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.

Download Data

https://www.dropbox.com/s/075d1qzbm5jjq9p/data_challenge.tar.gz?dl=0

Preliminary Checks

What does the data look like?

```
[1]  !head data_challenge/atacseq_peak_counts.txt
     !head data_challenge/atacseq_peaks.bed
     !awk 'END{print NR}' data_challenge/atacseq_peak_counts.txt
     !awk 'END{print NR}' data_challenge/atacseq_peaks.bed
     !head data_challenge/rnaseq_gene_counts.txt
     !awk 'END{print NR}' data_challenge/rnaseq_gene_counts.txt
```

```
peakid  s84     s85     s86     s93     s94     s95
1:10003-10507   278     195     292     255     287     284
1:20221-22634   66      56      90      67      66      120
1:28574-30038   184     71      139     157     153     160
1:37947-39588   28      15      42      81      70      128
1:90824-91498   15      13      23      17      17      29
1:107006-107256 2       1       3       8       6       20
1:127516-127818 5       2       6       13      14      21
1:136028-136767 16      14      15      11      16      20
1:137366-137718 11      7       9       15      13      23
1       10002   10507
1       20220   22634
1       28573   30038
1       37946   39588
1       90823   91498
1       107005  107256
1       127515  127818
1       136027  136767
1       137365  137718
1       138436  139449
173286
```

```
173285
featureid       s69     s70     s71     s75     s76     s77
ENSG00000000003 1       1       0       8       2       1
ENSG00000000005 0       0       0       0       0       0
ENSG00000000419 993     469     664     1172    491     685
ENSG00000000457 385     207     235     610     226     353
ENSG00000000460 849     436     522     1002    430     608
ENSG00000000938 26      6       12      75      43      44
ENSG00000000971 4       2       11      10      7       7
ENSG00000001036 3       1       2       1       0       0
ENSG00000001084 1767    861     995     1984    864     1124
58052
```

Do peak IDs match in bed/counts?

[2]

```
!awk 'NR==FNR{x[$1]; next} {y=$1":"$2+1"-"$3; if (y in x)
{present++;} else {absent++}} END{print "present = " present,
"absent = " absent}' data_challenge/atacseq_peak_counts.txt
data_challenge/atacseq_peaks.bed
```

```
present = 173285 absent =
```

Formatting into BED6

[3]

```
!awk 'BEGIN{OFS="\t"}{name=$1":"$2+1"-"$3; score=$3-$2;
strand="."; print $0,name,score,strand}'
data_challenge/atacseq_peaks.bed > output/atacseq_peaks.bed6
!awk 'BEGIN{OFS="\t"}NR>1{name=$1; score=$4-$3+1; strand=$6;
print $2,$3-1,$4,name,score,strand}'
data_challenge/rnaseq_annotation.txt > output/rnaseq_genes.bed6
```

GRCh38 genome index generated from the GENCODE GRCh38 Primary Assembly

[4]

```
!curl
"ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_
32/GRCh38.primary_assembly.genome.fa.gz" | gunzip -c | sed
's/chr//g' > external/GRCh38.fa && samtools faidx
external/GRCh38.fa
!awk 'BEGIN{OFS="\t"} {print $1,$2}' external/GRCh38.fa.fai |
sort -k1,1 -k2,2n > output/GRCh38.genome
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time
Current
```

```
                                 Dload  Upload   Total   Spent    Left
Speed
100  805M  100  805M    0     0  3976k      0  0:03:27  0:03:27 --:--:--
2361k
```

Are peaks supported by at least 2 replicates from a group?

[5]

```
!awk '{ for(i = 2; i <= 4; i++) {if($i>0){ctrl++}; } for(i = 5; i
<= NF; i++) {if($i>0){test++};} if(ctrl < 2 && test < 2)
{flag++}; test=0;ctrl=0}END{print "sites supported by less than 2
replicates per group:"flag}'
data_challenge/atacseq_peak_counts.txt
```

```
sites supported by less than 2 replicates per group:
```

Some regions are blacklisted, looking for overlaps and removing them
Original blacklist downloaded from
https://www.encodeproject.org/annotations/ENCSR636HFF/

[7]

```
!gunzip -c external/ENCFF419RSJ.bed.gz | sed 's/chr//g' | awk
'BEGIN{OFS="\t"}{name=$1":"$2+1"-"$3; score=$3-$2; strand=".";
print $0,name,score,strand}' > external/blacklist.bed
!bedtools intersect -wo -b output/atacseq_peaks.bed6 -a
external/blacklist.bed > output/atac_blacklisted.tsv
#If the peaks present more than 25 bp overlap with the black
listed sites they are removed:
!awk 'BEGIN{OFS="\t"} $13 > 25' output/atac_blacklisted.tsv |
bedtools subtract -A -a output/atacseq_peaks.bed6 -b - >
output/atacseq_fpeaks.bed6
!awk 'BEGIN{OFS="\t"} NR==FNR{if($13 > 25 ){bad[$10]};next}
{if($1 in bad){next}else{print $0}}' output/atac_blacklisted.tsv
data_challenge/atacseq_peak_counts.txt >
output/atacseq_fpeak_counts.txt
!awk 'END{print NR}' output/atacseq_fpeak_counts.txt
!awk 'END{print NR}' output/atacseq_fpeaks.bed6
```

```
173265
173264
```