the differentiated cells of the gland and hence are relatively easy to identify in sequencing data due to their high variant allele fraction (VAF)[49].

Eleven endometrial glands were microdissected from a tissue biopsy of a 60 year old with a *POLE*^{L424V} mutation and were individually whole genome sequenced. Substantial increases in SBS (148 SBS/year vs. 29 SBS/year in healthy individuals) (Fig. 3b) and ID (6/year vs. <1/year in normal individuals) mutation rates were observed. Somatic cancer "driver" mutations are common in normal healthy endometrium[49,269,270]. Cancer "driver" mutations were observed in all but one endometrial gland (10/11) from this individual. The landscape and burden of driver mutations was broadly comparable to endometrial glands from wild-type individuals[49]. However, endometrial drivers identified in this study were almost exclusively C>T mutations with 50% of endometrial drivers occurring at T$\underline{C}$G trinucleotides (mutated base underlined), which is a characteristic peak in the *POLE*-related SBS signature, SBS10b. This would potentially suggest that SBS10b may contribute to the generation of endometrial drivers and thus the elevated cancer risk in normal endometrial tissue from individuals with *POLE* mutations.

Next, to investigate mutagenesis in other cell types, multiple normal tissues were sampled from a 47 year old individual with a germline *POLE*^{L424V} mutation. Tissue samples were isolated using laser capture microdissection and subjected to low-input library preparation and whole genome sequencing. In skin, ageing (SBS1 and SBS5) and UV light signatures (SBS7a and SBS7b) were accompanied by substantial additional contributions of *POLE* signatures (SBS10a and SBS10b). The VAFs of mutations in samples from other tissues indicated that they were highly polyclonal hence few somatic mutations were identified and mutational signatures were not easily identifiable. Therefore, microdissected fragments of tissue were subjected to an alternative method, called NanoSeq, that permits the estimation of mutation rates and identification of mutational signatures from samples in which many cell clones or lineages are intimately mixed and histologically visible clonal units are not apparent[201]. In all tissues sampled - cerebral cortex, skin epidermis, artery, smooth and skeletal muscle - normal ageing mutational signatures (SBS1 and SBS5) and mutational signatures of defective *POLE* proofreading (SBS10a and SBS10b) were present. The DNA polymerase associated signatures were additive to normal ageing signatures, thus implying the presence of elevated mutation burdens in these tissues.

Recent studies have shown that cells in some post-mitotic tissues - including cortical neurons[201,202,221,271] and smooth muscle cells[201] from healthy individuals - acquire somatic mutations throughout adult life. In this experiment, samples from all post-mitotic tissues surveyed harboured substantial additional mutation burdens due to defective polymerase mutational processes. This is somewhat surprising as the DNA polymerase signatures are thought to arise due to faulty DNA synthesis during cell replication, which is not considered to occur in post-mitotic tissues. Potential explanations that might account for the presence of a replication related mutational signature in a post-mitotic tissue include: 1) neuronal cells dividing during adult life thus contributing to the accrual of replication-related somatic mutations, 2) defective DNA polymerases being involved in DNA repair outside of DNA replication. Thus somatic mutations may accumulate in the absence of cell division, although the possibility that mitotic cell types e.g. glia may have been present in the tissues sampled, cannot be excluded. It is also theoretically possible that the elevated mutation burdens observed in these post-mitotic tissues were acquired at a very rapid rate during the cell divisions that occur during normal development. Since these tissue samples were taken at a single point in time and longitudinal samples were not available, it is not possible to exclude this explanation.

Mutation burdens in blood (n=1 *POLE*^{L424V}, n=1 *POLD1*^{S478N}) were evaluated using NanoSeq. Mutation rates in blood samples were elevated compared to wild-type controls (Figure 3.8b). Mutation rate increases in blood (~4-fold) were similar to those in endometrium (~5-fold).

However, although 12% of individuals with *POLE* mutations develop endometrial cancer, the risk of blood malignancies is not increased[185]. Therefore, increased somatic mutation burdens do not appear to translate into an increased risk of blood malignancies in PPAP.

To investigate the possibility that the increased mutation rate might drive the pre-malignant clonal expansion of blood, termed age-related clonal haematopoiesis (ARCH), deep targeted sequencing of blood samples was undertaken. Age-related clonal haematopoiesis (ARCH) is a common condition caused by a characteristic set of somatically acquired driver mutations[272,273]. Twenty-two samples from fourteen individuals were sequenced to ~10,000 fold coverage for ARCH-associated driver mutations[251]. No evidence of ARCH-related mutations was observed at the standard 2% VAF threshold. The results are consistent with previous observations that haematological malignancies are not part of the clinical spectrum observed in these individuals[184,185,274] and support the broader clinical findings that, despite the elevated genome-wide mutation rate, individuals with *POLE* and *POLD1* mutations do not show an increased frequency of age-related phenotypes.

Mutations in human germ cells and their progenitors accrue throughout life in healthy individuals and may contribute to the development of disease in the progeny[50,275,276]. To investigate mutagenesis in germ cells with *POLE* and *POLD1* mutations, sperm samples from two individuals (n=1 *POLE*^{L424V}, n=1 *POLD1*^{S478N}) were sequenced using NanoSeq. In human sperm from healthy controls, ~3 mutations are acquired per year of adult life[275], compared to ~17 SBS/yr in sperm with a *POLE* mutation and ~15 in *POLD1* (Figure 3.8b). Corresponding to ~6-fold and ~5-fold increases in the SBS rate for *POLE* and *POLD1* respectively. An increased mutation rate in sperm may confer a higher risk of inherited disorders in the offspring of affected individuals. However, the incidence of developmental or *de novo* Mendelian disorders has not been comprehensively investigated in PPAP.

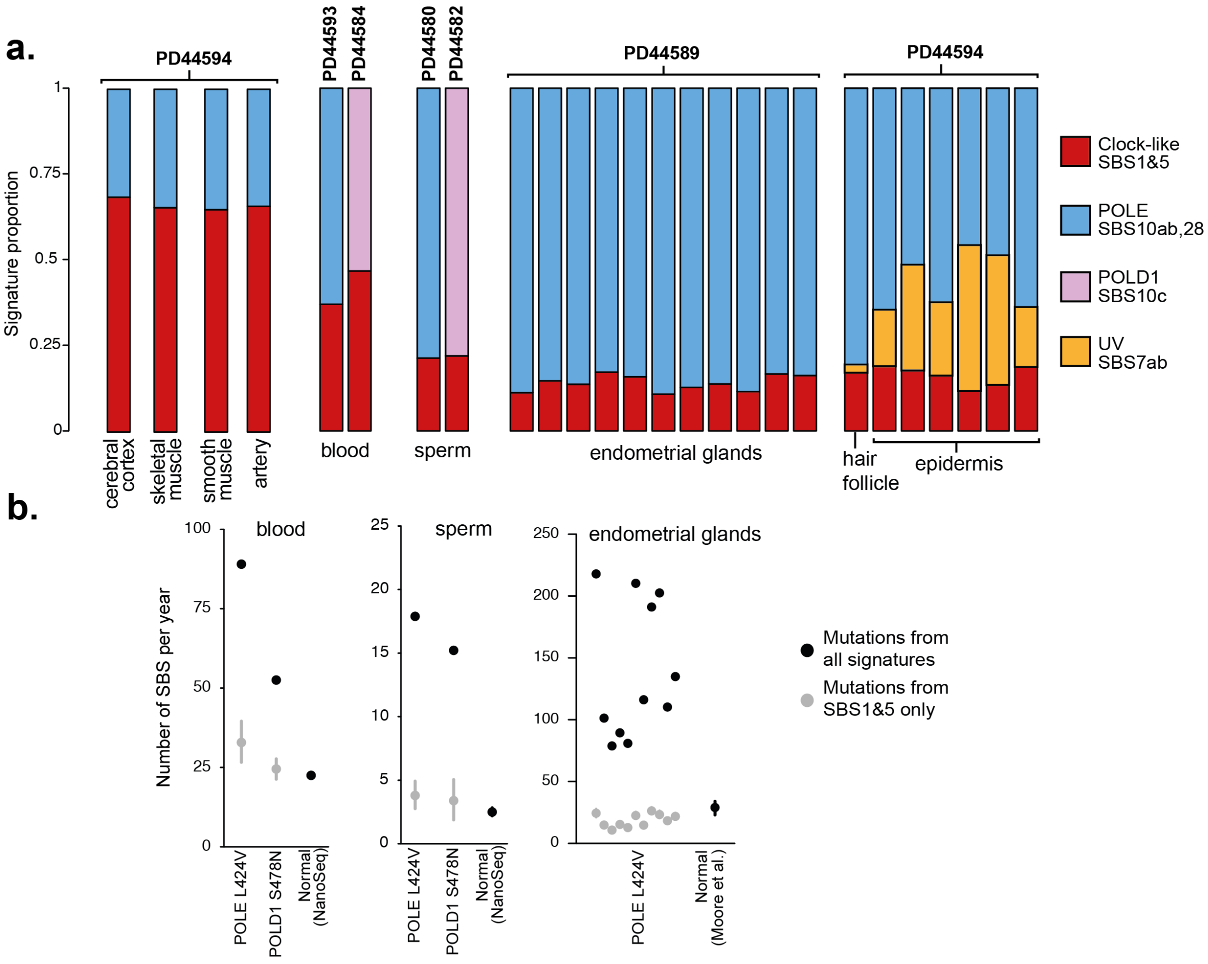

Figure 3.8 | Mutational signatures and mutation rates in other tissues

(a) Stacked bar plots showing the contribution of mutational signatures to mutagenesis. Red bars indicate mutation burden due normal ageing / clock-like mutational processes, SBS1 and SBS5. Blue and pink bars indicate mutagenesis due to defective *POLE* and *POLD1* respectively. Yellow bars indicate mutagenesis due to UV-associated mutational signatures, which are ubiquitous in normal human skin[50,203].

(b) Mutation rates in blood, sperm and endometrial glands. Black dots represent the mutation rates of SBS mutations per year due to all mutational signatures. Grey dots represent the mutation rate of SBS mutations for ageing-related signatures, SBS1 and SBS5. Bars represent 95% confidence intervals of mutation burden estimates.

Section 3 summary:

1) Mutational signatures associated with defective polymerase activity were identified in all normal tissues studied, implying the presence of ubiquitously elevated mutation rates.
2) These findings were observed both in tissues with a high cell turnover - skin, endometrium, blood and sperm - as well as tissues with limited or no cell turnover - cerebral cortex, muscle and arterial wall.
3) Elevated mutation rates in sperm were observed implying the presence of an increased mutation rate in the germline of families of affected individuals.

Section 4 – Mutagenesis during early embryogenesis and genomic distribution

Defective DNA polymerases cause elevated mutation burdens in early embryogenesis

Mutations accrue throughout development, from the first cell division onwards[205,208,275]. To assess whether the mutational processes of defective Polε and Polδ are present at the earliest stages of life, the mutational spectra of embryonic mutations were examined. Embryonic mutations were identified and mutational signatures were fitted to the mutation counts from each individual (Methods). For individuals with *POLE* mutations, SBS signatures (SBS10a and SBS10) accounted for a significant proportion of mutations present. In individuals with *POLD1* mutations, where SBS counts were lower, the number of ID mutations of T insertions at T homopolymer tracts was assessed. Insertion mutations were increased in most but not all individuals with *POLD1* mutations. Thus the mutational processes associated with defective *POLE* and *POLD1* are present in some individuals from the earliest stages of life. Variation in the burden of embryonic mutations was largely explained by the inheritance pattern of the polymerase mutation. Individuals who had a maternal inheritance pattern tend to have a higher burden of embryonic polymerase mutations than those who inherited the mutation from their father had lower embryonic mutation burdens. When the DNA polymerase mutation is maternally inherited, transcripts of the faulty *POLE*/*POLD1* gene are present from the first cell division. Since the embryo only starts to produce its own transcripts after the first few cell divisions, if the faulty allele is of paternal origin, the faulty protein is only produced after zygotic genome activation occurs thus protecting the cell for the first few divisions[277].

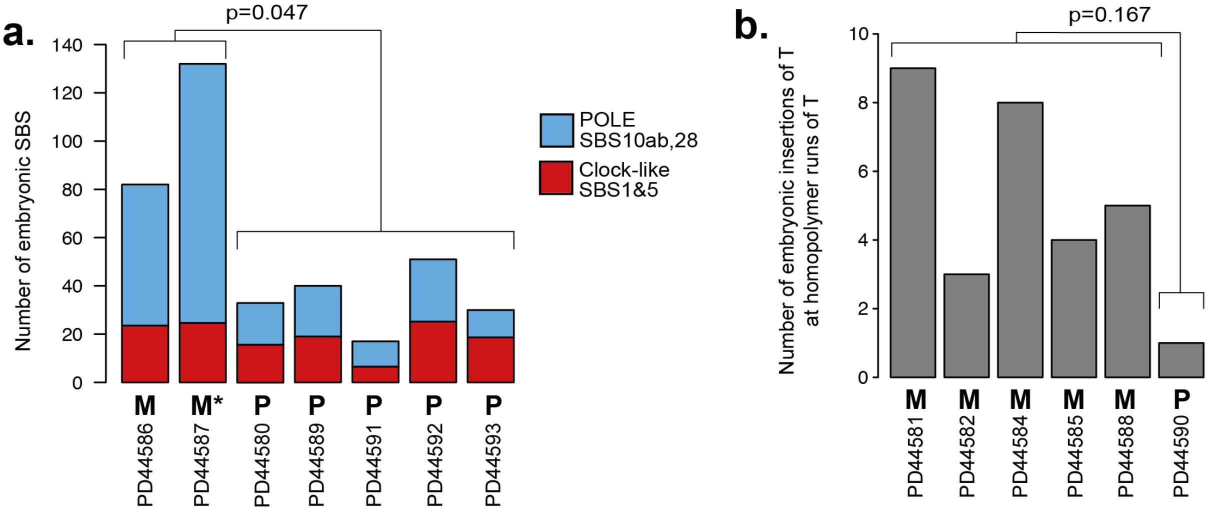

Figure 3.9 | DNA polymerase mutagenesis in embryonic development

(a) Stacked bar plots showing the SBS mutation burden and signature of embryonic mutations represented per-individual. Letters “M” and “P” indicate maternal and paternal inheritance of the mutant DNA polymerase allele. The asterisk represents inferred inheritance.

(b) Bar plots showing the ID mutation burden of embryonic mutations represented per-individual. Letters “M” and “P” indicate maternal and paternal inheritance of the mutant DNA polymerase allele.

Genomic distribution of mutations associated with *POLE* and *POLD1* mutations

To assess the potential impact of the elevated genome wide mutation burdens on protein-coding exons, the proportion of exonic SBS mutations attributable to activity of defective DNA polymerases was assessed. The relative burden of DNA polymerase-related mutational signatures was substantially lower in protein-coding exons than in intergenic and intronic regions of the genome (Wilcoxon signed-rank test $P = 6.1\times10^{-5}$) (Figure 3.10a-b). For example, in colonic crypts with $\textit{POLE}^{\mathrm{L424V}}$ mutations there was a sevenfold genome-wide increase and only a ~twofold increase in protein-coding exons. Mutation rates in untranslated regions i.e. introns were similar to those observed in intergenic regions. This relative sparing of exons may potentially be explained by a couple of factors 1) the fidelity of DNA polymerases is reduced in late-replicating regions and areas of the genome that are more distant from replication origins and 2) in addition to polymerase proofreading, replication-related errors are identified and repaired by the MMR machinery, which is more active in exonic regions than in non-coding regions of the genome[278]. A late-replication bias was observed with SBS10a but was not obvious with SBS10b, SBS10c and SBS10d. The effects of an elevated genome-wide mutation burden may therefore be somewhat mitigated by the exon-sparing associated with the polymerase mutational signatures[262].

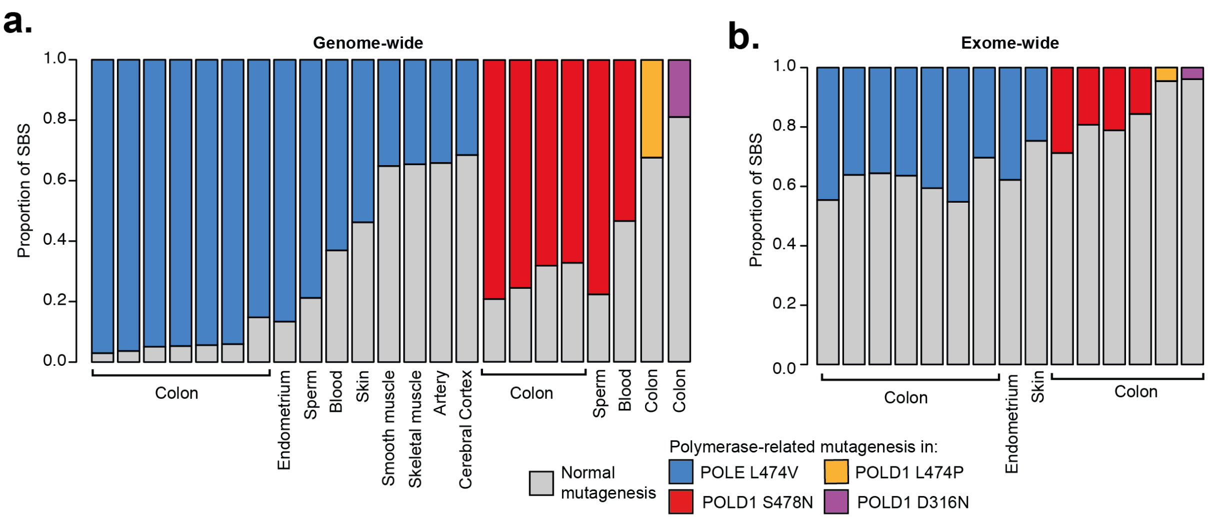

Figure 3.10 | Genomic distribution of mutations

(a) Stacked bar plot showing the genome-wide proportion of mutational signatures in normal samples. Each bar represents one individuals' tissue.

(b) Stacked bar plot showing the exome-wide proportion of mutational signatures in normal samples. Each bar represents one individuals' tissue. Exome-side SBS burdens were insufficient to estimate mutational signature extraction in several tissues.

Section 4 summary:

1. Elevated mutation burdens due to defective DNA polymerases are present in early embryogenesis.
2. Protein-coding regions show reduced polymerase-associated mutation burdens which may potentially mitigate the biological consequences of an elevated genome-wide mutation rate.

Chapter summary

In this chapter a survey of mutation burdens and mutational signatures in normal tissues from individuals with DNA polymerase mutations was undertaken. Elevated somatic mutation rates were observed in all tissues and all individuals surveyed. Differences in the extent to which mutation rates were increased, were observed between different cell and tissue types. The germline DNA polymerase mutations cause differentially increased somatic mutation rates and are associated with distinctive mutational signatures. The data from sperm show that the influence of constitutive defects in DNA polymerases is not limited to somatic tissues. The findings from this study offer an insight into the consequences of a ubiquitously increased somatic mutation rate. They can inform our understanding of the cancer risk in individuals with PPAP and also contribute to our understanding of the somatic mutation theory of ageing, both of which will be discussed in the final chapter of this thesis.

Chapter 4 - Mutation burdens and mutational signatures in normal cells with *MUTYH* mutations

Introduction

This chapter will describe the mutation burdens and mutational processes present in normal cells from individuals with inherited *MUTYH* mutations. As outlined in the introduction, MUTYH-associated polyposis (MAP) is an inherited cancer predisposition syndrome caused by biallelic germline mutations in the gene encoding MUTYH, a DNA glycosylase. Adenomas and cancers with *MUTYH* mutations have an elevated mutation burden[56-58] and it is plausible that elevated mutation rates are also present in normal tissues from individuals with MAP. Discovery of elevated mutation burdens in normal tissues may potentially explain the observed cancer risk in individuals with MAP. This chapter will address the following questions paralleling those raised in the previous chapter:

1. Are the increased somatic mutation rates observed in adenomas and cancers with *MUTYH* mutations also observed in normal cells with inherited *MUTYH* mutations?
2. If the mutation rates are increased? If so, what is the extent of the increase?
3. Do all cells in each tissue have increased mutation burdens or are some cells unaffected?
4. Is an increased mutation rate seen across all tissues in MAP?
5. If increased in all cell types, do all tissues have similar increases in their mutation rate or are there differences between tissue types?
6. Which mutational signatures are responsible for any increases in mutation rate in normal cells with *MUTYH* mutations?
7. Is the transformation from normal tissue to an adenoma accompanied by an increased mutation rate? If so, how great is the increase?
8. Which mutational processes are active in the process of neoplastic transformation?

Section 1 - Clinical information and mutation rates

Clinical information

Normal healthy tissues from ten individuals aged 16 to 79 years with biallelic germline *MUTYH* mutations were studied. All individuals had clinical features associated with MAP. Polyposis was present in all ten individuals, with between 16 and >100 colorectal adenomas, six had duodenal polyposis, five had colorectal cancer and one individual developed jejunal and pancreatic neuroendocrine cancer. Germline genotypes of the ten individuals studied included five individuals with homozygous missense mutations (three with *MUTYH*$^{\text{Y179C+/+}}$, one with *MUTYH*$^{\text{G286E+/+}}$), four compound heterozygous for missense mutations (*MUTYH*$^{\text{Y179C+/-G396D+/-}}$) and two siblings with homozygous nonsense mutations (*MUTYH*$^{\text{Y104*+/+}}$) thought to be the offspring of a consanguineous lineage.

Patient	Sex	Age	Germline mutation	Colonic adenoma burden	Colorectal cancer	Colorectal cancer age	Duodenal polyposis	Related to	Evidence of ageing phenotypes
PD44887	Female	56	c.536 G>A; p.Y179C and c.1187 G>A; p.G396D	Approx. 50 (56yr)	N	NA	N	N	none reported
PD44888	Female	49	c.312C>A; p.Tyr104* (HOM)	> 100 (46yr)	N	NA	N	PD44889 - sibling	none reported
PD44889	Female	39	c.312C>A; p.Tyr104* (HOM)	>50 (38yr)	N	NA	N	PD44888 - sibling	none reported
PD44890	Male	16	c.536 G>A; p.Y179C and c.1187 G>A; p.G396D	>100 (16yr)	N	NA	Y	N	none reported
PD44891	Male	69	c.536 G>A; p.Y179C and c.1187 G>A; p.G396D	16 (69yr)	Y	69 yr	UNK	N	none reported
PD50743	Female	79	c.857G>A; p.G286E (HOM)	>50 (78yr)	Y x 2	76 yr (rectal), 78 yr (colon)	Y	N	none reported
PD50744	Female	68	c.536 G>A; p.Y179C and c.1187 G>A; p.G396D	>100 (45yr)	Y	45 yr	Y	N	none reported
PD50745	Female	63	c.536 G>A; p.Y179C (HOM)	>50 (61yr)	N	NA	Y	N	none reported
PD50746	Female	58	c.536A>G; p.Y179C (HOM)	>20 (38 yr)	Y	38 yr	Y	N	none reported
PD50747	Male	59	c.536 A>G; p.Y179C (HOM)	"multiple" (46 yr)	Y	56 yr	Y	N	none reported

Table 4.1 | Demographics and clinical information of MAP cohort

Mutation rates in normal intestinal stem cells

To assess the mutation rate in individual intestinal stem cells, a similar approach was adopted to that employed in the previous chapter. Laser-capture microdissection was used to dissect 144 individual normal intestinal crypts (large intestine n=107 and small intestine n=37) from 10 individuals with germline *MUTYH* mutations. DNA libraries were prepared individually from each intestinal crypt and subjected to whole-genome sequencing (median ~28-fold coverage). The median variant allele fraction (VAF) of mutations called in intestinal crypts was 0.43, indicating that the cell population within each of the intestinal crypts studied arose from a single dominant clone.

The single base substitution (SBS) mutation burdens of individual crypts ranged from a median for each individual of 2,294 to 33,350 SBS mutations, equating to mutation rates of 92-1,446 SBS/year, 2-31-fold higher than normal crypts from wild-type individuals (~46 SBS/year)[46] (Figure 4.1b, linear mixed-effects model, R^2=0.98, 95% confidence interval, 1.5-2.5 and 29.8-33.1 respectively). Therefore, all normal crypts from all MAP individuals studied showed elevated somatic mutation rates (Figure 4.1a-b).

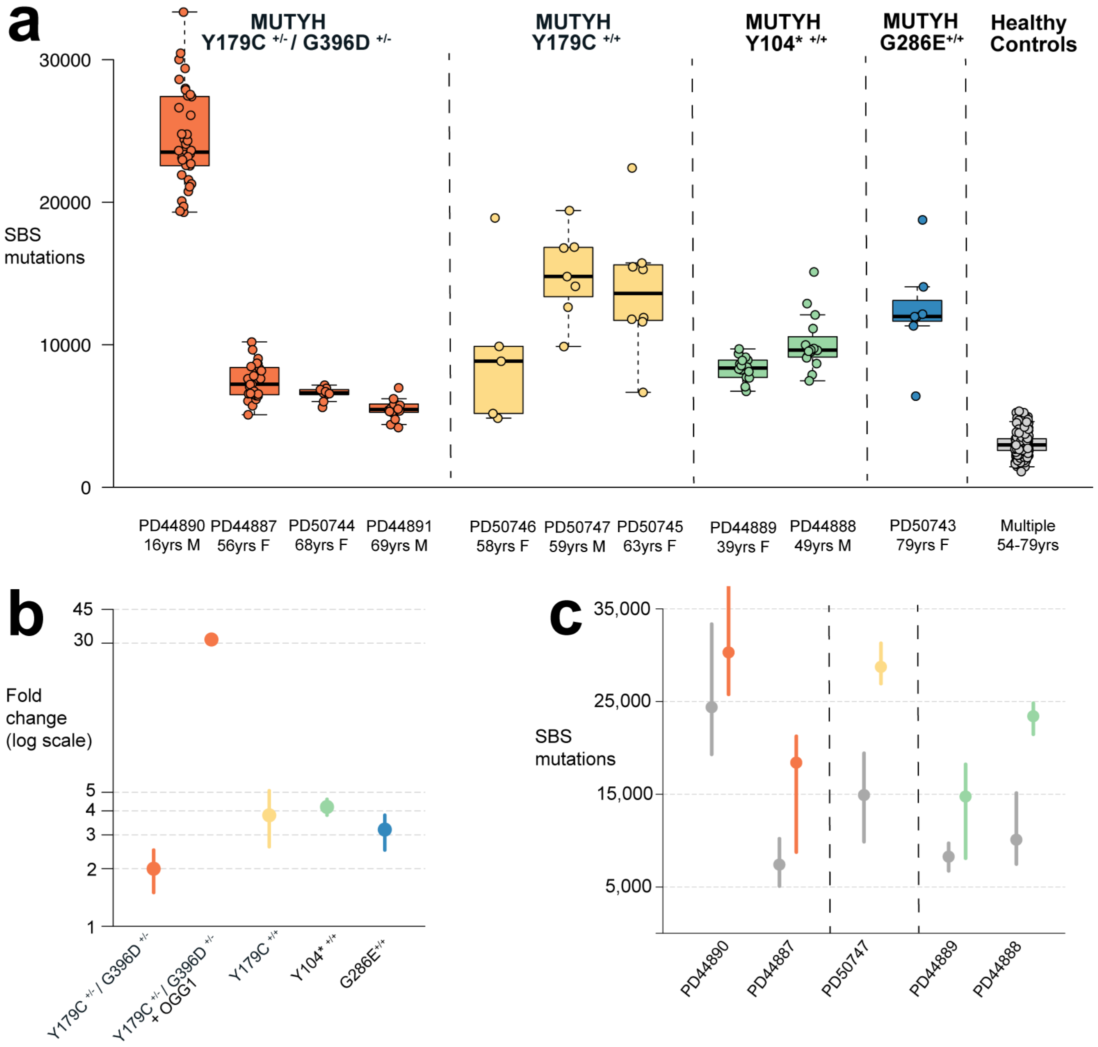

Fig 4.1 | Genome-wide somatic SBS mutation rate in normal and neoplastic cells from individuals with MUTYH mutations

(a) Genome-wide SBS mutation burden. Each dot represents one normal intestinal crypt. Colours correspond to germline *MUTYH* genotype. Boxplots show the median, interquartile range (IQR) and whiskers extend to the farthest point that is within 1.5x IQR. Healthy controls from individuals without *MUTYH* mutations are shown for comparison[46].

(b) Fold changes in genome-wide SBS mutation rate compared to wild-type controls. Dots represent the model estimate and bars represent the 95% confidence interval (linear mixed-effects model).

(c) Genome-wide SBS mutation burdens of paired normal intestinal crypts and adenomas from individuals with *MUTYH* mutations. Dots and bars from adenoma glands are coloured according to germline genotype. Dots and bars from paired normal intestinal crypts are coloured grey.

Substantial differences in the mutation rate were observed between individuals. Most strikingly, a ~31-fold increase in the SBS burden was observed in PD44890, a 16-year-old male with the *MUTYH*$^{Y179C+/- \ G396D+/-}$ genotype. The other nine individuals studied had mutation burdens that were ~two to fourfold higher than healthy controls (Figure 4.1b) (linear mixed-effects model, R^2=0.98). The cause for the large difference in mutation rate between individuals is not entirely clear. However, biallelic heterozygous germline mutations in the *OGG1* gene were identified in the germline of individual PD44890. One mutation, *OGG1*^{R46Q}, which was paternally inherited, is known to impair glycosylase activity and has previously been observed in two cases of kidney cancer[279,280]. The second, *OGG1*^{G308E}, was maternally inherited and is of uncertain significance as a colorectal cancer risk allele[281,282]. The *OGG1*^{G308E} mutation is not thought to directly impair the glycosylase activity of OGG1[283]. Whether one or both of these mutations is sufficient to increase the somatic mutation rate is unclear.

Inherited *OGG1* mutations are not known to be associated with predisposition to cancer. Since both *MUTYH* and *OGG1* are effectors of base excision repair and the resolution of 8-oxo-guanine (8-OG), it is possible that the combination of inherited mutations in both genes in the same individual contributes to the high mutation rates observed[284]. Experimental models with combined deletion of *MUTYH*, *OGG1* and *MTH1* exhibit a substantially elevated somatic mutation burden, early onset of cancer and a shortened lifespan**[284,285]**. In mice with homozygous loss of Ogg1 and Myh (Ogg1$^{-/-}$/Myh$^{-/-}$/Msh2$^{+/-}$) the cancer risk is many fold increased compared to heterozygous *OGG1* and *MUTYH* mutations (Ogg1$^{+/-}$/Myh$^{+/-}$/Msh2$^{+/-}$). Acknowledging the unknown contribution of the *MSH2* mutations, this would potentially imply that *OGG1* and/or *MUTYH* have a substantially greater contribution to cancer risk when both alleles of these two genes carry potentially inactivating mutations. In addition, evidence for the effect of the germline *OGG1* mutations on the somatic mutation rate in PD44890 is suggested by the presence of a distinctive mutational spectrum in this individual which will be discussed later in this chapter.

Significant differences in the SBS mutation rates were observed between the various *MUTYH* genotypes. The lowest mutation rates were seen in individuals with the *MUTYH*$^{Y179C+/- \ G396D+/-}$ genotype (excluding PD44890) (93 SBS/year, 95% confidence interval (C.I.) 68-116). Significantly higher mutation rates were observed in individuals with *MUTYH*$^{Y179C+/+}$ (177

SBS/year, 95% C.I. 121-236), *MUTYH*$^{\text{Y104*+/+}}$ (193 SBS/year, 95% C.I. 173-212) and *MUTYH*$^{\text{G286E+/+}}$ (145 SBS/year, 95% C.I. 117-172) (linear mixed-effects model, R2=0.98, *P*=10-10, *P*=10-7, *P*=10-23 and *P*=10-13 respectively). The mutation rate in PD44890 was substantially higher than in all other individuals, 1446 SBS/yr, (95% C.I. 1371-1520). The higher mutation rate in individuals with *MUTYH*$^{\text{Y179C+/+}}$ compared to those with the compound *MUTYH*$^{\text{Y179C+/- G396D+/-}}$ genotype correlates with greater polyp burdens and an earlier onset of colorectal cancer[158]. Furthermore, individual PD44890, who had the highest mutation rates, had a very early onset of severe disease, >100 adenomas and small bowel cancer (Table 4.1). Thus, the SBS mutation rate in normal intestinal crypts from individuals with MAP correlates with previously reported clinical features of disease severity.

In addition to the substantial variation in mutation burden that was attributable to the different *MUTYH* mutations, inter-individual heterogeneity was also observed. This was exemplified by the individuals with the compound heterozygous *MUTYH* genotype (*MUTYH*$^{\text{Y179C+/- G396D+/-}}$) who had marked differences in their mutation rates, potentially implying the contribution of additional factors including but not limited to the influence of modifying germline mutations.

Individuals with MAP have a substantially increased risk of colorectal cancer and, to a lesser extent, small intestine cancer. To investigate whether the somatic mutation rate might account for these differences in cancer risk, comparison was made between the mutation rates in small and large intestinal epithelium crypts. Since paired samples from a single individual were not available, analysis was conducted on intestinal crypts from individuals who shared the same germline mutation, *MUTYH*$^{\text{Y179C+/- G396D+/-}}$. Colonic crypts from individual PD44887 and PD44891 (n=38) were compared to duodenal crypts from individual PD50744 (n=9). Individual PD44890, who also had the *MUTYH*$^{\text{Y179C+/- G396D+/-}}$ germline mutation, was excluded from this analysis due to the potential influence of the germline *OGG1* mutations on the mutation rate. The median SBS mutation rate in colonic epithelium was 115 SBS/yr (range 60-182) and 98 SBS/yr (range 83-106) in the duodenal epithelium. There was no statistically significant difference between the mutation rate in colonic and small bowel crypts in this subset of the data (two-sided Wilcoxon test, *P*=0.193). This analysis is based on a small number of samples from a limited subgroup of individuals. However, the results suggest that the mutation rate in the large and small bowel from individuals with *MUTYH* mutations is very similar. Moreover, these results reflect the broader observation of similar mutation rates in the small and large intestinal epithelium of healthy individuals without germline *MUTYH* mutations[46,82].

Section 1 summary:

1. The mutation rate in normal intestinal epithelium was elevated in all crypts from all individuals which suggests that all intestinal crypts have increased mutation burdens.
2. The VAF of mutations implies that all cell types within the intestinal crypt carry increased burdens.
3. Marked heterogeneity was observed in mutation rates between individuals, some of which is attributable to the different germline *MUTYH* genotypes.
4. Increased mutation rates can be further modified by identifiable factors such as germline *OGG1* mutations and potentially by additional cryptic factors.
5. Relative increases in the mutation rate broadly correlate with the observed disease severity implying that the mutation rate may explain a substantial proportion of the elevated cancer risk.
6. Mutation rates in the colon and small intestine were broadly comparable.

Section 2 - Coding mutation rates in normal tissues with *MUTYH* mutations

Mutations in coding exons can alter protein structure and function, contribute to the development of cancer and may be implicated in the process of cellular ageing. Therefore, the mutation burdens in exons were examined to assess if a) the genome-wide increased mutation rate is also observed in coding exons and b) whether this could potentially give rise to protein altering mutations. Elevated mutation rates were observed in coding exons (~1.5-29-fold) (Figure 4.2). Increases in the exonic mutation rate were, however, slightly lower than the genome-wide mutation rate (Figure 4.2). There was no difference between the mutation rate in intergenic and intronic regions of the genome. The cause for the reduced mutation burden in exons is not entirely clear. Possible explanations include: differences in the underlying distribution of 8-OG across the genome; differential activity of base excision repair; or differential MUTYH activity in certain genomic regions.

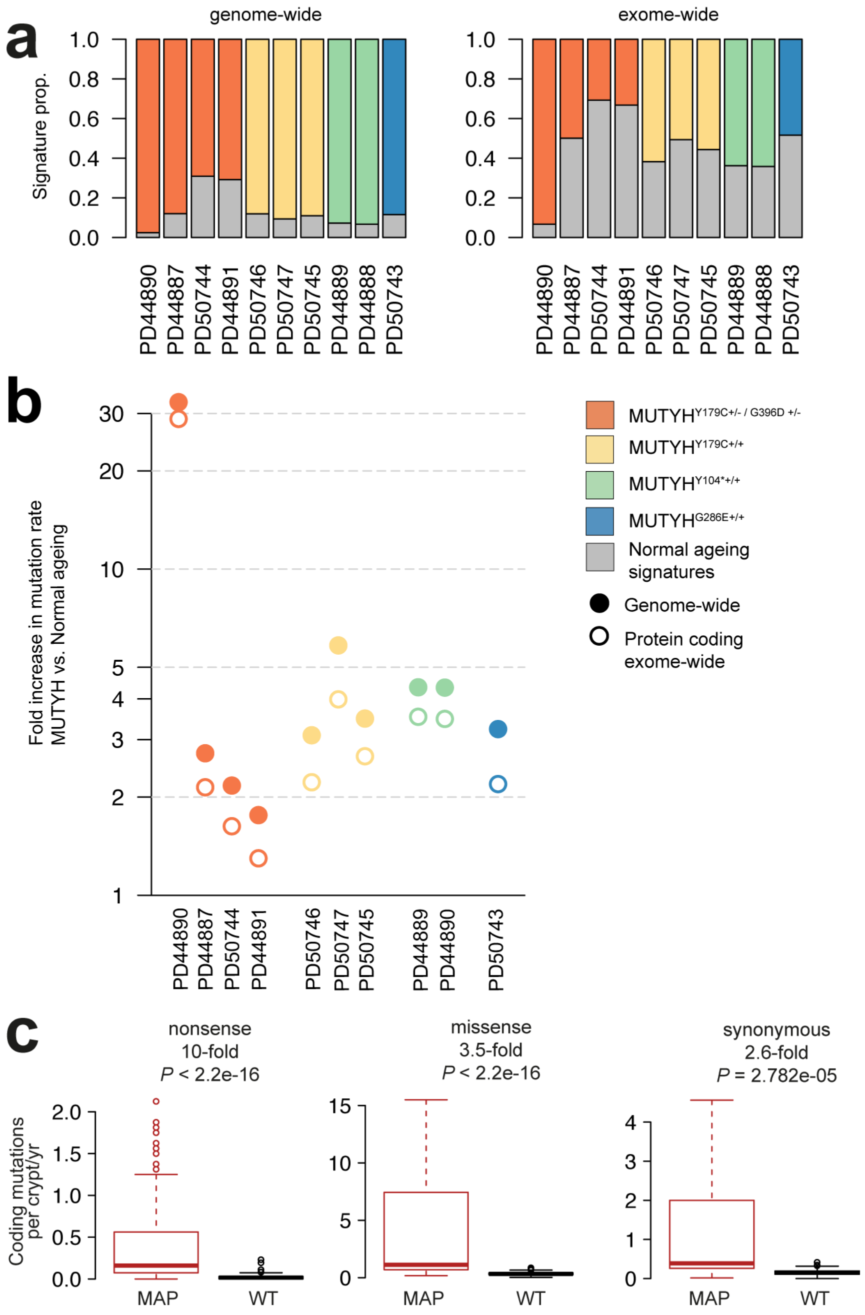

Figure 4.2 | Coding mutation burden in normal intestinal crypts from individuals with *MUTYH* mutations

(a) Proportion of mutagenesis in coding exons due to *MUTYH* mutational processes and normal age-related mutational processes. Each bar represents one individual and is coloured according to the germline genotype.

(b) Fold increase in genome-wide (filled circle) and coding (hollow circle) mutation rate compared to normal wild-type crypts. Each dot represents one individual and is coloured according to the germline genotype

(c) Coding mutation rate of nonsense, missense and synonymous mutations in normal crypts from individuals with *MUTYH* mutations (MAP) compared to normal wild-type (WT) crypts from healthy controls. Boxplots show the median, interquartile range (IQR) and the whiskers represent the furthest point within 1.5 times the IQR. Healthy controls from individuals without *MUTYH* mutations are shown for comparison[46].

Nonsense, missense and synonymous mutation rates were increased in intestinal crypts with *MUTYH* mutations compared to wild-type controls. There were 10x-fold more nonsense mutations than wild-type controls, 3.5x-fold more missense mutations than controls and 2.6x-fold more synonymous mutations than controls. Defective MUTYH generates mutational signatures characterised by C>A mutations (SBS18 and SBS36). C>A mutations at specific trinucleotide contexts may generate an increased proportion of truncating mutations as observed in the previous chapter in normal cells with germline POLE mutations. Cancers with somatic POLE mutations, bearing the mutational signature SBS10a, show an increased burden of C>A mutations at a TCT trinucleotides, resulting in the introduction of stop codons at glutamic acid residues which can result in protein truncation[263].

Mutation burdens in neoplastic cells

Next, neoplastic glands from 13 adenomas (*n*=5 individuals) were studied. SBS mutation burdens in adenoma glands were ~2 fold (range 1.2-2.5) higher than in matched normal samples sampled at the same time from the same individuals (Figure 4.1c). Therefore, the elevated mutation rate in normal intestinal crypts with *MUTYH* mutations is further increased during the process of neoplastic transformation in individuals with MAP. This is consistent with observations from sporadic neoplasia[83,265] in healthy individuals. Discussion of the mutational processes associated with this increased mutation rate occurs later in this chapter.

Other mutation types in intestinal cells with *MUTYH* mutations

Small insertion and deletion (ID) mutations accumulated at a rate of 2.1 ID/yr (linear mixed-effects model, 95% C.I. 1.2-3.0, *P*=<10-4), which is higher than in wild-type controls (1.3 ID/yr, linear mixed-effects model, C.I. 0.54-2.0, *P*=0.0011)[46]. The cause of this mildly elevated ID mutation burden is not entirely clear. However, in two individuals, potential causes of the increased ID rate were apparent. In PD44890, the high ID mutation rate (~6 ID/yr) was, at least in part, attributable to the mutagenic effects of exposure to the mutagenic toxin colibactin[46] and in PD50747, who had a rate of ~6 ID/yr , a previously unreported mutational signature, IDA, was identified. Structural rearrangements and copy number changes were found at similar frequencies in intestinal crypts with *MUTYH* mutations compared to those from wild-type healthy controls[46]. Thus, burdens of ID, CNV and SVs are broadly in keeping with wild-type tissues and do not explain the observed increase in cancer risk in MAP. In addition, telomere attrition was not accelerated in normal crypts with *MUTYH* mutations compared to wild-type controls (Methods).

Section 2 summary:

1. The coding mutation rate is increased in normal intestinal crypts with *MUTYH* mutations
2. The greatest relative increase in coding mutations was in nonsense mutations, which impair protein function and contribute to the inactivation of recessive genes including known cancer genes.
3. The structural variants and copy number mutation rates are not increased in cells with *MUTYH* mutations.

Section 3 - Mutational signatures in cells with inherited *MUTYH* mutations

Mutational signatures were extracted from the combined set of mutations identified in normal intestinal crypts and adenoma glands (Methods). *De novo* signatures were extracted using HDP[44]. HDP signature components were then decomposed / deconvoluted using a library of known COSMIC reference signatures[3,6]. Finally, the reference signatures identified in each sample were refitted to the mutation counts to identify the contribution of each mutational signature to each branch of the phylogenetic tree. See Methods for further details and for a description of the thresholds applied. Using this approach, nine *de novo* signature components HDP N0-N8 were identified. Of these, three components were identified that accounted for the majority of mutations (HDP N1-N3). N1-N3 predominantly comprised C>A mutations (Figure 4.3). Component N1 closely resembled reference signature SBS36, and N2 and N3 showed a strong resemblance to reference signature SBS18.

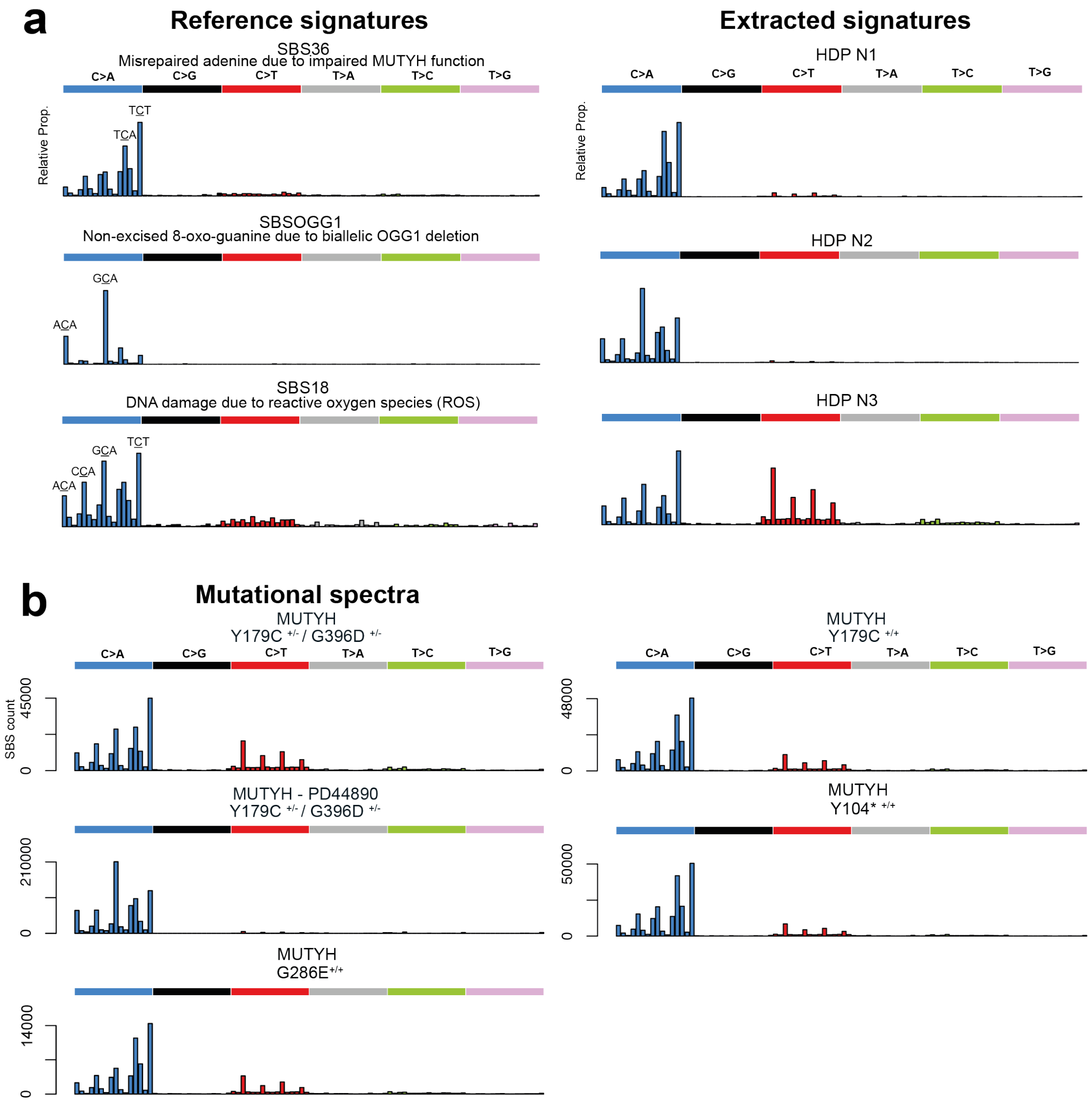

Figure 4.3 | Reference signature components, extracted signatures and mutational spectra

(a) Reference signatures SBS36, SBSOGG1[250] and SBS18. Extracted C>A signatures HDP N1-N3.

(b) Composite mutational spectra aggregated by germline *MUTYH* genotype.

Signature component N2 had dominant C>A peaks at ACA, GCA and TCA and TCT trinucleotide contexts (mutated base underlined) and closely resembled COSMIC reference signature SBS18 reaching the prospectively set threshold for similarity (cosine similarity ≥ 0.9). Thus, for the purpose of assessing signature contributions, SBS18 and SBS36 were used for refitting of all samples with N2 exposure. However, the predominant peak in HDP N2, at the C>A GCA trinucleotide is not a feature of the current reference SBS18, potentially suggesting that this component is a variant of SBS18, or perhaps, a novel mutational signature. To investigate whether N2 is an independent mutational process, *de novo* extraction was conducted using two additional mutational signature algorithms SigProfiler[6] and sigfit[43]. A signature component that resembled HDP N2 was extracted by both of these methods, lending support to the proposal that N2 represents an independent mutational process that is distinct from SBS18.

The N2 signature component was found exclusively in PD44890 in whom *OGG1* mutations were also identified. N2 bears semblance to the mutational spectrum generated through *OGG1* knockout in IPS cells, referred to hereafter as SBSOGG1[250]. To investigate whether HDP N2 is a composite of mutational signatures, additional deconvolution of HDP N2 was undertaken. HDP N2 was deconvoluted in a combinatorial approach using C>A signatures SBS18, SBS36 and SBSOGG1. Quality of the reconstitution was quantified using cosine similarity to the original HDP N2 component. This analysis showed that whilst good similarity metrics were attained with combination of the SBS18 and SBS36, optimal deconvolution was seen with the combination of SBS36 and SBSOGG1. Results are summarised below with the optimal solution highlighted in green (Table 4.2). These results would potentially support the proposal that the combination of *MUTYH* and *OGG1* germline mutations contributed to both the elevated mutation burden and the distinctive mutational signature observed in PD44890.

	SBS18	SBS36	SBSOGG1	Cosine similarity (original vs reconstituted)
SBS18 & SBS36	0.6	0.4	NA	0.9
SBS18 & SBS36 & SBSOGG1	0	0.52	0.48	0.98
SBS18 & SBSOGG1	0.52	NA	0.48	0.95
SBS36 & SBSOGG1	NA	0.52	0.48	0.98

Table 4.2 | Deconvolution of mutational signature component SBS N2

Output of HDP N2 signature deconvolution using an expectation maximisation method. Combinations of components used to deconvolute are shown in the left-hand column. Estimated contribution of each signature component is shown in columns titled SBS18, SBS36 and SBSOGG1. Cosine similarity of the original HDP N2 vs the reconstituted component is shown in the final column.

Mutational signatures present in normal cells with *MUTYH* mutations

Four reference mutational signatures were identified in all samples, SBS1, SBS5, SBS18 and SBS36 (Figure 4.4). SBS1 is due to the deamination of 5-methyl-cytosine, SBS5 is of unknown aetiology, SBS18 is thought to be caused by reactive oxidative species and SBS36 is caused by mispairing of adenine opposite 8-OG, usually the result of impaired or defective MUTYH glycosylase.

Three further reference mutational signatures, SBS88, SBS17b and SBS38 were identified in smaller numbers of samples. SBS88 is due to the mutagenic effects of a strain of *E.coli* that produces the genotoxin, colibactin. It was first identified in normal intestinal stem cells and typically occurs in the early decades of life[46,80]. SBS88 was principally identified in individual PD44890. SBS88 mutation burdens were in keeping with those observed in previous studies[46]. Thus, defective MUTYH does not appear to play a role in the genesis of SBS88. SBS17b is observed in normal tissues and cancers exposed to 5-FU/capecitabine chemotherapy and it also occurs sporadically in untreated neoplasms[6,60]. SBS17b was identified with varying proportions in normal intestinal crypts belonging to PD50747, who had previously undergone treatment with capecitabine. SBS17b burdens were further enriched in glands from two adenomas potentially implying that adenoma crypts exhibit a particular sensitivity to the mutagenic effect of this chemotherapy. Lastly, SBS38, is associated with the indirect effects of UV light, and is often accompanied by the UV light related signatures, SBS7a and SBS7b. Here, SBS38 was observed at very low levels in a small number of crypts without SBS7a/b exposure. SBS38 is a C>A signature and bears similarity to other C>A signatures identified. The presence of SBS38 may therefore represent an artefact generated through over splitting / deconvolution of signatures identified by HDP.

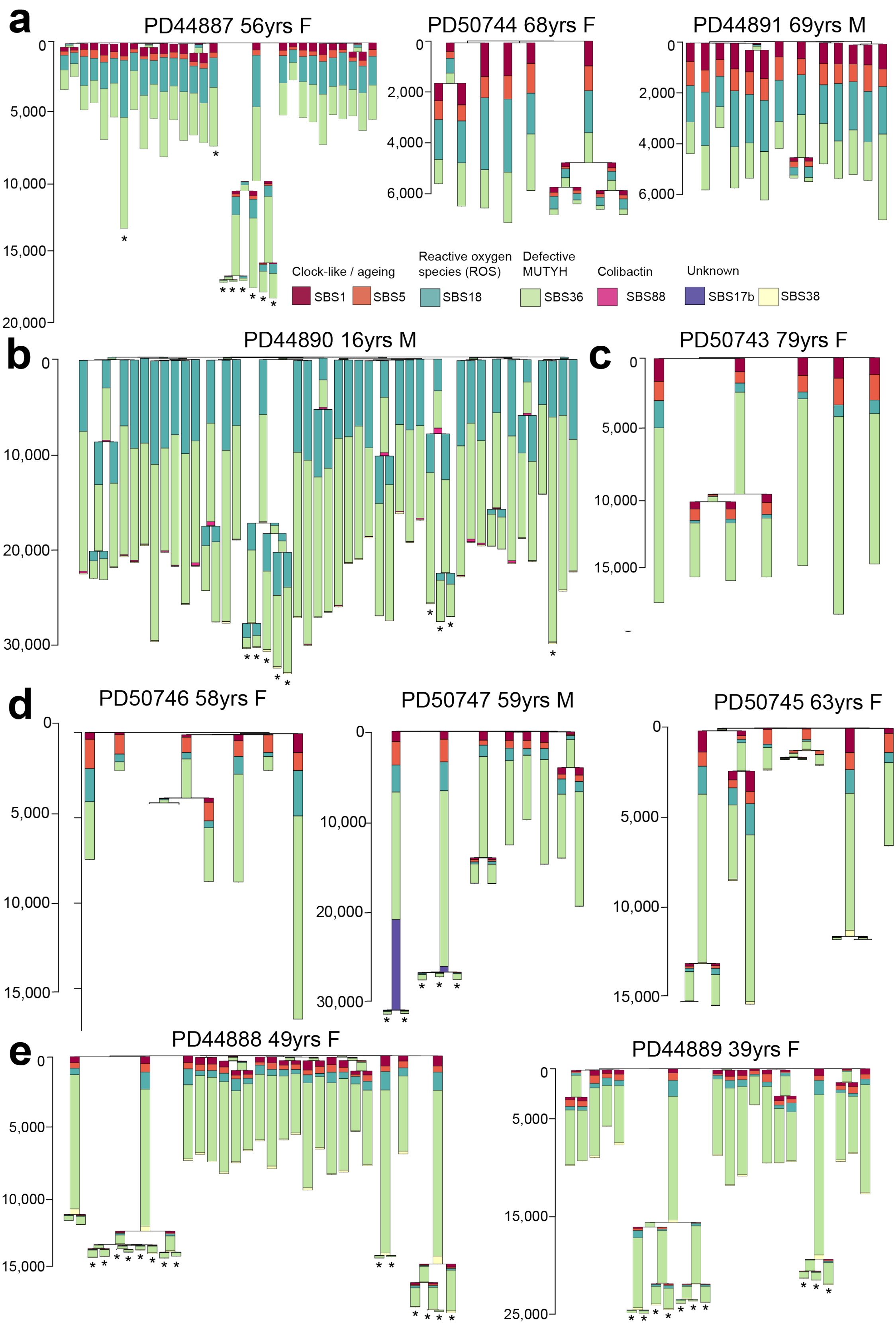

Figure 4.4 | Phylogenetic trees displaying SBS mutational signature exposure

Phylogenetic trees reconstructed from somatic mutations identified in individual intestinal crypts and neoplastic glands. Branch lengths are proportional to the number of somatic SBS mutations. Stacked bar plots are overlaid to show the contribution of mutational processes to each branch of the tree. Each tree represents the samples from one individual. Trees are organised by germline genotype: (a) *MUTYH*$^{\text{Y179C+/- G396D+/-}}$ (b) *MUTYH*$^{\text{Y179C+/- G396D+/-}}$ with *OGG1* germline mutations, (c) *MUTYH*$^{\text{G286E+/+}}$, (d) *MUTYH*$^{\text{Y179C+/+}}$ and (e) *MUTYH*$^{\text{Y104*+/+}}$. Adenoma glands bearing cancer driver mutations are indicated with an asterisk '*'.

The increased SBS mutation burdens in normal crypts from individuals with inherited *MUTYH* mutations were caused by mutational signatures SBS18 and SBS36. Burdens of SBS1 and SBS5 were not elevated in normal crypts with *MUTYH* mutations. In fact, the burden of signatures SBS1 and SBS5 identified in normal crypts with *MUTYH* mutations was lower than in wild-type healthy individuals (25 SBS/yr vs 46 SBS/yr). This is likely due to the presence of modest undeconvoluted amounts of SBS1 and SBS5 in the reference signatures for SBS18 and SBS36 thus leading to an underestimate of the true SBS1/5 burden during the signature fitting process (Methods).

Influence of germline *MUTYH* genotype on mutational signature composition

The proportion of MUTYH-associated mutational signatures, SBS18 and SBS36 varied between individuals with different germline *MUTYH* mutations (Figure 4.5). Intestinal crypts from individuals with the compound heterozygous mutation, *MUTYH*$^{Y179C+/- \ G396D+/-}$, had a higher proportion of SBS18 mutations whereas, those with the other germline genotypes, *MUTYH*$^{Y179C+/+}$, *MUTYH*$^{G286E+/+}$, *MUTYH*$^{Y104*+/+}$, had a greater burden of SBS36 mutations. The cause for this genotype-mutational signature association is unclear. However, the *MUTYH*^{G396D} has previously been associated with the preferential generation of SBS18 in neoplasia from individuals with MAP[176]. The *MUTYH*^{Y179C} mutation affects the MUTYH catalytic domain and mutations in this region alter the binding of MUTYH to DNA causing a near-complete loss of catalytic activity whereas the *MUTYH*^{G396D} mutation retains some catalytic function[149,156]. Therefore, the mutational signature SBS36 may arise due to complete or near-complete impairment of MUTYH function[149], whereas SBS18 may result from partial or incomplete loss of MUTYH function.

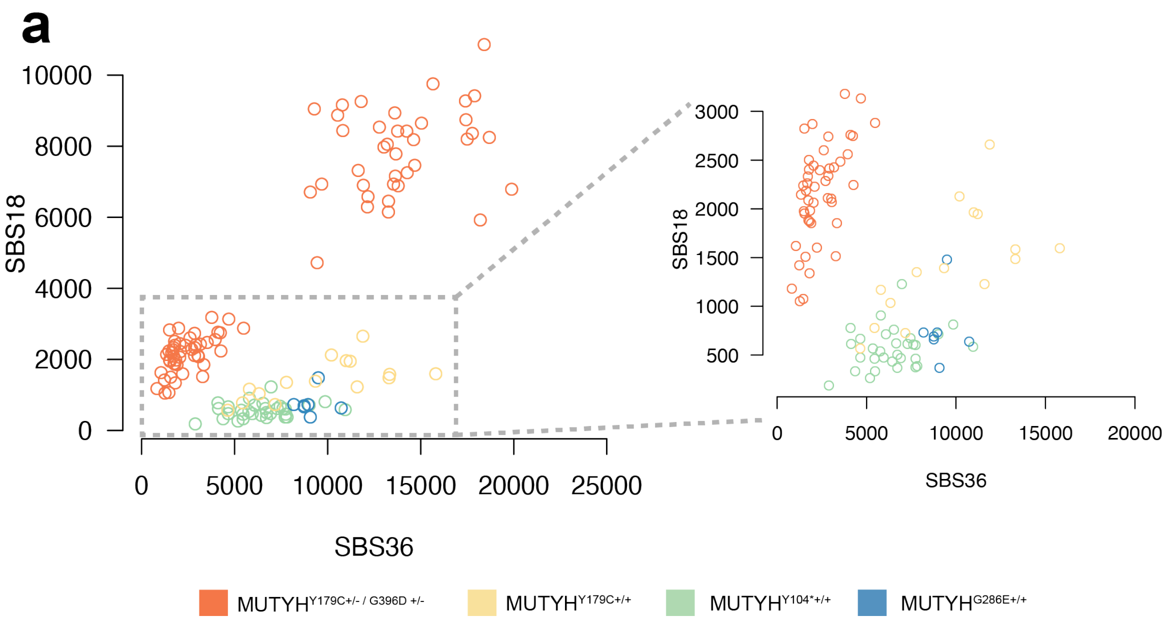

Figure 4.5 | MUTYH-associated mutational signatures in normal intestinal crypts

Scatter plot showing the burden of each MUTYH-related mutational signature. SBS36 mutation burden (x-axis) is plotted against the SBS18 (Y-axis) mutation burden. Each dot represents an individual normal intestinal crypt and is coloured according to the germline *MUTYH* genotype. The plot inset is shown without PD44890 to enable better resolution of samples with a lower mutation burden.

Mutational signatures in adenoma glands

The additional mutation burden in neoplastic glands and tissues was almost entirely due to the contribution of mutational signatures SBS18 and SBS36. The factors underlying the accelerated accumulation of MUTYH-associated signatures SBS18 and SBS36 in MAP adenomas are unclear. Potential contributors include: less efficient DNA repair in neoplastic cells, greater levels of oxidative DNA damage or impaired base-excision repair due to a shorter cell cycle.

Cancer driver mutations in normal and neoplastic cells with *MUTYH* mutations

Cancer driver mutations were identified using two methods. Firstly, the dN/dS method was applied to assess for genes under positive evolutionary selection[21]. However, no signals of positive selection were identified in normal crypts with *MUTYH* mutations. Secondly, mutations were annotated and categorised to identify likely or known hotspot mutations in oncogenes and truncating mutations in tumour suppressor genes. In total, 15% of normal crypts (22/144) bore potential cancer driver mutations which is more than double the rate observed in wild-type normal crypts (25/449) analysed using the same method[46]. A substantial proportion of the driver mutations identified were nonsense mutations (16/22), which is consistent with the broader exome-wide increase in nonsense mutations discussed earlier. The spectrum of driver mutations from normal epithelial crypts and adenomatous glands resembled mutational signatures SBS36 and SBS18 (Figure 4.6). Hence, the mutational processes associated with defective MUTYH glycosylase appear to promote the generation of potential cancer driver mutations in normal intestinal crypts from individuals with MAP. These data support previous observations that driver mutations in MAP neoplasms have a specific mutational spectrum attributable to the mutational processes of defective MUTYH[57,58,177].

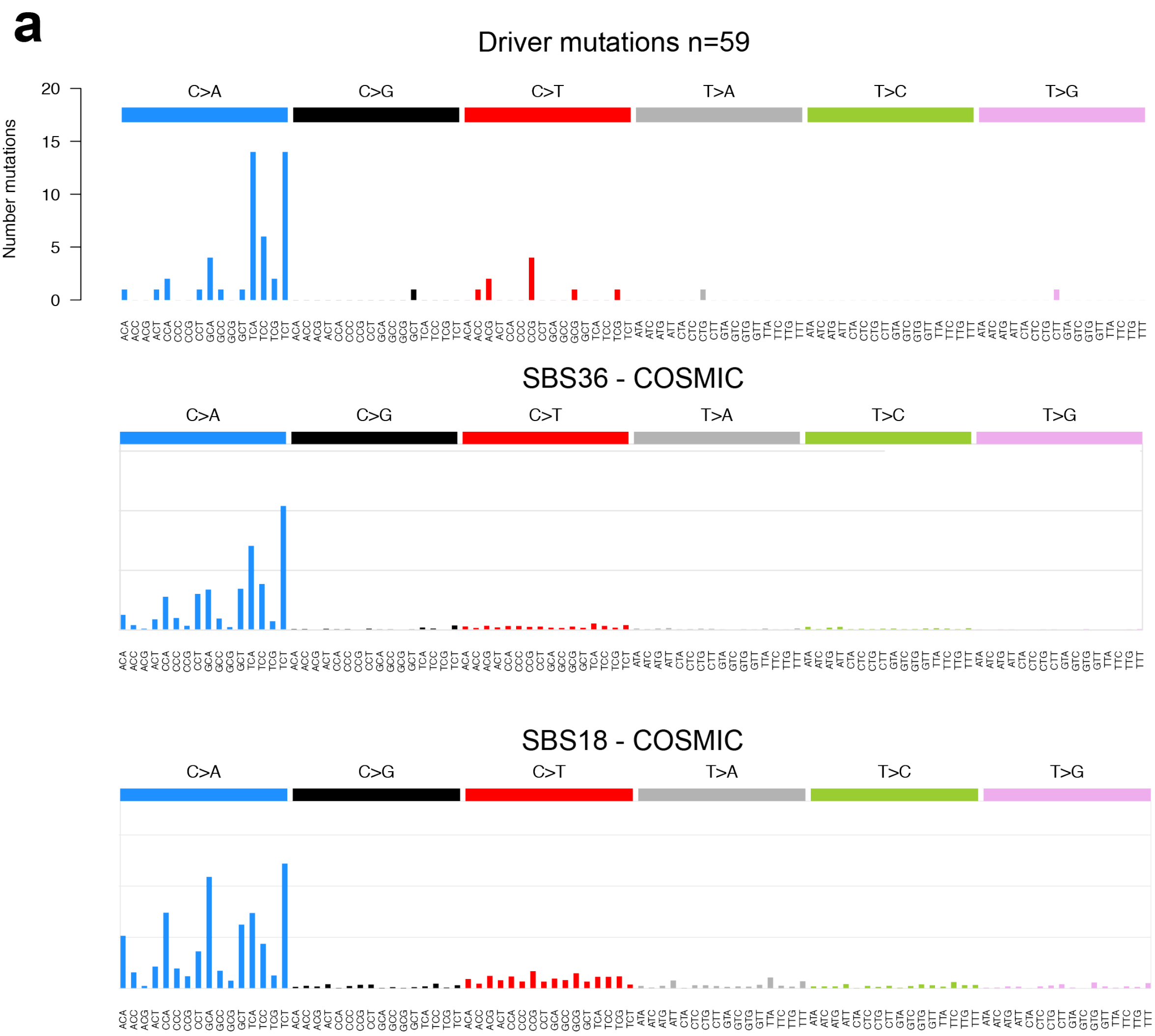

Figure 4.6 | Driver mutations in normal crypts and neoplastic glands with *MUTYH* mutations

(a) Mutational spectrum of cancer driver mutations identified in normal crypts and neoplastic glands with *MUTYH* mutations. Reference signatures SBS36 and SBS18 are shown below for comparison (https://cancer.sanger.ac.uk/signatures/)

Insertion and deletion mutational signatures in cells with *MUTYH* mutations

ID mutational signatures were extracted using HDP (Methods). In total, five indel signature components were identified (N0-N4). The predominant signature component, N1, comprised a mixture of ID1 and ID2. ID1 and ID2 are associated with strand slippage during DNA replication which results in insertions (ID1) and deletions (ID2) of T bases at T homopolymers. These two signatures are present in most human cancers and many normal tissues[46-48,50,197]. In this study, ID1 predominated in normal cells and ID2 in neoplastic cells.

A third ID signature, ID18, which is due to the mutagenic effects of colibactin and is associated with the presence of SBS88, was observed in a handful of samples from PD44890 (Figure 4.7). A further signature, IDA, was observed in one individual, PD50747, who had a history of previous capecitabine exposure. IDA was present in normal crypts (5% of mutations) and neoplastic crypts (20% of mutations). It is dominated by C insertions at C homopolymers and has not previously been described. Its aetiology is presently unclear but may be associated with the chemotherapy treatment. The absence of this signature for the other nine patients in this cohort would seem to suggest that the aetiology is unrelated to the germline *MUTYH* mutations.

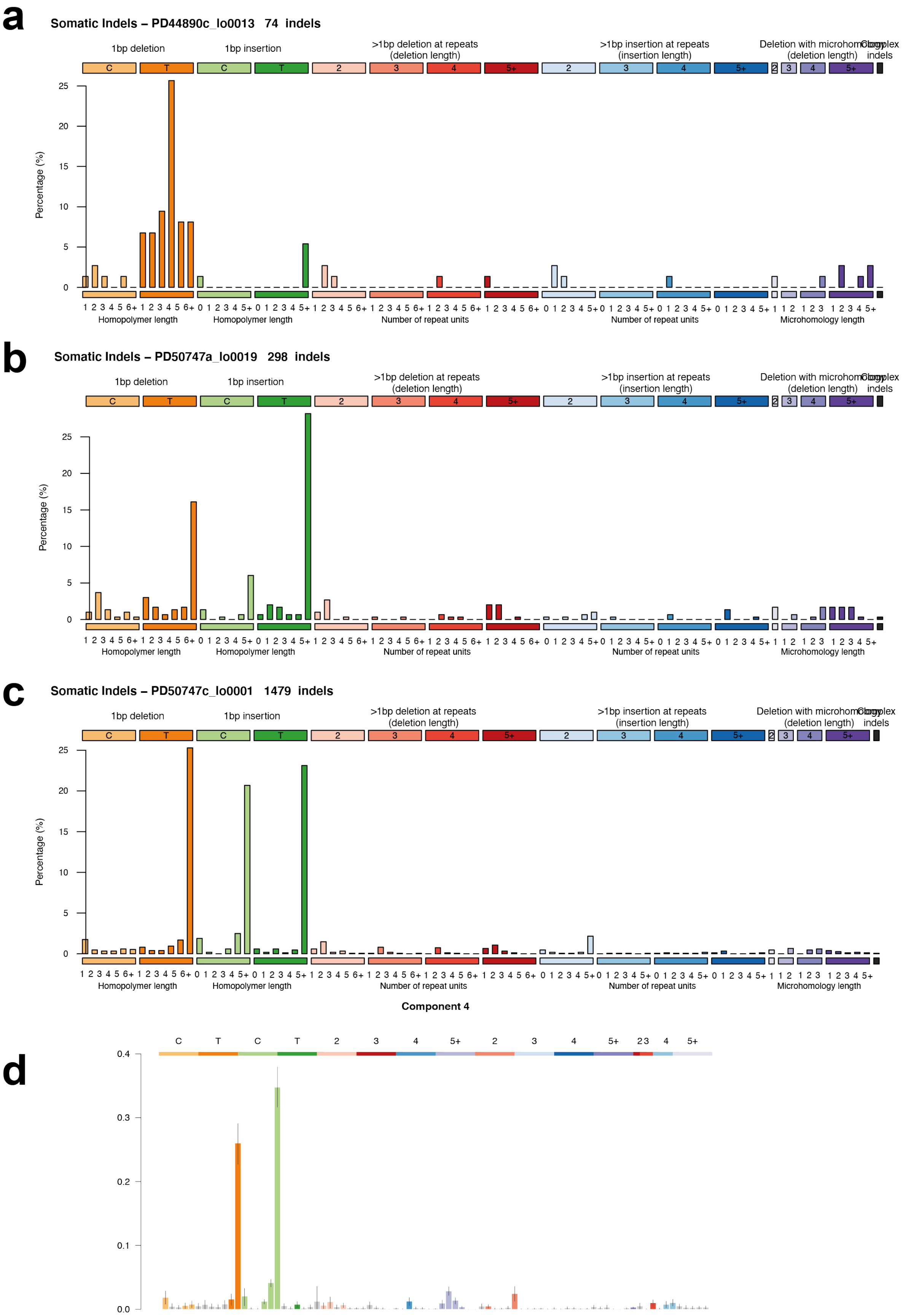

Figure 4.7 | Sporadic ID mutational signatures

(a) ID mutational spectrum showing a representative example of a normal crypt from individual PD44890 that has the colibactin-associated mutational signature, ID18.

(b) & (c) ID mutational spectra showing representative examples of a (b) normal crypt and (c) neoplastic gland from individual PD50747 with evidence of the mutational signature IDA.

(d) Plot showing the relative frequency of different ID signature contexts associated with HDP ID Component 4, referred to as IDA. Credibility intervals are represented by the thin grey bars.

Section 3 summary:

1. SBS18 and SBS36 are the predominant mutational signatures in normal intestinal cells with *MUTYH* mutations.
2. An additional SBS18-like signature was identifiable in cells with *MUTYH* and *OGG1* mutations.
3. The mutant alleles of *MUTYH* differentially affect the proportion of SBS mutational signatures present.
4. No *MUTYH*-specific ID mutational process was identified

Section 4 - Mutation burdens and mutational signatures in other normal tissues

Next, the mutation burden in two additional cell types was studied to assess whether the elevated mutation burdens observed in intestinal epithelium are also present in other tissues with *MUTYH* mutations. DNA samples extracted from whole blood, predominantly comprising peripheral white blood cells of the granulocyte lineage, and tissue lymphocytes from individuals with germline *MUTYH* mutations, were subjected to a novel modified duplex sequencing method[201]. This method enables estimation of mutation rates and mutational signatures from samples in which multiple lineages are mixed and single clonal units are not identifiable.

Increased mutation burdens were present in all blood samples from all individuals with *MUTYH* mutations (n=10 individuals) compared to wild-type controls[201] (n=9 healthy individuals age 20-80yrs)(25 SBS/yr vs 19 SBS/yr, linear mixed-effects model, R^2=0.89, *MUTYH*; 95% C.I., 19-31, *P*=10-7 and wild-type; 95% C.I., 14-24, *P*=10-6)(Figure 4.8a). The overall increase in mutation rate in blood was less pronounced than in the colon. In individuals with *MUTYH* mutations, the intestinal crypt mutation burden was on average ~13x-fold higher (95% C.I. 10-17, linear model) than in peripheral blood sampled from the same individual (Figure 4.8b). By contrast, lower relative increases of ~2.5x-fold were observed in wild-type healthy controls[46,201]. It would appear therefore, that there is a differential increase in mutation burden across tissues with *MUTYH* mutations that is not observed, to the same extent, in healthy controls. The cause for this is unclear, but may reflect higher levels of oxidative damage in the intestinal stem cells than in the bone marrow. Indeed, in wild-type individuals, the reactive-oxygen species related mutational signature, SBS18, is observed in colonic epithelium[46,82] but not in blood[35,207], implying the presence of greater quantities of unrepaired oxidative genome damage in intestinal vs. haematopoietic stem cells.

The mutation burden of tissue lymphocytes isolated from intestinal Peyer's patches (six samples from four individuals) was also increased in individuals with *MUTYH* mutations compared to healthy controls (53 SBS/yr vs 40 SBS/yr, linear mixed-effects model, R^2=0.68, *MUTYH*; 95% C.I., 21-85, *P*=0.01 and wild-type; 95% C.I., 13-66, *P*=0.01) (Figure 4.8d) further supporting a pervasive effect of defective MUTYH glycosylase activity even in tissues that are not known to be at risk of cancer in MAP. Similar to the findings from blood, the absolute and relative increase in mutation rate was less pronounced in lymphocytes compared with the intestine.

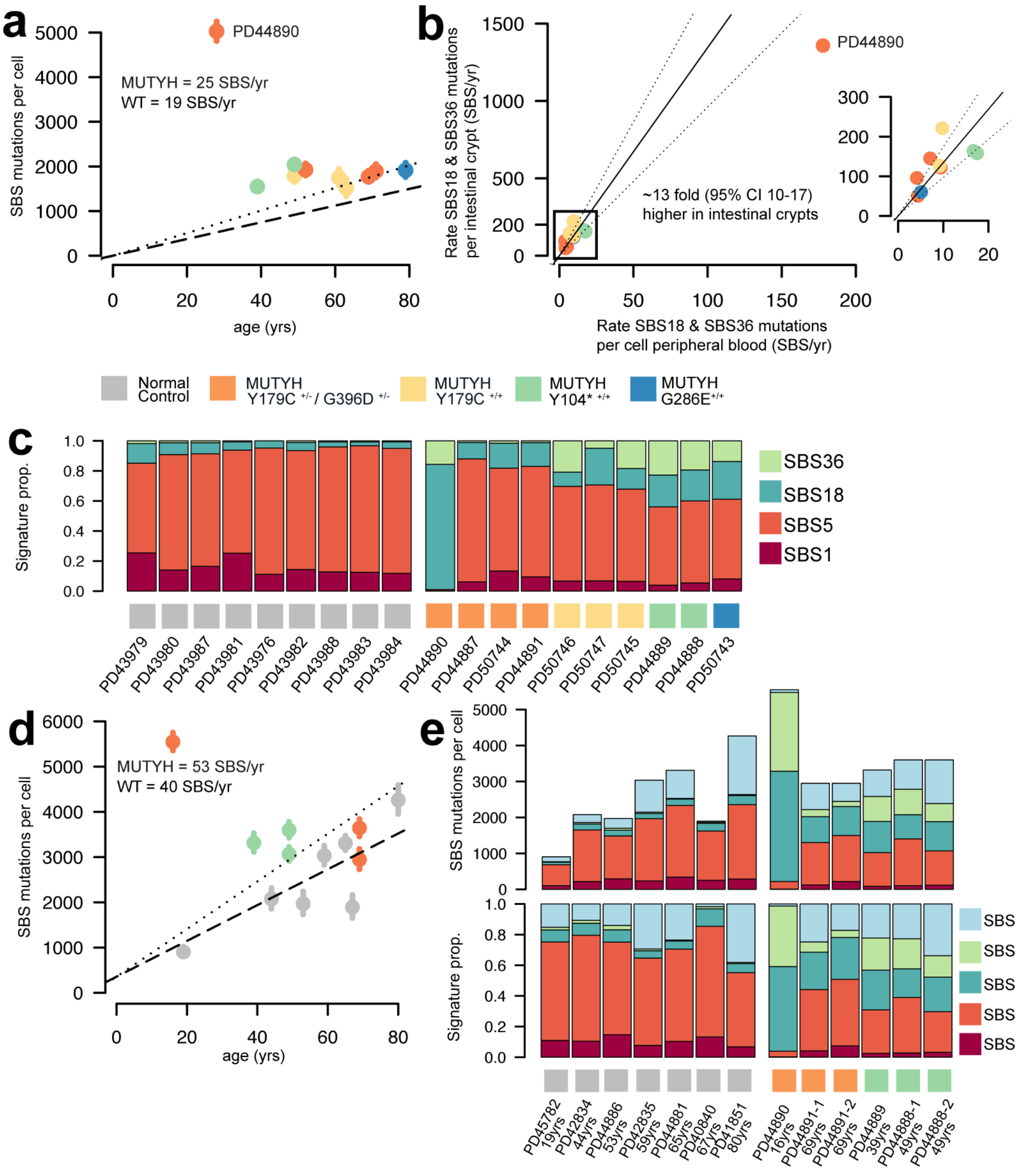

Figure 4.8 | Mutation burden and mutational signatures in blood and tissue lymphocytes from individuals with *MUTYH* mutations

(a) Somatic mutation burden (SBS mutations per cell) in peripheral blood samples (y-axis) plotted against the age of the individual (x-axis). Each point represents one sample and which are coloured according to the *MUTYH* germline mutation. Bars indicate the 95% confidence interval. Linear models fitted to the data are summarised by the regression lines (*MUTYH*, dotted line and wild-type, dashed line). Individual PD44890 who has additional germline *OGG1* mutations and is a clear outlier, was excluded from the modelling.

(b) Plot showing the rate of SBS18 and SBS36 mutations in peripheral blood (x-axis) and intestinal crypts (y-axis) (SBS/yr). Linear mixed-effects model is shown as the black line, dotted lines indicate the 95% confidence interval. Individual PD44890, who has additional germline *OGG1* mutations and is a clear outlier, was excluded from this analysis. Plot inset shows the mutation rate for the 9 individuals who did not carry germline *OGG1* variants.

(c) Stacked bar plots showing the proportion of mutational signature in peripheral blood samples for each individual. Coloured boxes below each bar indicate the germline *MUTYH* mutation.

(d) Plot showing the somatic mutation burden (SBS mutations per cell) in tissue lymphocytes (y-axis) plotted against the age of the individual (x-axis). Each point represents one sample and is coloured according to the *MUTYH* germline mutation. Bars indicate the 95% confidence interval. Linear models fitted to the data are displayed by regression lines, the *MUTYH* are shown by the dotted line and wild-type by the dashed line. Individual PD44890 who has additional germline *OGG1* mutations and is a clear outlier, was excluded from the analysis.

(e) Bar plots showing the proportion of mutational signature in tissue lymphocytes for each individual. Absolute mutation signature burden (above) and signature proportion (below). Coloured boxes below each bar indicate the germline *MUTYH* mutation.

The mutation burden of blood and lymphocytes from individual PD44890 was substantially higher than in other individuals in this cohort, which mirrors the stronger mutator phenotype observed in the intestinal epithelium from this individual (Figure 4.8b). The greater increases imply that the effect of the *MUTYH* mutations and additional modifying germline mutations are observed not only in colonic epithelium but also in blood, lymphocytes and potentially other tissues.

Mutational signatures SBS18 and SBS36 were responsible for the increased mutation burden in peripheral blood and tissue lymphocytes in individuals with *MUTYH* mutations (Figure 4.8c and Figure 4.8e). The proportion of mutations contributed by these signatures was, however, lower than in intestinal epithelium. A further mutational signature, SBS9, which is associated with the physiological maturation of B-cells[6,210], was observed in lymphocyte samples from both wild-type and MAP samples. The presence of SBS9 in these samples indicates that the tissue lymphocyte populations contained B-lymphocytes.

Section 4 summary:

1. Mutation rates are increased in peripheral blood and lymphocytes from individuals with *MUTYH* mutations.
2. Increases in peripheral blood and lymphocytes are smaller than in colonic epithelium.
3. Mutational signatures SBS18 and SBS36 are responsible for the additional mutation burdens observed.
4. Elevated mutation rates in peripheral blood broadly reflect the increased mutation rates observed in intestinal epithelium.

Chapter summary

This chapter describes the somatic mutation rates and mutational signatures in normal intestinal epithelium, blood and tissue lymphocyte cells from individuals with germline *MUTYH* mutations. Mutation rates were raised in all tissues studied suggesting that other, and possibly all, tissues from all individuals with *MUTYH* mutations have increased somatic mutation burdens. Of the tissues studied, the greatest increases were observed in the intestinal epithelium which is also the tissue type with the highest cancer risk in MAP[158,172]. Both mutation burden and mutational signatures were influenced by the different germline genotypes although some inter-individual heterogeneity was unexplained. Additional germline mutations in components of the BER pathway may act to further increase the mutation rate in normal cells from individuals with *MUTYH* mutations and may influence the mutational spectrum.

These data have potential implications for our understanding of cancer risk in MAP and may also contribute to our conception of the somatic mutation theory of ageing, both of which will be discussed in the final chapter of this thesis.

Lastly, the presence of an elevated mutation burden in peripheral blood mirrors the increases observed in colon. This would suggest that measurement of somatic mutation rates in blood samples could have a potential role in assessing the presence of a mutator phenotype and in the personalisation of disease surveillance. This too will be expanded upon in the final chapter.

Chapter 5 - Implications of a ubiquitously elevated mutation rate in normal tissues

Introduction

The data presented in this thesis demonstrate the ubiquitous presence of elevated somatic mutation burdens in normal tissues from individuals with PPAP and MAP. Every intestinal cell studied showed an increased mutation rate, implying that potentially all intestinal cells from all individuals have increased mutation burdens. Mutation burdens were further increased in early neoplasms. Finally, mutation burdens were increased in other tissues, including both those with a known elevated risk of cancer in these syndromes, and those with no known risk of cancer.

In this chapter, the following will be discussed:

1) Implications of an elevated mutation rate on cancer predisposition in PPAP and MAP
2) Comparison of observations in PPAP and MAP
3) Comparison of observations in PPAP and MAP with other known polyposis syndromes
4) Implications for our understanding of the somatic mutation theory of ageing

Implications of an elevated mutation rate on cancer predisposition in PPAP and MAP

Elevated mutation rates were observed in normal tissues from individuals with PPAP and MAP (Table 5.1). Several observations suggest that the increased mutation burden observed in normal tissues in these individuals may be responsible for the observed predisposition to cancer: 1) mutation rates were most increased in the colonic epithelium, the tissue with the highest risk of cancer, 2) the increased rate in normal intestinal epithelium was further elevated in the process of neoplastic transformation as evidenced by the increased mutation burdens in adenomas, 3) increased genome-wide mutation rates were associated with increased exonic mutation rates and the generation of greater numbers of potential cancer driver mutations and 4) driver mutations in normal tissues and adenomas showed the characteristic genome-wide mutational spectra associated with the *POLE*/*POLD1* and MAP mutational signatures. Lastly, relatively higher mutation rates in normal tissues are seen in individuals with specific germline *POLE*/*POLD1* and *MUTYH* genotypes that are associated with a higher polyp burden and cancer risk[158,185]. Taken together, these observations suggest that elevated mutation burdens in normal intestinal stem cells generate higher burdens of cancer-causing driver mutations which predisposes to the development of neoplasia in individuals with PPAP and MAP.

However, elevated mutation rates do not correspond to an increased risk of cancer in all tissues in PPAP and MAP. For example, mutation burdens were increased in peripheral blood yet there has been no reported increase in the incidence of haematological cancers in these two syndromes. Therefore, an increased somatic SBS / ID mutation rate does not appear to confer an overtly increased cancer risk in all tissues. The causes that underlie the difference in cancer incidence across tissues from individuals with these predisposition syndromes are unclear. Several factors could potentially explain some of the differences in cancer risk that are observed: 1) mutation rate increases in most tissues are relatively modest when compared to those in the most cancer-prone tissues in PPAP and MAP, 2) the C>A mutational processes in PPAP and MAP may not generate the mutational spectrum of the requisite set of cancer driver mutations in all tissue types i.e. many driver mutations are caused by C>T mutations, 3) other mutation types that are not increased in PPAP and MAP - such as copy number changes and structural rearrangements - may be required for the generation of cancer in some tissues, 4) a smaller overall number of stem cells in some tissues may mean that an increased mutation rate does not generate an observable increase in cancer driver mutations over the course of a life time, and 5) the different clonal structure in certain tissues may make them less susceptible to the effects of an increased mutation rate.

	PPAP	MAP
Colon	Increased	Increased
Small Intestine	Increased	Increased
Blood	Increased	Increased
Tissue B-lymphocytes	Not assessed	Increased
Endometrium	Increased	Not assessed
Skin	Increased	Not assessed
Sperm	Increased	Not assessed
Artery	Increased	Not assessed
Muscle	Increased	Not assessed
Cerebral Cortex	Increased	Not assessed

Table 5.1 | Summary of somatic mutation rate changes in normal tissues studied from individuals with PPAP and MAP

Comparison of mutation rates in PPAP and MAP

Differences were observed between the mutation rates in PPAP and MAP (Table 5.2). SBS mutation rates were higher in *POLE*^{L424V} than in the *MUTYH* genotypes. However, SBS mutation rates were broadly comparable between *POLD1*^{S478N} and *MUTYH*. The most striking difference, however, was observed in the ID mutation rates. All individuals with PPAP had an increased ID mutation rate whereas those with MAP had normal or near-normal ID mutation rates.

Coding exon SBS mutation burdens were also increased in both PPAP and MAP (Figure 5.1). However, despite large differences in the genome-wide mutation rate, the coding exon SBS mutation rates were comparable between the two conditions, albeit, with some variation according to the specific germline mutation (Figure 5.1). Coding exon ID mutation rates were elevated in PPAP but not in MAP.

	SBS Rates	ID rates	Other mutation types
POLE$^{\text{L424V}}$	~7-fold	~13-fold (range 6.5-13)	Unaffected
POLD1$^{\text{S478N}}$	~3-fold	~44-fold (range 22-88)	Unaffected
POLD1$^{\text{D316N}}$	~1.2-fold	~12-fold (range 6-24)	Unaffected
POLD1$^{\text{L474P}}$	~1.2-fold	~12-fold (range 6-24)	Unaffected
MUTYH$^{\text{Y179C+/- G396D+/-}}$	~2-fold	Unaffected	Unaffected
MUTYH$^{\text{Y179C+/- G396D+/- OGG1 +/+}}$	~31-fold	Unaffected	Unaffected
MUTYH$^{\text{Y179C+/+}}$	~4-fold	Unaffected	Unaffected
MUTYH$^{\text{Y104*+/+}}$	~4-fold	Unaffected	Unaffected
MUTYH$^{\text{G286E+/+}}$	~3-fold	Unaffected	Unaffected

Table 5.2 | Summary of genome-wide mutation rate changes in intestinal cells from individuals with *POLE*/*POLD1* and *MUTYH* mutations organised by germline genotype

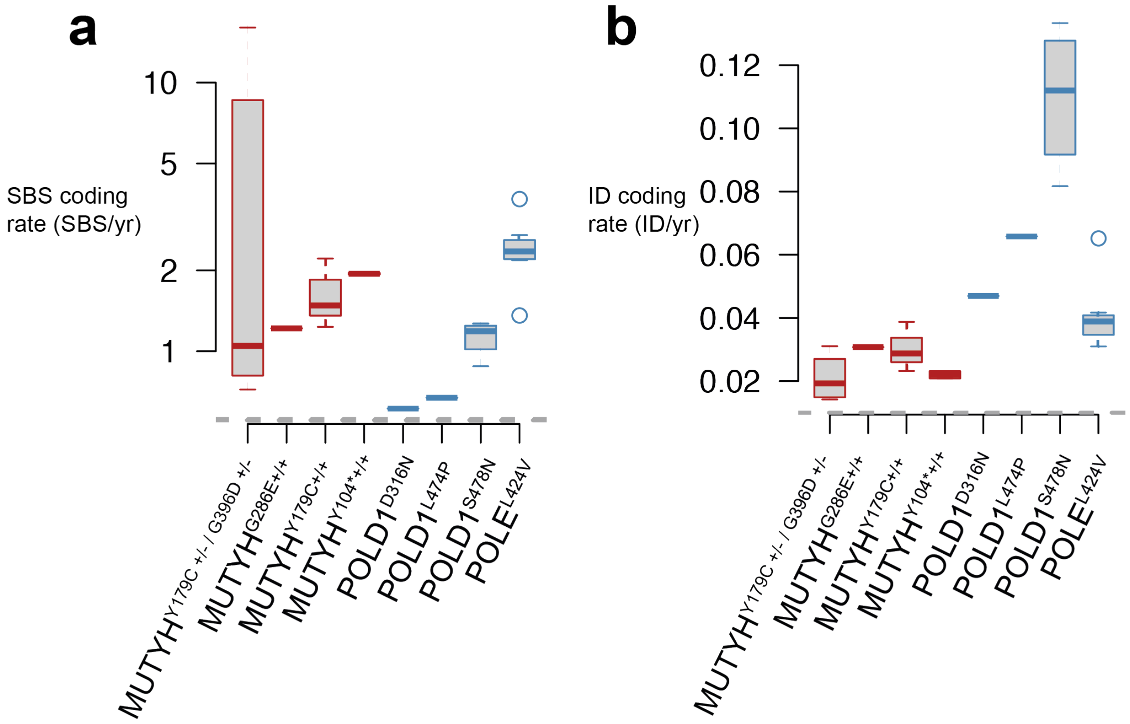

Figure 5.1 | Mutation rate in protein-coding exons

Somatic mutation rate of (a) SBS mutations and (b) ID mutations in protein-coding exons. Data is grouped according to the germline genotype (x-axis). Mutation rate per year per intestinal crypt is displayed on the y-axis. For the box and whisker plots, the central line represents the median, the box represents the inter-quartile range (IQR) from 1st to 3rd quartiles and the whiskers extend to the farthest point that is within 1.5x IQR . The large inter-quartile range in the *MUTYH*$^{Y179C+/- \, G396D +/-}$ grouping is due to the outlier, PD44890, who had a substantially higher mutation rate than the other 9 individuals with MAP.

Correlation of mutation rates and clinical severity in PPAP and MAP

An increased somatic mutation rate appears to explain the general observation of an increased cancer risk in individuals with PPAP and MAP. There are some differences in the clinical phenotype of these syndromes and further differences between the various germline genotypes. Whether the somatic coding exon mutation rate in normal tissue explains differences between the different genotypes, is presently unclear.

The lifetime incidence of colorectal cancer for individuals with the most severe PPAP (*POLE*^{L424V}) and MAP genotypes (*MUTYH*$^{Y179C+/+}$ and *MUTYH*$^{TRUNC+/+}$) is close to 100%[158,170,185]. There are, however, subtle differences in the age of onset of colorectal cancer and the intestinal polyp burdens. Individuals with PPAP develop colorectal cancer at a younger age than individuals with MAP, yet, have fewer polyps (median 12) than those with MAP (median 30-50). The exonic SBS mutation rates in PPAP and MAP are broadly similar hence do not provide an obvious explanation for these clinical observations.

An appreciable difference in the ID mutation rates was however observed between PPAP and MAP. The largest relative increases in the exonic ID mutation rate were seen in *POLD1*^{S478N} carriers (~11-fold) with smaller increases in *POLE*^{L424V} carriers (~4-fold). However, no substantial increase was observed in individuals with the various *MUTYH* genotypes. It is therefore possible that the elevated ID mutation rates in PPAP partially contribute to the earlier age of colorectal cancer onset seen in PPAP vs MAP. However, since the absolute increase in the ID mutation burden is modest, even in *POLD1*^{S478N}, when compared to the increase in the SBS mutation rate in *POLE* and *MUTYH* genotypes, this may mitigate the effect of large fold increases in the ID mutation rate.

Individuals with PPAP have substantially lower polyp burdens than those with MAP despite showing very similar SBS coding mutation burdens and higher ID mutation burdens. It is noteworthy that there is a reported genotype-phenotype correlation in MAP and that this concurs with an increased somatic mutation rate in experimental systems. In MAP, the *MUTYH*$^{Y179C+/+}$ and *MUTYH* truncating alleles are associated with the highest polyp burdens[158] and highest mutation rates *in vitro*[149]. In support of this observation, individuals with the

MUTYH$^{Y179C+/+}$ genotype or *MUTYH* truncating alleles had the highest mutation rate of all *MUTYH* genotypes in this study indicating a potential association between the mutation rate and polyp burden.

In the cohorts studied in this thesis, there was no evidence to suggest that inter-individual differences in the polyp burden were explained by variation in the somatic mutation rate. The only potentially exception was the 16 year old individual with biallelic *MUTYH* and *OGG1* mutations, who had the highest SBS somatic mutation rate of any individual we have studied. This individual demonstrated early onset intestinal polyposis with a relatively large number of polyps (~100) and developed multiple neoplasms. It appears plausible that the increased mutation rate conferred by the combination of these specific alleles, may have contributed to an earlier development of cancer and higher polyp burdens. However, a clear correlation between the somatic mutation rate and polyp burdens was not seen in other individuals in the PPAP and MAP cohorts, thus suggesting the presence of other, currently cryptic factors in modulating the effect of a raised mutation rate on the generation of polyps.

In summary, the observation of increased somatic mutation rates in normal tissues appears to explain the general observation of an increased burden of cancer in individuals with PPAP and MAP. The relative differences between PPAP and MAP, and the finer distinctions in phenotypes that accompany the various germline genotypes, were not clearly explained by the results of this study. These observations suggest that factors, other than the mutation rate, may influence the heterogeneity in polyp burdens observed in these cancer predisposition syndromes.

Comparison of observations in PPAP and MAP with other known intestinal cancer predisposition syndromes

In this section, the findings presented in this thesis will be placed in the broader context of other known intestinal cancer predisposition syndromes. Comparison will be made in two main domains:

1. Somatic mutation rates, intestinal polyposis and cancer risk
2. Spectrum of tissues affected by extra-intestinal cancers

Comprehensive assessment of the genome-wide somatic mutation rate in normal tissues has, until recently, been unfeasible. Therefore, it is not generally known if the somatic mutation rate is altered in normal tissues from individuals with other intestinal cancer predisposition syndromes. Increased somatic mutation rates in normal tissues are suspected in some syndromes based on the functions of the affected genes. Using the insights generated from the experiments in this thesis, the presence of an increased somatic mutation rate in normal tissues in PPAP and MAP was confirmed. By contrast, in a parallel study of normal tissues from individuals with Lynch syndrome, mutation rates were found to be comparable to normal epithelium from wild-type individuals[286].

Polyposis is a feature of many intestinal predisposition syndromes. In PPAP and MAP, it appears likely that the increased polyp risk is largely driven by a ubiquitously increased somatic mutation rate in normal epithelium. Similarly, in NTHL1 and MBD4 deficiency, which were not studied in this thesis, the increased polyp burdens are attributed to elevated mutation rates in normal intestinal epithelial tissue[110,194]. By contrast, in Lynch syndrome and FAP, increased somatic mutation rates are not thought to be universally present in normal tissues, rather, they are increased at a later stage in the process of neoplastic development[58,117,286]. In Lynch syndrome, increased mutation rates and microsatellite instability are present in a substantial proportion of intestinal adenomas but not in normal epithelium, suggesting that increased mutation rates occur during the process of neoplastic transformation[117,286]. In FAP, the mutation rate in normal epithelium and adenomas is not thought to be higher than in comparable samples from wild-type individuals[58]. Therefore, whilst increased somatic mutation burdens in normal intestinal epithelium may contribute to the elevated risk of polyposis in several predisposition syndromes, other mechanisms are likely to be operative in syndromes such as FAP and Lynch syndrome.

Mutation rate	PPAP	MAP	Lynch	FAP
Normal tissues	Increased	Increased	Unaffected	Unclear
Adenomas	Increased	Increased	Increased*	Unaffected**
Carcinoma	Increased	Increased	Increased	Unaffected**

Table 5.3 | Summary of mutation rate changes in normal, adenoma and carcinoma tissues in PPAP, MAP, Lynch syndrome and FAP

* Increased mutation rates occur in ~70-80% of adenomas – other adenomas are comparable to those from wild-type individuals

** Mutation rates in FAP adenomas and carcinomas are thought to be comparable to sporadic neoplasms from wild-type non-predisposed individuals

Based on our current knowledge of the genetics and molecular mechanisms of intestinal cancer predisposition and the insights generated in this study, the intestinal cancer predisposition syndromes can be grouped into four categories (Table 5.4):

1. Increased somatic mutation burdens in normal tissues due to inherited biallelic inactivation of a DNA repair protein e.g. *MUTYH, NTHL1* and *MMR* genes in constitutive mismatch repair deficiency
2. Increased somatic mutations burden in normal tissues due to an inherited heterozygous dominantly acting allele e.g. *POLE*/*POLD1*
3. Inheritance of a heterozygous inactivating mutation in a recessive DNA repair gene, which, upon loss of the second allele in a small subset of cells, generates increased somatic mutation burdens e.g. *MMR* genes in Lynch syndrome
4. Inheritance of a heterozygous inactivating mutation in genes responsible for stem cell homeostasis but not directly involved in DNA repair e.g. *APC* in FAP and *SMAD4*/*BMPR1A* and Juvenile Polyposis Syndrome. Loss of the second allele results in altered stem cell homeostasis and subsequently the generation of neoplasia. Increased mutation rates are not expected in normal tissues from these individuals.

		Cancer cells	
		Comparable to wild-type individuals	Increased
Normal cells	Comparable to wild-type individuals	Category 4 FAP/AFAP Juvenile Polyposis Syndrome	Category 3 Lynch Syndrome
	Increased		Category 1 MAP NTHL1 MBD4 Constitutive mismatch repair deficiency (CMMRD) Category 2 PPAP

Table 5.4 | Summary of proposed classification of intestinal cancer predisposition syndromes based on somatic mutation rates in normal and neoplastic tissues

Comparison with other intestinal polyposis syndromes – tissue distribution

PPAP and MAP are primarily associated with colorectal neoplasia; however, both syndromes also have an appreciable risk of cancer in other tissues (Table 5.5). After the colon, the next most commonly affected site in the intestinal tract is the duodenum. Duodenal polyposis affects a substantially smaller number of individuals than colorectal cancer in PPAP, MAP and other intestinal polyposis syndromes. Duodenal cancer affects up to 10% of individuals with PPAP[185], 1% of individuals with MAP[174], up to 10% of individuals with FAP[287] and up to 5% with Lynch syndrome[288]. By contrast, primary sporadic small bowel cancer affects less than 0.05% of the population[289]. The data presented in this thesis do not suggest an obvious reason for the large difference in cancer incidence in small and large bowel cancer in predisposed or healthy individuals. However, they show the presence of an elevated mutation rate in small bowel and would suggest that, much like the colon, this is a major cause for the relative increase in small bowel cancer risk in PPAP and MAP compared to healthy wild-type individuals.

Endometrial cancer is observed in PPAP, Lynch and NTHL1 but is not observed in MAP, FAP and MBD4 (Table 5.5). In normal endometrial epithelium with a *POLE* mutation, a substantially increased somatic mutation burden was observed. Increased mutation burdens might also reasonably be expected in other individuals with germline mutations in dominantly acting DNA repair genes. Indeed, in individuals with *NTHL1* germline mutations, endometrial cancers had an increased mutation burden compared to sporadic endometrial cancers from wild-type individuals, potentially implying that normal endometrial glands in affected individuals also bear an increased mutation rate[194]. Increased mutation rates may therefore be responsible for the increased cancer risk in endometrium in PPAP and possibly in *NTHL1* mutation carriers. However, in MAP, where increased mutation burdens were observed in all tissues and it would be reasonable to expect an increase in mutation rate in endometrium, no increase in endometrial cancer risk is observed[172]. Similarly, in MBD4 deficiency, where increased mutation burdens are seen in adenomas and blood neoplasms, increased mutation burdens in normal tissues including the endometrium might reasonably be expected and yet, no phenotypic evidence of endometrial cancer is observed, albeit the clinical data regarding the phenotype of individuals with MBD4-deficiency is limited[185].

	POLE	POLD1	MAP	FAP	Lynch	NTHL1	MBD4
Large bowel polyposis	Y	Y	Y	Y	N	Y	Y
Small bowel polyposis	Y	N	Y	Y	N	Y	N
Early onset colorectal cancer	Y	Y	Y	Y	Y	Y	Y
Early onset duodenal cancer	Y	N	Y	Y	Y	?	N
Endometrial cancer	Y	Y	N	N	Y	Y	N
Ovarian cancer	Y	N	N	N	Y	N	N
Breast cancer	Y	Y	N*	N	N*	Y	N
CNS cancer	Y	Y	N	N	Y	Y	N
Myeloid neoplasms	N	N	N	N	N	N	Y

Table 5.5 | Summary of tissue-specific cancer risk in a subset of polyposis syndromes

*Inconclusive evidence at present

Haematological malignancy in inherited cancer predisposition syndromes

Haematological malignancies are not a reported feature of either PPAP or MAP[172,184,185]. This is somewhat surprising, as increased somatic mutation rates were observed in blood samples from all individuals surveyed. Therefore, increased somatic mutation rates in blood from individuals with PPAP and MAP do not appear to confer an increased risk of haematological malignancy. Moreover, haematological malignancies are rarely seen in other intestinal polyposis syndromes. One exception, however, is the inherited germline condition associated with loss of function in the DNA glycosylase, MBD4. MBD4-deficiency causes ARCH, AML and is associated with colorectal polyposis and an increased risk of colorectal cancer[110]. Evidence for an increased mutation burden in normal tissues in MBD4-deficient individuals is suggested by the age-related increase in the mutation burden in intestinal adenomas and is supported by the observation of ~33-fold increases in mutation burden in MBD4 AML – compared to MBD4 wild-type AML - and at least 9-fold increase in MBD4-adenomas – compared to sporadic adenomas. It is not known why haematological malignancies occur in MBD4-deficiency and are absent from other inherited cancer predisposition syndromes associated with DNA repair deficiencies.

The reason for the absence of blood neoplasms in PPAP and MAP, and perhaps by extension other similar syndromes, is not entirely clear. However, several factors may potentially contribute to this observation, 1) the relatively modest elevation in the mutation rate of blood in PPAP and MAP may be insufficient to effect an increased cancer risk, 2) positive selection of mutations and subsequent clonal expansion in blood take many decades[207], thus the accrual of mutations at increased rate may be insufficient to confer an increased cancer risk during the individuals' lifetime, 3) the haematopoietic stem cell pool is estimated to comprise ~ 20,000-200,000 cells[207,211]. By contrast, the colon is populated with 15,000,000 intestinal crypts, each with at least one actively dividing stem cell[290]. Therefore, if the absolute number of stem cells in a tissue contributes to the cancer risk, this would be higher in colon than blood, 4) the mutational processes of defective POLE/POLD1 and MUTYH generate a specific, predominantly C>A, mutational spectra. By contrast, MBD4-deficiency generates a C>T spectrum at CpG dinucleotides and the majority of ARCH driver mutations are C>T mutations with many occurring at CpG islands[196]. Therefore, the accumulation of an elevated mutation burden of a specific mutational signature may not necessarily generate the spectrum of mutations required to generate blood neoplasia. For one or more of the above reasons, the MBD4-deficient cases may be an exception to the general rule that hypermutation in normal tissues does not confer an appreciable increase in haematological neoplasia.

Finally, it should be noted that several of the syndromes mentioned in this discussion are rare, which presents certain challenges for identifying and quantifying cancer risk. This is particularly the case when assessing the relative risk of extracolonic malignancies which tend to affect only a subset of individuals with each condition. Further phenotypic characterisation of larger groups of affected individuals may refine the known associations and potentially identify rare phenotypes that are not presently apparent.

Estimation of intestinal cancer risk using the somatic mutation rate in blood

In this thesis, the mutation rate in peripheral blood from individuals with PPAP and MAP was investigated using a modified highly error-corrected duplex sequencing method called NanoSeq[201]. Using this method, elevated mutation burdens and distinctive mutational signatures were identified in the peripheral blood of all individuals investigated. The principle aim of these experiments was to investigate whether the increased mutation burdens and associated mutational signatures observed in the intestinal epithelium were also present in blood and what this might tell us about the relative risk of cancer and the process of ageing in haematopoietic cells. However, an unanticipated additional insight from these experiments was that the mutation rate in peripheral blood is increased compared to wild-type individuals and the characteristic mutational signatures seen in intestinal cells are also observed in peripheral blood. Furthermore, there was a correlation between the mutation burdens observed in colon and those in blood sampled from same individuals at the same time, suggesting that the measurement of mutation rates and mutational signatures in blood may predict the mutation rate in the intestinal epithelium.

In the MAP cohort, the differences in the mutation rates and signatures observed in the various germline *MUTYH* mutations in colon, were also detectable in peripheral blood. Furthermore, the inter-individual heterogeneity present in individuals with the same germline *MUTYH* mutation, was also observable in the peripheral blood samples. Measurement of the mutation rate from peripheral blood may, therefore, give a fine grained read-out of the strength of an individual's mutator phenotype and potentially be used to estimate an individual's cancer risk and thus, personalise the surveillance strategy. This application of NanoSeq could potentially serve as a non-invasive method of risk stratification in individuals with MAP and other conditions associated with increased somatic mutation rates due to germline DNA repair gene mutations e.g. *POLE*, *POLD1* and *NTHL1*.

Finally, if single-molecule based methods can be adapted to identify other mutation types e.g. CNVs and structural rearrangements, measurement of mutation burdens in peripheral blood could also be applied to inherited syndromes characterised by the accumulation of these mutation types.

Implications for our understanding of the somatic mutation theory of ageing

The somatic mutation theory of ageing (SMT) proposes that the accumulation of mutations over the course of a lifetime in healthy cells and tissues is responsible for the phenotypic features associated with ageing[232,233,235,237,238,291]. In recent years, technological advances in molecular biology and DNA sequencing technologies have made it possible to interrogate the genome of normal cells and investigate a potential link between the accumulation of mutations and ageing in a way that was not previously possible. Using these technologies, studies investigating a broad range of tissues have shown that somatic mutations are continuously acquired in normal tissues with age[46,47,49,50,82,201,207,211]. These findings support the central assertion of the SMT, that mutation burdens increase with age in healthy human tissues paralleling the development of 1) common age-related diseases e.g. Type 2 diabetes, hypertension and dementia and 2) visible signs of ageing e.g. loss or greying of hair and alterations in skin. Furthermore, these studies observe that the somatic mutation rate is remarkably consistent across different individuals, implying that higher mutation rates may not be tolerated.

In this study of normal tissues from individuals with PPAP and MAP, increased mutation burdens were identified in all tissues from all individuals. Mutation rates were many fold higher than has been previously observed in non-predisposed healthy individuals. To aid the interpretation of the increased mutation burdens observed, “mutational ages” were calculated. The mutational age is the number of years a cell would need to live in order for it to acquire the number of mutations observed based on the rate of somatic mutation accumulation in wild-type healthy individuals. In the PPAP study, SBS mutational ages of up to 600 years old were seen in individuals with the $\textit{POLE}^{\text{L424V}}$ mutation and ID mutational ages of over 3000 years were observed in an individual with the $\textit{POLD1}^{\text{S478N}}$ mutation (Figure 5.2). In the MAP cohort, most SBS mutational ages were increased, albeit more modestly, ranging from ~80-200 years. However, in the 16 year old with *MUTYH* and *OGG1* mutations, the ~31 fold increases in SBS mutation rate in intestinal epithelium translates to a mutational age of ~500 years.

Despite the substantial increases in mutation rates and the ‘mutational ages’, individuals with PPAP and MAP do not show evidence of early-onset or accelerated diseases or phenotypes of ageing[172,184,185]. These observations would suggest therefore, that the increased mutation burdens in normal tissues from these individuals are, aside from their influence on cancer risk, largely tolerated. Therefore the data presented in this thesis indicates that normal cells and tissues can tolerate higher burdens of somatic mutations than was previously thought and increased mutation burdens in normal cells and tissues do not appear to contribute to the normal physiological process of ageing.

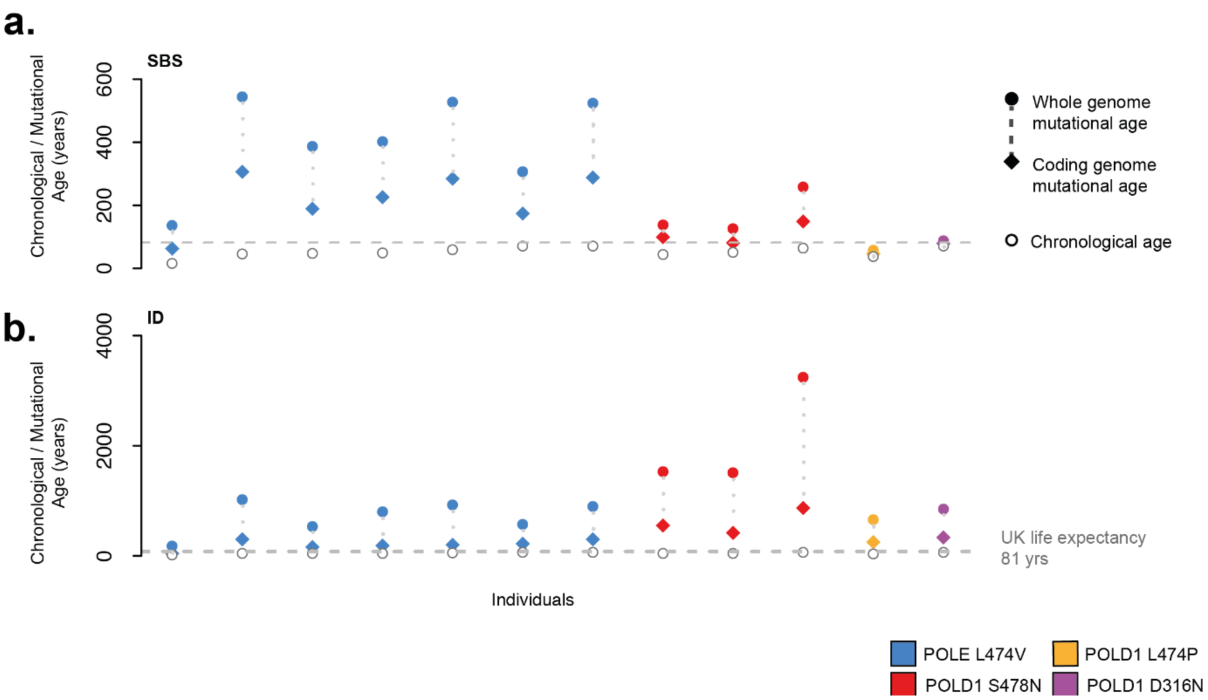

Figure 5.2 | Mutational ages of individuals with DNA polymerase mutations

Mutational ages of individuals with DNA polymerase mutations based on data from intestinal crypts. Whole-genome ages and coding genome ages are represented by the circle and diamond shapes respectively. Points are coloured by the germline mutation. Chronological ages are displayed as unfilled grey circles and the grey dashed line represents the median UK life expectancy. Panel (a) shows mutational ages due to SBS mutations and (b) ID mutations.

Several potential caveats may, however, temper this conclusion:

1. Tissue type

In this thesis, mutation rates were assessed in large and small intestinal epithelium, endometrium, blood, tissue lymphocytes, sperm, skin, artery, cerebral cortex, smooth and skeletal muscle. Whilst this represents a sizeable number of different cell and tissue types, many cell types from other tissues were not available to be assessed. If mutation burdens are relatively unaffected in the tissues that are responsible for ageing and age-related diseases, this may potentially explain the absence of ageing phenotype observed.

2. Mutation type

Increased burdens of small mutation types – SBS & ID – were observed in normal tissues. Burdens of larger structural rearrangements were, however, not increased. If large scale rearrangements contribute to the process of physiological ageing in healthy individuals, then this would not have been observed at an accelerated rate in individuals with PPAP and MAP. There is a potential rationale for structural rearrangements playing a role in ageing. Impaired genomic integrity is observed in several of the known progeroid syndromes including: Werner syndrome, Cockayne syndrome, Blooms syndrome, Ataxia-telangiectasia, Xeroderma Pigmentosum and Rothmund-Thomson Syndrome[292,293]. Whether the mechanism of ageing in these syndromes is representative of the physiological process of ageing that occurs in normal cells from healthy individuals is, however, presently unknown.

3. Genomic distribution

If the accumulation of somatic SBS and ID mutations contributes to the process of ageing, it is likely that an effect would be mediated through the accumulation of mutations in protein-coding exons. Thus, mutational processes that spare the accumulation of mutations on these parts of the genome may lessen the potential impact of a high genome-wide mutation rate. The substantially elevated genome-wide mutation rates observed in PPAP disproportionately affect intronic, intergenic and non-coding parts of the genome, thus relatively sparing protein-coding exons. Exon-sparing was, however, modest in the MAP cohort. Furthermore, in one individual, a 16 year old with additional *OGG1* mutations, ~30x-fold higher coding exon mutation rates were observed compared to wild-type controls. Despite the substantially elevated coding mutation rate, this individual did not display evidence of premature ageing.

4. Cellular damage

In this thesis, it is reported that elevated mutation burdens are not associated with overt evidence of premature ageing in PPAP and MAP. Indeed, despite 2x-fold or greater increases in the somatic mutation rate in most individuals, diseases associated with ageing such as heart disease, cognitive impairment and type 2 diabetes, are not observed more frequently or at an earlier age of onset than in the general population. However, the absence of overt organismal signs and features of ageing does not necessarily imply the absence of cellular ageing. It is plausible that elevated somatic mutation burdens could impact some cellular processes that do not manifest as an ageing phenotype. The observation of cellular alterations would not affect the conclusions regarding organismal ageing and the SMT, but could offer potential insight into the consequences of somatic mutation accumulation on the function of cells from different tissues.

Whilst the data presented in this thesis suggest that somatic mutations are unlikely to be the sole cause of ageing, it is possible that even higher somatic mutation burdens could alter cellular physiology and confer a premature ageing phenotype. The accumulation of 1,000s or 10,000s of additional SBS and ID mutations may have a relatively modest impact in the context of the genome which spans ~3.2 billion bases and ~22,000 genes. It is possible that mutation burden increases that are a magnitude higher i.e. ~100-fold or even ~1000-fold, may effect an ageing phenotype. The principle assertion of this thesis, that the accumulation somatic mutations is not responsible for physiological ageing in normal tissues from healthy individuals, would nevertheless stand.

Lastly, the mechanism by which somatic mutation accrual may lead to ageing is not well defined. If ageing is a direct consequence of protein damage caused by somatic mutations then such phenotypes may be observed earlier than if ageing depends on downstream processes such as pervasive clonal expansions which may take decades to emerge hence only presenting in later life.

Conclusion

In this thesis, investigation of individuals with germline *POLE*/*POLD1* and *MUTYH* mutations revealed the presence of ubiquitously elevated somatic mutation rates in normal healthy tissues. Differences in mutation rates and mutational signatures were observed between individuals reflecting the influence of germline mutations on the spectrum of somatic mutagenesis in normal tissues. These studies show that normal cells from affected individuals have different mutation rate increases, which may explain the presence of variable cancer risk. While increased mutation rates were observed in normal extra-intestinal tissues, it is presently unclear how this shapes the spectrum of extra-intestinal malignancies in individuals with cancer predisposition syndromes. Lastly, the observation that somatic mutations are tolerated by normal cells may inform the debate regarding the somatic mutation theory of ageing.

Future prospects

There are multiple themes emerging from the work outlined in this thesis that are deserving of further investigation. To outline three of them:

1. Understanding the factors responsible for the tissue or cell-type specific cancer risk in individuals with elevated somatic mutation rates

The results presented in this Thesis identify increased mutation rates across all cell and tissue types examined from individuals with PPAP and MAP. The degree of increase in the mutation rate does not, however, correlate directly with the cancer risk in all tissues. Understanding the general principles that govern the relationship between an increased mutation rate and cancer risk in different cell types may help improve our understanding of cancer predisposition and potentially uncover the reasons for differences in cancer prevalence across different tissues both in predisposed and non-predisposed individuals. To investigate this it would be of value to expand the number of individuals and tissues surveyed in individuals with MAP and PPAP. Then, to investigate a similar repertoire of tissue and cell types from individuals with other known cancer predispositions associated with DNA repair e.g. MBD4 deficiency and NTHL1 deficiency. Such an assessment would describe both the mutation rates and mutational signatures present across different diseases and their impact on tissue-specific cancer risk. Modelling of these relationships - taking into account other potentially relevant factors such as tissue structure, stem cell numbers and driver mutation spectra of cancers arising in each tissue - may contribute to our understanding of the ways in which somatic mutagenesis confers cancer risk.

2. Utilisation of somatic mutation rates to risk stratify individuals with inherited cancer predisposition associated with defective DNA repair

In this Thesis mutation burdens and mutational signatures in peripheral blood mirrored those observed in colonic epithelium. Differences were observed between the various germline mutations and between individuals with the same germline mutation. These observations suggest that measurement of mutagenesis in peripheral blood may be a useful method for individualised prediction of cancer risk. Extrapolation of these data using a larger cohort of individuals with MAP paired with in-depth clinical phenotyping would allow the potential viability of this as a translational biomarker to be assessed. Since higher mutation rates are generally associated with a greater risk of colorectal cancer in the MAP cohort, it is possible that measurement of the somatic mutation rate could be used to risk stratify individuals and hence guide surveillance or, in the future, timing of administration of prophylactic or preventative drug therapies.

A substantial number of individuals with multiple polyps or early onset or metachronous colorectal cancers have no identifiable predisposing germline mutation. Screening of the mutation rate and mutational spectrum in blood may identify individuals with an increased mutation rate and or an altered mutational spectrum which could be used to ascertain if a previously undescribed constitutive DNA repair defect is present. Whole exome or genome sequencing could then be used to identify putative germline mutations in DNA repair genes for further investigation.

3. Investigation of the contribution of somatic mutations to the process of ageing in normal healthy tissues

Substantial advances have been made in the ability to characterise somatic mutagenesis across different normal cell and tissue types in recent years. The findings of these studies are largely consistent with and in support of the SMT. However, the findings from this study appear to be incompatible with the current conception of the SMT. It is however possible that an elevated burden of somatic mutations could cause alterations on a cellular level which are not overtly manifest in the human phenotype. Therefore, it would be valuable to undertake work to investigate the impact of high mutation burdens on a cellular level. This could be undertaken using tissues from individuals with known germline hypermutators or by modelling high mutation rates in model systems. Assessment of the impact of mutation rates on the function of multiple cell types could help to better understand how an increased mutation rate may impact cellular physiology and thus ageing. Furthermore, if the accrual of somatic mutations does not contribute to ageing in healthy cells and tissues, this type of experiment would provide robust support for this important negative observation.

Publications arising from work presented in this thesis

Robinson, P.S., Coorens, T.H.H., Palles, C. *et al.* Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet* 53, 1434–1442 doi:10.1038/s41588-021-00930-y (2021)

Robinson, P.S. et al. Inherited *MUTYH* mutations cause elevated somatic mutation rates and distinctive mutational signatures in normal human cells. *bioRxiv*, doi:10.1101/2021.10.20.465093 (2021)

References

1 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724, doi:10.1038/nature07943 (2009).

2 ICGC TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93, doi:10.1038/s41586-020-1969-6 (2020).

3 Forbes, S. A. *et al.* COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research* **45**, D777-D783, doi:10.1093/nar/gkw1121 (2017).

4 Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

5 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

6 Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101, doi:10.1038/s41586-020-1943-3 (2020).

7 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

8 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

9 Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729, doi:10.1038/ng.128 (2008).

10 Ley, T. J. *et al.* DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72, doi:10.1038/nature07485 (2008).

11 Kim, H. *et al.* Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* **52**, 891-897, doi:10.1038/s41588-020-0678-2 (2020).

12 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).

13 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

14 Knudson, A. G., Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).

15 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705, doi:10.1038/s41568-018-0060-1 (2018).

16 Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

17 Inoue, K. & Fry, E. A. Haploinsufficient tumor suppressor genes. *Adv Med Biol* **118**, 83-122 (2017).

18 Lobry, C., Oh, P. & Aifantis, I. Oncogenic and tumor suppressor functions of Notch in cancer: it's NOTCH what you think. *J Exp Med* **208**, 1931-1935, doi:10.1084/jem.20111855 (2011).

19 Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *British Journal of Cancer* **8**, 1-12, doi:10.1038/bjc.1954.1 (1954).

20 Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal

cancers. *Proceedings of the National Academy of Sciences* **112**, 118-123, doi:10.1073/pnas.1421839112 (2015).

21 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).

22 Martinez-Jimenez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* **20**, 555-572, doi:10.1038/s41568-020-0290-x (2020).

23 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e318, doi:10.1016/j.cell.2018.02.060 (2018).

24 Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81-84, doi:10.1038/nature14173 (2015).

25 Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-364, doi:10.1038/nature14221 (2015).

26 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

27 Schneider, G., Schmidt-Supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat Rev Cancer* **17**, 239-253, doi:10.1038/nrc.2017.5 (2017).

28 Koren, S. *et al.* PIK3CA(H1047R) induces multipotency and multi-lineage mammary tumours. *Nature* **525**, 114-118, doi:10.1038/nature14669 (2015).

29 Xu, B. *et al.* The significance of dynamin 2 expression for prostate cancer progression, prognostication, and therapeutic targeting. *Cancer Med* **3**, 14-24, doi:10.1002/cam4.168 (2014).

30 Trochet, D. & Bitoun, M. A review of Dynamin 2 involvement in cancers highlights a promising therapeutic target. *J Exp Clin Cancer Res* **40**, 238, doi:10.1186/s13046-021-02045-y (2021).

31 Tremblay, C. S. *et al.* Loss-of-function mutations of Dynamin 2 promote T-ALL by enhancing IL-7 signalling. *Leukemia* **30**, 1993-2001, doi:10.1038/leu.2016.100 (2016).

32 Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122-128, doi:10.1038/s41586-019-1907-7 (2020).

33 Drost, J. *et al.* Sequential cancer mutations in cultured human intestinal stem cells. doi:10.1038/nature14415 (2015).

34 Coorens, T. H. H. *et al.* Embryonal precursors of Wilms tumor. *Science* **366**, 1247-+, doi:10.1126/science.aax1323 (2019).

35 Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Reports* **25**, 2308-+, doi:10.1016/j.celrep.2018.11.014 (2018).

36 Mitchell, T. J. *et al.* Timing the Landmark Events in the Evolution of Clear Cell Renal Cell Cancer: TRACERx Renal. *Cell* **173**, 611-623 e617, doi:10.1016/j.cell.2018.02.020 (2018).

37 Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *New England Journal of Medicine* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).

38 Palin, K. *et al.* Contribution of allelic imbalance to colorectal cancer. *Nat Commun* **9**, 3664, doi:10.1038/s41467-018-06132-1 (2018).

39 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).

40 Solimini, N. L. *et al.* Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337**, 104-109, doi:10.1126/science.1219580 (2012).

41 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).

42 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246-259, doi:10.1016/j.celrep.2012.12.008 (2013).

43 Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* (2020).

44 Roberts, N. D. Patterns of somatic genome rearrangement in human cancer. *Univesity of Cambridge, Doctoral Thesis* (2018).

45 Papaemmanuil, E. *et al.* Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221, doi:10.1056/NEJMoa1516192 (2016).

46 Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532-537, doi:10.1038/s41586-019-1672-7 (2019).

47 Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538-542, doi:10.1038/s41586-019-1670-9 (2019).

48 Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266-272, doi:10.1038/s41586-020-1961-1 (2020).

49 Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640-646, doi:10.1038/s41586-020-2214-z (2020).

50 Moore, L. *et al.* The mutational landscape of human somatic and germline cells. *Nature* **597**, 381-386, doi:10.1038/s41586-021-03822-7 (2021).

51 COSMIC. *Mutational Signatures (v3.2)*, <https://cancer.sanger.ac.uk/signatures/> (2021).

52 Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560-561, doi:10.1038/287560a0 (1980).

53 Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836 e816, doi:10.1016/j.cell.2019.03.001 (2019).

54 Brady, S. W. *et al.* Pan-neuroblastoma analysis reveals age- and signature-associated driver alterations. *Nat Commun* **11**, 5183, doi:10.1038/s41467-020-18987-4 (2020).

55 Haradhvala, N. J. *et al.* Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nature Communications* **9**, doi:10.1038/s41467-018-04002-4 (2018).

56 Pilati, C. *et al.* Mutational signature analysis identifies MUTYH deficiency in colorectal cancers and adrenocortical carcinomas. *J Pathol* **242**, 10-15, doi:10.1002/path.4880 (2017).

57 Viel, A. *et al.* A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20**, 39-49, doi:10.1016/j.ebiom.2017.04.022 (2017).

58 Rashid, M. *et al.* Adenoma development in familial adenomatous polyposis and MUTYH -associated polyposis: Somatic landscape and driver genes. *Journal of Pathology* **238**, 98-108, doi:10.1002/path.4643 (2016).

59 Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis* **30**, 763-770, doi:10.1093/mutage/gev073 (2015).

60 Christensen, S. *et al.* 5-Fluorouracil treatment induces characteristic T>G mutations in human cancer. *Nature Communications* **10**, 4571, doi:10.1038/s41467-019-12594-8 (2019).

61 Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks(+)E. coli. *Nature* **580**, 269-+, doi:10.1038/s41586-020-2080-8 (2020).

62 Boot, A. *et al.* Characterization of colibactin-associated mutational signature in an Asian oral squamous cell carcinoma and in other mucosal tumor types. *Genome Res* **30**, 803-813, doi:10.1101/gr.255620.119 (2020).

63 Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Article Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell*, 1282-1294, doi:10.1016/j.cell.2019.02.012 (2019).

64 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).

65 Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & De Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163**, 1641-1654, doi:10.1016/j.cell.2015.11.054 (2015).

66 Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-121, doi:10.1038/s41586-019-1913-9 (2020).

67 Maclachlan, K. H. *et al.* Copy number signatures predict chromothripsis and clinical outcomes in newly diagnosed multiple myeloma. *Nat Commun* **12**, 5172, doi:10.1038/s41467-021-25469-8 (2021).

68 Vohringer, H., Hoeck, A. V., Cuppen, E. & Gerstung, M. Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. *Nat Commun* **12**, 3628, doi:10.1038/s41467-021-23551-9 (2021).

69 Sanders, M. A. *et al.* Life without mismatch repair. *bioRxiv* (2021).

70 Ritsma, L. *et al.* Intestinal crypt homeostasis revealed at single-stem-cell level by in vivo live imaging. *Nature* **507**, 362-365, doi:10.1038/nature12972 (2014).

71 Snippert, H. J. *et al.* Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. *Cell* **143**, 134-144, doi:10.1016/j.cell.2010.09.016 (2010).

72 Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003-1007, doi:10.1038/nature06196 (2007).

73 Potten, C. S., Kellett, M., Roberts, S. A., Rew, D. A. & Wilson, G. D. Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut* **33**, 71-78, doi:10.1136/gut.33.1.71 (1992).

74 Khankari, N. K. *et al.* Association between Adult Height and Risk of Colorectal, Lung, and Prostate Cancer: Results from Meta-analyses of Prospective Studies and Mendelian Randomization Analyses. *PLoS Med* **13**, e1002118, doi:10.1371/journal.pmed.1002118 (2016).

75 St Clair, W. H. & Osborne, J. W. Crypt fission and crypt number in the small and large bowel of postnatal rats. *Cell Tissue Kinet* **18**, 255-262, doi:10.1111/j.1365-2184.1985.tb00655.x (1985).

76 Greaves, L. C. *et al.* Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. *Proceedings of the National Academy of Sciences* **103**, 714-719, doi:10.1073/pnas.0505903103 (2006).

77 Li, Y. Q., Roberts, S. A., Paulus, U., Loeffler, M. & Potten, C. S. The crypt cycle in mouse small intestinal epithelium. *J Cell Sci* **107 (Pt 12)**, 3271-3279 (1994).

78 Bruens, L., Ellenbroek, S. I. J., Rheenen, J. V. & Snippert, H. J. In Vivo Imaging Reveals Existence of Crypt Fission and Fusion in Adult Mouse Intestine. *Gastroenterology* **153**, 674-677.e673, doi:10.1053/j.gastro.2017.05.019 (2017).

79 Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909-918 e908, doi:10.1016/j.stem.2018.04.020 (2018).

80 Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**, 672-684 e611, doi:10.1016/j.cell.2020.06.036 (2020).

81 CRUK. *Early Diagnosis Data Hub*, <https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/> (2021).

82 Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260-264, doi:10.1038/nature19768 (2016).

83 Roerink, S. F. *et al.* Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457-+, doi:10.1038/s41586-018-0024-3 (2018).

84 Brown, K. F. *et al.* The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. *Br J Cancer* **118**, 1130-1141, doi:10.1038/s41416-018-0029-6 (2018).

85 CRUK. *Bowel cancer statistics*, <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer> (2021).

86 CRUK. *Early Diagnosis Data Hub*, <https://crukcancerintelligence.shinyapps.io/EarlyDiagnosis/> (2021).

87 ONS. *Cancer survival in England: adult, stage at diagnosis and childhood – patients followed up to 2018*, <https://www.ons.gov.uk/releases/cancersurvivalinenglandadultstageatdiagnosisandchildhoodpatientsfollowedupto2018> (2019).

88 Bowel Cancer UK. *Take the test. July marks 10 year anniversary of bowel cancer screening in England*, <https://www.bowelcanceruk.org.uk/news-and-blogs/news/take-the-test.-july-marks-10-year-anniversary-of-bowel-cancer-screening-in-england/> (2016).

89 National Institute for Health and Care Excellence. Pembrolizumab for untreated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency [NICE Technology appraisal guidance 709]. (2021).

90 Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N Engl J Med* **319**, 525-532, doi:10.1056/NEJM198809013190901 (1988).

91 National Institute of Clinical Excellence. Cetuximab and panitumumab for previously untreated metastatic colorectal cancer [NICE Technology appraisal guidance TA439]. (2017).

92 Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608-611, doi:10.1038/nature07602 (2009).

93 Fearon, E. F. & Vogelstein, B. for Colorectal Tumorigenesis. **61**, 759-767 (1990).

94 Deng, G. *et al.* Proximal and distal colorectal cancers show distinct gene-specific methylation profiles and clinical and molecular characteristics. *European Journal of Cancer* **44**, 1290-1301, doi:10.1016/j.ejca.2008.03.014 (2008).

95 Gupta, B. *et al.* Identification of High-Risk Aberrant Crypt Foci and Mucin-Depleted Foci in the Human Colon With Study of Colon Cancer Stem Cell Markers. *Clinical Colorectal Cancer* **16**, 204-213, doi:10.1016/j.clcc.2016.09.001 (2017).

96 Inoue, A. *et al.* B-RAF mutation and accumulated gene methylation in aberrant crypt foci (ACF), sessile serrated adenoma / polyp (SSA / P) and cancer in SSA / P. *British Journal of Cancer* **112**, 403-412, doi:10.1038/bjc.2014.545 (2014).

97 Takayama, T. *et al.* Analysis of K-ras, APC, and β-catenin in aberrant crypt foci in sporadic adenoma, cancer, and familial adenomatous polyposis. *Gastroenterology* **121**, 599-611, doi:10.1053/gast.2001.27203 (2001).

98 Nishisho, I. *et al.* Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665-669, doi:10.1126/science.1651563 (1991).

99 Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235-237, doi:10.1038/359235a0 (1992).

100 Lamlum, H. *et al.* APC mutations are sufficient for the growth of early colorectal adenomas. *Proceedings of the National Academy of Sciences* **97**, 2225-2228, doi:10.1073/pnas.040564697 (2000).

101 Fearon, E. R., Hamilton, S. R. & Vogelstein, B. Clonal analysis of human colorectal tumors. *Science* **238**, 193-197, doi:10.1126/science.2889267 (1987).

102 Jones, A. M. *et al.* Analysis of copy number changes suggests chromosomal instability in a minority of large colorectal adenomas. *J Pathol* **213**, 249-256, doi:10.1002/path.2234 (2007).

103 Bolhaqueiro, A. C. F. *et al.* Ongoing chromosomal instability and karyotype evolution in human colorectal cancer organoids. *Nat Genet* **51**, 824-834, doi:10.1038/s41588-019-0399-6 (2019).

104 Gala, M. & Chung, D. C. Hereditary colon cancer syndromes. *Semin Oncol* **38**, 490-499, doi:10.1053/j.seminoncol.2011.05.003 (2011).

105 Snowsill, T. *et al.* A systematic review and economic evaluation of diagnostic strategies for Lynch syndrome. *Health Technol Assess* **18**, 1-406, doi:10.3310/hta18580 (2014).

106 Yurgelun, M. B. *et al.* Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer. *J Clin Oncol* **35**, 1086-1095, doi:10.1200/JCO.2016.71.0012 (2017).

107 Campos, F. G., Figueiredo, M. N. & Martinez, C. A. Colorectal cancer risk in hamartomatous polyposis syndromes. *World J Gastrointest Surg* **7**, 25-32, doi:10.4240/wjgs.v7.i3.25 (2015).

108 Raskin, L. *et al.* Targeted sequencing of established and candidate colorectal cancer genes in the Colon Cancer Family Registry Cohort. *Oncotarget* **8**, 93450-93463, doi:10.18632/oncotarget.18596 (2017).

109 Valle, L. *et al.* Update on genetic predisposition to colorectal cancer and polyposis. *Mol Aspects Med* **69**, 10-26, doi:10.1016/j.mam.2019.03.001 (2019).

110 Palles, C. *et al.* Germline loss-of-function variants in the base-excision repair gene <em>MBD4</em> cause a Mendelian recessive syndrome of adenomatous colorectal polyposis and acute myeloid leukaemia. *bioRxiv* (2021).

111 Hampel, H. *et al.* Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). *N Engl J Med* **352**, 1851-1860, doi:10.1056/NEJMoa043146 (2005).

112 Aarnio, M., Mecklin, J. P., Aaltonen, L. A., Nystrom-Lahti, M. & Jarvinen, H. J. Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome. *Int J Cancer* **64**, 430-433, doi:10.1002/ijc.2910640613 (1995).

113 Aarnio, M. *et al.* Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer* **81**, 214-218, doi:10.1002/(sici)1097-0215(19990412)81:2<214::aid-ijc8>3.0.co;2-l (1999).

114 Bonadona, V. *et al.* Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA* **305**, 2304-2310, doi:10.1001/jama.2011.743 (2011).

115 Stoffel, E. *et al.* Calculation of risk of colorectal and endometrial cancer among patients with Lynch syndrome. *Gastroenterology* **137**, 1621-1627, doi:10.1053/j.gastro.2009.07.039 (2009).

116 Lynch, H. T., Snyder, C. L., Shaw, T. G., Heinen, C. D. & Hitchins, M. P. Milestones of Lynch syndrome: 1895-2015. *Nat Rev Cancer* **15**, 181-194, doi:10.1038/nrc3878 (2015).

117 Ahadova, A. *et al.* Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. *Int J Cancer* **143**, 139-150, doi:10.1002/ijc.31300 (2018).

118 Lagerstedt Robinson, K. *et al.* Lynch syndrome (hereditary nonpolyposis colorectal cancer) diagnostics. *J Natl Cancer Inst* **99**, 291-299, doi:10.1093/jnci/djk051 (2007).

119 Kastrinos, F. *et al.* Phenotype comparison of MLH1 and MSH2 mutation carriers in a cohort of 1,914 individuals undergoing clinical genetic testing in the United States. *Cancer Epidemiol Biomarkers Prev* **17**, 2044-2051, doi:10.1158/1055-9965.EPI-08-0301 (2008).

120 Binder, H. *et al.* Genomic and transcriptomic heterogeneity of colorectal tumours arising in Lynch syndrome. *J Pathol* **243**, 242-254, doi:10.1002/path.4948 (2017).

121 Hemminki, A. *et al.* Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. *Nat Genet* **8**, 405-410, doi:10.1038/ng1294-405 (1994).

122 Monahan, K. J. *et al.* Guidelines for the management of hereditary colorectal cancer from the British Society of Gastroenterology (BSG)/Association of Coloproctology of Great Britain and Ireland (ACPGBI)/United Kingdom Cancer Genetics Group (UKCGG). *Gut* **69**, 411-444, doi:10.1136/gutjnl-2019-319915 (2020).

123 Moller, P. *et al.* Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut* **66**, 464-472, doi:10.1136/gutjnl-2015-309675 (2017).

124 Seppala, T. *et al.* Colorectal cancer incidence in path_MLH1 carriers subjected to different follow-up protocols: a Prospective Lynch Syndrome Database report. *Hered Cancer Clin Pract* **15**, 18, doi:10.1186/s13053-017-0078-5 (2017).

125 Giardiello, F. M. *et al.* Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-society Task Force on colorectal cancer. *Am J Gastroenterol* **109**, 1159-1179, doi:10.1038/ajg.2014.186 (2014).

126 Parsons, R. *et al.* Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75**, 1227-1236, doi:10.1016/0092-8674(93)90331-j (1993).

127 Smyrk, T. C., Watson, P., Kaul, K. & Lynch, H. T. Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. *Cancer* **91**, 2417-2422 (2001).

128 Gatalica, Z. *et al.* Comprehensive tumor profiling identifies numerous biomarkers of drug response in cancers of unknown primary site: analysis of 1806 cases. *Oncotarget* **5**, 12440-12447, doi:10.18632/oncotarget.2574 (2014).

129 Gatalica, Z. *et al.* Programmed cell death 1 (PD-1) and its ligand (PD-L1) in common cancers and their correlation with molecular cancer type. *Cancer Epidemiol Biomarkers Prev* **23**, 2965-2970, doi:10.1158/1055-9965.EPI-14-0654 (2014).

130 Felton, K. E., Gilchrist, D. M. & Andrew, S. E. Constitutive deficiency in DNA mismatch repair: is it time for Lynch III? *Clin Genet* **71**, 499-500, doi:10.1111/j.1399-0004.2007.00801.x (2007).

131 Aretz, S. *et al.* Frequency and parental origin of de novo APC mutations in familial adenomatous polyposis. *Eur J Hum Genet* **12**, 52-58, doi:10.1038/sj.ejhg.5201088 (2004).

132 Gayther, S. A. *et al.* Regionally clustered APC mutations are associated with a severe phenotype and occur at a high frequency in new mutation cases of adenomatous polyposis coli. *Hum Mol Genet* **3**, 53-56, doi:10.1093/hmg/3.1.53 (1994).

133 Bisgaard, M. L., Fenger, K., Bulow, S., Niebuhr, E. & Mohr, J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. *Hum Mutat* **3**, 121-125, doi:10.1002/humu.1380030206 (1994).

134 Bulow, S. *et al.* Duodenal surveillance improves the prognosis after duodenal cancer in familial adenomatous polyposis. *Colorectal Dis* **14**, 947-952, doi:10.1111/j.1463-1318.2011.02844.x (2012).

135 Groen, E. J. *et al.* Extra-intestinal manifestations of familial adenomatous polyposis. *Ann Surg Oncol* **15**, 2439-2450, doi:10.1245/s10434-008-9981-3 (2008).

136 Campos, F. G. Surgical treatment of familial adenomatous polyposis: dilemmas and current recommendations. *World J Gastroenterol* **20**, 16620-16629, doi:10.3748/wjg.v20.i44.16620 (2014).

137 Spigelman, A. D., Williams, C. B., Talbot, I. C., Domizio, P. & Phillips, R. K. Upper gastrointestinal cancer in patients with familial adenomatous polyposis. *Lancet* **2**, 783-785, doi:10.1016/s0140-6736(89)90840-4 (1989).

138 Samadder, N. J. *et al.* Association of Sulindac and Erlotinib vs Placebo With Colorectal Neoplasia in Familial Adenomatous Polyposis: Secondary Analysis of a Randomized Clinical Trial. *JAMA Oncol* **4**, 671-677, doi:10.1001/jamaoncol.2017.5431 (2018).

139 Kim, B. & Giardiello, F. M. Chemoprevention in familial adenomatous polyposis. *Best Pract Res Clin Gastroenterol* **25**, 607-622, doi:10.1016/j.bpg.2011.08.002 (2011).

140 Giardiello, F. M. *et al.* Treatment of colonic and rectal adenomas with sulindac in familial adenomatous polyposis. *N Engl J Med* **328**, 1313-1316, doi:10.1056/NEJM199305063281805 (1993).

141 Giardiello, F. M. *et al.* Primary chemoprevention of familial adenomatous polyposis with sulindac. *N Engl J Med* **346**, 1054-1059, doi:10.1056/NEJMoa012015 (2002).

142 Cruz-Correa, M., Hylind, L. M., Romans, K. E., Booker, S. V. & Giardiello, F. M. Long-term treatment with sulindac in familial adenomatous polyposis: a prospective cohort study. *Gastroenterology* **122**, 641-645, doi:10.1053/gast.2002.31890 (2002).

143 Hyer, W. *et al.* Management of Familial Adenomatous Polyposis in Children and Adolescents: Position Paper From the ESPGHAN Polyposis Working Group. *J Pediatr Gastroenterol Nutr* **68**, 428-441, doi:10.1097/MPG.0000000000002247 (2019).

144 Cooke, M. S., Evans, M. D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* **17**, 1195-1214, doi:10.1096/fj.02-0752rev (2003).

145 Balaban, R. S., Nemoto, S. & Finkel, T. Mitochondria, oxidants, and aging. *Cell* **120**, 483-495, doi:10.1016/j.cell.2005.02.001 (2005).

146 Rosenquist, T. A., Zharkov, D. O. & Grollman, A. P. Cloning and characterization of a mammalian 8-oxoguanine DNA glycosylase. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 7429-7434, doi:DOI 10.1073/pnas.94.14.7429 (1997).

147 Mcgoldrick, J. P., Yeh, Y. C., Solomon, M., Essigmann, J. M. & Lu, A. L. Characterization of a Mammalian Homolog of the Escherichia-Coli Muty Mismatch Repair Protein. *Molecular and Cellular Biology* **15**, 989-996 (1995).

148 Cheng, K. C., Cahill, D. S., Kasai, H., Nishimura, S. & Loeb, L. A. 8-Hydroxyguanine, an Abundant Form of Oxidative DNA Damage, Causes G -> T and a -> C Substitutions. *Journal of Biological Chemistry* **267**, 166-172 (1992).

149 Komine, K. *et al.* Functional Complementation Assay for 47 MUTYH Variants in a MutY-Disrupted Escherichia coli Strain. *Hum Mutat* **36**, 704-711, doi:10.1002/humu.22794 (2015).

150 Ruggieri, V. *et al.* Loss of MUTYH function in human cells leads to accumulation of oxidative damage and genetic instability. *Oncogene* **32**, 4500-4508, doi:10.1038/onc.2012.479 (2013).

151 Wooden, S. H., Bassett, H. M., Wood, T. G. & McCullough, A. K. Identification of critical residues required for the mutation avoidance function of human MutY (hMYH) and implications in colorectal cancer. *Cancer Letters* **205**, 89-95, doi:10.1016/j.canlet.2003.10.006 (2004).

152 Kundu, S., Brinkmeyer, M. K., Livingston, A. L. & David, S. S. Adenine removal activity and bacterial complementation with the human MutY homologue (MUTYH) and Y165C, G382D, P391L and Q324R variants associated with colorectal cancer. *DNA Repair* **8**, 1400-1410, doi:10.1016/j.dnarep.2009.09.009 (2009).

153 Parker, A. R. *et al.* Cells with pathogenic biallelic mutations in the human MUTYH gene are defective in DNA damage binding and repair. *Carcinogenesis* **26**, 2010-2018, doi:10.1093/carcin/bgi166 (2005).

154 Sampson, J. R., Jones, S., Dolwani, S. & Cheadle, J. P. MutYH (MYH) and colorectal cancer. *Biochem Soc Trans* **33**, 679-683, doi:10.1042/BST0330679 (2005).

155 Sampson, J. R. *et al.* Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* **362**, 39-41, doi:10.1016/S0140-6736(03)13805-6 (2003).

156 Al-Tassan, N. *et al.* Inherited variants of MYH associated with somatic G : C -> T : A mutations in colorectal tumors. *Nature Genetics* **30**, 227-232, doi:10.1038/ng828 (2002).

157 Collaborative Group on Duodenal Polyposis in, M. A. P. *et al.* Duodenal Adenomas and Cancer in MUTYH-associated Polyposis: An International Cohort Study. *Gastroenterology* **160**, 952-954 e954, doi:10.1053/j.gastro.2020.10.038 (2021).

158 Nielsen, M. *et al.* Analysis of MUTYH genotypes and colorectal phenotypes in patients With MUTYH-associated polyposis. *Gastroenterology* **136**, 471-476, doi:10.1053/j.gastro.2008.10.056 (2009).

159 Out, A. A. *et al.* Leiden Open Variation Database of the MUTYH gene. *Hum Mutat* **31**, 1205-1215, doi:10.1002/humu.21343 (2010).

160 Dolwani, S. *et al.* Analysis of inherited MYH/(MutYH) mutations in British Asian patients with colorectal cancer. *Gut* **56**, 593, doi:10.1136/gut.2006.094532 (2007).

161 Ali, M. *et al.* Characterization of mutant MUTYH proteins associated with familial colorectal cancer. *Gastroenterology* **135**, 499-507, doi:10.1053/j.gastro.2008.04.035 (2008).

162 Goto, M. *et al.* Adenine DNA glycosylase activity of 14 human MutY homolog (MUTYH) variant proteins found in patients with colorectal polyposis and cancer. *Hum Mutat* **31**, E1861-1874, doi:10.1002/humu.21363 (2010).

163 Croitoru, M. E. *et al.* Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. *J Natl Cancer Inst* **96**, 1631-1634, doi:10.1093/jnci/djh288 (2004).

164 Farrington, S. M. *et al.* Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *Am J Hum Genet* **77**, 112-119, doi:10.1086/431213 (2005).

165 Jenkins, M. A. *et al.* Risk of colorectal cancer in monoallelic and biallelic carriers of MYH mutations: a population-based case-family study. *Cancer Epidemiol Biomarkers Prev* **15**, 312-314, doi:10.1158/1055-9965.EPI-05-0793 (2006).

166 Tenesa, A. *et al.* Association of MUTYH and colorectal cancer. *Br J Cancer* **95**, 239-242, doi:10.1038/sj.bjc.6603239 (2006).

167 Webb, E. L., Rudd, M. F. & Houlston, R. S. Colorectal cancer risk in monoallelic carriers of MYH variants. *Am J Hum Genet* **79**, 768-771; author reply 771-762, doi:10.1086/507912 (2006).

168 Lubbe, S. J., Di Bernardo, M. C., Chandler, I. P. & Houlston, R. S. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol* **27**, 3975-3980, doi:10.1200/JCO.2008.21.6853 (2009).

169 Theodoratou, E. *et al.* A large-scale meta-analysis to refine colorectal cancer risk estimates associated with MUTYH variants. *Br J Cancer* **103**, 1875-1884, doi:10.1038/sj.bjc.6605966 (2010).

170 Win, A. K. *et al.* Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. *Gastroenterology* **146**, 1208-1211 e1201-1205, doi:10.1053/j.gastro.2014.01.022 (2014).

171 Ma, X., Zhang, B. & Zheng, W. Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Gut* **63**, 326-336, doi:10.1136/gutjnl-2012-304121 (2014).

172 Vogt, S. *et al.* Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. *Gastroenterology* **137**, 1976-1985 e1971-1910, doi:10.1053/j.gastro.2009.08.052 (2009).

173 Nielsen, M. *et al.* Duodenal carcinoma in MUTYH-associated polyposis. *J Clin Pathol* **59**, 1212-1215, doi:10.1136/jcp.2005.031757 (2006).

174 Thomas, L. E. *et al.* Duodenal Adenomas and Cancer in MUTYH-associated Polyposis: An International Cohort Study. *Gastroenterology* **160**, 952-954 e954, doi:10.1053/j.gastro.2020.10.038 (2021).

175 Sutcliffe, E. G. *et al.* Multi-gene panel testing confirms phenotypic variability in MUTYH-Associated Polyposis. *Fam Cancer* **18**, 203-209, doi:10.1007/s10689-018-00116-2 (2019).

176 Georgeson, P. *et al.* Evaluating the utility of tumour mutational signatures for identifying hereditary colorectal cancer and polyposis syndrome carriers. *Gut*, doi:10.1136/gutjnl-2019-320462 (2021).

177 Jones, S. *et al.* Increased frequency of the k-ras G12C mutation in MYH polyposis colorectal adenomas. *Br J Cancer* **90**, 1591-1593, doi:10.1038/sj.bjc.6601747 (2004).

178 Morrison, A., Araki, H., Clark, A. B., Hamatake, R. K. & Sugino, A. A third essential DNA polymerase in S. cerevisiae. *Cell* **62**, 1143-1151, doi:10.1016/0092-8674(90)90391-Q (1990).

179 Pursell, Z. F., Isoz, I., Lundstrom, E. B., Johansson, E. & Kunkel, T. A. Yeast DNA polymerase epsilon participates in leading-strand DNA replication. *Science* **317**, 127-130, doi:10.1126/science.1144067 (2007).

180 Shinbrot, E. *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res* **24**, 1740-1750, doi:10.1101/gr.174789.114 (2014).

181 Li, H.-D., Zhang, H. & Castrillon, D. H. Polymerase-mediated ultramutagenesis in mice produces diverse cancers with high mutational load Graphical abstract The Journal of Clinical Investigation. *J Clin Invest* **128**, 4179, doi:10.1172/JCI122095 (2018).

182 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).

183 Hodel, K. P. *et al.* POLE Mutation Spectra Are Shaped by the Mutant Allele Identity, Its Abundance, and Mismatch Repair Status. *Mol Cell* **78**, 1166-1177 e1166, doi:10.1016/j.molcel.2020.05.012 (2020).

184 Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics* **45**, 136-143, doi:10.1038/ng.2503 (2013).

185 Palles, C. *et al.* The clinical features of polymerase proof-reading associated polyposis (PPAP) and recommendations for patient management. *Fam Cancer*, doi:10.1007/s10689-021-00256-y (2021).

186 Wang, F. *et al.* Evaluation of POLE and POLD1 Mutations as Biomarkers for Immunotherapy Outcomes Across Multiple Cancer Types. *JAMA Oncol* **5**, 1504-1506, doi:10.1001/jamaoncol.2019.2963 (2019).

187 Schmid, J. P. *et al.* Polymerase ε1 mutation in a human syndrome with facial dysmorphism, immunodeficiency, livedo, and short stature ("FILS syndrome"). *Journal of Experimental Medicine* **209**, 2323-2330, doi:10.1084/jem.20121303 (2012).

188 Weedon, M. N. *et al.* An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy. *Nature Genetics* **45**, 947-950, doi:10.1038/ng.2670 (2013).

189 Lessel, D. *et al.* POLD1 Germline Mutations in Patients Initially Diagnosed with Werner Syndrome. *Hum Mutat* **36**, 1070-1079, doi:10.1002/humu.22833 (2015).

190 Pelosini, C. *et al.* Identification of a novel mutation in the polymerase delta 1 (POLD1) gene in a lipodystrophic patient affected by mandibular hypoplasia, deafness, progeroid features (MDPL) syndrome. *Metabolism* **63**, 1385-1389, doi:10.1016/j.metabol.2014.07.010 (2014).

191 Elouej, S. *et al.* Exome sequencing reveals a de novo POLD1 mutation causing phenotypic variability in mandibular hypoplasia, deafness, progeroid features, and lipodystrophy syndrome (MDPL). *Metabolism* **71**, 213-225, doi:10.1016/j.metabol.2017.03.011 (2017).

192 Shastry, S. *et al.* A novel syndrome of mandibular hypoplasia, deafness, and progeroid features associated with lipodystrophy, undescended testes, and male

hypogonadism. *J Clin Endocrinol Metab* **95**, E192-197, doi:10.1210/jc.2010-0419 (2010).

193 Rivera, B., Castellsague, E., Bah, I., van Kempen, L. C. & Foulkes, W. D. Biallelic NTHL1 Mutations in a Woman with Multiple Primary Tumors. *N Engl J Med* **373**, 1985-1986, doi:10.1056/NEJMc1506878 (2015).

194 Weren, R. D. *et al.* A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* **47**, 668-671, doi:10.1038/ng.3287 (2015).

195 Grolleman, J. E. *et al.* Mutational Signature Analysis Reveals NTHL1 Deficiency to Cause a Multi-tumor Phenotype. *Cancer Cell* **35**, 256-266 e255, doi:10.1016/j.ccell.2018.12.011 (2019).

196 Sanders, M. A. *et al.* MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood* **132**, 1526-1534, doi:10.1182/blood-2018-05-852566 (2018).

197 Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75-82, doi:10.1126/science.aba8347 (2020).

198 Yamaji, Y. *et al.* Incidence and recurrence rates of colorectal adenomas estimated by annually repeated colonoscopies on asymptomatic Japanese. *Gut* **53**, 568-572, doi:10.1136/gut.2003.026112 (2004).

199 Mirzaie, A. Z., Khakpour, H., Mireskandari, M., Shayanfar, N. & Fatahi, L. Investigating The Frequency of Serrated Polyps/Adenomas and Their Subtypes in Colonic Polyp Samples. *Med Arch* **70**, 198-202, doi:10.5455/medarh.2016.70.198-202 (2016).

200 Knudsen, A. L., Bisgaard, M. L. & Bulow, S. Attenuated familial adenomatous polyposis (AFAP). A review of the literature. *Fam Cancer* **2**, 43-55, doi:10.1023/a:1023286520725 (2003).

201 Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature*, doi:10.1038/s41586-021-03477-4 (2021).

202 Bae, T. *et al.* Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **555**, 550-555 (2018).

203 Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886, doi:10.1126/science.aaa6806 (2015).

204 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science* **917**, 911-917 (2018).

205 Coorens, T. H. H. *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387-392, doi:10.1038/s41586-021-03790-y (2021).

206 Coorens, T. H. H. *et al.* Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80-85, doi:10.1038/s41586-021-03345-1 (2021).

207 Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *bioRxiv* (2021).

208 Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85-90, doi:10.1038/s41586-021-03548-6 (2021).

209 Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422-425, doi:10.1038/nature13448 (2014).

210 Machado, H. E. *et al.* Genome-wide mutational signatures of immunological diversification in normal lymphocytes. *bioRxiv* (2021).

211 Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473-+, doi:10.1038/s41586-018-0497-0 (2018).

212 Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Reports* **25**, 2308-2316.e2304, doi:10.1016/j.celrep.2018.11.014 (2018).

213 Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606, doi:10.1038/nprot.2014.170 (2014).

214 Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513, doi:10.1073/pnas.1208715109 (2012).

215 Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc Natl Acad Sci U S A* **113**, 9846-9851, doi:10.1073/pnas.1607794113 (2016).

216 Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**, 269-285, doi:10.1038/nrg.2017.117 (2018).

217 Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci U S A* **110**, 1999-2004, doi:10.1073/pnas.1221068110 (2013).

218 Tomasetti, C. & Vogelstein, B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78-81, doi:10.1126/science.1260825 (2015).

219 Tang, J. *et al.* The genomic landscapes of individual melanocytes from human skin. *Nature* **586**, 600-605, doi:10.1038/s41586-020-2785-8 (2020).

220 Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312-317, doi:10.1038/s41586-018-0811-x (2019).

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

222 Kuijk, E. *et al.* Early divergence of mutational processes in human fetal tissues. *Sci Adv* **5**, eaaw1271, doi:10.1126/sciadv.aaw1271 (2019).

223 Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, doi:10.1126/science.aaw0726 (2019).

224 Li, R. *et al.* A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398-403, doi:10.1038/s41586-021-03836-1 (2021).

225 Rouhani, F. J. *et al.* Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells. *PLoS Genet* **12**, e1005932, doi:10.1371/journal.pgen.1005932 (2016).

226 Behjati, S. *et al.* Mutational signatures of ionizing radiation in second malignancies. *Nat Commun* **7**, 12605, doi:10.1038/ncomms12605 (2016).

227 Pich, O. *et al.* The evolution of hematopoietic cells under cancer therapy. *Nat Commun* **12**, 4803, doi:10.1038/s41467-021-24858-3 (2021).

228 Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. **917**, 911-917 (2018).

229 Nanki, K. *et al.* Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *Nature* **577**, 254-259, doi:10.1038/s41586-019-1844-5 (2020).

230 Kakiuchi, N. *et al.* Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**, 260-265, doi:10.1038/s41586-019-1856-1 (2020).

231 Ng, S. W. K. *et al.* Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature*, doi:10.1038/s41586-021-03974-6 (2021).

232 Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194-1217, doi:10.1016/j.cell.2013.05.039 (2013).

233 Vijg, J. & Dong, X. Pathogenic Mechanisms of Somatic Mutation and Genome Mosaicism in Aging. *Cell* **182**, 12-23, doi:10.1016/j.cell.2020.06.024 (2020).

234 Schumacher, B., Pothof, J., Vijg, J. & Hoeijmakers, J. H. J. The central role of DNA damage in the ageing process. *Nature* **592**, 695-703, doi:10.1038/s41586-021-03307-7 (2021).

235 Szilard, L. On the Nature of the Aging Process. *Proc Natl Acad Sci U S A* **45**, 30-45, doi:10.1073/pnas.45.1.30 (1959).

236 Morley, A. A. Is ageing the result of dominant and co-dominant mutations? *J Theor Biol* **98**, 469-474, doi:10.1016/0022-5193(82)90131-x (1982).

237 Burnet, F. M. Intrinsic mutagenesis: a genetic basis of ageing. *Pathology* **6**, 1-11, doi:10.3109/00313027409077150 (1974).

238 Curtis, H. J. A composite theory of aging. *Gerontologist* **6**, 143-149, doi:10.1093/geront/6.3_part_1.143 (1966).

239 Orgel, L. E. Ageing of clones of mammalian cells. *Nature* **243**, 441-445, doi:10.1038/243441a0 (1973).

240 Kirkwood, T. B. & Holliday, R. The evolution of ageing and longevity. *Proc R Soc Lond B Biol Sci* **205**, 531-546, doi:10.1098/rspb.1979.0083 (1979).

241 Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *bioRxiv* (2021).

242 Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat Protoc*, doi:10.1038/s41596-020-00437-6 (2020).

243 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinformatics* **56**, 15 10 11-15 10 18, doi:10.1002/cpbi.20 (2016).

244 Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15 17 11-15 17 12, doi:10.1002/0471250953.bi1507s52 (2015).

245 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

246 Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinformatics* **56**, 15 19 11-15 19 17, doi:10.1002/cpbi.17 (2016).

247 Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* **17**, 66, doi:10.1186/s13059-016-0924-1 (2016).

248 Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101**, 1566-1581, doi:10.1198/016214506000000302 (2006).

249 Pich, O. *et al.* The mutational footprints of cancer therapies. *Nature Genetics* **51**, 1732-+, doi:10.1038/s41588-019-0525-5 (2019).

250 Zou, X. Q. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer*, doi:10.1038/s43018-021-00200-0 (2021).

251 Abelson, S. *et al.* Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400-404, doi:10.1038/s41586-018-0317-6 (2018).

252 Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718, doi:10.1038/nature21703 (2017).

253 Farmery, J. H. R., Smith, M. L., Diseases, N. B.-R. & Lynch, A. G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep* **8**, 1300, doi:10.1038/s41598-017-14403-y (2018).

254 Sieverling, L. *et al.* Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat Commun* **11**, 733, doi:10.1038/s41467-019-13824-9 (2020).

255 Feuerbach, L. *et al.* TelomereHunter - in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, 272, doi:10.1186/s12859-019-2851-0 (2019).

256 Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144, doi:10.1016/j.cell.2010.09.016 (2010).

257 Albertson, T. M. *et al.* DNA polymerase ε and δ proofreading suppress discrete mutator and cancer phenotypes in mice. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 17101-17104, doi:10.1073/pnas.0907147106 (2009).

258 Goldsby, R. E. *et al.* High incidence of epithelial cancers in mice deficient for DNA polymerase delta proofreading. *Proc Natl Acad Sci U S A* **99**, 15560-15565, doi:10.1073/pnas.232340999 (2002).

259 Fortune, J. M. *et al.* Saccharomyces cerevisiae DNA Polymerase δ: High fidelity for base substitutions but lower fidelity for single-and multi-base deletions. *Journal of Biological Chemistry* **280**, 29980-29987, doi:10.1074/jbc.M505236200 (2005).

260 Schmitt, M. W., Matsumoto, Y. & Loeb, L. A. High fidelity and lesion bypass capability of human DNA polymerase delta. *Biochimie* **91**, 1163-1172, doi:10.1016/j.biochi.2009.06.007 (2009).

261 Korona, D. A., Lecompte, K. G. & Pursell, Z. F. The high fidelity and unique error signature of human DNA polymerase ε. *Nucleic Acids Research* **39**, 1763-1773, doi:10.1093/nar/gkq1034 (2011).

262 Fang, H. *et al.* Mutational processes of distinct POLE exonuclease domain mutants drive an enrichment of a specific TP53 mutation in colorectal cancer. *PLoS Genet* **16**, e1008572, doi:10.1371/journal.pgen.1008572 (2020).

263 Temko, D. *et al.* Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. *J Pathol* **245**, 283-296, doi:10.1002/path.5081 (2018).

264 Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Research* **28**, 654-665, doi:10.1101/gr.230219.117 (2018).

265 Lin, S.-H. *et al.* The somatic mutation landscape of premalignant colorectal adenoma. *Gut*, gutjnl-2016-313573, doi:10.1136/gutjnl-2016-313573 (2017).

266 Zhou, Z. X., Lujan, S. A., Burkholder, A. B., Garbacz, M. A. & Kunkel, T. A. Roles for DNA polymerase delta in initiating and terminating leading strand DNA replication. *Nat Commun* **10**, 3992, doi:10.1038/s41467-019-11995-z (2019).

267 Risio, M., Coverlizza, S., Ferrari, A., Candelaresi, G. L. & Rossini, F. P. Immunohistochemical study of epithelial cell proliferation in hyperplastic polyps, adenomas, and adenocarcinomas of the large bowel. *Gastroenterology* **94**, 899-906, doi:10.1016/0016-5085(88)90545-8 (1988).

268 Tanaka, M. *et al.* Evidence of the monoclonal composition of human endometrial epithelial glands and mosaic pattern of clonal distribution in luminal epithelium. *American Journal of Pathology* **163**, 295-301, doi:Doi 10.1016/S0002-9440(10)63653-X (2003).

269 Lac, V. *et al.* Oncogenic mutations in histologically normal endometrium: the new normal? *J Pathol* **249**, 173-181, doi:10.1002/path.5314 (2019).

270 Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep* **24**, 1777-1789, doi:10.1016/j.celrep.2018.07.037 (2018).

271 Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350** (2015).

272 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

273 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).

274 Bellido, F. *et al.* Open POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and / or polyposis : review of reported cases and recommendations for genetic testing and surveillance. *Genet Med* **18**, 325-332, doi:10.1038/gim.2015.75 (2016).

275 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nature Genetics* **48**, 126-133, doi:10.1038/ng.3469 (2015).

276 Goriely, A. & Wilkie, A. O. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am J Hum Genet* **90**, 175-200, doi:10.1016/j.ajhg.2011.12.017 (2012).

277 Schulz, K. N. & Harrison, M. M. Mechanisms regulating zygotic genome activation. *Nat Rev Genet* **20**, 221-234, doi:10.1038/s41576-018-0087-x (2019).

278 Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics* **49**, 1684-1692, doi:10.1038/ng.3991 (2017).

279 Audebert, M., Radicella, J. P. & Dizdaroglu, M. Effect of single mutations in the OGG1 gene found in human tumors on the substrate specificity of the Ogg1 protein. *Nucleic Acids Res* **28**, 2672-2678, doi:10.1093/nar/28.14.2672 (2000).

280 Audebert, M. *et al.* Alterations of the DNA repair gene OGG1 in human clear cell carcinomas of the kidney. *Cancer Res* **60**, 4740-4744 (2000).

281 Smith, C. G. *et al.* Role of the oxidative DNA damage repair gene OGG1 in colorectal tumorigenesis. *J Natl Cancer Inst* **105**, 1249-1253, doi:10.1093/jnci/djt183 (2013).

282 Kinnersley, B. *et al.* Re: Role of the oxidative DNA damage repair gene OGG1 in colorectal tumorigenesis. *J Natl Cancer Inst* **106**, doi:10.1093/jnci/dju086 (2014).

283 Blons, H. *et al.* Frequent allelic loss at chromosome 3p distinct from genetic alterations of the 8-oxoguanine DNA glycosylase 1 gene in head and neck cancer. *Mol Carcinog* **26**, 254-260 (1999).

284 Xie, Y. *et al.* Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. *Cancer Res* **64**, 3096-3102, doi:10.1158/0008-5472.can-03-3834 (2004).

285 Ohno, M. *et al.* 8-oxoguanine causes spontaneous de novo germline mutations in mice. *Sci Rep* **4**, 4689, doi:10.1038/srep04689 (2014).

286 Lee, B. C. H. *et al. Mutational landscape of normal epithelial cells in Lynch Syndrome patients* (2021).

287 Brosens, L. A., Keller, J. J., Offerhaus, G. J., Goggins, M. & Giardiello, F. M. Prevention and management of duodenal polyps in familial adenomatous polyposis. *Gut* **54**, 1034-1043, doi:10.1136/gut.2004.053843 (2005).

288 Hammoudi, N. *et al.* Duodenal tumor risk in Lynch syndrome. *Dig Liver Dis* **51**, 299-303, doi:10.1016/j.dld.2018.10.005 (2019).

289 Lillemoe, K. & Imbembo, A. L. Malignant neoplasms of the duodenum. *Surg Gynecol Obstet* **150**, 822-826 (1980).

290 Calabrese, P. & Shibata, D. A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates. *Bmc Cancer* **10**, doi:Artn 3
10.1186/1471-2407-10-3 (2010).

291 Morley, A. A. The somatic mutation theory of ageing. *Mutation Research DNAging* **338**, 19-23, doi:10.1016/0921-8734(95)00007-S (1995).

292 Navarro, C. L., Cau, P. & Levy, N. Molecular bases of progeroid syndromes. *Hum Mol Genet* **15 Spec No 2**, R151-161, doi:10.1093/hmg/ddl214 (2006).

293 Vidak, S. & Foisner, R. Molecular insights into the premature aging disease progeria. *Histochem Cell Biol* **145**, 401-417, doi:10.1007/s00418-016-1411-1 (2016).