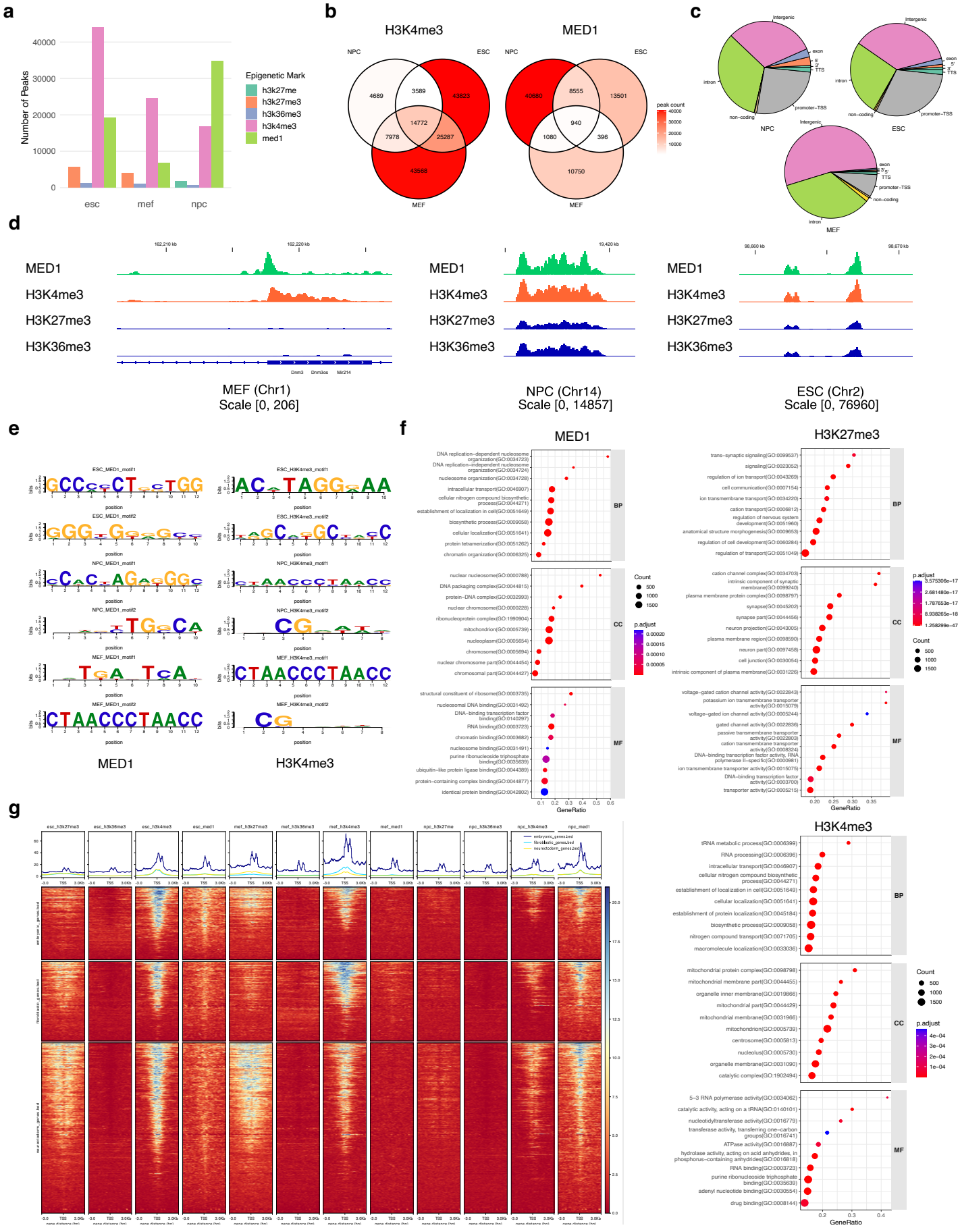


Figure 1. Epigenetic-seq Analysis

(A) Bar chart showing the number of peaks identified in each ChIP-seq sample across three cell types. Different colors indicate distinct epigenetic marks.
(B) Three-way Venn diagram showing the overlap of MED1 and H3K4me3 peaks across three cell types. There are cell-type-specific and shared peaks for both.
(C) Pie chart showing the genome-wide distribution of peaks for MED1 across threee cell types. MED1 predominantly binds to intergenic regions, promoters (TSS), and introns. The proportions of these genomic features vary among the three cell types.
(D) Genome browser tracks for MED1 and H3K4me3 binding in three representative loci from ESCs, MEFs, and NPCs, respectively.
(E) Top 2 DNA motifs identified by HOMER for MED1 and H3K4me3 binding sites in ESCs, MEFs, and NPCs.
(F) CistromeGO analysis summary for the regions of MED1 binding, H3K4me3, and H3K27me3. Shown are the top 10 enriched terms for biological process (BP), cellular component (CC), and molecular function (MF).
(G) Heatmaps and average profiles showing the ChIP-seq signal intensity of MED1, H3K4me3, H327me3, and H3K36me3 over three gene sets: neurectodermal, fibroblastic, and embryonic genes. Each row represents one gene locus aligned at the transcription start site (TSS), and the signal is visualized over ±3 kb.
Abbreviations: ESC, embryonic stem cell; MEF, mouse embryonic fibroblast; NPC, neural progenitor cell; MED1, Mediator 1; H3K4me3, histone H3 lysine 4 trimethylation; H3K27me3, histone H3 lysine 27 trimethylation; H3K36me3, histone H3 lysine 36 trimethylation.

12212859 Sijie Li

script

preprocess

1. align
   - do batch (all the steps are similar, only shown in the align steps)

```
for fasta in ./fqs/*.fq.gz
do
    base=$(basename ${fasta})
    out=$(echo $base | sed 's/.rp1.fq.gz//')
    sbatch -J "${out}_bowtie" --export=ALL,f=${fasta},out=${out} align.slurm
done
```

   - slurm

```
bowtie2 --end-to-end --no-unal \
-x bowtie2-indeces/mm10 \
-U ${f} -p 8 -S ${out}.sam
```

2. peak calling

```
samtools view -bS ${f} | samtools sort -o ${out}.sorted.bam
samtools index ${out}.sorted.bam
macs2 callpeak -t ${out}.sorted.bam -f BAM \
    -g mm -n ${out} --outdir macs2_output/${out}/
```

3. makewig

```
mkdir wiggle
bamCoverage -b ${out}.sorted.bam -o wiggle/${out}.bw \
 --extendReads 200 -p 8
```

4. HOMER annotation

```
mkdir -p HOMER_anno
annotatePeaks.pl macs2_output/${out}/${out}_peaks.narrowPeak \
    mm10 > HOMER_anno/${out}.txt
mkdir -p HOMER_motif
findMotifsGenome.pl macs2_output/${out}/${out}_peaks.narrowPeak \
    mm10 HOMER_motif/${out}/
```

Panel A

count_peak.sh

```
for peak in macs2_output/*/*.narrowPeak
do
    num_peak=$(wc -l < ${peak})
    base=$(basename ${peak})
    out=$(echo $base | sed 's/_peaks.narrowPeak//')
    echo ${out}\\t${num_peak} >> peak_count.tsv
done
```

bar chart

```
peak_counts <- read.table("peak_count.tsv")
colnames(peak_counts) <- c("label","count")
peak_counts$celltype <- sub("^(.*)_(.*)$", "\\1", peak_counts$label)
peak_counts$mark <- sub("^(.*)_(.*)$", "\\2", peak_counts$label)
peak_counts$celltype <- factor(peak_counts$celltype, levels = c("esc", "mef", "npc"))
# color blind friendly color
mark_colors <- brewer.pal(n = length(unique(peak_counts$mark)), name = "Set2")
pA <- ggplot(peak_counts, aes(x = celltype, y = count, fill = mark)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  ...
ggsave("PanelA_peak_barplot.pdf", plot = pA, width = 6, height = 4, units = "in")
```

Panel B

bedtool: intersect--FAIL

follow the official guide of bedtool intersect

- I first use **summit** to identify overlap, but it is **too strict**! (only when the summit overlap!)

- thus, I then use .narrowPeak result to identify overlap
  - if **≥ 1 bp overlap**--consider an intersection
  - use `-u` to report the fact at least one overlap was found in B (do not record the overlapping region)
    - do not use in the first intersect when determining the triple overlapping.
  - I do not add multiple databases after `-b` for it will return any overlapping happens in these databases (like a union not intersection)

```
bedtools intersect -u -a macs2_output/mef_med1/mef_med1_peaks.narrowPeak \
-b macs2_output/esc_med1/esc_med1_peaks.narrowPeak > intersect/med1/mef_vs_esc_med1.bed

bedtools intersect -a macs2_output/npc_med1/npc_med1_peaks.narrowPeak \
-b macs2_output/esc_med1/esc_med1_peaks.narrowPeak\
| bedtools intersect -u -a macs2_output/mef_med1/mef_med1_peaks.narrowPeak \
-b - > intersect/med1/all_med1.bed
```

Same for other pairs and markers.

count intersections

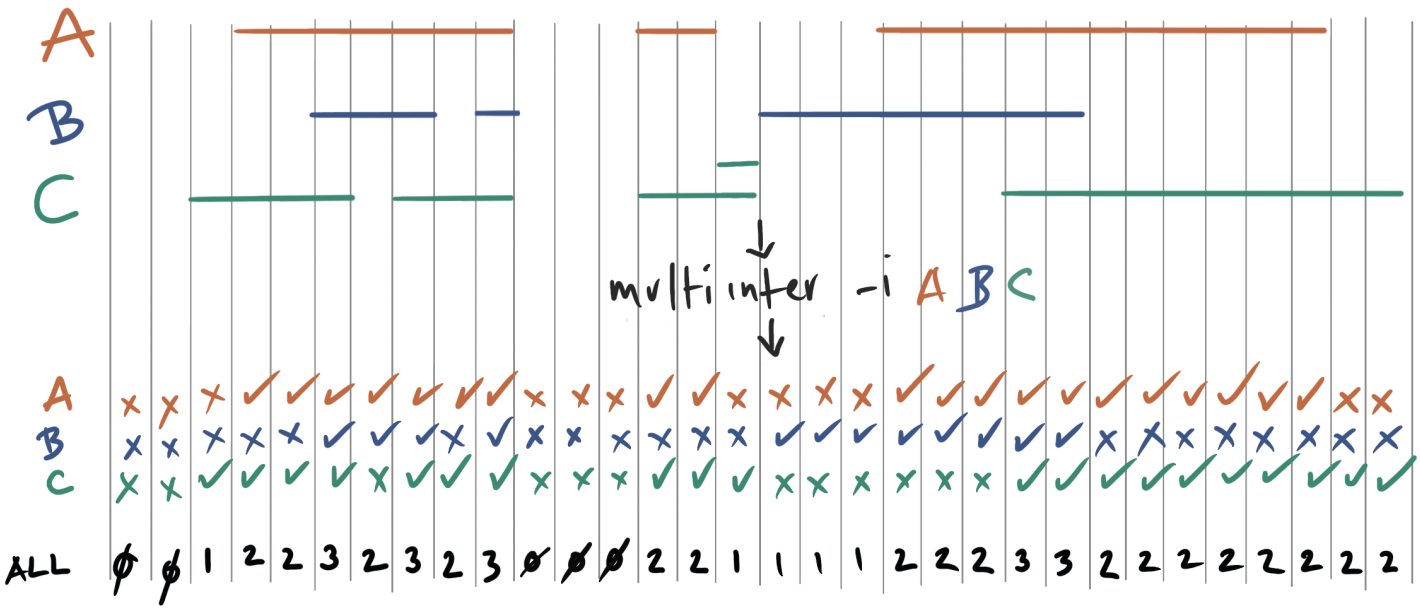
```
for peak in ./macs2_output/*med1/*_med1_peaks.narrowPeak
do
base_peak=$(basename ${peak})
out=$(echo $base_peak | sed 's/_peaks.narrowPeak//')
echo $out $(wc -l < $peak)  >> overlap.tsv # use "<" to avoid file directory when wc
done

for overlap in intersect/med1/*.bed
do
base=$(basename ${overlap})
out=$(echo $base | sed 's/.bed//')
echo $out $(wc -l < $overlap) >> overlap.tsv
done
```

Same for h3k4me3

BUT, the repeating intersect will cause problem!

bedtool: multiinter

make sure hte `.narroPeak` result already sorted




```
bedtools multiinter -i macs2_output/npc_med1/npc_med1_peaks.narrowPeak \
macs2_output/esc_med1/esc_med1_peaks.narrowPeak \
macs2_output/mef_med1/mef_med1_peaks.narrowPeak > intersect/multiinter
```

Venn plot

ID generation

`ggVennDiagram` take in ID (those with the same id are considered in the intersection), thus, I have to number the region ovelapping label (using abcd within AWK)

```
#!/bin/sh
mkdir -p intersect/Venn
for intersect in intersect/multiinter_*
do
base=$(basename $intersect)
out=$(echo $base | sed 's/multiinter_//')
awk 'BEGIN{a=0;b=0;c=0;d=0}{
if ($5 == "1"){print "NPC"a;a++}
else if ($5 == "1,2"){print "NPC_ESC"b;b++}
else if ($5 == "1,3"){print "NPC_MEF"c;c++}
else if ($5 == "1,2,3"){print "all"d;d++}}' $intersect \
> intersect/Venn/NPC_${out}_Venn
# same for the other cell types
done
```

- I first use `VennDiagram` , but it is too ugly! Use `ggVennDiagram` instead.

- Use the third column of nultiinter result as label (indication which cell have this peak region)

```
NPC_med1 <- scan("./Venn/NPC_med1_Venn",what = character())
ESC_med1 <- scan("./Venn/ESC_med1_Venn",what = character())
MEF_med1 <- scan("./Venn/MEF_med1_Venn",what = character())

med1_list <- list(
  NPC = NPC_med1,
  ESC = ESC_med1,
  MEF = MEF_med1)
pB_med1 <- ggVennDiagram(med1_list, label_alpha = 0, label = "count") +
  scale_fill_gradient(low = "white", high = "red",name ="peak count")
ggsave("panelB_MED1_ggvenn.pdf", plot = pB_med1, width = 5, height = 5)
```

Same for h3k4me3.

Panel C

count number of genome feature

```
mkdir anno_MED1_distribution
for anno in *_med1.txt
do
base=$(basename $anno)
out=$(echo $base | sed 's/_med1.txt//')
awk 'BEGIN{getline}{print $8}' $anno | sort | uniq -c > anno_MED1_distribution/${out}
done
```

pie plot

```
esc_df <- read.table("anno_MED1_distribution/esc")
esc_dis <- esc_df[,1]
names(esc_dis) <- esc_df[,2]
annotation_colors <- brewer.pal(n = length(esc_dis), name = "Set2")
pdf("panelC_ESC_pie.pdf", width=4, height=4)
pie(esc_dis, main="ESC",cex=0.5,col = annotation_colors)
dev.off()
```

same for other cells.

Panel D

- find from the intersection of the markers peak (use bedtools intersect)
- mind the range (intensity of the peak) when turning on the autoscale in IGV

Panel E

```
motiffile <- read.table("HOMER_motif/esc_med1/homerResults/motif1.motif",comment.char =
A <- motiffile[,1]
C <- motiffile[,2]
G <- motiffile[,3]
T <- motiffile[,4]
pcm <- rbind(A,C,G,T)
ESC_MED1_motif1 <- new("pcm", mat=as.matrix(pcm), name="ESC_MED1_motif1")
opar<-par(mfrow=c(6,1))
plot(ESC_MED1_motif1)
...
par(opar)
```

same for other motifs, cell types, and markers.

Panel F

merge peak

merge the peak of the marker across the three type of cells.

```
mkdir merged_peaks
# MARKER="med1"
# MARKER="h3k4me3"
MARKER="h3k27me3"
OUTPUT="merged_peaks/${MARKER}_all_celltypes.bed"
> $OUTPUT # if there already have this file--clear it!
for peak in macs2_output/*${MARKER}/*${MARKER}_peaks.narrowPeak
do
  echo "Merging $peak"
  cat $peak >> $OUTPUT
done
```

then load it onto CistromeGO

dot plot

```
bp <- read.table("./med1/1750520304_CistromeGO_go_bp_result.txt", sep = "\t", comment.c
cc <- read.table("./med1/1750520304_CistromeGO_go_cc_result.txt", sep = "\t", comment.c
mf <- read.table("./med1/1750520304_CistromeGO_go_mf_result.txt", sep = "\t", comment.c
bp$type <- "BP"
cc$type <- "CC"
mf$type <- "MF"
library(dplyr)
library(tidyr)
library(stringr)
# GeneRatio = b/B
go_data <- bind_rows(bp, cc, mf) %>%  # vertically
  separate(N.n.B.b, into = c("N", "n", "B", "b"), sep = ",\\s*|,", convert = TRUE) %>%
  mutate(GeneRatio = b / B,
         Term = str_wrap(Term, width = 50))  # avoid too long
top_go <- go_data %>%
  group_by(type) %>%
  slice_min(order_by = FDR, n = 10)
# visualization
library(ggplot2)
p <- ggplot(top_go, aes(x = GeneRatio, y = reorder(Term, GeneRatio))) +
  geom_point(aes(size = Gene_number, color = FDR)) +
  facet_grid(type ~ ., scales = "free_y", space = "free") +
  scale_color_gradient(low = "red", high = "blue", name = "p.adjust") +
  scale_size_continuous(name = "Count") +
  theme_bw(base_size = 12) +
  theme(
    strip.text.y = element_text(angle = 0, face = "bold", size = 12),
    strip.background = element_rect(fill = "grey90", color = NA),
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(size = 10),
    axis.title.x = element_text(size = 12),
    panel.spacing = unit(1, "lines")
  ) +
  xlab("GeneRatio") +
  ylab(NULL)
ggsave("PanelF_MED1.pdf", plot = p, width = 7, height = 10)
```

Same for other markers

Panel G

using deeptools

```
computeMatrix reference-point \
  -S wiggle/*.bw \
  -R ~/data/epigenetic/gene_beds/*.bed \
  --referencePoint TSS \
  -b 3000 -a 3000 \
  --skipZeros \
  -p 4 \
  -out matrix_TSS.tab.gz

plotHeatmap \
  -m matrix_TSS.tab.gz \
  -out heatmap_TSS.pdf \
  --whatToShow "plot, heatmap and colorbar" \
  --colorMap RdYlBu \
  --refPointLabel "TSS"

plotProfile \
  -m matrix_TSS.tab.gz \
  -out plotProfile_TSS.pdf \
  --perGroup \
  --legendLocation upper-right \
  --refPointLabel "TSS"
```

legend

Abbreviations: ESC, embryonic stem cell; MEF, mouse embryonic fibroblast; NPC, neural progenitor cell; MED1, Mediator 1; H3K4me3, histone H3 lysine 4 trimethylation; H3K27me3, histone H3 lysine 27 trimethylation; H3K36me3, histone H3 lysine 36 trimethylation.
Panel A. Bar chart showing the number of peaks identified in each ChIP-seq sample across three cell types. Different colors indicate distinct epigenetic marks.
Panel B. Three-way Venn diagram showing the overlap of MED1 and H3K4me3 peaks across three cell types. There are cell-type-specific and shared peaks for both.
Panel C. Pie chart showing the genome-wide distribution of peaks for MED1 across threee cell types. MED1 predominantly binds to intergenic regions, promoters (TSS), and introns. The proportions of these genomic features vary among the three cell types.
Panel D. Genome browser tracks for MED1 and H3K4me3 binding in three representative loci from ESCs, MEFs, and NPCs, respectively.

Panel E. Top 2 DNA motifs identified by HOMER for MED1 and H3K4me3 binding sites in ESCs, MEFs, and NPCs.
Panel F. CistromeGO analysis summary for the regions of MED1 binding, H3K4me3, and H3K27me3. Shown are the top 10 enriched terms for biological process (BP), cellular component (CC), and molecular function (MF).
Panel G. Heatmaps and average profiles showing the ChIP-seq signal intensity of MED1, H3K4me3, H327me3, and H3K36me3 over three gene sets: neurectodermal, fibroblastic, and embryonic genes. Each row represents one gene locus aligned at the transcription start site (TSS), and the signal is visualized over ±3 kb.