RESULTS CHAPTER 2

MUTATIONAL LANDSCAPE OF NORMAL COLON

Introduction to this chapter

I.1. Colonic stem cell numbers and clonal dynamics in health

I.1.a. Monoclonal origin of crypts

The human colon undergoes extraordinary amounts of cellular turnover. Its luminal surface area is approximately 3,300cm^2 and it is renewed every 3-4 days (Potten et al. 1992). 15 million invaginations of the epithelium (or crypts) (Boman and Huang 2008), each containing about 2,000 cells (Potten et al. 1992), form the regenerative unit of the colon and house four key functional cell types: enterocytes, goblet cells, crypt base secretory cells, and enteroendocrine cells. Rarer cell types include tuft, microfold, and cup cells.

All of these differentiated cell types derive ultimately from stem cells that sit at the base of the crypt. The propensity of cells at the base of mouse crypts[1] to phagocytose cellular debris allowed an early form of functional lineage tracing: when they were fed tritiated thymidine, phagosomes in crypt basal cells were rapidly labelled. The tritiated thymidine label could only later be detected in different mature cell types higher up the crypt, suggesting that basal cells were multipotent progenitors (Cheng and Leblond 1974).

Not only do all cells in the crypt have a stem cell as a recent ancestor, but they can all retrace their line of descent to the same ancestral stem cell. Crypts from mice, chimaeric for either the H2 antigen, detectable with a monoclonal antibody, or a carbohydrate polymorphism stained with a lectin, were clonal for a given marker (Ponder et al. 1985). Similarly, detection of a Y-chromosome-linked marker in XX-XY chimaeric mice showed the shared common origin of all

[1] Here, as in much of the discussion on stem cell biology that follows, the experiment was performed on small intestine rather than colon. The stem cell biology of the colon is thought to be similar, and when there are notable differences between the two organs I will draw attention to them.

cell types in a gastric gland (including neuroendocrine cells, which had not been demonstrated previously) (Thompson et al. 1990). In these chimaeras, however, large patches of epithelium shared a marker, and so the finding could have been a consequence of embryonic mosaicism rather than adult stem cell dynamics. That crypts derive from a single adult cell was confirmed using mice heterozygous for an inactivating mutation of a lectin-binding protein. ENU random mutagenesis resulted in loss of the wild type copy – and consequently the lectin stain – in a small proportion of cells, such that the progeny of a single cell could be traced (Winton et al. 1988). Treatment with ENU in adulthood resulted initially in ribbons of cells lacking staining emanating from one mutated cell, but after a chase period it only left wholly stained or unstained crypts (Winton and Ponder 1990). Similarly, treatment of mice with a mutagen resulted in sporadic loss of the X-linked G6PD biochemical marker in a small proportion of colonic crypts, but, importantly, in all the cells of affected crypts (Griffiths et al. 1988).

I.1.b. Neutral drift dynamics

The ancestral stem cell need not still be alive, and could have multiple self-renewing descendants (Potten and Loeffler 1990). More recent experiments have shown that a number of extant stem cells replace one another through a process of neutral drift, resulting, over time, in monoclonal conversion. By inducibly labelling less than 2% of cells with a marker at a defined point it was possible to determine that half of small intestinal crypt cells in a mouse were clonal within eight weeks, and, furthermore, that the trajectory to monoclonality was consistent with neutral dynamics (López-García et al. 2010). The same result was derived using a confetti system in mice, under the control of Lgr5, a stem cell marker (Barker et al. 2007). Initially multicoloured crypts became monochrome over a period of 1-6 months in a manner consistent with neutral drift (Snippert et al. 2010). In humans, the monoclonality of colonic crypts has been demonstrated by staining for loss of the cytochrome oxidase (CCO) protein, which is encoded by the mitochondrial genome and so – due to the higher mutation rate of the mitochondrial genome and the process of drift to homoplasmy such that only one mutation is necessary – is reasonably frequently inactivated (Baker et al. 2014). This process of neutral drift is made possible, in part, by the relative non-

quiescence of intestinal stem cells: bromodeoxyuridine measurements indicate that the cell cycle time at the base of human colonic crypts is of the order of 30 hours (Potten et al. 1992).

I.1.c. Quantification of stem cell numbers and the rate of drift

The number of stem cells and the rate of neutral drift are important parameters to understand cancer risk, both because stem cells are believed to be the cell of origin of cancers and because the number of stem cells per crypt affects the probability that a driver mutation will be able to colonise a crypt (see General Introduction). Furthermore, the time to monoclonality of a crypt is of technical interest in this dissertation. As described below, in our study we sequenced individual colonic crypts, which allowed us to detect mutations that were present in every cell in the crypt. Thus, in effect we recover the genome of the most recent common ancestor (MRCA) of the crypt. If this common ancestor existed a very long time ago, we might be significantly underestimating the mutation burden of stem cells at the time of resection.

In mice, inducible labelling strategies have begun to unpick these parameters. Kozar et al. used a continuous labelling approach to mark a small proportion of cells, effectively allowing the output of single cells to be monitored. An out of frame reporter, under a house-keeping promoter, was placed next to a CA[30] microsatellite tract such that slippage of this highly mutable stretch would infrequently place the reporter back in frame (Kozar et al. 2013). Both wholly- and partially-labelled crypts were observed. The former increased linearly with mouse age, while the latter remained constant. The proportion of each of these informs on the stem cell number and replacement rate (if the rate is known at which the reporter is activated), allowing the inference of approximately seven stem cells per crypt in mouse colon, with most crypts drifting to monoclonality within a few months (Lopez-Garcia 2010). Interestingly, fewer stem cells per crypt were observed in the small intestine, which may contribute to the decreased cancer incidence.

Estimates in humans are more controversial. Coalescent modelling of variation in methylation patterns in human crypts, proposed to act as a somatically heritable but mutable mark, suggested the presence of at least eight stem cells and that the time to the MRCA of the crypt in humans is between 15 and 40 years (Nicolas et al. 2007). Substantial uncertainty exists in these estimates: a small number of CpG sites were assayed in a small number of cells per crypt, and the

model relies on assumptions on the kinetics of methylation and demethylation as well as the shape of the phylogeny of how differentiated cells are related to the common ancestor of the crypt. Cytochrome oxidase (CCO) staining has also been used to estimate these parameters (Baker et al. 2014). The observation that a ribbon of CCO^- staining varied in width as it rose from the base of the crypt to the lumen led the authors to conjecture that these 'wiggles' might be a read-out of symmetrical stem cell divisions at the base. This led to the estimation of approximately six functional stem cells and a rapid stem cell replacement rate, with monoclonal conversion times approaching three weeks (calculated in Nicholson et al. 2018). Ingenious though this idea was, it does not fit particularly well with our understanding of transit times within the crypt. Given that it takes about one week for a cell to migrate from the crypt base to the gut lumen in humans, and cells at the crypt base divide every few days (and not all of these need be symmetric cell divisions (Kozar et al. 2013)), it is arguable whether a ribbon of CCO^- staining captures a long enough time period to assay neutral drift. Only the bottom half of crypts were examined, and sometimes over eight wiggles per crypt are reported, which might indicate that the wiggles are rather a result of the behaviour of cells in the transit-amplifying compartment or that the software used to detect them is overly sensitive. In addition, due to mitochondrial heteroplasmy and neutral drift of mitochondria within the cell, the amount of cellular CCO can, in theory, fall below detectable levels and then recover. Finally, the continuous clonal labelling approach used in mice by Kozar et al. was applied to humans by staining for loss of the mPAS protein due to spontaneous somatic mutations (Nicholson et al. 2018). The median time to monoclonal conversion was estimated to be 6.3 years, and the number of stem cells to be between five and 10. A final caveat should be added to all the models of neutral drift discussed so far, whether in mouse or human, in that they treat all stem cells as having an equal chance of survival. However, 3D intravital imaging of a confetti mouse has indicated that stem cells further from the centre of the crypt base are more likely to be lost by differentiation (Ritsma et al. 2014).

I.1.d. Crypt fission

Crypts themselves occasionally divide to produce two daughter crypts (a process termed crypt fission) throughout life. The relevance of crypt fission to our understanding of colorectal

cancer is that neoplasia occurs, initially, through a process of crypt fission that allows a driver mutation to extend beyond the borders of the crypt in which it arose. A low crypt fission rate in normal colon probably reduces the rate of clonal expansion of mutations that could contribute to malignant transformation, and so reduces the probability of a 'second hit'. An understanding of the dynamics and regulation of crypt fission is, therefore, of central importance in our model of the evolution to cancer.

In humans, patches of multiple crypts that are CCO^- are observed, and both their frequency and size increase with age (Greaves et al. 2006). Sequencing the CCO gene in the two arms of a CCO^- bifurcating crypt showed that they shared the same inactivating mutation, indicating that fission rather than fusion was occurring (Greaves et al. 2006). Modelling of the CCO^- patch size as a simple birth process allowed an estimate of a fission event every 36 years, while modelling of crypt fission rates based on protein stains results in an estimate of crypts dividing on average every 140 years (Nicholson et al. 2018). It is unclear that a simple birth process is appropriate, as that would result in an increase in crypt number over the course of life, which – in mice at least – does not seem to be the case (Bruens et al. 2017). Recently, *in vivo* imaging has provided evidence for crypt fusion events in the mouse small intestine (Bruens et al. 2017), which could serve to control crypt numbers. Crypt exhaustion may additionally occur. Nonetheless, simulations indicated that the inclusion of crypt fusion has a negligible effect on the estimation of fission rates (Nicholson et al. 2018). Presumably, the ability of a crypt to fission has evolved as a regenerative response to damage. Indeed, inflammatory conditions often increase the number of crypts seen in fission.

I.2. The genomics of colorectal cancer

Cancer is a late product of somatic evolution, representing the end-point of the adenoma-carcinoma sequence (General Introduction). There now follows a brief overview of the genomic landscape of colorectal cancers. Cancers are the winners of somatic evolution, and their features provide clues as to the forces that govern natural selection in the colon. Two features warrant discussion: the mutation burden of colorectal cancers, and the features and numbers of driver mutations per cancer.

I.2.a. Mutational processes in colorectal cancer

Colorectal adenocarcinomas are one of the most highly mutated cancer types. They are surpassed only by melanomas, lung cancers – which are associated with exposure to the potent mutagens of ultraviolet radiation and tobacco exposure, respectively – and oesophageal cancers (Alexandrov et al. 2018). Colorectal cancers can be separated into two groups based on their mutation burden: so-called 'hypermutators' account for ~15% of cancers with a median of ~30 mutations per million coding bases, while the other ~85% typically have ~3 mutations per million coding bases (Cancer Genome Atlas Network et al. 2012). These two groups are generally associated with different histopathology, age-, site-, and gender-specific incidence, clinical features, and patterns of driver mutations. As alluded to in the General Introduction, hypermutators, with their high burdens of point mutations and short insertions and deletions (indels), tend to have far fewer copy number alterations (Cancer Genome Atlas Network et al. 2012), which is suggestive of two evolutionary paths to colorectal cancer, both of which are facilitated by an increase in mutation rate.

The analysis of mutational signatures in colorectal cancer provides a window into the origin of these mutations. Deconvolution of 60 whole colorectal cancer genomes, analysed in conjunction with thousands of other tumours as part of the Pan Cancer Analysis of Whole Genomes, revealed the activity of myriad mutational processes across the cohort: 13 single base substitution (SBS), 10 doublet base substitution (DBS), and four insertion and deletion (ID) signatures (Alexandrov et al. 2018) (figure 2.1). Most colorectal cancers had three to five SBS signatures, two to three ID signatures, and four to five DBS signatures. Signatures can be divided into those that affected almost all cancers and sporadic signatures that affected only a subset of cancers.

Common signatures might represent processes that are active in normal colorectal stem cells or that are necessarily associated with the process of transformation; without sequencing normal tissues, they cannot be told apart. Signatures that are common in colorectal cancers include SBS1, SBS5, SBS18, DBS2, DBS4, DBS6, DBS9, ID1, and ID2. Both SBS1 and SBS5 are found in almost all cancers sequenced to date. SBS1 accounts for a median of ~3,000 mutations per colorectal cancer genome (Alexandrov et al. 2018). It is characterised by C to T mutations in an N<u>C</u>G context (the mutated base is underlined), and is thought be due to the hydrolytic deamination of 5-methylcytosine to uracil (Rideout et al. 1990). In replication, an A is paired with the uracil,

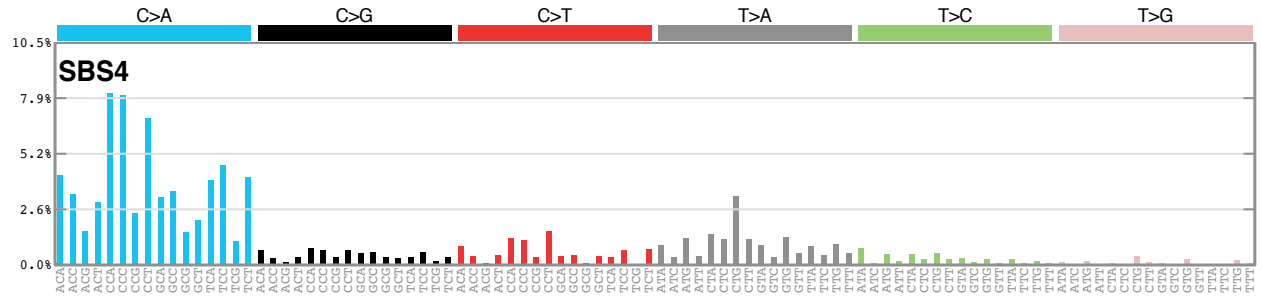
and in the following round of replication, a T is paired with that A, such that the mutation is fixed as a T. As such, this is a process that is likely to occur spontaneously in all cells, but whose rate could be increased in cancers if mutations are more likely to be fixed by DNA replication rather than repaired. Cancers that arise from less mitotically-active tissues tend to have lower rates of SBS1 and on this basis it has been proposed that SBS1 acts as a mitotic clock (Alexandrov et al. 2018, Alexandrov et al. 2015). SBS5 is also found in almost all tissues sequenced, but is of unknown cause; its relatively featureless trinucleotide profile provides few clues as to its aetiology. SBS18 is present in a large fraction of colorectal cancers, but not all. It should be noted, though, that the complex nature of cancer genomes, with multiple mutational processes with overlapping trinucleotide profiles active in a given cancer, makes the extraction of mutational processes relatively complicated. SBS18 may well truly be present in all cancers but not always be detected because some of its mutations could be misattributed to a different signature. SBS18 is characterised by C to A mutations and has been linked to the activity of reactive oxygen species attacking guanines to form 8-Oxoguanine, which, if not excised, can pair with an A (Viel et al. 2017). ID1 and ID2 are, respectively, single base insertions and deletions of a single T in a polyT tract, postulated to be due to replication slippage, and DBS2, DBS4, DBS6, and DBS9 are of unknown origin, although some have been observed in normal mouse cells and DBS2 and DBS4 have been noted to correlate with the age of cancer diagnosis (Alexandrov et al. 2018).

Sporadic mutational processes detected in the PCAWG cohort of 60 colorectal cancers include: SBS10a, SBS10b, SBS15, SBS17a, SBS17b, SBS28, SBS37, SBS44, SBS45, DBS5, DBS8, DBS10, DBS11, and ID14. Some have a known cause. For example, SBS15, SBS44, DBS7, DBS10, as well as a marked increase of ID1 and ID2, are associated with loss of DNA mismatch repair. Cases with these signatures are the ‘hypermutators’ described above; it is notable that hypermutation is the result of a strong increase in a small number of processes rather than a generalised increase in all processes. Mutations in the proof-reading domain of polymerase epsilon are associated with vast numbers of mutations due to SBS10a and SBS10b. A large number of sporadic mutation processes active in colorectal cancer, however, are still of unknown cause (Alexandrov et al. 2018).

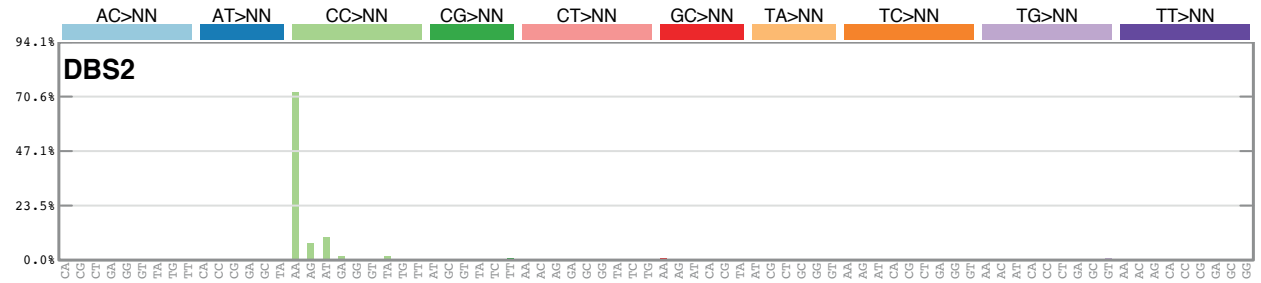
Larger structural changes have also begun to be classified thoroughly. Most colorectal cancers have of the order of a hundred structural variants, most of which are complex events, large

Figure 2.1. Categories for signature decomposition, reproduced with permission from Alexandrov et al. (2018). An example signature is shown for each of single bases substitutions (**a**), doublet base substitutions (**b**), and small insertions and deletions (**c**), in order to show the categories into which every signature is separated. In figures to follow in this chapter the category labels are often removed due to space constraints, but all are plotted with the same order and colouring as here.

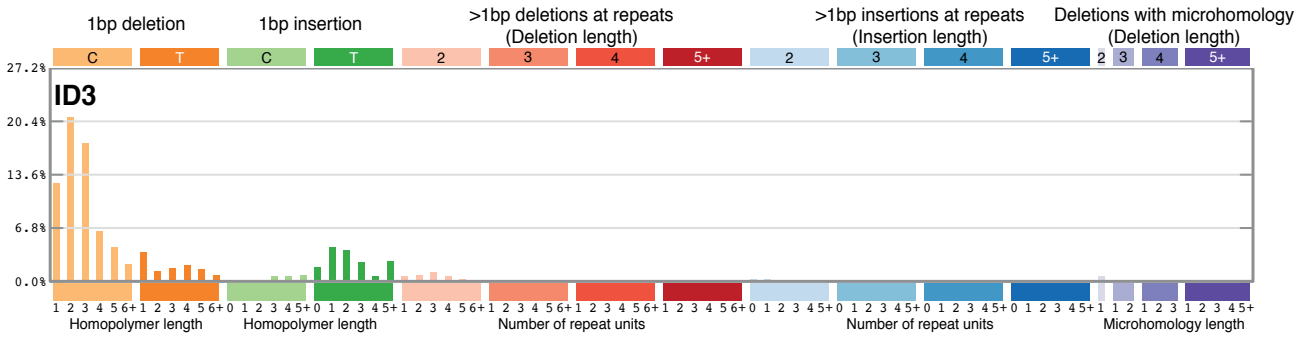
a


C>A
C>G
C>T
T>A
T>C
T>G
SBS4
10.5%
7.9%
5.2%
2.6%
0.0%

b


AC>NN
AT>NN
CC>NN
CG>NN
CT>NN
GC>NN
TA>NN
TC>NN
TG>NN
TT>NN
DBS2
94.1%
70.6%
47.1%
23.5%
0.0%

c


1bp deletion
1bp insertion
>1bp deletions at repeats (Deletion length)
>1bp insertions at repeats (Insertion length)
Deletions with microhomology (Deletion length)
ID3
27.2%
20.4%
13.6%
6.8%
0.0%
Homopolymer length
Homopolymer length
Number of repeat units
Number of repeat units
Microhomology length

deletions, or tandem duplications (Li et al. 2017), although, as noted above, hypermutated cancers have very few or none.

Thus, the mutational processes associated with colorectal cancer are diverse and variable across tumours. From sequencing cancers alone, it is unclear whether this is a reflection of the process of transformation or of the diversity in mutational processes operative across normal colorectal epithelium.

I.2.b. Driver mutations in colorectal cancer

I.2.b.i. Features of driver mutations

A comprehensive discussion of all driver mutations in colorectal cancer is beyond the scope of this chapter. Here, rather, I attempt to sketch out succinctly two of the molecular pathways that are frequently deregulated, and whose nature informs our understanding of somatic evolution in the colon. I will not discuss driver mutations that are thought to act by increasing mutation rate, since the debate about whether an increased mutation rate is of importance has been covered in the General Introduction. Studies modelling the effect of driver mutations in normal tissues are discussed in section I.5. of this chapter.

The vast majority of colorectal cancers harbour mutations that de-regulate the Wnt signalling pathway. Wnt ligands are absent from the microenvironment of differentiated colonocytes. A complex formed of APC, AXIN, and GSK3B phosphorylates β-catenin, targeting it for degradation. At the crypt base, however, Wnt ligands are present, and these bind to the Wnt receptor on crypt stem cells, signalling to inhibit the degradation of β-catenin. β-catenin may then shuttle to the nucleus, where, it binds the TCF4 protein and turns it from transcriptional repression to transcriptional activation of a wide array of genes, including *CMYC* and *CCND1* (reviewed in Fearon 2011 and in Bienz and Clevers 2000). Inactivating mutations have been documented in colorectal cancers in *APC* (homozygous truncating mutations in 70-80% of sporadic tumours and heterozygous loss of function mutations inherited in the fully penetrant colorectal cancer syndrome Familial Adenomatous Polyposis) and *AXIN2* (truncating mutations that are frequently heterozygous (Segditsas and Tomlinson 2006)). Both result in a failure to degrade β-catenin.

Mutations are also observed in *TCF7L2* (which encodes TCF4). While it is counterintuitive that loss of TCF4 should promote colorectal cancer, it has been proposed that loss of the repressive effect of TCF4 on some genes is a large part of the effect of complexing with β-catenin (Fearon 2011). It seems that these mutations of the Wnt signalling pathway essentially revert cells to a stem cell-like phenotype: disruption of the β-catenin-TCF4 complex in colorectal cancer cell lines halted proliferation and increased the expression of genes associated with differentiation (van der Wettering et al. 2002). Kinzler and Vogelstein proposed that *APC* is a gatekeeper gene, which they defined as follows: 'gatekeeper genes are responsible for maintaining a constant cell number in renewing cell populations [...] Mutation of the gatekeeper leads to a permanent imbalance of cell division over cell death' (Kinzler and Vogelstein 1996).

The second pathway that warrants brief discussion is that of Ras signalling. This pathway mediates pro-proliferative signalling downstream of receptor tyrosine kinases such as EGFR. KRAS, BRAF, PIK3CA, and PTEN all form a part of this pathway and their genes are mutated in colorectal cancers at frequencies of 40%, 10% (mostly in hypermutators), 20%, and 10% respectively (Fearon 2011). Interestingly, *KRAS* mutations are frequently found in hyperplastic polyps with little chance of progressing to cancer, indicating that they might drive proliferation sufficiently powerfully to create a macroscopic lesion without necessarily setting a cell firmly on the road to cancer.

To simplify the features of the most frequent driver mutations in colorectal cancer, then, it seems that these act through a combination inhibiting differentiation and driving proliferation. A plethora of other driver mutations in colorectal cancer, however, have a broad range of effects.

I.2.b.ii. Numbers of driver mutations

The number of driver mutations needed for colorectal cancer is a primary determinant of cancer risk (Armitage and Doll 1954, Nordling 1953, General Introduction) and contextualises the observation of driver mutations in normal colon: if only two driver mutations were necessary to cause cancer we would be far more alarmed by the observation of a single driver mutation in normal colorectal epithelium than if 20 were required. A distinction should be drawn between the number of driver mutations *observed* in colorectal cancers, and the number of driver mutations

necessary to cause a cancer. Ongoing evolution within a cancer will increase the number of mutations found in a cancer over the bare minimum necessary. Both will be discussed here, but the latter quantity is the one of greatest interest to us.

Tabulating the driver mutations that were known in 1990, Fearon and Vogelstein estimated that most cancers have four to five driver mutations (Fearon and Vogelstein 1990). This was probably an underestimate, as only small parts of the genome were assayed and driver mutations have been discovered since. A more recent analysis of 52 colorectal cancer whole genomes that identified coding and non-coding driver mutations with a method that was largely based on prior knowledge found a mean of 7.4 driver point mutations and 2.5 driver genomic rearrangements per colorectal cancer (Sabarinathan et al. 2017). There was substantial variation, however, as almost 20% of the cancers had three or fewer drivers, and another ~20% had 10 or more. Approaches that do not rely on prior knowledge are perhaps more robust and comprehensive. Martincorena et al. developed a dNdS method to detect driver point mutations in the coding regions of cancer genomes. This method compares the ratio of nonsynonymous to synonymous mutations in genes, under the assumption that nonsynonymous mutations code for proteins and are therefore subject to selection, while synonymous mutations have no functional impact. Thus, an excess of nonsynonymous mutations indicates positive selection of the gene, while a depletion indicates negative selection. The trinucleotide composition of the gene, the mutational processes active in the tumour, and local mutation rate were taken into account. Using this method, a mean of 10 positively-selected coding point mutations per tumour was estimated from hundreds of colorectal cancer exomes (Martincorena et al. 2017). Perhaps a similar analysis, restricted to the tumours that are estimated to have transformed recently, might be informative for the minimum number of mutations necessary to cause cancer, although some driver events, including rearrangements, noncoding mutations, and epigenetic alterations would be missed.

Estimates of the minimum number of driver mutations necessary were first derived using cancer incidence data. It was observed that the incidence of cancer rose proportionately to four to six times the power of age, which was interpreted as indicating that five to seven rate-limiting events (often assumed to be driver mutations) were necessary to cause a cancer (Nordling 1953, Armitage and Doll 1954, General Introduction). This method assumes that driver mutations are independent events; if each driver mutation induces a clonal expansion or increases the mutation rate, and thus increases the probability of another mutation affecting one cell of the mutated clone,

the number of driver mutations could be lower. Realisation of this, along with the discovery of recessive oncogenesis (Knudson 1971) led to the elaboration of a multi-stage clonal expansion model. In this model, a number of mutations, each considered to be an independent rare event, must accumulate in a cell before it undergoes a clonal expansion, and then another high frequency event must occur (another mutation in any one of the many mutated cells). Fitting this model to colorectal cancer incidence data, trying different numbers of pre-clonal expansion mutations, showed that the most likely number of pre-clonal expansion mutations was just two (Luebeck and Moolgavkar 2002). These two mutations were posited to correspond each to the inactivation of one allele of the APC gene. Revealing though this is, assumptions are made that could be incorrect: the first mutation may already cause a small clonal expansion and neutral drift within the crypt is not taken into account.

Tomasetti et al. observed that patients with Lynch syndrome have a 114-fold risk of cancer over the general population, but that the microsatellite unstable tumours that these patients get only have eight times as many mutations as microsatellite stable sporadic tumours. From this they inferred that three driver mutations are sufficient to cause cancer (Tomasetti et al. 2014). This analysis ignores the fact that the mutation rate is only accelerated in some of the cells in the colon of Lynch syndrome patients, those that have lost the second copy of a mismatch repair gene. It would hold better for patients born with biallelic mismatch repair, who get colorectal cancer much younger than Lynch syndrome patients (Wimmer and Kratz 2010). Secondly, the analysis does not consider the different pathways to cancer taken by mismatch repair deficient *versus* proficient cancers; for instance, the former often have driver mutations in *TGFBRII* and *BAX*, which contain microsatellite tracts and so must be mutated disproportionately more rapidly than other driver mutations.

An orthogonal approach to estimate the number of driver mutations per patient is to induce different combinations of driver mutations in cancer models. For example, intestinal organoids (*in vitro* clonal expansions that recapitulate crypt organisation) engineered to have driver mutations in *APC*, *TP53*, *KRAS*, and *SMAD4* grew independently of niche factors and developed a morphology similar to organoids derived from invasive carcinomas (Drost et al. 2015).

In summary, then, we still do not know conclusively the minimum number of driver mutations needed for a colorectal cancer. The weight of the evidence, however, seems to point to a handful of mutations, probably between two and seven, and thus fewer than the observed counts.

I.3. Colorectal cancer incidence

Some features of colon cancer incidence deserve a brief mention. The age incidence of colorectal cancer has been discussed in the context of driver mutations. Three other features of cancer incidence, however, raise questions of colon biology.

Firstly, while in the West colonic cancer is a common disease, with a lifetime risk of about 5% (Cancer Research UK Bowel Cancer Incidence Statistics), other parts of the world have much lower rates. Comparing extremes, there is a 10-fold age-standardised incidence difference between central Africa and Oceania. While this may be partly genetic, countries in Eastern Europe or in Asia that have recently adopted a more Western diet have seen rapid increases in incidence. A number of risk factors have been identified, including smoking, alcohol, and eating processed meats, as well as composition of the microbiome, infections with *Fusobacterium spp.*, and inflammatory diseases like ulcerative colitis. Conversely, low-dose aspirin has been shown to have a protective effect (Brenner et al. 2014). All of these could presumably alter one or both of the mutation rate and selection pressures in the colon, but their precise mechanism of action is as yet incompletely understood.

Secondly, differences in cancer incidence between parts of the gut are intriguing. Most strikingly, in the UK, the incidence of adenocarcinoma in the large bowel is approximately 60-fold higher than in the small bowel (Cancer Research UK Bowel Cancer Incidence Statistics, Aparicio et al. 2014), despite their similar stem cell biology. I am not aware of a simple explanation for this from the perspective of comparative evolution. Relative to other primates the colon has reduced in size even more than the small intestine: the colon takes up about 20% of the gut, whereas in other large primates it is closer to 50% (Milton 1987). While the function of the human colon is largely limited to water reabsorption, in primates such as gorillas the colon plays a major role as a fermenting chamber. Short chain fatty acids produced as a result of bacterial fermentation of fibre provides over half of a gorilla's calories, compared to less than 10% in a human (Popovich et al. 1997). Combined with the geographic variability in colon cancer risk, we can speculate that dietary changes associated with ever more energy-rich foods might be responsible for the high incidence of colorectal cancer observed nowadays in the Western world.

Thirdly, even within the colon there are substantial differences in cancer incidence. In men, cancer incidence is approximately twice as high in the sigmoid colon as in the caecum (23% v. 12%), while in women it is only slightly higher (20% v. 17%) (Cancer Research UK Bowel Cancer Incidence Statistics). Clinical factors may play a role in these incidence rates, since left-sided cancers typically cause symptoms earlier, but the difference is nonetheless striking and, to my knowledge, unexplained. These differences could be caused by small differences in the rate of known mutational processes, different frequencies of activity of sporadic mutational processes, differences in stem cell dynamics (such as stem cell numbers per crypt), or other factors.

The sequencing of normal colonic and small intestinal stem cells can begin to resolve the role of somatic mutations in these curious discrepancies in cancer incidence.

I.4. Current understanding of somatic mutations in normal colon

I.4.a. Mutation rates

In recent years there has been a flurry of interest in the mutational processes in normal tissues, including in normal colon. The mutations in normal colon can be inferred from cancer genomes, which represent the sum of the mutations that occurred before and after transformation. Assuming a similar lag time across colorectal adenocarcinomas between the departure from normal mutational processes and resection, the number of mutations due to normal mutational processes should correlate with patient age. Examination of this relationship in colon revealed that only signature 1 correlated with age, with a rate of 23 (95% CI 19-28) mutations per year (Alexandrov et al. 2015).[2] Interestingly, the curve passes through the origin. An increase in the mutation rate of signature 1 during carcinogenesis would shift the curve up, whereas a lag between diagnosis and the time to the most recent common ancestor of the tumour in whom mutations can be called would shift the curve down. While a novel insight, the authors note that 'Peering through the "cracked lens" of cancer genomes may obscure or distort the estimates of clock-like mutation

[2] Please note that this is the COSMIC version of signature 1, which is not composed exclusively of C to T at CpG mutations but has some background in other contexts that resemble those of signature 5. Using the PCAWG SBS1, the number of mutations attributed to this signature would be lower, and signature 5 may be found to accumulate linearly.

rates of normal cells that are progenitors of the cancers' (Alexandrov et al. 2015). Sequencing the genomes of normal cells provides far greater resolution and allows the investigation of non-clock-like processes.

Normal organoids derived from 21 single colonic stem cells from six patients ranging in age from nine to 67 years old (although none was between the ages of 15 and 53) provided the first clear insight into the mutations in normal colonic stem cells (Blokzijl et al. 2016). Small intestinal organoids were derived as well and, remarkably, both had the same mutation rate of 36 mutations per year (95% CI 26.9-50.6 for colon and 25.8-43.6 for small bowel), which indicates that the model of cancer risk being mostly due to stem cell divisions (Tomasetti and Vogelstein 2015, General Introduction) does not explain the difference between cancer incidence in the large and small bowel. Interestingly, four out of 15 colonic organoids that could be assessed were found to have structural variants, including a complex translocation and a trisomy of chromosome 13, while small deletions were found in three out of 14 small intestinal organoids. Similarly, high density SNP arrays on individual human colonic crypts showed the presence of deletions and amplifications, which increased in prevalence with age, with detectable copy number changes in one in seven crypts from a 78 year-old. (Hsieh et al. 2013).

Blokzijl and colleagues found three single base substitution signatures to be operative in colonic organoids: signatures 1, 5, and 18. Numbers are not provided in the text, but judging from the figures, signature 1 accumulated at a rate of about 25 (95% confidence interval ~18-38),[3] signature 5 at about 10, and signature 18 at about five mutations per year. Similar numbers were found in the small bowel, with a little less signature 1 and more signature 5. The rate of accumulation for signature 18 is not significantly different from 0, and signature 18 was found to be enriched in sequential *in vitro* cultures, which led the authors to ascribe it to an oxidative process during organoid culture. For all signatures in colon the slope of the curve of mutation burden versus age cuts the y intercept near to, but slightly above, the origin, which hints at a period of transiently increased mutation rate. The authors remark on the lack of interindividual variability in mutational processes, but with only six individuals they were unlikely to capture mutational processes that are not ubiquitous. No driver mutations were found, but one disadvantage of the organoid culture system is that it can select against driver mutations in certain tissues; indeed, in the colon, wild-

[3] Again, this is the COSMIC version of signature 1. Using the PCAWG definition, probably fewer mutations would be attributed to SBS1 and more to SBS5.

type organoids will outcompete mutant ones unless niche factors are removed (van de Wetering et al. 2015). Despite its relatively limited power, this study is seminal in that clean whole genomes of single normal cells were seen for the first time. The discrepancy between the mutation burden of ~3,000 mutations in a colonic stem cell from a 60 year-old and the average mutation burden of 10,000-20,000 mutations in a non-hypermutated colorectal cancer demonstrates that cancers have an elevated mutation burden over their normal counterparts.

Nonetheless, until normal and tumour from the same people are studied, it remains theoretically possible that those people who get cancer have an elevated mutation rate all around their bowel. In an attempt to resolve this, organoids were derived from APCmin/+ mouse adenomatous crypts and normal crypts and exome sequenced (Lugli et al. 2017). The rate of acquisition of point mutations was found to be ~11 fold higher in adenomas, although small numbers of mutations were captured: only 71 mutations were seen in total across seven normal organoids and 15 tumour-derived organoids. The caveats of organoids (as discussed above) remain, and mouse intestinal organoids have been shown to have different mutational spectra to human ones, with an enrichment of C to A and fewer C to T mutations (Blokzijl et al. 2016, Behjati et al. 2014). Furthermore, while this indicates that the *people* who get cancer need not have a generally increased mutation rate, it is possible that the *cells* that become cancerous could, even prior to transformation, have had an increased mutation rate. A phylogenetic analysis of tumours and comparison to normal tissues from the same patients could resolve this.

An orthogonal approach to quantify mutation rates in normal tissues despite their polyclonality is to perform very deep and highly error-corrected sequencing, such that mutations in individual molecules of DNA can be called reliably (General Introduction). Analysis of normal colonic epithelium from 11 individuals showed an increase in mutation rate with age, reaching ~3,500 mutations per genome in people over the age of 40 (Hoang et al. 2016), which is similar to the number found by sequencing organoids. It should be noted that this assays mutations across all cells in the epithelium, some of which will have arisen during the process of differentiation, while organoids only report the mutations in stem cells. The similarity between these two estimates then indicates that there is no dramatic increase in mutation rates over the course of differentiation.

I.4.b. Driver mutations in normal colon

Relatively little is known about the frequency of driver mutations in normal human colon, largely due to difficulty in their detection. To my knowledge, two approaches have been used so far: PCR-based methods to detect specific mutations, which are only practical for assaying hotspots, and staining for tumour suppressor proteins. This detects homozygous mutations that result in a loss of expression rather than merely a loss of function.

KRAS hotspot mutations have been detected through PCR-based methods since 1998 in a number of studies, but most were non-quantitative or only semi-quantitative and frequently failed to detect *KRAS* mutations in normal mucosa (discussed in Parsons et al. 2010). Parsons and colleagues used allele-specific competitive blocker polymerase chain reaction (ACB-PCR) to quantify *KRAS* codon 12 GTT and GAT mutations in 89 samples of colonic mucosa (Parsons et al. 2010). Mutant *KRAS* was detected in all samples of normal mucosa, and it was estimated that 1 in 3,500 normal cells contained a KRAS codon 12 mutation. This is 60 times more frequent than the mutation is expected to occur by chance (Tomasetti et al. 2013), indicating positive selection. *KRAS* mutations were more frequent in the sigmoid colon, concordant with the observation that sigmoid tumours more frequently have *KRAS* mutations. Interestingly, the frequency of *KRAS* codon 12 GTT mutations was found to be higher in adenomas than in carcinomas, indicating that some *KRAS* mutations may promote the transition to malignancy more effectively than others. No correlation of variant allele fraction was observed with the patients' age, which may be a result of the relatively small age span covered (50 to 80 years old). A more recent study using targeted sequencing on a larger cohort of patients validated the presence of *KRAS* mutations in normal colon (Nicholson et al. 2018). Similarly, ACB-PCR investigation of *PIK3CA* found that the H1047R mutation was above the detection threshold of $1x10^{-5}$ in 20 out of 20 normal samples, whereas the E545K mutation was not detected in any of the 20 samples (Parsons et al. 2017).

Recently, staining for four putative tumour suppressor proteins located on the X chromosome (such that loss of one allele was sufficient to inactivate all copies of the protein) was performed for 186 patients across an age range (Nicholson et al. 2018). One of these, STAG2, was found to be lost in most patients at a mean frequency of about one in 7,000 crypts in a 60 year-old. The mechanisms by which these driver mutations might colonise colorectal epithelium are discussed below.

I.5. Quantitative insights into the earliest stages of colorectal cancer evolution

In order to form a tumour, driver mutations that occur in colonic stem cells must be able first to sweep through the crypt and second to spread beyond it. Quantitative analyses of the effects of driver mutations in mouse models, and more recently in humans, have begun to elucidate how this occurs.

I.5.a. Driver mutations skew the odds of stem cell competition

Under a model of neutral drift, neutral mutations that arise in a single stem cell have a probability of becoming fixed that is inversely proportional to the number of stem cells per crypt. A mutation that decreases the probability that the cell in which it occurs will be lost from the crypt is more likely to be able to go on to form a tumour. Mouse models indicate that this is a property of common colorectal cancer driver mutations in *Apc* and *Kras* (Vermeulen et al. 2013, Snippert et al. 2014). Vermeulen et al. (2013) induced driver mutations and a coloured tag in mice at infrequent levels, such that only one cell in a crypt would be mutated. Quantifying the growth of the labelled cells allowed a model of the benefit that the driver mutation conferred. *Kras* G12D mutant stem cells outcompeted their wild-type neighbours 80% of the time. Snippert et al. (2014) induced both *Kras* G12D and the confetti reporter at a low frequency in Lgr5 positive cells, and similarly found that the mutant cells had a higher chance of sweeping through a crypt. An EdU pulse showed that *Kras* mutant cells were cycling faster, consistent with our understanding of the Ras pathway in driving proliferation.

Vermeulen et al. also studied *Apc* and *P53*. Interestingly, *Apc* +/- cells outcompeted *Apc* +/+ cells, but *Apc* -/- cells outcompeted *Apc* +/- cells. This shows that mutations of tumour suppressor gene alleles are not necessarily independent events, as they are frequently modelled to be. The selective advantage of *Apc* mutations within the crypt is consistent with its suggested role in controlling the balance stem cell self-renewal and differentiation (section I.4.b.). As discussed in the General Introduction, *P53* mutations were only found to be advantageous when colitis was

induced. The concept of a context-specific driver mutation is both fascinating and daunting in that it adds another layer of complexity to the construction of quantitative models of cancer.

In humans, comparing STAG2 to a neutral mark showed that the proportion of crypts that had lost the expression of STAG2 in all cells was increased relative to the proportion of crypts where STAG2 was lost in only a fraction of cells, indicating more rapid clonal sweeps (Nicholson et al. 2018). STAG2 loss was estimated to increase the probability of replacing a wild type neighbour from 0.5 to 0.99.

I.5.b. Driver mutations increase the rate of crypt fission

Crypt fission is rare physiologically (Baker et al. 2014, Nicholson et al. 2018), so unless a driver mutation can increase this rate, it is likely to remain entombed in its own crypt. It seems that as well as giving a selective advantage within the crypt, canonical driver mutations also promote clonal expansion beyond the crypt.

Snippert et al.'s (2014) multi-coloured labelling method allowed them to analyse the dynamics of crypts that had been fully colonised by *Kras*. In the presence of *Kras* mutations, adjacent crypts were more likely to be the same colour, which indicated an excess of crypt fission events. It was estimated that *Kras* G12D increased the rate of crypt fission 30-fold. In humans, the discrepancy between the rate at which *KRAS* mutations should occur by chance and their allele fraction in bulk epithelium indicates that they must increase the rate of crypt fission approximately 10-fold (Nicholson et al. 2018). Similarly, STAG2-negative patches of epithelium tended to be larger than patches of epithelium negative for neutral marks. Modelling the growth of these patches with age showed that 0.7% of crypts with neutral marks fissioned per year, whereas 2.15% of crypts with STAG2 loss fissioned per year.

Theoretically, early driver mutations need not cause crypt fission. In ulcerative colitis (a risk factor for colorectal cancer), crypt fission rates are increased, presumably as part of a wounding response, and so driver mutations could hitch-hike out of the crypt; once the clone was big enough, a second driver mutation that did allow a disruption of the tissue architecture would be more likely to strike the clone. Nonetheless, it seems probable that most colorectal cancer drivers that tend to occur early in the adenoma-carcinoma sequence will promote crypt fission as

it would provide a strong fitness advantage. Many known tumour suppressor genes seem to have evolved at the time when our ancestors became multicellular (Domazet-Loso and Tautz 2010), and it has been suggested that they served the purpose of limiting the selfish behaviour of cells in metazoa. With this in mind it is perhaps not so surprising that their deregulation results in an atavism that involves concomitant proliferation and disruption of tissue architecture.

Thus, we begin to be able to describe in a quantitative manner how driver mutations can colonise the colorectal epithelium. Much remains unanswered, however. To list but a few questions: what are the actual mechanisms behind a competitive advantage within the crypt, and what governs crypt fission? Are mutations that allow cells to outcompete their wild type neighbours necessarily advantageous to cancer? What are the effects of combinations of driver mutations? And what other driver mutations lurk in normal colonic epithelium?

Results

R.1. Study design

We aimed to investigate the landscape of somatic mutations in normal colon. A small number of normal colonic organoids had previously been sequenced (Blokzijl et al. 2016, section I.4.a), which indicated the activity of only three mutational signatures in normal colon and little variation in between different samples. We designed an experiment to explore the variety in mutation burden, mutational processes, and frequency of driver mutations in the normal colonic mucosa between different people and between different crypts within one person. We set out to exploit the stem cell architecture of the colon as a clonal unit by laser capture microdissection of single crypts, followed by sequencing. The advantages of laser capture microdissection over organoids (which were used by Blokzijl and colleagues (2016)) are the following: the method is more easily scaled to analysing hundreds of samples; spatial information on the location of the crypts is retained, allowing the investigation of processes such as crypt fission; there is no selection of crypts in culture, allowing an unbiased discovery of driver mutations; and no mutations are acquired *in vitro*. On the other hand, one downside of bulk sequencing whole crypts is that only

mutations in the most recent common ancestor of the crypt are called. The time to the most recent common ancestor of the crypt is likely to be within the decade before resection (section I.1.c.). This should be borne in mind in analyses of mutation burden, but is not problematic for the discovery of mutational signatures.

R.2. Development of a protocol to sequence individual crypts

At the time when this experiment was begun (Autumn 2015), 500ng of DNA were required for Illumina to guarantee sequencing success, and sequencing was rarely performed with less than 200ng. Colonic crypts, each with ~2,000 cells (~12ng of DNA) of which only a fraction are obtained in a given section, were far below this threshold. We developed a pipeline to allow the sequencing of single colonic crypts. Peter Ellis developed a more sensitive library preparation method (Methods), while I, with advice from Robert Osborne, optimised the thickness of sections, the choice of fixative, the staining protocol, and the lysis method. All experiments were performed on fresh frozen colonic tissue, initially from a mouse, and later from a human.

The thickness of sections was chosen to be the largest possible that still allowed the dissection of single crypts. In very thick sections, if crypts are visualised longitudinally it may be that a fragment of another crypt is hidden behind the back wall of the crypt that is being dissected, which would result in a polyclonal sample. The spacing between crypts may vary depending on factors such as mucosal oedema. Images of *en face* crypt sections from a number of mucosal samples, however, showed that 30 micron sections rarely resulted in capturing one crypt behind another. The staining regimen was chosen to be the simplest that still allowed crypts to be visualised, the rationale being that any unnecessary chemicals might damage DNA. Furthermore, as these experiments require very long days, any time that can be saved is valuable. Crypts were therefore stained only with Gill's haematoxylin and no eosin. With 30 micron sections and staining with haematoxylin, crypts could clearly be seen as clonal units (figure 2.2a). The images are much less attractive than those in textbooks due to a combination of the thickness of the section (pathology sections are often only 4 microns thick), the use of fresh frozen tissue (whereas pathology sections are formalin-fixed and paraffin-embedded), the absence of a coverslip (since

this cannot be used in laser capture microdissection), and the fact that sections were only stained with haematoxylin (rather than haematoxylin and eosin as in standard pathology).

The best fixative and lysis methods were evaluated jointly by the quantification of libraries made by Peter Ellis. Four fixatives were tested: acetone, paraformaldehyde, methanol, and ethanol. Three different lysis methods were assayed: alkaline lysis, protease lysis, and chaotropic lysis (RLT). Below are the results for half a plate testing these combinations. From these results and repeat experiments that confirmed them, fixation with methanol and protease lysis were chosen.

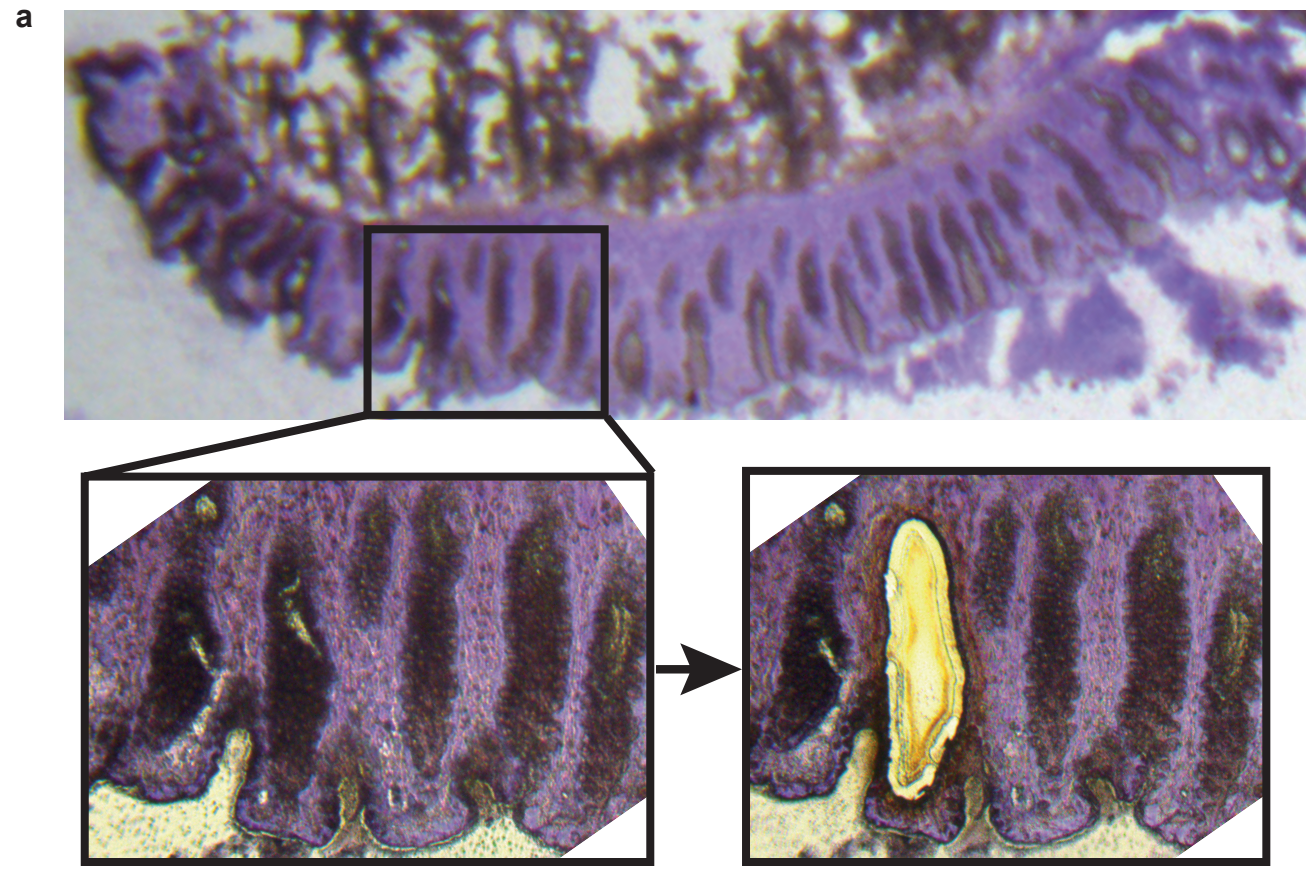
	7	8	9	10	11	12
A	blank	blank	blank	blank	blank	blank
B	Alk EtOH 1 crypt	Alk PFA 1 crypt	Prot EtOH 1 crypt	Prot PFA 1 crypt	RLT EtOH 1 crypt	RLT PFA 1 crypt
C	Alk EtOH 1 crypt	Alk PFA 1 crypt	Prot EtOH 1 crypt	Prot PFA 1 crypt	RLT EtOH 1 crypt	RLT PFA 1 crypt
D	Alk EtOH 1 crypt	Alk PFA 1 crypt	Prot EtOH 1 crypt	Prot PFA 1 crypt	RLT EtOH 1 crypt	RLT PFA 1 crypt
E	Alk ace 1 crypt	Alk MeOH 1 crypt	Prot ace 1 crypt	Prot MeOH 1 crypt	RLT ace 1 crypt	RLT MeOH 1 crypt
F	Alk ace 1 crypt	Alk MeOH 1 crypt	Prot ace 1 crypt	Prot MeOH 1 crypt	RLT ace 1 crypt	RLT MeOH 1 crypt
G	Alk ace 1 crypt	Alk MeOH 1 crypt	Prot ace 1 crypt	Prot MeOH 1 crypt	RLT ace 1 crypt	RLT MeOH 1 crypt
H	blank	blank	blank	blank	blank	blank

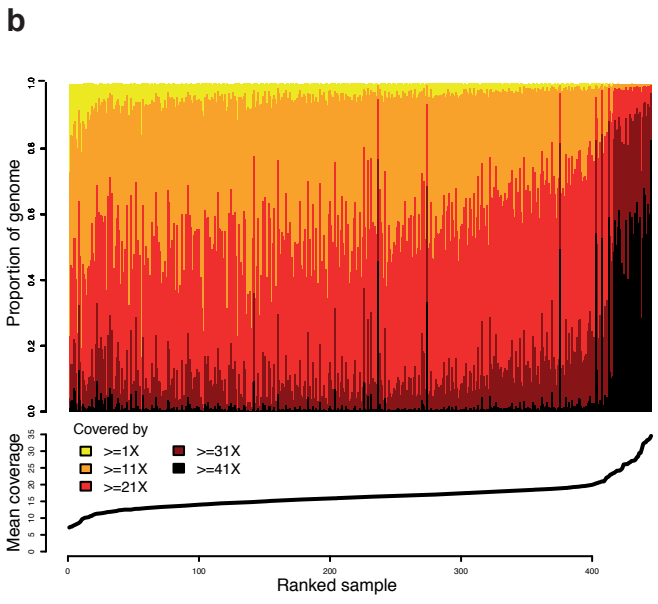
A	0.076	0.12	0.113	0.12	0.14	0.167
B	0.233	0.114	0.12	1.71	0.125	0.168
C	0.086	0.092	0.098	2.55	0.139	0.133
D	0.079	0.08	1.504	0.198	0.121	0.152
E	0.144	0.204	0.693	28.895	0.118	0.152
F	0.073	0.248	11.444	18.849	0.121	0.135
G	0.156	0.096	19.788	6.247	0.12	0.173
H	0.107	0.1	0.11	0.111	0.125	0.127

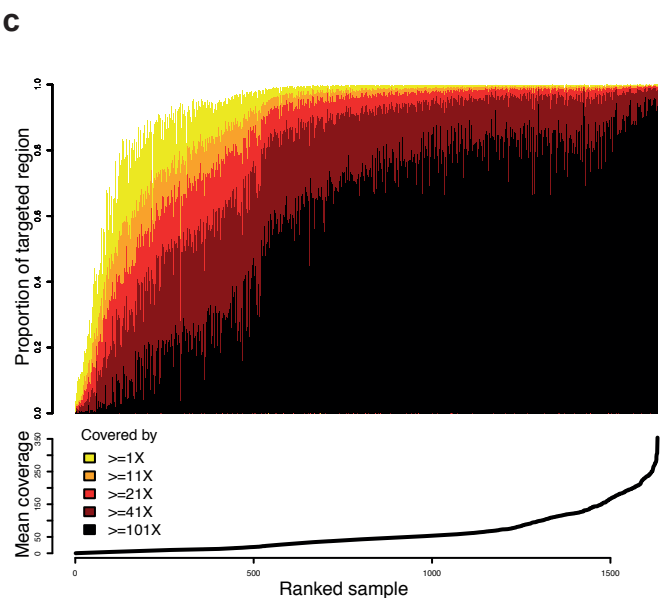
Table 2.1. Quantification of libraries to test fixation and lysis condition. The top panel shows the layout of this half of the plate. The bottom panel shows library preparation results, in ng/ml. Alk, alkaline lysis; Prot, protease lysis; RLT, chaotropic lysis; EtOH, ethanol fixation; ace, acetone fixation; PFA, paraformaldehyde fixation; MeOH, methanol fixation.

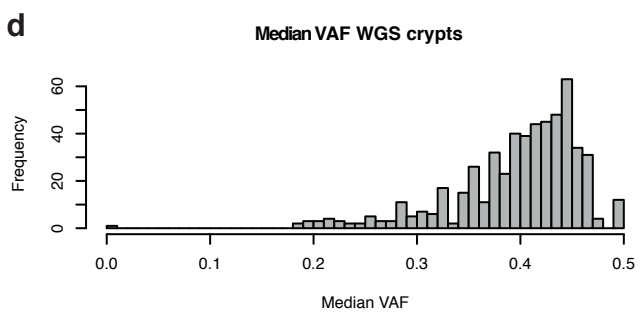
With these sample preparation methods and the library construction protocol developed by Peter Ellis, 11 single colonic crypts from one sample were sequenced at 1-2x coverage each. Even coverage across the genome was observed, and sequencing metrics were acceptable. Pooling the crypts allowed known germline mutations in this patient to be recovered. Sequencing at higher coverage (~15X per crypt) allowed somatic mutations to be called using our standard algorithms

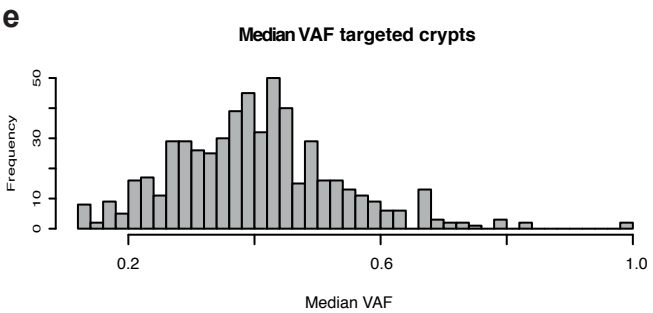
Figure 2.2. Laser capture microdissection of crypts. a, a representative image of a section of colonic tissue, with a magnified inset showing the section before and after dissection of a crypt. **b-c,** the coverage of crypts that underwent whole genome (**b**) and targeted (**c**) sequencing. Crypts are ordered by their mean depth (shown below), and for each crypt the proportion that is covered by a certain read depth is shown as a stacked barplot. **d-e**, their respective VAF (which is half of the clonal fraction). **f-h**, the distribution of coverage over exonic regions of putative colorectal cancer driver genes, from combining both whole genome and targeted data. **f,** the mean coverage across all samples of each gene. **g**, the number of crypts in which each gene was covered by an average of >7 reads, and **h**, the number of crypts in which each gene was covered by an average of >9 reads. Please note that the ordering of genes in each figure is different.

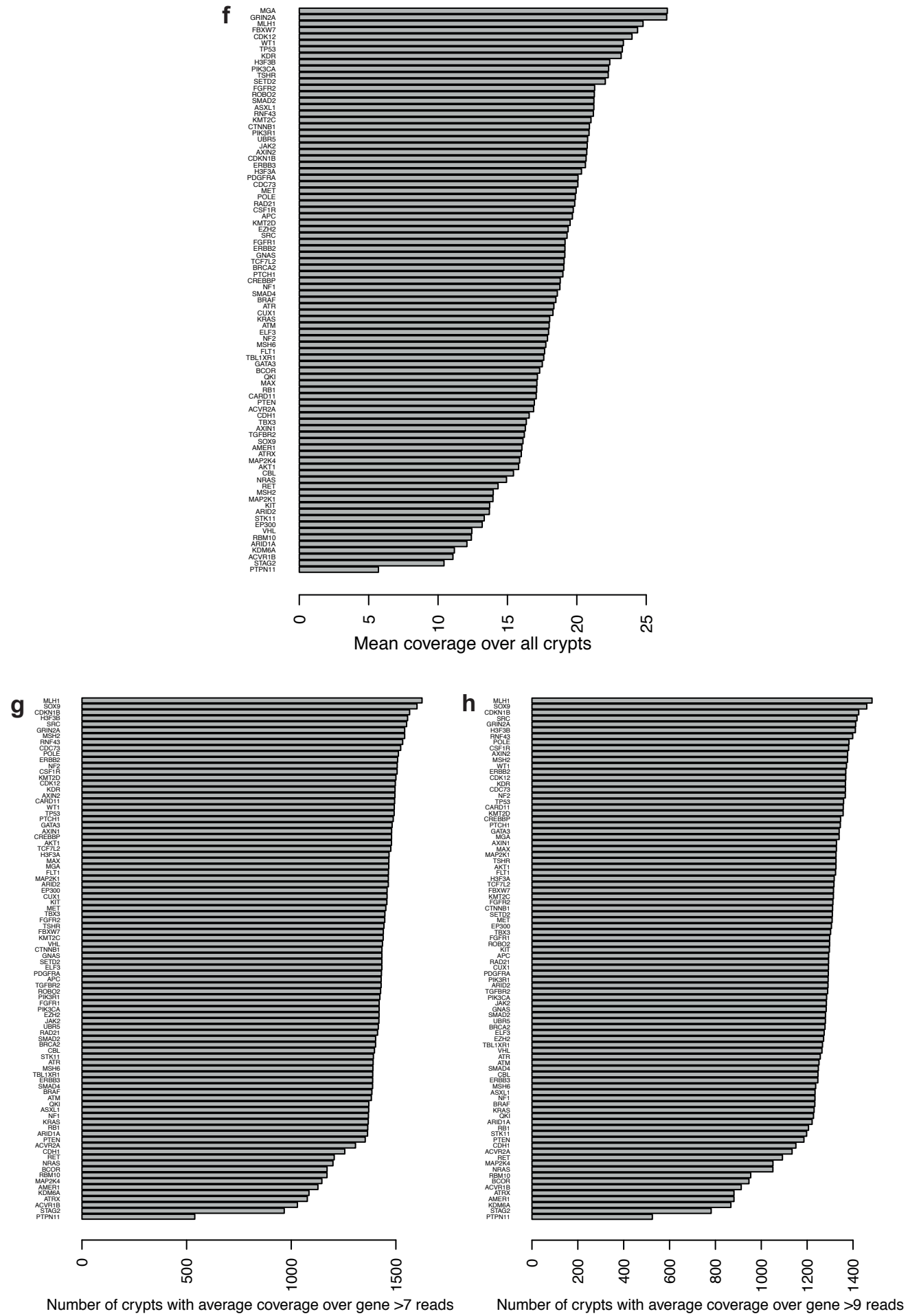

a


b
Proportion of genome
Covered by
>=1X
>=11X
>=21X
>=31X
>=41X
Mean coverage
Ranked sample


c
Proportion of targeted region
Covered by
>=1X
>=11X
>=21X
>=41X
>=101X
Mean coverage
Ranked sample


d
Median VAF WGS crypts
Frequency
Median VAF


e
Median VAF targeted crypts
Frequency
Median VAF


f
MGA
GRIN2A
MLH1
FBXW7
CDK12
WT1
TP53
KDR
H3F3B
PIK3CA
TSHR
SETD2
FGFR2
ROBO2
SMAD2
ASXL1
RNF43
KMT2C
CTNNB1
PIK3R1
UBR5
JAK2
AXIN2
CDKN1B
ERBB3
H3F3A
PDGFRA
CDC73
MET
POLE
RAD21
CSF1R
APC
KMT2D
EZH2
SRC
FGFR1
ERBB2
GNAS
TCF7L2
BRCA2
PTCH1
CREBBP
NF1
SMAD4
BRAF
ATR
CUX1
KRAS
ATM
ELF3
NF2
MSH6
FLT1
TBL1XR1
GATA3
BCOR
QKI
MAX
RB1
CARD11
PTEN
ACVR2A
CDH1
TBX3
AXIN1
TGFBR2
SOX9
AMER1
ATRX
MAP2K4
AKT1
CBL
NRAS
RET
MSH2
MAP2K1
KIT
ARID2
STK11
EP300
VHL
RBM10
ARID1A
KDM6A
ACVR1B
STAG2
PTPN11
0
5
10
15
20
25
Mean coverage over all crypts
g
MLH1
SOX9
CDKN1B
H3F3B
SRC
GRIN2A
MSH2
RNF43
CDC73
POLE
ERBB2
NF2
CSF1R
KMT2D
CDK12
KDR
AXIN2
CARD11
WT1
TP53
PTCH1
GATA3
AXIN1
CREBBP
AKT1
TCF7L2
H3F3A
MAX
MGA
FLT1
MAP2K1
ARID2
EP300
CUX1
KIT
MET
TBX3
FGFR2
TSHR
FBXW7
KMT2C
VHL
CTNNB1
GNAS
SETD2
ELF3
PDGFRA
APC
TGFBR2
ROBO2
PIK3R1
FGFR1
PIK3CA
EZH2
JAK2
UBR5
RAD21
SMAD2
BRCA2
CBL
STK11
ATR
MSH6
TBL1XR1
ERBB3
SMAD4
BRAF
ATM
QKI
ASXL1
NF1
KRAS
RB1
ARID1A
PTEN
ACVR2A
CDH1
RET
NRAS
BCOR
RBM10
MAP2K4
AMER1
KDM6A
ATRX
ACVR1B
STAG2
PTPN11
0
500
1000
1500
Number of crypts with average coverage over gene >7 reads
h
MLH1
SOX9
CDKN1B
SRC
GRIN2A
H3F3B
RNF43
POLE
CSF1R
AXIN2
MSH2
WT1
ERBB2
CDK12
KDR
CDC73
NF2
TP53
CARD11
KMT2D
CREBBP
PTCH1
GATA3
MGA
AXIN1
MAX
MAP2K1
TSHR
AKT1
FLT1
H3F3A
TCF7L2
FBXW7
KMT2C
FGFR2
CTNNB1
SETD2
MET
EP300
TBX3
FGFR1
ROBO2
KIT
APC
RAD21
CUX1
PDGFRA
PIK3R1
ARID2
TGFBR2
PIK3CA
JAK2
GNAS
SMAD2
UBR5
BRCA2
ELF3
EZH2
TBL1XR1
VHL
ATR
ATM
SMAD4
CBL
ERBB3
MSH6
ASXL1
NF1
BRAF
KRAS
QKI
ARID1A
RB1
STK11
PTEN
CDH1
ACVR2A
RET
MAP2K4
NRAS
RBM10
BCOR
ACVR1B
ATRX
AMER1
KDM6A
STAG2
PTPN11
0
200
400
600
800
1000
1200
1400
Number of crypts with average coverage over gene >9 reads

that have been developed for cancer genomes. Subsequently, bioinformatic filters were developed by Mathijs Sanders to remove artefacts that are due to this library preparation method (Methods).

It should be noted that some wells are empty because the crypt does not fall into the well. Electrostatic attractions often cause a dissected segment to stick to the underside of the slide. Visual inspection of adjacent wells never showed a crypt that had gone into the wrong well. In general, 30-40% of dissected crypts resulted in libraries with over 5ng/ul, which was chosen as the threshold to proceed to sequencing for most experiments.

R.3. Samples

Samples were obtained from four cohorts in order to cover as broad an age span as possible (Methods). 42 patients aged 11 to 78, 27 of whom had no diagnosis of colorectal disease and 15 of whom had been found to have a colorectal adenocarcinoma, were investigated. Wherever possible, biopsies from the caecum, transverse, and sigmoid colon were taken, as well as terminal ileum in a subset of cases. From these samples I dissected >5,000 crypts, of which 2,035 were sequenced: 571 were whole genome sequenced at ~15X coverage (figure 2.2b), and 1,464 underwent targeted sequencing using a bait-set of known cancer genes. Inspection of the allele fractions from the whole genomes showed that most crypts were 80-90% clonal (figure 2.2d), with some contamination which is likely to be stromal. The clonality of crypts that underwent targeted sequencing should be the same, but its assessment is less accurate as few mutations are called per genome (figure 2.2e). Targeted sequencing is more sensitive to low amounts of input DNA because of the additional step of bait hybridisation. Targeted sequencing on such small quantities of DNA was at the limit of current technical capabilities and variable coverage was achieved (figure 2.2c). CaVEMan (the algorithm used to call substitutions) requires – in most cases – three reads to call a mutation in a diploid genome. With a depth of eight reads, a clonal sample will achieve this many mutant reads 85% of the time (based on the binomial distribution). We therefore considered that only sites covered by at least eight reads could be genotyped with acceptable accuracy. Pindel (the algorithm used to call small insertions and deletions), however, requires four mutant reads to call a mutation. With a depth of 10 reads, a clonal sample will achieve this many mutant reads 83% of the time, and so we only considered that we could accurately genotype indels where sites were

covered by at least 10 reads. Because of the different coverage requirements for calling substitutions vs small indels, we estimate that we were adequately powered to call substitutions in 1,403 and indels in 1,046 of all crypts (genomes and targeted combined).

R.4. Driver mutations

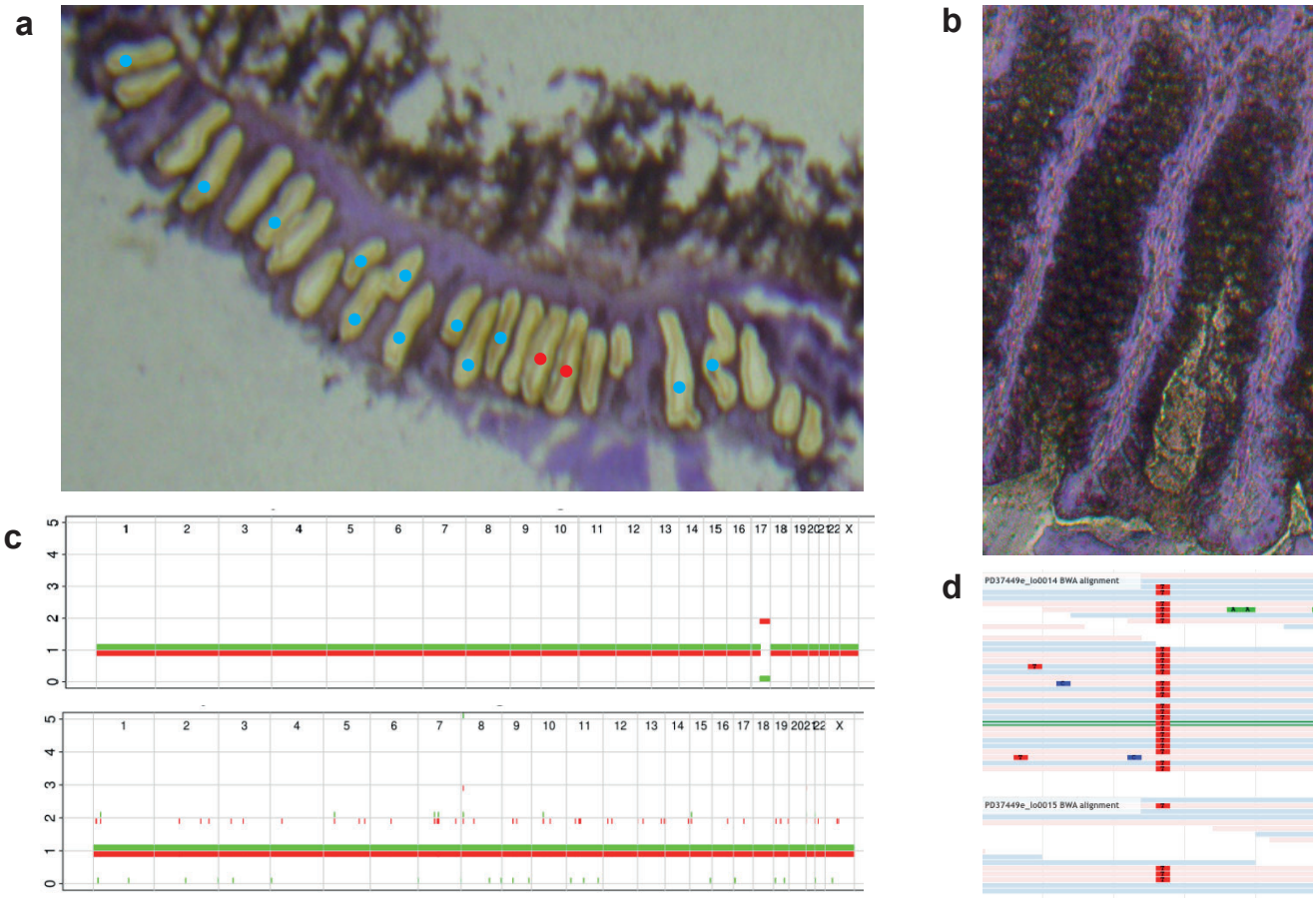
Two approaches were taken to discover driver mutations in normal crypts: first a modelling method to detect positive or negative selection of genes, and second, manual annotation based on prior knowledge.

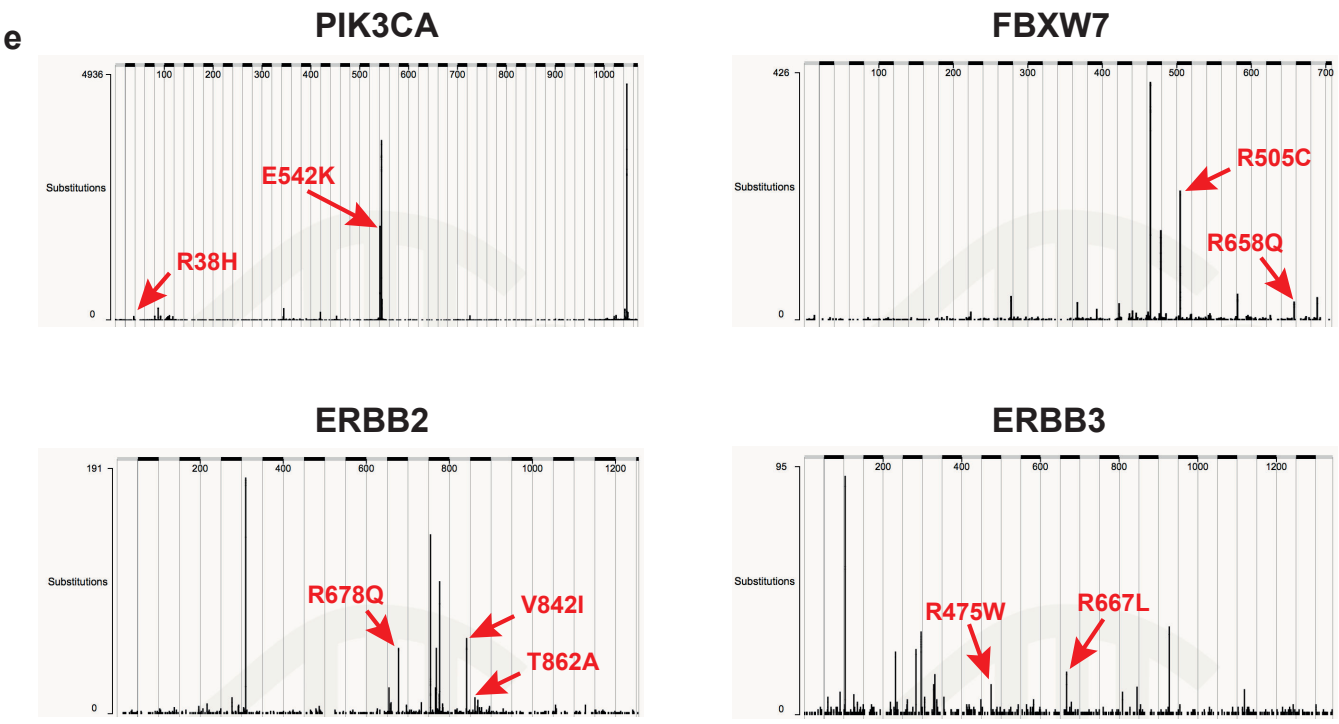
A dNdS approach (section I.2.b.ii., Methods, Universal patterns) was used to detect positive or negative selection. Two separate analyses were performed: a genome-wide analysis using only the whole genomes, and an analysis restricted to 90 putative colorectal cancer driver genes (Appendix B) that were included in our bait-set, using combined genomes and targeted sequences. In both analyses, the 95% CI for the global dNdS spanned 1, which indicates that the vast majority of the mutations in our dataset are selectively neutral. In the genome-wide analysis, no genes were significantly mutated. In the analysis of 90 genes, however, there was evidence of positive selection of two genes: *AXIN2* (three truncating mutations, adjusted p value 0.004), and *STAG2* (two truncating mutations, adjusted p value 0.038).

AXIN2 is a negative regulator of the WNT signalling pathway (section I.2.b.i.). *AXIN2* is inactivated in 2.3% of colorectal adenocarcinomas and smaller proportions of other cancer types (Forbes et al. 2017). One of the three *AXIN2* nonsense mutations was present in two adjacent crypts that were closely related genetically, sharing 1,606 SBS1 mutations; the *AXIN2* mutation must have occurred in their common ancestor which then underwent crypt fission. In one sister crypt, but not the other, the *AXIN2* mutation was rendered homozygous by copy number neutral loss of heterozygosity of chromosome 17q (figure 2.3a-d). This suggests that while loss of one copy of *AXIN2* already confers a growth advantage and may have contributed to the crypt fission, loss of a second copy could provide a further advantage and aid the expansion of the mutant clone, as has been shown for other tumour suppressors in mouse models (section I.5.a.). This provides evidence for ongoing clonal evolution in normal colon.

STAG2 is a component of the cohesin complex, which has roles in sister chromatid cohesion, DNA repair, and regulation of gene expression and chromatin structure (Hill et al. 2016). Although *STAG2* loss has been associated with aneuploidy in solid tumours (Solomon et al. 2011), this is not always the case (Hill et al. 2016, Balbas-Martinez et al. 2013, Taylor et al. 2014), and we do not observe this here. It is inactivated in 0.9% of colonic adenocarcinomas, and more frequently in other tumour types (Forbes et al. 2017). In our dataset, both *STAG2* nonsense mutations occurred in men, so no wild type copies of this X chromosome recessive cancer gene would remain in these cells. STAG2 loss has previously been shown to confer a proliferative advantage in human colon (section I.5.).

Figure 2.3. Driver mutations in normal colon. a-d, driver mutation in *AXIN2*. **a,** a section (after dissection) in which an inactivating *AXIN2* mutation was found. Red dots represent crypts with the *AXIN2* mutation. Blue dots represent crypts that could be assessed and were found not to have the mutation. Crypts without dots failed sequencing and could not be assessed. **b**, the two crypts with the *AXIN2* mutations prior to dissection did not appear different to any other crypts. **c**, copy neutral loss of heterozygosity (CNN-LOH) of one of the crypts 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 in each of the crypts. The mutation is coloured red. 25 out of 29 reads support the mutation in the crypt that has CNN-LOH; the four reads that do not are presumably the result of stromal contamination. **e**, putative driver missense mutations in oncogene hotspots. The number of substitutions catalogued in COSMIC (Forbes et al. 2017) are shown on the y axis at each position along the gene, with the mutations observed in our cohort highlighted.


a
b
c
d
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
PD37449e_lo0014 BWA alignment
PD37449e_lo0015 BWA alignment


e
PIK3CA
Substitutions
E542K
R38H
FBXW7
R505C
R658Q
ERBB2
R678Q
V842I
T862A
ERBB3
R475W
R667L

Additional potential driver mutations were identified by manual curation based on the known cancer genes in colorectal cancer and their distinctive patterns of mutation. Nine canonical missense hotspot mutations in the dominantly acting cancer genes *PIK3CA* (E542K, R38H (a minor hotspot)), *ERBB2* (R678Q, V842I, T862A), *ERBB3* (R475W, R667L), and *FBXW7* (R505C, a major hotspot, and R658Q, a minor hotspot) were observed (figure 2.3e). Given the specificity of these mutation hotspots, the majority of these are likely to be driver mutations and confer some growth advantage. As with *AXIN2*, the *PIK3CA* E542K mutation was also in two adjacent crypts which shared 2,516 SBS1 mutations and had 93 and 208 private SBS1 mutations, implying a recent crypt fission event. Indeed, these two crypts shared more mutations than any other pair of crypts in our dataset.

Finally, a series of heterozygous truncating mutations in the recessive colorectal cancer genes *ARID2*, *ATM* (two mutations), *ATR*, *BRCA2*, *CDK12* (two mutations), *CDKN1B*, *RNF43* (two mutations), *TBL1XR1*, and *TP53* were found. They were not associated with loss of heterozygosity, and no crypt had more than one driver mutation. It is likely that some of these did not confer any selective advantage. Nonetheless, in mouse colon heterozygous nonsense mutations of the *Apc* tumour suppressor gene can confer a selective advantage (Vermeulen et al. 2013), and indeed the *AXIN2* mutations for which we have compelling evidence of driver function were mostly heterozygous. Even if not currently advantageous, these mutations could set the scene for future clonal expansions, through loss of the remaining wild type copy or a change in microenvironment as has been observed for *P53* (General Introduction, section I.5.a. of this chapter).

On the basis of these findings, treating the *AXIN2*, *STAG2*, and dominant cancer gene hotspot mutations as drivers (i.e. all but the heterozygous mutations in recessive cancer genes) we estimate that at least 1% of normal colorectal crypts in a 50-60 year old carries a driver mutation. We are underpowered to detect a change in driver frequency with age. Since there are ~15 million crypts in the colon, ~150,000 crypts carry a driver. In the over 70s, ~40% of people have an adenoma on colonoscopy (Corley et al. 2013) and ~5% of people develop colorectal cancer over their lifetime (Cancer Research UK bowel cancer incidence statistics), and some of these may arise from more recently-acquired driver mutations. Therefore, only an extremely small proportion of these crypts with driver mutations becomes a macroscopically detectable adenoma (< 1/375,000) or carcinoma (< 1/3,000,000) within the following few decades.

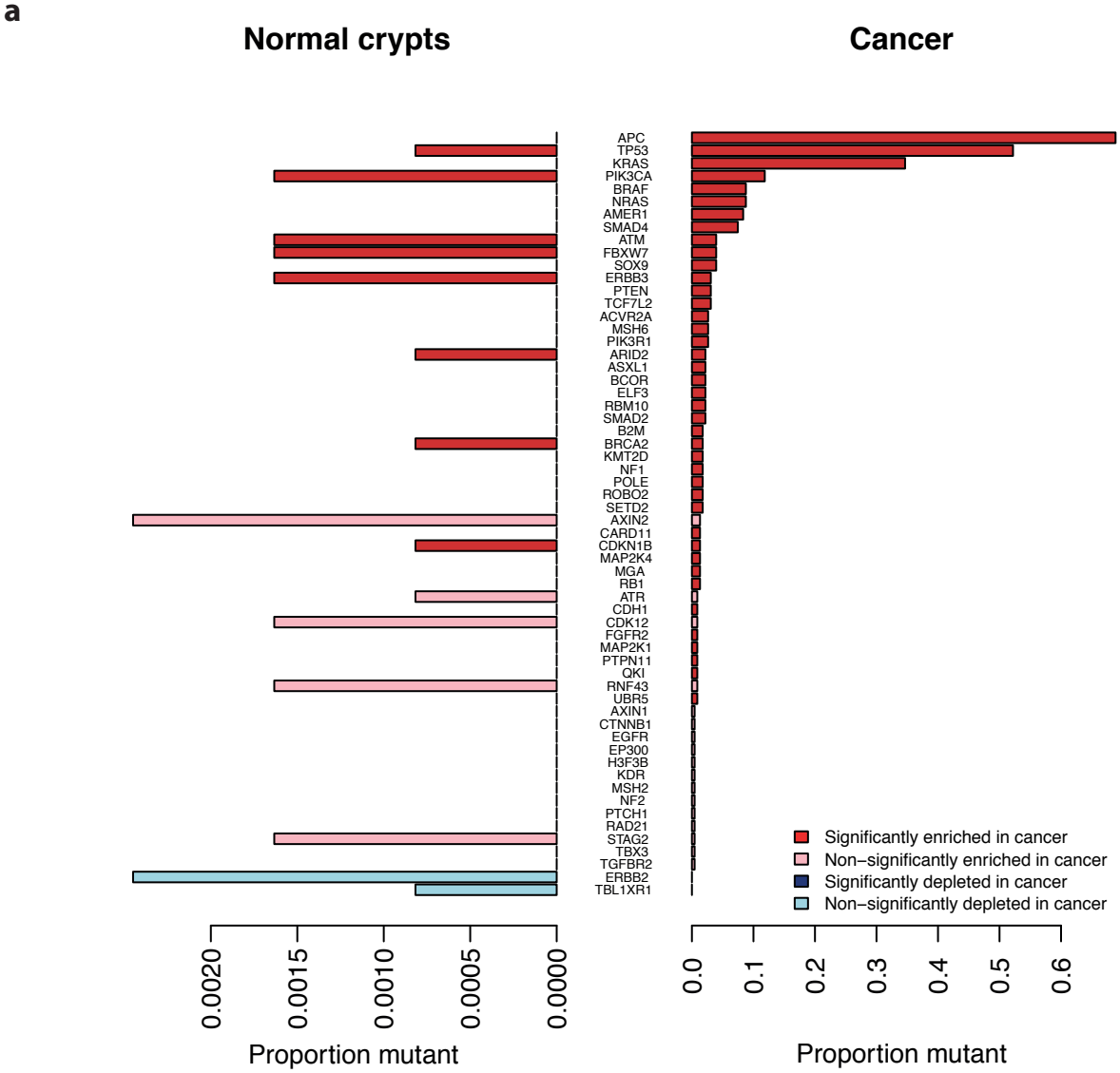
Given that they are so common in colorectal cancer and absent from our dataset, we might conjecture that crypts with driver mutations in *APC* or *KRAS* as a first hit might have a higher chance of progressing. Using the mutation rate observed in our data, we can estimate that even if they were selectively neutral, 12,000 crypts from a 50-60 year old would have inactivated one copy of *APC*, and five crypts would have both copies inactivated. As even heterozygous *Apc* mutations confer a selective advantage in mouse models (Vermeulen et al. 2013), the true frequency is likely to be higher. PCR-based analysis of bulk epithelium has shown that 1 in 3,500 epithelial cells bears *KRAS* G12D (Parsons et al. 2010), which would indicate a few thousand crypts with this mutation per colon. Thus, even for these mutations, the probability of progression to cancer must remain low.

Comparison of the frequency of particular cancer gene mutations between normal epithelium and colorectal cancers extends the adenoma-carcinoma sequence and informs on the properties of driver mutations. Mutations reported in 260 cancers (Cancer Genome Atlas Research Network et al. 2012) were annotated for driver mutations using the same criteria as for the manual annotation of driver mutations in normal tissues. The pattern of driver mutations is different between cancer and normal tissues ($p=0.003$ by randomization test, figure 2.4). In colorectal cancer, mutations in *APC*, *KRAS* and *TP53* are common, accounting for 56% of base substitution and indel drivers but are comparatively rare among normal crypts with driver mutations (1 out 14 drivers). By contrast, mutations in, for example, *ERBB2* and *ERBB3* are relatively common in normal crypts with drivers (5/14) but rare in colorectal cancer (7/631). It is, therefore, not simply that the genes found in cancers are found at lower frequencies in normal tissues, but rather some account for a greater proportion of driver mutations in cancer than they do in normal tissues.

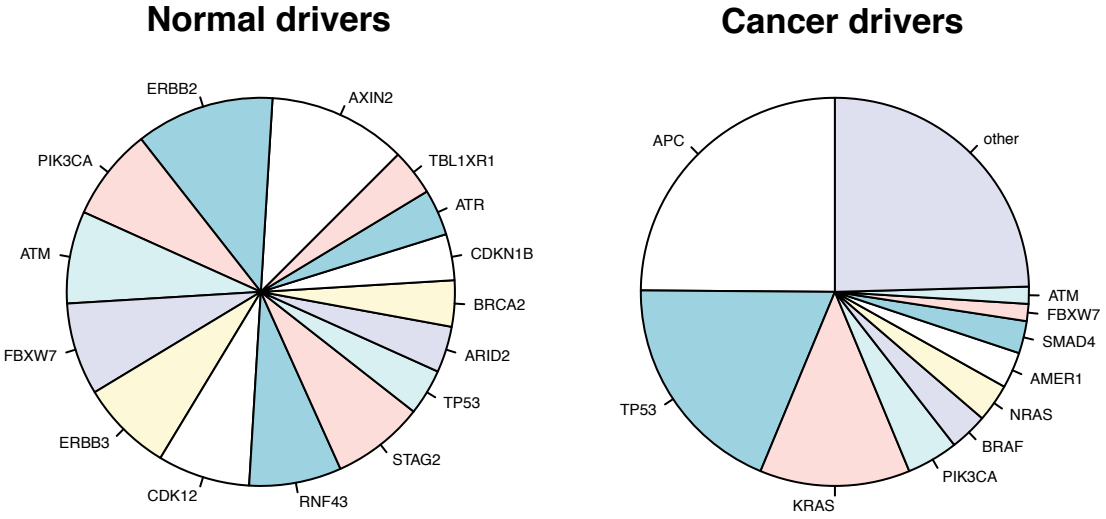
A mutation may be enriched in a cancer for four reasons: (1) the mutation itself has itself promoted progression of the lesion; (2) the mutation provides a selective advantage only in a neoplastic microenvironment; (3) the mutation occurs more frequently in the cancer due to an increased mutation rate; or (4), the mutation itself increases the mutation rate and so the probability of one of its descendants acquiring a mutation with properties (1) or (2). Mutations that provide a selective advantage in normal tissues but do not confer any of the above properties will be found at equal frequencies in cancer and in normal tissue.

Figure 2.4. Comparison of driver frequency in normal colon and colorectal cancer. The frequency of driver mutations in colorectal cancer is derived using data from Cancer Genome Atlas Network (2012). **a**, the proportion of crypts or cancers with driver mutations in each gene found in either of the two groups. **b**, the proportion of driver mutations in each gene in normal and cancer.

a


Normal crypts
Cancer
APC
TP53
KRAS
PIK3CA
BRAF
NRAS
AMER1
SMAD4
ATM
FBXW7
SOX9
ERBB3
PTEN
TCF7L2
ACVR2A
MSH6
PIK3R1
ARID2
ASXL1
BCOR
ELF3
RBM10
SMAD2
B2M
BRCA2
KMT2D
NF1
POLE
ROBO2
SETD2
AXIN2
CARD11
CDKN1B
MAP2K4
MGA
RB1
ATR
CDH1
CDK12
FGFR2
MAP2K1
PTPN11
QKI
RNF43
UBR5
AXIN1
CTNNB1
EGFR
EP300
H3F3B
KDR
MSH2
NF2
PTCH1
RAD21
STAG2
TBX3
TGFBR2
ERBB2
TBL1XR1
Significantly enriched in cancer
Non−significantly enriched in cancer
Significantly depleted in cancer
Non−significantly depleted in cancer
0.0020
0.0015
0.0010
0.0005
0.0000
Proportion mutant
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Proportion mutant

b


Normal drivers
ERBB2
AXIN2
PIK3CA
TBL1XR1
ATR
ATM
CDKN1B
BRCA2
FBXW7
ARID2
TP53
ERBB3
STAG2
CDK12
RNF43
Cancer drivers
APC
other
ATM
FBXW7
SMAD4
AMER1
TP53
NRAS
BRAF
PIK3CA
KRAS

Thus, *APC*, *KRAS*, and *TP53* mutations, which are present orders of magnitude more frequently in cancer than in normal tissues (figure 2.4a) will confer properties (1)-(4), while mutations in *ERBB2* or *STAG2*, which are not significantly enriched in cancers, may merely provide a selective advantage in normal tissues without promoting cancer development. The distinction must therefore be drawn between mutations that are under positive selection and those that actually promote cancer development, although there will be significant overlap between the two categories.

Uneven coverage over the genome presents some difficulties when estimating the frequency of driver mutations. Ideally, every locus in every gene would be covered by many reads in every crypt, and the true frequency of driver mutations would need no correction. Second best, all genes would be covered equally well, but all genes would not be callable in a certain proportion of crypts. The frequency of driver mutations could then be estimated by dividing the number of drivers by the number of crypts in which they were callable. Third best, coverage would be even within each gene, but some genes would be better covered than others. If there were a sufficient frequency of drivers in each gene, we could estimate the true frequency of drivers per gene. Let us imagine that gene A achieves sufficient coverage to call mutations in 1,000 crypts and 10 drivers are found in it, and gene B achieves sufficient coverage to call mutations in 2,000 crypts and 100 drivers are found in it, we would say that the 1% of crypts have gene A drivers and 5% of crypts have gene B drivers. Assuming that these were occurring in different crypts, we would conclude that 6% of crypts have driver mutations.

The first challenge of our data is that the driver frequency is so low that we cannot estimate a per-gene driver frequency. All we can do to derive a meaningful estimate is to pool our driver mutations. Thus, if we can detect one gene A and one gene B driver mutation in our cohort, and the mean number of crypts in which we can call mutations accurately over all base pairs is 1,500, our best estimate is that two in 1,500 crypts has a driver mutation.

The second challenge is that 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. All we can do here is treat different *parts of a gene* with different coverage like the different *genes* in the section above; that is to say, to take an average of the number of crypts in which they are covered. The approach that we took, therefore, 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.

Note that for this reason of uneven coverage, this approach is less suitable to estimating the frequency of mutations in a given gene. Particular driver mutations, may be under-represented in our cohort. If one part of gene A is covered by a very low number of crypts, and that it is this part where most driver mutations occur, we will underestimate the frequency of driver mutations in gene A. Similarly, other driver mutations may be over-represented. This should be borne in mind when considering figure 2.4. The true frequencies of driver mutations in these genes may, in time, reveal themselves to be different to those that we have estimated from imperfect data here. We nonetheless present this figure as a preliminary indication of the landscape of driver mutations in normal colon.

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. Without prior knowledge of which are under-represented and which are over-represented, using a mean is a valid approach. We stress the highly simplified nature of this approach. It is our resort because the frequency of driver mutations in our dataset is so low. The value that we derive of 1% of colonic crypts bearing a driver should be taken as a first ballpark estimate to guide further investigation. Further studies of larger numbers of crypts will be required to achieve greater accuracy.

R.5. Mutational processes and rates in the colons of different people

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 from the whole genomes and estimated their contribution to the mutation burden of each crypt.

We first derived phylogenies of how crypts from each patient were related to one another and assigned every mutation to a branch of a phylogeny. This allowed us to treat every branch as a sample in signature extraction. This has the dual advantage of avoiding double counting of mutations in signature extraction (as mutations shared between two crypts are only counted once with this method, while they would be counted twice if every crypt were treated as a sample), and of allowing us to time mutational signatures over the course of life, since mutations shared by two samples must have occurred before mutations that are private to one of them.

The mutational signatures that are extracted from an analysis are dependent on the samples that went into it. If all samples have perfectly correlated contributions of different processes, these will only be extracted as one signature. A cohort of normal genomes from a single tissue runs this risk. We wanted to be able to frame our signature extraction results in the context of previous work in cancers, in order to allow comparisons with different studies. We therefore performed a signature extraction using a hierarchical Dirichlet process (HDP) (Roberts et al. 2015, Roberts et al. 2018), providing the algorithm with the catalogue of mutational signatures extracted from the Pan Cancer Analysis of Whole Genomes (PCAWG) (Alexandrov et al. 2018). 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 (figure 2.5). 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 is hereafter presented as its expectation maximisation deconvolution (Methods). 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 DBSN1 was deconvoluted in PCAWG DBS2, PCAWG DBS4, PCAWG DBS6, PCAWG DBS9, and PCAWG DBS11; and the HDP IDN1 was deconvoluted into PCAWG ID1, PCAWG ID2, and PCAWG ID5 (figure 2.6). 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 (2016), and a version of the NNMF algorithm used by Alexandrov and colleagues (Alexandrov et al. 2018). These all produced comparable results (figure 2.8).

Following expectation maximisation deconvolution, we found nine SBS, six DBS, and five ID signatures in our dataset (figure 2.7). Of these, 14 were known and six were novel (denoted by an "N" in their nomenclature, e.g. "SBSN1"). The signatures can be divided into those that are common and those that are rare depending on whether they are present in more or less than 85% of crypts. The common signatures are: SBS1, SBS5, SBS18, DBS2, DBS4, DBS6, DBS9, DBS11, ID1, ID2, and ID5. The rare signatures are: SBS2, SBS13, four novel SBS signatures (SBSN1 – SBSN4), DBS8, and two novel ID signatures (IDN2 and IDN3). The correlation with age of every signature is shown in figure 2.10.

The mutational signatures extracted are dependent both on the process causing mutations and on the trinucleotide composition of the genome. As variable coverage was achieved across the genome of crypts (figure 2.2), it was theoretically possible that some of our lower coverage crypts may be altering the profile of the signatures that were extracted. Mutational signatures were historically extracted from exome data (Alexandrov et al. 2013), and when they were updated to include genomes (Alexandrov et al. 2018), mutational processes largely remained the same. Those that changed in profile did so as a result of including additional samples with very simple trinucleotide profiles that allowed the NMF algorithm to draw out the salient features of the signature more clearly, rather than because of including other parts of the genome with a different trinucleotide composition. This would suggest that the trinucleotide profile of parts of the genome that are less well covered in some crypts would have less of an effect on the signatures extracted than one might expect. So as to detect changes in mutational signature composition that might be due to coverage, we ordered crypts by increasing coverage, and plotted the proportional contribution of different signatures to them (figure 2.9a). No obvious systematic differences with coverage are observed, and all novel signatures are seen in crypts with both good and bad coverage. Second, we compared the raw trinucleotide profiles of a representative selection of six of our lowest-depth crypts (all with average depth <10X), with the trinucleotide profile of six normal colonic organoids (see section R.7.), all of which were sequenced at 30-40X. There are no obvious differences between the trinucleotide profiles of these high- and low-coverage samples (figure 2.9b).

Figure 2.5. Signatures of mutational processes in normal colon. Signatures extracted by HDP are shown, with the trinucleotide context of a sample that contains a large proportion of the relevant signature shown underneath. Signatures are presented as in figure 2.1.

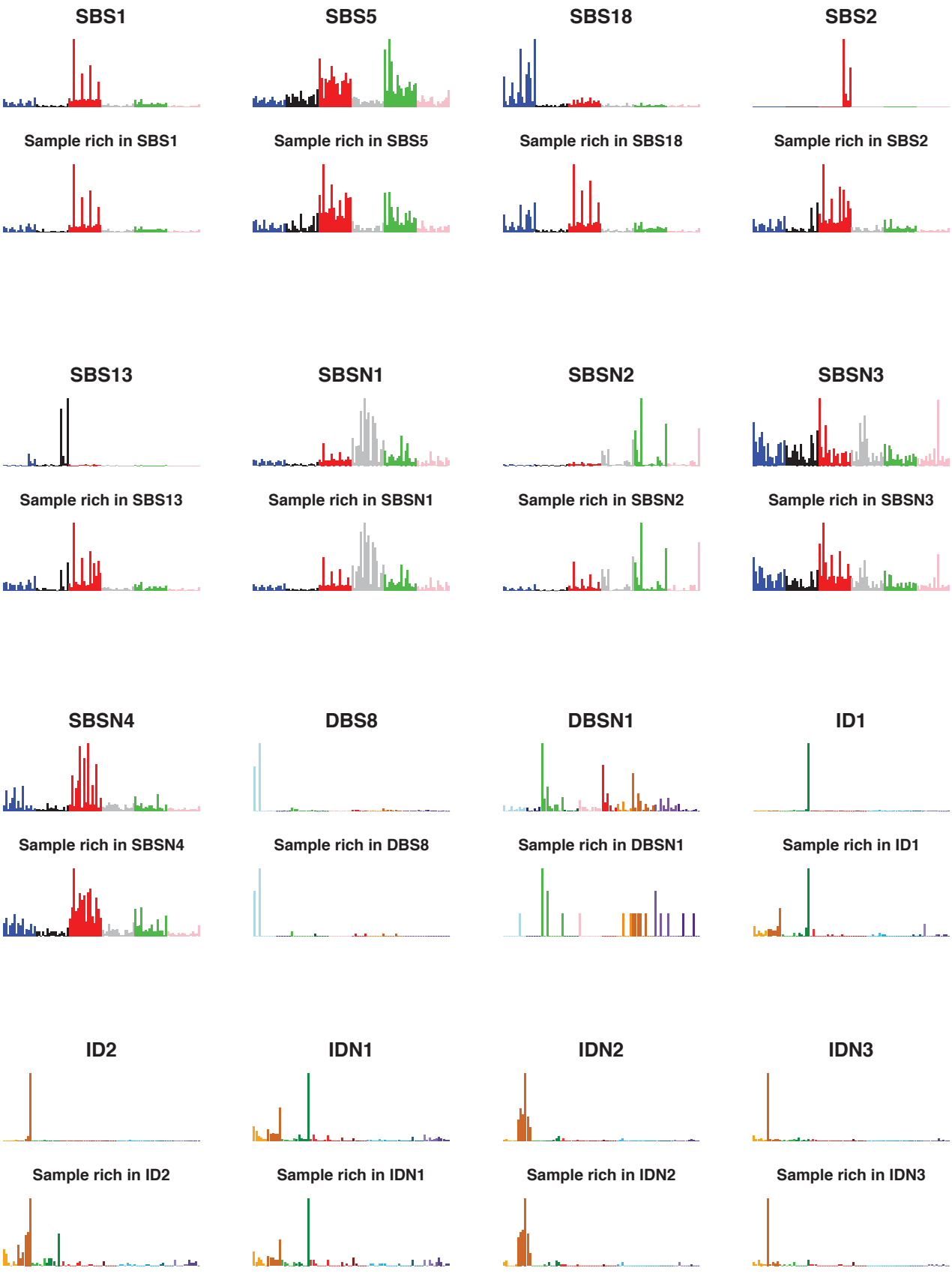

SBS1
SBS5
SBS18
SBS2
Sample rich in SBS1
Sample rich in SBS5
Sample rich in SBS18
Sample rich in SBS2
SBS13
SBSN1
SBSN2
SBSN3
Sample rich in SBS13
Sample rich in SBSN1
Sample rich in SBSN2
Sample rich in SBSN3
SBSN4
DBS8
DBSN1
ID1
Sample rich in SBSN4
Sample rich in DBS8
Sample rich in DBSN1
Sample rich in ID1
ID2
IDN1
IDN2
IDN3
Sample rich in ID2
Sample rich in IDN1
Sample rich in IDN2
Sample rich in IDN3

Figure 2.6. Expectation maximisation (EM) decomposition of HDP signatures. Three signatures were decomposed by EM. For each signature, the original HDP version in shown on the top left, the PCAWG signatures that are deemed by EM 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.

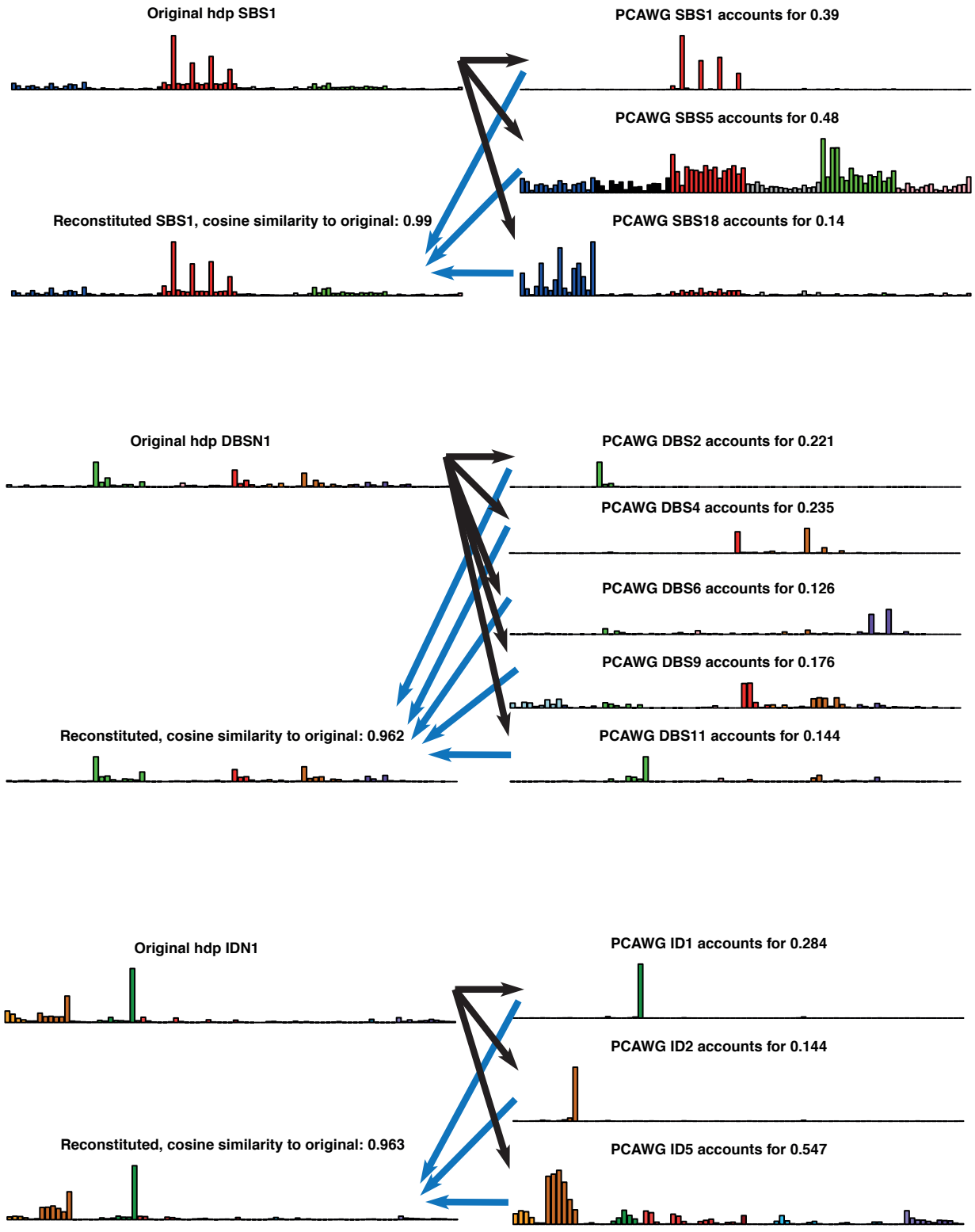

Original hdp SBS1
PCAWG SBS1 accounts for 0.39
PCAWG SBS5 accounts for 0.48
Reconstituted SBS1, cosine similarity to original: 0.99
PCAWG SBS18 accounts for 0.14
Original hdp DBSN1
PCAWG DBS2 accounts for 0.221
PCAWG DBS4 accounts for 0.235
PCAWG DBS6 accounts for 0.126
PCAWG DBS9 accounts for 0.176
Reconstituted, cosine similarity to original: 0.962
PCAWG DBS11 accounts for 0.144
Original hdp IDN1
PCAWG ID1 accounts for 0.284
PCAWG ID2 accounts for 0.144
Reconstituted, cosine similarity to original: 0.963
PCAWG ID5 accounts for 0.547

Figure 2.7. Final set of signatures in normal colon, following EM decomposition. These are the set that are used in analyses.

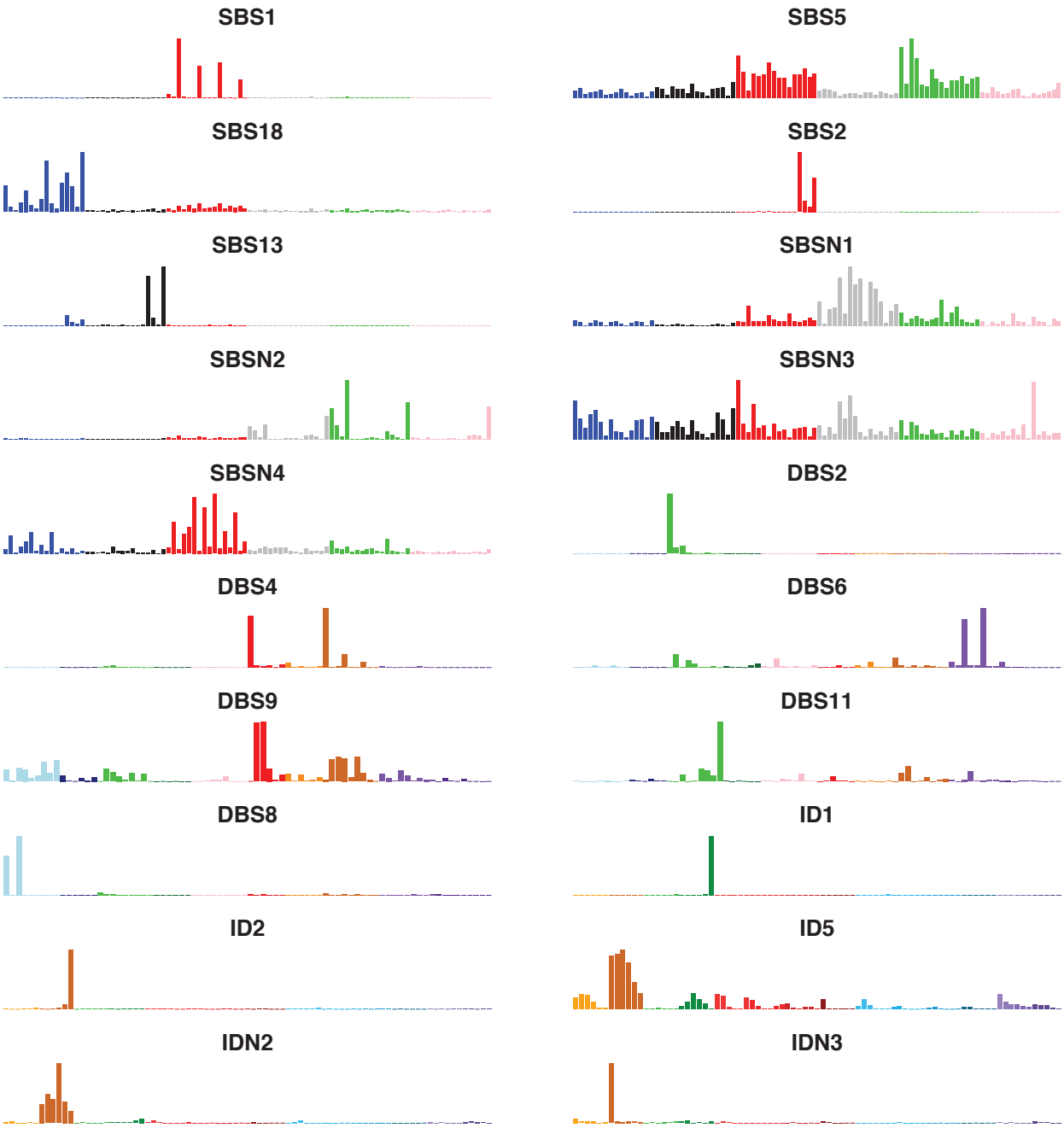

SBS1
SBS5
SBS18
SBS2
SBS13
SBSN1
SBSN2
SBSN3
SBSN4
DBS2
DBS4
DBS6
DBS9
DBS11
DBS8
ID1
ID2
ID5
IDN2
IDN3

Figure 2.8. 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 (Blokzijl et al. 2016).

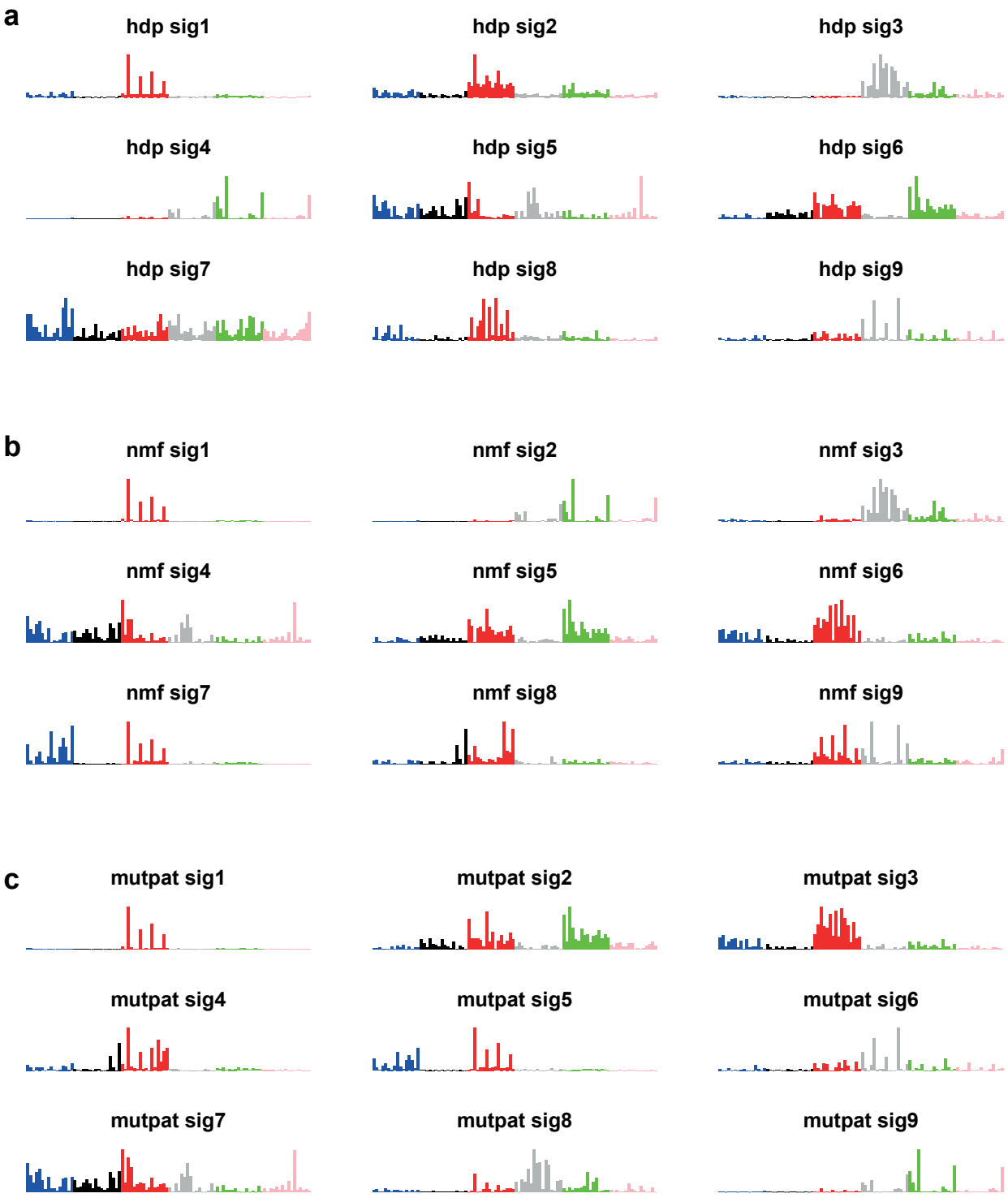
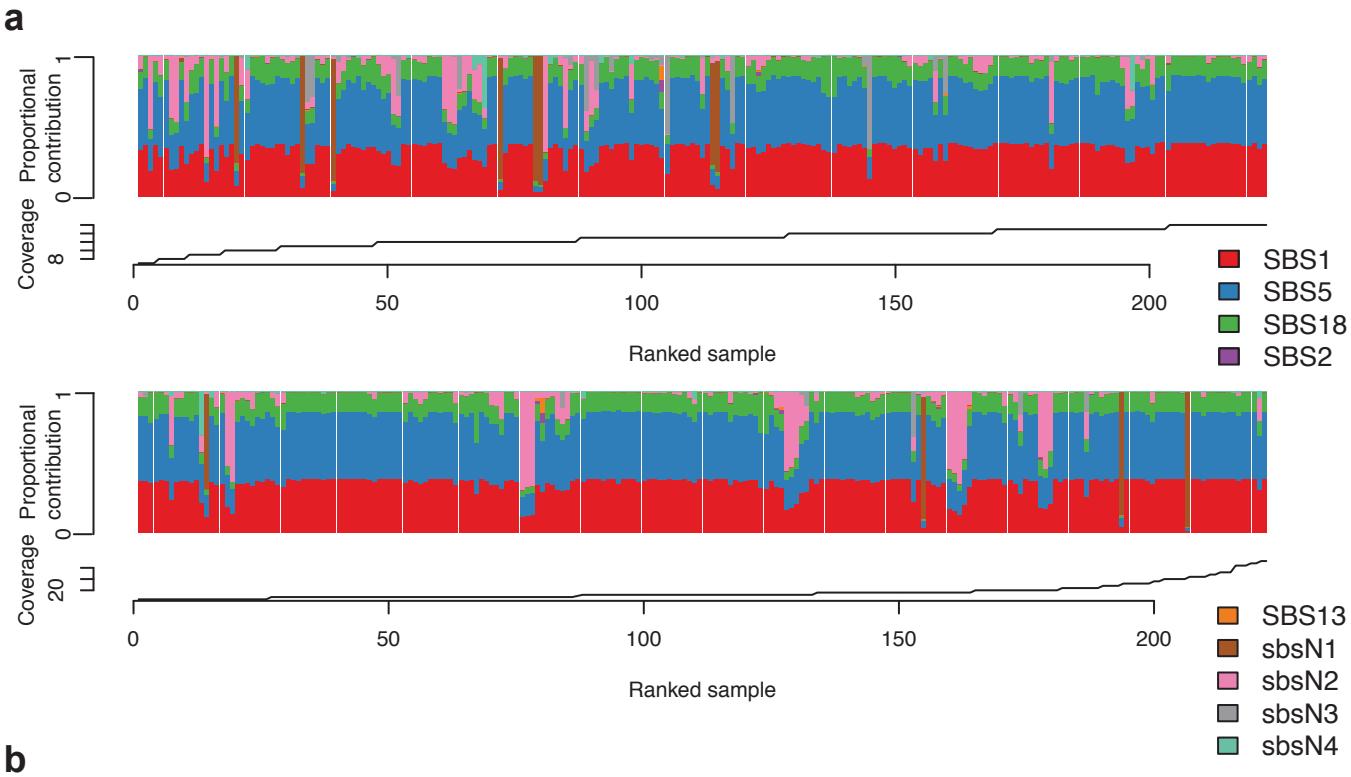

a
hdp sig1
hdp sig2
hdp sig3
hdp sig4
hdp sig5
hdp sig6
hdp sig7
hdp sig8
hdp sig9
b
nmf sig1
nmf sig2
nmf sig3
nmf sig4
nmf sig5
nmf sig6
nmf sig7
nmf sig8
nmf sig9
c
mutpat sig1
mutpat sig2
mutpat sig3
mutpat sig4
mutpat sig5
mutpat sig6
mutpat sig7
mutpat sig8
mutpat sig9

Figure 2.9. Mutational signatures with coverage. a, crypts are ranked by their median coverage, with coverage increasing from left to right. The proportional contribution of each signature is presented a stacked barplot. **b,** the trinucleotide profile of six crypts with high coverage are boxed in red, and the trinucleotide profile of six crypts with low coverage are boxed in blue. The high coverage samples are normal colonic organoids (see section R7) all sequenced at >30X, whereas the low coverage samples are laser capture microdissected crypts all with coverage <10X.


a
Proportional contribution
Coverage
Ranked sample
SBS1
SBS5
SBS18
SBS2
SBS13
sbsN1
sbsN2
sbsN3
sbsN4

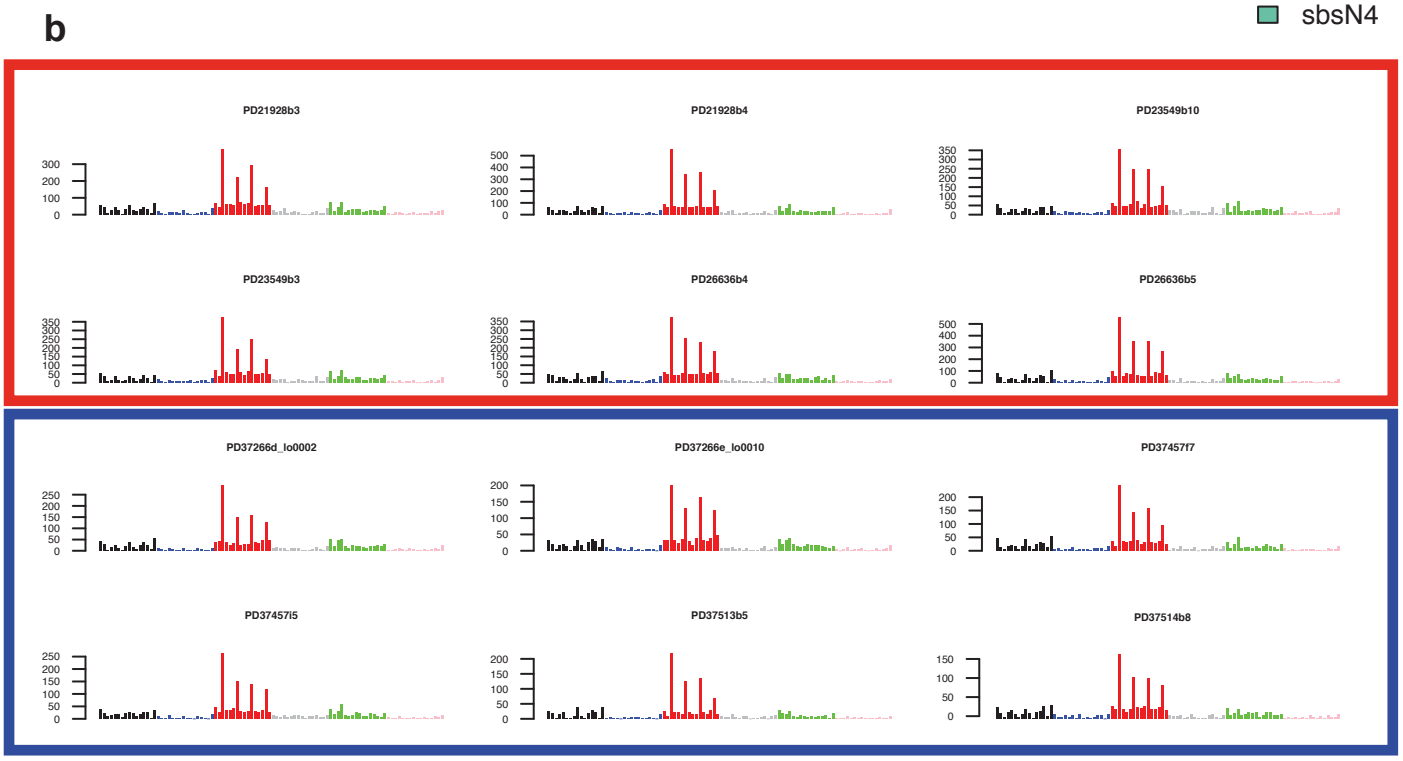
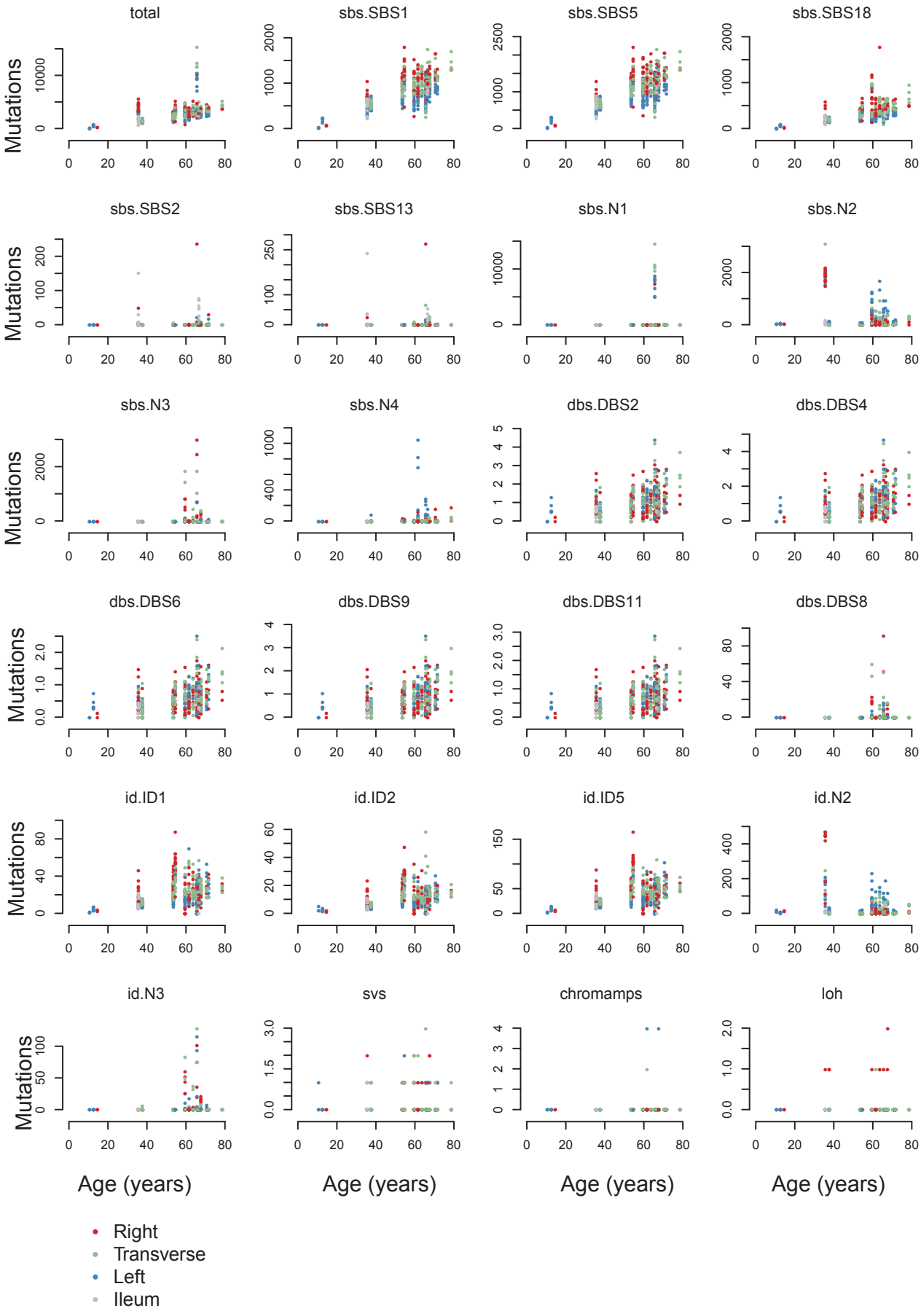

b
PD21928b3
PD21928b4
PD23549b10
PD23549b3
PD26636b4
PD26636b5
PD37266d_lo0002
PD37266e_lo0010
PD37457f7
PD37457l5
PD37513b5
PD37514b8

Figure 2.10. Mutation burden v. age for every signature. For each signature, the total number of mutations in every crypt attributed to each signature is plotted against the age of the patient from whom the crypt derives. Points are coloured by the anatomical location from of the crypt. Mutation burden is adjusted for the callable proportion of the genome (Methods section 12).


total
sbs.SBS1
sbs.SBS5
sbs.SBS18
sbs.SBS2
sbs.SBS13
sbs.N1
sbs.N2
sbs.N3
sbs.N4
dbs.DBS2
dbs.DBS4
dbs.DBS6
dbs.DBS9
dbs.DBS11
dbs.DBS8
id.ID1
id.ID2
id.ID5
id.N2
id.N3
svs
chromamps
loh
Mutations
Age (years)
Right
Transverse
Left
Ileum

R.5.a. Common mutational signatures

All three of SBS1, SBS5, and SBS18 were found in all crypts from all patients and correlated linearly with their age, indicating that the underlying mutational processes are continuously active in all individuals throughout life. There was, however, substantial variation in mutation burdens for each of these signatures between crypts from the same individual. These were partly explained by differences between sites in the colon: for all three signatures, mutations accumulated at the fastest rate in right colon (ascending and caecum), then transverse colon, and slowest in left colon (caecum and sigmoid). Terminal ileum accumulated mutations at a similar rate to the left colon (figure 2.11). Even taking site differences into account, however, there was still substantial spread of mutation burden, and in the case of SBS18 one strong outlier. For example, the mean ratio of SBS1 burden between the most and least mutated crypts from the same site of one person is 1:1.3, and the mean ratio between the SBS1 mutation burden of crypts from the same site in two people in their 60s is 1:1.4. The cause of this variability is unknown, but possible explanations include differences in cell division rate, exposure to locally acting mutagens, and variability in the time to the most recent common ancestor of each crypt. Interestingly, bromodeoxyuridine staining of colonic crypts showed substantial inter- and intrapatient variability in cell division rates (Potten et al. 1992). For SBS1 and SBS5, which appear to accumulate in a clock-like manner, the x axis intercept is of approximately five to 10 years of age. With a linear mutation rate the x axis intercept would represent the time to the most recent common ancestor of the crypt, and this is similar to previous estimates of the duration of monoclonal conversion in humans (section I.1.c.).

DBS 2, 4, 6, 8, and 11 were all extracted by HDP as one mutational processes that was present in almost all crypts and deconvoluted into these signatures by expectation maximization. These processes must therefore be tightly correlated in normal colonic crypts. As numbers of doublet base substitutions are small, we are underpowered to detect differences between sites.

Similarly, ID1, ID2, and ID5 were all extracted as one ubiquitous signature by HDP, and deconvoluted into their constituents by expectation maximization. They correlate with the age of diagnosis of the patient, which is consistent with their proposed aetiology in replication slippage, and show the same ordering of mutation burden between sites as SBS1, SBS5, and SBS18 (figure 2.11).

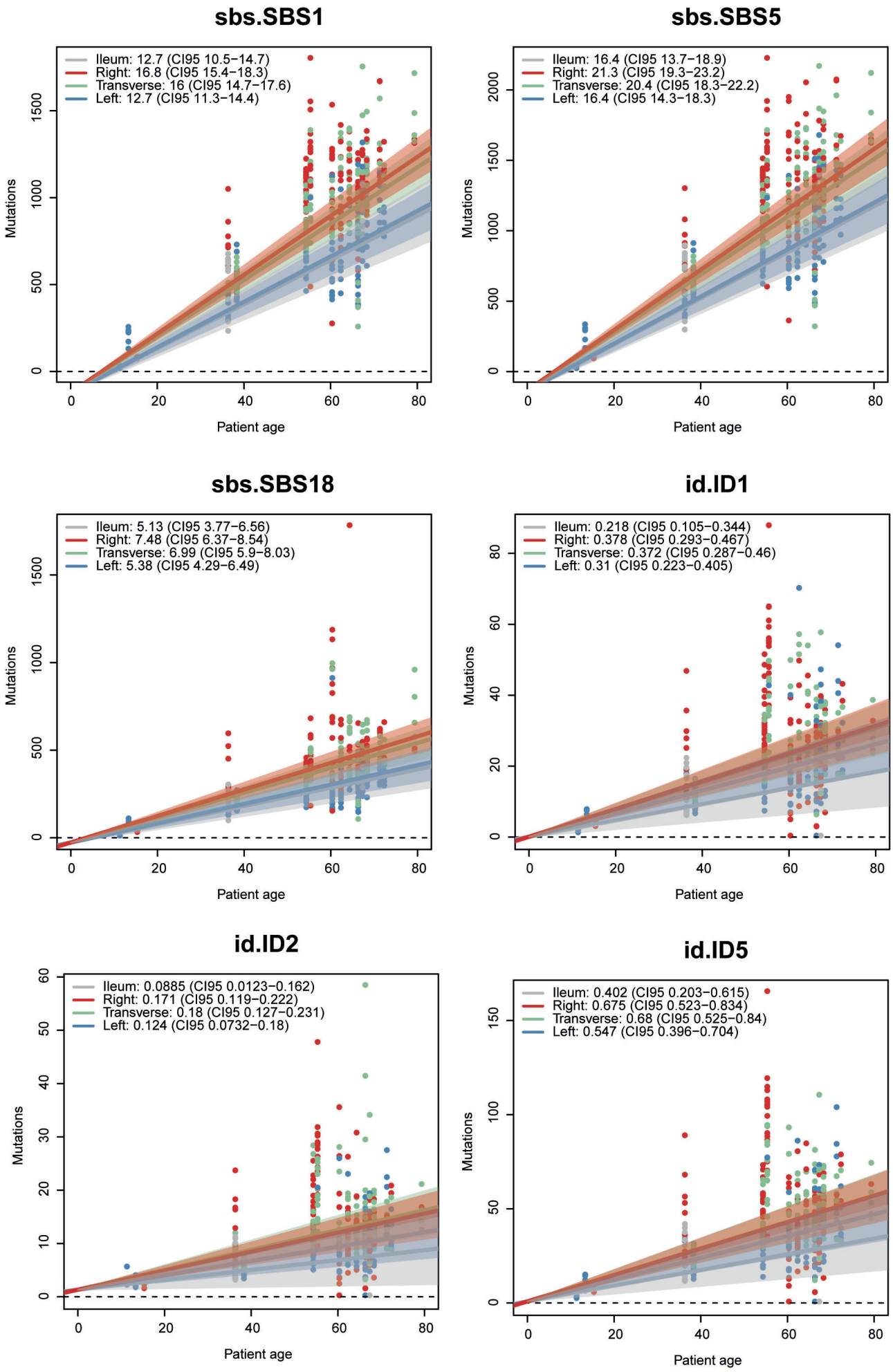
Thus, the signatures that we found to be common have in some shape or form been described before. The size of the cohort and the low complexity of normal genomes relative to cancerous ones allow us to gain novel insights.

R.5.b. Rare mutational signatures

Many mutational processes, both known and novel, only affected subsets of crypts. SBS2 and SBS13 are known signatures predominantly characterised by C>T and C>G mutations respectively at T<u>C</u>A and T<u>C</u>T trinucleotides. They are thought to be due to activity of the APOBEC family of cytidine deaminases and often occur together (Alexandrov et al. 2013, Roberts et al. 2013). The C>T mutations may be a result of cytidine deamination, while the C>G and C>A (and possibly some C>T) mutations may be the errors of translesion polymerases following excision of uracils by repair machinery (Helleday et al. 2014). SBS2 and SBS13 were clearly observed in a colonic crypt from one individual and in an ileal crypt from another, each with >100 mutations of SBS2/13; smaller contributions may be present in other crypts. Therefore, APOBEC DNA-editing of the human genome occurs in gastrointestinal stem cells, to our knowledge the first time that it has been shown in normal cells (beyond the physiological role of AID, an APOBEC family member, in lymphocytes). The wider sequence context of the mutations suggests that APOBEC3A is the major contributing enzyme (Chan et al. 2015).

A novel signature (described as novel here, as it was unknown at the time that we found it, although it has since been observed in oral squamous cell carcinomas (Boot et al. 2018)), SBSN2, characterised predominantly by T>C mutations at A<u>T</u>A, A<u>T</u>T, and T<u>T</u>T, and T>G mutations at T<u>T</u>T was detected in 30% of crypts. In the most affected crypts, it accounted for 3,000 mutations, doubling the mutation burden. SBSN2 exhibits strong transcription strand bias, with 2.5 times as many T>A mutations occurring on the untranscribed as on the transcribed strand. Transcription strand bias is often due to transcription coupled nucleotide excision repair (TC-NER) acting on DNA with covalently bound bulky adducts and distortion of the helical structure. Assuming that this is the case, damage to adenine by a carcinogen may underlie SBSN2. SBSN2 exhibited a highly variable mutational burden between individuals and between crypts from the same individual that was not attributable to age. Examination of the branches of phylogenies in which

Figure 2.11. 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 patient as a random effect. Confidence intervals were determined by bootstrapping. Mutation burden is adjusted for the callable proportion of the genome (Methods section 12).


sbs.SBS1
Ileum: 12.7 (CI95 10.5−14.7)
Right: 16.8 (CI95 15.4−18.3)
Transverse: 16 (CI95 14.7−17.6)
Left: 12.7 (CI95 11.3−14.4)
Mutations
Patient age
sbs.SBS5
Ileum: 16.4 (CI95 13.7−18.9)
Right: 21.3 (CI95 19.3−23.2)
Transverse: 20.4 (CI95 18.3−22.2)
Left: 16.4 (CI95 14.3−18.3)
Mutations
Patient age
sbs.SBS18
Ileum: 5.13 (CI95 3.77−6.56)
Right: 7.48 (CI95 6.37−8.54)
Transverse: 6.99 (CI95 5.9−8.03)
Left: 5.38 (CI95 4.29−6.49)
Mutations
Patient age
id.ID1
Ileum: 0.218 (CI95 0.105−0.344)
Right: 0.378 (CI95 0.293−0.467)
Transverse: 0.372 (CI95 0.287−0.46)
Left: 0.31 (CI95 0.223−0.405)
Mutations
Patient age
id.ID2
Ileum: 0.0885 (CI95 0.0123−0.162)
Right: 0.171 (CI95 0.119−0.222)
Transverse: 0.18 (CI95 0.127−0.231)
Left: 0.124 (CI95 0.0732−0.18)
Mutations
Patient age
id.ID5
Ileum: 0.402 (CI95 0.203−0.615)
Right: 0.675 (CI95 0.523−0.834)
Transverse: 0.68 (CI95 0.525−0.84)
Left: 0.547 (CI95 0.396−0.704)
Mutations
Patient age

SBSN2 occurred showed that whenever SBSN2 could be timed, it always occurred early in life (figure 2.12 f, h, z, aa, am, ao). Assuming a constant mutation rate of SBS1 (as indicated by the age correlation observed), SBSN2 occurs in the first ten years of life. It cannot be the result of one brief mutagenic burst, since it is found in sequential branches (figure 2.12ao). In addition, SBSN2 clustered spatially. In a thirty-six year-old patient, crypts from all around the colon were affected, but not crypts from the ileum. In other patients, crypts from the left colon were much more affected than crypts from other parts of the colon. This was not due simply to the sharing of mutations between crypts. SBSN2 correlates with IDN2, another novel signature characterised by deletion of single T in a run of Ts with a mode length of four (compare ochre bars in ID trees and pink bars in SBS trees in figure 2.12). The initiating events for this relatively common mutational process are unknown, but our data indicate an extrinsic, locally-acting mutagen that affects children. Many causes are possible, including diet, infections, and microbiome composition.

SBSN3 was predominantly characterised by C>T substitutions at A<u>C</u>A, T>A at C<u>T</u>N, and T>G at G<u>T</u>G. It was present in five individuals and in these in a subset of crypts (figure 2.12 e, aa, af, ai, aj). Like SBSN2, SBSN3 was active early in life in the two patients in whom we could time it (figure 2.12 aa, ai), and even when mutations were not shared we could detect spatial clustering. The cause of SBSN3 is unknown, but here again we have evidence of an early, locally-acting process. SBSN3 correlated with DBS8 and IDN3 (compare dark grey in SBS trees, dark grey in DBS trees, and green in ID trees). DBS8 is composed of AC > CA and AC > CT mutations, thus representing some of the same set of base changes as SBSN3. DBS8 has been reported in rare hypermutated cancers with no obvious cause for their hypermutation (Alexandrov et al. 2018). IDN3 is dominated by the deletion of a single T with no other Ts surrounding it.

All crypts from the left, right, and transverse colon of a 66 year-old man were dominated by a mean of 8,567 mutations due to a new signature, SBSN1. This signature is characterised by T>A substitutions with a transcriptional strand bias that is again consistent with damage to adenines. The mutation burden in his colorectal epithelium was three- to five-fold higher than expected for his age, and thus by extrapolation equivalent to that of a 200-300 year-old. This individual had a rich and unusual clinical history: initially diagnosed with a large anaplastic lymphoma in 1994 and treated with six cycles of CHOP (cyclophosphamide, doxorubicin, vincristine, prednisolone), the diagnosis was subsequently revised to Hodgkin's lymphoma, and in 2002 he was treated with three cycles of Chl-VPP / PABlOE (chlorambucil, vinblastine,

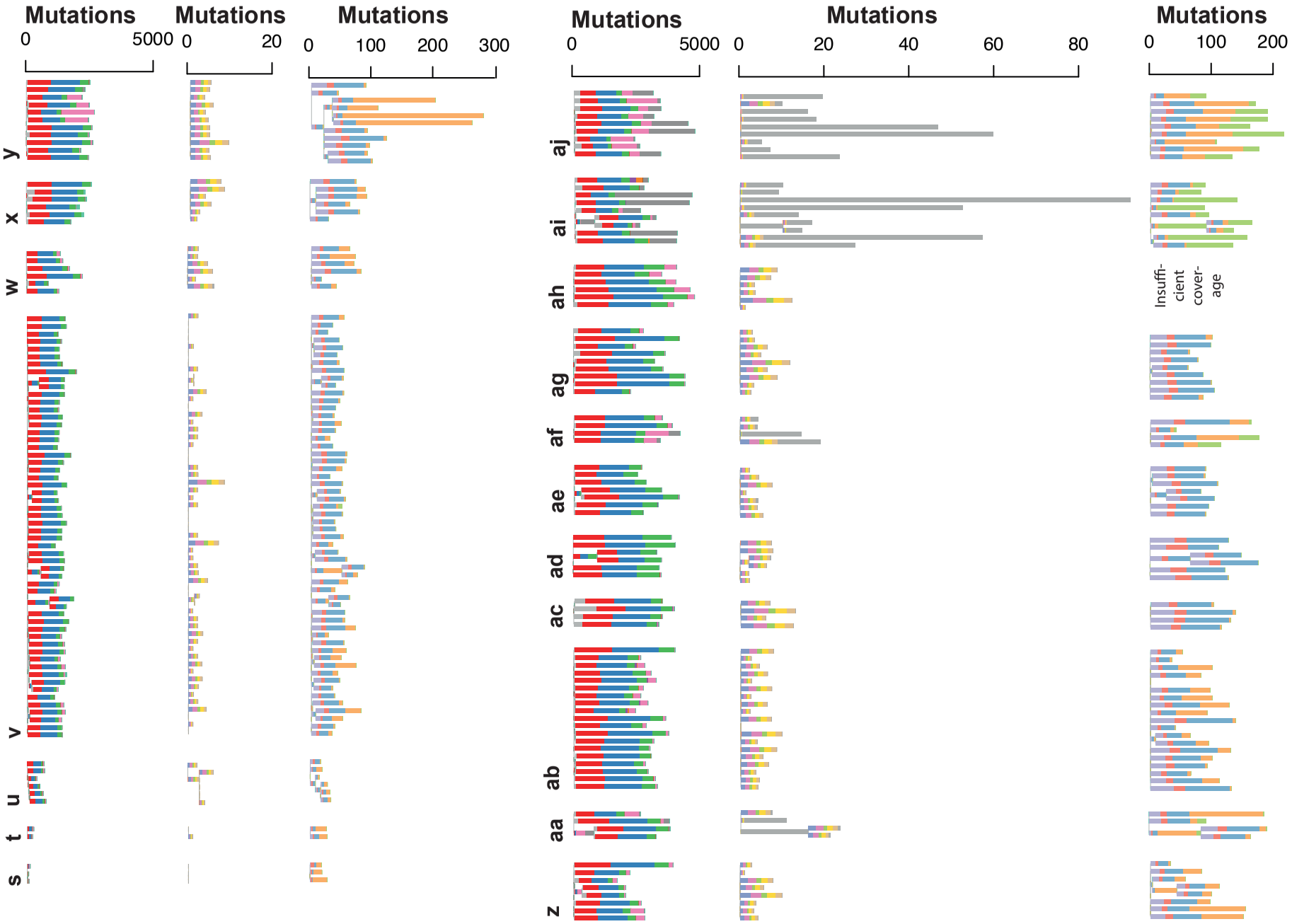
procarbazine and prednisolone, alternating with prednisolone, doxorubicin, bleomycin, vincristine and etoposide). In 2011 he had a positive faecal occult blood test, for which he had a colonoscopy (when the biopsies that we used were taken) and a caecal adenocarcinoma was found. Two years later this gentleman died from a diffuse large B cell lymphoma.

SBSN1 bears a strong resemblance to SBS25 (cosine similarity 0.9), and they share the same transcription strand bias. Signature 25 has previously been detected only in two lymphoma cell lines derived from lymphomas from patients who had received chemotherapy. The case history of one of these patients is available (Wolf et al. 1996): he too had been treated with a cocktail of chemotherapies, with overlap of several drugs (cyclophosphamide, doxorubicin, vincristine/vinblastine, prednisolone, procarbazine, bleomycin). He had also been treated with an experimental ricin-coupled anti-CD25 immunotoxin. Signature 25 was not detected in any of 23,829 cancers in PCAWG and so represents a very rare mutational process. The vast majority of PCAWG cancers were primaries that had not been exposed to chemotherapy, and so it seems likely that chemotherapy is responsible, either through the action of one cytotoxic agent, a combination of them, or their interaction with a germline polymorphism. Cyclophosphamide has been observed to cause T>A mutations in the chicken DT40 cell line (Szilkrist et al. 2016); further work testing different chemotherapies on human cells would be required to identify the causal agent. To our knowledge, this is the first report of the mutagenic consequence of chemotherapy in normal human cells *in vivo*.

To determine whether this process could have played a part in the development of this gentleman's adenocarcinoma, we obtained the biopsy of his tumour taken during the same colonoscopy as the normal samples. The tumour was necrotic and individual crypts could not be distinguished. We therefore bulk-sequenced it. This complicates comparisons of mutation burden between the tumour and normal sample, since the time to the most recent common ancestor of the tumour and the normal samples need not be the same. The tumour was not closely related to any of the individual crypts. As with most tumours, it had an excess of mutational processes not seen in the normal crypts, including single base substitutions, small insertions and deletions, and copy number changes. The T>A mutational process was present in the clonal peak of mutations, but not subclonally, indicating that exposure to chemotherapy predated the last clonal sweep of the tumour; indeed, it is quite plausible that it could have occurred while the tumour was still a normal crypt. After adjusting for copy number changes, a similar number of T>A mutations were present

Figure 2.12. Crypt phylogenies. For every patient, 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. 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. In most cases, crypts from the same individual are distant from one another, and so we would not expect to see late branching events.

a
b
c
d
e
f
g
h
i
j
Mutations
0
5000
Mutations
0
20
Mutations
0
100
200
300
k
l
m
n
o
p
q
r
PIK3CA
E542K
Mutations
0
5000
Mutations
0
20
Mutations
0
100
200


Mutations
0
5000
Mutations
0
20
Mutations
0
100
200
300
Mutations
0
5000
Mutations
0
20
40
60
80
Mutations
0
100
200
s
t
u
v
w
x
y
z
aa
ab
ac
ad
ae
af
ag
ah
ai
aj
Insuffi-
cient
cover-
age

ak
al
am
an
ao
ap
AXIN2
loss
Mutations
0
5000
10000
15000
Mutations
0
20
Mutations
0
100
200
300
400
500
600
Insuffi-
cient
cover-
age
sbs_X0
sbs_SBS1
sbs_SBS5
sbs_SBS18
sbs_SBS2
sbs_SBS13
sbs_N1
sbs_N2
sbs_N3
sbs_N4
dbs_X0
dbs_DBS2
dbs_DBS4
dbs_DBS6
dbs_DBS9
dbs_DBS11
dbs_DBS8
id_X0
id_ID1
id_ID2
id_ID5
id_N2
id_N3

in the trunk of the tumour to the normal crypts. The tumour was driven by a clonal *KRAS* G12D hotspot substitution and two inactivating mutations (one substitution and one indel) in *APC*. These were not due to T>A mutations, and so the mutagenic activity of this chemotherapy cannot be held directly responsible for the development of the tumour. It is possible, however, that the selection pressure of chemotherapy could have provided the conditions for a pre-malignant clone to expand and progress.

R.5.c. Large scale genomic rearrangements in normal colon

Colorectal cancers frequently bear a large number of genomic rearrangements (Li et al. 2017), and rearrangements were found in four out of 15 organoids (Blokzijl et al. 2016). In our larger cohort, we sought evidence of large copy number changes in 449 crypts that had >10X coverage and >0.3 VAF sufficient for us to call copy number changes accurately. In stark contrast to colorectal cancers, 82% of evaluable crypts had no large-scale genomic rearrangements. Remarkably, however, five crypts had seven whole chromosome copy number increases affecting the same three chromosomes. We observed: amplification of both copies of chromosome 3; trisomy 3 and trisomy 9; trisomy 7; amplification of both copies of both chromosome 7 and of the X chromosome; and amplifications of both copies of both chromosome 7 and chromosome 9 (figure 2.13a). We also observed an amplification of the X chromosome. All amplifications increased the copy number of the chromosome by one or two copies. We did not observe any chromosomal deletions. While regions of chromosome 7 are often amplified in colorectal cancer (Cancer Genome Atlas Network et al. 2012), we are not aware of frequent amplifications of chromosomes 3 and 9. In addition, we found large regions of copy neutral loss of heterozygosity in 12 crypts, affecting 1p, 6p, 7p, 8q, 9q, 10q (twice), 17p, 17q, 18q, 21q, 22q, and the X chromosome (e.g. figure 2.3c).

Five of these copy number changes could be timed reliably by using their allele fractions to assign mutations to a copy number state. The count of mutations at each copy state can be used to estimate when the copy number change occurred. This is because mutations that occur before a chromosomal gain will be on two copies, whereas those that occurred after it will be on one copy. Timing these seven changes showed that they all occurred between the ages of 20 and 51 (figure

2.13b). Two gains in the same crypt were timed independently and found to occur at the same time, suggesting that they occurred as one event. This analysis requires the assumption of a constant mutation rate per length of DNA over life, which seems reasonable given that the signatures that are responsible for the majority of mutations in normal colorectal genomes correlate linearly with the age of the patients from whom they are derived. Trinucleotide plots of the mutations assigned to each copy number state were inspected and showed no dramatic change in mutational profile.

Structural variants were detected by abnormally mapping reads. We observed 48 deletions, 18 tandem duplications, four translocations, and two inversions. All were private to a single crypt, except for one which was shared between two adjacent crypts from one patient, which share a very distant common ancestor: this must reflect a deletion that occurred in the embryo or in early childhood.

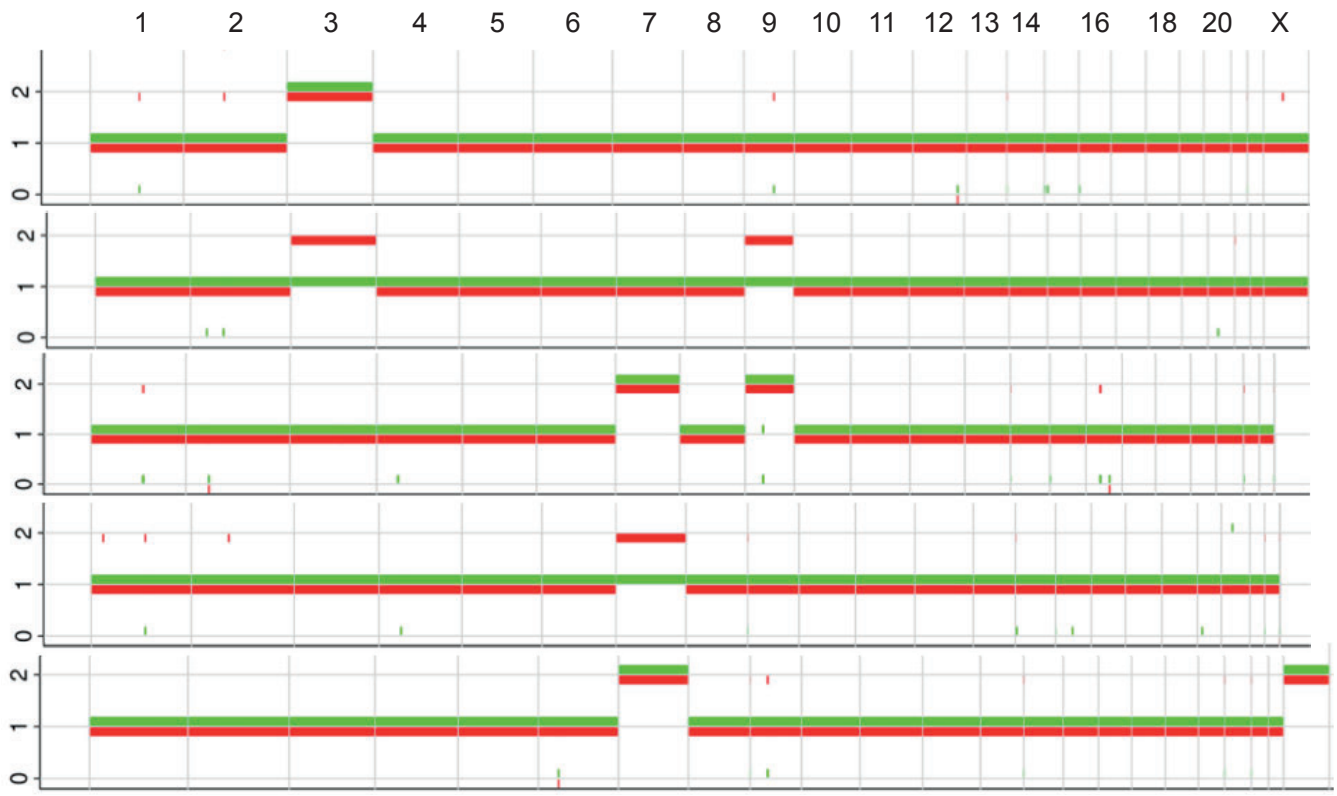
R.6. Comparison of mutational signatures in our cohort and PCAWG colorectal cancers

The total mutation burden in most crypts that we sequenced is of the order of 3,000 mutations per genome, substantially less than the mutation burden of 10,000-20,000 mutations per genome in even non-hypermutated colorectal adenocarcinomas. We sought to explain the source of the mutational excess in cancers by comparing the number of mutations due to each mutational signature in PCAWG colorectal adenocarcinomas and our normal tissues (figure 2.14). A number of complicating factors should be borne in mind: tumours and normal crypts have a different time to their MRCA; patients were of different ages, although in both cohorts most samples come from patients in their 50s and 60s; they were sequenced on different platforms and mutations called with different filters; and signatures were extracted separately (although the same set of reference signatures was used).

As anticipated, the total mutation burden in colorectal cancers is always higher than in normal crypts, with the exception of those from the patient who had been exposed to chemotherapy. The mutation burden of the near-ubiquitous signatures SBS1, SBS5, SBS18, DBS2, DBS4, DBS6, DBS9, ID1, and ID2 was higher in tumour samples than in normal samples, suggesting an acceleration of normal mutational processes. Some rare mutational processes were

Figure 2.13. Copy number changes in normal colon. 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.

a


1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 X
0 1 2

b

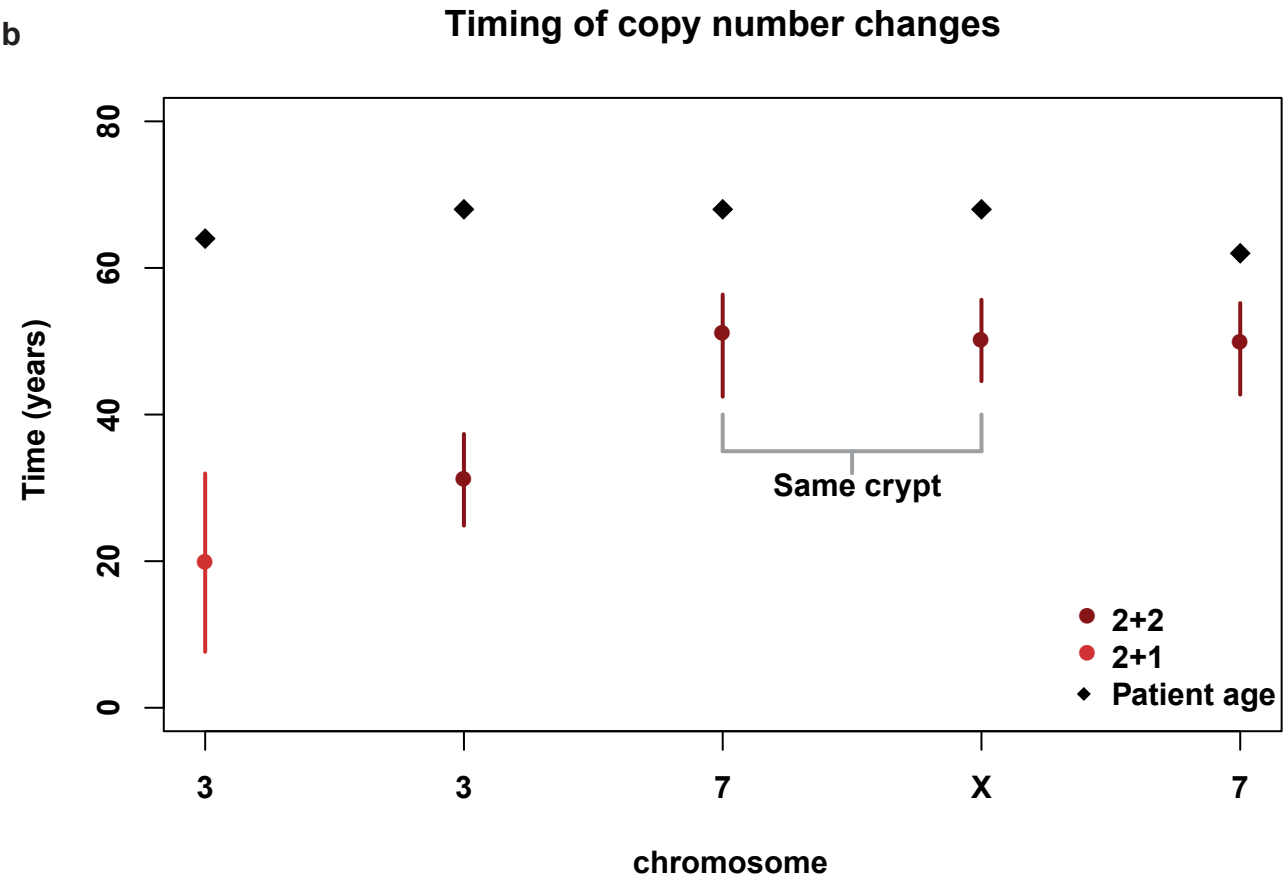
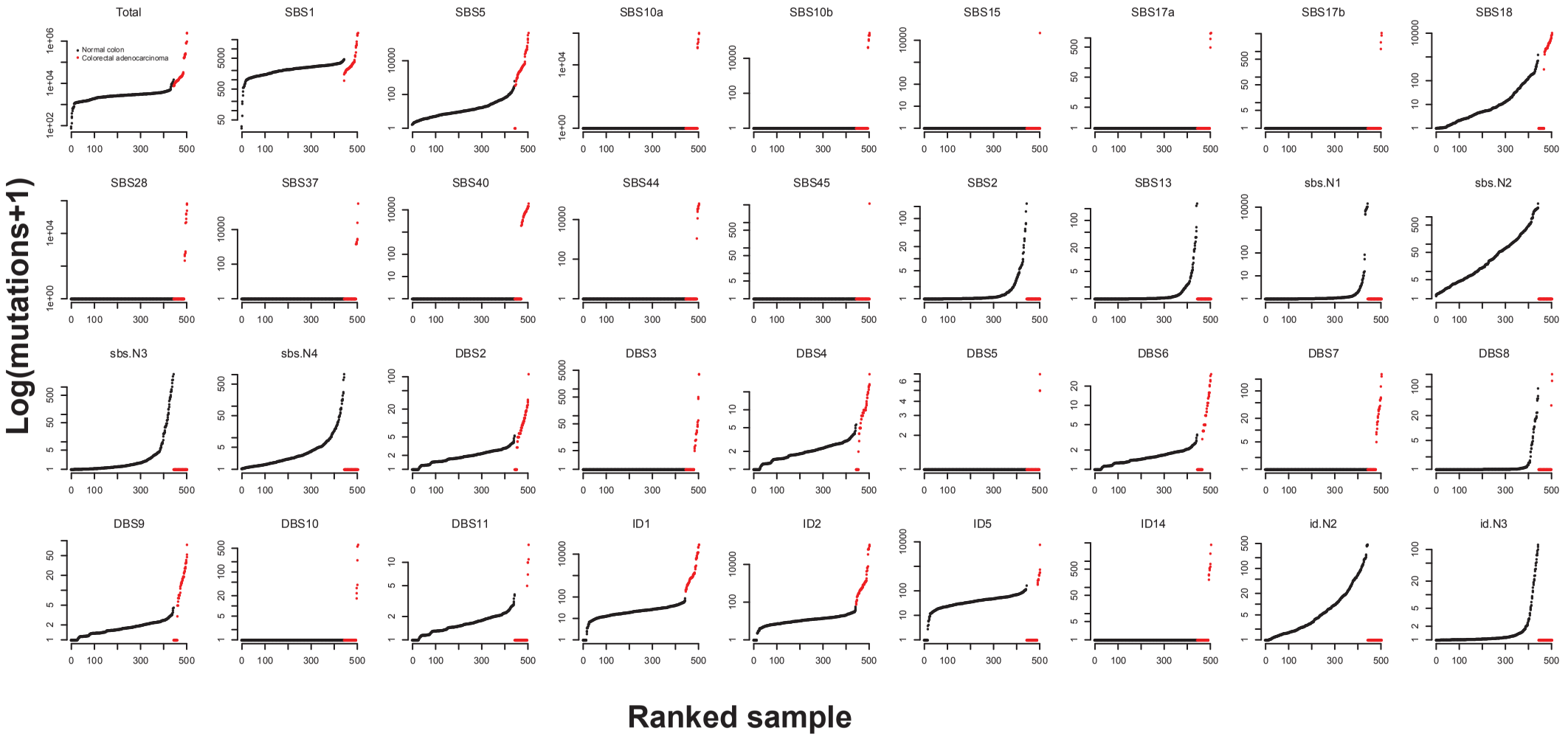

Timing of copy number changes
Time (years)
0 20 40 60 80
Same crypt
2+2
2+1
Patient age
3 3 7 X 7
chromosome

Figure 2.14. Mutation burden of each signature in normal crypts versus cancer. For every 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. (2018).


Total
SBS1
SBS5
SBS10a
SBS10b
SBS15
SBS17a
SBS17b
SBS18
SBS28
SBS37
SBS40
SBS44
SBS45
SBS2
SBS13
sbs.N1
sbs.N2
sbs.N3
sbs.N4
DBS2
DBS3
DBS4
DBS5
DBS6
DBS7
DBS8
DBS9
DBS10
DBS11
ID1
ID2
ID5
ID14
id.N2
id.N3
Normal colon
Colorectal adenocarcinoma
Log(mutations+1)
Ranked sample

only found in tumours, such as SBS10a or SBS10b and SBS17. Perhaps if a larger number of normal crypts had been sequenced these mutational processes would eventually have been found in normal tissues too; nonetheless, sporadic mutational processes do seem to be enriched in cancers relative to normal tissue. Conversely, all the novel signatures were (by definition) found only in normal tissues. No mutational processes were found in all tumours and only in tumours. Thus, there is no process that is specific to and intrinsically linked with malignancy.

R.7. Mutational processes and rates in the progression from normal to cancer

The additional mutational processes detected in cancers could have occurred during the process of transformation or have preceded it: it could be that rare colonic cells with a naturally higher mutation burden are more likely to become cancerous. Furthermore, as explained above, the comparison of bulk tumour samples with normal colonic crypts is complicated by the difference in the time to the most recent common ancestor of each sample. We reasoned that the former difficulty could be resolved by reconstructing the mutational life history of the tumour through multi-region sequencing, and the latter by sequencing single cells derived from tumour and normal epithelium. We set out to achieve this by analysing organoids derived from three patients undergoing a resection for colorectal cancer. For each patient, organoids were derived from single cells from a tumour and from individual normal crypts five centimetres away from the tumour. From the first patient we whole genome sequenced four normal and seven tumour organoids, from the second five normal and 11 tumour organoids, and from the third four normal and eight tumour organoids. This study was set up by Sam Behjati and Sophie Roerink and organoids were generated by the Clevers group at the Hubrecht Institute.

R.7.a. Comparison of mutational processes in single cells from cancer and normal colon

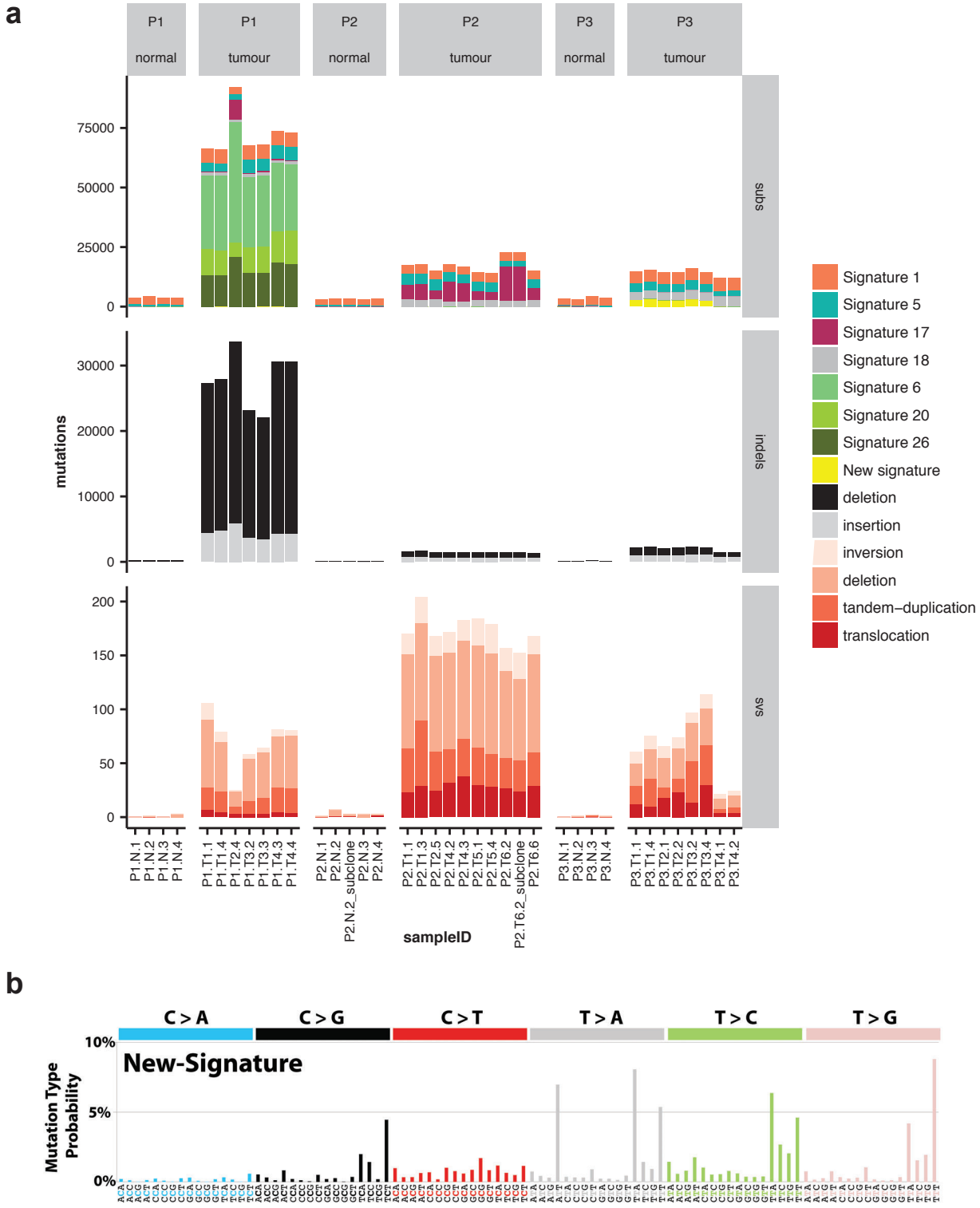
For all three patients, all tumour organoids had many more mutations than any of the normal organoids. Subcloning of organoids showed that the mutation burden acquired *in vitro* was

minimal, as had been shown by others (Blokzijl et al. 2016). Patient 1 had a mismatch repair deficient tumour. Methylation profiling (analysed by Matthew Young) showed that this was due to hypermethylation of the *MLH1* promoter. This patient had a mean of 3,792 substitutions in normal clones but 72,398 substitutions in tumour clones. Patients 2 and 3, who had mismatch repair proficient tumours, had means of 3,172 and 3,621 substitutions in their normal organoids, and of 22,291 and 14,209 substitutions, respectively, in their tumour-derived organoids (figure 2.15a). Similarly, the mean burden of indels in normal organoids for patients 1-3 was 227, 130, and 167, while the mean indel burden in their tumour organoids was 27,893, 1,485, and 2,021. There was a mean of one structural variant in normal organoids, compared to a mean of 71, 176, and 67 from the tumours of patients 1-3. Thus, as one would expect from the comparison of mutation burden between normal colonic crypts and cancer genomes, the mutation burden is increased in cancer relative to normal tissue even within the same patient. Tumour organoids were derived from single cells, while normal organoids were derived from single crypts. More than one stem cell from a crypt might have contributed to the organoid, and so the most recent common ancestor of normal organoids may predate the resection, while that of tumour organoids will not. Nonetheless, the difference is likely to be less than a decade and cannot explain the discrepancy in mutation burden between tumour and normal organoids.

To investigate the origin of the increased burden of mutations in cancer relative to normal, signatures of mutational processes were extracted for substitutions by Ludmil Alexandrov using NNMF. Signatures are referred to here by their numbers in COSMIC (Forbes et al. 2017).[4] All the mutations in normal clones were attributed to signatures 1 and 5. Additional mutational signatures were found in the tumours (figure 2.15). These included: signatures associated with mismatch repair deficiency in patient 1 (signatures 6, 20, and 26); signature 17 in patients 1 and 2, which is of unknown cause and is found in a minority of colorectal cancers; signature 18 in patients 2 and 3, which is thought to be associated with reactive oxygen species; and a novel signature (figure 2.15b), found in only six out of eight cancer-derived organoids in patient 3. 100 mutations that were likely to be due to this novel signature were validated by capillary sequencing (data not shown).

[4] Please note that signature profiles here reflect the TCGA and cosmic versions, which means the number of mutations attributed to each signature cannot be compared directly to analyses that used the PCAWG catalogue.

Figure 2.15. Mutation burden of colonic organoids. a, barplots of the mutation burden in every organoid. For substitutions, mutations are broken down by signature, according to the COSMIC catalogue (NB not the PCAWG catalogue), whereas indels are classified as insertion or deletion, and structural variants as inversion, deletion, tandem-duplication, or translocation. For each patient the burden in normal and tumour organoids are shown. **b**, the trinucleotide spectrum of the novel signature (yellow in **a**).


a
P1 normal
P1 tumour
P2 normal
P2 tumour
P3 normal
P3 tumour
subs
indels
svs
mutations
75000
50000
25000
0
30000
20000
10000
200
150
100
50
P1.N.1
P1.N.2
P1.N.3
P1.N.4
P1.T1.1
P1.T1.4
P1.T2.4
P1.T3.2
P1.T3.3
P1.T4.3
P1.T4.4
P2.N.1
P2.N.2
P2.N.2_subclone
P2.N.3
P2.N.4
P2.T1.1
P2.T1.3
P2.T2.5
P2.T4.2
P2.T4.3
P2.T5.1
P2.T5.4
P2.T6.2
P2.T6.2_subclone
P2.T6.6
P3.N.1
P3.N.2
P3.N.3
P3.N.4
P3.T1.1
P3.T1.4
P3.T2.1
P3.T2.2
P3.T3.2
P3.T3.4
P3.T4.1
P3.T4.2
sampleID
Signature 1
Signature 5
Signature 17
Signature 18
Signature 6
Signature 20
Signature 26
New signature
deletion
insertion
inversion
deletion
tandem−duplication
translocation
b
C > A
C > G
C > T
T > A
T > C
T > G
10%
5%
0%
Mutation Type Probability
New-Signature

While the majority of the discrepancy between tumour and normal could be attributed to additional mutational signatures, tumour cells also had an increased burden of signatures 1 and 5 relative to the normal cells. Given signature 1's proposed correlation with mitoses, this suggests that cancer cells have undergone more cell divisions. Indeed, increased rates of Ki67 staining, a marker of proliferation, are observed in the resected cancer specimens (staining performed by collaborators; data not shown). However, the increase in signature 1 is not sufficient to explain the discrepancy in the number of mutations between tumour and normal. The fact that cancer cells have additional mutational signatures that are not seen in normal cells shows that additional mutational processes are operative, which must mean an increase in mutation rate per cell division. Thus, cancer cells, even in MMR proficient patients, show an increase in mutation rate per cell division. The cause of this increased mutation rate per cell division warrants further investigation. Some cancer-specific signatures, as in mismatch-repair deficiency, may be a result from the loss of a repair process that occurs in normal cells. Others may be the result of new mutagenic exposures. For example, as signature 18 has been associated with oxidative damage (Viel et al. 2017), its acceleration in all three tumours relative to normal tempts one to speculate that it might be a result of the change in metabolism associated with tumour growth (Warburg et al. 1927).

R.7.b. Reconstructing the mutational histories of tumours with phylogenies

Phylogenies were constructed for the samples derived from each patient (Methods), and mutations were assigned to each branch. Signatures were extracted treating every branch as an independent sample (Methods), which allowed us to see the mutational processes operative at different stages of tumour evolution (figure 2.16). For patient 2, in the trunk of whose tumour a whole genome duplication was present, mutations in the trunk were further separated according to whether they occurred before or after the whole genome duplication (Methods).

In all patients, the trunk of the tumour contained a greater proportion of signatures 1 and 5 than later branches, suggesting that the additional mutational signatures in cancer cells were not present through their whole life history. This was clearest in patient 2. While signature 1 accounted for approximately 60% of the mutations before the whole genome duplication, afterwards it only accounted for 30%, with signature 17 becoming much more prominent. In patient 3, a subclonal

signature was observed (yellow in figure 2.16c), indicating ongoing diversification of mutational processes through tumour evolution. This could be a result of a heritable change (such as loss of another repair protein) or may reflect the particular microenvironment of different tumour regions, since the phylogeny covaries with the spatial structure of the tumour. Likewise, rearrangements and small indels increase markedly in rate beyond the trunk of the tumour (figure 2.17 and 2.18). In patient 3, the rate of rearrangement acquisition can be seen to vary subclonally too, as the clade of the tree that had acquired *TP53* mutations had far more rearrangements than the clade that had not.

The fact that the genomic landscape of the trunk of the tumour phylogenies resembles that of normal colorectal stem cells indicates that there was nothing particularly special about the mutational processes of the cell-of-origin of cancers. It suggests that cancers derive from cells that were exposed to normal mutational processes for much of life until an event, such as one or more driver mutations, caused a change in the active mutational processes. In patient 1, the trunk of the tumour contains almost no signatures of mismatch repair deficiency-associated processes (although there was an increase in indels), despite the fact that all organoids have hypermethylated the *MLH1* promoter, and so this event presumably did occur in the trunk of the tumour. This suggests that the acquisition of aberrant mutational processes was rapidly followed by intratumoral growth and diversification. It could be that both are a result of the *BRAF* V600E mutation, given that Ras pathway mutations are associated with growth (see introduction to this chapter) and that this particular mutation frequently coincides with mismatch repair deficiency (Fearon 2011).

To understand the transition from normal to cancer, we attempted to time the onset of the cancer-specific mutational signatures (Methods). This analysis requires several assumptions, as we can only time mutations relative to signature 1 and cannot relate them directly to real time. As there is more signature 1 in tumour samples than in normal samples, it cannot be used as a real-time clock in the tumour, although we use it in this way in the normal organoids. To estimate the onset of cancer-specific mutational signatures, we assumed that the rate of signature 1 started increasing at approximately the same time as the onset of new mutational signatures. This may be valid if both were associated with the acquisition of driver mutations. In all three patients, signature 18 seemed to be the first cancer-specific mutational signature to become operative. Using the ratio of signature 1 to signature 18 in the next branch after our first timepoint (branches after the most recent common ancestor (MRCA) in patients 1 and 3, and the branch between the WGD and the

Figure 2.16. Phylogenies of colonic organoids. For each patient, the anatomical location in the colon of the tumour is shown, with the phylogeny underneath. White circles represent organoids derived from normal tissue, while filled coloured circles represent organoids derived from tumour, with the colour matching the colour of the biopsy site in the schematic above the phylogeny. Phylogenies are depicted as in figure 2.12 with branch lengths proportional to numbers of mutations, and signature contributions overlaid as stacked barplots.

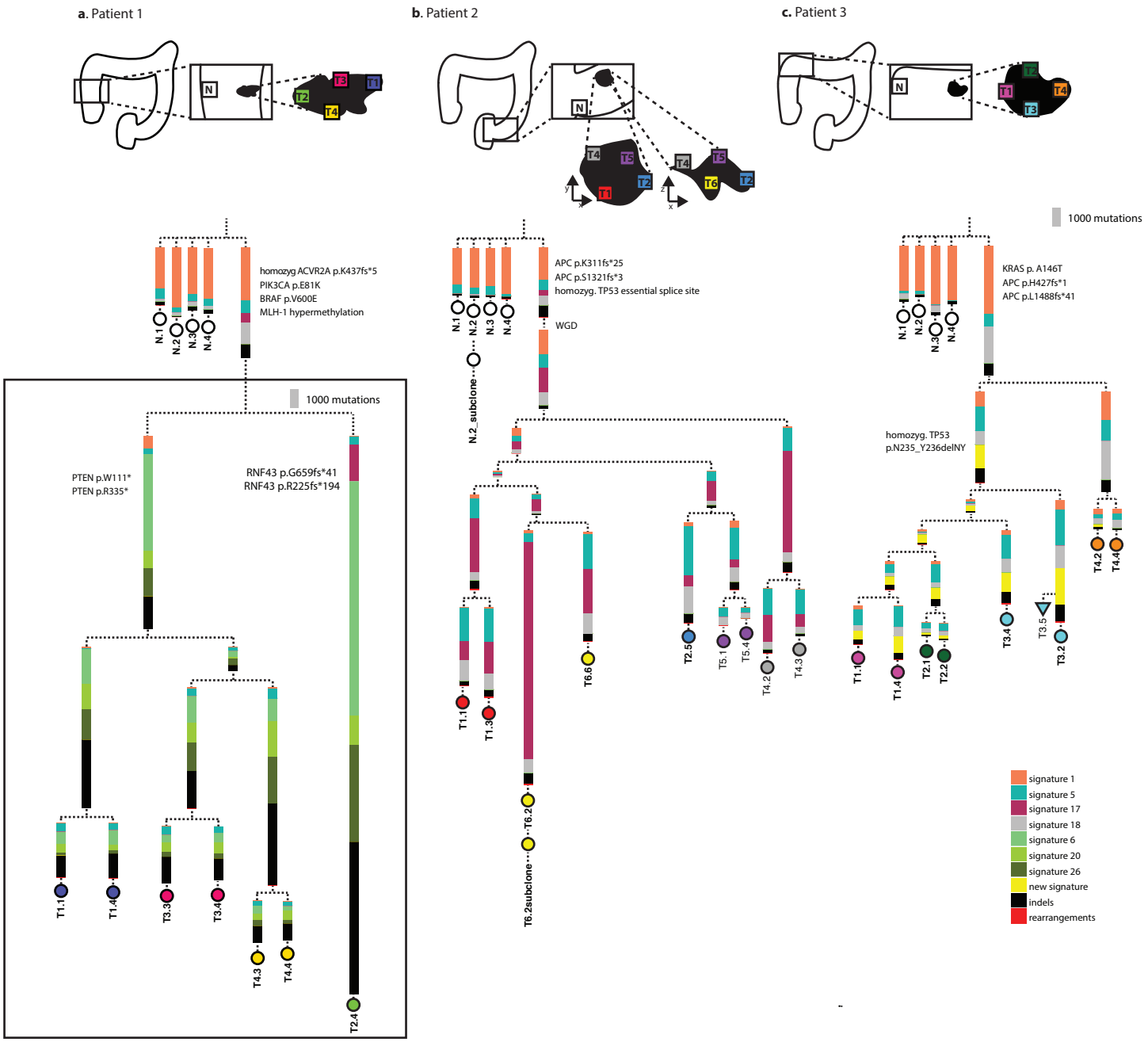

a. Patient 1
b. Patient 2
c. Patient 3
N
T1
T2
T3
T4
T5
T6
1000 mutations
homozyg ACVR2A p.K437fs*5
PIK3CA p.E81K
BRAF p.V600E
MLH-1 hypermethylation
PTEN p.W111*
PTEN p.R335*
RNF43 p.G659fs*41
RNF43 p.R225fs*194
N.1
N.2
N.3
N.4
T1.1
T1.4
T3.3
T3.4
T4.3
T4.4
T2.4
APC p.K311fs*25
APC p.S1321fs*3
homozyg. TP53 essential splice site
WGD
N.2_subclone
T6.2subclone
T1.3
T6.6
T2.5
T5.1
T5.4
T4.2
T4.3
T6.2
KRAS p. A146T
APC p.H427fs*1
APC p.L1488fs*41
homozyg. TP53
p.N235_Y236delNY
T2.1
T2.2
T3.2
T3.5
signature 1
signature 5
signature 17
signature 18
signature 6
signature 20
signature 26
new signature
indels
rearrangements

Figure 2.17. Phylogenies of colonic organoids showing indel burden. Phylogenies are shown as in figure 2.16, but showing indels only as figure 2.16 was dominated by substitutions.

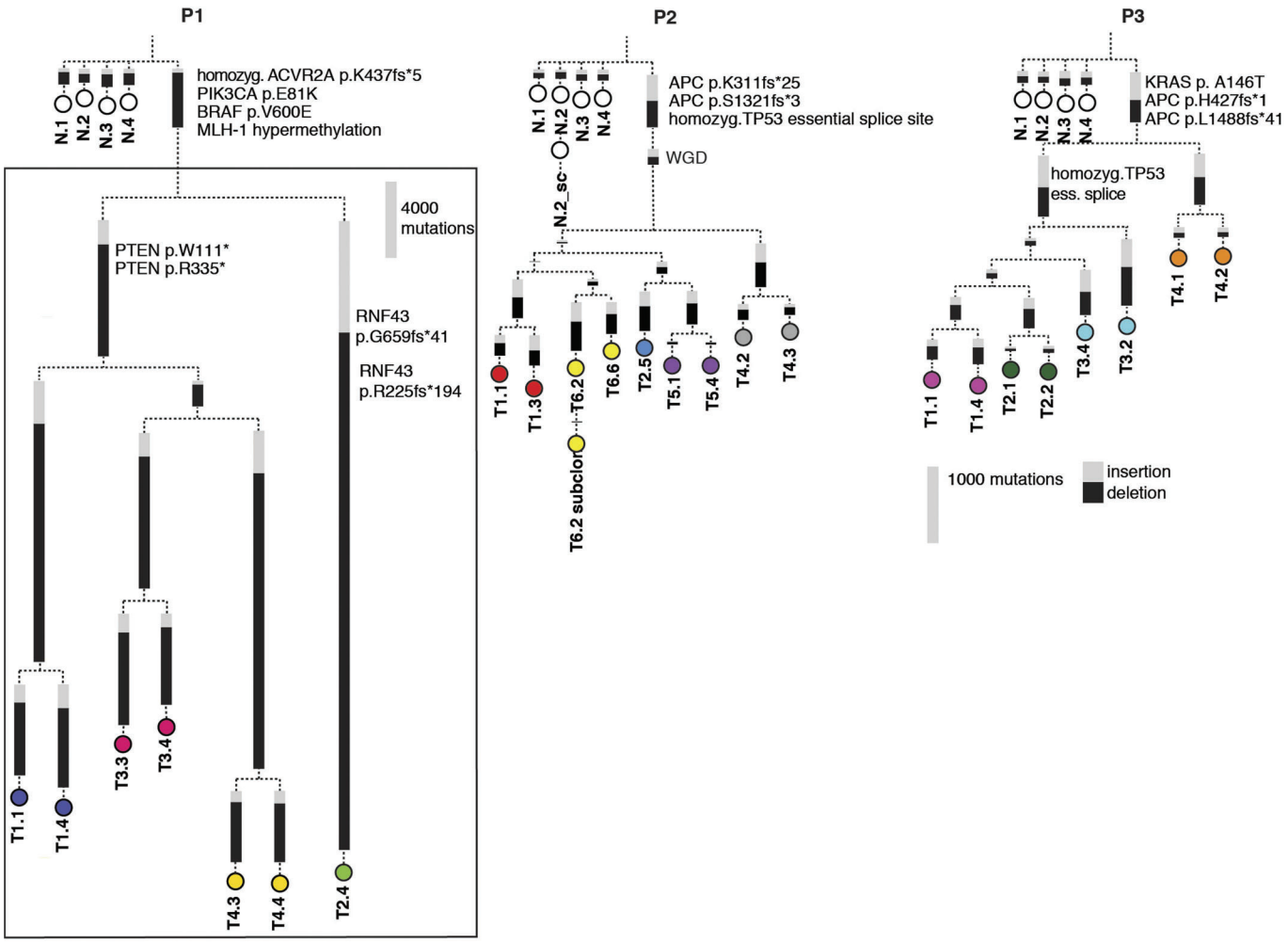
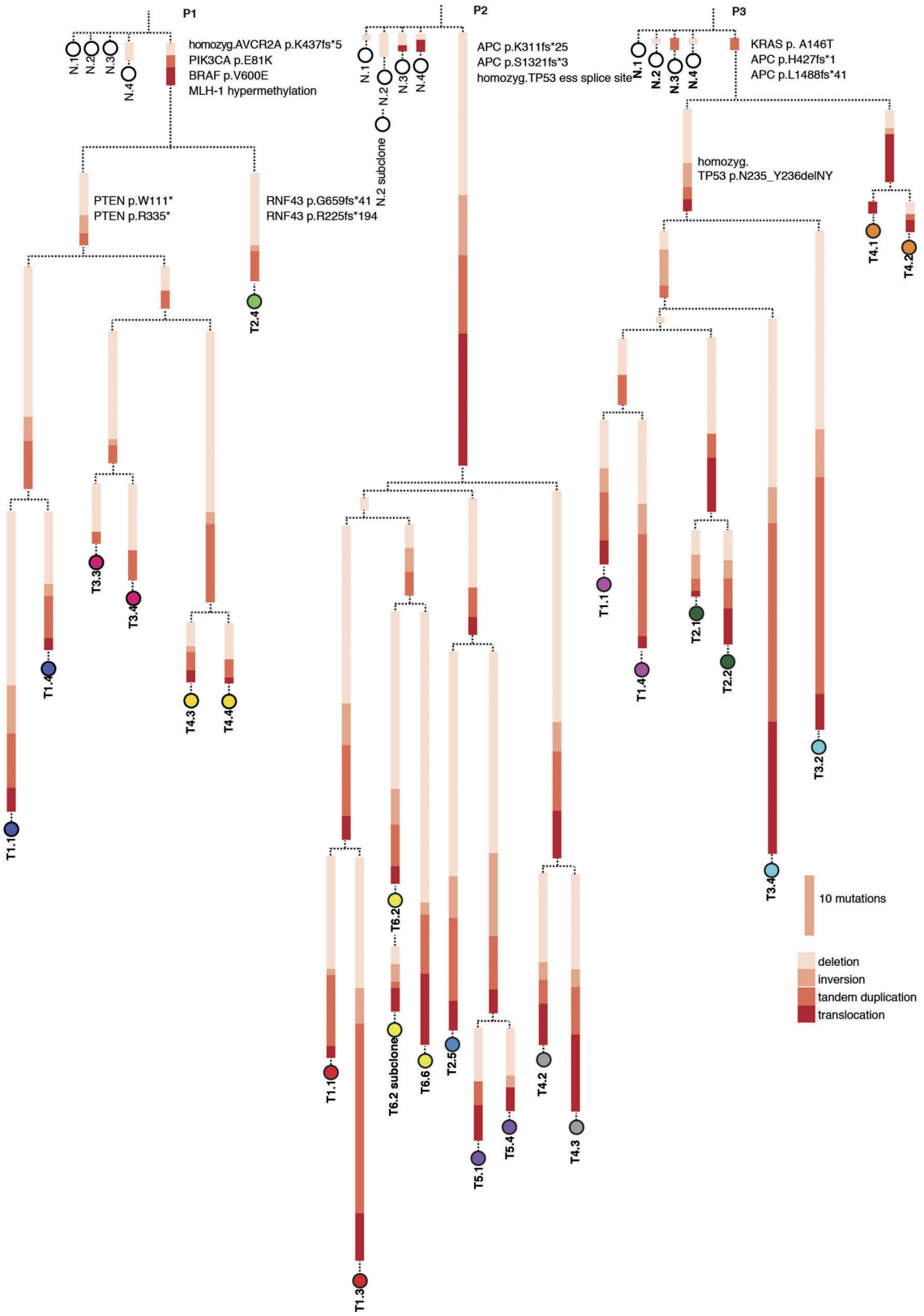

P1
N.1
N.2
N.3
N.4
homozyg. ACVR2A p.K437fs*5
PIK3CA p.E81K
BRAF p.V600E
MLH-1 hypermethylation
4000
mutations
PTEN p.W111*
PTEN p.R335*
RNF43
p.G659fs*41
RNF43
p.R225fs*194
T1.1
T1.4
T3.3
T3.4
T4.3
T4.4
T2.4
P2
N.1
N.2
N.3
N.4
N.2_sc
APC p.K311fs*25
APC p.S1321fs*3
homozyg.TP53 essential splice site
WGD
T1.1
T1.3
T6.2
T6.2 subclone
T6.6
T2.5
T5.1
T5.4
T4.2
T4.3
P3
N.1
N.2
N.3
N.4
KRAS p. A146T
APC p.H427fs*1
APC p.L1488fs*41
homozyg.TP53
ess. splice
T4.1
T4.2
T1.1
T1.4
T2.1
T2.2
T3.4
T3.2
1000 mutations
insertion
deletion

Figure 2.18. Phylogenies of colonic organoids showing rearrangement burden. Phylogenies are shown as in figure 2.16, but showing rearrangements only as figure 2.16 was dominated by substitutions.


P1
N.1
N.2
N.3
N.4
homozyg.AVCR2A p.K437fs*5
PIK3CA p.E81K
BRAF p.V600E
MLH-1 hypermethylation
PTEN p.W111*
PTEN p.R335*
RNF43 p.G659fs*41
RNF43 p.R225fs*194
T1.1
T1.4
T3.3
T3.4
T4.3
T4.4
T2.4
P2
N.1
N.2
N.3
N.4
N.2 subclone
APC p.K311fs*25
APC p.S1321fs*3
homozyg.TP53 ess splice site
T1.1
T1.3
T6.2
T6.2 subclone
T6.6
T2.5
T5.1
T5.4
T4.2
T4.3
P3
N.1
N.2
N.3
N.4
KRAS p. A146T
APC p.H427fs*1
APC p.L1488fs*41
homozyg.
TP53 p.N235_Y236delNY
T1.1
T1.4
T2.1
T2.2
T3.4
T3.2
T4.1
T4.2
10 mutations
deletion
inversion
tandem duplication
translocation

MRCA in patient 2) we calculated the number of signature 1 mutations present in the cell when signature 18 started. Assuming a linear acquisition of signature 1 in the trunk of the tumour at the same rate as in the normal samples, we could translate this signature 1 burden into the age at which signature 18 began. As the ratio of other signatures to signature 1 increases down the tree it is likely that that the rate of acquisition of signature 18 relative to signature 1 increases over time. If this is the case, the estimates that we obtain can be seen as a lower bound of age. Following this line of reasoning, in patient 1 (aged 67) signature 18 began in the last 24 years, in patient 2 (aged 68) in the last 20 years, and in patient 3 (aged 56) in the last 22 years.

Summary of results in this chapter

Our understanding of the process of somatic mutation, and its consequences, in normal cells remains rudimentary. This has largely been because of the difficulties of detecting somatic mutations in normal cells. A protocol to sequence individual colonic crypts isolated by laser capture microdissection was derived and applied to sequence hundreds of crypts from 42 different individuals. We characterised the genomic landscape of these normal cells, quantifying their mutation burden, the signatures of the mutational processes that have affected them, and the spectrum of driver mutations found in the colons of healthy people.

Putative driver mutations were found to occur in approximately 1% of crypts, and so number in the hundreds of thousands in the colons of middle-aged healthy individuals. The spectrum of driver mutations was different to that observed in cancers, with some of the most common colorectal cancer driver mutations absent from our data. Caution should be exercised with these estimates, however, as we are limited by a low frequency of driver mutations and variable coverage within and between genes; further studies (guided by this initial estimate) will be required to establish the frequency of driver mutations more accurately.

Striking variability in somatic mutation rates in normal cells was found between different people and between different cells within the same person. This was due both to variation in the number of mutations due to ubiquitous mutational processes and to the presence of sporadic mutational processes that only affected certain individuals or certain crypts within one individual. One novel signature, which quintupled the normal mutation burden, could be linked to a previous

exposure to chemotherapy. Others were of unknown aetiology, but accounted for thousands of mutations in focal patches of the colons of children. APOBEC mutations, which have never previously been reported in normal cells, were found in a small subset of crypts.

The genomes of colorectal cancers show mutational signatures that were not found even in a large cohort of normal cells. Those mutational processes that were found in all normal cells are universally accelerated in cancer. Phylogenetic analysis of the genomes of cancer cell-derived organoids reveals that the cell ancestral to the cancer was for most of its life subject to the same mutational processes as its neighbours that did not transform. In the process of transformation to cancer, this clone both accelerated normal mutational processes and acquired novel ones.

These findings will be discussed in the Discussion chapter.