nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-020-00775-6

ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells of origin

Ronen Sadeh[1,2,12], Israa Sharkia[1,2,12], Gavriel Fialkoff[1,2], Ayelet Rahat[2], Jenia Gutin[1,2], Alon Chappleboim[1,2], Mor Nitzan[1], Ilana Fox-Fisher[3], Daniel Neiman[3], Guy Meler[1], Zahala Kamari[1,2], Dayana Yaish[4], Tamar Peretz[5], Ayala Hubert[5], Jonathan E. Cohen[5,6], Azzam Salah[5], Mark Temper[5], Albert Grinshpun[5], Myriam Maoz[5], Samir Abu-Gazala[7], Ami Ben Ya'acov[8], Eyal Shteyer[8], Rifaat Safadi[9], Tommy Kaplan[1], Ruth Shemer[3], David Planer[10], Eithan Galun[4], Benjamin Glaser[11], Aviad Zick[5], Yuval Dor[3] and Nir Friedman[1,2] ✉

Cell-free DNA (cfDNA) in human plasma provides access to molecular information about the pathological processes in the organs or tumors from which it originates. These DNA fragments are derived from fragmented chromatin in dying cells and retain some of the cell-of-origin histone modifications. In this study, we applied chromatin immunoprecipitation of cell-free nucleosomes carrying active chromatin modifications followed by sequencing (cfChIP-seq) to 268 human samples. In healthy donors, we identified bone marrow megakaryocytes, but not erythroblasts, as major contributors to the cfDNA pool. In patients with a range of liver diseases, we showed that we can identify pathology-related changes in hepatocyte transcriptional programs. In patients with metastatic colorectal carcinoma, we detected clinically relevant and patient-specific information, including transcriptionally active human epidermal growth factor receptor 2 (HER2) amplifications. Altogether, cfChIP-seq, using low sequencing depth, provides systemic and genome-wide information and can inform diagnosis and facilitate interrogation of physiological and pathological processes using blood samples.

Blood contains cfDNA fragments derived from dying cells[1]. cfDNA has a half-life of ~15 min[2] and, therefore, represents events that occurred close to sampling time. cfDNA analysis is used for assessment of fetus chromosomal aberrations, graft rejection, monitoring tumor dynamics and targeted treatment[3–7]. These applications rely on genetic differences between the host and the tissue of interest. Analysis of CpG methylation in cfDNA is emerging as an alternative independent of genetic alteration[5,8–11]. CpG methylation profiles are determined during differentiation and are stable afterwards and, thus, are highly informative about cell identity (for example, liver or lung). However, genetic and methylation-based approaches do not report on recent transcriptional events, as mutations and methylation changes occur over developmental time scales.

The basic repeating unit of chromatin is the nucleosome, which is a histone-DNA complex encompassing ~150 base pairs (bp) of DNA[12]. Histone proteins are subject to multiple covalent modifications, which are involved in nearly all aspects of messenger RNA (mRNA) biogenesis[13–16]. Histone modification patterns reflect recent events related to chromatin regulation and activity of RNA polymerase[13,15], and different combinations of such modifications mark the location and activity of non-coding regions, enhancers, promoters and gene bodies[17–22]. Chromatin immunoprecipitation and sequencing (ChIP-seq) enables genome-wide mapping of histone modifications and provides detailed understanding of the regulatory activity within cells[17–19,23–27].

Upon cell death, the genome is fragmented, and chromatin, mostly in the form of nucleosomes, is released into the circulation as cell-free nucleosomes (cf-nucleosomes)[28–30] that retain some histone modifications[31–33]. We reasoned that capturing and DNA sequencing of modified nucleosomes from plasma might inform on DNA-related activities, including transcription, within the cells of origin (Fig. 1a). This currently inaccessible epigenetic information extends beyond cfDNA modalities examined to date[4–11,34–43].

In this study, we performed chromatin immunoprecipitation and sequencing of cell-free nucleosomes directly from human plasma (cfChIP-seq). We show that cfChIP-seq recapitulates the original genomic distribution of modifications associated with transcriptionally active promoters, enhancers and gene bodies, demonstrating that plasma nucleosomes retain the epigenetic information of their

[1]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel. [2]The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. [3]Institute for Medical Research Israel-Canada, The Hebrew University-Hadassah Medical School, Jerusalem, Israel. [4]The Goldyne Savad Institute for Gene Therapy, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [5]Sharett Institute of Oncology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [6]The Wohl Institute for Translational Medicine, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [7]Department of Surgery, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [8]The Juliet Keidan Institute of Pediatric Gastroenterology Institute, Shaare Zedek Medical Center, Jerusalem, Israel. [9]The Liver Unit, Institute of Gastroenterology and Liver Diseases, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [10]Department of Cardiology, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [11]Dept of Endocrinology and Metabolism Service, Hadassah-Hebrew University Medical Center, Jerusalem, Israel. [12]These authors contributed equally: Ronen Sadeh, Israa Sharkia. ✉e-mail: nir.friedman@mail.huji.ac.il

cells of origin. We applied cfChIP-seq to ~250 samples from more than 100 individuals, including 61 self-declared healthy donors; four patients with acute myocardial infarction (AMI); 29 patients with autoimmune, metabolic or viral liver diseases; and 56 patients with metastatic colorectal carcinoma (CRC). We identified bone marrow megakaryocytes, but not erythroblasts, as major contributors to the cfDNA pool in healthy donors. We show pathology-related changes in hepatocytes chromatin and connect it to changes in transcriptional programs in these cells. In patients with CRC, we detect the disease with high sensitivity and demonstrate that cfChIP-seq can identify subgroups of patients with distinct cancer-related transcriptional programs and with potential implications for diagnosis and treatment.

Results

ChIP-seq of cf-nucleosomes from plasma. We devised a protocol for cf-nucleosome ChIP-seq from <2 ml of plasma (Methods) that overcomes the extremely low concentration of cf-nucleosomes and high concentration of native antibodies in plasma (Fig. 1a,b). Briefly, we covalently immobilized ChIP antibodies to paramagnetic beads, which can be incubated directly in plasma, avoiding competition with native antibodies. Additionally, we used an on-bead adaptor ligation[26,44–46], where barcoded sequencing DNA adaptors are ligated directly to chromatin fragments before the isolation of DNA.

We performed cfChIP-seq on multiple plasma samples from healthy individuals with antibodies targeting marks of accessible/active promoters (H3K4me3 or H3K4me2), enhancers (H3K4me2 or H3K4me1) and gene body of actively transcribed genes (H3K36me3) (Fig. 1c). cfChIP-seq profiles with different antibodies show the expected patterns (Fig. 1c,d and Extended Data Fig. 1a).

Several lines of evidence suggest that cfChIP-seq is highly specific. 1) cfChIP-seq signal is consistent with reference ChIP-seq in tissues[25], evident by the agreement of peaks (Fig. 1c and Extended Data Fig. 1b), in the average pattern around promoters and enhancers (Fig. 1d and Extended Data Fig. 1c) and in quantitative comparison of the signal across multiple genomic locations, such as all promoters, ($r > 0.8$; Fig. 1e and Extended Data Fig. 1d). Essentially all promoters that are ubiquitously marked by H3K4me3 (housekeeping) in reference ChIP-seq are enriched for this mark in cfChIP-seq (9,795/10,505 promoters, $P < 10^{-1,000}$). Focusing on non-housekeeping gene promoters, there is significant overlap (1,324/2,311 promoters, $P < 10^{-288}$) with promoters from monocytes and neutrophils that are the major contributors to the cfDNA pool[5,11] (Fig. 1f). 2) Performing cfChIP-seq with a mock antibody resulted in dramatically lower yield (Supplementary Table 1). 3) The level of non-specific reads are mostly similar to or lower than standard ChIP-seq (Methods and Extended Data Fig. 1e).

Several avenues of evidence rule out the possibility that the cfChIP-seq signal is derived from in-tube lysis during sample handling. 1) We identified 676 promoters carrying H3K4me3 that are absent in ChIP-seq from white blood cells (leukocytes; Fig. 1f). These include promoters of genes that are expressed specifically in bone marrow residing megakaryocytes (below). 2) Fragment size distributions of cfChIP-seq correspond to DNA wrapped around mono- and di-nucleosomes (Fig. 1g and Extended Data Fig. 2a), consistent with apoptotic or necrotic cell death, but not with cell lysis, which results in much larger (>10-kb) fragments[47]. 3) In patients, we detect disease-related chromatin from remote tissues, including heart, liver and colon (below).

Together, these results strongly suggest that cfChIP-seq assays cf-nucleosomes originating from cells that have died in vivo and preserved the endogenous patterns of active histone methylation marks within them.

cfChIP-seq detects pathology-related origin of cf-nucleosomes. We find that self-reported healthy donors show highly similar cfChIP-seq profiles (Extended Data Fig. 2b and Supplementary Note). We contrasted these to samples from a patient with metastatic CRC, in whom a large fraction of the cfDNA was expected to be of tumor origin[11,34] (Supplementary Table 1). For the CRC sample, we observed many regions showing statistically significant increases in H3K4me3 (1,562 regions), H3K4me2 (2,473 regions) and H3K36me3 (5,122 regions) (Methods and Supplementary Table 2). Genes associated with these regions include several classic CRC markers, such as *CCAT1* (colorectal cancer-associated transcript 1)[48], *CDX1* and *EPCAM* (Fig. 2a). In addition, we observed increased H3K4me3 signal at the promoter of *EGFR-AS1* that is involved in EGFR addiction[49].

We used data from The Cancer Genome Atlas (TCGA) and Genome-Tissue Expression (GTEx) projects[50,51] to generate cancer-specific signatures of genes whose expression is significantly higher in tumors compared to normal tissues (Methods and Supplementary Table 3). Testing for overlaps, we found that the set of genes with high H3K4me3 signal in the cancer sample had a significant overlap (303 of 739 genes; hypergeometric test, $q < 10^{-90}$) with colorectal adenocarcinoma (COAD) genes but only a negligible overlap with non-gastrointestinal cancers genes (Methods and Extended Data Fig. 3a).

Tissue-specific enhancers are also detected by cfChIP-seq. Using the Roadmap Epigenomics compendium chromatin annotations, we assigned cell types to distal enhancers (Methods). Comparing H3K4me2 signal in healthy samples to cancer samples, we observed significant differences in colon-specific enhancers, which are barely present in healthy samples (Extended Data Fig. 3b).

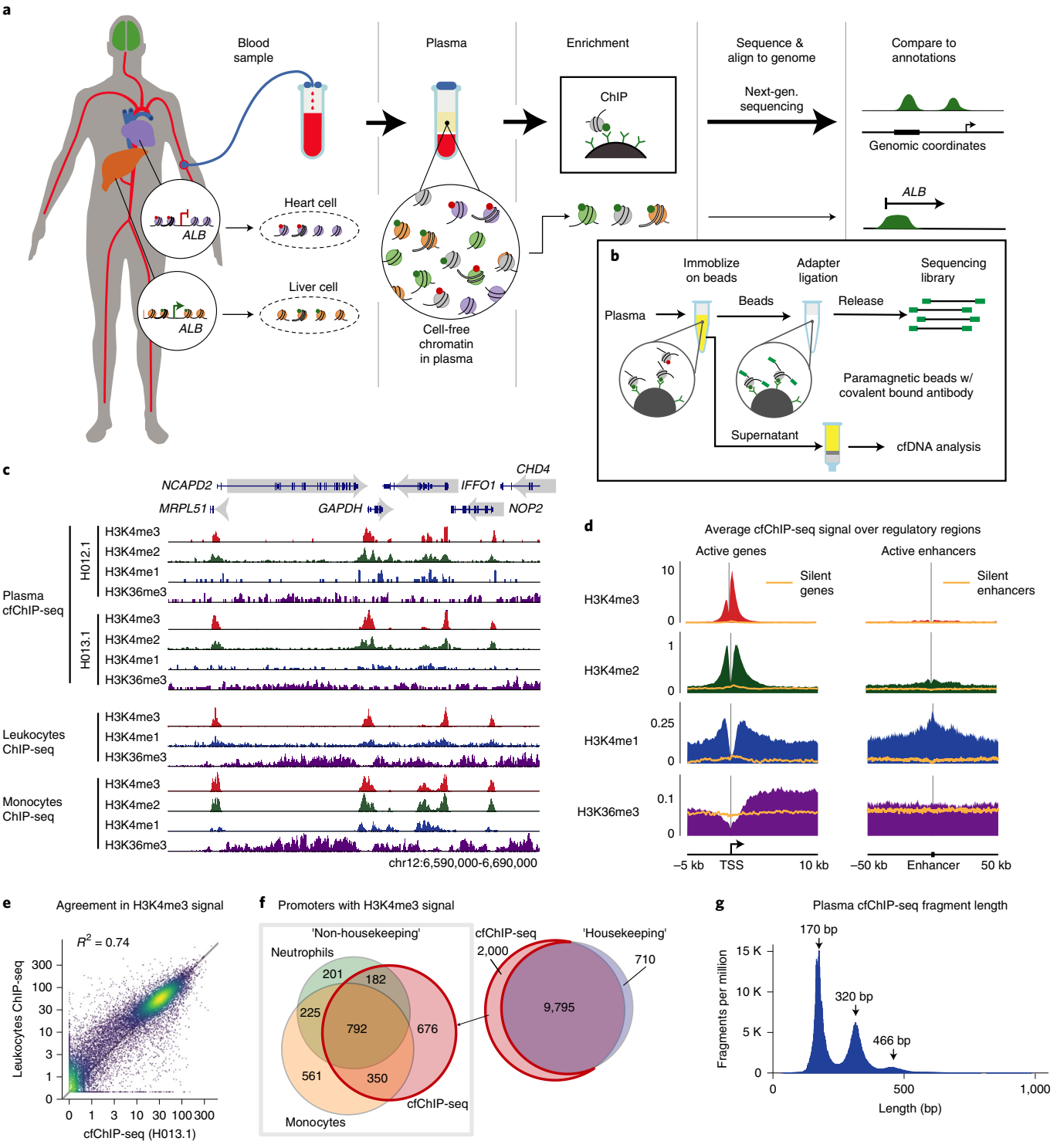
Tri-methylation of H3 lysine 36 (H3K36me3) requires active transcription elongation to be deposited and is indicative of gene activity[14]. Indeed, we observed the typical enrichment of H3K36me3 cfChIP-seq signal at gene bodies (Fig. 1d and Extended

Fig. 1 | Chromatin immunoprecipitation from plasma. a, cfChIP-seq method outline. Chromatin fragments from different cells are released to the bloodstream. These fragments are immunoprecipitated and sequenced. **b**, cfChIP-seq protocol. Antibodies are covalently bound to paramagnetic beads. Target fragments are immunoprecipitated directly from plasma. After washing, on-bead ligation is performed to add indexed sequencing adapters to the fragments. The indexed fragments are released and amplified by PCR to generate sequencing-ready libraries. **c**, Genome browser view of cfChIP-seq signal on a segment of chromosome 12. Top tracks are cfChIP-seq signals from two healthy donors. The lower tracks are published ChIP-seq results from human white blood cells (leukocytes)[25]. In each group, we show four tracks corresponding to four histone marks: H3K4me3 (red), H3K4me2 (green), H3K4me1 (blue) and H3K36me3 (purple). **d**, Meta-analysis of cfChIP-seq signal over active promoters and enhancers. The orange line denotes the average of corresponding negative control regions (inactive genes and enhancers), providing an estimate of the background. Scale of all graphs is in coverage of fragments per million. **e**, Comparison of normalized H3K4me3 coverage of cfChIP-seq from a healthy donor against ChIP-seq from leukocytes[25]. Each dot corresponds to a single gene. *x* axis: healthy cfChIP-seq sample; *y* axis: leukocytes ChIP-seq. **f**, Analysis of promoters of RefSeq genes with a significant cfChIP-seq signal (Methods) in healthy donors. cfChIP-seq captures most housekeeping promoters (ones that are marked in most samples in the reference compendium). The remaining 2,000 non-housekeeping genes in cfChIP-seq show large overlaps with non-housekeeping promoters marked in neutrophils and monocytes, the two cell types that contribute most to cfDNA in healthy donors. **g**, Size distribution of sequenced cfChIP-seq fragments shows a clear pattern of mono- and di-nucleosome fragment sizes: *x* axis: fragment length in base pairs (bp); *y* axis: number of fragments per million in 1-bp bins.

Data Fig. 3c), and the signal in healthy donors correlated with leukocyte RNA sequencing levels (Extended Data Fig. 3d). Comparing the H3K36me3 signal from healthy donors to that of the patient with cancer, we observed 5,416 genes that are hyper H3K36 tri-methylated by at least four-fold in the cancer sample (Fig. 2a).

Examining the genes with increased H3K36me3 signal in this cancer sample, we distinguish among three main classes. Class I includes ~3,400 genes marked by H3K36me3 and H3K4me3 in healthy and cancer samples (*DHX9*; Fig. 2b). Class II contains ~1,300 genes similarly marked by H3K4me3 but that differ in their H3K36me3 signal, which provides additional information beyond H3K4me3 (*SAP18* and *SKA3*; Fig. 2b). Finally, 141 Class III genes are marked with both signals only in the cancer sample (*VWA2*; Fig. 2b). Contrasting the set of highly expressed COAD signature genes with these three classes, we observe that each class captures different parts of these sets (Fig. 2c).

Altogether, these results demonstrate the ability of cfChIP-seq to probe the state of various genomic features, including promoters, enhancers and gene bodies in the tissue of origin. Moreover, cfChIP-seq detects functional changes in samples from a patient



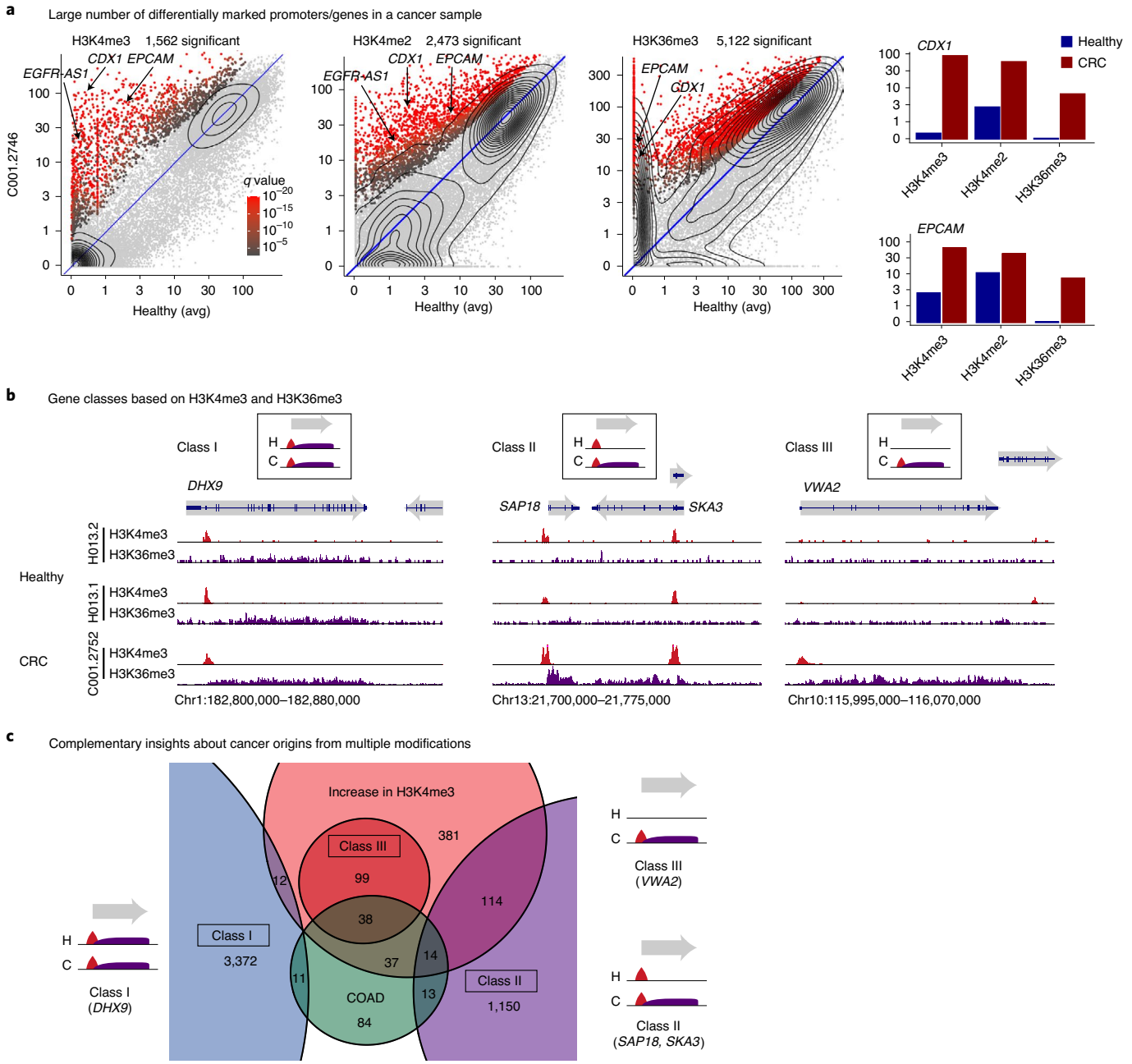


Fig. 2 | cfChIP-seq of multiple marks is informative on gene expression. a, Detection of genes with significant high coverage in a sample from a patient with CRC (C001; Supplementary Table 4). For each gene, we compared mean normalized coverage in a reference healthy cohort (*x* axis) against the normalized coverage in the cancer sample (*y* axis). For H3K36me3, the signal is normalized by gene length. Significance test whether the observed number of reads is significantly higher than expected based on the distribution of values in healthy samples (Methods). The levels of two genes in these comparisons are shown on the bar chart (right). **b**, Browser views of genes that demonstrate different H3K4me3 and H3K36me3 classes. Class I: genes marked by both marks in healthy and cancer samples. Class II: genes marked by H3K4me3 in healthy and cancer samples but with H3K36me3 only in the cancer sample (gain of H3K36me3). Class III: genes marked by both marks only in the cancer sample (gain of both marks). **c**, Venn diagram (zoom-in view) showing the relations of genes from the three classes in **b** with the set of genes that show increased H3K4me3 and the set of genes previously identified to be highly expressed in COAD (Methods).

with cancer, which are consistent with independent studies of this cancer type.

cfChIP-seq of H3K4me3 correlates with gene expression. To systematically evaluate the extent to which cfChIP-seq reflects gene expression patterns in the cells of origin, we focused on the H3K4me3 mark, because the signal is highly concentrated at promoters and is predictive of gene expression levels[27,52,53].

We quantified the relationship between promoter H3K4me3 and gene expression levels using 56 Roadmap Epigenomics samples with matched gene expression and H3K4me3 ChIP-seq profiles. For each gene, we compared the expression levels of the gene to promoter H3K4me3 ChIP-seq signal across all samples (Methods). We found that, for a large group of genes (10,150/14,313 genes), H3K4me3 ChIP-seq signal was significantly correlated with expression levels of the gene (Pearson $0.28 \leq r \leq 0.99$; Fig. 3a). The remaining genes

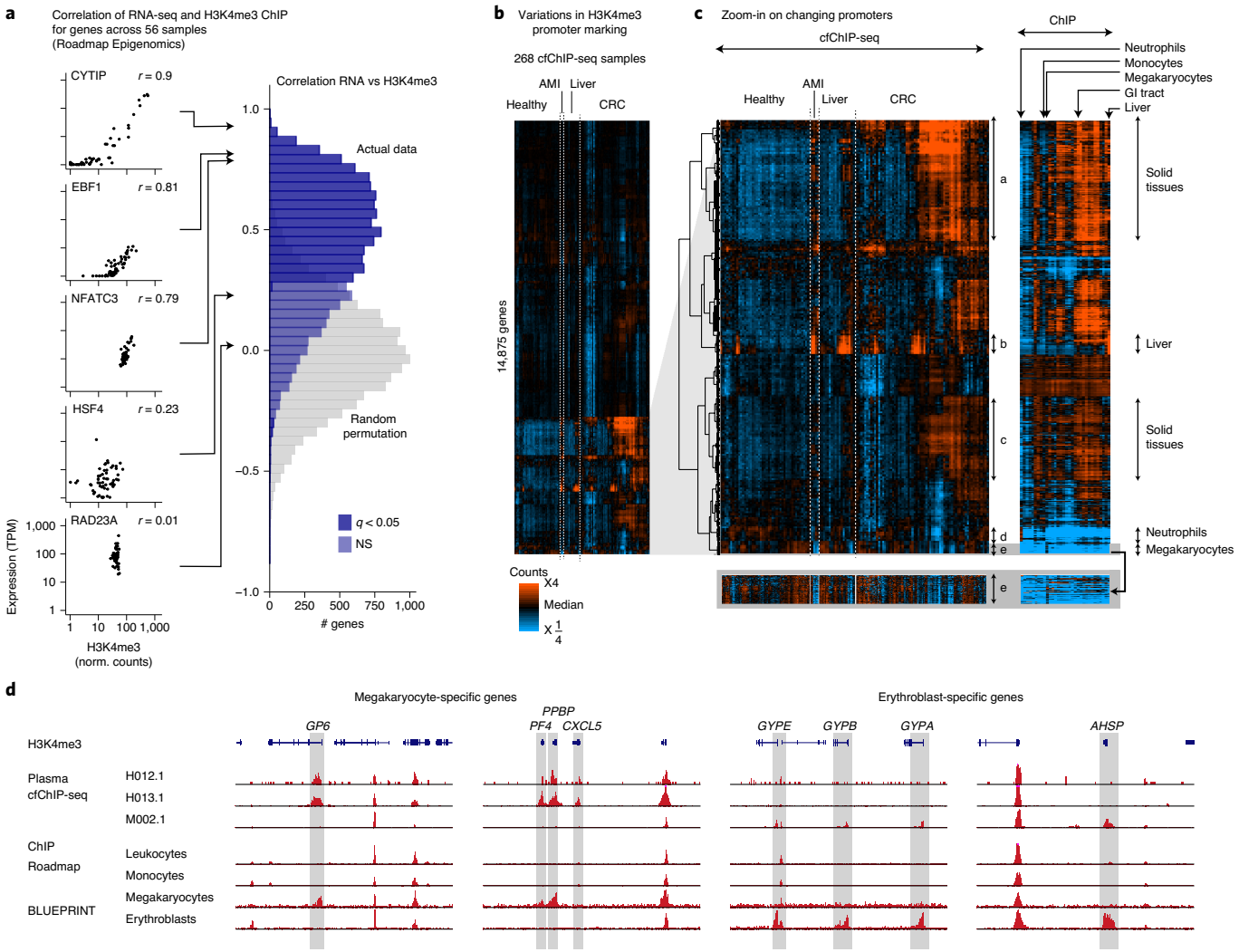


Fig. 3 | H3K4me3 cfChIP-seq signal is correlated with expression levels. a, Gene-level analysis of the correlation in expression level and H3K4me3 signal across 56 Roadmap Epigenomics samples[25] with matching profiles of both expression and H3K4me3 ChIP-seq. For each gene, we computed the Pearson correlation of its normalized expression levels and normalized H3K4me3 levels across the samples. Right, a histogram of the correlations on all RefSeq genes (significance with respect to random correlation, shown in gray). Left, examples of genes with different correlation values. **b**, Heat map showing patterns of the relative H3K4me3 cfChIP-seq coverage on promoters of 14,875 RefSeq genes. The normalized coverage on the gene promoter (Methods) was log-transformed ($\log_2$(1+coverage)) and then adjusted to zero mean for each gene across the samples. The samples include cfChIP-seq samples from a compendium that includes healthy donors, patients with AMI, patients with liver disease and patients with CRC. **c**, Zoom-in on the bottom cluster of **c**. The right panel shows the H3K4me3 ChIP-seq from tissues and cell types from Roadmap Epigenomics[25] and BLUEPRINT[58]. Specific clusters of genes are marked by arrows. **d**, Genome browser view for megakaryocyte- and erythroblast-specific genes. Shown is cfChIP-seq from two healthy individuals (H012.1 and H013.1) and a patient with AMI who exhibited enhanced erythropoiesis (M002.1). Also shown are two ChIP-seq profiles from the Roadmap Epigenomics reference atlas and two samples from the BLUEPRINT project of cord-blood-derived megakaryocytes and erythroblasts. NS, not significant.

are either genes that have high H3K4me3 levels in their promoters in most samples (housekeeping, 1,616/4,163 genes; for example, *RAD23A*) or genes with low levels of expression in all tissues (1,299/4,163 genes).

Next, we examined the relation between expression levels and cfChIP-seq H3K4me3 signal. Comparison of H3K4me3 cfChIP-seq signal at promoters shows a good agreement with RNA levels in cells known to contribute to the cfDNA pool (Pearson $R^2 = 0.40$; Extended Data Fig. 4a–d), consistent with similar comparisons in matched H3K4me3 ChIP-seq signal and RNA levels[25].

These results show that H3K4me3 cfChIP-seq signal is informative of gene expression levels in tissues of origin.

cfChIP-seq survey of diverse physiological and pathologic conditions. Do cfChIP-seq profiles reflect the underlying physiology? We performed H3K4me3 cfChIP-seq on 268 samples from a diverse cohort of individuals (Supplementary Table 4), including 88 samples from 61 healthy donors (ages 23–66 years), eight samples from four patients with AMI, 38 samples from 33 patients with a range of liver-related pathologies and 135 samples from 56 patients with metastatic CRC. The cfDNA content of these patients was expected to be significantly different owing to changes in the contributing tissue of origin. For example, we expected to detect cfDNA from cardiomyocytes after AMI[39], cfDNA from colon tumors in patients with CRC[54,55] and an increase in hepatocyte cfDNA in various liver pathologies[42].

We performed hierarchical clustering of 14,875 RefSeq gene promoters that have a noticeable signal in at least one sample (Fig. 3b and Methods). We found that 10,177 genes show relatively small differences among samples. These tend to be highly expressed

housekeeping genes with CpG islands at their promoters (Extended Data Fig. 4e,f). The remaining 4,698 genes display a rich tapestry of patterns (Fig. 3c).

Platelet progenitor cfDNA in healthy donors. Our analysis identified a cluster with a clear signal in healthy donors (Cluster e; Fig. 3c) that is enriched for megakaryocytes-specific genes such as *GP6* and *PF4* (25/144 genes in the cluster are in the REACTOME 'Platelet activation, signaling and aggregation', $P < 2 \times 10^{-25}$). However, we are not aware of previous reports of megakaryocytes as a source of cfDNA. Conversely, previous analysis of cfDNA CpG methylation identified erythroblasts as major (20–40%) contributors of cfDNA[11,56]. However, erythroblast-specific promoters are largely absent in healthy samples (Fig. 3d) but were detected in a sample from a patient with severe hypoxia (for example, *GYPA*, *GYPB* and *ASHP*; Fig. 3d). These results suggest that platelet progenitors (megakaryocytes), but not erythrocyte progenitors, are major contributors to the cfDNA pool in healthy donors. The possible source of the discrepancy is lineage adjacency of erythrocytes and megakaryocytes that are both derived from a common hematopoietic progenitor[57] and, thus, might have similar CpG methylation patterns. This observation highlights the potential of gene expression oriented information as provided by cfChIP-seq in detecting events that are otherwise indistinguishable.

cfChIP-seq detects cfDNA cell of origin. To detect the compositions of cells and tissues that contribute to the cfDNA pool, we used published ChIP-seq data to define cell type/tissue-specific signatures as promoters that have high signal only in one cell type[25,58] (Fig. 4a, Methods and Supplementary Table 5). In healthy donors, we observed a strong signal of neutrophils, monocytes and megakaryocytes and a lower but significant signal of liver, in agreement with published cfDNA methylation analysis[11] (Fig. 4b). In contrast, we did not observe significant signals in signatures of other tissues (Fig. 4b).

As controlled test cases for cell type detection, we considered pathologies involving organ damage. One such case is AMI, which involves the ongoing death of cardiomyocytes. In contrast to samples from healthy donors, a cardiomyocyte signal is clearly detected in samples from patients with AMI (Fig. 4c). Furthermore, we see good agreement among the strength of the cfChIP-seq heart signature, the levels of troponin measured in the blood and the estimate of heart cfDNA by CpG methylation[39] (Fig. 4c). In addition, a significant increase in heart signature signal is observed immediately after percutaneous coronary intervention (PCI) (Fig. 4d), as previously reported by assaying cfDNA methylation[39].

This example highlights the sensitivity of the method. We detected a significant heart signal in patient M003.1, who had very low troponin levels and 0.25% contribution of heart cfDNA. The sensitivity of detection depends on the number of informative nucleosomes in the signature of interest, the specific capture rate of modified nucleosomes and the non-specific capture of background cfDNA (Methods and Extended Data Fig. 5a). Our analysis shows that sensitivity of 0.1% can be readily achieved with a biologically relevant signature size (Extended Data Figs. 5b and 6, Methods and Supplementary Note).

In the case of partial hepatectomy, we observed dramatic changes in the cfChIP-seq signal of liver signature after the operation, as expected, which decayed to basal levels after 1 week (Fig. 4e). These changes are consistent with measurement of the liver marker alanine aminotransferase (ALT). A noticeable difference is the faster drop in the cfChIP-seq liver signal compared to ALT, likely reflecting the shorter half-life of cfDNA (<1 h) relative to ALT (~47 h)[59]. We found excellent agreement between liver cfChIP-seq signature levels and liver cfDNA estimates ($R^2 = 0.96$; Extended Data Fig. 7a).

An advantage of cfChIP-seq is that it is not limited to a set of preselected markers and, hence, can provide an unbiased view of the contributions of different cell types to the cfDNA pool. We evaluated the panel of cell-type-specific signatures across all cfChIP-seq samples (Fig. 4f and Supplementary Table 6) and detected signatures of monocytes, neutrophils and remote organs (for example, liver and bone marrow megakaryocytes) in all samples. The observed decrease in the relative level of leukocyte signatures in samples that show increased cfDNA load is consistent with a smaller proportion of cfDNA from these cells. For example, AMI patient M004.1 had a cfDNA concentration of 21 ng ml^{-1}, and 35% of his cfDNA originated from heart based on CpG methylation analysis.

These results demonstrate that cfChIP-seq signal reflects differences in the tissue-of-origin composition. Ongoing pathological processes are reflected in signal changes corresponding to the affected tissue.

cfChIP-seq signal reflects patient-specific transcriptional program activity. To test whether cfChIP-seq can reveal specific transcriptional programs within the tissue of origin, we evaluated the H3K4me3 cfChIP-seq signal in gene sets representing different cellular processes, protein complexes and transcriptional responses and targets of transcription factors based on ChIP studies[60–63] (Fig. 4g and Methods). We tested for changes in the signal of a gene set compared to the mean and variance of a reference healthy cohort of 26 samples (Methods) that uncovered multiple gene sets that differed from the expected signal in healthy donors (Supplementary Table 7).

For example, in patient M002.1, cfChIP-seq identified a strong increase in the signal of heme biosynthesis ($q < 10^{-9}$) and a strong decrease in granulocytes pathway ($q < 10^{-9}$), consistent with the results discussed above (Fig. 4d). Another example is the increased interferon signature in patient M004, who suffered severe heart damage as reflected by the levels of troponin and cfChIP-seq heart markers (Fig. 4c). Induction of interferon response can promote a fatal response to AMI[64]. The induction of interferon-mediated immune response is accompanied by increased cfChIP-seq signal in targets of STAT2 and other immune-related transcription factors. In addition, consistent with the massive amount of cardiomyocyte cfDNA in patient M004, we observed a significant increase in targets of MYOD1 and MYOG, which are two factors involved in cardiomyocyte development.

Detection of pathology-specific liver signals. The dynamic nature of active histone marks suggests that cfChIP-seq might inform on intra-tissue pathology-related alterations in gene expression. Many of the gene programs enriched in our samples are related to liver function (Fig. 4g); thus, we decided to test this hypothesis on liver hepatocytes. We assembled a cohort of patients with verified liver-related diagnosis and/or patients showing increased liver contribution, including patients at different stages of non-alcoholic fatty liver disease/non-alcoholic steatohepatitis (NAFLD/NASH, $n = 15$), patients with autoimmune hepatitis (AIH, $n = 3$), patients post liver transplant ($n = 5$), patients with infection associated with liver injury ($n = 1$), patients with AMI-associated liver injury ($n = 1$) and patients with partial hepatectomy ($n = 2$) (Supplementary Table 4).

We estimated the percentage of liver-derived chromatin in each sample using the Roadmap Epigenomics liver H3K4me3 ChIP-seq sample as a reference of liver tissue (Fig. 5a and Methods). The estimates range from ~2% in healthy samples to 44% in liver samples, consistent with a CpG methylation-based estimate of liver cfDNA quantity ($R^2 = 0.87$; Extended Data Fig. 7b). For example, in sample L001.1 from a patient with acute AIH, 44% of the cfDNA was liver derived, and 942 genes were significantly increased compared to healthy reference (Fig. 5b and Supplementary Table 2).

To understand whether this increase in liver genes signal is universal to all liver pathologies, we compared L001.1 with

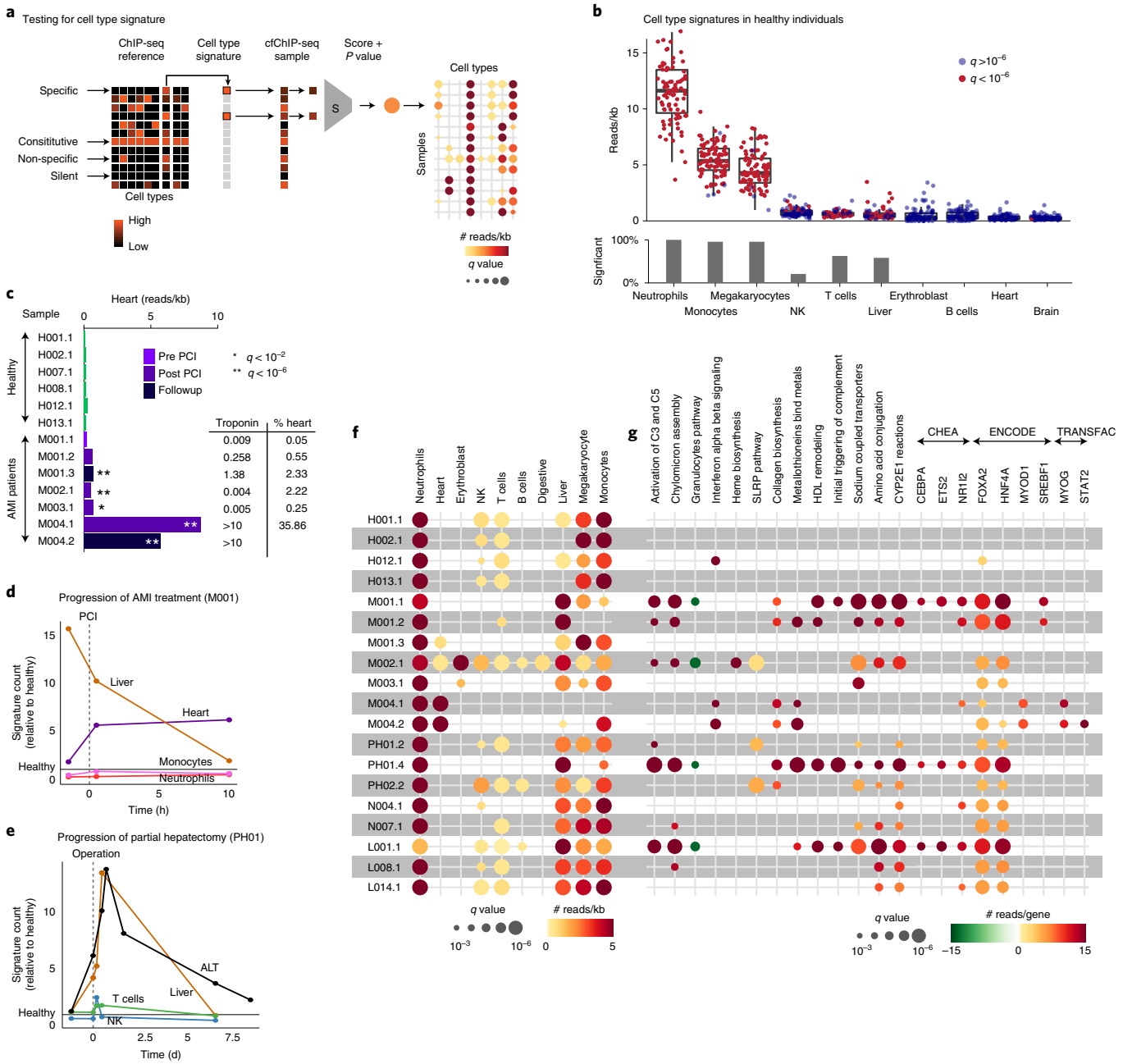


Fig. 4 | cfChIP-seq identifies cell type and program-specific expression patterns. a, Using the compendium of ChIP-seq profiles, we define for each cell type a signature consisting of the locations that are high only in the target cell type. Given a cfChIP-seq profile, we sum the signal at signature locations and test against the null hypothesis of non-specific background signal (Methods). **b**, Evaluation of average signal for cell type signatures in 88 healthy samples from 61 donors. Top, distribution of signature values (normalized reads/kb). Each dot is a sample. Box limits: 25–75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5× interquartile range from the hinge. Dots marked in red indicate values significantly higher than background levels (Methods). Bottom, percent of samples with significant signal for each signature. **c**, H3K4me3 cfChIP-seq signal in heart-specific locations in samples from representative healthy donors and patients with AMI (Supplementary Table 4) tested with respect to background levels (Methods). Inset, measured troponin levels and percent cfDNA from cardiomyocytes as estimated using DNA CpG methylation markers[39] from the same blood draws. **d**, Changes in signature strength in a patient with AMI (M001) before/after PCI. Signature levels are normalized to the mean in healthy donors. **e**, Changes in cfChIP-seq liver signature (brown line) and ALT levels (liver damage biomarker, black line) in samples of a patient who underwent partial hepatectomy (PH01). **f**, Heat map showing significance of selected cell type signatures in selected healthy donors and patients (Supplementary Table 6). Circle radius represents statistical significance (FDR-corrected *q* value), and the color represents read density (normalized reads per kb; Methods). **g**, Heat map showing significance of selected gene sets from curated database of transcriptional programs[60] and transcription factor targets[85–87] (Methods and Supplementary Table 7) tested against the null hypothesis of healthy baseline (Methods). Circle radius represents statistical significance (FDR-corrected *q* value), and the color represents the average read number (normalized reads per genes) compared to healthy baseline (Methods). NK, natural killer.

M001.1—a sample from a patient with AMI that had similar estimated levels of liver contribution (41%). As expected, many liver-specific genes were similarly increased in both samples (Fig. 5c; dark gray circles); however, pronounced differences in hundreds of genes were observed between the two samples (Fig. 5c and Supplementary Table 8). L001.1 was enriched for genes involved

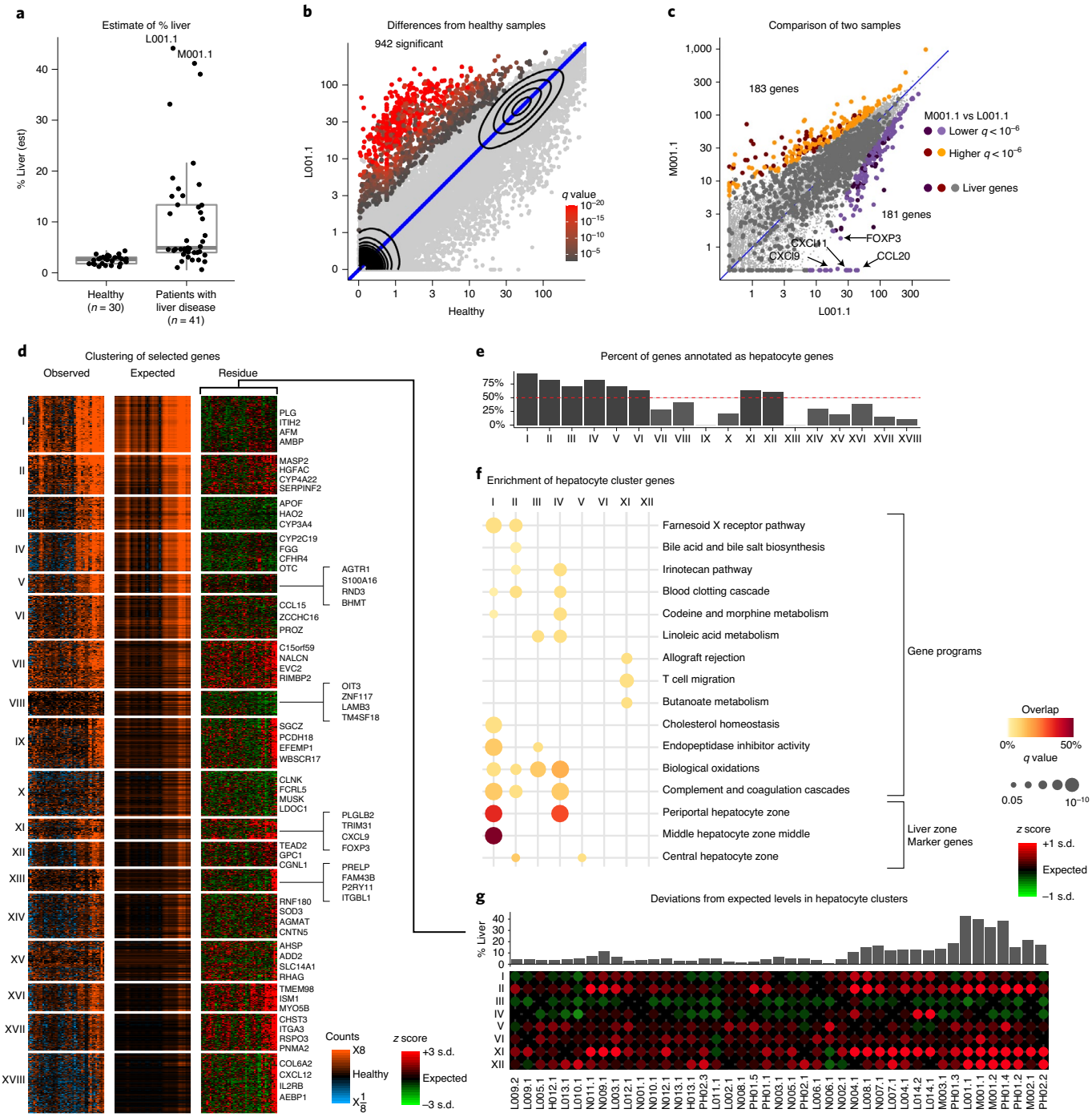


Fig. 5 | cfChIP-seq detects changes in liver-specific transcriptional programs. a, Estimate of % liver contribution to a healthy reference cohort and a cohort of patients with various liver pathologies (Supplementary Table 4). Box limits: 25–75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5× interquartile range from the hinge. **b**, Evaluation of differentially marked genes in a sample of a patient with acute AIH (L001) as in Fig. 2a. **c**, Differentially marked genes between two samples with similarly high liver contribution: L001.1 (acute AIH) and M001.1 (AMI-induced liver damage). For each gene, we compared the observed levels (L001.1, *x* axis; M001.1, *y* axis) and tested against the null hypothesis that the two values were sampled from the same distribution (Methods). Dark circles, genes that are significantly different in liver ChIP-Seq (Roadmap Epigenomics) compared to healthy reference. **d**, Clustering of 1,320 genes that are significantly higher in one of the samples in the liver cohort compared to healthy baseline. Left, values compared to healthy baseline. Middle, expected level assuming healthy liver signal with sample-specific % liver contribution. Right, *z* score of observed value from expected value. Listed 3–4 representative genes per cluster (right). Sample order in each heat map is identical and matches the order in **g**. **e**, Percent of genes in each cluster of **d** that are annotated as hepatocyte genes[67]. Clusters above the 50% threshold (red dashed line) are considered of hepatocyte origin. **f**, Enrichment analysis of hepatocyte clusters (Clusters I–VI, XI and XII). Hypergeometric test for significant overlap with gene programs from curated databases[88] and marker genes of hepatocyte zones[68]. Circle radius, FDR-corrected *q* values of hypergeometric enrichment test; circle color, fraction of overlap. **g**, Top, percent of liver contribution in each sample. Bottom, deviations from expected values for each sample in each of the hepatocyte clusters (average *z* score for each sample on cluster genes).

in interferon-gamma signaling (EnrichR, $q<3\times10^{-20}$), immune system (EnrichR, $q<1.9\times10^{-11}$), MHC class II protein complex (EnrichR, $q<6.4\times10^{-7}$) and allograft rejection (EnrichR, $q<3.7\times10^{-6}$), consistent with the autoinflammatory state of this patient. We also detected a stronger signal in genes associated with AIH, such as the ones encoding the transcription factor *FOXP3*, and the interferon-gamma-induced chemokines *CXCL9*, *CXCL11* and *CCL20* (refs.[65,66]). Several of these genes (dark colors), such as genes encoding proteins involved in complement and coagulation pathways (for example, *CFH* and *C4BPA*), are liver specific, demonstrating the potential of cfChIP-seq in detecting intra-organ transcriptional changes.

To get a more systematic view of differences in liver-specific expression programs among samples, we focused on 1,320 genes with significantly higher than expected cfChIP-seq signal in at least one of the liver samples (Fig. 5d, left). For each gene, we calculated the expected signal based on the estimated liver contribution of that sample (Fig. 5d, middle; Methods) and the *z* score to quantify the extent of deviation of the observed signal from the expected value, accounting for both sampling noise and the variability between healthy donors (Fig. 5d, right).

This analysis identified different types of gene clusters. In some clusters (for example, Clusters I–V, % variance explained (PVE) > 45%), the expected signal explains most of the variation among samples, suggesting that most of the signal in them is due to contribution from liver cells. In other clusters, such as Cluster XV, the signal is not explained by the amount of liver contribution (PVE < 5%), and, indeed, many of the genes in this cluster are expressed specifically in erythrocyte progenitors (for example, *ASHP* and *HBD*; 37/78 genes, $q<10^{-12}$). In some clusters (for example, Clusters VII, XI, XII and XIV; 30% > PVE > 10%), the amount of liver contribution partially explains the observed differences, suggesting that they are either differentially expressed in the liver among the individuals or originate from a mixture of several different tissues.

To better understand the contribution of liver-specific transcriptional programs, we focused on clusters where at least 50% of the genes were annotated as hepatocyte genes[67] (Fig. 5e; Clusters I-VI, XI and XII). We performed enrichment analysis of the gene sets in each cluster (Fig. 5f). As expected, we saw strong enrichments for many liver-related terms (Supplementary Table 9). Some clusters showed strong enrichments only to specific terms. For example, the genes of Cluster I were enriched for genes involved in the process of cholesterol homeostasis (9/111 genes, $q<4\times10^{-8}$), and the genes in Clusters I and IV were enriched with genes of the complement and coagulation cascade (14/111 genes, $q<3\times10^{-15}$, and 11/77 genes, $q<2\times10^{-12}$, respectively).

Next, we examined a single-cell RNA sequencing atlas of human liver cells[68] that identifies marker genes for hepatocytes at different liver zones on a functional axis from the portal vein (input to the liver from the gastrointestinal tract) to the central vein (output from the liver)[69]. Testing our gene clusters against these marker genes, we saw that Clusters I and IV were enriched for marker genes of periportal hepatocyte zones, Cluster I also for genes of middle hepatocyte zones and Clusters II and V for genes of the central hepatocyte zone (Fig. 5f). These could indicate either increased cell death in the relevant zone or global changes in liver metabolism toward the relevant metabolic regime.

Examining the deviations in the signal of clusters among samples allows us to identify sample-specific changes in hepatocyte-specific transcriptional programs (Fig. 5g). For example, we saw high levels of Cluster I genes in patients with immune-related pathology (for example, L001, L004, L008, L014 and N004) and high levels of Cluster IV genes in a subset of these patients (N004 and L014). Thus, although these clusters were both enriched for the periportal zone markers (Fig. 5f), they captured transcriptional programs that were differential among patients in the liver cohort.

Together, these results demonstrate the ability of cfChIP-seq to detect cell states within a remote tissue (liver) and within a specific cell type (hepatocytes).

Analysis of CRC by cfChIP-seq. We analyzed a collection of samples from an ongoing longitudinal study, following patients with metastatic CRC before and during treatment, including patients with undetectable or minimal disease at the time of sampling (135 samples from 56 patients; Supplementary Table 4).

Samples from within the CRC cohort showed much higher cfChIP-seq signal variability than observed among healthy donors (Fig. 3c). Closely collected samples from a single patient showed higher similarity than samples collected far apart (Fig. 6a), suggesting that, to a large extent, the variability among cancer samples is due to differences in the underlying patient molecular state[70]. Differences between CRC samples and healthy reference are apparent when examining relevant signatures (Fig. 6b). We selected a subset of COAD genes (based on analysis of TCGA gene expression data of COAD) that are not observed at all in a reference cohort of healthy donors and used them as a CRC signature (189 genes). We calibrated these scores to the range of 0–1, representing a rough proxy of tumor load. Using this signature, we classified CRC samples with area under the curve = 0.94 (Fig. 6c and Extended Data Fig. 8a).

We observed large differences in CRC signature magnitude among patients and during treatment of the same patient, consistent with the course of therapy (Fig. 6d)[70]. We also detected differences that appeared to result from disease progression (Extended Data Fig. 8b)—for example, an increased liver signal in C010.1943 versus C010.3743 (ARCHS4 tissue $q<2.9\times10^{-16}$), reflecting chemotherapy-induced liver damage[71], intra-tumor variation or immune-related signaling, such as the enrichment for interferon-gamma genes in C010.3743 versus C010.1943 (REACTOME $q<2.8\times10^{-6}$; Extended Data Fig. 8b).

cfChIP-seq detects molecular variability among patients with CRC. Identification of cancer-specific transcriptional programs can assist treatment choice[72,73]. A comparison of samples from different patients with similar CRC signature levels revealed differences in hundreds of genes (Fig. 6e). These differences can be due to contribution of additional tissues (for example, enrichment for liver genes in C001.2752 versus C040.3606, ARCHS4 tissue $q<4.1\times10^{-9}$), whereas others might reflect inter-tumor transcriptional differences—for example, enrichments for Wnt/calcium/cyclic GMP pathway in C040.3606 versus C001.2752 (BioPlanet, $q<0.00012$) and for cell adhesion molecules in C025.2815 versus C001.2752 (BioPlanet, $q<0.0045$). Additional examples include *EGFR-AS1* and the CRC marker *CCAT1* (ref. [74]) (Fig. 6e and Extended Data Fig. 9a). Transcription of *EGFR-AS1* modulates the splicing of EGFR and might affect anti-EGFR treatment[49]. When examining all samples, we identified variation in genes associated with immune activity, such as the checkpoint receptors *CD160*, *TIGIT* and *PDL1* (*CD274*) (Extended Data Fig. 9b), suggesting that we might detect tumor-related immune signals.

To identify major cfChIP-seq signature subtypes, we tested the gene set compendium (above) against samples with relatively high cancer loads (56 samples from 29 patients, where CRC signature > 0.15). We found 680 (of 7,538) gene sets that had informative signals in these samples (Supplementary Table 10, Methods and Extended Data Fig. 9c). We used these to initialize an iterative process to identify signatures that distinguish among sample subgroups (Methods), resulting in five gene signatures that capture the main behaviors in the original set of programs (Fig. 6f). Signatures A–C capture cancer gene expression programs, and signatures D and E capture duplication events.

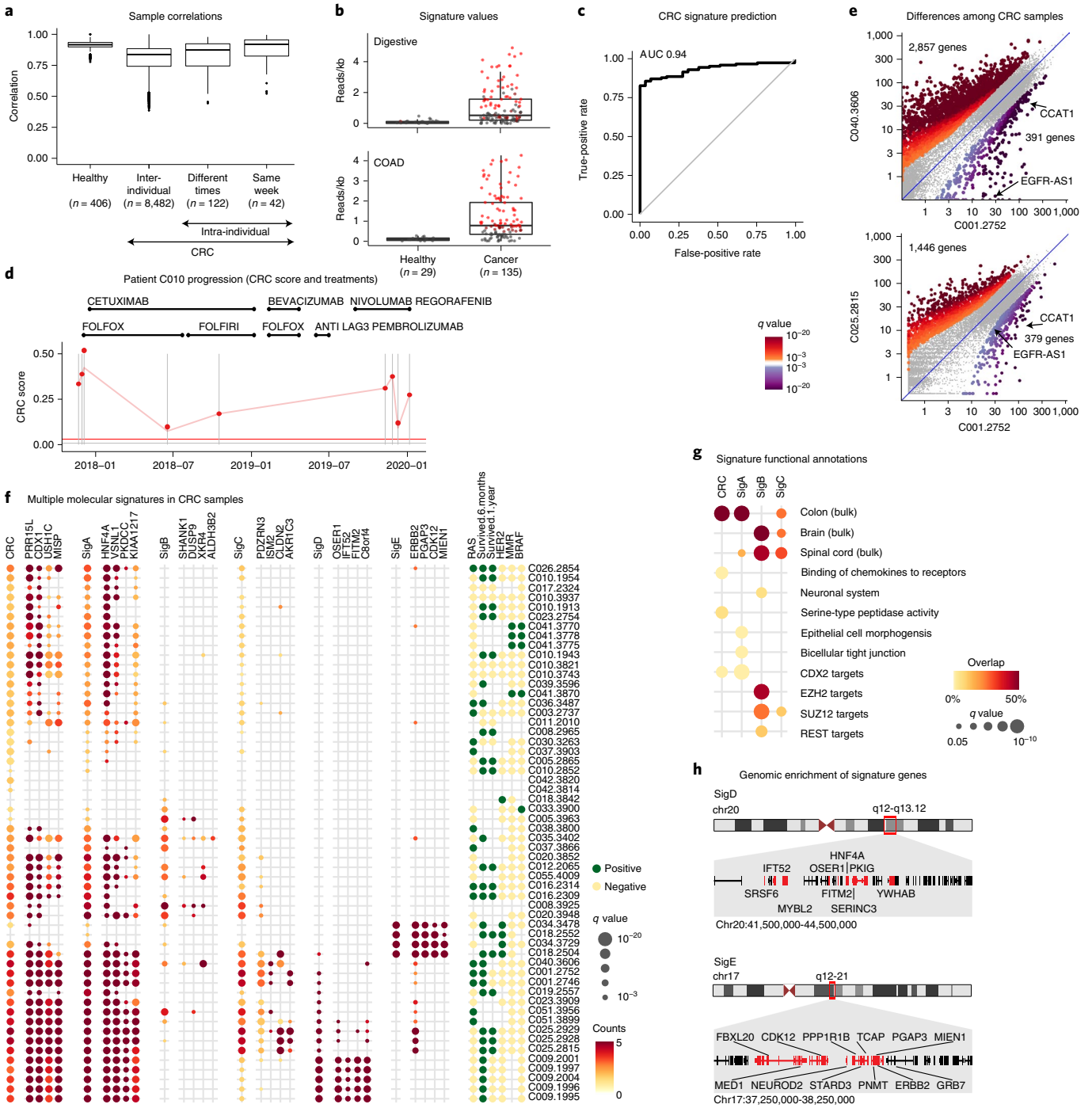


Fig. 6 | cfChIP-seq identifies molecular heterogeneity in patients with CRC. a, Pairwise comparisons (Pearson correlation, *y* axis) among samples: healthy donors; different patients with CRC; the same patient with CRC more than 1 week apart; and the same patient with CRC less than 1 week apart. Box limits: 25–75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5× interquartile range from the hinge. **b**, Signature differences between healthy and CRC samples. Top, signature of digestive tissue. Bottom, COAD gene signature. Box plots show distribution of signal (reads/kb, *y* axis) in each group. Each sample is a dot; red = significantly above background (Digestive) or healthy baseline (COAD). Box limits: 25–75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5× interquartile range from the hinge. **c**, Classification accuracy of patients with CRC versus healthy donors with CRC signature. Fraction false positive (*x* axis) versus fraction true positive (*y* axis). Diagonal line, expected curve for random classification. **d**, CRC signature progression during a single patient treatment. Top, treatment history as a function of time (*x* axis). Bottom, CRC signature strength (*y* axis) for different time points. **e**, Differences among samples with high CRC signature strength. For each gene, we compared coverage in the two samples (*x* axis and *y* axis). Significance test of whether the two values are sampled from the same distribution (Methods). **f**, Signature and representative gene levels are shown for five signatures identified by our analysis and the CRC signature. Circle color, increase in counts per gene above healthy baseline; circle radius, significance of this increase (Methods). Rightmost panel displays major clinical parameters—RAS, BRAF mutations and HER2 amplification—MMR deficiency and survival after 6 months and 1 year after the sample was taken. **g**, Functional enrichment of signatures. Representative enrichment from an unbiased testing of signature genes against large annotations database[88] using FDR-corrected hypergeometric test (Supplementary Table 11). **h**, Genome regions containing SigD and SigE genes. Marked in red are genes from each signature in the specific genomic loci. AUC, area under the curve.

The scores of the largest signature (SigA) were highly correlated with the CRC scores, although there was only a partial overlap between the two (Extended Data Fig. 9d). This signature was enriched with genes associated with colon (ARCHS4 tissue, $q < 10^{-64}$) and targets of CDX2, a transcription factor active in CRC (TRRUST, $q < 10^{-9}$) (Fig. 6g and Supplementary Table 11). The second signature (SigB) differentiated a small subset of the high-tumor-load samples and was enriched for genes in neuronal-associated terms (brain, ARCHS4 tissue, $q < 10^{-39}$) and polycomb repressive complex (PRC) and REST targets (ENCODE and ChEA, SUZ12 $q < 10^{-22}$, EZH2 $q < 10^{-22}$, REST $q < 10^{-7}$). REST represses neuronal genes in colon epithelium and is often deleted in CRC tumors[75]. This could indicate de-repression and misregulation of neuronal genes due to loss of polycomb/REST activity or indicate involvement of neuronal phenotypes in these tissues[76]. The third signature (SigC) selected a larger subset of samples, which included most of the samples selected by SigB, although there was little overlap of genes between the two signatures (Extended Data Fig. 9d).

We compared these signatures to the consensus molecular subtypes (CMS) classification of CRC tumors[77]. We examined the behavior of these signatures in 198 labeled CRC tumor gene expression profiles in the TCGA database[51] (Extended Data Fig. 9e). This analysis showed that SigA genes tend to have lower expression in CMS1 tumors, whereas SigB genes tend to have higher expression in CMS4. CMS4 tumors are characterized by upregulation of epithelial-to-mesenchymal transition and cancer stem cell-like phenotype and have been shown to have low EZH2 expression[78], which is consistent with the REST and PRC de-repression observed in SigB (Fig. 6g).

Ten of the 19 genes in SigD and 13 of the 17 genes in SigE were clustered around regions of known genomic duplications at chr20q13.12 and chr17q12-q21, respectively (Fig. 6h)[79,80]. The chr20q13.12 amplification has been previously reported in CRC and includes *HNF4A*, a gene encoding a transcription factor with increased activity in CRC[79]. The chr17q12-q21 includes the gene *ERBB2* and is known as the HER2 amplicon that appears in multiple types of cancer and with 4% prevalence in CRC[79]. Consistently, SigE is high in samples with identified HER2 amplifications (Fig. 6f), suggesting that cfChIP-seq detects this massive genomic amplification event. Unlike genomic copy number, H3K4me3 cfChIP-seq signal further increases the confidence that these copy number variations involve active transcription in the amplified regions. Detection of HER2 amplification in colon cancer has practical implications, as it is a predictive marker for prolonged survival of patients treated with HER2 inhibitors[81].

Altogether, these results show that a single cfChIP-seq blood test has the potential to detect the variability in patients with CRC related to the load of the tumor (CRC score), the contribution of additional tissues (for example, liver damage and immune cells) and gene expression inter-tumor heterogeneity.

Discussion

Here we introduce cfChIP-seq to infer the transcriptional programs of dying cells by genome-wide mapping of plasma cf-nucleosomes carrying specific histone marks. cfChIP-seq was performed on plasma cf-nucleosomes with four histone marks associated with active transcription (H3K4me1, H3K4me2, H3K4me3 and H3K36me3) for probing active or paused enhancers and promoters and gene body-associated transcriptional elongation. We further performed in-depth promoter-centric analysis on a large cohort of 61 healthy donors and 89 patients, including 135 samples from patients with metastatic CRC.

Beyond determining the cells of origin, cfChIP-seq can detect differences in patient- and disease-specific transcriptional programs—for example, among individuals with different etiology of increased liver cfDNA (Fig. 5). Our analysis shows that, even at this early stage, cfChIP-seq is highly sensitive in detecting signatures of interest, including cancer-specific signatures (Figs. 4 and 6 and Extended Data Figs. 5, 6 and 8). A unique feature of cfChIP-seq is that the immunoprecipitation step generates a biologically relevant reduced representation of the genome. This allows us to perform genome-wide unbiased analysis without the need for pre-selecting markers and with low sequencing depth.

Most current cfDNA-based methods rely on detecting genomic alterations in cfDNA to quantify the contribution of cfDNA from cells with altered genomic sequence, such as fetus, transplant or mutated genes in tumors[4–7]. These methods are blind to events that involve turnover and death of somatic cells. More recent approaches leverage epigenetic information in cfDNA. Extremely deep sequencing of total cfDNA to identify nucleosomes and transcription factors positions[35,82] and occupancy[34] reflect tissue of origin and gene expression. However, they rely on detecting changes in coverage over target regions, with a signal of each tissue/cell type imposed on the background of all other tissues/cell types[35]. An alternative modality is assaying cfDNA CpG methylation along the sequence[8–11,36,39–42]. DNA methylation serves as a stable epigenetic memory and is largely unchanged upon dynamic cellular responses. As such, it is highly informative regarding cell lineage but much less about transient changes in expression. Current assays of DNA methylation sequence both the methylated and unmethylated cfDNA, requiring deep sequencing of pre-selected sites to detect events with small representation in cfDNA.

Many cellular processes, including cancerous transformation, involve large changes in transcriptional programs that are intimately connected with specific histone modifications. Therefore, assaying chromatin marks in cf-nucleosomes provides rich and complex information beyond current methodologies.

We exploited the wealth of knowledge about gene expression for interpreting cfChIP-seq results. For example, observation of cfChIP-seq signal from genes encoding platelet-specific proteins (for example, GP6 and GP9), but not erythrocyte-specific proteins (for example, HBB), in healthy donors led us to identify megakaryocytes as major cfDNA contributors in healthy donors. Similarly, using existing annotations of liver expression programs, we identified the genes that represent hepatocyte contribution to the signal. We then used marker genes identified in a recent liver single-cell RNA sequencing atlas[68] to detect contributions from different liver zonation expression programs in each of the patients. Finally, in our analysis of the CRC cohort, we used a large collection of gene sets[60] as the basis for identifying signatures that classify molecular phenotypes of the samples.

These examples demonstrate the potential of using a single histone mark focused at gene promoters. There are potential advantages to combining multiple chromatin marks. Using H3K36me3 cfChIP-seq, which marks active elongation, we can better distinguish between a poised state and actual transcription. Parallel analysis of enhancer chromatin marks such as H3K4me1/2 can provide more precise understanding of the regulatory program that activated the genes. It is often the case that the same gene is regulated by multiple enhancers that are responsible for its activation in a specific cell type or transcriptional response. The main challenge in harnessing this information is the partial knowledge of enhancer–gene interactions in multiple tissues[83].

In addition to transcription, chromatin state is also intimately related to other chromatin-templated processes, such as cell cycle progression and DNA damage and repair. The potential for observing such processes with a non-invasive assay can revolutionize understanding of basic questions in human physiology and pathology. Here we demonstrate its ability to probe the active and poised genes in cells of origins, but, to fully harness the potential of this assay, we need a deeper understanding of the processes of cell death in health and disease and a more detailed understanding of

epigenetic footprint of transcription that would allow us to better exploit the transcriptional profiles currently collected in a large number of projects. Finally, deconvolving the superimposed signals from multiple cell populations is a central challenge for improved interpretation[84].

Altogether, cfChIP-seq is a highly informative and minimally invasive assay that opens up a wide range of opportunities for studying basic questions in human physiology that have been inaccessible until now.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-00775-6.

Received: 20 September 2019; Accepted: 17 November 2020;
Published online: 11 January 2021

References

1. Mandel, P. Les acides nucleiques du plasma sanguin chez l'homme. *CR Acad. Sci. Paris* **142**, 241–243 (1948).
2. Lo, Y. M. et al. Rapid clearance of fetal DNA from maternal plasma. *Am. J. Hum. Genet.* **64**, 218–224 (1999).
3. De Vlaminck, I. et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. *Sci. Transl. Med.* **6**, 241ra77 (2014).
4. Schwarzenbach, H., Hoon, D. S. & Pantel, K. Cell-free nucleic acids as biomarkers in cancer patients. *Nat. Rev. Cancer* **11**, 426–437 (2011).
5. Sun, K. et al. Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. *Proc. Natl Acad. Sci. USA* **112**, E5503–E5512 (2015).
6. Lu, J.-L. & Liang, Z.-Y. Circulating free DNA in the era of precision oncology: pre- and post-analytical concerns. *Chronic Dis. Transl. Med.* **2**, 223–230 (2016).
7. Wan, J. C. et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat. Rev. Cancer* **17**, 223–238 (2017).
8. Lehmann-Werman, R. et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc. Natl Acad. Sci. USA* **113**, E1826–E1834 (2016).
9. Guo, S. et al. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. *Nat. Genet.* **49**, 635–642 (2017).
10. Kang, S. et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol.* **18**, 53 (2017).
11. Moss, J. et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat. Commun.* **9**, 448142 (2018).
12. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98**, 285–294 (1999).
13. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707–719 (2007).
14. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77–88 (2007).
15. Berger, S. L. The complex language of chromatin regulation during transcription. *Nature* **447**, 407 (2007).
16. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178 (2015).
17. Heintzman, N. D. et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
18. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
19. Heintzman, N. D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
20. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).
21. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
22. Lawrence, M., Daujat, S. & Schneider, R. Lateral thinking: how histone modifications regulate gene expression. *Trends Genet.* **32**, 42–56 (2016).
23. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636–640 (2004).
24. Visel, A. et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
25. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
26. Lara-Astiaso, D. et al. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
27. Weiner, A. et al. High-resolution chromatin dynamics during a yeast stress response. *Mol. Cell* **58**, 371–386 (2015).
28. Holdenrieder, S. et al. Nucleosomes in serum of patients with benign and malignant diseases. *Int. J. Cancer* **95**, 114–120 (2001).
29. Holdenrieder, S. et al. Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. *Clin. Chem.* **51**, 1544–1546 (2005).
30. Rumore, P. M. & Steinman, C. R. Endogenous circulating DNA in systemic lupus erythematosus. Occurrence as multimeric complexes bound to histone. *J. Clin. Invest.* **86**, 69–74 (1990).
31. Gezer, U. et al. Characterization of H3K9me3- and H4K20me3-associated circulating nucleosomal DNA by high-throughput sequencing in colorectal cancer. *Tumour Biol.* **34**, 329–336 (2013).
32. Bauden, M. et al. Circulating nucleosomes as epigenetic biomarkers in pancreatic cancer. *Clin. Epigenetics* **7**, 106 (2015).
33. Deligezer, U. et al. H3K9me3/H4K20me3 ratio in circulating nucleosomes as potential biomarker for colorectal cancer. *Circulating Nucleic Acids in Plasma and Serum* 97–103 (Springer, 2011).
34. Ulz, P. et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* **48**, 1273–1278 (2016).
35. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* **164**, 57–68 (2016).
36. Xu, R.-H. et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. *Nat. Mater.* **16**, 1155–1161 (2017).
37. Haller, N., Tug, S., Breitbach, S., Jörgensen, A. & Simon, P. Increases in circulating cell-free DNA during aerobic running depend on intensity and duration. *Int. J. Sports Physiol. Perform.* **12**, 455–462 (2017).
38. Ramachandran, S., Ahmad, K. & Henikoff, S. Transcription and remodeling produce asymmetrically unwrapped nucleosomal intermediates. *Mol. Cell* **68**, 1038–1053 (2017).
39. Zemmour, H. et al. Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. *Nat. Commun.* **9**, 1443 (2018).
40. Li, W. et al. CancerDetector: ultrasensitive and non-invasive cancer detection at the resolution of individual reads using cell-free DNA methylation sequencing data. *Nucleic Acids Res.* **46**, e89 (2018).
41. Shen, S. Y. et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579–583 (2018).
42. Lehmann-Werman, R. et al. Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. *JCI Insight* **3**, e120687 (2018).
43. Cristiano, S. et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385–389 (2019).
44. Gutin, J. et al. Fine-resolution mapping of TF binding and chromatin interactions. *Cell Rep.* **22**, 2797–2807 (2018).
45. Singh, S. S. et al. Widespread suppression of intragenic transcription initiation by H-NS. *Genes Dev.* **28**, 214–219 (2014).
46. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
47. Mizuta, R. et al. DNase γ is the effector endonuclease for internucleosomal DNA fragmentation in necrosis. *PLoS ONE* **8**, e80223 (2013).
48. Ozawa, T. et al. CCAT1 and CCAT2 long noncoding RNAs, located within the 8q.24.21 'gene desert', serve as important prognostic biomarkers in colorectal cancer. *Ann. Oncol.* **28**, 1882–1888 (2017).
49. Tan, D. S. W. et al. Long noncoding RNA EGFR-AS1 mediates epidermal growth factor receptor addiction and modulates treatment response in squamous cell carcinoma. *Nat. Med.* **23**, 1167–1175 (2017).
50. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
51. Cancer Genome Atlas Research Networket al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
52. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA* **107**, 2926–2931 (2010).
53. Liu, C. L. et al. Single-nucleosome mapping of histone modifications in *S. cerevisiae*. *PLoS Biol.* **3**, e328 (2005).
54. Swarup, V. & Rajeswari, M. R. Circulating (cell-free) nucleic acids—a promising, non-invasive tool for early detection of several human diseases. *FEBS Lett.* **581**, 795–799 (2007).

55. Leon, S. A., Shapiro, B., Sklaroff, D. M. & Yaros, M. J. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res.* **37**, 646–650 (1977).
56. Lam, W. K. J. et al. DNA of erythroid origin is present in human plasma and informs the types of anemia. *Clin. Chem.* **63**, 1614–1623 (2017).
57. Deutsch, V. R. & Tomer, A. Megakaryocyte development and platelet production. *Br. J. Haematol.* **134**, 453–466 (2006).
58. Stunnenberg, H. G., International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1897 (2016).
59. Giannini, E. G., Testa, R. & Savarino, V. Liver enzyme alteration: a guide for clinicians. *CMAJ* **172**, 367–379 (2005).
60. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
61. Drew, K. et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.* **13**, 932 (2017).
62. Giurgiu, M. et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
63. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41**, D793–D800 (2013).
64. King, K. R. et al. IRF3 and type I interferons fuel a fatal response to myocardial infarction. *Nat. Med.* **23**, 1481–1487 (2017).
65. Czaja, A. J. Chemokines as orchestrators of autoimmune hepatitis and potential therapeutic targets. *Aliment. Pharmacol. Ther.* **40**, 261–279 (2014).
66. Mercer, F. & Unutmaz, D. The biology of FoxP3: a key player in immune suppression during infections, autoimmune diseases and cancer. *Adv. Exp. Med. Biol.* **665**, 47–59 (2009).
67. Lachmann, A. et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**, 1366 (2018).
68. Aizarani, N. et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature* **572**, 199–204 (2019).
69. Jungermann, K. & Katz, N. Functional specialization of different hepatocyte populations. *Physiol. Rev.* **69**, 708–764 (1989).
70. Reinert, T. et al. Analysis of circulating tumour DNA to monitor disease burden following colorectal cancer surgery. *Gut* **65**, 625–634 (2016).
71. Tannapfel, A. & Reinacher-Schick, A. Chemotherapy associated hepatotoxicity in the treatment of advanced colorectal cancer (CRC). *Z. Gastroenterol.* **46**, 435–440 (2008).
72. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional addiction in cancer. *Cell* **168**, 629–643 (2017).
73. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
74. Nissan, A. et al. Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. *Int. J. Cancer* **130**, 1598–1606 (2012).
75. Coulson, J. M. Transcriptional regulation: cancer, neurons and the REST. *Curr. Biol.* **15**, R665–R668 (2005).
76. Rademakers, G. et al. The role of enteric neurons in the development and progression of colorectal cancer. *Biochim. Biophys. Acta Rev. Cancer* **1868**, 420–434 (2017).
77. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
78. Koppens, M. A. J. et al. Large variety in a panel of human colon cancer organoids in response to EZH2 inhibition. *Oncotarget* **7**, 69816–69828 (2016).
79. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
80. Ferrari, A. et al. A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. *Nat. Commun.* **7**, 12222 (2016).
81. Sartore-Bianchi, A. et al. Dual-targeted therapy with trastuzumab and lapatinib in treatment-refractory, KRAS codon 12/13 wild-type, HER2-positive metastatic colorectal cancer (HERACLES): a proof-of-concept, multicentre, open-label, phase 2 trial. *Lancet Oncol.* **17**, 738–746 (2016).
82. Ulz, P. et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat. Commun.* **10**, 4666 (2019).
83. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).
84. Shen-Orr, S. S. & Gaujoux, R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Curr. Opin. Immunol.* **25**, 571–578 (2013).
85. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
86. Matys, V. et al. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
87. Lachmann, A. et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–2444 (2010).
88. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

Methods

Patients. All clinical studies were approved by the relevant local ethics committees. The study was approved by the Ethics Committees of the Hadassah-Hebrew University Medical Center of Jerusalem. Informed consent was obtained from all individuals or their legal guardians before blood sampling.

Sample collection. Blood samples were collected in VACUETTE K3 EDTA tubes, transferred immediately to ice, and 1× protease inhibitor cocktail (Roche) and 10 mM EDTA were added. The blood was centrifuged (10 min, 1,500*g*, 4 °C); the supernatant was transferred to fresh 14-ml tubes and centrifuged again (10 min, 3,000*g*, 4 °C); and the supernatant was used as plasma for ChIP experiments. The plasma was used fresh or flash frozen and stored at −80 °C for long-term storage.

Bead preparation. Fifty micrograms of antibody were conjugated to 5 mg of epoxy M270 Dynabeads (Invitrogen) according to manufacturer instructions. The antibody–beads complexes were kept at 4 °C in PBS with 0.02% azide solution.

AB	Company	Catalog number
IgG	Cell Signaling	2729S
H3K4Me1	Diagenode	C15410194
H3K4Me2	Diagenode	C15410035
H3K4Me3	Diagenode	C15410003
H3K36Me3	Diagenode	C15410192

Immunoprecipitation, next-generation sequencing library preparation and sequencing. 0.2 mg of conjugated beads (~2 µg of antibody) were used per cfChIP-seq sample. The antibody–beads complexes were added directly into the plasma (1–2 ml of plasma) and allowed to bind to cf-nucleosomes by rotating overnight at 4 °C. The beads were magnetized and washed eight times with blood wash buffer (50 mM Tris-HCl, 150 mM NaCl, 1% Triton X-100, 0.1% sodium deoxycholate, 2 mM EDTA, 1× protease inhibitor cocktail) and three times with 10 mM Tris pH 7.4. All washes were done with 150 µl of the washing buffer on ice by shifting the beads from side to side on a magnet. Do not use a vacuum to remove supernatant during washes in buffers that do not contain detergents.

On-beads chromatin barcoding and library amplification were done as previously described[26,44] except for the DNA elution and cleanup step where the beads were incubated for 1 h at 55 °C in 50 µl of chromatin elution buffer (10 mM Tris pH 8.0, 5 mM EDTA, 300 mM NaCl, 0.6% SDS) supplemented with 50 units of proteinase K (Epicenter), and the DNA was purified by 0.9X SPRI cleanup (AMPure XP, Agencourt). The purified DNA was eluted in 25 µl of EB (10 mM Tris pH 8.0), and 23 µl of the eluted DNA was used for polymerase chain reaction (PCR) amplification with KAPA HotStart polymerase (16 cycles). The amplified DNA was purified by 0.8X SPRI cleanup and eluted in 12 µl of EB. The eluted DNA concentration was measured by Qubit, and the fragment size was analyzed by tapestation visualization. Note: if adapter dimers are substantially visible by tapestation after library amplification, we recommend pooling samples and performing additional 0.8X SPRI DNA cleanup or separating the pooled samples on a 4% agarose gel (E-Gel EX Agarose Gels, 4%, Invitrogen) and gel purification of fragments larger than adapter dimers (>150 bp). DNA libraries were paired-end sequenced by Illumina NextSeq 500.

Sequence analysis. Reads were aligned to the human genome (hg19) using bowtie2 (2.3.4.3) with 'no-mixed' and 'no-discordant' flags. We discarded fragments with low alignment scores (-q 1) and duplicate fragments. See Supplementary Table 1 for read number, alignment statistics and numbers of unique fragments for each sample. BAM files were processed by samtools (1.7) to BED files and by R (4.0.2) scripts (see 'Code availability').

Roadmap Epigenomics atlas. We downloaded aligned read data from the Roadmap Epigenomics consortium database. For our analysis, we discarded pre-natal, embryonic stem cell and cell line samples, resulting in 64 tissues and cell types (Supplementary Table 12). The aligned read files were then processed with the same scripts as cfChIP-seq samples—that is, all steps from numbers of reads mapped to each genomic window, background estimation and normalization.

Tumor type gene signatures. We downloaded RNA sequencing data from the UCSC Toil RNAseq Recompute Compendium[89], which includes samples from the TCGA and GTEx projects. We defined the set of genes that were over-expressed in a tumor type to satisfy three requirements: 1) significantly higher expression in tumor samples compared to the corresponding tissue samples (*t*-test, $q < 0.001$ after false discovery rate (FDR) correction); 2) significantly higher expression compared to all healthy samples (*t*-test, $q < 0.001$ after FDR correction); and 3) median expression in the tumor was higher than the median expression in each of the healthy samples.

Expected healthy expression level. To best emulate expression profiles of healthy individuals in the analysis of Extended Data Fig. 4a, we performed an in silico mix of the four cell types that contribute the most to cfDNA[11]: neutrophils, 32%; monocytes, 32%; megakaryocytes, 20%; and natural killer cells, 5%. The gene expression for these cell types was downloaded from the BLUEPRINT consortium.

Transcription start site location catalog. We downloaded the Roadmap Epigenomics ChromHMM annotation of all consolidated tissues. Using these annotations, we constructed a catalog of potential functional sites (enhancers, transcription start sites (TSS) and genes). We extended the catalog to include 3-kb regions centered on the TSS of annotated transcripts in the UCSC gene database and the Ensembl transcript database. We used the combined catalog to define regions along the genome. We used a different version of the catalog for analysis of each antibody, to match the mark. For H3K4me3 analysis, we used only TSS; for H3K36me3 analysis, we used only gene bodies; and for H3K4me2, we had annotations of TSS and enhancers. In each version of the catalog, the remaining mappable genome regions were assigned to background and tiled at 5-kb windows. See Supplementary Note for more detailed procedures.

We quantified the number of reads covering each region in the catalog in each of our samples and atlas samples. We estimated a locally adaptive model of non-specific reads along the genome for each of the samples and extracted counts that represent a specific ChIP signal in the catalog for each sample (Supplementary Note). These were then normalized (Supplementary Note) and scaled to 1 million reads in the reference healthy samples.

Estimating capture rates. To estimate capture rates of cfChIP-seq, we used our prior knowledge of the genomic distribution of H3K4me3 marked nucleosomes, which are highly localized at TSS (Fig. 1d), to distinguish between non-specific capture (in regions without TSS) and specific capture (in TSS that are known to be constitutively marked by H3K4me3). We used this idea in two different approaches (see Supplementary Note for more details).

In the global approach, we compared input to output of the cfChIP-seq assay. At the input end, we estimated the total number of nucleosomes that are present in the sample using the input cfDNA, which provides an upper bound on the number of nucleosomes it can contain (with each nucleosome, ~200 bp of DNA). We also estimated the percent of these that are modified, which, for H3K4me3, tends to be ~1–2%. At the output end, we estimated how many of the unique fragments are background and how many are signal (see above). We then divided #signal fragments in output by #modified nucleosomes in input to get the specific capture rate and, similarly, #background fragment in output by total #nucleosomes to get the non-specific capture rate.

In the local approach, we compared expected input coverage to output coverage. Using input cfDNA amounts, we can estimate the number of alleles (genomes) that cover each position. We then examined two types of regions, one as 'high-signal', where we assume that ~100% of the nucleosomes are modified (for example, promoters of constitutive genes), and the other one as 'no-signal', where 0% of the nucleosomes are modified (for example, background regions). The coverage we observed in the cfChIP-seq output is due only to non-specific capture in the no-signal region and due to both specific and non-specific capture in the high-signal region.

In both methods, we take into account an estimate of the sequencing depth that influences the number of observed reads (Supplementary Note). We estimated the probability of specific capture (above) to a range between 0.01% and 0.1% across dozens of H3K4me3 cfChIP-seq experiments (Extended Data Fig. 6b).

Sensitivity analysis. The ability of cfChIP-seq to detect rare molecular events in the cfDNA pool is dictated by several factors: the number of informative nucleosomes in the sampled plasma, the capture rate of target nucleosomes and the signal-to-noise ratio of the assay (Extended Data Fig. 5a). The number of informative nucleosomes in the plasma is proportional to the size of the genomic region in question and the amount of cells of interest that had shed their nucleosomes to the blood (Extended Data Fig. 6a). For example, we defined the cardiomyocyte-specific signature as 366 nucleosomes that are marked with H3K4me3 only in cardiomyocytes (Extended Data Fig. 6a). Detection of any H3K4me3 nucleosome from these regions is indicative of cardiomyocyte presence. Assuming a 1% contribution of cardiomyocyte to a cf-nucleosomes pool of ~1,000 genomes per milliliter, we expect ~36,600 informative nucleosomes in a 10-ml plasma sample.

We estimate capture rate as discussed above. We further assume independence of the concentration of plasma nucleosomes and capture rate (Extended Data Fig. 6c). We then define 'detectable' if the probability of capturing sufficient molecules to reject the null hypothesis of background capture is higher than 0.95 (Supplementary Note).

To evaluate these predictions, we titrated male-derived plasma into female-derived plasma. We evaluated the sensitivity for genomic signatures of different sizes at male-specific locations on the Y chromosome (Extended Data Fig. 6d,e), concluding that cfChIP-seq can detect the presence of male chrY DNA plasma when it constitutes 1.5% of the genomes in the plasma (Extended Data Fig. 6e), consistent with our estimates based on the parameters of the specific experiment (Extended Data Fig. 6f,g).

Tissue signatures. To define tissue-specific signatures of a specific modification, we examined binned representation of the atlas according to our catalog. For each

tissue, we defined a signature of unique windows with signal in one of the samples of the target tissue and without coverage in all others (Supplementary Note).

Gene-level analysis. For each gene, we defined the set of windows that match the gene (TSS in H3K4me3/2 and gene body in H3K36me3). The signal for a gene is the aggregate signal background over windows associated with it (Supplementary Note).

Comparison to RNA sequencing. The comparison of H3K4me3 ChIP to RNA sequencing was performed as follows. RNA expression (normalized transcripts per million (TPM)) was downloaded from the Roadmap Epigenomics Project. Normalized cfChIP-seq coverage per gene in the matching sample was taken from the Roadmap Epigenomics Atlas (above). We examined RefSeq genes that appeared in both datasets. For each gene, we computed Pearson correlation between log(TPM + 1) and log(ChIP-seq coverage + 1) values across all 56 tissue/cell types that had matched RNA sequencing and H3K4me3 ChIP-seq data.

Estimating healthy mean and variance. To define the healthy reference of signal per gene, we estimated the mean and variance of each gene in a set of 26 reference samples (Supplementary Table 1). The observed variation among the samples is due to the combination of biological variability and sampling noise. Thus, to estimate mean/variance, we used a maximum likelihood approach that models the sampling noise of each sample and identifies the mean/variance that best matches this model (Supplementary Note).

Statistical analysis. We use several custom-designed statistical tests in our analysis. In all analyses, we corrected for multiple hypotheses using FDR and estimate *q* values (R function p.adjust()).

- Comparison to background (Figs. 1f, 4b,c,f and 6b and Extended Data Fig. 6e). We test whether the total sum over a collection of windows (a signature, promoter windows of a gene, etc.) is larger than we would expect from the background signal. Formally, we examine whether we can reject the null hypothesis that the number of reads in the windows of interest is Poisson distributed according to estimated background rate at these windows (Supplementary Note).
- Comparison to reference (healthy) (Figs. 2a, 4g, 5b and 6f and Extended Data Fig. 9a,b). We test whether the total sum over a collection of windows is higher than we would expect according to mean and variance in healthy donor reference. We estimate two sample-specific parameters: 1) background rate (discussed above) and 2) a scaling factor that rescales average expectations to the sequencing depth of the specific sample (Supplementary Note). Together, these define the distribution of total reads in these windows under the null hypothesis that the individual is from the healthy population. We compute the *P* value of the actual number of observed reads in the gene windows using a two-tailed test, testing for the probability of having this number or higher and this number or lower according to the null hypothesis (Supplementary Note). Note: in the analysis of Extended Data Fig. 9a,b, we use a variant of this test where the reference is modified according to the % tumor estimated for the sample.
- Comparison of two samples (Figs. 5c and 6e and Extended Data Fig. 8b). We test whether we can reject the null hypothesis that the values observed for the same gene (or collection of windows) in two samples are from the same distribution. For each sample, we estimate the background rate and scaling factor (as above). Under the null hypothesis, they share the same normalized mean, which is scaled differently in each sample and added to the sample-specific background estimate. Under the alternative hypothesis, they have different means. These are nested hypotheses, and, thus, we use the likelihood ratio test.

Pathways and transcription factor targets. We downloaded a large collection of gene expression signatures representing different cellular processes, protein complexes and transcriptional responses from the MSigDB collection[60]. We downloaded transcription factor targets from the Harmonizome database[90]. These include targets from ENCODE[85], TRANSFAC[86] and CHEA[87].

Estimation of liver percentage. We used a linear regression model that matches the observed counts of select representative genes to a sum of contribution of healthy-wo-liver and healthy liver. Briefly, we used the Roadmap Epigenomics Atlas 'liver' (E066) as 100% liver. We assumed that the mean healthy profile contains about ~3% liver contribution and so defined the healthy-wo-liver as the result of subtracting 3% of liver profile from the healthy sample. We then identified the set of distinguishing genes as those that are close to zero in healthy-wo-liver and high in liver and those that are high in healthy-wo-liver and low in liver. These were used as input features for robust linear regression (R rlm() function) that estimates the linear combination of liver and healthy-wo-liver profiles that is closest to the observed profile. The weights (linear regression coefficients) are normalized to sum to one, and the contribution of liver is taken as % liver in the sample.

Cancer signatures. We tested a compendium of gene programs from multiple sources against high-scoring CRC samples. Gene programs that had significant enrichment above/below healthy reference in at least three CRC samples but less than two-thirds of all the CRC samples were selected for the next step. The pattern of significantly above/below enrichments was clustered (Extended Data Fig. 9c). Each cluster of gene programs corresponds to a classification of the CRC samples (significant versus non-significant). For each such cluster, we identified the genes that had significantly higher signal in the positive class of CRC samples compared to remaining CRC samples. The differential genes defined a new gene signature. These were clustered based on their classifications of samples and combined into non-overlapping sets of gene signatures (Supplementary Note).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data collected in this study were deposited in the European Genome-phenome Archive (EMBL-EBI) repository with accession code EGAS00001004913. BED files and browser tracks are available in the Zenodo repository: https://doi.org/10.5281/zenodo.3967253.

Browser tracks can be viewed by the UCSC genome browser.

- Session: http://genome.ucsc.edu/s/nirfriedman/cfChIP-seq
- Track hub: http://www.cs.huji.ac.il/~nir/Hubs/cfChIP-seq/hub.txt

Additional data from public repositories are listed here:

The datasets are as follows: UCSC known genes (AH5036); Ensembl transcripts (AH5046); genomic annotations (AH5040): AnnotationHub (http://bioconductor.org/packages/release/bioc/html/AnnotationHub.html); consolidated ChIP-seq: Roadmap Epigenomics (https://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/); mRNA-seq: Roadmap Epigenomics (https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/57epigenomes.RPKM.pc.gz); consolidated ChromHMM calls: Roadmap Epigenomics (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/all.mnemonics.bedFiles.tgz).

Code availability

R code for processing cfChIP-seq data is available at https://github.com/nirfriedman/cfChIP-seq.git.

References

89. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
90. Rouillard, A. D. et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, baw100 (2016).

Acknowledgements

We thank N. Kaminski, J. Moss, E. Pikarsky, N. Rajewsky, O.J. Rando, A. Regev and members of the Friedman lab for discussions and comments on this manuscript. We thank L. Friedman for help with illustrations and graphics. This work was supported by the European Research Council's AdG Grants 340712 'ChromatinSys' (to N.F.) and 786575 'RxmiRcanceR' (to E.G.); the Israel Science Foundation's I-CORE program grant 1796/12 (to T.K. and N.F.) and grants 2612/18 (to N.F.), 3020/20 (to A.G.), 2473/17 (to E.G.) and 486/17 (to E.G.); Israel Ministry of Science and Technology grant 3-14352 (to A.G.); National Institutes of Health grants RM1HG006193 (to N.F.) and CA197081-02 (to E.G.); Deutsche Forschungsgemeinschaft SFB841 (to E.G.); and DKFZ-MOST grant (to E.G.).

Author contributions

R. Sadeh and N.F. developed the concept. R. Sadeh, N.F. and I.S. designed the experiments with help from E.G., B.G., A.Z. and Y.D. R. Sadeh developed the cfChIP-seq method with help from A.R. and I.S. I.S., R. Sadeh and A.R. performed cfChIP-seq experiments. N.F. and G.F. developed analytical tools with help from J.G., M.N., G.M. and T.K. N.F., R. Sadeh, G.F., I.S. and J.G. analyzed the data. I.F.F, D.N. and R. Shemer performed the cfDNA methylation assays. Z.K. collected healthy donor samples. D.Y., T.P., A.H., J.E.C., A.S., M.T., A.G., M.M., S.A.G., A.B.Y., E.S., R. Safadi, D.P., E.G., B.G. and A.Z. provided clinical insights, recruited patients and collected patient samples. N.F., R. Sadeh, J.G., G.F. and A.C. wrote the paper with input from all authors.

Competing interests

A patent application for cfChIP-seq has been submitted by the Hebrew University of Jerusalem. R. Sadeh, I.S., J.G. and N.F. are founders of Senseera.

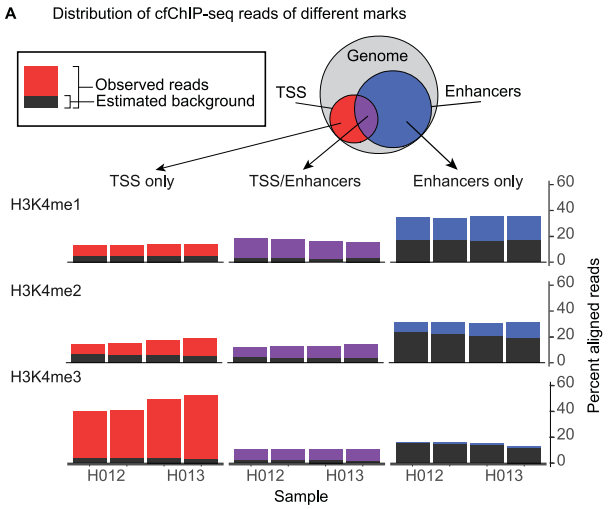
Additional information

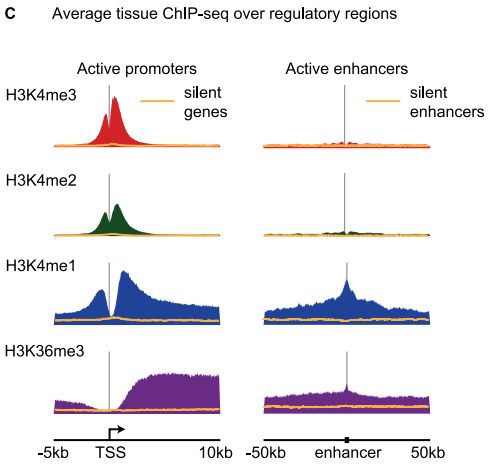
Extended data is available for this paper at https://doi.org/10.1038/s41587-020-00775-6.

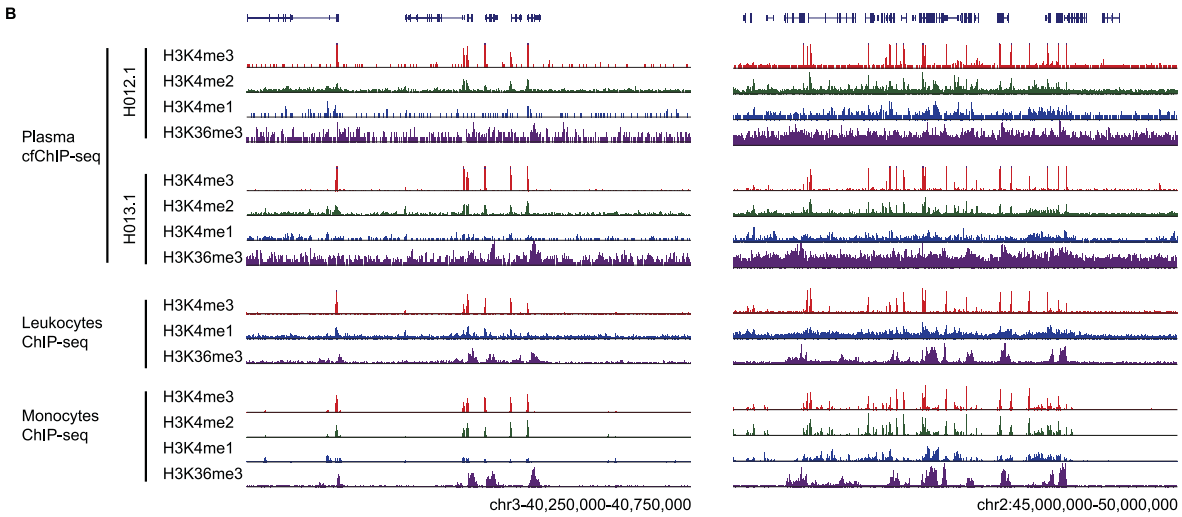
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-00775-6.

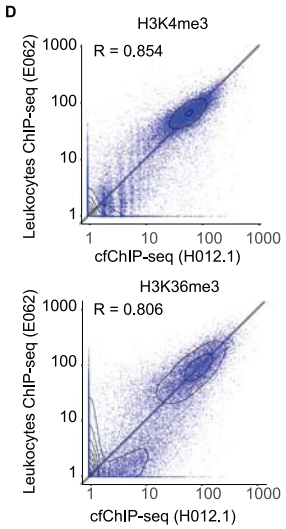
Correspondence and requests for materials should be addressed to N.F.

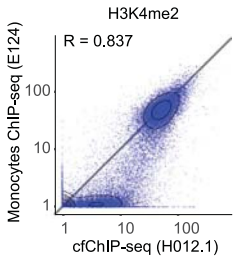
Reprints and permissions information is available at www.nature.com/reprints.

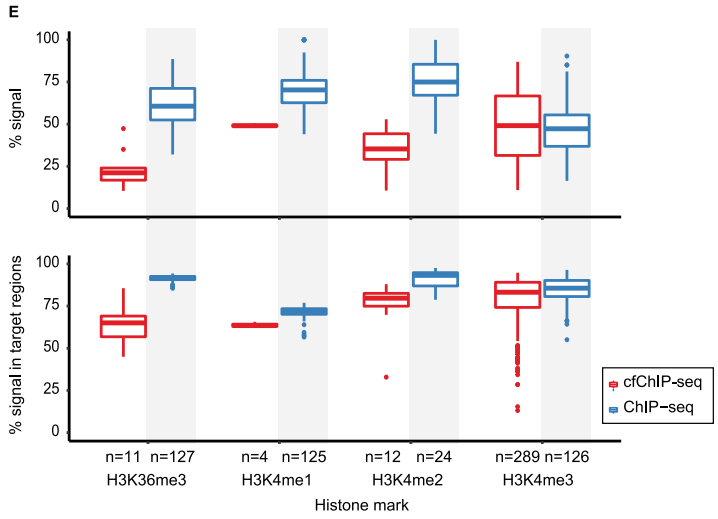






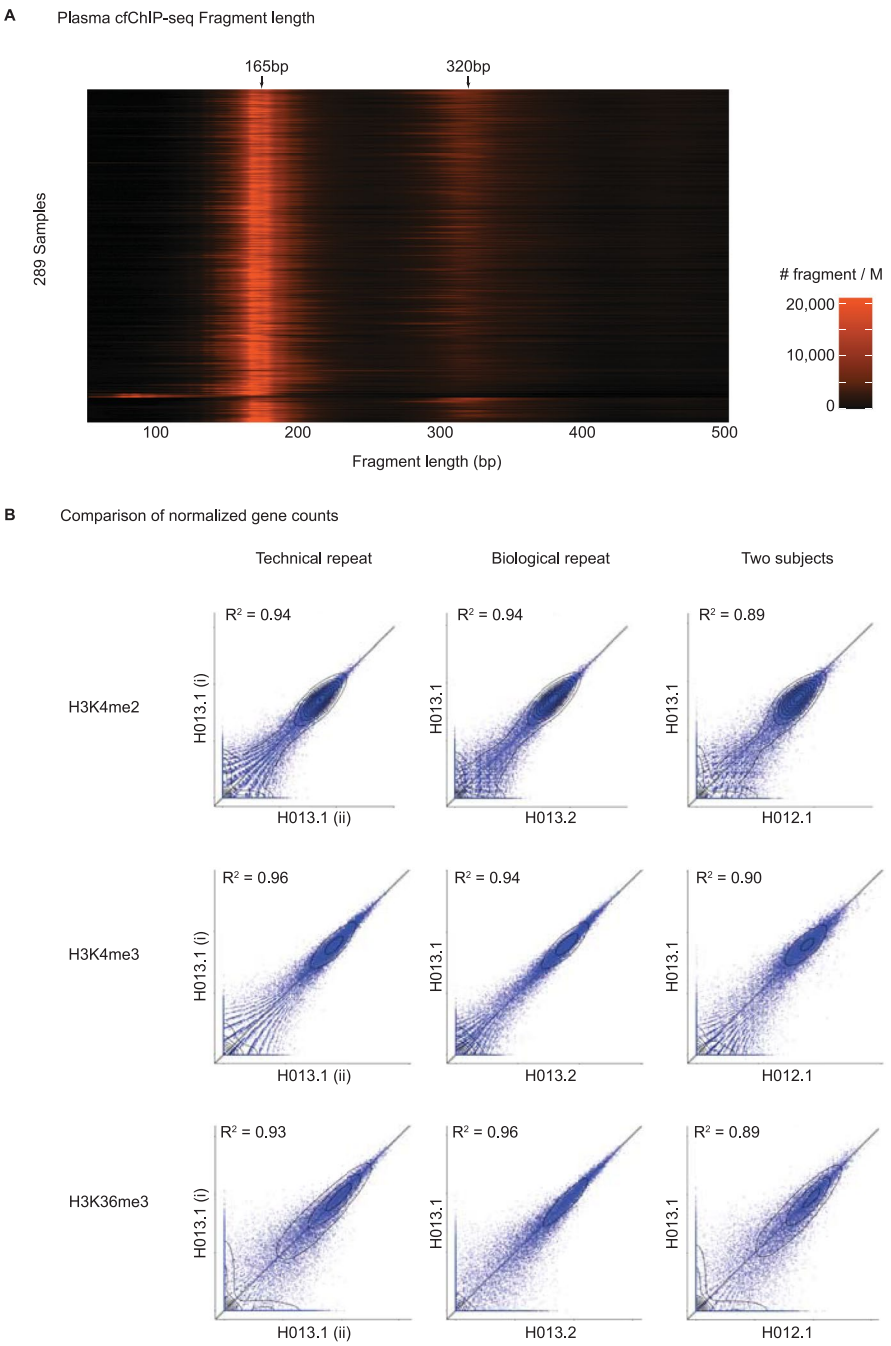




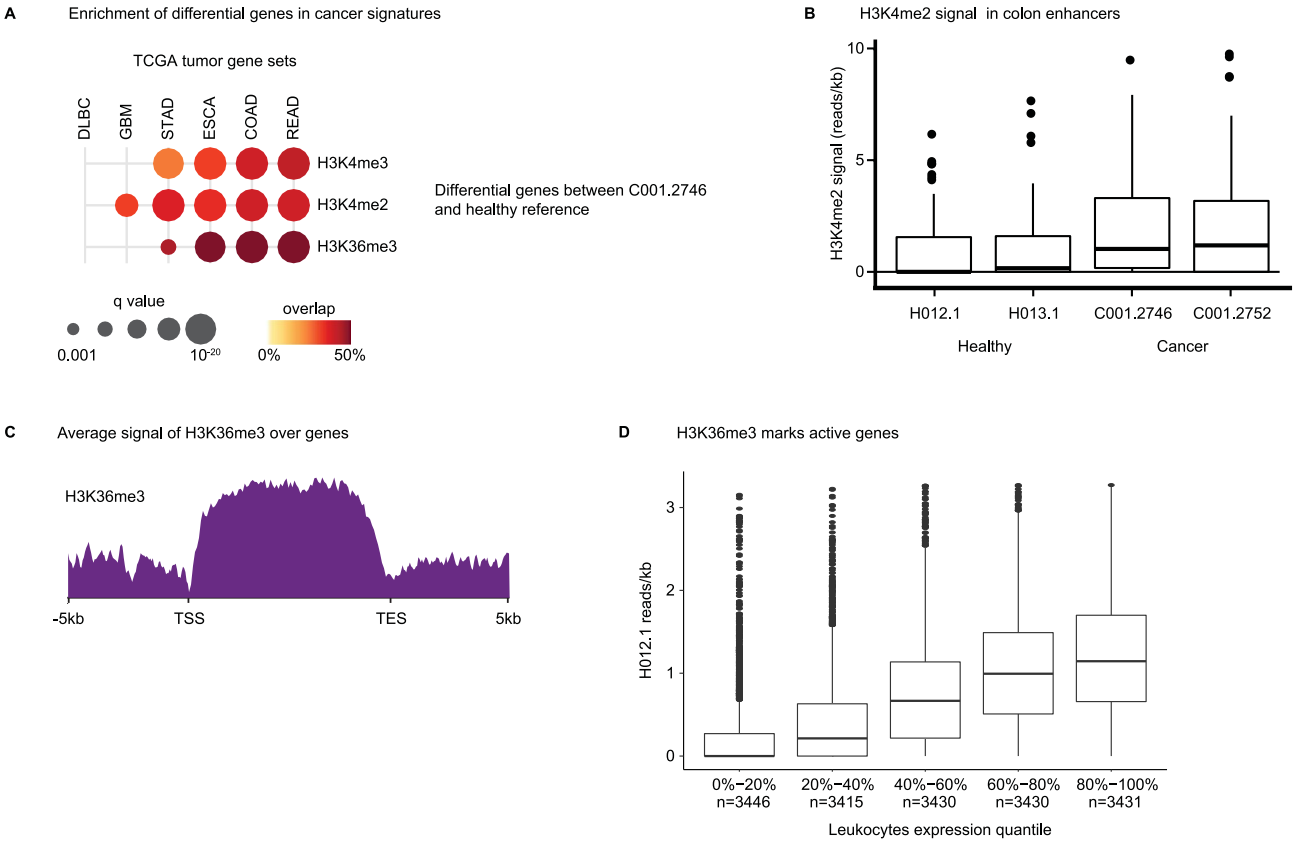


Extended Data Fig. 1 | See next page for caption.

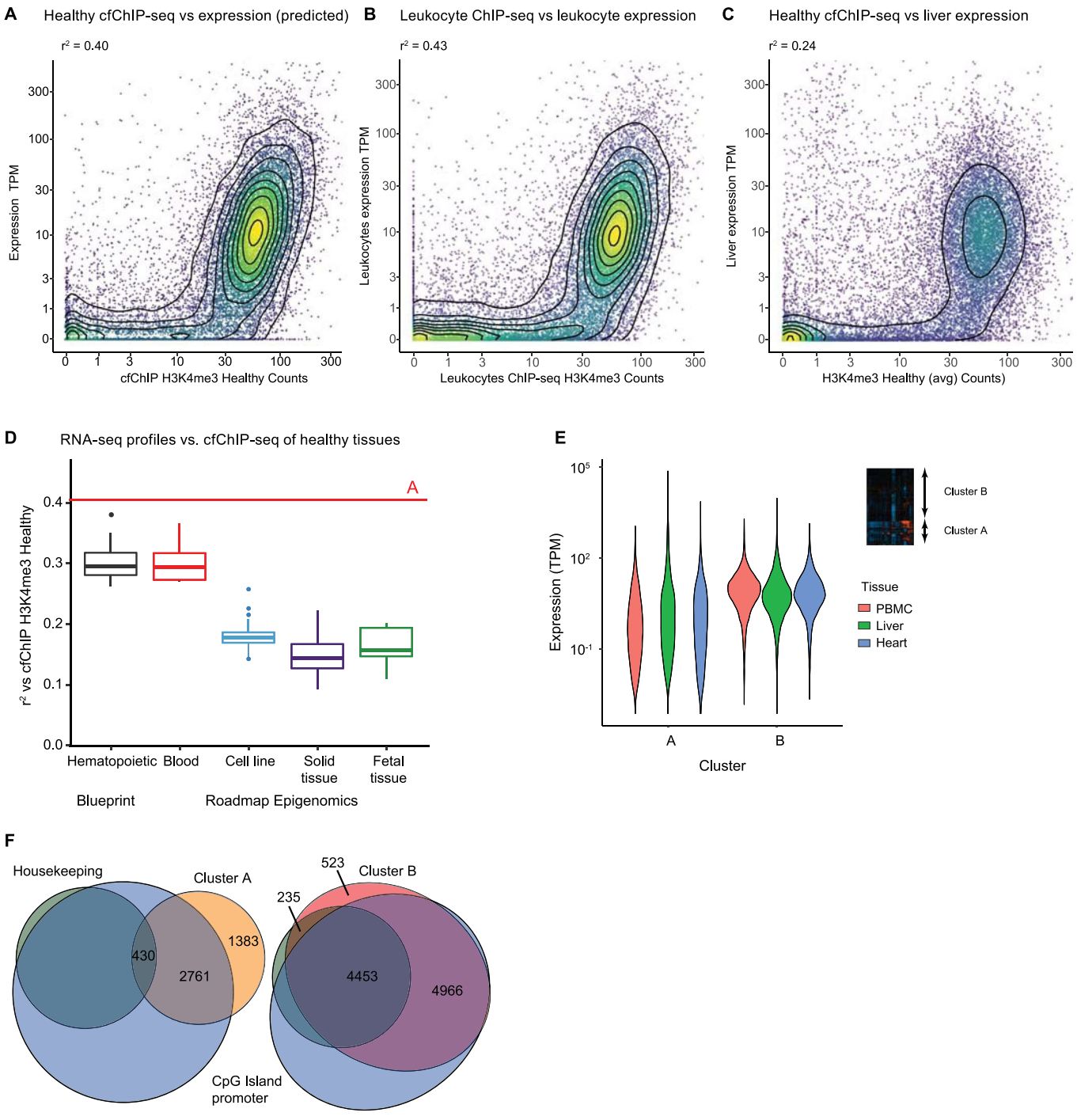
Extended Data Fig. 1 | Supporting data for Fig. 1. a, Distribution of reads for cfChIP-seq with different antibodies on four samples (H012.1, H012.2, H013.1, and H013.2). We divided the genome into regions that contain (putative) TSS based on our catalogue (see below) and (putative) Enhancers. Since there are regions that are marked as both (in different tissues), we consider the intersection separately. For each subset we show the fraction of reads mapped to the region. Within each bar, the fraction estimated as background (based on our background model, Methods) is marked in dark gray. **b**, Genome browser view (as in Fig. 1c). **c**, Metaplots (as in Fig. 1d) of ChIP-seq samples from the Roadmap Epigenomics compendium. **d**, Scatter plots showing signal levels from cfChIP-seq versus Leukocyte ChIP-seq of H3K4me3, H3K4me2, and H3K36me3 (similar to Fig. 1e). **e**, Estimation of the amount of specific reads in cfChIP-seq. Top panel: box plot of the estimate of % reads that are above background levels for all the cfChIP-seq samples analyzed in the manuscript (Supplementary Table 1) compared to selected ChIP-seq samples from Roadmap Epigenomics compendium. Bottom panel: percent of the signal above background that is in the expected genomic locations (i.e H3K4me1 and H3K4me2 - promoters and enhancers, H3K4me3 - promoters, H3K36me3 - gene bodies). For comparison, the same analysis pipeline was applied to selected Roadmap Epigenomic ChIP-seq samples against the same marks. Box limits: 25% –75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge.



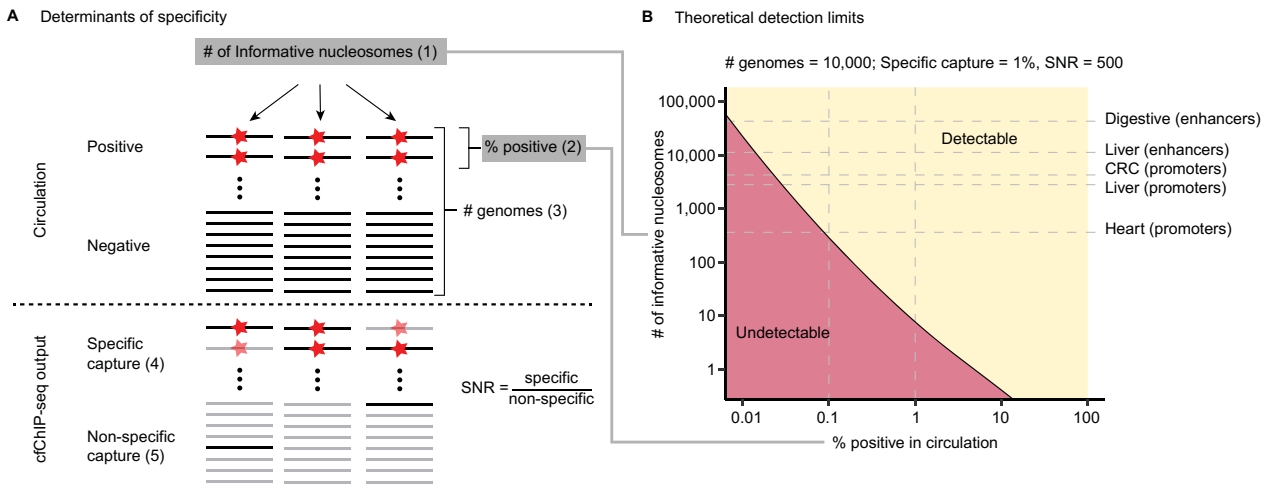
Extended Data Fig. 2 | Supporting data for Fig. 1. a, Fragment length distribution for all samples in this manuscript. Each row represents a histogram of fragment length of a specific sample. Color represents the number of fragments/million with that length (RPM). **b**, Reproducibility of the cfChIP-seq assay. Shown are technical repeats, biological repeats (two samples from the same donor) and comparison of two different donors for three histone marks. Each dot is a gene, and values are normalized counts at the gene promoter (H3K4me2/3) or body (H3K36me3).



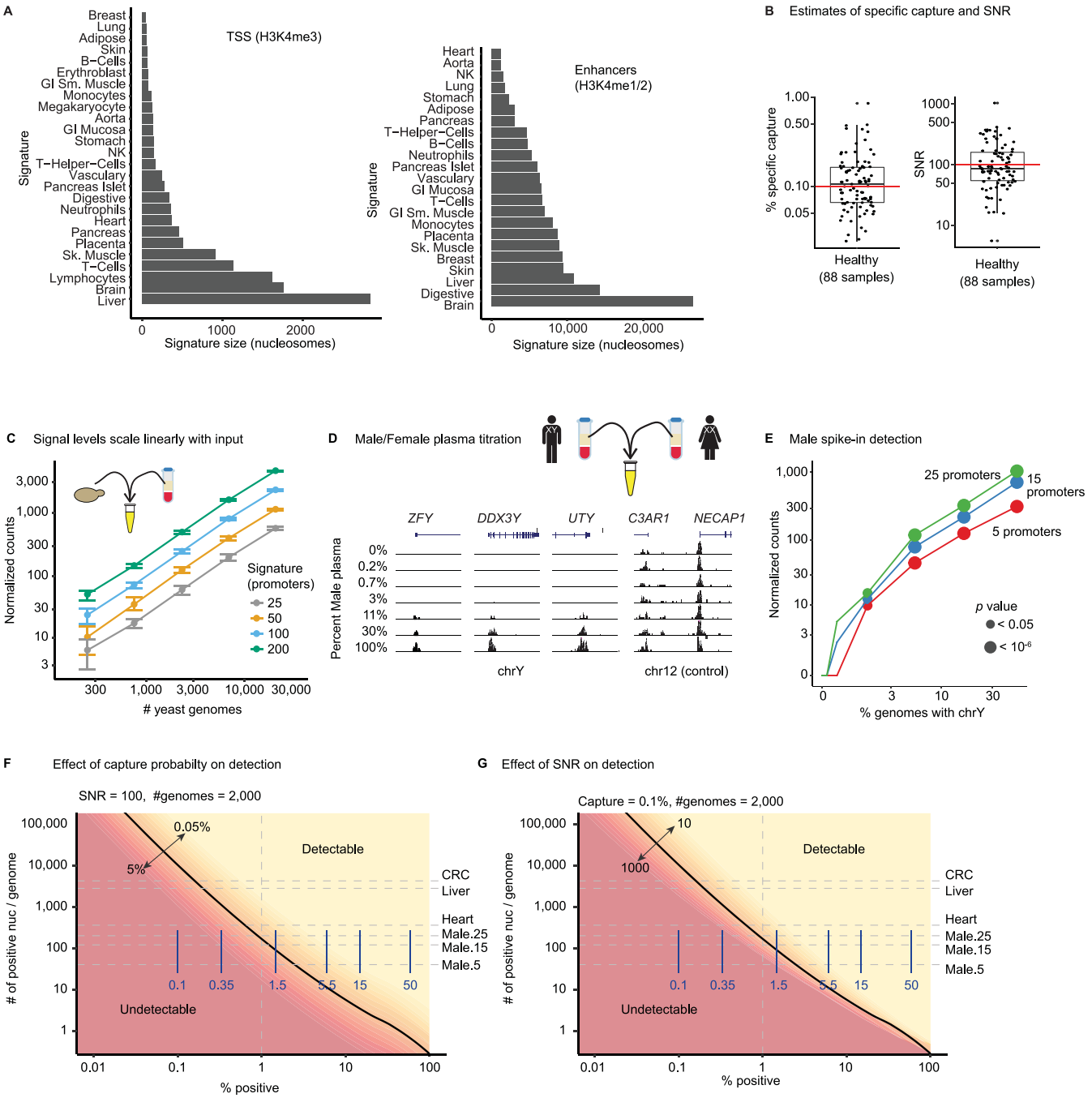
Extended Data Fig. 3 | Supporting data for Fig. 2. a, Testing gene sets defined by highly expressed in different cancer types (TCGA, Methods) against genes with higher signal in a CRC tumor sample (Fig. 2a). Hypergeometric test with FDR corrected q-values. **b**, Levels of H3K4me2 coverage over colon-specific enhancers (y-axis) in healthy donors and in CRC cancer samples. Box limits: 25% –75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge, n = 144. **c**, Average coverage of H3K36me3 across gene bodies (meta gene). **d**, Coverage of H3K36me3 cfChIP-seq over gene bodies in a healthy donor (H012.1) for genes at different leukocyte expression quantiles. Box limits: 25% –75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge.



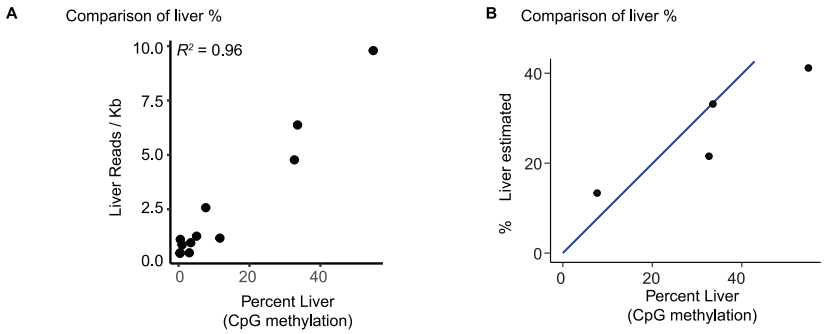
Extended Data Fig. 4 | Supporting data for Fig. 3. a, Comparison of H3K4me3 cfChIP-seq signal from a healthy donor (H012.1) with expected gene expression levels, based on the expression in cells contributing to cfDNA in healthy subjects (Methods). Each dot is a gene. x-axis: normalized number of H3K4me3 reads in gene promoter. y-axis: expected expression in number of transcripts/million (TPM). **b**, Comparison (as in A) of Leukocytes H3K4me3 ChIP-seq signal vs. Leukocytes gene expression levels (both for Roadmap Epigenomic sample E062). **c**, Comparison (as in A) of H3K4me3 cfChIP-seq signal from a healthy donor (H012.1) vs. Liver gene expression levels (Roadmap Epigenomics sample E066). **d**, Summary of correlations of healthy cfChIP-seq levels against different expression patterns from Roadmap Epigenomics and BLUEPRINT. For each category of expression profiles we plot the boxplot of r2 values. Red line denotes the correlation against the predicted expression mixture of cells contributing to cfDNA pool (panel A). Box limits: 25% –75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge. **e**, Comparison of the expression levels of genes in two clusters of Fig. 3c (see inset). Cluster A contains 4,690 genes that change between samples, and Cluster B contains 10,177 genes that do not change between samples. Violin plots show the distribution of expression levels in three tissues - PBMC, Heart, and Liver, from the Roadmap Epigenomics expression data. **f**, Overlap of both clusters with the set of genes with CpG island promoters (blue) and housekeeping genes (green; based on analysis of GTEX compendium, see Methods). For clarity we show each cluster in a separate Venn diagram.



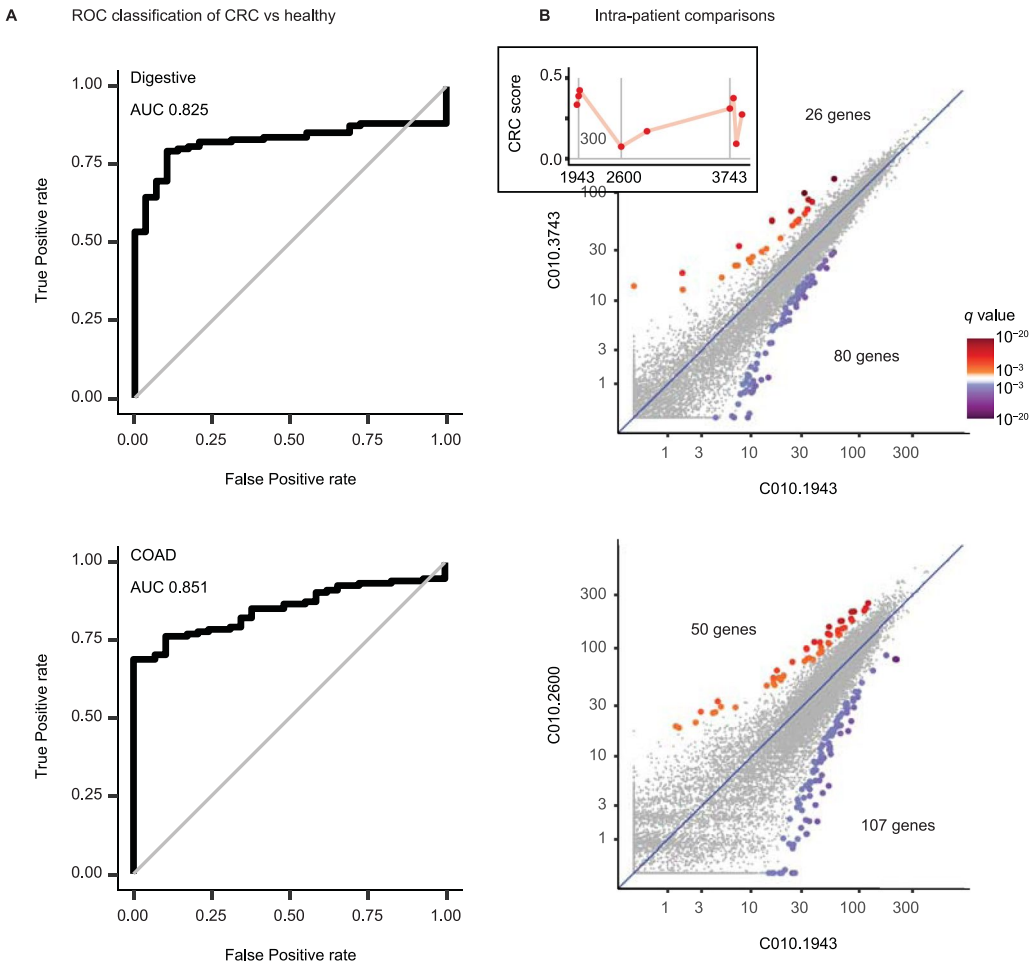
Extended Data Fig. 5 | cfChIP-seq is highly sensitive. a, Schematics of the parameters involved in determining cfChIP-seq sensitivity. 1. Number of informative nucleosomes is the total number of signature-specific nucleosomes in the plasma that carry a mark of interest; 2. The percent contribution of the signature-positive cells to the circulation; 3. Total number of genomes in circulation; 4. The specific capture probability of marked nucleosomes by the cfChIP-seq assay; and 5. The non-specific capture probability of nucleosomes (background). The signal to noise ratio (SNR) is the ratio of the specific to non-specific capture probabilities. **b**, Simulation analysis of event detection power as a function of percent positive (x-axis) and number of informative locations (y-axis). Detection is defined as 95% probability of assay results (capture & sequencing) that reject the null hypothesis of background signal with $p < 0.05$ (Poisson test, Methods). Simulation assumes number of genomes = 10,000 (10 ml plasma of healthy donor), capture probability of 1%, and SNR of 500 (Methods, Supplementary Note). The size of several example signatures are shown.



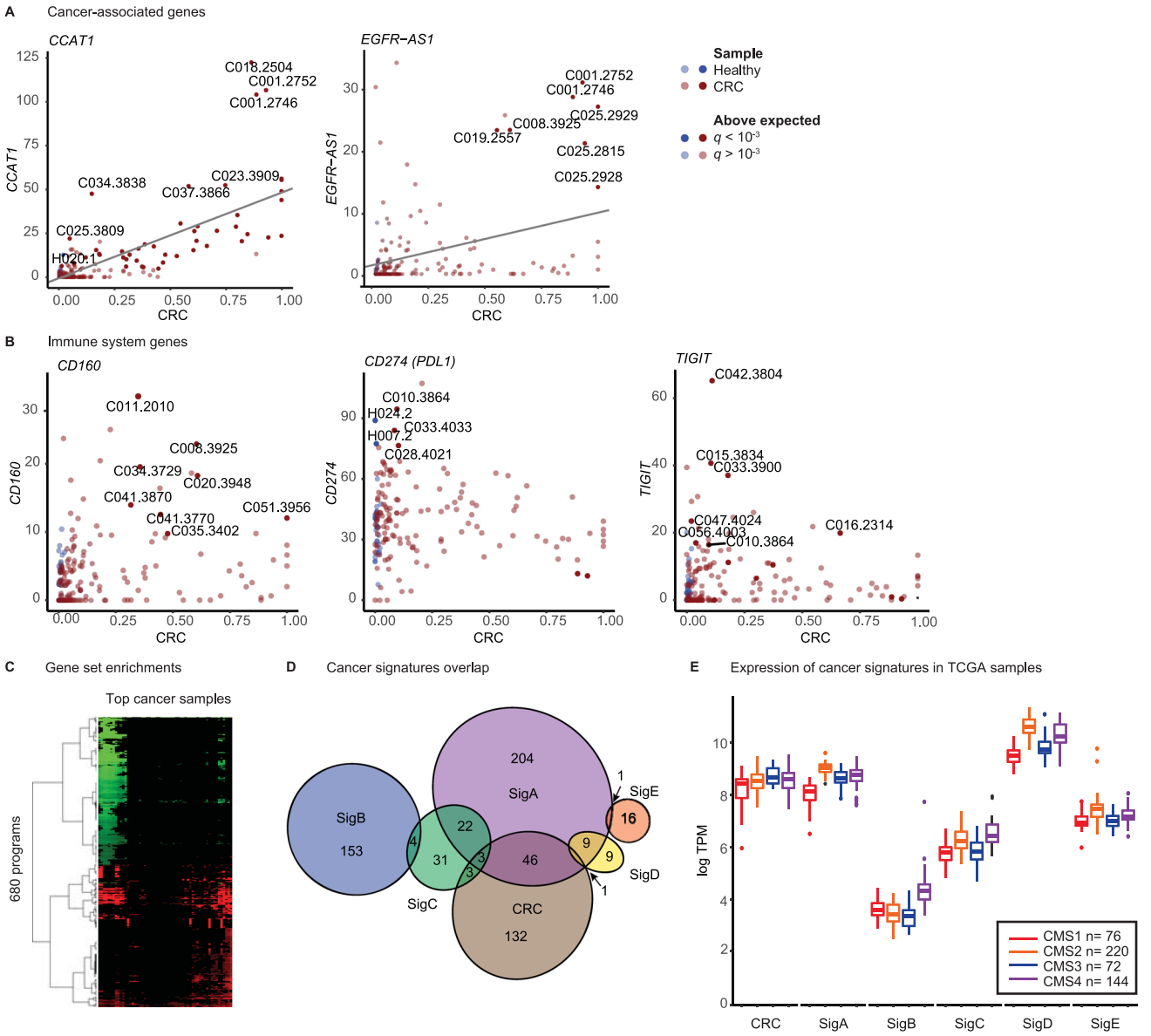
Extended Data Fig. 6 | Sensitivity analysis. a, Total sizes (in nucleosomes) of TSS (Left) and Enhancer (Right) signatures of various cell types. **b**, Estimates of specific capture rate and of SNR (specific capture / non-specific capture) over 88 healthy samples, assuming 1000 genomes/ml and 2 ml input. Box limits: 25% -75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge. **c**, Signal level is linear with input. Plasma of a healthy donor was spiked in with different amounts of yeast nucleosomes (x-axis). The number of counts observed (y-axis) for signatures of different sizes. Error bars show 20-80% range over 100 different sampled signatures of the given size. **d**, Genome browser of chrY male-specific promoters (left) and a representative autosomal region (right) in the male/female titration experiment. **e**, Test of sensitivity using male spike-in. Plasma of healthy female and male donors were titrated at different ratios. Detection of male-specific promoters as a function of percent of chrY genomes in the sample (x-axis). Shown are the number of counts (y-axis) and significance (circle radius) of signal above background distribution (Methods). **f**, Simulation study of the effect of capture probability on detection. The blue marks denote the concentrations used in the male-female titration experiment which had capture probabilities ~0.1% and SNRs of ~500-800. **g**, Simulation study of the effect of SNR levels on detection probability.



Extended Data Fig. 7 | cfChIP-seq liver signal is proportional to % liver cfDNA. a, % Liver as estimated using DNA CpG methylation markers vs. signature strength. **b**, % Liver as estimated using DNA CpG methylation markers vs. estimate of % liver in Fig. 5a.



Extended Data Fig. 8 | Supporting data for Fig. 6. a, Evaluation of classification of CRC samples vs. healthy samples using Digestive (Top) and COAD (Bottom) signature scores (as Fig. 6c). **b**, Intra-patient comparisons (as Fig. 6e). Inset: time samples drawn on the patient timeline (Fig. 6d).



Extended Data Fig. 9 | Supporting data for Fig. 6. a, Levels of CRC associated genes in different samples. Each point is a sample plotted with % CRC (x-axis) vs. normalized number of reads of the gene (y-axis). Solid points - the signal of the gene is significantly above the expectation given % CRC (Methods). **b**, Example of immune-related genes in CRC samples. Same as (A). **c**, Clustering of gene set enrichment in CRC samples (see Supplementary Table 11). **d**, Venn diagram of overlaps between cancer gene signatures that were identified in our analysis. **e**, Evaluation of cancer signatures in CRC samples from TCGA, grouped by their CMS subtype. Box limits: 25% -75% quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge.

nature research

Corresponding author(s): Nir Friedman

Last updated by author(s): Nov 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	bcl2fastq 2.18.0.12
Data analysis	bowtie2 2.3.4.1; samtools 1.7; bedtools 2.26 ; R 4.0.2; ; analysis code available at https://github.com/nirfriedman/cfChIP-seq.git

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	In all experiments, cfChIP-seq samples were removed from consideration if they did not meet pre-established QC standard: the number of unique aligned reads was below 200K or the %signal was below 20%. Exceptions to this rule in samples where %signal * unique reads > 200K.
Replication	We performed biological and technical repeats. Please see discussion of reproducibility in Supplementary Note for full set of results. All replication attempts were reported.
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

IgG - Cell signalling #2729S
H3K4Me1 - Diagenode #C15410194
H3K4Me2 - Diagenode #C15410035
H3K4Me3 - Diagenode #C15410003
H3K36Me3 - Diagenode #C15410192

Validation

IgG - Cell signalling #2729S was validated as a negative control for ChIP by the manufacturer, and showed extremely low signal compared to anti H3, Rbp1 CTD, or anti H3K9me2.

Anti H3K4me3 C15410003 from Diagenode was shown by the manufacturer to be highly specific in ChIP-qPCR and ChIP-seq in HeLa cells by showing high ChIP signal at promoters of highly expressed genes such as EIF4A2 and GAPDH and a very low signal at promoters of inactive genes such as MYOD1. Specificity was also demonstrated by testing against IgG background. https://www.diagenode.com/en/p/h3k4me3-polyclonal-antibody-premium-50-ug-50-ul

Anti H3K36me3 C15410192 from Diagenode was shown by the manufacturer to be highly specific in ChIP-qPCR and ChIP-seq in HeLa cells by showing high ChIP signal at coding regions of actively transcribed genes as a positive control compared to promoter regions and coding regions of inactive genes as a negative control. Specificity was also demonstrated by testing against IgG background. https://www.diagenode.com/en/p/h3k36me3-polyclonal-antibody-premium-50-mg-42-ml

Anti H3K4me1 C15410194 from Diagenode was shown by the manufacturer to be highly specific in ChIP-qPCR and ChIP-seq in K562 cells by showing high ChIP signal at positive regions surrounding the ACTB and GAS2L1 genes and negative controls at the promoters of promoters of EIF4A2 and GAPDH genes. Specificity was also demonstrated by testing against IgG background. https://www.diagenode.com/en/p/h3k4me1-polyclonal-antibody-premium-50-mg-54-ml

Anti H3K4me2 from Diagenode was tested in ChIP-qPCR and ChIP-seq on HeLaS3 chromatin by the manufacturer. The antibody shows high signal at regions surrounding the promoters of highly transcribed genes as positive control and low signal at promoter centers. Specificity was also demonstrated by testing against IgG background. https://www.diagenode.com/en/p/h3k4me2-polyclonal-antibody-classic-50-mg-42-ml

Human research participants

Policy information about studies involving human research participants

Population characteristics	See Supplemental Table 3 for all details
Recruitment	Self reported subjects were recruited through a call within our institute. Patients were recruited through the physician treating them
Ethics oversight	The study was approved by the Ethics Committees of the Hebrew University-Hadassah Medical Center of Jerusalem. Informed consent was obtained from all subjects or their legal guardians before blood sampling.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	FASTQ and BAM files for all the cfChIP-seq samples performed in this study.
Genome browser session (e.g. UCSC)	https://genome.ucsc.edu/s/nirfriedman/cfChIP-seq

Methodology

Replicates	We performed several biological and technical repeats on healthy samples. Full details on these are Table S1
Sequencing depth	We aimed for 5-10M reads/sample. In some cases the numbers were much larger (difference in library size). Table S1 lists sequencing statistics for each sample, including estimate of sequencing capture rates.
Antibodies	See antibodies above
Peak calling parameters	no peak calling - we used quantative read counting in pre-defined regions. See Methods and Supplemental Note for details.
Data quality	See Supplemental Note
Software	See Supplemental Note and https://github.com/nirfriedman/cfChIP-seq.git