

Chip-seq - Analysis

Aurélien Ginolhac

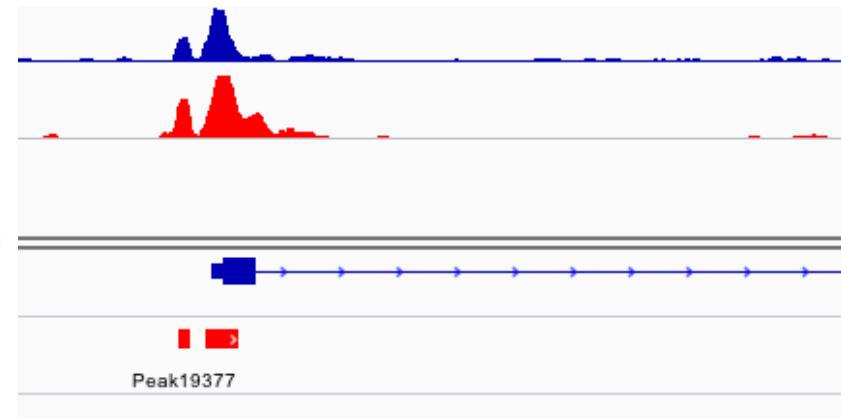
aurelien.ginolhac@uni.lu

Bioinformatics analysis

Sequence file
fastq

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACCTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

Peak file



what this
course is
about

Steps

Filter poor-quality reads (optional)

- Remove sequences with poor-quality bases
- Remove sequences with adapter sequence or other contaminants



Align reads to the genome

- Many aligners to choose from
- Allele-aware aligners
- Speed and memory considerations



Filter artefacts and reads aligning to multiple locations

- Remove duplicate sequences
- PCR artefacts
- Eliminate non-unique aligning reads (but this masks segmental duplications)



Call narrow/broad peaks (ChIP-seq, DNase-seq, FAIRE-seq)

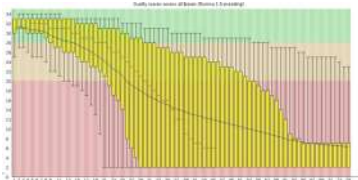
- Settings may vary based on the type of peak
- Highly dependent on threshold settings

Furey 2012.
Nat. Genet. Rev.

Steps, graphics

TGCATGAAAGTCTGTAAGGGGTA

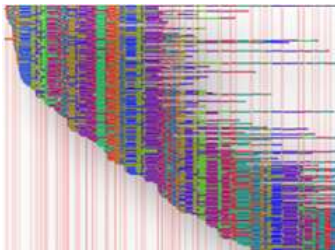
Quality control



Cleaning

TGCATGAAAGTCTGTAAGGGGTA
XXXXXXXX

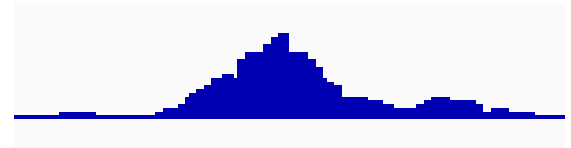
Mapping



Differential peak calling

Motif discovery

Peak calling



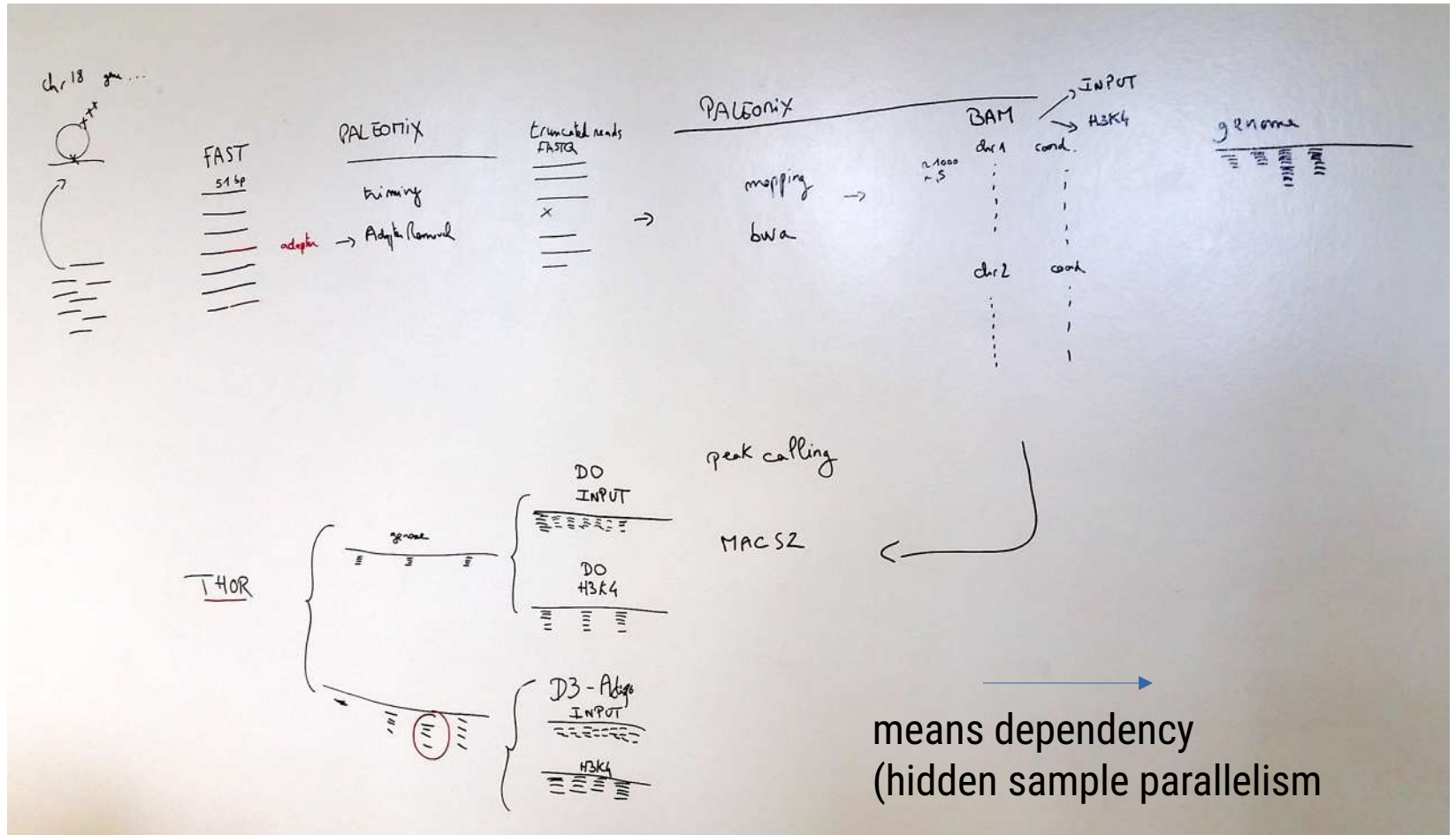
Signal normalization

1 input / 1 IP

Controls



White board implementation



Computer implementation

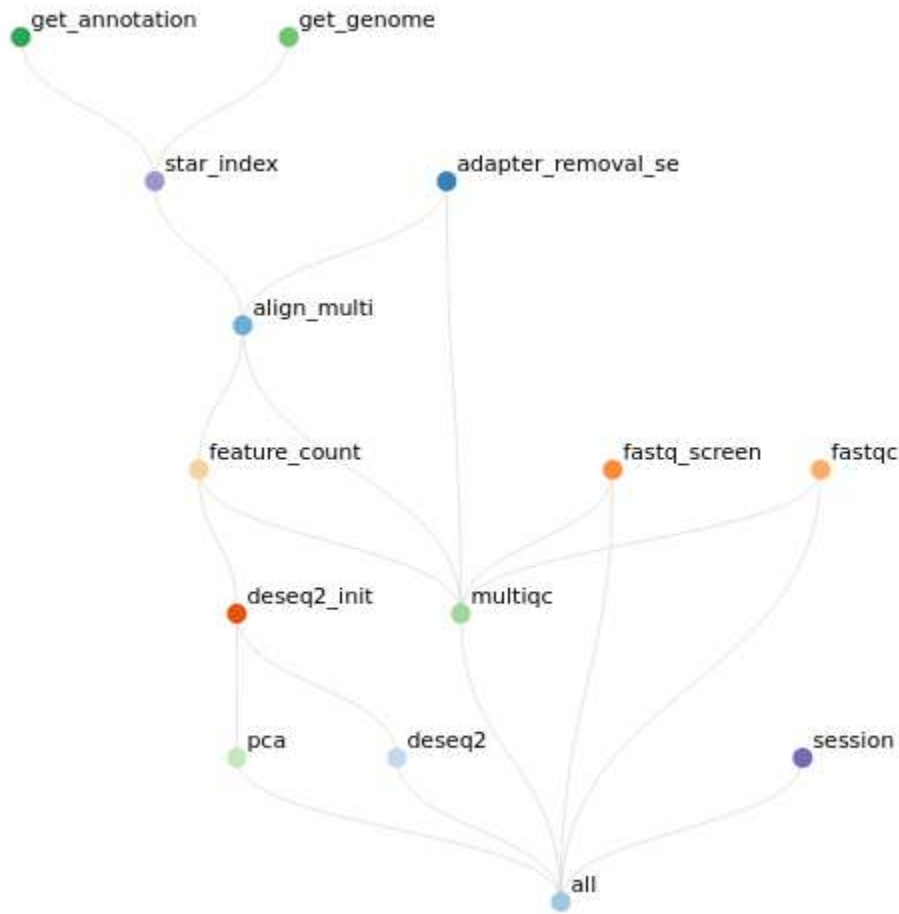
```
#!/bin/csh
# Created by: Chandra Sekhar Pedamallu @ DFCI, The Broad Ins
# Date: June 2016
# Full PathSeq pipeline
# time $pdir/FullPathSeq_June2016.sh $bamloc/SN218_Run0771_L
#
set start_time=`date +%s`

@ noargs=$#
#####PLEASE SET THESE ENVIRONMENTAL
# Program Settings
set Institute="BROAD"
[...]
if($noargs < 4) then
    echo "Please check your arguments"
    echo "Usage : ./FullPathSeq_xxxx.sh <Input file in BAM d
    echo "Example : ./FullPathSeq_xxxx.sh unmappedreads.10K.
    exit
endif
#####INPUT FILES#####
# Present Directory
set pdir = `pwd`
rm $pdir"/clean.files"
rm -r $pdir"/Commands/"
[...]
```

Issues

- mix of code and parameters
- common actions are mingled
- software/input defined as **absolute** paths
- comments are instructions
- software dependencies not included (csh!)
- on HPC, no admin rights
- any issue implies to start over
- version in filename (see the usage with **xxxx**)
- file management (here: `rm recursive!`)
- requires many effort to port over

What we need, dependencies



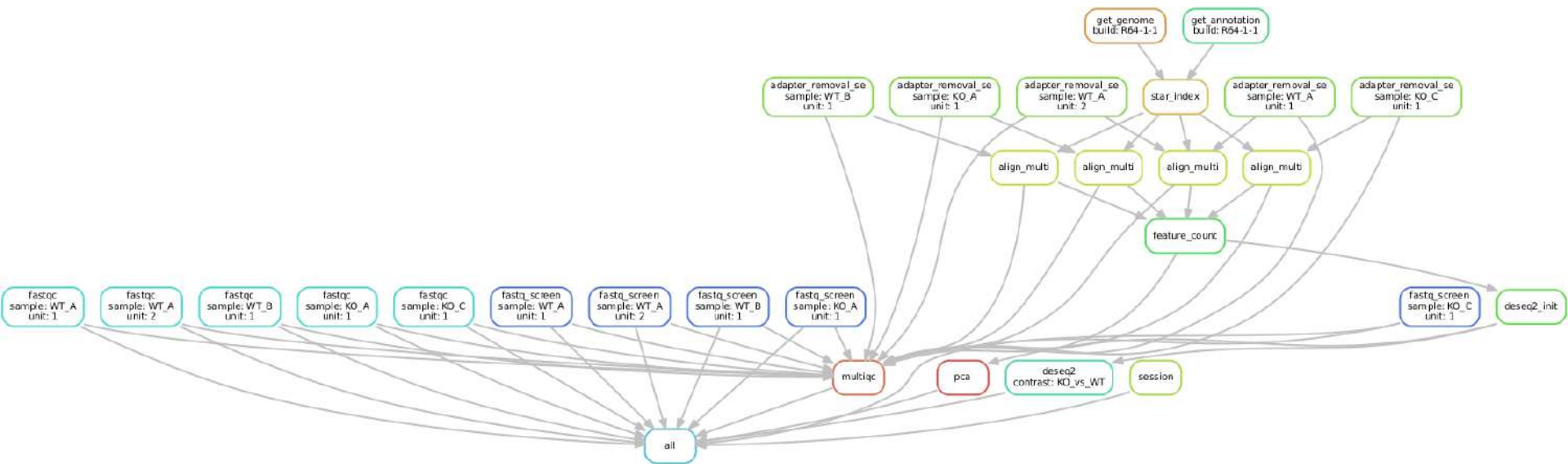
Solved issues

- run only what needs to be, stop micromanaging your analyses
- use **relative** paths
- clear instructions in a README
- code in one folder, users edit 3 text files
- software installed in singularity image
- makes seamless deployment on HPC
- singularity images are versioned
- code and reports also versioned
- ongoing work in temporary folders
- multi-platform (Windows, MacOS, GNU/Linux)

Source: Snakemake RNA-seq workflow,
<https://gitlab.lcsb.uni.lu/aurelien.ginolhac/snakemake-rna-seq>



- This is the workflow being **used**: not just a diagram
- **Explicit** dependencies
- **Parallelization**: independent branches, like `fastqc` and `adaptor_removal` and `samples`



Why you don't need/want to install any software?

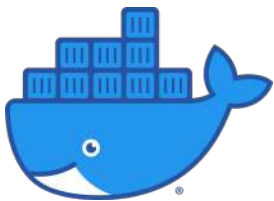
- it's boring
- On HPC one doesn't have admin rights
- modules prepared by the HPC teams are great, but more specific software are missing

Docker

- is great but requires admin rights
- singularity is kind of docker for High Performance Computers

Of course, someone has to install software, it doesn't have to be you

– Aurélien adapted from **Jenny Bryan**






 ginolhac / snake-chip-seq


Image for the ChIP-seq snakemake template 

 Last pushed: a month ago

Tags and Scans

 VULNERABILITY SCANNING - DISABLED [Enable](#)

This repository contains 1 tag(s).

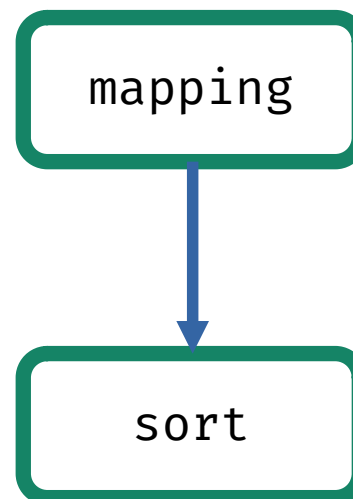
TAG	OS	PULLED	PUSHED
 0.1		4 days ago	a month ago

Snakemake basic

```
rule mapping:
  input: "A.fastq.gz"
  output: "A.bam"
  shell:
    """
    bowtie2 -x GRCm38 -U {input} | \
    samtools sort - > {output}
    """

rule sort:
  input: "A.bam"
  output: "A.bam.bai"
  shell:
    """
    samtools index {input}
    """
```

Deduced DAG



Snakemake workshop



Johannes Köster
@johanneskoester



Registration for the 2021 edition of the Boston [#Snakemake](#) tutorial days (21st and 22nd Sep) is open now. It will be virtual, via Zoom, participation is free, and of course also open to non Bostonians :-):
koesterlab.github.io/bsd2021.html

- free
- by Snakemake creator/maintainer
- online

<https://koesterlab.github.io/bsd2021.html>

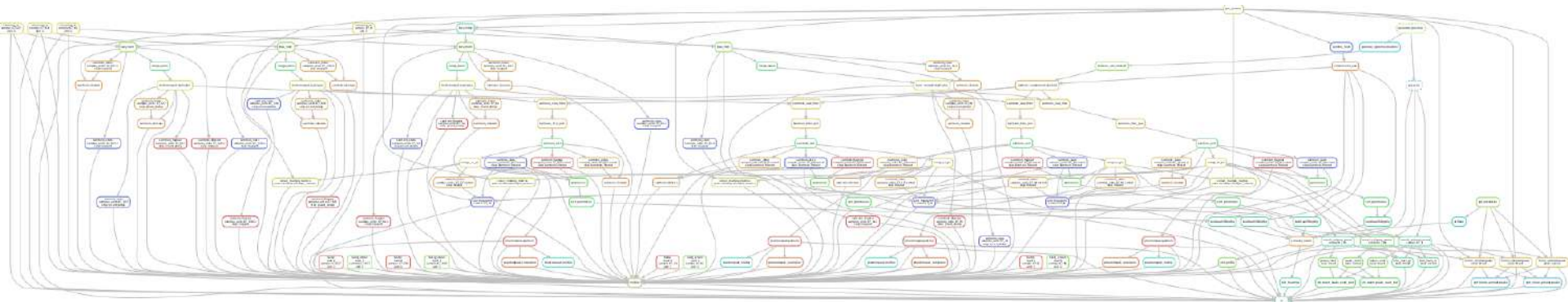
LCSB biocore: Sarah Peter



<https://r3.pages.uni.lu/school/snakemake-tutorial/>

ChIP-seq template

Direct Acyclic Graph



<https://ginolhac.github.io/chip-seq/>



ChIP-seq tutorials



Search

ChIP-seq tutorials

[Home](#)

[Command line, basics](#)

[Setup snakemake](#)

[Workflow](#)

[Mapping](#)

[Peak calling](#)

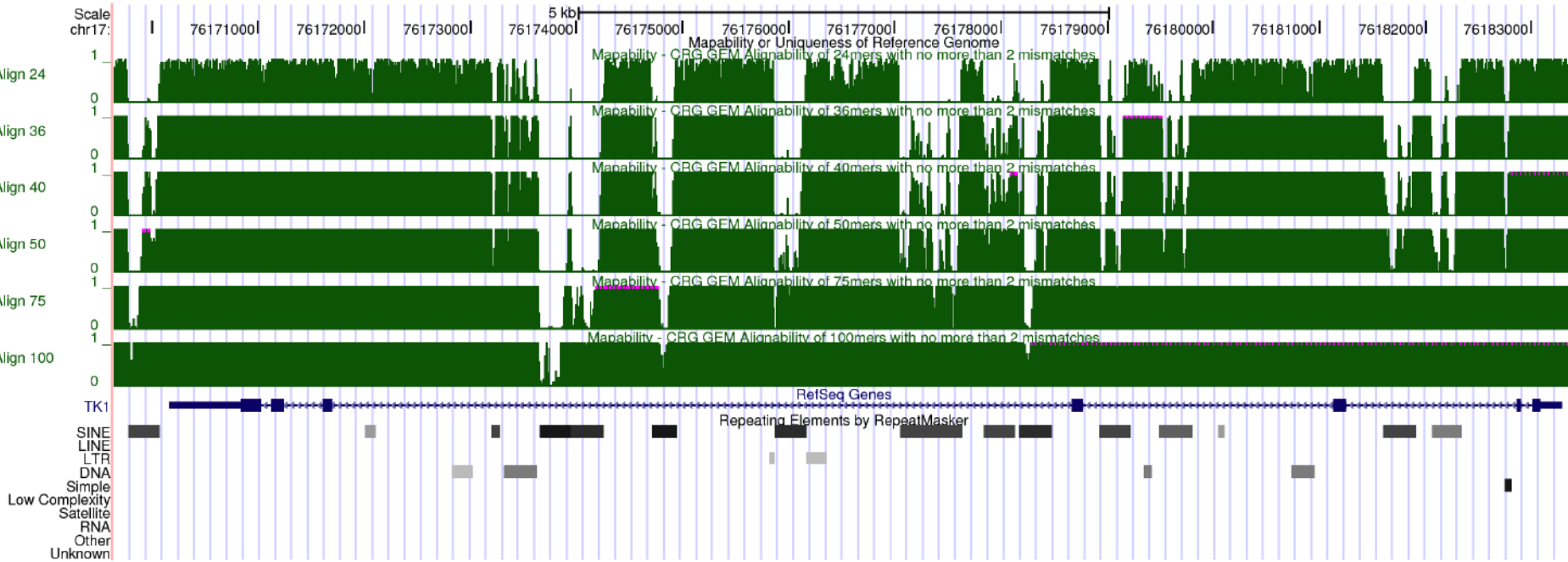
[Contact](#)

ChIP-seq practical session

Running all analyses is computationally intensive and despite the power of the current laptops, jobs should be run on high-performance clusters (HPC).

Moreover, bioinformatic analyses involve many inter-dependent steps that need to be coherently run by a workflow manager such as [snakemake](#)

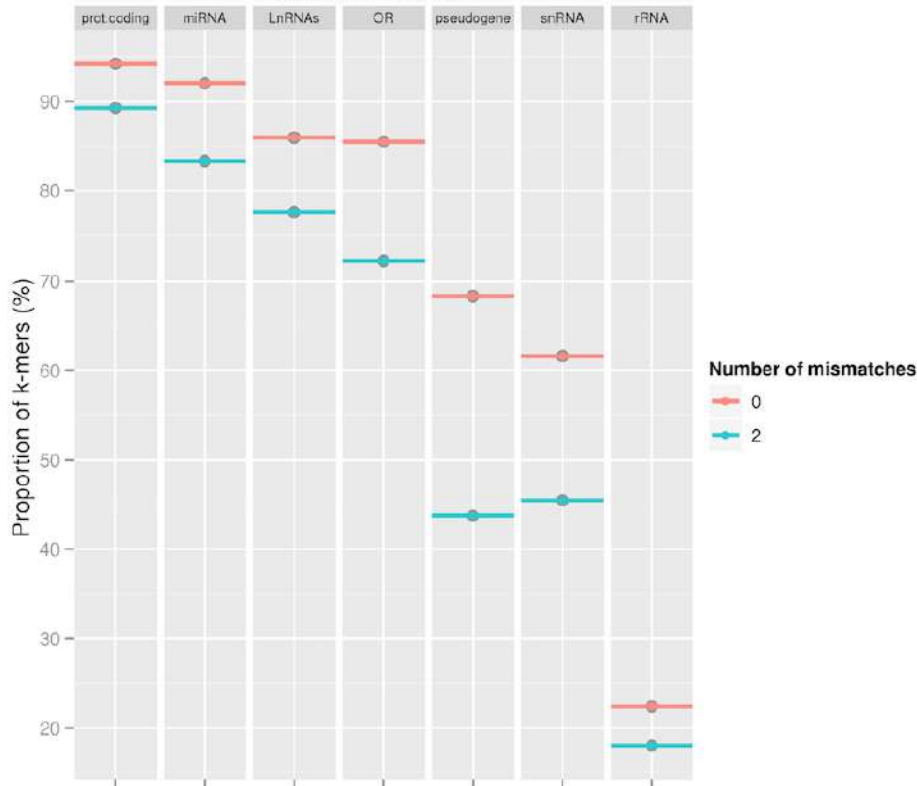
Mappability, causes



Mappability, consequences

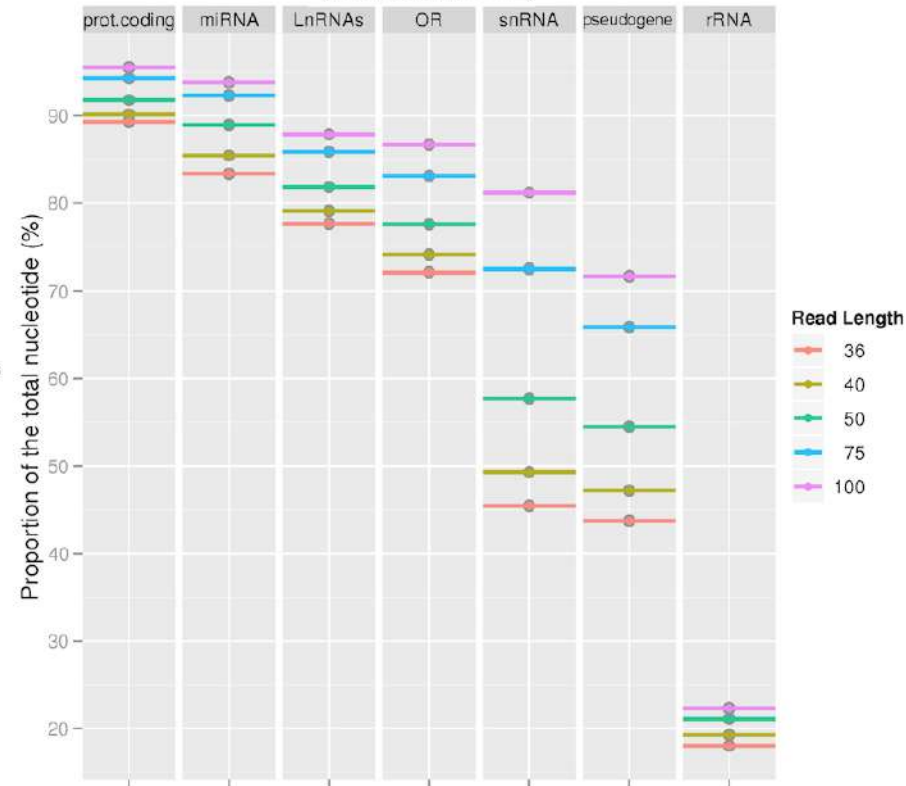
Mismatches

Unique mappings - variable mismatches
(fixed 36-mers)



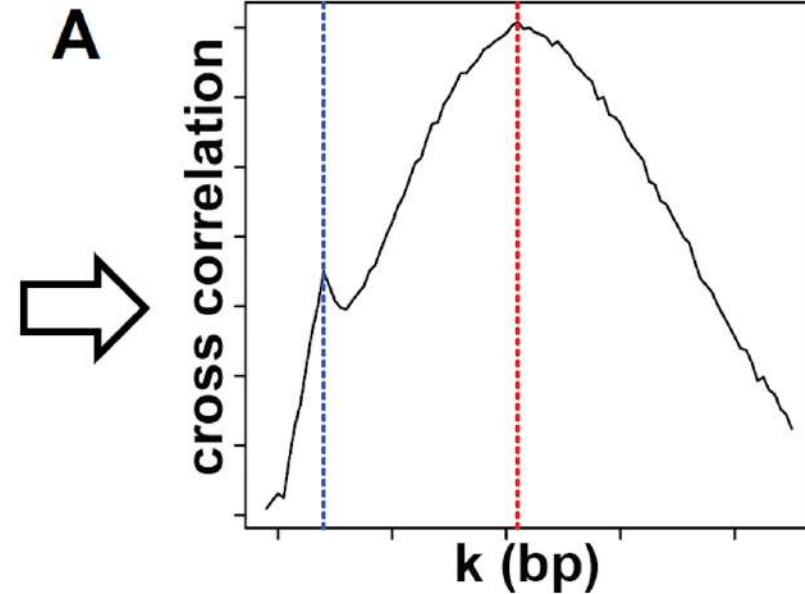
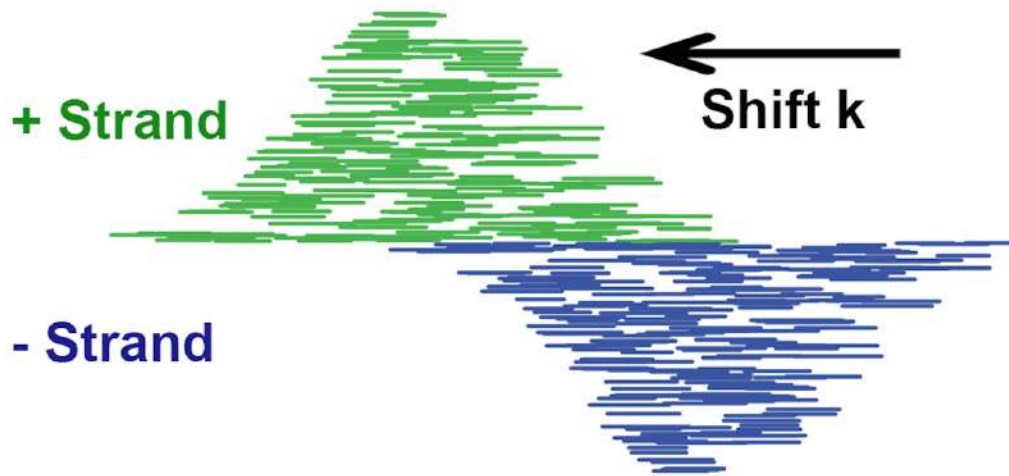
Read Length

Unique mappings - variable k-mer sizes
(mismatches ≤ 2)

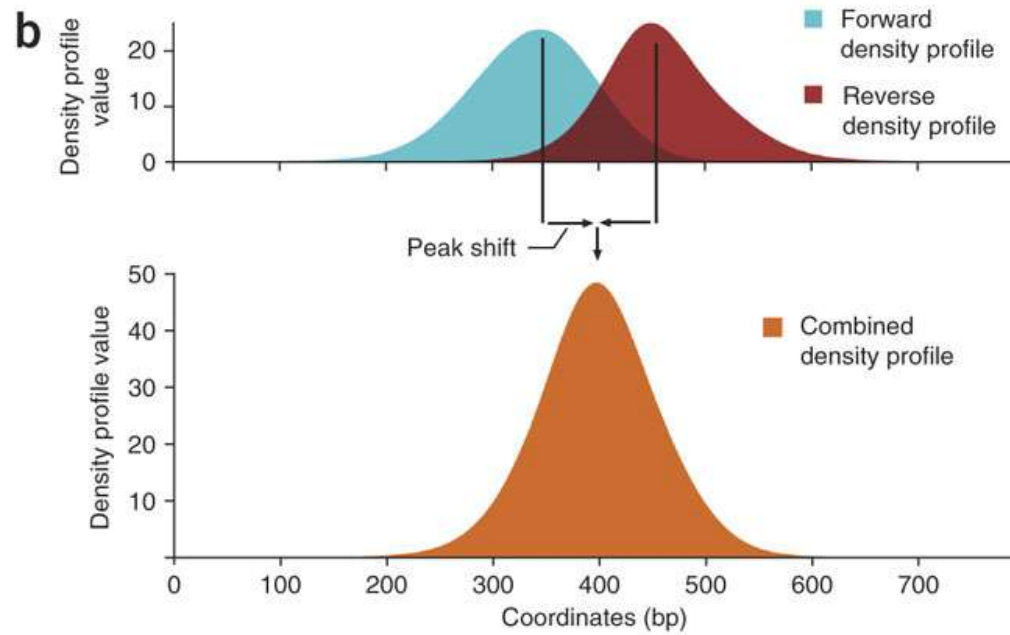
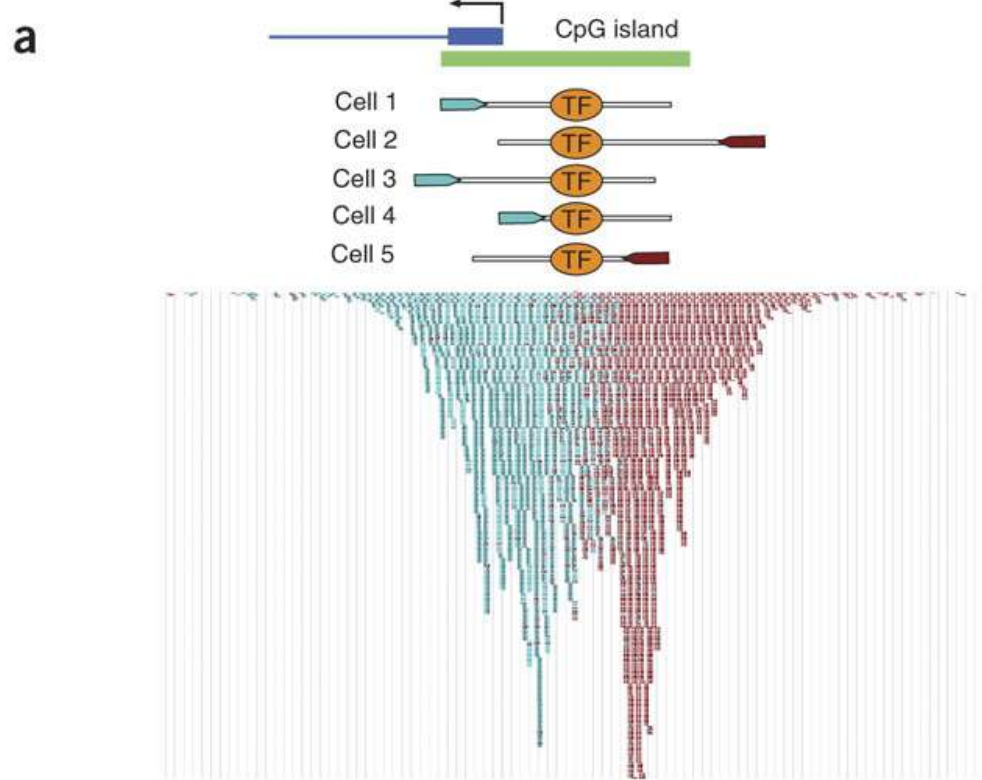


Uniquely mapped

Peal calling, infer the shift size



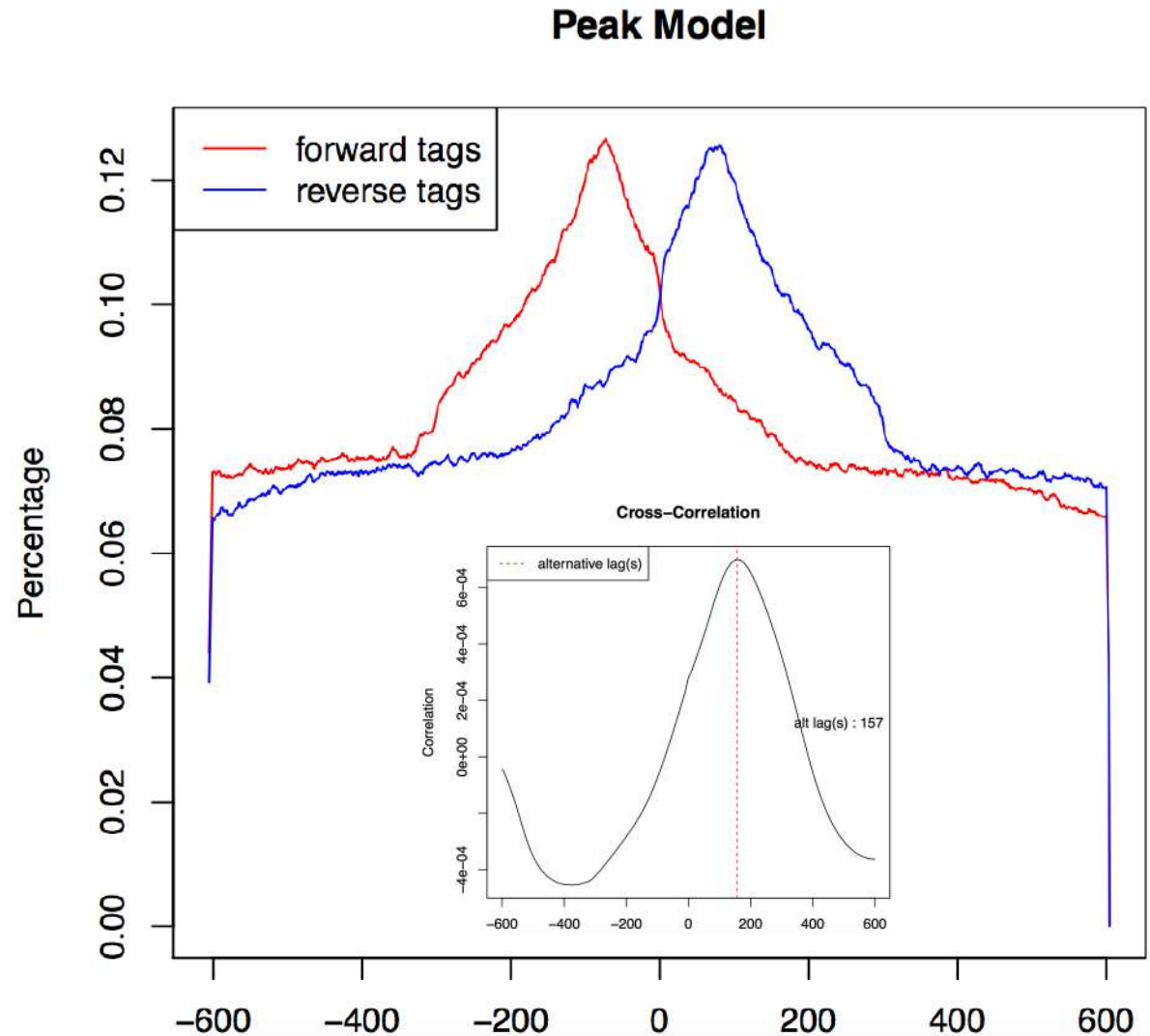
Bailey et al. 2013.
PLoS Comp. Biol.



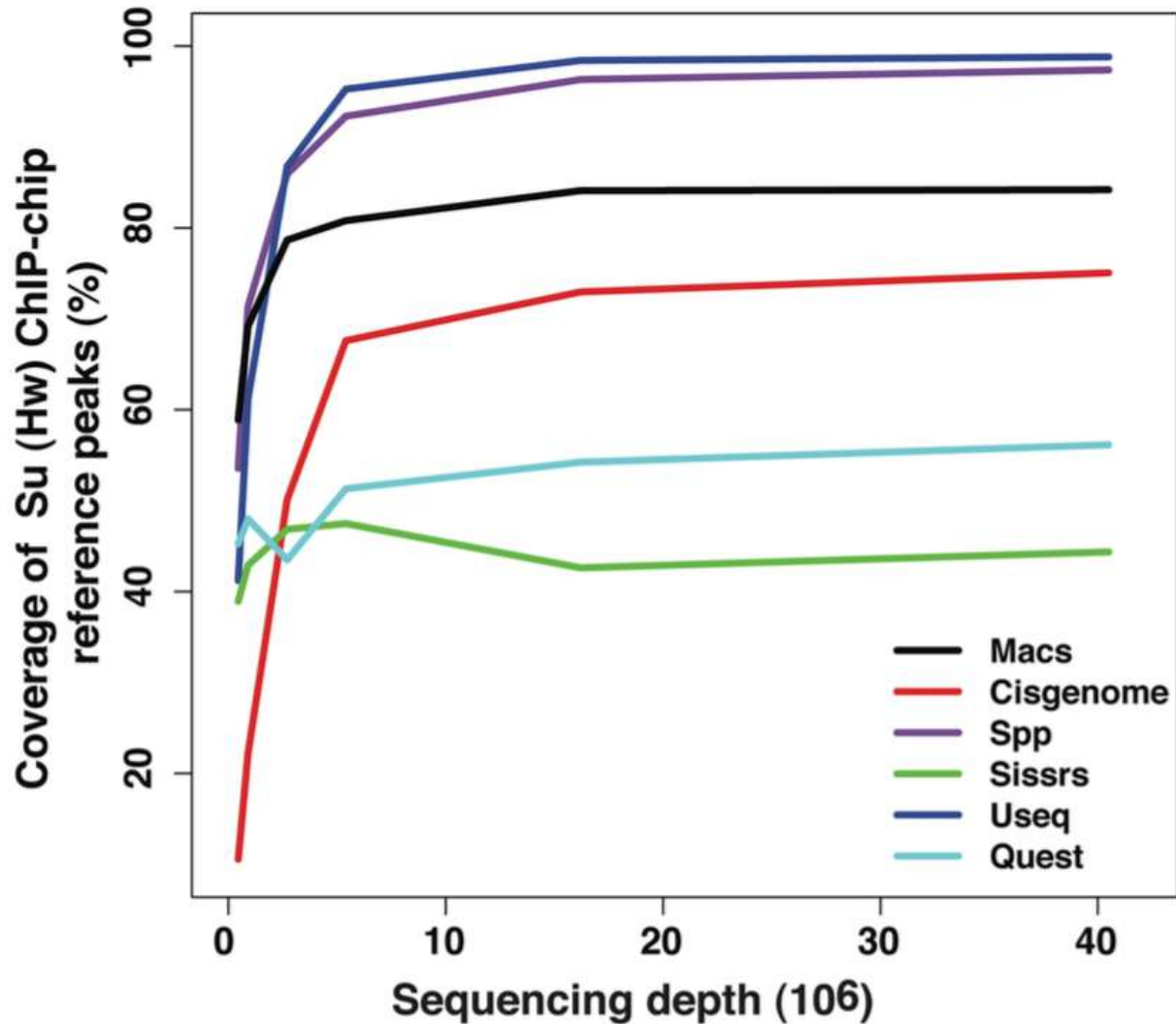
Valouev et al. 2008.
Nat. Methods

Shift modeling: MACS2

Given a sonication size (*bandwidth*) and a high-confidence fold-enrichment (*mfold*), **MACS** slides 2 *bandwidth* windows across the genome to find regions with tags more than *mfold* enriched relative to a random tag genome distribution



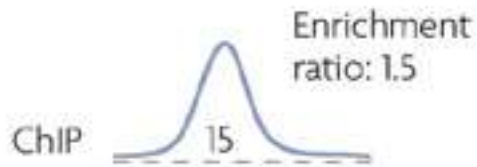
Sequencing depth



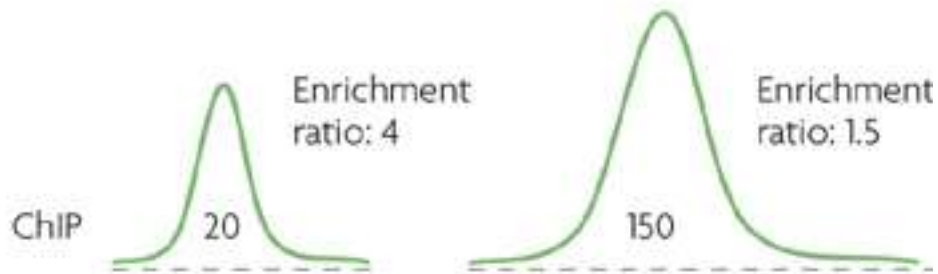
Chen et al. 2013.
Nat. Methods

Sequencing depth

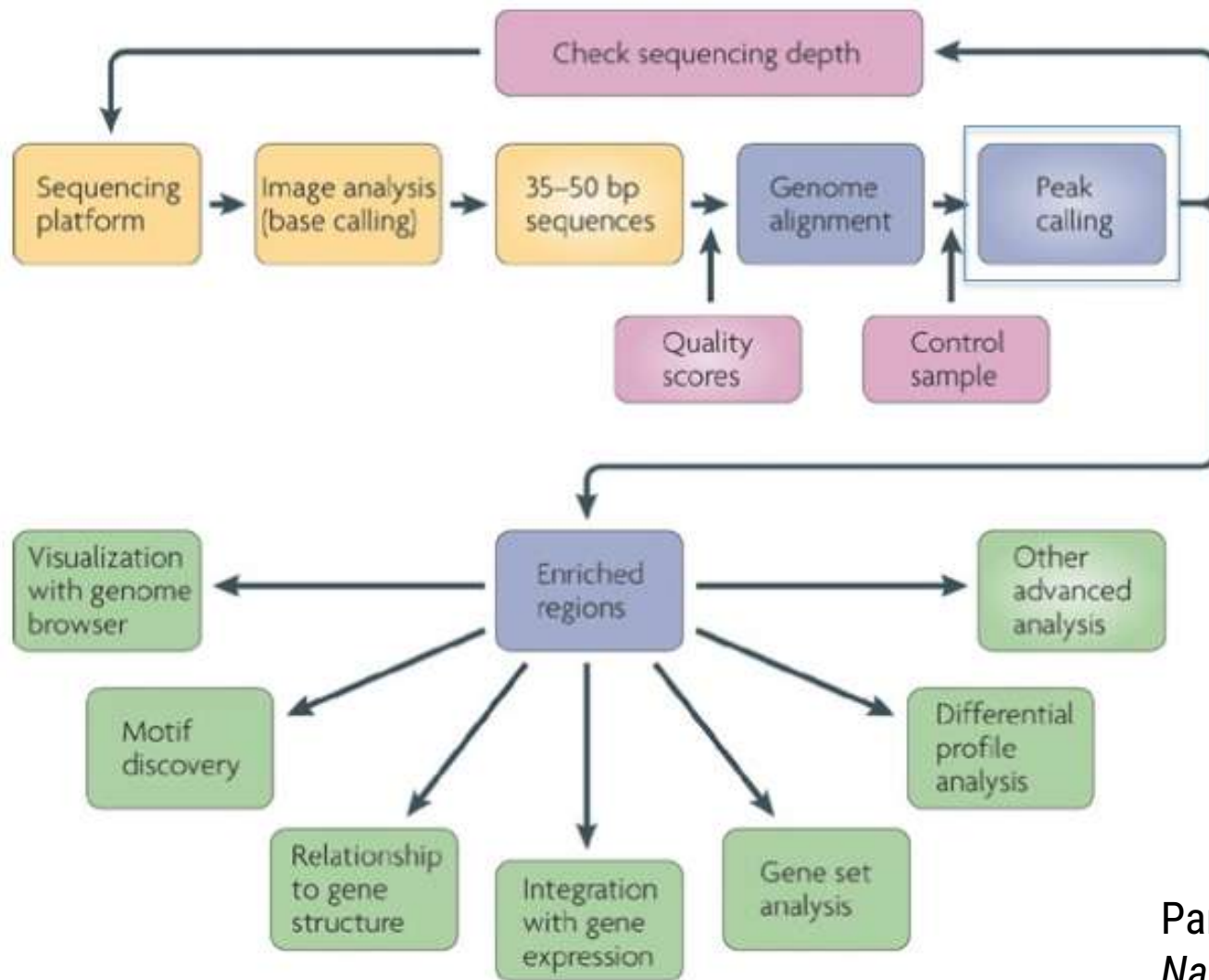
Ba Not statistically significant



Bb Statistically significant



Park et al. 2009.
Nat. Rev. Genet.



Park et al. 2009.
Nat. Rev. Genet.