
Supplementary information

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

In the format provided by the
authors and unedited

Supplemental note to Sadeh et al. “ChIP-seq of plasma cell-free nucleosomes identifies gene expression programs of the cells-of-origin”

Assay reproducibility	2
TSS location catalogue	5
Enhancer location catalogue	6
Gene body location catalogue	6
Processing of sequencing files	7
Estimating background signal	7
Gene-level signal and normalization	10
Estimate of cfChIP-seq capture efficiency	11
Estimation of sequencing efficiency	16
Sensitivity Analysis	19
Defining tissue-specific signatures	22
Statistical tests	22
Cancer programs	25
References	26

Assay reproducibility

To assess the reproducibility of cfChIP-seq, we performed technical and biological replicates from several subjects. Specifically, two healthy individuals (H012 and H013, one female and one male) were sampled in two biological repeats, after finalizing the development of the experimental and analysis methods. These samples were not part of our method development as objective test cases. This experiment was run once and we used the two samples without any cherry picking.

Reproducibility can be assayed at multiple levels. The simplest is to compare the raw cfChIP-seq signal of samples. We tiled the genome with 2kb tiles, and counted the number of reads mapped onto each tile. As might be expected, most of the tiles had zero observations, but some had higher values. See Figure S1 for some examples.

This comparison shows that some samples have differences in yields (either ChIP or sequencing efficiency, see below). In general the R^2 values for H3K4me3 are high (~0.8-0.95), for H3K4me2 are somewhat lower (~0.6-0.8) and much lower for H3K4me1 and H3K36me3. This is consistent with the sharply concentrated nature of H3K4me3 peaks, and the much more diffuse patterns of H3K4me1 and H3K36me3.

A more relevant comparison of samples is by examining the estimated counts per gene. These estimates remove non-specific background and normalize samples to reduce artifacts of differential yield levels. Indeed, comparing normalized gene signal profiles shows much better agreement between healthy samples (Extended Data Figure 2B).

When examining all 88 healthy samples of H3K4me3, we can see that normalized gene correlations are much better than raw 2kb counts (compare Figure S2A vs Figure S2B). In particular biological repeats, tend to be of higher correlations.

Comparison of raw read counts in 2K windows tiling the genome

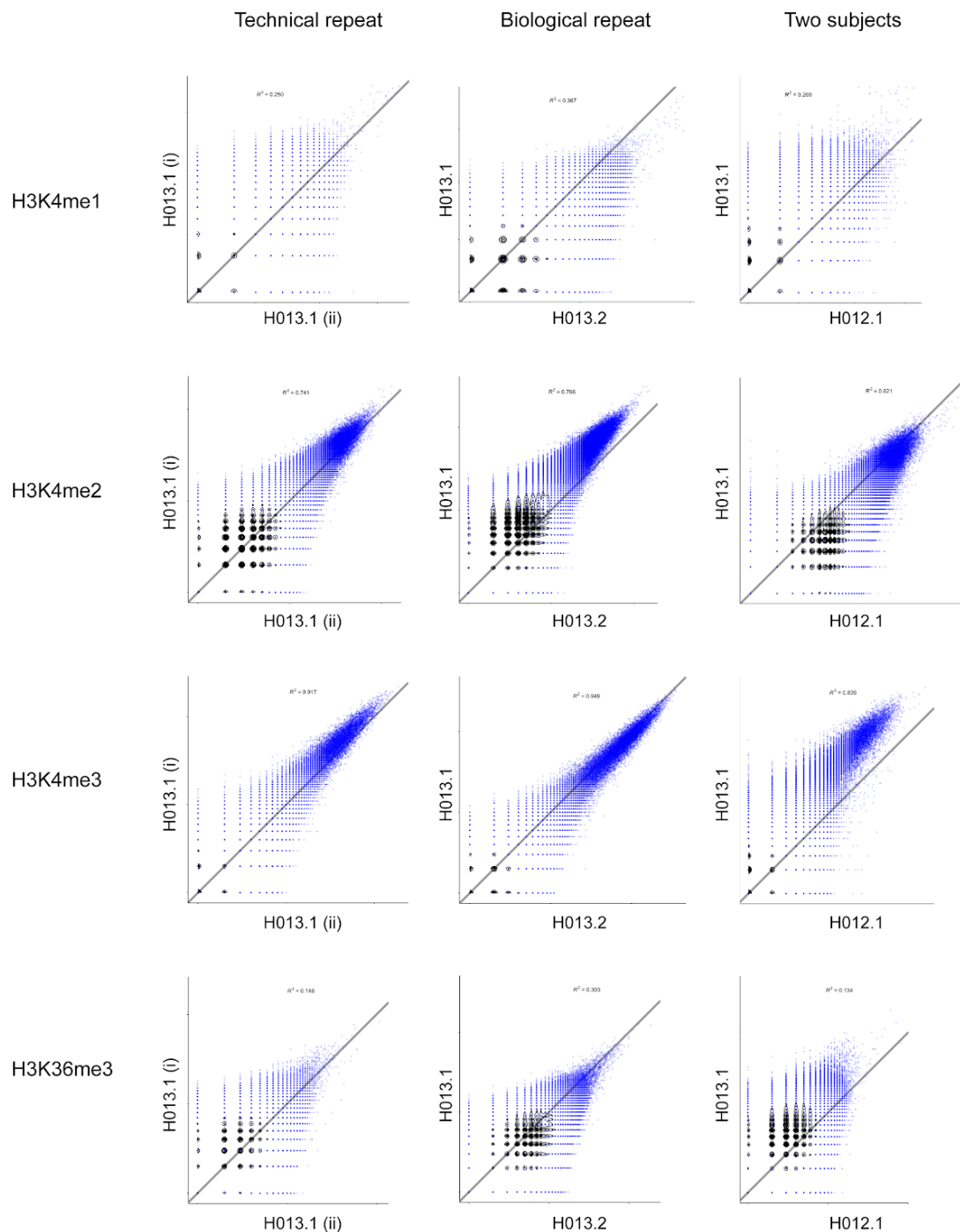


Figure S1: Comparison of signal in 2kb tiles along the genome.

For each modification (rows) we compared two technical repeats (left column), biological repeats (middle column) and two donors (right column). Each plot shows log(1+count) in one sample to the other. Gray line indicates the x=y line.

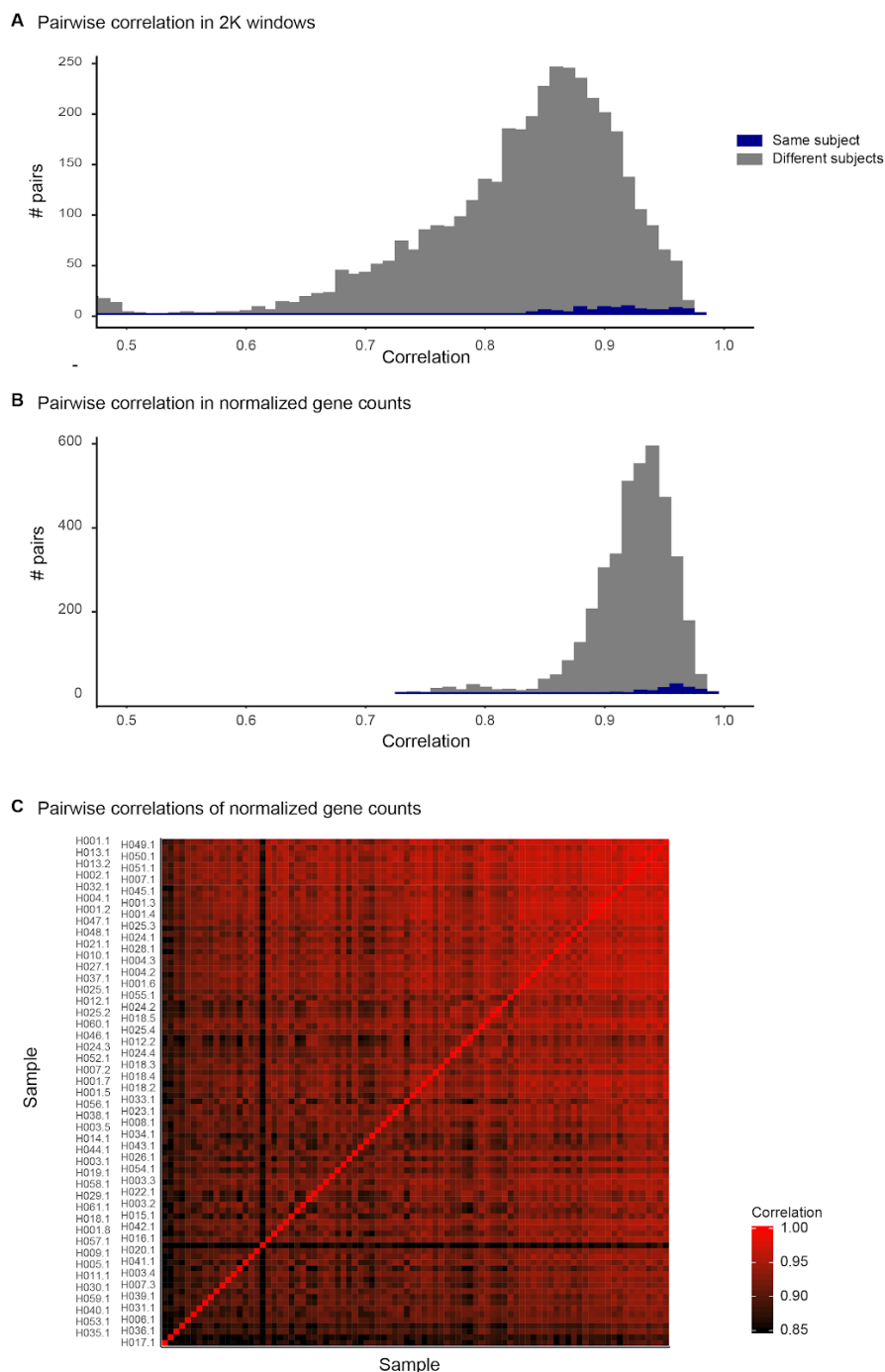


Figure S2: Reproducibility of H3K4me3 profiles.

- A. Histogram of pairwise correlations between healthy samples at the level of 2kb tiles.
- B. Histogram of pairwise correlations between normalized gene counts.
- C. Heatmap showing pairwise correlations of gene counts.

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 the Roadmap Epigenomics website (see Data Availability). UCSC Browser known gene annotations and ENSEMBL transcript annotations were downloaded (see Data Availability).
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 TSS 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-seq 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 cell types we observe. Given observations of H3K4me3 signal at enhancers¹, these might be signatures of active enhancers, but this would require additional verification.

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 in case of collisions). 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 background windows.

The resulting catalogue was saved as a BED file (TSS.bed). This catalogue was used in all analysis of H3K4me3 cfChIP-seq 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, 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 unit length of ChromHMM calls). Nonetheless, these windows cover 40% of the genome.

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

Gene body location catalogue

The construction of 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 an annotated gene or transcript.

In Step 3 we added the 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 additional 4,301 and 861 gene windows based on UCSC genes and ENSEMBL transcripts, respectively, that were not covered by ChromHMM. These numbers are 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-seq.

Processing of sequencing files

Base calling was performed with bcl2fastq (2.18/2.20). Paired-end reads were mapped to the human genomes ('hg19' assembly) using bowtie2 with "no-mixed" and "no-discordant" flags discarding 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 BioConductor "GenomicRanges" countOverlaps() function. Specifically, we assigned a fragment to a window if its center is within the window. In other words, all fragments were resized to length 1 while maintaining their center.

Estimating background signal

Every ChIP procedure has a non-specific background signal. In the case of cfChIP-seq the background is due to 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 located in TSS and gene 5' regions. Thus, reads in other locations represent background. To account for TSS that are not annotated in our TSS catalogue, we reasoned that some small fraction of background windows might contain real signal, and thus we ignored the ones with the highest values in estimation of background.

In more detail we did the following. We created a vector with the coverage of all "background" windows of size $\geq 4\text{Kb}$ (323,237 out of 485,245). The vast majority of these were 5Kb long. We then applied the following procedure:

```
estimateBackground(X)           // X vector of values
  T ← quantile(95, X)           // find the 95th quantile of X
  X ← X[X ≤ T]                  // restrict ourselves to values below T
   $\hat{\lambda} = \arg \max_{\lambda} \prod_{i=1}^{|X|} P_{\lambda}(x_i | x_i \leq T)$  // maximum likelihood of truncated poisson
  Return  $\hat{\lambda}/5$                 // convert to reads/Kb (the median length of
                                // background windows is 5KB)
```

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 example of local effect is the sex chromosomes that appear at 50% levels in male samples, and 100% (chrX) and 0% (chrY) in female samples. These were not the only local effects - some regions showed higher levels of background. This could be due to segmental duplications (mostly at regions close to centromeres and telomeres) or accessibility issues. Moreover, in cancer samples there were patient specific deviations that most likely reflect chromosomal aberrations.

To overcome these issues, we devised a localized background rate estimate. We repeated the estimation procedure described above, in different genomic resolutions:

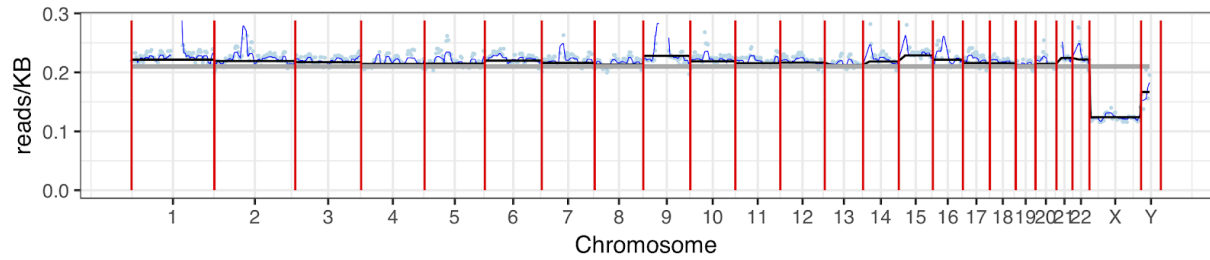
1. Genome-wide background level.
2. Chromosome-specific background.
3. Tiles of 10Mb covering each chromosome at offsets of 2.5Mb.
4. Tiles of 5Mb covering each chromosome at offsets of 1.25Mb.

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).

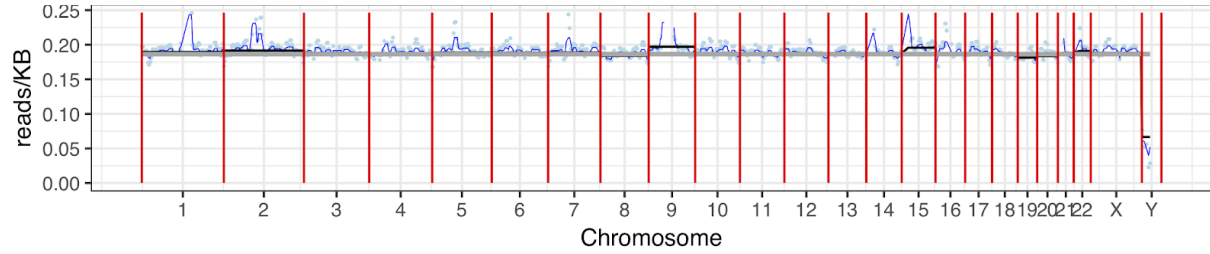
The result is an estimate of background coverage rate at overlapping tiles of 5Mb. To get a single estimate, for each location we take the maximum of the estimate of the tiles covering it (typically 4 tiles). We chose the maximum as we reasoned that over-estimation of background might reduce the estimated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

Example of the background estimate for a healthy male sample is in Figure S3A. 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 (Figure S3B) is similar, except for the sex chromosomes. When we examine a patient with cancer (Figure S3C), 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).

A. Healthy male



B. Healthy female



C. Cancer sample

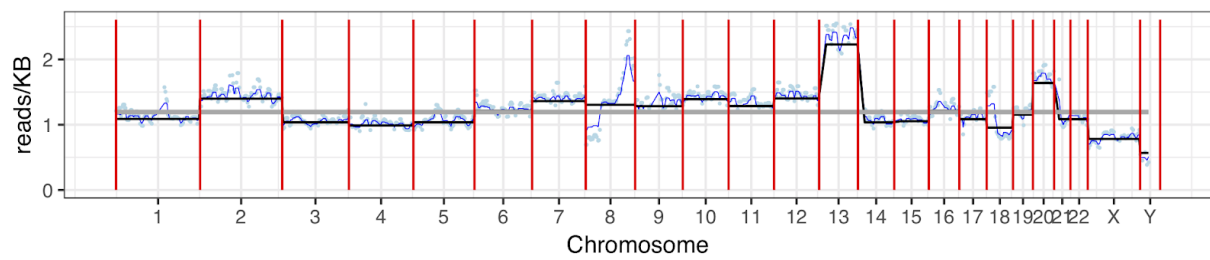


Figure S3: Variability in background coverage. Three samples are shown. For each we plot the estimated background at different resolutions: The gray line displays the genome-wide estimate, the black lines the chromosome-specific estimate, the blue line the larger tile-based estimates, and the light blue points the local estimate (based on the smaller tiles)

Gene-level signal and normalization

We define the following notation:

- $C[w, s]$ - read counts observed in window w in sample s
- $\hat{\lambda}[w, s]$ - estimated background rate in window w in sample s

Using these, we define

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

To be the background rate for window w in sample s . The rates we estimate are in units for counts/kb, and $width(w)$ returns the width of window w in units of kb, and so the result is a rate in units of counts.

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 (taking into account the local rate and the width of each window).

More precisely,

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

where W_g is the set of windows assigned to gene g .

The null assumption is that the coverage $C[g, s]$ is distributed as a Poisson with parameter $B[g, s]$. Thus, we argued that values much larger than expected are signals. We define the raw signal at gene g as

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

To normalize the effect of different coverages, 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 signals in a set of reference healthy samples (Supplementary Table 1). The precise choice of significance level did not change the normalization.

We applied quantile normalization (Bioconductor `normalize.quantiles()`) between the average signal of the housekeeping genes in the healthy reference set to the signal of the housekeeping genes in each one of the acquired cfChIP-seq samples. This resulted in normalized values for

housekeeping genes in each sample. However, it does not assign values for all other genes. We thus used linear regression to 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

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

Estimate of cfChIP-seq capture efficiency

To get a rough estimate of the cfChIP-seq 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 human 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 $\sim 1.65 * 10^7$ nucleosomes and we compute

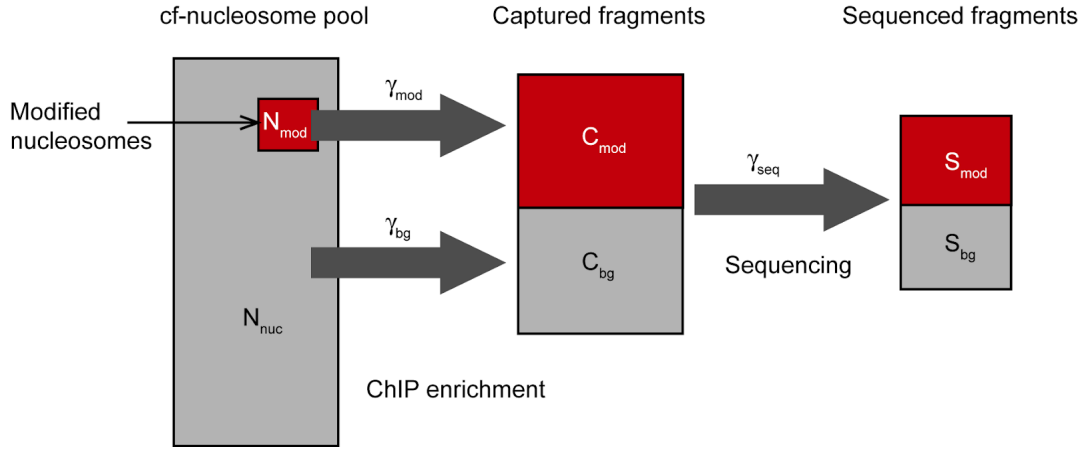
$$N_{nuc} = N_{genome} * 1.65 * 10^7$$

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

The **global method** is described in Figure S4A. 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:

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

A. Global method



B. Local method

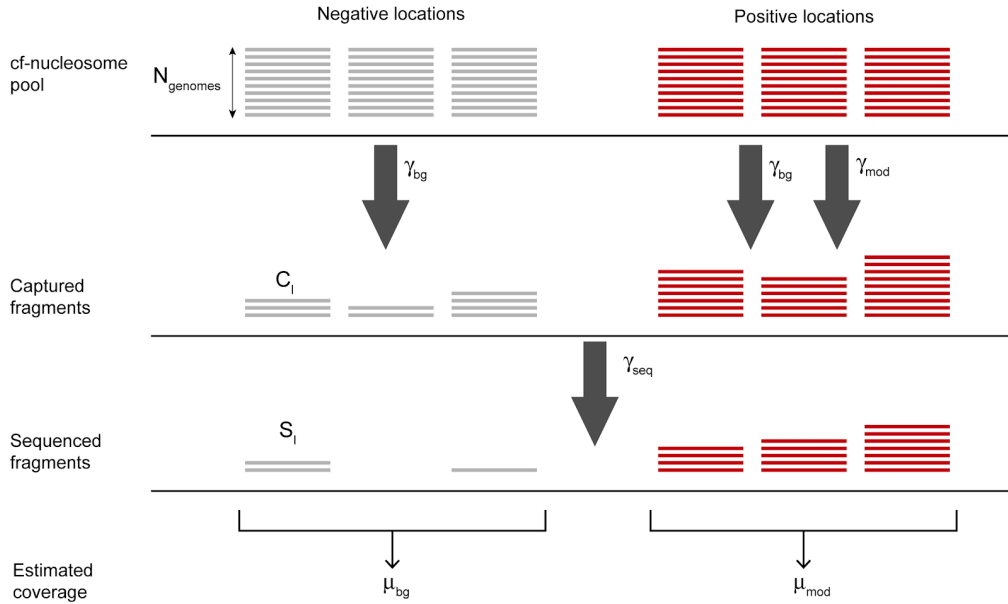


Figure S4: Two strategies to estimate capture rates.

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 cells² and other quantitative studies of modifications. Single molecule imaging³ in multiple cell types show that H3K4me3 levels are fairly consistent at 2% while quantitative MS suggest less than 1%⁴. Based on H3K4me3 ChIP-seq of blood cells we estimate the percent to be:

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

Estimation of other marks is roughly as follows:

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

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

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

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 γ_{bg} in terms of reads/kb (see section above). According to this logic:

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

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 calculate

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

Putting these all together we get:

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

$$\gamma_{mod} \approx (S_{total} - \gamma_{bg} * 3.3 * 10^6) / (\gamma_{seq} * p_{mod} * N_{genome} * 1.6 * 10^7)$$

And

$$\gamma_{bg} \approx \alpha * 3.3 * 10^6 / (\gamma_{seq} * N_{genome} * 1.6 * 10^7)$$

The alternative **local method** (Figure S4B) 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 constitutively marked genes (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-seq 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

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

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 locations to be the coverage in the 95'th percentile of non background regions

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

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

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

And

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

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. However, deviations from the assumption would imply that

the estimate is an underestimate of the actual capture rate.

Supplementary Table 1 contains the estimates using both methods in the samples where we quantified DNA content.

Estimation of sequencing efficiency

This problem has been examined in the literature⁵ and in various tools (e.g., Picard's EstimateLibraryComplexity). Here we derive a simple method that provides 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 are sequenced. Due to the large amplification and the numbers of initial fragments (~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, due to the small amount of input material in cfChIP, 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 amplification artifacts.

We can summarize the observed duplicate frequency as

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

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 that 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

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

We can find the maximum likelihood

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

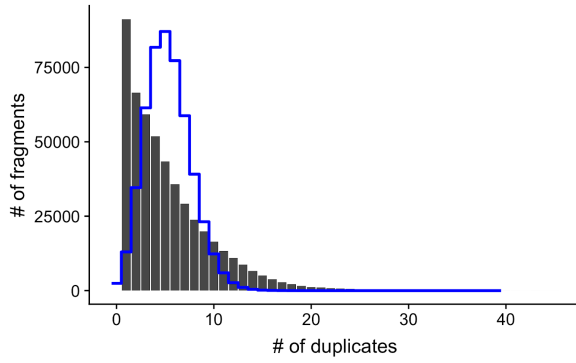
using line search. Once we have this parameter we can estimate the number of total fragments (including the ones we did not see). Briefly, based on the Poisson model, we conclude that

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

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

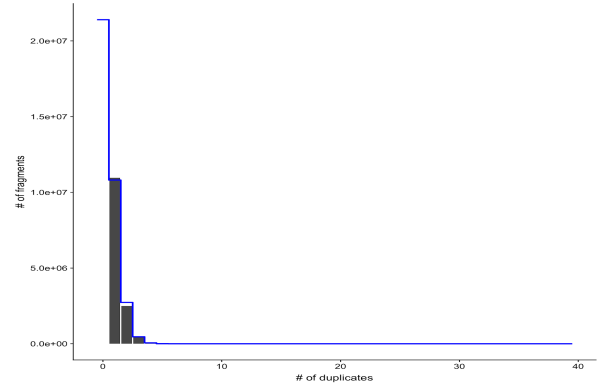
$$\gamma_{seq} = n_{unique}/n_{total} = 1 - e^{-\lambda^*} \approx 1 - e^{-\hat{\lambda}}$$

In many cases the actual distribution of duplicates does not match a Poisson distribution. For example Figure S5A.



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%



C001.2752 - 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.

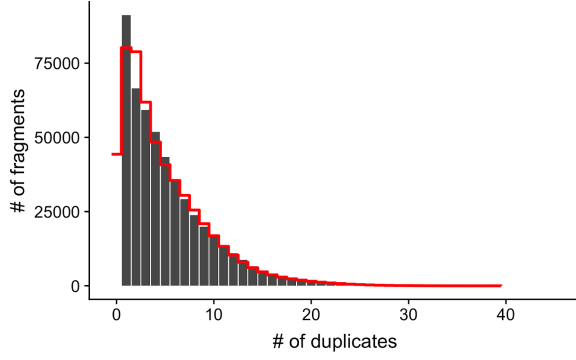
Estimate of γ_{seq} is 39.7%

Figure S5: Two examples of sequencing depth estimation

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:

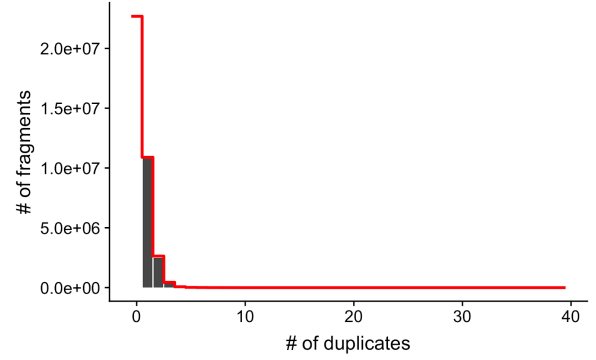
$$p_{mix}(k | \lambda, \rho_1, \dots, \rho_M) = \sum_{m=1}^M \rho_m * p_{poisson}(k | m * \lambda)$$

Where $\rho_1, \dots, \rho_M$ 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 (Figure S6). We can gain better sense of these by examining the individual components that make up the mix (Figure S6 lower row). In H012.1 the mixture is made up of several components while in C001.2752 the mixture is dominated by the first component.



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%



C001.2752 - here the mixture fit is close to the Poisson fit above.

Estimate of γ_{seq} is 38.3%

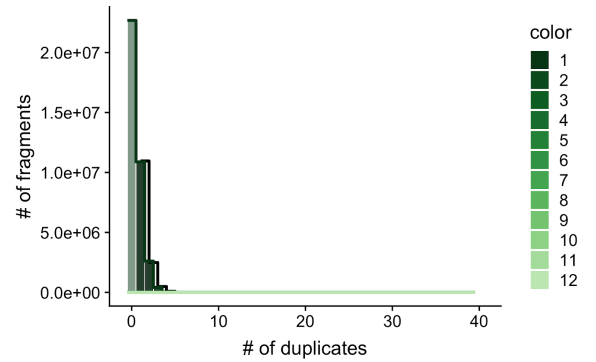
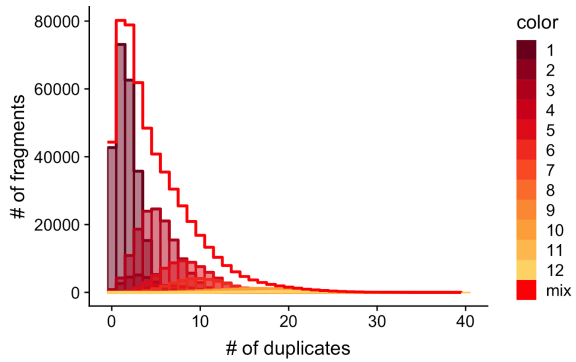


Figure S6: Two examples of sequencing depth estimation with mixture model. Below are the mixture components of the estimate.

Our conclusion (based on these two examples and more) is that for shallow sequenced samples (e.g., similar to C001.2752), 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:

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

Sensitivity Analysis

Suppose we are given:

- N - number of genomes in the cfDNA pool,
- k - number of informative nucleosomes per genome,
- p - percent of target cells in the cfDNA pool,
- γ_{mod} - specific capture probability, and
- γ_{bg} - non-specific capture probability. ($\gamma_{bg} = \gamma_{mod}/SNR$).

Example: Suppose we defined a signature of 200 nucleosomes that are marked with H3K4me3 only in cardiomyocytes (real numbers in 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 $\sim 1,000$ genomes/ml, we expect $\sim 2,000$ informative nucleosomes in a 1ml plasma sample. Thus, in this example: $N = 1,000$, $k = 200$, $p = 0.01$

The number of captured molecules among the k nucleosome positions is the sum of specific capture and non specific capture. The specific capture is a random variable

$$X_{mod} \sim B(pkN, \gamma_{mod})$$

That is, binomial distribution of capturing each of the kN nucleosomes with probability γ_{mod} . Similarly the non-specific capture is also binomially distributed.

$$X_{bg} \sim B(kN, \gamma_{bg})$$

Assuming that this processes are independent we capture

$$X = X_{mod} + X_{bg}$$

molecules. Note that in both equations we use kN as the number of potential molecules. Which means that we can trade off between the number of genomes (e.g., amount of plasma) and the number of informative locations.

Example (continued): Assume $\gamma_{mod} = 0.01$ and $\gamma_{bg} = 0.0001$. This means that

$X_{mod} \sim B(2,000,0.01)$ and $X_{bg} \sim B(200,000,0.0001)$. The resulting distributions for X_{bg}, X_{mod}, X are shown in Figure S7.

We now can ask if we observe x molecules, would we reject the null hypothesis that we saw only non-specific capture. Formally we can define the event that contains all such values of x :

$$D = \{ x : P_{(kN, \gamma_{bg})}(X_{bg} \geq x) < 0.05 \}$$

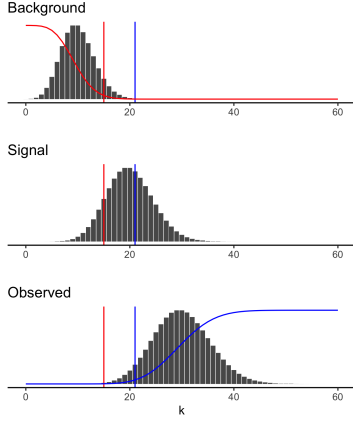
We say that the event is detectable with the parameters $(N, k, p, \gamma_{mod}, \gamma_{bg})$ if

$$P_{(kN, p, \gamma_{mod}, \gamma_{bg})}(X \in D) > 0.95$$

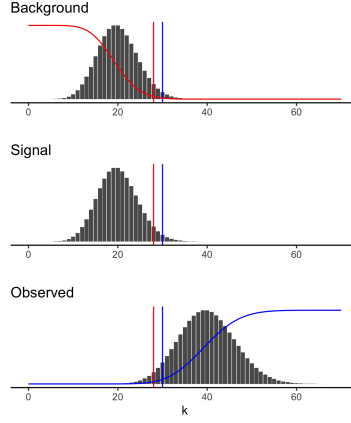
That is, it is likely that we would capture number of molecules x that allows us to reject the null hypothesis.

Example (continued): As shown in Figure S7, for $N = 1,000$, $k = 200$, $p = 0.01$, $\gamma_{mod} = 0.01$, and $\gamma_{bg} = 0.0001$ the probability of observing a value that would reject the null hypothesis is larger than 0.95. However, if we double the background rate to $\gamma_{bg} = 0.0002$ or half the size of the signature to $k = 100$, then this is no longer the case. Changing the parameters in the opposite direction increases the probability of detection.

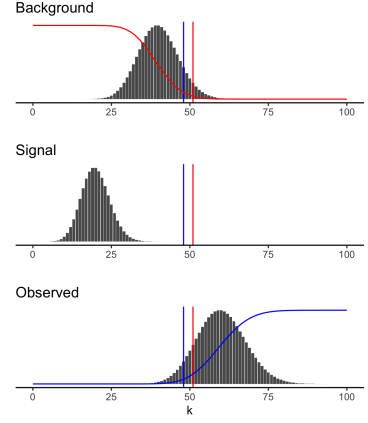
$$k = 200, \gamma_{bg} = 0.00005$$



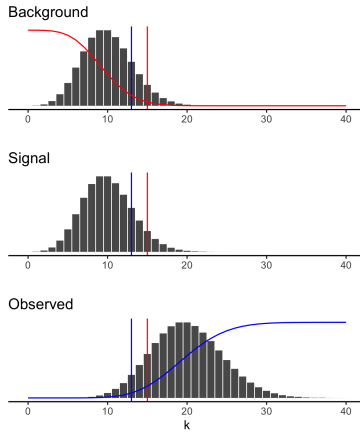
$$k = 200, \gamma_{bg} = 0.0001$$



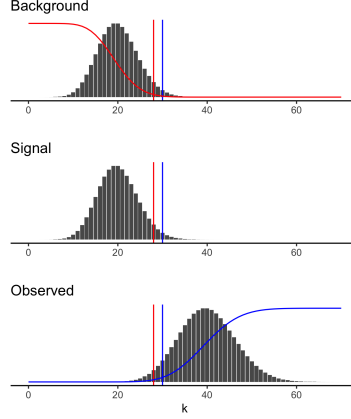
$$k = 200, \gamma_{bg} = 0.0002$$



$$k = 100, \gamma_{bg} = 0.0001$$



$$k = 200, \gamma_{bg} = 0.0001$$



$$k = 300, \gamma_{bg} = 0.0001$$

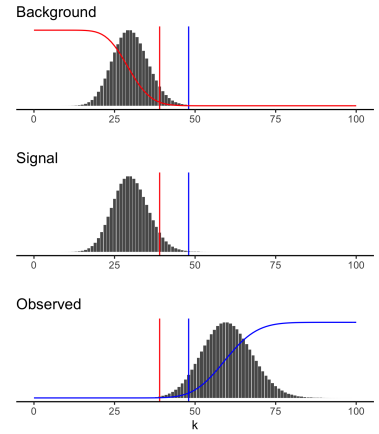


Figure S7: Examples of distribution of background and signal levels for different capture values with $N = 1,000$, $p = 0.01$, and $\gamma_{mod} = 0.01$. For each choice of parameters, we plot the distribution of background signal, of specific signal, and of the observed count. The red graph in the Background histogram shows the reverse cumulative distribution function of the background and the blue graph in the Observed histogram shows the cumulative distribution function of the observed values. The vertical red line denotes the threshold value for rejecting the null hypothesis of background distribution (reverse cumulative distribution function is equal to 0.05). The vertical blue line denotes the 5% of the observed signal (cumulative distribution is equal to 0.05). If the blue line is larger (to the right) of the red line, then we are in a detectable regime. That is, with high probability we would observe a value that would reject the null hypothesis.

Defining tissue-specific signatures

Using the Roadmap Epigenomics metadata table we defined sets of Roadmap samples that belong to a tissue or group of tissues (see Supplementary Table 12). 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.

Recall that $N[w, s]$ is the normalized signal of window w in sample s .

We then defined for each group of Roadmap samples the set of specific windows, as windows w meeting 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 Supplementary Table 5).

Statistical tests

Recall that

- $B[w, s]$ - background estimate for window w in sample s
- $C[w, s]$ - actual read counts observed in window w in sample s
- $\nu[s]$ - normalization factor for sample s

Comparison to background test. To test whether a gene or a signature is present above background in a sample, we use 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 associated with a gene or a tissue-specific signature as above.

Estimating mean and variance. To perform a test of signal vs behavior in a reference set of samples (e.g., samples from the healthy cohort), we need to estimate the mean and variance of the signal in the reference set. Ideally, we would want an estimate in units of normalized counts. This estimate is confounded by the sampling noise in each sample which can be a major source of variability, especially for windows with low signal. We thus built an estimation procedure that takes this variability into account.

Given a set of windows W of interest, let

- $s_1, \dots, s_n$ be the set of reference samples
- $c_i \leftarrow \sum_{w \in W} C[w, s_i]$ be the observed count numbers of W in sample s_i
- $\lambda_i \leftarrow \sum_{w \in W} B[w, s_i]$ be the estimated background rates of W in sample s_i
- $\xi_i = (\nu[s_i])^{-1}$ be the inverse of global normalization factor for sample s_i

To estimate the distribution of values, we assume that for each sample

$$C_i \sim \text{Poisson}(\xi_i N_i + \lambda_i)$$

Where $N_1, \dots, N_n \sim P(N)$ are i.i.d. samples of normalized counts from the unknown distribution we want to estimate.

To understand the implications of this model, using linearity of expectations we get

- $E[C_i] = \xi_i E[N_i] + \lambda_i$

This matches how we estimate normalized counts (recall that we defined normalized counts as $N_i \leftarrow (C_i - B_i)/\xi_i$ above).

Using the law of total variation we get that

- $\text{Var}[C_i] = \xi_i^2 \text{Var}[N_i] + \xi_i E[N_i] + \lambda_i$

which implies that the observed counts variance is due both to the variance of N_i and additional variance due to the observation process.

For the estimate we assume that N_i is distributed according to a Gamma distribution with mean μ_N and variance σ_N^2 . One often treats the Gamma distribution as the counterpart of normal distributions for non-negative variables. Indeed if $\mu_N \gg \sigma_N$ then the Gamma distribution is

almost a normal distribution.

With this assumption, we can use classical results to show that

$$C_i \sim NB(\mu = \xi_i \mu_N, \sigma^2 = \xi_i^2 \sigma_N^2 + \xi_i \mu_N) \quad \text{when } \lambda_i = 0$$

Where $NB(\mu, \sigma^2)$ is the Negative Binomial distribution with mean μ and variance σ^2 . That is,

$$P(Z = k) = \Gamma(k + \frac{\mu^2}{\sigma^2 - \mu}) \Gamma(k + 1)^{-1} \Gamma(\frac{\mu^2}{\sigma^2 - \mu})^{-1} (\frac{\sigma^2 - \mu}{\sigma^2})^k (\frac{\mu}{\sigma^2})^{\mu^2 / (\sigma^2 - \mu)} \quad \text{if } Z \sim NB(\mu, \sigma^2)$$

When $\lambda_i > 0$ this is not the precise distribution of C_i . However, if λ_i is small we can assume that the Negative Binomial is a close approximation to the distribution:

$$C_i \sim NB(\mu = \xi_i \mu_N, \sigma^2 = \xi_i^2 \sigma_N^2 + \xi_i \mu_N + \lambda_i)$$

With this assumption we define the likelihood of the parameters we want to estimate

$$L(\mu_N, \sigma_N^2) = \prod_i P_{NB}(c_i; \mu = \xi_i \mu_N, \sigma^2 = \xi_i^2 \sigma_N^2 + \xi_i \mu_N + \lambda_i)$$

We then use an optimization procedure to search for the parameters that maximize the likelihood. (R optim() procedure).

Comparison of a single sample to a reference. Here we test whether the observed signal of a set of windows W in sample s is higher or lower than expected in a reference set of samples (e.g., healthy samples).

We used the following procedure to perform a two-tailed test against the null assumption specified by the given mean and variance.

computePValueAgainstReference(W, s, μ, σ^2)

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

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

$$\mu_X \leftarrow \mu / \nu[s] + \lambda$$

$$\sigma_X^2 \leftarrow \sigma^2 / \nu[s]^2 + \mu_X$$

if($x > \mu_X$) Return $\frac{1}{2} P_{NB(\mu_X, \sigma_X^2)}(X \geq x)$ // NB p-value

if($x < \mu_X$) Return $\frac{1}{2} P_{NB(\mu_X, \sigma_X^2)}(X \leq x)$ // NB p-value

As in the test above the set of windows W can be the set associated with a gene or a signature or a group of genes. The difference is that we need to estimate the mean and variance of signal over W in the reference cohort.

Comparison of two groups of samples. Given two groups of samples, A_1 and A_2 reject the null hypothesis that the signal over a set of windows W is from the same underlying distribution.

We use a *likelihood ratio test* (LRT) where we compare the likelihood under the scenario that all samples in both sets have the same distribution to the alternative where we have a different distribution in each set. The statistics is

$$\chi = 2 (\log \hat{L}(W, A_1 \cup A_2) - (\log \hat{L}(W, A_1) + \log \hat{L}(W, A_2)))$$

with

$$\hat{L}(W, A) = \max_{\mu, \sigma^2} L(\mu_N, \sigma_N^2 : W, A)$$

where $L(\mu_N, \sigma_N^2 : W, A)$ is the likelihood function defined above with respect to the event defined by W on samples in A .

Under the null assumption the distribution of the statistic χ is approximately a Chi-squared distribution with 2 degrees of freedom, which allows to compute the p-value.

Cancer programs

To define cancer signatures from a set of gene programs, we perform the following procedure:

1. For each program j in the set of gene programs, let W_j be the set of windows for the set of genes in the program. Estimate the mean and variance μ_j, σ_j^2 from the healthy cohort. Then define the p-value of the profile on each sample as

$$p[j, s] = \text{computePValueAgainstReference}(W_j, s, \mu_j, \sigma_j^2)$$

After computing p-values of all programs against all relevant cancer samples, we define $q[j, s]$ as the q-value of program j in sample s after FDR correction.

2. We define the matrix

$$A[j, s] = 1\{q[j, s] < 10^{-6}\}$$

where $1\{\text{condition}\}$ is the indicator function that is equal to 1 when the condition holds and 0 otherwise. We then filter matrix rows (programs) to those that are informative

about differences among the samples:

$$A' \leftarrow A \left[\left\{ j : 2 < \sum_s A[j, s] < \frac{2}{3}M \right\}, \right] \text{ where } M \text{ is the number of samples.}$$

3. We cluster the rows of A' , and identify for each cluster $k \in \{1, \dots, K\}$ the consensus over samples as the set of samples where a majority of the programs in the cluster are significant:

$$A_k = \left\{ s : \sum_{j \in \text{Cluster } k} A'[j, s] > \frac{1}{2} |\text{Cluster } k| \right\}$$

4. For each cluster k find the set G_k of genes that are differentially marked between A_k and the remaining samples (see above). Use FDR correction to adjust p-values and select genes with $q < 0.001$
5. Repeat steps 1-4 with gene sets $G_1, \dots, G_K$
6. Return the (revised) gene sets $G_1, \dots, G_K$ as signatures.

References

1. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
2. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
3. Shema, E. *et al.* Single-molecule decoding of combinatorially modified nucleosomes. *Science* **352**, 717–721 (2016).
4. Zheng, Y., Huang, X. & Kelleher, N. L. Epiroteomics: quantitative analysis of histone marks and codes by mass spectrometry. *Curr. Opin. Chem. Biol.* **33**, 142–150 (2016).
5. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nat. Methods* **10**, 325–327 (2013).