The landscape of somatic mutation in normal colorectal epithelial cells

Henry Lee-Six[1], Sigurgeir Olafsson[1], Peter Ellis[1], Robert J. Osborne[1], Mathijs A. Sanders[1,2], Luiza Moore[1], Nikitas Georgakopoulos[3], Franco Torrente[4], Ayesha Noorani[5], Martin Goddard[6], Philip Robinson[1], Tim H. H. Coorens[1], Laura O'Neill[1], Christopher Alder[1], Jingwei Wang[1], Rebecca C. Fitzgerald[5], Matthias Zilbauer[4,7], Nicholas Coleman[8], Kourosh Saeb-Parsy[3], Inigo Martincorena[1], Peter J. Campbell[1], Michael R. Stratton[1]*

1. Wellcome Sanger Institute, Hinxton, UK
2. Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands
3. Department of Surgery and Cambridge NIHR Biomedical Research Centre, Cambridge Biomedical Campus, Cambridge, UK
4. Department of Paediatric Gastroenterology, Hepatology, and Nutrition, Addenbrooke's, Cambridge, UK
5. Medical Research Council Cancer Unit, Hutchison/Medical Research Council Research Centre, University of Cambridge, Cambridge, UK, and Cambridge University Hospitals NHS Trust, Hills Road, Cambridge, UK
6. Department of Pathology, Papworth Hospital NHS Trust, UK
7. University Department of Paediatrics, University of Cambridge, UK
8. Department of Pathology, University of Cambridge, Cambridge, UK and Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
*e-mail: mrs@sanger.ac.uk

Abstract

The colorectal adenoma-carcinoma sequence has provided a paradigmatic framework for understanding the successive somatic genetic changes and consequent clonal expansions leading to cancer. As for most cancer types, however, understanding of the earliest phases of colorectal neoplastic change, which may occur in morphologically normal tissue, is comparatively limited. Here, we whole genome sequenced hundreds of normal crypts from 42 individuals. Signatures of multiple mutational processes were revealed, some ubiquitous and continuous, others only found in some individuals, in some crypts or during certain periods of life. Likely driver mutations were present in ~1% of normal colorectal crypts in middle-aged individuals, indicating that adenomas and carcinomas are rare outcomes of a pervasive process of neoplastic change across morphologically normal colorectal epithelium. Colorectal cancers exhibit substantially elevated mutation burdens relative to normal cells. Sequencing normal colorectal cells provides quantitative insights into the genomic and clonal evolution of cancer.

Introduction

Sequencing of >20,000 cancers has identified the repertoire of driver mutations in cancer genes converting normal cells into cancer cells and revealed the mutational signatures of the underlying biological processes generating somatic mutations[1,2]. Cancers are, however, end stages of an evolutionary process operating within cell populations and commonly arise through the accumulation of multiple driver mutations engendering a series of clonal expansions. Understanding this progression has depended on identifying somatic mutations in morphologically abnormal neoplastic proliferations representing intermediate stages between normal and cancer cells.[3]

As for most cancer types, however, the earliest stages of progression to colorectal cancer remain less well understood. The driver mutation that first sets a colorectal epithelial cell on the path to cancer is likely caused by mutational processes operative in normal cells, of which there is limited understanding. The nature and numbers of the earliest neoplastic clones with

driver mutations, which conceivably are morphologically indistinguishable from normal cells, are similarly unclear. In large part, these deficiencies are due to the technical challenge of identifying somatic mutations in normal tissues, which are composed of myriad microscopic cell clones. Several different approaches have been adopted to address this[4-14], revealing signatures of common somatic mutational processes in normal cells of the small and large intestine, liver, blood, skin, and nervous system. Thus far, however, studies have not been of sufficient scale to characterise variation in signature activity or detect less frequent processes[4-14]. Remarkably high proportions of normal skin, oesophageal, and endometrial epithelial cells have been shown to be members of clones already carrying driver mutations[10,11,15,16], and large mutant clones have been detected in blood[17-20]. The extent of this phenomenon in the colon, an organ with a high cancer incidence, has not been investigated.

Colonic epithelium is a contiguous cell sheet organised into ~15,000,000 crypts each composed of ~2,000 cells[21]. Towards the base of each crypt resides a small number of stem cells ancestral to the maturing and differentiated cells in the crypt[22]. These stem cells stochastically replace one another through a process of neutral drift[23,24] such that all stem cells, and thus all cells, in a crypt derive from a single ancestor stem cell that existed in recent years[25-27]. The somatic mutations that were present in this ancestor are thus found in all ~2,000 descendant cells and can be revealed by DNA sequencing of an individual crypt. These stem cells are thought to be the cells of origin of colorectal cancers[28]. To characterise the earliest stages of colorectal carcinogenesis, somatic mutation burdens, mutational signatures, clonal dynamics, and the frequency of driver mutations in normal colorectal epithelium were explored by sequencing individual colorectal crypts.

Results

Somatic mutations and mutational signatures

2,035 individual colonic crypts from the normal epithelium of 42 individuals aged 11 to 78, of whom 15 had a history of colorectal cancer and 27 did not (Methods, Supplementary Table 1), were isolated using laser capture microdissection and sequenced. The distribution of mutation allele fractions from whole genome sequencing of 571 individual crypts showed that crypts were derived from a single ancestral stem cell (Extended Data Fig. 1d), and simulations indicated that ~90% of mutations called were fully clonal (Supplementary Results 2). There was substantial variation in mutation burdens between individual crypts, ranging from 1,508 to 15,329 for individuals in their sixties, which was not obviously attributable to technical factors. To explore the biological basis of this variation we extracted mutational signatures and estimated the contribution of each to the mutation burden of every crypt (Methods, Supplementary Results 1).

Nine single base substitution (SBS), six doublet base substitution (DBS), and five small indel (ID) mutational signatures were found. Of these, 14 closely matched (Methods) a known reference signature (SBS1, SBS2, SBS5, SBS13, SBS18, DBS2, DBS4, DBS6, DBS8, DBS9, DBS11, ID1, ID2, and ID5, nomenclature as in Alexandrov et al[1]) and six did not (SBSA, SBSB, SBSC, SBSD, IDA, and IDB) (Fig. 1, Extended Data Fig. 2-4). Thus, new mutational signatures were extracted despite extensive prior analysis of cancers, perhaps due to masking by the comparative complexity of signature mixtures present in cancer genomes.

Ubiquitous mutational signatures

11 signatures (SBS1, SBS5, SBS18, DBS2, DBS4, DBS6, DBS9, DBS11, ID1, ID2, and ID5) were found in >85% of crypts and are here termed "ubiquitous". All have been previously described[1].

SBS1 is characterised by C>T substitutions at NCG trinucleotides (the mutated base is underlined) and is likely due to deamination of 5-methylcytosine. Its mutation load correlated linearly with age (Fig. 2). There was, however, variation in SBS1 mutation burdens between crypts from the same individual (p=2.25e-27). This was due, in part, to different SBS1 mutation rates in different colonic sectors, with mean rates across individuals of 16.8 mutations per year (95% CI 15.2-18.3) in the right (ascending and caecum), 16.1 (95% CI 14.4-17.5) in the transverse, and 12.8 (95% CI 11.1-14.4) in the left (descending and sigmoid) colon. The SBS1 mutation rate in the terminal ileum was 12.7 (95% CI 10.6-14.9) (Supplementary Results 1). SBS5 is a flat, featureless signature of unknown cause and SBS18 is characterised by C>A mutations, which may be due to DNA damage by reactive oxygen species[29,30]. Their mutation burdens correlated with age, with the same ordering of sector differences as SBS1 (p=9.89e-26 for SBS5, p=5.43e-22 for SBS18). Even after taking anatomical location and age into account, differences in mutation burden remained between different crypts, notably for SBS18 (Fig. 2, Extended Data Fig. 9, Extended Data Fig. 6al). Combining ubiquitous SBS mutational signatures, and averaging over anatomical sites, the mutation rate was 43.6 mutations per year, comparable with previous estimates[4].

DBS2, DBS4, DBS6, DBS9, and DBS11 were tightly correlated in all colonic crypts. ID1, ID2, and ID5, which are characterised by insertions and deletions of a single T and may be the consequence of slippage during DNA replication, all accumulated linearly with age with the same order of sector differences as SBS1 (p=1.66e-05 for ID1, p=4.53e-06 for ID2, and p=4.53e-06 for ID5) (Supplementary Results 1, Extended Data Fig. 5).

The correlations of ubiquitous signatures with age indicate that the mutational processes underlying them operate throughout life, in all individuals and all colorectal stem cells. However, the results also suggest that differences in physiology and/or microenvironment (and potentially age of the most recent common ancestor of crypts[27]) between different sectors of the colon cause measurable differences in somatic mutation rates.

Sporadic mutational signatures

Nine signatures (SBS2, SBS13, SBSA-D, DBS8, and IDA-B) were present only in a subset of individuals and/or a subset of crypts and are termed "sporadic". All were novel, except for SBS2, SBS13 and DBS8. SBS2 and SBS13 are characterised by C>T and C>G mutations at TCN, are likely due to APOBEC cytidine deaminases and usually occur together[31,32]. They were unequivocally present in only two crypts (a colonic crypt (Extended Data Fig. ai) and an ileal crypt (Extended Data Fig. ao) from different individuals), occurring together and each accounting for over 150 mutations. To our knowledge, this is the first report that APOBEC DNA-editing of the human genome occurs in normal cells *in vivo*. The sequence context of these mutations in normal colon suggests that APOBEC3A is the major contributing enzyme[33].

Four SBS signatures that do not match the reference set, SBSA-D, were found in normal colorectal cells (SBSA has recently been reported in an oral squamous carcinoma[34]). SBSA is characterised by T>C at ATA, ATT, and TTT, and T>G at TTT. Its mutation burden correlated closely with that of IDA, in which single T deletions in short runs of Ts (with a mode of four) predominate, suggesting that they are due to the same underlying mutational

process. SBSA was detectable in 29/42 individuals, often accounting for thousands of mutations in just a subset of crypts. It clustered spatially in the colon, with crypts from the same biopsy carrying the signature even though the mutations themselves were not shared (Supplementary Results 1, Extended Data Fig. 9). 2.5-fold more T>C mutations occurred when the T was on the transcribed than on the untranscribed strand. Transcriptional strand bias is often due to transcription coupled nucleotide excision repair acting on DNA damaged by exogenous exposures causing covalently bound bulky adducts, but can also be caused by transcription coupled DNA damage[35]. Assuming either is the case, damage to adenine underlies SBSA. To investigate the timing of SBSA, phylogenetic trees of mutations were constructed and the mutational signatures in each branch established (Fig. 3, Extended Data Fig. 6). SBSA was confined to early branches of these phylogenies (when these were available for analysis) (Fig. 3b, Extended Data Fig. 6 f, h, z, aa, am, ao, aq). Using the number of SBS1 mutations as indicators of real time, the mutational process underlying SBSA appears to be active before 10 years of age (Supplementary Methods, Extended Data Fig. 6aq). SBSA may therefore be caused by an extrinsic, locally acting and patchily distributed mutagenic insult occurring during childhood.

SBSB was characterised by C>T at A<u>C</u>A, T>A at C<u>T</u>N, and T>G at G<u>T</u>G and was present in subsets of crypts from four individuals (e, aa, ai, and aj in Extended Data Figure 6), accounting for variable numbers of substitutions, with a maximum of 3,002 in one crypt. (Fig. 3c, Extended Data Fig. 6ai). In the two individuals in whom it could be timed (Extended Data Fig. 6 aa, ai, aq), it appeared – as with SBSA – to be most active in the first decade of life. SBSB correlated with DBS8 and IDB (Fig. 3c, Extended Data Fig. 9), suggesting that they are caused by the same underlying mutational process. DBS8 is composed of AC>CA and AC>CT mutations and has previously been reported in rare hypermutated cancers with no obvious cause[1]. IDB is dominated by deletion of a single T with no other Ts surrounding it.

SBSC is characterised by one C>T mutation in CC dinucleotides. It primarily affects three crypts, with 1,050, 827, and 695 mutations respectively, from the left colon of one individual with an unremarkable history (Extended Data Fig. 9, Extended Data Fig. 6m, Supplementary Table 1).

All crypts from a 66 year-old man carried many thousands of mutations of SBSD (Figure 3d, Extended Data Fig. 6ap), characterised by T>A substitutions with a transcriptional strand bias compatible with damage to adenine. This individual had been treated with multiple chemotherapeutic agents (cyclophosphamide, doxorubicin, vincristine, prednisolone, chlorambucil, bleomycin and etoposide) for lymphoma and subsequently developed caecal adenocarcinoma. SBSD resembles SBS25 (cosine similarity 0.9), previously found in Hodgkin lymphoma cell lines from two chemotherapy-treated patients[31,36]. To our knowledge this is the first time that the mutational consequences of chemotherapy have been demonstrated in normal human cells *in vivo*. The mutation burden in his colorectal epithelium was 3-5 fold higher than expected for his age, thus by extrapolation equivalent to that of a 200-300 year-old.

Copy number changes and structural variants

Copy number changes and/or structural variants were found in 80 out of 449 (18%) evaluable crypts. Five crypts exhibited eight whole chromosome copy number increases which affected the same three chromosomes – 3, 7 and 9 – as well as the X chromosome (Extended Data Fig. 7a). Thus, copy number increases clustered in certain crypts and tended to affect certain

chromosomes. No whole chromosome losses were observed. Arm-level chromosome 7 copy number increases are common in colorectal cancers[37] and adenomas[38]. Chromosome 3 and 9 copy number increases are seen in colorectal cancers, but are almost as frequently deleted[37]. Copy number neutral loss of heterozygosity (CNN-LOH) was observed in 12 crypts, affecting chromosomes 1p, 6p, 7p, 8q, 9q, 10q (twice), 17p, 17q, 18q, 21q and 22q (Extended Data Fig. 7c). CNN-LOH is frequently observed in colorectal cancers, although the specific changes that we observe here are not recurrent features[39]. Five copy number changes could be timed and all were estimated to have occurred in adulthood (Extended Data Fig. 7b). Two changes that affected the same crypt appeared to be synchronous (Supplementary Results 1). Structural variant analysis detected 48 large deletions, 18 tandem duplications, four translocations, and two inversions (Extended Data Fig 7d, Supplementary Results 1). All were private to a single crypt, except for one deletion which was present in two adjacent crypts sharing few mutations, indicating that it occurred during gestation or early childhood.

Driver mutations

Driver mutations are those that confer a selective advantage during cancer evolution. To search for drivers in normal colon, the whole genome sequences of 571 crypts were supplemented with targeted sequencing of 90 known colorectal cancer genes (Supplementary Table 4) in additional crypts. In total, substitutions in these genes were evaluable in 1,403 crypts and indels in 1,046. Statistical analysis revealed evidence of positive selection on the recessive cancer genes *AXIN2* (three truncating mutations, adjusted q value 0.004) and *STAG2* (two truncating mutations, adjusted q value 0.038) indicating that these mutations are likely drivers. Additional likely drivers were identified in cancer genes with canonical missense hotspot mutations. Nine hotspot mutations in *PIK3CA* (E542K, R38H), *ERBB2* (R678Q, V842I, T862A), *ERBB3* (R475W, R667L), and *FBXW7* (R505C, R658Q) were observed (Extended Data Fig. 8). Given the specificity of these hotspot mutations, most are likely to be drivers. In addition, heterozygous truncating mutations were found in the recessive cancer genes *ARID2*, *ATM* (two), *ATR*, *BRCA2*, *CDK12* (two), *CDKN1B*, *RNF43* (two), *TBL1XR1*, and *TP53* (Supplementary Table 5). There was no statistical evidence for selection of truncating mutations in the set of 90 colorectal cancer genes overall. The possibility that some have conferred clonal growth advantage, however, is not excluded. None of the analysed crypts carried more than one putative driver.

23 pairs of adjacent crypts shared over 100 SBS1 mutations and thus were likely to have been generated by postnatal crypt fission. Two pairs carried driver mutations (one *AXIN2* nonsense mutation and one *PIK3CA* E542K), although the association of driver mutations with crypt fission is not significant (p=0.17). In one sister crypt the *AXIN2* mutation was rendered homozygous by CNN-LOH of 17q, revealing ongoing clonal evolution in normal colon (Fig. 4, Fig. 3b).

On the conservative assumption that just the *AXIN2* and *STAG2* truncating mutations and the missense hotspot mutations in *PIK3CA*, *ERBB2*, *ERBB3* and *FBXW7* are drivers, ~1% of normal colorectal crypts (~150,000 crypts) in a 50-60 year old (the mean age of crypts assessed for drivers in our cohort was 53 years old) carries a driver mutation. Since in the over 70s ~40% of people have an adenoma on colonoscopy[40] and ~5% of people develop colorectal cancer over their lifetime[41] (and some of these may arise from more recently-acquired driver mutations) only an extremely small proportion of these crypt microneoplasms becomes a macroscopically detectable adenoma (< 1/375,000) or carcinoma (< 1/3,000,000) within the following few decades.

Clonal dynamics of normal epithelium

The distribution of allele fractions of mutations within the crypt informs on the dynamics of stem cell turnover within the crypt. We estimate that the average number of years to the most recent common ancestor of crypts is 5.5 years (CI95 for the mean: 1-10.5 years), similar to previous estimates[14]. Our data are compatible with previous estimates of 7 active stem cells and 1.3 stem cell replacements per year[14] or with 5 stem cells and 0.6 stem cell replacements per year[42], but we cannot exclude a larger number of stem cells turning over more frequently (Extended data figure 10, Supplementary Results 2). The microdissection approach also allowed investigation of the clonal structure of colonic epithelium beyond the crypt. By comparing the genetic relatedness of crypts with their spatial relatedness, we estimate that crypts fission at a mean rate of once every 27 years (CI95: 15.9-47.6 years) (Extended data figure 10, Supplementary Results 2).

Comparisons with colorectal cancer

There are marked differences between the genomes of normal colorectal stem cells and those of colorectal cancers. The total mutation burdens of substitutions (10,000-20,000) and indels (1,000-2,000) found in most colorectal carcinomas[1] (excluding those with hypermutator phenotypes in which it is usually >10-fold more) are higher than the ~3,000 substitutions and 300 indels found in most normal crypts from 50-60 year-old individuals (Extended Data Figure 11a). These differences may be underestimated as the most recent common ancestor of cancers likely predates that of normal crypts. The high mutation burdens and associated mutational signatures of DNA mismatch repair deficiency and/or polymerase ε/δ mutations were not found in any normal colorectal crypts but are present in ~20% colorectal cancers. Equally striking is the difference between the 0-4 structural changes per normal crypt (with the majority having none (Supplementary Results 1)) and the 10s to 100s per colorectal cancer[43]. In all these respects, the genomes of normal crypts with driver mutations were similar to those of normal crypts without drivers (Extended Data Fig. 9).

There was no difference in the burden either of sporadic or ubiquitous mutational processes between the crypts of individuals with and without a colorectal cancer (Supplementary Results 1). If differences in mutational processes in normal cells do underlie why some people develop colon cancer and others do not, these mutational processes must affect only a small proportion of crypts in the colon, or only exert subtle effects on the mutation rate such that we could not detect differences between the two groups. The increased base substitution and indel mutation loads in cancers are due to a combination of higher burdens of the ubiquitous mutational signatures found in normal crypts, additional signatures thus far found exclusively in cancers (confirming previous reports[5,44]) and larger numbers of copy number changes and structural variation (Extended Data Figure 11a). The causes of some of these additional mutations in cancer are known (for example, defective mismatch repair and polymerase ε/δ mutations) but the majority are uncertain.

The relative frequencies of mutated cancer genes differ between colorectal adenomas/carcinomas and normal colorectal cells ($p=0.003$, Supplementary Results 1, Extended Data Figure 11a). In colorectal cancer, mutations in *APC*, *KRAS* and *TP53* are common[37], accounting for 56% of base substitution and indel drivers (Supplementary Methods) but are comparatively rare among normal crypts with driver mutations (1/14). By contrast, mutations in, for example, *ERBB2* and *ERBB3* are common in normal crypts with drivers (5/14) but rare in colorectal cancer (7/631). In the case of *APC* (but not *KRAS* and perhaps not *TP53*), biallelic inactivation may be required to confer a strong growth advantage, which helps to explain why *APC* may be mutated less frequently in normal colon

than *ERBB2/3* that require a single hit to do so. The results suggest that mutations in *APC*, *KRAS* and *TP53* confer higher likelihoods of conversion to adenoma and carcinoma than mutations in *ERBB2* and *ERBB3* whereas the latter confer higher likelihoods of stem cells colonising crypts. There was no detectable difference in the frequency of driver mutations between individuals in our cohort who had colorectal cancer relative to those who did not (Supplementary Results 1).

Discussion

This study has characterised all classes of somatic mutation in hundreds of normal colorectal epithelial stem cells. Our experimental design allows us to gain insights into different facets of the earliest stages of the clonal evolution of colorectal cancers, namely the range of mutational processes, the frequency of driver mutations, and the clonal dynamics of colonic stem cells.

A substantial repertoire of base substitution and indel mutational processes is operative, some ubiquitous and some sporadic, together with relatively infrequent copy number changes and genome rearrangements. APOBEC DNA-editing occurs in normal colon, albeit only in rare cells. Many signatures, however, are of unknown aetiology and some appear to be acquired early in life. The presence of five times the age-standard mutation load in all colorectal cells, and potentially many other tissues, in an individual who had undergone chemotherapy provides new insight into the impact of such exposures and raises questions pertaining to its relationship with chemotherapy's relatively modest impact on cancer risk[45].

The earliest stages of colorectal cancer development have been revealed in this manuscript. They are characterised by numerous crypts carrying driver mutations, of which only a very small fraction ever manifests as macroscopic neoplasms. Certain mutated cancer genes appear to foster this pervasive and invisible wave of microneoplastic change whereas others particularly engender progression to colorectal adenoma and cancer. The conversion of these early microneoplasms to more advanced stages of colorectal neoplasia is associated with acquisition of elevated mutational loads, composed of base substitutions, indels, structural variants and copy number changes. More extensive studies of colorectal epithelium will enable characterisation of the rarer intermediate stages between these early clones and small adenomas, and refine understanding of the development of the subset of microneoplasms with higher likelihoods of becoming carcinomas.

The proportion of normal colorectal epithelial cells with driver mutations (1%) is, however, substantially lower than that of other normal tissues so far studied, notably skin (30%)[10], oesophagus (>50%)[16]. This may be due, at least in part, to the modular structure of glandular epithelia. The small number of stem cells within a crypt diminishes the probability that a cell with a driver mutation will outcompete its wild-type neighbours. Moreover, even if it does colonise the crypt, a mutant stem cell is entombed in it unless it can overcome the largely unknown forces that govern clonal expansion through crypt fission. The lower driver burden of colon relative to endometrium[11,15], which is also glandular, remains to be explored.

Fundamental questions are being addressed with respect to differences in cancer incidence rates between tissues. The somatic mutation burden in colon and ileum is similar despite the substantially higher cancer incidence rate in colon (as previously noted[4]) and therefore does not appear to account for this difference. Whether the total burden of microneoplastic change

across the colon and in other tissues more closely correlates with these differences is yet to be determined.

Finally, this study provides a reference perspective on the mutational signatures and driver mutations in normal colon against which disease states of inflammatory, genetic, neoplastic, degenerative and other aetiologies can be compared. Similar surveys conducted across the range of normal cell types will inform on the universal process of somatic evolution in the human body in health and disease.

ACKNOWLEDGEMENTS

This work was supported by the Wellcome Trust. We thank Paul Scott, Jo Fowler, David Fernandez-Antoran, and Yvette Hooks for their advice with histology and laser capture microdissection, and Moritz Gerstung for his advice on statistics. We thank the Sanger Institute Research and Development Facility for their help sequencing microbiopsies. The authors would like to thank the staff of WTSI Sample Logistics, Genotyping, Pulldown, Sequencing and Informatics facilities for their contribution. We further thank Krishnaa Mahbubani, Rogier ten Hoopen, Cinzia Scarpini, and the Phoenix study team of Nicola Grehan, Irene Debiram-Beecham, Jason Crawte, Tara Nukcheddy Grant, Pierre Lao-Sirieix, and Andy Hindmarsh for their help with sample collection. Access to transplant organ donor samples was provided by the Cambridge Biorepository for Translational Medicine. Ayesha Noorani was funded through an MRC Clinical Research Fellowship. The autopsy cohort was funded through this and an MRC core grant (RG84369) and an NIHR Research Professorship (RG67258) to Rebecca Fitzgerald. We thank the Human Research Tissue Bank, which is supported by the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre, from Addenbrooke's Hospital. Additional infrastructure support was provided from the CRUK funded Experimental Cancer Medicine Centre in Cambridge. Finally, we thank all the individuals who contributed samples to this study.

AUTHOR CONTRIBUTIONS

MRS and HLS designed the study and wrote the manuscript with contributions from all the authors. KSP, NC, MZ, RCF, NG, FT, AN, MG, and LM recruited patients and obtained samples. PE, RO, HLS, and LM devised the protocol to laser capture microdissect and sequence colonic crypts. HLS prepared sections, microdissected, and lysed colonic crypts. PR contributed to laser capture microdissection. PE and CA made libraries. HLS performed most of the data curation and statistical analysis. SO estimated the rate of crypt fission. MAS devised filters for substitution calling. JW performed in-house NMF signature extraction. TC and PR contributed to statistical analyses. LON provided technical assistance. PJC and IM oversaw statistical analyses. MRS supervised the study.

COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

1. Alexandrov, L.B. et al. The repertoire of mutational signatures in human cancer. Preprint at: https://www.biorxiv.org/content/early/2018/05/15/322859 (2018).
2. Sabarinathan, R. et al. The whole genome panorama of cancer drivers. Preprint at: https://www.biorxiv.org/content/early/2017/12/23/190330 (2017).
3. Fearon E.R. & Vogelstein B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-767 (1990).
4. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
5. Roerink S.F. et al., Intra-tumour diversification in colorectal cancer at the single cell level. *Nature* **556,** 457-462 (2018).
6. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
7. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
8. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
9. Lee-Six H. et al., Population dynamics of normal human blood inferred from somatic mutations. *Nature* https://doi.org/10.1038/s41586-018-0497-0 (2018).
10. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
11. Suda, K. et al. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777-1789 (2018).
12. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
13. Hoang, M.L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *PNAS* **113**, 9846-9851 (2016).
14. Nicholson, A. et al. Fixation and spread of somatic mutations in adult human colonic epithelium. *Cell Stem Cell* **22**, 909-918 (2018).
15. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. Preprint at: https://www.biorxiv.org/content/10.1101/505685v1(2018).
16. Martincorena, I. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911-917 (2018).
17. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
18. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
19. McKerrell, T. et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep.* **10**, 1239–1245 (2015).
20. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477-2487 (2015).
21. Potten, C.S. et al. Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut.* **33**, 71-78 (1992).
22. Cheng, H. & Leblond, C.P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine. V. Unitarian Theory of the origin of the four epithelial cell types. *Am J Anat.* **141**, 537-561 (1974).
23. Lopez-Garcia, C. et al. Intestinal stem cell replacement follows a pattern of neutral drift. *Science* **330**, 822-825 (2010).

24. Snippert, H.J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134-144 (2010).
25. Griffiths D.F. et al. Demonstration of somatic mutation and colonic crypt clonality by X-linked enzyme histochemistry. *Nature* **333**, 461-463 (1988).
26. Winton, D.J., and Ponder, B.A. Stem-cell organization in mouse small intestine. *Proc. Biol. Sci.* **241**, 13-18 (1990).
27. Kozar, S. et al. Continuous clonal labeling reveals small numbers of functional stem cells in intestinal crypts and adenomas. *Cell Stem Cell* **13**, 626-633 (2013).
28. Barker, N. et al. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608-611 (2009).
29. Rouhani, F.J. et al. Mutational history of a human cell lineage form somatic to induced pluripotent stem cells. *PLoS Genet.* **12**, e1005932 (2016).
30. Viel, A. et al. A specific mutational signature associated with DNA 8-Oxoguanine persistence in MUTYH-defective colorectal cancer. *EBioMedicine* **20**, 39-49 (2017).
31. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415-421 (2013).
32. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).
33. Chan, K. et al. An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet.* **47**, 1067-1072 (2015).
34. Boot, A. et al. Mutational signature analysis of Asian OSCCs reveals novel mutational signature with exceptional sequence context specificity. Preprint at: https://www.biorxiv.org/content/early/2018/07/19/368753.1 (2018).
35. Haradhvala, N. J. et al. Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538-549 (2016).
36. Wolf, J. et al. Peripheral blood mononuclear cells of a patient with advanced Hodgkin's lymphoma give rise to permanently growing Hodgkin-Reed Sternberg cells. *Blood* **87**, 3418-3428 (1996).
37. The Cancer Genome Atlas Network, Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012)
38. Bomme, L. et al. Cytogenetic analysis of colorectal adenomas: karyotypic comparisons of synchronous tumors. *Cancer Genet Cytogenet* **106(1)**:66-71 (1998).
39. Andersen, C.L. et al. Frequent occurrence of uniparental disomy in colorectal cancer. *Carcinogenesis* **28(1)**: 38-48 (2007).
40. Corley, D.A., et al. Variation of adenoma prevalence by age, sex, race, and colon location in a large population: implications for screening and quality programs. *Clin. Grastroenterol. Hepatol.* **11**, 172-180 (2013).
41. Cancer Research UK, Bowel Cancer Incidence Statistics, https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence#heading-Seven (Accessed August 2018).
42. Stamp, C. et al. Predominant asymmetrical stem cell fate outcome limits the rate of niche succession in human colonic crypts. *EBioMedicine* **31**: 166-73 (2018).
43. Li, Y. et al. Patterns of structural variation in human cancer. Preprint at: https://www.biorxiv.org/content/early/2017/08/27/181339 (2017).
44. Lugli, N. et al. Enhanced Rate of Acquisition of Point Mutations in Mouse Intestinal Adenomas Compared to Normal Tissue. *Cell Reports* **19**, 2185-2192 (2017).
45. Travis, L.B. Therapy-associated solid tumors, *Acta Oncologica*, **41**, 323-333 (2002).

FIGURE LEGENDS

Figure 1. Mutational signatures present in normal colon. a, an example SBS, DBS, and ID signature showing the categories into which mutations are divided. Later figures are shown in the same format. **b**, the complement of signatures in normal colonic epithelium. Known signatures are labelled according to their nomenclature in PCAWG, while novel signatures are labelled with letters. SBS, single base substitution; DBS, doublet base substitution; ID, small insertion or deletion.

Figure 2. Mutation burden *versus* age for every signature. For every signature, the median (horizontal bar) and range (vertical bar) in mutation burden for all the crypts from each individual are shown. Each individual is coloured differently. n=445 crypts from 42 individuals.

Figure 3. Crypt phylogenies. For selected individuals (**a-d**), each phylogeny is shown three times, with branch lengths proportional to SBS (top), DBS (middle), and ID (bottom) mutation counts. A stacked barplot of the signatures contributing to each branch is superimposed. The ordering of signatures along branches is for visualisation purposes. "X0" indicates mutations that could not confidently be assigned to any signature. Phylogenies for all individuals are shown in Extended Data Fig. 6. Selected phylogenies are dominated by ubiquitous signatures (**a**), SBSA and IDA (**b**), SBSB, DBS8, and IDB (**c**), or SBSD for the individual exposed to chemotherapy (d).

Figure 4. An *AXIN2* inactivating mutation. (**a**) a section after dissection. Red dots represent crypts with the *AXIN2* mutation, blue dots those without it. Crypts without dots failed sequencing. (**b**) crypts with the mutation appeared no different to others. (**c**) CNN-LOH of one crypt over the *AXIN2* locus. The copy number state (y axis) for every chromosome is shown, with one allele coloured red and the other green. (**d**) Jbrowse image of reads supporting the *AXIN2* mutations. The mutation is red. 25/29 reads support it in the crypt with CNN-LOH; the four that do not presumably represent stromal contamination.

EXTENDED FIGURE LEGENDS

Extended Data Figure 1. Laser capture microdissection of crypts. (**a**) a representative image of a section of colonic tissue, with a magnified inset showing the section before and after dissection of a crypt. (**b-c**), the coverage of crypts that underwent whole genome (**b**) and targeted (**c**) sequencing. (**d-e**), their respective VAF (which is half of the clonal fraction). (**f-g**), substitutions (**f**) and indels (**g**) removed by filtering steps and their mutational spectrum, arranged as in Figure 1.

Extended Data Figure 2. HDP signature extraction results. Results of signature extraction using an HDP with conditioning on signatures known to be active in colorectal cancer. For each signature, the extracted signature and the profile of a sample that has a strong contribution of that signature are shown. Signatures are presented as in Fig. 2. The HDP extraction was followed by deconvolution by Expectation Maximisation (Methods, Extended Data Fig. 3) to produce the version of signatures presented in the main text. HDP, Hierarchical Dirichlet Process.

Extended Data Figure 3. Expectation maximisation decomposition of HDP signatures. Three signatures were decomposed. For each panel, the original HDP version in shown on the top left, the PCAWG signatures that are deemed to contribute at least 10% of mutations to it on the right, and the reconstituted signature built by combining the PCAWG signatures on the bottom left. The cosine similarity of the reconstituted signature to the original is shown in the title to the reconstituted signature plot. HDP, Hierarchical Dirichlet Process; PCAWG, Pan Cancer Analysis of Whole Genomes.

Extended Data Figure 4. Validation of single base substitution signatures. Other methods of signature extraction were run to test the robustness of signature decomposition. **a**, HDP without pre-conditioning on PCAWG. **b**, In-house NNMF without pre-conditioning on PCAWG. **c**, NNMF implemented by the MutationalPatterns R package (Methods). HDP, Hierarchical Dirichlet Process; PCAWG, Pan Cancer Analysis of Whole Genomes; NNMF, Non-Negative Matrix Factorisation.

Extended Data Figure 5. Linear modelling of signature accumulation. For signatures that appeared to show a linear accumulation with age, the mutation rate per site was determined using mixed models, with age and site as fixed effects, and individual as a random effect. Confidence intervals were determined by bootstrapping. n=445 crypts from 42 individuals. Solid lines represent the mean slope of the regression and shaded areas its 95% confidence intervals.

Extended Data Figure 6. Crypt phylogenies. (**a-ap**) For every individual, the phylogeny of crypts is shown three times: on top, with branch lengths proportional to the number of single base substitutions; in the middle, with branch lengths proportional to the number of doublet base substitutions; on the bottom, with branch lengths proportional to the number of small insertions and deletions. Scale bars are shown on the right-hand side. A stacked barplot of the mutational signatures that contribute to each branch is overlaid over every branch. "X0" indicates mutations that could not confidently be assigned to any signature. Please note that the ordering of signatures along a given branch is just for visualisation purposes: we cannot distinguish the timing of different signatures along a branch. (**aq**), Cumulative burden of SBSA (top panel) and SBSB (bottom panel) is plotted relative to the cumulative burden of SBS1 in order to time these mutational processes throughout life. Informative clades are shown (from patients with labelling as in the other panels), with every node and tip of the clade plotted in the space of the cumulative number of mutations due to a given signature that have occurred up until that node in the tree. Lines represent the branching structure of the tree.

Extended Data Figure 7. Copy number changes and structural variants in normal colon. 449 crypts had sufficient coverage to be evaluated. (**a**) whole chromosome amplifications in five crypts. The copy number state (y axis) for each allele, one coloured red, and one coloured green, is shown. Chromosomes are labelled along the top of the graph. (**b**) timing of copy number changes throughout life. Vertical bars represent 95% confidence intervals determined by bootstrapping. Horizontal bars represent the most likely time of the copy number change, as defined by mutationTimeR (Supplementary Results 1) (**c**), crypts with loss of heterozygosity. For each chromosome with an LOH event, the copy number across the whole chromosome is shown in the top part of the panel, with the total copy number in black and the minor allele copy number in blue. Underneath are shown example single nucleotide polymorphisms that support the LOH. In each case, reads from the crypt in question are shown above, and reads from its matched normal are shown underneath. Thus in

the first image, the wild-type state (below) is heterozygous for a T (red) single nucleotide polymorphism, whereas in the crypt in question (above), this polymorphism has now become homozygous. Small deviations from a fully homozygous state are likely due to stromal contamination. (**d**) reads supporting structural variants in normal colon are shown.

Extended Data Figure 8. Gain of function driver mutations in normal colon. Putative driver missense mutations in oncogene hotspots. The number of substitutions catalogued in COSMIC are shown on the y axis at each position along the gene, with the mutations observed in our cohort highlighted.

Extended Data Figure 9. Occurrence matrix of signatures and driver mutations in crypts. For all crypts that were whole genome sequenced to sufficient depth and for crypts that underwent targeted sequencing and in which driver mutations were found, the signatures and driver mutations are shown. Each vertical column represents a crypt. The individual to which each crypt belongs is indicated by alternating colours in the top bar. The site to which each crypt belongs is shown underneath. The matrix is coloured by contribution of each signature to each crypt, normalised for each signature; thus the crypt with the largest contribution of a given signature is coloured purple, and the crypt with the smallest contribution is coloured white. Crypts in which the signatures could not be assessed, either because they underwent targeted sequencing or the coverage was poor, are coloured grey. Driver mutations, including heterozygous mutations in tumour suppressor genes, are indicated by a black bar.

Extended Data Figure 10. Stem cell dynamics of normal colon (Supplementary Results 2). (**a**) stem cell number and stem cell replacement rate in normal human colonic crypts, as estimated by approximate Bayesian computation. Every point represents a simulation. Points are coloured according to their similarity to the observed data: the most similar 0.1% are coloured dark red, and so on, until the least similar simulations are blue. (**b**) approximate Bayesian computation of the crypt fission rate in the human colon. The prior distribution of the crypt fission rate used to simulate many biopsies of the colon is shown above. The unit for the crypt fission rate is fissions per crypt per year. The posterior distribution of the crypt fission rate parameter estimated by neural network regression on the simulations is shown below. (**c-d**), evidence of crypt fusion in human colon. In each figure, above is shown a phylogeny that depicts the genetic relationships between selected crypts. Dotted blue lines show low allele fraction mutations that are shared between crypts in a manner that is incompatible with phylogeny that is dictated by the clonal mutations. Underneath each crypt in the phylogeny is an image depicting its position in the section. Sections are labelled according to their z-stacked order. Beneath this, the allele fraction of mutations on each branch of the phylogeny in each crypt is shown. To the right hand side is the trinucleotide context of the mutations that occurred on the branch.

Extended Data Figure 11. Comparison of the mutational signatures and driver landscape of normal crypts and colorectal adenocarcinomas. (**a**) a comparison of the burden of mutations due to every mutational signature found in either group. For each signature, the (mutation burden+1) of every sample is shown on the y axis on a log scale. Normal colon and cancer samples are ordered within their groups. Colorectal adenocarcinoma signature attributions and burden are from Alexandrov et al.[1]. 60 cancers are compared with 472 normal crypts. (**b**) the proportion of driver mutations in each gene in normal colon and colorectal cancer. The frequency of driver mutations in cancer was derived using data from The Cancer Genome Atlas Network[43] (Supplementary Methods).

SUPPLEMENTARY METHODS

Human tissues

We obtained healthy colonic biopsies from four cohorts (Supplementary Table 1). The first represents seven deceased organ donors ranging in age from 36 to 67, from whom colonic and small intestinal biopsies were taken at the time of organ donation (REC 15/EE/0152). The second represents individuals aged 60 to 72 who were having a colonoscopy following a positive faecal occult blood test as part of the Bowel Cancer Screening Programme (Ethical approval 08-H0308-13); we selected 16 who were not found to have either an adenoma or a carcinoma on colonoscopy, and 15 who were found to have a colorectal carcinoma (the normal biopsies that we use were distant from these lesions). The third cohort represents three paediatric patients who underwent routine colonoscopy to exclude inflammatory bowel disease and who were found to have a completely normal intestinal mucosa macroscopically and histologically (REC 12/EE/0482). The final cohort included one 78 year-old gentleman with oesophageal cancer who underwent a warm autopsy (REC 13/EE/0043). This gentleman had been treated with palliative chemotherapy of Epirubicin, Oxaliplatin and Capecitabine within the three months before the autopsy; given that monoclonal conversion within crypts is on the order of years, mutations due to these chemotherapies are likely to be private to a small proportion of stem cells per crypt and so are unlikely to be detected. All samples were obtained with informed consent and studies approved by East of England Research Ethics Committees.

Laser capture microdissection of colonic crypts

Fresh frozen biopsies were embedded in optimal cutting temperature (OCT) compound. 30 micrometre sections were fixed in methanol for five minutes, washed three times with phosphate-buffered saline, and stained with Gill's haematoxylin for 20 seconds. Crypts were isolated by laser capture microdissection, and collected in separate wells of a 96-well plate. They were lysed using the Arcturus PicoPure Kit (Applied Biosystems) according to the manufacturer's instructions. DNA library prep then proceeded without clean-up or quantification.

Library preparation

Two library preparation methods were used for laser capture microdissected (LCM) material: in initial experiments sonication was used to fragment DNA, and later, an enzymatic fragmentation method was implemented as it could make libraries from even lower input. Comparison of the two methods showed no difference in mutation calls once post-processing filters (described below) had been implemented. All samples in this study were processed using an Agilent Bravo Workstation (Option B; Agilent Technologies).

For sonication libraries, LCM lysate (20 μl) was mixed with 100 μl TE buffer (Ambion; 10 mM Tris-HCl, 1 mM EDTA) and DNA was fragmented using focused acoustics (Covaris LE220; Covaris, Inc.). Fragmented DNA was mixed with 80 μl Ampure XP beads (Beckman Coulter). Following a 5 min binding reaction and magnetic bead separation, genomic DNA was washed twice with 75% ethanol. Beads were resuspended in 20 μl nuclease-free water (Ambion) and processed immediately for DNA library construction. Each sample (20 μl) was mixed with 2.8 μl of NEBNext Ultra II End Prep Reaction Buffer, 1.25 μl of NEBNext Ultra

II End Prep Enzyme Mix (New England BioLabs) and incubated on a thermal cycler for 30 min at 20°C then 30 min at 65°C. Following DNA fragmentation and A-tailing, each sample was incubated for 20 min at 20°C with a mixture of 30 µl ligation mix and 1 µl ligation enhancer (New England BioLabs), 0.9 µl nuclease-free water (Ambion) and 0.1 µl duplexed adapters (100 uM; 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3', 5'-phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'). Adapter-ligated libraries were purified using Ampure XP beads by addition of 65 µl Ampure XP solution (Beckman Coulter) and 65 µl TE buffer (Ambion). Following elution and bead separation, DNA libraries (21.5 µl) were amplified by PCR by addition of 25 µl KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 1 µl PE1.0 primer (100 µM; 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3') and 2.5 µl iPCR-Tag (40 µM; 5'-CAAGCAGAAGACGGCATACGAGATXGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC-3') where 'X' represents one of 96 unique 8-base indexes The sample was then mixed and thermal cycled as follows: 98 °C for 5 min, then 12 cycles of 98 °C for 30 s, 65°C for 30 s, 72 °C for 1 min and finally 72 °C for 5 min. Amplified libraries were purified using a 0.7:1 volumetric ratio of Ampure Beads (Beckman Coulter) to PCR product and eluted into 25 µl of nuclease-free water (Ambion). DNA libraries were adjusted to 2.4 nM and sequenced on the HiSeq X platform (illumina) according to the manufacturer's instructions with the exception that we used iPCRtagseq (5'-AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC-3') to read the library index.

For enzymatic fragmentation, LCM lysate (20 ul) was mixed with 50 ul Ampure XP beads (Beckman Coulter) and 50 µl TE buffer (Ambion; 10 mM Tris-HCl, 1 mM EDTA) at room temperature. Following a 5 min binding reaction and magnetic bead separation, genomic DNA was washed twice with 75% ethanol. Beads were resuspended in 26 µl TE buffer and the bead/genomic DNA slurry was processed immediately for DNA library construction. Each sample (26 µl) was mixed with 7 µl of 5X Ultra II FS buffer, 2 µl of Ultra II FS enzyme (New England BioLabs) and incubated on a thermal cycler for 12 min at 37°C then 30 min at 65°C. Following DNA fragmentation and A-tailing, each sample was incubated for 20 min at 20°C with a mixture of 30 µl ligation mix and 1 µl ligation enhancer (New England BioLabs), 0.9 µl nuclease-free water (Ambion) and 0.1 µl duplexed adapters (100 uM; 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3', 5'-phos-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG-3'). Adapter-ligated libraries were purified using Ampure XP beads by addition of 65 µl Ampure XP solution (Beckman Coulter) and 65 µl TE buffer (Ambion). Following elution and bead separation, DNA libraries (21.5 µl) were amplified by PCR by addition of 25 µl KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 1 µl PE1.0 primer (100 µM; 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T-3') and 2.5 µl iPCR-Tag (40 µM; 5'-CAAGCAGAAGACGGCATACGAGATXGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATC-3') where 'X' represents one of 96 unique 8-base indexes The sample was then mixed and thermal cycled as follows: 98 °C for 5 min, then 12 cycles of 98 °C for 30 s, 65°C for 30 s, 72 °C for 1 min and finally 72 °C for 5 min. Amplified libraries were purified using a 0.7:1 volumetric ratio of Ampure Beads (Beckman Coulter) to PCR product and eluted into 25 µl of nuclease-free water (Ambion). DNA libraries were adjusted to 2.4 nM and sequenced on the HiSeq X platform (Illumina) according to the manufacturer's instructions with the exception that we used iPCRtagseq (5'-AAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTC-3') to read the library index.

Whole genome sequencing

We generated paired end sequencing reads (150bp) using Illumina XTEN® machines resulting in ~15x coverage per sample. In 94% of the whole genome crypts included for statistical analysis, over 90% of the callable genome was covered by more than 10 reads. Sequences were aligned to the human reference genome (NCBI build37) using BWA-MEM.

Targeted sequencing

A 2.3 MB capture panel was designed in-house to pull down genes that are known or suspected to play a role in neoplasia. We performed custom RNA bait design following the manufacturer's guidelines (SureSelect, Agilent). Samples were multiplexed on flow cells and subjected to paired end sequencing (75-bp reads) using Illumina HiSeq2000 machines. One 96-well plate of samples was sequenced on each lane, but as tissue recovery was variable, a range of coverage was achieved. Sequences were aligned to the human reference genome (NCBI build37) using BWA-align.

Data Availability

Whole genome and targeted sequencing data are deposited in the European Genome Phenome Archive (EGA) with EGA accession EGAD00001004192 and EGAD00001004193. Images of microdissections and the physical distances between crypts are available on Mendeley Data by searching for the title of this article. All other data is available from the authors on request.

Code Availability

Code for statistical analyses is provided as part of the supplement. Custom R scripts and their input data for signature analysis are available on GitHub at https://github.com/HLee-Six/colon_microbiopsies. All other code is available from the authors on request.

Calling substitutions

Substitution calling was broken down into three steps: mutation discovery; filtering to produce a list of clean sites; and genotyping, where the presence or absence of every mutation in every sample is evaluated.

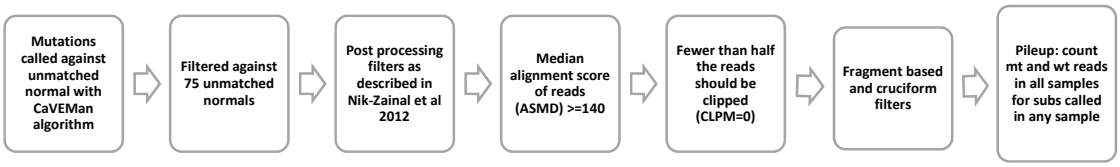
First, mutations were initially discovered using the Cancer Variants through Expectation Maximisation (CaVEMan) algorithm[46]. CaVEMan uses a naïve Bayesian classifier to derive the probability of all possible genotypes at each nucleotide. CaVEMan copy number options were set to major copy number 5 and minor copy number 2 for normal clones, as in our experience this maximises sensitivity. The algorithm was run using an unmatched normal in order to be able to derive phylogenies: had another sample from the same individual been treated as a matched normal, early embryonic mutations would have been treated as germline and discarded, resulting in incorrect trees.

Second, a number of post-processing filters were applied (Extended Data Figure 2). These included filtering against a panel of 75 unmatched normal samples to remove common single nucleotide polymorphisms, post-processing as described previously[32] and two filters (only applied to whole genome sequencing data) designed to remove mapping artefacts associated with BWA-MEM: the median alignment score of reads supporting a mutation should be greater than or equal to 140, and fewer than half of these reads should be clipped. The library preparation protocol for microbiopsies produced shorter library insert sizes than standard methods. Reads could therefore overlap, resulting in double counting of mutant reads. Fragment-based statistics were generated to prevent the calling of variant supported by a low

number of fragments. Variants were annotated by ANNOVAR[47] and fragment-based statistics (fragment coverage, number of fragments supporting the variant, fragment-based allele fraction) were calculated for each variant after the exclusion of marked PCR duplicates. In the rare event of discordance in the called base at the variant position between overlapping paired-end reads, the base with the highest quality score was selected. Fragment-based statistics were calculated separately for high quality fragments (alignment score ≥ 40 and base scores ≥ 30). Variants supported by at least three high quality fragments were retained and used for the next stage of variant filtering. Inspection of variants specific to LCM experiments revealed that the vast majority were present within inverted repeats capable of forming hairpin structures, that they were supported by reads with very similar alignment start position (and so not marked as PCR duplicates), and were primarily located close to the alignment start within the supporting reads. Commonly these variants coincided with other proximal variants (1-30 bp), but filtering based on variant proximity would also remove actual kataegis events. *In silico* modelling of the potential hairpin showed that the variants were aligning to each other in the stem of the structure, but could not form a base pair, while all other bases could. The artefacts are likely the consequence of erroneous processing of cruciform DNA (existing either prior to DNA isolation or formed during library preparation) by the enzymatic digestion protocol applied. We have considered modelling the hairpin structures to filter these variants, but given the fact that read clustering (i.e., similar alignment position) serves as a hallmark for these artefacts, we opted to use the proximity of the variant to the alignment start, and the standard deviation (SD) and median absolute deviation (MAD) of the variant position within the supporting reads, as features for filtering. These statistics were calculated separately for positive and negative strand aligned reads. In case the variant was supported by a low number of reads (i.e., 0-1 reads) for one of the strands, the filtering was based only on the statistics generated for the other strand. Per variant, if one of the strands had too few reads supporting, it was required for the other strand that either: (I) there should be ≤ 90% supporting reads to report the variant within the first 15% of the read starting from the alignment start, or (II) the statistics MAD > 0 and SD > 4. Per variant, if both strands were supported by sufficient reads it was required for both strands separately that either: (I) there should be ≤ 90% supporting reads to report the variant within the first 15% of the read, (II) the statistics MAD > 2 and a SD > 2, or (III) that the other strand should have the statistics MAD > 1 and SD > 10 (i.e., the variant is retained if the other strand demonstrates strong measures of variance). In our experience, the proposed strategy vastly reduces the number of artefactual variants while retaining all other variants, as assessed by running the last filtering step on WGS data from non-LCM experiments.

Third, mutations were genotyped in every sample. A pileup of all the samples from a given individual was constructed, counting the number of mutant and wild type reads in every sample over every site that had been called in any sample from that person. Only reads with a mapping quality of 30 or above and bases with a base quality of 30 or above were counted. After applying these filters, mutations were genotyped based on the number of mutant and wild type reads at each locus. Mutations were called based on a variant allele fraction (VAF) > 0.2, a depth > 7, and at least 4 mutant reads. If the depth over a locus was less than seven in a given sample, or if there was more than one mutant read but the other criteria were not met, the genotype was set to NA for tree construction purposes. Loci that were set to NA in more than one third of the samples were removed for construction of the phylogeny. Positions were called as germline if they were either called as present or NA in all of the samples from a given individual.

1.2% of all mutations were present in the coding regions of the genome. All mutations in coding regions are provided (Supplementary Table 3).



Calling short insertions and deletions (indels)

As for substitutions, calling of indels was broken down into mutation discovery, filtering, and genotyping. Mutations were called with the Pindel algorithm[48] using an unmatched normal. Post processing filters were applied as in Nik-Zainal et al.[32], and the number of mutant and wild-type reads was tabulated as above. The same dataset-specific filters were applied as for substitutions. Indels were then genotyped based on a VAF>0.2, a depth of at least 10, and support of at least 5 mutant reads.

Calling structural variants

Genomic rearrangements were called using the BRASS algorithm (https://github.com/cancerit/BRASS). Abnormally paired read pairs from WGS were grouped and filtered by read remapping. Read pair clusters with ≥50% of the reads mapping to microbial sequences were removed, as were rearrangements where the breakpoint could not be reassembled. Candidate breakpoints were matched to copy number breakpoints defined by ASCAT (see below) within 10kb. Only structural variants where the two breakpoints were more than 1000 base pairs apart were considered. Structural variants were called against a matched normal skin or blood sample when available and against another crypt from the same individual with good coverage when not.

Calling copy number

Copy number changes were called using the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm[49]. The same matched normal sample was used as for calling structural variants. For additional validation of copy number changes in normal colon, the QDNAseq algorithm[50] was run. ASCAT uses both the read depth and ratios of heterozygous single nucleotide polymorphisms to determine an allele-specific copy number, while the QDNAseq relies solely on variations in sequencing coverage. To call amplifications and deletions in the colonic microbiopsy cohort, only those that were both called by ASCAT and showed a clear departure from the background log2ratio by QDNAseq were retained. To call copy neutral loss of heterozygosity in this cohort, all such events called by ASCAT were checked visually on Jbrowse[51] to verify an imbalance of parental snps. Only crypts with >10X coverage, for which copy number changes could be reliably detected, were used.

Detection of driver variants and positive selection

Driver mutations were detected both through an unbiased dNdS method and through manual annotation. For these analyses, the CaVEMan and Pindel calls were used without post-processing filters (such as requiring a VAF cutoff of >0.2) in order to maximise our sensitivity. All putative driver variants were visually inspected using Jbrowse[51], and so we could afford a higher false positive rate in the mutation discovery phase.

dNdScv[52] was used to conduct three tests: first, using only the whole genome sequencing data, an analysis of selection over all genes; second, using combined whole genome and targeted sequencing data, over all the genes covered by the bait-set; and finally, using again this combined dataset, over 90 selected cancer genes (Supplementary Table 4, Supplementary Results 1).

Manual annotation of driver variants based on prior knowledge complemented this. A list of 90 colorectal cancer genes (appendix) curated from the literature that were also covered by the bait-set were intersected with the list of substitutions and indels from combined whole genome and targeted sequencing. Mutations were annotated as putative drivers if they were either missense mutations that fell in an oncogene hotspot (based on visualisation of the distribution of mutations in the gene on COSMIC[53]), or if they were truncating mutations that fell in a tumour suppressor gene.

Structural variants that might act as drivers were assessed by intersection of genes involved in each structural variant with the twelve genes involved in gene fusions that have been reported in colorectal cancer in COSMIC (*VTI1A, TCF7L2, TPM3, NTRK1, PTPRK, RSPO3, ETV6, NTRK3, EIF3E, RSPO2, C2orf44*, and *ALK*). No fusion genes were found. None of the genes involved in structural variants in our data overlapped with the list of 90 cancer genes used for assessing substitutions and indels, and nor were there any genes that were affected by more than one structural variant. No high-level copy number amplifications were observed and there were no homozygous deletions.

Please note that the driver frequency is low, such that we cannot estimate a per-gene driver frequency. All we can do to derive a meaningful estimate is to pool our driver mutations. In addition, coverage may fluctuate even within a gene. Some portions of a gene may be well covered in 1,000 crypts, and others in 2,000 crypts. The approach that we took was to calculate, for the average exonic base pair in our 90 cancer genes, the number of crypts in which that base pair was covered by >=8 reads (for substitutions) and by >=10 reads (for indels). 64% of all bases in the targeted panel across all crypts are covered by >=8 reads, which equates to a number of callable bases equivalent to having sequenced ~1,400 crypts with perfect coverage over every base in every crypt. This average number of crypts in which all base pairs achieve good coverage becomes the denominator for calculating the driver mutation frequency (with the number of drivers observed in the dataset as the numerator). A similar approach can be taken with indels. Our estimate of 1% uses a global correction, on the assumption that under-representation and over-representation will even itself out when estimating the total frequency of driver mutations in the whole dataset.

Estimation of frequency of driver mutations in cancer

Publically-available colorectal cancer mutation calls were obtained from The Cancer Atlas Network[37]. Driver mutations were annotated manually in the same way as in our dataset: only mutations that fell in the 90 genes that we had selected were considered, and they were annotated as putative drivers if they were either missense mutations that fell in an oncogene hotspot (based on visualisation of the distribution of mutations in the gene on COSMIC[53]), or if they were truncating mutations that fell in a tumour suppressor gene.

Construction of phylogenies

Phylogenies are used in this analysis for timing mutations. The most informative branches in this case are the long branches shared by a small number of crypts, which are very robust to all tree construction methods. Trees were built using maximum parsimony using substitutions

called as described above. For every individual, the input matrix of mutation calls was bootstrapped 100 times. Phylogenies were constructed for each replicate using the Wagner method of the Mix programme from the Phylip suite of tools[54]. The consensus phylogeny was constructed from 100 bootstrap runs using the extended majority rule method for the Consense programme from the Phylip suite of tools[54].

Across all phylogenies, a mean of 10% and a median of 1.5% of mutations per tree did not fit the trees perfectly. Phylogenies with more crypts had more mutations that fitted imperfectly. Consider a mutation that is really present in 50 crypts. Even with 15X coverage over the site in every sample, and with every crypt completely clonal, if we simulate resampling of mutant reads from the binomial distribution (with size of 15 and probability of 0.5), 17% of the time the mutation will have fewer than the 3 reads required to call it in at least one sample. Variation in sequencing depth, clonality, and sequencing errors would further decrease the probability of calling the mutation perfectly in every sample. Nodes across all our phylogenies had mean bootstrapping values of 0.77 and median bootstrapping values of 0.99. Branches at the very top of the phylogenies, likely representing embryonic cell divisions, are supported by only a few mutations and have lower support because in a given bootstrap sample the couple of mutations that support this node may be omitted. Longer shared branches almost always have bootstrapping values of 1. These longer shared branches are those that are most important to our analyses, because they are the most informative when timing mutational signatures relative to one another and because they represent postnatal crypt fission events. In order to increase further our confidence in our phylogenies, we validated them by reconstructing them with indels. To do this, the same procedure as for substitutions was followed for indel matrices. As there were fewer indels than substitutions, nodes in indel phylogenies were generally reconstructed with lower confidence than in substitution phylogenies, but they broadly agree. 85% of nodes reconstructed with >=90% confidence in the indel tree were present with exactly the same set of descendants in the substitution trees. Any errors in the phylogenies should be relatively minor and not affect our downstream analyses.

The phylogeny inference programme used provided the topology of the tree but not the assignment of mutations. Mutations from the input matrix of genotypes therefore have to be re-assigned to branches. In order to assign a set of mutation calls with no false negative and no false positives to a tree, each branch of the tree was considered in turn. If a mutation was called in all the descendants of a given branch, and in no samples that were not descendants of the branch, mutations were assigned to that branch.

Some colonic microbiopsies suffered from low coverage and stromal contamination. For this reason, we did not expect mutations to fit the tree perfectly, as a mutation that was truly present in a colony might be missed if too few supporting reads are found. Mutations were only assigned to the tree in order to determine the mutational processes active at a particular time. We reasoned that it was preferable to assign only mutations that fit the tree perfectly and adjust the branch lengths based on the power to call mutations at a given branch, rather than attempting to assign mutations that fit the tree imperfectly. Using the clonality and coverage of all descendants of a branch, the proportion of true substitutions or indels on the branch that would be first discovered (whether by CaVEMan or Pindel) and then genotyped as present according to the criteria described above was calculated. The observed branch length was then adjusted by dividing by this proportion. Adjustment proportions can be found in Supplementary Table 7. This was done for both substitutions and indels, but not for structural variants and for larger copy number changes due to a lack of data: most branches

have no large variants and so could not be extended appropriately. Rearrangements and copy number changes were assigned to phylogenies manually.

Extraction of mutational signatures

Mutational signatures were extracted using the mutations assigned to every branch of a phylogeny as a 'sample'. This allows better discrimination of mutational processes that may occur at different times within the same cell. Mutations were categorised following the method used by the Mutational Signatures working group of the Pan Cancer Analysis of Whole Genomes (PCAWG)[1]. Single base substitutions were categorised into 96 classes according the identity of the pyrimidine mutated base pair, and the base 5' and 3' to it. Doublet base substitutions were categorised into 78 classes according to the identity of the reference and alternative bases. Indels were classified according to whether they were an insertion or a deletion, the identity of the inserted/deleted base, the length of the mononucleotide tract in which they occurred, or the degree of homology with the surrounding sequence into 83 classes (Fig. 1a).

Signatures were extracted using a hierarchical Dirichlet Process[55,56]. Code and the input mutations are provided at https://github.com/HLee-Six/colon_microbiopsies. First, the algorithm was conditioned on the set of mutational signatures that have found to be operative in colorectal cancers in PCAWG[1]: SBS1, SBS2, SBS3, SBS5, SBS13, SBS16, SBS17a, SBS17b, SBS18, SBS25 (included although it is not found in colorectal cancer because the similarity with the mutational profile with crypts from one individual had been previously noted), SBS28, SBS30, SBS37, SBS40, SBS41, SBS43, SBS45, SBS49, DBS, DBS3, DBS4, DBS6, DBS7, DBS8, DBS9, DBS10, DBS11, ID1, ID2, ID3, ID4, ID5, ID6, ID7, ID8, ID10, and ID14. This allows simultaneous discovery of new signatures and matching to known ones. Nine single base substitution (SBS), two doublet base substitution (DBS), and five indel (ID) signatures were discovered (Extended Data Fig. 2). Despite pre-conditioning, signatures that were perfectly correlated in all samples were still amalgamated. This occurred, for example, with signatures 1, 5, and 18. Therefore, expectation maximisation was used to deconvolute all HDP signatures into known PCAWG signatures. If a signature reconstituted from the components that expectation maximisation extracted (only including PCAWG signatures that accounted for at least 10% of mutations in each sample to avoid over-fitting) had a cosine similarity to the HDP signature of more than 0.95, the signature was presented as its expectation maximisation deconvolution. Three HDP signatures met these criteria: the HDP SBS1 signature was deconvoluted into a mixture of PCAWG SBS1, PCAWG SBS5, and PCAWG SBS18; the HDP DBSA was deconvoluted in PCAWG DBS2, PCAWG DBS4, PCAWG DBS6, PCAWG DBS9, and PCAWG DBS11; and the HDP IDC was deconvoluted into PCAWG ID1, PCAWG ID2, and PCAWG ID5 (Extended Data Fig. 3). To test the robustness of this signature analysis, other signature extraction methods were used: HDP with no pre-conditioning, the non-negative matrix factorisation (NNMF) method used by Blokzijl and colleagues[4], and a version of the NNMF algorithm used by Alexandrov and colleagues[1]. These all produced comparable results (Extended Data fig. 4).

Timing SBSA and SBSB throughout life

Five patients had informative clades with branchpoints that allowed us to time SBSA. Plotting the cumulative amount of SBSA vs SBS1 at each node in these clades (Extended Data Fig. 6aq), we observed that for each the rate of accumulation of SBSA relative to SBS1

was high in early branchpoints, and then slowed down almost to zero on all branches but for one (a branch of patient ao, where it continues to be acquired, albeit at a slow rate).

We can take the inflexion point on the graph of cumulative SBSA vs SBS1 to be the upper limit of the point in time when SBSA slowed down. This provides an upper bound because:

1. When we observe the presence of a signature on a branch, we know that the causative process must have been active at some point during the lifetime of the branch, but we cannot say when on the branch it occurs; it might have ended long before the branch did.
2. If the time to the most recent common ancestor of the crypt is longer than 0, the age at which this stopped would be earlier.
3. If the SBS1 mutation rate is increased in early life, as it may be during the rapid growth of the embryo, the age at which the inflexion point occurs would have been younger.

Using these five informative clades, and assuming a clock-like but personalised rate of SBS1 accumulation (i.e. each patient can accumulate SBS1 at their own constant rate), we found that the upper bound of the age at which SBSA slowed was: 9.7 years (patient h); 7.1 years (patient z); 2.4 years (patient am); 20.1 years (patient aa); and 9.6 years (patient ao). There are no branches that begin after 10 years of age with a high ratio of SBSA to SBS1.

The most informative branchpoint is the earliest inflexion point; the estimate of 2.4 years from patient aa, is therefore, perhaps our best estimate. Nonetheless, we did not want to base our statement on a single patient, and so 10 years was given in the text as four patients had branches that ended before 10 years of age.

A similar argument can be made for SBSB (Extended Data Fig. XXX). For SBSB, however, only two clades were informative. The estimated upper bounds of age for SBSB activity were 2.4 years old (this was the same inflexion point as for SBSA in patient aa), and 6.4 years old (patient ai). For patient ai, a reasonable amount of SBSB is still acquired after this branchpoint. If the ratio of accumulation of SBSB vs SBS1 continued at the same rate as before this branchpoint, the number of SBSB mutations seen in the terminal branches would have been observed by age 8.2 years old.

Telomere length analysis

Telomere length was estimated from whole genome sequencing data using Telomerecat. Telomerecat is a python based software package for telomere length estimation from short-read whole genome sequence data[57]. It functions by classifying paired-end reads as either fully or partially telomeric based on the canonical hexamer TTAGGG, and uses that ratio to estimate an average telomere length. Notably, Telomerecat measures telomeres and also accounts for interstitial telomeric repeats. It is ploidy and species agnostic (assuming that the telomere hexamer is the canonical mammalian signature of $TTAGGG_n$). Telomerecat has four main stages; 1) Identification of all telomeric or partially telomeric read pairs and creation of a subsetted bam file containing only these reads 2) Classification of telomeric read-pairs into intratelomeric, boundary or junction-spanning or intrachromosomal 3)Error correction of boundary or junction-spanning read pairs 4) Estimation of telomere length based on the ratio of intratelomeric and boundary / junction spanning read pairs.

Telomerecat has been validated on whole genome DNA sequencing files from both tumour and normal samples[57]. Its results show concordance with an established method of telomere length measurement; the mean Telomere Restriction Fragment (mTRF) technique. Alternative packages are available notably: Computel[58], Telseq[59] and TelomereHunter[60]. They all have respective strengths and have been benchmarked through their methods publications. We have opted for Telomerecat as it provides a base pair resolution estimate of telomere length whilst providing correction for variations in sequencing depth in a ploidy agnostic manner.

We have run Telomerecat on 445 crypt bam files with good coverage and clonality to generate telomere length estimates. Telomerecat was threaded across 10 cores and 100 simulation cycles were requested per run. Values displayed are the median telomere length across all chromosomes in that samples measured in base pairs.

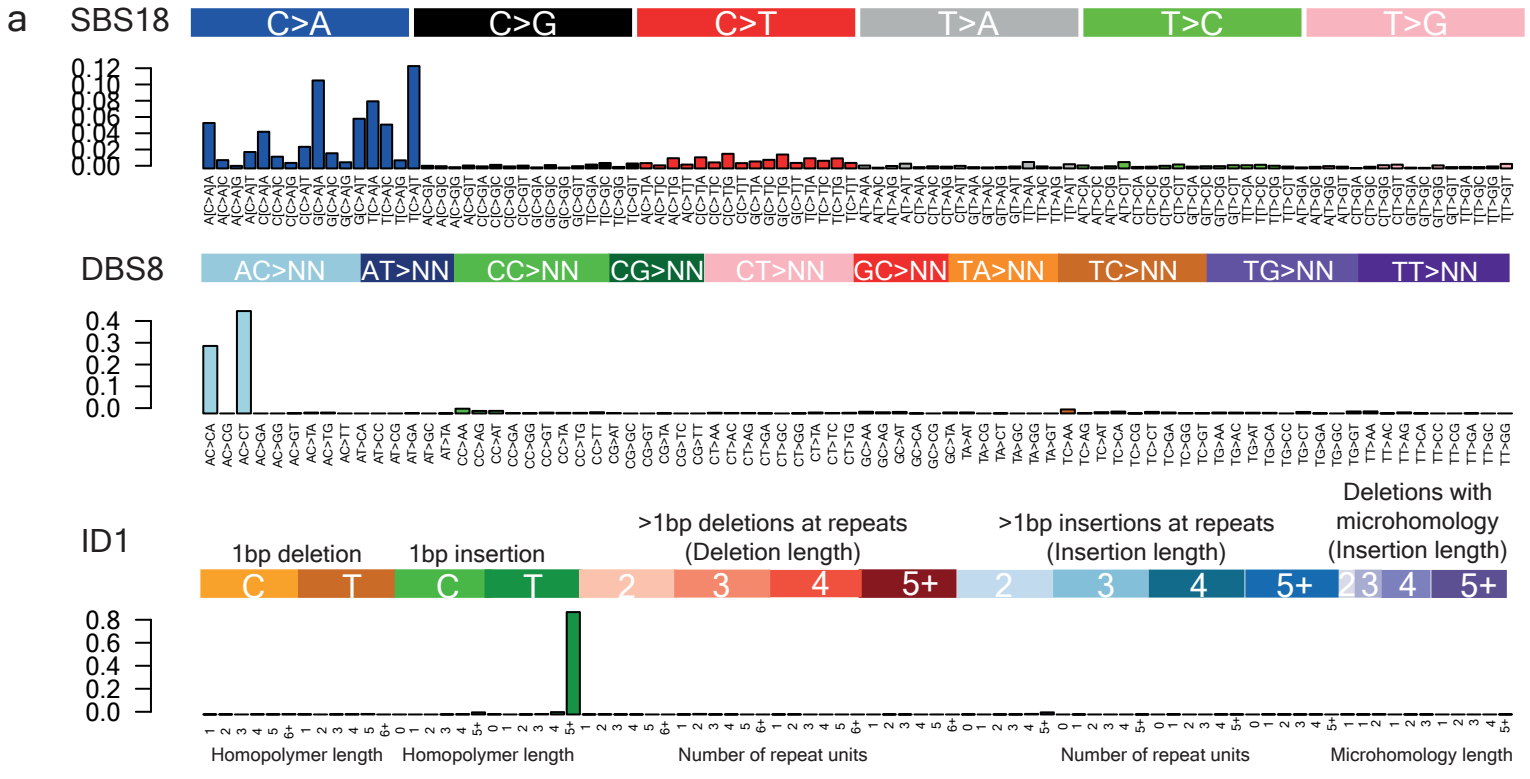
Statistical analyses

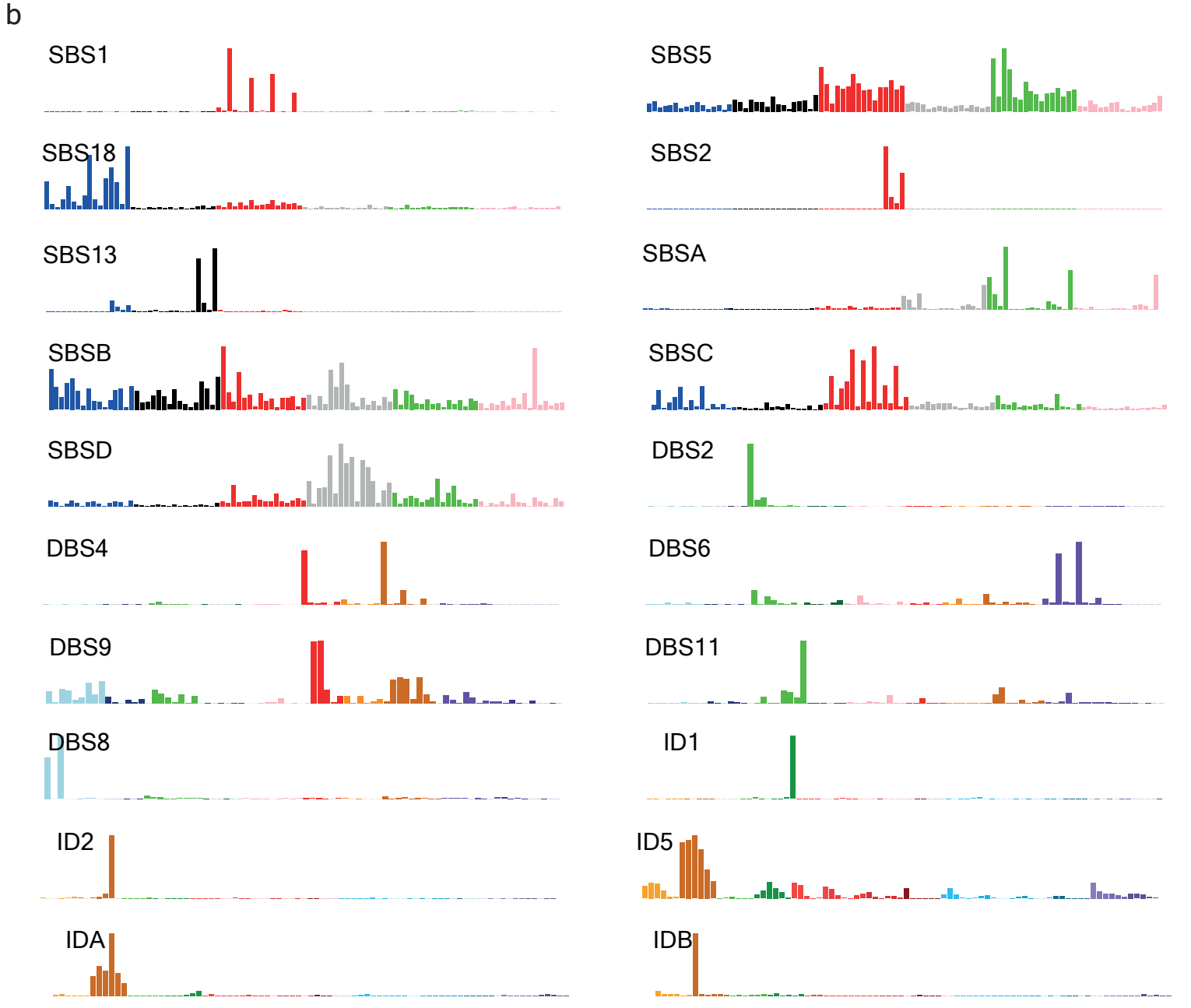
All statistical analyses were performed in R (Supplementary Results 1, Supplementary Results 2). Code can be found at https://github.com/HLee-Six/colon_microbiopsies.

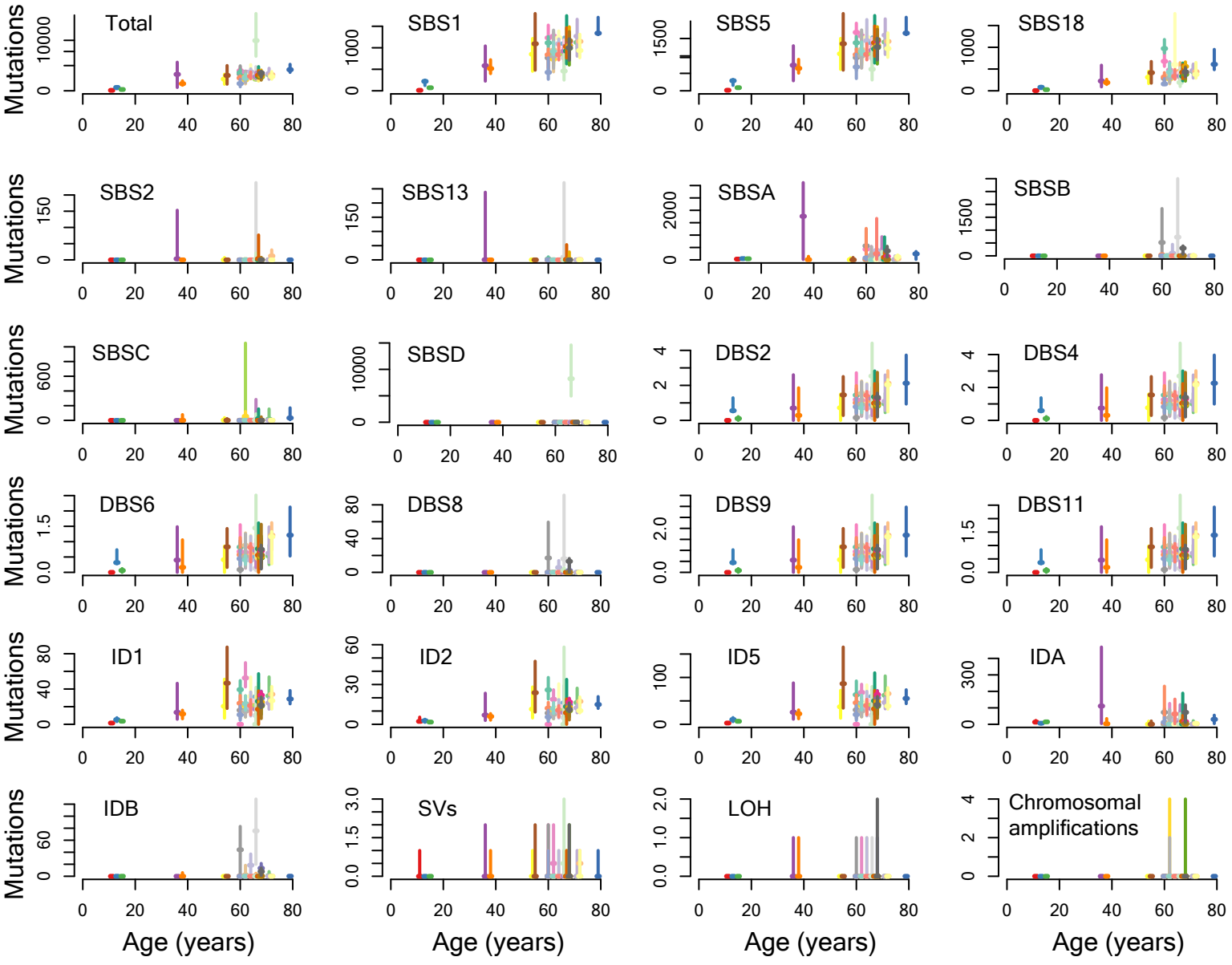
REFERENCES FOR SUPPLEMENTARY METHODS

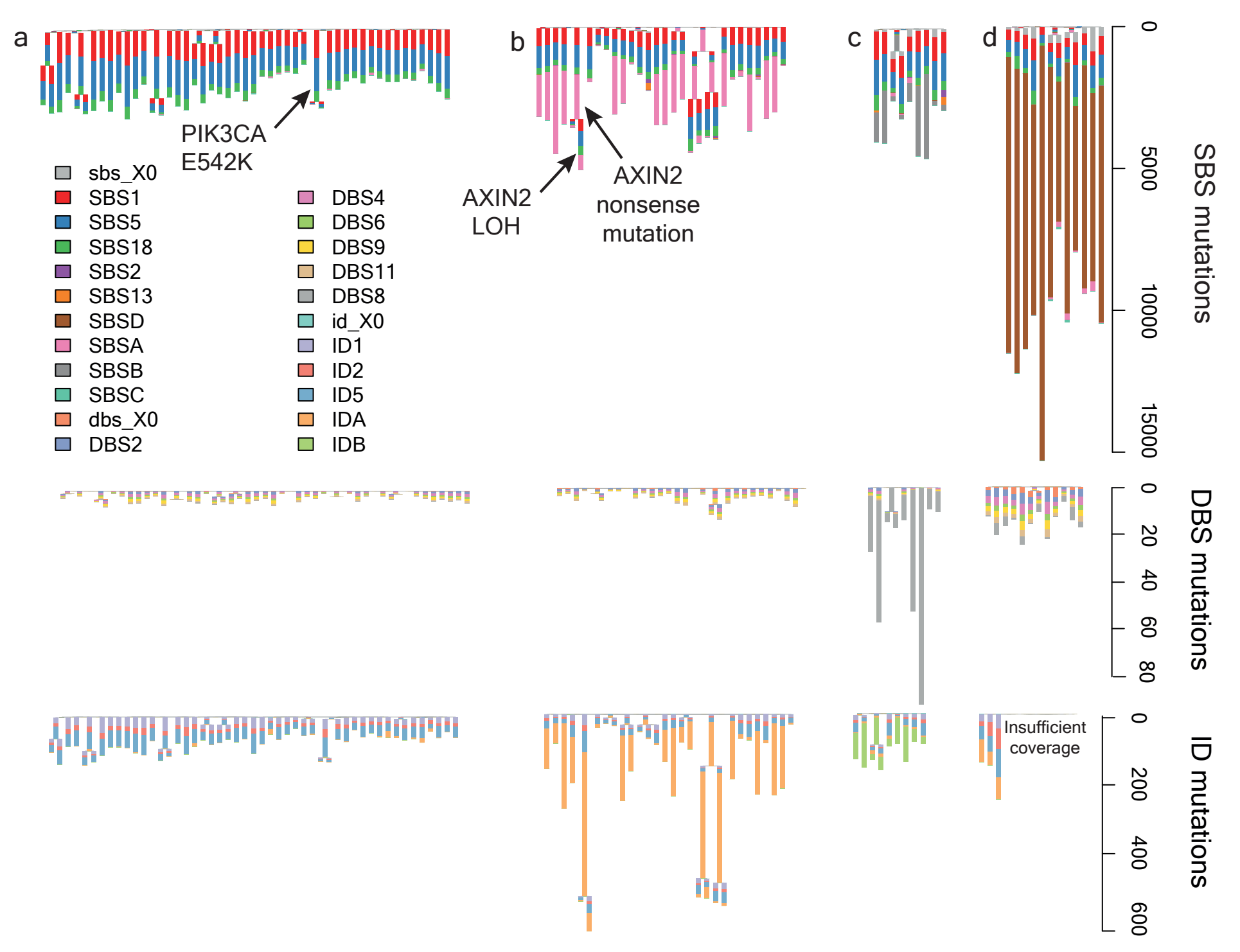
46. Jones, D. et al. cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
47. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, (2010).
48. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
49. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
50. Scheinin, I. et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research*, **24**, 2022–2032 (2014).
51. Buels R *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* doi: 10.1186/s13059-016-0924-1 (2016).
52. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
53. Forbes S.A. et al. COSMIC: somatic cancer genetic sat high-resolution. *Nucleic Acids Res.* **45**, D777-D783 (2017).
54. Felsenstein, J. PHYLIP — Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
55. Roberts, N. *Patterns of somatic genome rearrangement in human cancer.* PhD thesis, Univ Cambridge, UK (Wellcome Trust Sanger Institute, 2018).
56. Nicola Roberts, R pkg for Hierarchical Dirichlet Process, https://github.com/nicolaroberts/hdp (Accessed August 2018).

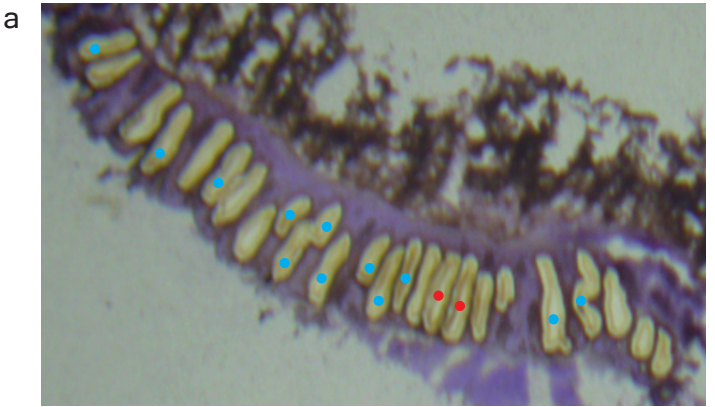
57. Farmery, J. H. R., Smith, M. L., Bioresource, N., Diseases, R. & Lynch, A. G. Telomerecat : A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. 1–17 (2018).
58. Nersisyan, L. & Arakelyan, A. Computel: Computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One* **10,** 1–14 (2015).
59. Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. **42,** 7–10 (2018).
60. Feuerbach, L., Sieverling, L., Deeg, K. I., Ginsbach, P. & Hutter, B. TelomereHunter : telomere content estimation and characterization from whole genome sequencing data. (2016).

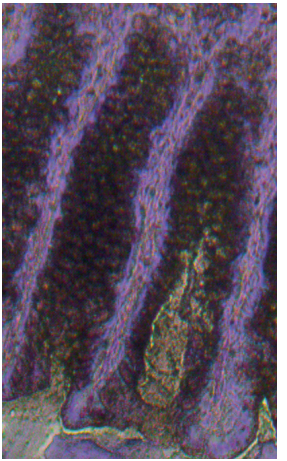

a
SBS18
C>A
C>G
C>T
T>A
T>C
T>G
DBS8
AC>NN
AT>NN
CC>NN
CG>NN
CT>NN
GC>NN
TA>NN
TC>NN
TG>NN
TT>NN
ID1
1bp deletion
1bp insertion
>1bp deletions at repeats (Deletion length)
>1bp insertions at repeats (Insertion length)
Deletions with microhomology (Insertion length)
Homopolymer length
Homopolymer length
Number of repeat units
Number of repeat units
Microhomology length


b
SBS1
SBS5
SBS18
SBS2
SBS13
SBSA
SBSB
SBSC
SBSD
DBS2
DBS4
DBS6
DBS9
DBS11
DBS8
ID1
ID2
ID5
IDA
IDB


Mutations
Total
SBS1
SBS5
SBS18
SBS2
SBS13
SBSA
SBSB
SBSC
SBSD
DBS2
DBS4
DBS6
DBS8
DBS9
DBS11
ID1
ID2
ID5
IDA
IDB
SVs
LOH
Chromosomal amplifications
Age (years)


a
b
c
d
PIK3CA
E542K
AXIN2
LOH
AXIN2
nonsense
mutation
sbs_X0
SBS1
SBS5
SBS18
SBS2
SBS13
SBSD
SBSA
SBSB
SBSC
dbs_X0
DBS2
DBS4
DBS6
DBS9
DBS11
DBS8
id_X0
ID1
ID2
ID5
IDA
IDB
SBS mutations
0
5000
10000
15000
DBS mutations
0
20
40
60
80
ID mutations
0
200
400
600
Insufficient
coverage


a


b


c
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5


d
PD37449e_lo0014 BWA alignment
PD37449e_lo0015 BWA alignment