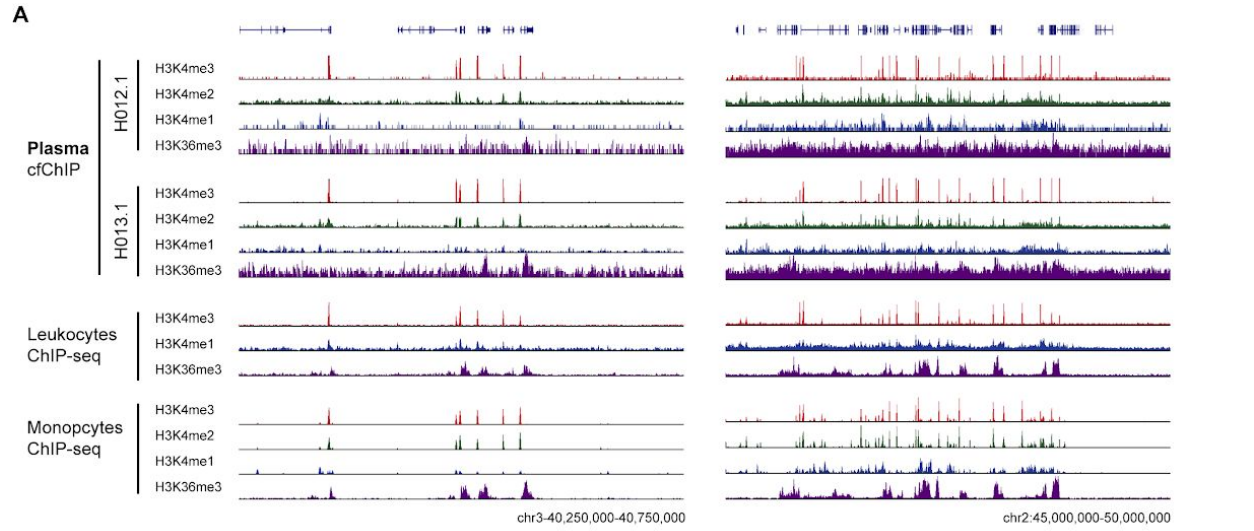
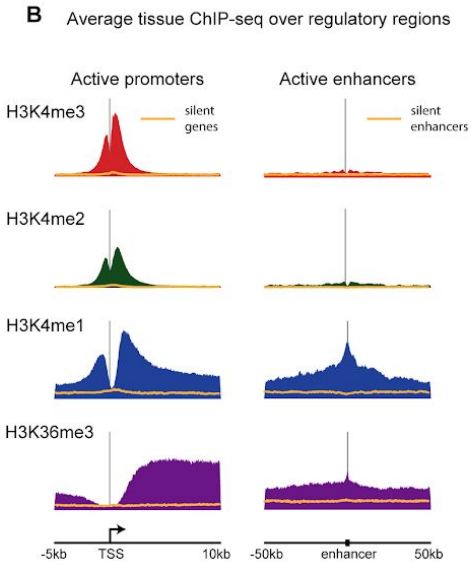
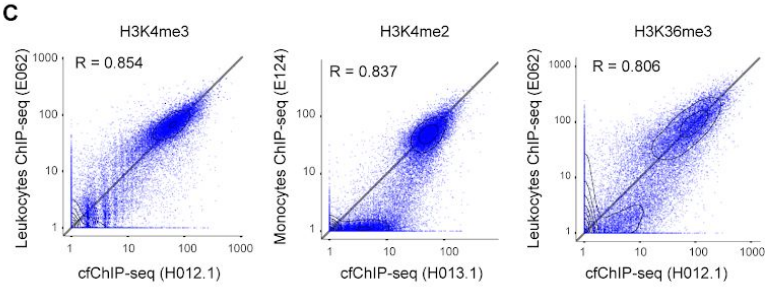
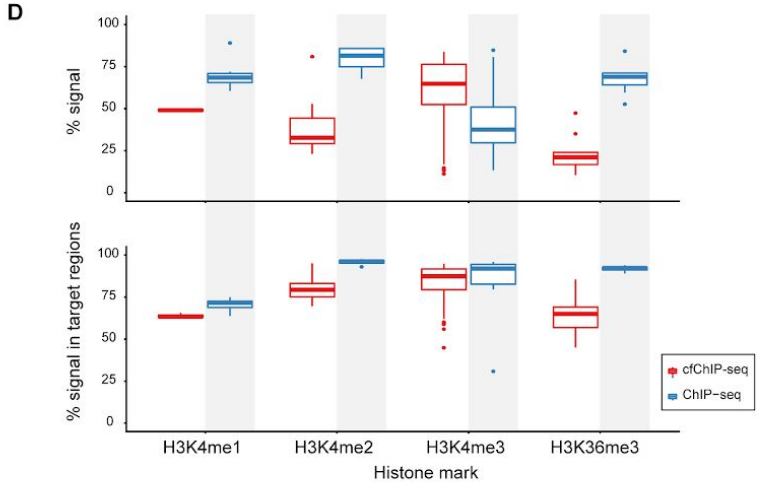
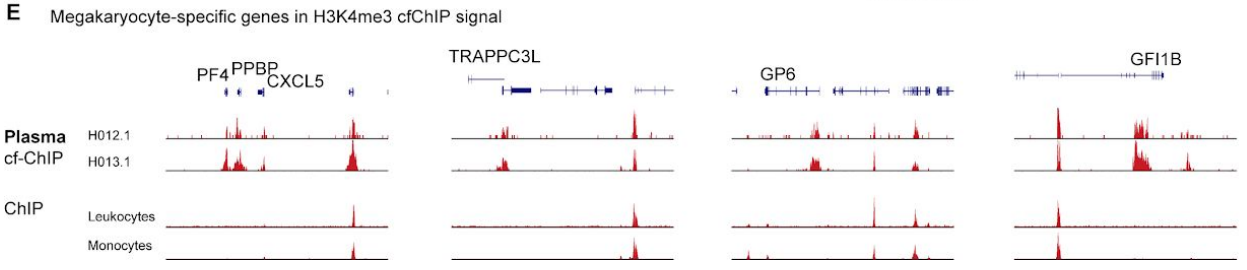
Supplemental Figures, Table captions, and Note


A
Plasma cfChIP
H012.1
H013.1
H3K4me3
H3K4me2
H3K4me1
H3K36me3
Leukocytes ChIP-seq
Monopcytes ChIP-seq
chr3-40,250,000-40,750,000
chr2:45,000,000-50,000,000


B
Average tissue ChIP-seq over regulatory regions
Active promoters
Active enhancers
H3K4me3
H3K4me2
H3K4me1
H3K36me3
silent genes
silent enhancers
-5kb
TSS
10kb
-50kb
enhancer
50kb

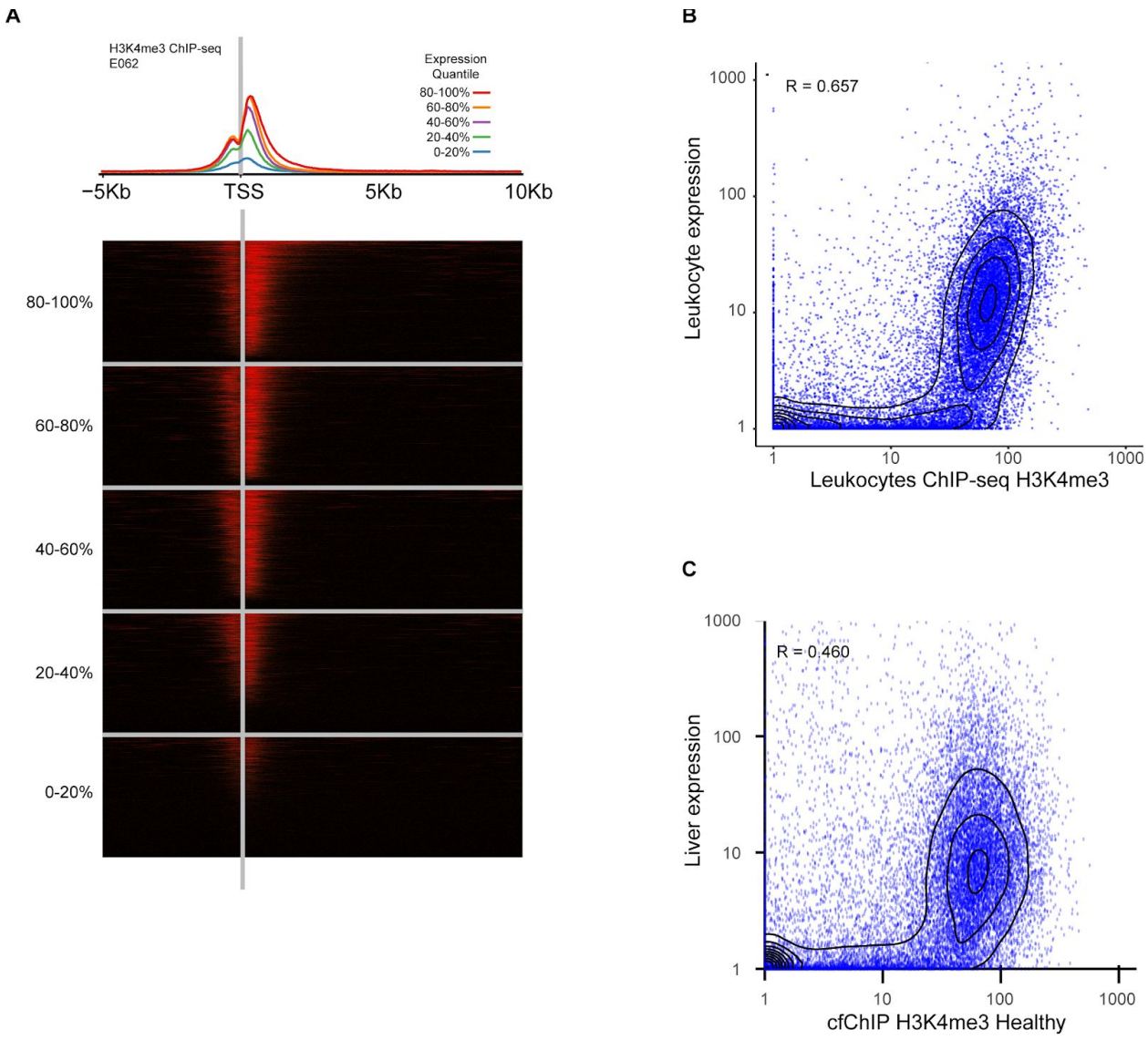

C
H3K4me3
H3K4me2
H3K36me3
R = 0.854
R = 0.837
R = 0.806
Leukocytes ChIP-seq (E062)
Monocytes ChIP-seq (E124)
cfChIP-seq (H012.1)
cfChIP-seq (H013.1)


D
% signal
% signal in target regions
cfChIP-seq
ChIP-seq
H3K4me1
H3K4me2
H3K4me3
H3K36me3
Histone mark


E
Megakaryocyte-specific genes in H3K4me3 cfChIP signal
PF4
PPBP
CXCL5
TRAPPC3L
GP6
GFI1B
Plasma cf-ChIP
H012.1
H013.1
ChIP
Leukocytes
Monocytes

Supplemental Figure S1

A. Genome browser view (as in Figure 1C).
B. Metaplots (as in Figure 1D) collected for ChIP-seq samples from (Roadmap Epigenomics Consortium et al. 2015).
C. Scatter showing signal from cfChIP-seq versus Leukocyte ChIP-seq of H3K4me3, H3K4me2, and H3K36me3 (similar to Figure 1E)
D. 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 (Table S1) compared to selected ChIP-seq samples from (Roadmap Epigenomics Consortium et al. 2015) . Bottom panel: percent of the signal above background that is in 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.
E. Examples of genome browser view for megakaryocytes promoters.

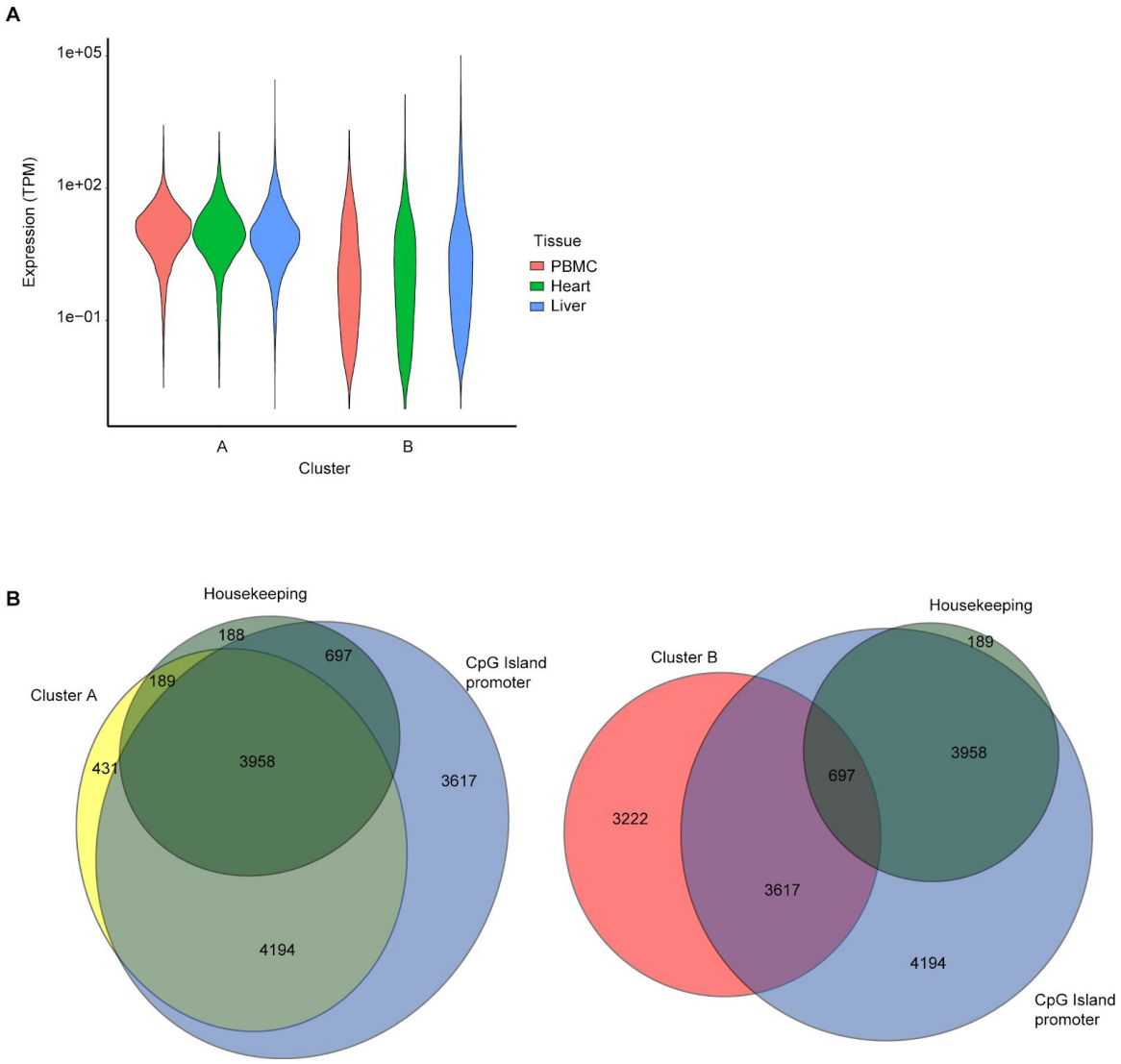


Supplemental Figure S2

A. Metaplot and heatmap (as in Figure 2A) of H3K4me3 Leukocytes ChIP-seq (roadmap sample E062) (Roadmap Epigenomics Consortium et al. 2015).

B. Comparison (as in Figure 2B) of ChIP-seq of Leukocytes vs. expression in Leukocytes (roadmap sample E062).

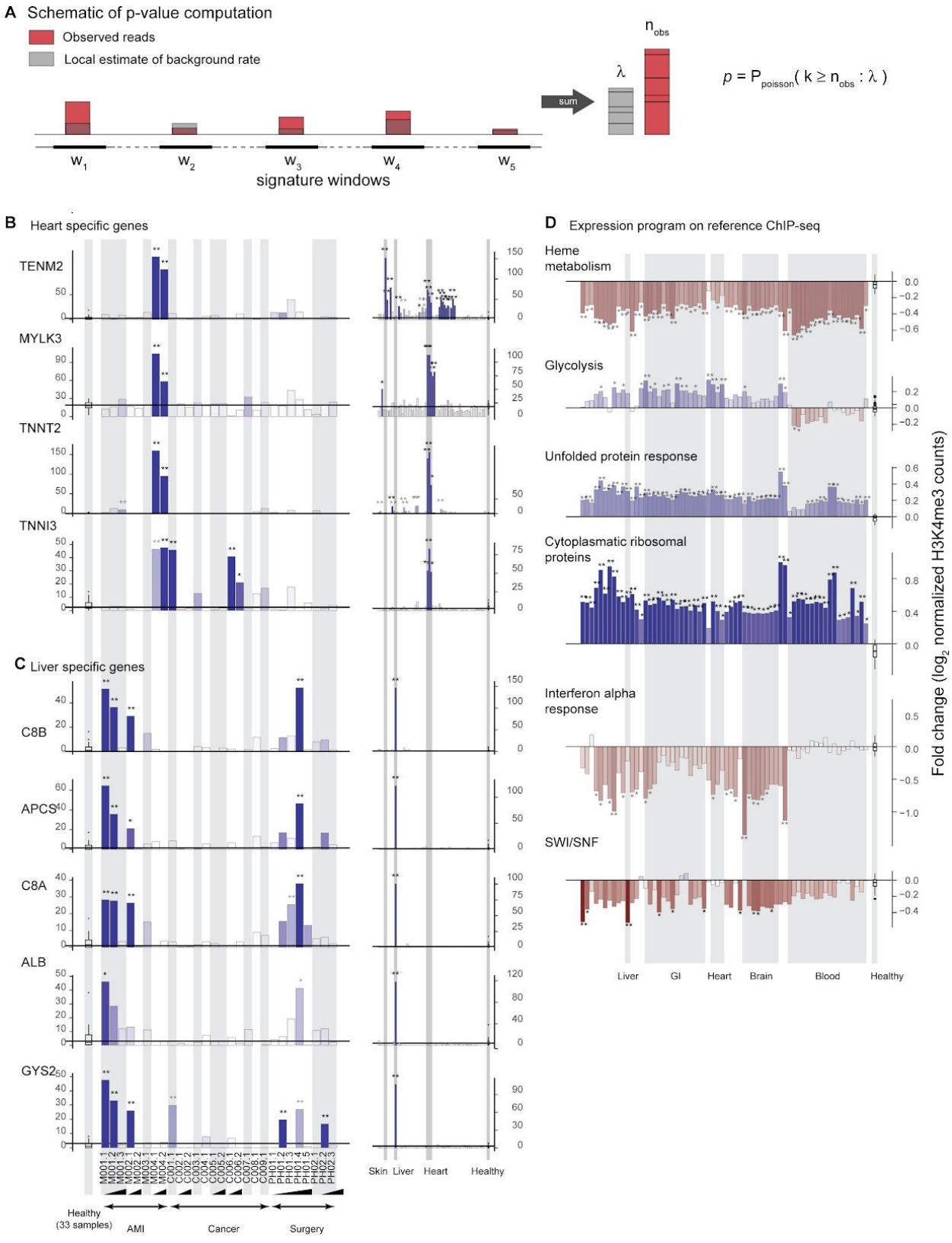
C. Comparison (as in Figure 2B) of H3K4me3 cfChIP signal from a healthy subject (H012.1) against expression levels of genes in Liver.



Supplemental Figure S3

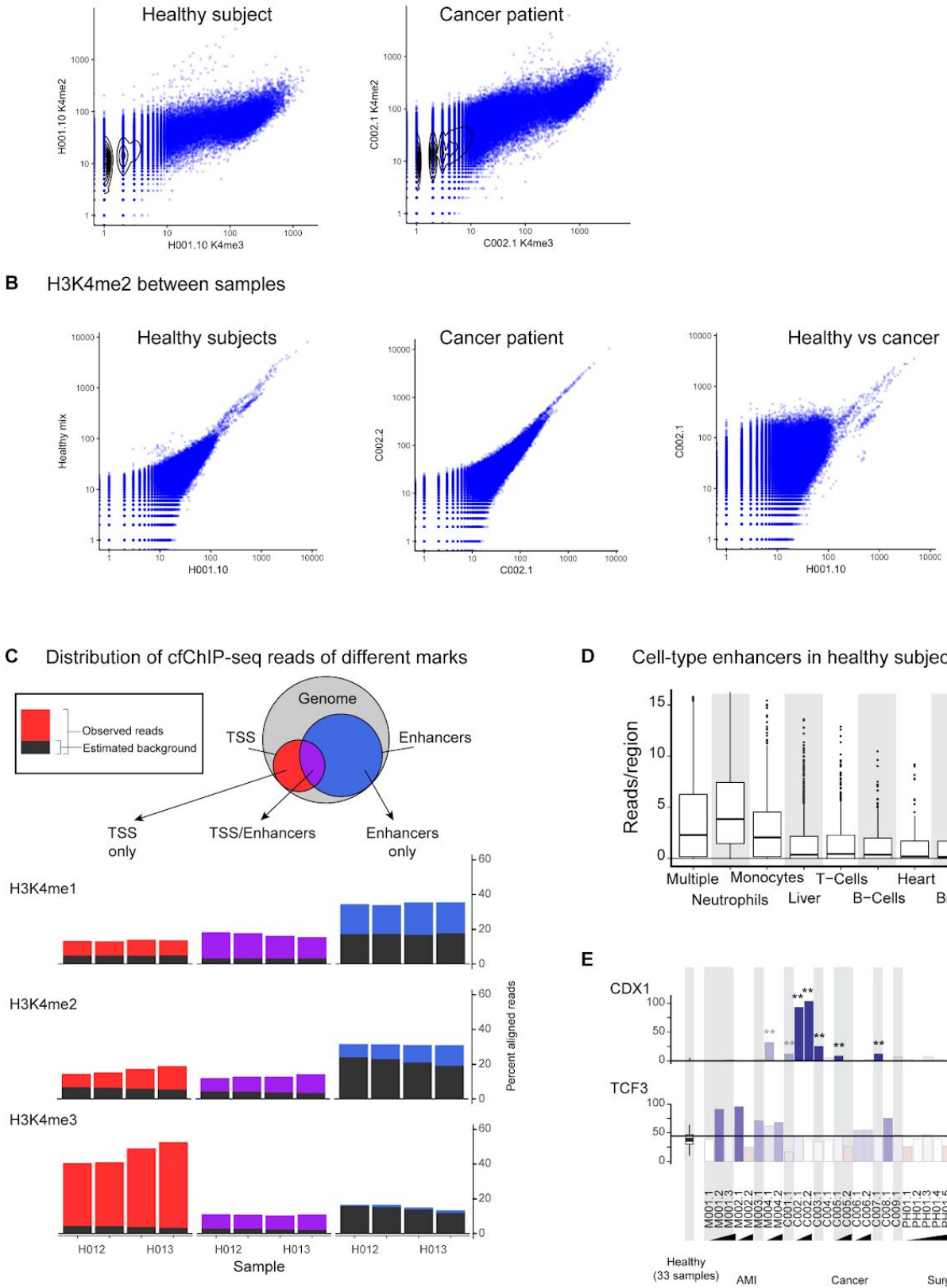
A. Comparison of the expression levels of genes in two clusters of Figure 3A (see inset). Cluster A are 9,376 genes that do not change between samples, and Cluster B are 8,374 genes that exhibit changes between pattern. Violin plots show the distribution of expression levels in three tissues - PBMC, Heart, and Liver, from the Roadmap Epigenomics expression data (Roadmap Epigenomics Consortium et al. 2015).

B. Overlap of both clusters with the set of genes with CpG island promoters (blue) and housekeeping genes (based on analysis of GTEX compendium (GTEx Consortium 2015), see Methods). For clarity we show each cluster in a separate Venn diagram.


A Schematic of p-value computation
Observed reads
Local estimate of background rate
sum
n_{obs}
λ
$p = P_{poisson}(k \geq n_{obs} : \lambda)$
w_1
w_2
w_3
w_4
w_5
signature windows
B Heart specific genes
TENM2
MYLK3
TNNT2
TNNI3
C Liver specific genes
C8B
APCS
C8A
ALB
GYS2
Skin Liver Heart Healthy
Healthy (33 samples)
AMI
Cancer
Surgery
D Expression program on reference ChIP-seq
Heme metabolism
Glycolysis
Unfolded protein response
Cytoplasmatic ribosomal proteins
Interferon alpha response
SWI/SNF
Liver GI Heart Brain Blood Healthy
Fold change (log_2 normalized H3K4me3 counts)

Supplemental Figure S4

A. Concept for signature scoring process.
B. Examples of heart-specific genes in cfChIP samples (left) and in Roadmap Epigenomics reference samples (right). Presentation is as in Figure 5D).
C. Same as B, but with liver-specific genes.
D. Evaluation of expression programs (as in Figure 5A) on Roadmap Epigenomics reference samples.

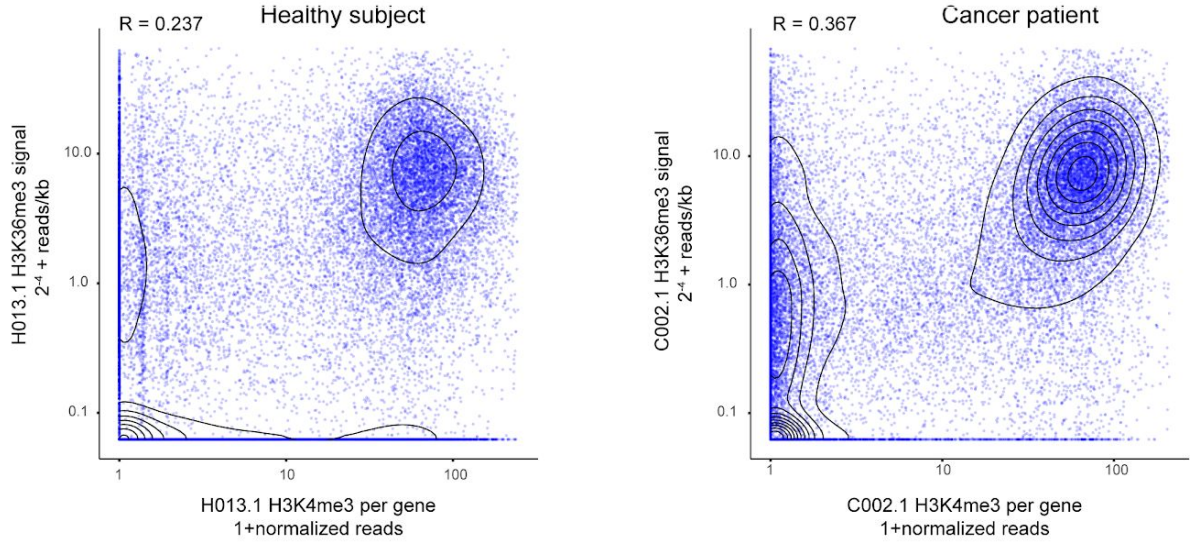

A
Comparison between H3K4me2 and H3K4me3 on the same sample
Healthy subject
Cancer patient
H001.10 K4me2
H001.10 K4me3
C002.1 K4me2
C002.1 K4me3
B
H3K4me2 between samples
Healthy subjects
Cancer patient
Healthy vs cancer
Healthy mix
H001.10
C002.2
C002.1
C
Distribution of cfChIP-seq reads of different marks
Observed reads
Estimated background
Genome
TSS
Enhancers
TSS only
TSS/Enhancers
Enhancers only
H3K4me1
H3K4me2
H3K4me3
Percent aligned reads
H012
H013
Sample
D
Cell-type enhancers in healthy subjects
Reads/region
Multiple Neutrophils
Monocytes
Liver
T-Cells
B-Cells
Heart
Brain
E
CDX1
TCF3
Healthy (33 samples)
AMI
Cancer
Surgery

Supplemental Figure S5

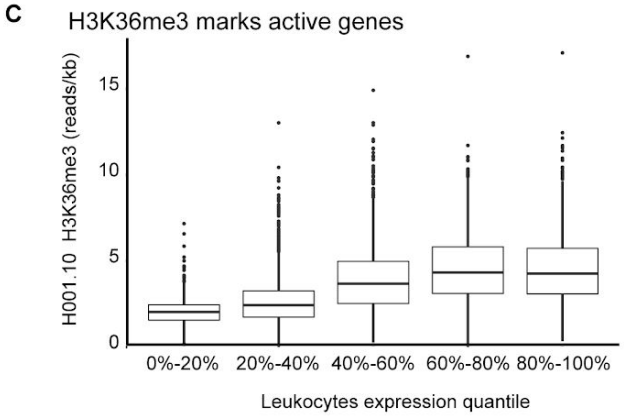
A. Comparison of H3K4me2 with H3K4me3 signal per gene in a healthy subject and cancer patient
B. Comparison of H3K4me2 signal per gene between samples.
C. Distribution of reads for cfChIP-seq with different antibodies on four samples (H012.1, H012.2, H013.1, and H013.2). We divided the genome into regions 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 region. Within each bar, the fraction estimated as background (based on our background model, Methods) is marked in dark gray.
D. Box plot for cfChIP signal in tissue-specific enhancers (as in Figure 6B)
E. Distribution of CDX1 and TCF3 in healthy samples and different patient samples (as in Figure 5D).

A Average singal of H3K36me3 over genes

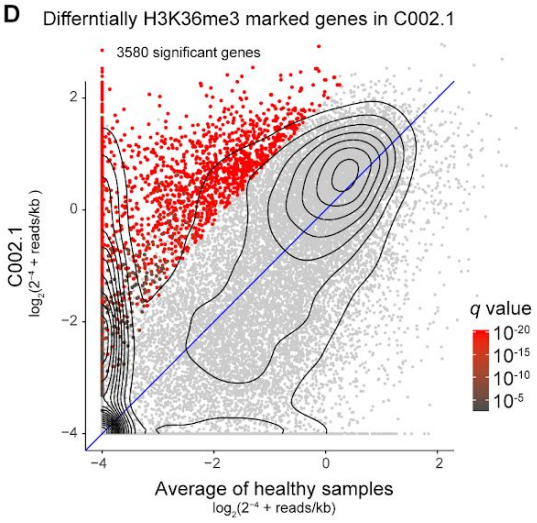
B Comparison of H3K36me3 on gene body and H3K4me3 on their promoters



C



D



Supplemental Figure S6

A. Metaplot showing average H3K36me3 levels along gene body.
B. Comparison of H3K4me3 and H3K36me3 levels per gene (H3K4me3 on promoters, H3K36me3 on gene body).
C. Level of average H3K36me3 levels along genes increases with expression level.
D. Comparison of H3K36me3 levels per gene between C002.1 and healthy reference.

Supplemental Tables

Table S1: Sequencing statistics for each sample

Sequencing statistics per sample, yield, etc.

Table S2: Subject clinical information

Subject basic info, times of collection

Table S3: Cluster annotations

EnrichR (Kuleshov et al. 2016) results for each of the clusters of Figure 3. Fields are as follows. Database: database name from the EnrichR databases, Cluster: cluster number, Term: name of enriched term, Overlap: X/Y X genes from cluster in Y genes in term, N: number of genes in the cluster, Adjusted.P.value: $-\log_{10}$ p from EnrichR, Q.value: corrected p-value as $-\log_{10}$ q, Genes: genes in the overlap.

Table S4: Cell type signatures

Each line lists one genomic window with the following fields: Signature: name of the signature (e.g., Liver, or T-cells), Chr: chromosome, Start: start location, End: end location, Gene: gene name if there is a gene (or genes) associated with the location, Type: TSS or background., Browser: browser address line that contains the region + 10kb to each direction, Signal: maximal normalized signal in this window in tissues that belong to the positive class, Background: maximal normalize signal in tissues that do not belong to the positive class.

Table S5: Full analysis of signatures vs samples

Each line list one sample. The first group of columns list the signature value (normalized reads/kb) of the signature in the sample. The second group of columns (header XX-qvalue) lists the $-\log_{10}$ q of the test whether this value is above background.

Table S6: Compendium of pathways/complexes

Table S7: Analysis of pathways/complexes

Z-scores pathway changes.

Table S8: Significantly high genes in each sample

Each line lists a gene in one sample. Fields are: Sample: name of sample, Gene: name of gene, Observed

normalized counts in sample ($\log_2(1+x)$ transformed), Healthy normalized counts in healthy reference group ($\log_2(1+x)$ transformed), FoldChange: difference between observed and healthy, p-Value: $-\log_{10} p$ for a test that observed number of reads is higher than expected value according to healthy reference, Q-Value: $-\log_{10}$ q corrected p-value, Browser window: browser coordinates around gene.

Table S9: Tumor signatures

Table S10: Tumor enrichment

$-\log_{10}$ q-values for hypergeometric enrichment test for highly expressed genes in samples (rows) against tumor types (columns). All q-values above 0.001 (e.g., $-\log_{10} q < 3$) are set to zero.

Table S11: Data sources

Table S12: Roadmap samples used

Supplementary Note

Estimate of cfChIP-seq efficiency

To get a rough estimate of the cfChIP procedure efficiency, we consider two alternative estimates: a **global** method and a **local** method.

Both start by estimating the amount of material in the plasma.

Assuming one diploid cell contains 6.6pg of DNA, we can use the measured amount of cfDNA in each sample to compute the number of genomes:

$$N_{genome} = 2 * V * C / 0.0066$$

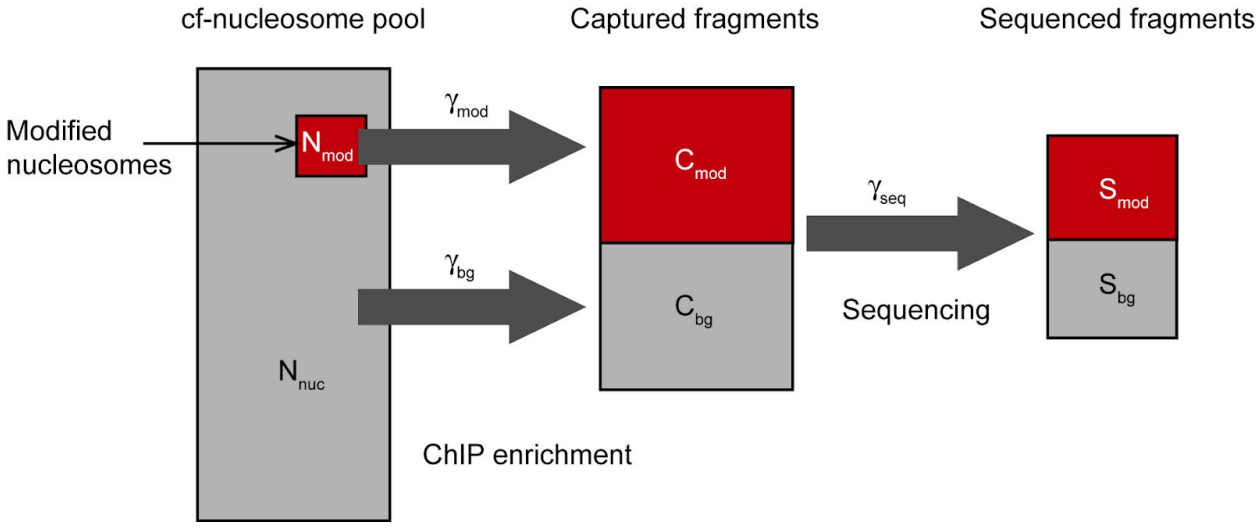
Where V is the sample volume (in ml), and C is the cfDNA concentration (in ng/ml).

Assuming an average nucleosome consists of ~200bp (including linker), each genome is packed by ~ $1.65*10^7$ nucleosomes and we compute

$$N_{nuc} = N_{genome} * 1.65 * 10^7$$

This is likely an overestimate of the amount of input material, since not necessarily all cfDNA is packed in nucleosomes.

The **global method** is described in the following scheme:



To estimate the total number of nucleosomes marked by a certain modification, we assume that the same percent of the genome is marked in every cell. Denote this to be (p_{mod}) we have:

$$N_{mod} = N_{nuc} * p_{mod}$$

If we estimate the number of marked nucleosomes we captured in total (C_{mod}) we can then estimate the yield as

$$\gamma_{mod} = C_{mod}/N_{mod}$$

If we also estimate the number of unmarked (background) nucleosomes we captured (C_{bg}) we can also estimate the rate of non-specific capture by comparing it to the total input

$$\gamma_{bg} = C_{bg}/N_{nuc}$$

This method raises two issues. 1) How do we estimate the percent of nucleosomes in a cell marked by every modification, and 2) how do we estimate the fraction of sequenced reads that originated from modified nucleosomes vs. background reads.

For the percent of marked nucleosomes we use a rough estimate based on ChIP-seq of specific blood cells and other quantitative studies of modifications. Single molecule imaging (Shema et al. 2016) in multiple cell types show that H3K4me3 levels are fairly consistent at 2% while quantitative MS suggest less than 1% (Zheng, Huang, and Kelleher 2016). Based on H3K4me3 ChIP-seq of blood cells we estimate the percent to be:

$$p_{K4me3} = 0.01 \quad (1\%)$$

Estimation of other marks is roughly as follows:

$$p_{K4me2} = 0.025 \quad (2.5\%)$$

$$p_{K4me1} = 0.10 \quad (10\%)$$

$$p_{K36me3} = 0.17 \quad (17\%)$$

The second question is how to estimate the amount of captured modified nucleosomes. Here we use genomic areas that are not expected to contain marked nucleosomes (based on ChIP-seq of multiple tissues) to estimate the observed background rate α in terms of reads/kb. (In fact we use a more nuanced estimate, see below on the precise estimation procedure.) According to this logic:

$$S_{bg} = \alpha * L_{genome}/1000 \quad \text{and} \quad S_{mod} = S_{total} - S_{bg}$$

Where S_{total} is the total number of unique fragments we recovered and $L_{genome} = 3.3 * 10^9$ is the length of the genome.

Estimating the number of captured fragments from the number of sequenced fragments requires considering the sequencing efficiency --- what fraction of the captured fragments (ones that reached the last PCR step) were actually sequenced. We estimate this efficiency by examining the distribution of duplicate reads in our sequencing output (more below). Given the estimate

γ_{seq} we estimate

$$C_{mod} = S_{mod}/\gamma_{seq}$$

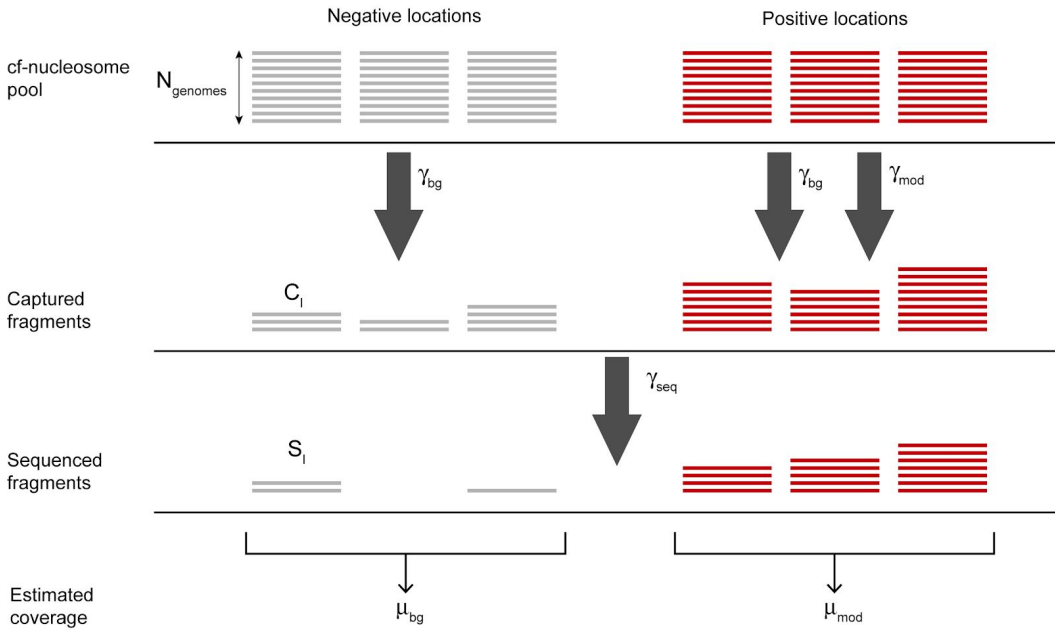
Putting these all together we get:

$$\gamma_{mod} \approx \left(S_{total} - \alpha * 3.3 * 10^6\right) * 0.0066/\left(\gamma_{seq} * p_{mod} * V * C * 2 * 1.6 * 10^7\right)$$

And

$$\gamma_{bg} \approx \alpha * 3.3 * 10^6 * 0.0066/\left(\gamma_{seq} * V * C * 2 * 1.6 * 10^7\right)$$

The alternative **local method** reasons that every modification has genomic areas which tend to be fully marked and others that are completely unmarked. For instance, nucleosomes flanking promoters of "housekeeping genes" are expected to be marked by H3K4me3 in all cells, while areas distant from genes and enhancers are expected to be unmarked.. The coverage at these areas depends on the rates aforementioned and thus can be used to estimate cfChIP efficiency.



More precisely, let l be a genomic location, and S_l the number of unique sequenced fragments that overlap with it. Then for an unmarked location:

$$E[S_l] = N_{genome} * \gamma_{bg} * \gamma_{seq} \quad \text{(negative location)}$$

We emphasize the expectation, since at each specific location the number of fragments will be distributed around this mean.

In a positive location, where we expect that all input nucleosomes are marked, the expectation is

$$E[S_I] = N_{genome} * (\gamma_{mod} + \gamma_{bg}) * \gamma_{seq} \quad \text{(positive location)}$$

The sum $\gamma_{mod} + \gamma_{bg}$ is due to the fact that a fragment can be captured in a specific or non-specific manner.

To estimate capture rate, we define the coverage in positive location to be the coverage in the 95'th percentile of non background regions

$$Cp = quantile(95, C_{mod})$$

And the coverage in negative locations μ_{bg} to be the mean coverage at background regions. Using these we conclude that

$$\gamma_{bg} \approx \mu_{bg} / \left(N_{genome} * \gamma_{seq}\right)$$

And

$$\gamma_{mod} \approx Cp / \left(N_{genome} * \gamma_{seq}\right) - \gamma_{bg}$$

Note that this approach circumvents the need to estimate the total number of modified nucleosomes in cells. However, it does require to define positive locations that are assumed to be modified in all cells --- and while there are locations with strong ChIP-seq peaks in multiple cell types that seem to meet that requirement, we do not have any current experimental evidence that support such a claim.

Table S1 contains the estimates using both methods in the samples where we quantified DNA content.

Estimation of sequence efficiency

This problem has been examined in the literature (Daley and Smith 2013) and in various tools (e.g., Picard's EstimateLibraryComplexity). Here we derive a simple method that provide an initial estimate, although more complex ones exist.

We can view the sequencing protocols as starting with a small set of molecules that have the sequencing adapters. These are amplified (16 rounds of PCR) and some fraction of the amplified fragments is sequenced. Due to the large amplification and the numbers of initial fractions (~millions), we can think of the number of times we see a specific input molecule to be Poisson distributed with a parameter λ which summarizes the rate at which we see this specific fragment when collecting the actual number of sequences.

In processing the sequenced reads, we view duplicates as artifacts of the sequencing and not as two identical input molecules. Indeed, comparing two technical repeats from the same plasma sample, we find negligible overlap, suggesting that most duplicates are technical artifacts.

We can summarize the observed duplicate frequency as

$$\widehat{p}_k = n_k / n_{unique}$$

where n_k are the number of fragments that were duplicated k times, and n_{unique} is the number of unique fragments observed. Estimation of sequencing efficiency can be reduced to estimating n_0 the number of fragments we did not observe.

The simplest approach is to assume duplicates are due to a Poisson process with unknown parameter λ. To estimate this parameter we need a slightly modified procedure, as we do not observe fragments with $k = 0$. This results in a truncated likelihood function

$$l_{poisson}(\lambda) = \sum_{k=1} n_k \log(p(k|\, k > 0,\, \lambda) = \sum_{k=1} n_k \left(\log \tfrac{1}{k!} + k \log \lambda - \log(1 - e^{-\lambda})\right)$$

We can find the maximum likelihood

$$\widehat{\lambda} = \arg\max_{\lambda} l_{poisson}(\lambda)$$

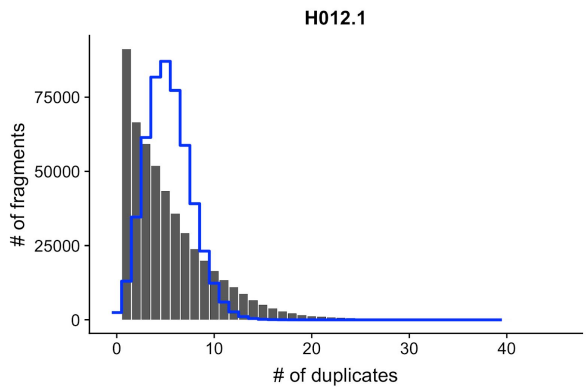
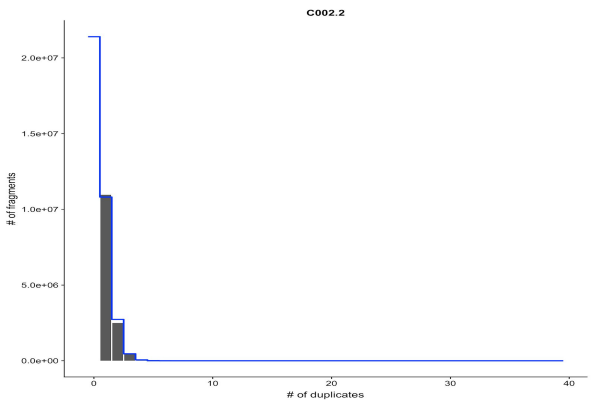
using line search. Once we have this parameter we can estimate the number of total fragments (including the ones we did not see). Briefly, based on the Poisson model, we conclude that

$$n_{unique} = n_{total} * p(k > 0 \mid \lambda^*) = n_{total} * (1 - e^{-\lambda^*})$$

Where λ^* is the unknown real rate. Since sequencing efficiency is the fraction of observed fragments

$$\gamma_{seq} = n_{unique}/n_{total} = 1 - e^{-\lambda^*} \approx 1 - e^{-\widehat{\lambda}}$$

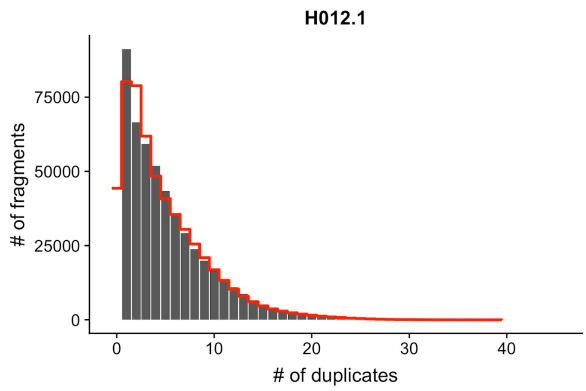
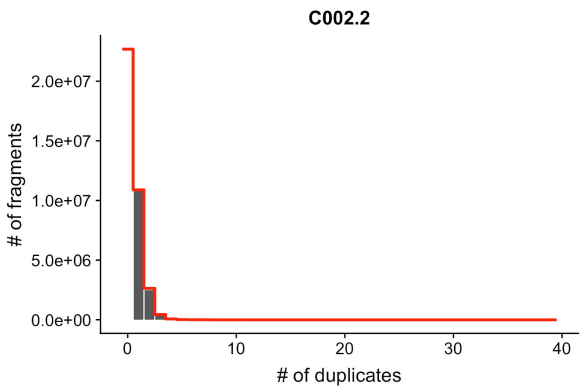
In many cases the actual distribution of duplicates does not match a Poisson distribution. For example

	
H012.1 - Poisson fit (blue) does not match observed duplication frequency (gray). Note that this sample was sequenced deeply, and the estimate is that there are few unseen sequences. Estimate of γ_{seq} is 99.5%	C002.2 - Poisson fit matches nicely the observed duplication frequency. This sample was sequenced much more shallow and most fragments were seen only once. The estimate is that there are many more unseen sequences. Estimate of γ_{seq} is 39.7%

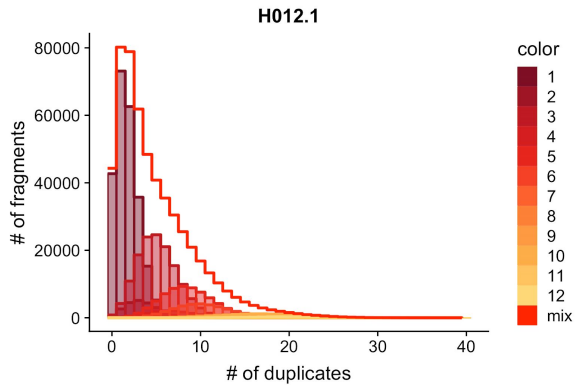
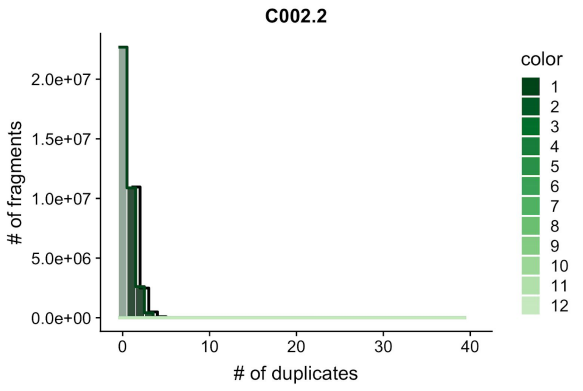
We reasoned that this discrepancy might be due to some fragments undergoing early duplication(s), thus amplified more than the rest. However, since the actual rate depends on decisions downstream of amplification (e.g., loading libraries onto sequencer etc), we need to assume that the rate of reproducing each copy of these “early winners” is the same as the rest of the fragments. This means that we assume a mixture of Poissons:

$$p_{mix}(k \mid \lambda,\ \rho_1 ..., \rho_M) \ = \ \sum_{m=1}^{M} \rho_m * p_{poisson}(k \mid m * \lambda)$$

Where $\rho_1 ..., \rho_M$ are mixture weights that estimate the proportion of fragments that are amplified once, twice, thrice and so on. Fitting this mixture model requires performing numerical optimization to find parameters that maximize the likelihood. For our two examples we get:

	
H012.1 - with mixture fit (red line). This is much closer to the empirical distribution, and makes a substantially different estimate of the number of unseen fragments. Estimate of γ_{seq} is 91.4%	C002.2 - here the mixture fit is close to the Poisson fit above. Estimate of γ_{seq} is 38.3%

We can gain better sense of these by examining the individual components that make up the mix.

	
H012.1 - the mixture is made up of several components	C002 - the mixture is dominated by the first component

Our conclusion (based on these two examples and more) is that for shallow sequenced samples (e.g., similar to C002.2), the two estimators of efficiency are similar, while for deeply sequenced samples (e.g., similar to H012.1), there are differences in the estimates. Since in these cases the mixture model is closer to the observed duplication frequency we consider it as the more accurate estimate.

To be clear, the estimate of sequencing efficiency according to this model is similar to above:

$$\gamma_{seq} \approx p_{mix}(k > 0 \mid \lambda, \rho_1, ..., \rho_M) = 1 - \sum_{m=1}^{M} \rho_m * e^{-m*\lambda}$$

TSS location catalogue

We constructed a TSS catalogue through the following steps. All steps were carried on the "hg19" human genome assembly:

1. We downloaded ChromHMM calls for 111 tissues and cell types throughout the human genome from Roadmap Epigenomics website (Table S7). UCSC Browser known gene annotations and ENSEMBL transcript annotations were downloaded (Table S7).

2. We filtered all genomic ranges that were marked with states "1_TssA" or "2_TssAFlnk", and merged adjacent ranges that were marked as either state in exactly the same set of tissues. We call these "ChromHMM TSS windows". We found 640,744 such windows. Each ChromHMM TSS window was assigned gene name(s) through the following steps:

    a. If it was within 2.5Kb of one or more TSSs in UCSC known gene annotation, it was assigned the name of these genes.

    b. If not, we searched for an ENSEMBL transcript start within 2.5Kb. Again if such were found the TSS window received the gene name associated with the transcript.

    c. All other TSS windows remained without a name

    This procedure assigned names to 190,552 windows, and 450,192 windows remained unnamed. In some cases the unnamed TSS windows can be seen as alternative starts of an annotated gene (either upstream or downstream of the annotated TSS), but in many other cases these were far from transcript starts. In general, after correcting for length, the rate of seeing a read from H3K4me3 cfChIP in unnamed windows is 6-fold smaller than that of named windows. This suggests that most of the unnamed windows are not observed in our data. In terms of coverage, named TSS windows cover 4.46% of the genome, and unnamed ones an additional 7.03%. Given the estimate of 1-2% of nucleosomes in a cell carrying the H3K4me3 mark, we conclude that most of these putative promoters are not used in any given cell.

3. To include transcripts that are not represented in the TSS catalogue, we examined all genes in the UCSC known gene database and all transcripts in the ENSEMBL database.

For each gene we defined a TSS window of size 3Kb centered on the TSS. We discarded all such windows that overlapped with a TSS window described in step 2. In total this step added 11,518 and 36,166 TSS windows from UCSC known genes and ENSEMBL transcripts, respectively.

4. We created windows that tile the remaining genomic regions between TSS windows. For each TSS window without an adjacent TSS window, we created “flanking” regions of size 1Kb (or less). This resulted in 475,727 flanking windows (as some of the TSS windows are adjacent to each other, depending on ChromHMM calls in different tissues). The remaining uncovered regions were tiled with “background” regions of size 5Kb (or less). In total there were 485,245 such windows.

The resulting catalogue was saved as BED file (TSS.bed). This catalogue was used in all analysis of H3K4me3 cfChIP and re-analysis of reference ChIP-seq data.

Enhancer location catalogue

The construction of enhancer catalogue was similar to that of the TSS catalogue, except for Step 2 above where we included the enhancer states “6_EnhG” and “7_Enh”. Similarly to the TSS catalogue construction, here too we merged consecutive enhancer windows that are called in the same tissues. In Step 3 we only added additional TSSs (as above). Step 4 was left unchanged, except for the fact that we considered regions that were neither TSS nor Enhancers.

The resulting catalogue contained 2,345,831 enhancer windows,, most of which were relatively short: 91.9% of the windows are 1kb or shorter, and 49.2% are only 200bp long (the basic length of a ChromHMM call). Nonetheless, the these windows covers 40% of the genome.

We used this for the analysis of H3K4me1 and H3K4me2 cfChIP.

Gene body location catalogue

The construction this catalogue was similar to that of the TSS catalogue, except that in Step 2 we collected only the transcribed states “3_TxFlnk”, “4_Tx”, and “5_TxWk”. We named regions if they overlapped with a annotated gene or transcript.

In Step 3 we added UCSC known gene database and all transcripts in the ENSEMBL database and added the gene body, if it did not overlap with regions from Step 2. The resulting gene regions are often very long, and thus before Step 4, all gene regions were tiled with 5kb size windows. Step 4 was left unchanged, except it considered regions that were not genes.

The resulting catalogue contained 1,021,467 gene windows based on ChromHMM calls, and an additional 4,301 and 861 gene windows based on UCSC genes and ENSEMBL transcripts that were not covered by ChromHMM. This is consistent with the fact that most genes were expressed in one of the tissues covered by the Roadmap Epigenomics project. Most of the

genome (71%) was covered by these windows --- recall that these include both exonic and intronic regions.

We used this catalogue for the analysis of H3K36me3 cfChIP.

Processing of sequencing files

Base calling was performed with bcl2fastq (2.18). Paired-end reads were mapped to the human genomes ('hg19' assembly) using bowtie2 with "no-mixed" and "no-discordant" flags descarding reads with quality 0. BEDPE files (start and end of every fragment) were obtained using BEDtools "bamtobed" with "bedpe" flag, discarding duplicate fragments.

BEDPE files were converted to coverage counts over windows in the catalogue using BEDtools "intersect" command and also using BioConductor "GenomicRanges" countOverlaps() function. Both methods count for each window the number of sequenced fragments that overlap with the window.

Estimating background signal

Every ChIP procedure has non-specific background signal. In the case of cfChIP the background is due to some forms of non-specific binding of DNA and chromatin fragments to the beads-antibody complex. Our experience showed that the background levels varied between samples and batches of bead-antibody ligation. Moreover, the sequencing depth varied between samples, and in a deeply sequenced sample the number of background reads increases. Thus, it was important to estimate background signal levels to be able to contrast them with actual signal.

We initially applied a simple minded procedure for removing background in H3K4me3 signal. We reasoned that virtually all of the specific signal in H3K4me3 is at TSS and gene 5' regions. Thus, reads in other locations represent background. To account for TSSs that are not annotated in our TSS catalogue, we reasoned that some small fraction of background windows might contain real signal, and thus we removed the ones with the highest values.

In more detail we did the following. We created a vector with the coverage of all "background" windows of size >= 4Kb (323,237 out of 485,245). And applied the following procedure:

estimateBackground (X)

$T \leftarrow quantile(95,\ X)$ // find the 95th quantile of X

$X \leftarrow X[X \leq T]$ // restrict ourselves to values below T

$\hat{\lambda} = \arg\max_{\lambda} \prod_{i=1}^{|X|} P_{\lambda}(x_i \mid x_i \leq T)$ // maximum likelihood of truncated poisson

Return $\hat{\lambda}/5$ // convert to reads/Kb

This procedure was relatively robust to the choice of quantile for removing outlier windows.

However, in some samples the Poisson distribution was not a good fit for background values. Further examination revealed that much of this discrepancy was due to local background effects. One local effect is the sex chromosomes that appear in 50% levels in males, and 100% (X) and 0% (Y) in female. These were not the only local effects - some regions showed higher levels of background. This could be due to segmental duplication (regions close to centromeres and telomeres) or accessibility issues. Moreover, in cancer samples there were clear aberrations that were patient specific.

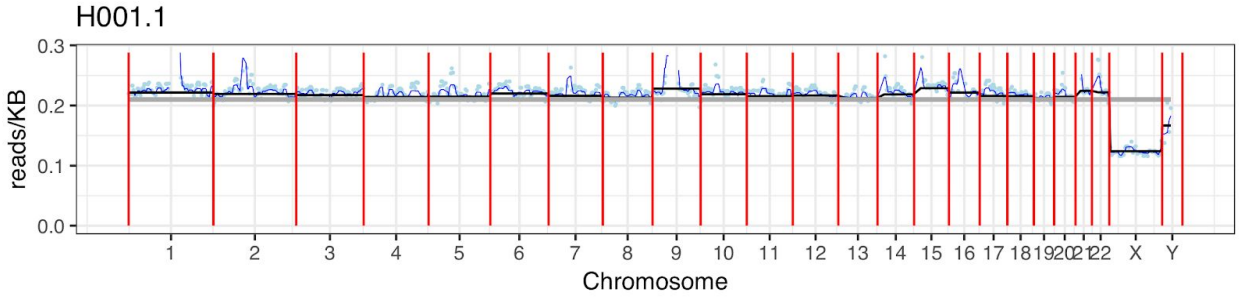
To overcome these issues, we devised a localized background rate estimate. We used the above estimation procedure, but in successive levels of resolution.

1. Genome-wide background level.
2. Chromosome-specific background.
3. Tiles of 10Mb covering each chromosome at offsets of 2.5Mb.
4. Tiles of 5Mb covering each chromosome at offset of 1.25Mb.

The estimate at each level used the estimate of the previous level as a prior (using pseudo-counts of 1000 windows in levels 2 and 3, and 500 windows in level 4).

The result is an estimate of background coverage rate at overlapping tiles of 5Mb. To get a single estimate, for each location we take the maximum of the estimate of the tiles covering it (typically 4 tiles). We choose the maximum as we reasoned that over-estimate of background might reduce the estimated signal but would reduce the number of background artifact.

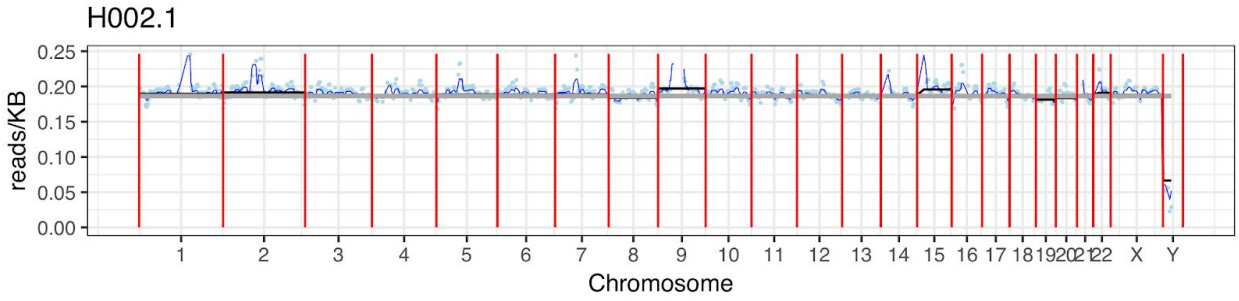
Example of the background estimate for a healthy sample (male)



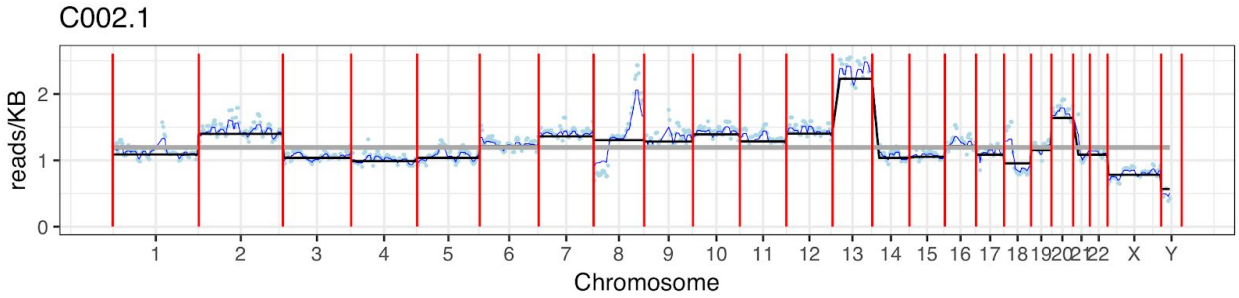
The gray line displays the genome-wide estimate, the black lines the chromosome-specific estimate, the blue line the larger tile-based estimates, and the light blue points the local estimate (based on the smaller tiles)

In this example the chrX background is lower than autosomal chromosomes (slightly more than half) and chrY background is a bit lower. Many locations in chrY are orthologous to ones in chrX leading to skewed estimates. Other deviations occur close to centromeres where we find the background level to be higher in some chromosomes (e.g., chr1, chr9).

A healthy female sample is similar, except for the sex chromosomes.



When we examine a patient with cancer, the background estimate is much more variable, presumably reflecting chromosomal aberrations in the tumor (e.g., duplication in chr13 and in an arm of chr8).



Gene-level signal and normalization

For each gene we assigned a set of TSS windows that are annotated with the gene name. For each sample we computed the actual total coverage over the windows assigned to the gene and the expected mean of background reads over these windows (using possibly different local rate at each window and the window size).

More precisely,

$$C[g,s] = \sum_{w \in W_g} C[w,s]$$

$$B[g,s] = \sum_{w \in W_g} \hat{\lambda}[w,s] * width(w)$$

where W_g is the set of windows assigned to gene g, $C[w,s]$ the coverage of window w in sample s, and $\hat{\lambda}[w,s]$ the estimated background rate of window w in sample s.

The null assumption is that the coverage C_g is distributed as a Poisson with parameter B_g. Thus, we argued that values much larger than expected are signal. We define the raw signal at gene g as

$$S[g,s] = C[g,s] - B[g,s] \quad \text{if } C[g,s] \geq B[g,s] + 2\sqrt{B[g,s]} \text{ and 0 otherwise}$$

Thus we consider $C[g,s]$ to be a real signal if it is larger than two standard deviations from the mean of the background level for the gene.

Applying this procedure for each sample generates a matrix of counts for each gene in each sample. We also include in this matrix the samples from Roadmap Epigenomics data of H3K4me3 ChIP which were processed in the same manner.

To normalize the effect of different coverage, we reasoned that the signal at promoters of "housekeeping" autosomal genes should be similar in different samples. We defined these genes as ones with highly significant signal in a set of reference healthy samples. The precise choice of significance level did not change the normalization.

On the matrix of raw signal samples X housekeeping genes we applied quantile normalization (Bioconductor normalize.quantiles()). This resulted in normalized values for housekeeping genes in each sample. However, it does not assign values for all other genes. We thus, estimate a multiplicative normalization factor for each sample to best match quantile-normalized values to raw values. For most samples the relation between the two was linear.

The scaling factors were rescaled so that the total normalized signal (below) at the set of reference healthy samples will be a million on average.

Using these normalization factors, $\nu[s]$ we computed for each sample the normalized gene levels

$$N[g,s] = \nu[s] * S[g,s]$$

Using the same normalization procedure we also normalized the coverage at each window in each sample

$$N[w,s] = \nu[s] * \max(C[w,s] - B[w,s], 0)$$

Defining tissue-specific signature

Using the Roadmap Epigenomics metadata table we defined sets of Roadmap samples that belonged to a tissue or group of tissues (see Table S8). These definitions included some redundancies. For example, the group Lymphocytes included B-Cells, T-Cells, and NK samples, and thus subsumed each of these groups.

We then defined for each group the set of specific windows, as windows w passing the following criteria:

1. The window w is on an autosomal chromosome
2. In at least one of the atlas samples in the group, $N[w,s] \geq 35$

3. In all atlas samples outside the group, $N[w,s] < 15$
4. In all windows w' within 1Kb of w, $N[w,s] < 15$

The last condition is added as we noticed that often when a gene is expressed there is "spill over" to neighboring windows.

Groups for which we found less than 4 specific windows were considered to be without signature. For all other groups, we define the signature as the set of specific windows (see Table S2).

Statistical tests

We use two main tests in the manuscript.

Detection test. To test whether a gene or a signature is present above background in a sample, we used a Poisson distribution. More specifically:

computeDetectionPValue (W, s)

$\lambda \leftarrow \sum_{w \in W} B[w,s]$

$x \leftarrow \sum_{w \in W} C[w,s]$

Return $P_\lambda(X \geq x)$ // Poisson p-value

Here W is a set of windows, it can be the windows associated with a gene or tissue-specific signature as above.

High coverage test. To test whether the observed signal of a set of genes is higher than expected in healthy samples, we used a reference of healthy subjects to define the expected normalized signal of the gene $H[g]$ as the average of $N[g,s]$ in the reference samples.

We then used the following procedure

computeOverExpressionPValue(G,s)

$\lambda \leftarrow \sum_{g \in G} B[g,s] + \frac{1}{\mathrm{v}[s]} * \sum_{g \in G} H[g]$

$x \leftarrow \sum_{g \in G} C[g,s]$

Return $P_\lambda(X \geq x)$ // Poisson p-value

The main difference from the previous test is that we included the contribution of healthy samples after we transform from normalized units to the units of the specific sample. The second difference is that we work at the level of genes.

References

Daley, Timothy, and Andrew D. Smith. 2013. “Predicting the Molecular Complexity of Sequencing Libraries.” *Nature Methods* 10 (4): 325–27.

GTEx Consortium. 2015. “Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans.” *Science* 348 (6235): 648–60.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. “Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.” *Nucleic Acids Research* 44 (W1): W90–97.

Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, et al. 2015. “Integrative Analysis of 111 Reference Human Epigenomes.” *Nature* 518 (7539): 317–30.

Shema, Efrat, Daniel Jones, Noam Shoresh, Laura Donohue, Oren Ram, and Bradley E. Bernstein. 2016. “Single-Molecule Decoding of Combinatorially Modified Nucleosomes.” *Science* 352 (6286): 717–21.

Zheng, Yupeng, Xiaoxiao Huang, and Neil L. Kelleher. 2016. “Epiproteomics: Quantitative Analysis of Histone Marks and Codes by Mass Spectrometry.” *Current Opinion in Chemical Biology* 33 (August): 142–50.