ARTICLE

https://doi.org/10.1038/s41586-018-0024-3

Intra-tumour diversification in colorectal cancer at the single-cell level

Sophie F. Roerink[1,13], Nobuo Sasaki[2,11,13], Henry Lee-Six[1,13], Matthew D. Young[1], Ludmil B. Alexandrov[3,4,5], Sam Behjati[1,6], Thomas J. Mitchell[1,7], Sebastian Grossmann[1], Howard Lightfoot[1], David A. Egan[8,12], Apollo Pronk[9], Niels Smakman[9], Joost van Gorp[10], Elizabeth Anderson[1], Stephen J. Gamble[1], Chris Alder[1], Marc van de Wetering[2], Peter J. Campbell[1], Michael R. Stratton[1]* & Hans Clevers[2]*

Every cancer originates from a single cell. During expansion of the neoplastic cell population, individual cells acquire genetic and phenotypic differences from each other. Here, to investigate the nature and extent of intra-tumour diversification, we characterized organoids derived from multiple single cells from three colorectal cancers as well as from adjacent normal intestinal crypts. Colorectal cancer cells showed extensive mutational diversification and carried several times more somatic mutations than normal colorectal cells. Most mutations were acquired during the final dominant clonal expansion of the cancer and resulted from mutational processes that are absent from normal colorectal cells. Intra-tumour diversification of DNA methylation and transcriptome states also occurred; these alterations were cell-autonomous, stable, and followed the phylogenetic tree of each cancer. There were marked differences in responses to anticancer drugs between even closely related cells of the same tumour. The results indicate that colorectal cancer cells experience substantial increases in somatic mutation rate compared to normal colorectal cells, and that genetic diversification of each cancer is accompanied by pervasive, stable and inherited differences in the biological states of individual cancer cells.

Recent studies have explored genetic diversification within cancer cell populations by identifying mutations shared by subpopulations of cells within the cancer clone[1–6]. In principle, however, the extent of intra-tumour genetic diversity is most comprehensively revealed by single-cancer-cell DNA sequencing, which can potentially identify all mutations, including those that arose in the distant past, those that occurred very recently, and those not present in any other cell[7,8]. Despite recent advances, however, single-cell genome sequencing remains dependent on prior whole-genome amplification, which is associated with incomplete genome coverage and artefactual mutations[9]. Diversification of epigenomic, transcriptomic, proteomic and metabolic states, and of functional states such as resistance to anticancer therapy, may also occur during expansion of the neoplastic cell population[10–13]. Methylation, gene expression and drug response data have previously been obtained from multiple cells from individual tumours, but the collection of all these features together with accurate genome information from the same single cells has not been reported, to our knowledge[14–19]. The origins of epigenomic and transcriptomic diversification are unclear, and there is little insight into whether these are transient or stable.

One experimental approach that enables comprehensive, systematic and integrated exploration of intra-tumour diversification is to derive immortal cell lines from multiple single cells from the same cancer[20]. These serve as proxies for the single cells from which they originate and can be subjected to extensive and multimodal characterization, thereby revealing all aspects of intra-tumour diversification retained during in vitro growth. We have recently developed protocols for derivation of clonal organoids from normal and neoplastic colorectal stem cells, and we use this strategy here to compare single cells from normal and neoplastic colorectal epithelium[21–24].

Comparison of the number of mutations in single cancer cell genomes with that in individual normal cells from the same tissue may also reveal whether alterations in somatic mutation rate and mutational process have been experienced by neoplastic cells. Substantial increases in mutation rate are known to occur during the development of cancers with DNA mismatch repair deficiency or mutations in genes encoding DNA polymerases[25]. Whether mutation rate increases are common in cancers without these specific abnormalities is, however, currently unknown and a matter of controversy[26–30].

Clonal organoid derivation

Colorectal cancers from three previously untreated individuals (P1, P2 and P3) were each dissected into 4–6 pieces (Extended Data Fig. 1). Organoid cultures were derived from cell suspensions made separately from each piece and were maintained for up to one week without passage. Subsequently, individual organoids were disaggregated and flow-sorted to obtain single cells from which clonal cancer organoids were established. For each individual, organoids were also derived from single crypts in normal colorectal epithelium from the same resection specimens. A crypt derives from a single stem cell that has been estimated to exist several months before crypt isolation[24,31].

The coding regions of 360 known cancer genes were sequenced in all normal and cancer-derived clonal organoids for likely driver mutations and a subset were whole-genome sequenced. Somatic mutations

[1]Cancer Ageing and Somatic Mutations Programme, Wellcome Trust Sanger Insitute, Hinxton, UK. [2]Hubrecht Institute, University Medical Center Utrecht and Princess Maxima Center, Utrecht, The Netherlands. [3]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA. [4]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [5]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. [6]Department of Paediatrics, University of Cambridge, Cambridge, UK. [7]Academic Urology Group, Department of Surgery, Addenbrooke's Hospitals NHS Foundation Trust, University of Cambridge, Cambridge, UK. [8]Cell Screening Core, Department of Cell Biology, Center for Molecular Medicine, University Medical Centre Utrecht, Utrecht, The Netherlands. [9]Department of Surgery, Diakonessenhuis, Utrecht, The Netherlands. [10]Department of Pathology, Diakonessenhuis, Utrecht, The Netherlands. [11]Present address: Department of Gastroenterology, Keio University School of Medicine, Tokyo, Japan. [12]Present address: Core Life Analytics, Utrecht, The Netherlands. [13]These authors contributed equally: Sophie F. Roerink, Nobuo Sasaki, Henry Lee-Six. *e-mail: mrs@sanger.ac.uk; h.clevers@hubrecht.eu

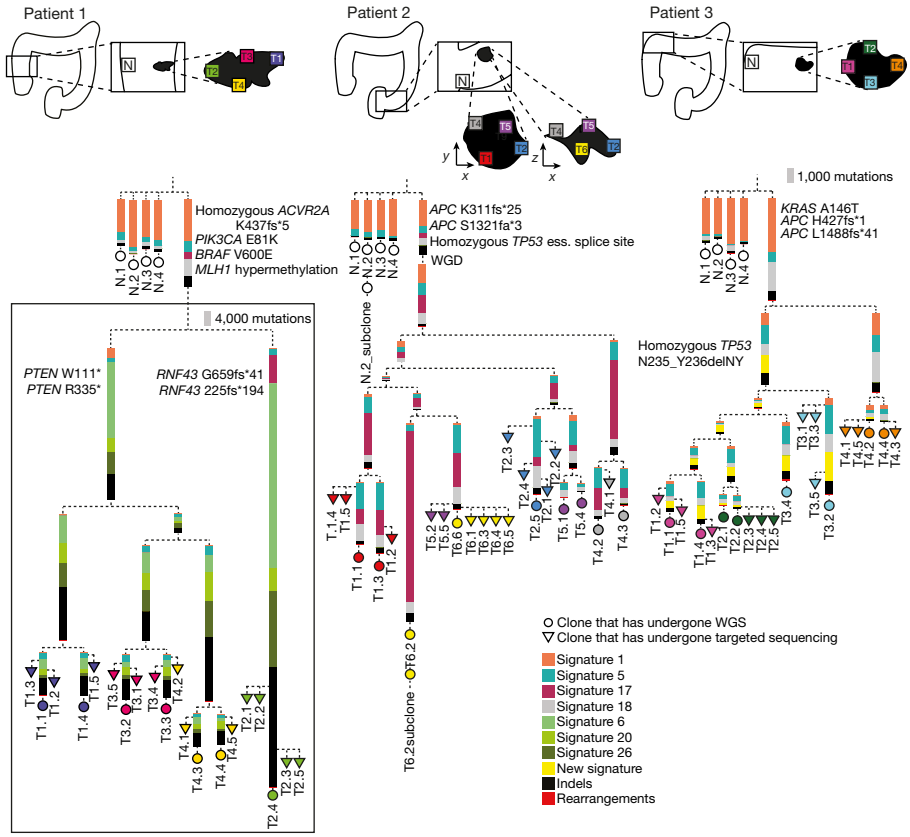


Fig. 1 | Mutation patterns during colorectal cancer evolution. Multiregion sampling of each colorectal cancer is illustrated by coloured labels (T) and normal tissue sampling by white labels (N). Phylogenetic trees from three individuals have been constructed using somatic mutations in clonal colorectal organoids derived from normal and cancer cells. Organoids underwent whole-genome sequencing (circles) or targeted cancer gene panel sequencing (triangles). The lengths of branches are proportional to mutation numbers and each mutation type and mutation signature is indicated by a different colour. Driver mutations and a whole genome duplication (WGD) are indicated in the phylogenetic tree.

were identified by comparison with the sequences of DNA extracted from pieces of normal colorectal tissue. The overwhelming majority of somatic mutations identified in this way are likely to have occurred in vivo and not during in vitro culture (Extended Data Fig. 2c). Clonal organoids were also subjected to analysis of DNA methylation state at 470,000 CpG sites, RNA sequencing (RNA-seq), and were assessed for response to several anticancer therapeutics.

Phylogenetic trees of somatic mutations

Using the catalogues of somatic mutations from clonal organoids, we derived cell phylogenetic trees for each individual (Fig. 1a–c and Extended Data Figs. 3–6). The structures of the trees generally recapitulated the geographic origins of the clonal organoids within each cancer, with more closely related branches originating from the same tumour pieces. However, organoids within each piece continued to exhibit extensive genetic diversification: for example, in organoid clones isolated from tumour section P3.T3, at least 40% of mutations were not shared with other clones from this piece (Extended Data Fig. 1o).

In the trunk of the cancer cell phylogenetic tree of P1, we identified likely driver mutations in *BRAF* (V600E), *PIK3CA* (E81K) and *ACVR2A* (protein-truncating small indel). All cancer clones from this individual also showed microsatellite instability characteristic of DNA mismatch repair deficiency and hypermethylation of the *MLH1* promoter (the likely cause of this instability) (Extended Data Fig. 7). In addition, there were likely driver truncating mutations in *PTEN* and *RNF43* that were restricted to subsets of branches of the tree. In P2, two protein-truncating *APC* mutations and a homozygous splice site *TP53* mutation were present in the cancer trunk. In P3, a *KRAS* mutation (A146T) and two truncating *APC* mutations were present in the cancer trunk, and a *TP53* in-frame deletion was present in a subset of the branches. No driver mutations or *MLH1* methylation were observed in clonal organoids derived from normal colon epithelium from the three patients.

Mutation load in normal and cancer cells

A mean of 3,792 base substitutions was found in normal organoid clones derived from P1, 3,172 from P2 and 3,621 from P3 (Fig. 2), as previously reported[24]. The mean number of base substitutions in cancer-derived clones was higher in all three individuals: 72,398 in P1, 22,291 in P2 and 14,209 in P3. There were also substantial differences in the number of small indels and genome rearrangements. A mean of 227 small indels was observed in clones derived from normal colorectal epithelium from P1, 130 from P2 and 167 from P3. By comparison, the mean number of indels was 27,893 in cancer clones from P1, 1,485 from P2 and 2,021 from P3. There was a mean of one genome rearrangement in clonal organoids derived from normal colorectal epithelial cells contrasting with means in cancer-derived clonal organoids of 71 rearrangements from P1, 176 from P2 and 67 from P3. As the normal and cancer clones are derived from cells obtained from each individual at the same times, increases in somatic base substitution, small indel and genome rearrangement mutation rates are likely to have occurred in the lineages from fertilized egg to cancer cell, including in the two cancers that were proficient in DNA mismatch repair. Most of the additional mutation loads in cancer cells were acquired in the branches of the cancer phylogenetic tree rather than the trunk and therefore occurred following the last dominant clonal expansion within the cancer cell population.

Mutational signatures

We extracted mutational signatures and estimated the contributions of each signature to each segment of the phylogenetic trees. Eight base substitution mutational signatures were found (referred to according to the nomenclature in COSMIC http://cancer.sanger.ac.uk/cosmic/signatures).

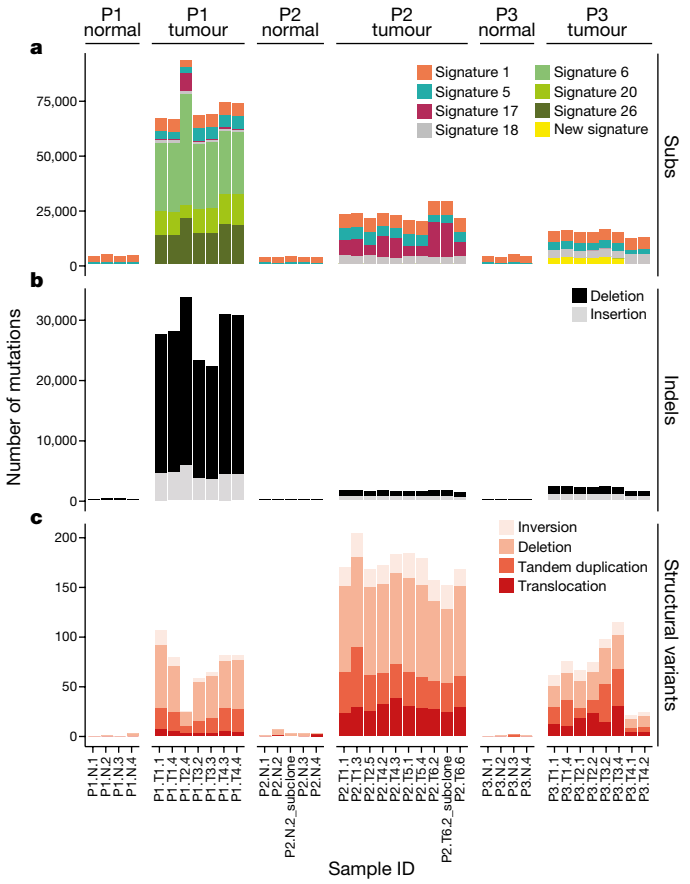


Fig. 2 | Total mutation burden in normal colorectal and colorectal cancer cells from three individuals. a, Substitutions have been further subdivided by contributions of mutational signatures. **b**, Indels have been subclassified as short insertions and deletions. **c**, Structural variations have been subdivided into deletions, inversions, tandem duplications and translocations.

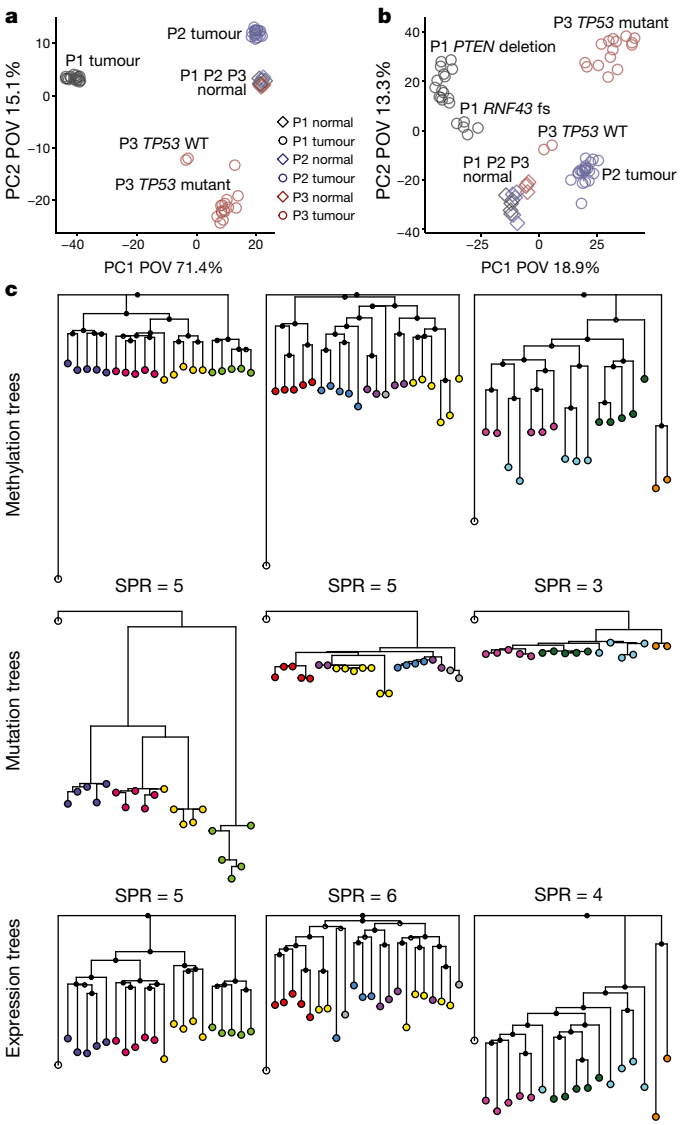


Fig. 3 | Diversification of methylation and transcriptome state during expansion of the neoplastic cell population. a, Clustering analysis of methylation state in each clonal organoid using 450 K Infinum arrays ($n = 70$ clones). PC, principal component; POV, percentage of variance. **b**, Clustering analysis of transcriptome state in each clonal organoid using RNA-seq ($n = 73$ clones). **c**. Phylogenetic trees based on methylation data (top), mutation data (middle) and expression data (bottom). Distances between mutation tree and methylation or expression tree topologies are expressed as subtree prune and regraft distance (SPR).

These include the previously described signatures 1, 5, 6, 17, 18, 20 and 26 and a signature that has not been previously encountered[32] (see also Supplementary Notes section 5). Each mutational signature can be regarded as the outcome of a mutational process, which includes components of DNA damage or modification, DNA repair (or absence of it) and DNA replication, with each component potentially influencing the profile of the signature. Signature 1 is likely to result from deamination of 5-methylcytosine to thymine and has been reported to act in a 'clock-like' manner, with mutations accumulated continuously over the lifetime of an individual at different rates in different tissues. The number of signature 1 mutations is proposed to correlate with the number of mitotic divisions[33]. Signature 5 is of uncertain origin and also shows accumulation of mutations in a clock-like manner, with different rates in different tissues, although the rates do not correlate with those of signature 1[33]. Signatures 1 and 5 are found in most human cancers and probably in most normal cells[24,33].

Signature 1 dominated and, with signature 5, accounted for the large majority of mutations in normal colorectal stem cells[24] (Fig. 1). Signature 1 also dominated in the trunks of the cancer phylogenetic trees, presumably reflecting, at least in part, the long segment of normal cell lineage from the fertilized egg to the cell in which the first cancer driver mutation was acquired. However, in each of the three cancer trunks, signatures 17 and/or 18 also showed contributions (which were not detectable in the normal clonal organoids). Signatures 17 and 18 have been found in many cancer types and have uncertain mechanisms, although signature 18 may be related to DNA damage induced by reactive oxygen species and/or by deficiency of base excision repair.

A different pattern of signature contributions was seen in the branches of the three cancer phylogenetic trees (Fig. 1). In P1, the mutations were predominantly of signatures 6, 20, 26 and indels; in P2, signatures 5, 17, 18, and indels; and in P3, signatures 5, 18, indels and a new signature predominantly characterized by T > G, T > A and T > C mutations at NTA and NTT trinucleotides (the mutated base is underlined; Extended Data Fig. 2d). The last signature occurred in all clones carrying a small in-frame deletion in *TP53*, but its relationship to this putative driver mutation is unknown. With regard to structural changes, P1 cancer clones carried deletions, inversions and tandem duplications, but few translocations, and their copy number profiles were relatively flat (Fig. 1 and Extended Data Figs. 5, 6). P2 cancer clones carried all types of rearrangement accompanied by changes in copy number profiles. In P3, *TP53* mutant clones carried abundant rearrangements of all types resulting in aberrant copy number states, whereas rearrangements were approximately tenfold less common in *TP53* wild-type clones with few copy number changes. The numbers of

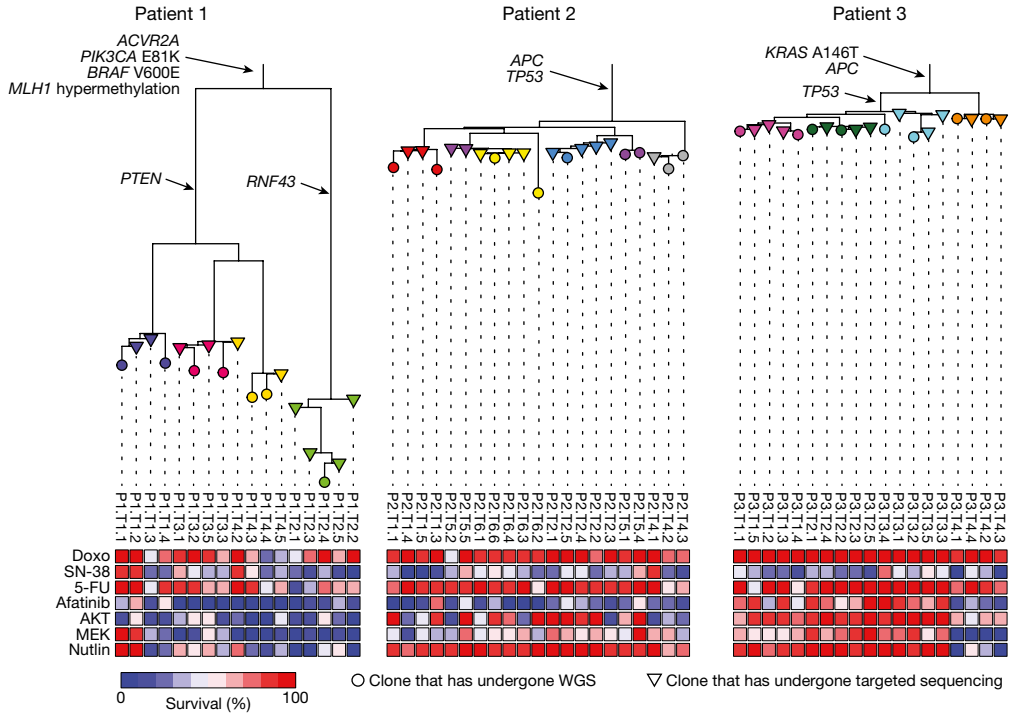


Fig. 4 | Individual clones show diverse responses to drugs commonly used in colorectal cancer treatment. Phylogenetic trees show the genetic structure of each tumour, with branch lengths representing mutation numbers. Coloured labels correspond to individual tumour segments as in Fig. 1. Mean survival in two independent experiments is displayed for exposure to doxorubicin (doxo), SN-38 (active metabolite of irinotecan), 5-FU, afatinib (an EGFR inhibitor), AKT inhibitor VIII (AKT), MEK1/2 inhibitor III (MEK) and nutlin-3a (nutlin; a stabilizer of TP53). Concentrations displayed represent intra- and inter-individual variation in the response to each drug; full dose response ranges are shown in Extended Data Fig. 10a.

mutations of several signatures differed markedly between individual branches, indicating varying contributions of mutational processes in different parts of the cancer.

There were more signature 1 mutations in each cancer organoid than in normal organoids from the same individuals (Fig. 1 and Extended Data Fig. 2e). Assuming that the clock-like correlation between the number of signature 1 mutations and number of mitoses undergone in normal cells is maintained at the same rate during neoplastic cell proliferation, we estimate that cancer cells from individual P1 have undergone 1.9 (±0.5) (s.d.) times as many mitoses as normal cells, cancer cells from individual P2 2.5 (±0.2) times as many, and from individual P3 1.7 (±0.2) times as many. An alternative explanation for the increase in signature 1 mutations in cancer cells is increased DNA methylation in cancer cells. However, cancer organoids were generally hypomethylated compared to normal cells (Extended Data Fig. 8b). In the distal branches of the phylogenetic trees (that is, during the most recent phases of cancer growth) the base substitution mutation rates per mitosis (as estimated by the total number of mutations divided by the number of signature 1 mutations) were markedly increased compared to normal cells (estimated 100-fold in P1, which is DNA mismatch-repair deficient, and tenfold in P2 and P3, which are mismatch-repair proficient). Thus, assuming that these estimates of past mitoses undergone are correct, the increases in base substitution, indel and genome rearrangement mutation rates over time also represent increases in mutation rates per mitosis.

Methylome and transcriptome

Epigenetic changes may be part of, and contribute to, the biological diversification of intra-tumour cell populations. To explore this possibility, we examined the methylation status of 470,000 CpG sites in the normal and tumour-derived clonal organoids. Organoids derived from normal stem cells from P1, P2 and P3 clustered together in principal component analyses, albeit with normal clones from each individual closer to each other than to normal clones from other individuals (Fig. 3 and Extended Data Fig. 8a). Clonal organoids from each colorectal cancer clustered together, with the exception of the two *TP53* wild-type clones of P3, but separately from those derived from the other cancers and from normal organoids. Thus, the methylation states of normal colorectal stem cells from different individuals were relatively similar, but tumours from different individuals had developed divergent epigenetic states. For the P1 tumour, this conformed to the pattern of global hypermethylation previously termed CpG island methylator phenotype (CIMP)[25]. To investigate intra-tumour diversification of transcriptome state, we performed RNA-seq of normal colorectal epithelium and cancer-derived organoid clones. Clustering by gene expression profiles correlated well with clustering by methylation, with normal organoids from all individuals clustering together, while separate clusters existed for each cancer (Fig. 3b and Extended Data Fig. 9a). For each cancer we constructed phylogenetic trees based on methylation and gene expression (Fig. 3c). The topologies of the methylation and expression trees were remarkably similar to the mutation-based trees. Thus, diversification of methylation and transcriptome state occurred within each cancer and this was apparently heritable, stable and independent of the tumour microenvironment, as it persisted after organoid culture in vitro.

Diversification of drug responses

The clonal cancer organoids were exposed in vitro to a set of drugs used to treat colorectal cancer, including the chemotherapeutic agents 5-fluorouracil (5-FU), doxorubicin and 7-ethyl-10-hydroxycamptothecin (SN-38, the active metabolite of irinotecan), and the targeted agents afatinib (an epidermal growth factor receptor (EGFR) inhibitor), nutlin-3a (a stabilizer of TP53), a MEK1/2 inhibitor and an AKT inhibitor. Different organoids from the same cancer displayed substantial and reproducible differences in IC_{50} values of up to 1,000-fold (Fig. 4 and Extended Data Fig. 10a, b), for both chemotherapeutic agents and targeted therapies. Some differences were attributable to particular somatic mutations. Notably, nutlin-3a exerted much greater growth

inhibition of *TP53* wild-type than mutant clones in P3 tumour organoids. Additionally, truncating mutations in *RNF43*, a recessive cancer gene encoding a negative regulator of the WNT pathway[34], rendered cells highly sensitive to the WNT secretion/porcupine inhibitor IWP2 (Extended Data Fig. 10c). The remaining variation in drug response did not, however, clearly relate to the geographical zones of origin or to the phylogenetic trees of each cancer. There were several examples of marked differences in drug sensitivity between closely related clones. For example, P2.T4.1 showed marked resistance to SN-38 compared to the other P2.T4 clones, whereas P3.T1.5 showed distinct sensitivity to 5-FU compared to all other clones from this individual (Fig. 4 and Extended Data Fig. 10a). The mechanism underlying this diversification in biological behaviour is unclear, but there was no obvious correlation with the degree of mutational diversification.

Discussion

Previous studies have addressed particular aspects of intra-cancer diversification by profiling the transcriptome, DNA copy number state and functional responses of individual cells[14–16,35]. To our knowledge, this is the first systematic and integrated analysis at genetic, epigenetic, transcriptomic and functional levels of multiple single-cell-derived clones from human cancers to incorporate high-quality and comprehensive description of essentially all somatic mutations present in multiple single cells. All three cancers studied, including the two DNA mismatch-repair proficient cases, clearly exhibited higher mutation burdens than normal colorectal stem cells. These are likely to result from increased mutation rates experienced during the lineages from fertilized egg to colorectal cancer cell. More comparisons of normal and colorectal cancer cells, and similar comparisons for other classes of cancer, are required for corroboration but it seems likely that increases in somatic mutation rates are common during the development of human cancers. These increases are predominantly due to recruitment of mutational processes that are inactive or marginally active in normal cells, and which dominate at later stages in the evolution of the cancer cell population. The roles of these processes in generating driver mutations, however, are unclear, as they may have started before or after acquisition of the early driver mutations in the trunks of the cancer phylogenetic trees[5]. The mechanisms underlying the increases in mutation rate in the DNA mismatch-repair proficient cancers are, for the most part, unknown. These increases may be due to somatic genetic or epigenetic changes (although these are not currently obvious), to metabolic stress attendant upon the elevated mitotic rate and other features of the neoplastic state, or to effects of the cancer cell microenvironment. Alongside intra-tumour mutational diversification, diversification of methylation state, transcriptome state and biological responses to therapeutics occur. While some methylation and transcriptome changes that occurred in vivo may have been lost during in vitro growth we were able to capture methylation and transcriptome changes that followed the evolution of the cancer through the mutational phylogenetic tree, which appeared to be stable and, at least partly, independent of the tumour cell microenvironment, as they persisted after cells were removed from the tumour. Diversification of methylation and transcriptome states and of drug responses are likely to result partly from driver mutations in cancer genes, but other, currently unknown, genetic and/or epigenetic mechanisms may be involved[36]. Future studies analysing the genomes, methylomes and transcriptomes of primary cells of a cancer will be needed to reveal all genetic, epigenetic and transcriptional variation occurring between cancer cells in vivo. Nevertheless, this study has shown the strength of the organoid system in stably retaining these characteristics and enabling functional assays on clones derived from individual cells. The analysis indicates that all three colorectal cancers contained cells resistant to most of the drugs commonly used to treat the disease. Differential drug responses between clones that are closely related both genetically and epigenetically suggest that resistance mechanisms can arise late in tumorigenesis.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0024-3.

Received: 26 January 2017; Accepted: 5 March 2018;
Published online 11 April 2018.

1. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
2. de Bruin, E. C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
3. Gundem, G. et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353–357 (2015).
4. Sottoriva, A. et al. A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).
5. Uchi, R. et al. Integrated multiregional analysis proposing a new model of colorectal cancer evolution. *PLoS Genet.* **12**, e1005778 (2016).
6. Suzuki, Y. et al. Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer. *Mol. Oncol.* **11**, 124–139 (2017).
7. Navin, N. E. The first five years of single-cell cancer genomics and beyond. *Genome Res.* **25**, 1499–1507 (2015).
8. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
9. Leung, M. L. et al. Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Res.* **27**, 1287–1299 (2017).
10. Brocks, D. et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports* **8**, 798–806 (2014).
11. Oakes, C. C. et al. Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov.* **4**, 348–361 (2014).
12. Mazor, T. et al. DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. *Cancer Cell* **28**, 307–317 (2015).
13. Caiado, F., Silva-Santos, B. & Norell, H. Intra-tumour heterogeneity — going beyond genetics. *FEBS J.* **283**, 2245–2258 (2016).
14. Stevens, M. M. et al. Drug sensitivity of single cancer cells is predicted by changes in mass accumulation rate. *Nat. Biotechnol.* **34**, 1161–1167 (2016).
15. Dubach, J. M. et al. Quantitating drug–target engagement in single cells *in vitro* and *in vivo*. *Nat. Chem. Biol.* **13**, 168–173 (2017).
16. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
17. Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
18. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
19. Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods* **13**, 229–232 (2016).
20. Meyer, M. et al. Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity. *Proc. Natl Acad. Sci. USA* **112**, 851–856 (2015).
21. van de Wetering, M. et al. Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161**, 933–945 (2015).
22. Sato, T. et al. Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. *Gastroenterology* **141**, 1762–1772 (2011).
23. Behjati, S. et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
24. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
25. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
26. Tomlinson, I. P., Novelli, M. R. & Bodmer, W. F. The mutation rate and cancer. *Proc. Natl Acad. Sci. USA* **93**, 14800–14803 (1996).
27. Loeb, L. A. Human cancers express a mutator phenotype: hypothesis, origin, and consequences. *Cancer Res.* **76**, 2057–2059 (2016).
28. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
29. Navin, N. E. Cancer genomics: one cell at a time. *Genome Biol.* **15**, 452 (2014).
30. Sieber, O. M., Heinimann, K. & Tomlinson, I. P. M. Genomic instability—the engine of tumorigenesis? *Nat. Rev. Cancer* **3**, 701–708 (2003).
31. Snippert, H. J. et al. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
32. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
33. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
34. Koo, B.-K. et al. Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. *Nature* **488**, 665–669 (2012).
35. Dalerba, P. et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nat. Biotechnol.* **29**, 1120–1127 (2011).
36. Landau, D. A. et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell* **26**, 813–825 (2014).

Acknowledgements We thank I. Martincorena, R. van Boxtel, J. Truszkowski, H. Francies and M. Garnett for discussion of our findings. This work was supported by funding from the Wellcome Trust (098051), Stichting Vrienden van het Hubrecht and KWF (SU2C-AACR-DT1213 and HUBR KWF 2014-6917). Individual authors were supported as follows: S.F.R., Louis-Jeantet Foundation; N.S., JSPS Overseas Research Fellowships; H.L.-S., Wellcome Trust Non-clinical PhD Studentship; S.B., Wellcome Trust Intermediate Clinical Research Fellowship and St. Baldrick's Foundation Robert J. Arceci Innovation Award; P.J.C., Wellcome Trust Senior Research Fellowship in Clinical Science.

Reviewer information *Nature* thanks M. Lawrence and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions A.P., N.Sm. and J.v.G. provided the samples and pathology information. N.Sa. generated organoid cultures and performed drug sensitivity assays. D.A.E. provided assistance with drug sensitivity assays. E.A., S.J.G. and C.A. provided technical assistance. S.F.R. and H.L.-S. analysed and interpreted the sequencing data. H.L.-S. derived phylogenies. L.B.A. performed signature analysis. S.F.R. and M.D.Y. analysed and interpreted methylation and expression data. H.L., S.G. and P.J.C. contributed to statistical analyses. T.J.M. performed phylogeny analysis from the tissue biopsies. S.B. and M.v.d.W. contributed to the study design. S.F.R. generated the figures. M.R.S. and H.C. directed the study. M.R.S. wrote the manuscript with contributions from all authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0024-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0024-3.

Correspondence and requests for materials should be addressed to M.R.S. or H.C.

METHODS

Human tissues. Tissue material was obtained from The Diakonessen Hospital, Utrecht. From the resected colon segment, both normal and tumour tissues were isolated. The isolated tumour tissue was subdivided into 4–5 segments. Normal tissue was taken at least 5 cm away from the tumour. All samples were obtained with informed consent and the study was approved by the ethical committee of The Diakonessen Hospital, Utrecht.

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Human organoid culture. Human normal and tumour colon organoids were established and maintained as described from isolated colonic epithelium[21,22]. In brief, long-term normal colonic organoid culture required human intestinal stem cell medium (HISM) composed of advanced DMEM/F12 (AdMEM) with penicillin/streptomycin, 10 mM HEPES, 1×GlutaMAX, 1×B27 (Invitrogen) and 1 μM *N*-acetylcysteine (SIGMA), supplemented with 50 ng ml^{-1} human recombinant EGF (Peprotech), 0.5 μM A83-01 (Tocris), 3 μM SB202190 (SIGMA), 1 μM nicotinamide (SIGMA), 10 nM prostaglandin E2,Wnt3A-conditioned medium (CM) (50% final concentration), Noggin-CM (10% final concentration), and R-Spondin1-CM (10% final concentration). Tumour organoids were cultured in medium containing only EGF, Noggin-CM, R-Spondin1-CM and A83-01.

Establishment of clonal organoids. For clonal organoids from normal crypts, isolated single crypts were embedded in 10 μl Matrigel and cultured in HISM medium. For clonal tumour organoid cultures, tumour cell suspensions were cultured for 7–14 days in HISM without Wnt3A-CM. Then, 10–15 individual organoids were picked and separately dissociated into single cells by TryPLE express (Thermo Fisher), washed and suspended in AdMEM containing propidium iodide (PI). Forty-eight single cells were sorted into tumour organoid medium (HISM plus 10 μM ROCK inhibitor Y-27632 (Tocris BioScience); no Wnt-CM) from each tumour organoid. Sorting was based on FCS area/FCS peak and PIneg/FCS area using a Moflo machine (Beckman Coulter). Sorted cells were spun down at 1,000*g* at 4 °C for 5 min, after which single cells were each embedded into 10 μl of basement membrane extract (BME, Amsbio) and seeded into 96-well plates at a ratio of 1 cell per well. The gel was left to solidify in a 37 °C incubator after HISM (no Wnt3A-CM) was added. Y-27632 was added to the medium for the first week after sorting. For each original tumour organoid, a single clonal organoid was selected and expanded for further study and for preparing frozen stocks. Culturing times and plating efficiencies are listed for each organoid in Supplementary Datafile S1.

Histology procedures. Tissues were fixed in 4% formaldehyde solution overnight and embedded in paraffin. Sections were subjected to haematoxylin and eosin (H&E) and immunohistochemistry staining. The Ki67 antibody (MONX10283, Monosan) was used at 1:250 dilution.

Organoid CellTiter-Glo viability assay. Tumour organoids were cultured for 5–10 days after being trypsinized into single cells in HISM without Wnt-3a-CM. The organoids were mechanically dissociated by pipetting before being resuspended in 5% BME/growth medium (25×10^3 organoids per ml). Before seeding, 10 μl BME was dispensed into 384-well plates, and then 30 μl growth medium containing organoids was dispensed into each plate (at about 750 organoids per well. Drug screening was carried out using nutlin-3a, afatinib, MEK1/2 inhibitor III, AKT inhibitor VIII, 5-FU, doxorubicin, and SN-38. Drug dilutions were performed in two series: 1) stepwise 2 fold-dilutions from 20 μM to 19.5 nM; and 2) stepwise 2 fold-dilutions from 15 μM to 29.3 nM. The measurements for these two dilution series were combined into a single curve. All drugs were dispensed by a HP-D300 automated liquid dispenser (TECAN). Samples were incubated for 6 days at 37 °C, and cell viability was measured by CellTiter-Glo 3D kit (Promega) on a SpectraMax M5e (Molecular Devices). Cell viability measurements were performed in duplicate wells for each clone. Survival ratios in drug-treated organoids were normalized to the average survival in a DMSO control. Each experiment was repeated on a different day. To assess variability between technical and biological replicates we calculated the area under the curve (AUC) for each survival curve. AUC values are calculated using the trapezoid method and are divided by the area covered by 100% survival on the *y*-axis and the range of the $\log_{10}$ concentrations on the *x*-axis (Extended Data Fig. 10b).

DNA and RNA extraction. DNA and RNA were concomitantly extracted from frozen tissue samples or organoid cultures using AllPrep DNA/RNA minikit (Qiagen 80204).

Whole-genome sequencing. From each individual, 7–10 tumour-derived clones were selected for whole-genome sequencing (WGS), as well as 4–5 normal-derived clones. We generated paired end sequencing reads (150 bp) using Illumina XTEN machines, resulting in ± 30× coverage per sample. Sequences were aligned to the human reference genome (NCBI build37) using BWA-MEM. Sequencing statistics are listed in Supplementary Datafile S2.

Cancer gene panel sequencing. All WGS sequenced clones and 40 additional tumour-derived clones were subjected to targeted sequencing. An in-house developed cancer gene panel (CGPv3) was used, designed to pull down 360 genes that are known or suspected to play a role in cancer[3,37]. The panel targets genes from the Cancer Gene Census (COSMIC), genes recurrently amplified or overexpressed in cancer and candidate cancer genes such as kinases from the MAP kinase signalling pathway. We performed custom RNA bait design following the manufacturer's guidelines (SureSelect, Agilent) and previously described workflows to create pulldown libraries from native genomic DNA[3,37]. Samples were multiplexed on flow cells and subjected to paired end sequencing (75-bp reads) using Illumina HiSeq2000 machines, resulting in more than 700× coverage for the target design for tumour-derived samples and more than 2,000× coverage for normal tissue-derived samples. Sequences were aligned to the human reference genome (NCBI build37) using BWA-align.

RNA sequencing. From each individual, RNA was isolated from all tumour and normal derived organoid clones and subjected to RNA-seq analysis. RNA-seq libraries were prepared according to previously described workflows and sequenced on Illumina HiSeq2000 machines[38]. Between four and seven barcoded samples were pooled per library. Sequenced reads were aligned to the human reference genome (NCBI build37) with Bowtie/TopHat.

Methylation arrays. Infinium HumanMethylation450 BeadChip arrays were used to characterize the methylation status of more than 450,000 CpG sites for all clones.

Mutation discovery. All somatic changes in whole genome and targeted data were analysed with mutation calling pipelines developed in house (available at https://github.com/cancerit).

Substitutions. Single-base somatic substitutions were identified using CaVEMan[39] and a number of post-processing filters were applied. For each patient, the only germline reference available was healthy colorectal tissue more than 5 cm from the tumour, consisting of epithelial and connective tissue. In order to allow the discovery of a field effect[40,41] that might have spread into the matched normal sample, we ran CaVEMan using an unmatched normal reference. Germline SNPs were removed by comparison to a panel of 75 unrelated normal samples. Additional post-processing filters were applied to these mutation calls as described[42]. For XTEN/BWA-mem aligned data we added two filters to the pipeline for the median alignment score (ASMD) ≥ 140 and the clipping index (CLPM) $= 0$, meaning that fewer than half of the reads should be clipped.

We then tabulated the number of mutant and wild-type reads for every mutation discovered in every sample, including the adjacent colorectal tissue. We considered only mutations that were covered by ten reads in all related clones, and mutations that were seen on at least two reads in each direction. As the adjacent normal colorectal tissue is not entirely composed of epithelium we reasoned that if there were a field effect the somatic mutations in this tissue should not be fully clonal. We therefore deducted germline mutations on the basis that they were fully clonal in the bulk normal, while mutations that were subclonal in the bulk were not removed from the analysis. To define a mutation as subclonal in the bulk, the probability of finding the observed number of mutant reads or fewer given the sequencing coverage had to be less than 0.005, based on the binomial distribution with a probability of 0.5 for autosomes. Mutations that failed to meet this criterion were considered to be germline and were removed.

Indels. Indels were called using Pindel[43,44] using the adjacent bulk colorectal tissue as a matched normal. Post processing filters were the same as for substitutions except that ASMD ≥ 140 and CLPM $= 0$ were not used.

Copy number. Copy number profiles were constructed for WGS samples by ASCAT[45,46], using adjacent healthy colorectal tissue as a matched normal. Copy number profiles of WGS analysed samples were visualized with the plotHeatmap function of the R package 'copynumber'[47].

Rearrangements. Rearrangements were called using healthy adjacent colorectal tissue as a matched normal. Abnormally paired read pairs from WGS were grouped and filtered by read remapping using 'Brass' (https://github.com/cancerit/BRASS). Read pair clusters with 50% or more of the reads mapping to microbial sequences were removed. Candidate breakpoints were matched to copy number breakpoints defined by ASCAT within 10 kb. Rearrangements not associated with copy number breakpoints or with a copy number change of less than 0.3 were removed.

Phylogenetic tree construction. The phylogeny of single-cell-derived organoids for each patient was constructed from substitutions called in WGS data. For each patient, substitutions that had been discovered by CaVEMan in any organoid from that patient were called as present or absent in each organoid using the algorithm Shearwater[48]. This algorithm compares allele frequencies of variants to a background error model derived from sequencing thousands of samples from unrelated studies on the same platform. Sequencing errors are known to occur at different frequencies across different sites of the genome[48]. By obtaining a comprehensive view of the number of variant calls at each position in unrelated normal genomes, we built an error model for each nucleotide change for each position. This method has previously been used to find variants at a low frequency[49]. Here, we compared the observed frequency of each variant to our error model. After correcting for multiple testing, only variants that were significantly mutated over the error

model were kept, using a *q* value cutoff of 0.05. Although most true variants largely exceed this threshold, this procedure maximizes the chance of retaining variants in genomic regions that have undergone copy number changes, lowering the apparent variant allele fraction (VAF), since the Shearwater algorithm was designed to detect subclonal variants. As a further stringent filter to minimise false positive calls, variants had to be supported by at least three mutant reads to be considered by the algorithm. In this way every mutation called in an individual was genotyped as being present or absent in each sample. Phylogenies were constructed using this binary matrix of mutations present or absent in each sample. Private mutations were discarded from tree building as they are uninformative. A fake outgroup with no mutations called was generated for each individual. Phylogenies were constructed using the Phylip suite of tools[50]. The programme seqboot was used to generate 100 bootstrap replicates of each dataset by resampling the mutations with replacement. Phylogenies were then reconstructed for each bootstrap replicate by maximum parsimony using the Mix programme using the Wagner method, using the fake outgroup as a root. The jumble = 10 option was used, randomizing the order of the input samples 10 times for each bootstrap replicate. Finally, the programme Consense was used to build a consensus of all the trees that had been built for each patient, using the majority rule (extended) option. This reports, for each node in the most parsimonious tree, how many of the trees that had been built contain a node that partitions the samples into the same two groups. All nodes in each tree that relate tumour samples to each other were supported by all bootstrap replicates (Supplementary Notes).

Assignment of somatic changes in WGS to the phylogenetic tree. *Substitutions.* Substitutions were called as present or absent in each organoid as described above. To assign these mutations to the tree, each branch of the tree was considered in turn. If a mutation was called in all the organoids that were descendants of a given branch, and in no organoids that were not descendants of the branch, mutations were assigned to that branch. Ignoring private mutations, which necessarily fit any tree, 97.7% of shared mutations fitted the tree structure from patient 1 perfectly, 89.7% fitted the tree from patient 2 perfectly, and 88.1% fitted the tree from patient 3 perfectly. The lower concordance with the tree for patients 2 and 3 reflects the increased copy number changes that have occurred in these phylogenies. Examination of the copy number state at loci where there were discordant mutations showed that the majority could be explained by deletions of those mutations in a subclone. Substitutions that did not fit the tree perfectly were therefore assigned to the most recent common ancestor of the samples in which they were called.

All substitution calls and their assignment to branches of the tree, as well as substitutions that did not fit the tree perfectly and their associated copy number states are listed in Supplementary Data file S4.

Indels. Indels were called as being present or absent in each sample based on a variant allele fraction (the proportion of mutant reads at a locus) cutoff. The variant allele fraction cutoff was chosen for each patient based on a histogram of the variant allele fraction to separate the sequencing noise distribution from the distribution of true mutations. Variant allele fraction cutoffs were chosen as 0.15 for patients 1 and 3, and 0.11 for patient 2. Indels were then assigned to branches of the tree that they fitted perfectly. Indels that were assigned to the tree, along with their assignments, as well as indels that did not fit the tree are provided in Supplementary Data file S4.

Rearrangements. The same rearrangement may be called in related samples with slightly different breakpoints. To identify rearrangements that had been sequenced in related clones as the same, both the upstream and the downstream breakpoints had to fall within 500 bp of each other. The majority of rearrangements fitted the tree. Visualization of discordant rearrangements using IGV[51] showed that often an overlapping rearrangement meant that the rearrangement was lost in a clone. All rearrangement calls that could and could not be assigned to the tree can be found in Supplementary Data file S4.

Timing substitutions and indels relative to a whole genome duplication (WGD). A whole genome duplication was observed in the trunk of the tumour for patient 2.

Substitutions. For substitutions, we aimed first to obtain an accurate estimate of the timing of WGD in molecular time, and second to time as many substitutions as possible relative to the WGD in order to perform signature analysis on them.

To obtain an estimate of the timing of the WGD, we examined substitutions in regions with two copies of one allele and none of the other (as determined by ASCAT). In these regions, if a mutation occurred before the WGD on that allele, it will be at copy number 2. If it occurred afterwards, it will be at copy number 1. One hundred and eighty substitutions occurred at copy number 2, and 67 at copy number 1. As at least half of the mutations that occurred before the WGD have been lost in such regions (as there is loss of one allele), the WGD can be estimated to have occurred at 84% ($(180 \times 2)/(180 \times 2 + 67)$) of molecular time in the trunk of the tumour (95% confidence interval of 80.8–87.6% calculated by bootstrapping 10,000 times).

Second, we wanted to time substitutions in regions with a greater range of copy number states for mutational signature analysis. To do this, for every truncal substitution in every tumour clone from patient 2, the copy number segment (as called by ASCAT) in which that mutation fell was defined. Mutations could be timed only in samples in which there was a minor copy number of 0 and a major copy number greater than 1. Fortunately, because of the extensive copy number changes in this tumour, all mutations fell in a region that met these criteria in at least one sample. For a given mutation that fell in such a copy number segment in a given sample, the VAF in that sample of known germline single nucleotide polymorphisms (SNPs) that fell in that segment (that necessarily occurred before the WGD) and the VAF of somatic mutations assigned to branches further down the tree (that necessarily occurred after the WGD) was examined. If, in a given sample, a mutation had a VAF greater than 90% of the VAFs of the mutations that were known to occur further down tree it was considered to have occurred before the WGD, whereas if it had a VAF less than 90% of the VAFs of the SNPs it was considered to have occurred after the WGD. If there was any overlap between the 90th percentiles of the SNPs and the later mutations, or if the mutation fitted neither of these criteria, it was considered uninformative and was not used in the signature analysis. This accounted for 9,094 mutations (out of a total of 12,623 assigned to the trunk) that were not used in signature analysis. There is no reason to believe that mutations that were excluded for these reasons should be attributable to different mutational signatures than those that could be included, and indeed their trinucleotide mutation contexts are similar (data not shown). For each mutation, then, the number of samples in which it had been counted before and after the WGD was tallied. If a mutation was called as occurring before the WGD in some samples and after the WGD in others, the mutation was designated as conflicting and excluded from the analysis. Eighty-two mutations fell into this category, and the remaining 3,447 could be timed unambiguously relative to the WGD and used in the signature analysis. In Fig. 1 we extrapolated the preWGD and postWGD fractions and their relative signature components to all mutations identified in the clonal trunk of P2. Mutations that were included in the signature analysis, those that were excluded as being uninformative, and those that were excluded as being conflicting, are reported in Supplementary Datafile S4.

Indels. For indels, we simply aimed to estimate the proportion that occurred before and after the WGD, and so for this analysis we restricted ourselves to regions of the genome with copy number 2 + 0. An analogous approach to timing substitutions was taken, although rather than considering the distribution of germline indels and indels further down the tree, a hard VAF cutoff of 0.85 (which separated the bimodal distribution of indel VAFs in these loci) was used to define mutations as occurring before or after the WGD.

Rearrangements. We timed rearrangements relative to the WGD in patient 2 by using the copy number step associated with deletions and tandem duplications in the trunk of this tumour, as determined by inspection of the change in read counts at breakpoints. The ratio of tandem duplications and deletions that had occurred before rather than after the WGD was extrapolated to give a ratio for all the rearrangements in the trunk, assuming that the relative proportion of different rearrangement classes stayed the same after the WGD.

Driver mutations. Driver mutations in TP53 and APC were timed relative to the WGD in patient 2. The TP53 mutation was at VAF 1 in a region that was 2 + 0 in all samples, indicating that it occurred before the WGD. There were mutations in both alleles (which we will call mutation 1 and mutation 2) of *APC*. P2.T4.2 and P2.T5.1 both had the *APC* locus called as 2 + 2, and both mutations were at VAF 0.5. P2.T1.1, P2.T1.3, and P2.T6.2 were 2 + 1 in the *APC* region. Mutation 1 was at VAF 0.67 and mutation 2 at 0.33. In P2.T2.5 the region was also called as 2 + 1, but mutation 1 was at VAF 0.33 and mutation 2 at VAF 0.67. This shows biallelic inactivation of *APC* before the WGD.

Assignment of samples to the tree based on targeted sequencing. Samples for which both WGS and targeted sequencing were available were used to estimate the sensitivity in the targeted data for finding substitutions that were identified in each branch by the WGS data. The targeted capture was designed against 360 cancer genes; in addition, all off-target reads that covered substitutions identified by WGS were considered. For example, in clone P1.T1.1, a fraction of 0.09 of all substitutions in the ultimate branch P1.T1.1 was found in the targeted data. Samples for which only targeted sequencing was available were then assigned to the ultimate branch of the tree with which they shared most substitutions. For example, clone P1.T1.4 shared a fraction of 0.04 with branch P1.T1.1 and negligible fractions with other branches. To estimate the time point at which P1.T1.4 branched off, we divided the shared fraction 0.04 in clone P1.T1.4 by the sensitivity in clone P1.T1.1 0.09. Thus, we estimated that P1.T1.4 shares 0.04/0.09 = 43% of mutations with branch P1.T1.1. The proportion of mutations shared with each branch are listed in Supplementary Notes section 2.

Driver analysis. To classify driver events in substitutions, indels and rearrangements we used the following criteria: 1) deleterious mutations in genes identified in CRC by TCGA[25]; 2) all other known oncogenes carrying a canonical

activating mutation; and 3) tumour suppressor genes with loss of function, and/or carrying two deleterious mutations. A more inclusive approach for identifying functional mutations is listed in Supplementary Datafile S3 and described in the Supplementary Notes.

Mutational signature analysis. Signature extraction based on non-negative matrix factorization was performed as previously described[32,52]. Mutations in trinucleotide context were grouped according to branches of the phylogenetic tree or according to sample. Datasets were combined with data from 560 breast cancer genomes to increase performance of the NNMF procedure[53].

Expression analysis. Clustering analyses were based on FPKM values calculated with the Cufflinks algorithm[54]. To select informative genes for clustering we applied the following filters: FPKM > 1; coefficient of variation across all samples > 0.7 or absolute difference > 5. Subsequently, FPKM values were $\log_2$ transformed and normalized. Normalization across samples was applied by subtracting the median expression value from individual expression values. Normalization across genes was applied by subtracting the gene's median expression from individual expression values. Normalized values were subjected to principle component analysis (PCA).

For each tumour clone we calculated a set of genes differentially expressed compared to all normal clones pooled together. Genes were called as differentially expressed if they had an FDR-corrected P value less than 0.05, resulting from a likelihood ratio test using a negative binomial generalized linear model fit with the R package 'edgeR'[55]. Raw counts were input into the edgeR model, along with normalization offsets calculated using the TMM method[56]. To construct the expression-based phylogenetic trees, we calculated Euclidean distances based on all genes that were differentially expressed in at least one tumour clone from that individual. Trees were inferred by the minimum evolution method, with the fast-me.bal function in the R-package 'ape'[57].

Methylation analysis. Methylation arrays were processed using the R package minfi[58]. We excluded three samples that failed standard QC metrics (more than 1,000 failed probes). We then excluded any probe from the analysis if it contained a variant identified in one of the samples, had a detection P value $> 1 \times 10^{-10}$ in $> 10\%$ samples or one sample with $P > 0.01$, occurred at a location known to cross-hybridize with another genomic location, or where there was a known SNP at the CpG targeted by the probe. The remaining probes were then normalized either using the 'preprocessRaw' function when comparing all samples together or using the SWAN quantile-normalization method when comparing clones from a particular tumour[59]. The latter method is appropriate when the number of methylated probes is expected to be roughly constant across all samples, which was the case for each tumour. For all comparisons, M values ($\log_2$ ratio of methylated to unmethylated probes) were calculated from the normalized probe intensities. Samples were clustered using PCA. For computational reasons, we used only the 1,000 most variable probes for PCA.

Probes that were differentially methylated between tumour and normal cells were identified using an F-test with variance shrinking and a false discovery rate of 0.01 via the 'dmpFinder' function comparing each tumour clone to all normal derived clones[58]. To construct the methylation-based phylogenetic trees, we calculated Euclidean distances based on probes that were differentially methylated in at least one tumour clone from that individual. For computational reasons, we used a q-value cutoff of 1×10^{-8} for selection of informative probes. Trees were inferred by the minimum evolution method, with the fast-me.bal function in the R-package 'ape'[57].

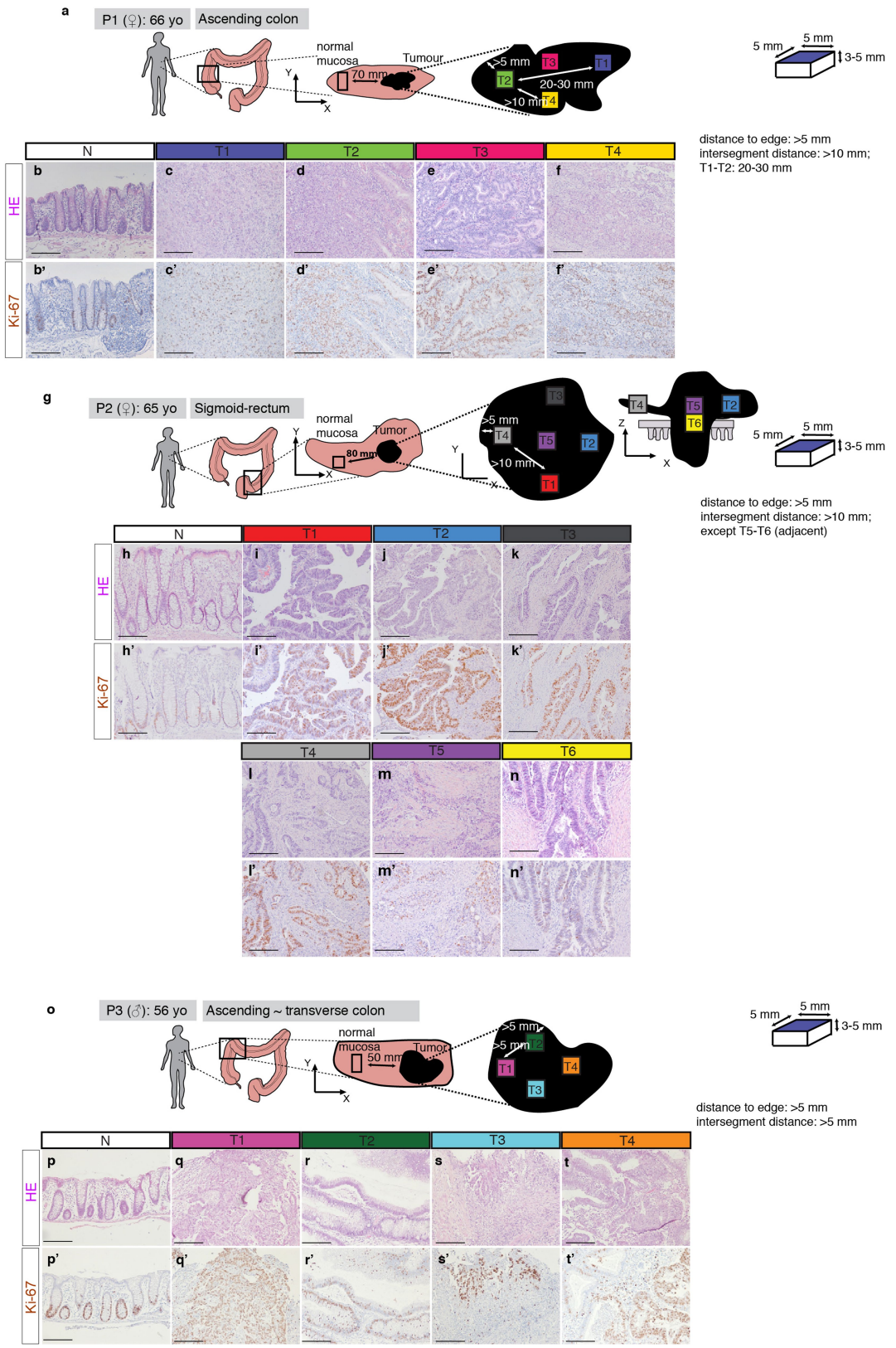
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Mutation calling pipelines developed in house are available at https://github.com/cancerit. The Shearwater algorithm for deriving a background error model is available at: https://www.bioconductor.org/packages/devel/bioc/vignettes/deepSNV/inst/doc/shearwaterML.html. The software for signature analysis used in this manuscript is available at: https://www.mathworks.com/matlabcentral/fileexchange/38724-wtsi-mutational-signature-framework.

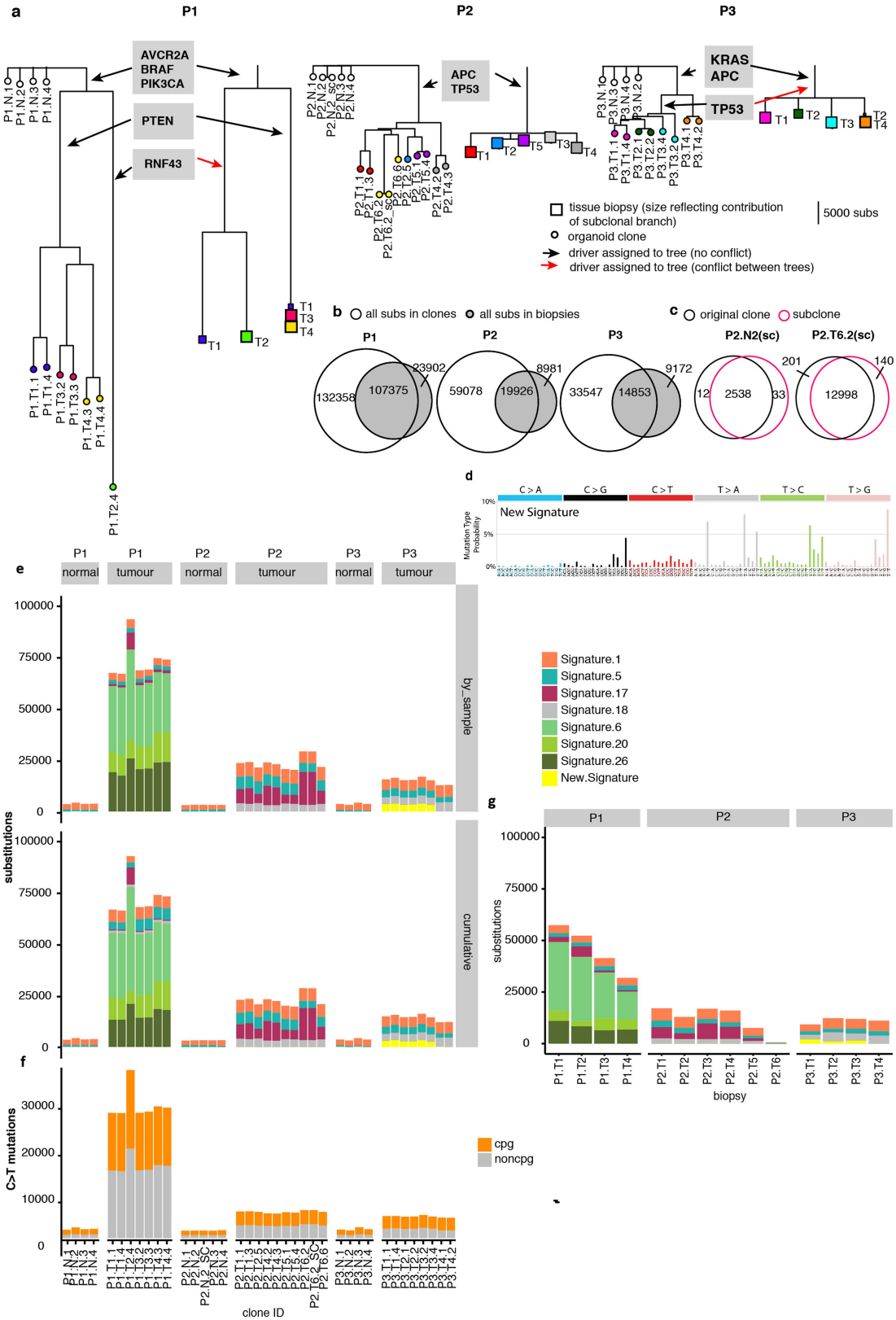
Custom R scripts developed for the analyses and visualizations in this manuscript are available from the authors on request.

Data access. Sequencing data have been deposited at the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) under accession numbers EGAS00001000869 (targeted sequencing), EGAS00001000985 (RNA-seq) and EGAS00001000881 (WGS). RNA sequencing data of these organoid clones has also been described elsewhere[60].

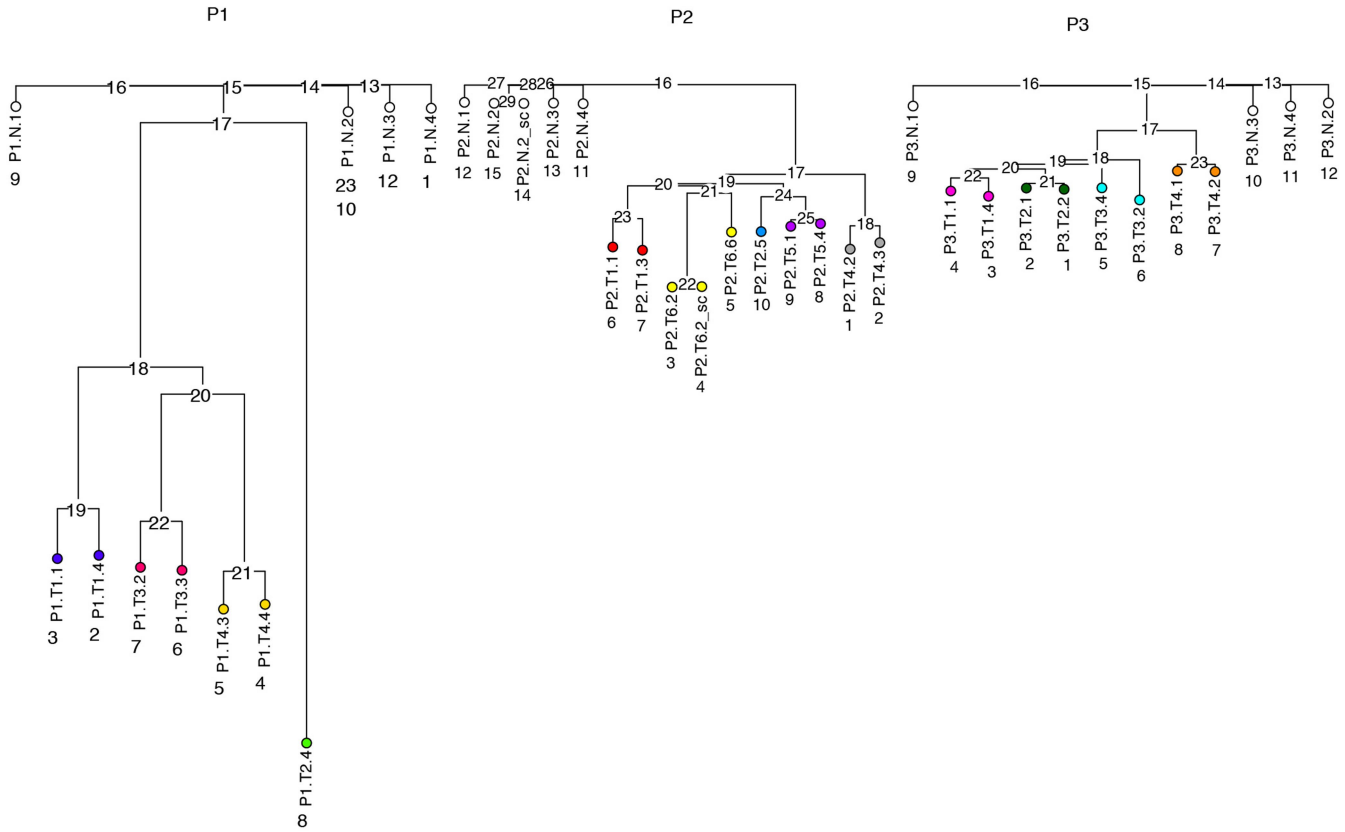
37. Yates, L. R. et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).
38. Iorio, F. et al. A landscape of pharmacogenomic interactions in cancer. *Cell* **166**, 740–754 (2016).
39. Jones, D. et al. cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
40. Lochhead, P. et al. Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. *Mod. Pathol.* **28**, 14–29 (2015).
41. Luo, Y., Yu, M. & Grady, W. M. Field cancerization in the colon: a role for aberrant DNA methylation? *Gastroenterol. Rep. (Oxf.)* **2**, 16–20 (2014).
42. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
43. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
44. Raine, K. M. et al. cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
45. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1–15.9.17 (2016).
46. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
47. Nilsen, G. et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).
48. Gerstung, M., Papaemmanuil, E. & Campbell, P. J. Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics* **30**, 1198–1204 (2014).
49. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
50. Felsenstein, J. PHYLIP — Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
51. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
52. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
53. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
54. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
55. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
56. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
57. Desper, R. & Gascuel, O. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.* **9**, 687–705 (2002).
58. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
59. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).
60. Borten, M. A., Bajikar, S. S., Sasaki, N., Clevers, H. & Janes, K. A. Automated brightfield morphometry of 3D organoid populations by OrganoSeg. *Sci. Rep.* (in the press).



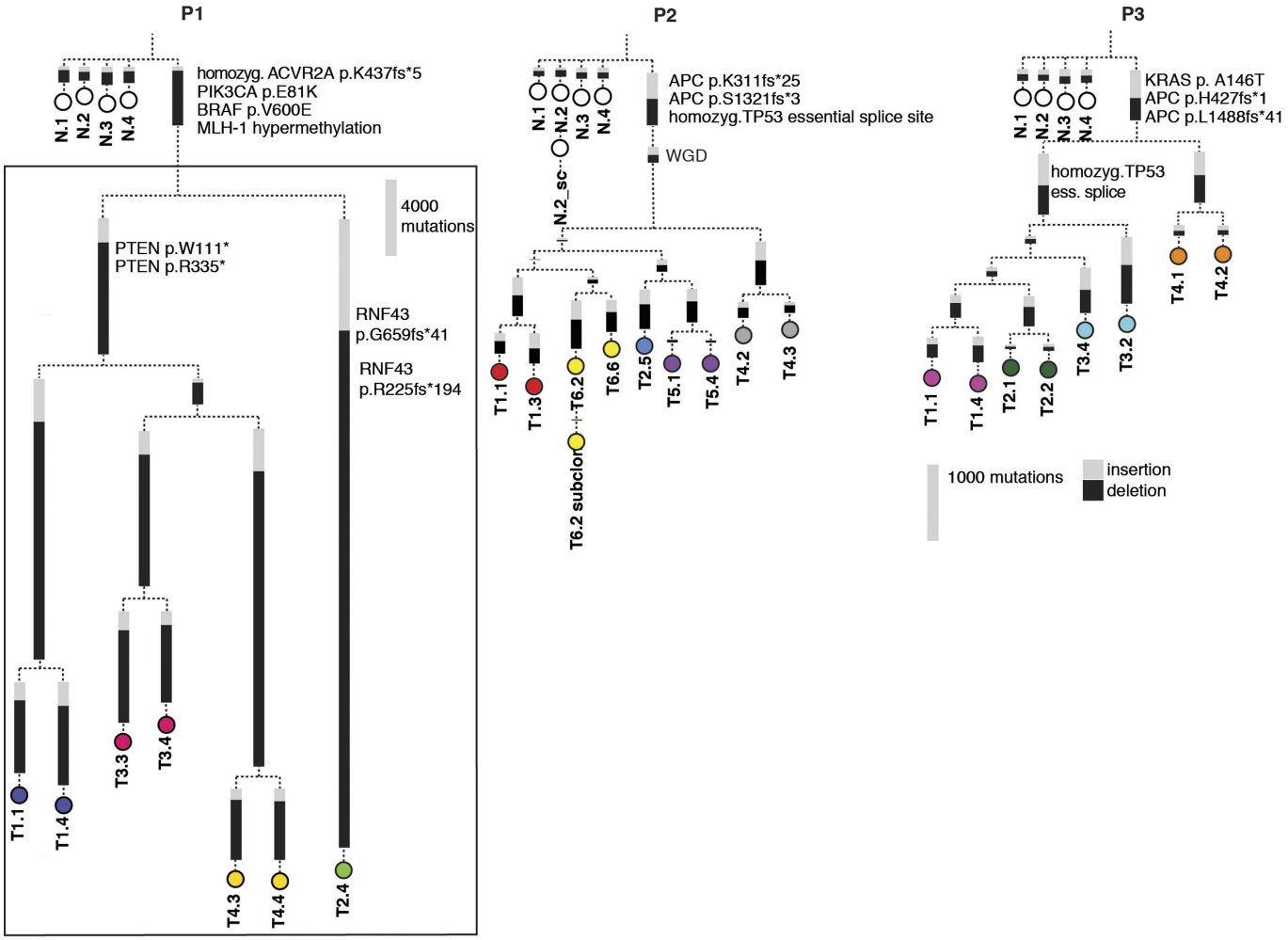
Extended Data Fig. 1 | Origin of clonal organoids analysed in this study. Specimens were derived from the ascending colon of a 66-year-old woman (**a**–**f**), sigmoid rectum of a 65-year-old woman (**g**–**n**) and ascending transverse colon of a 56-year-old man (**o**–**t**), respectively. From each tumour, 4–6 segments were resected (sized 5 × 5 × 3–5 mm. All sections except T3 from P2 resulted in viable clonal organoids. **b**–**f**, **h**–**n**, **p**–**t**, Haematoxylin and eosin staining and Ki67 immunohistochemistry show cell morphology for individual tumour sections. Scale bars: 200 μm.



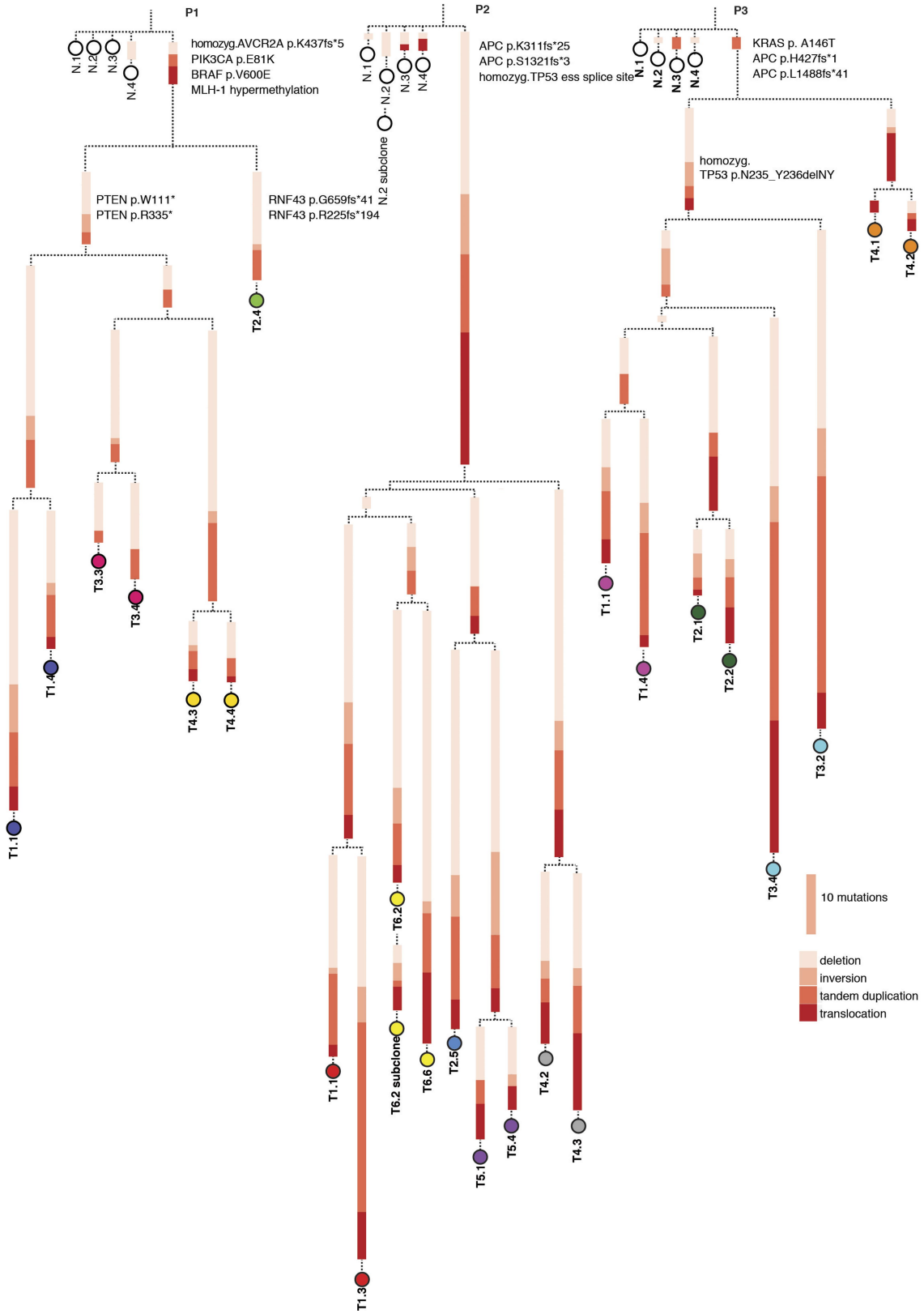
Extended Data Fig. 2 | Substitution analysis. a, Comparison of phylogeny reconstructions from WGS analysis of clonal organoids (left) and subclonal analysis of the original tissue biopsies (right) from individuals P1–P3. The analysis of clonal organoids allows a very detailed phylogeny, exact placement of driver mutations and analysis of cell-to-cell differences. **b**, Venn diagrams depicting overlap between substitutions identified by the organoid approach and the tissue biopsy approach. **c**, Venn diagrams depicting overlaps between clones P2.N3 and P2.T6.2 and their respective subclones (see Methods). Only a small proportion of the total mutations is added during culturing in both normal and tumour organoids. **d**, New signature identified in this study in tumour organoid samples from P3, characterized by T > G, T > A and T > C mutations at NTA and NTT trinucleotides (mutated bases underlined). **e**, Contribution of each of the identified mutation signatures to individual samples. Top (by_sample), results of signature extraction from all substitutions identified in each sample (Supplementary Notes). Bottom, proportion in each sample derived by adding up proportions in the branches of the phylogenetic tree that make up that sample (identical to Fig. 1). **f**, Numbers of C > T mutations by CpG context. **g**, Signature analysis of substitutions identified in the original tissue biopsies.



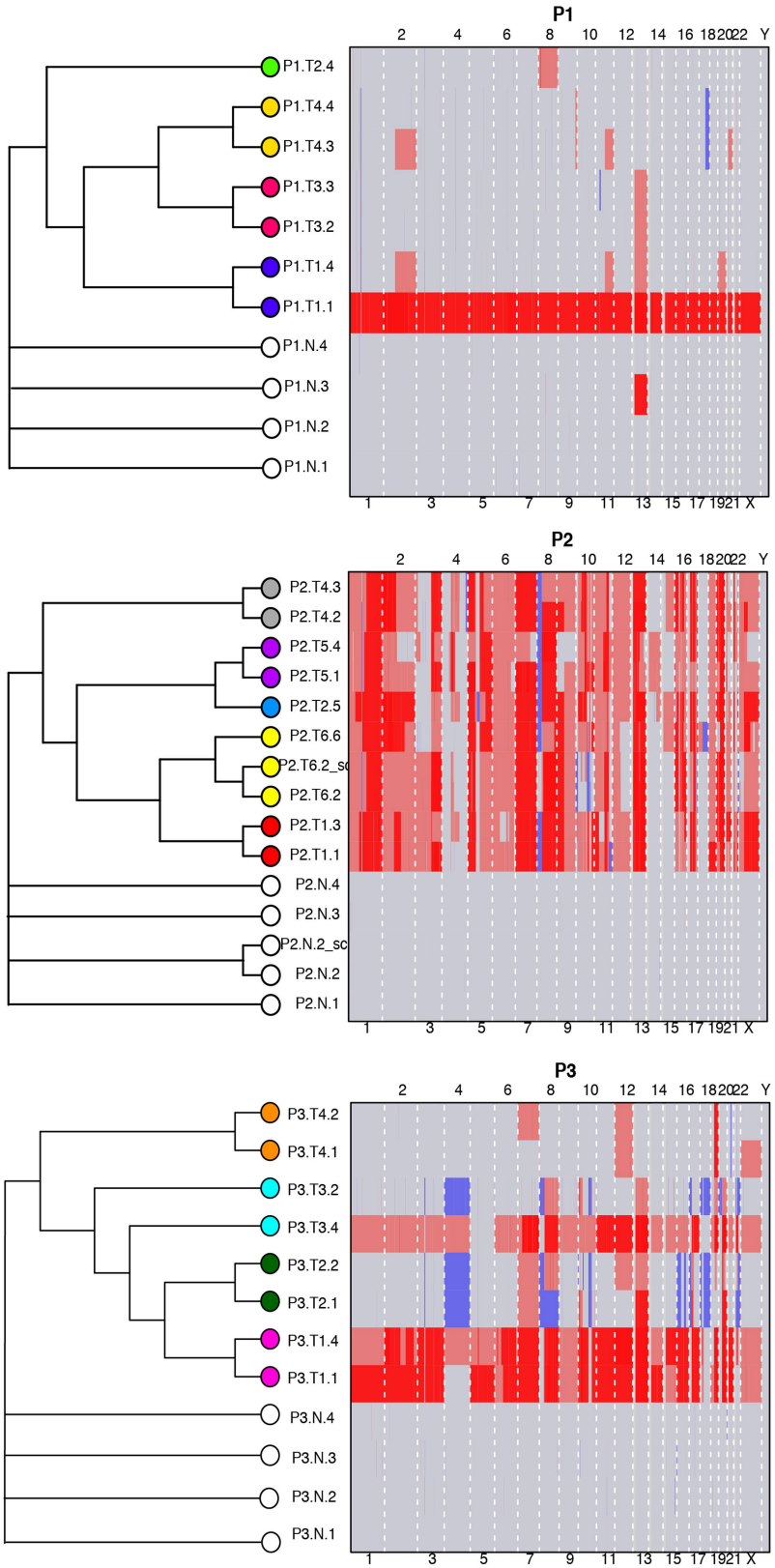
Extended Data Fig. 3 | Phylogenetic trees for clones that have been analysed by WGS. Branch lengths represent total mutation numbers; labels of nodes and tips in the tree correspond with labels in Supplementary Data files S3–S5.



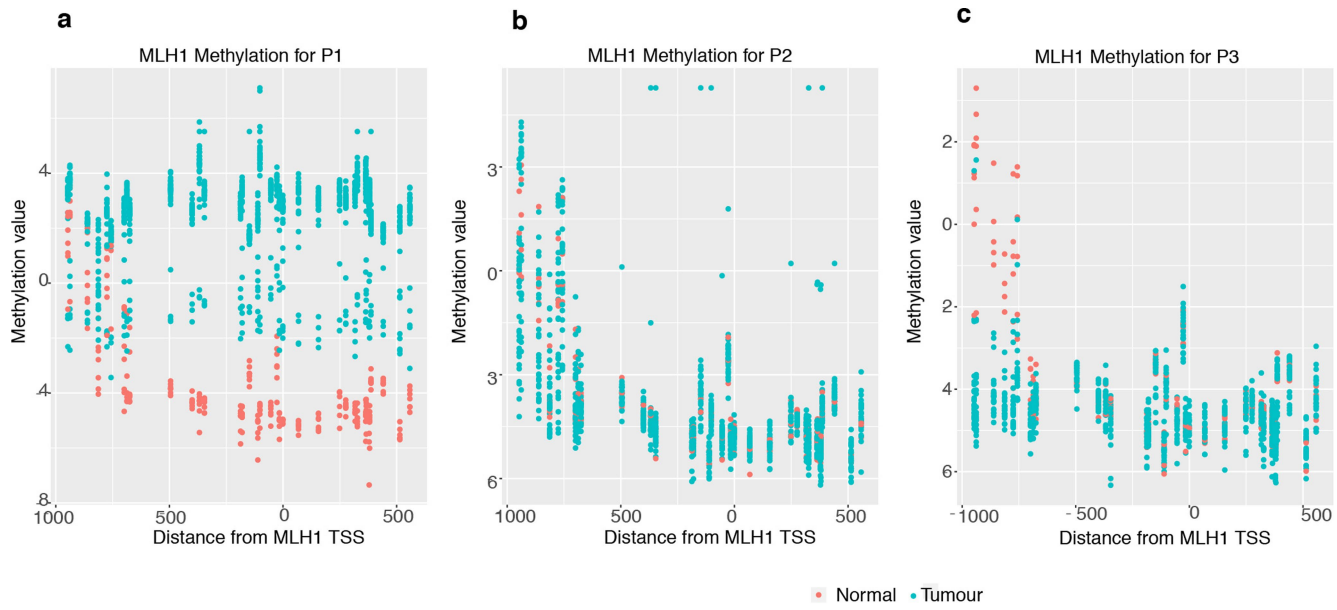
Extended Data Fig. 4 | Phylogenetic trees for indels. Phylogenies for three individuals with branch lengths representing indel numbers, further subdivided in insertions and deletions. Boxed area for P1 shows the high number of indels in this patient, who displays microsatellite instability (MSI) in all tumour clones in a different scale.

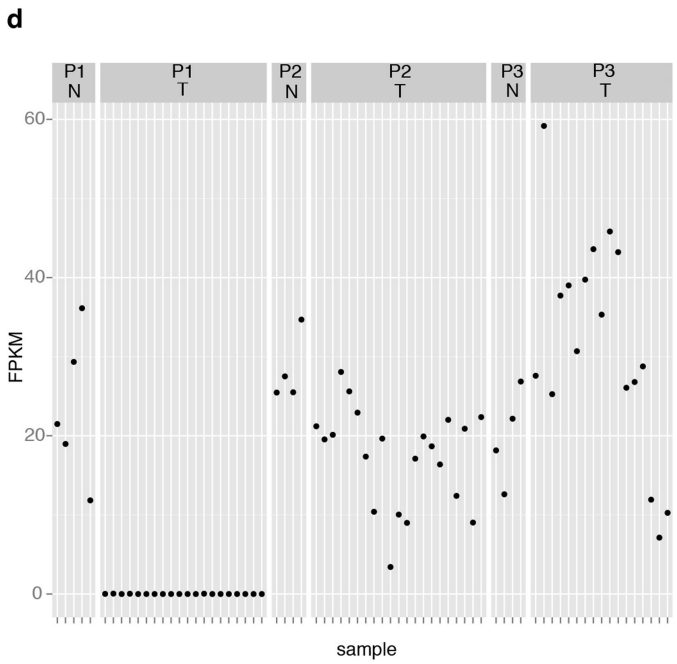


Extended Data Fig. 5 | Phylogenetic trees for rearrangements. Phylogenies for three individuals with branch lengths representing rearrangement numbers, further subdivided into deletions, inversions, tandem duplications and translocations.

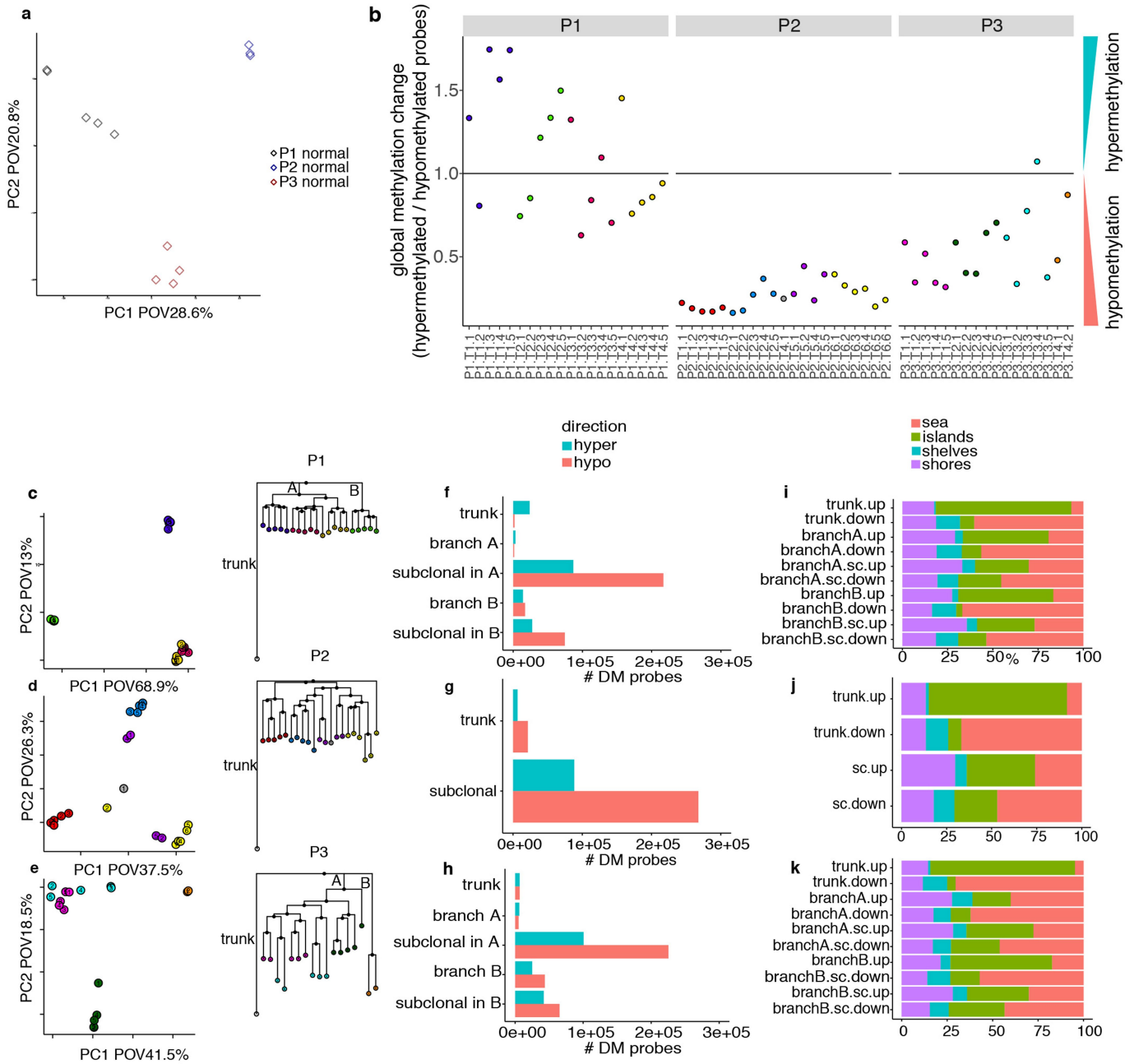


Extended Data Fig. 6 | Copy number analysis. Copy number profiles of all clones that have been WGS analysed, displayed as a heatmap (amplification in red, loss in blue). The structures of the phylogenetic trees are displayed on the left; branch lengths are not scaled.

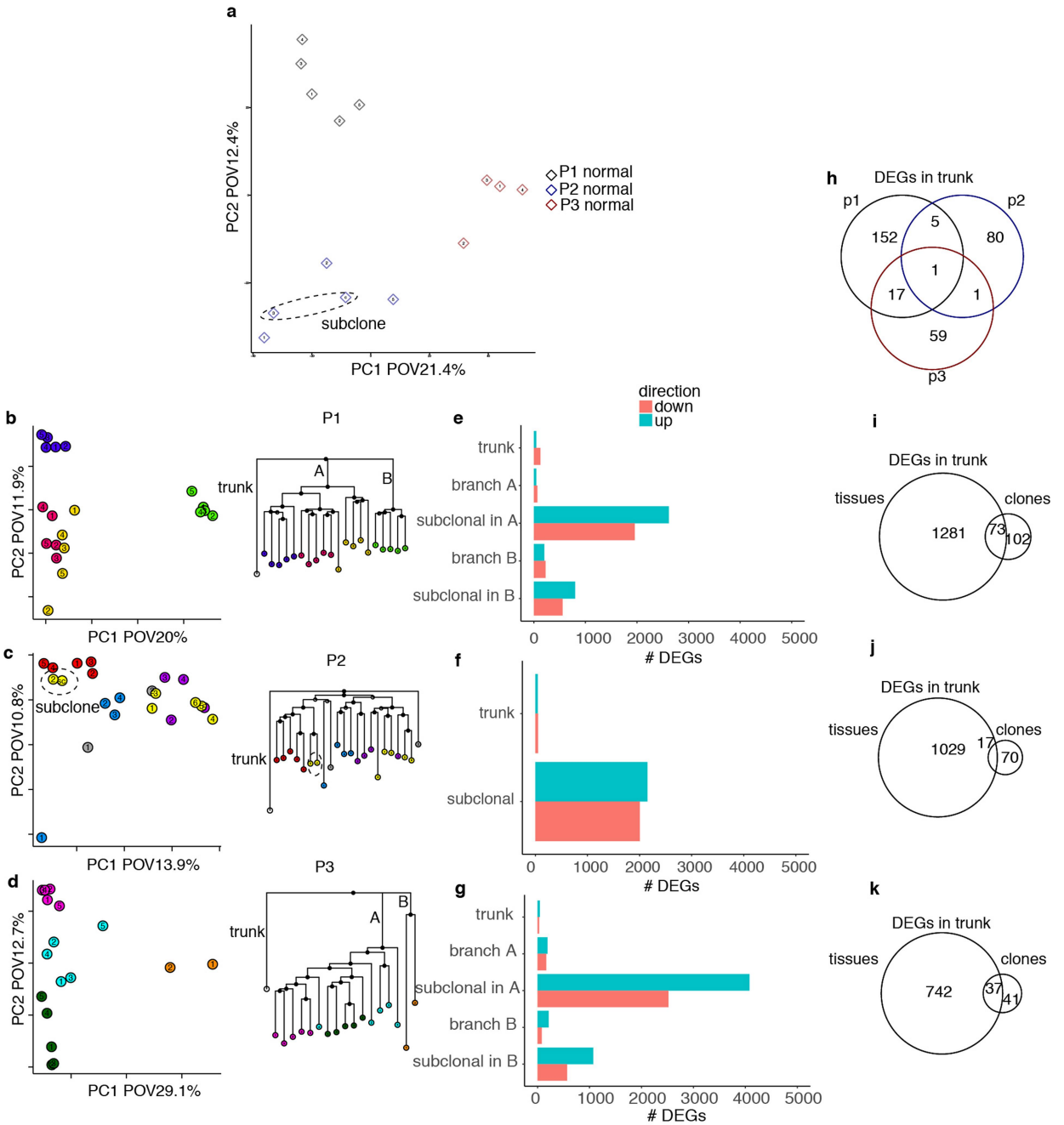




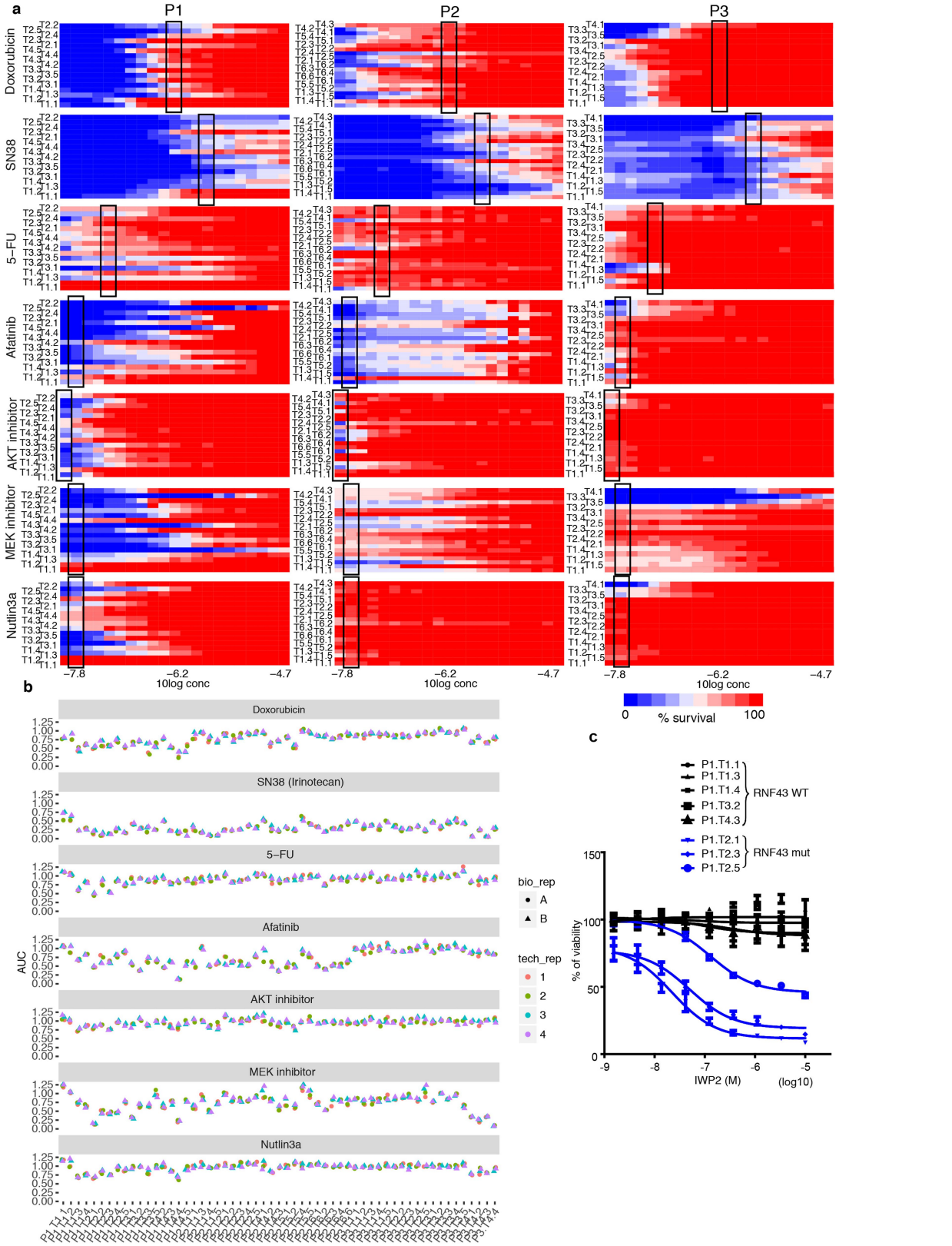
Extended Data Fig. 7 | *MLH1* hypermethylation in P1. a–c, Methylation pattern of the *MLH1* gene for tumour and normal clones for three individuals, showing hypermethylation in proximity to the transcription start site (TSS) for P1 tumour clones compared to normal clones. **d**, Expression of *MLH1* in all clones; *MLH1* transcript could not be detected in tumour clones from P1.



Extended Data Fig. 8 | Methylation analysis. a, Clustering of methylation data by PCA showing normal-derived organoids from three individuals ($n = 12$ biologically independent samples). **b**, Global methylation change in each tumour clone, expressed as the ratio of hypermethylated probes to hypomethylated probes. Hyper- and hypomethylation are assessed by comparing to the baseline methylation levels in the normal-derived clones (indicated with line at $y = 1$). **c**–**e**, Left, clustering of methylation data by PCA of tumour organoids from each individual, displaying the first two principal components. Clones from different segments are shown in different colours as in Extended Data Fig. 2. Right, phylogenetic trees based on expression data (as in Fig. 3b) with the main branches used for our methylation analysis indicated. **c**, P1, $n = 20$ biologically independent samples. **d**, P2, $n = 21$ biologically independent samples. **e**, P3, $n = 17$ biologically independent samples. **f**–**h**, Direction of methylation changes during tumour development. Methylation changes were assigned to either the branch of the tumour or the main subclonal branches (indicated in the phylogenetic trees in **e**). **i**–**k**, Relative proportion of probes in CpG islands, shores, shelves and seas that were differentially methylated in different branches (Supplementary Notes section 6).



Extended Data Fig. 9 | Expression analysis. a, PCA based on expression pattern of normal organoids from each individual, displaying the first two principal components ($n = 13$). A subclone and its ancestral clone are circled. **b**–**d**, Left, PCA of tumour clones from each individual. Clones derived from different segments are shown in different colours as in Figs. 2–4. A subclone derived from a tumour clone from P2 and its ancestor clone are circled. Right, phylogenetic trees based on expression data (as in Fig. 4b) with the main branches used for our expression analysis indicated. **b**, P1, $n = 20$ biologically independent samples. **c**, P2, $n = 22$ biologically independent samples. **d**, P3, $n = 17$ biologically independent samples. **e**–**g**, Global analysis of expression changes attributed to the trunk of the tree, the main branches or subclonal variation. **h**, Venn diagram displaying the differentially expressed genes that were attributed to the trunk of each tumour. Differentially expressed genes determined by a likelihood ratio test using a negative binomial generalized linear model fit (FDR < 0.05). **i**–**k**, Comparison of differentially expressed genes identified in the organoid clones of each patient versus the original tissue sections. Only genes that were significantly altered in all clones or all biopsies from each individual are considered.



Extended Data Fig. 10 | Drug response data. Dose response data for seven drugs, tested on organoid clones from three individuals. Twenty-one concentrations were tested for each drug, ranging from 14.7 nm to 20 μM. Mean survival from two duplicate experiments is displayed in a heatmap. The concentration displayed in Fig. 4 is outlined with a black box in each panel. **b**, Reproducibility of drug response data. Each measurement was performed twice (technical replicate) and each experiment was performed in duplicate (biological replicate). For each biological or technical replicate the area under the curve (AUC) is shown. **c**, Dose–response curves after 6 days of treatment with IWP2 (Wnt secretion inhibitor) for clonal tumour organoids derived from P1. *RNF43* mutant clones are responsive, whereas *RNF43* wild-type (WT) clones are resistant. Data points and error bars represent the mean and s.d. of four independent technical replicates from two independent experiments.

nature research

Corresponding author(s): Michael R Stratton(mrs@sanger.ac.uk); Hans Clevers (h.clevers@hubrecht.eu)

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

Describe how sample size was determined.

No statistical methods were used to predetermine sample size.

2. Data exclusions

Describe any data exclusions.

We excluded one normal derived clone from P1 from our analysis. Although all characteristics (mutation numbers and signatures) were in the range of the other normals, we could not confirm that this clone was an independent sample or a subsample from P1.N.2. None of our conclusions would be affected by either including or excluding this sample.

3. Replication

Describe whether the experimental findings were reliably reproduced.

Drug response testing has been replicated and reliably reproduced (extended data figure 10)

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Not relevant to this study; we describe each case on itself and do not make assumptions on groups that they represent.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

Investigators were blinded to each cloneID, but not individual's ID during data collection and analysis. The reason for this is that samples from the three individuals were collected at different timepoints - ie after they underwent surgery.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

alignment:
BWA-aln version 0.5.9-r16+rugo (targeted data)
BWA-mem version 0.7.12-r1039 (whole genome data)
variant calling:
Caveman 1.11.0 (available from https://github.com/cancerit)
Pindel 2.1.0 (available from https://github.com/cancerit)
Brass 3.0.4 (available from https://github.com/cancerit)
ASCAT 1.5.1 (available from https://github.com/cancerit)
copynumber 1.16.0 (R package available from https://bioconductor.org/packages/release/bioc/html/copynumber.html)
ShearwaterML: R package deepSNV version 1.21.4 available from https://www.bioconductor.org/packages/devel/bioc/vignettes/deepSNV
Phylogeny analysis:
phylip suite of tools: phylip-3.695 (available from http://evolution.genetics.washington.edu/phylip.html)
ape 4.1 (R package available from https://cran.r-project.org/web/packages/ape/index.html)
Signature extraction: available from https://www.mathworks.com/matlabcentral/fileexchange/38724-wtsi-mutational-signature-framework
RNAseq analysis:
Bowtie 0.12.7
TopHat 1.3.3
Cufflinks 1.0.2
GoSeq 1.30.0 (Rpackage available from https://bioconductor.org/packages/release/bioc/html/goseq.html)
edgeR 3.20.2 (R package available from http://bioconductor.org/packages/release/bioc/html/edgeR.html)
Methylation analysis:
minfi 1.24.0 (R package available from https://bioconductor.org/packages/release/bioc/html/minfi.html)

Custom Rscripts to create the figures of this manuscript have been developed in R-3.1 for this manuscript and are available upon request from the authors).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

All unique materials are available from the authors upon request.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

Ki67 antibody (MONX10283, Monosan) was used and extensively validated for use in human colorectal tissue in previous work (eg Drost et al, Nature 2015)

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used

c. Report whether the cell lines were tested for mycoplasma contamination.

no eukaryotic cell lines were used

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

no commonly misidentified cell lines were used

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	No animals were used

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Samples from 3 individuals have been studied: P1 is a 66 yo male with a tumour in the ascending colon P2 is a 65 yo male with a tumour in the sigmoid rectum P3 is a 56 yo female with a tumour in the ascending - transverse colon