nature medicine

Chromatin immunoprecipitation from fixed clinical tissues reveals tumor-specific enhancer profiles

Paloma Cejas[1–3], Lewyn Li[1,2], Nicholas K O'Neill[2], Melissa Duarte[1,2], Prakash Rao[1,2], Michaela Bowden[4], Chensheng W Zhou[4], Marta Mendiola[3,5], Emilio Burgos[5], Jaime Feliu[3], Juan Moreno-Rubio[6], Héctor Guadalajara[7], Víctor Moreno[8], Damián García-Olmo[7], Joaquim Bellmunt[2], Stephanie Mullane[2], Michelle Hirsch[9], Christopher J Sweeney[2], Andrea Richardson[9], X Shirley Liu[1,10], Myles Brown[1,2], Ramesh A Shivdasani[1,2] & Henry W Long[1,2]

Extensive cross-linking introduced during routine tissue fixation of clinical pathology specimens severely hampers chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) analysis from archived tissue samples. This limits the ability to study the epigenomes of valuable, clinically annotated tissue resources. Here we describe fixed-tissue chromatin immunoprecipitation sequencing (FiT-seq), a method that enables reliable extraction of soluble chromatin from formalin-fixed paraffin-embedded (FFPE) tissue samples for accurate detection of histone marks. We demonstrate that FiT-seq data from FFPE specimens are concordant with ChIP-seq data from fresh-frozen samples of the same tumors. By using multiple histone marks, we generate chromatin-state maps and identify *cis*-regulatory elements in clinical samples from various tumor types that can readily allow us to distinguish between cancers by the tissue of origin. Tumor-specific enhancers and superenhancers that are elucidated by FiT-seq analysis correlate with known oncogenic drivers in different tissues and can assist in the understanding of how chromatin states affect gene regulation.

The occurrence of recurrent mutations in chromatin-modifying genes in many human cancers[1] underscores the importance of chromatin in disease pathogenesis, and distinctive chromatin features in tumors correlate with clinical outcomes. Innumerable blocks of FFPE tissue are available, with associated clinical information, to investigate how particular chromatin states might reflect or underlie clinical features. However, ChIP-seq analysis of archived FFPE human clinical samples has, to date, met with limited success in specimens from individuals with cancer. This is, in part, because the prolonged exposure of clinical specimens to formalin results in excessive chemical cross-linking, which limits the isolation of soluble DNA–protein complexes. Although one published protocol for ChIP–qPCR (ref. 2) demonstrated enrichment of specific genomic regions in some clinical FFPE samples, genome-wide ChIP-seq analysis has never been shown for clinical FFPE tumor samples, and no other groups have applied this protocol.

Here we present a robust method to extract soluble chromatin from human clinical samples that have been preserved for long periods of time (median 8 years; range 2–15 years; **Supplementary Table 1**) in different hospitals and countries. Despite the presumably discordant conditions for sample preservation, this method enabled us to readily assess various histone marks and to generate chromatin-state maps comparable to those from cell lines and fresh-frozen (FF) tumors. As compared directly to the protocol that was previously reported for processing FFPE samples[2], FiT-seq shows higher resolution and sensitivity. FiT-seq analysis for samples of colorectal, breast and bladder carcinomas and for seminomas identified enhancer patterns and superenhancers that distinguish each tumor type. FiT-seq analysis of colon cancers revealed substantial chromatin remodeling as compared to that in matched normal tissues.

RESULTS

A ChIP-seq protocol for archival FFPE tissue specimens

We first studied the activation-associated histone mark histone H3 dimethyl Lys4 (H3K4me2), which appears at promoters and enhancers[3], in FFPE samples of primary colorectal cancer (CRC) by using the previously published method pathology tissue–chromatin immunoprecipitation (PAT–ChIP)-seq (ref. 2). To release soluble chromatin, the protocol uses micrococcal nuclease digestion, followed by sonication and immunoprecipitation (IP). We observed low chromatin yields (typically <1 μg from 20 10-μm sections); more

[1]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Medical Oncology, Hospital La Paz Institute for Health Research (IdiPAZ), Madrid, Spain. [4]Center for Molecular Oncologic Pathology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [5]Department of Pathology, Hospital La Paz Institute for Health Research (IdiPAZ), Madrid, Spain. [6]Precision Oncology Laboratory, Infanta Sofia University Hospital, Madrid, Spain. [7]Department of Surgery, Jimenez Diaz Foundation, Madrid, Spain. [8]Clinical Research Unit, Jimenez Diaz Foundation, Madrid, Spain. [9]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [10]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to R.A.S. (ramesh_shivdasani@dfci.harvard.edu) or H.W.L. (henry_long@dfci.harvard.edu).

Received 20 January; accepted 16 March; published online 25 April 2016; doi:10.1038/nm.4085

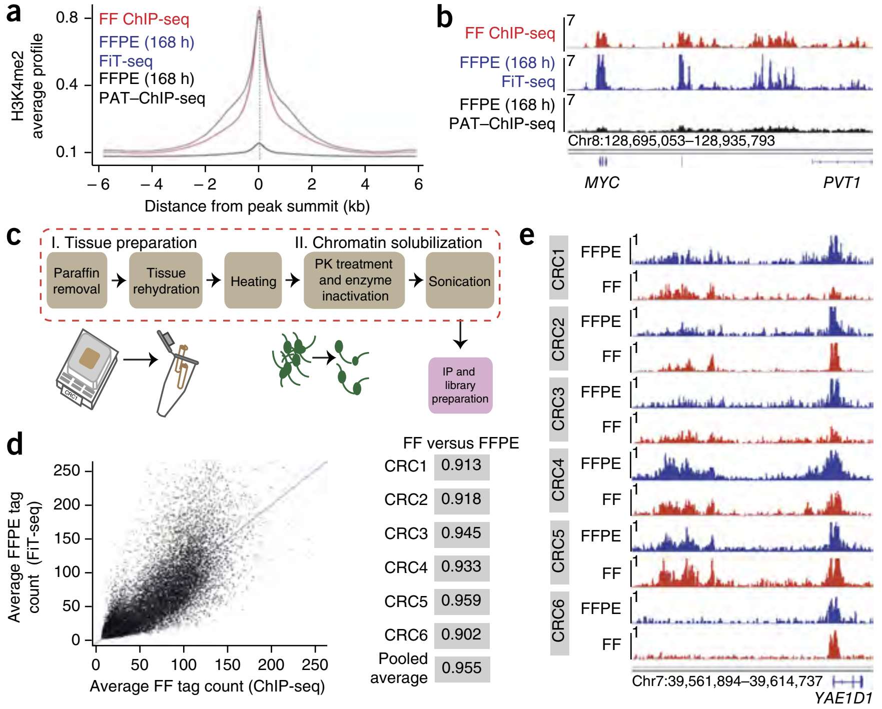

Figure 1 Development of FiT-seq, and its effect on strength and resolution of histone mark signals. (**a**) Average distribution of H3K4me2 signals in mouse xenografts of MCF-7 breast cancer cells either fresh-frozen (FF) and processed with conventional ChIP-seq or fixed in formaldehyde for 168 h and subsequently processed with the PAT–Chip protocol or using FiT-seq. The summits of the top 5,000 ChIP-seq peaks from the FF xenograft were used as the centers to build aggregate plots of the data from one replicate of each sample class. (**b**) Corresponding integrative genome viewer (IGV) tracks at a single representative, well-known locus (*MYC*), illustrating FiT-seq signals and comparing them to ChIP-seq signals obtained from FF MCF-7 xenograft samples. Each *y* axis indicates the number of reads per million per base pair (rbm). (**c**) Key steps in the FiT-seq protocol. (**d**) Correlation plot of average H3K4me2 signals in experiments on six pairs of FFPE and FF samples of primary CRCs. Each pair was taken from the same surgical specimen. All wiggle files from each sample type (FF or FFPE) were pooled for the analysis. Right column shows Spearman correlations for each individual pair and for the average signal from the pooled samples (for which the Pearson correlation was 0.876) represented in the correlation plot. (**e**) Representative IGV tracks at an arbitrary locus (Yae 1 domain containing 1; *YAE1D1*), illustrating concordance of H3K4me2 signals from FF samples analyzed by ChIP-seq (red) and matched FFPE samples analyzed by FiT-seq (blue) for six CRC specimens. The *y* axis represents number of reads per million per base pair (rbm).

notably, the resulting ChIP-seq signals showed extremely poor resolution of peaks, particularly outside of transcription start sites (TSSs or promoters), and gave broader and weaker signals than the same sites in matched FF tissues (**Supplementary Fig. 1a**,**b**). Low chromatin yields and poor signal resolution probably reflect excessive tissue fixation, a problem expected with archival tissue specimens. To examine the effects of fixation on ChIP-seq signals in a controlled fashion, we used mouse xenografts of MCF-7 breast cancer cells that were fixed for variable periods of time in 4% clinical-grade formaldehyde. In relation to the corresponding FF specimens, signal resolution was markedly reduced in samples that were fixed for 168 h (**Fig. 1a**,**b**), whereas samples that were fixed for 48 h gave results similar to those of their FF counterparts (**Supplementary Fig. 1c**), without any special processing. Only long fixation times mimicked the effects on chromatin we had observed in archival clinical specimens.

Because the DNA in FFPE samples is already fragmented, we reasoned that protein–protein cross-links may be primarily responsible for the broad peaks (**Supplementary Fig. 1a**). Although heat can reverse formaldehyde-induced cross-links, the temperatures used in heat-induced antigen retrieval for immunohistochemistry (IHC)[4,5] would also reverse protein–DNA cross-links, hence preventing effective ChIP. Because conservation of tissue architecture is dispensable for ChIP-seq[6], we reasoned that controlled protease treatment might help resolve the excessive number of cross-links introduced by long exposure to formaldehyde while preserving the chromatin complexes of greatest interest. Brief proteinase K (PK) digestion did not materially affect the IP of H3K4me2 from chromatin isolated from cultured cells (**Supplementary Fig. 1d**), and treatment of clinical FFPE specimens with PK at ambient temperature in a buffer containing 0.1% sodium dodecyl sulfate (SDS) substantially increased the yield of soluble chromatin.

PK treatment before sonication of tumor xenografts that were fixed for 168 h dramatically improved the resolution of signals from ChIP-seq as compared to that from PAT–ChIP-seq—signal levels now resembled those from matched FF tissue (**Fig. 1a**,**b**), thus validating a simple FiT-seq protocol in which PK treatment precedes sonication for the recovery of material for ChIP analyses. By using the MACS2 software to call peaks (using default parameters; Online Methods), we identified 49,964 peaks from FiT-seq data, as compared to 845 peaks from PAT–ChIP-seq data (**Supplementary Table 2**). Many more regions identified by FiT-seq overlapped with peaks identified in the matched FF tumor xenograft (**Supplementary Table 2**), and they showed high local sequence conservation (**Supplementary Fig. 1e**), a characteristic of high-confidence ChIP-seq sites for active histone marks[7]. We applied this methodology to a CRC sample and achieved consistent results with 5 or 10 min of PK digestion; we observed not only sharp ChIP-seq peaks, as expected for H3K4me2-marked histones but also considerably higher average signals as compared to PAT–ChIP-seq analysis done on the same sample (**Supplementary Fig. 1f**,**g**). Gently heating chromatin in the absence of a proteolysis step modestly improved signal resolution in FFPE samples (**Supplementary Fig. 1f**,**g**), and the best results came from heating samples at 40 °C for 1 h followed by brief PK digestion, resulting in our complete FiT-seq protocol (**Fig. 1c**). This protocol produced highly reproducible data in replicate experiments (Spearman correlation = 0.98; **Supplementary Fig. 1h**).

To determine how well FiT-seq matches the performance of ChIP-seq on FF tumor specimens, we examined six cases of CRC (CRC1–CRC6) in which some pieces of tumor had been frozen soon

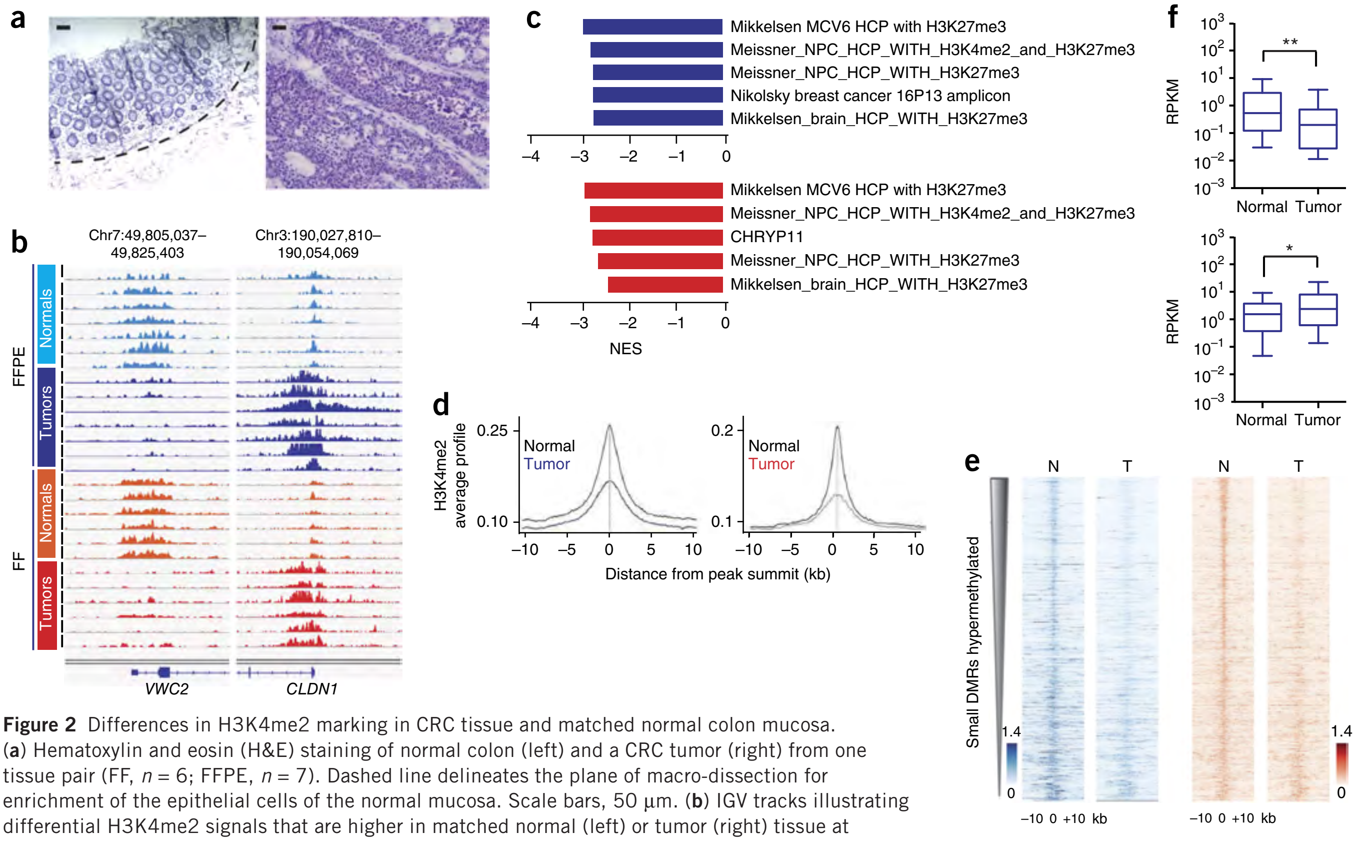

Figure 2 Differences in H3K4me2 marking in CRC tissue and matched normal colon mucosa. (**a**) Hematoxylin and eosin (H&E) staining of normal colon (left) and a CRC tumor (right) from one tissue pair (FF, $n = 6$; FFPE, $n = 7$). Dashed line delineates the plane of macro-dissection for enrichment of the epithelial cells of the normal mucosa. Scale bars, 50 µm. (**b**) IGV tracks illustrating differential H3K4me2 signals that are higher in matched normal (left) or tumor (right) tissue at representative TSSs from FiT-seq analysis of FFPE specimens ($n = 7$) and ChIP-seq analysis of FF specimens ($n = 6$). All y axis scales are 0–1 rbm. (**c**) GSEA of promoters with reduced H3K4me2 in tumors. Graphs show the most enriched gene sets and corresponding normalized enrichment scores (NES) derived from FF (red) or FFPE (blue) specimens. (**d**) Average distribution of H3K4me2 at 5,800 regions known to have DNA hypermethylation in CRCs, as compared to normal colon[12]. Signals pooled from seven FFPE tumor (blue) and seven matched FFPE normal colonic mucosa (black) samples (left) and from six FF tumor (red) and six matched FF normal colonic mucosa (black) samples (right) are shown with respect to DMR summits. (**e**) Heat maps of pooled H3K4me2 signals at the same 5,800 DMRs in **d** from normal (N) and tumor (T) samples for FFPE (FiT-seq; blue) (left) and FF (ChIP-seq; red) (right) pairs. Colored bars represent rbm. (**f**) Box plot (whiskers show the smallest and largest values) of mRNA levels (as determined by RNA-seq) associated with 100 promoters showing the highest loss (top) or gain (bottom) of H3K4me2 in CRC specimens, as compared to normal colon mucosa. $^{**}P = 0.0012$, $^{*}P = 0.01$; by the Mann–Whitney U-test.

after surgery, whereas other pieces had been preserved as FFPE blocks (**Supplementary Table 1**). H3K4me2 peaks identified either by FiT-seq in FFPE specimens or by ChIP-seq in FF specimens had comparable fractions that mapped to different genomic regions (promoters, 16% versus 22%; exons, 8% versus 8%; and intergenic regions 76% versus 70%; for FF and FFPE samples, respectively). Matched samples showed high correlation of H3K4me2 at both TSSs and distant enhancers (**Fig. 1d**,**e** and **Supplementary Fig. 1i**). Spearman correlations for pairwise comparisons of FiT-seq and FF ChIP-seq samples ranged from 0.902 to 0.959, with a Spearman correlation of 0.955 and a Pearson correlation of 0.876 for the average (pooled) signal. Thus, FiT-seq is a reliable method to map histone modifications in archived clinical specimens.

FiT-seq reveals biological features of CRC

To determine whether FFPE samples harbor the same underlying chromatin features as those in FF samples, we studied CRC tumor specimens and normal colonic mucosa that were removed from individuals during surgery, dissecting each under microscopic guidance to enrich for epithelial cells (**Fig. 2a**). Because normal tissues corresponding to the six frozen CRC specimens (CRC1–CRC6) were not available, we obtained six frozen tumor and normal samples from a different cohort of patients and then examined H3K4me2 in these FF samples (CRC9–CRC14) and in samples from seven FFPE cases (CRC2–CRC8); each of these 13 tumors was paired with its corresponding normal mucosa (**Supplementary Table 1**). The resulting data from all of the samples (**Fig. 2b**) showed that the median number of called peaks was comparable in FFPE and FF samples for both tumor (44,825 by FiT-seq and 56,683 by ChIP-seq) and normal mucosa (32,826 by FiT-seq and 29,667 by ChIP-seq). Differences in peak numbers between matched tumor and normal tissues were not significant in either set ($P < 0.05$ by Student's t-test). Thus, although individual sites may be marked differently in tumor and normal cells, the overall degree of H3K4me2 marking is similar.

To further compare CRCs and matched normal colonic mucosa, we extracted reads that mapped to within 2 kb of TSSs, using the HTSeq algorithm[8], and identified regions that were differentially marked with H3K4me2, using the DEseq package[9] (**Supplementary Table 3**; Online Methods). This approach yielded 714 and 1,140 differentially marked promoters ($P < 0.01$; fold change > 2; **Supplementary Table 3**) in the FiT-seq and ChIP-seq data sets, respectively. By considering all of the promoter regions that are marked more strongly with H3K4me2 in normal mucosa than in tumors, we found that gene set enrichment analysis (GSEA) showed striking commonality between FF and FFPE specimens, and that there was significant enrichment (normalized enrichment score (NES) = 3.2; false-discovery rate (FDR) = 0) for

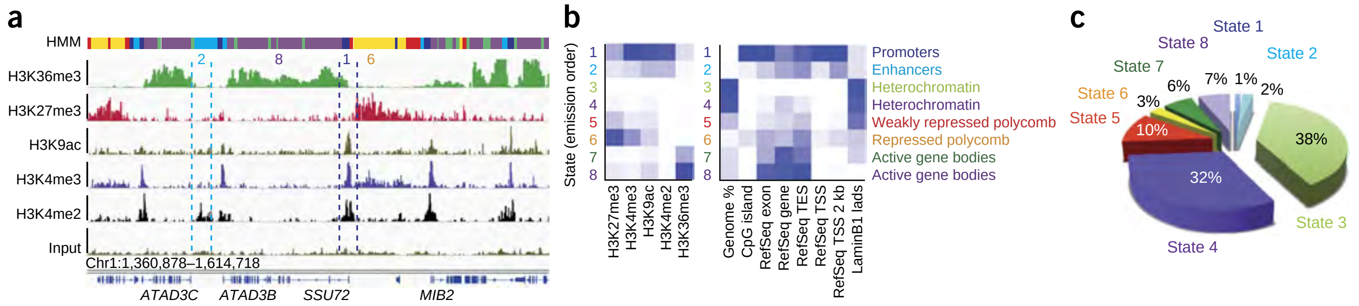

Figure 3 Additional histone marks mapped by FiT-seq. (**a**) IGV tracks at a representative region and hidden Markov modeling (HMM) of chromatin states (colored bars above the tracks) based on FiT-seq data from five distinct histone marks in an FFPE bladder cancer sample (*y* axis scale: 0–1 rbm). *ATAD3C*, ATPase family, AAA domain containing 3C; *ATAD3B*, ATPase family, AAA domain containing 3B; *SSU72*, SSU72 homolog, RNA polymerase II CTD phosphatase; *MIB2*, mindbomb E3 ubiquitin protein ligase 2. (**b**) Heat map showing the combinations of marks that delineate genomic regions into discrete, well-characterized states (left) and the association of each chromatin state with its presumed biological role, on the basis of enrichment in well-annotated genome regions characterized by RefSeq and the University of California, Santa Cruz (UCSC) hg19 genome (right) for the bladder cancer specimen in **a**. (**c**) Fractional representation of each chromatin state in the genome of the bladder cancer specimen in **a**.

gene sets that represent targets of Polycomb repressor 2 (PRC2) in embryonic stem cells[10] (**Fig. 2c** and **Supplementary Table 4**). These same promoters are known to have hypermethylated DNA in CRC and in other carcinomas[11].

To investigate this relationship further, we considered 5,800 differentially DNA-methylated regions (DMRs) that have been previously shown to have hypermethylated DNA in CRCs[12] and plotted the pooled H3K4me2 signal from the seven FFPE cases and the six FF cases (**Fig. 2d**; Online Methods). H3K4me2 signal density at DMRs was decreased at comparable levels in the pools from both the FF and FFPE tumor samples, as compared to those from the matched normal tissues (**Fig. 2d**,**e**). These regions are known to have reduced promoter H3K4me3 in CRC cell lines[13], and the effect that we observed is specific to CRC, as H3K4me2 signals from an FFPE sample of breast cancer were not decreased at the same promoters (**Supplementary Fig. 2a**,**b**). To further validate the specificity and accuracy of FiT-seq, we used a different group of 5,800 TSSs that does not overlap with the DMRs and found similar H3K4me2 density in CRCs and normal colonic mucosa in both the FFPE and FF specimens (**Supplementary Fig. 2c**).

For promoters showing greater H3K4me2 signals in tumors than in paired normal mucosa, both FFPE and FF samples showed similar enrichment for gene sets, such as those involved in cell cycle (KEGG_CELL_CYCLE and REACTOME_CELL_CYCLE), as well as multiple sets representing transcriptional deregulation in primary CRCs as compared to normal mucosa (SABATES_COLORECTAL_ADENOMA_UP and GRADE_COLON_AND_RECTAL_CANCER_UP; **Supplementary Table 4**). Thus, enrichment for key gene sets that might be expected in most CRCs was as evident by FiT-seq on FFPE samples as by conventional ChIP-seq on FF tumors.

To determine whether the changes in histone marks that were detected by FiT-seq mirror alterations in gene expression, we analyzed tissue sections from the same FFPE tissue blocks by RNA-seq. These RNA-seq data highly correlated with RNA-seq data from an independent collection of fresh-frozen CRC samples in The Cancer Genome Atlas[14] (**Supplementary Fig. 2d**). Genes with the largest reduction in promoter H3K4me2 abundance in tumor FFPE specimens were expressed at significantly lower levels than in the normal mucosa, whereas genes with the largest increases in promoter H3K4me2 were expressed at higher levels (**Fig. 2f**).

Most *cis*-regulatory activity occurs through distant enhancers[15], which are challenging to unambiguously assign to genes. In lieu of such assignments, we examined DNA sequence motifs that were enriched among distant elements (>2 kb from TSSs) and that showed stronger H3K4me2 marking in tumors than in normal colon mucosa. Analyses using both FFPE and FF specimens showed that the sequence motif recognized by the transcription factor AP1 was the most significantly enriched in tumors, consistent with the role of this transcription factor in the response to epidermal growth factor (EGF) signaling in CRC. Moreover, nine of the top ten motifs identified in the FF sample cohort were also seen to be significantly enriched in samples from the FFPE cohort, including motifs that have well-known associations with intestinal epithelial function, such as those recognized by hepatocyte nuclear factor (HNF) 1, HNF4, T cell factor (TCF) 3 and TCF4 (**Supplementary Table 5**). Thus, FiT-seq analysis of clinical CRC specimens accurately maps histone marks that show robust correlations with hypermethylated promoters, aberrant gene activity and enrichment of transcription factor motifs.

FiT-seq analysis produces chromatin-state maps

To extend the utility of FiT-seq beyond the analysis of H3K4me2, we mapped additional and informative histone marks that established consortia (such as the Encyclopedia of DNA Elements (ENCODE) Consortium and the NIH Roadmap Epigenomics Mapping Consortium) have used to characterize epigenomes. Using the FiT-seq protocol without any additional modifications, we obtained data for the presence of the repressive H3K27me3 mark and the activation-associated H3K4me3 and acetylated lysine 9 of histone H3 (H3K9ac) marks at promoters, the H3K36me3 along gene bodies and the H3K4me1 mark at enhancer elements. The results using specimens from individuals with bladder cancer (**Fig. 3a**) and CRC (**Supplementary Fig. 3a**) showed the expected distribution of these marks with respect to known genomic features. The bladder cancer sample showed the predicted inverse correlation between H3K27me3 and H3K36me3 (**Fig. 3a**), and the CRC sample showed intergenic regions that could similarly be identified by the presence of H3K4me2 and H3K4me1 (**Supplementary Fig. 3a**). We used the ChromHMM algorithm[16] to analyze this data-rich genome-wide information from a bladder cancer to delineate chromatin states that corresponded to well-characterized regions, including promoters (state 1—overlap of H3K4me3, H3K4me2 and H3K9ac), enhancers (state 2—marked by H3K4me2 or H3K4me1, but not by H3K4me3), active gene bodies (states 7 and 8—marked by H3K36me3), Polycomb-repressed genes (states 5 and 6—marked by H3K27me3) and areas lacking

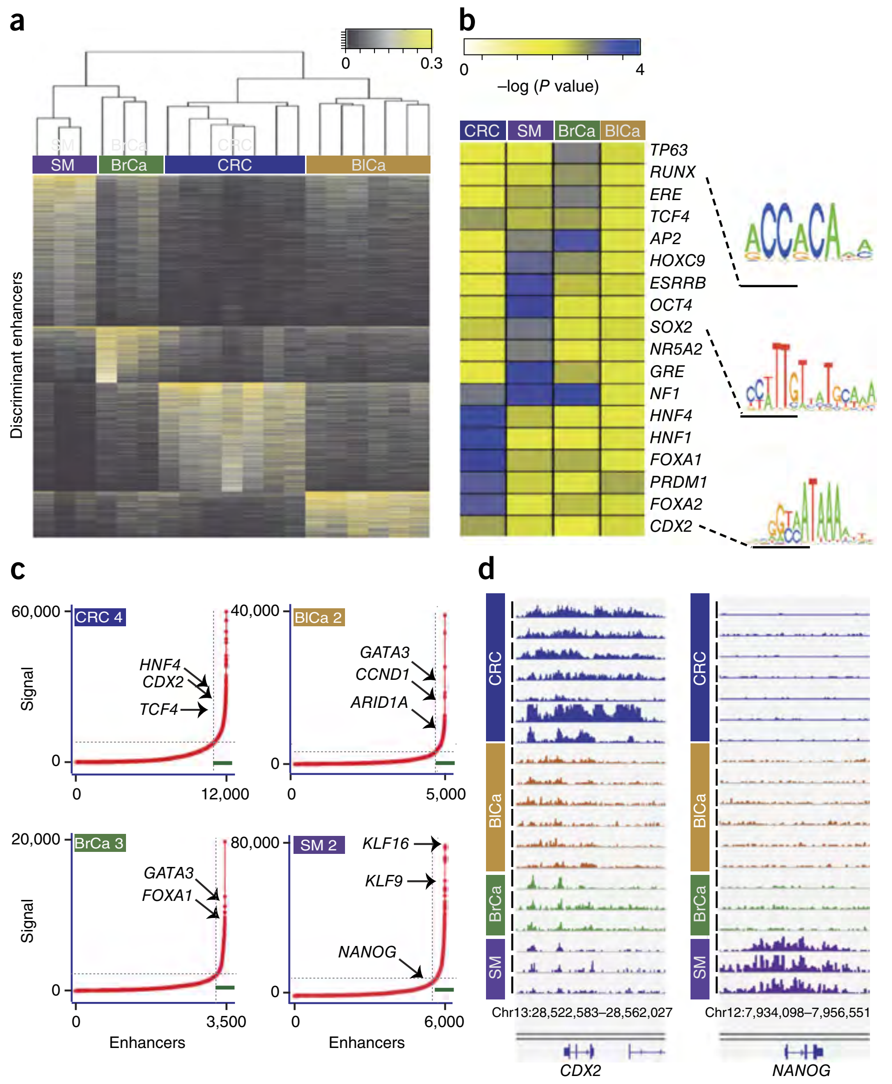

Figure 4 Tumor-specific enhancers and H3K4me2-defined superenhancers identified in multiple cancer types. (**a**) Unsupervised hierarchical clustering of genome-wide H3K4me2 signals for the top 25% most-variable peaks located >2 kb from TSSs (presumptive enhancers) across FFPE samples from four different human tumor types—seminoma (SM), breast cancer (BrCa), bladder cancer (BlCa) and CRC. (**b**) Heat map of *P* values (as determined by hypergeometric test) for representative transcription-factor-binding motifs enriched at the centers of enhancers that are associated specifically with each tumor type. The sequence logos for three representative motifs are shown. (**c**) Signal distribution at H3K4me2-marked enhancers detected in individual representative FFPE samples of each tumor type; superenhancers for selected genes, as defined by this mark, are indicated by arrows. The green bars identify the subset of enhancers known as superenhancers. (**d**) IGV tracks showing H3K4me2-defined superenhancers near the *CDX2* and Nanog homeobox (*NANOG*) loci in CRC and SM samples, respectively. Scale for all tracks, 0–1 rbm.

any of these signals, which mainly represent heterochromatin (states 3 and 4) (**Fig. 3b**). The chromatin states inferred from the combinations of different histone marks are enriched in regions that are expected on the basis of external annotations, and the fractions of each state in the genome are compatible with those typically observed in tissues[17] (**Fig. 3b**,**c** and **Supplementary Fig. 3b**). For example, CpG islands are enriched at promoters (state 1), and genes annotated in the National Center for Biotechnology Information (NCBI) Reference Sequence database (RefSeqGenes) but not TSSs are enriched at active gene bodies (states 7 and 8), which emphasizes accurate identification of boundaries for individual histone marks. Thus, informative chromatin states are adequately preserved in archived FFPE tissues, and FiT-seq captures genome-wide occurrence of histone marks representing diverse regions and activities.

FiT-seq identifies tumor-type-specific enhancers and superenhancers

To demonstrate the application of FiT-seq to additional cancer types, we mapped H3K4me2-marked enhancers in cases of estrogen receptor (ER)-positive breast cancer, non-muscle-invasive (stage T1) bladder cancer and seminomas (**Supplementary Table 1**). Across all 19 FFPE samples, including the seven CRCs studied above, an unsupervised analysis of marked regions located more than 2 kb away from known TSSs readily separated tumors by the tissue of origin (**Fig. 4a** and **Supplementary Fig. 4a**). Thus, the distinctive epigenomes of primary tumors are preserved in archived specimens and enhancer analysis by FiT-seq could have value in defining the origins of occult tumors. Additionally, analysis of variance (ANOVA) analysis identified the most characteristic enhancers in each tumor type (**Supplementary Table 6**), and sequence motifs enriched in each group bind transcription factors known to function in the corresponding tissue (**Fig. 4b** and **Supplementary Table 7**). In particular, HNF proteins, caudal type homeobox 2 (CDX2) and TCF4 are instrumental in both intestinal-lineage identity and CRC pathogenesis[18–20], and they are highly enriched in CRC-specific enhancers, whereas enhancers specific to ER$^+$ breast cancer are enriched for ER and Runt-related transcription factor (RUNX) motifs[21]. By contrast, FFPE-preserved seminoma specimens showed enrichment of motifs for OCT4 and SOX2, core embryonic stem cell factors that are also expressed in germ cell tumors[22], as well as for ESRRB and NR5A2, which maintain pluripotency[23,24]. Early-stage bladder cancers showed the fewest distinctive enhancers, with consequently little enrichment of sequence motifs.

Finally, tissue-specific superenhancers have been defined in cancers and normal tissues on the basis of particularly extensive H3K27ac and H3K4me1 marks and the binding of the general factor mediator complex subunit 1 (MED1)[25]. Because H3K4me2 has been used for enhancer identification[3,26,27], we used FiT-seq data of this mark in diverse tumor types as a proxy to identify superenhancers using the published algorithm[25]. The results showed the same highly skewed distribution of ChIP-seq signal intensities (**Fig. 4c**) that are typically seen with related histone marks, such as H3K4me1 and H3K27ac[25], where the levels of H3K4me2 at sites beyond the inflection point substantially exceeded those at other enhancers, often by an order of magnitude. This approach identified hundreds of H3K4me2-defined candidate superenhancers in each tumor type (**Fig. 4c**,**d** and **Supplementary Table 8**). An abundance of these superenhancers coincided with lineage-specific genes for each tumor type, including *CDX2* in CRC and GATA binding protein 3 (*GATA3*) in breast cancer[28] (**Fig. 4d** and **Supplementary Fig. 4b**), both of which are known to harbor classical superenhancers in cancer cell lines[25]. Similarly, the loci encoding Krüppel-like factors, which help to maintain pluripotent stem cells[29], showed superenhancer features in every seminoma examined (**Supplementary Fig. 4b**). Although no sequence motif was significantly enriched among typical (non-super-) enhancers in bladder cancer (**Fig. 4b**), four of the six cases showed superenhancers near key bladder-cancer-associated oncogenes, such as *GATA3* and cyclin D1 (*CCND1*); the top-ranked H3K4me2-defined superenhancer in one case mapped to fibroblast growth factor 3 (*FGFR3*) (**Supplementary Fig. 4b**), a driver oncogene in ~12% of bladder cancers[30]. Thus, FiT-seq can identify tumor-specific enhancers in fixed tissues from a range of human cancers and provide insight into their key transcriptional regulators.

DISCUSSION

FFPE tissue samples represent an enormous source of material for informative epigenome analyses, but they typically resist extraction of soluble chromatin, an effect we could mimic best with long (168 h) formaldehyde fixation times in xenografts of human cancer cell lines. The FiT-seq protocol described here is simple, requiring heat treatment followed by brief, controlled proteolytic digestion of FFPE tissue slices to markedly increase both the strength and resolution of histone ChIP-seq signals. The processing did not materially affect H3K4me2 ChIP-seq signals, either in raw chromatin or on material processed for FiT-seq, and archive duration was not a limiting factor; samples stored for 10 years gave signals comparable to those of more recently preserved specimens. In contrast to the PAT–ChIP-seq method, resolution of FiT-seq signals was comparable to those from FF samples, and soluble chromatin released by FiT-seq can be used to interrogate various histone marks that cover major classes of genomic regions.

FiT-seq can identify tissue- and tumor-specific chromatin states and enhancers in at least four tumor types. In CRC specimens in particular, FiT-seq analysis identified tumor-specific epigenome reprogramming and revealed an inverse association between H3K4me2 and DNA hypermethylation at hundreds of genes that show reduced expression in CRCs as compared to normal colon mucosa. Although this relationship between H3K4me2 and DNA hypermethylation has been previously reported in various contexts[13], ours is the first demonstration of this effect in pairs of primary CRCs and matched normal tissues that were preserved as FF or FFPE samples. These data highlight the interplay of DNA hypermethylation and reduced promoter H3K4me2 as epigenetic processes in CRC; whereas DMRs are not globally associated with altered gene expression[12], the highest losses and gains in promoter H3K4me2 associate with corresponding changes in gene expression. FiT-seq also revealed tumor-specific enhancers and H3K4me2-defined superenhancers that distinguished cancers by their tissue of origin, making it a method that might be used an alternative to transcriptional profiling in classifying tumors of uncertain origin. It is, however, unclear whether such kinds of superenhancers represent epigenome features specific to the tumor or to the tissue.

FiT-seq has the potential to extend the current understanding of cancer epigenomes and, combined with DNA and RNA analysis of the same samples, to determine how specific mutations affect functional chromatin states. Such studies, previously limited to the few tumors frozen selectively for research purposes, can now be extended to the vast number of clinical specimens that are preserved as FFPE blocks. FiT-seq may also help to decipher the functions of alleles that are associated with an increased risk of disease, as determined in genome-wide association studies, most of which are located in noncoding regions. Finally, FiT-seq analysis can contribute to the understanding of tumors that lack clear genetic drivers, the identification of novel cancer subtypes and the development of clinical biomarkers.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene Expression Omnibus: data from this study are deposited under accession code GSE73398.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by the DFCI–Novartis Drug Discovery Program (R.A.S.), the US NIH grant P50CA127003 (R.A.S.), the Susan F. Smith Center for Women's Cancers (M. Brown), grants from the Susan G. Komen and the Breast Cancer Research foundations (M. Brown), the Medical Oncology Translational Grant Program from the Dana-Farber Cancer Institute (C.J.S.), a grant from Instituto de Salud Carlos III the Spanish Economy and Competitiveness Ministry (grant PI13-01818; P.C.), the ConSEPOC–Comunidad de Madrid (grant S2010/BMD-2542; P.C.), and fellowships from the Asociación Española Contra el Cáncer (Programa de Formación Avanzada en Oncología 2010; P.C.) and the Fundación Caja Madrid (P.C.). We thank E. Díaz and M. de Miguel for technical assistance.

AUTHOR CONTRIBUTIONS

P.C., R.A.S., M. Brown and H.W.L. conceived and designed the study; L.L., P.C., H.W.L. and N.K.O'N. performed computational and statistical analyses; M.D., P.R., M. Bowden and C.W.Z. acquired data; E.B. and M.M. reviewed CRC pathology; J.F., J.M.-R., H.G., V.M., S.M., J.B., D.G.-O., C.J.S., M.H. and A.R. supervised collection of tumor samples; X.S.L., M. Brown, H.W.L. and R.A.S. provided overall supervision; and P.C., R.A.S. and H.W.L. drafted the manuscript, with input from all of the authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17–37 (2013).
2. Fanelli, M. *et al.* Pathology tissue–chromatin immunoprecipitation, coupled with high-throughput sequencing, allows the epigenetic profiling of patient samples. *Proc. Natl. Acad. Sci. USA* **107**, 21535–21540 (2010).
3. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
4. Shi, S.R., Imam, S.A., Young, L., Cote, R.J. & Taylor, C.R. Antigen retrieval immunohistochemistry under the influence of pH using monoclonal antibodies. *J. Histochem. Cytochem.* **43**, 193–201 (1995).
5. Shi, S.R., Liu, C., Perez, J. & Taylor, C.R. Protein-embedding technique: a potential approach to standardization of immunohistochemistry for formalin-fixed, paraffin-embedded tissue sections. *J. Histochem. Cytochem.* **53**, 1167–1170 (2005).
6. Yamashita, S. Heat-induced antigen retrieval: mechanisms and application to histochemistry. *Prog. Histochem. Cytochem.* **41**, 141–200 (2007).
7. Shin, H., Liu, T., Duan, X., Zhang, Y. & Liu, X.S. Computational methodology for ChIP-seq analysis. *Quant. Biol.* **1**, 54–70 (2013).
8. Anders, S., Pyl, P.T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
9. Anders, S. & Huber, W. Differential expression analysis for sequence-count data. *Genome Biol.* **11**, R106 (2010).
10. Lee, T.I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301–313 (2006).
11. Ben-Porath, I. *et al.* An embryonic stem-cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat. Genet.* **40**, 499–507 (2008).
12. Hansen, K.D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43**, 768–775 (2011).

13. Balasubramanian, D. *et al.* H3K4me3 inversely correlates with DNA methylation at a large class of non-CpG-island-containing start sites. *Genome Med.* **4**, 47 (2012).
14. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
15. Carroll, J.S. *et al.* Genome-wide analysis of estrogen-receptor-binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).
16. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
17. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
18. Verzi, M.P. *et al.* TCF4 and CDX2, major transcription factors for intestinal function, converge on the same *cis*-regulatory regions. *Proc. Natl. Acad. Sci. USA* **107**, 15157–15162 (2010).
19. Verzi, M.P. *et al.* Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Dev. Cell* **19**, 713–726 (2010).
20. Hatzis, P. *et al.* Genome-wide pattern of TCF7L2 (TCF4) chromatin occupancy in colorectal cancer cells. *Mol. Cell. Biol.* **28**, 2732–2744 (2008).
21. van Bragt, M.P., Hu, X., Xie, Y. & Li, Z. RUNX1, a transcription factor mutated in breast cancer, controls the fate of ER-positive mammary luminal cells. *eLife* **3**, e03881 (2014).
22. Santagata, S., Ligon, K.L. & Hornick, J.L. Embryonic stem cell transcription factor signatures in the diagnosis of primary and metastatic germ cell tumors. *Am. J. Surg. Pathol.* **31**, 836–845 (2007).
23. Zhang, X., Zhang, J., Wang, T., Esteban, M.A. & Pei, D. Esrrb activates Oct4 transcription and sustains self-renewal and pluripotency in embryonic stem cells. *J. Biol. Chem.* **283**, 35825–35833 (2008).
24. Heng, J.C. *et al.* The nuclear receptor Nr5a2 can replace Oct4 in the reprogramming of murine somatic cells to pluripotent cells. *Cell Stem Cell* **6**, 167–174 (2010).
25. Hnisz, D. *et al.* Superenhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
26. Zentner, G.E. & Scacheri, P.C. The chromatin fingerprint of gene enhancer elements. *J. Biol. Chem.* **287**, 30888–30896 (2012).
27. Ong, C.T. & Corces, V.G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12**, 283–293 (2011).
28. Kouros-Mehr, H., Slorach, E.M., Sternlicht, M.D. & Werb, Z. GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell* **127**, 1041–1055 (2006).
29. Jiang, J. *et al.* A core Klf circuitry regulates self-renewal of embryonic stem cells. *Nat. Cell Biol.* **10**, 353–360 (2008).
30. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).

ONLINE METHODS

Clinical material and xenografts. Six pairs of frozen CRC specimens and their corresponding normal colonic mucosa, and FFPE samples from three cases each of breast cancer and seminoma (**Supplementary Table 1**), were obtained from collections at the Brigham and Women's Hospital, Boston, MA (USA). Seven pairs of CRCs and matched normal colon mucosa, preserved as FFPE and matched FF specimens for six of these seven cases, were obtained from Hospital La Paz, Madrid (Spain). Six FFPE samples of bladder cancer were obtained from Hospital del Mar, Barcelona (Spain) and Xarxa de Bancs de Tumors sponsored by Pla Director d'Oncologia de Catalunya. The following committees approved the research on human tissues: Institutional Review Board, Dana-Farber and Harvard Cancer Center, Boston, MA; Institutional Review Board, Hospital del Mar, Barcelona, Spain; Ethics Committee for Clinical Investigation, La Paz University Hospital, Madrid, Spain. All tissues were stained with hematoxylin and eosin (H&E) and reviewed by pathologists to confirm the histologic diagnosis. Mouse xenografts of MCF-7 breast cancer cells were fixed for 24 or 168 h in 4% buffered formalin before further processing to prepare FFPE blocks. The cells tested negative for mycoplasma before starting the xenograft.

Fixed-tissue chromatin immunoprecipitation (FiT-seq). We used microscopy-guided macrodissection to obtain >70% epithelial enrichment of colorectal tumor and normal mucosal specimens. Ten sections of 10-µm thickness for FFPE and FF tumor specimens and 20 sections of macrodissected normal mucosa were used for ChIP analyses. FFPE tissue sections were washed with xylene three times to remove all traces of paraffin and then rehydrated in an ethanol/water series (95/5, 80/20, 70/30, 50/50, 20/80, 0/100). The samples were then heated for 1 h at 40 °C in a 0.1% SDS buffer (50 mM Tris-HCl, 10 mM EDTA). Samples were subsequently treated with 40 ng/µl proteinase K (Roche) for 5 or 10 min, followed by protease inactivation with 2 µg/ul of the serine protease inhibitor AEBSF (AG Scientific). The resulting material was sonicated for 40 min in 0.1% SDS buffer in a Covaris E210 instrument (duty cycle: 20%; intensity: 8; cycle burst: 200). Soluble chromatin was quantified fluorometrically using a Q-bit (ThermoFisher Scientific), and ChIP was performed using agarose bead columns (Active Motif) with 10 µg of antibodies against H3K4me2 (Millipore 07-030), H3K4me1 (Diagenode C15410194), H3K4me3 (Diagenode C15410003), H3K36me3 (Abcam ab9050), H3K27me3 (Cell Signaling C36811), and H3K9ac (Abcam ab4441) in 1 ml total volume. DNA libraries were prepared using ThruPLEX-FD kits (Rubicon Genomics) using the manufacturer's protocols and sequenced on a Nextseq 500 instrument (Illumina). For PAT–ChIP-seq we followed the protocol as previously published[2]. Briefly, the key steps in the protocol included a deparaffinization and rehydration step followed by 30 min RNase A (Sigma-Aldrich) treatment and a micrococcal nuclease digestion (0.1 U per 1ug of chromatin) (Sigma-Aldrich) step before sonication. The IP was done using Sepharose beads (Active Motif) using the H3K4me2 antibody (Millipore 07-030). Library preparation and sequencing was done as described above.

ChIP on fresh-frozen (FF) tissue specimens. For nucleosome-resolution ChIP-seq (samples CRC9–CRC14 and CM9–CM14), FF sections were washed in phosphate-buffered saline (PBS) and treated with 0.2 U micrococcal nuclease (Sigma-Aldrich) for 8 min at 37 °C. Alternatively (for all other frozen samples), sections were washed in PBS, cross-linked with 1% formaldehyde for 10 min at RT and quenched with 125 mM glycine for 5 min, at ambient temperature. The cross-linked material was subsequently sonicated for 40 min on a Covaris E-210. Soluble chromatin was immunoprecipitated with 10 µg H3K4me2 antibody (Millipore 07-030) as described previously[19]. DNA libraries were prepared using ThruPLEX-FD kits (Rubicon Genomics) and sequenced on a Nextseq 500 instrument (Illumina).

Statistical analyses. *ChIP-seq.* Sequenced DNA tags were aligned with Bowtie against the reference genome Hg19 and peaks were called using MACS v2.0.10 (https://github.com/taoliu/MACS) with default cutoffs ($q < 0.01$). Numbers of unique reads and called peaks are tabulated in **Supplementary Table 1**. Enhancers were identified as H3K4me2-marked regions located >2 kb upstream or >2 kb downstream from TSSs. Big wiggle files and conservation plots were also generated by MACS.

Wiggle tracks were visualized using the integrative genomics viewer[31]. Heat maps were prepared using deepTools[32]. Conservation plots were obtained with the Conservation Plot (version 1.0.0) tool available in Cistrome[33]. Functional annotation of H3K4me2 regions was performed using GREAT[34], using the default association rule setting (basal plus extension; proximal 5 kb upstream 1 kb downstream, plus distal: up to 1,000 kb) and the whole human genome as background. All motif analysis was done with the HOMER v4.7.2. software[35] (http://homer.salk.edu/homer), using default parameters. Gene set enrichment analysis was performed with GSEA software[36] against the MSigDB database of gene signatures. We used ChromHMM[16] to perform hidden Markov modeling and, by default, chromatin states were analyzed at 200-bp intervals and a fold threshold of 10.

To compare CRCs and matched normal colonic mucosa, reads found within 2 kb of TSSs were extracted using HTSeq, and differential regions were identified by the DEseq package[9]. For the distal enhancer comparison we considered the union of all peaks from each sample type (FF normal and tumor; FFPE normal and tumor). For a comparison across cancer types, H3K4me2 peaks from all data sets were merged to create a union set and then filtered to exclude peaks present at TSSs (retaining peaks found in intergenic regions). Normalized counts of bases found in this union set, in bases per million reads per base, were extracted using bamliquidator (https://github.com/BradnerLab/pipeline/wiki/bamliquidator).

The samples were arranged by unsupervised hierarchical clustering in R using the Spearman correlations of the top 25% most-variable peaks (**Supplementary Fig. 4**). In **Figure 4a**, the features on the y axis were quantile-normalized and clustered by k-means in R. To identify enhancers associated specifically with each tumor type in **Figure 4b**, a union set of intergenic peaks over all tumor types was created, read counts were extracted using HTSeq and the software DEseq was used to identify intergenic peaks that were differentially marked in each tumor type relative to the other three types combined. These differential intergenic peaks were then input into the software HOMER for motif analysis.

RNA-seq. Two sections of 10 µm for FFPE tumors and four sections for the corresponding macrodissected normal mucosa were used. FFPE sections were deparaffinized and rehydrated, and total RNA was extracted using Allprep FFPE kit (Qiagen, Hilden, Germany). RNA concentration was measured by the Quant-iT RiboGreen RNA assay (ThermoFisher), and quality was assessed on an Agilent 2100 Bioanalyzer, using the Agilent RNA 6000 Nano kit. Ribosomal and mitochondrial RNA were removed using biotinylated, target-specific oligonucleotides combined with Ribo-Zero rRNA removal beads and the TruSeq Stranded Total RNA kit (Illumina). Libraries were quantitated by Picogreen assay (Life Technologies) and library quality control was performed using the Agilent High-sensitivity DNA kit (Agilent). Single-end 75-bp sequences were obtained on a NextSeq 500 Instrument (Illumina). Sequence tags were mapped to the reference genome Hg19 using TopHat v2.0.6, and transcript levels were calculated on data from triplicate samples as fragments per kb per 10^6 mapped reads (FPKM) using Cufflinks v2.0.2. Differential expression was determined with CuffDiff, using Chi-square tests with 1 degree of freedom and two-tailed P values to assess statistical significance[37]. Normalized RNA-seq data from other CRCs were obtained from TCGA (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) and used to assess correlation with RNA-seq data generated from FFPE specimens.

31. Robinson, J.T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
32. Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191 (2014).
33. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).
34. McLean, C.Y. *et al.* GREAT improves functional interpretation of *cis*-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
35. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
36. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
37. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578 (2012).