**ATAC-seq PIPELINE**

(SCRIPTS: /NGS/working\_projects/AtacSeq/scripts)

ATAC articles + forum

<http://www.the-scientist.com/?articles.view/articleNo/44772/title/Reveling-in-the-Revealed/>

<https://www.biostars.org/p/233574/>

<https://sites.google.com/site/atacseqpublic/?pli=1>

<https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-7-33>

QC:

Fastqc Command: $fastqc 2\_merged\_fastq\_files/\*.fq.gz --noextract -o 3\_qc\_pretrim -t 64

MultiQC command $multiqc -f -ip -o 4\_multiqc\_pretrim 3\_fastqc\_pretrim

Trimming:

remove Nextera primer sequences ([2017 paper](https://elifesciences.org/content/6/e21883" \l "SD24-data))

trim\_galore.pl

Mapping

alignment: grid\_bowtie2.pl (-X 2000 --no-mixed –no-discordant) ([reason for -X 2000](http://lab.loman.net/2013/05/02/use-x-with-bowtie2-to-set-minimum-and-maximum-insert-sizes-for-nextera-libraries/))

convert to bam:

1) Identify flags present -

cut -f2 mswatd1\_fixed.sam >mswatd1.flags

sort -u mswatd1.flags -o sorted\_mswatd1.flags

77 1st in pair, read unmapped, mate unmapped

141 2nd in pair, read unmapped, mate unmapped

83 1st in pair, mapped in proper pair, read reverse strand

163 2nd in pair, mapped in proper pair, mate reverse strand

99 1st in pair, mapped in proper pair, mate reverse strand

147 2nd in pair, mapped in proper pair, read reverse strand

2) grid\_sam2sortedbam\_master.pl

sort: sort by name

fixmate: fixes mates + filter out the unmapped reads (flag 77 + 141)

view: remove MAPQ<30 (0.999% probability of a correct alignment)

sort: sorts by coordinates + print as bam file

Filter out dups and chrM

remove duplicates: grid\_remdup\_master.pl

index bam file: for i in \*.bam; do samtools index -b $i; done;

remove mito reads: requires .bai files

for i in \*mito.bam; do bedtools bamtobed -i $i >/NGS/working\_projects/AtacSeq/data/10\_beds/$i.bed; done

rename bams: for i in \*.bam; do mv $i ${i/.dedup.bam.no-mito.bam/.nodup.nomito.bam}; done

delete the .bai files: rm \*.bai

re-index bam file: for i in \*.bam; do samtools index -b $i; done;

get stats in bash: for i in \*.bam; do samtools flagstat $i > $i.stats; done;

collect number of properly paired mapped reads: (needed later)

1) grep '0 mapped' \*stats >>sequencing\_depth.txt

2) sed -e 's/.nodup.nomito.bam.stats:/\t/' sequencing\_depth.txt | \

sed -e 's/ + 0 mapped (100.00% : N//' | \

sed -e 's/\/A)//' >depth.txt

3) rm sequencing\_depth.txt

rename the files: $ for i in \*.bam; do mv $i ${i/.nodup.nomito/}; done

$ for i in \*.bam.bai; do mv $i ${i/.nodup.nomito/}; done

$ for i in \*.bam.stats; do mv $i ${i/.nodup.nomito/}; done

Create beds, adjust for TN5 and remove blacklists

create beds: for i in \*mito.bam; do bedtools bamtobed -i $i >/NGS/working\_projects/AtacSeq/data/10\_beds/$i.bed; done

adjust beds for TN5: add 4 to ‘+’ strand and subtract 5 from ‘‐’ strand

for i in \*.bed;

do awk -F $'\t' 'BEGIN {OFS = FS}{if ($6 == "+") {$2 = $2 + 4; $3 = $3 + 4}

else if ($6 == "-") {$2 = $2 - 5; $3 = $3 - 5} print $0}' $i >$i.adj; done

remove blacklists: for i in \*.adj; do bedtools intersect -v -a $i -b \

/NGS/musRefs\_10/blacklists/mitochondrial.homologue.blacklist.mm10.sorted.bed \

/NGS/musRefs\_10/blacklists/signal.artifact.blacklist.mm10.sorted.bed \

> ${i/.nodup.nomito.bam.bed.adj/}.bed; done

The ATAC-seq authors created a [mitochondrial blacklist](https://www.biostars.org/p/178574/) (found [here](https://sites.google.com/site/atacseqpublic/atac-seq-analysis-methods/mitochondrialblacklists-1)) which represents high signal regions on the nuclear genome caused by read sequence homology with the mitochondrial genome. A signal artifact blackist has also been created by ENCODE (found [her1e](http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/mm10-mouse/))

[IDR analysis pipeline](https://sites.google.com/site/anshulkundaje/projects/idr)

**PEAK CALLING FOR VIEWING IN IGV/UCSC AND DOWNSTREAM ANALYSIS**

PEAK CALLING (macs2 callpeak)

Peakcalling just gives you a region where transposase cutting events are enriched and in which "you can dig deeper". NucleoATAC software then takes a bed with peaks and infers NFR and nucleosome positions. So MACS peaks are (by themselves) not suitable for base-pair-resolved analysis. If you just want to say which regions are accessable or not, then MACS peaks should do the job. If in-depth analysis, such as NFR and nucleosome positions or TF footprints are to be obtained, you need specialized software/ custom approaches, such as NucleoATAC, but also excellent data quality and good sequencing depth.

MACS resources: [readme](https://github.com/taoliu/MACS), [forum](https://groups.google.com/forum/" \l "!forum/macs-announcement), [ATAC](https://groups.google.com/forum/" \l "!topic/macs-announcement/4OCE59gkpKY)

File format/nomodel/shift/extsize discussions: [macs forum](https://groups.google.com/forum/" \l "!topic/macs-announcement/4OCE59gkpKY), [macs github](https://github.com/taoliu/MACS/issues/145), [bioconductor](https://support.bioconductor.org/p/86594/" \l "94037)

- To pre-shift before MACS2: bedtools slop ‐i filename.bed ‐g genomefile ‐l 75 ‐r ‐75 ‐s

-s detects strand orientation to add/subtract 75 appropriately

-g genomefile defining the length of each chromosome (/NGS/musRefs\_10/mm10\_chr\_size.txt)

Options:

--treatment [filename] Required, input file format is specified by -f

--format BED Must use BED. BEDPE not compatible with nomodel/shift/extsize.

BAM files can’t be adjusted for TN5 9bp insert so must use a BED file.

--nomodel Prevents MACS from building the shifting model required for chip-seq data

--shift -100 MACS removes bias in chip-seq data by moving the 5’ position of all reads --shift

controls how far the move is. ATAC/DNA-seq has no such bias so this move must be controlled using –-shift and then negated with –-extsize. When set to -100, the 5’position is moved 100 bases upstream.

--extsize 200 Designed to work with ATAC/DNA-seq to compensate for shift. Always = -2(shift).

e.g. --shift -100 --extsize 200. MACS extends the 3’ end of all reads to a final

fixed size fragment length of 200bps to smooth the pileup signals (this is called a smoothing window). So the read is now centred around the original 5’ position which is the cut site in the open chromatin. The size should be large enough to contain the transcription factor binding region (so don’t limit it to fragment length!)

--gsize mm recommended mappable genome size. -g ce lets MACS2 consider C elegans genome as background. Set it as 'dm' for fly or 'hs' for human (so mm for mouse)

--keep-dup all “All” keeps every tag – otherwise you lose some that start at same 5’ position

--bdg --SPMR Normalisation + stores fragment pileup, control lambda, -log10pvalue, and

-log10qvalue in bedGraph files. Generate pileup signal file of 'fragment pileup per million reads' in bedGraph format. From [biostars](https://www.biostars.org/p/168450/): SPMR normalizes by fragment pileup per million reads to allow transformation to bigwig

--qvalue 0.001 Default q = 0.05. set explicitly in case not set correctly (advice of forums)

--call-summits MACS will use a more sophisticated signal processing approach to find subpeak

summits in each enriched peak region. Highly recommended to detect adjacent binding events. The output subpeaks of a big peak region will have the same peak boundaries, and different scores and peak summit positions.

Alternatively use –-broad –-broad-cutoff

--broad Calls broad peaks by linking nearby highly enriched regions.

--broad-cutoff 0.001 Cutoff for broad region. Only works with --broad. Is a q value unless -p is set. Default: 0.1. Set explicitly in case not set correctly (advice of forums)

-outdir name the outdirectory. Not sure if written as 5e-2 or 0.05

--name sample name to be used as prefix for the output files (e.g. mswatd0)

Call a consensus peak set: (run command from inside bed file folder)

macs2 callpeak --format BED --treatment fgwatd0.bed fmgwat.bed fswatd7.bed mgwatd7.bed mswatd2.bed fgwatd-1.bed fmswat.bed fswatt4.bed mgwatt4.bed mswatd4.bed fgwatd1.bed fswatd0.bed mgwatd0.bed mmgwat.bed mswatd7.bed fgwatd2.bed fswatd-1.bed mgwatd-1.bed mmswat.bed mswatt4.bed fgwatd4.bed fswatd1.bed mgwatd1.bed mswatd0.bed fgwatd7.bed fswatd2.bed mgwatd2.bed mswatd-1.bed fgwatt4.bed fswatd4.bed mgwatd4.bed mswatd1.bed --nomodel --shift -100 --extsize 200 --qvalue 0.05 --cutoff-analysis --gsize mm --bdg --SPMR --keep-dup all

Narrow peaks: --outdir [path]/narrow/consensus --name narrow --call-summits

Broad peaks (for nucleoatac): --outdir [path]/broad/consensus --name broad --broad --broad-cutoff 0.001

Call individual peak sets: grid\_ind\_peak\_calling.pl

Create bigwigs (runtime 24hrs): grid\_bigwigs\_master.pl

[Converting bed to bigbed](https://wiki.hpcc.msu.edu/display/Bioinfo/Converting+BED+to+bigBED+Format)

**ALTERNATIVE**

Call peaks on individual samples: grid\_peak\_calling.pl

Merge output: cat file1 file2 file3 .... | \

sort -k1,1 -k2,2n | \

bedtools merge -i stdin - o collapse -c 4 > all\_peaks.bed

where c is the number of the column containing the peak ids

output: chr1 20 65 peakFile-3, peakFile-1, peakFile-2

Filter overlapping peaks (see below): getTopPeaks.pl

Filter out overlapping peaks for downstream analysis

From sams paper:

- The number of DHS sites identified in pre-adipocytes (day 0) and adipocytes (day 6) is surprisingly low (9988 and 11 936 DHS sites, respectively) compared with what is found for the 4-h time point (31 750 DHS sites) and what has previously been reported for other cell types (The ENCODE Project Consortium, 2007; Boyle et al, 2008). So they consider 31740 as normal for a timepoint → that is DHSs in Dnase data → it doesnt include nucleosomes!! that is why you have much more!

- A putative peak is defined as a 150-bp window. When peaks overlap, the peak with the highest z-score is chosen.

ATAC-seq authors [2015](http://www.nature.com/nature/journal/v523/n7561/extref/nature14590-s1.pdf) paper:

- Using the filtered peak set, peak **summits** were extended +/-250 bps. The top 50,000 non-overlapping 500bp summits (by log10qvalue), which we refer to as accessibility peaks were used for all downstream analysis.

Get non-overlapping peaks (all or top 100,000):

1) script: getTopPeaks.pl

2) command: sort -k5n consensus\_narrow\_summits.noverlap.bed | tail -n 100000 | sort -k1,1 -k2,2n >consensus.narrow.noverlap.top.sorted.peaks.bed

Get the raw counts per sample:

bedtools multicov -q 30 -p -bams fgwatd-1.nodup.nomito.bam fgwatd0.nodup.nomito.bam fgwatt4.nodup.nomito.bam fgwatd1.nodup.nomito.bam fgwatd2.nodup.nomito.bam fgwatd4.nodup.nomito.bam fgwatd7.nodup.nomito.bam fmgwat.nodup.nomito.bam fswatd-1.nodup.nomito.bam fswatd0.nodup.nomito.bam fswatt4.nodup.nomito.bam fswatd1.nodup.nomito.bam fswatd2.nodup.nomito.bam fswatd4.nodup.nomito.bam fswatd7.nodup.nomito.bam fmswat.nodup.nomito.bam mgwatd-1.nodup.nomito.bam mgwatd0.nodup.nomito.bam mgwatt4.nodup.nomito.bam mgwatd1.nodup.nomito.bam mgwatd2.nodup.nomito.bam mgwatd4.nodup.nomito.bam mgwatd7.nodup.nomito.bam mmgwat.nodup.nomito.bam mswatd-1.nodup.nomito.bam mswatd0.nodup.nomito.bam mswatt4.nodup.nomito.bam mswatd1.nodup.nomito.bam mswatd2.nodup.nomito.bam mswatd4.nodup.nomito.bam mswatd7.nodup.nomito.bam mmswat.nodup.nomito.bam -bed /NGS/working\_projects/AtacSeq/data/11\_peaks/narrow/consensus/consensus\_narrow\_summits.noverlap.bed >/NGS/working\_projects/AtacSeq/data/11\_peaks/narrow/consensus/all\_peaks\_raw\_counts.txt

add header:

sed -i '1i

chr\tstart\tend\tid\tscore\t

fgwatd-1\tfgwatd0\tfgwatt4\tfgwatd1\tfgwatd2\tfgwatd4\tfgwatd7\tfmgwat\tfswatd-1\t

fswatd0\tfswatt4\tfswatd1\tfswatd2\tfswatd4\tfswatd7\tfmswat\tmgwatd-1\t

mgwatd0\tmgwatt4\tmgwatd1\tmgwatd2\tmgwatd4\tmgwatd7\tmmgwat\tmswatd-1\t

mswatd0\tmswatt4\tmswatd1\tmswatd2\tmswatd4\tmswatd7\tmmswat\t' counts.txt

Normalise the raw counts by number of properly mapped reads (diffbind normalisation more sophisticated)

You need sequencing\_depth.txt from the output erlier filtering

For each value x in sample y: x/(number of properly mapped reads in y/1,0000,000) = RPM

Get the log2 of each RPM for easier comparison

**DOWNSTREAM ANALYSIS**

**Differential Binding**

Principle idea:

1. Call peaks with macs2

2. Use the narrow peak file, which has the summit positions and extend the summit on either sides by 250bp

3. Count the reads for each replicate

4. Feed the matrix to DESeq2

Scripts: grid\_diffbind\_master.pl (see script for explanation of output)

processDiffReports.pl

diffHeat.r

Creating a master report: (not done)

cat \*report.txt >master\_report.txt

sed 's/\"//g' master\_report.txt > noquotes.txt

head -n 1 noquotes.txt >master\_report.txt

tail -n +2 noquotes.txt | sort -k1,1 -k2,2n >>master\_report.txt

Useful links:

[2016](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4857123/), [2017](https://elifesciences.org/content/6/e21883" \l "SD24-data), [2017](http://www.nature.com/ni/journal/vaop/ncurrent/full/ni.3683.html" \l "supplementary-information) (Papers using diffbind on ATAC-seq data)

[don't use bdgdiff](https://github.com/jknightlab/ATACseq_pipeline/blob/master/macs2_diff/macs2_diff_workflow.md)

[diffbind counts v macs2 scores](https://support.bioconductor.org/p/67486/)

[normalisation of dba.count scores](https://support.bioconductor.org/p/67367/)

[oxford academic tool compare](https://academic.oup.com/bib/article/17/6/953/2453197/A-comprehensive-comparison-of-tools-for" \l "47712114)

[using diffbind on broad peaks](https://support.bioconductor.org/p/76216/)

[explains why report widths column is not uniform](https://support.bioconductor.org/p/95844/)

DESEQ2 without replicates

“Honestly, I don't think it's much use to run DESeq2 without replicates. The software allows for this analysis with large caveats (see the relevant section in results) that such an analysis is only vaguely exploratory. But you could just as well plot the log ratios of peak counts across conditions and skip running DESeq2.” - Micheal Love

**NUCLEOSOME POSITIONING AND OCCUPANCY ANALYSIS:**

Nucleoatac implements a mixture model based on the overall fragment length distribution of a given sample. The occupancy measures how likely a given fragment is to come from nucleosome-free (NFR) DNA or nucleosome-protected DNA. This metric is normalized between 0 and 1. What this prediction is doing is identifying gaps between two called nucleosome positions with low nucleosome occupancy. This will miss nfrs where there isn't a sufficiently strong signal from a flanking nucleosome to estimate occupancy accurately. It is important to remember that lack of nucleosome signal or call on its own does not imply lack of nucleosome; without sufficient reads (i.e. sufficient accessibility at the linkers) NucleoATAC won't be able to detect a nucleosome. This becomes especially important when considering a set of peaks that is a merge of several peak sets where there may be some peaks with very low coverage in some of the samples.

Useful links: [readme](http://nucleoatac.readthedocs.io/en/latest/), [ATAC-seq forum](https://sites.google.com/site/atacseqpublic/home?pli=1.+T) (needs registration), [FAQs](https://github.com/GreenleafLab/NucleoATAC/issues?utf8=✓&q=)

[recommended sequencing depth](https://github.com/GreenleafLab/NucleoATAC/issues/50), [optimizing runtime](https://github.com/GreenleafLab/NucleoATAC/issues/43)

[2016](https://elifesciences.org/content/5/e20148) and [2017](http://www.nature.com/nature/journal/v541/n7636/full/nature20781.html" \l "methods) papers using nucleoatac

[using a consensus peak set](https://github.com/GreenleafLab/NucleoATAC/issues/56)

[original nucleoatac paper](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4617971/)

PWM file:

The default Human.PWM.txt file should be pretty similar for any species as it is simply reflecting sequence preferences of nextera Tn5 on naked dna. The ones included are from Human and Yeast genomic dna samples and are both very similar. If not using nextera Tn5 (i.e. using some other transposase) then a new pwm should be created based on a genomic dna sample.

Required files:

1. sorted + indexed BAM files (.bam + .bam.bai)

2. indexed reference genome fasta file (.fa + .faidx)

3. processed broadPeak file (see below)

Processing the broadPeak file:

1. must leave the file as position sorted

2. may need bed file format (first 6 columns only)

3. Extend peaks by 100bps in both directions to capture nucleosomes flanking NFRs making nucleoasome calling easier () especially if sequencing is not very deep. ([advice from authors](https://github.com/GreenleafLab/NucleoATAC/issues/46))

bedtools slop -i broadPeakfile -g mm10\_chr\_size.txt -l 100 -r 100 -s >extended.broadPeak

4. Regions should not overlap so it is advisable to use bedtools merge on these regions

bedtools merge -i extended.broadPeak > final.broadPeak

Script: grid\_nucleoatac.pl

Step 1 OCC OUTFILES: Computing Occupancy and Nucleosomal Insert Distribution

- occ.bdg: Bedgraph track with nucleosome occupancy score using the fragment sizes

This is also a low-resolution estimate of nucleosome position

Range 0-1 so easier to compare across loci (doesn’t depend on read depth)

NB: Unreliable outside open chromatin regions (low signal:noise ratio)

Signal track more gives more precise nucleosome position

- occ.lower\_bound: Bedgraph track with lower bound estimate for nucleosome occupancy score

- occ.upper\_bound: Bedgraph track with upper bound estimate nucleosome occupancy score

Upper and lower bound are estimates of the upper and lower limits in nucleosome occupancy based on the number of reads but without accounting for systematic biases. If upper and lower are far apart this indicates not enough reads at that locus to accurately predict occupancy.

- nuc\_dist.txt/.eps: Text file and EPS plot with estimate of fragment size at nucleosomes

peaks in 140-200bp are from DNA protected by 1 nucleosome

peaks in >200by are from DNA protected by >1 nucleosome

peaks in <50bp are either a)TF or b) nucleosome protected DNA

- fragmentsizes.txt: Text file with fragment size distribution within input peaks

- occ\_fit.txt/.eps: Text file and EPS plot of model for NFR and nucleosomal distributions

Green are red are models, Blue is observed, Orange is smoothed fit

Waves are due to double helix DNA structure. [explained in detail](https://github.com/GreenleafLab/NucleoATAC/issues/18)

- occpeaks.bed: Peaks from nucleosome occupancy track -- low resolution nucleosome calls. (1) chrom (2) dyad position (0-based) (3) dyad position (1-based)

(4) occupancy (5) occupancy lower bound (6) occupancy upper bound

(7) # of reads in 121 bp window

Step 2 VPROCESS OUTFILES: Processing Vplot - - - Vmat + Vmat.eps: Text file and EPS plot with normalized V-plot for cross-correlation

Step 3 NUC OUTFILES: Obtaining nucleosome signal and calling position

- nucpos.bed: Nucleosome dyad calls text file

(1) chrom (2) dyad position (0-based) (3) dyad position(1-based)

(4) z-score (5) nucleosome occupancy estimate

(6) lower bound for nucleosome occupancy estimate

(7) upper bound for nucleosome occupancy estimate

(8) log likelihood ratio (9) normalied nucleoatac signal value

(10) cross-correlation signal value before normalization

(11) number of potentially nucleosome-sized fragments

(12) number of fragments smaller than nucleosome sized

(13) "fuzziness" (measure of how wide signal peak is)

- nucpos.redundant.bed: Includes nucleosome position calls that were within the minimum separation for non-redundant calls

- signal.(smooth).bdg: (Smoothed) Bedgraph track with normalized cross-correlation signal

The smoothed track includes only positive signal

This is a high-resolution estimate of nucleosome position but depends on coverage

so is not a good measure of occupancy (see occ.bdg)

Tall peak indicates canonical nucleosome position

Small oval peaks represents less consistent nucleosomes position

- occ.bdg: Bedgraph track with nucleosome occupancy score

Step 4 MERGE OUTFILES: Making combined nucleosome position map

- nucmap\_combined.bed: Combines low resolution nucleosome calls from occ (occ.bdg?) and higher

resolution calls from nuc (occ.bdg?) to create more comprehensive map

Step 5 NFR OUTFILES: Calling NFR position

- nfrpos.bed: NFR position (1) chrom, (2) left boundary (0-based),

(3) right boundary (1-based), (4) mean occupancy,

(5) minimum upper bound occupancy, (6) insertion density, (7) bias density

Other OUTFILES:

- ins.bdg: A tabix-indexed bedgraph file with insertion densities (subcommand pyatac ins)

Convert bed files for viewing on UCSC track: nucleoatacToBigBed.pl

Convert bedgraphs to bigwigs for viewing in IGV:

* for i in \*bedgraph; do bedtools slop -i $i -g /NGS/musRefs\_10/mm10\_chr\_size.txt -b 0 | bedClip stdin /NGS/musRefs\_10/mm10\_chr\_size.txt $i.clip; done;
* for i in \*clip; do bedGraphToBigWig $i /NGS/musRefs\_10/mm10\_chr\_size.txt $i.bw; done;

Read into R: [nucleoatac.r](https://github.com/GreenleafLab/NucleoATACR/)

Peak annotation

Using chipPeakAnno: [chipPeakAnno](https://www.bioconductor.org/packages/release/bioc/html/ChIPpeakAnno.html), [Manual](https://www.bioconductor.org/packages/release/bioc/manuals/ChIPpeakAnno/man/ChIPpeakAnno.pdf), [User Guide](https://bioconductor.org/packages/release/bioc/vignettes/ChIPpeakAnno/inst/doc/ChIPpeakAnno.html), [Workflow](https://www.bioconductor.org/packages/devel/bioc/vignettes/ChIPpeakAnno/inst/doc/pipeline.html), [alt\_Workflow](http://www.bioconductor.org/help/workflows/chipseqDB/" \l "using-the-chippeakanno-package)

Scripts: chipPeakAnno\_MIN.r

Tissue Specificity

The TisSpec algorithm was used to determine highly tissue specific genes of the gonadal fat pad from [ENCODE](https://www.encodeproject.org/experiments/ENCSR000BYZ/) expression data on the assumption that it is also GWAT. The gene set was cross-matched with genes with significantly different peaks at promoter regions between gp\_SWAT and gp\_GWAT

From PMID 21427703 – Sams paper

- limited correlation between chromatin marks and chromatin peaks

- a peak can occur well before the TF for it is transcribed

- peaks can precede transcriptional activation of genes by days and remain long after expression is reduced

- 25000 active marks (ac27) do not have a corresponding peak

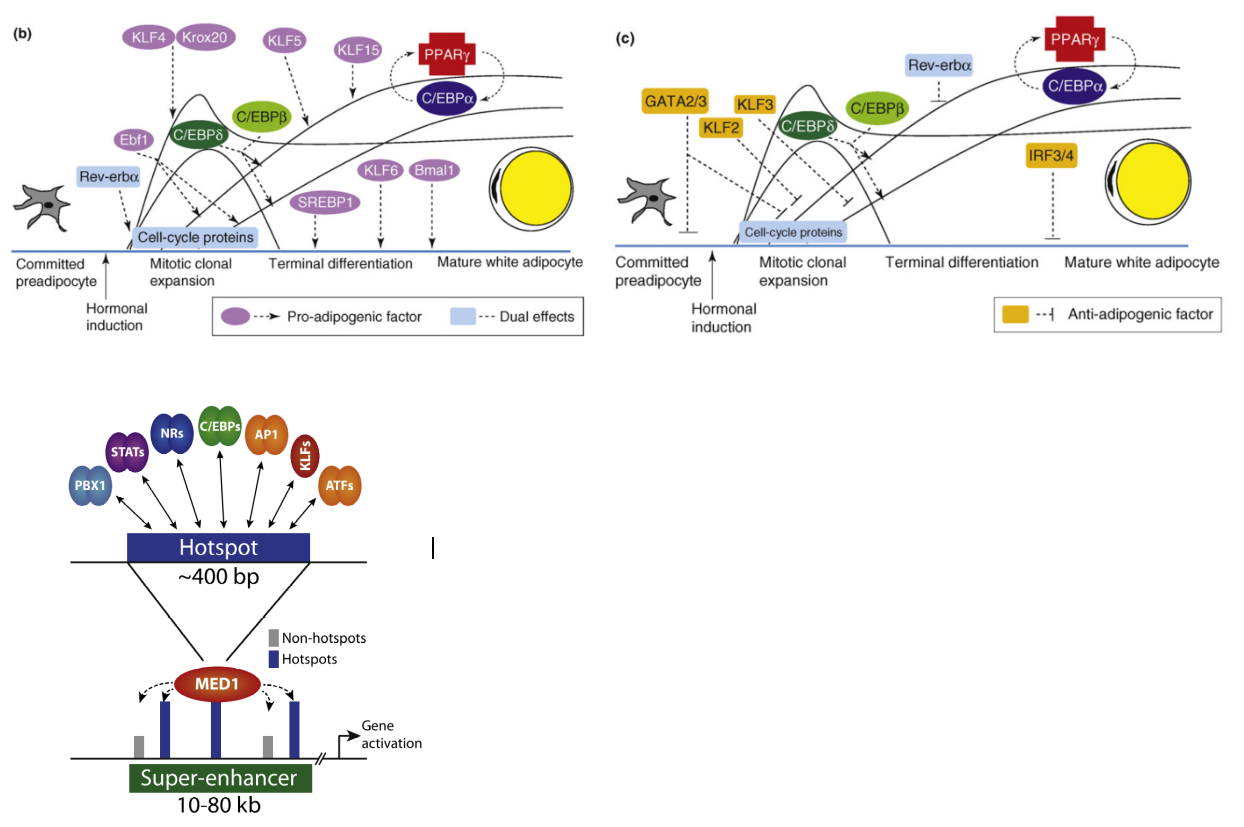
Perl script: diffspec.pl

GENES OF INTEREST:

* TBX15 BED: this is the region that’s associated with body fat distribution, it’s in the intron of Tbx15 at chr3:99,279,503-99,280,343 TBX15\_fine\_mapped\_region. Promoters of interest are Wars2, Tbx15, Spag17,
* FTO BED: this is the region that corresponds to human rs1421085. This is the sequence we deleted in mouse, also in an regulatory region in the intron of Fto at chr8:91,374,370-91,374,451 FTO\_82bp\_deletion. Promoters of interest are Rpgrip, Fto, Irx3, Irx5,

CONTROLS + Important adipogenesis genes: (PMID 19269847)

* Tbx15 – should only be in SWAT
* Fabp4 (r) – induced late in adipogenesis
* Zfp423 – a marker of adipocytes
* PdgfRa – a marker of preadipocytes only
* Plin1 – (aka perilipin) a marker of mature adipocytes
* Adipoq – (aka adiponectin) a marker of mature adipocytes
* Wnt10b – inhibits adipogenesis (should not be expressed)
* KLF4+5 – transactivates Cebpg (should be expressed early)
* Arntl – clock gene required for adipogenesis
* Pparg – a marker of adipogenesis (absent in mature) – highly tissues specific to adipocytes
* Cebpg – a marker of adipocytes (absent in mature)
* Gata2+3 – initiates terminal differentiation (should be expressed)
* Prdm16 – initiates differentiation of BAT (should not be expressed in WAT)
* Fbox1 – should not be expressed at all?
* Prkag1 – metabolic housekeeping gene – should be expressed everywhere
* Gapdh – housekeeping gene – should be expressed everywhere
* Canx – housekeeping gene – should be expressed everywhere
* Ubc – housekeeping gene – should be expressed everywhere



Super-enhancers (SEs):

10-80kb

Highly enriched near early induced genes (4hrs)

Much more Med1 than in normal enhancers

Also contains motifs for the ESC master TFs Oct4, Sox2 and Nanog

Enriched with motifs for lineage determining TFs (LDTFs) (i.e. Pparg, Cebp)

GSEA of genes associated with SEs → should return biological processes that describe the cell type

At 4hrs find clusters of Med1 AND open chromatin

Hotspots:

Locations in superenhancers highly enriched for...

- At least 5 of 15 hotspot defining genes

- Enriched in Klf4 and Esrrb binding sites

- ME1/AC27 histone marks than normal enhancers

- Contains Pparg:RXR heterodimer response element motif

- Nucleosome remodelling → widening of NFRs (nucleoatac)

Targets:

93% occur within same topological domain

95% overlap with their associated superenhancer

Most targets are genes that encode transcription factors

Higher expression than targets of normal enhancers

Lower xpression during differentiation

Interesting Finds

* Missing Gene: chr11:83,263,497-83,298,341 chromhmm shows promoter, atac shows open chromatin – look for ORFs? Is it actually an enhancer not a promoter? (PMID 19269847)

IDEAS

* Nucleosome remodelling: Determine the chromatin landscape of Pparg and Cebp and find out if LDTFs (e.g. CEBP, Pparg) + SDTF binding result in widening of NFRs? (PMID 25650801). Widening may also be due to CTCF or NF-y binding (PMID 25565413). Also Can you show that nucleosome remodelling during differentiation is enriched at enhancers associated with pluripotency and differentiation? (pmid 25158628, 20208536) - [http://cistrome.org/db/#/](http://cistrome.org/db/" \l "/) → view chip data in UCSC
* Do the master regulators Pparg and Cebp target similar sets of genes as other factors in adipogenesis – if so - are the effects agonistic or antagonistic? (PMID 19269847)
* Superenhancers (10-80kb) have a large amount of MED1 binding and are highly enriched nearly near early induced genes (4hrs). They should be enriched with motifs for LDTFs (Pparg, Cebp in pre-adipocytes). Identify the collection of super-enhancers of MSWAT: 1 - identify enriched motifs in open chromatin at 4hrs. 2 - GSEA the motifs. 3 - Map the motif positions on the genome. 4 – Identify superenhancers by clusters of MED1. 3 – GSEA the clusters. 5 - Identify hotspots in superenhancers by internal clusters of motifs for at least 5 of the 15 genes in PMID 24857652 – include MED1, PolII, Pparg and C/Ebp? Is the Pparg:RXR heterodimer response element present (PMIDs 18981474, 19269847)? Examine for widening of NFRs within hotspots. Repeat same for FSWAT. Repeat same for M/FGWAT. What are the differences? What is common? Do these superenhancer sites match those found in pmid 24857652. If so, can the same method be applied to SCN data? Use <http://cistrome.org/BETA/index.html> to identify the targets of the super-enhancers
* K means clustering of peaks during adipogenesis to get temporal profiles (PMID 21427703) – may need grouping of time points for replicates (-1 with 0, 4h+1,2+4,7+m) – the biggest jump is in the first 4 hours (PMID 21427703)
* Run motif analysis with meme on the 6000 most intense peaks induced at 4h (PMID 19269847) – should show enrichment of motif for the “hotspot” genes: RXR, C/Ebpb, C/Ebpq, Nr3c1 (aka: GR) and Stat5a motifs
* Recent ATAC paper: mendelley → “layer specific chromatin accessability”
* [footprinting](https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-7-33)
  + centipede <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3959825/pdf/nihms554473.pdf>

<https://github.com/slowkow/CENTIPEDE.tutorial>

<https://rajanil.github.io/msCentipede/>

<https://www.biostars.org/p/234906/>

* + meme
* [combining atac and chipseq](https://www.biostars.org/p/221826/) - Overlap with existing ChIP-seq datasets
* [CRE review with good ideas](http://onlinelibrary.wiley.com/doi/10.1002/wsbm.1374/full)
* TFFM for motif discovery: [moses lab blog](http://www.moseslab.csb.utoronto.ca/blog/?p=40), [R package](https://academic.oup.com/bioinformatics/article/32/10/1555/1743236/TFBSTools-an-R-bioconductor-package-for), mentioned in this [paper](http://onlinelibrary.wiley.com/doi/10.1002/wsbm.1374/full)
* ATAC omics tools
* Frp, Tchkonia et al 2006 (identification of depot-specific…) - Get the list of developmental genes whose expression differentiates between depots →> Get the list of peaks that target these genes → are there significant differences between the peaks of different depots

