

Chip-seq - Analysis

Aurélien Ginolhac

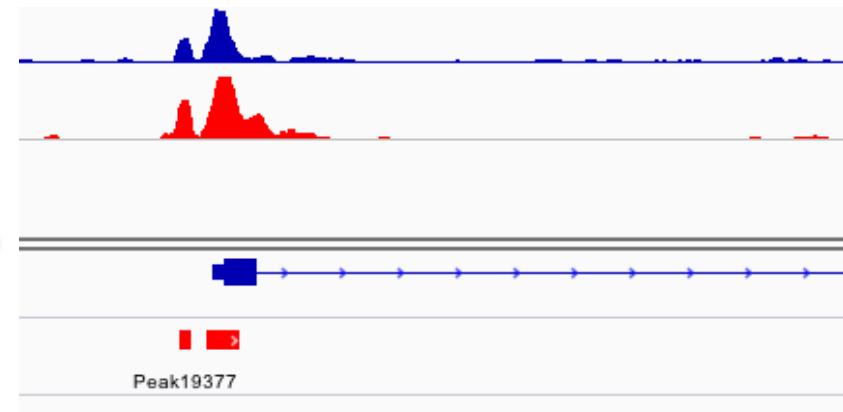
aurelien.ginolhac@uni.lu

Bioinformatics analysis

Sequence file
fastq

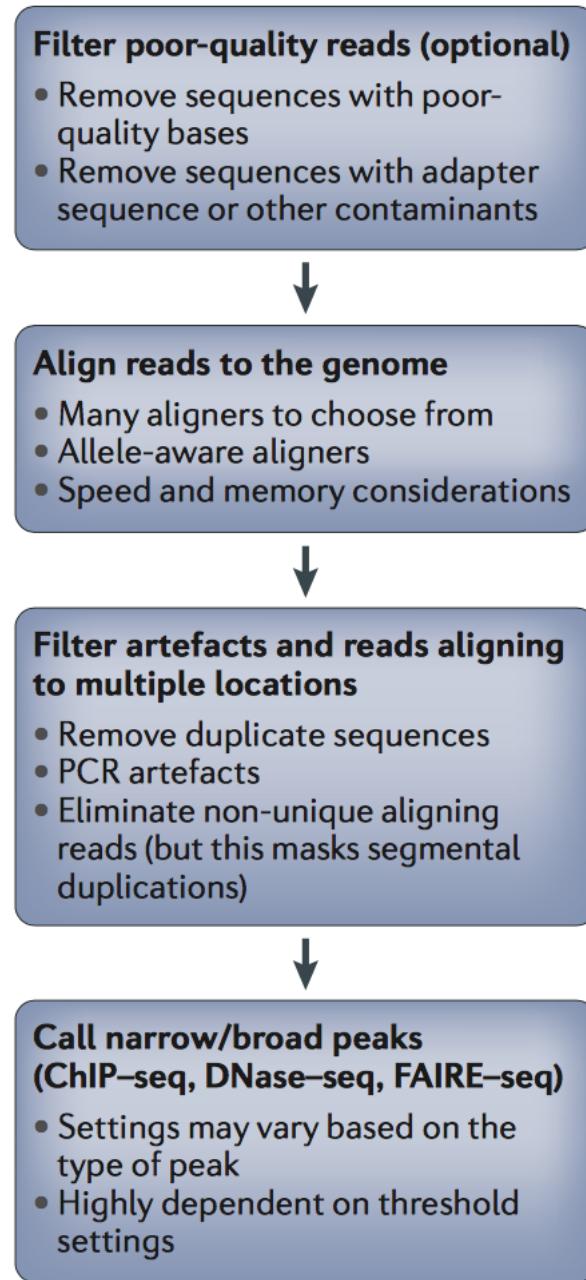
```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?; >52; >:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTCCGTGTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBBAB@B9B@=BABA@A:@693:@B=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

Peak file



what this
course is
about

Steps

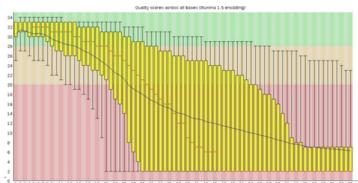


Furey 2012.
Nat. Genet. Rev.

Steps, graphics

TGCATGAAAGTCTGTAAAGGGGTA

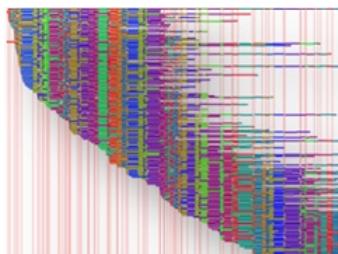
Quality control



Cleaning

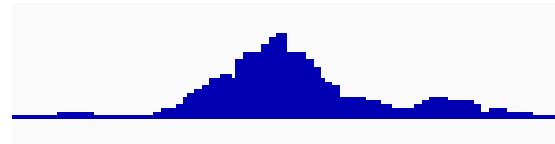
TGCATGAAAGTCTGTAAAGGGGTA
XXXXXX

Mapping



Differential peak calling

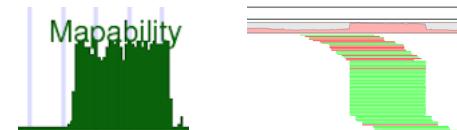
Peak calling



Signal normalization

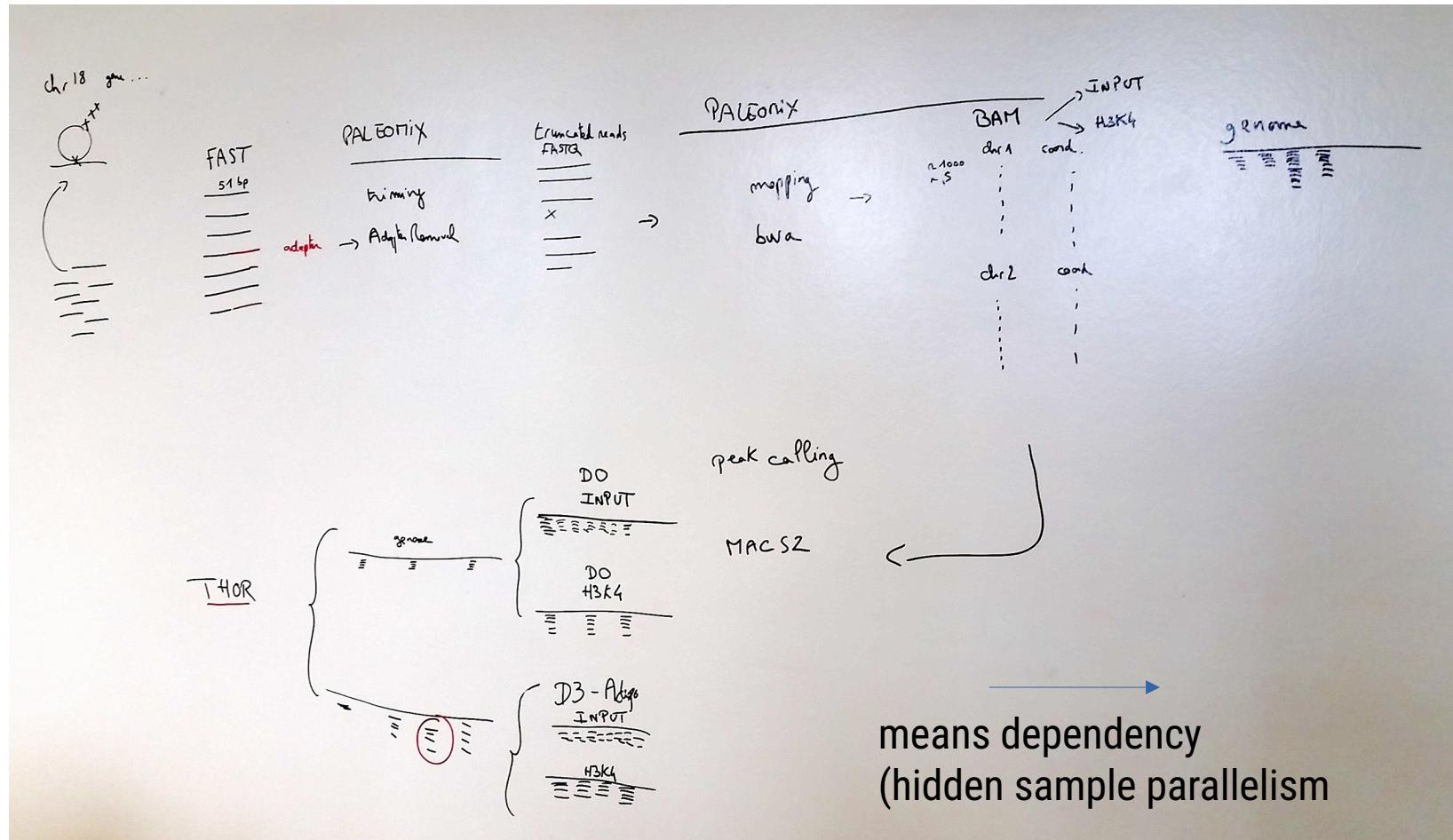
1 input / 1 IP

Controls



Motif discovery

White board implementation



Computer implementation

Issues

```
#!/bin/csh
# Created by: Chandra Sekhar Pedamallu @ DFCI, The Broad Inst
# 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/"
[...]
```

- 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

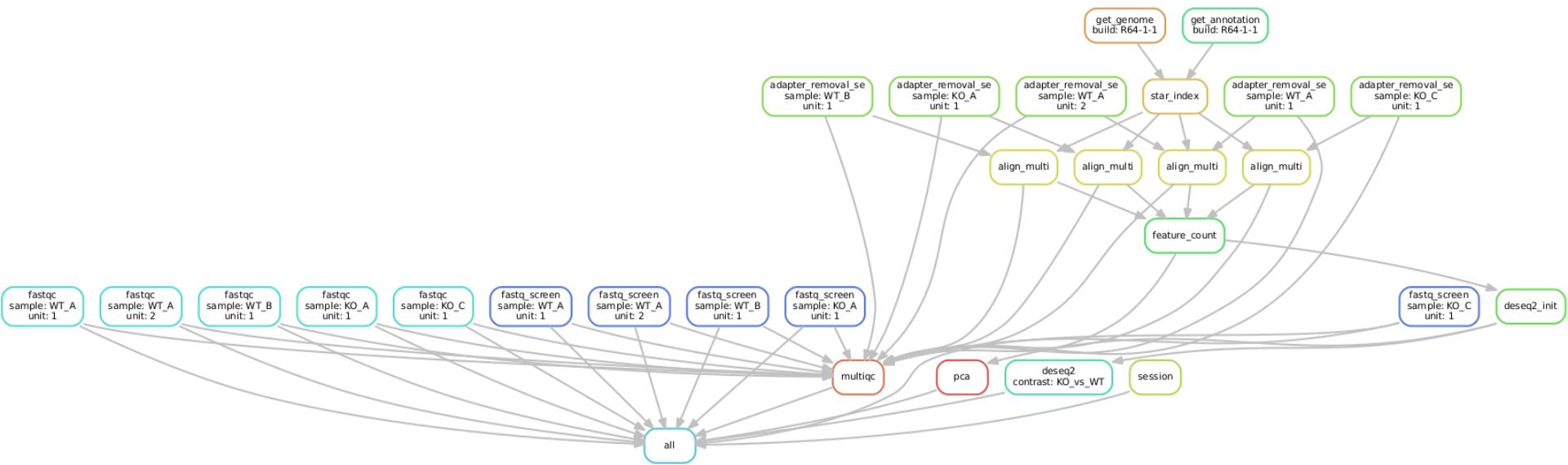
- run only what needs to be, stop micromanaging your analyses
- use **relative** paths
- clear instructions in a README
- code in one folder, users edit 3 text files
- software installed in `singularityimage`
- makes seamless deployment on HPC
- singularity images are versioned, code and reports too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



Python

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

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



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

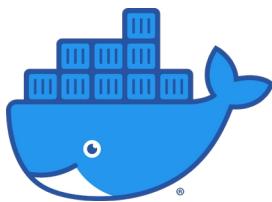
- it's boring
- On HPC one doesn't have admin rights
- modules prepared by the HPC teams are great, but more specific software are missing

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

This repository contains 1 tag(s).

VULNERABILITY SCANNING - DISABLED
[Enable](#)

TAG

OS

PULLED

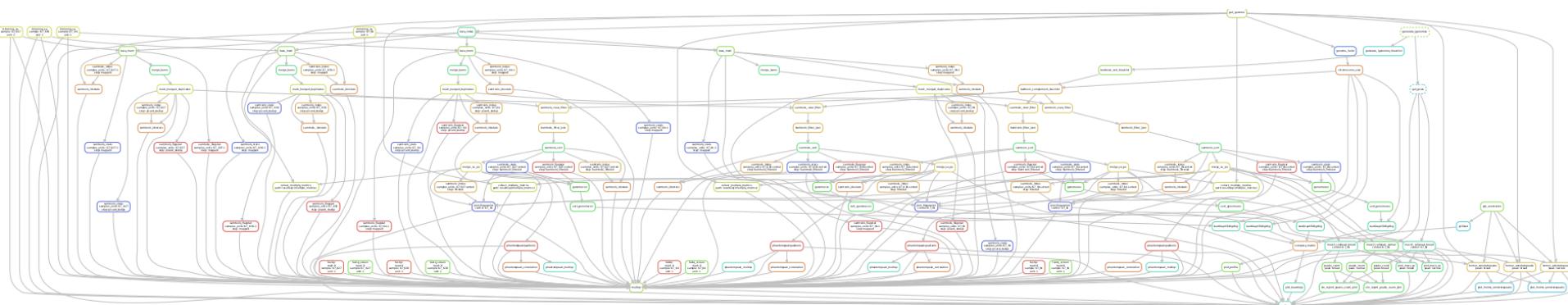
PUSHED

0.1



4 days ago

a month ago



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

The screenshot shows the left sidebar of a documentation site. At the top is a blue header bar with the title "ChIP-seq tutorials" and a search bar below it. Below the header is a dark grey sidebar containing several links:

- Home
- ChIP-seq practical session
- log in iris
- TMUX
- monitoring the resources used

Below this is another dark grey section labeled "Setup" containing:

- Command line, basics
- QC

Docs » Home

[Edit on GitHub](#)

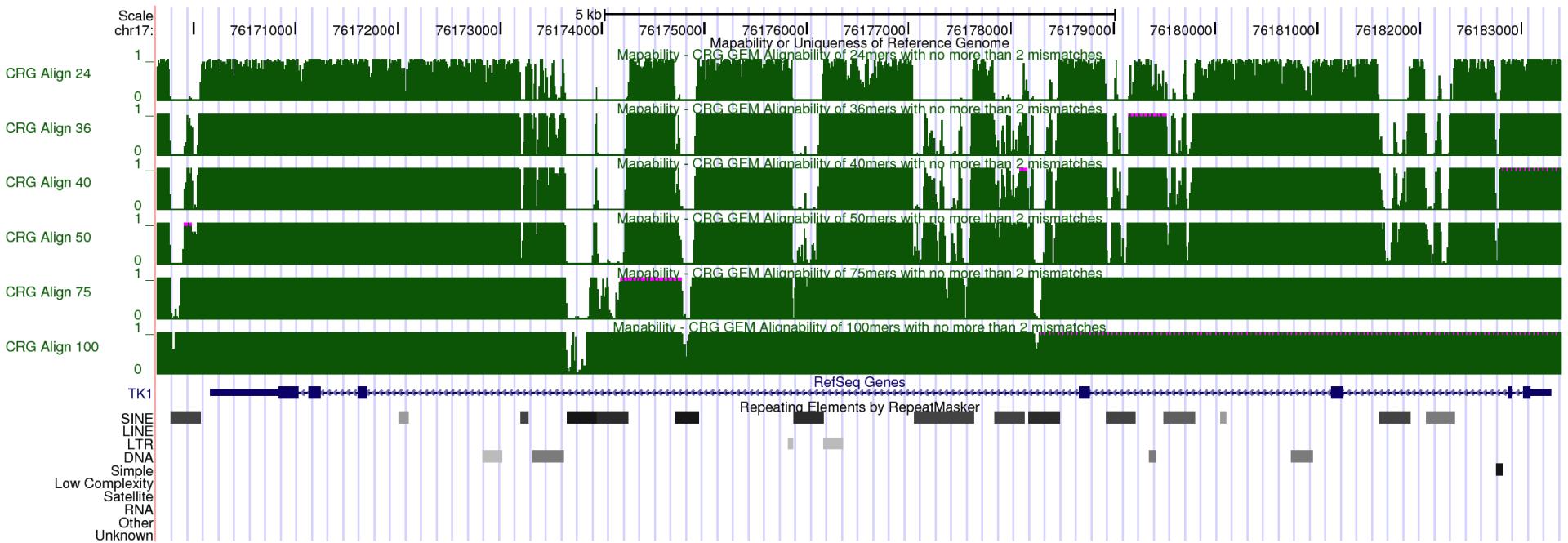
ChIP-seq practical session

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

log in `iris`

`iris` is one of the [High Performance Computer \(HPC\) of the UNI](#).

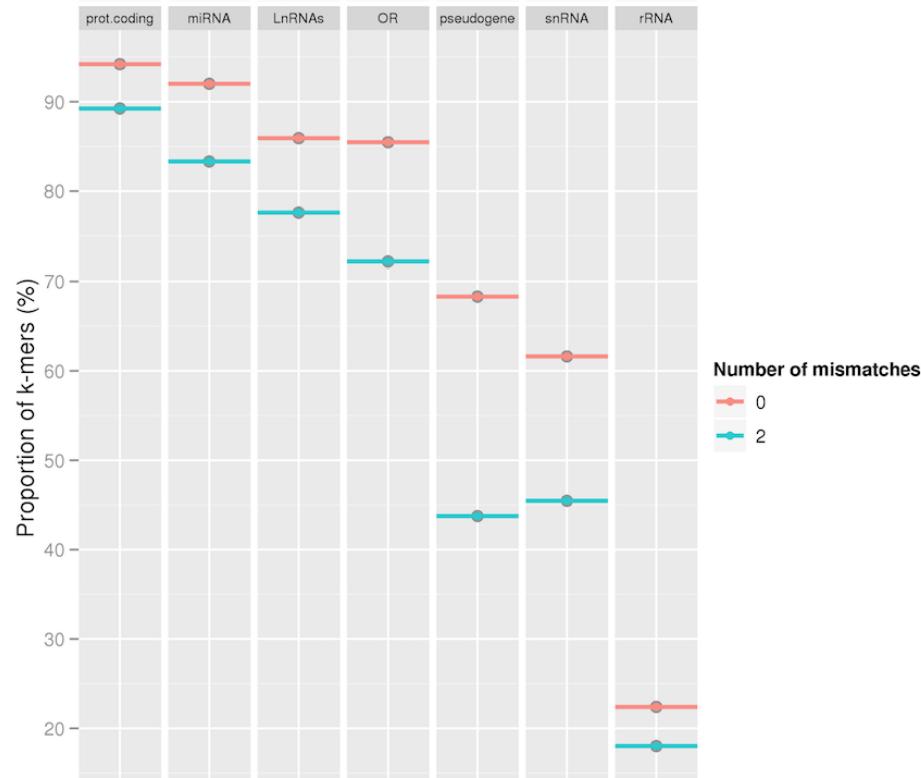
Mappability, causes



Mappability, consequences

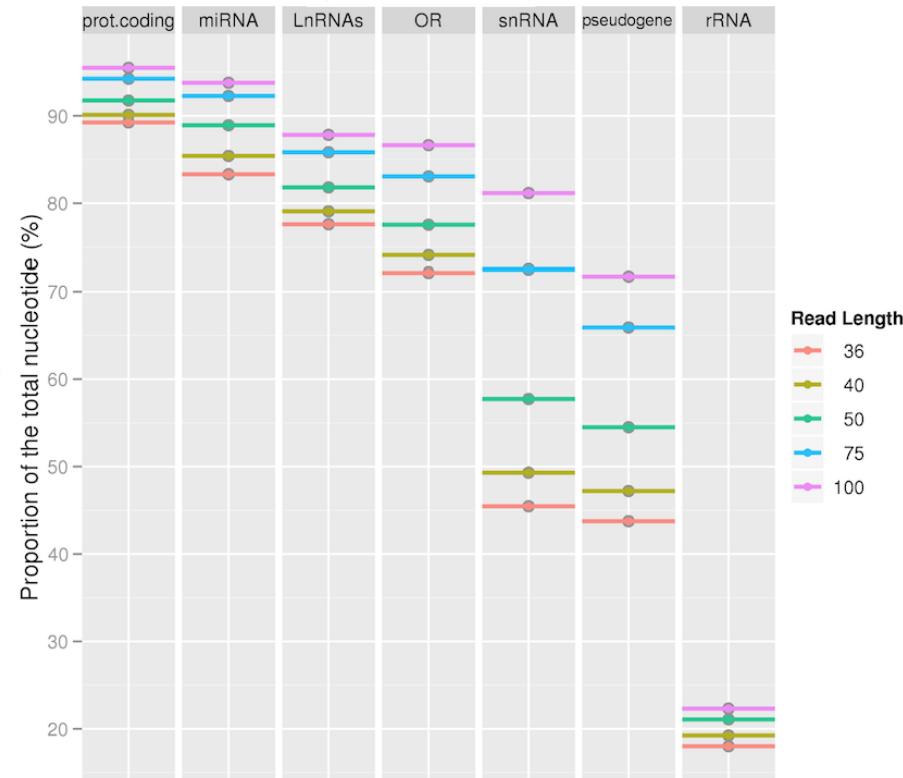
Mismatches

Unique mappings - variable mismatches
(fixed 36-mers)



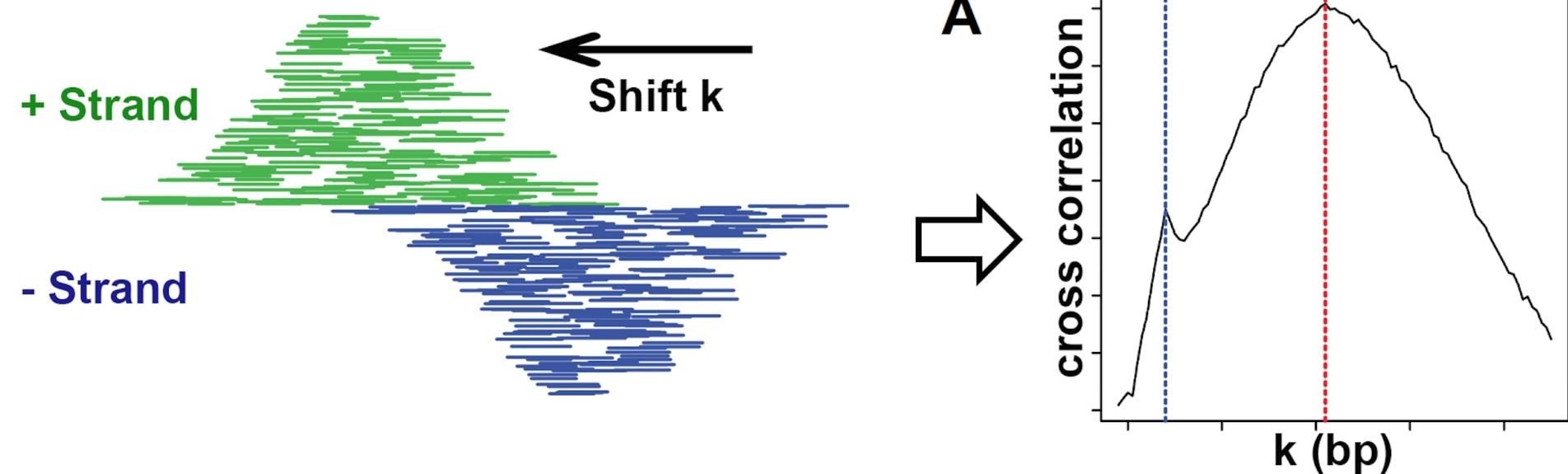
Read Length

Unique mappings - variable k-mer sizes
(mismatches<=2)

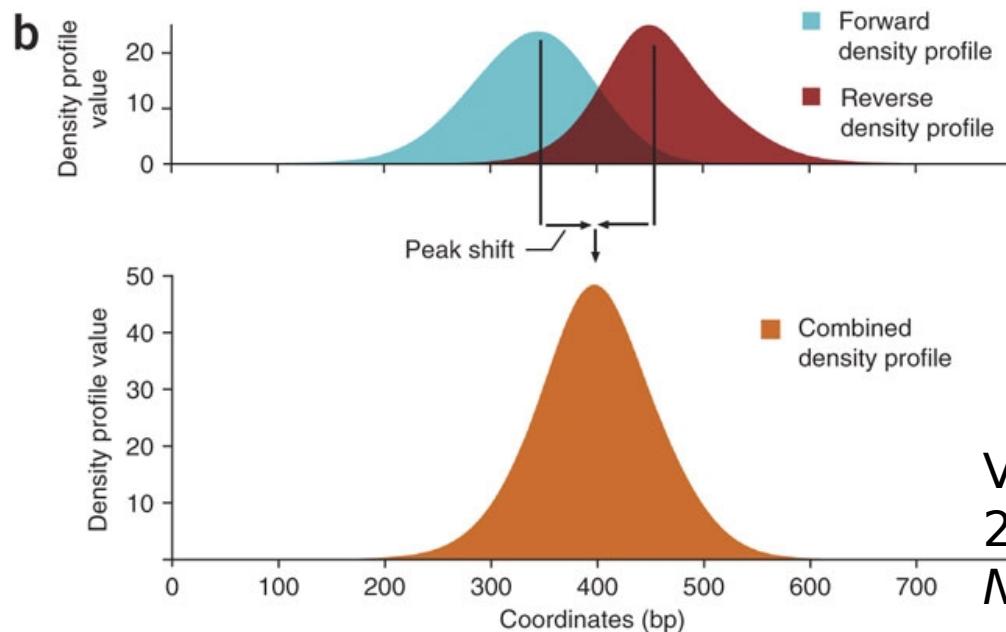
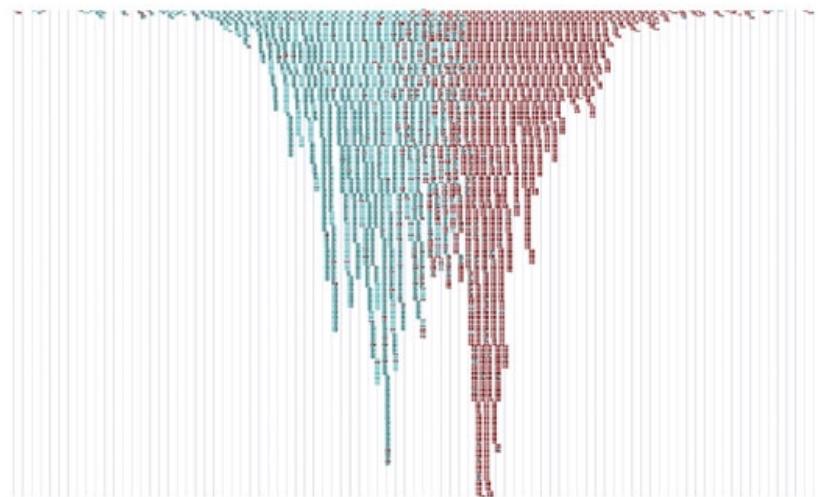
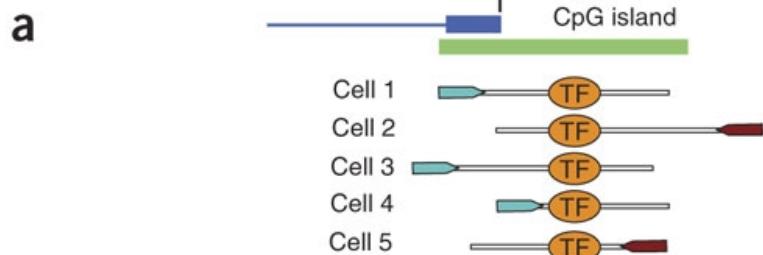


Derrien et al.
2013.
PLoS ONE

Peal calling, infer the shift size



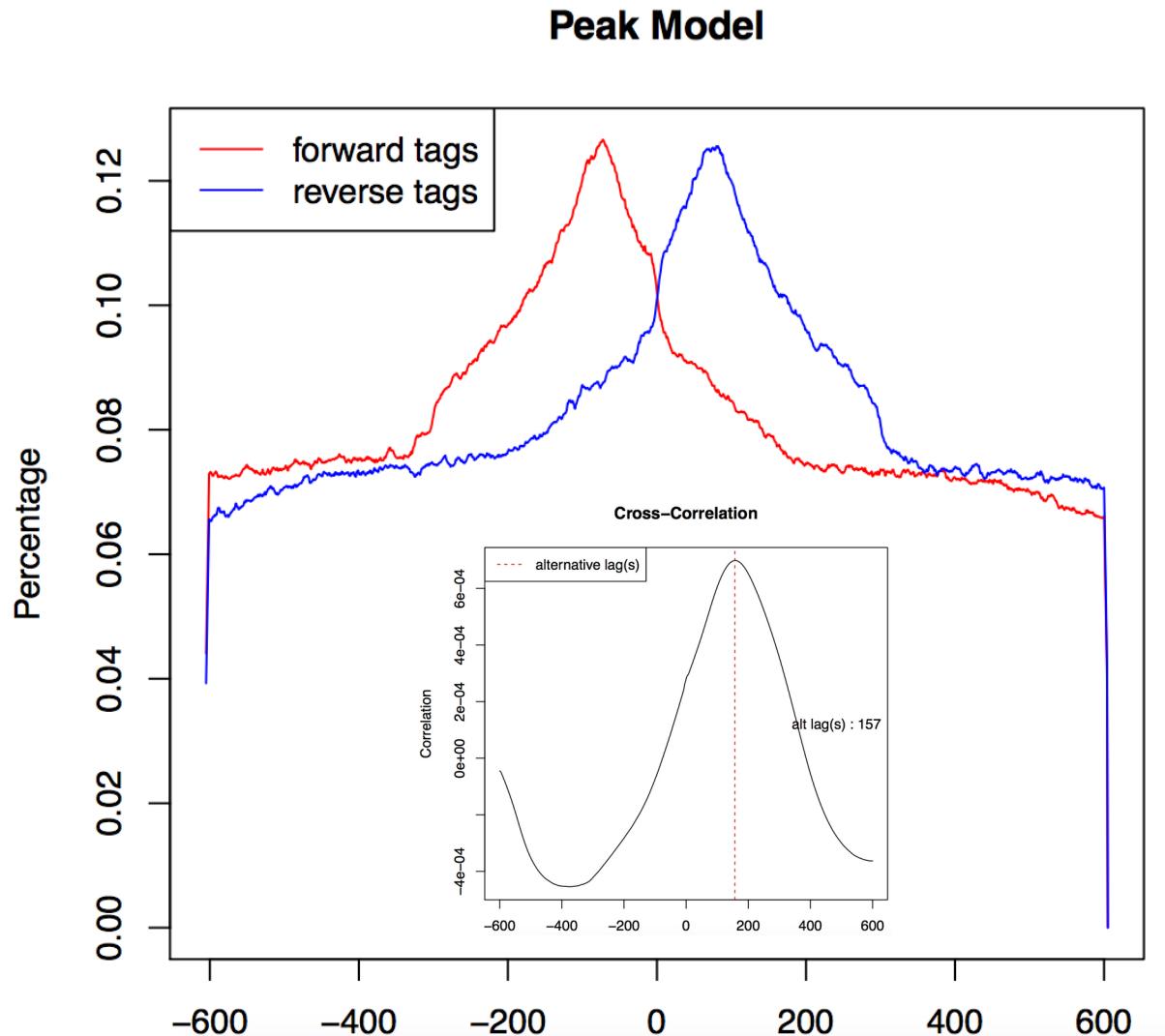
Bailey et al.
2013.
PLoS Comp. Biol.



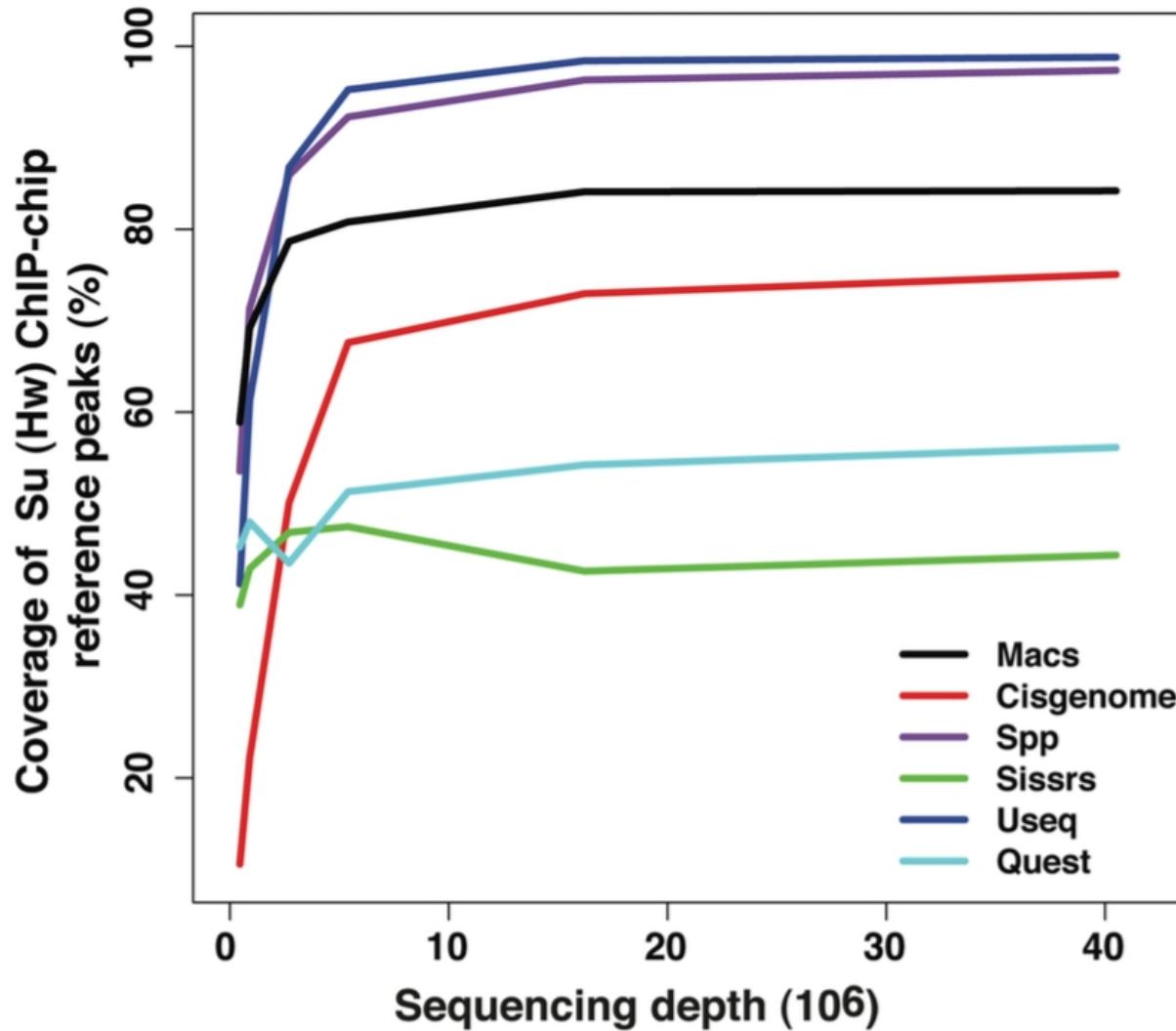
Valouev et al.
2008.
Nat. Methods

Shift modeling: MACS2

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



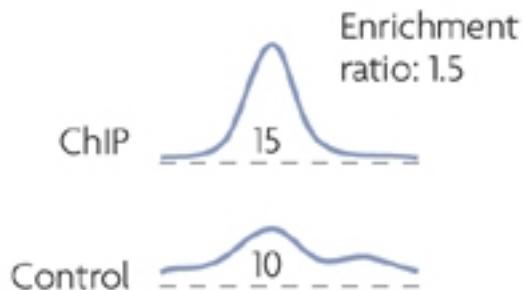
Sequencing depth



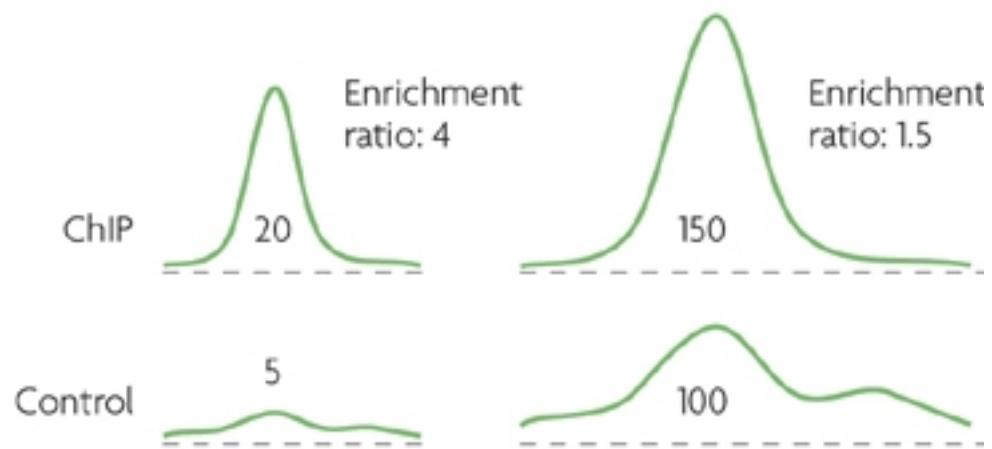
Chen et al. 2013.
Nat. Methods

Sequencing depth

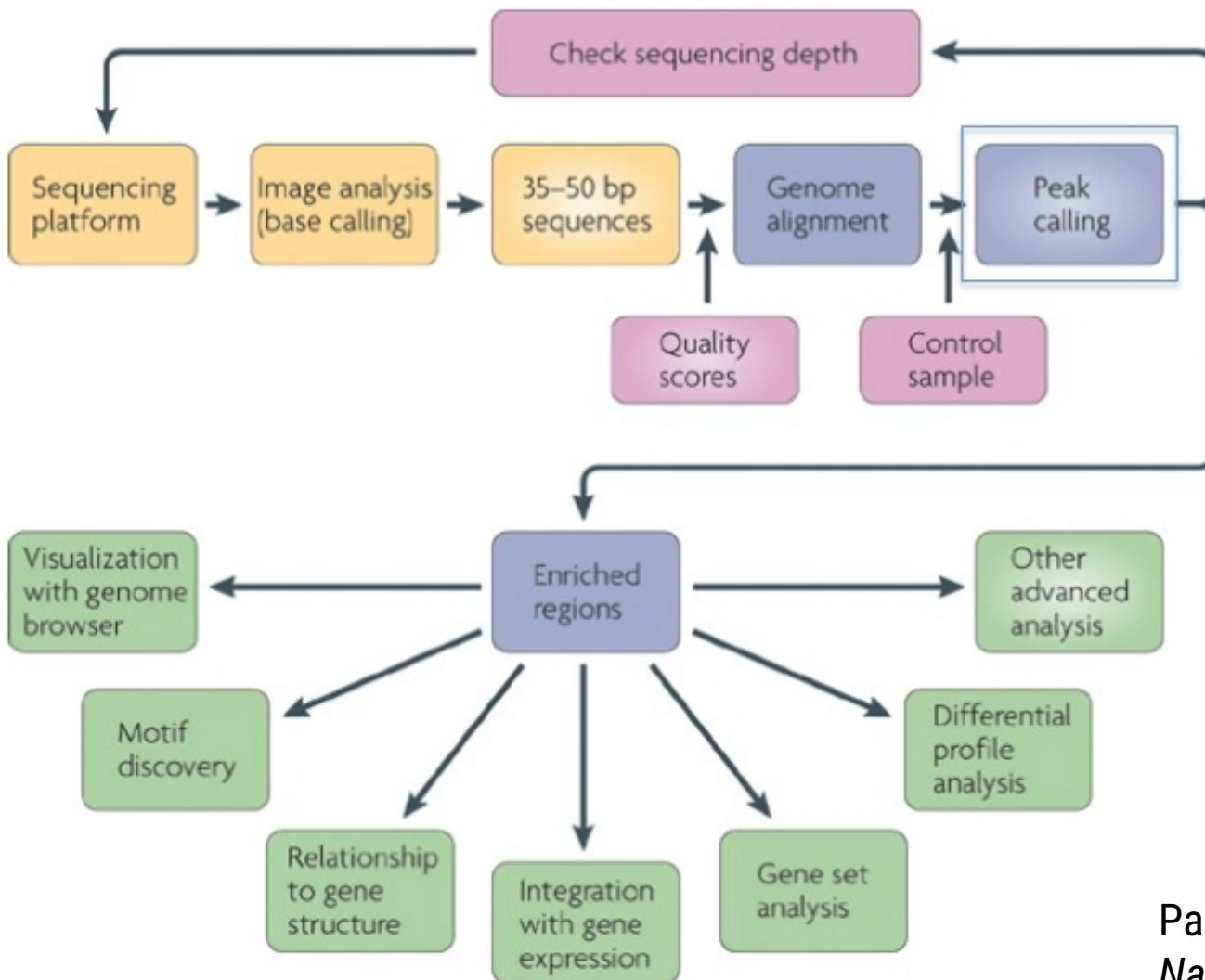
Ba Not statistically significant



Bb Statistically significant



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



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