Re-analysis of a CD4 ChIP-Seq data set with csaw

Ryan C. Thompson
Salomon Lab
The Scripps Research Institute

May 6, 2016

Outline

- Intro to T-cells and experimental design
- ChIP-Seq overview
- Consensus peak-calling wih IDR
- Previous promoter-oriented analysis (published soon)
- Initial QC and analysis of whole genome analysis
- Genomic region blacklists

CD4 T-cell activation and memory formation

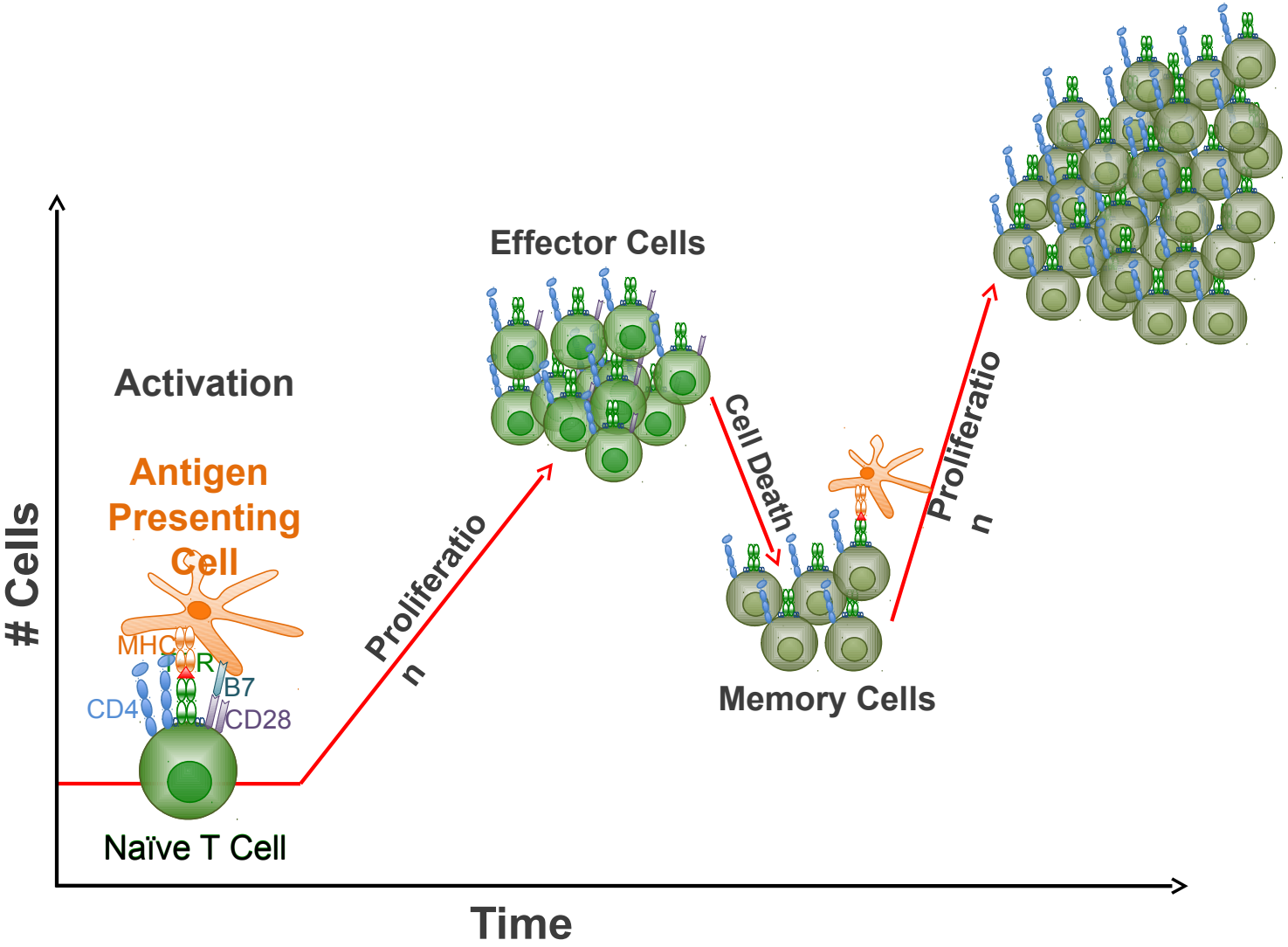

Figure 1: CD4 T-cell response to successive infections

Experimental design

- Isolate and culture naïve & memory cells from 4 donors
- Activate cells and take samples at pre-activation (day 0) and at days 1, 5, and 14 post-activation
- RNA-seq and ChIP-seq on all samples
- ChIP using antibodies against H3K4Me2, H3K4Me3, HeK27me3 (and input)

Experimental design

- Isolate and culture naïve & memory cells from 4 donors
- Activate cells and take samples at pre-activation (day 0) and at days 1, 5, and 14 post-activation
- RNA-seq and ChIP-seq on all samples
- ChIP using antibodies against H3K4Me2, H3K4Me3, HeK27me3 (and input)

- **Data analysis?**
- Profit

How ChIP-Seq works, more or less

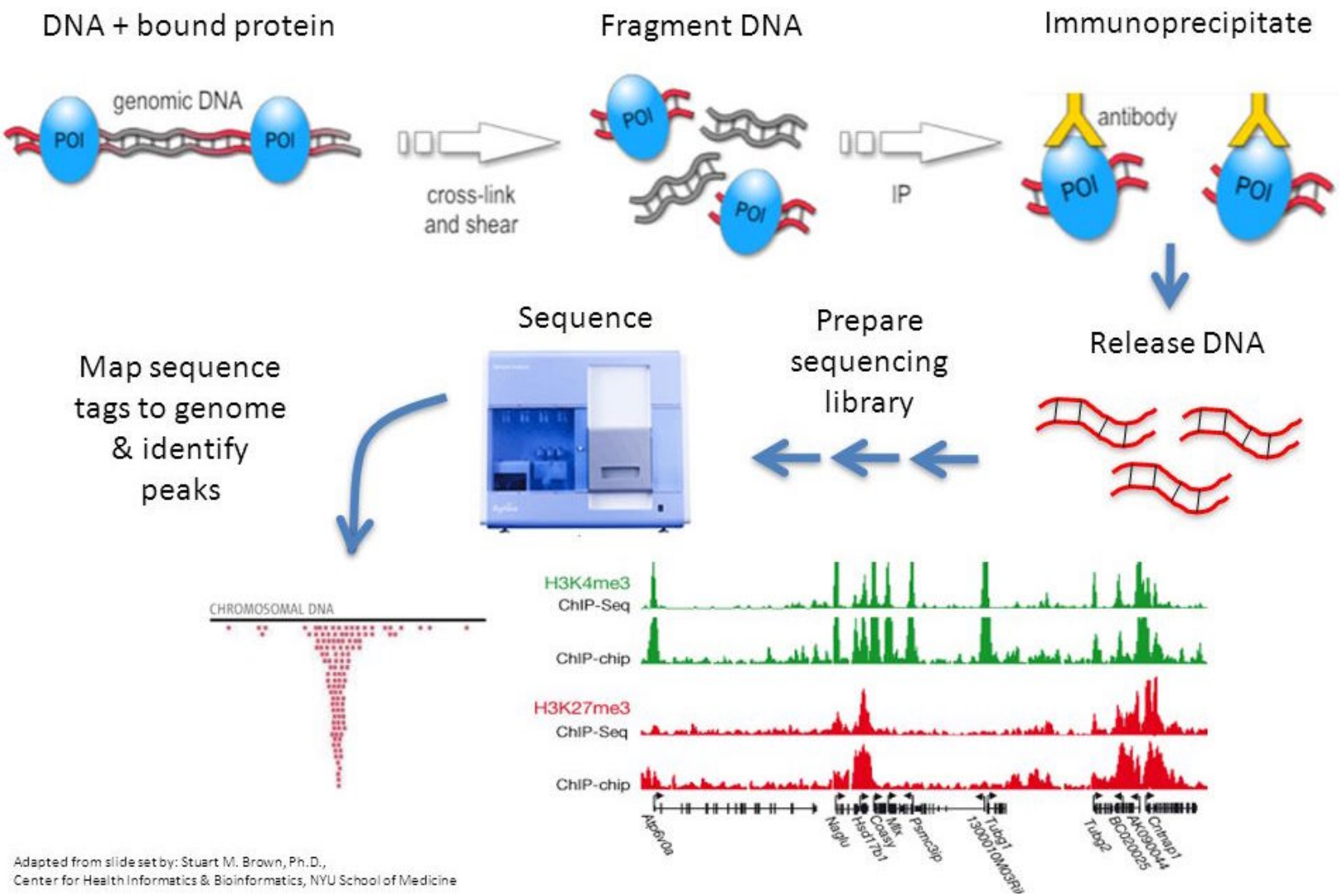

Figure 2: Overview of ChIP-Seq workflow

Promoter-oriented analysis

First steps

- Map with bowtie2
- Call peaks with MACS
- Determine consensus *biologically* reproduced peaks using Irreproducible Discovery Rate (like FDR but comparing consistency between two lists)

First steps

- Map with bowtie2
- Call peaks with MACS
- Determine consensus *biologically* reproduced peaks using Irreproducible Discovery Rate (like FDR but comparing consistency between two lists)

- Process RNA-seq as usual: tophat $\rightarrow$ htseq-count $\rightarrow$ edgeR

IDR workflow

IDR helps choose a significance threshold at which peaks are reproducible between biological replicates.

- Call peaks in each individual sample
- Run IDR on each pairof samples to determine p-value $\rightarrow$ IDR mapping
- Since multiple samples give more information than any pair, take the smallest relationship as an upper bound for the overall IDR
- Combine all samples and call peaks again
- Filter combined-sample peak calls at the p-value corresponding to the chosen IDR threshold to obtain consensus peaks

IDR workflow

IDR helps choose a significance threshold at which peaks are reproducible between biological replicates.

- Call peaks in each individual sample
- Run IDR on each pairof samples to determine p-value $\rightarrow$ IDR mapping
- Since multiple samples give more information than any pair, take the smallest relationship as an upper bound for the overall IDR
- Combine all samples and call peaks again
- Filter combined-sample peak calls at the p-value corresponding to the chosen IDR threshold to obtain consensus peaks

- (Should do saturation analysis, but I didn't)

Promoter analysis

The original analysis focused on gene promoters, and comparing promoter histone behavior with gene expression. It also focused mainly on H3K4me2/3.

- Determine effective promoter radius by looking at distribution of nearest TSS-to-peak distances
- Define promoter regions as TSS $\pm$ radius
- Merge overlapping promoters for the same gene (for genes with multiple TSS)
- Count reads in promoter regions
- Perform "differential binding" analysis on promoter counts using edgeR, similarly to RNA-seq
- Look at RNA-seq DE vs promoter ChIP-Seq DB
- Look at RNA-seq vs peak presence/absence in promoter

Determining promoter radius

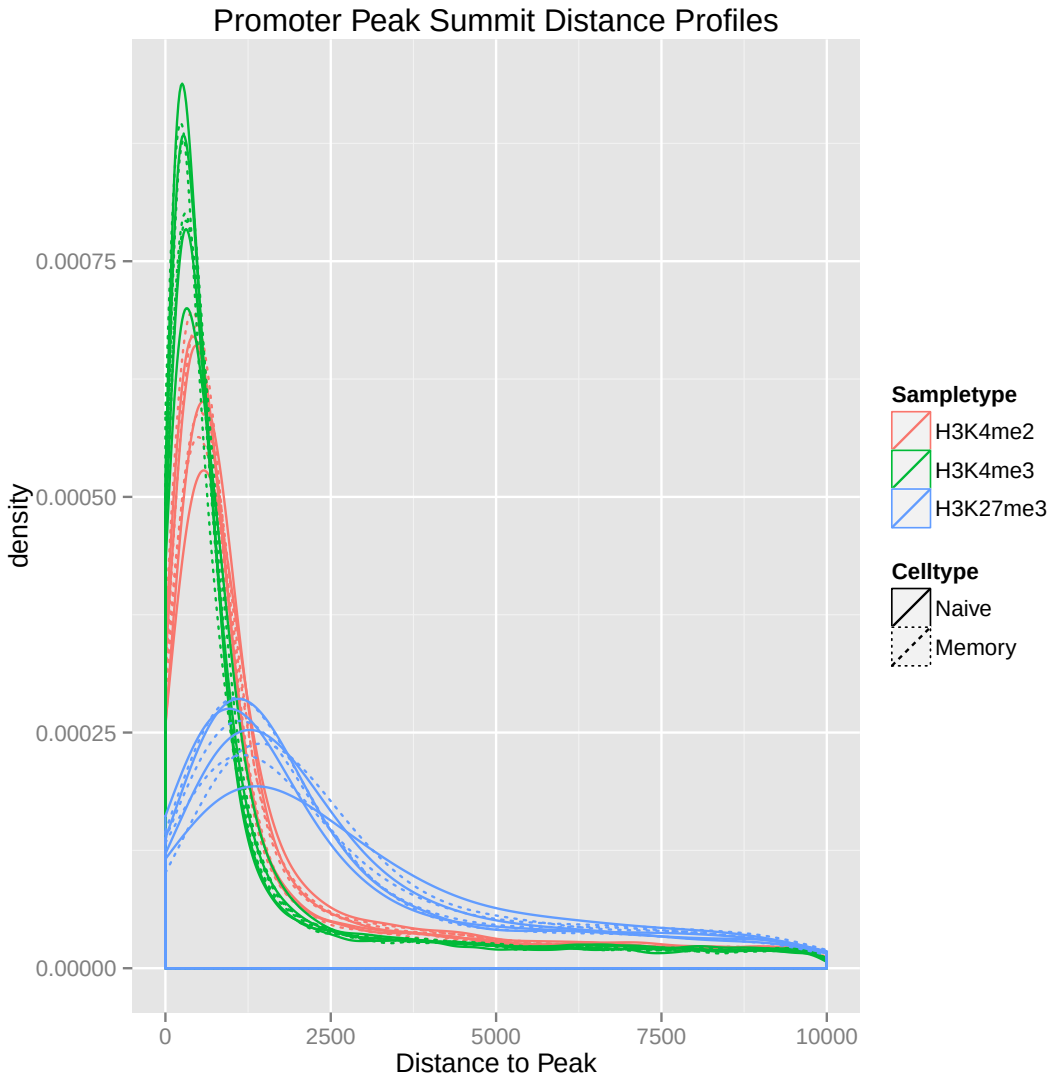

Figure 3: Distribution of distances from TSS to nearest peak summit

Static expression distribusion vs peak status

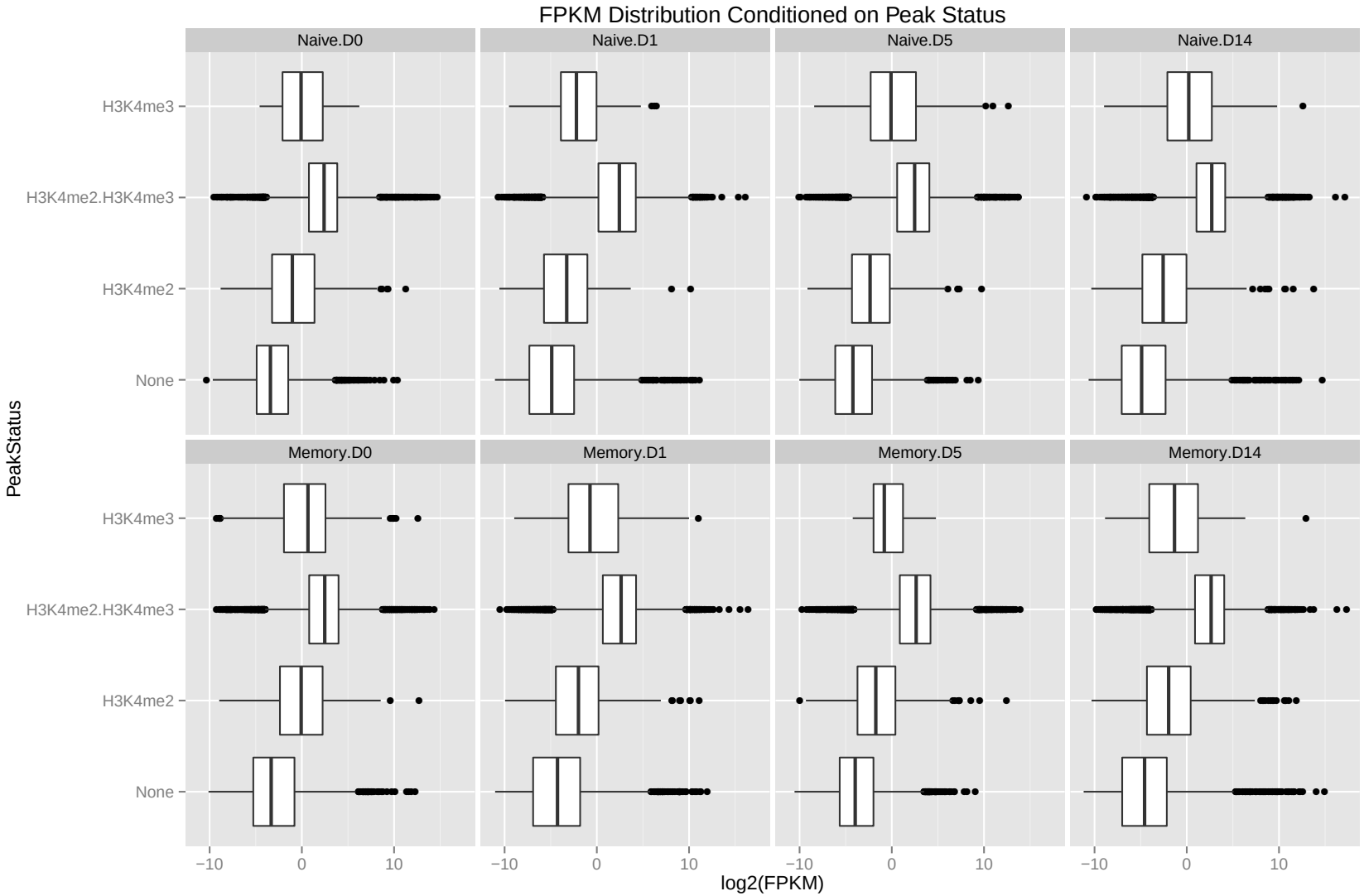

Figure 4: Gene expression distribution by promoter peak status

Static expression distribusion vs peak status

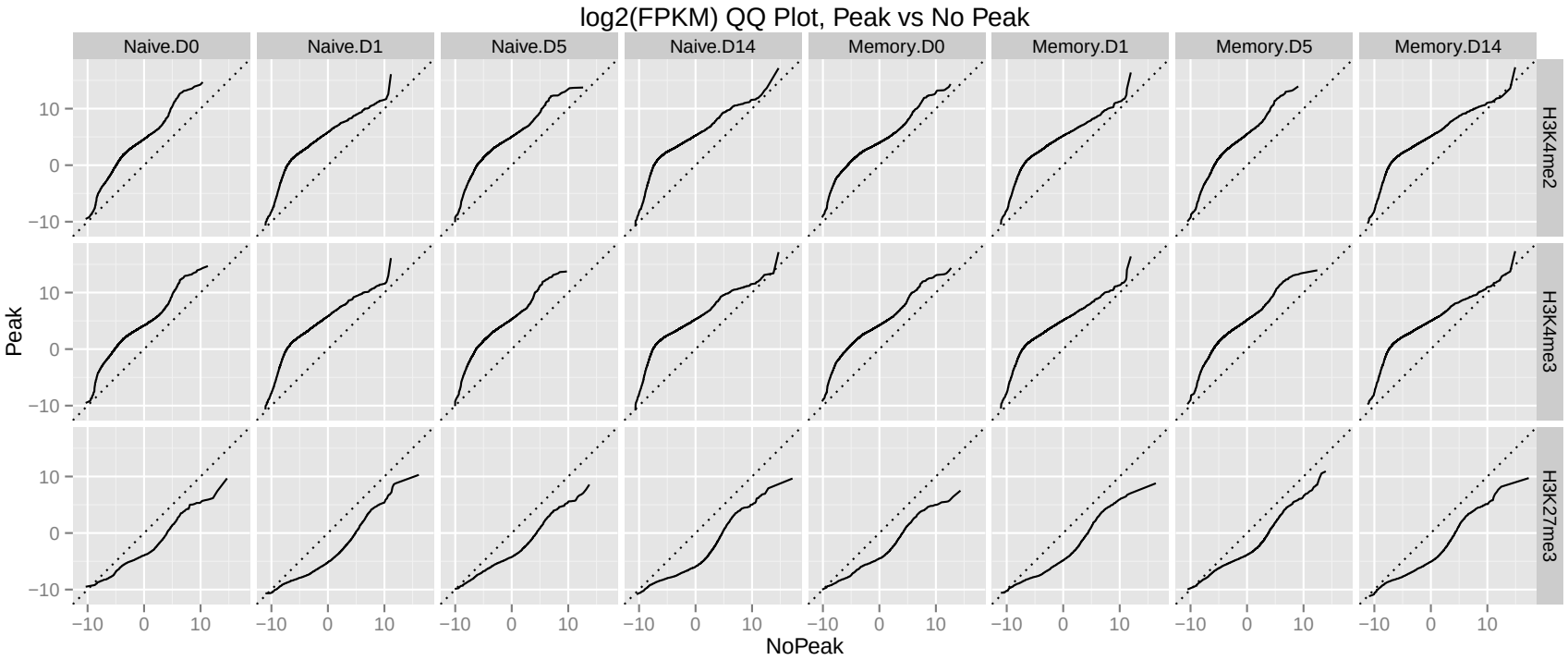

Figure 5: Gene expression QQ plots, peak vs no peak

Histone and RNA interaction dynamics

- The above plots only compared promoter peak status and expression at the same time points.

Histone and RNA interaction dynamics

- The above plots only compared promoter peak status and expression at the same time points.
- What if we look at, e.g. initial peak status vs expression change over time, or vice versa?

Initial peak status predicts 24h RNA change in naïve cells

Figure 3

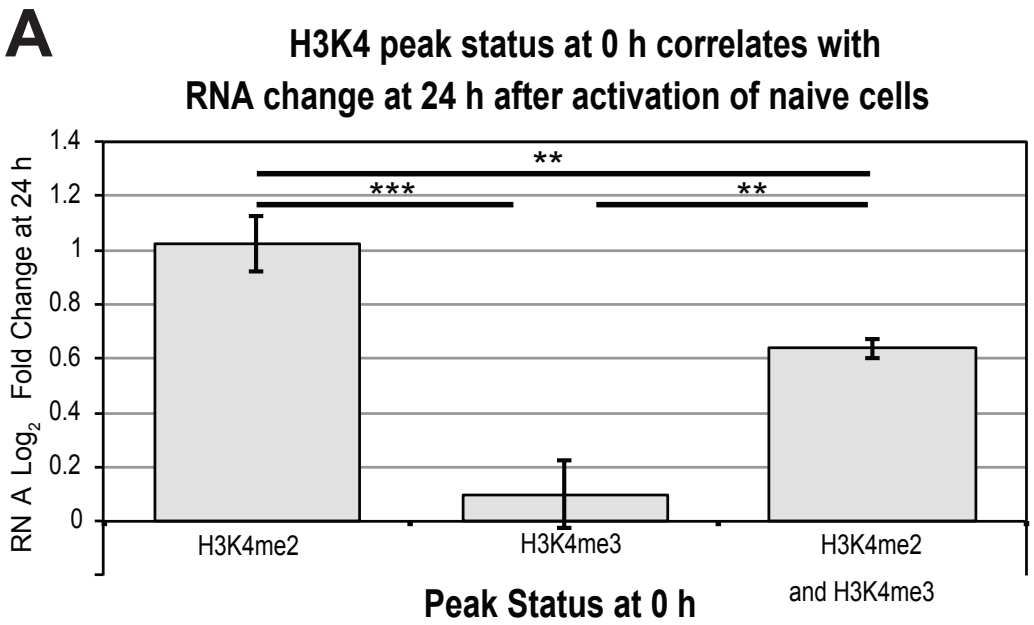

Figure 6: RNA 24h change vs peak status

Initial promoter HeK4me3/2 ratio predicts 24h RNA change in naïve cells

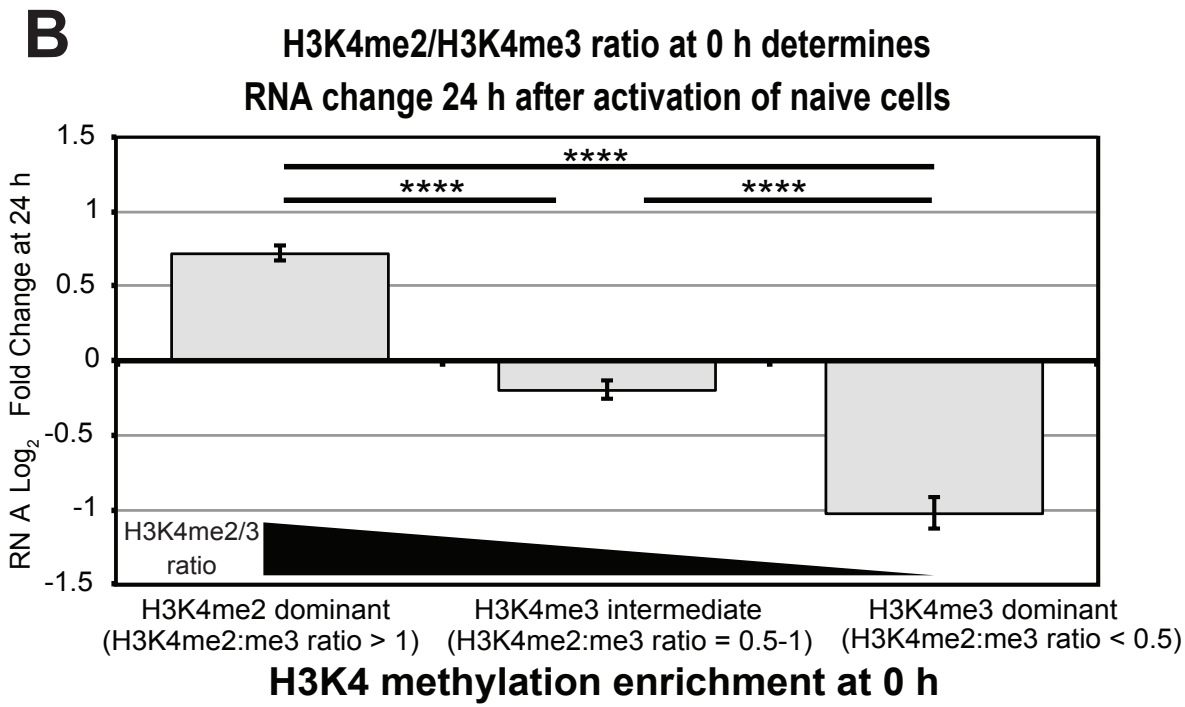

Figure 7: RNA 24h change vs peak status

Different relationship between peaks and expression in memory

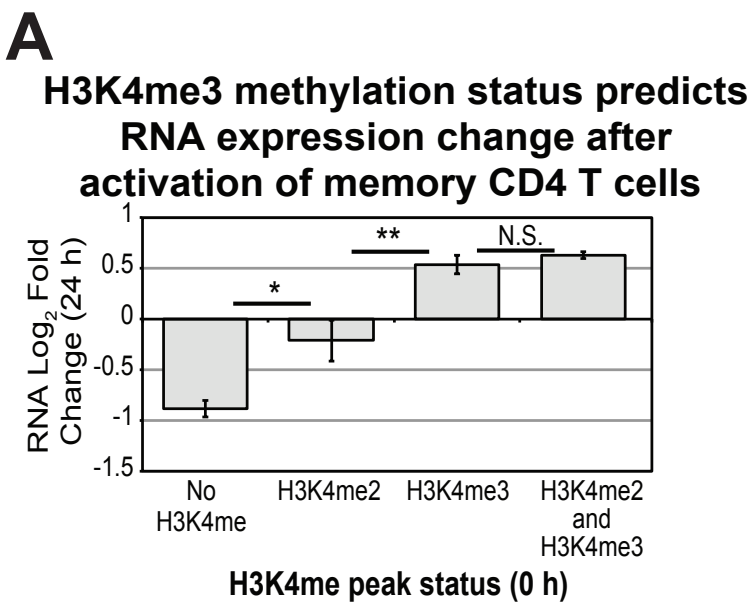

Figure 8: RNA 24h change vs peak status

Different relationship between peaks and expression in memory

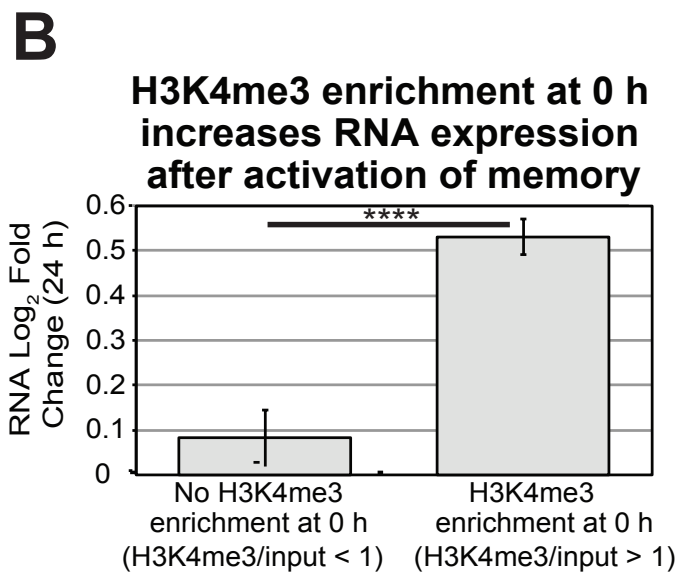

Figure 9: RNA 24h change vs peak status

Genome-wide analysis

Starting point

- Using existing bam files
- Using existing peak calls

Starting point

- Using existing bam files
- Using existing peak calls

- Analyze using the csaw Bioconductor package (csaw = "ChIP-Seq Analysis with Windows")
- QC Steps
  - Look at strand cross-correlation plots to determine fragment length
  - Look at coverage plots centred on local maxima to determine protein footprint size
  - Look at sample-vs-sample MA plots to determine proper normalization

Starting point

- Using existing bam files
- Using existing peak calls

- Analyze using the csaw Bioconductor package (csaw = "ChIP-Seq Analysis with Windows")
- QC Steps
  - Look at strand cross-correlation plots to determine fragment length
  - Look at coverage plots centred on local maxima to determine protein footprint size
  - Look at sample-vs-sample MA plots to determine proper normalization

- Analyze window counts with edgeR
- Combiond significance of adjacent windows to increase power

Determining fragment length from strand cross-correlation

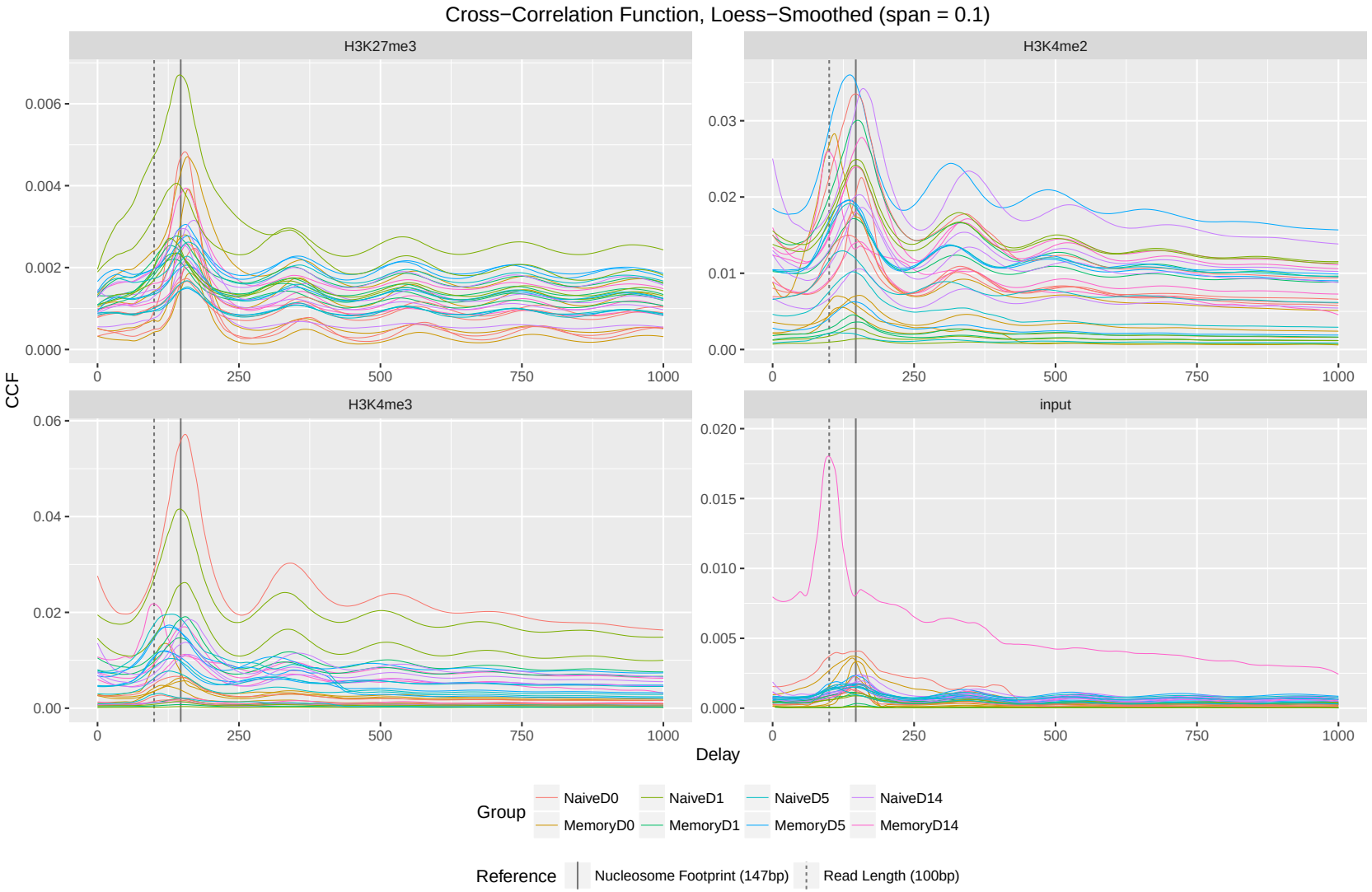

Figure 10: Strand Cross-correlation plots

Determining peak width by profiling coverage around local maxima

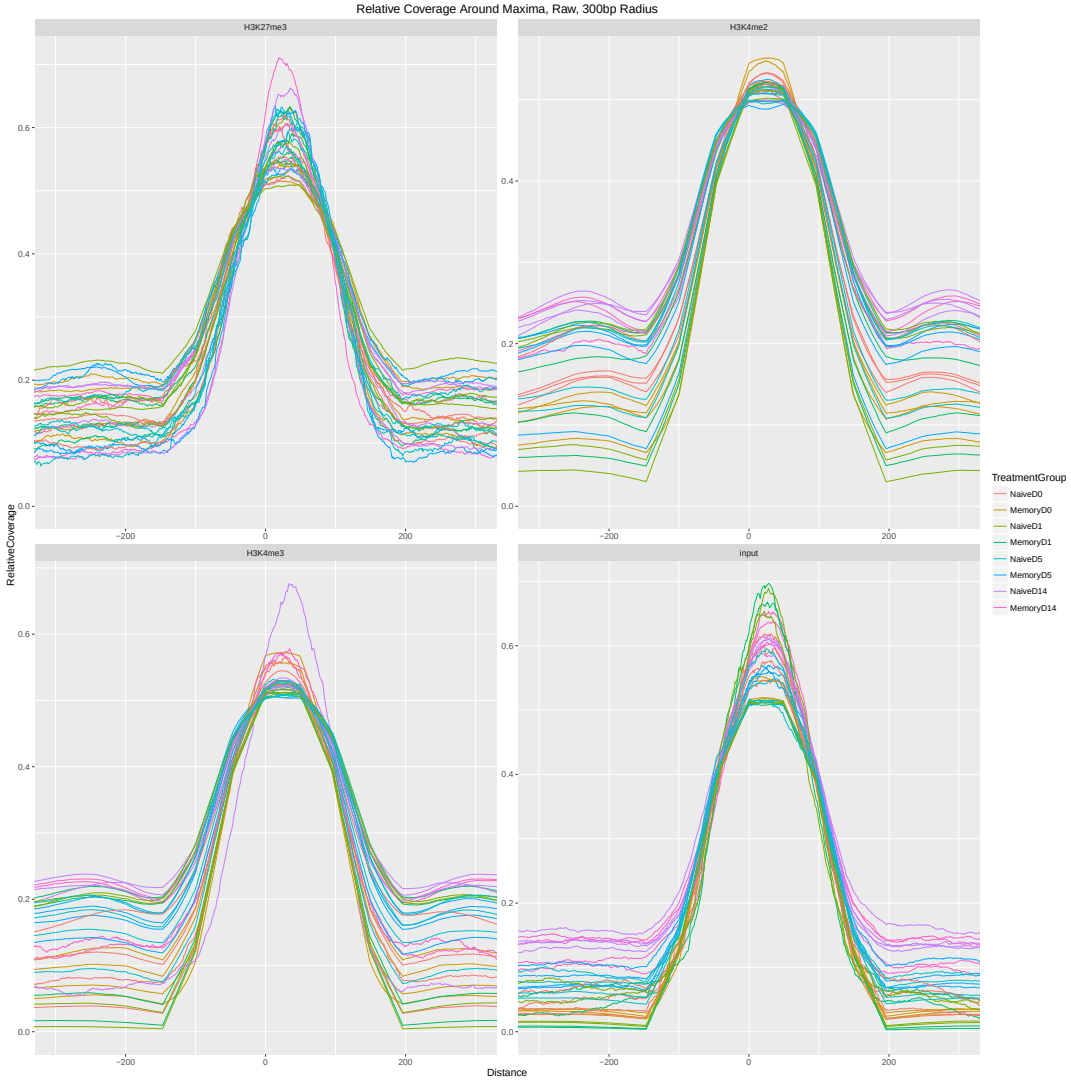

Figure 11: Local maxima coverage plots

Windows?

Figure 12: No, not that kind of windows

Window counting

- Tile windows of footprint size across the genome at desired resolution
- Extend each read's 3′ end to the fragment length
- Count fragments overlapping each window (non-unoque is OK)
- Also count the total reads assigned
- Separately repeat with large bins of 10kb, requiring each read to map to a single bin (10kb bins will be used for background normalization)

Normalizing window counts

- Multiple ways to normalize
- Composition (i.e. background) normalization: run TMM (standard edgeR norm. method) on 10kb bins
- Efficiency bias normalization: Select only high-abundance windows and run TMM
- Peak-overlap normalization: Select only windows that overlap called peaks and run TMM

Normalizing window counts

- Multiple ways to normalize
- Composition (i.e. background) normalization: run TMM (standard edgeR norm. method) on 10kb bins
- Efficiency bias normalization: Select only high-abundance windows and run TMM
- Peak-overlap normalization: Select only windows that overlap called peaks and run TMM

Do these give different results? How to choose which is correct?

High-abundance windows correlate with peaks

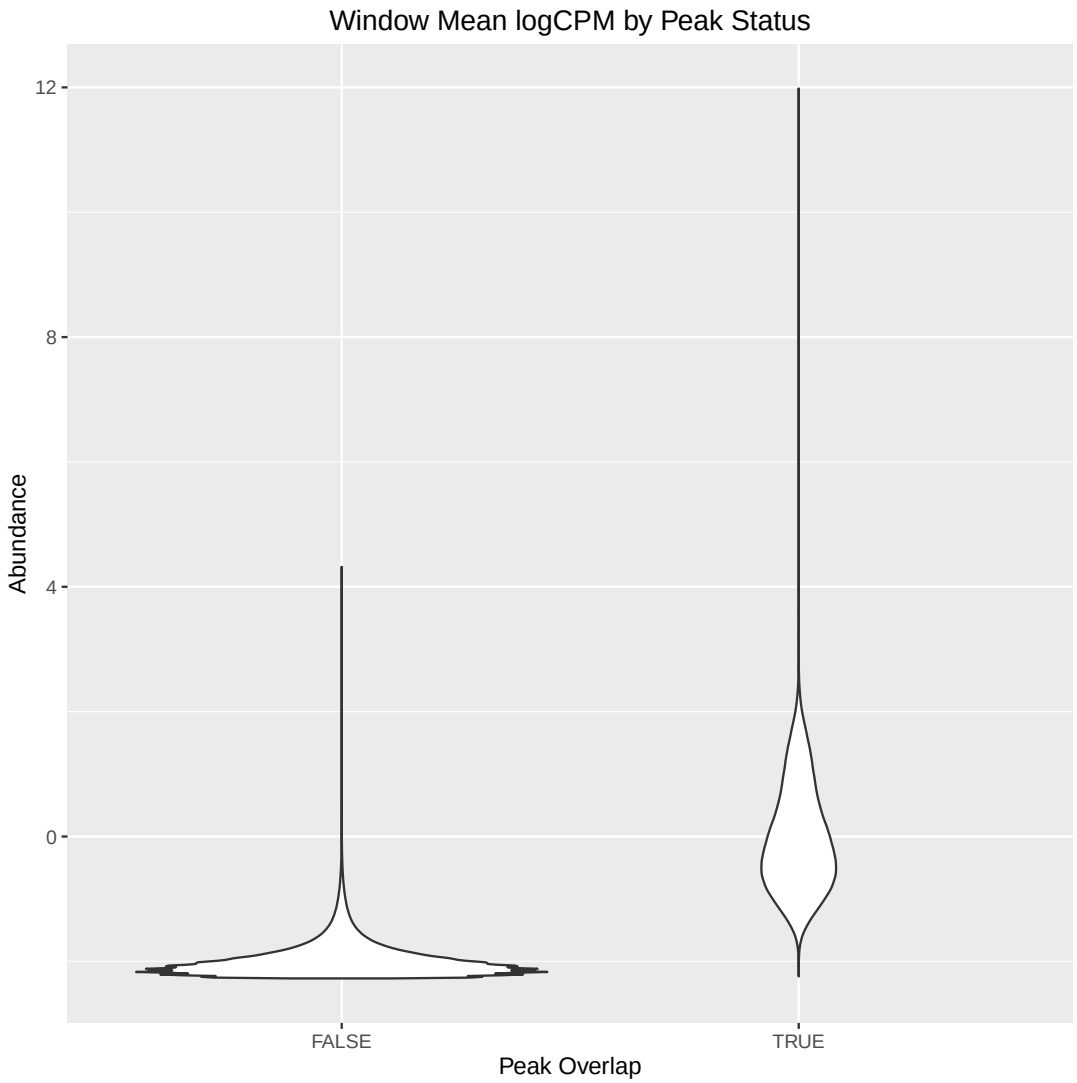

Figure 13: Window abundance vs peak overlap

ChIP-Seq MA plots are bimodal

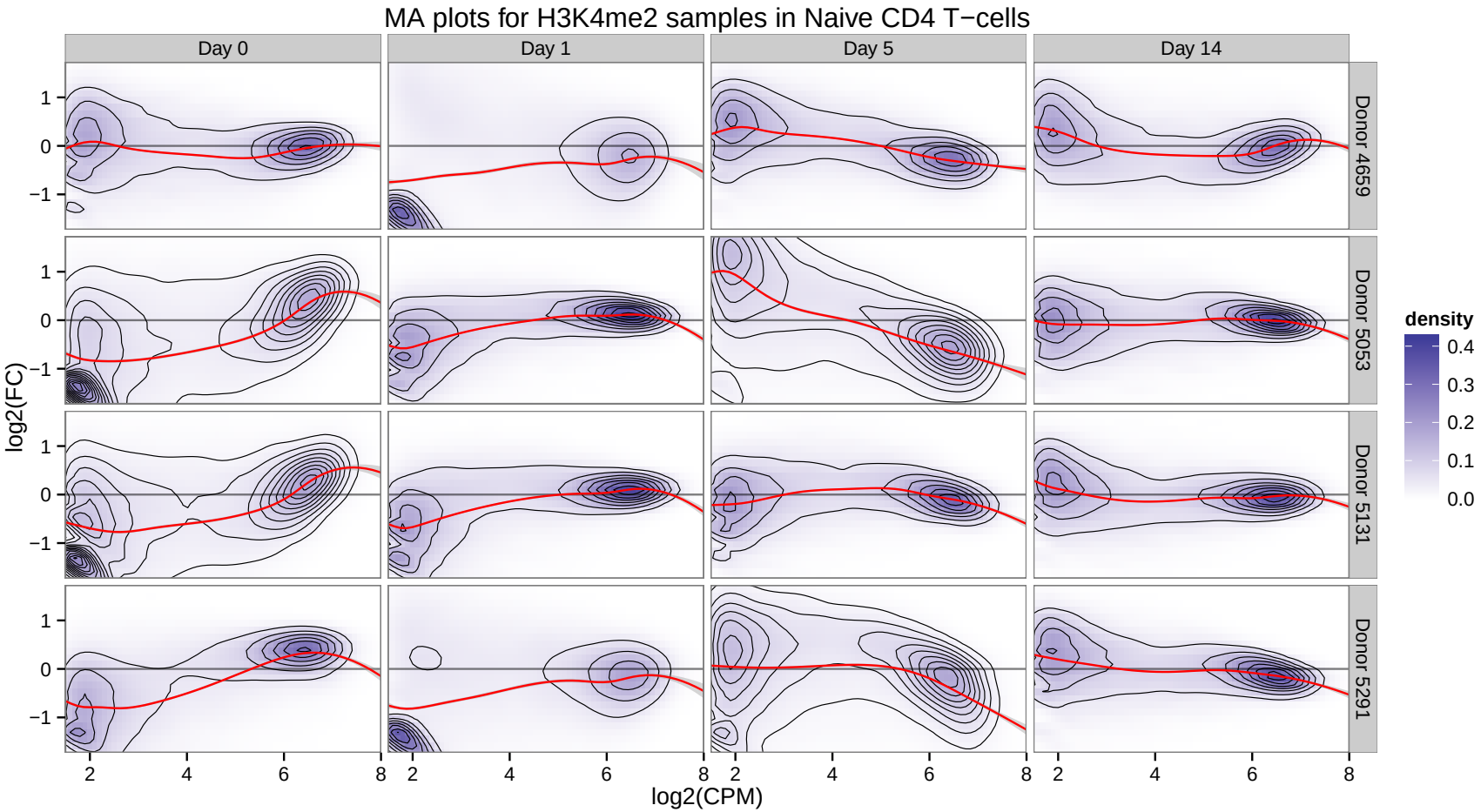

Figure 14: Promoter MA Plots

Normalization factors line up with one of the modes

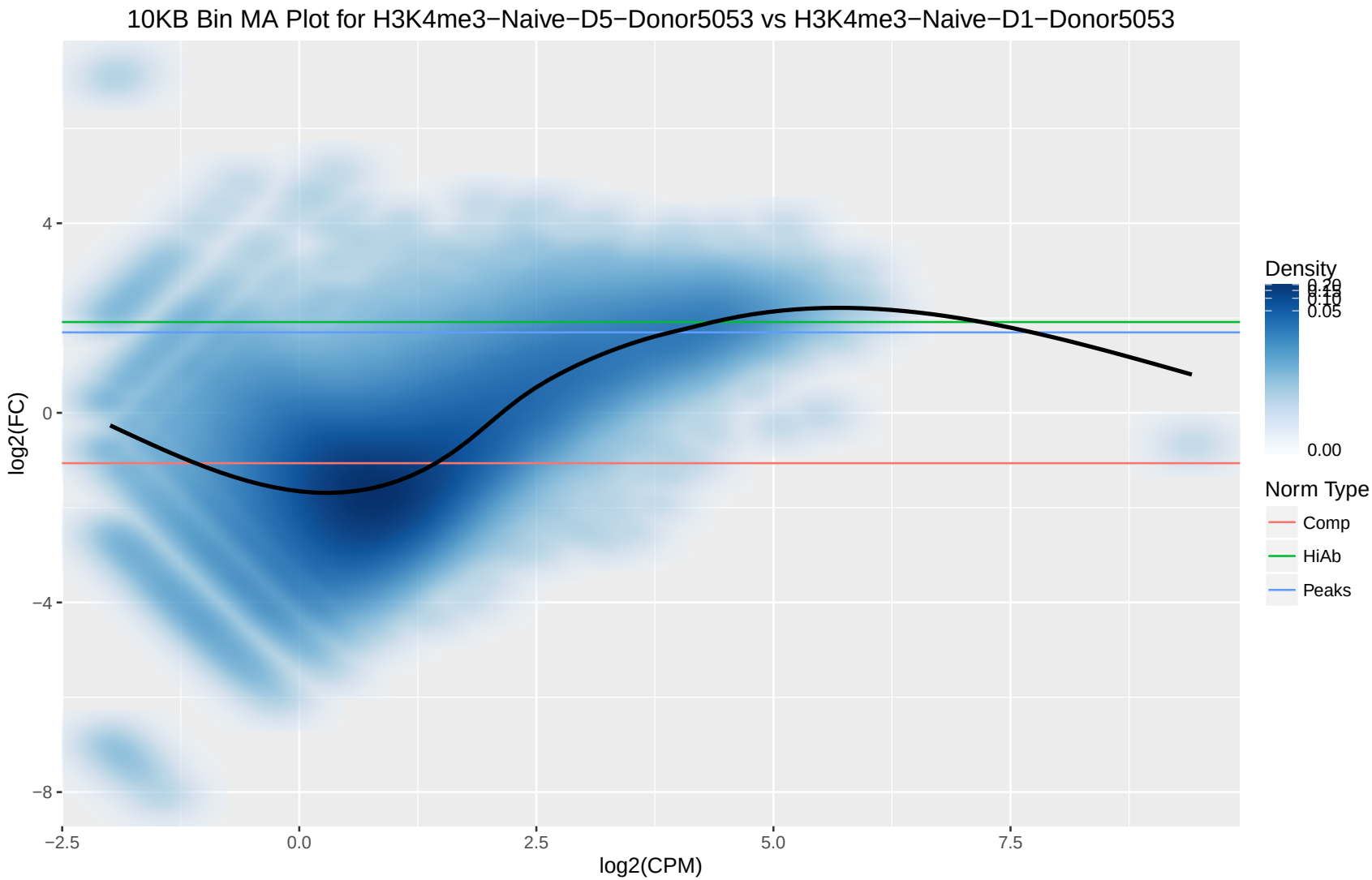

Figure 15: 10kb bin MA Plot

Choosing a normalization

- Genuine global changes in histone modification are possible
- Global changes can also be explained by variations in ChIP enrichment efficiency (efficinecy bias)
- Normalizing to the signal eliminates global changes; normalizing to the BG preserves them
- We need to rule out efficiency bias before we can use the BG normalization
- This can be done by making sure technical replicates have consistent efficiency

Choosing a normalization

- Genuine global changes in histone modification are possible
- Global changes can also be explained by variations in ChIP enrichment efficiency (efficinecy bias)
- Normalizing to the signal eliminates global changes; normalizing to the BG preserves them
- We need to rule out efficiency bias before we can use the BG normalization
- This can be done by making sure technical replicates have consistent efficiency

- Or we could just give up and use that loess curve

Blacklist regions

- Some areas of the genome have spurious extreme high coverage (over 100x the rest of the genome's coverage) even in input samples
- These regions tend to have higher proportions of multi-mapping reads
- Some of them are known repeats, but others have no obvious sequence features to explain the coverage as a mapping artifact
- Peak-calling and other analyses don't work in these regions
- Reads mapping to these regions must be excluded from downstream analyses

Blacklist regions

- Some areas of the genome have spurious extreme high coverage (over 100x the rest of the genome's coverage) even in input samples
- These regions tend to have higher proportions of multi-mapping reads
- Some of them are known repeats, but others have no obvious sequence features to explain the coverage as a mapping artifact
- Peak-calling and other analyses don't work in these regions
- Reads mapping to these regions must be excluded from downstream analyses

- What happens if you don't do this?

CCF Plot with blacklist exclusion

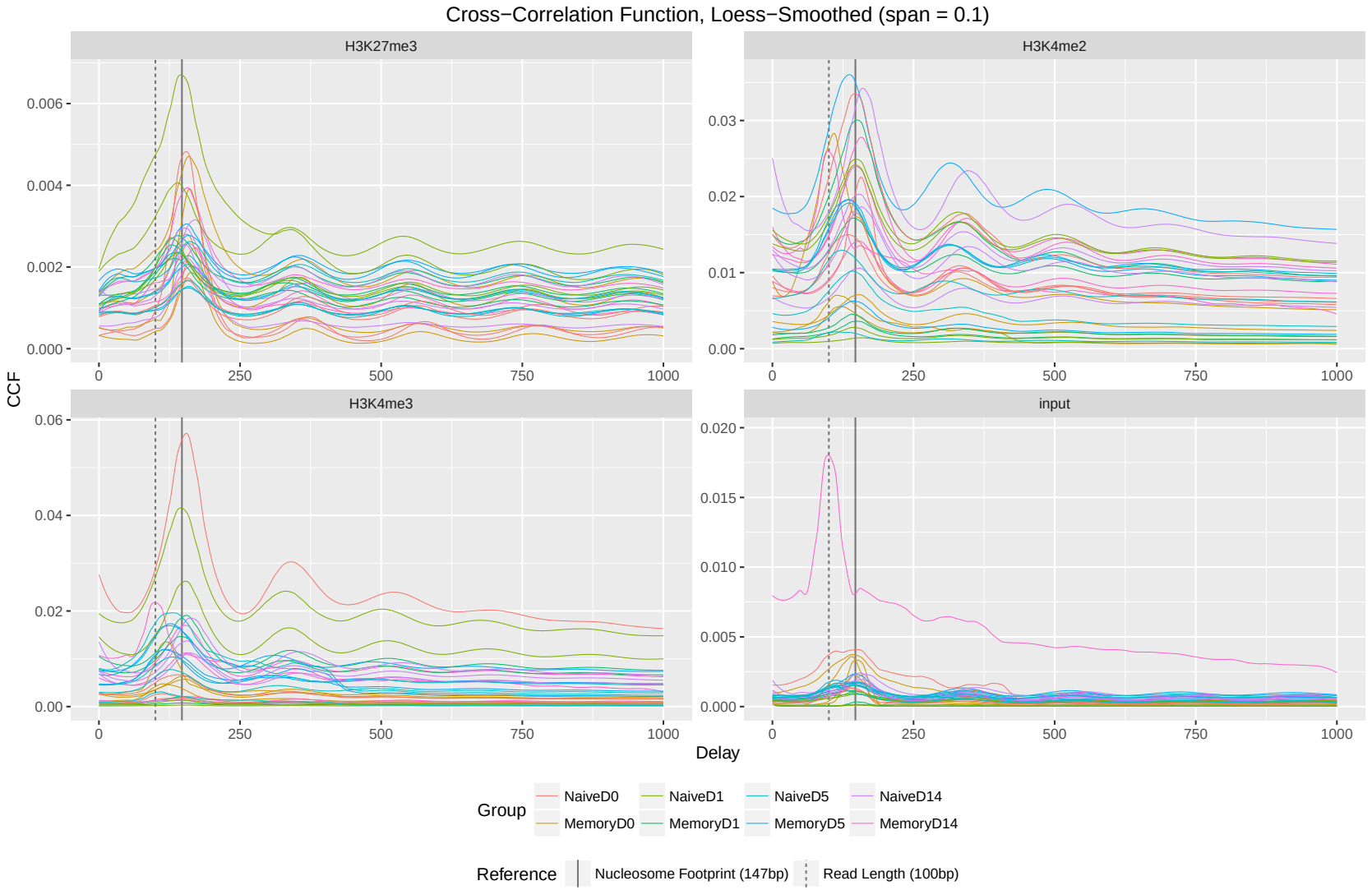

Figure 16: Strand Cross-correlation plots with blacklisting

CCF Plot WITHOUT blacklist exclusion

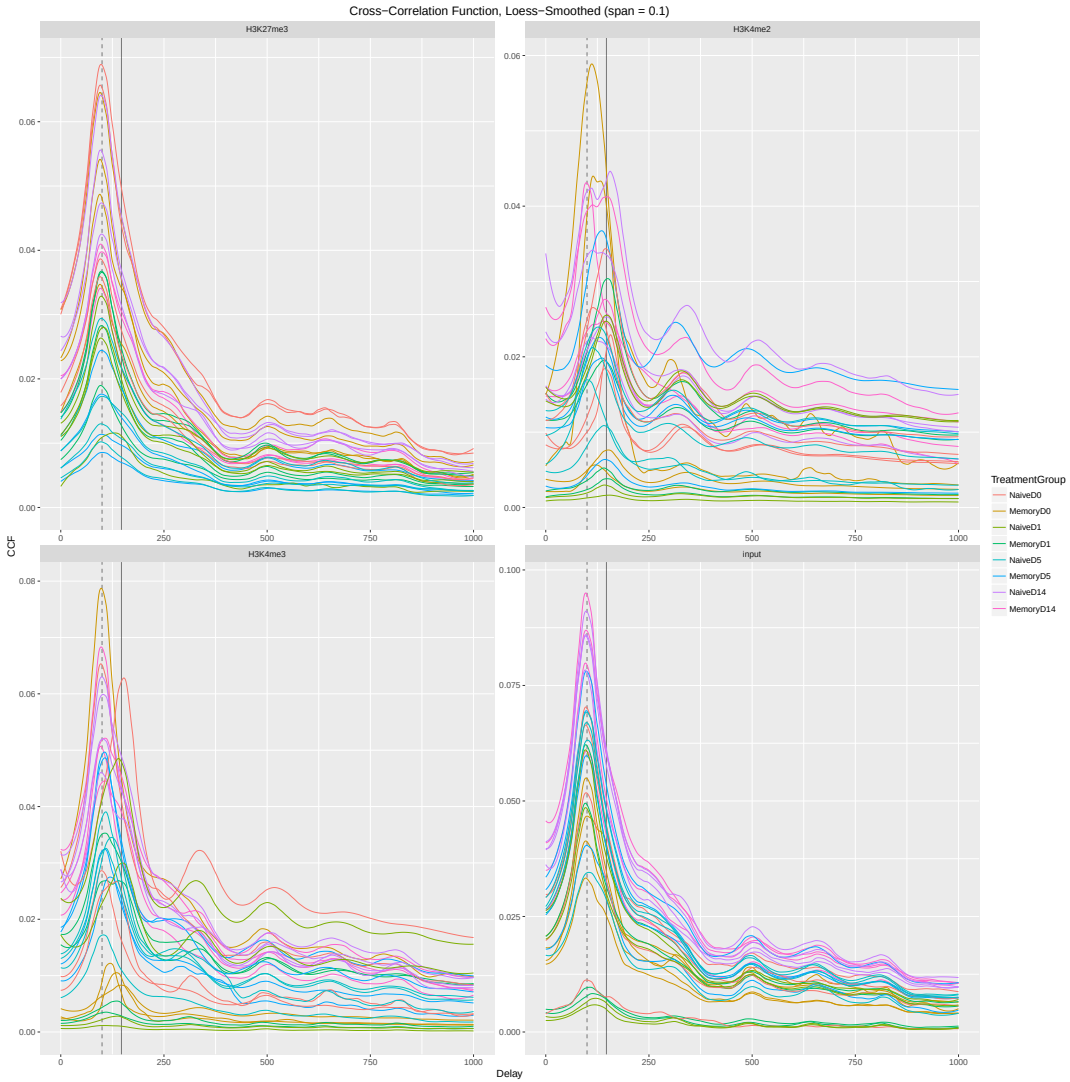

Figure 17: Strand Cross-correlation plots without blacklisting

MA plot with blacklist exclusion

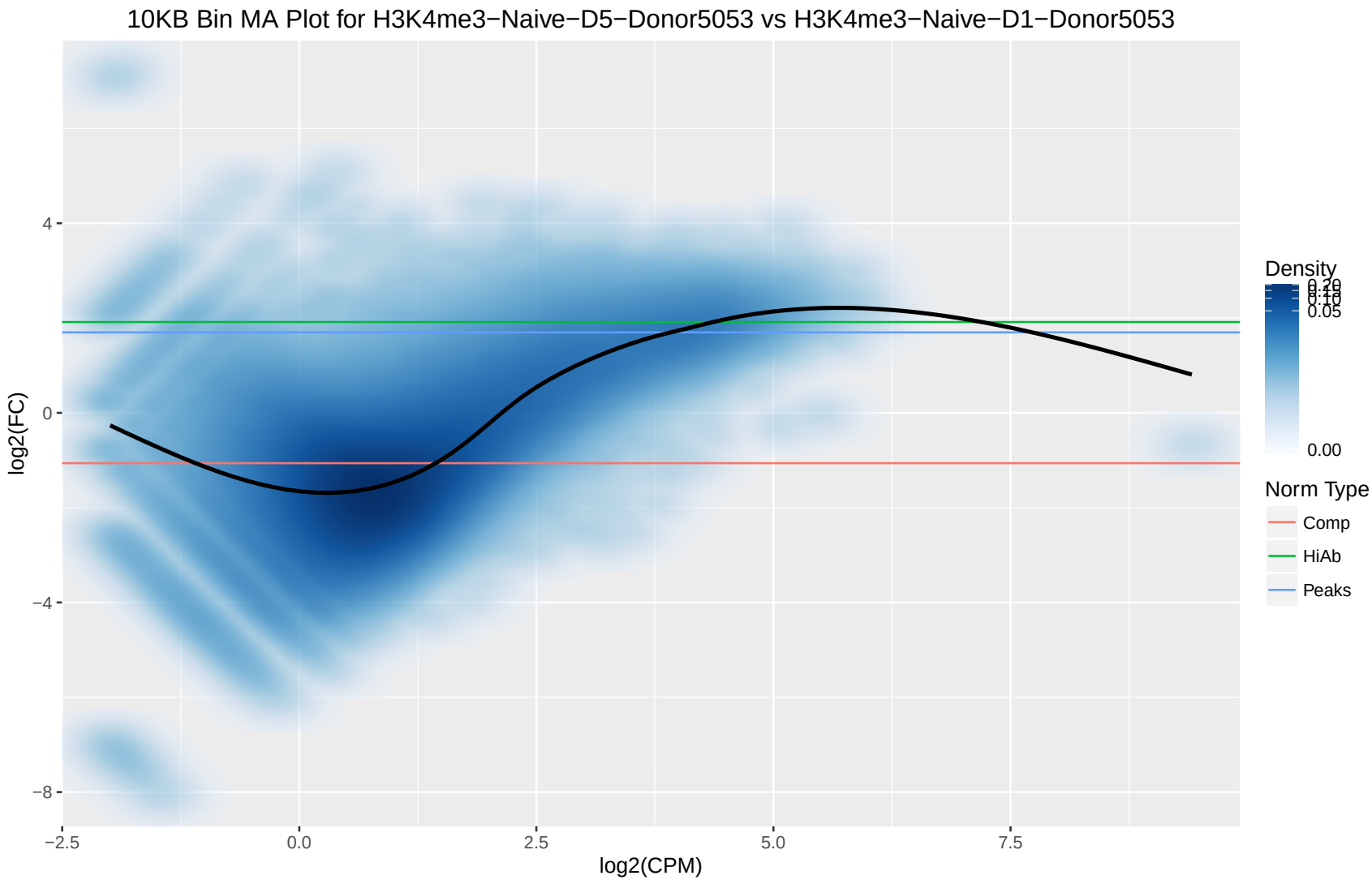

Figure 18: 10kb bin MA Plot with blacklisting

MA plot WITHOUT blacklist exclusion

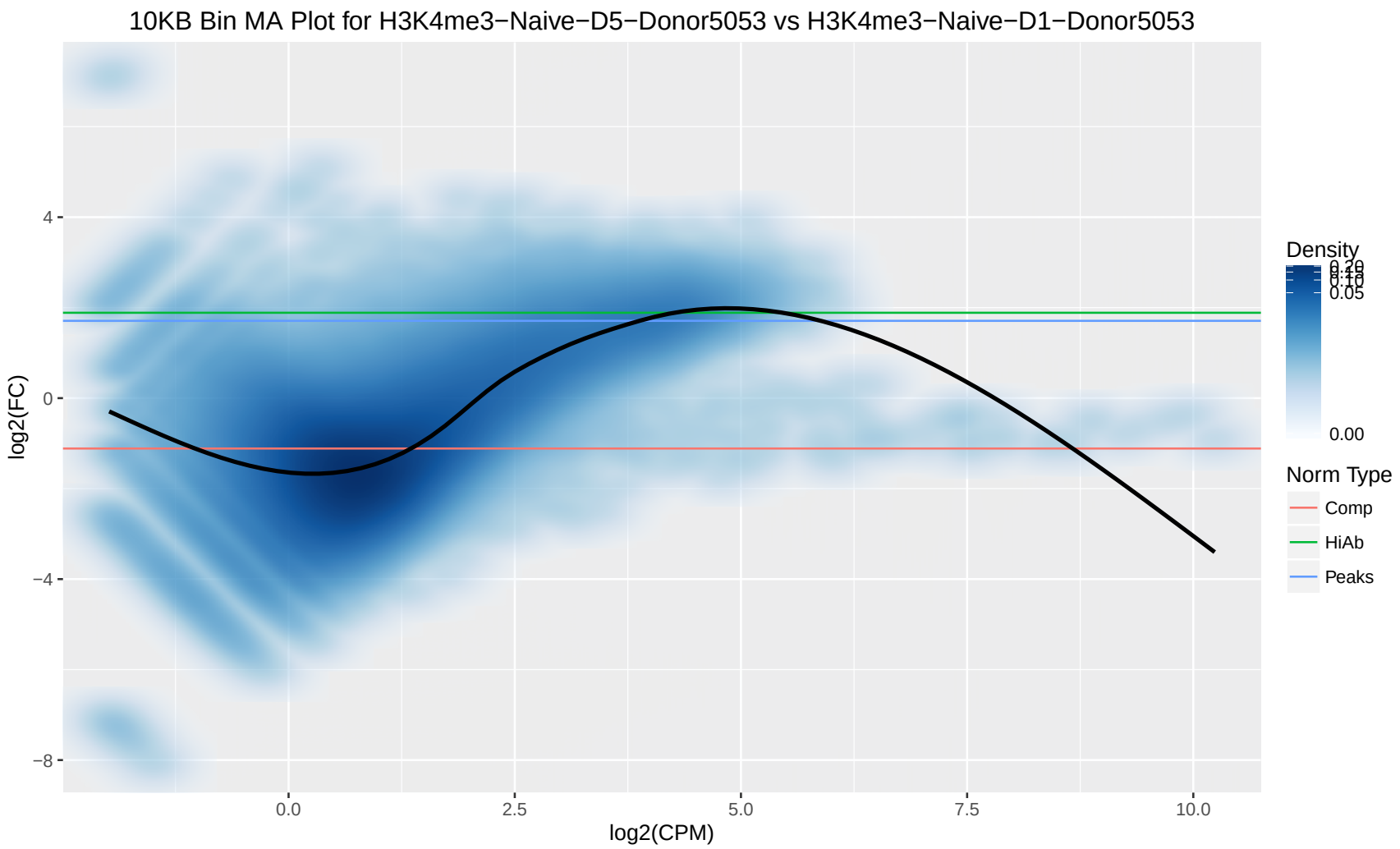

Figure 19: 10kb bin MA Plot without blacklisting

Any Questions?