Chip-seq - Analysis

Aurélien Ginolhac

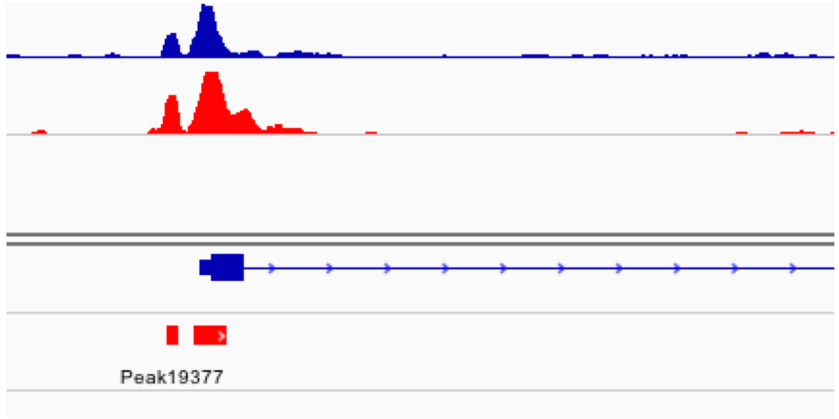
aurelien.ginolhac@uni.lu

Bioinformatics analysis

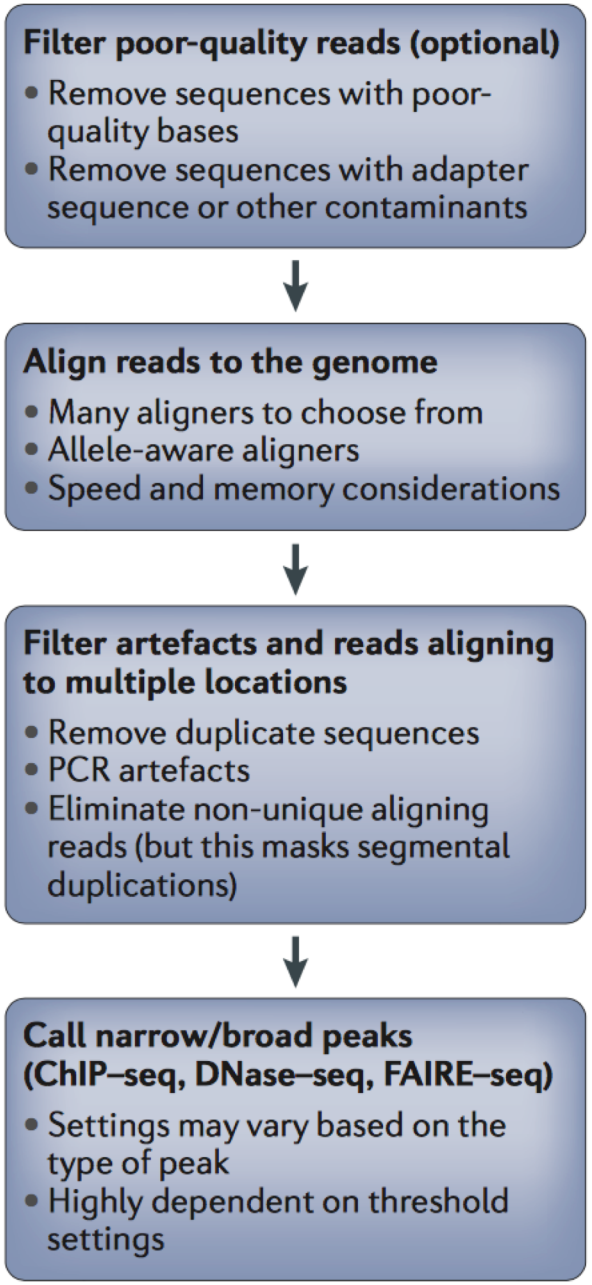
Sequence file

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+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
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

Peak file



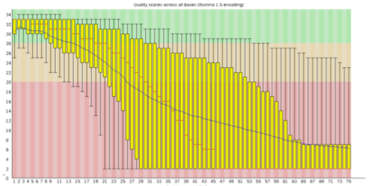
Steps



Furey 2012.
Nat. Genet. Rev.

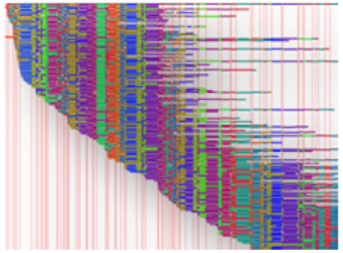
Steps

Quality control



Cleaning

Mapping



Differential peak calling

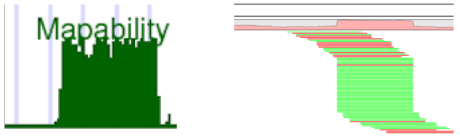
Motif discovery

Peak calling

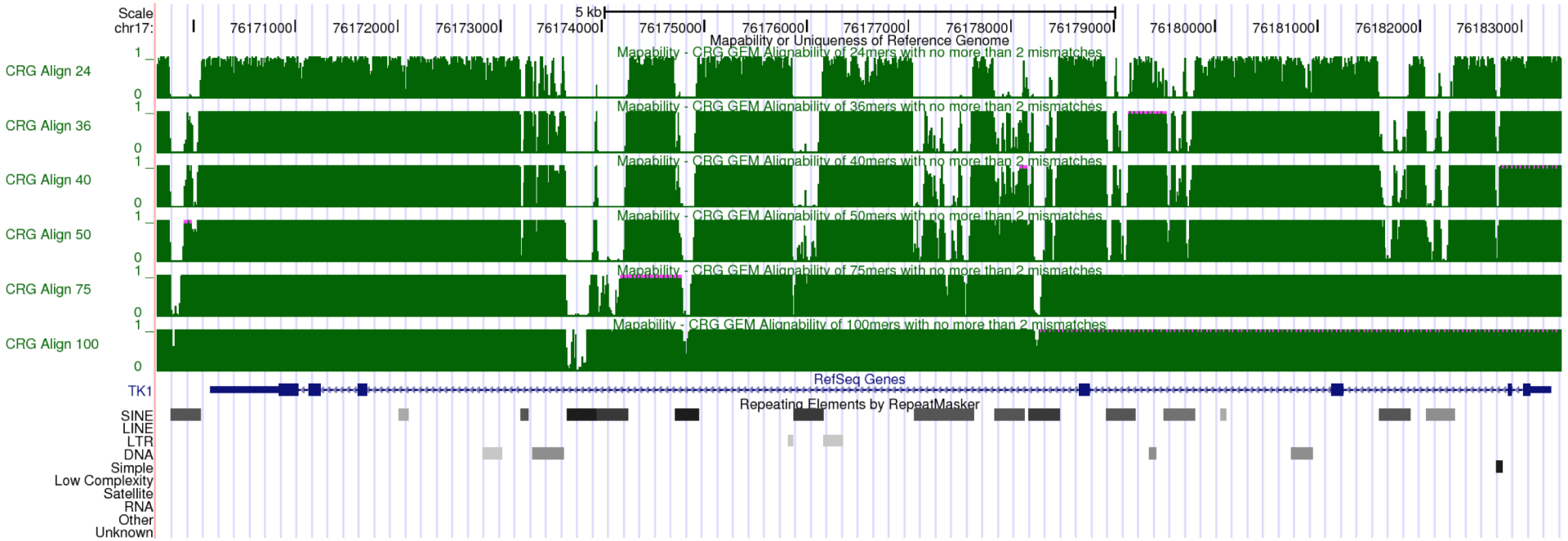
Signal normalization

1 input / 1 IP

Controls



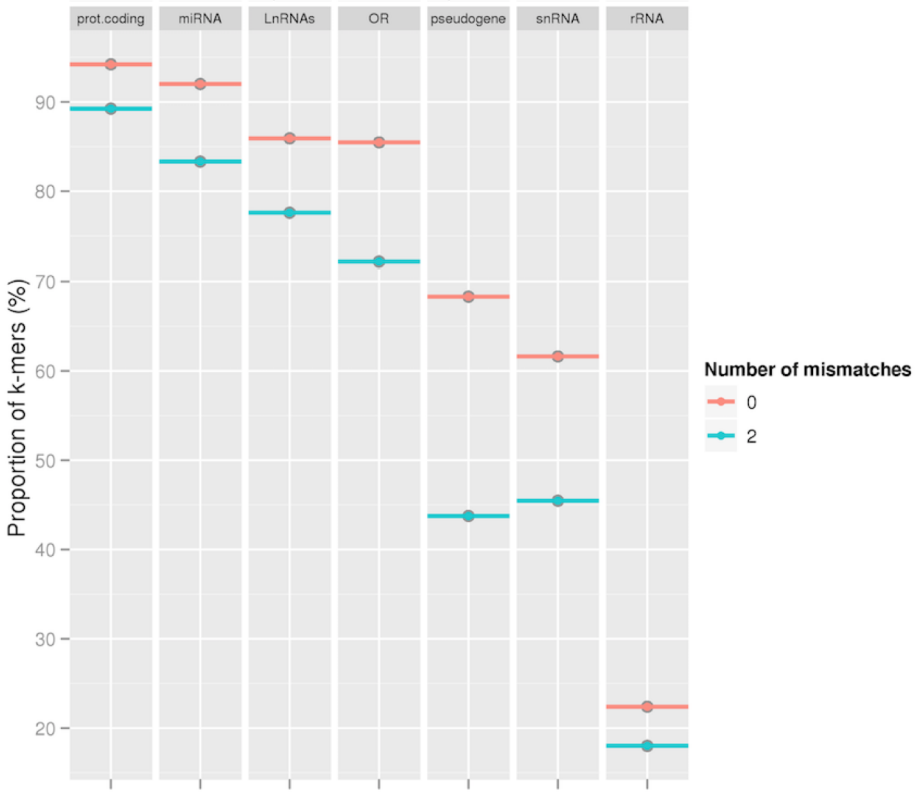
Mappability, causes



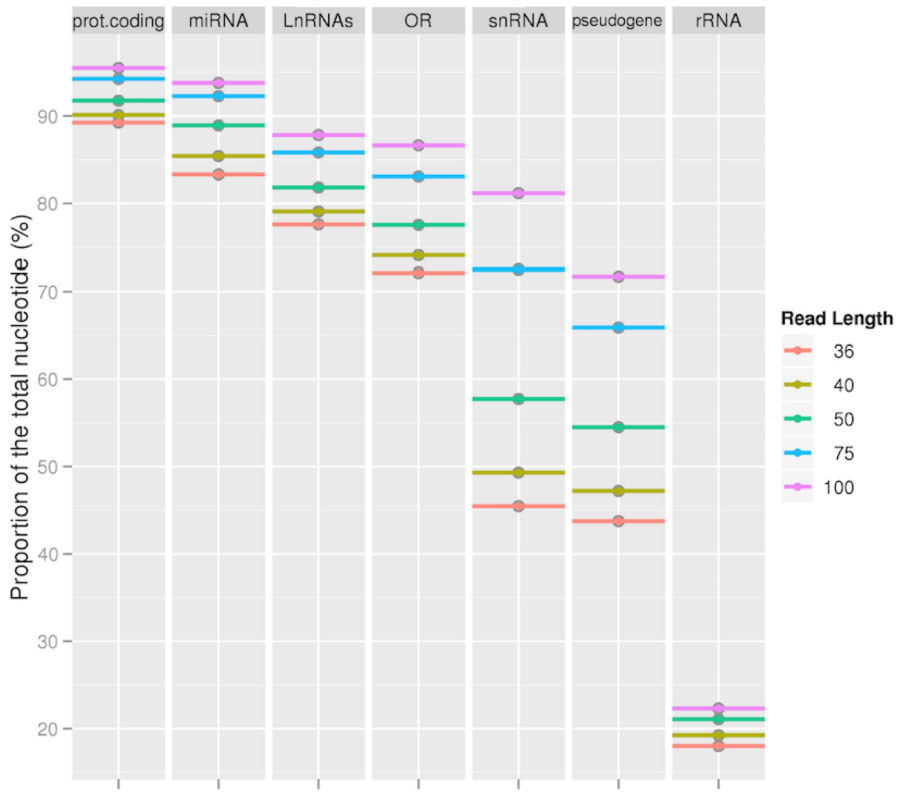
Mappability, consequences

Uniquely mapped

Mismatches

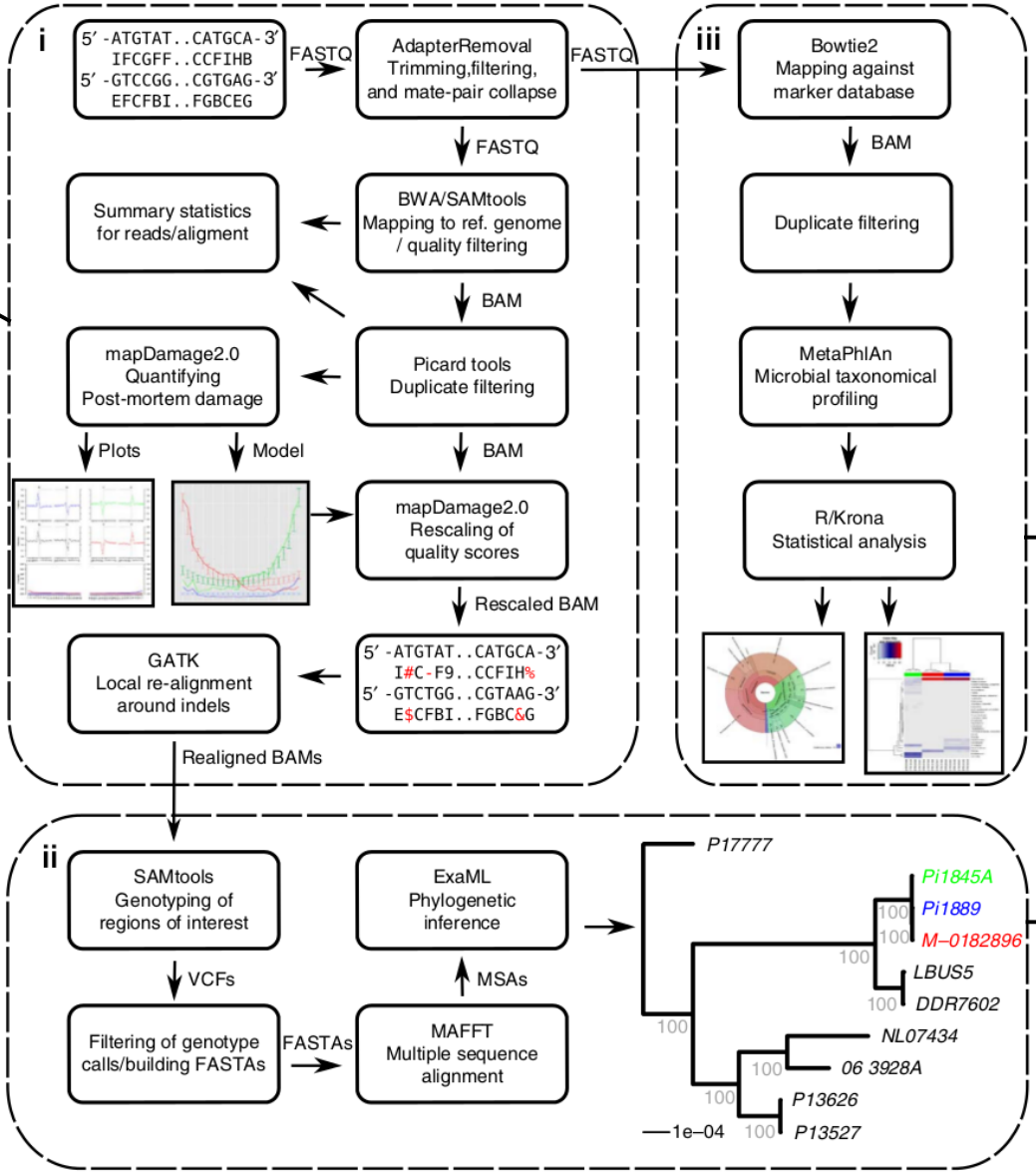


Read Length



Derrien et al. 2013.
PLoS ONE

PALEOMIX



BAM pipeline

Metagenomic pipeline, not described today

Phylogenetic pipeline

Schubert et al. 2014. *Nat. Protocols*

PALEOMIX, usage

```
PALEOMIX - pipelines and tools for NGS data analyses.
Version: v1.0.1

Usage: paleomix <commad> [options]

Commands:
    paleomix help             -- Displays this message.

Pipelines:
    paleomix bam_pipeline     -- Pipeline for trimming and mapping of NGS reads.
    paleomix phylo_pipeline   -- Pipeline for genotyping and phylogenetic
                                 inference from BAMs.

BAM/SAM tools:
    paleomix cleanup          -- Reads SAM file from STDIN, and outputs sorted,
                                 tagged, and filter BAM, for which NM and MD
                                 tags have been updated.
    paleomix coverage         -- Calculate coverage across reference sequences
                                 or regions of interest.
    paleomix depths           -- Calculate depth histograms across reference
                                 sequences or regions of interest.
    paleomix duphist          -- Generates PCR duplicate histogram; for use with
                                 the 'Preseq' tool.
    paleomix rmdup_collapsed  -- Filters PCR duplicates for collapsed paired-
                                 ended reads generated by the AdapterRemoval
                                 tool.

VCF/GTF/BED/Pileup tools:
    paleomix create_pileup    -- Creates tabixed indexed pileup for a (sparse)
                                 set of BED coordinates.
    paleomix gtf_to_bed       -- Convert GTF file to BED files grouped by
                                 feature (coding, RNA, etc).
    paleomix sample_pileup    -- Randomly sample sites in a pileup to generate a
                                 FASTA sequence.
    paleomix vcf_filter       -- Quality filters for VCF records, similar to
                                 'vcfutils.pl varFilter'.
    paleomix vcf_to_fasta     -- Create most likely FASTA sequence from tabix-
                                 indexed VCF file.

Misc tools:
    paleomix cat              -- Generalized cat command for gz, bz2 and
                                 uncompressed files.
```

PALEOMIX, scheduling

```
Scheduling:
  --bowtie2-max-threads=BOWTIE2_MAX_THREADS
                        Maximum number of threads to use per BWA instance [4]
  --bwa-max-threads=BWA_MAX_THREADS
                        Maximum number of threads to use per BWA instance [4]
  --max-threads=MAX_THREADS
                        Maximum number of threads to use in total [48]
  --dry-run             If passed, only a dry-run in performed, the dependency
                        tree is printed, and no tasks are executed.
```

Uses all cores by default; easy to set to use *e.g.* half of the cores on a per-server basis

Config file with default values can be created
~/.pypeline/bam_pipeline.ini

Hands On with PALEOMIX

Two exercises,

1) run an example on simulated reads

options

reference, FASTA

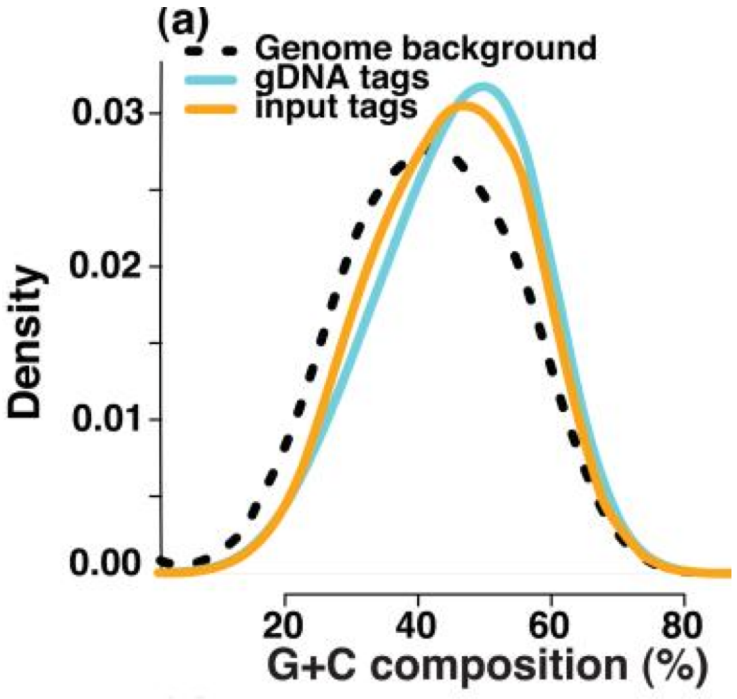
data, FASTQ

Makefile (YAML format)

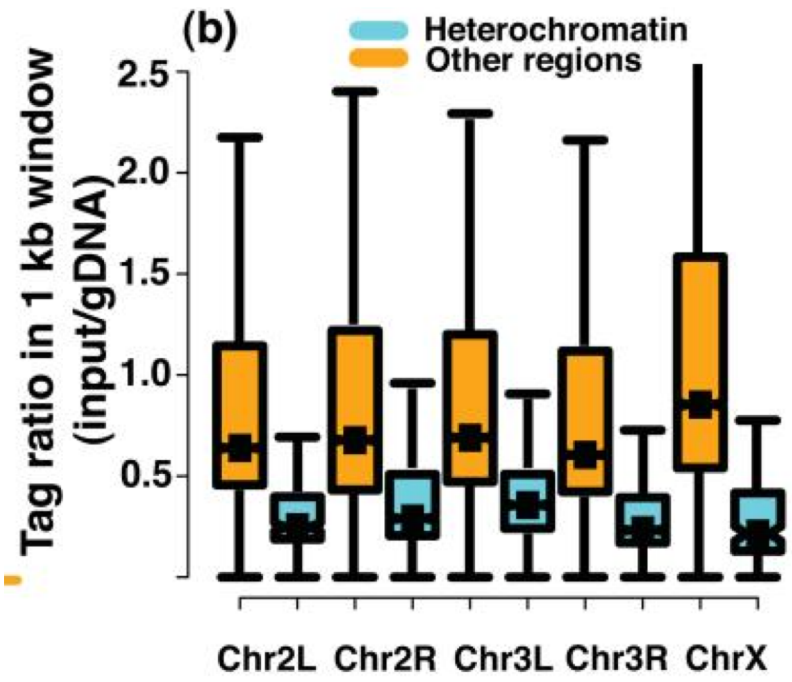
```
Options:
  Platform: Illumina
  QualityOffset: 33
  SplitLanesByFilenames: yes
  CompressionFormat: bz2
  Aligners:
    Program: BWA
    BWA:
      MinQuality: 30
      FilterUnmappedReads: yes
      UseSeed:    yes
    Bowtie2:
      MinQuality: 0
      FilterUnmappedReads: yes
      --very-sensitive:
  PCRDuplicates: filter
  RescaleQualities: yes
  mapDamage:
    --downsample: 100000
  Features:
    - Realigned BAM  # Generate indel-realigned BAM using the GATK Indel realigner
    - mapDamage      # Generate mapDamage plot for each (unrealigned) library
    - Coverage       # Generate coverage information for the raw BAM (wo/ indel realignment)
    - Depths         # Generate histogram of number of sites with a given read-depth
    - Summary        # Generate target summary (uses statistics from raw BAM)

Prefixes:
  rCRS:
    Path: 000_prefixes/rCRS.fasta
    Label: "mitochondrial"

ExampleProject:
  Synthetic_Sample_1:
    ACGATA:
      Lane_1: 000_data/ACGATA_L1_R{Pair}_*.fastq.gz
    GCTCTG:
      Lane_1: 000_data/GCTCTG_L1_R1_*.fastq.gz
```

Input *versus* IP, biases

GC content

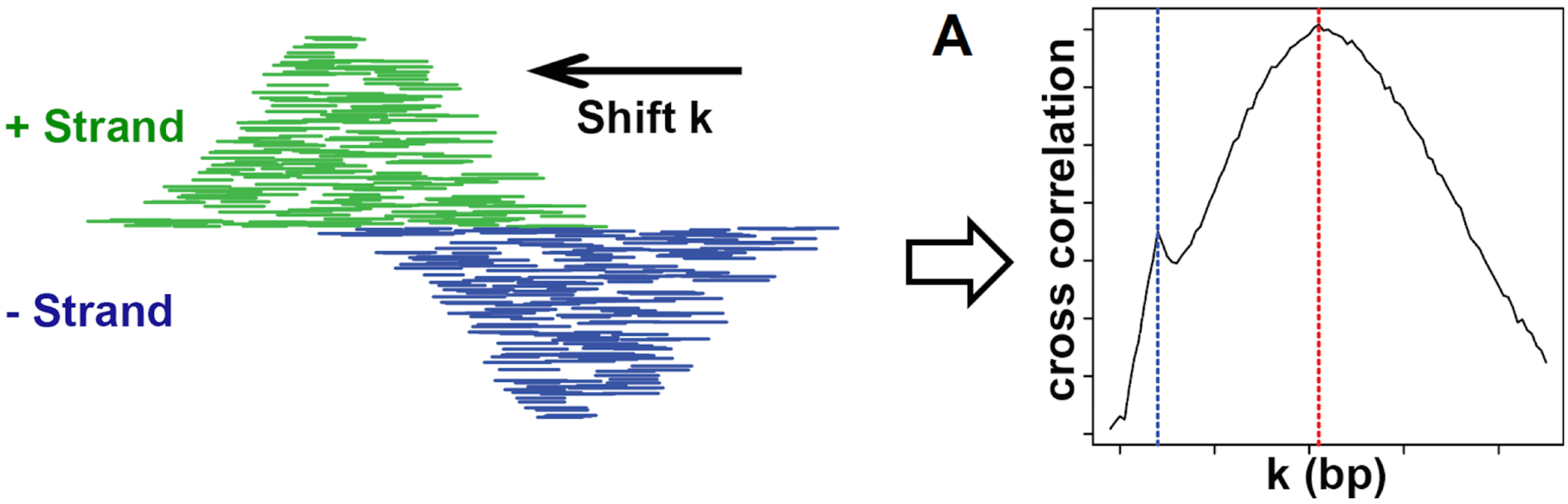


fragmentation

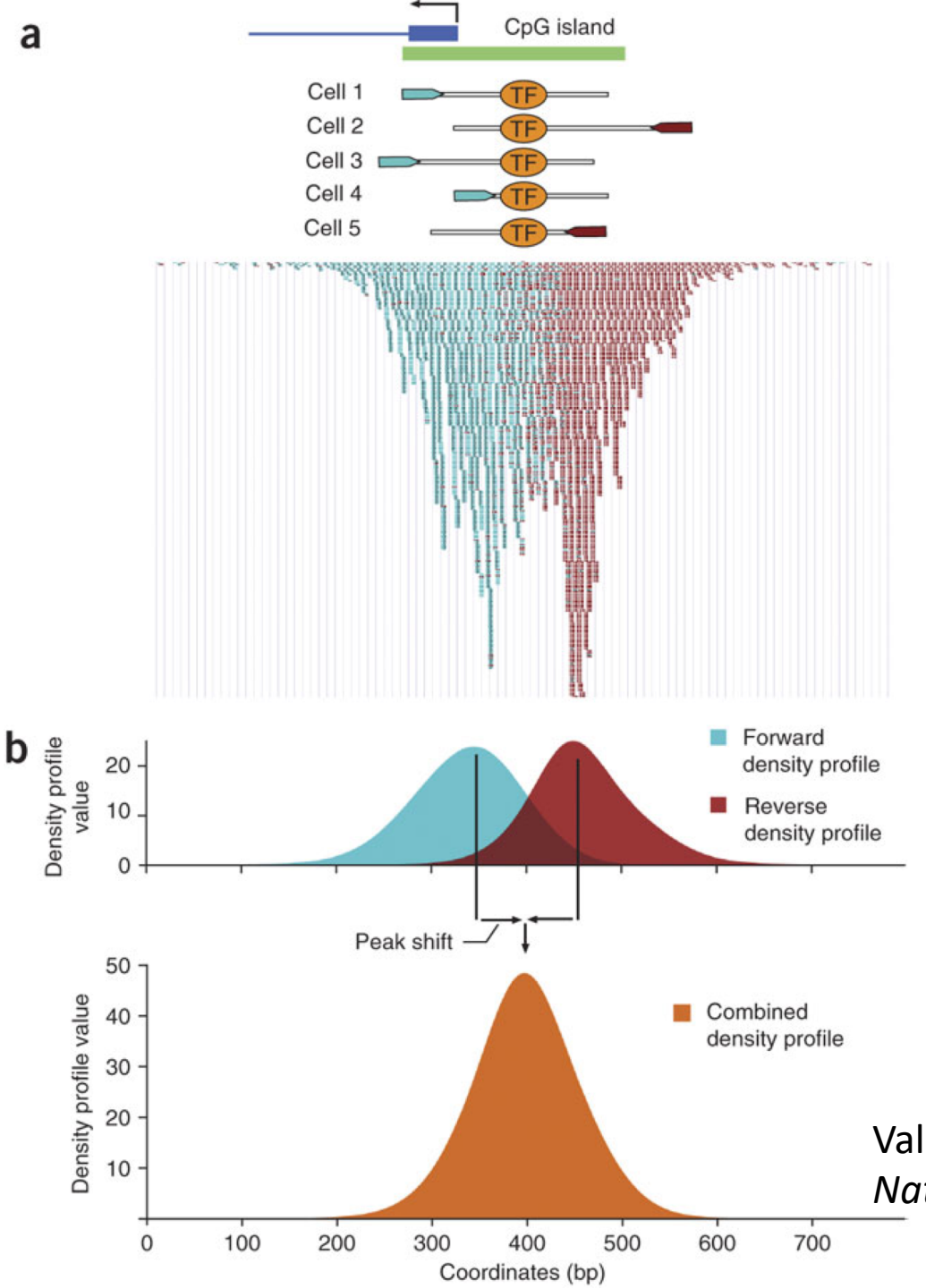


Chen et al. 2013.
Nat. Methods

Peal calling, infer the shift size



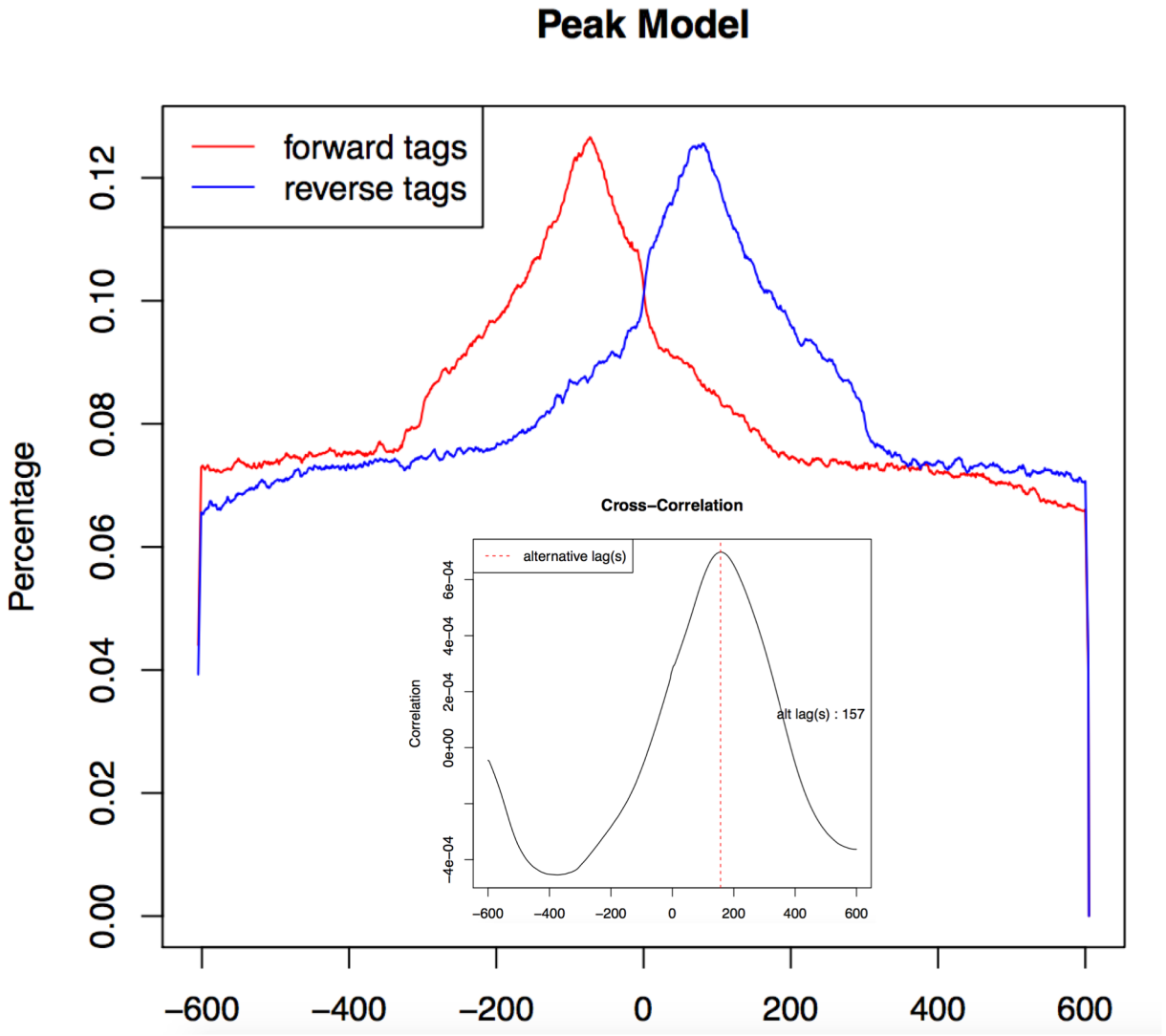
Bailey et al. 2013.
PLoS Comp. Biol.



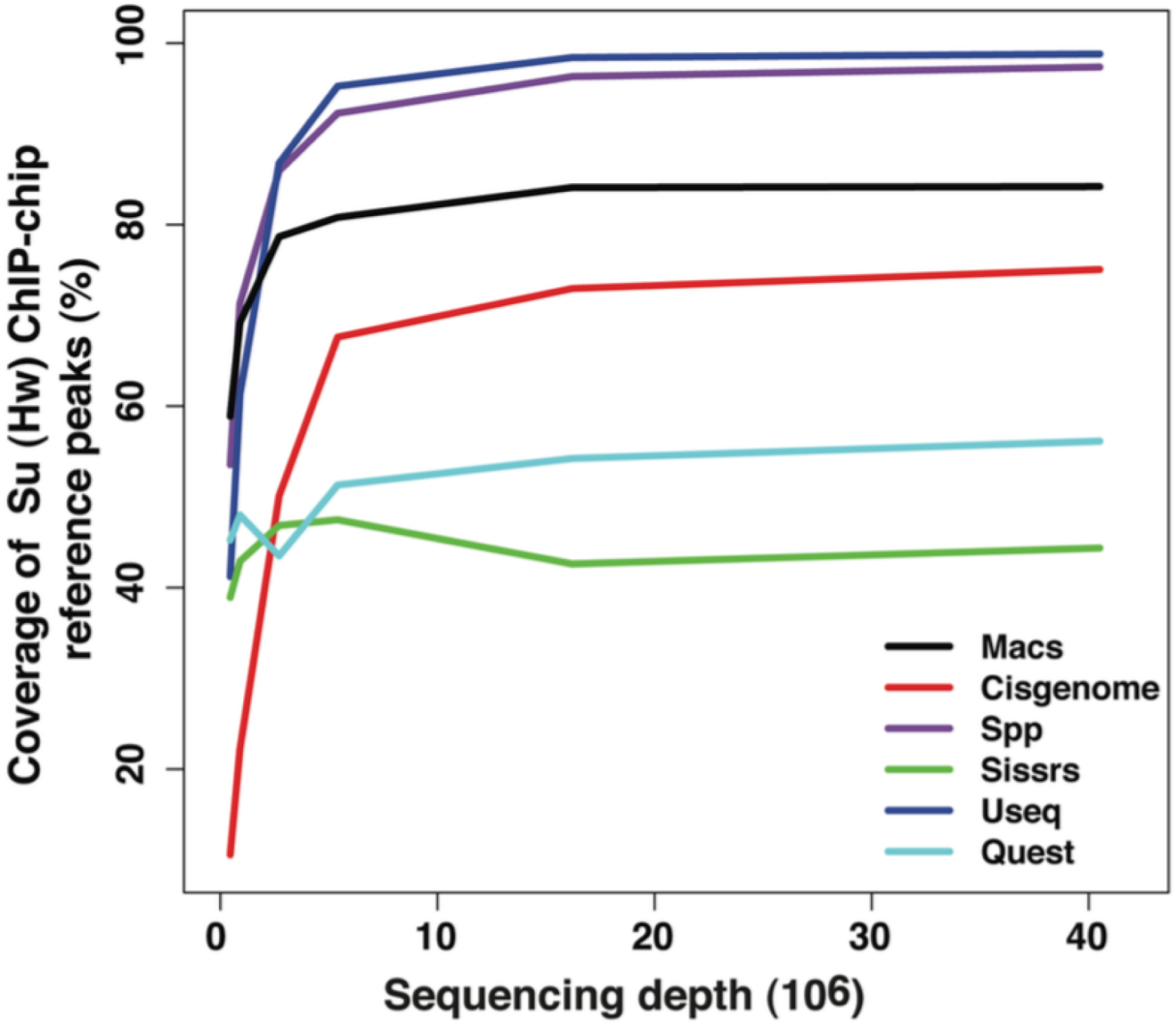
Valouev et al. 2008.
Nat. Methods

Shit 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

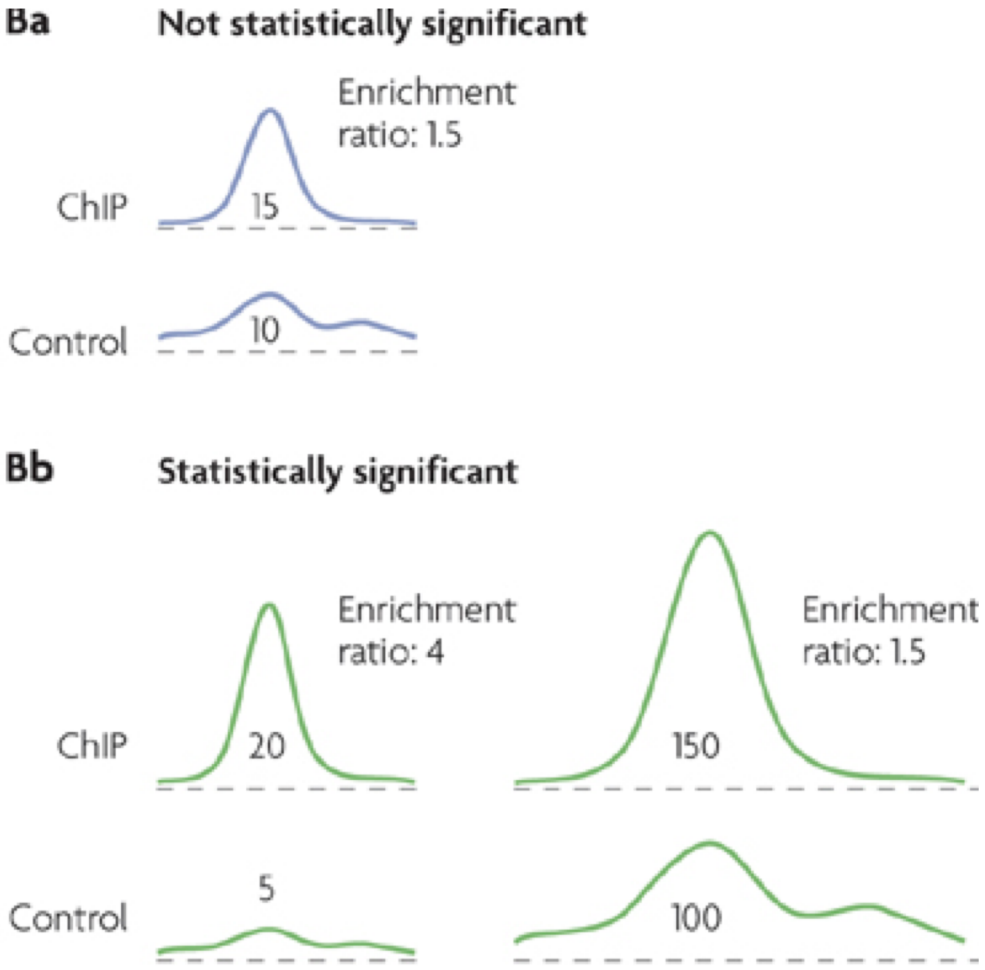


Sequencing depth



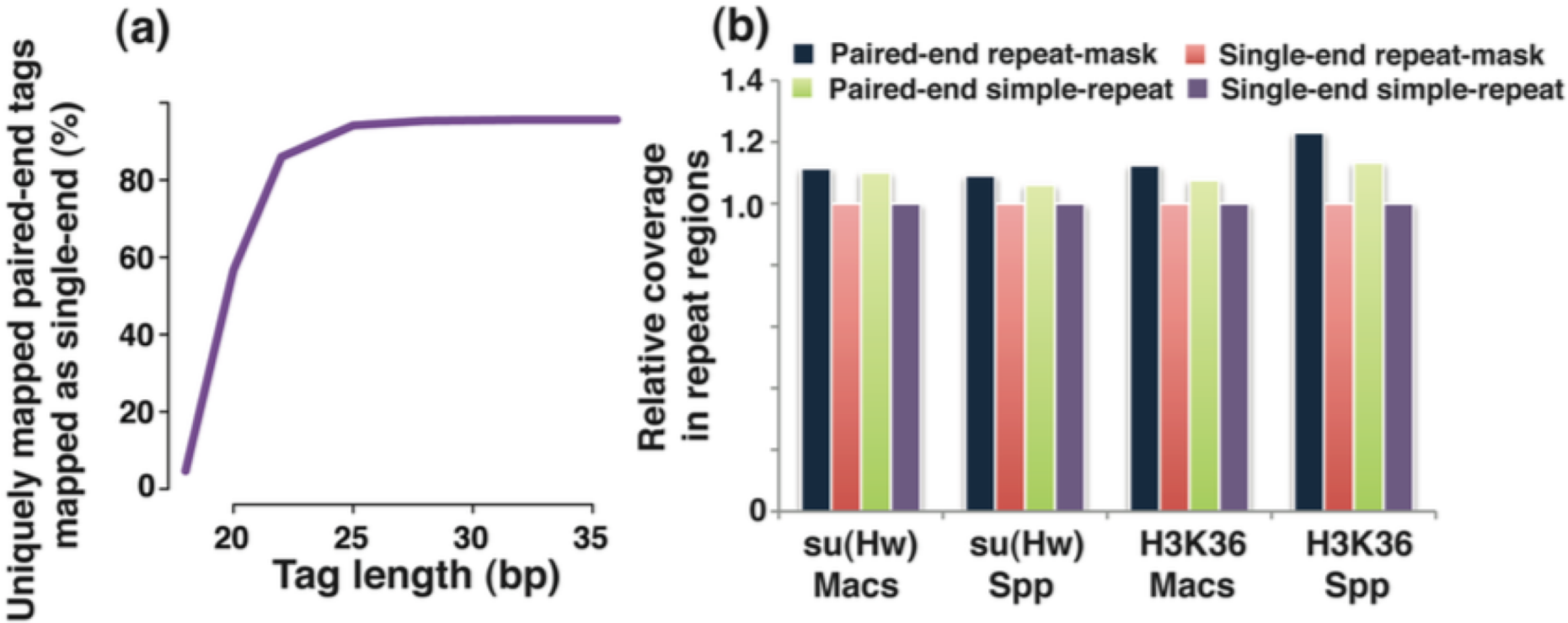
Chen et al. 2013.
Nat. Methods

Sequencing depth



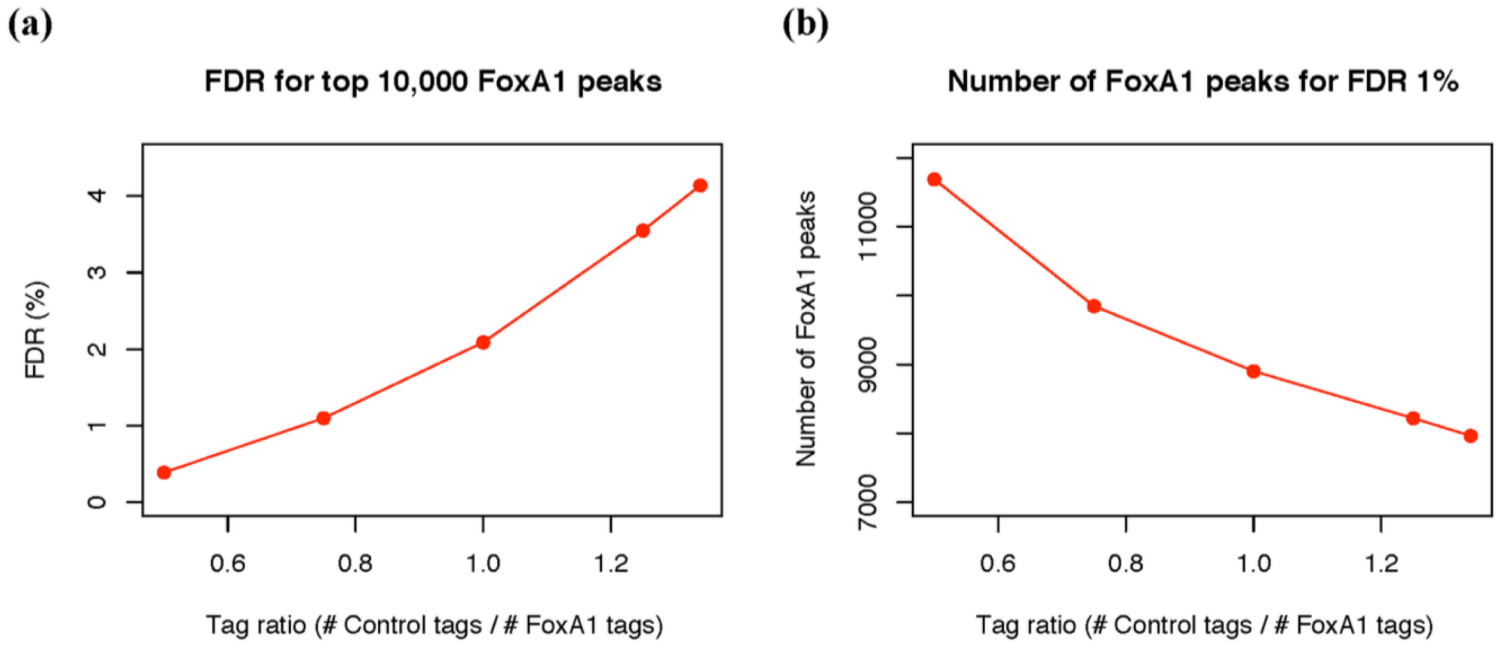
Park et al. 2009.
Nat. Rev. Genet.

Sequencing Paired/Single reads



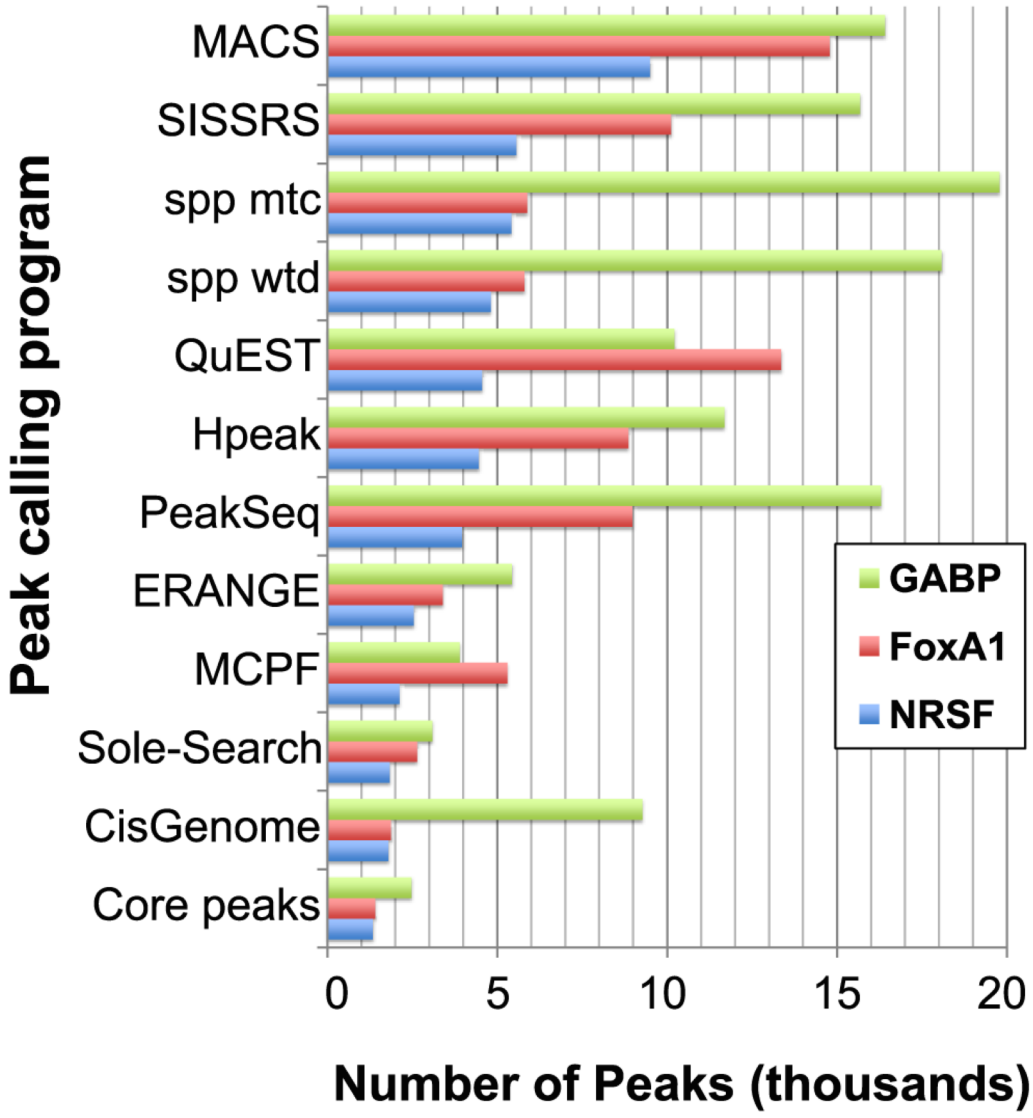
Chen et al. 2013.
Nat. Methods

Figure S5. The influence of unbalanced tag numbers between ChIP and control experiments in ChIP-Seq data analysis. With the increase of tag ratio between control and FoxA1 ChIP experiments, (a) to identify the same number of FoxA1 peaks results in higher FDR, and (b) less FoxA1 peaks are identified under the same FDR cutoff. The analysis is based on random sampling of control tags.



Zhang et al. 2008.
Gen. Biol.

Many different software

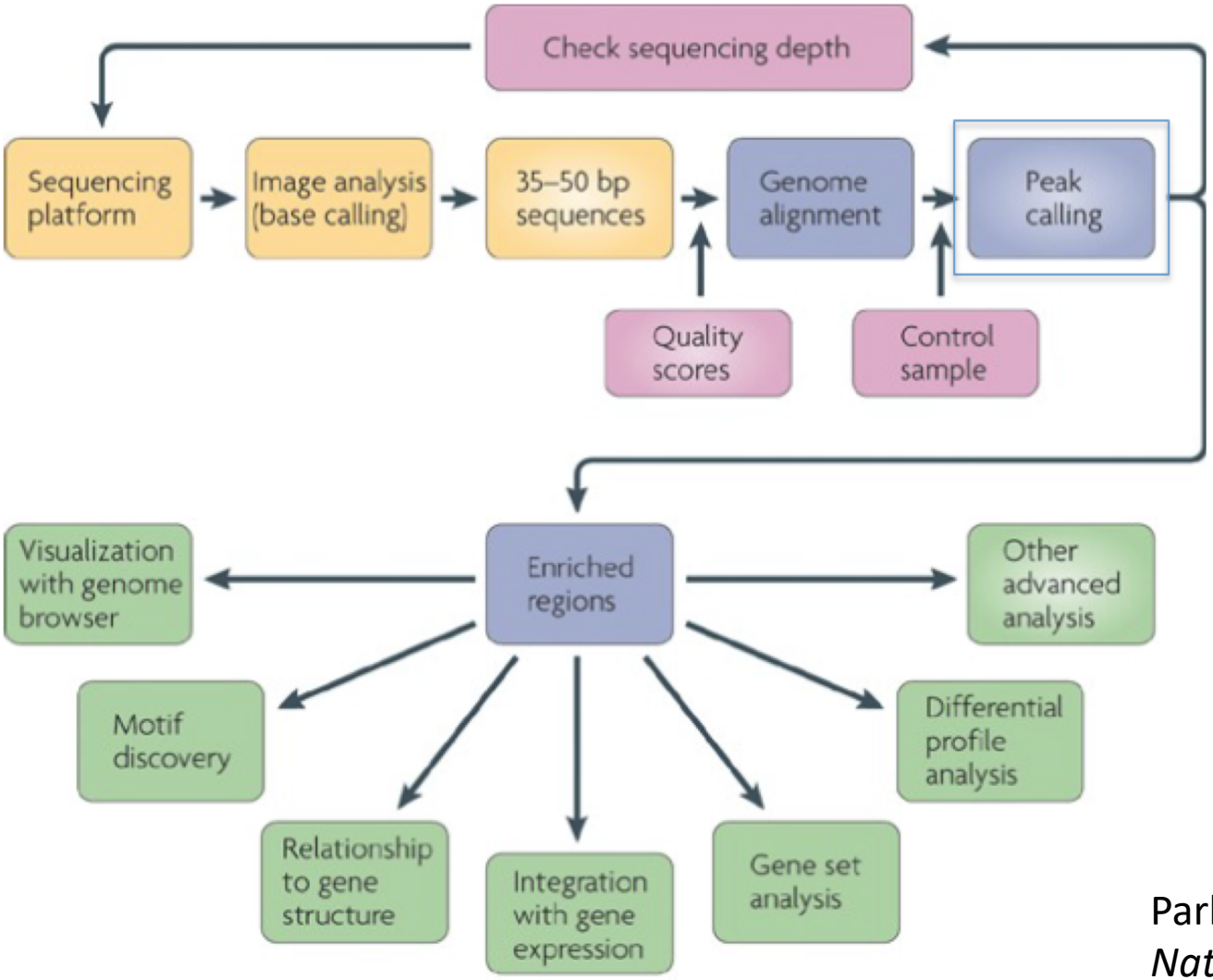


Wilbanks & Facciotti 2010.
PLoS ONE

Many different softwares

FoxA1	CisGenome	Sole-Seach	ERANCE	MCPF	wtd	mtc	Hpeak	PeakSeq	SISSRS	QuEST	MACS
CisGenome	X	58	52	33	30	30	21	21	18	13	12
Sole-Search	82	X	67	47	44	44	30	29	27	18	18
ERANGE	96	86	X	58	56	55	38	38	34	22	23
MCPF	93	95	90	X	74	72	56	57	52	31	35
wtd	93	98	95	81	X	93	65	63	55	36	39
mtc	93	97	94	80	95	X	66	63	56	37	39
Hpeak	100	100	100	94	99	99	X	86	79	51	55
PeakSeq	100	100	100	96	98	96	88	X	80	50	58
SISSRS	96	100	98	97	96	96	88	88	X	54	61
QuEST	94	91	89	78	84	85	77	73	73	X	60
MACS	99	100	100	99	99	99	92	96	91	67	X

Wilbanks & Facciotti 2010.
PLoS ONE



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

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