Recruitment Data Challenge

The Bioinformatics & Biostatistics Group @ The Francis Crick Institute

Introduction

Here you will find the data from an RNA-Seq and ATAC-Seq experiment. Both experiments have the same design. There is a treatment and control group each containing three replicates making a total of six samples per experiment. The data files are defined as follows (all files are tab delimited text files):

RNA-Seq Data

- **rnaseq_design.txt**: Sample ids and corresponding condition labels.
- **rnaseq_gene_counts.txt**: Raw (not normalised) gene-level read counts for each sample.
- **rnaseq_annotation.txt**: Gene level annotation.

ATAC-Seq Data

- **atacseq_design.txt**: Sample ids and corresponding condition labels.
- **atacseq_peak_counts.txt**: Raw (not normalised) ATAC-Seq peak level counts for each sample.
- **atacseq_peaks.bed**: A bed file defining the peak loci

All sequence data were aligned to the human genome reference hg38.

The Challenge

The treatment here is thought to activate a transcriptional program via remodelling of the chromatin architecture. The aim here is to:
1. Identify genes that may be regulated in this fashion.
2. Identify the possible transcriptional programs involved.
3. Present candidate transcription factors that may be responsible for the underlying regulation.

Please produce a 20 minute presentation detailing your exploration of the data, your analysis approach and findings?

Analysis

Strategy

1. Identify genes with significant changes in expression.
2. Identify zones with significant changes in accessibility.
3. Detect hotspots in accessibility changes over gene regulatory areas of differentially expressed genes.
4. Detect enriched TF motifs in zones presenting accessibility changes.
5. Detect enriched TF motifs in hotspots.
6. Perform GO Analysis to put genes in context.

AME Analysis of Motif Enrichment

Defining windows for ROIs

High Access / Upregulated

- Window 0: -2000 to 0
- Window 1: -4250 to -3250
- Window 2: -8000 to -4500
- Window 3: -9750 to -8500

[1]
```
!awk 'BEGIN{OFS="\t";offset5=-2000;offset3=0}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_hiprom_win0.fasta
```

[2]
```
!awk 'BEGIN{OFS="\t";offset5=-4250;offset3=-3250}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_hiprom_win1.fasta
```

[3]
```
!awk 'BEGIN{OFS="\t";offset5=-8000;offset3=-4500}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
```

```
not in genome"}}' output/GRCh38.genome output/upregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_hiprom_win2.fasta
```

[4]
```
!awk 'BEGIN{OFS="\t";offset5=-9750;offset3=-8500}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_hiprom_win3.fasta
```

High Access / Downregulated

- Window 0: -2000 to 0
- Window 1: -9500 to -4500

[5]
```
!awk 'BEGIN{OFS="\t";offset5=-2000;offset3=0}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/dwregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/dwregby_hiprom_win0.fasta
```

[6]
```
!awk 'BEGIN{OFS="\t";offset5=-9500;offset3=-4500}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/dwregby_hiprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/dwregby_hiprom_win1.fasta
```

Low Access / Upregulated

- Window 0: -3000 to 0
- Window 1: -6000 to -4500
- Window 2: -9750 to -8500

[7]
```
!awk 'BEGIN{OFS="\t";offset5=-3000;offset3=0}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_loprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_loprom_win0.fasta
```

[8]
```
!awk 'BEGIN{OFS="\t";offset5=-6000;offset3=-4500}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_loprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_loprom_win1.fasta
```

[9]
```
!awk 'BEGIN{OFS="\t";offset5=-9750;offset3=-8500}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/upregby_loprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/upregby_loprom_win2.fasta
```

Low Access / Downregulated

- Window 0: -3000 to 0
- Window 1: -9750 to -8250

[10]
```
!awk 'BEGIN{OFS="\t";offset5=-3000;offset3=0}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/dwregby_loprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/dwregby_loprom_win0.fasta
```

[11]
```
!awk 'BEGIN{OFS="\t";offset5=-9750;offset3=-8250}
NR==FNR{chr[$1]=$2;next} {if($1 in chr){lim=chr[$1];if($6=="+")
{left=$2+offset5;right=$2+offset3}else{left=$3-offset3;right=$3-
offset5}; if(left<0){left=0}else if(right>lim)
{right=lim}else{$2=left;$3=right}; print}else{print "error "$1"
not in genome"}}' output/GRCh38.genome output/dwregby_loprom.bed
| bedtools intersect -a - -b output/atac_hi.bed | sort -k1,1 -
k2,2n | bedtools getfasta -name -s -fi external/GRCh38.fa -bed -
| awk '/^>/{$0=$0"_"(++i)}1' > output/dwregby_loprom_win1.fasta
```

All High Access Regions
All Low Access Regions
All Regions

[12]
```
!bedtools getfasta -name -fi external/GRCh38.fa -bed
output/atac_hi.bed > output/atac_hi.fasta
!awk '{/>/&&++a||b+=length()}END{print b/a,a}'
output/atac_hi.fasta
!bedtools getfasta -name -fi external/GRCh38.fa -bed
output/atac_lo.bed > output/atac_lo.fasta
!awk '{/>/&&++a||b+=length()}END{print b/a,a}'
output/atac_lo.fasta
!bedtools getfasta -name -fi external/GRCh38.fa -bed
output/atac_diffx.bed > output/atac_diffx.fasta
!awk '{/>/&&++a||b+=length()}END{print b/a,a}'
output/atac_diffx.fasta
!bedtools getfasta -name -fi external/GRCh38.fa -bed
output/atacseq_fpeaks.bed6 > output/atac_fpeaks.fasta
!awk '{/>/&&++a||b+=length()}END{print b/a,a}'
output/atac_fpeaks.fasta
```

```
741.084 16971
929.923 11618
817.825 28589
685.638 173264
```

Generating backgrounds for AME

Creating a random background

[13]
```
!bedtools random -l 100 -n 100000 -seed 42 -g
output/GRCh38.genome > output/atac_bgd.bed 2> /dev/null
```

```
!bedtools getfasta -name -fi external/GRCh38.fa -bed
output/atac_bgd.bed > output/atac_bgd.fasta
!awk '{/>/&&++a||b+=length()}END{print b/a,a}'
output/atac_bgd.fasta
```

```
100 100000
```

AME: General and Differential Landscape - Enriched Motifs

[14]
```
!ame --o output/AME/total_VS_bgd --scoring avg --method fisher --
hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_bgd.fasta output/atac_fpeaks.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/total_VS_bgd'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (173264).
MOTIF: 1 SEQ: 273264/273264
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 273264/273264
```

[15]
```
!ame --o output/AME/diffx_VS_total --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_fpeaks.fasta output/atac_diffx.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/diffx_VS_total'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
```

```
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (28589).
MOTIF: 1 SEQ: 201853/201853
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 201853/201853
```

AME: High/Low Accessibility Motifs Enrichment

Note: I have selected Human Only Motifs from JASPAR
http://jaspar.genereg.net
File is provided in external/JASPAR2018_HUMAN_meme.txt

[16]
```
!ame --o output/AME/hi_VS_diffx --scoring avg --method fisher --
hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_diffx.fasta output/atac_hi.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/hi_VS_diffx'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (16971).
MOTIF: 1 SEQ: 45560/45560
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 45560/45560
```

[17]
```
!ame --o output/AME/lo_VS_diffx --scoring avg --method fisher --
hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_diffx.fasta output/atac_lo.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/lo_VS_diffx'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (11618).
MOTIF: 1 SEQ: 40207/40207
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 40207/40207
```

AME Regional Enrichment with Up/Down Regulated Genes

High Access / Upregulated Only Window 0 has hits

- Window 0: -2000 to 0
- Window 1: -4250 to -3250
- Window 2: -8000 to -4500
- Window 3: -9750 to -8500

[18]
```
!ame --o output/AME/uphiwin0_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/upregby_hiprom_win0.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/uphiwin1_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/upregby_hiprom_win1.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/uphiwin2_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/upregby_hiprom_win2.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/uphiwin3_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/upregby_hiprom_win3.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/uphiwin0_VS_hi'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (85).
MOTIF: 1 SEQ: 17056/17056
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 17056/17056
```

High Access / Downregulated (No hits)

- Window 0: -2000 to 0
- Window 1: -9500 to -4500

[19]
```
#!ame --o output/AME/dwhiwin0_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/dwregby_hiprom_win0.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/dwhiwin1_VS_hi --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_hi.fasta output/dwregby_hiprom_win1.fasta
external/JASPAR2018_HUMAN_meme.txt
```

Low Access / Upregulated No Hits

- Window 0: -3000 to 0
- Window 1: -6000 to -4500
- Window 2: -9750 to -8500

[20]
```
#!ame --o output/AME/uplowin0_VS_lo --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_lo.fasta output/upregby_loprom_win0.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/uplowin1_VS_lo --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
```

```
output/atac_lo.fasta output/upregby_loprom_win1.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/uplowin2_VS_lo --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_lo.fasta output/upregby_loprom_win2.fasta
external/JASPAR2018_HUMAN_meme.txt
```

Low Access / Downregulated No Hits

- Window 0: -3000 to 0
- Window 1: -9750 to -8250

[21]
```
#!ame --o output/AME/dwlowin0_VS_diffx --scoring avg --method
fisher --hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --
control output/atac_diffx.fasta output/dwregby_loprom_win0.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/dwlowin0_VS_lo --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_lo.fasta output/dwregby_loprom_win0.fasta
external/JASPAR2018_HUMAN_meme.txt
#!ame --o output/AME/dwlowin1_VS_lo --scoring avg --method fisher
--hit-lo-fraction 0.25 --evalue-report-threshold 0.05 --control
output/atac_lo.fasta output/dwregby_loprom_win1.fasta
external/JASPAR2018_HUMAN_meme.txt
```

```
Using average odds scoring.
Added external/JASPAR2018_HUMAN_meme.txt to motif_sources which now has
1 file names.
Motif file name is external/JASPAR2018_HUMAN_meme.txt.
Writing results to output directory 'output/AME/dwlowin0_VS_diffx'.
E-value threshold for reporting results: 0.05
Checking alphabets in 1 motif files.
Loading motifs from file 'external/JASPAR2018_HUMAN_meme.txt'
Loading primary sequences.
Loading control sequences.
Not in partition maximization mode. Fixing partition at the number of
primary sequences (3).
MOTIF: 1 SEQ: 28592/28592
Sorting sequences by sequence PWM score to get PWM ranks; breaking ties
to put negatives first.
Leaving sequences sorted by PWM score.
Optimizing over sequence PWM score threshold.
MOTIF: 639 SEQ: 28592/28592
```

Clean Up Results

[22]
```
!awk 'BEGIN{OFS="\t"} $1 !~ "#" {split($4, a, "::"); for (k in a)
{print $1,$3,a[k],$7,$8};delete a}'
output/AME/hi_VS_diffx/ame.tsv | sed 's/'\(.*\)'//g' >
output/hi_TFMs.tsv
!awk 'BEGIN{OFS="\t"} $1 !~ "#" {split($4, a, "::"); for (k in a)
{print $1,$3,a[k],$7,$8};delete a}'
output/AME/lo_VS_diffx/ame.tsv | sed 's/'\(.*\)'//g' >
output/lo_TFMs.tsv
!awk 'BEGIN{OFS="\t"} $1 !~ "#" {split($4, a, "::"); for (k in a)
{print $1,$3,a[k],$7,$8};delete a}'
output/AME/uphiwin0_VS_hi/ame.tsv | sed 's/'\(.*\)'//g' >
output/uphiprox_TFMs.tsv
```