The co-evolution of the genome and epigenome in colorectal cancer

Timon Heide[1,2,], Jacob Househam[1,3,*], George D Cresswell[1], Inmaculada Spiteri[1], Claire Lynn[1], Maximilian Mossner[1,3], Chris Kimberley[3], Javier Fernandez-Mateos[1], Bingjie Chen[1], Luis Zapata[1], Chela James[1], Iros Barozzi[4,5], Ketevan Chkhaidze[1], Daniel Nichol[1], Vinaya Gunasri[1,3], Alison Berner[3], Melissa Schmidt[3], Eszter Lakatos[1,3], Ann-Marie Baker[1,3], Helena Costa[6], Miriam Mitchinson[6], Rocco Piazza[7], Marnix Jansen[6], Giulio Caravagna[1,8], Daniele Ramazzotti[7], Darryl Shibata[9], John Bridgewater[10], Manuel Rodriguez-Justo[6], Luca Magnani[4], Trevor A Graham[1,3,§], Andrea Sottoriva[1,2,§]*

[1] Evolutionary Genomics and Modelling Lab, Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK
[2] Computational Biology Research Centre, Human Technopole, Milan, Italy
[3] Evolution and Cancer Lab, Centre for Genomics and Computational Biology, Barts Cancer Institute, Queen Mary University of London, London, UK
[4] Department of Surgery and Cancer, Imperial College London, London, UK
[5] Centre for Cancer Research, Medical University of Vienna, Vienna, Austria
[6] Department of Pathology, University College London, London, UK
[7] Department of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy
[8] Department of Mathematics and Geosciences, University of Triest, Triest, Italy
[9] Department of Pathology, University of Southern California Keck School of Medicine, Los Angeles, CA, 90033, USA
[10] UCL Cancer Institute, University College London, London, UK

[*] equal contribution
[§] Correspondence to: trevor.graham@icr.ac.uk and andrea.sottoriva@fht.org

Abstract

Colorectal malignancies are a leading cause of cancer death. Despite seminal genomic studies, DNA alterations alone do not fully explain malignant evolution. Here we investigate the co-evolution of the genome and epigenome of colorectal tumours at single-clone resolution using spatial multi-omics of individual glands. We collected 1,373 samples from 30 primary cancers and 9 concomitant adenomas and generated 1,212 chromatin accessibility profiles, 527 whole-genomes and 297 whole-transcriptomes. We found positive selection for DNA mutations in chromatin modifier genes and recurrent somatic chromatin accessibility alterations (SCAAs), including in regulatory regions of cancer drivers devoid of genetic mutations. Genome-wide alterations in transcription factor binding accessibility involved *CTCF*, downregulation of interferon, and increased accessibility for *SOX* and *HOX*, suggesting developmental genes involvement. SCAAs were heritable and distinguished adenomas from cancers. Mutational signature analysis showed the epigenome influencing DNA mutation accumulation. This study provides a map of (epi)genetic tumour heterogeneity, with fundamental implications for understanding colorectal cancer biology.

Introduction

Clonal evolution, fuelled by intra-tumour heterogeneity, drives tumour initiation, progression and treatment resistance[1,2]. Much is known about the genetic evolution and intra-tumour heterogeneity of colorectal malignancies[3–5]. Although genetic heterogeneity is widespread[6], epigenetic changes are also responsible for phenotypic variation between cancer cells[7–10]. Epigenetic profiling of chromatin accessibility in colon cancer has been performed in seminal studies in cell lines[11] and human samples[12]. However, current investigations are limited to single bulk samples and some also lack normal controls[13]. Moreover, how cancer genomes and epigenomes concomitantly evolve and shape intra-tumour genetic and epigenetic heterogeneity remains unexplored.

Measuring genome-epigenome co-evolution in a quantitative manner requires multi-omic profiling at single clone resolution and accurate spatial sampling of human neoplasms, as well as matched normal tissue. Colorectal cancers (CRCs) are organised into glandular structures, reminiscent of the crypts of the normal intestinal epithelium[14]. Normal crypts are tube-like invaginations where cell proliferation is driven by a relatively small number of stem cells at the base[15–18] and cancer glands are thought to have the same architecture[19]. This implies that all cells within a gland share a recent common ancestor and are a few cell divisions apart: thus, glands are largely clonal populations that, through cell proliferation, copy DNA with relatively high fidelity. Ultimately, the gland can be thought of as a natural "whole-genome amplification machine" that can be exploited to perform multi-omics at single clone resolution. Indeed, single crypt and gland genomic profiling has been long used to study clonal dynamics in both normal[20–22] and cancer cells[4,23–28]. We have developed a new method to concomitantly profile single nucleotide variants (SNVs), copy number alterations (CNAs), chromatin accessibility with ATAC-seq[29] and full transcriptomes with RNA-seq from the same individual gland or crypt. Here we present the results of multi-region single gland multi-omics of 1,373 samples from 39 lesions arising in 30 patients, with 23-57 tumour samples per patient (median=43).

Results

Single gland multi-omics

We prospectively collected fresh resection specimens from 30 stage I-III primary colorectal cancers and 9 concomitant adenomas belonging to 30 patients referred for surgery at the University College London Hospital (Figure 1A, Table S1 for clinical information, Methods Section 1.1). Single gland isolation was performed from normal and neoplastic tissue (Figure 1B, Methods Section 1.2), followed by separation of nuclei from cytosol (Figure 1C). Leftover fragments that remained after gland isolation were retained to assess how representative glands are of the bulk they originated from. We referred to those samples, consisting of a few tens of glands, as "minibulks". We used the nuclei to perform whole-genome sequencing and chromatin accessibility profiling with ATAC-seq and the cytosol to perform full transcriptome RNA-seq (Figure 1D, Methods Section 2.1). We verified that cytosolic RNA expression in our normal colon tissue controls was highly correlated with whole-cell RNA expression from the TCGA cohort[5] (Figure S2).

Our tumour spatial sampling strategy was designed to measure clonal evolution at multiple scales. We first sampled four spatially distant regions of a given cancer (regions A,B,C,D) located close to the tumour edge, one distant normal epithelium region (region E), and concomitant adenomas if present (region F,G,H). A bulk sample was collected from each region and was spatially annotated in the original resection specimen (Figure 1E and S1).

Commented [AS1]: Comment 1.7

Commented [AS2]: Comment 1.8

Each piece was cut into 4 subregions (e.g., A1-A4, B1-B4, …) as shown in Figure 1E (bottom-right). We then collected and profiled 12-40 (median=37) individual tumour glands and 3-18 (median=4) minibulks per patient, a few healthy crypts and a minibulk from the matched normal, as well as blood when available (Figure 1F and S3). C542 sample F was originally labelled as adenoma but confirmed to be part of the cancer upon histopathology re-evaluation (Figure S1).

ATAC-seq was performed in 18-61 samples per patient (median=42, Table S2, Methods Section 2.2), deep whole-genome sequencing (WGS, median depth 35x) in 3-15 samples per patient (median=8.5), and low-pass whole genome sequencing (lpWGS, median depth 1x) in 1-22 samples per patient (median=8) – see Table S3 and Methods Section 2.3. For a proportion of samples (n=382/1,373) both WGS and ATAC-seq data were available (Figure 1G). We also generated a total of 623 whole-transcriptomes, of which 297 were of high quality to be used for analysis (1-40 samples per patient, median=7, Methods Section 2.4) with many also overlapping the WGS dataset, the ATAC-seq dataset or both (Figure 1H). In addition, we ran methylation arrays on 8 samples (Methods Section 2.5). We identified copy number alterations (CNAs), somatic single nucleotide variants (SNVs), indels (Indels), and ATAC peaks for all samples (Methods Section 3).

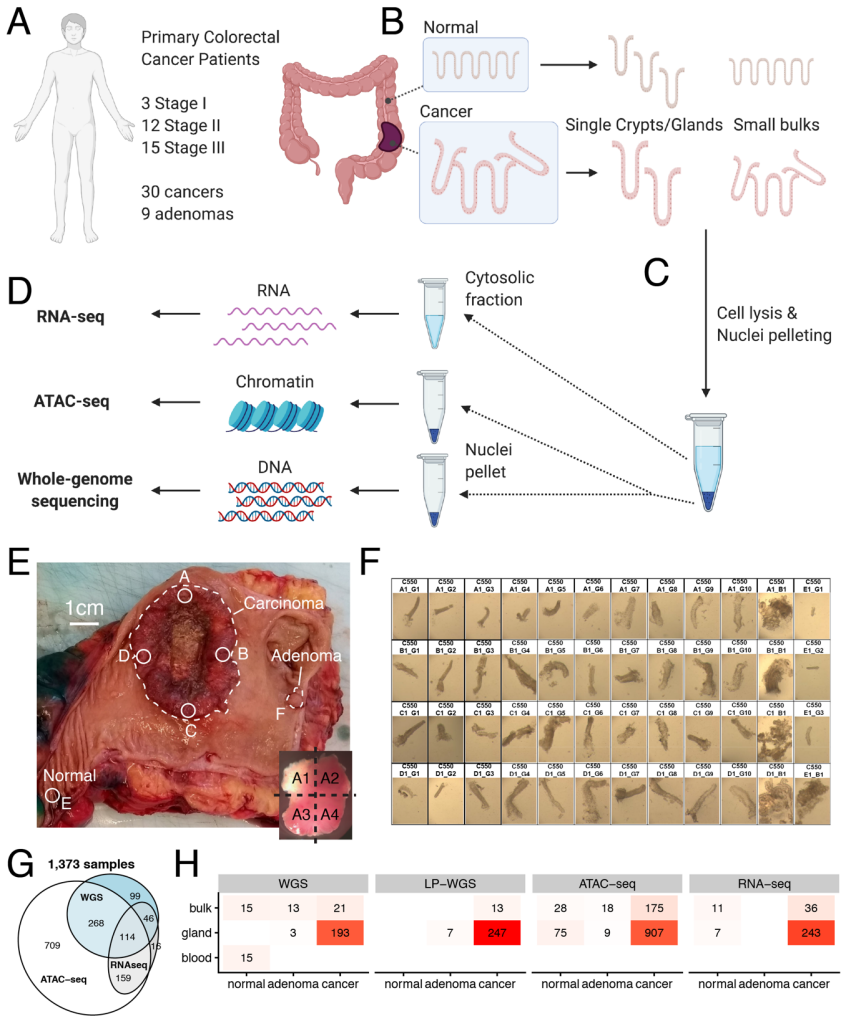

Figure 1. Spatial single gland multi-omics. ***(A)*** *Fresh colectomy specimens from 30 stage I-III colorectal cancer patients were used to collect tissue from 30 cancers and 9 adenomas.* ***(B)*** *Single glands and small bulks ('minibulks') were isolated from normal and neoplastic samples.* ***(C)*** *From each sample we performed cell lysis followed by nuclei pelleting.* ***(D)*** *Cytosolic fractions were used for RNA-seq whereas nuclei were used for whole-genome sequencing and ATAC-seq.* ***(E)*** *We identified separate regions of the cancer: A,B,C,D, a distant normal sample: E, and adenomas if present: F,G,H. Each sample was split into 4 fragments (inlet square).* ***(F)*** *From each fragment we collected individual glands (marked as _G) as well as minibulks, agglomerates of a few dozen crypts (marked as _B).* ***(G)*** *We performed multi-omics using whole-genome sequencing, ATAC-seq and RNA-seq on the same sample, achieving a good level of overlap between assays.* ***(H)*** *For each assay we had representative samples from normal, adenoma and cancer.*

Genetic mutations affecting the epigenome

We first assessed the landscape of genetic alterations in our cohort. Six cases in the cohort were characterised by microsatellite instability (MSI, Methods Section 3.8), as reported in Figure 2A, leading to significantly higher SNV and InDel burdens (Figure 2B). Copy number alterations recapitulated previous datasets[3,5], with microsatellite stable (MSS) cases displaying high aneuploidy and largely diploid MSI cases (Figure S4). As previously described[3], adenoma samples showed a lower degree of aneuploidy than MSS carcinomas, except for two outliers (Figure S5). Recurrent CN loss of canonical tumour suppressor genes, such as *APC*, *PTEN*, *TP53* and *SMAD4,* was confirmed. Focal amplifications were found in *FGFR1* (2 cases) and *MYC* (1 case). Recurrent cancer driver events in colorectal cancers were recapitulated in this dataset, with stereotypical mutations in *APC*, *KRAS* and *TP53* (Figure 2C and S6). Except for a single case (C539), mutations in these three genes were invariably clonal. The mutational profiles of the adenomas were consistent with earlier studies, specifically compared to Lin *et al.* 2018[30], no differences were observed for APC (4/8 vs 73/135, p-value=1, Fisher's Exact Test) or KRAS (2/8 vs 13/135, p-value=0.20, Fisher's Exact Test). We observed a slightly larger incidence of TP53 mutations (2/8 vs 4/135, p-value=0.037, Fisher's Exact Test). Compared to Cross *et al.* 2018[3], no major differences were detected (TP53: p-value=1, KRAS: p-value=0.33, APC: p-value=0.029, PIK3CA: p-value=1, Fisher's Exact Test).

To investigate the influence of genetic mutations on the epigenome, we specifically examined somatic mutations in chromatin modifier genes, namely the lysine demethylases (*KDM*), lysine acetyltransferases (*KAT*), lysine methyltransferases (*KMT*) and SWI/SNF (*ARID1A*) families (see Figure 2D for MSS cases and Figure S7 for all). Evolutionary selection on chromatin modifier genes was assessed by dN/dS[31,32] (Methods Section 3.10). Clonal (occurring in all cancer cells) truncating mutations in chromatin modifier genes in MSS cases showed clear signs of positive selection, with dN/dS significantly >1 (Figure 2E, arrow). Subclonal chromatin modifier mutations were present but positive selection was not detected, with dN/dS≈1 (Figure 2E). No evidence of positive selection for chromatin modifier gene mutations was detected in MSI cancers, although their high mutational burden may limit the power of detection. Hence, clonal truncating mutations in chromatin modifiers were found in 6/24 MSS cases (25%) and all MSI cases, with few recurrently mutated genes, suggesting a convergent pattern of selection for inactivation of chromatin modifiers in CRC.

Commented [AS3]: Comment 1.4a

Commented [AS4]: Comment 1.12

Commented [AS5]: Comment 1.1

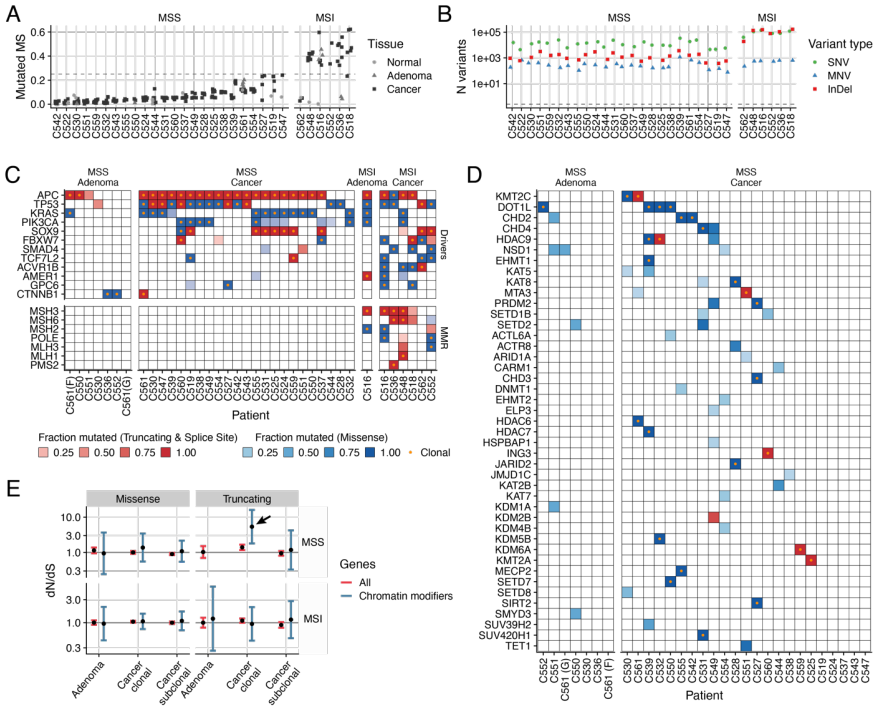

Figure 2. DNA alterations in canonical cancer drivers and chromatin modifier genes. ***(A)*** *Microsatellite instability per case.* ***(B)*** *Mutational burden by type of mutation (InDel: small deletion or insertion, MNV: multiple nucleotide variant, SNV: single nucleotide variant).* ***(C)*** *Recurrently mutated colorectal cancer driver genes, with orange dot indicating whether the mutation is clonal.* ***(D)*** *Truncating mutations and Indels in chromatin modifier genes in MSS cases.* ***(E)*** *dN/dS analysis of clonal and subclonal chromatin modifier mutations in MSS and MSI cancers and adenomas. Bars are 95% CI.*

Focal chromatin alterations are recurrent, largely clonal and hit known driver genes

Recurrent genetic events in cancer driver genes clearly demonstrate the role of somatic alterations in cancer, but how common epigenetic changes of chromatin accessibility in CRC are? We examined the landscape of somatic chromatin accessibility alterations (SCAAs) in our cohort. We identified peaks in the ATAC-seq data for each region of a cancer using MACS2[33] and compared each peak size in the tumour versus normals (see Figure S8), normalising for the effect of copy copy number alterations (Figure S9-11), to identify significant SCAAs (Figure 3A, Methods Section 3.11). We found highly recurrent SCAAs in both promoters (Figure 3B) and putative enhancers (Figure 3C) of several genes of interest, including many previously associated to cancer. We note that these levels of recurrence are as high if not higher than many genetic driver mutations (see Figure 2C).

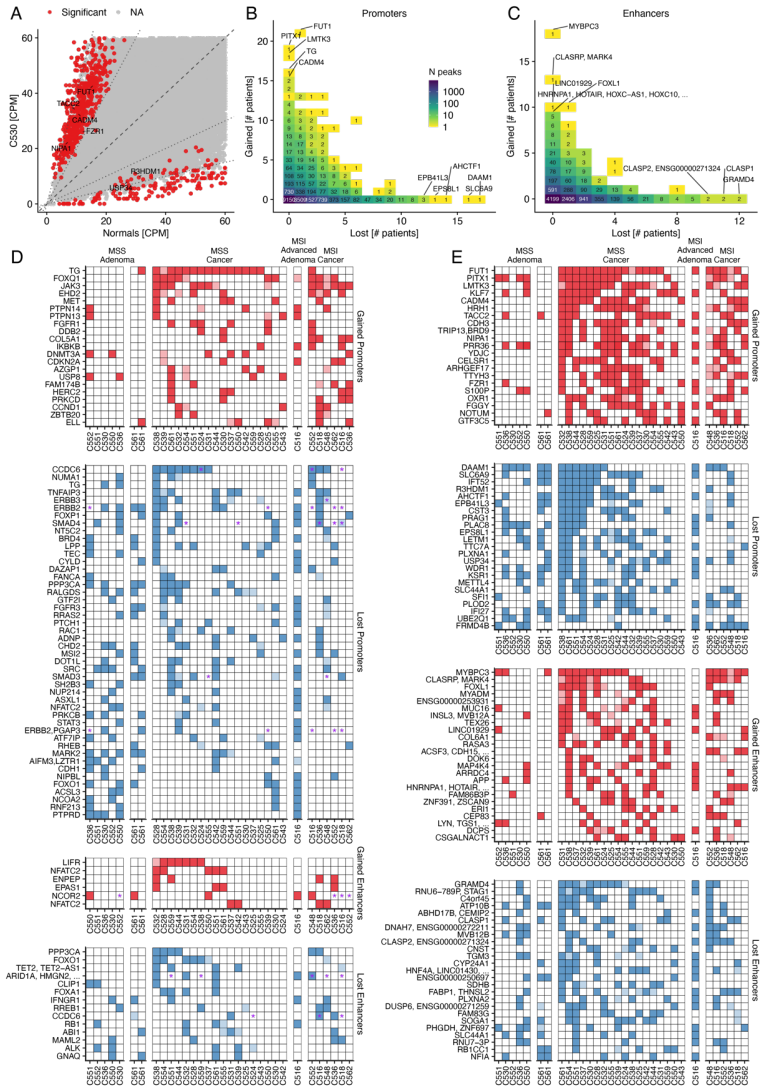
A
Significant
NA
C530 [CPM]
Normals [CPM]
B
Promoters
Gained [# patients]
Lost [# patients]
N peaks
C
Enhancers
MYBPC3
CLASRP, MARK4
HNRNPA1, HOTAIR, HOXC-AS1, HOXC10, ...
CLASP2, ENSG00000271324
GRAMD4
D
MSS Adenoma
MSS Cancer
MSI Advanced Adenoma
MSI Cancer
Gained Promoters
Lost Promoters
Gained Enhancers
Lost Enhancers
E
F
JAK3 promoter gain
chr19:17642053–17642553
G
CCDC6 promoter loss
chr10:59605006–59805506
H
FOXL1 (GH16J086618) enhancer gain
chr16:86619226–86619726
I
NXPH1 (GH07J008389) enhancer gain
chr7:8389807–8390307
J
# of rec. SCAAs (gain)
Advanced Adenoma
0.00079
per case
Adenoma
Cancer (pure)
K
# of rec. SCAAs (loss)
0.5
L
chr19:48752815–48753315
FUT1 (Promoter), p = 0.00078753 (open)
RNA CPM
Tumour closed
Tumour open
M
chr2:88144817–88145317
FABP1 (Enhancer), p = 0.00016266 (closed)

Figure 3. Somatic chromatin accessibility alterations (SCAAs) in cancers and adenomas. ***(A)*** *Example of SCAAs detected in cancer C530 versus normal. Significantly altered peaks in red.* ***(B)*** *Recurrence of lost and gained chromatin accessibility in promoters* ***(C)*** *and enhancers.* ***(D)*** *SCAAs hitting known cancer driver genes occurring in ≥3 patients. Stars indicate DNA mutation in reported colorectal cancer driver gene.* ***(E)*** *Summary of the 20 most recurrent SCAAs in promoter and putative enhancers of genes not previously associated with cancer through DNA mutation. Subclonal changes are marked in shaded squares.* ***(F)*** *Clonal somatic peak gained at the JAK3 promoter in cancer C551.* ***(G)*** *Recurrent promoter loss of accessibility of colorectal cancer driver CCDC6, example from C524.* ***(H)*** *FOXL1 enhancer gain of accessibility was found in regions B and C of C524 but not in other regions.* ***(I)*** *Example of somatic peak in NXPH1 enhancer gain found in the cancer but not in the concomitant adenomas of C561. All heterogeneous peaks were identified accounting for purity differences.* ***(J)*** *SCAA burden of adenomas versus carcinomas for gain of accessibility versus* ***(K)*** *loss of accessibility.* ***(L)*** *For a proportion of promoters and* ***(M)*** *enhancers, we were able to confirm changes in gene expression.*

Recurrent SCAAs were identified in known cancer driver genes previously identified by genetic studies (Figure 3D, list in Table S4, shown are events occurring in ≥4 patients). Many of these genes were devoid of genetic mutations in our cohort (see purple stars in Figure 3D), confirming that SCAAs are an alternative modality for driver gene (in)activation. We also found recurrent SCAAs in genes that were not previously associated with tumourigenesis by means of genetic mutation (Figure 3E, shown are the 20 most recurrent events per group, Figures S12-15 for all). We then leveraged our spatial multi-region profiling strategy to assess intra-tumour SCAA heterogeneity. The signal from ATAC peaks is notoriously difficult to compare between samples because it is confounded by variability in purity and transcription start site enrichment (TSSe). We used our matched WGS to identify clonal (truncal) DNA mutations present in all cancer samples and assessed the frequency of these variants in the reads from ATAC-seq to obtain an estimate of sample purity (Methods Section 3.11.2). Samples from each region were treated as pseudo-"biological replicates", and compared the signal between different cancer regions and the corresponding normal while accounting for purity (Methods Section 3.11.5). 24/30 cancers and 9/9 adenomas had sufficient samples with enough purity for the analysis. We focused on the 20 most recurrently altered loci per category (promoter/enhancer, gained/lost), as well as those associated to CRC driver genes from the IntoGen list[34] found in ≥4 patients (Table S4). We found that for the largest majority of these events (695/730, 95.2%) we had no evidence that they could be subclonal, indicating that the majority of SCAAs are clonal epigenetic changes in the malignancy (Figure 3D,E, see shading). Clonality of all SCAAs is reported in Figure S16-S19.

Amongst the recurrently altered and almost invariably clonal epigenetic changes, we found *JAK3* promoter gain of accessibility in 11/24 cancers (Figure 3F), we well as loss of chromatin accessibility in CRC tumour suppressor gene *CCDC6* in both promoter (11/24 cancers) and enhancer regions (3/24 cancers), see example case C524 in Figure 3G. Notably, mutations *CCDC6* are infrequent in CRC (3/30 cases in our cohort, annotated as purple star in Figure 3D). Furthermore, *ARID1A* enhancer loss was observed in 4 cancers and 1 adenoma (Figure S19), with only 2 of those cases also bearing a mutation in this gene. Alterations in multiple other putative CRC drivers were also found, such as *SMAD3* and *SMAD4* promoter loss, *NCOR2* enhancer gain, and *FBXW7* enhancer loss. *NFATC2* and *LIFR* cancer driver genes that were not reported in colorectal cancer were found epigenetically altered in our cohort, and in the absence of DNA mutations. Of interest, we

Commented [AS6]: Comment 1.16

found typically-clonal promoter SCAAs in *FOXQ1* in 11/24 patients, a known oncogene reported to be involved in colorectal cancer tumourigenicity[35], angiogenesis and macrophage recruitment during progression[36]. Although most recurrent SCAAs were clonal in the cancer, a proportion of SCAAs were found to be subclonal and confined to one or more regions. This was exemplified by *FOXL1* enhancer gain (10/24 patients – 42%) in Figure 3H occurring only in regions B and C of cancer C524.

We note that ATAC peaks called in our dataset showed strong overlap with peaks from the TCGA dataset (single-sample and lacking normals[13]) and the ENCODE normal colon tissue dataset[37], both reanalysed with our pipeline (Figure S20). Due to unmatched normal controls however, in these orthogonal single bulk sample datasets it is not possible to distinguish chromatin changes occurred in the cancer versus those already present in the normal colon (e.g., to determine the somatically-changed status of the peak), and indeed most of the signal of chromatin accessibility comes from the tissue of origin of the sample[13].

Somatic changes in chromatin accessibility distinguish adenomas from cancers

After observing strong evidence for that chromatin accessibility changes contribute to tumourigenesis, we then sought to define the evolutionary trajectories of SCAAs and determine their role in adenoma-carcinoma transition whilst considering the possibility that these changes may be a product of normal tissue aging. We examined the stage of tumour development when SCAAs occurred. Out of the 834 SCAAs found in ≥6 patients in cancers with available concomitant adenomas, only 141 (16.9%) were also detected in the matched adenoma, suggesting that most SCAAs likely occurred at the onset of malignant transformation, hence after neoplastic growth initiation but before subclonal diversification (as they were also largely clonal). Such events are exemplified by the gain of accessibility of *NXPH1* enhancer (4/24 patients – 17%), which was present in each region of the cancer but not in any of the concomitant three adenomas (Figure 3I, all events in Figure S21). Indeed the lower SCAA burden of adenomas compared to cancers was not dependent on purity or read depth (Figure S22A,B). In a power analysis where we explicitly normalised for coverage (Figure S22C), we found a significantly lower burden of recurrent gain of accessibility SCAAs (>10 patients) between adenomas and carcinomas (Figure 3J). No difference was found in the burden of loss of accessibility between adenomas and carcinomas (Figure 3K). We note that the only advanced adenoma in our cohort that was found co-locating with the cancer (C516) indeed showed the SCAA gain burden of a carcinoma (Figure 3J). It was previously noted that there were limited differences between adenomas and carcinomas in colorectal cancer at the level of point mutations in driver genes, and instead major differences at the level of chromosomal instability[3]. Here we found also differences in epigenetic rewiring between adenomas and cancers. Moreover, the higher burden of SCAA gains in cancers supports the idea that carcinogenesis involves an increased genome-wide chromatin accessibility.

To elucidate whether the observe SCAAs resulted from patient-specific alterations, we applied the same approach that we used for the analysis of cancer samples to the normal glands of each patient. Here we found very few SCAAs in individual normal glands, demonstrating that, supporting that the SCAAs we observed in the tumours were indeed somatic alterations. However, in a small subset of SCAAs alterations were also identifiable in multiple glands of the same patient (see Figure S23A). Still, very few of these changes were recurrent across cases (see Figure S23B) and no correlation of the frequency of chromatin alterations in normal glands and the frequency of the reported SCAAs was

Commented [AS7]: Comment 1.3 and 3.2

Commented [AS8]: Comment 3.5

observed (see Figure S23C). Plausibly germline genetic variation could cause chromatin accessibility alterations in normal tissue.

Impact of SCAAs on gene expression

We next assessed the impact of SCAAs on gene expression using matched RNA-seq (e.g., Figure 3L,M). Over 11.9% of promoters (45/379) and 14.1% of enhancers (10/71) with recurrent SCAAs (>5 patients) showed signs of altering the expression of associated genes (Figure S24, FDR<0.01, Table S4, Methods Section 3.11.4 and 3.14). We note that chromatin accessibility measures the *potential* for transcription, indicating priming for future expression or a remnant 'scar' of important past transcription. Therefore, more chromatin changes than those that correlate with expression in our analysis may actually be important for tumour evolution. Moreover, the power to detect expression changes was limited by the recurrence of a given SCAA in the cohort, incomplete matched RNA data, and the lack of information about other factors influencing transcription such as methylation, post-translational modifications, or just *trans*-regulation by other genes. To further probe the impact of somatic mutations on SCAAs, we analyzed SNVs that we found were associated with changes in *cis* gene expression in our associated paper[38] and found that some of these SNVs co-occurred with a change in chromatin accessibility at the locus (Figure S25).

Commented [AS9]: Comment 3.1

Commented [AS10]: Comment 1.2b and 3.2

Transcription factor accessibility analysis reveals global epigenetic reprogramming

Beyond focal changes in chromatin accessibility in promoters and enhancers, we investigated whether chromatin architecture could have genome-wide influence on transcriptional control. To examine this, we analysed the genome-wide accessibility of transcription factor (TF) binding sites for 870 transcription factors[37] using publicly available TF motif and ChIP-seq data (Methods Section 3.12). We piled-up the ATAC reads for all binding sites of a given TF across the genome and plotted read count versus the distance from the centre of the TF motif and the length of each read, producing a characteristic signature of TF accessibility for a given sample, which also encodes the footprint of the TF complex itself in the cancer (Figure 4A and S26) and normal (Figure 4B). The normalised tumour-difference showed the somatic change in accessibility Figure 4C. As many transcription factors bind to similar loci, we considered only largely non-overlapping TF annotations to ensure a single locus could not drive the signal of multiple TFs (Figures S27 & S28). These analyses showed pervasive genome-wide rewiring of TF chromatin accessibility in CRCs (Figure 4D, see Methods section 3.12 for details).

Unsupervised clustering of somatic TF binding signatures produced three major clusters. The first major cluster (green cluster, Figure 4D) was associated with downregulation of interferon signalling through loss of chromatin accessibility in loci putatively bound by transcription factors from the IRF (interferon-regulatory factor) family, suggesting suppression of immune signalling. Reactome and GO analysis indicated that the signal was significantly enriched for downregulation of interferon-γ (FDR=4.2e-6) and interferon α/β (FDR=3e-8), as well as downregulation of cell differentiation (FDR=5e-5) (Figure 4E). The signal was even stronger in MSI cancers, which are heavily infiltrated by immune cells (p=0.012, Fisher's Exact Test).

Commented [AS11]: Comment 1.2a

The second major cluster (blue cluster, Figure 4D) contained two distinct subgroups of patients with differential chromatin accessibility for *CTCF*. CCCTC-Binding Factor (CTCF) is a key player in chromatin insulation, determining looping and TAD (Topological Associating Domain) formation. Most cases were characterised by loss of CTCF binding site accessibility, particularly in MSI cancers. A smaller group showed increased *CTCF*

accessibility. *CTCF* chromatin accessibility alterations were previously noted in single bulk cancer sample[39], *CTCF* somatic mutations are frequent in CRC[40], and indeed a mouse model of chronic CTCF hemizygosity led to higher cancer incidence and dysregulation of oncogenic pathways[41].

The third major cluster (red cluster, Figure 4D) showed increased chromatin accessibility for TFs involved in stem cell differentiation and pluripotency (GO: *'positive regulation of stem cell differentiation'* – FDR=2.5e-4, and *'mesenchymal stem cell differentiation'* – FDR=9e-4; KEGG: *signalling pathways regulating pluripotency of stem cells* – FDR=0.047), as well as TFs involved in development, such as the *HOX*, *FOX* and *SOX* families (UniProt: *'homeobox'* – FDR=2.7e-40, *'developmental protein'* – FDR=1.7e-21). The chromatin accessibility of this cluster of TFs was higher in cancer in most cases, suggesting possible reactivation of developmental genes in colorectal cancer tumorigenesis (Figure 4F). The expression of the TFs involved in this cluster is reported in Figure S29.

Interestingly, matched RNA-seq data showed that gene expression of HLA genes was significantly reduced in both MSS and MSI cancers with respect to normals (Figure 4G) consistent with the downregulation of interferon signalling as highlighted by the signal in the "green" cluster.

Commented [AS12]: Comment 2.2

We also noted a small cluster characterised by increased accessibility of *SNAI1* and *SNAI2* transcription factor binding sites, two genes involved in Epithelial to Mesenchymal Transition - EMT[42]. This cluster was significantly enriched with cases showing truncating mutations in chromatin modifier genes (p=0.047, Fisher's Exact Test), consistently with previously reported regulation of EMT by chromatin modulators[43]. Although more patients are needed, we cannot exclude that there could be additional subgroups of patients with distinct TF accessibility patterns beyond the *CTCF* subgroup (blue cluster).

Commented [AS13]: Comment 1.14

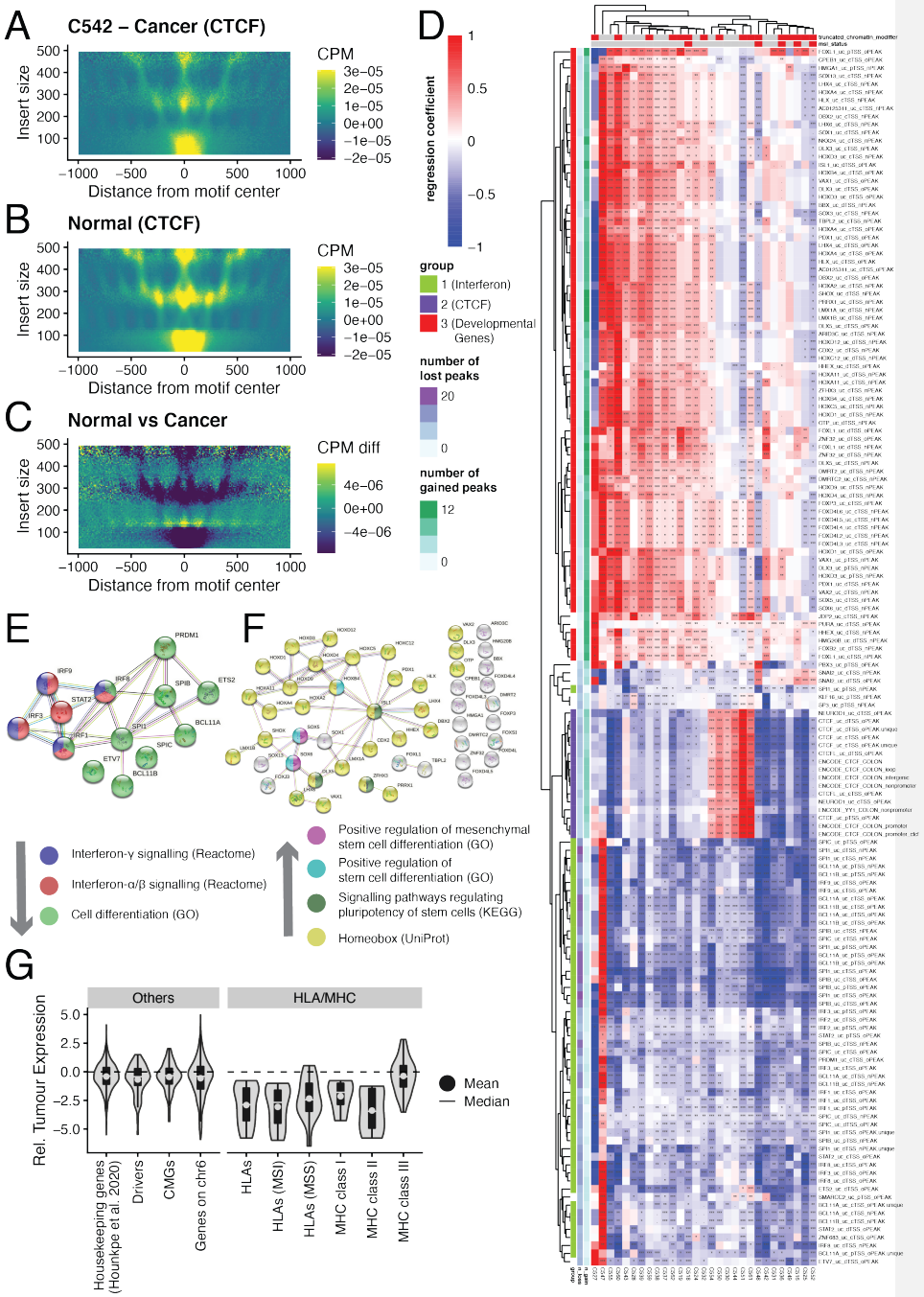
A
C542 – Cancer (CTCF)
Insert size
Distance from motif center
CPM
B
Normal (CTCF)
C
Normal vs Cancer
CPM diff
D
regression coefficient
group
1 (Interferon)
2 (CTCF)
3 (Developmental Genes)
number of lost peaks
number of gained peaks
E
F
Interferon-γ signalling (Reactome)
Interferon-α/β signalling (Reactome)
Cell differentiation (GO)
Positive regulation of mesenchymal stem cell differentiation (GO)
Positive regulation of stem cell differentiation (GO)
Signalling pathways regulating pluripotency of stem cells (KEGG)
Homeobox (UniProt)
G
Others
HLA/MHC
Rel. Tumour Expression
Housekeeping genes (Hounkpe et al. 2020)
Drivers
CMGs
Genes on chr6
HLAs
HLAs (MSI)
HLAs (MSS)
MHC class I
MHC class II
MHC class III
Mean
Median

Figure 4. Transcription factor binding site accessibility is rewired in tumours. (A) *TF binding site accessibility (in this example CTCF) is computed by summing the signal of ATAC-seq reads centred at the binding site, plotted against read length.* ***(B)*** *The same is done for the normal controls.* ***(C)*** *Signal from the normal is subtracted from the signal from the cancer to assess differential accessibility. TF accessibility for CTCF is reduced in this example as demonstrated by fewer ATAC cuts at the binding site in the cancer.* ***(D)*** *The differential signal is then regressed against TSSe and purity to identify TF binding accessibility altered in tumours. Results here for the three major clusters of differentially accessible TF loci (heatmap colour is regression coefficient, star indicates significance). Major cluster identity is denoted by left-most column.* ***(E)*** *String-db analysis of the green TF cluster highlights downregulation of interferon signalling.* ***(F)*** *String-db analysis of the red cluster indicates upregulation of stem cell differentiation and activity of developmental genes such as the homeobox family.* ***(G)*** *Relative tumour expression of HLA genes versus other gene groups.*

Binding sites of developmental TFs with increased accessibility are demethylated

We further attempted to corroborate the increased accessibility to TF involved in development. Changes in chromatin accessibility can be accompanied by changes in DNA methylation, with heterochromatin regions often being methylated and vice-versa for open chromatin regions. This is particularly the case for regions that are permanently silenced after development[44]. We tested whether SCAAs identified at TF binding sites (Figure 4D) were reflected in the methylation of the same loci. We performed methylation profiling on a subset of 8 samples using Illumina EPIC 850k methylation arrays (1x sample from C516, 2x samples from C518, 2x samples from C560 and 3x samples from C561 – see Section 3.13 for detail). First, we report that C518 is a likely a CpG island methylator phenotype case according to established markers[45] (Figure S30). Comparing the methylation of TF binding annotations in cluster 3 (Figure 4F), we found that methylation in these regions was significantly lower than in normal tissue, supporting the finding that these sites were accessible (Figure S31A). This was particularly clear for TF binding sites of DLX5, HOXA4, HOXB4, ISL1, SOX5 and SOX6 (Figure S31B), suggesting stable reactivation of regulatory regions involved in developmental genes. We note that this was not due to a general pattern of global hypomethylation, as methylation in genes which are usually normally highly methylated in normal were also high in cancer (Figure S32).

Commented [AS14]: Comment 3.4

Commented [AS15]: Comment 3.7

Chromatin changes are stable and heritable, and can be a substrate for Darwinian clonal selection

Epigenetic alterations, and in particular chromatin modifications, are responsible for cell identity in all tissues, but it remains unclear whether epigenetic changes in cancer are stable during tumour evolution. Seminal studies have begun unravelling epigenetic heritability in blood cancers[46,47], suggesting that stable SCAAs could provide a heritable substrate for Darwinian selection to operate. For most detected SCAAs, if the peak was differentially accessible in one region of the tumour, it was also differentially accessible in other distant regions. Because we sampled opposing tumour sides, each sampled region likely only has a most common recent ancestor many thousands of cell divisions ago (Figure 3E&F). Hence, we argue that most SCAAs we detect are likely *clonal* or have high 'clonality', i.e. they are shared by large proportions of cancer cells. This can occur either through convergence of different lineages to the same SCAAs, or through evolution by common descent. Given the number of putatively clonal SCAAs, as well as the distance and probably difference in microenvironment of the distinct regions of each cancer, we argue that the most parsimonious explanation is, as for species evolution, evolution by

Commented [AS16]: Comment 1.17

Commented [AS17]: Comment 1.6

common descent, rather than convergence of many different lineages to the same overall epigenetic pattern.
To further test the heritability of epigenetic alterations we specifically compared SCAAs within versus between tumour regions (Figure S33A). In a majority of patients (23/29), ANOVA controlling for TSSe and total read count, showed that samples from the same region were significantly less divergent in terms of SCAAs than samples from different regions (Figure S33B). Moreover, a direct correlation between genetic distance and epigenetic distance was found in 8/29 cases (the power of this analysis is limited by small sample numbers), after controlling for purity (example in Figure S33C). This was not the case for all patients, either because of lack of a correlation or not enough data (example in Figure S33D). Thus, chromatin profiles were heritable and followed, at least in part, genetic divergence (Figure S33B; see coefficients of the ANOVA analysis per region in Figure S34), thus providing additional evidence that common descent is the reason of SCAAs common to multiple samples of the same tumour, not convergence. Genome wide TF SCAAs (Figure 4) showed similar evidence of heritability (Figure S35), suggesting that such rewiring of the chromatin existed in a common ancestor of all the samples and was inherited during tumour growth. There were however some interesting exceptions where different regions showed distinct SCAA profiles. For example, C548 showed homogeneous loss of accessibility to *CTCF* binding sites at loop loci. In C543 both promoter and loop binding sites of *CTCF* were altered and in a heterogeneous manner, with region displaying differential organisation of the chromatin (Figure S35).

Commented [AS18]: Comment 1.6 and 3.2

Mutational signatures affecting the epigenome

There is a 'growing appreciation of the multidimensional nature of mutation signatures beyond the 96 channel representation and across different regions of the genome, especially in relation to replicating timing and 3D genome organisation[48]. However, the relation between mutational signatures and epigenetic features remains poorly studied due to lack of matched data. Here we examined the feedback between epigenome and transcription status and mutational processes[49,50] through tumour evolution. We performed *de novo* signature discovery using a methodology robust to over-fitting[51], detecting six mutational signatures across our cohort (Figure 36A):

- SparseSignature1, corresponding to COSMIC signature 1 of C>T deamination at methylated CpG sites
- SparseSignature2, corresponding to COSMIC signatures 2+13 caused by APOBEC enzymes
- SparseSignature3, corresponding to COSMIC clock-like signature 5
- SparseSignature4, corresponding to COSMIC signature 17a+b of unknown aetiology
- SparseSignature5, corresponding to COSMIC signature 9+41 also of unknown aetiology
- SparseSignature6, corresponding to COSMIC signature 44 caused by mismatch repair deficiency.

Genome wide signature activity divided the cohort into five distinct clusters of patients (Figure S36B,C). The two major clusters consisted of MSS (cluster 1) and MSI cases (cluster 2). Cluster 3 contained only patient C549, which was strongly enriched with the APOBEC signature. Cluster 4 with patients C561 and C539 had high activity of SparseSignatures 4 and 5 of unknown aetiology. Cluster 5 with patients C518 and C548 had higher SparseSignature 3 (clock-like signature in CRC). We assessed changes in mutational process activity over time by comparing inferred activity between clonal and

subclonal mutations (Figure 5A). SPS1 (deamination) was dominant in MSS cases throughout tumour evolution, and in MSI cancers SPS6 (mismatch repair) was also dominant throughout. Interestingly, SPS2 (APOBEC), SPS4 and SPS5 (unknown) were enriched subclonally level in cases where they were active, demonstrating activity late in tumour evolution.

Mutations in chromatin modifier genes, or in transcription factor binding sites, can determine the characteristics of the epigenome. Conversely, chromatin architecture determines how the cancer genome accumulates mutations due to its effect on different mutational processes and activity of DNA repair genes[53,54]. To examine the latter, we compared mutational signature burdens between epigenetic regulatory regions identified with the ATAC-seq data: active/inactive promoter (e.g., chromatin open/closed), active/inactive enhancer, intergenic and coding, as well as using matched RNA-seq data to differentiate between typically expressed and not expressed genes.

SparseSignature1 (cytosine deamination) was 2-4 fold higher in closed chromatin regions of the genome (inactive promoters and enhancers) for both clonal and subclonal mutations, consistent with the need for methyl-cytosine (enriched in inactivated regulatory regions) to be present in order for it to become deaminated and produce the associated mutational signature (Figure 5C). Analogous differences were observed in the coding regions of the genome between genes expressed versus not expressed genes in the normal: specifically, genes that were “switched on” in tumour after being off in normal carried an intermediate load of C>T deamination mutations that were likely accumulated in the normal tissue before carcinogenesis when the locus had inaccessible chromatin, before the mutation rate was reduced when the chromatin opened and gene expression was induced (Figure 5C). Similar dynamics were observed for SparseSignature4 (Figure 5D) and SparseSignature5 (Figure 5E)[55]. The activity of the mismatch repair signature in MSI cases was more uniformly distributed across the genome (Figure S38).

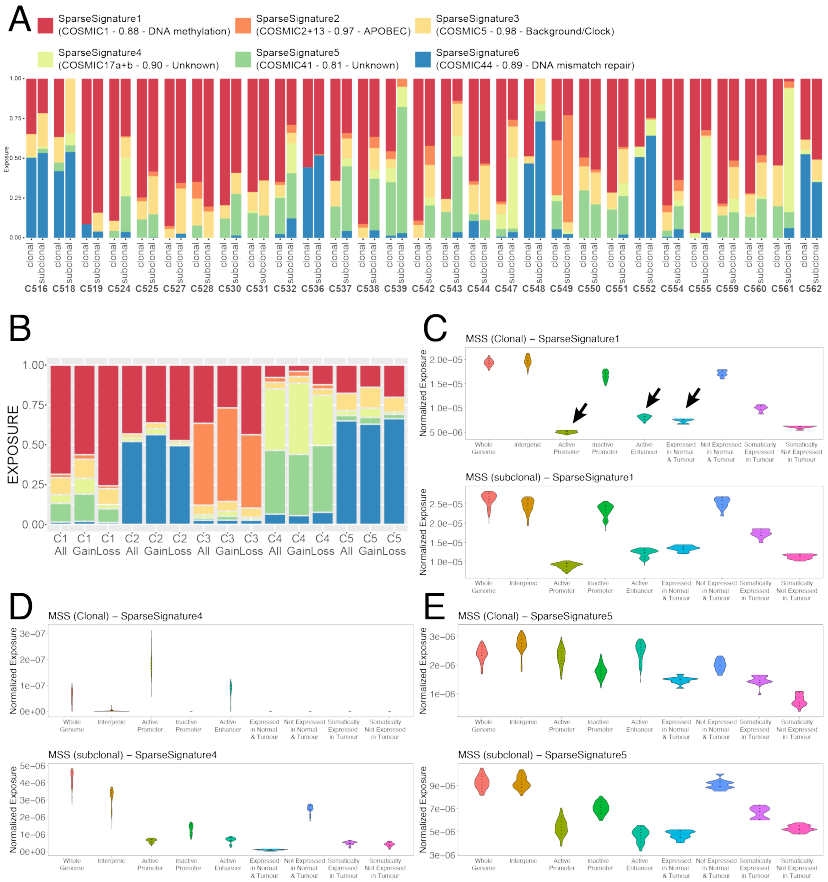

Figure 5. DNA mutational signatures and the epigenome. (A) *Clonal and subclonal mutational signature composition for each patient.* ***(B)*** *Proportion of each signature for every cluster responsible for generating loss or gain of CTCF binding affinity in our cohort.* ***(C)*** *The epigenome influences accumulation of deamination signature 1 in distinct regions, both for clonal and subclonal mutations.* ***(D)*** *Signature SparseSignature4, mostly present subclonally, is also influenced by the epigenome status.* ***(E)*** *Signature SparseSignature5, particularly at the subclonal level, is again depleted in active regions as SparseSignature1.*

We hypothesised that different mutational processes may also differentially alter TF binding site affinity, as an example mechanism of how mutational processes can directly influence the cancer epigenome. It has been previously documented that point mutations can disrupt *CTCF* binding sites[40]. We selected *CTCF* sites with somatic mutations that were predicted by deltaSVM[52] to cause significant loss or gain of binding and assessed the relative contribution of each mutational signature to these *CTCF* mutations across the five mutational signature clusters. In MSS cancers (cluster 1), mutations predicted to cause loss of binding had a signature that was consistent with the background mutational signature acting on the genome (cosine similarity = 0.977; Figure S39A), and the same

was true for gains (cosine similarity = 0.919; Figure S39B). In MSI cancers (cluster 2), SparseSignature6 (mismatch repair, Figure S39C) was consistent with causing gain of *CTCF* binding affinity (cosine similarity = 0.925). In C549, the only case with high levels of SparseSignature4 (COSMIC signature 17, Figure S39D), such signature was also a source of mutations causing gain of affinity (cosine similarity = 0.977). These results suggest that CpG deamination causes the largest proportion of mutations altering *CTCF* binding in MSS cancers, with a higher tendency of generating loss of binding (Figure 5B). In MSI cases, the mismatch repair signature is also a dominant factor in causing altered binding of *CTCF*, with preference for generating increased affinity (Figure 5B). When considering the abundance of any given mutational signature in the genome, we found that 4% and 8% of signature 1 mutations cause respectively gain and loss of CTCF binding, whereas 5% and 8% of signature 6 mutations cause respectively gain and loss of CTCF binding (see all in Figure S40).

Commented [AS19]: Comment 2.3

Discussion

The contribution of epigenetic events to cancer evolution is recognised as highly significant[10,56], but has remained understudied[8]. Recently, a pan-cancer analysis revealed the chromatin accessibility profile of multiple cancer types, but the lack of appropriate matched normal control precluded proper identification of cancer-specific events, as opposed to tissue specific and 'cell of origin' chromatin profiles which remained the dominant signal in the data[13]. Studies with normal tissue references have identified complex patterns of SCAAs within CRCs[11,12], but have not been able to assess the evolutionary dynamics that led to these chromatin changes. Here, we show that genetic and epigenetic modification of cancer-associated genes occurs independently but recurrently in CRCs, and that epigenome alterations likely control important tumour cell phenotypes, including immune escape. Further, we find that chromatin alterations are stable and heritable, providing a substrate for Darwinian selection to act, and interrelatedly, chromatin alterations influence the accumulation of somatic genetic alterations that can also drive evolution[57,58]. Currently, genomics detects driver alterations or mutational processes that inform on drug sensitivity but is blind to potentially clinically-actionable biology governed by the epigenome. The observation that epigenetic changes occur in regulatory regions of known cancer driver genes in the absence of somatic mutations argues for the importance of epigenomics for genomic medicine. Certainly, the interaction between somatic mutations and SCAAs remains challenging to unravel. Although multiple studies have investigated the effects of somatic mutations in chromatin modifier genes, for instance linking mutations with increase transcriptional heterogeneity[59], identifying the direct (*cis*) functional effects on the chromatin from DNA variants remains difficult. Our multiomic dataset provides some clear examples of a genome-epigenome relationship: we observed somatic mutations associated with changed *cis* gene expression where there was also a chromatin accessibility change. Follow-up work is required to explore the functional impact of epigenetic alterations in cancer driver genes and other loci.

We also observed that the epigenomes of adenomas and carcinomas are distinct. The lower prevalence of SCAAs in adenomas and, at the same time, the clonality of most SCAAs in carcinomas, suggest that many cancer SCAAs may occur at the onset of malignant transformation. This is important because, besides broad copy number alterations, mostly non-focal chromosomal arm gains or losses of unknown significance, there is little difference in driver alterations between benign adenomas and malignant carcinomas[3]. Moreover, there is no validated prognostic genetic alteration that predicts

Commented [AS20]: Comment 1.13

recurrence in colorectal cancer. The seminal paper of Johnstone et al.[12] showed that chromatin topology changed over time in ageing colon tissue, including in transformed tissues, and showed a link between altered chromatin patterns and patient outcomes. This is consistent with our finding of a decisive role for SCAAs in cancer biology. We acknowledge that our multi-omic analysis was based on the analysis of tumour glands, and it possible that the biology could differ in the rare CRCs that completely lack glands.

Commented [AS21]: Comment 1.5

Commented [AS22]: Comment 1.10

One of the most intriguing results was the evidence of reactivation of developmental genes during tumourigenesis. Those genes are usually silenced in somatic tissue, and the reactivation of these gene families and their involvement in tumourigenesis has been postulated before in the context of glioblastoma tumourigenesis[55] as enabler of growth and adaption. We identified a group of TFs with decreased accessibility that were related to interferon signaling as well as cell differentiation, suggesting the possible activation of early progenitor-like transcriptional programmes or de-differentiation. On the other hand, we also found a group of TFs that had increased accessibility and was highly enriched with homeobox genes, e.g., SOX5 and SOX6, that are directly involved in early development cell differentiation. We speculate that we may detect a combination of processes that aim at avoiding full differentiation reprogramming cell fate through the involvement of developmental genes. We hypothesize the process may lead to an 'early progenitor' phenotype that is proliferative (unlike a fully stem-like phenotype) but does not differentiate completely[60]. Further functional work is warranted.

Commented [AS23]: Comment 2.1

Our spatially resolved multi-omic analysis of primary colorectal cancers shows non-genetic determinants of cancer cell biology and clonal evolution.

Methods

1.1 Sample collection
Primary tumour tissue and matched blood samples were prospectively collected from patients undergoing curatively-intentioned surgery at University College London Hospital (UCLH). All patients gave informed consent for collection of their materials to the UCLH Cancer Biobank (REC approval 15/YH/0311). Four regions of each primary cancer were sampled by punch biopsy or scalpel dissection, at notionally 12, 3, 6 and 9 o'clock positions around the tumour periphery. Tissue was slow-frozen to -80C, using a Mr Frosty Freezing Container (Thermofisher) in 1ml of a buffered media (MEM supplemented with 5% FBS and 0.5% 5mM HEPES buffer, diluted with 10% DMSO) in a 1.8ml Nunc Cryotube (Sigma- Aldrich) immersed in isopropanol to preserve chromatin structure.

1.2 Gland isolation
Each biopsy was manually dissected into a series of smaller pieces of tissue (notionally 1mm2) with attempts made to record the relative spatial location of smaller piece within the larger biopsy. A tissue piece was selected for gland extraction and placed in a 50µl HBSS supplemented with Protease inhibitors: 1 tablet in 50ml dH2O as directed (C0mplete Protease Inhibitor Cocktail, Sigma- Aldrich) and RNASE inhibitor 1U/µl (Protector RNase Inhibitor, Sigma- Aldrich) and kept on wet ice until needed. A clean glass slide was placed into a 10cm Petri dish and 500ul of PBS supplemented with RNAse and protease inhibitors was pipetted on top of the slide. The petri dish was then transferred to stage of a dissecting microscope. Tissue pieces were manually dissociated under the microscope

using two 16G needles, where individual glands glands were pulled away from the tissue mass. An additional epithelial “minibulk” sample was collected for every specimen that comprising approximately a total of 10-20 crypts/glands. Each gland or bulk specimen was transferred into a 1.5ml Eppendorf tube containing to a total volume of 50µl cell lysis buffer (10mM Tris-HCl pH 7.4, 10mM NaCl, 3mM MgCl2, 0.1% IGEPAL CA-630 supplemented with Protease inhibitors: 1 tablet in 50ml dH2O as directed (C0mplete Protease Inhibitor Cocktail, Sigma- Aldrich) and RNASE inhibitor 1U/µl (Protector RNase Inhibitor , Sigma- Aldrich) and incubated on ice for 10-45min. Bulk samples were collected in a final volume of 100µl cell lysis buffer. We found longer or warmer incubations decreased RNA quality and yields and negatively affected chromatin structure. While selecting the 30 cases included in our study we rejected only a single additional case due to being unable to isolate any glands, confirming that retainment of glandular structures is pervasive in CRC.

Commented [AS24]: Comment 1.10

2.1 Chromatin, DNA and RNA separation

Each tube containing an individual gland or bulk was lightly vortexed, transferred to a pre-chilled centrifuge and spun at 500g for 10 min holding the temperature at 4C. This produced a cell nuclei pellet at the bottom of the tube, with the cytosolic fraction present within the supernatant. For RNA extraction: 45µl of the supernatant was transferred into a new tube containing 300ul of Trizol (taking care not to disturb the pellet). Trizol lysates could be stored at -20oC if not processed immediately or at -80oC for long term storage. For extraction of nuclear material: the nuclei pellet was resuspended in residual cell lysis buffer. 2.5µl of suspension (roughly half the remaining suspension) was transferred into another tube for subsequent DNA extraction, that could be frozen if required. The remaining suspension was immediately used for preparation of ATAC-seq libraries, as we found subsequent handling or storage compromised library quality.

2.2 Preparation of ATAC-seq libraries.

Tubes containing the nuclei suspension (roughly 2.5µl) were kept on wet ice. 2.5µl of 2X TD buffer and 0.25µl of Tn5 transposes (Illumina) was added to each tube (resulting in a final volume of approximately 5µl) before incubation at 37°C for 30 min. Purification was performed reaction using AMpure XP SPRI beads (Beckman Coulter), 10µl (2x sample volume) of room temperature beads were added to each tube and mixed by pipetting 10 times, then incubated at RT for 1 minute. The tube was placed on a magnetic plate and beads allowed to settle for 3 minutes. Once clear the supernatant was discarded. With the tube still on the magnet, 200 µL of 80% ethanol was added and incubated at RT for 30 seconds, ethanol supernatant was discarded, this step was repeated for a total of 2 ethanol washes. The tube was removed from the magnet and 12µl of 10mM Tris buffer added to each tube and mixed by pipetting 10 times, then incubated at RT for 1 minute, The tubes were placed on a magnetic plate and the beads allowed to settle for 3 minutes. Once clear 10µl of supernatant containing purified DNA was transferred to a fresh tube for immediate library preparation or stored at -20C for later use.

For library preparation: the transposed sample was supplemented with 1µl 10 µM of Nextera i7 PCR primer, 1µl of 10 µM Nextera i5 PCR primer (Illumina) and 12.5µl of NEBNext Q5 High-Fidelity 2x PCR Master Mix (New England Biolabs). PCR amplification was performed, with initital eleongation at 72C for 5 minutes,then initial denaturation at 98C for 30 seconds, and then for glands 14 cycles (10 cycles for bulks) of the following: 10 seconds of denaturation at 98C, annealing step at 63C for 30 seconds followed by 72C for 1 minute.

Following amplification, samples were purified with 2X SPRI beads and eluted in 20-30µl of 10mM Tris Buffer, pH 8. Samples were screened using the Agilent Tapestation 4200 and HSD1000 screentapes. Only those which showed fragment size distribution with peaks at multiples of ~147bp, indicating intact nucleosomal structure within the nuclei, were sent for sequencing.

2.3 Preparation of whole genome sequencing libraries
DNA fractions were extracted using the Zymo QuickDNA Microprep plus kit according to the manufacturer's instructions. Only samples with a total DNA yield higher than 10ng were taken forward for WGS library preparation. Libraries were prepared using the NEBnext Ultra II FS kit according to manufacturer's instructions. A short enzymatic fragmentation step of 5 minutes was performed, and 5 PCR cycles are used for library enrichment. After purification, libraries were quantified by Qubit and run on the Agilent Tapestation using HSD1000 screentapes. Samples with sufficient library DNA yield and characteristic fragment size distribution (~200-500bp) were further subjected to either low-pass (~1x coverage) or deep (~35x coverage) WGS.

2.4 RNA library preparation
The cytoplasmic fractions of each sample in the form of Trizol lysates were used for RNA extraction using the Directzol kit (Zymo R2052). Modifications to the manufacturer's protocol were introduced to increase the total RNA yields. Firstly, we passed the initial trizol/ethanol mix twice through the spin column. Secondly, we eluted the RNA using two 25ul volumes of water instead of just one 50ul elution. The optional DNase step was used.

Agilent Tapestation QC showed low RIN scores for most samples (<3) and so was not used to exclude samples for library preparation. Libraries were prepared using the Illumina TruSeq RNA Exome kit (compatible with low quality input material) according to the manufacturer's instructions.

2.5 Methylation arrays
DNA methylation array analyses were carried out on selected bulk samples with sufficient DNA yield. Genomic DNA was bisulphite converted using Zymo EZ DNA Methylation kit. A 50µl reaction containing 2.5-100ng of DNA was incubated in the dark using a modified conversion protocol; 95°C for 30 seconds then 50°C for 60 min, for 16 cycles then hold at 4°C. The full 8ul elute of converted DNA was repaired using the Infinium HD FFPE Restore Kit (Illumina). All 8ul of the bisulphite converted DNA for each sample was analysed on the Illumina Human MethylationEPIC BeadChip (Illumina). Processing was carried out by the University College London Genomics Core Facility according to standard protocol.

2.6 Sequencing
Sequence libraries were multiplexed and sequenced on an Illumina Novaseq, typically using S2 flow cells. Read length and depth was varied as required by library composition. Sequencing was performed by the Institute of Cancer Research Tumour Profiling Unit (TPU).

3.1 Whole-genome sequencing – alignment
Contaminating adapter sequences were removed using Skewer v0.2.2[61]. Adapter sequences were 'AGATCGGAAGAGC' and 'ACGCTCTTCCGATCT', with a maximum error rate of 0.1, minimum mean quality value of 10 and a minimum read length of 35 after

trimming using options "-l 35 -r 0.1 -Q 10 -n". The trimmed and filtered reads from each sequencing run and library where separately aligned to the GRCh38 reference assembly of the human genome[62] using the BWA-MEM algorithm v0.7.17[63] Following the GATK best practices and the associated set of tools v4.1.4.1[64–66] reads were sorted by coordinates (GATK SortSam), merged independent sequencing runs or libraries generated from the same tissue sample and marked duplicated reads using GATKs MarkDuplicates. The structure of the final bam files was verified using GATKs ValidateSamFile.

3.2 ATAC-seq – alignment
Adapter sequences were removed with Skewer v0.2.2[61] using the full-length adapter sequences below with the option "-m any.

Adapter sequences:
CTGTCTCTTATACACATCTCCGAGCCCACGAGACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
CTGTCTCTTATACACATCTGACGCTGCCGACGANNNNGTGTAGATCTCGGTGGTCGCCGTATCATT

The reads of each sequencing run and library were aligned to the GRCh38 reference genome using Bowtie2 v2.3.4.3[67] with the options "--very-sensitive -X 2000" set. After sorting the reads with samtools v1.9[68] reads mapping to non-canonical chromosomes and mitochondria (chrM) were removed (GATK PrintReads followed by RevertSam and SortSam). After of independent libraries of each sample, we removed duplicated reads using GATKs MarkDuplicates and removed all reads mapping to multiple-locations (multi-mappers). The final bam files were validated with GATK's ValidateSamFile.

3.3 Detection of germline variants
HaplotypeCaller v4.1.4.1 with the GATK package[69] was used to identify germline variants from the reference normal samples in each patient (buffy coats or adjacent normal tissue) using know germline variant annotations from the build 146 of the dbSNP database[70] separately for each chromosome. Resulting VCF files were then merged with GATK MergeVcfs. Variant recalibration was performed with gatk VariantRecalibrator with options set according to GATK best practices[70–73] and applied to VCF files using gatk ApplyVQSR with the options "-mode SNP -ts-filter-level 99.0" and "-mode INDEL -ts-filter-level 99.0" respectively. All germline variant calls marked as "PASS" were retained.

3.4 Verification of sample-patient matches
For all samples we excluded the possibly of sample mismatch by comparing germline variants identified in normal tissue to neoplasia samples of a given patient. The reads of each read-group were extracted with samtools view using options '-bh {input_bam} -r {read_group_id}' and GATK's CheckFingerprint tool was applied to extract statistics on sample-patient matches[74]. For virtually all high-purity samples without extensive loss of heterozygosity, we were able to confirm that the samples were obtained from the expected patient, for the latter group we inspected copy-number profiles (see below) to confirm that these matched the remaining samples.

3.5 Copy number analysis

Deep whole genome sequencing
Coverage of genomic loci relative to matched normal tissue samples (buffycoats or adjacent normals) were extracted with methods provided in the sequenza v2.1.2 package

for R[75] and binned in non-overlapping windows of 10^6 bp. B-allele frequencies (BAF) of germline mutations determined with the GATK HaplotypeCaller (see above) for each patient were added to these binned files. Joint segmentation on BAFs and read depth counts across all samples from a given tumour were used to determine a set of breakpoints to use for the subsequent analysis. Specifically, GC content bias correction from was applied using the 'gc.norm' method from sequenza v2.1.2 and positions with non-unique mappability (i.e., < 1), as determined by the approach of QDNAseq v.3.8[76], in windows of 50 bp were removed. Piecewise constant curves were fitted for each chromosome arm using the multipcf function (gamma = 80) from the copynumber v1.22.0 package for R[77]. The per-patient set of break points, binned depth-ratio and BAF data were then inputted into the sequenza algorithm (version 2.1.2) to determine allele specific copy-numbers, ploidy Ψ and purity ρ estimates[75]. The initial parameter space searched was restricted to $\{\rho \mid 0.1 \leq \rho \leq 1\}$ and $\{\Psi \mid 1 \leq \Psi \leq 7\}$. Upon manual review of the results, we identified several samples with unreasonable fits (cases where calls suggested extremely variable ploidy values across samples). For these samples, we manually identified alternative solutions consistent with the other samples and somatic variant calls.

Low-pass whole genome sequencing

Low-pass WGS bam files were processed using QDNAseq[76] to convert read counts in 500kb bins across the autosomes of hg38 into log2ratio data. Data normalisation was performed in accordance with the QDNAseq workflow, except for outlier smoothing (smoothOutlierBins function) which was seen to artificially depress signal from highly amplified bins. Bins for hg38 were also generated according to QDNAseq instructions. Log2 ratio values in each bin were normalised by subtracting the median log2 ratio from all log2 ratios per sample. Samples in a patient were segmented jointly using the multipcf function in the R package copynumber (gamma = 10)[77] and the mean segment log2ratio was calculated across the bins.

Absolute copy number status was calculated using the approach taken by ASCAT[78]. Using the ASCAT equation to describe logR ratios, we took an integer ploidy value Ψ_t in the tumour *t* as determined by paired deep WGS in each case and searched a range of purities from 0.1 to 1 (and assumed gamma was 1 as is the case in sequencing data). For each purity (ρ) value we calculated the continuous copy number status of each bin and calculated the sum of squared differences of these values to the nearest positive integer of the modulus. Purity estimates were given by local minima (goodness of fit to integer copy number values, measured as the sum of square distances) across the purity range considered. The absolute copy number state for each bin was taken as the closest integer value calculated using this purity. If no local minimum is found the purity is assumed to be 1. If the best solution produced negative copy number states at some loci, these were set to copy number zero to avoid impossible copy number states. In two patients per sample ploidies were determined by manual adjustment due to integer ploidy values producing poor fits.

3.6 SNV detection

Somatic mutations were first called for each tumour sample separately against matched blood derived or adjacent normal tissue samples with Mutect2 (version 4.1.4.1) using options "–af-of-alleles-not-in resource 0.0000025 –germline-resource af-onlygnomad.hg38.vcf.gz"[69,79] Variants detected in any tumour sample (marked PASS, coverage AD 10 in both normal and tumour, at least 3 variant reads in the tumour, 0 variant reads in the normal, reference genotype in normal and non-reference genotype in cancer) were merged into a single list of "candidate mutations". The multi-sample caller

Platypus v0.8.1.1[80] was then used to recall variants at each candidate mutation position in all samples of the patient. In practice, this meant that the pipeline leverage information across samples to improve the sensitivity of variant calling. The platypus output of joint variant calls was then filtered to only keep high quality variants with flags ”PASS”, ”alleleBias”, ”QD” or ”Q20”, in canonical chromosomes (i.e., not in decoy), a minimum number of reads NR>5 in all samples, a genotyping quality GQ>10 in all samples, a reference genotype (i.e., 0/0) in the normal reference and a non-reference genotype (i.e., 0/1 or 1/1) in at least one tumour sample.

To alleviate concerns of false-negative calls of mutations in important driver alterations, we generated a second set of variant calls for the identification of known driver mutations and dNdS analysis (see details below) to which we did not apply the second step of filtering.

3.7 SNV annotation

Somatic variants were annotated and candidate driver genes of colorectal cancers reported by[3] and IntOGen[34] as well as pan-cancer driver genes reported[32] and[81] filtered with the Variant Effect Predictor v93.2[82].

3.8 MSI status detection

The identification of microsatellite instability (MSI) colorectal cancers was performed with the MSIsensor v0.2[83]. We first determined the position of microsatellites sites by applying the msisensor scan method to the GRCh38 reference assembly and subset these to the first chromosome. In a second step we identified the fraction of mutated microsatellites in each sample using the msisensor msi method with default options. Generally, in known MSI cases (e.g., those identified by mutation burden and mutational signature) more than 30% of microsatellites were mutated and we used this as a critical value to classify cases as MSS and MSI. One exception was C562, where the low purity of the samples led to a low msisensor score. However, this case was clinically classified as MSI by pathological reports and it had a relatively high indel burden leading to the conclusion that it was MSI.

3.9 Extraction of reads supporting variants

Using the VCF files from both somatic and germline variant calling, we extracted the number of reads supporting the reference and alternate alleles as well as the total number of reads covering the sites from WGS, LP-WGS and ATAC-seq samples using python and the pysam library[68], pysam version 0.15.2, samtools version 1.9.

3.10 dN/dS analysis

dndscv package for R[32] was used for dN/dS analysis. Per-patient variant calls were obtained from the vcf files[84] and lifted over to the hg19 reference genome using the rtracklayer package for R[85] Variants were divided into clonal mutations (i.e., present in all samples) and subclonal mutations (i.e., present in a subset of samples) present in the cancer and a set of mutation present in any of the adenoma samples. MSI and MSS patients were treated separately. dndscv was applied separately to each of the four sets (MSI/MSS & clonal/subclonal) (using default parameters apart from deactivated removal of cases due to number of variants). Further, dN/dS values for a set of 167 chromatin modifier genes were extracted.

3.11 ATAC-seq

3.11.1 ATAC peak calling analysis

Extraction of cut-sites
For the detection of cut-sites (hereafter “peaks” where read density was high) bed-files of ATAC-seq cut-sites were produced. Aligned reads were sorted by read name using “samtools sort -n{bam}”, all proper reads pairs (i.e., reads mapped to the same chromosome and with correct read orientation) were isolated using ”samtools view -bf 0x2” and finally converted to the bed format using ”bedtools bamtobed -bedpe -mate1 -i{bam}”. Equivalent to[86] the start site of reads was shifted to obtain the cut sites: specifically, forward reads were shifted by -4 bases and reverse reads by +5 bases. ATAC-seq reads spanning nucleosomes have an insertion size periodicity of multiples of 200 bp and reads in regions of open-chromatin have insertion sizes smaller than 100 bp[86]. For this reason, in line with previous studies, ATAC-seq reads were divided into a set of nucleosome-free reads (insertion size ≤ 100) and a set of nucleosome associated reads (180 ≤ insertion size ≤ 620).

Peak detection
Peaks were called separately for each tumour region using MACS2 v2.21[87] using ”macs2 callpeak -f BED -g hs –shift -75 –extsize 150 –nomodel –call-summits –keep-dup all -p 0.01” with the concatenated and sorted bed read files of nucleosome-free cut-sites of all samples as input. A set of normal peaks (across patients) were also called using the concatenated normal sample bed files (i.e., region "E" samples) and per adenoma peak calls using all adenoma bulk samples as input.

Filtering and concatenation of peaks
Strict filtering of per-region peak calls was applied (extended by 250 bp, q-value of 0.1%, enrichment of 4.0, maximum number of peaks 20,000). Iterative merging was then applied, using a method equivalent to that used by[11] on per-region peak calls of individual patients (per-tumour peaks set) as well as across all cancer samples and pan-patient normal peak calls (pan-patient peak set). This procedure resulted in a total of N = 343,240 peaks, of which filtered N = 67,215 peaks called in >2 tumour regions or the panel of normal. The ChIPseeker v2.14.0 package for R[88] was used to annotate peaks based on their genomic location. For peaks that were not proximal to known promotor regions (1000 bp), overlaps with known Enhancer elements reported in the double-elite annotations of the GeneHancer database was examined[89]. The general distribution of these features in the genome and overlaps of peaks with those reported by[13].

Extraction of cut-sites in peaks
Read counts for each peak in the final set were collated using bedtools[90] as follows: ”bedtools coverage -a bed peaks -b bed cut sites -split -counts -sorted”.

3.11.2 Purity estimation for ATAC-seq and accounting for copy number alterations

Commented [AS25]: Comment 1.4b

Clonal variants identified by paired WGS sequencing (clonal variants were those present in all samples from the cancer) were used to estimate sample-specific ATAC-seq purity. First, variants in intervals with identical (clonal) copy-number states (i.e., A/B states) and regions of closed chromatin, were identified from WGS data. Copy-number values c_i and mutation multiplicity m_i of each variant site i were obtained from the WGS data. For a mutation at site i covered by $n_{s,i}$ reads in sample s the number of reads k_i containing the alternate allele is expected to follow a binomial distribution with the likelihood:

$$B\left(k_i \middle| p_{s,i}, n_{s,i}\right) = \binom{n_{s,i}}{k_i} p_{s,i}^{k_i} \left(1 - p_{s,i}\right)^{n_{s,i} - k_i}$$

where the expected success probability $p_{s,i}$ is a function of the sample purity as, the number of mutated alleles in the tumour cells $m_{s,i}$, the total copy-number of the mutated site in the tumour cells $c_{s,i}$ and the copy-number in contaminating normal cells CN=2

$$p_{s,i} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + (1-\rho_s)c_n} = \frac{\rho_s m_{s,i}}{\rho_s c_{s,i} + 2 - 2\rho_s}$$

A maximum-likelihood estimate of the sample purity ρ_s was then obtained by minimising the negative-log-likelihood across all N mutated sites:

$$l(\rho_s) = \sum_{i=0}^{N} -log\left(B\left(k_i | p_{s,i}, n_{s,i}\right)\right)$$

To account for the influence of copy number alterations on the read counts, the signal observed at a locus should be given by $S = S_N \frac{2(1-\rho)+\pi\rho}{2(1-\rho)+\psi\rho}$, where S_N is the signal of the reference allele, ρ the purity of the sample, π the copy-number of the locus and ψ the ploidy of the tumour. For pooled samples we calculate the average of S weighted by the total number of reads across samples. Indeed, CNAs were affecting the read depth at the locus (example in Figure S9A, CNA coefficient from the ATAC data vs CNA profile from WGS in black line). The correlation between the expected and observe coefficients is reported in Figure S9B, whereas the relationship between CN status and coefficient is reported in Figure S9C.

Commented [AS26]: Comment 1.4b

However, it is important to consider that in general CNAs are causing relatively small changes in the ATAC-seq signals compared to bona fide SCAA (example in Figure S10, with significant SCAAs indicated in red). This was demonstrated by the strong correlation of the recurrence number in the model with CN adjustment versus the one without (Figure S11). Where this approach was most relevant was in the identification of lost chromatin accessibility in regions with a CN gain and gained chromatin accessibility in regions with a CN loss.

3.11.3 Identification of recurrently altered peaks across patients

Analysis was restricted to samples with purity ρ > 0.4. Peaks proximal (≤ 1000 bp) to a transcription start site (i.e., promotors) and those more distant to a TSS (i.e., putative enhancers) were considered separately to account for the possibility of differential dispersion. Whereas we relied on proximity for promoters, we used the GeneHancer database for enhancers[89]. An overdispersed Poisson model was fitted to each peak *edgeR* v3.30.3[91,92], and per sample set normalisation factors were calculated using the TMMwsp method[93], estimated a global dispersion estimate across sets from all cancers and compared each set of pure glands (per-patient) against a large pool-of-normal tissue ATAC-seq samples. Recurrently altered peaks were identified as those that were significantly altered at a level of p ≤ 0.01 in at least 4/26 (i.e., 20%) of cases.

Commented [AS27]: Comment 3.3

3.11.4 Identification of associated changes in gene expression

The basic processing of matched RNA-seq data is described in the associated manuscript TRANSCRIPTOME. A subset of 27,699 peaks that where either adjacent to a known transcription start site (TSS) of a gene[94] or overlapped a previously characterised enhancer element described in the GeneHancer database[89] were identified. Of these

456/27699 (≅ 1.65%) were recurrently altered. Changes in gene-expression of genes associated with these sites were tested for using *DESeq2*[95] to compare coefficients of the fitted beta-binomial regression model (design: *~Patient*, with all normal samples as 'Normal') with the contrast argument being a list of vectors containing the significant and non-significant patient sets.
For promotors, a one-tailed hypothesis test was applied by setting the *altHypothesis* argument to 'less' (for closed peaks) or 'greater' (for opened peaks). For enhancers a two-tailed hypothesis test on all associated genes was applied by setting the *altHypothesis* argument to 'greaterAbs'. P-values were from all tests were adjusted for multiple hypothesis testing using FDR method[96] associations at FDR<0.1% were reported. For the visualisation of gene expression values, the average gene expression values across samples from a given cancer and all normal samples on variance stabilised (log-transformed) FPM values (counts per million reads in gene) were calculated.

3.11.5 Identification of subclonal changed is recurrently altered peaks
Subclonality was assessed only for a set of recurrent somatic accessibility changes, comprising recurrent events affecting driver genes and the top 25 most recurrent in each of the of the 4 categories: gained promoter, lost promoter, gained enhancer, lost enhancer (total of 521 sites assessed).

Our previous analyses recognised that sample purity was highly correlated with tumour piece (regions A-D). To distinguish subclonal chromatin accessibility alterations from variability in ploidy, regression to account for purity was performed. Specifically, a log ratio test from DESeq2[97] was used to compare a "full model" *~purity + region* to a reduced model *~purity*. Samples from the same region were used as biological replicates. Events were considered putatively subclonal when the adjusted p-value was below 0.05 and if the direction of log fold change from analysis of matched bulk tissues was correlated with that observed in individual samples. In the case of gained events, subclonal events were filtered out if MACS peak-calling (see above) had not called a peak within 500 bp of the location of the putative gain event (this removed 33 sites). For losses, 5/45 subclonal events were removed as the log fold change was in the wrong direction.

For visualisation of peaks, coverage per region was calculated 1 kb upstream and 1kb downstream from the centre of the peak. Coverage was normalised per million reads in peaks and was plotted using functions from *GenomicRanges*[97] and *Gviz*[98].

3.12 TF Binding site prediction
The *motifmatchr* package for R[99], a reimplementation of the C++ library MOODS[100,101], was used to identify binding sites for all human TF motifs defined in a curated version of the CIS-BP database[102]. The list of predicted binding sites was filtered using a minimum significance value of $p \leq 10^{-6}$, followed by removal of binding sites in centromeric regions and non-autosomal (i.e., sex and non-canonical) chromosome. After this initial filtering predicted binding sites were split into six distinct groups based on i) there distance to the next TSS (proximal: $d \leq$ 2000 bp, close: 2000 bp < $d \leq$ 10,000 bp, distal d > 10,000 bp) and ii) whether they overlapped with a peak observed in the ATAC-seq data. For a number of TF homotypic clustering of binding sites in specific intervals was observed; to account for this binding sites that where closer than $d \leq$ 1000 bp to the next predicted binding site of the same TF were removed.

Extraction of signal values

For each of the TF sets described above, the counts of insertions around the centre of the TF binding site (±1000 bp) as well as the insertion size of the read pair (i.e., the distance to the second nick) for each sample[97] were tabulated. The insertion-sizes (rows) were binned into intervals of 5 bp and divided by total count of reads with an equivalent size in the entire genome. After this the background signal was estimated to be the average number of insertions 1000 bp – 750 bp from the centre of TF binding site per insertion size and subtracted from the counts. The difference between these "normalised and background corrected TF signals" in each sample and a pool of normal samples was calculated and integrated across the central region of the TF binding sites (insertion size [25;120], distances [-100 bp;100 bp]) as a summary statistic. Regression analysis linear regression was used to identify associations with purity estimates and in this context signals were found to correlated with TSSe (for both nucleosome-free and all reads). For this reason, an additional term was added to the regression model of each TF to correct for this effect: *signal ~ tsse*tsse$_{nf}$ + purity:patient* where tsse and *tsse$_{nf}$* are the TSSe differences of the sample and the pooled-normal samples) and weighted each observation by the square root of the number of reads in the sample. A second linear model in which a region-specific effect of the purity: *signal ~ tsse*tsse$_{nf}$ + purity:region* was considered was also fitted to the data. For both models, the statistical significance of the 'purity' coefficient was determined. The estimates of the coefficients were also used as a patient specific summary for subsequent analysis.

Cluster analysis
The analysis was focused on the 150 TF for which a significant association with the tumour cell content (i.e., the purity) and TF signal was most frequently observed. With the aim to identify general patterns in these data, a clustering analysis was conducted (hierarchical clustering with Euclidean distance and complete linkage). This method identified three major groups of TFs, and to each of these, analysis with String-DB[103] was applied to identify significantly overrepresented pathways.

3.13 Methylation arrays analysis

A reference normal dataset methylation array dataset was downloaded from[104] that including normal tissue sampled adjacent to colorectal cancers that was profiled using the HumanMethylation450 BeadChip array (Illumina).

Here, 8 bulk samples from 4 patients (C516, C518, C560 and C561) were profiled MethylationEPIC BeadChip (Infinium) microarray according to manufactorer's instructions.

The ChAMP R package pipeline[105] was used to analyse the methylation beadarray data. Probes that had a detection of P > 0.01 and probes with <3 beads in at least 5% of samples per probe, probes that were on the X or Y chromosome, all SNP-related probe as well as all multi-hit probes were all removed. Subset-within-array normalization to was used to correct for biases resulting from type 1 and type 2 probes on the array. After QC and normalization, beta values were calculated for further comparison.

To compare the methylation patterns between our samples and the reference normal dataset, the overlapped probes of all samples located in the region of distal to TSS (dTSS), close to TSS (cTSS) and proximal to TSS (pTSS) in both on ATAC peak (oPEAK) and not on ATAC peak (nPEAK) were compared.

3.14 Processing of RNA-seq

After initial quality control with FastQC[106] and default adaptor trimming with Skewer[61], paired-end reads were aligned to GRCh38 reference genome and version 28 of the Gencode GTF annotation using the STAR 2-pass method[107]. Read groups were added with Picard v.2.5.0[108]. Per gene read counts were produced with htseq-count that is incorporated into the STAR pipeline[42].

Filtering of RNA samples

Raw gene counts were first filtered for reads uniquely assigned to non-ribosomal protein-coding genes located on canonical chromosomes (chr1-22, X and Y). If samples had less than 5M of these 'usable' reads they were re-sequenced to improve coverage. Where possible, the same library preparation pool was sent again for sequencing. These `top-ups' proved to be true technical replicates, since the resulting gene expression of the re-sequenced samples clustered very closely to their original samples on both a sample-sample heatmap and a principal component analysis (PCA). It was therefore determined that the fastqs of these samples could simply be merged at the start of the pipeline. In cases where resequencing was required but insufficient library remained, a new library was prepared and the sequencing run that produced the highest read was used in subsequent analysis. For 8 samples, the sequencing of the second library contained too few reads to enable downstream analysis. 6/8 samples showed per gene read counts that were very similar between libraries 1 and 2 (Spearman's rank correlation between replicates was significantly higher than the mean; Wilcoxon one-way rank test; FDR<0.01) and so read counts were combined across libraries, the other 2/8 samples were discarded. Samples were also discarded if matched DNA-sequencing revealed a tumour purity of less than 0.05.

Commented [AS28]: Comment 1.11

Gene expression normalisation and filtering

The number of non-ribosomal protein coding genes on the 23 canonical chromosome pairs used for quality control was 19,671. Raw read counts uniquely assigned to these genes were converted into both transcripts per million (TPM) and variance-stabilising transformed (vst) counts via DESeq2[95].

A list of expressed genes (n=11,667) was determined by filtering out genes for which less than 5% of tumour samples had at least 10TPM. In order to concentrate on tumour epithelial cell gene expression, genes were further filtered out if they negatively correlated with purity as estimated from matched DNA sequencing data (see associated manuscript EPIGENOME for methodology of purity estimation). Specifically, for the 157 tumour samples that had matched DNA-sequencing and therefore accurate purity estimates, a linear mixed effects model of Exp (vst) ~ Purity + (1|Patient) was compared via a chi-squared test to Exp ~ (1|Patient). Genes which had a negative coefficient for Purity in the first model and an FDR adjusted p-value less than 0.05, suggesting that Purity significantly affected the expression, were filtered out. This led to a filtered list of 11,401 expressed genes.

3.15 Mutational signatures analysis

Mutational signatures analysis was performed with SparseSignatures[51]. This method uses LASSO regularization[109] to reduce noise in the signatures, controlled by a regularization parameter lambda (λ). It implements a procedure based on bi-cross-validation[110] to select the best values for both the regularization parameter λ and the number of signatures.

Deconvolution using a maximum of 10 signatures was performed and values of λ of 0.000, 0.025, 0.050 and 0.100 were tested. Optimal parameters were selected based on the median bi-cross-validation error estimated over 1000 iterations, resulting in an optimal estimate with minimum cross-validation median error when 6 signatures were fitted and λ=0.025. A second analysis with SigProfiler[111], with default parameters and a total of 1000 iterations, confirmed the existence of these signatures.
Signatures based clustering was performed considering the 6 signatures solution by SparseSignatures; the low-rank signatures exposure matrix given as an output by the tool was used to compute the pairwise similarity matrix for each patient as 1 - cosine similarity of their exposures. Clustering was then performed on the similarity matrix by k-means with 6 clusters explaining all the variance. Although from a statistical perspective, clusters C3 and C4 are defined by a few samples (and explain 3/4% of the variance), from the biological perspective, we have evidence that in these patients the distribution of mutations resembles very different signatures and mutational processes (Figure S36A).

Commented [AS29]: Comment 1.15

Mutational signatures exposures were also analysed across epigenetic regions. Mutations were first grouped in clonal or subclonal across whole genome and then in different genomic regions (as described above). Signatures activities in each region was estimated by Jackknife sampling[112]. Specifically, data from each patient were partitioned based on their clusters as defined above, and repeated Jackknife sampling performed 100 times independently for each of the 3 clusters (including a random sample of 90% of the tissue samples each time). For each iteration the mutations within each genomic region were used to computed a data matrix normalised against trinucleotide count (across the 96 channels) in the whole genome versus region specific counts, and signatures assignments then performed on the normalized data by LASSO[51,109]. Finally, relative signature activities estimated over the 100 Jackknife samples were normalized based on total size of each region. Moreover, since clusters C3 and C4 represent rare and very distinct mutational patterns, we excluded these samples from the estimation of mutational processes in the epigenetic regions by Jackknife and instead we focused on MSS (cluster 1) vs MSI (clusters 2 and 5) tumour, as samples in clusters C3 and C4 would have probably biased the Jackknife estimation for these two groups.

Supplementary Figure Legends

Figure S1. Colectomy specimen collection images. Resection specimens were collected from UCLH and sampled with the supervision of a pathologist. Spatial information on different regional samples was retained and indicated in the images. A, B, C, D are cancer regions. E is distant normal epithelium. Eventual concomitant adenomas are reported as F, G, H, etc.

Figure S2. Correlation between gene expression in TCGA normal colon samples vs our normal samples.

Figure S3. Gland and bulk collection from each tumour region. We collected individual glands from cancer and normal samples from different regions of each tumour. We also collected ‘minibulks’, composed by agglomerate of a few dozen glands. Each sample was imaged individually.

Figure S4. Copy number alteration profiles. We estimated absolute copy number alterations for each sample in each patient, both for deep WGS and low-pass WGS.

Figure S5. Chromosomal differences between adenomas and carcinomas. (A) Ploidy and **(B)** PGA (Percentage Genome Altered) of adenomas vs carcinomas, separated by MSI/MSS status. **(C)** Comparison of the two values.

Figure S6. Single nucleotide variant profiles. We called point mutations and Indels in each sample and identified clusters of mutations found at the same frequency in the same samples. Values in Cancer Cell Fraction (CCF) are represented.

Figure S7. Mutations in chromatin modifier genes for all samples.

Figure S8. Images of all the normal samples used for ATAC-seq reference.

Figure S9: Association between CNA and the observed ATAC-seq signals. (A) As expected, association between CNA and the coefficients of the negative-binomial regression with edgeR was found. Black line indicates CNA profile determined by WGS. **(B)** This relationship was generally explained by the expected effect of CNA under a given ploidy and purity. The black crosses show the average signal of sides with a given rounded CN and the black line the expected relationship. **(C)** A consistent variability of the signal at different CN states was observed.

Figure S10. Identified SCAAs after adjustment for CNAs. A subset of highly recurrent SCAAs are labelled in the figure. Significantly differential SCAAs are show in red.

Figure S11. Number of significant SCAA before and after adjustment of the test for the expected effects of CNA alterations.

Figure S12. Peak densities for promoter gained loci in Figure 3D.

Figure S13. Peak densities for promoter lost loci in Figure 3D.

Figure S14. Peak densities for enhancer gained loci in Figure 3E.

Figure S15. Peak densities for enhancer lost loci in Figure 3E.

Figure S16. Peak densities for promoter gained peaks found subclonal from Figure 3D,E.

Figure S17. Peak densities for promoter lost peaks found subclonal from Figure 3D,E.

Figure S18. Peak densities for enhancer gained peaks found subclonal from Figure 3D,E.

Figure S19. Peak densities for enhancer lost peaks found subclonal from Figure 3D,E.

Figure S20. Comparison of peak calling in our cohort from reanalysed TCGA ATAC-seq data.

Figure S21. Peak densities for peaks found in cancers but not in concomitant adenomas in Figure 3D,E.

Figure S22. SCAA burden of adenomas vs carcinomas. (A) Purity of adenomas and carcinomas are comparable, excluding the differences in chromatin accessibility are due to cellularity. **(B)** Coverage differences are appreciable between cancers and adenomas, however when adjusted for number of reads in peaks **(C)** it is the case that SCAA burden is significantly higher in carcinomas.

Figure S23. SCAAs identified in individual normal glands. (A) Heatmap of recurrent losses and gains promoter SCAAs identified in normal glands. This figure is equivalent to that shown in Figure 3 of the main manuscript. The last column of each patient shows if reads showed significantly differentially accessibility in a pool of all normal glands of patient. **(B)** Shows the distribution of losses and gains for all peaks. **(C)** Shows the lack of correlation of the recurrence of SCAAs in normal glands and the cancers.

Figure S24. Gene expression differences for all the recurrent peaks that correlated with gene expression.

Figure S25. Some of the eQTL somatic variants from a related analysis (Househam, Heide et al.) were associated with changes in chromatin accessibility at the locus. Here we show the significant examples of this phenomenon for each patient, with the SNV phylogenetic tree (left) versus changes in chromatin accessibility (right).

Figure S26. Transcription Factor binding sites density plots for annotations in Figure 4D.

Figure S27. Overlapping of TF annotations in Figure 4D.

Figure S28. Correlation of the TF signal in Figure 4D between all versus only unique loci.

Figure S29. Gene expression of TFs from cluster 1 of heatmap in Figure 4A.

Figure S30. Methylation levels of CIMP (CpG Island Methylation Phenotype) markers of cancers and normal.

Figure S31. Demethylation in reactivated TF binding sites. (A) We selected genomic regions in cluster 3 (enriched in developmental genes like SOX and HOX families) and verified their methylation status with CpG methylation arrays in EPICC samples versus normal. **(B)** In particular regions corresponding to binding sites of DLX5, HOXA4, HOXB4, ISL1, SOX5 and SOX6 showed decreased methylation in cancer vs normal.

Figure S32. Methylation levels of cancer vs normal for housekeeping genes and genes that are usually methylated in normal. These results exclude a global hypomethylation pattern in the cancers.

Figure S33. Heritability of chromatin accessibility. (A) We compared ATAC distance (euclidean on promoter peaks) between glands from the same region (within-region) and glands of different regions (between-regions) to evaluate divergence of chromatin against space and genetic distance. **(B)** For the large majority of patients within-region ATAC

distance is significantly lower than between region, indicating heritability of the chromatin that follows the spatial and phylogenetic structure of the tumour. Here we plot the F statistics of the ANOVA model on TSSe, number of reads, and region. **(C)** The distances between and within regions (left) and correlations with the genetic distance (right). **(D)** Cases in which no correlation with the genetic distances existed data were often from low purity samples or sparse.

Figure S34. Coefficients of the ANOVA model for the correlation between genetic and epigenetic distance for each region.

Figure S35. Clonality of TF binding site accessibility. A significant proportion of TF binding site accessibility changes were 'clonal' within the tumour, with distant regions showing the same pattern, again testimony of the heritability of chromatin accessibility. In this example CTCF loop and promoter loci in C548. However, there were some exceptions, as in this example of C543.

Figure S36. Mutational signature discovery with SparseSignatures. (A) Mutational signature discovery with sparse signatures identified 6 signatures in our cohort. **(B)** Principal Component Analysis divided the patients into 5 clusters depending on contribution from each signature. **(C)** Signature activity varied between clusters

Figure S37. Mutational signature deconvolution with SigProfiler.

Figure S40. Accumulation of different mutational signatures in distinct epigenetic regions.

Figure S39. Predicted versus observed mutational signatures that cause gain and loss of CTCF.

Figure S40. Proportion of each signature contributing to mutations affecting CTCF binding.

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
3. The S:CORT Consortium *et al.* The evolutionary landscape of colorectal tumorigenesis. *Nat. Ecol. Evol.* **2**, 1661–1672 (2018).
4. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nat. Genet.* **47**, 209–216 (2015).

5. The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
6. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
7. Mazor, T., Pankov, A., Song, J. S. & Costello, J. F. Intratumoral Heterogeneity of the Epigenome. *Cancer Cell* **29**, 440–451 (2016).
8. Black, J. R. M. & McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nat. Rev. Cancer* **21**, 379–392 (2021).
9. Nam, A. S., Chaligne, R. & Landau, D. A. Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. *Nat. Rev. Genet.* **22**, 3–18 (2021).
10. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
11. Akhtar-Zaidi, B. *et al.* Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. *Science* **336**, 736–739 (2012).
12. Johnstone, S. E. *et al.* Large-Scale Topological Changes Restrain Malignant Progression in Colorectal Cancer. *Cell* **182**, 1474-1489.e23 (2020).
13. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
14. Humphries, A. & Wright, N. A. Colonic crypt organization and tumorigenesis. *Nat. Rev. Cancer* **8**, 415–424 (2008).
15. Baker, A.-M. *et al.* Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon. *Cell Rep.* **8**, 940–947 (2014).
16. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).
17. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal Stem Cell Replacement Follows a Pattern of Neutral Drift. *Science* **330**, 822–825 (2010).

18. Snippert, H. J. *et al.* Intestinal Crypt Homeostasis Results from Neutral Competition between Symmetrically Dividing Lgr5 Stem Cells. *Cell* **143**, 134–144 (2010).
19. Merlos-Suárez, A. *et al.* The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse. *Cell Stem Cell* **8**, 511–524 (2011).
20. Yatabe, Y., Tavare, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl. Acad. Sci.* **98**, 10839–10844 (2001).
21. Nicolas, P., Kim, K.-M., Shibata, D. & Tavaré, S. The Stem Cell Population of the Human Colon Crypt: Analysis via Methylation Patterns. *PLoS Comput. Biol.* **3**, e28 (2007).
22. Shibata, D. Inferring human stem cell behaviour from epigenetic drift: Somatic cell mitotic clocks. *J. Pathol.* **217**, 199–205 (2009).
23. Humphries, A. *et al.* Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proc. Natl. Acad. Sci.* **110**, E2490–E2499 (2013).
24. Kang, H. *et al.* Many private mutations originate from the first few divisions of a human colorectal adenoma: Co-clonal expansion. *J. Pathol.* **237**, 355–362 (2015).
25. Siegmund, K. D., Marjoram, P., Tavaré, S. & Shibata, D. Many colorectal cancers are "flat" clonal expansions. *Cell Cycle* **8**, 2187–2193 (2009).
26. Tsao, J.-L. *et al.* Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl. Acad. Sci.* **97**, 1236–1241 (2000).
27. Tsao, J.-L. *et al.* Colorectal Adenoma and Cancer Divergence. *Am. J. Pathol.* **154**, 1815–1824 (1999).
28. Cross, W. *et al. Stabilising selection causes grossly altered but stable karyotypes in metastatic colorectal cancer*. http://biorxiv.org/lookup/doi/10.1101/2020.03.26.007138 (2020) doi:10.1101/2020.03.26.007138.
29. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, (2015).

30. Lin, S.-H. *et al.* The somatic mutation landscape of premalignant colorectal adenoma. *Gut* **67**, 1299–1305 (2018).
31. Zapata, L. *et al.* Negative selection in tumor genome evolution acts on essential cellular functions and the immunopeptidome. *Genome Biol.* **19**, 67 (2018).
32. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
33. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
34. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
35. Kaneda, H. *et al.* FOXQ1 Is Overexpressed in Colorectal Cancer and Enhances Tumorigenicity and Tumor Growth. *Cancer Res.* **70**, 2053–2063 (2010).
36. Tang, H. *et al.* Forkhead Box Q1 Is Critical to Angiogenesis and Macrophage Recruitment of Colorectal Cancer. *Front. Oncol.* **10**, 564298 (2020).
37. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
38. Househam, J. *et al. Phenotypic plasticity and genetic control in colorectal cancer evolution*. http://biorxiv.org/lookup/doi/10.1101/2021.07.18.451272 (2021) doi:10.1101/2021.07.18.451272.
39. Fang, C. *et al.* Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome Biol.* **21**, 247 (2020).
40. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
41. Aitken, S. J. *et al.* CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biol.* **19**, 106 (2018).

42. Chae, Y. K. *et al.* Epithelial-mesenchymal transition (EMT) signature is inversely associated with T-cell infiltration in non-small cell lung cancer (NSCLC). *Sci. Rep.* **8**, 2918 (2018).

43. Serresi, M. *et al.* Functional antagonism of chromatin modulators regulates epithelial-mesenchymal transition. *Sci. Adv.* **7**, eabd7974 (2021).

44. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

45. Ogino, S. CpG island methylator phenotype (CIMP) of colorectal cancer is best characterised by quantitative DNA methylation analysis and prospective cohort studies. *Gut* **55**, 1000–1006 (2006).

46. Gaiti, F. *et al.* Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).

47. Fennell, K. A. *et al.* Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature* **601**, 125–131 (2022).

48. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* 1–11 (2020) doi:10.1038/s41588-020-0708-0.

49. Australian Pancreatic Cancer Genome Initiative *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

50. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

51. Lal, A., Liu, K., Tibshirani, R., Sidow, A. & Ramazzotti, D. De novo mutational signature discovery in tumor genomes using SparseSignatures. *PLOS Comput. Biol.* **17**, e1009119 (2021).

52. Lee, D. *et al.* A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).

53. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).

54. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).

55. Liau, B. B. *et al.* Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell* **20**, 233-246.e7 (2017).

56. Suva, M. L., Riggi, N. & Bernstein, B. E. Epigenetic Reprogramming in Cancer. *Science* **339**, 1567–1570 (2013).

57. Pich, O. *et al.* Somatic and Germline Mutation Periodicity Follow the Orientation of the DNA Minor Groove around Nucleosomes. *Cell* **175**, 1074-1087.e18 (2018).

58. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).

59. Hinohara, K. *et al.* KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance. *Cancer Cell* **34**, 939-953.e9 (2018).

60. Buczacki, S. J. A. *et al.* Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature* **495**, 65–69 (2013).

61. Jiang, H., Lei, R., Ding, S.-W. & Zhu, S. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**, 182 (2014).

62. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).

63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

64. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

65. Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinforma.* **43**, (2013).

66. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

68. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

69. Poplin, R. *et al. Scaling accurate genetic variant discovery to tens of thousands of samples*. http://biorxiv.org/lookup/doi/10.1101/201178 (2017) doi:10.1101/201178.

70. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

71. Frazer, K. A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).

72. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

73. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006).

74. Javed, N. *et al.* Detecting sample swaps in diverse NGS data types using linkage disequilibrium. *Nat. Commun.* **11**, 3697 (2020).

75. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).

76. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* **24**, 2022–2032 (2014).

77. Nilsen, G. *et al.* Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* **13**, 591 (2012).

78. Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci.* **107**, 16910–16915 (2010).

79. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).

80. WGS500 Consortium *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

81. Tarabichi, M. *et al.* A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**, 144–155 (2021).

82. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

83. Niu, B. *et al.* MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics* **30**, 1015–1016 (2014).

84. Obenchain, V. *et al.* VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).

85. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).

86. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

87. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

88. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

89. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).

90. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

91. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
92. 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).
93. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
94. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2019).
95. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
96. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
97. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
98. Hahne, F. & Ivanek, R. Visualizing Genomic Data Using Gviz and Bioconductor. in *Statistical Genomics* (eds. Mathé, E. & Davis, S.) vol. 1418 335–351 (Springer New York, 2016).
99. Schep, A. University, S. motifmatchr: Fast Motif Matching in R. *Bioconductor Version Release 312* (2021).
100. Korhonen, J., Martinmaki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
101. Pizzi, C., Rastas, P. & Ukkonen, E. Finding Significant Matches of Position Weight Matrices in Linear Time. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 69–79 (2011).
102. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).

103. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

104. Fennell, L. *et al.* Integrative Genome-Scale DNA Methylation Analysis of a Large and Unselected Cohort Reveals 5 Distinct Subtypes of Colorectal Adenocarcinomas. *Cell. Mol. Gastroenterol. Hepatol.* **8**, 269–290 (2019).

105. Tian, Y. *et al.* ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982–3984 (2017).

106. A quality control analysis tool for high throughput sequencing data. https://github.com/s-andrews/FastQC.

107. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

108. Picard Tools - By Broad Institute. http://broadinstitute.github.io/picard/.

109. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).

110. Owen, A. B. & Perry, P. O. Bi-cross-validation of the SVD and the nonnegative matrix factorization. *Ann. Appl. Stat.* **3**, (2009).

111. PCAWG Mutational Signatures Working Group *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

112. Efron, B. & Stein, C. The Jackknife Estimate of Variance. *Ann. Stat.* **9**, (1981).

Acknowledgments

This study was principally supported by funding from the Medical Research Council (MR/P000789/1 to A.S.) and the Wellcome Trust (202778/Z/16/Z to T.A.G. and 202778/B/16/Z to A.S.). A.S. and T.A.G. were also supported by Cancer Research UK (A22909 and A19771) and the National Institute of Health (NCI U54 CA217376 to D.S., T.A.G. and A.S). This work was also supported by a Wellcome Trust award to the Centre for Evolution and Cancer at the ICR (105104/Z/14/Z). We thank the ICR's Tumour Profiling Unit, specifically Nik Matthews, Pradeep Ramigiri, Ioannis Assiotis, Kerry Fenwick and Ritika Chauan for their support in the sequencing efforts. D.R. was partially supported by a

Bicocca 2020 Starting Grant and by a Premio Giovani Talenti dell'Università degli Studi di Milano-Bicocca. L.M. is supported by Cancer Research UK (A23110).

Conflicts of interest

The authors declare no conflict of interest.

Data availability

Analysed data are available on Mendeley: https://data.mendeley.com/datasets/dvv6kf856g/2. Sequence data have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001005230. Further information about EGA can be found on https://ega-archive.org.

Code availability

Complete scripts to replicate all bioinformatic analysis and perform simulations and inference are available at: https://github.com/sottorivalab/EPICC2021_data_analysis. Further exploration of the ATAC-seq data shown in Figure 3 can be done using a Shiny-App accessible at https://theide.shinyapps.io/EPICC_shiny_app/.

Author contributions

T.H. analysed and interpreted the data, with focus on ATAC and WGS data. J.H. analysed and interpreted the data, with focus on RNA data. G.D.C. performed copy number analysis. I.S. devised multi-omics protocol, collected the samples and generated the data. C.K. collected the samples and contributed to data generation. C.L. contributed to ATAC data analysis. M.M., J.F.M., A.M.B., H.C., M.M., contributed to data generation. B.J. analysed methylation array data. L.Z.O. contributed to dN/dS data analysis. C.J., E.L., G.Car. contributed to data analysis. D.N. and K.C. contributed to signature analysis. A.B. generated methylation array data. I.B. contributed to analysis of CTCF binding site mutations. M.J. contributed to tissue collection. D.R. performed mutational signature analysis. D.S. contributed to experimental design and data interpretation. J.B. contributed to sample collection coordination. M.R.J. supervised sample collection. L.M. contributed to result interpretation. T.A.G. and A.S. conceived, obtained funding, supervised the study and wrote the manuscript.