Week 3 – Thursday – CHIP-seq

Today we’re going to go through background of CHIPseq

Then we will read through someone elses pipeline and see if we can understand it

Background:

Crosslink proteins to dna, lyse cells, pulldown, enrich for protein of interest, dna then sequence (illumine) ie. Transcription factors or peaks – ie chromatin marks that will give you info on functional state

Andrea has a SNP within a promoter region, think it influences TF binding – but she doesn’t know the TF. Jasper (based on motif in seq) – detect anything that binds the motif.

Can provide 1-2000 bp for promoter region

Bias: CHIP can be tricky on small/inconstant numbers of cells.

Library fragmentation – (sonication -biases open chrom)

Tagmentation (inset seq into genome into genome – sequence bias)

Nucleosomes -mnase

Antibody that you use has to be specific – use several antibodies for epitope covering etc

Controls

Antibody, no chrom and vv – but what’s the point in sequencing

Input control – chromatin not immunoprecipitated – fragmentation bias is covered here. So you can get same overrepresented places where fragmentation has happened so more reads. But you need high sequencing as you’re covering a lot of the genome

Lost internet.

Normalise across conditions (spike in) – if everything’s same pattern but lower in one condion than another – specific to experimental design

Need to just identifiy where in the genome the reads are coming from. They use PE as better for duplicate removal – shorter read length

Greater for input. – depth depends on your experiment

Encode symposium – give guidelines based on their data

Workflow

fastq – fastqc

mapping – going to use bowtie2 – doesn’t take splicing into account

mapping qc – samtools/picard

peak calling – including macs2

QC of peak calling – Chipseeker is R so probably will use something else

Motif analysis – Homer/beam

Peak annotation – chip seeker-tels you where nearest gene is, whether peaks fall into gene/enhancer more than you’d expect

DEseq2 – DGE for differential binding

MAPPing qc

Sequencing of same fragment multiple times – might see a peak where there isn’t one – remove duplicates

Q – mitochondrial duplicates

Reads that are not properly paired – doesn’t make sense basically

Peak calling is basically seeing the pileup of reads – more than control in certain regions (and more than surrounding region)

Pooling samples – increase calling ability – look at overall properties of dataset.

Downsampling can help with this – but can be complicated

Impossible to normalised based on number of reads? Can try different ways –

MACS2 – industry standard – we will use this one today.

Check reproducibility – ie bedtools, overlap peaks between replicates

Encode project – irrep discovery rate – correlation of ranks between replicates – where does this fall away? Throw away weaker peaks. Controversional – only works with 2 replicates. David Sims doesn’t like it

ChipSeq – black listed – things that just always come up

Peak annotation – homer/great

Nearest neighbour. Promoters are easy to – but enhancers can be a long way from target genes so who knows. Caveeats of this – ie closest gene to chipseq peak may not be correct

Counting is same as DGE

How many reads overlap peaks

Want 3 replicates at least – diffbind – provides a wrapper – but limited with design of DGE comparison

Motif analysis – online interfaces, running on CL

Meme suite and homer good for this. – take sequences under peaks (ie under summit of peak) program will look for repeated kmer and will try and find TF motifs – tells you where within the peak the motifs are – association and cobinding of different TF factors

Viewing peaks – should just look at data first, just initially look at bam peaks, memory intensive etc. Convert bam to bed coverage/bigwig – breaks itno region and gives a coverage summary

Bigwig – compressed/indexed – less memory intensive

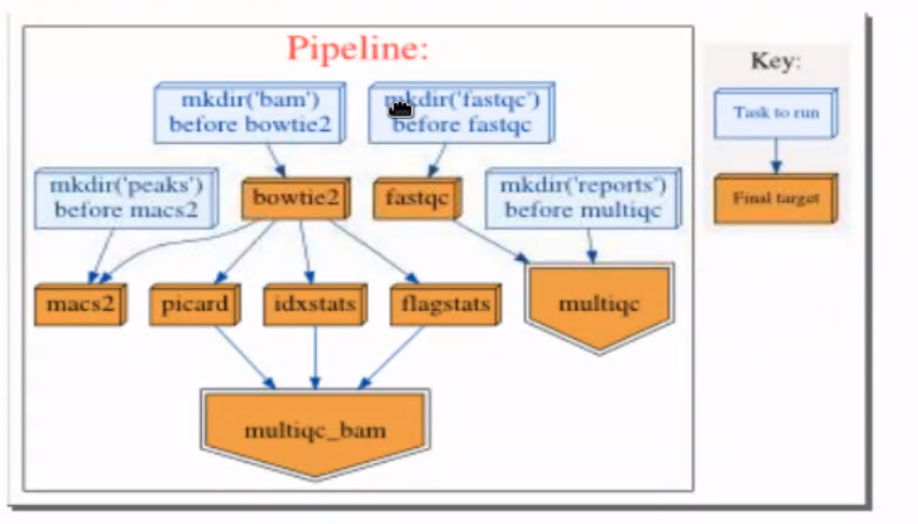
Exercise: we are going to now write the pipeline. Dataset is internal. All have an input control. We have them on the sever

Goes as far as peak calling – git pull the file off Github – chipseq\_pipeline

I’ve started to annotate this now

Shared week 3 – data

We’re going to look at the plot first.

Fastqc/multiqc

Mapping then QC – picard. Idx and flag stats. Draws thm all into multiqc

Then goes to macs2

Add remove duplicate step

Add Filtering – get rid of bad reads

Add merge step – merge controls and treatments

Fell over on macs2 – our py3 conda hasn’t been added into our environemt so I did conda install macs2

Also conda install homer

#Adding a deduplication before peakcalling – before macs2

Running just the multiqc\_bam   
python script multiqc\_bam -v 5

Picard

Samtools idxstats chr x or y tells you whether samples are male or female.

Duplicates – we need to get rid of some of it

Duplicates are reads that map in same coordinates – both pairs map to same cooridinates as another pair. PCR – amplification bias. Throw em away as assume they’re coming from same piece of genomic DNA. UMI are only real way to see if they come from an independent DNA molecule

All of mapping QC gets done at same time – but would be nice to add this info into the multiQC report.

MarkDuplicates -

Filter stage

Post duplicate removal – filter the bam file. – get rid of unpaired things, secondary alignment, mitochondrial reads

<http://broadinstitute.github.io/picard/explain-flags.html>

sam flags page

<https://davetang.org/wiki/tiki-index.php?page=SAMTools>

Had to do the bam indexing manually here – the pipeline didn’t work for some reason.

Currently the macs2 has been done on the bowtie2 data,

Initially after bowtie that we also do the remove duplicate, and filtering on this then

So we’ve done the macs (peak calling) on the output of the bowtie2- which is the alignment (without removing the duplicates or doing any of the filtering)

* Should we also do the broad peaks?
* Filter blacklisted peaks (Bedtools)
* •Count reads under peaks (Bedtools)
* •Compare replicate overlap (Bedtools)
* •Annotate peaks (Homer)
* •Motif analysis (Homer)
* •View outputs (IGV)

Ok we first need to filter the blacklisted peaks

Blacklist regions – gives you a whitelisted list – can do before or after – better to do before as you don’t want the blacklisted regions to affect how the model looks in the end

ls -lh \*filter\_read\* #shows you these files and memory etc

first going to count peaks under the reads using bedtools

I’m not sure about the homer stuff – we had to do which homer on command line to find where the genome thing is located

Basically need the full path to run homer.

annotatePeaks.pl peaks/rep1\_dex\_peaks.narrowPeak mm10