ChIP-seq Analysis

BaRC Hot Topics - Feb 23th 2016

Bioinformatics and Research Computing

Whitehead Institute

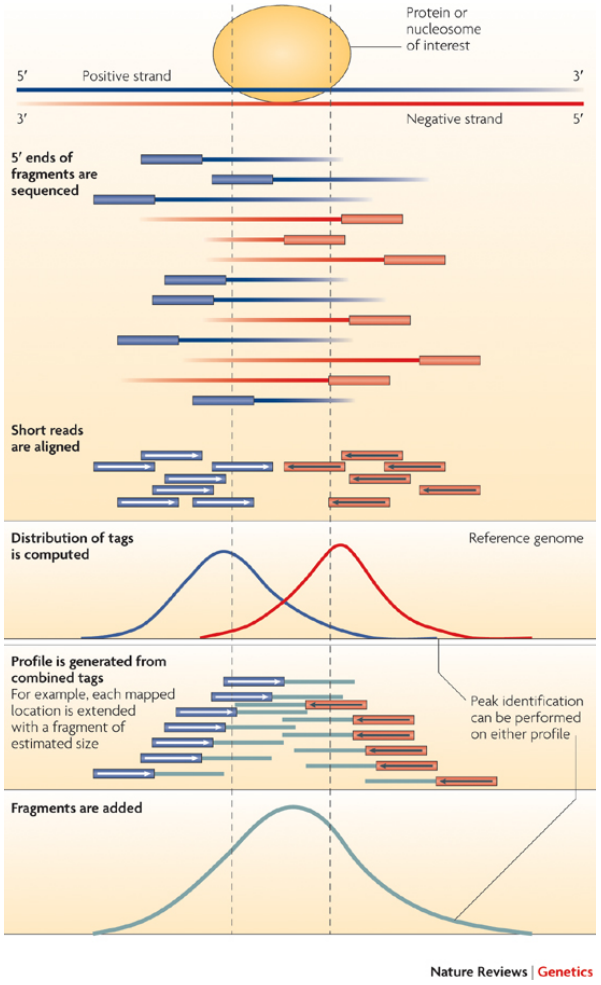
http://barc.wi.mit.edu/hot_topics/

Whitehead Institute

Outline

- ChIP-seq overview
- Experimental design
- Quality control/preprocessing of the reads
- Mapping
  - Map reads
  - Remove unmapped reads (optional) and convert to bam files
  - Check the profile of the mapped reads (strand cross-correlation analysis)
- Peak calling
- Linking peaks to genes
- Visualizing ChIP-seq data with ngsplot

ChIP-Seq overview



Steps in data analysis

1. Quality control

2. Mapping
   Treat IP and control the same way (preprocessing and mapping)
3. Peak calling
   i) Read extension and signal profile generation
   ii) Peak assignment
4. Peak analysis and interpretation
   i) Find genes next to peaks
   ii) Infer possible biological consequences of the binding

Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

Experimental design

- Include a control sample.
- If the protein of interest binds to repetitive regions, using paired–end sequencing may reduce the mapping ambiguity. Otherwise single reads should be fine.
- Include at least two biological replicates. If you have replicates you may want to use the parameter IDR *"irreproducible discovery rate"*. See us for details.
- If only a small percentage of the reads maps to the genome, you may have to troubleshoot your ChIP protocol.

Illumina data format

- Fastq format:

  http://en.wikipedia.org/wiki/FASTQ_format

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh
```

/1 or /2 paired-end

→ @seq identifier

→ seq

→ +any description

→ seq quality values

Input qualities	Illumina versions
--solexa-quals	<= 1.2
--phred64	1.3-1.7
--phred33	>= 1.8

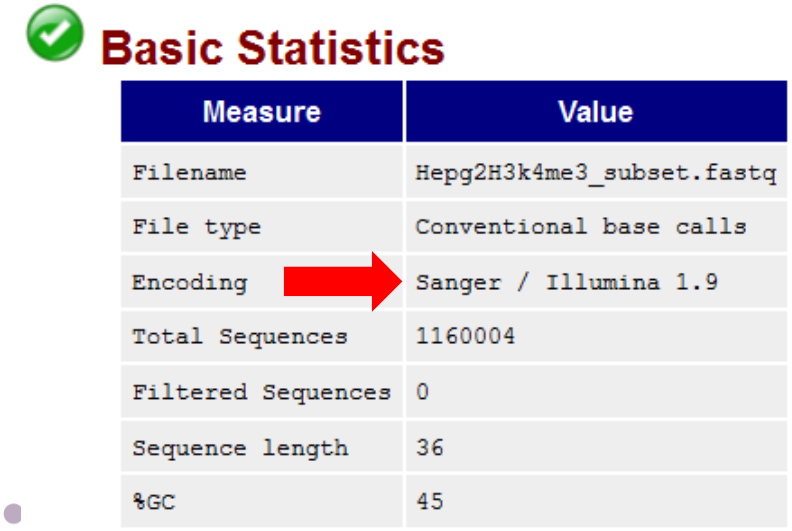
Check read quality with fastqc

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

1. Run fastqc to check read quality

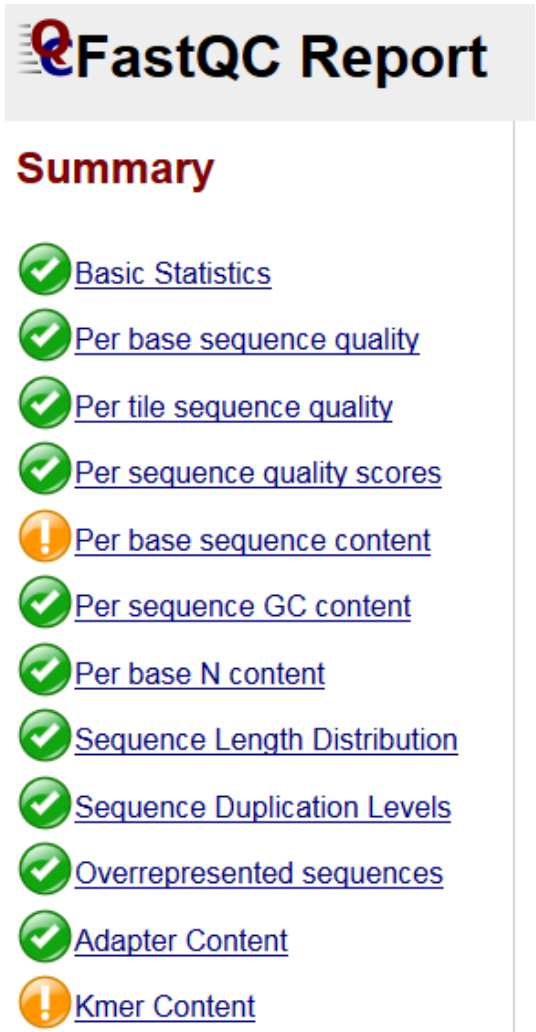
   bsub fastqc sample.fastq

2. Open output file: "fastqc_report.html"

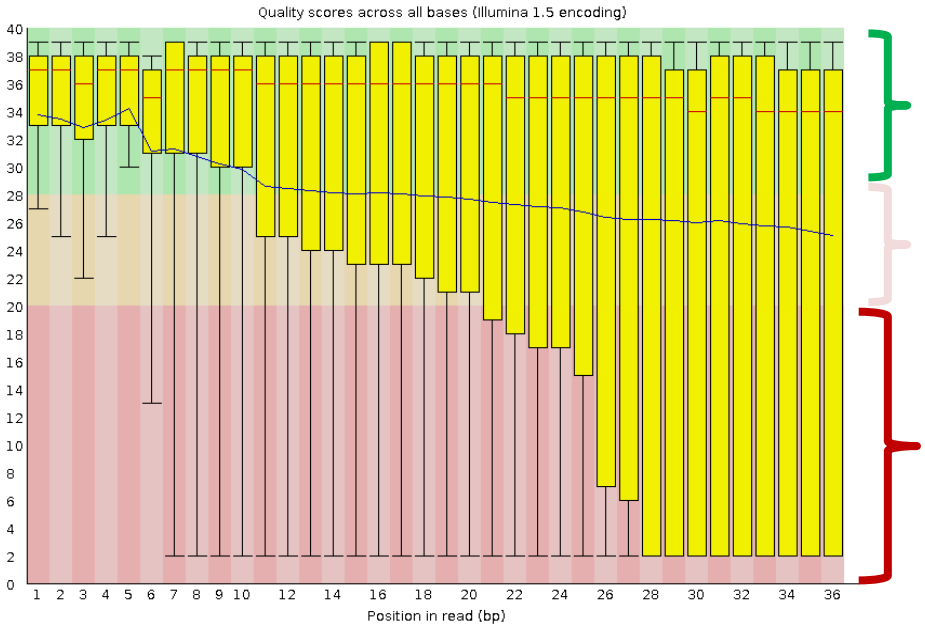


Basic Statistics

Measure	Value
Filename	Hepg2H3k4me3_subset.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1160004
Filtered Sequences	0
Sequence length	36
%GC	45



FastQC: per base sequence quality



Red: median
Blue: mean
Yellow: 25%, 75%
Whiskers: 10%, 90%

Quality value = −10 * log10 (error probability)
Quality = 10 => error rate = 10% => base call has 90% confidence
Quality = 20 => error rate = 1% => base call has 99% confidence
Quality = 30 => error rate = 0.1% => base call has 99.9% confidence

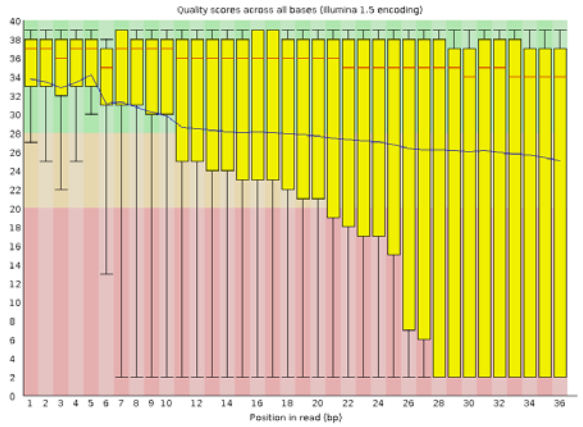
Preprocessing tools

- **Fastx Toolkit** (http://hannonlab.cshl.edu/fastx_toolkit/)
  - FASTQ/A Trimmer: Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).
  - FASTQ Quality Filter: Filters sequences based on quality
  - FASTQ Quality Trimmer: Trims (cuts) sequences based on quality
  - FASTQ Masker: Masks nucleotides with 'N' (or other character) based on quality

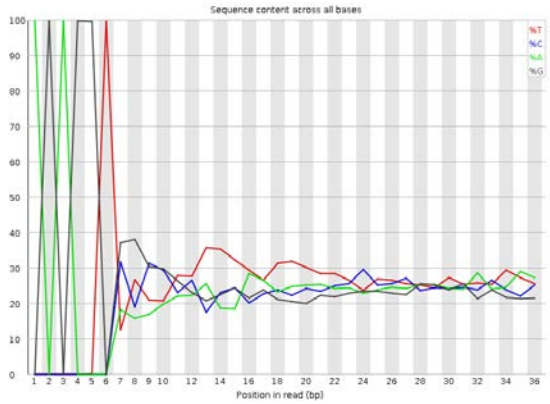
  (for a complete list go to the link above)
- **cutadapt** to remove adapters

  (https://code.google.com/p/cutadapt/)

What preprocessing do we need?



Bad quality -> Use "FASTQ Quality Filter" and/or "FASTQ Quality Trimmer"



Flagged Kmer Content: About 100% of the first six bases are the same sequence -> Use "FASTQTrimmer"

Sequence	Count	Percentage	Possible Source
TGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTTAGGCA	7360116	82.88507591015895	RNA PCR Primer, Index 3 (100% over 40bp)
GCGAGTGCGGTAGAGGGTAGTGGAATTCTCGGGTGCCAAG	541189	6.094535921273932	No Hit
TCGAATTGCCTTTGGGACTGCGAGGCTTTGAGGACGGAAG	291330	3.2807783416601866	No Hit
CCTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACTTAGG	210051	2.365464495397192	RNA PCR Primer, Index 3 (100% over 38bp)

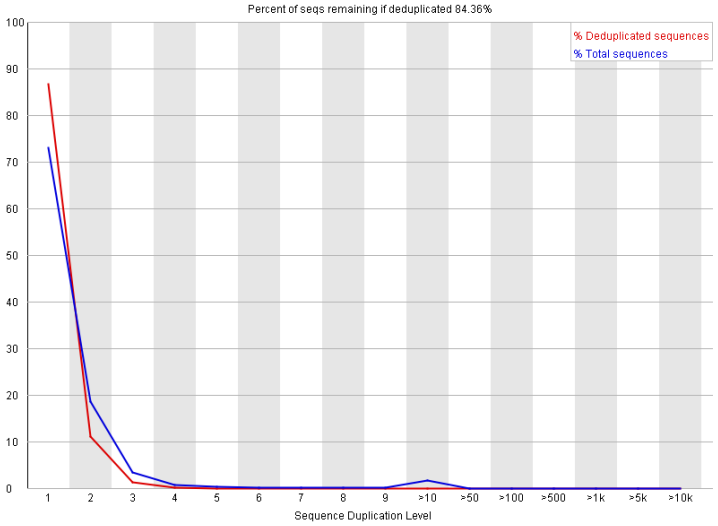
Overrepresented sequences -> If the over represented sequence is an adapter use "cutadapt"

Recommendation for preprocessing

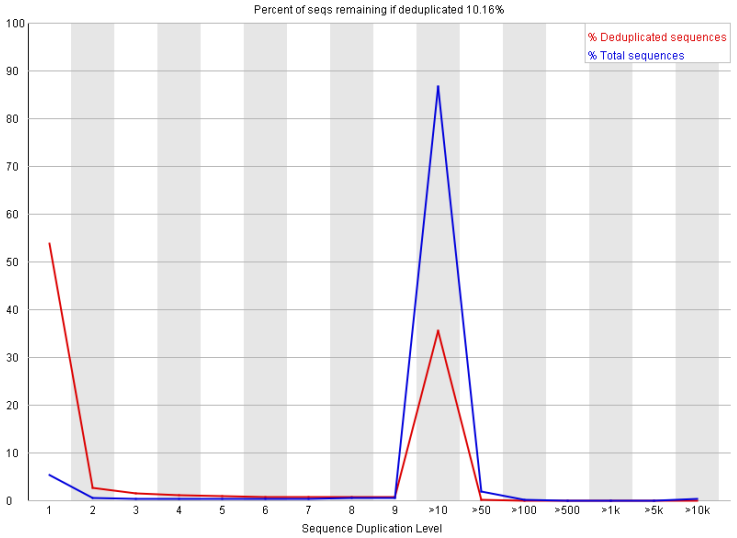
- Treat IP and control samples the same way during preprocessing and mapping.
- Watch out for preprocessing that may result in very different read length in the different samples as that can affect mapping.
- If you have paired-end reads, make sure you still have both reads of the pair after the processing is done.
- Run fastqc on the processed samples to see if the problem has been removed.

Library Complexity

- Refers to the fraction of unique fragments present in a given library.
- One proxy for library complexity is to look at the sequence duplication levels on the FastQC report:



% Complexity
85.6



% Complexity
4.95

Library Complexity II

- Low library complexity may be an indicator that:
  - A new sample and a new library should be prepared.
  - We have to find a better Ab to perform the IP.
  - We can not sequence the same sample anymore because we will not find new sequences.
- In certain experimental settings we may expect a low library complexity. *i.e.* We are profiling a protein that binds to a small subset of the genome.

Mapping
Non-spliced alignment software

- Bowtie:
  - bowtie 1 vs **bowtie 2**
    - For reads >50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory than Bowtie 1.
    - Bowtie 2 supports gapped alignment, it makes it better for snp calling. Bowtie 1 only finds ungapped alignments.
    - Bowtie 2 supports a "local" alignment mode, in addition to the “end-to-end" alignment mode supported by bowtie1. However *we don’t recommend "local" alignment mode for mapping of ChIP-seq data*.
- BWA:
  - refer to the BaRC SOP for detailed information

Local genomic files needed for mapping

tak: /nfs/genomes/

- Human, mouse, zebrafish, *C.elegans*, fly, yeast, etc.
- Different genome builds
  - mm9: mouse_gp_jul_07
  - mm10: mouse_mm10_dec_11
- human_gp_feb_09 vs human_gp_feb_09_no_random?
  - human_gp_feb_09 includes *_random.fa, *hap*.fa, etc.
- Sub directories:
  - bowtie
    - Bowtie1: *.ebwt
    - Bowtie2: *.bt2
  - fasta: one file per chromosome
  - fasta_whole_genome: all sequences in one file
  - gtf: gene models from Refseq, Ensembl, etc.

Example commands:
Mapping the reads and removing unmapped reads

```
bsub bowtie2 --phred33-quals -N 1 -x
/nfs/genomes/human_gp_feb_09_no_random/bowtie/hg19   -U
Hepg2Control_subset.fastq -S
Hepg2Control_subset_hg19.N1.sam
```

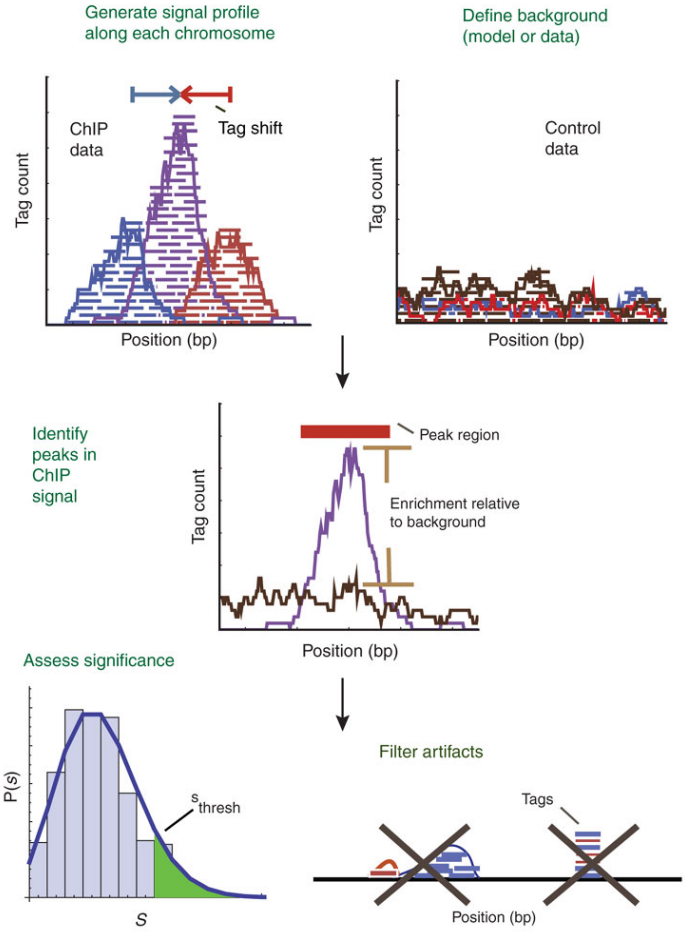
Optional: filter reads mapped by quality mapping score

```
samtools view -bq 10 file.bam > filtered.bam
```

Peak calling

i) Read extension and signal profile generation
- strand cross-correlation can be used to calculate fragment length

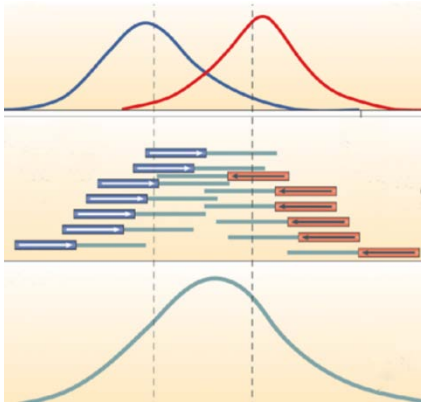
ii) Peak evaluation
- Look for fold enrichment of the sample over input or expected background
- Estimate the significance of the fold enrichment using:
  - Poisson distribution
  - negative binomial distribution
  - background distribution from input DNA
  - model background data to adjust for local variation (MACS)

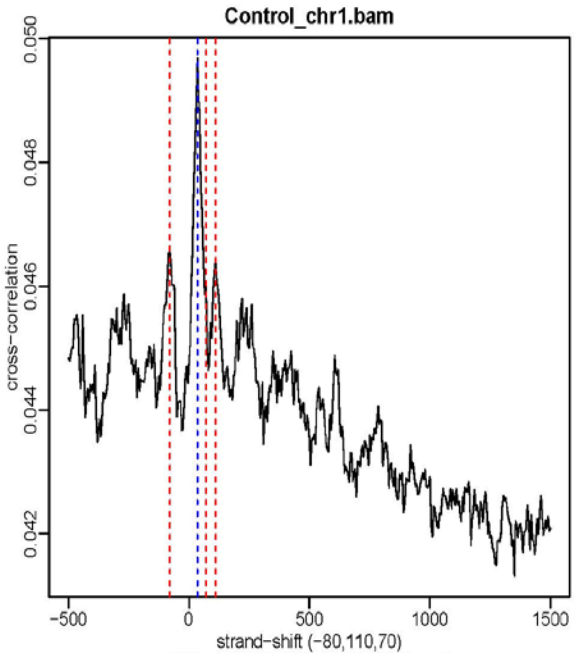


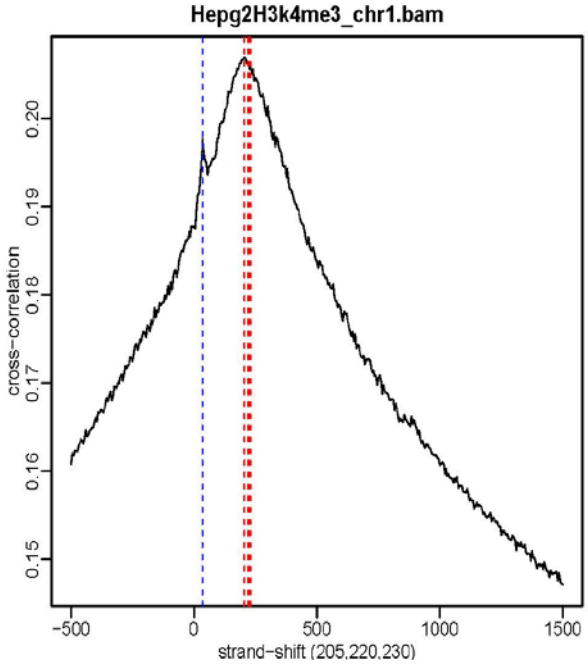
Pepke, S. *et al*. Computation for ChIP-seq and RNA-seq studies, *Nat Methods. Nov. 2009*

Estimation of the fragment length: Strand cross-correlation analysis

Example command:
/nfs/BaRC_Public/phantompeakqualtools/run_spp.R -c=H3k4me3_chr1.bam -savp -out=H3k4me3_chr1.run_spp.out







Peak calling: MACS

- MACS can calculate the fragment length but we will use a different program and give MACS the fragment length as an input parameter.
- It uses a Poisson distribution to assign p-values to peaks. But the distribution has a dynamic parameter, local lambda, to capture the influence of local biases.
- MACS default is to filter out redundant tags at the same location and with the same strand by allowing at most 1 tag. This works well.
- -g: You need to set up this parameter accordingly:

  Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts: 'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for *C. elegans* (9e7) and 'dm' for fruit fly (1.2e8), Default:hs
- For broad peaks like some histone modifications it is recommended to use `--nomodel` and if there is not input sample to use `--nolambda`.

Example of MACS command

MACS command

bsub macs2 callpeak -t H3k4me3_chr1.bam -c Control_chr1.bam --name H3k4me3_chr1 -f BAM -g hs --nomodel -B --extsize "size calculated on the strand croscorrelation analysis“

PARAMETERS

- -t TFILE Treatment file
- -c CFILE Control file
- --name NAME Experiment name, which will be used to generate output file names. DEFAULT: “NA”
- -f FORMAT Format of tag file, “BED” or “SAM” or “BAM” or “BOWTIE”. DEFAULT: “BED”
- --nomodel skips the step of calculating the fragment size.
- -B create a begraph
- --extsize EXTSIZE The arbitrary extension size in bp. When nomodel is true, MACS will use this value as fragment size to extend each read towards 3' end, then pile them up. **You can use the value from the strand cross-correlation analysis**

MACS output

Output files:

1. Excel peaks file ("_peaks.xls") contains the following columns

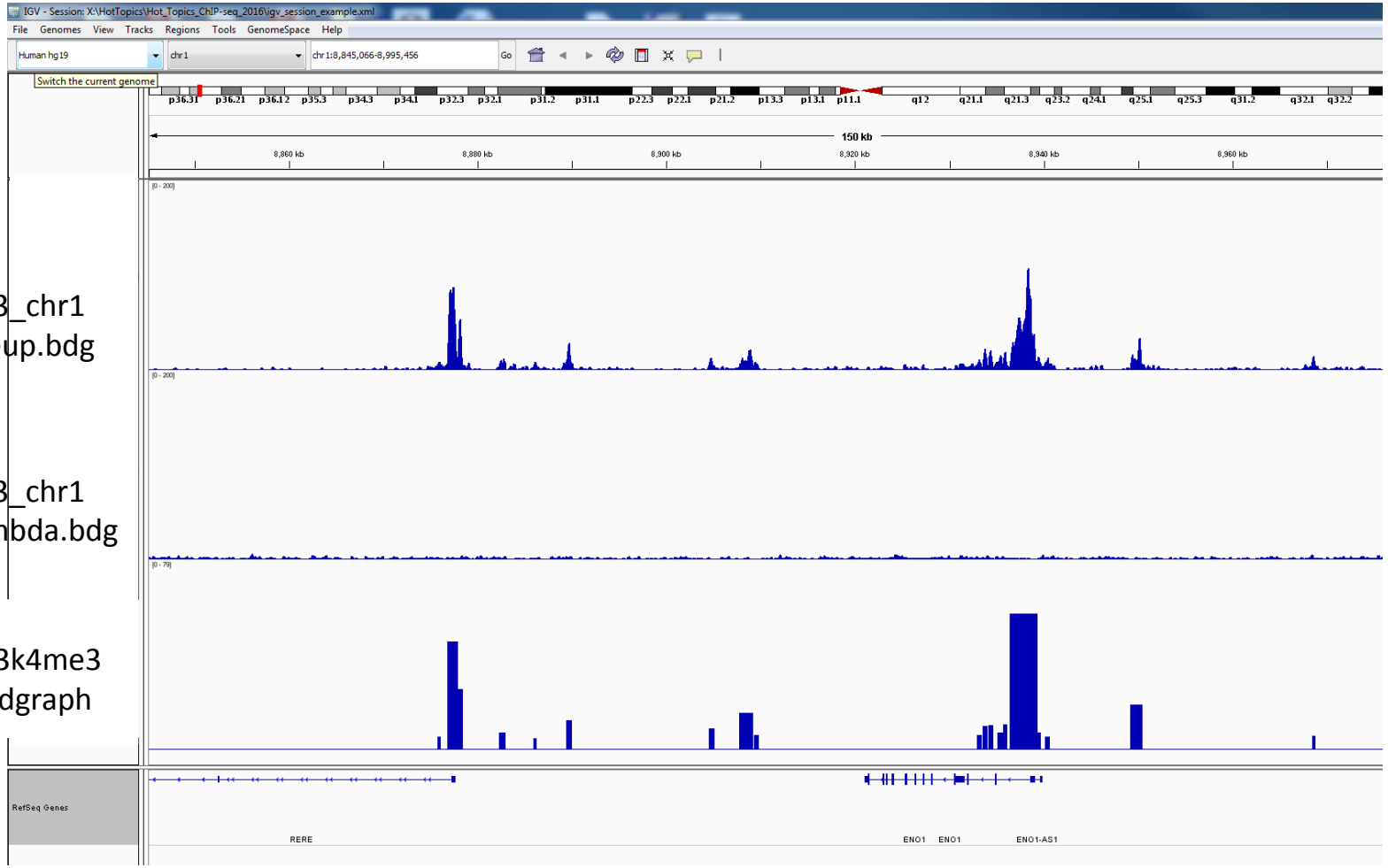
   `Chr, start, end, length, abs_summit, pileup, -LOG10(pvalue), -LOG10(qvalue), name`

2. "_summits.bed": contains the peak summits locations for every peaks. The 5th column in this file is -log10qvalue
3. "_peaks.narrowPeak" is BED6+4 format file. Contains the peak locations together with peak summit, fold-change, pvalue and qvalue.

To look at the peaks on a genome browser you can upload one of the output bed files or you can also make a bedgraph file with columns (step 6 of hands on):

`chr, start, end, fold_enrichment`

Visualize peaks in IGV



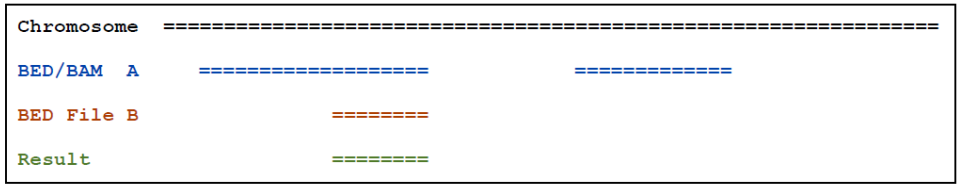
Other recommendations

- Look at your mapped reads and peaks in a genome browser to verify peak calling thresholds
- Optional: remove reads mapping to the ENCODE and 1000 Genomes blacklisted regions

  https://sites.google.com/site/anshulkundaje/projects/blacklists

Linking peaks to genes: Bed tools

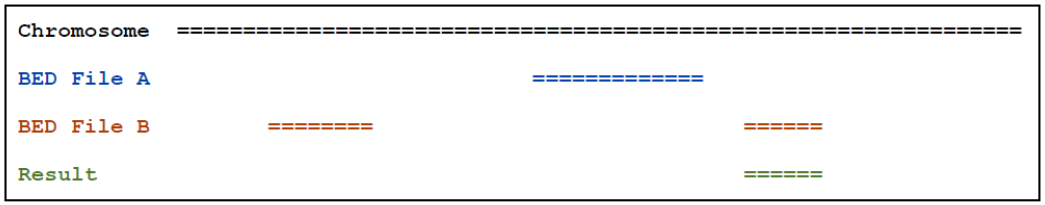
intersectBed



slopBed

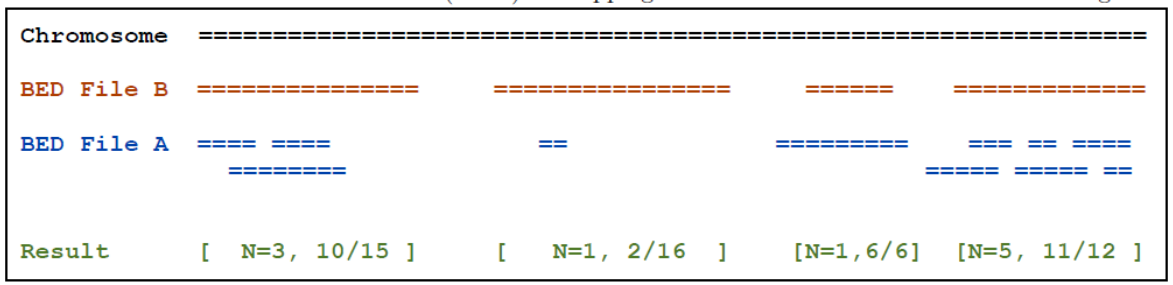
```
$ cat A.bed
chr1  5     100
chr1  800   980

$ cat my.genome
chr1  1000

$ slopBed -i A.bed -g my.genome -b 5
chr1  0     105
chr1  795   985

$ slopBed -i A.bed -g my.genome -l 2 -r 3
chr1  3     103
chr1  798   983
```

closestBed



coverageBed

Below are the number of features in A (N=...) overlapping B and fraction of bases in B with coverage.



groupBy

It groups rows based on the value of a given column/s and it summarizes the other columns

Linking peaks to nearby genes

- Take all genes and add 3Kb up and down with slopBed

```
slopBed -b 3000 -i GRCh37.p13.HumanENSEMBLgenes.bed -g
    /nfs/genomes/human_gp_feb_09_no_random/anno/chromInfo.txt >
    HumanGenesPlusMinus3kb.bed
```

- Intersect the slopped genes with peaks and get the list of unique genes overlapping

```
intersectBed -wa  -a HumanGenesPlusMinus3kb.bed  -b peaks.bed
    | awk '{print $4}' | sort -u > Genesat3KborlessfromPeaks.txt
```

```
intersectBed -wa  -a HumanGenesPlusMinus3kb.bed  -b peaks.bed | head -3
chr1   45956538    45968751    ENSG00000236624_CCDC163P
chr1   45956538    45968751    ENSG00000236624_CCDC163P
chr1   51522509    51528577    ENSG00000265538_MIR4421
```

Link peaks to closest gene

For each region find the closest gene and filter based on the distance to the gene

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed |head
chr1    20870   21204   H3k4me3_chr1_peak_1     5.77592 chr1    14363   29806   ENSG00000227232_WASH7P  0
chr1    28482   30214   H3k4me3_chr1_peak_2     374.48264       chr1    29554   31109   ENSG00000243485_MIR1302-10      0
chr1    28482   30214   H3k4me3_chr1_peak_2     374.48264       chr1    14363   29806   ENSG00000227232_WASH7P  0
#the next two steps can also be done on excel
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c
    6,7,8, -o distinct,distinct,distinct | head -3
ENSG00000227232_WASH7P  0       chr1    14363   29806
ENSG00000243485_MIR1302-10      0       chr1    29554   31109
ENSG00000227232_WASH7P  0       chr1    14363   29806

closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c 6,7,8, -o
    distinct,distinct,distinct | awk  'BEGIN {OFS="\t"}{ if ($2<3000)  {print $3,$4,$5,$1,$2} } ' | head -5
chr1    14363   29806   ENSG00000227232_WASH7P  0
chr1    29554   31109   ENSG00000243485_MIR1302-10      0
chr1    14363   29806   ENSG00000227232_WASH7P  0
chr1    134901  139379  ENSG00000237683_AL627309.1      0
chr1    135141  135895  ENSG00000268903_RP11-34P13.15   0
```

Link peaks to closest gene (1 command)

For each region find the closest gene and filter based on the distance to the gene

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed |
groupBy -g 9,10  -c 6,7,8, -o distinct,distinct,distinct | awk  'BEGIN
{OFS="\t"}{ if ($2<3000)  {print $3,$4,$5,$1,$2} }'  >
closestGeneAt3KborLess.bed
```

closestBed

-d print the distance to the feature in -b

groupBy

-g columns to group on

-c columns to summarize

-o operation to use to summarize

Comparing ChIP-seq across samples

*i.e. C*o-localization or differential binding

To compare two samples you can use :

1. intersectBed (finds the subset of peaks **common** in 2 samples **or unique** to one them)
2. macs2 bdgdiff (find peaks present only in one of the samples)

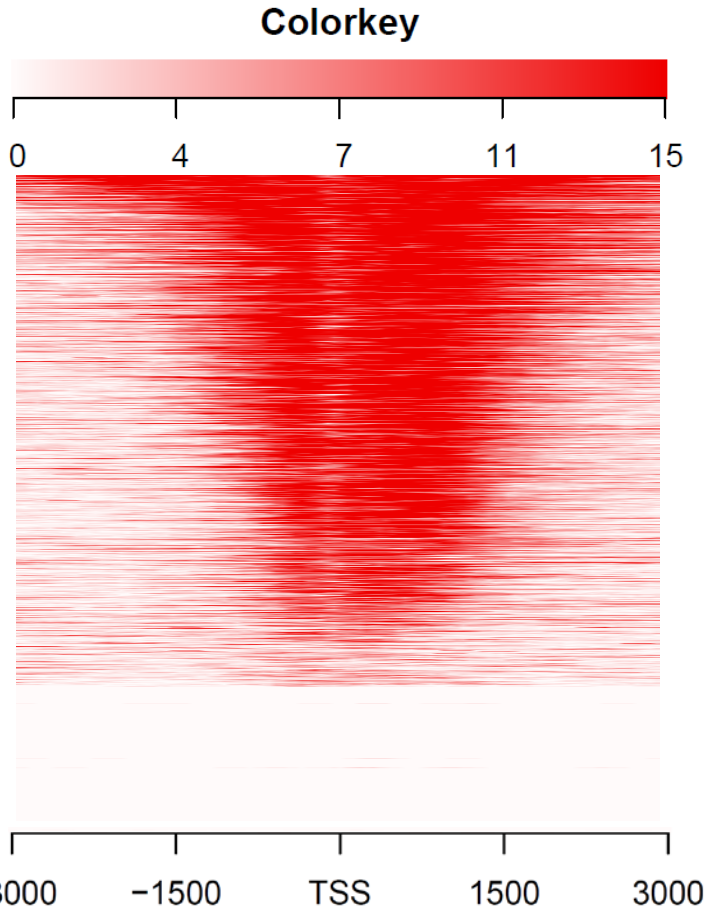
If more than 2 samples follow:

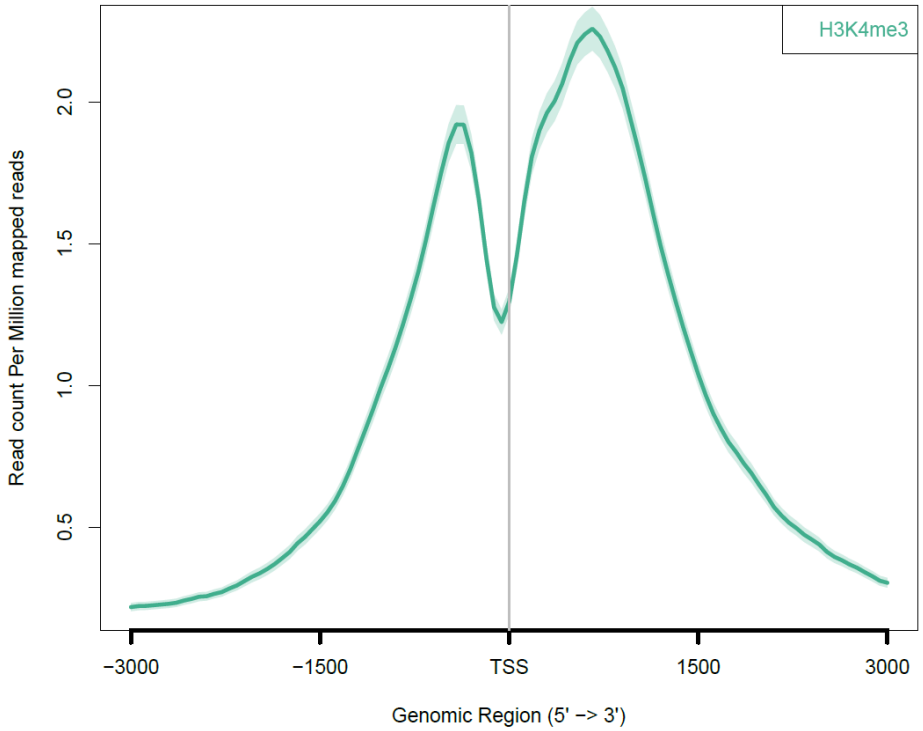
/nfs/BaRC_Public/BaRC_code/Perl/compare_bed_overlaps

Visualizing ChIP-seq reads with ngsplot

See Hot Topics: ngsplot

bsub ngs.plot.r -G hg19 -R tss -C H3k4me3_chr1.bam -O H3k4me3_chr1.tss -T H3K4me3 -L 3000 -FL 300





References

- Reviews and benchmark papers:

  ChIP-seq: advantages and challenges of a maturing technology (Oct 09)
  (http://www.nature.com/nrg/journal/v10/n10/full/nrg2641.html)

  Computation for ChIP-seq and RNA-seq studies (Nov 09)
  (http://www.nature.com/nmeth/journal/v6/n11s/full/nmeth.1371.html)

  Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput. Biol*. 2013

  A computational pipeline for comparative ChIP-seq analyses. *Nat. Protoc*. 2011

  ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res. 2012.*

  Identifying and mitigating bias in next-generation sequencing methods for chromatin biology

  Nature Reviews Genetics 15, 709–721 (2014) Meyer and Liu.

- Quality control and strand cross-correlation:

  http://code.google.com/p/phantompeakqualtools/

- MACS:

  Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 2008
  http://liulab.dfci.harvard.edu/MACS/index.html

  Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011

  http://onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0214s34/pdf

- Bedtools:

  https://code.google.com/p/bedtools/
  http://bioinformatics.oxfordjournals.org/content/26/6/841.abstract

- ngsplot:

  https://code.google.com/p/ngsplot/ Shen, L.*, Shao, N., Liu, X. and Nestler, E. (2014) BMC Genomics, 15, 284.

Other resources

- Previous Hot Topics

  Quality Control and Mapping Reads

  http://jura.wi.mit.edu/bio/education/hot_topics/NGS_QC_mapping_Feb2015/NGS_QC_Mapping2015_1perPage.pdf
- SOPs

  http://barcwiki.wi.mit.edu/wiki/SOPs/chip_seq_peaks
- ENCODE data

  http://genome.ucsc.edu/ENCODE/