Published in final edited form as:
Nat Biotechnol. 2021 May 01; 39(5): 586–598. doi: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], **Israa Sharkia**[#1,2], **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], **Nir Friedman**[1,2,*]

[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 Medical Center

[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

Correspondence to: Nir Friedman.

[*]lead contact: nir.friedman@mail.huji.ac.il.

Author contributions

R.S. and N.F. developed the concept; R.S., N.F., and I.S., designed the experiments with help from E.G., B.G., A.Z., and Y.D.; R.S., developed the cfChIP-seq method with help from A.R. and I.S.; I.S., R.S., 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.S., G.F., I.S., and J.G. analysed 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.S., J.G., G.F., and A.C. wrote the paper with input from all authors.

Competing interests statement

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

Reporting summary

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

Europe PMC Funders Author Manuscripts

[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

[#] These authors contributed equally to this work.

Abstract

Cell-free DNA 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. Here, we apply 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 show that we can identify pathology- related changes in hepatocyte transcriptional programs. In metastatic colorectal carcinoma patients, we detected clinically relevant, and patient-specific information, including transcriptionally active HER2 amplifications. Altogether, cfChIP-seq using low sequencing depth, provides systemic and genome-wide information, can inform diagnosis, and facilitate interrogation of physiological and pathological processes using blood samples.

Blood contains cell-free DNA (cfDNA) fragments derived from dying cells[1]. cfDNA has a half-life of ~15min[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 (e.g., 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, a histone-DNA complex encompassing ~150bp of DNA[12]. Histone proteins are subject to multiple covalent modifications, which are involved in nearly all aspects of 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 may inform on DNA-related activities,

including transcription, within the cells of origin (Figure 1A). This currently inaccessible epigenetic information extends beyond cfDNA modalities examined to date[4–11,34–43].

Here, we perform 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 cells of origin. We applied cfChIP-seq to ~250 samples from more than one hundred subjects including 61 self-declared healthy donors, four patients with acute myocardial infarction, 29 patients suffering from autoimmune, metabolic, or viral liver diseases and 56 metastatic colorectal carcinoma (CRC) patients. 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 CRC patients we detect the disease with high sensitivity and demonstrate that cfChIP-seq can identify subgroups of CRC patients with distinct cancer-related transcriptional programs, and with potential implications to diagnosis and treatment.

Results

ChIP-seq of cf-nucleosomes from plasma

We devised a protocol for cf-nucleosome ChIP-seq from <2ml of plasma (Methods) which overcomes the extremely low concentration of cfnucleosomes and high concentration of native antibodies in plasma (Figures 1A and 1B). Briefly, we covalently immobilized ChIP antibodies to paramagnetic beads, which can be incubated directly in plasma avoiding competition with native antibodies. Additionally, we use an on-bead adaptor ligation[26,44–46], where barcoded sequencing-DNA adaptors are ligated directly to chromatin fragments prior to 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) (Figure 1C). cfChIP-seq profiles with different antibodies show the expected patterns (Figures 1C, 1D and Extended Data Figure 1A).

Several lines of evidence suggest that cfChIP-seq is highly specific: (a) cfChIP-seq signal is consistent with reference ChIP-seq in tissues[25], evident by the agreement of peaks (Figure 1C, Extended Data Figure 1B), in the average pattern around promoters and enhancers (Figure 1D, Extended Data Figure 1C), and in quantitative comparison of the signal across multiple genomic locations, such as all promoters, ($R > 0.8$ Figure 1E, Extended Data Figure 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, $p < 10^{-1000}$). Focusing on non-housekeeping gene promoters, there is significant overlap (1,324/2,311 promoters, $p < 10^{-288}$) with promoters from monocytes and neutrophils who are the major contributors to the cfDNA pool[5,11] (Figure 1F). (b) Performing cfChIP-seq with a mock antibody resulted in dramatically lower yield (Supplementary Table 1). (c) The level

of non-specific reads are mostly comparable to, or lower than standard ChIP-seq (Methods and Extended Data Figure 1E).

Several avenues of evidence rule out the possibility that the cfChIP-seq signal is derived from in-tube lysis during sample handling. (a) We identified 676 promoters carrying H3K4me3 that are absent in ChIP-seq from white blood cells (leukocytes; Figure 1F), these include promoters of genes that are expressed specifically in bone marrow residing megakaryocytes (below). (b) Fragment size distributions of cfChIP-seq correspond to DNA wrapped around mono- and di-nucleosomes (Figure 1G, and Extended Data Figure 2A), consistent with apoptotic or necrotic cell death, but not with cell lysis, which results in much larger (>10kb) fragments[47]. (c) 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 cell free nucleosomes

We find that self-reported healthy donors show highly similar cfChIP-seq profiles (Extended Data Figure 2B, Supplemental Note). We contrasted these to samples from a metastatic CRC patient, where a large fraction of the cfDNA is 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, 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* (Figure 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, Supplementary Table 3). Testing for overlaps we find that the set of genes with high H3K4me3 signal in the cancer patient has a significant overlap (303 of 739 genes, hypergeometric test, $q < 10^{-90}$) with colorectal adenocarcinoma genes (COAD), but only a negligible overlap with non-gastrointestinal cancers genes (Methods, Extended Data Figure 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 Figure 3B).

Tri-methylation of H3 lysine 36 (H3K36me3) requires active transcription elongation to be deposited, and is indicative of gene activity[14]. Indeed, we observe the typical enrichment of H3K36me3 cfChIP-seq signal at gene bodies (Figure 1D and Extended Data Figure 3C) and

the signal in healthy donors correlates with leukocyte RNA-seq levels (Extended Data Figure 3D). Comparing the H3K36me3 signal from healthy donors to that of the cancer patient, we observe 5,416 genes that are hyper H3K36 tri-methylated by at least 4-fold in the cancer sample (Figure 2A).

Examining the genes with increased H3K36me3 signal in this cancer sample, we distinguish between three main classes. Class I includes ~3,400 genes marked by H3K36me3 and H3K4me3 in healthy and cancer samples (*DHX9*, Figure 2B). Class II contains ~1,300 genes similarly marked by H3K4me3 but differ in their H3K36me3 signal, which provides additional information beyond H3K4me3 (*SAP18* and *SKA3*, Figure 2B). Finally, 141 Class III genes are marked with both signals only in the cancer sample (*VWA2*, Figure 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 (Figure 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 cancer patient which are consistent with independent studies of this type of cancer.

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 since 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 find that for a large group of genes (10,150/14,313 genes), H3K4me3 ChIP-seq signal is significantly correlated with expression levels of the gene (Pearson $0.28 \leq r \leq 0.99$; Figure 3A). The remaining genes are either genes that have high H3K4me3 levels in their promoters in most samples (housekeeping, 1,616/4,163 genes, e.g., *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 Figures 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 pathological conditions

Do cfChIP-seq profiles reflect the underlying physiology? We performed H3K4me3 cfChIP-seq on 268 samples from a diverse cohort of subjects (Supplementary Table 4) including 88

samples from 61 healthy donors (ages 23-66), 8 samples from four patients with acute myocardial infarction (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 among these patients is expected to be significantly different due to changes in the contributing tissue of origin. For example, we expect to detect cfDNA from cardiomyocytes following AMI[39], cfDNA from colon tumors in CRC patients[54,55], and an increase in hepatocyte cfDNA in various liver pathologies[42].

We performed hierarchical clustering of 14,875 RefSeq genes promoters that have a noticeable signal in at least one sample (Figure 3B, Methods). 10,177 genes show relatively small differences among samples. These tend to be highly expressed housekeeping genes with CpG-island at their promoters (Extended Data Figures 4E and 4F). The remaining 4,698 genes display a rich tapestry of patterns (Figure 3C).

Platelet progenitor cfDNA in healthy donors

Our analysis identified a cluster with a clear signal in healthy donors (Cluster e, Figure 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", $p < 2x10^{-25}$). However, we are not aware of previous reports of megakaryocytes as a source of cell-free DNA. 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 (Figure 3D) but were detected in a sample from a patient suffering from severe hypoxia (e.g., *GYPA*, *GYPB* and *ASHP*, Figure 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 who are both derived from a common hematopoietic progenitor[57], and thus may 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 indistinguishable otherwise.

cfChIP-seq detects cfDNA cell of origin

To detect the compositions of cells/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] (Figure 4A, Methods, Supplementary Table 5). In healthy donors, we observe a strong signal of neutrophils, monocytes, and megakaryocytes, and a lower but significant signal of liver, in agreement with published cfDNA methylation analysis[11] (Figure 4B). In contrast, we did not observe significant signals in signatures of other tissues (Figure 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 cardiomyocytes signal is clearly detected in samples from AMI patients (Figure 4C). Furthermore, we see good agreement between 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] (Figure 4C). In addition, a significant increase in heart signature signal is observed immediately post percutaneous coronary intervention (PCI) (Figure 4D), as previously reported by assaying cfDNA methylation[39].

This example highlights the sensitivity of the method. We detected a significant heart signal in M003.1 who has 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, Extended Data Figure 5A). Our analysis shows that sensitivity of 0.1% can be readily achieved with a biologically relevant signature size (Extended Data Figures 5B and 6, Methods, Supplemental Note).

In the case of partial hepatectomy, we observed dramatic changes in the cfChIP-seq signal of liver signature following the operation, as expected, which decayed to basal levels after a week (Figure 4E). These changes are consistent with measurement of the liver marker 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 hour) relative to ALT (~47 hours)[59]. We find excellent agreement between liver cfChIP-seq signature levels and liver cfDNA estimates (R^2=0.96, Extended Data Figure 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 (Figure 4F and Supplementary Table 6) and detect signatures of monocytes, neutrophils, and remote organs (e.g., 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 21ng/ml 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 programs 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 based on gene expression studies, and targets of transcription factors based on ChIP studies[60–63] (Figure 4G, 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) which uncovered multiple gene sets that differ from the expected signal in healthy donors (Supplementary Table 7).

For example, in M002.1 cfChIP-seq identifies 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 (Figure 4D). Another example is the increased interferon

signature in M004, who suffered a severe heart damage as reflected by the levels of troponin and cfChIP-seq heart markers (Figure 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 M004, we observed a significant increase in targets of MYOD1 and MYOG, two factors involved in cardiomyocyte development.

Detection of pathology-specific liver signals

The dynamic nature of active histone marks suggested that cfChIP-seq may inform on intra-tissue pathology-related alterations in gene expression. Many of the gene programs enriched in our samples are related to liver function (Figure 4G), thus we decided to test this hypothesis on liver hepatocytes. We assembled a cohort of subjects with verified liver-related diagnosis and/or subjects showing increased liver contribution including subjects at different stages of Nonalcoholic fatty liver disease/Nonalcoholic steatohepatitis (NAFLD/NASH, n=15), autoimmune hepatitis (AIH, n=3), post liver transplant (n=5), infection associated with liver injury (n=1), AMI-associated liver injury (n=1) and partial hepatectomy patients (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 (Figure 5A, Methods). The estimates range from ~2% in healthy samples to 44% in liver patients, consistent with a CpG methylation based estimate of liver cfDNA quantity ($r^2 = 0.87$, Extended Data Figure 7B). For example, in sample L001.1 from an acute AIH patient 44% of the cfDNA was liver-derived and 942 genes were significantly increased compared to healthy reference (Figure 5B, Supplementary Table 2).

To understand whether this increase in liver genes signal is universal to all liver pathologies, we compared L001.1 with M001.1, a sample from an AMI patient that has similar estimated levels of liver contribution (41%). As expected, many liver-specific genes are similarly increased in both samples (Figure 5C, dark gray circles), however pronounced differences in hundreds of genes were observed between the two samples (Figure 5C, Supplementary Table 8). L001.1 is enriched for genes involved in interferon gamma signaling (EnrichR, $q < 3x10^{-20}$), immune system (EnrichR, $q < 1.9x10^{-11}$), MHC class II protein complex (EnrichR, $q < 6.4x10^{-7}$), and allograft rejection (EnrichR, $q < 3.7x10^{-6}$), consistent with the autoinflammatory state of this patient. We also detect 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*[65,66]. Several of these genes (dark colors) such as genes encoding proteins involved in complement and coagulation pathways (e.g., *CFH* and *C4BPA*) are liver specific, demonstrating the potential of cfChIP-seq in detecting intraorgan transcriptional changes.

To get a more systematic view of differences in liver-specific expression programs between samples, we focused on 1,320 genes with significantly higher than expected cfChIP-seq signal in at least one of the liver samples (Figure 5D, left panel). For each gene, we calculated the expected signal based on the estimated liver contribution of that sample

(Figure 5D, middle panel, 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 (Figure 5D right panel).

This analysis identified different types of gene clusters. In some clusters (e.g Clusters I-V % variance explained (PVE) > 45%), the expected signal explains most of the variation between 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 (e.g., *ASHP* and *HBD*; 37/78 genes, $q < 10^{-12}$). In some clusters (e.g 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 subjects, 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 are annotated as hepatocyte genes[67] (Figure 5E, Clusters I-VI, XI, XII). We performed enrichment analysis of the gene sets in each cluster (Figure 5F). As expected we see strong enrichments for many liver related terms (Supplementary Table 9). Some clusters show strong enrichments only to specific terms. For example, the genes of Cluster I are enriched for genes involved in the process of cholesterol homeostasis (9/111 genes, $q < 4x10^{-8}$) and the genes in Clusters I and IV are enriched with genes of the complement and coagulation cascade (14/111 genes, $q < 3x10^{-15}$, and 11/77 genes, $q < 2x10^{-12}$, respectively).

Next, we examined a single-cell RNA-seq 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 see that Clusters I and IV are 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 (Figure 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 between samples allows us to identify sample-specific changes in hepatocyte-specific transcriptional programs (Figure 5G). For example, we see high levels of Cluster I genes in patients with immune-related pathology (e.g., 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 are both enriched for the periportal zone markers (Figure 5F) they capture transcriptional programs that are differential among subjects 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).

Europe PMC Funders Author Manuscripts

Analysis of colorectal-cancer by cfChIP-seq

We analyzed a collection of samples from an ongoing longitudinal study following metastatic CRC patients 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 (Figure 3C). Closely collected samples from a single patient show higher similarity than samples collected far apart (Figure 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 (Figure 6B). We selected a subset of COAD-genes (based on analysis of the TCGA gene expression data of colorectal adenocarcinoma) 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 AUC=0.94 (Figure 6C and Extended Data Figure 8A).

We observed large differences in CRC signature magnitude between patients and during treatment of the same patient, consistent with the course of therapy (Figure 6D)[70]. We also detect differences that appear to result from disease progression (Extended Data Figure 8B), e.g. an increased liver signal in C010.1943 vs. C010.3743 (ARCHS4 tissue $q < 2.4\text{x}10^{-12}$) reflecting chemotherapy-induced liver damage[71], intratumor variation, or immune-related signaling such as the enrichment for interferon gamma genes in C010.3743 vs. C010.1943 (REACTOME $q < 2.8\text{x}10^{-6}$, Extended Data Figure 8B).

cfChIP-seq detects molecular variability among colorectal-cancer patients

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 (Figure 6E). These differences can be due to contribution of additional tissues (e.g enrichment for liver genes in C001.2752 vs. C040.3606, ARCHS4 tissue $q < 4.1\text{x}10^{-9}$), while others may reflect intertumor transcriptional differences, for example enrichments for Wnt/calcium/cyclic GMP pathway in C040.3606 vs. C001.2752 (BioPlanet, $q<0.00012$) and for Cell adhesion molecules (CAMs) in C025.2815 vs. C001.2752 (BioPlanet, $q<0.0045$). Additional examples include, *EGFR-AS1*, and the CRC marker *CCAT1*[74] (Figure 6E and Extended Data Figure 9A). *EGFR-AS1* regulates the splicing of EGFR and may affect anti-EGFR treatment[49]. When examining all samples, we identify variation in genes associated with immune activity such as the checkpoint receptors *CD160*, *TIGIT*, and *PDL1* (*CD274*) (Extended Data Figure 9B), suggesting that we may 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 (out of 7,538) gene sets that had informative signals in these samples (Supplementary Table 10, Methods, Extended Data Figure 9C). We

used these to initialize an iterative process to identify signatures that distinguish between sample subgroups (Methods) resulting with five gene signatures that capture the main behaviors in the original set of programs (Figure 6F). Signatures A-C capture cancer gene expression programs and signatures D-E capture duplications events.

The scores of the largest signature (SigA) are highly correlated with the CRC scores, although there is only a partial overlap between the two (Extended Data Figure 9D). This signature is enriched with genes associated with Colon (ARCHS4 tissue, $q < 10^{-64}$) and targets of CDX2, a transcription factor active in CRC (TRRUST, $q < 10^{-9}$) (Figure 6G and Supplementary Table 11). The second signature (SigB) differentiates a small subset of the high tumor load samples and is 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 derepression 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) selects a larger subset of samples, which includes most of the samples selected by SigB although there is little overlap of genes between the two signatures (Extended Data Figure 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 Figure 9E). This analysis shows that SigA genes tend to have lower expression in CMS1 tumors, while SigB genes tend to have higher expression in CMS4. CMS4 tumors are characterized by upregulation of epithelial to mesenchymal transition (EMT) 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 (Figure 6G).

Ten out of nineteen genes in SigD and 13 of the 17 genes in SigE are clustered around regions of known genomic duplications at chr20q13.12, and chr17q12-q21, respectively (Figure 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 (Figure 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 CRC patients related to the load of the tumor (CRC score), the contribution of additional tissues (e.g., liver damage, 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 cell-free 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 metastatic colorectal cancer patients.

Beyond determining the cells of origin, cfChIP-seq can detect differences in patient and disease-specific transcriptional programs, e.g. between subjects with different etiology of increased liver cfDNA (Figure 5). Our analysis shows that even at this early stage, cfChIP-seq is highly sensitive in detecting signatures of interest, including cancer-specific signatures (Figures 4 and 6, and Extended Data Figures 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 preselecting 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 preselected 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 exploit 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 (e.g GP6, GP9), but not erythrocyte-specific proteins (e.g., 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-seq atlas[68] to detect contributions from different liver zonation

expression programs in each of the subjects. 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 combine 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 our 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 our 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 which opens up a wide range of opportunities for studying basic questions in human physiology that have been inaccessible until now.

Materials and Methods

Patients

All clinical studies were approved by the relevant local ethics committees. 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.

Sample collection

Blood samples were collected in VACUETTE® K3 EDTA tubes, transferred immediately to ice and 1X protease inhibitor cocktail (Roche) and 10mM EDTA were added. The blood was centrifuged (10 minutes, 1500 × g, 4°C), the supernatant was transferred to fresh 14ml tubes, centrifuged again (10 minutes, 3000 × 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 storage.

Bead preparation

50μg of antibody were conjugated to 5mg of epoxy M270 Dynabeads (Invitrogen) according to manufacturer instructions. The antibody-beads complexes were kept at 4°C in PBS, 0.02% azide solution.

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

Immunoprecipitation, NGS library preparation, and sequencing

0.2mg of conjugated beads (~2μg 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 8 times with blood wash buffer (BWB: 50mM Tris-HCl, 150mM NaCl, 1% Triton X-100, 0.1% Sodium DeoxyCholate, 2mM EDTA, 1X protease inhibitors cocktail), and three times with 10mM Tris pH 7.4. All washes were done with 150ul of the washing buffer on ice by shifting the beads from side to side on a magnet. Do not use vacuum to remove supernatant during washes in buffers that do not contain detergents.

On-beads chromatin barcoding and library amplification was done as previously described[26,44] except for the DNA elution and cleanup step where the beads were incubated for 1 hour at 55°C in 50μl of chromatin elution buffer (10mM Tris pH 8.0, 5mM EDTA, 300mM NaCl, 0.6% SDS) supplemented with 50 units of proteinase K (Epicenter), and the DNA was purified by 0.9 X SPRI cleanup (Ampure xp, agencourt). The purified DNA is eluted in 25 μl EB (10mM tris pH 8.0) and 23 μl of the eluted DNA were used for PCR amplification with Kapa hotstart polymerase (16 cycles). The amplified DNA was purified by 0.8 X SPRI cleanup and eluted in 12 μl EB. The eluted DNA concentration was measured by Qubit and the fragments size was analyzed by tapestation visualization. Note: If adapter dimers are substantially visible by tapestation post library amplification, we recommend pooling samples and performing additional X 0.8 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 (>150bp). 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, ESC, and cell-line samples, resulting with 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, normalization, etc.

Tumor-type Gene Signatures

We downloaded RNA-seq data from the UCSC Toil RNAseq Recompute Compendium[85] which include samples from the TCGA and GTEx projects. We defined the set of genes that are 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 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 is 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 Figure 4A, we performed *in silico* mix of the four cells types that contribute the most to cfDNA[11]: neutrophils, 32%; monocytes 32%; megakaryocytes 20%; and NK cells 5%. The gene expression for these cell types was downloaded from the BLUEPRINT consortium.

TSS location catalogue

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

We quantified the number of reads covering each region in the catalogue 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 catalogue for each sample (Supplemental Note). These were then normalized (Supplemental Note) and scaled to 1M 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 transcription start sites (Figure 1D), to distinguish between non-specific capture (in regions without TSSs)

Europe PMC Funders Author Manuscripts

and specific capture (in TSSs that are known to be constitutively marked by H3K4me3). We use this idea in two different approaches (see Supplemental Note for more details).

In the **global approach**, we compare input to output of the cfChIP-seq assay. At the input end, we estimate 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 ~200bp of DNA). We also estimate the percent of these that are modified, which for H3K4me3 tend to be ~1-2%. At the output end, we estimate how many of the unique fragments are background and how many are signal (see above). We then divide #signal fragments in output by #modified nucleosomes in input to get specific capture rate, and similarly #background fragment in output by total #nucleosomes to get non-specific capture rate.

In the **local approach**, we compare expected input coverage to output coverage. Using input cfDNA amounts we can estimate the number of alleles (genomes) that cover each position. We then examine two types of regions, one “high-signal” where we assume that ~100% of the nucleosomes are modified (e.g., promoters of constitutive genes) and the other one as “no-signal” where 0% of the nucleosomes are modified (e.g., background regions). The coverage we observe 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 which influences the number of observed reads (Supplemental Note). We estimate the probability of specific capture (above) to range between 0.01% and 0.1% across dozens of H3K4me3 cfChIP-seq experiments (Extended Data Figure 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 (SNR) of the assay (Extended Data Figure 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 Figure 6A). For example, we defined the cardiomyocyte-specific signature as 366 nucleosomes that are marked with H3K4me3 only in cardiomyocytes (Extended Data Figure 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/ml, we expect ~36,600 informative nucleosomes in a 10ml plasma sample.

We estimate capture rate as discussed above. We further assume independence of the concentration of plasma nucleosomes and capture rate (Extended Data Figure 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 (Supplemental 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 Figures 6D and 6E), 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 Figure 6E) consistent with our estimates based on the parameters of the specific experiment (Extended Data Figures 6F and 6G).

Tissue Signatures

To define tissue specific signatures of a specific modification, we examined binned representation of the atlas according to our catalogue. 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 (Supplemental 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 (Supplemental Note).

Comparison to RNA-seq

The comparison of H3K4me3 ChIP to RNA-seq was performed as follows. RNA expression (normalized TPM) was downloaded from 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-seq 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 (Supplemental Note).

Statistical analysis

We use several custom designed statistical tests in our analysis. In all analysis we correct for multiple hypotheses using False Discovery Rate (FDR) and estimate q-values (R function p.adjust()).

- **Comparison to background** (Figures 1F, 4B, 4C, 4F, 6B, and Extended Data Figure 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 (Supplemental Note).

Europe PMC Funders Author Manuscripts

- **Comparison to reference (healthy)** (Figures 2A, 4G, 5B, 6F, and Extended Data Figures 9A and 9B). 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. In addition, 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 (Supplemental Note). Together, these define the distribution of total reads in these windows under the null-hypothesis that the subject 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 (Supplemental Note). Note: in the analysis of Extended Data Figures 9A and 9B we use a variant of this test where the reference is modified according to the %tumor estimated for the sample.
- **Comparison of two samples** (Figures 5C, 6E, and Extended Data Figure 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 likelihood ratio test (LRT).

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 Harmonizome database[86]. These include targets from ENCODE[87], TRANSFAC[88], and CHEA[89].

Estimation of liver percentage

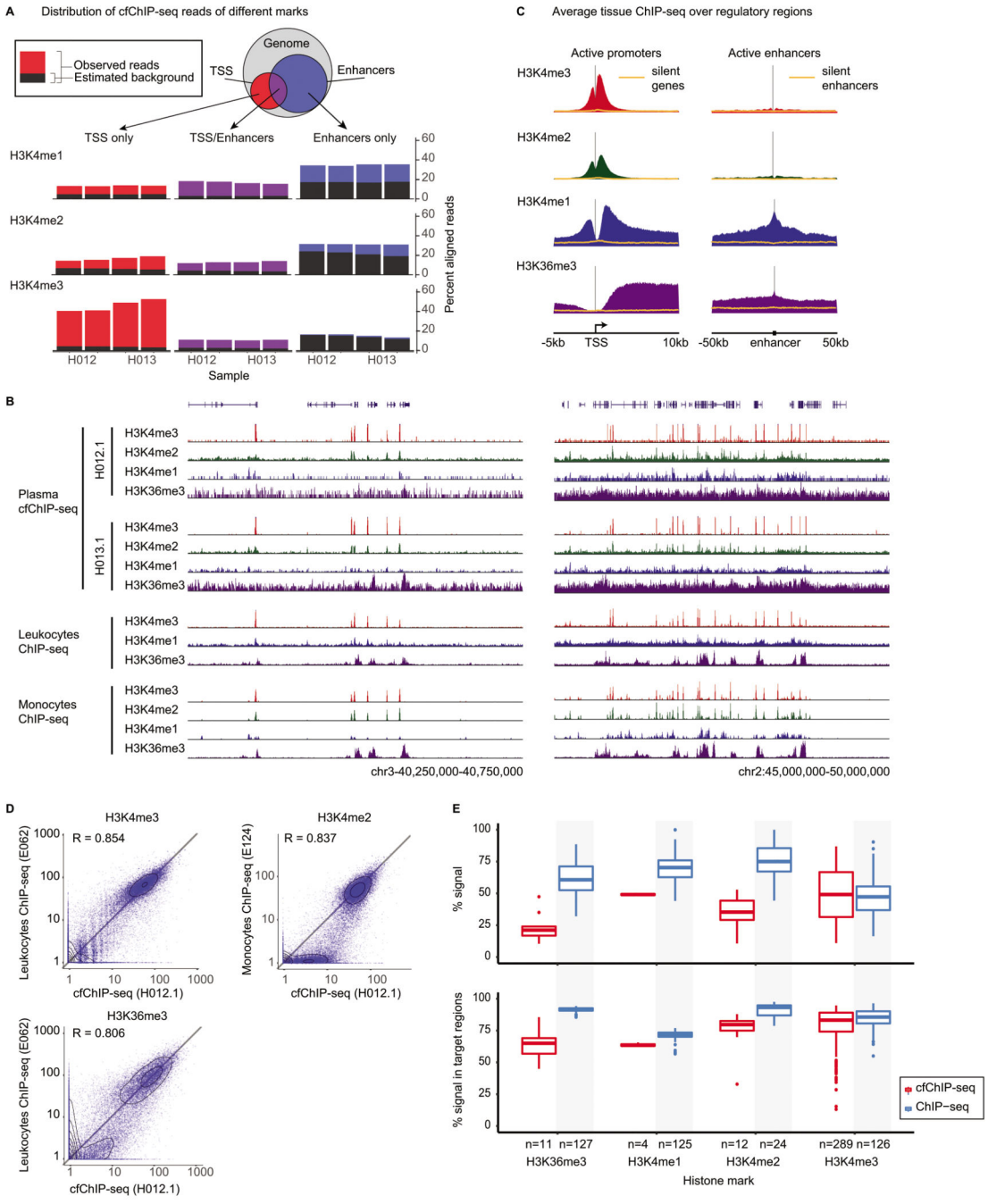
We used a linear regression model that matches the observed counts of a select representative genes to a sum of contribution of healthy-wo-liver and healthy liver. Briefly, we use the Roadmap Epigenomics Atlas “liver” (E066) as 100% liver. We assume that the mean healthy profile contains about ~3% liver contribution, and so define the healthy-wo-liver as the result of subtracting 3% of liver profile from the healthy sample. We then identify the set of distinguishing genes as those that are close to 0 in healthy-wo-liver and high in liver and those that are high in healthy-wo-liver and low in liver. These are 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.

Europe PMC Funders Author Manuscripts

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 3 CRC samples but less than $^{2}/_{3}$ of all the CRC samples were selected for the next step. The pattern of significantly above/below enrichments were clustered (Extended Data Figure 9C). Each cluster of gene programs corresponds to a classification of the CRC samples (significant vs. non-significant). For each such cluster we identified the genes that have significantly higher signal in the positive class of CRC samples compared to remaining CRC samples. The differential genes define a new gene-signature. These were clustered based on their classifications of samples, and combined into non-overlapping sets of gene signatures (Supplemental Note).

Extended Data



Extended Data 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

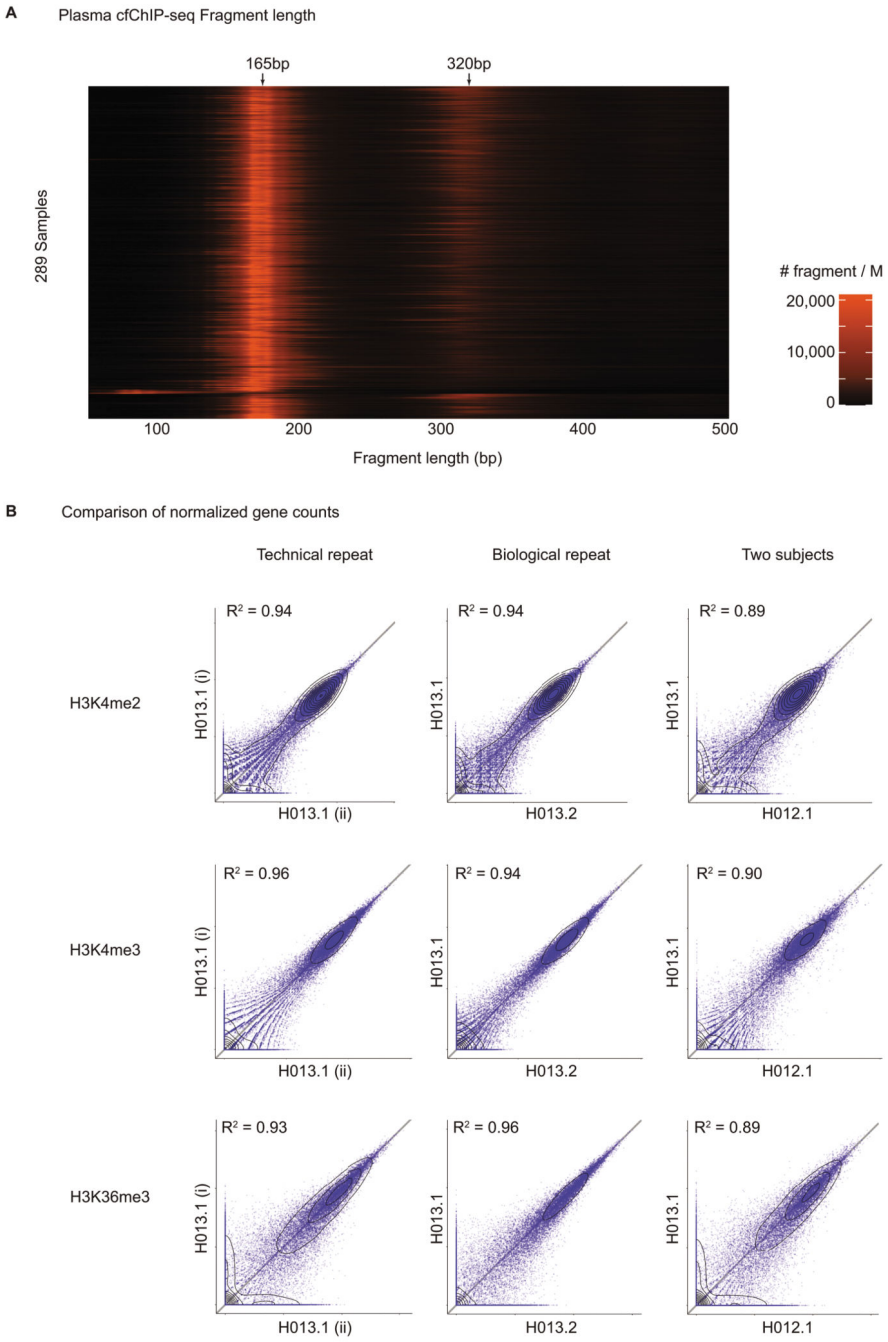
fraction estimated as background (based on our background model, Methods) is marked in dark gray.

B. Genome browser view (as in Figure 1C).

C. Metaplots (as in Figure 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 Figure 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.

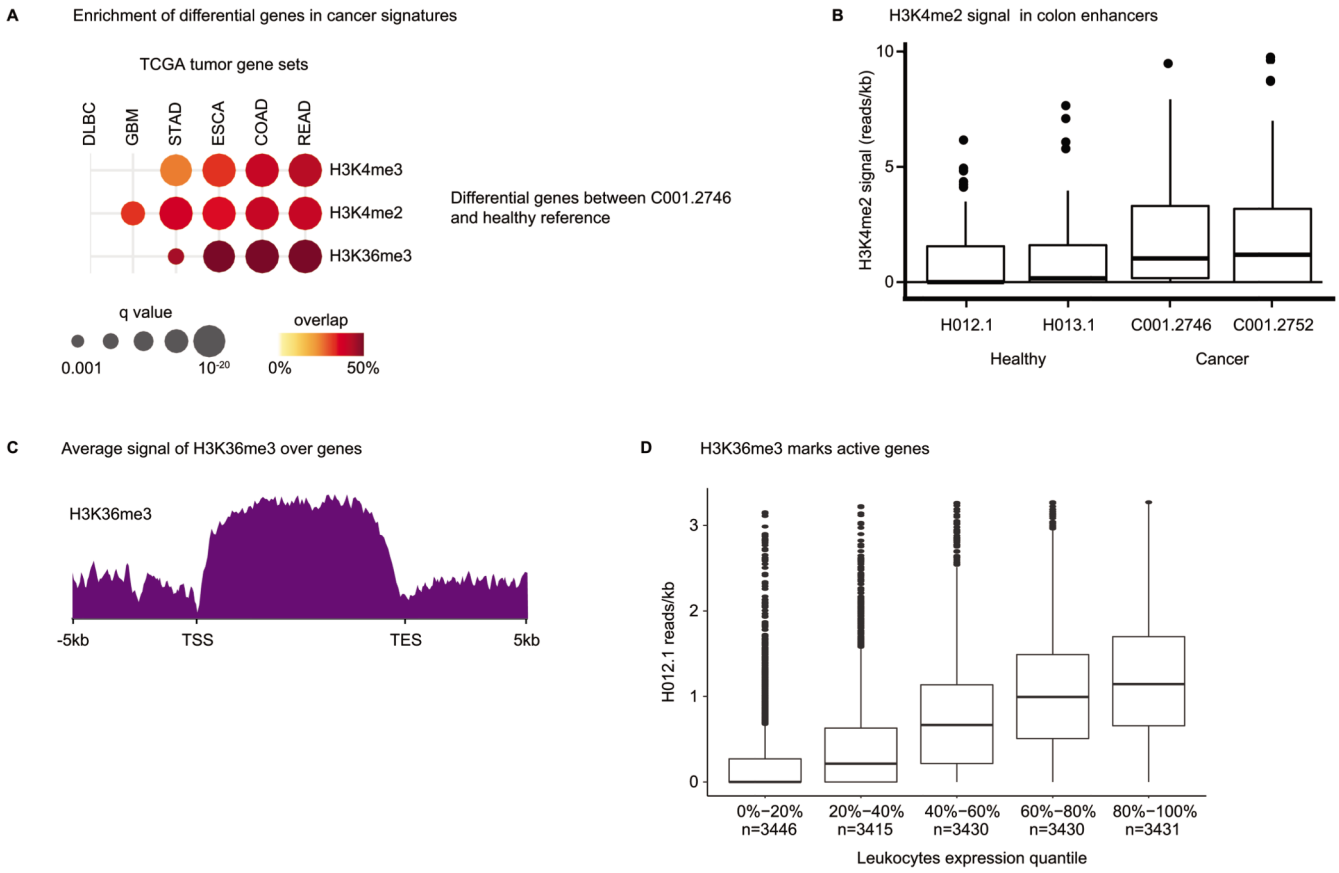


Extended Data Fig. 2.

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).

Europe PMC Funders Author Manuscripts



Extended Data Fig. 3.

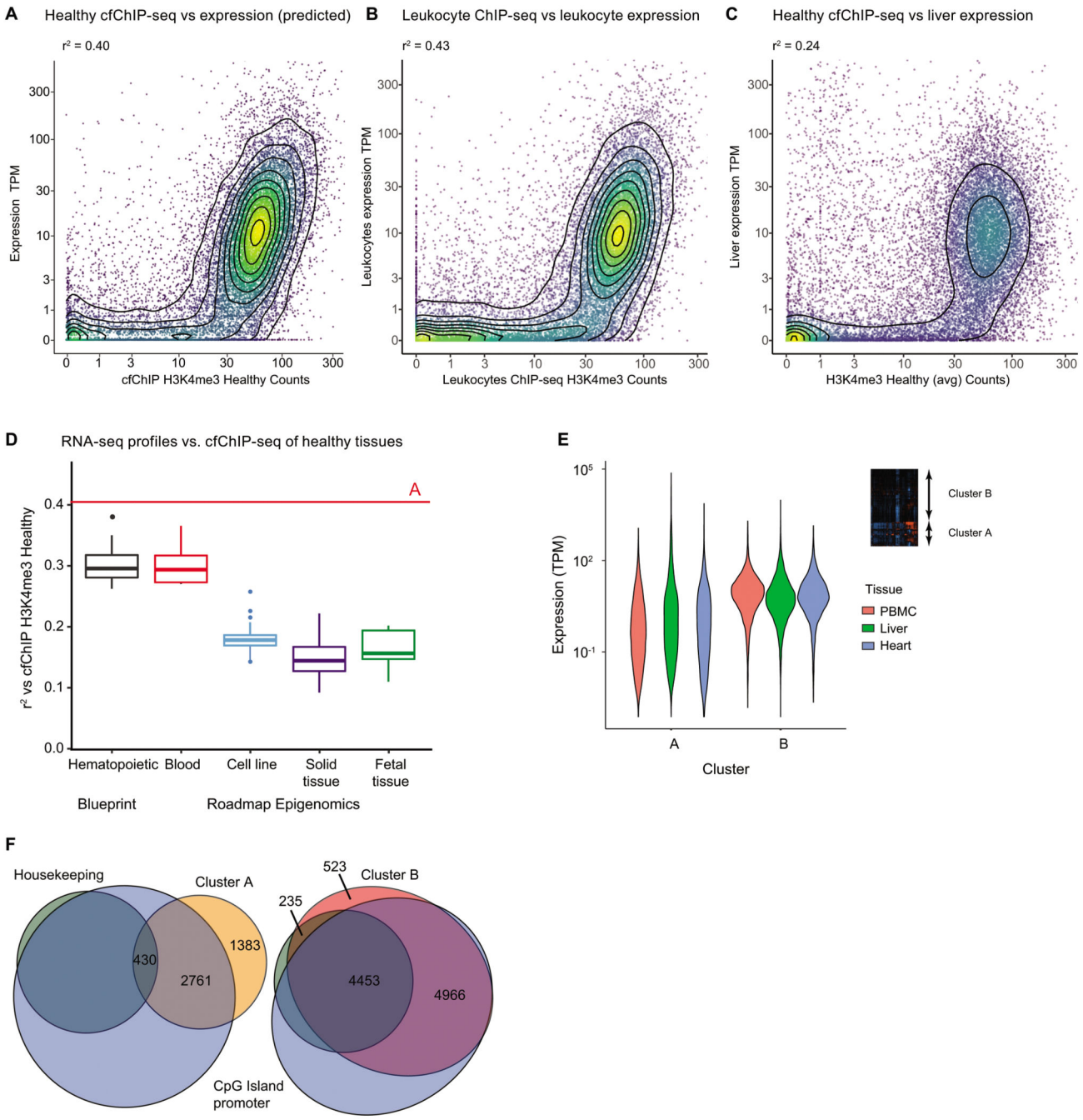
A. Testing gene sets defined by highly expressed in different cancer types (TCGA, Methods) against genes with higher signal in a CRC tumor sample (Figure 2A). Hypergeometric test with FDR corrected q-values.

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.

Europe PMC Funders Author Manuscripts

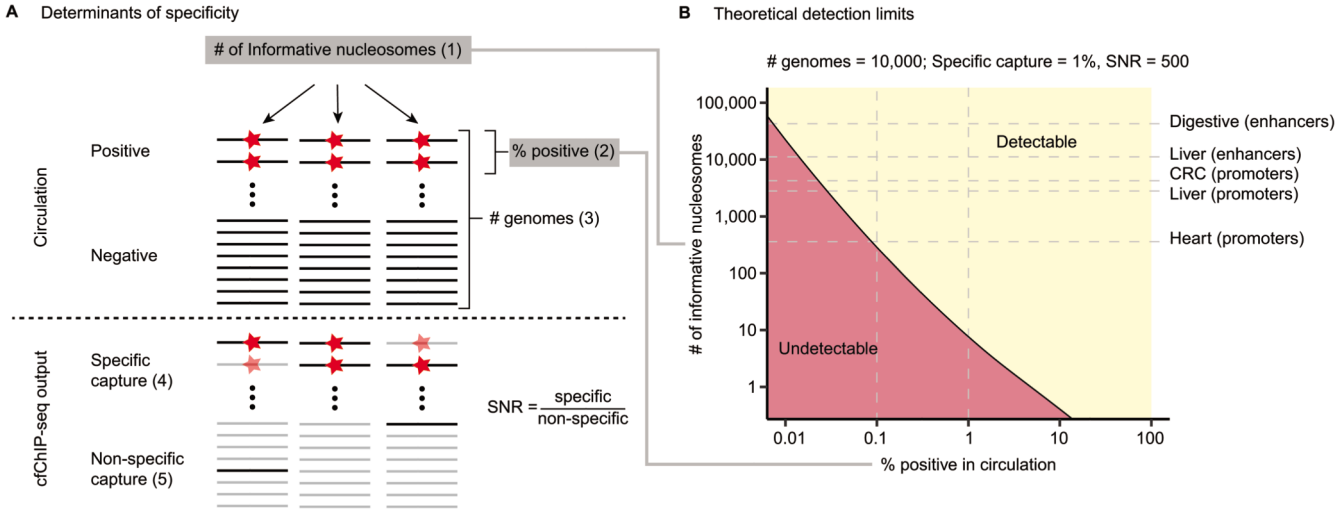


Extended Data Fig. 4.

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).

Europe PMC Funders Author Manuscripts

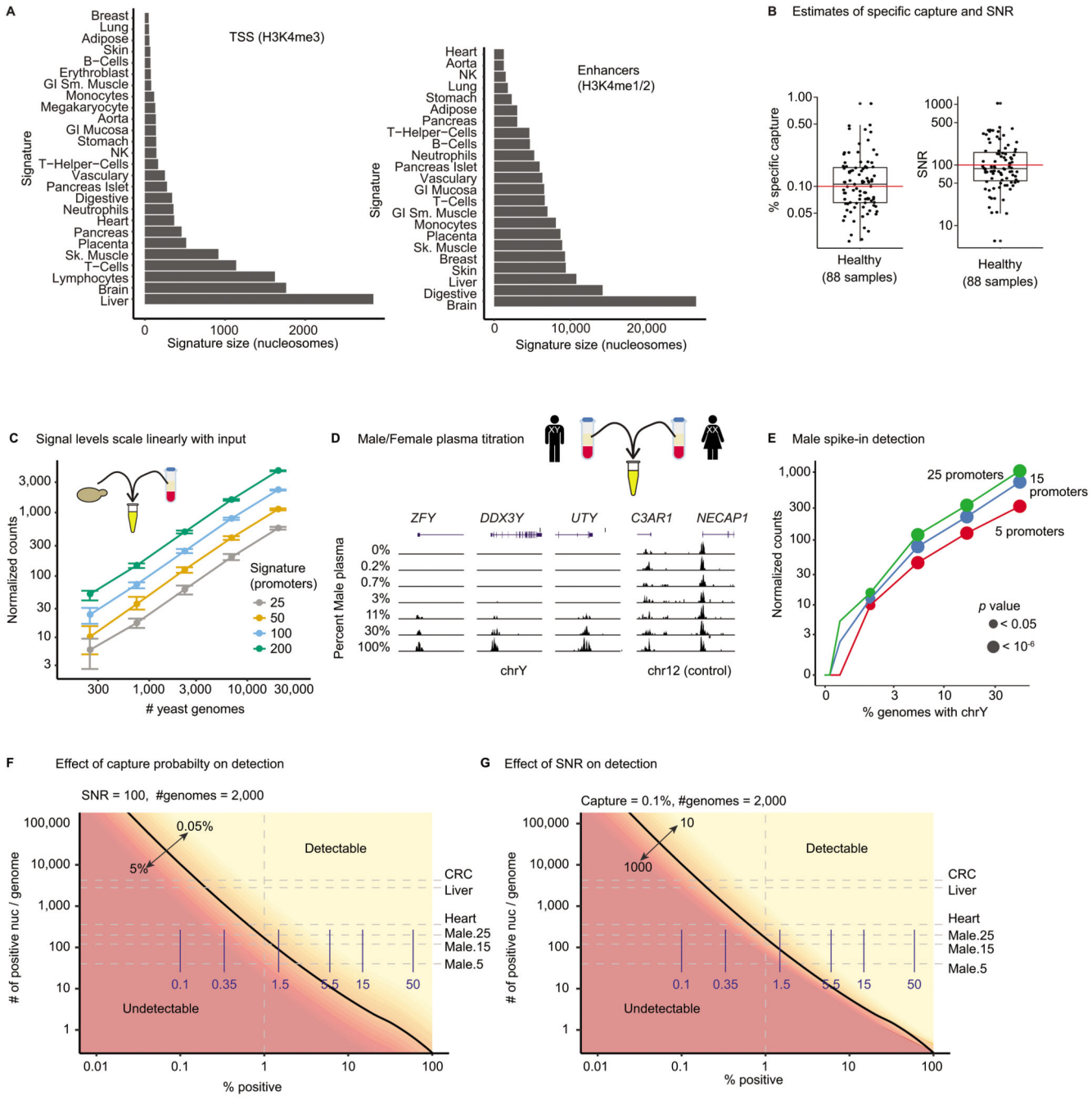
C. Comparison (as in A) of H3K4me3 cfChIP-seq signal from a healthy donor (H012.1) vs. Liver gene expression levels (Roadmap Epigenomics sample E066).
D. Summary of correlations of healthy cfChIP-seq levels against different expression patterns from Roadmap Epigenomics and BLUEPRINT. For each category 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 Figure 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.

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, Supplemental Note). The size of several example signatures are shown.



Extended Data Fig. 6.

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 2ml input. Box limits: 25% - 75%

Europe PMC Funders Author Manuscripts

quantiles, middle: median, upper (lower) whisker to the largest (smallest) value no further than 1.5 * inter-quartile range from the hinge.

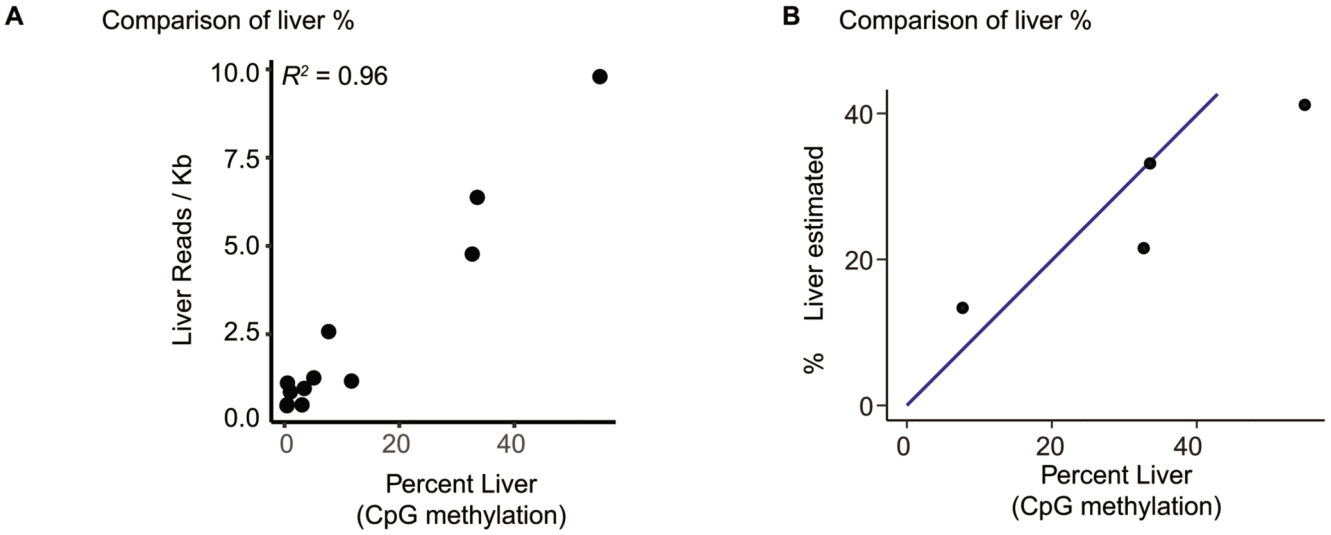
C. Signal level is linear with input. Plasma of a healthy donor was spiked in with different amounts of yeast nucleosomes (x- axis). The number of counts observed (y-axis) for signatures of different sizes. Error bars show 20-80% range over 100 different sampled signatures of the given size.

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.

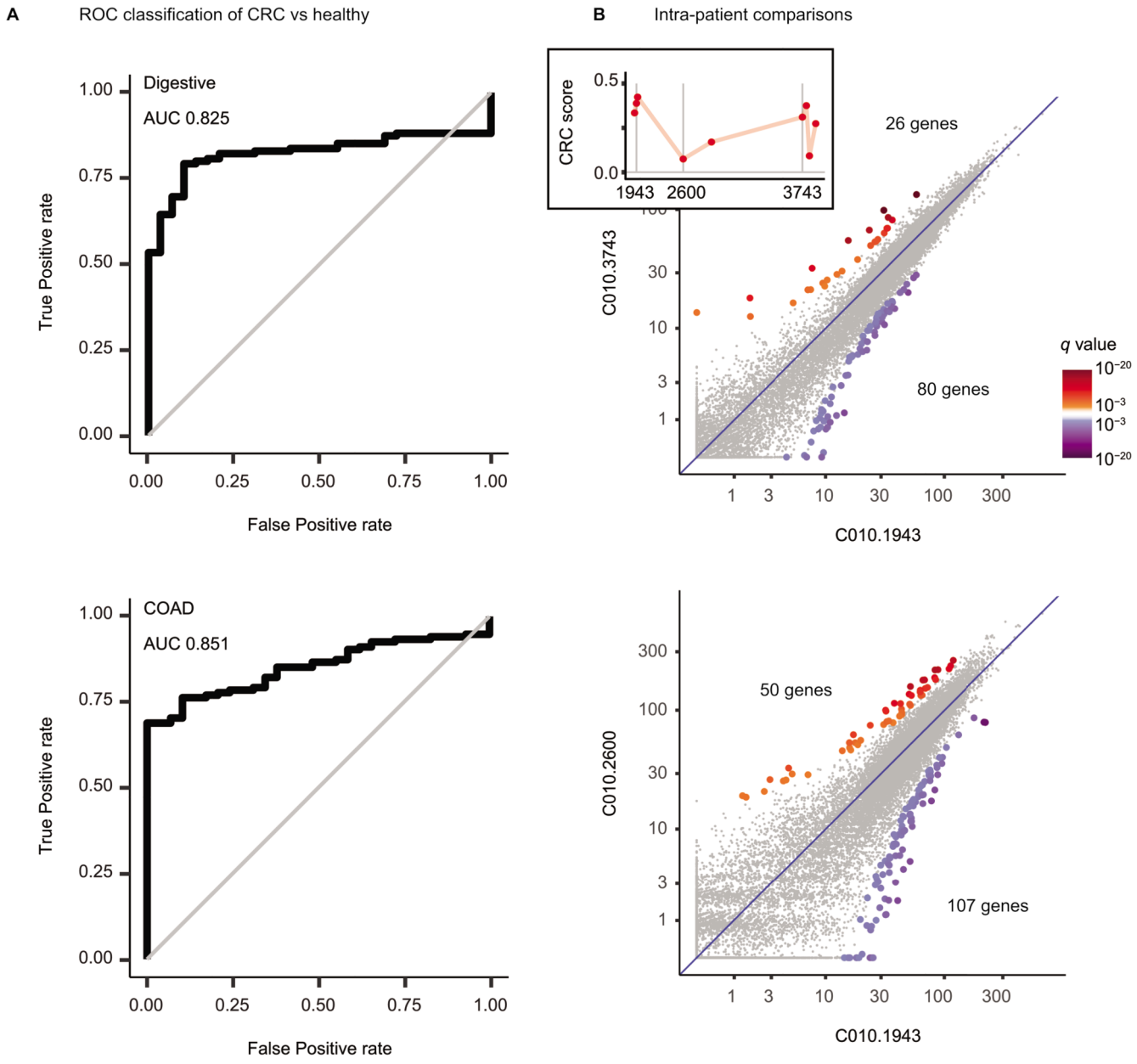


Extended Data Fig. 7.

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 Figure 5A.

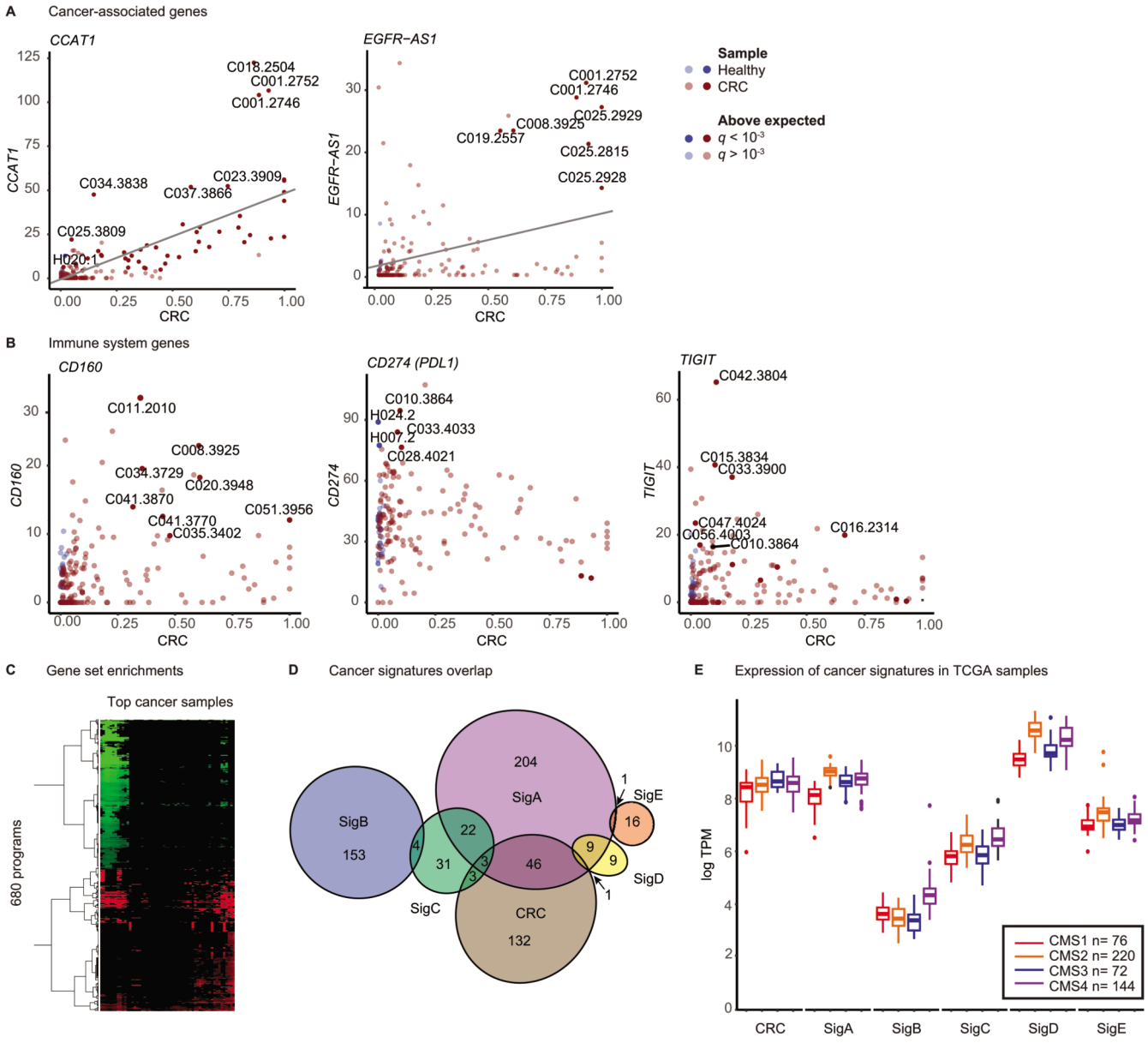
Europe PMC Funders Author Manuscripts



Extended Data Fig. 8.

A. Evaluation of classification of CRC samples vs. healthy samples using Digestive (Top) and COAD (Bottom) signature scores (as Figure 6C).

B. Intra-patient comparisons (as Figure 6E). Inset: time samples drawn on the patient timeline (Figure 6D).



Extended Data Fig. 9.

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.

Europe PMC Funders Author Manuscripts

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank N. Kaminski, J. Moss, E. Pikarsky, O.J. Rando, N. Rajewsky, 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: European Research Council's AdG Grants 340712 "ChromatinSys" (NF) and 786575 "RxmiRcanceR" (EG); Israel Science Foundation's I-CORE program grant 1796/12 (TK and NF) and grants 2612/18 (NF), 3020/20 (AG), 2473/17 (EG), and 486/17 (EG); Israel Ministry of Science and Technology grant 3-14352 (AG); NIH Grants RM1HG006193 (NF) and CA197081-02 (EG); Deutsche Forschungsgemeinschaft (DFG) SFB841 (EG); DKFZ-MOST grant (EG).

Data availability

Data collected in this study was deposited to the EGA (EMBL-EBI) repository (ACESSIONXXX). BED files and browser tracks are available at Zenodo repository DOI:10.5281/zenodo.3967253.

Browser tracks can be views by 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 as listed here:

Dataset	From	link
UCSC known genes (AH5036)	AnnotationHub	http://bioconductor.org/packages/release/bioc/html/AnnotationHub.html
ENSEMBL transcripts (AH5046)	AnnotationHub	http://bioconductor.org/packages/release/bioc/html/AnnotationHub.html
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
GTEx and TCGA RNA-seq	Toil RNAseq Recompute	https://xena.ucsc.edu/
Curated gene sets	MSigDB	http://software.broadinstitute.org/gsea/msigdb/collections.jsp
Oncogenic signature gene sets	MSigDB	http://software.broadinstitute.org/gsea/msigdb/collections.jsp
Immunologic signature gene sets	MSigDB	http://software.broadinstitute.org/gsea/msigdb/collections.jsp
Protein complexes	CORUM	http://mips.helmholtz-muenchen.de/corum/
CPDB Pathway Compendium	ConsensusPathDB	http://cpdb.molgen.mpg.de/

Dataset	From	link
BLUEPRINT RNA-seq (E-MTAB-3819)	EBI expression atlas dataset	https://www.ebi.ac.uk/gxa/experiments/E-MTAB-3819/Results
BLUEPRINT RNA-seq (E-MTAB-3827)	EBI expression atlas dataset	https://www.ebi.ac.uk/gxa/experiments/E-MTAB-3827/Results
CHEA Transcription Factor Targets	Harmonizome	http://amp.pharm.mssm.edu/Harmonizome/
ENCODE Transcription Factor Targets	Harmonizome	http://amp.pharm.mssm.edu/Harmonizome/
TRANSFAC Curated Transcription Factor Targets	Harmonizome	http://amp.pharm.mssm.edu/Harmonizome/

Code availability

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

References

1. Mandel P. Les acides nucleiques du plasma sanguin chez l'homme. CR Acad Sci Paris. 1948; 142:241–243.
2. Lo YM, et al. Rapid clearance of fetal DNA from maternal plasma. Am J Hum Genet. 1999; 64:218–224. [PubMed: 9915961]
3. De Vlaminck I, et al. Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection. Sci Transl Med. 2014; 6:241–77.
4. Schwarzenbach H, Hoon DS, Pantel K. Cell-free nucleic acids as biomarkers in cancer patients. Nat Rev Cancer. 2011; 11:426–437. [PubMed: 21562580]
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 U S A. 2015; 112:E5503–12. [PubMed: 26392541]
6. Lu J-L, Liang Z-Y. Circulating free DNA in the era of precision oncology: Pre- and post-analytical concerns. Chronic Diseases and Translational Medicine. 2016; 2:223–230. [PubMed: 29063046]
7. Wan JC, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nat Rev Cancer. 2017; 17:223–238. [PubMed: 28233803]
8. Lehmann-Werman R, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci U S A. 2016; 113:E1826–34. [PubMed: 26976580]
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. 2017; 49:635–642. [PubMed: 28263317]
10. Kang S, et al. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol. 2017; 18:53. [PubMed: 28335812]
11. Moss J, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018
12. Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell. 1999; 98:285–294. [PubMed: 10458604]
13. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007; 128:707–719. [PubMed: 17320508]
14. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. A chromatin landmark and transcription initiation at most promoters in human cells. Cell. 2007; 130:77–88. [PubMed: 17632057]

15. Berger SL. The complex language of chromatin regulation during transcription. Nature. 2007; 447:407. [PubMed: 17522673]
16. Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol. 2015; 16:178. [PubMed: 25650798]
17. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]
18. Barski A, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007; 129:823–837. [PubMed: 17512414]
19. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–112. [PubMed: 19295514]
20. Lenhard B, Sandelin A, Carninci P. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet. 2012; 13:233–245. [PubMed: 22392219]
21. Calo E, Wysocka J. Modification of enhancer chromatin: what, how, and why? Mol Cell. 2013; 49:825–837. [PubMed: 23473601]
22. Lawrence M, Daujat S, Schneider R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends Genet. 2016; 32:42–56. [PubMed: 26704082]
23. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306:636–640. [PubMed: 15499007]
24. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]
25. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–330. [PubMed: 25693563]
26. Lara-Astiaso D, et al. Immunogenetics. Chromatin state dynamics during blood formation. Science. 2014; 345:943–949. [PubMed: 25103404]
27. Weiner A, et al. High-resolution chromatin dynamics during a yeast stress response. Mol Cell. 2015; 58:371–386. [PubMed: 25801168]
28. Holdenrieder S, et al. Nucleosomes in serum of patients with benign and malignant diseases. Int J Cancer. 2001; 95:114–120. [PubMed: 11241322]
29. Holdenrieder S, et al. Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR. Clin Chem. 2005; 51:1544–1546. [PubMed: 16040855]
30. Rumore PM, Steinman CR. Endogenous circulating DNA in systemic lupus erythematosus. Occurrence as multimeric complexes bound to histone. J Clin Invest. 1990; 86:69–74. [PubMed: 2365827]
31. Gezer U, et al. Characterization of H3K9me3- and H4K20me3-associated circulating nucleosomal DNA by high-throughput sequencing in colorectal cancer. Tumour Biol. 2013; 34:329–336. [PubMed: 23086575]
32. Bauden M, et al. Circulating nucleosomes as epigenetic biomarkers in pancreatic cancer. Clin Epigenetics. 2015; 7:106. [PubMed: 26451166]
33. Deligezer, U, , et al. H3K9me3/H4K20me3 Ratio in Circulating Nucleosomes as Potential Biomarker for Colorectal CancerCirculating Nucleic Acids in Plasma and Serum. Springer Netherlands; 2011. 97–103.
34. Ulz P, et al. Inferring expressed genes by whole-genome sequencing of plasma DNA. Nat Genet. 2016; 48:1273–1278. [PubMed: 27571261]
35. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. Cell. 2016; 164:57–68. [PubMed: 26771485]
36. Xu R-H, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater. 2017; 16:1155–1161. [PubMed: 29035356]
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. 2017; 12:455–462. [PubMed: 27617389]

38. Ramachandran S, Ahmad K, Henikoff S. Transcription and Remodeling Produce Asymmetrically Unwrapped Nucleosomal Intermediates. Molecular Cell. 2017; 68:1038–1053.e4. [PubMed: 29225036]
39. Zemmour H, et al. Non-invasive detection of human cardiomyocyte death using methylation patterns of circulating DNA. Nat Commun. 2018; 9:1443. [PubMed: 29691397]
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. 2018; 46:e89. [PubMed: 29897492]
41. Shen SY, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature. 2018; 563:579–583. [PubMed: 30429608]
42. Lehmann-Werman R, et al. Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. JCI Insight. 2018; 3
43. Cristiano S, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. Nature. 2019; doi: 10.1038/s41586-019-1272-6
44. Gutin J, et al. Fine-Resolution Mapping of TF Binding and Chromatin Interactions. Cell Rep. 2018; 22:2797–2807. [PubMed: 29514105]
45. Singh SS, et al. Widespread suppression of intragenic transcription initiation by H-NS. Genes Dev. 2014; 28:214–219. [PubMed: 24449106]
46. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011; 147:1408–1419. [PubMed: 22153082]
47. Mizuta R, et al. DNase γ is the effector endonuclease for internucleosomal DNA fragmentation in necrosis. PLoS One. 2013; 8
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. 2017; 28:1882–1888. [PubMed: 28838211]
49. Tan DSW, et al. Long noncoding RNA EGFR-AS1 mediates epidermal growth factor receptor addiction and modulates treatment response in squamous cell carcinoma. Nat Med. 2017; 23:1167–1175. [PubMed: 28920960]
50. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–660. [PubMed: 25954001]
51. Cancer Genome Atlas Research Network. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–1120. [PubMed: 24071849]
52. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010; 107:2926–2931. [PubMed: 20133639]
53. Liu CL, et al. Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol. 2005; 3:e328. [PubMed: 16122352]
54. Swarup V, Rajeswari MR. Circulating (cell-free) nucleic acids–A promising, non-invasive tool for early detection of several human diseases. FEBS Lett. 2007
55. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. Cancer Res. 1977; 37:646–650. [PubMed: 837366]
56. Lam WKJ, et al. DNA of Erythroid Origin Is Present in Human Plasma and Informs the Types of Anemia. Clin Chem. 2017; 63:1614–1623. [PubMed: 28784691]
57. Deutsch VR, Tomer A. Megakaryocyte development and platelet production. British Journal of Haematology. 2006; 134:453–466. [PubMed: 16856888]
58. Stunnenberg HG, Hirst M, International Human Epigenome Consortium. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016; 167:1897.
59. Giannini EG, Testa R, Savarino V. Liver enzyme alteration: a guide for clinicians. CMAJ. 2005; 172:367–379. [PubMed: 15684121]
60. Liberzon A, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015; 1:417–425. [PubMed: 26771021]

61. Drew K, et al. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. Mol Syst Biol. 2017; 13:932. [PubMed: 28596423]
62. Giurgiu M, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2019; 47:D559–D563. [PubMed: 30357367]
63. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res. 2013; 41:D793–800. [PubMed: 23143270]
64. King KR, et al. IRF3 and type I interferons fuel a fatal response to myocardial infarction. Nat Med. 2017; 23:1481–1487. [PubMed: 29106401]
65. Czaja AJ. Review article: chemokines as orchestrators of autoimmune hepatitis and potential therapeutic targets. Aliment Pharmacol Ther. 2014; 40:261–279. [PubMed: 24890045]
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. 2009; 665:47–59. [PubMed: 20429415]
67. Lachmann A, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018; 9:1366. [PubMed: 29636450]
68. Aizarani N, et al. A human liver cell atlas reveals heterogeneity and epithelial progenitors. Nature. 2019; 572:199–204. [PubMed: 31292543]
69. Jungermann K, Katz N. Functional specialization of different hepatocyte populations. Physiol Rev. 1989; 69:708–764. [PubMed: 2664826]
70. Reinert T, et al. Analysis of circulating tumour DNA to monitor disease burden following colorectal cancer surgery. Gut. 2016; 65:625–634. [PubMed: 25654990]
71. Tannapfel A, Reinacher-Schick A. [Chemotherapy associated hepatotoxicity in the treatment of advanced colorectal cancer (CRC)]. Z Gastroenterol. 2008; 46:435–440. [PubMed: 18461519]
72. Bradner JE, Hnisz D, Young RA. Transcriptional Addiction in Cancer. Cell. 2017; 168:629–643. [PubMed: 28187285]
73. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]
74. Nissan A, et al. Colon cancer associated transcript-1: a novel RNA expressed in malignant and pre-malignant human tissues. Int J Cancer. 2012; 130:1598–1606. [PubMed: 21547902]
75. Coulson JM. Transcriptional regulation: cancer, neurons and the REST. Curr Biol. 2005; 15:R665–8. [PubMed: 16139198]
76. Rademakers G, et al. The role of enteric neurons in the development and progression of colorectal cancer. Biochim Biophys Acta Rev Cancer. 2017; 1868:420–434. [PubMed: 28847715]
77. Guinney J, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015; 21:1350–1356. [PubMed: 26457759]
78. Koppens MAJ, et al. Large variety in a panel of human colon cancer organoids in response to EZH2 inhibition. Oncotarget. 2016; 7:69816–69828. [PubMed: 27634879]
79. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
80. Ferrari A, et al. A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. Nat Commun. 2016; 7
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. 2016; 17:738–746. [PubMed: 27108243]
82. Ulz P, et al. Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. Nat Commun. 2019; 10
83. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet. 2019; doi: 10.1038/s41576-019-0128-0
84. Shen-Orr SS, Gaujoux R. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. Curr Opin Immunol. 2013; 25:571–578. [PubMed: 24148234]
85. Vivian J, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017; 35:314–316. [PubMed: 28398314]

86. Rouillard AD, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016

87. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

88. Matys V, et al. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–D110. [PubMed: 16381825]

89. Lachmann A, et al. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010; 26:2438–2444. [PubMed: 20709693]

90. Kuleshov MV, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44:W90–7. [PubMed: 27141961]

Europe PMC Funders Author Manuscripts

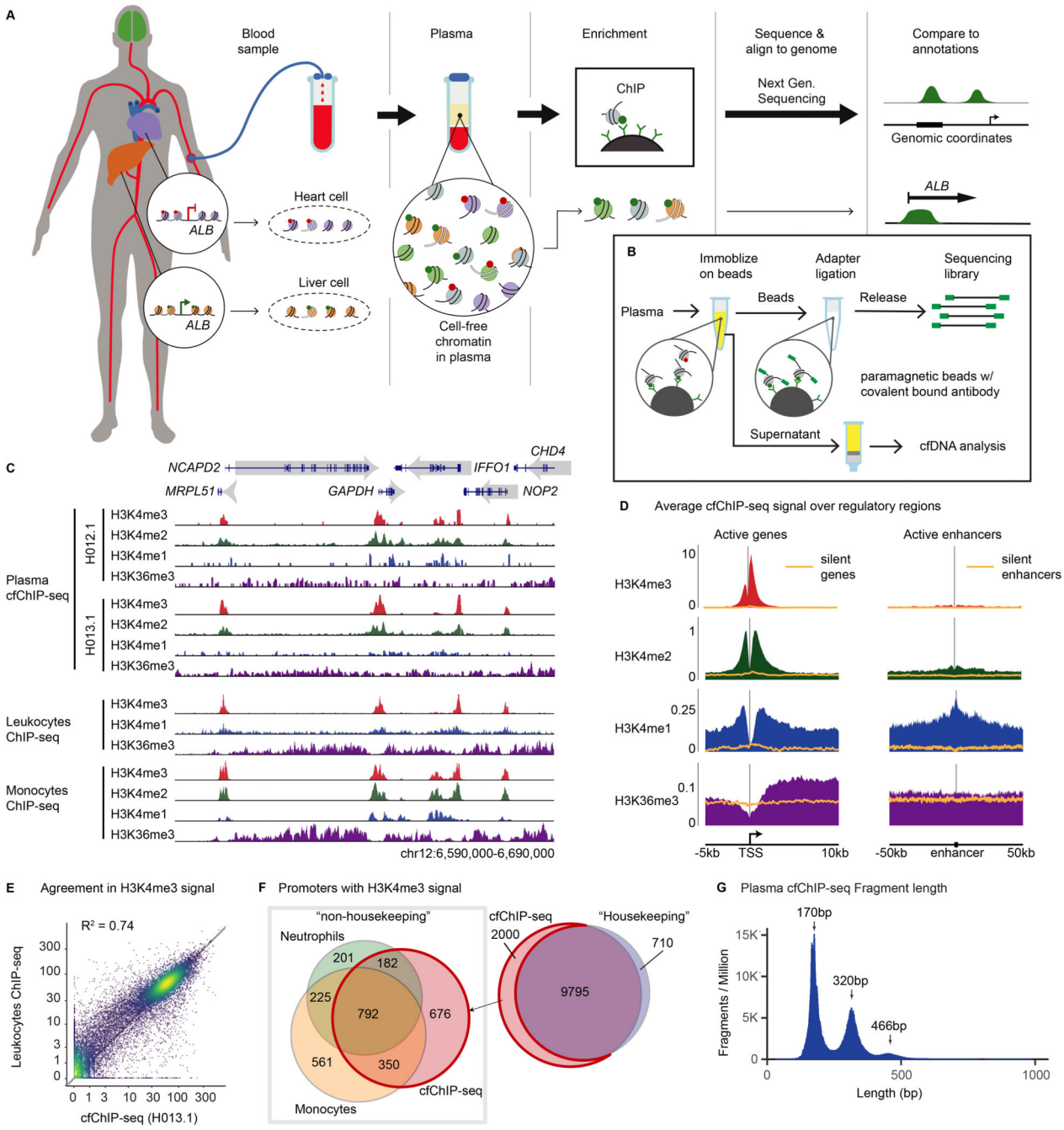


Figure 1. Chromatin Immunoprecipitation from plasma

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 2000 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.

Europe PMC Funders Author Manuscripts

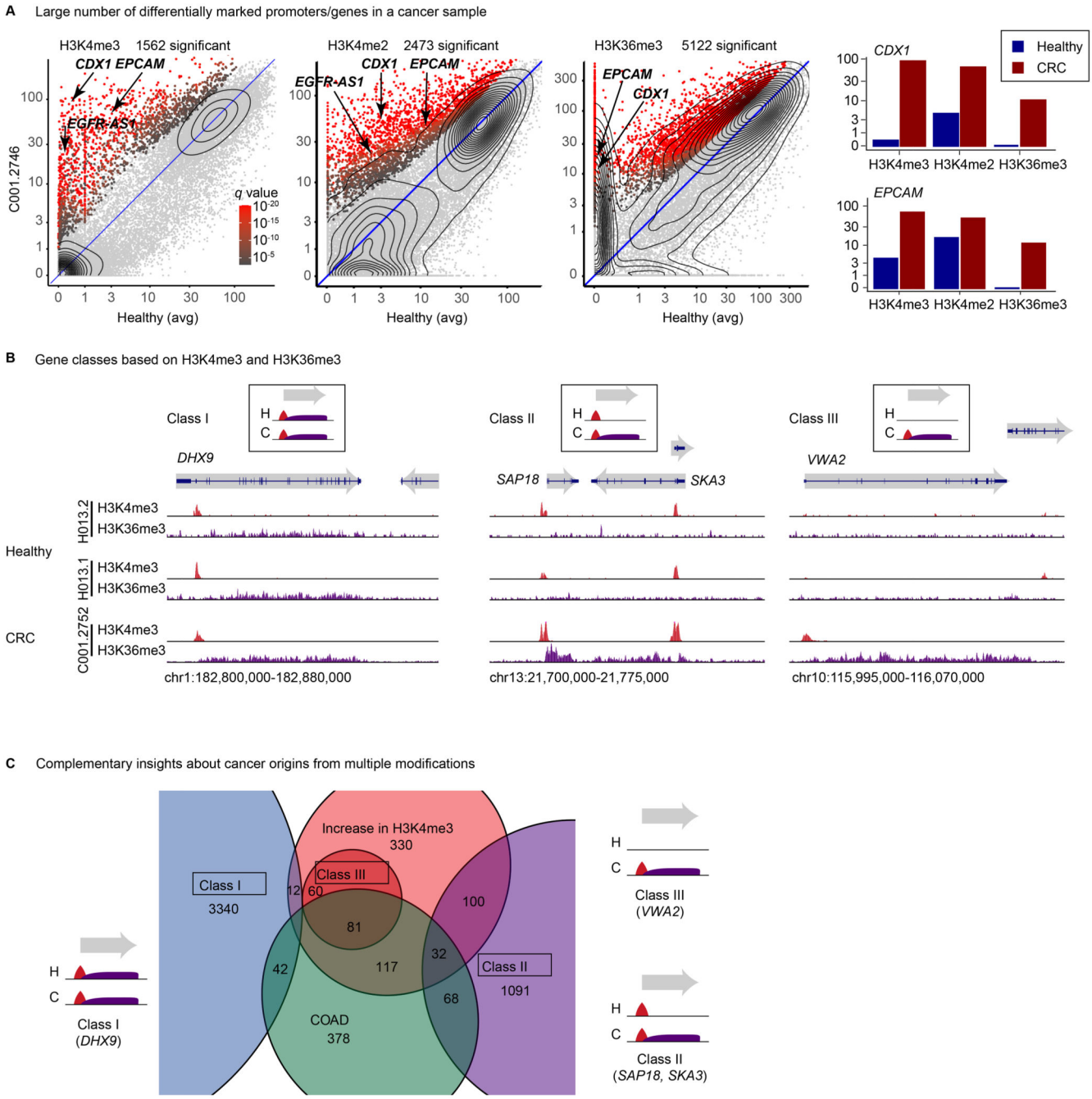


Figure 2. cfChIP-seq of multiple marks is informative on gene expression

A. Detection of genes with significant high coverage in a sample from a colorectal cancer (CRC) patient (C001, Supplementary Table 4). For each gene we compare 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 three genes in these comparisons are shown on the bar chart (right panel).

Europe PMC Funders Author Manuscripts

B. Browser views of genes that demonstrate different H3K4me3 and H3K36me3 classes. Class I: genes marked by both marks in healthy and cancer patient samples. Class II: genes marked by H3K4me3 in healthy and cancer samples, but with H3K36me3 only in the cancer patient sample (gain of H3K36me3). Class III: genes marked by both marks only in the cancer patient 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 colorectal adenocarcinoma (COAD, Methods).

Europe PMC Funders Author Manuscripts

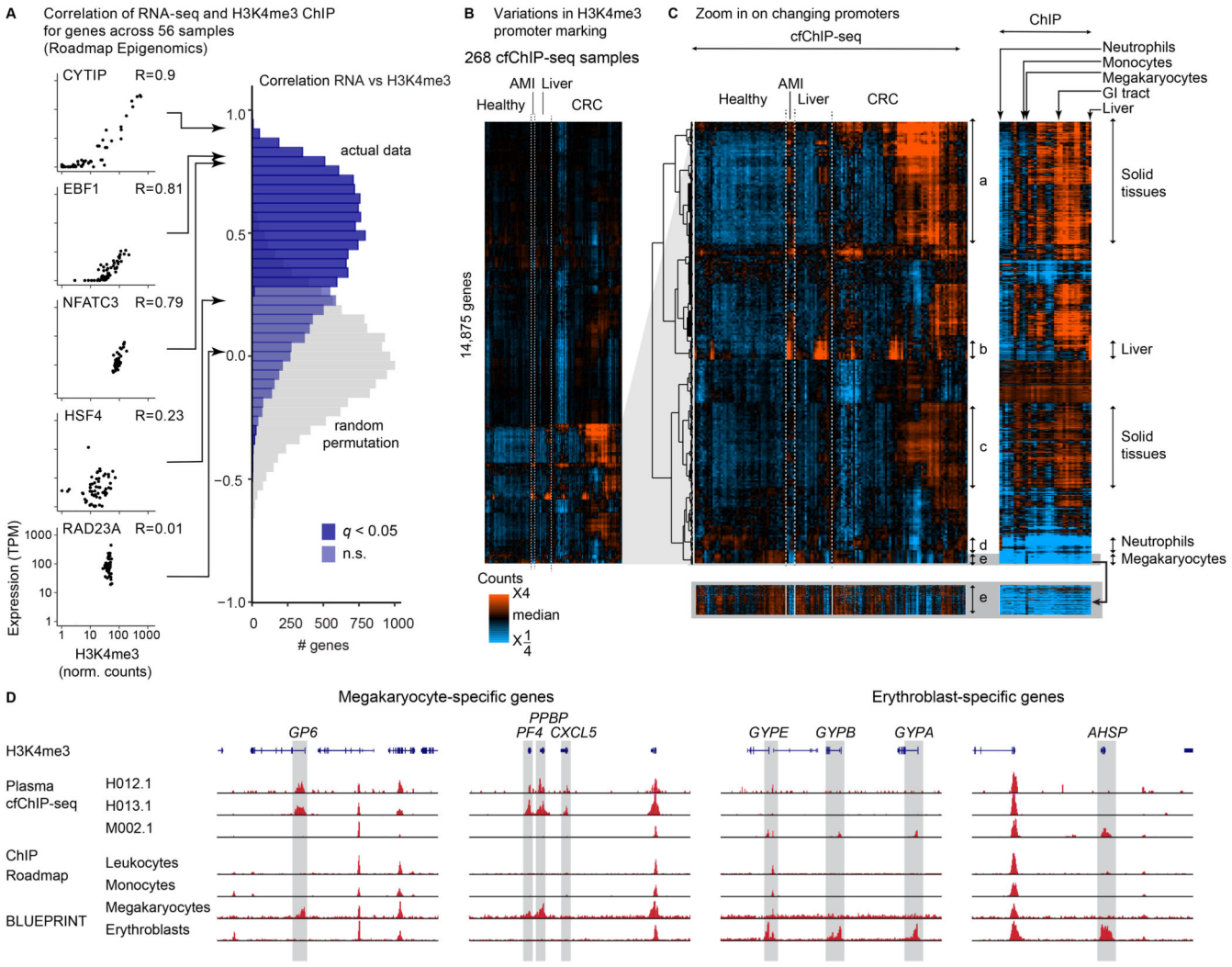


Figure 3. H3K4me3 cfChIP-seq signal is correlated with expression levels

A. Gene level analysis of the correlation in expression level and H3K4me3 signal across 56 Roadmap Epigenomic 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. Shown on the right is a histogram of the correlations on all RefSeq genes (significance w.r.t. to random correlation, shown in gray). Left: examples of genes with different correlation values.

B. Heatmap 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 (log2(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, acute myocardial infarction (AMI) patients, liver disease patients and CRC patients.

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.

Europe PMC Funders Author Manuscripts

D. Genome browser view for megakaryocyte- and erythroblast specific genes. Shown is cfChIP-seq from two healthy samples (H012.1 and H013.1) and an AMI subject who exhibited enhanced erythropoiesis (M002.1). Also shown are two ChIP-seq profiles from the Roadmap Epigenetic reference atlas, and two samples from the BLUEPRINT project of cord-blood derived megakaryocytes and erythroblasts.

Europe PMC Funders Author Manuscripts

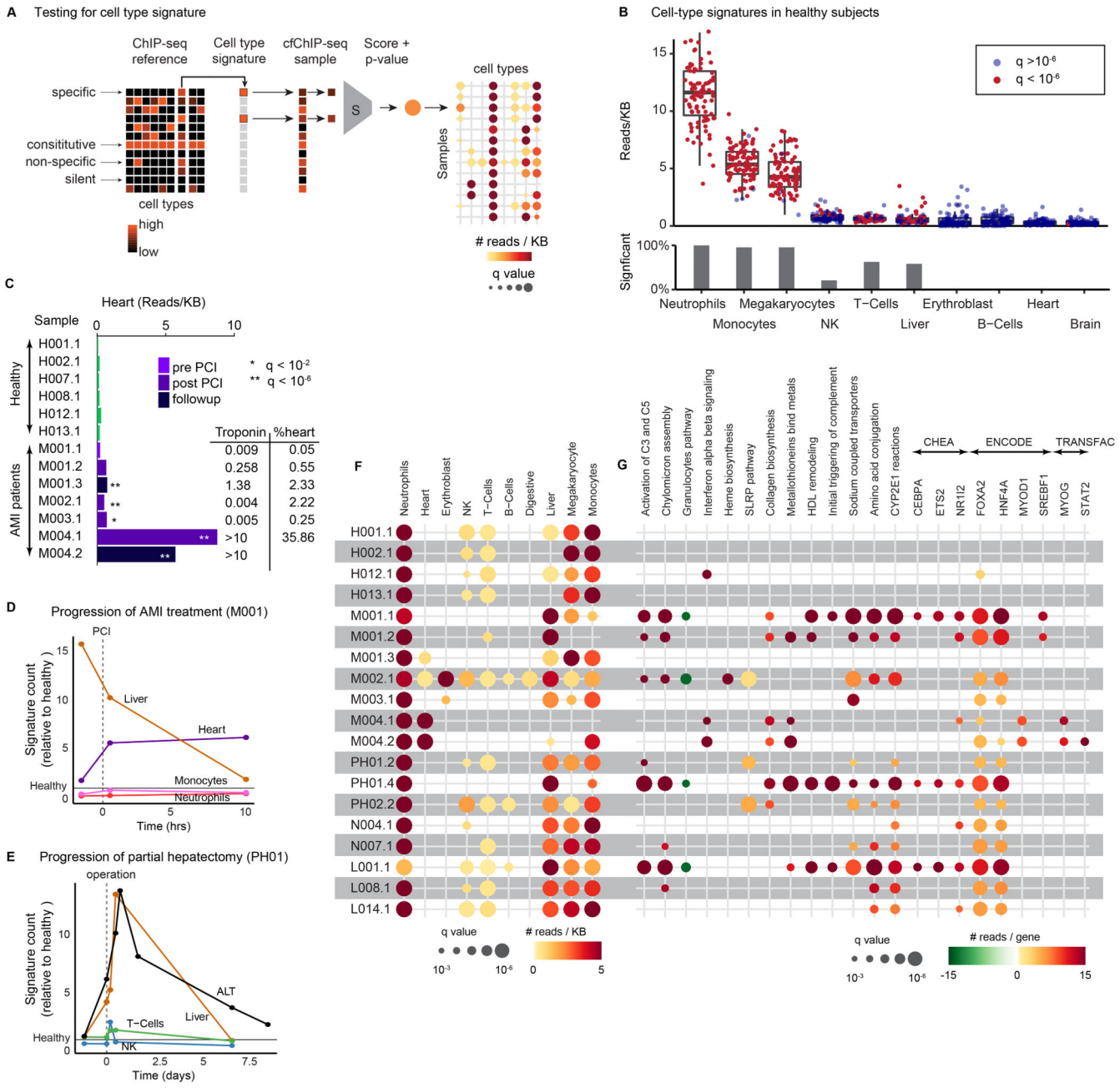


Figure 4. cfChIP-seq identifies cell-type and program specific expression patterns

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 * inter-quartile range from the hinge. Dots marked in red

Europe PMC Funders Author Manuscripts

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 acute myocardial infarction (AMI) patients (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 an AMI (M001) patient before/after PCI. Signatures 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 that underwent partial hepatectomy (PH01).

F. Heatmap 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. Heatmap showing significance of selected gene sets from curated database of transcriptional programs[60] and transcription factor targets[87–89] (Methods; 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).

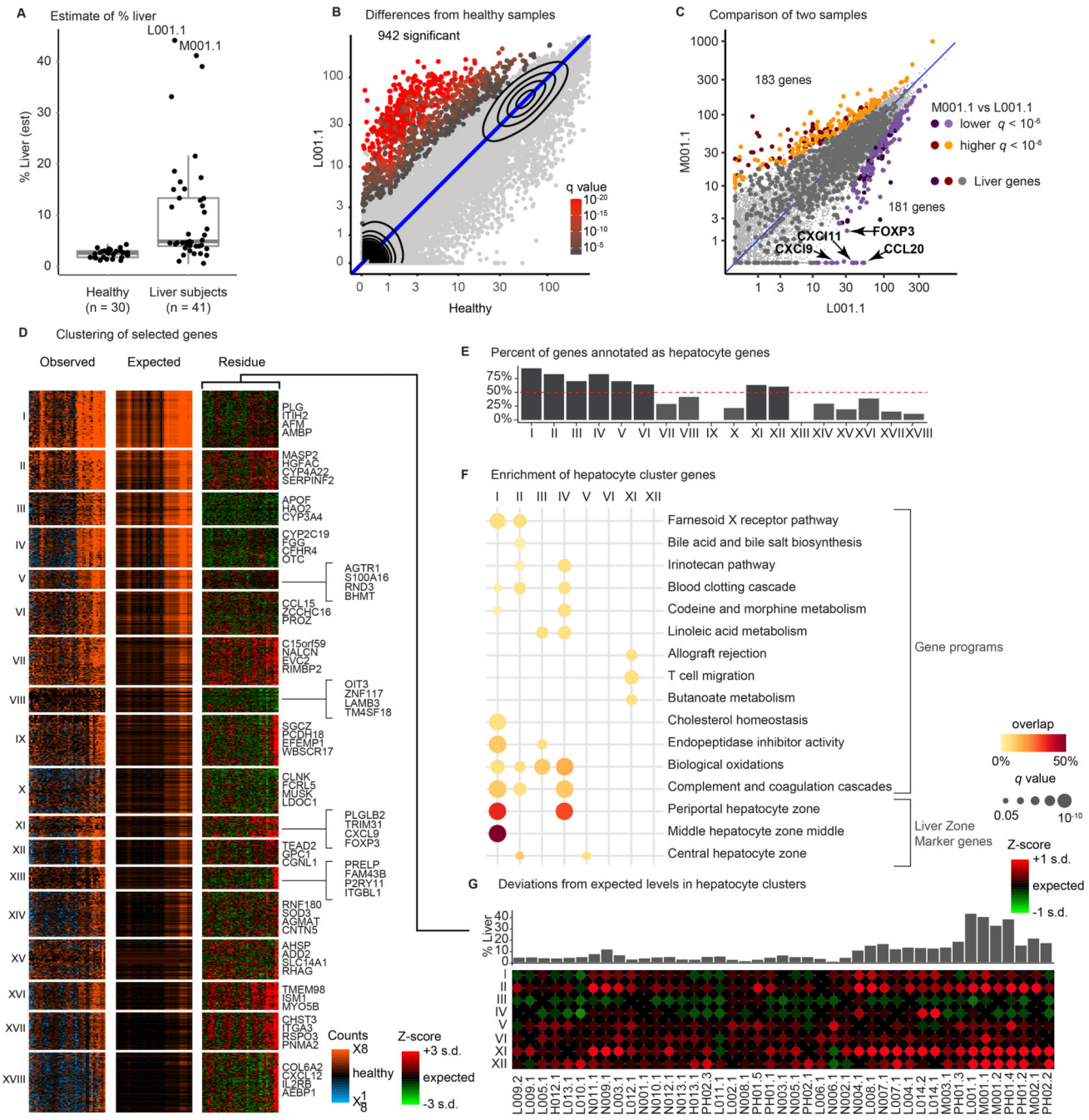


Figure 5. cfChIP-seq detects changes in liver-specific transcriptional programs

A. Estimate of %liver contribution to healthy reference cohort and a cohort of subjects 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 * inter-quartile range from the hinge.

B. Evaluation of differentially marked genes in a sample of an acute AIH subject (L001) as in Figure 2A.

Europe PMC Funders Author Manuscripts

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 compare the observed levels (L001.1, x-axis; M001.1 y-axis) and test 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 heatmap 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 origins.

F. Enrichment analysis of hepatocyte clusters (Clusters I-VI, XI, and XII). Hypergeometric test for significant overlap with gene programs from curated databases[90] 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).

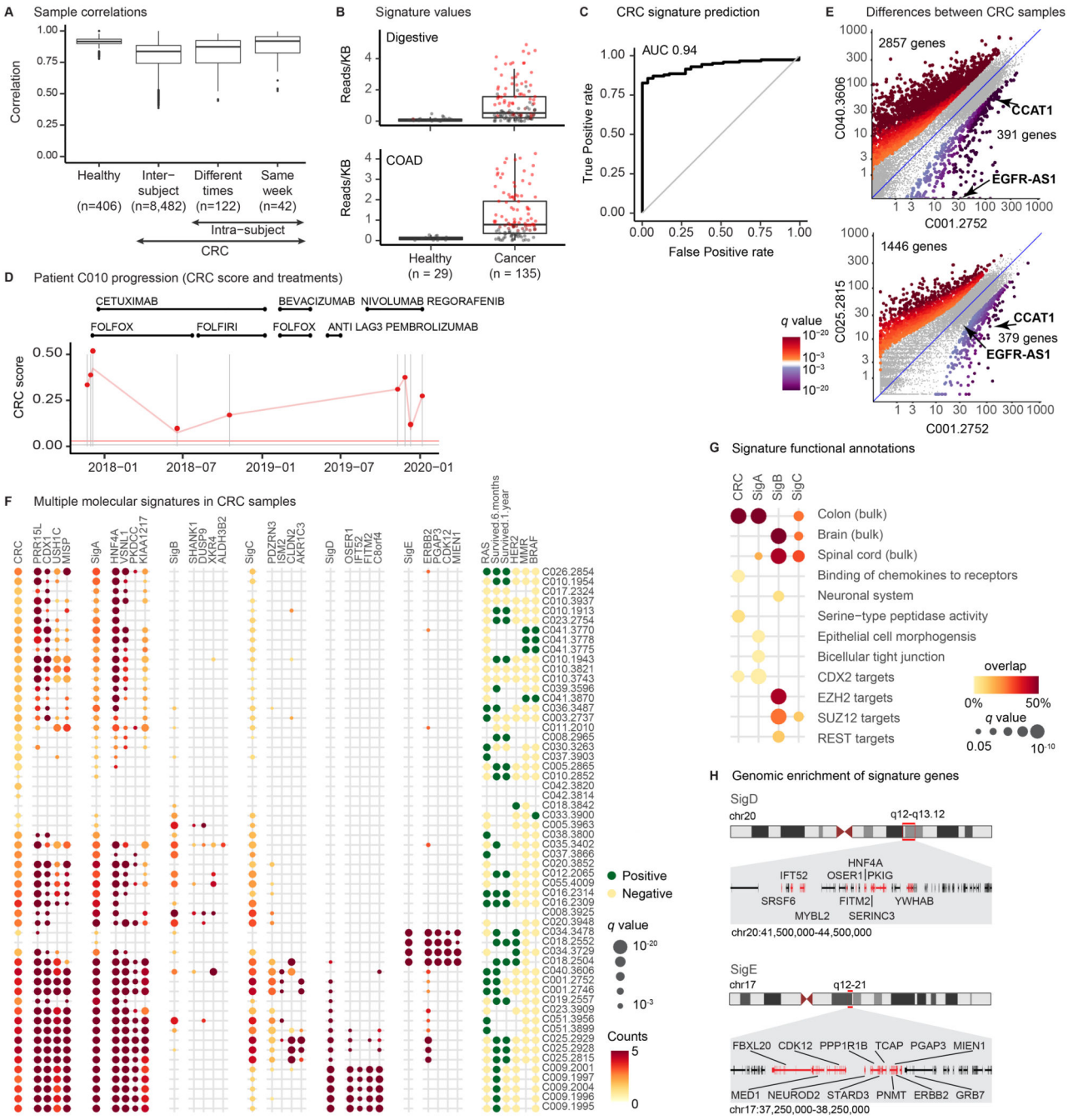


Figure 6. cfChIP-seq identifies molecular heterogeneity in colorectal carcinoma patients

A. Pairwise comparisons (pearson correlation, y-axis) of between samples: healthy donors; different CRC patients; the same CRC patient more than a week apart; the same CRC patient less than a week apart. 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.

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-

Europe PMC Funders Author Manuscripts

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 * inter-quartile range from the hinge.

C. Classification accuracy of CRC patients vs healthy donors with CRC signature. Fraction false positive (x-axis) vs fraction true positives (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 between samples with high CRC signature strength. For each gene we compare coverage in the two samples (x-axis and y-axis). Significance test 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/gene above healthy baseline, circle radius: significance of this increase (Methods). Rightmost panel displays major clinical parameters: RAS, BRAF mutations, 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[90] 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.