

Session: Next Generation Sequencing (NGS)

ChIP-seq and ATAC/DNase-seq data analysis

OUTLINE PART2 –

0. Intersection of ATAC-seq and CTCF ChIP-seq and GWAS SNPs

The data for the practical sessions is located on this path:

/datapath/poznan_ChIPseq_analysis/gwas_intersection/

Intersection of open chromatin regions, TF bound regions, and Disease associated SNPs

Regulatory regions have rapidly been gaining interest as a focus for disease-based research. Protein coding mutations are simpler to work with as a single variant can cause significant damage to a protein gene. Regulatory regions however often have a degree of redundancy built in, either within the region itself or due to secondary regions i.e. for instance TFs can often tolerate a nucleotide change in their binding motif.

Here we are going to use the open chromatin regions from ATAC-seq data, intersect them with CTCF ChIP-seq peaks. Those regions that are both open chromatin and CTCF will then be intersected with disease associated variants from a GWAS (Genome-wide Association Studies) catalog. CTCF peaks in general are fairly invariant across cell lines, and thus a given region has the potential to be of interest across a wide set of cells and tissues.

Note that the GWAS catalog of SNPs contains only a few of all the potential SNPs associated with the listed disease. The reported SNP is just one in a haplotype block of SNPs, we don't have all the SNPs in that haplotype block. You can only get those through genome wide sequencing of patients, whereas genome wide association studies are done on microarrays and are limited to only a few SNPs in a block.

References for those who are interested:

1) PMID: 26719772 "Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome", 2015, Tak YG and Farnham PJ

2) ENSEMBL lists some alternative sources for variants

http://www.ensembl.org/info/genome/variation/sources_documentation.html

3) <http://www.broadinstitute.org/mammals/haploreg/haploreg.php>

A web-based tool (for human-based queries) where one can provide a genomic region or a SNP, and the tool returns SNPs in the region (or SNPs in linkage disequilibrium with the provided SNP) and also returns which TFBSs are altered by the SNPs.

Exercise:

Data is in: /datapath/compugen2016/part2_tf/gwas_intersection/
Remember to use the appropriate path for the data in the below commands.

Intersect CTCF peaks with open chromatin

Find CTCF open chromatin regions that intersect GWAS SNPs = potential regulatory SNP

Predict TFBS locations for CTCF

Ask if GWAS SNPs intersect the TFBSs

A)

Use bedtools to intersect the ATAC-seq peaks from MACS2 and CTCF ChIP-seq data (ENCODE), and save the CTCF regions to a file.

Replace variables with your path and filenames.

```
bedtools intersect -wo -a $ctcfdata -b macs2_atacseq_peaks.narrowPeak >
ctcfopenchromatin.txt
```

B)

Next intersect the results with the GWAS file. This file was downloaded from

<https://www.ebi.ac.uk/gwas/home> and processed so that the chr and coordinate data are in the first 3 columns. There is a lot of information on each line about the SNP and where it came from.

```
bedtools intersect -wo ctcfopenchromatin.txt -b $gwasdata >
ctcfopenchromatin_gwas.txt
```

If you look inside the result files, you will see the `-wo` option appends the information from the second file to the CTCF peak that it intersected with.

```
head -10 ctcfopenchromatin.txt
head -2 ctcfopenchromatin_gwas.txt
```

C)

Count how many CTCF peaks intersect with open chromatin regions, and how many CTCF peaks there were to start with. It is remotely possible a CTCF region might have overlapped more than one ATAC region, therefore make sure to only count unique CTCF regions. Then count how many intersected a GWAS SNP.

```
wc -l $ctcfdata # the ctcf peaks file
cut -f1-3 ctcfopenchromatin.txt | sort | uniq | wc -l # use only the unique identifier
of chromosome and coordinate, to make sure we aren't counting any CTCF peaks twice
```

```
cut -f1-3 ctcfopenchromatin_gwas.txt | sort | uniq | wc -l
```

As you can see, filtering to get a small subset to look at is not necessary, as there aren't too many intersections!

D)

Make a bed file of these CTCF peaks and fetch the sequences for these regions using bedtools and a genome fasta file

```
awk -F '\t' '{OFS="\t"; print $1,$2,$3,"","","+"}' ctcfopenchromatin_gwas.txt |
sort -k1,4,1 -k2,2n | uniq > ctcf_gwas.bed
bedtools getfasta -fi $genomefasta -bed ctcf_gwas.bed > ctcf_gwas.fa # genome
fasta file is hg19_main_female.fa
```

E)

To predict binding site locations you will use the webtool RSAT and the CTCF PFM from JASPAR. (We present in Appendix B (a separate document) a little overview of the score thresholding question/problem).

Copy the CTCF PFM from JASPAR into a text file, or take it from the `pfm_vertebrates_JASPAR2014.txt` file in the `motif_matrices/` directory.

RSAT: <http://embnet.ccg.unam.mx/rsa-tools/> -> Pattern matching -> "matrix-scan (quick)"

Upload the sequences, and the ctcf matrix. Nothing else needs to be done (unless you want to include your email address). Select "Go". The returned results will have passed the weight score threshold of 1. We do expect all sequences to have at least one motif, as CTCF is a very easy TF to ChIP and it has a very strong PFM.

The results provide a link to a **gff** file, of which only lines with "site" in the 3rd column will be kept. The 7th column shows which strand matched the PFM. The last column shows the TFBS string. Save this file.

To if the GWAS SNPs overlap a potential CTCF TFBS inside the peak, we will first extract the coordinates from the gff file and then do a bedtools intersect again, against the GWAS file.

this awk command prints only things with "site" in column 3 and it splits the first field in the file into chr and coordinates.

```
awk -F '\t' '{OFS="\t"; if($3~/site/){split($1,arr,":|-|\\(|)"); print  
arr[1],arr[2],arr[3], $2,$3,$4,$5,$6,$7,$8,$9} }' gff_file.gff > ctcf.tfbs.txt
```

```
head -3 ctcf.tfbs.txt # always a good idea to check what a new file looks like before using it
```

F)

Generate a file of TFBS coordinates to intersect with GWAS. The gff file provides the start and end of the TFBS as a distance from the end of the peak sequence (col. 6 and 7). e.g. peak chr10 21147257 21147617.... -258 -240. You can get the TFBS coordinates using these numbers (e.g. start=21147617+(-258)+1 and end=21147617+(-240)+1)

```
awk -F '\t' '{OFS="\t"; split($11,arr,"="); print $1,$3+$6+1,  
$3+$7+1,$8,arr[3],"+"}' ctcf.tfbs.txt > ctcf.tfbscoord.txt
```

```
bedtools intersect -wo -a ctcf.tfbscoord.txt -b $gwasdata >  
ctcf.tfbscoord.gwas.txt
```

make a bed file of the TFBS coordinates and the SNP coordinate to upload into IGV

```
awk -F '\t' '{OFS="\t"; print $1,$2,$3,$32,$4,$6,$9-1,$9}'  
ctcf.tfbscoord.gwas.txt | sort -k1.4,1 -k2,2n | uniq > ctcf.tfbscoord.bed
```

```
awk -F '\t' '{OFS="\t"; print $1,$8,$9,$9-$2+1,$4,$6,$32}'  
ctcf.tfbscoord.gwas.txt | sort -k1.4,1 -k2,2n | uniq > ctcf.gwascoord.bed #the  
4th field is the position of the SNP relative to the start of the TFBS
```

```
head ctcf.tfbscoord.bed  
head ctcf.gwascoord.bed
```

G)

In IGV open the two bed files you generated. Because the TFBS binding sites are so small you should be able to see in IGV the actual sequence string (above the RefSeq track).

As an easy visualization of where the SNP intersects the CTCF motif, use the JASPAR CTCF sequence logo we provided as a JPG. You may be able to drag it over the IGV window and stretch it to make it align with the sequence there (the "fwd" logo is for +ve strand, and the "rev" logo is for -ve strand). Not exactly high throughput visualization, but there aren't many SNPs to look at.

The “name” under the TFBS track is the SNP ID.

The “name” under the SNP track is a number, and it should indicate which column in the sequence logo the SNP is located in. Are any of the SNP’s in a strong information position of the CTCF logo? What letter is the GWAS SNP?

Zoom out to view the TFBS and SNP relative to neighbouring genes.

What genes are neighbours to the potential regulatory SNP located at the strongest position in a CTCF TFBS? Use gene cards to look up their functions. If CTCF regulated these genes, does it look like damaging the CTCF binding site would impact regulation of something interesting?

Yes, alternatively to IGV you could have simply read the number in column 4 of `ctcf.gwascoord.bed` and compared it to the sequence logo position.

H)

Use `grep` to find the information in the GWAS catalog file for the SNPs that are interesting (SNPs that hit a strong information position in the CTCF motif).

Example:

```
grep "rs11656696-A" GWAS_catalog_associations_201609_chrCoord.txt
```

Note: CTCF has a long, strong, sequence logo. It might take more than one SNP to really damage the ability of CTCF to bind to a site. Experiments would be the only way to know, unless someone has already shown that CTCF has reduced affinity for a certain sequence (you could perhaps search the Uniprobe oligomers).

End of section 3.