Epigenetics and ChIP-seq

CSAMA 2015, Brixen
16. 06. 2015.

Aleksandra Pekowska
aleksandra.pekowska@embl.de

EMBL
40 YEARS | 1974–2014

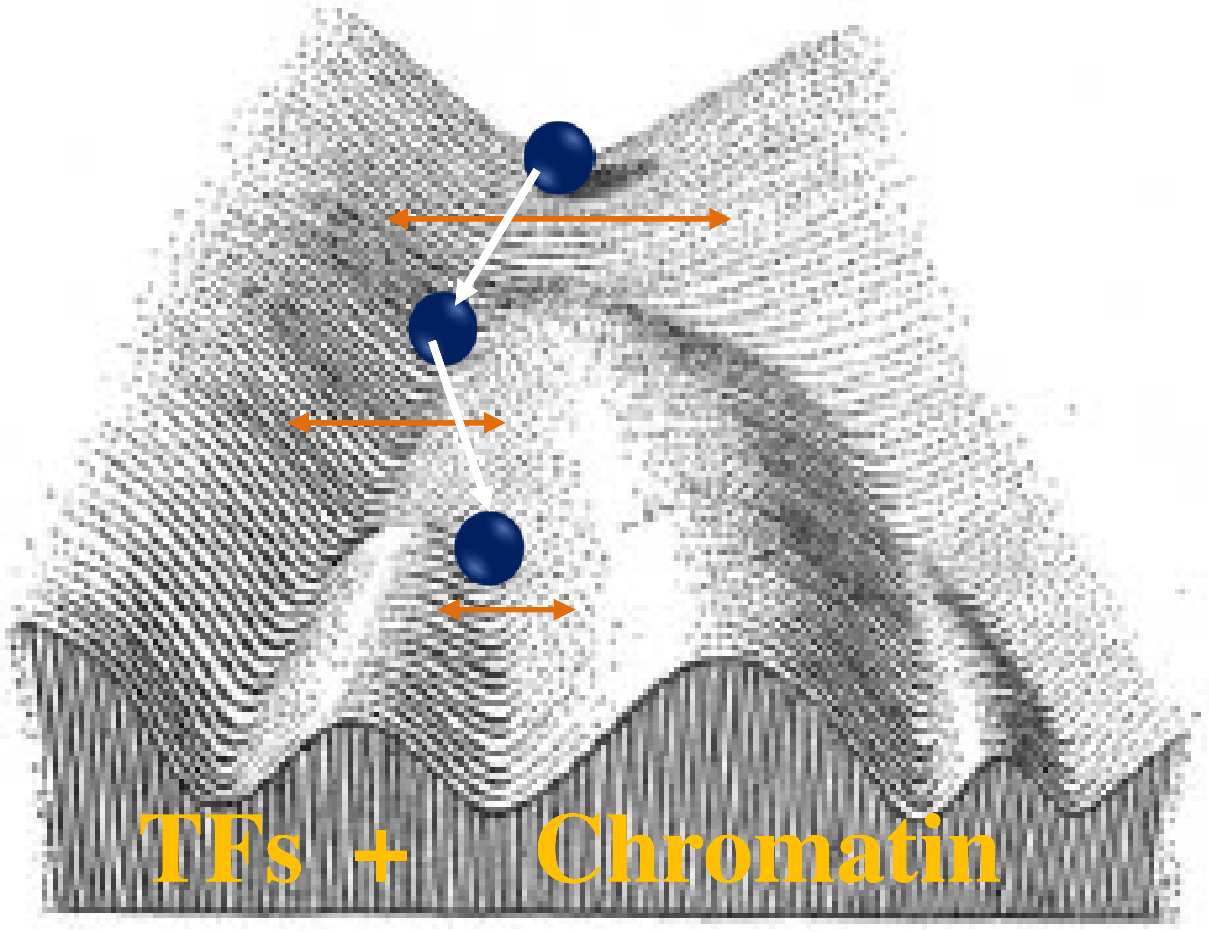
Outline of the lecture

Purpose: introduce basic steps and key considerations in ChIP-seq analysis

1. **Epigenetics - fundamental concepts**
2. **The ChIP-seq method**
3. **What kind of information can we obtain from ChIP-seq?**
4. **Study design**
5. **ChIP-seq analysis workflow:**
    a. Preprocessing
    b. Quality controls
    c. Isolation of enriched regions
    d. Analysis of enriched regions
    e. Visualization
    f. Average profiles
    g. Comparative analysis of enriched regions

Epigenetics - inheritance, but not as we know it

Non-genic memory of function transmitted from generation to generation (A. Bird)

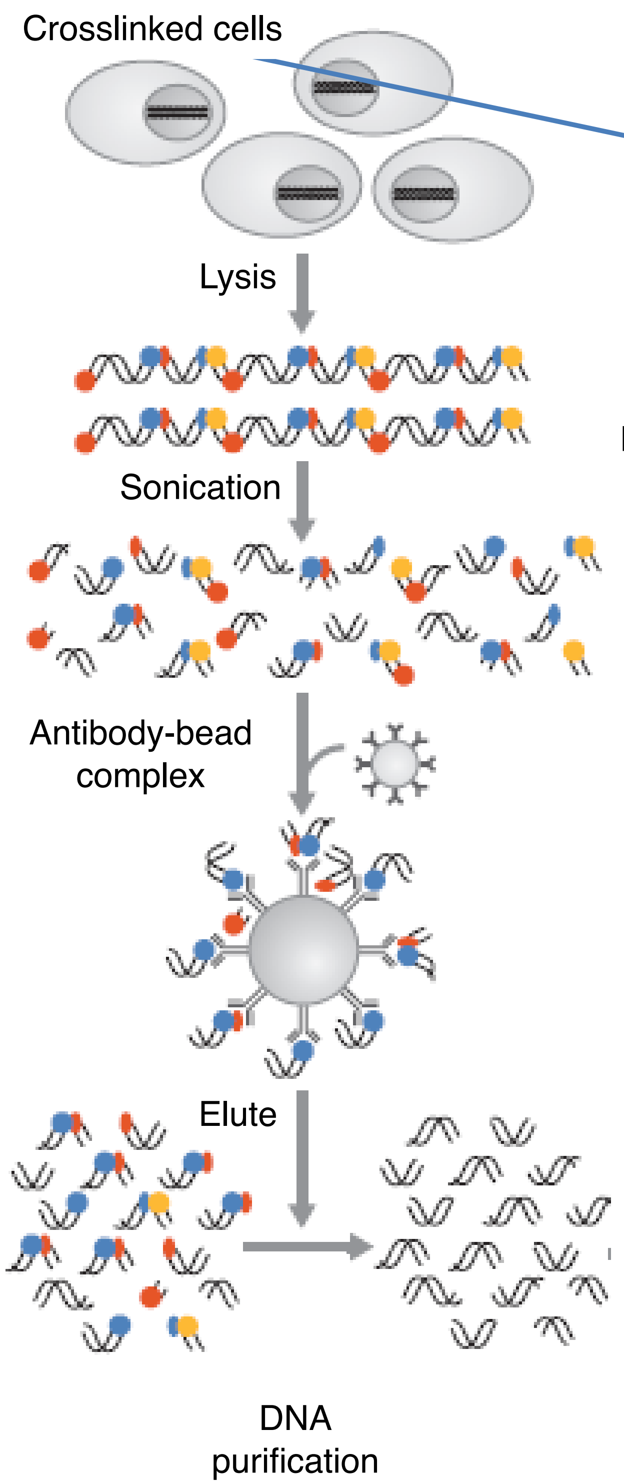


Adapted from Conrad Hal Waddington (1942)

Factors which are analysed:

- DNA methylation
- nucleosome occupancy
- **histone modifications**
- transcription factors
- RNA-polymerases
- chromatin modifying enzymes

Chromatin Immunoprecipitation



Crosslinked cells

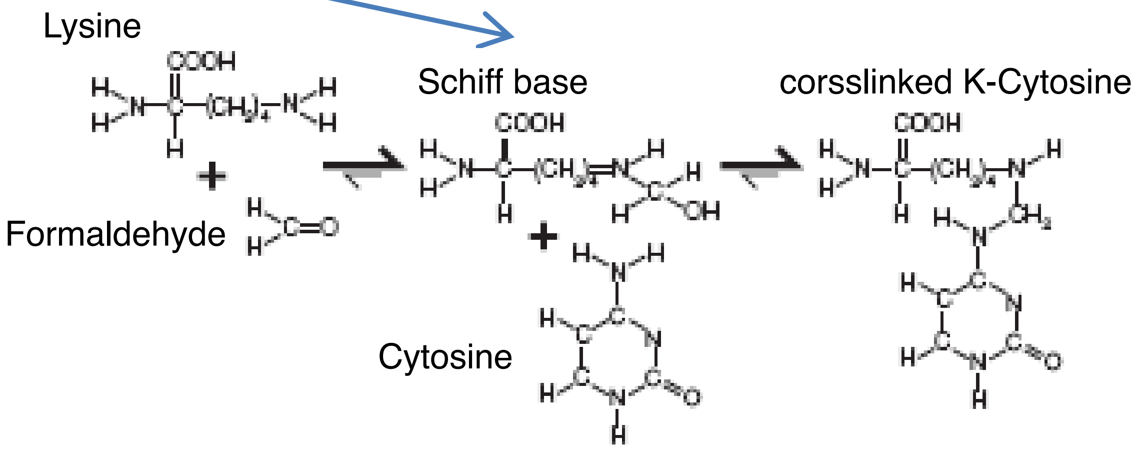
Lysis

Sonication

Antibody-bead complex

Elute

DNA purification



Lysine

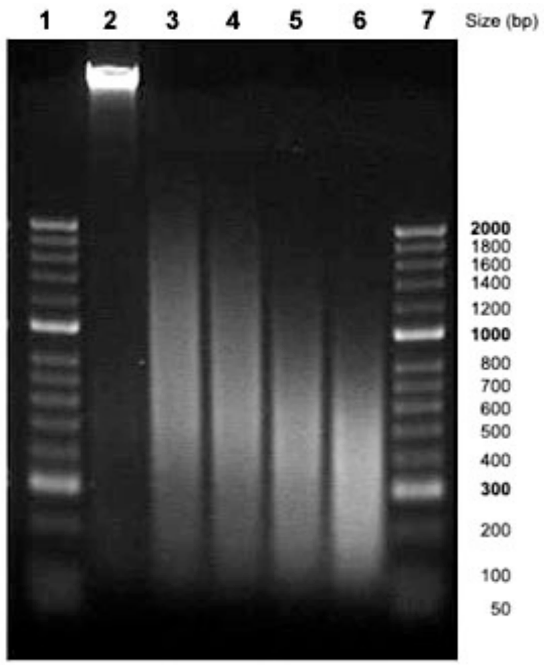
Formaldehyde

Schiff base

corsslinked K-Cytosine

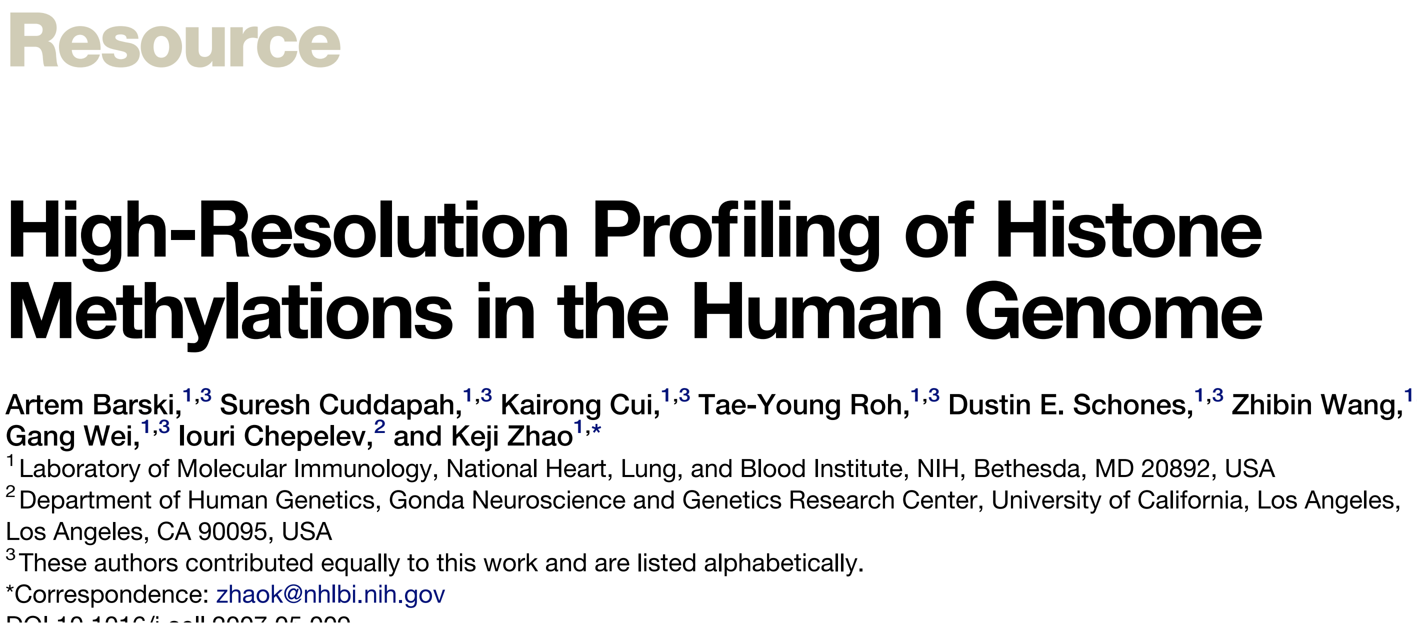
Cytosine

Sonication



Adapted from Massie 2008, Ran 2003, Park 2009.

What kind of information can we obtain from the ChIP-seq experiments ?



Resource

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1] Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]

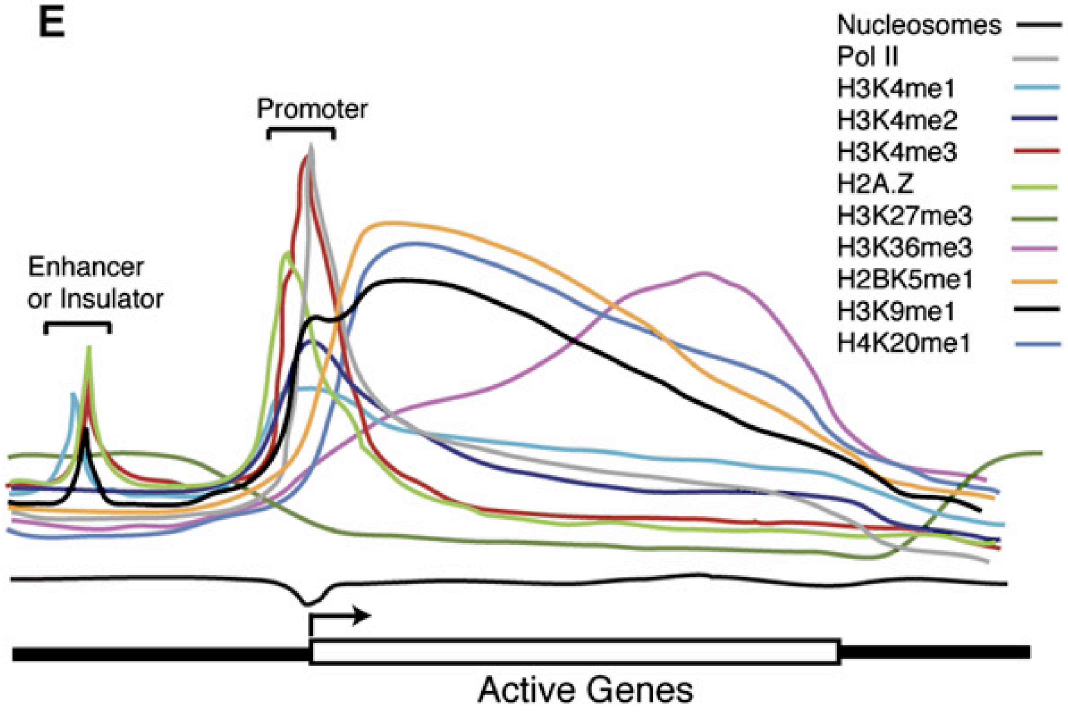
[1] Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA

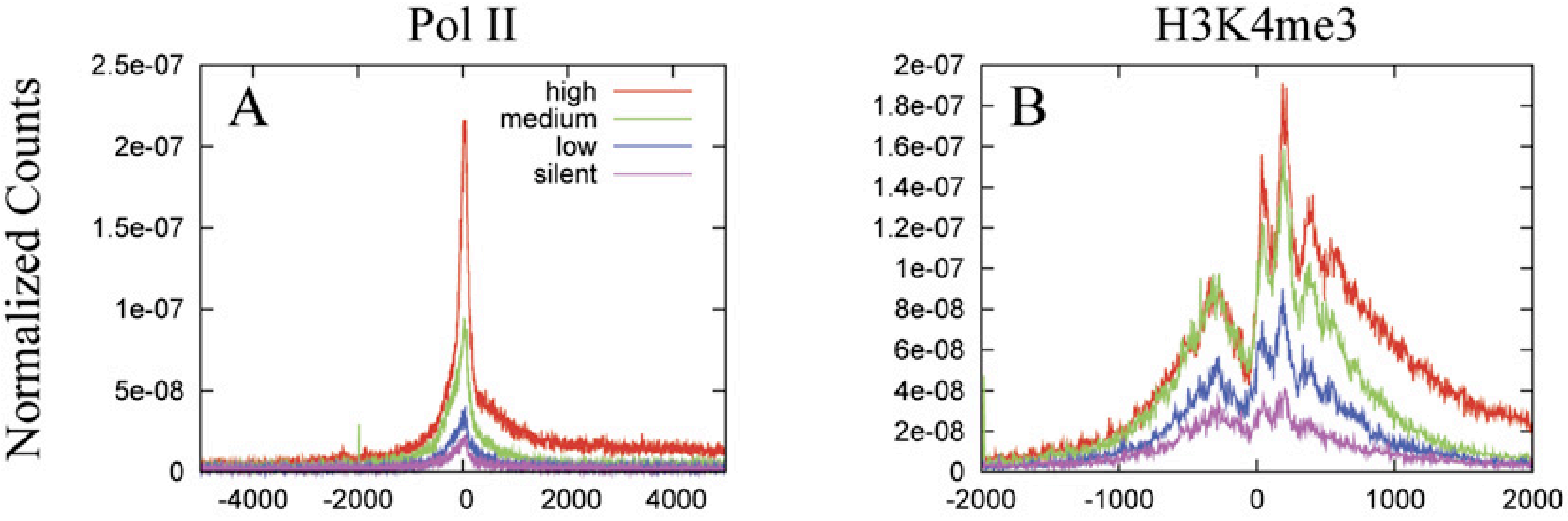
[2] Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles, Los Angeles, CA 90095, USA

[3] These authors contributed equally to this work and are listed alphabetically.

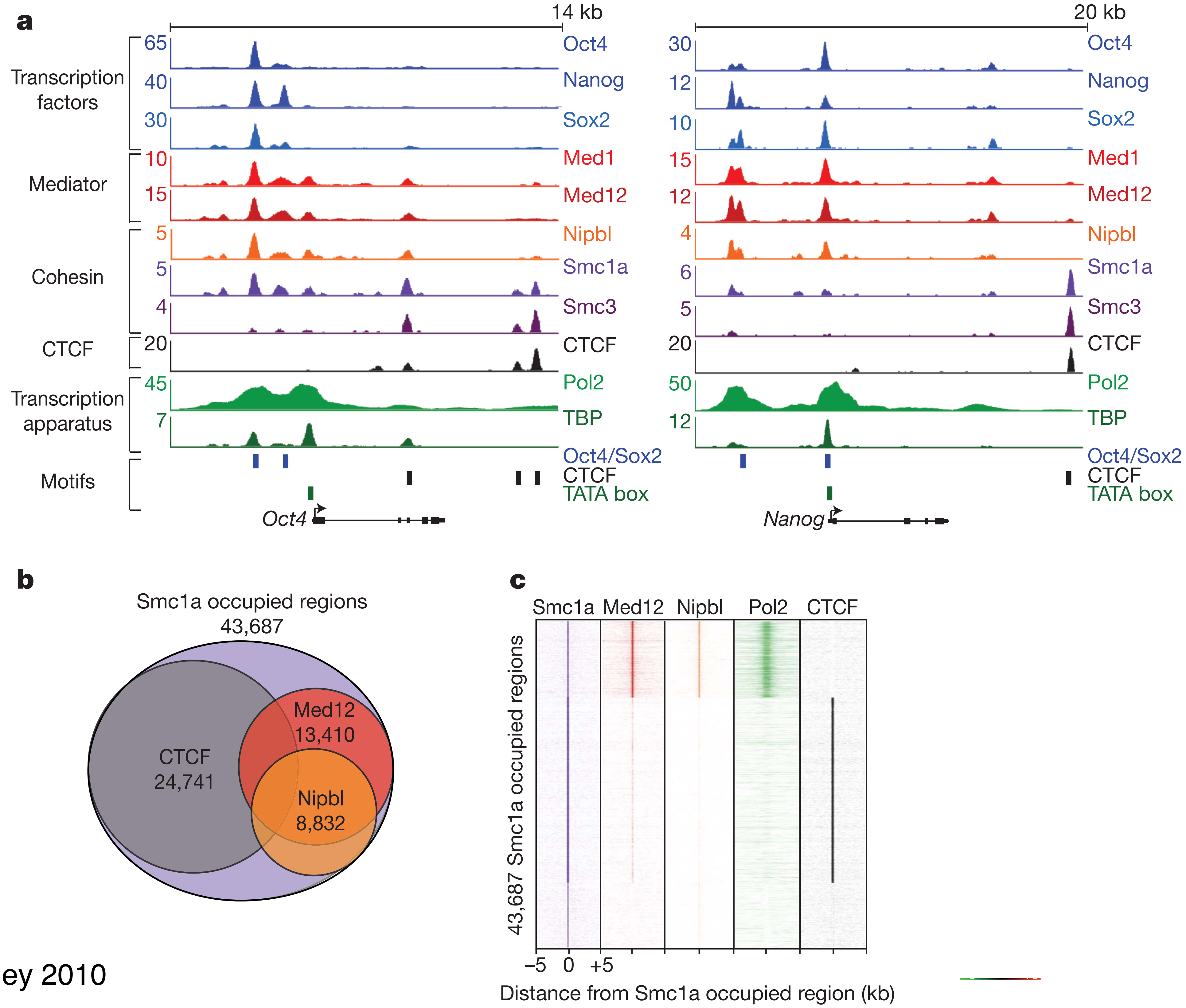
*Correspondence: zhaok@nhlbi.nih.gov

DOI 10.1016/j.cell.2007.05.009



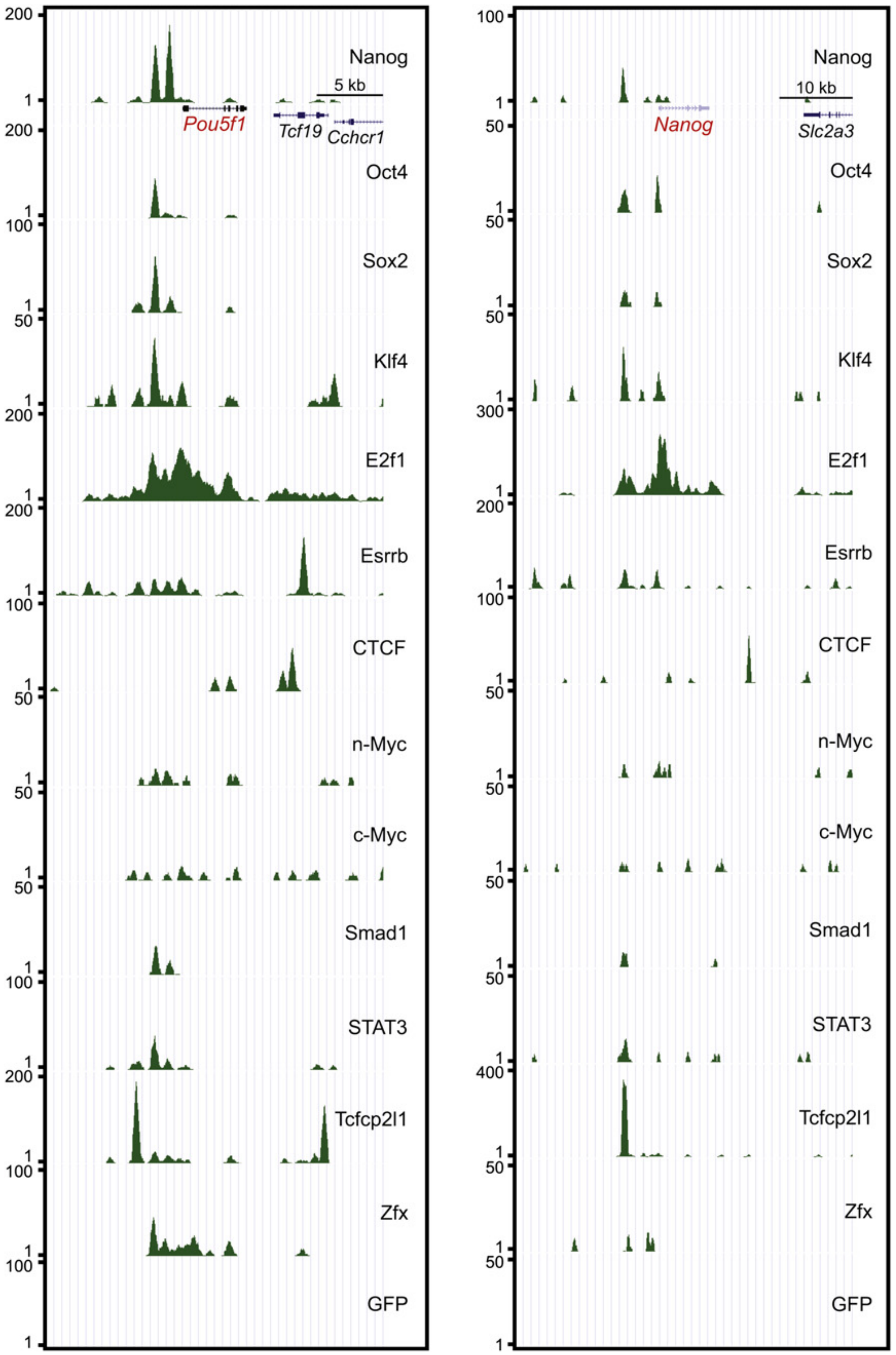


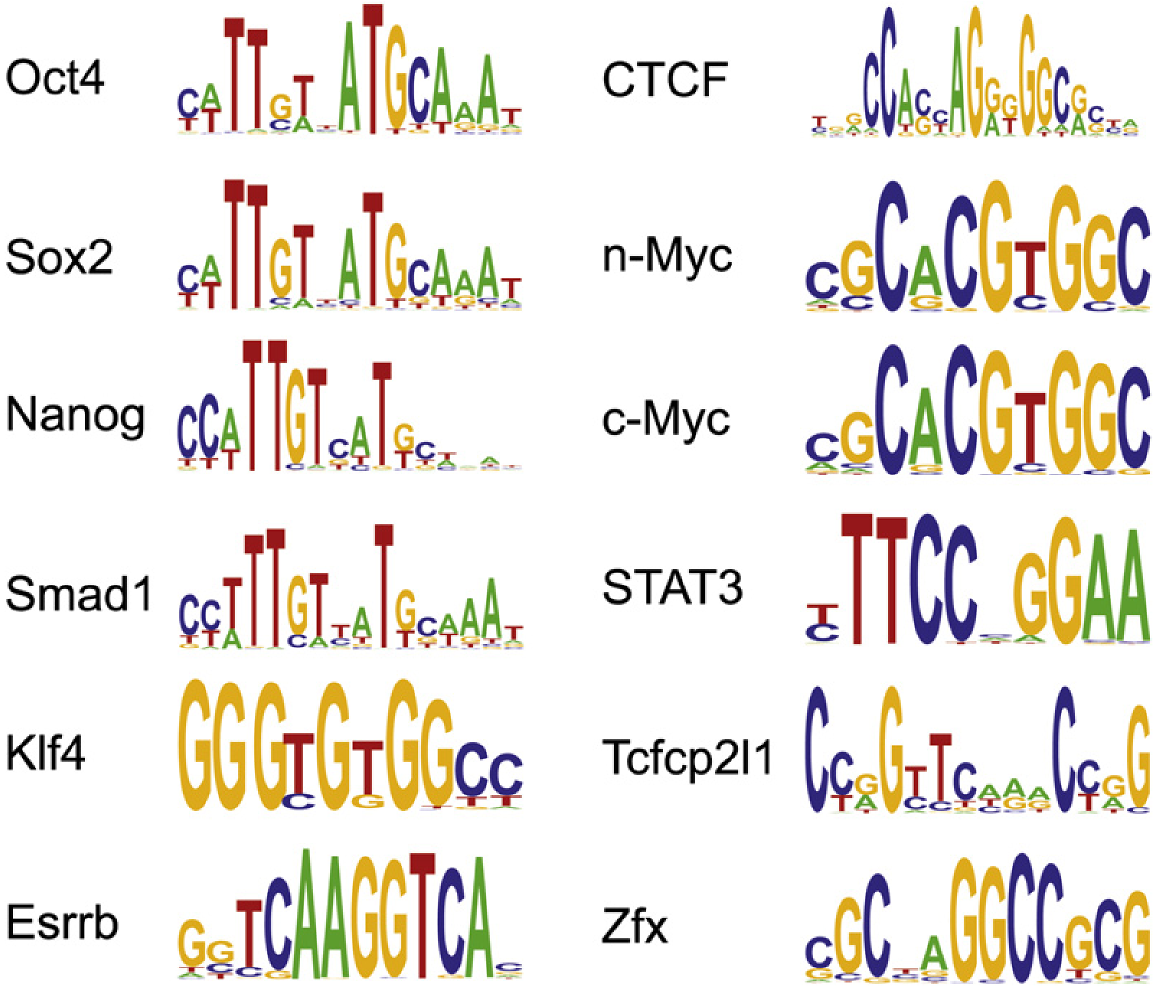
What kind of information can we obtain from the ChIP-seq experiments ?



Kagey 2010

What kind of information can we obtain from the ChIP-seq experiments ?





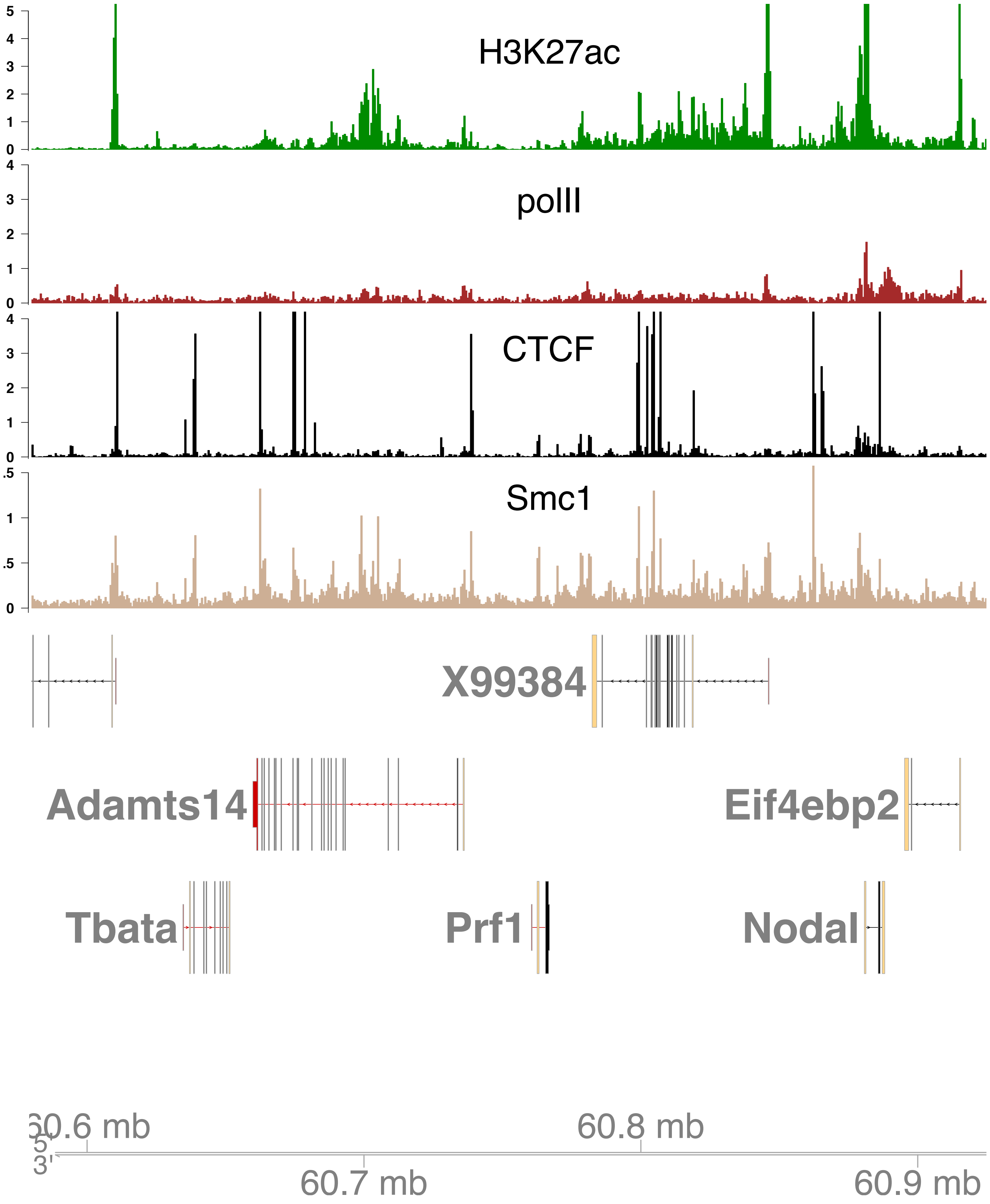
Chen 2008

To summarize - the most frequent tasks are:

1. Visualization along the genome
2. Peak finding and analysis (localization, co-occurrences, motifs)
3. Heatmaps of signal and average profiles at various genomic *loci*

But before we start the analysis... ChIP-seq: considerations for study design

- Distribution of modification - number of sequenced reads
- Paired vs. single end sequencing - fragment length estimation
- IgG control (pros and cons)
- Input control
- Biological replication!

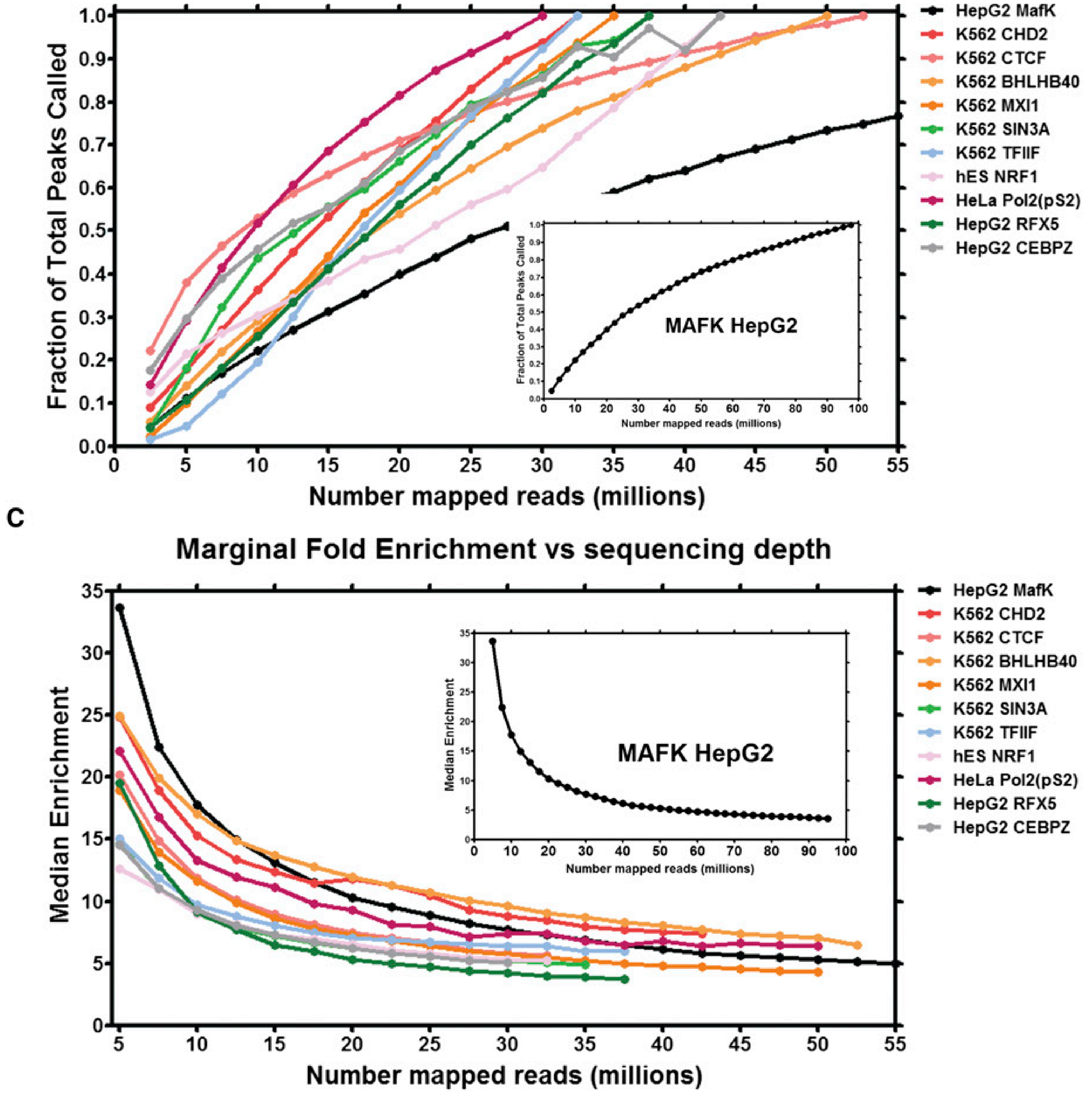


ChIP-seq profiles

- peaks vs. large domains
- signal to noise ratio

Data from:
Creyghton 2010
Kagey 2010

ChIP-seq: sequencing depth matters



Landt 2012

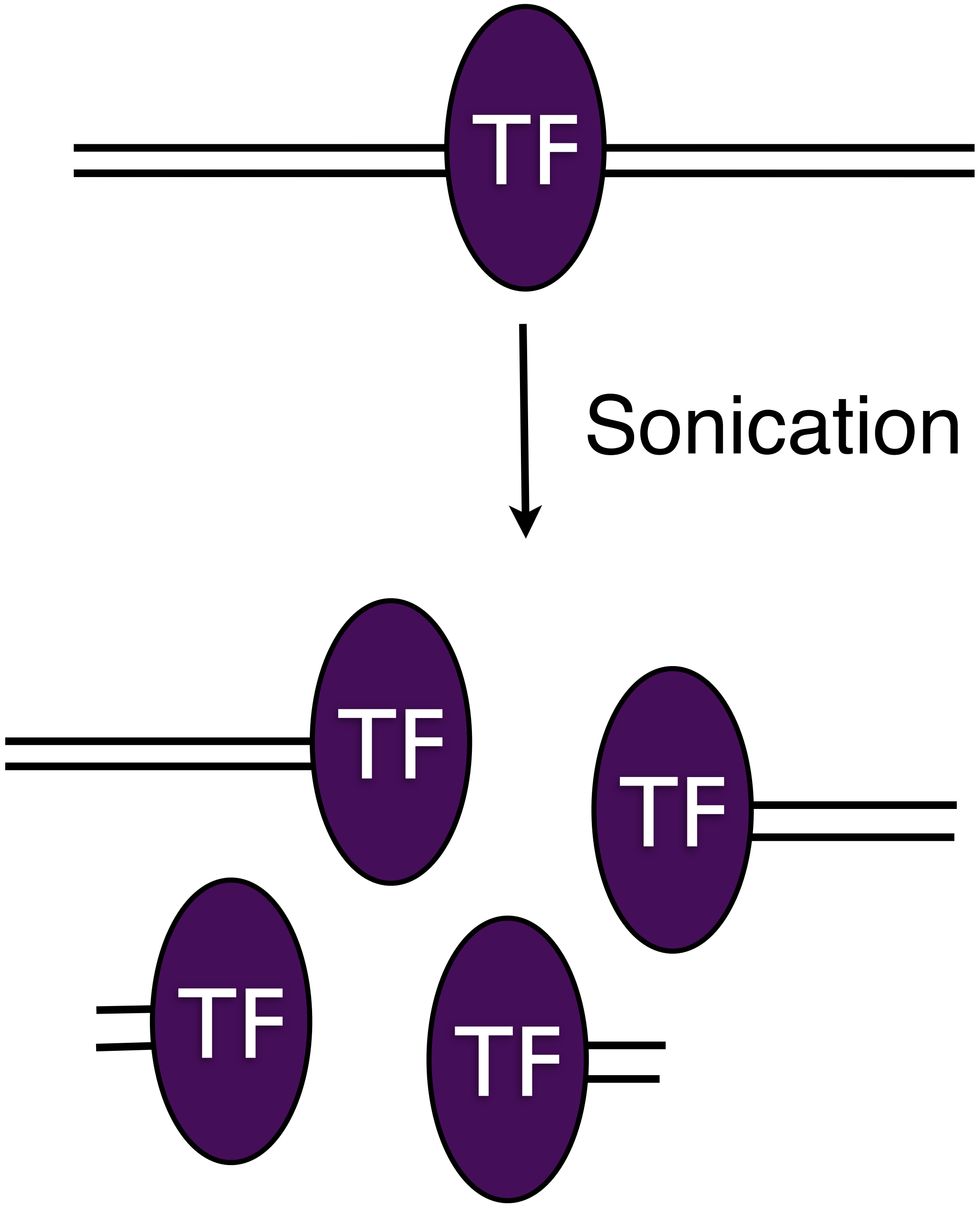
ENCODE consortium guidelines

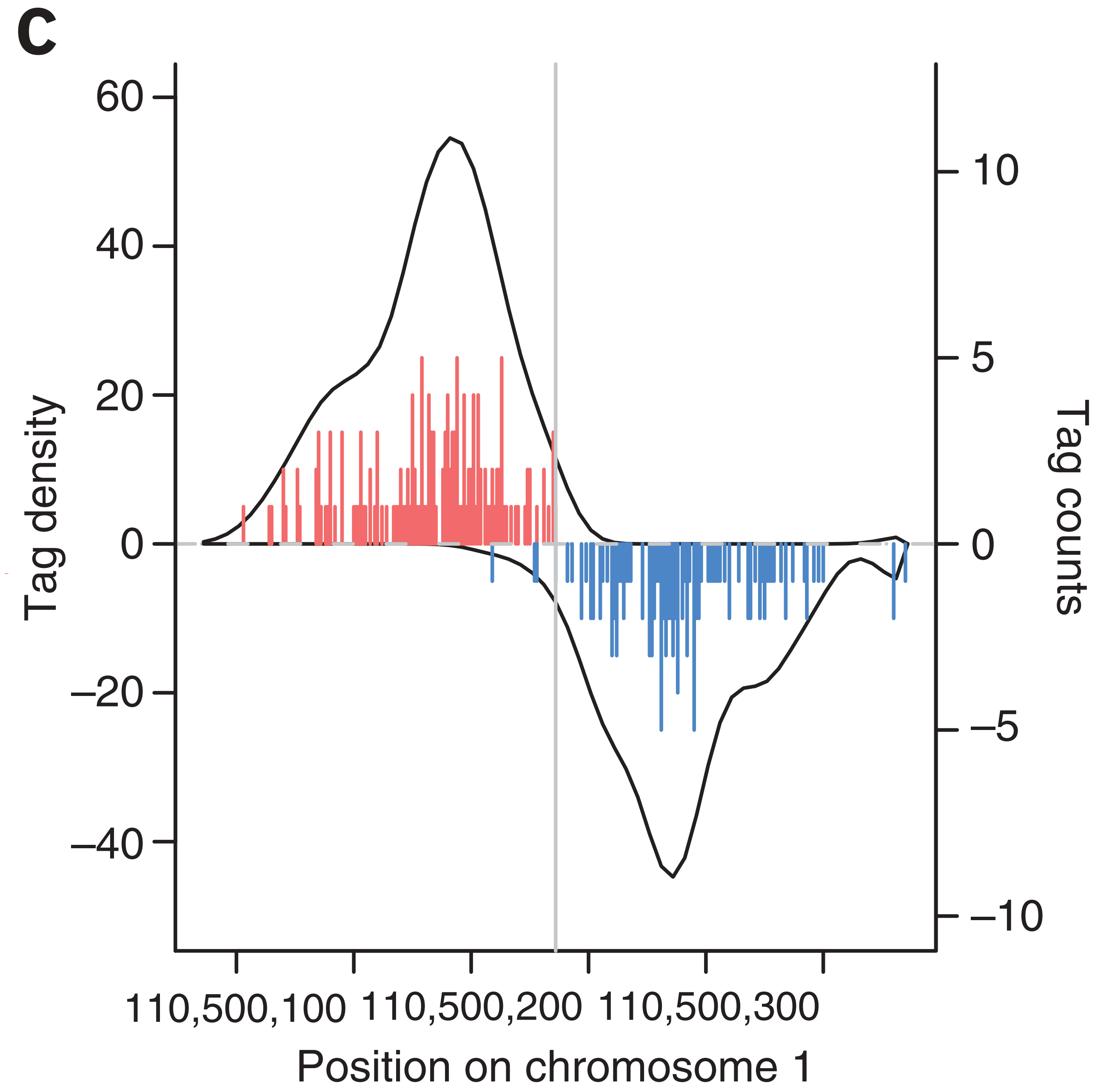
For mammalian genomes such as human and mouse:

1. > 20M aligned reads for broad marks
2. > 10M aligned reads for TFs

Paired vs. single end sequencing

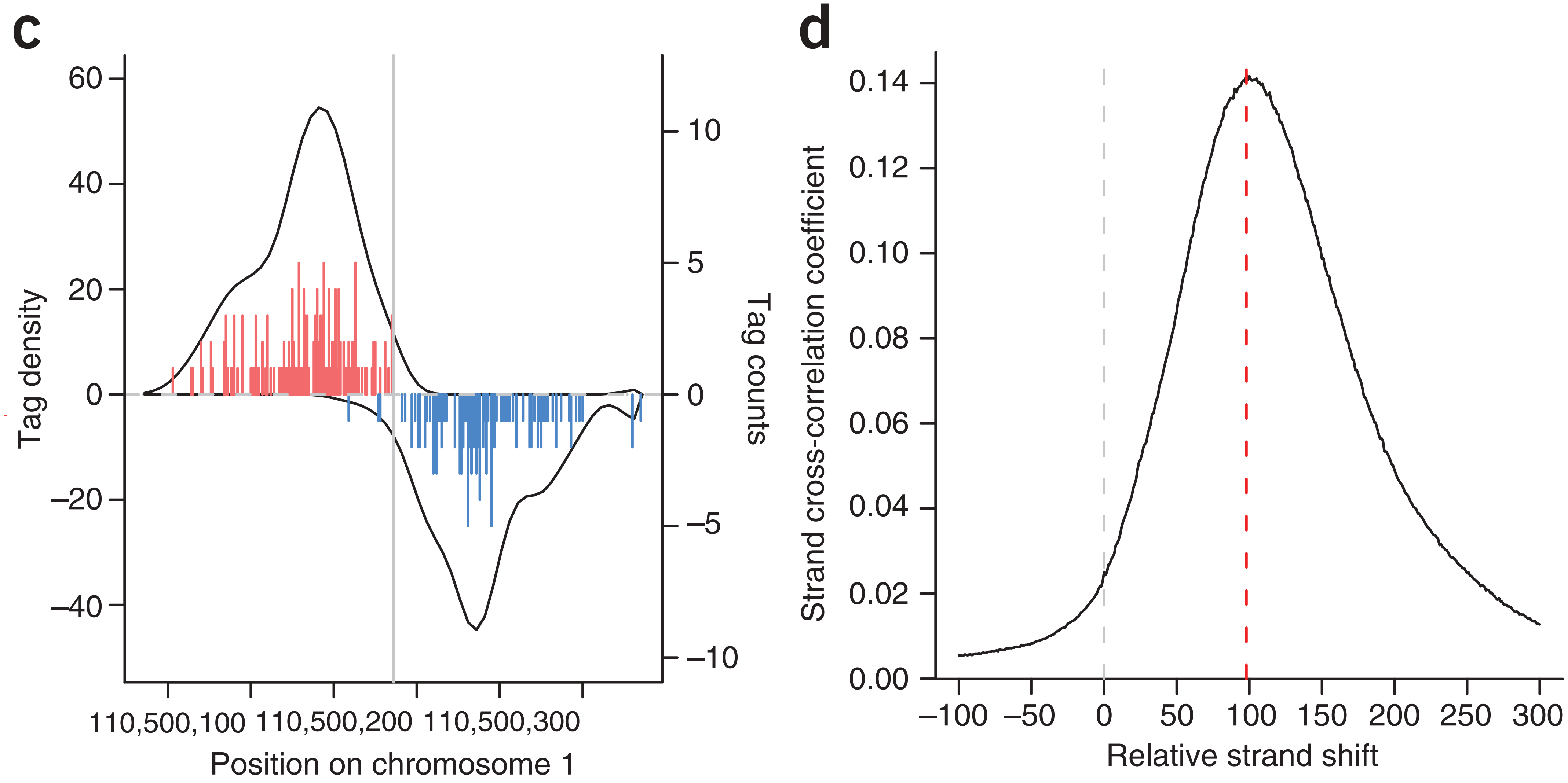
- paired end sequencing is always useful (nucleosome positioning) however not absolutely necessary





Kharchenko 2008

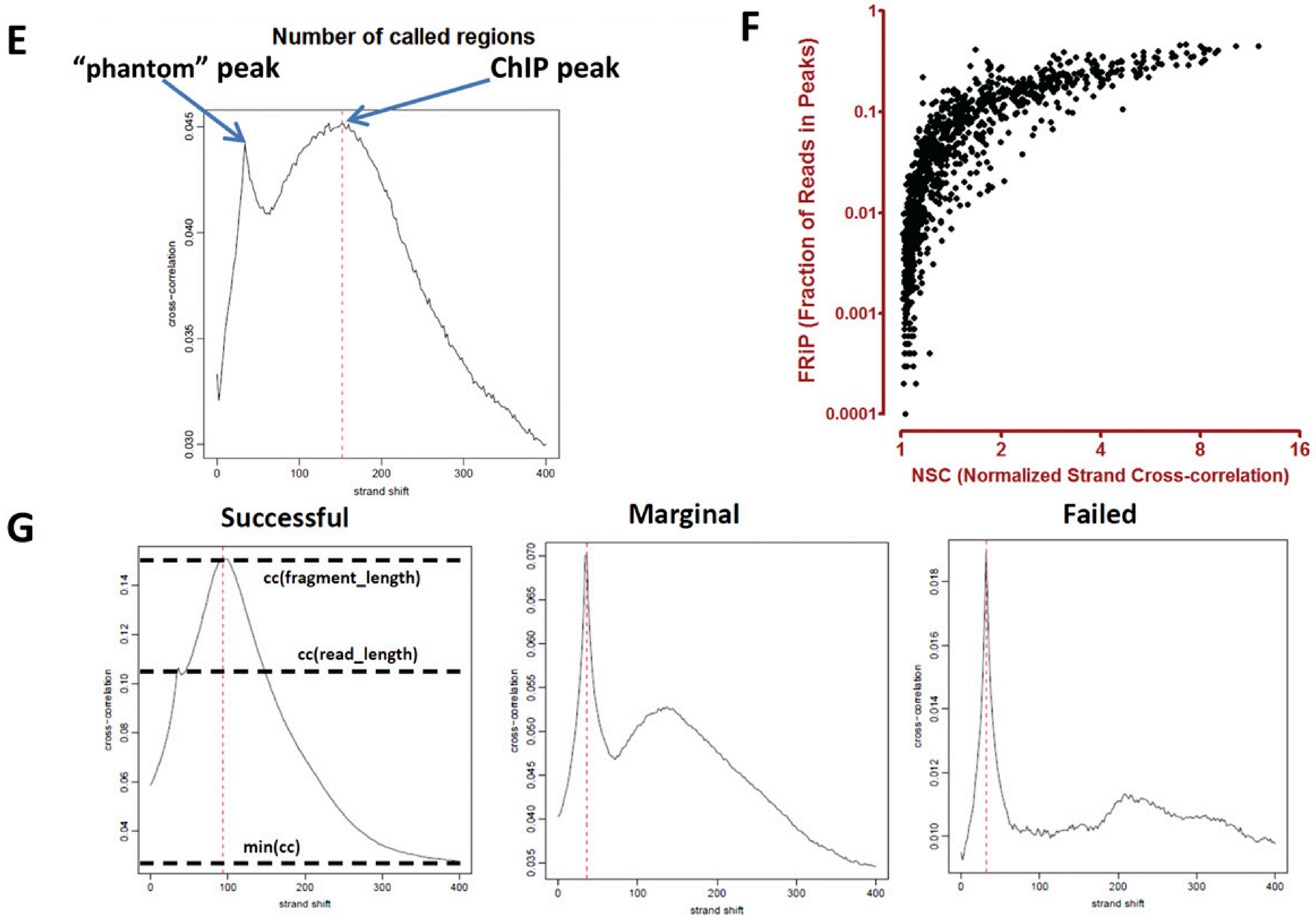
The estimation of the length of the ChIP fragments



Kharchenko 2008

- Binning - visualization and signal distribution analysis
- Quality control check
- **Peak finding**

Fragment length estimation - quality controls



$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length)\ - min(cc)}$$

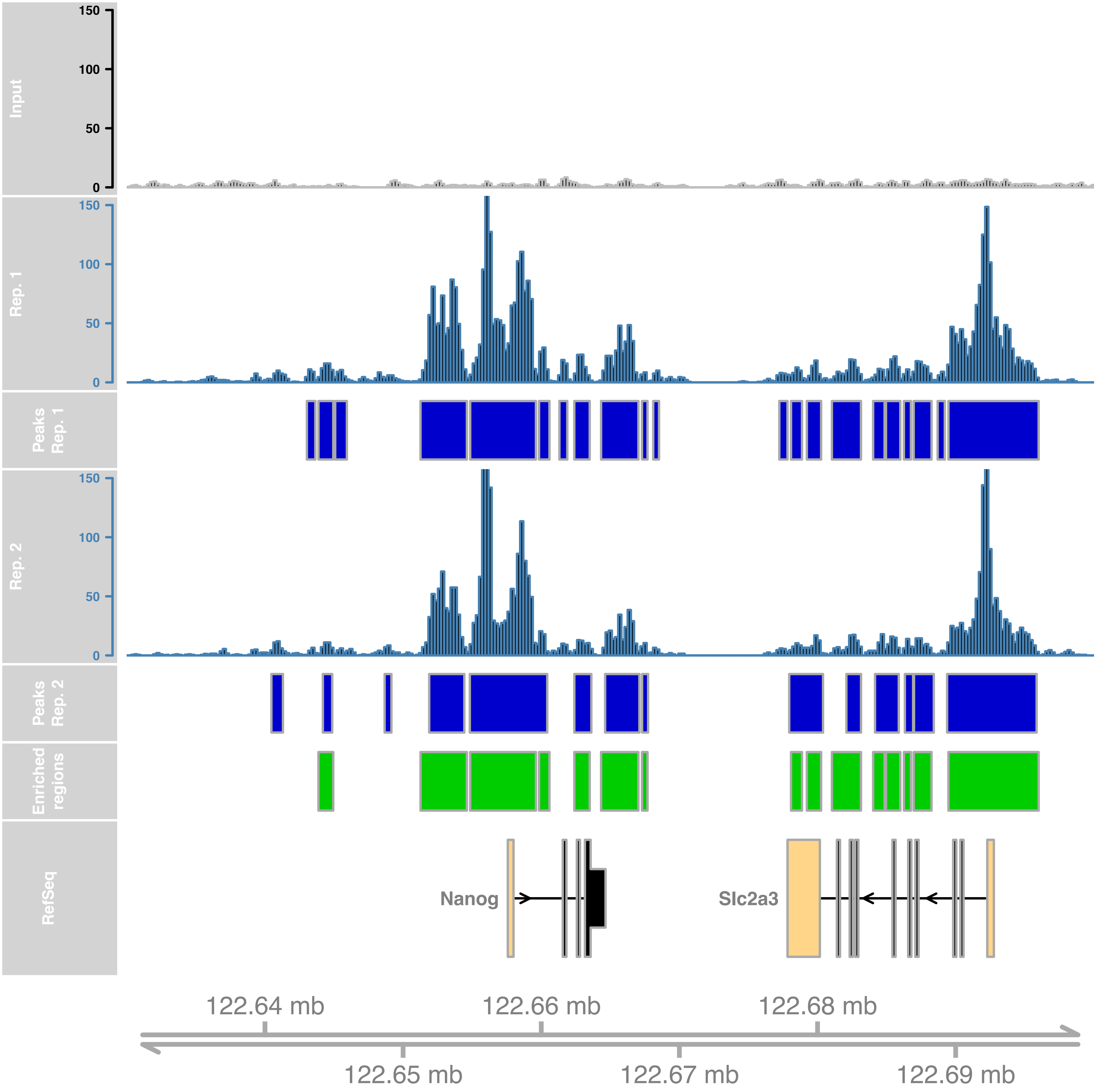
Landt 2012

Figure 4. (Legend on next page)

ChIP-seq: considerations for study design

- IgG control (pros and cons)
- Input control
- Biological replication

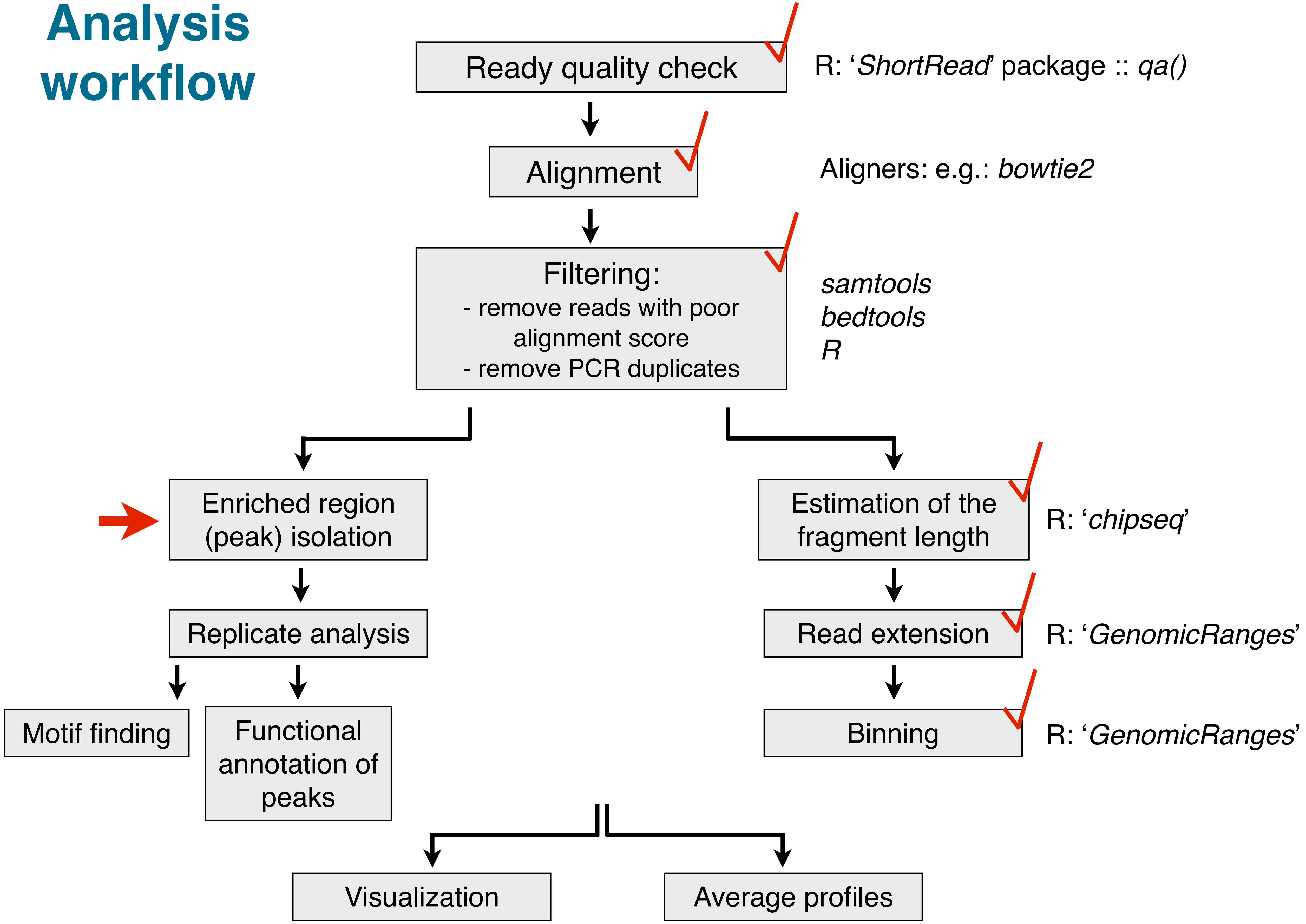
Finding enriched regions



<u>Enriched regions</u> (‘<u>peaks</u>’) - regions with signal which is significantly higher than the background - input or IgG

<u>Input reads</u> - background reads’ distribution exhibits a degree of clustering that is significantly greater than expected from a homogenous Poisson process (*P*-value< 10^{-6}, Kharchenko et al., 2008)

Analysis workflow



Ready quality check

R: ‘*ShortRead*’ package :: *qa()*

Alignment

Aligners: e.g.: *bowtie2*

Filtering:
- remove reads with poor alignment score
- remove PCR duplicates

samtools
bedtools
R

Enriched region (peak) isolation

Replicate analysis

Motif finding

Functional annotation of peaks

Estimation of the fragment length

R: ‘*chipseq*’

Read extension

R: ‘*GenomicRanges*’

Binning

R: ‘*GenomicRanges*’

Visualization

Average profiles

Model-based analysis of ChIP-seq (MACS)

Method

Model-based Analysis of ChIP-Seq (MACS)

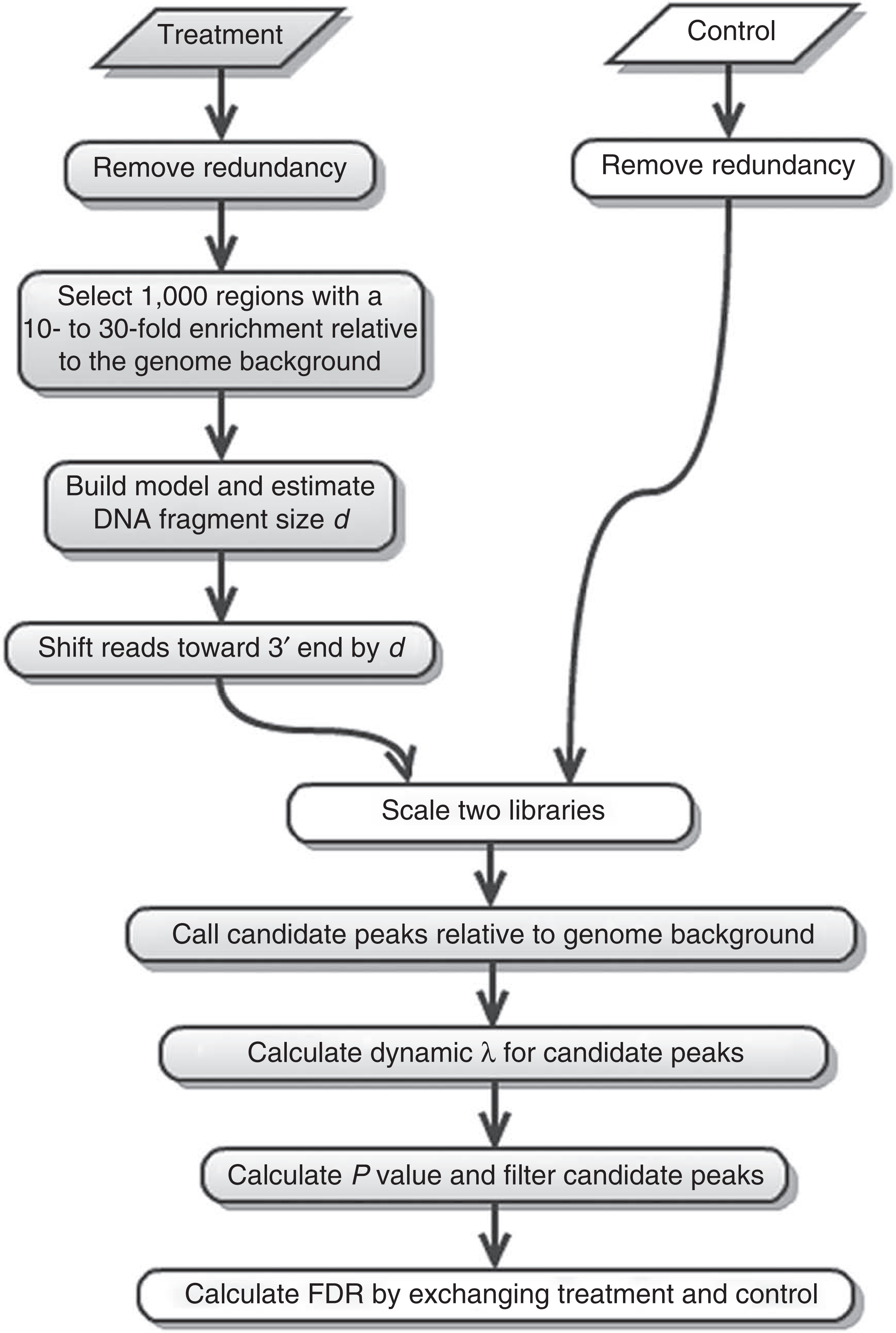
Yong Zhang[¤*], Tao Liu[¤*], Clifford A Meyer[*], Jérôme Eeckhoute[†], David S Johnson[‡], Bradley E Bernstein[§¶], Chad Nusbaum[¶], Richard M Myers[¥], Myles Brown[†], Wei Li[#] and X Shirley Liu[*]

- removes PCR duplicates

- *d* is estimated by picking highly enriched regions and looking at the distance between modes of positive and negative strand read pileups. Reads are extended towards this midpoint (building peak model)

- Sliding window of 2*d* to find significantly enriched bins using λ_{local}. We obtain enrichment P-value

- eFDR by swapping control and treatment



Treatment
Control
Remove redundancy
Remove redundancy
Select 1,000 regions with a 10- to 30-fold enrichment relative to the genome background
Build model and estimate DNA fragment size *d*
Shift reads toward 3′ end by *d*
Scale two libraries
Call candidate peaks relative to genome background
Calculate dynamic λ for candidate peaks
Calculate *P* value and filter candidate peaks
Calculate FDR by exchanging treatment and control

Several examples of peak callers

SICER - designed to deal with histone type data

PeakSeq, chromHMM ...

Peak callers in Bioconductor

OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

MOSAiCS - suitable for TF and histone modification data

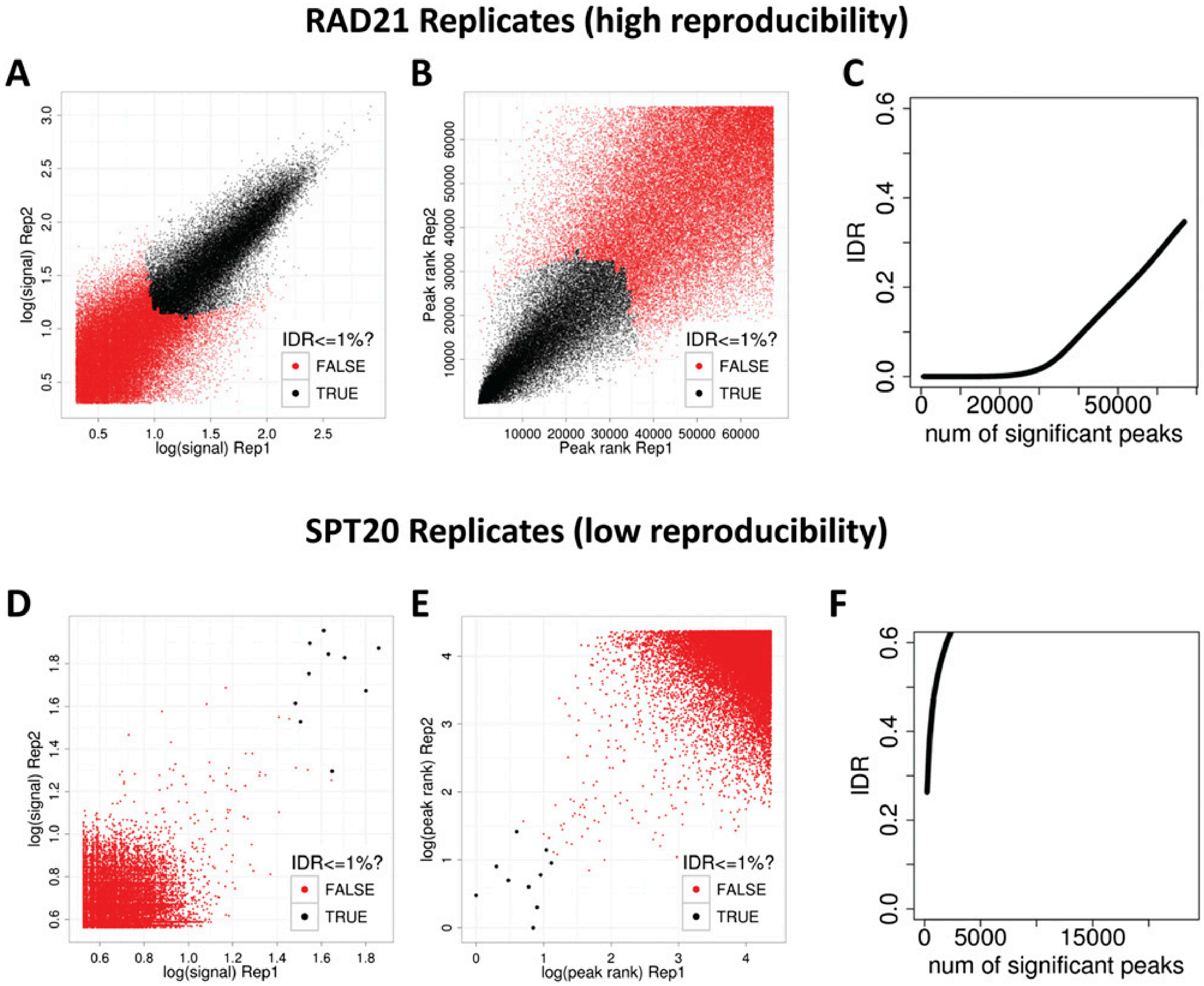
BayesPeak - suitable for TFs and histone modifications displaying peak-like signal

ChIPseqR - suitable for nucleosome positioning analysis

PICS CSAR NarrowPeaks CSSP ....

Peak processing - quality controls

- how do we decide whether samples and peaks are OK?



Landt 2012

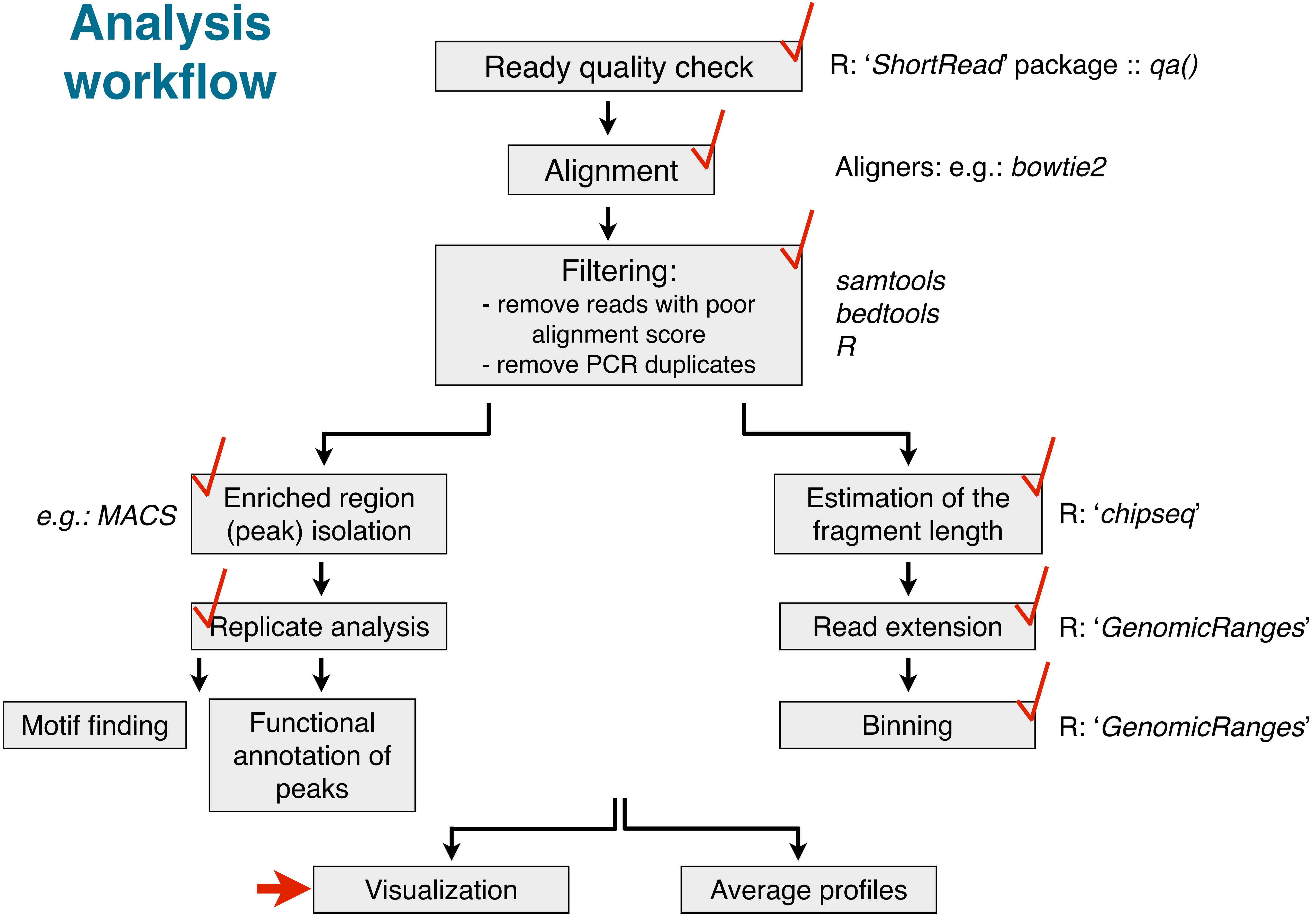
The irreproducible discovery rate (**IDR**, Li 2011) - rank peaks and assess for consistency

↓

Distinct and strong peaks are often called by most of peak finding software

Low strength peaks are often noisy

Analysis workflow



Ready quality check — R: '*ShortRead*' package :: *qa()*

Alignment — Aligners: e.g.: *bowtie2*

Filtering:
- remove reads with poor alignment score
- remove PCR duplicates

samtools
bedtools
R

Enriched region (peak) isolation — *e.g.: MACS*

Replicate analysis

Motif finding

Functional annotation of peaks

Estimation of the fragment length — R: '*chipseq*'

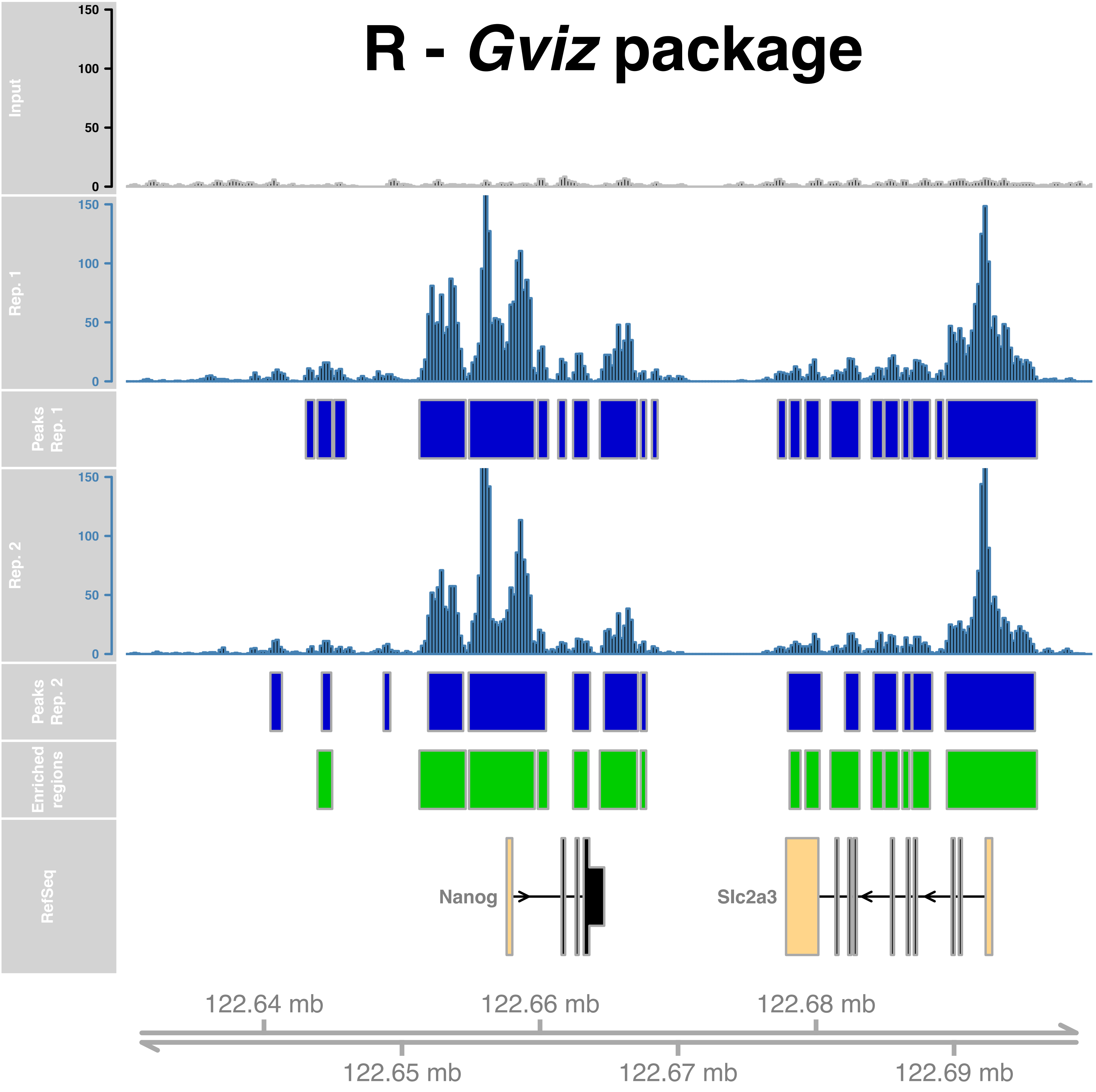
Read extension — R: '*GenomicRanges*'

Binning — R: '*GenomicRanges*'

Visualization

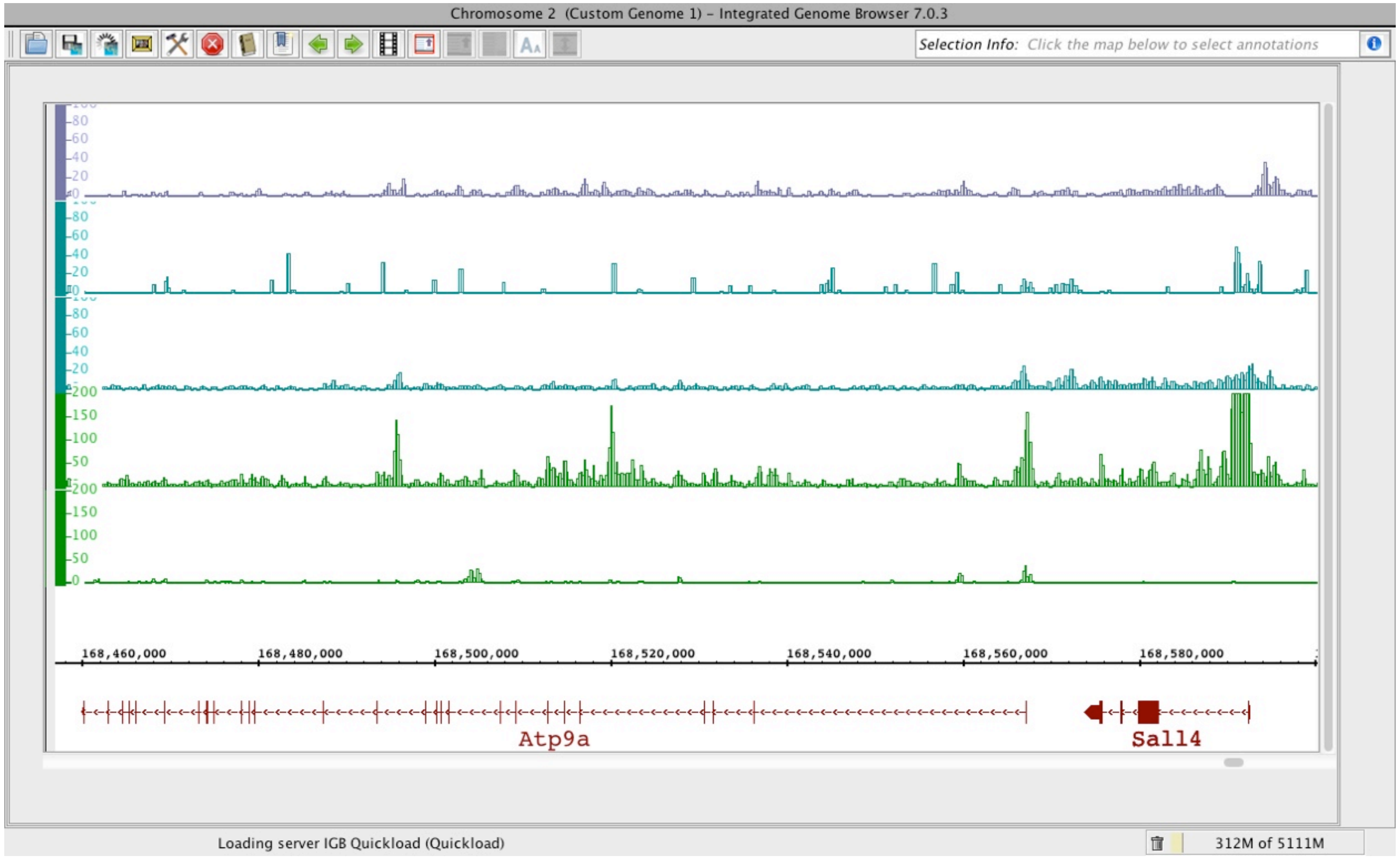
Average profiles

Visualization - seeing is believing



Visualization - other tools

IGB - Integrated Genome Browser - http://bioviz.org/igb/index.html


Chromosome 2 (Custom Genome 1) – Integrated Genome Browser 7.0.3
Selection Info: Click the map below to select annotations
80
60
40
20
0
200
150
100
50
168,460,000
168,480,000
168,500,000
168,520,000
168,540,000
168,560,000
168,580,000
Atp9a
Sall4
Loading server IGB Quickload (Quickload)
312M of 5111M

IGV - Integrative Genomics Viewer https://www.broadinstitute.org/igv/

Visualization - file formats

Binned or not data $\xrightarrow{R}$

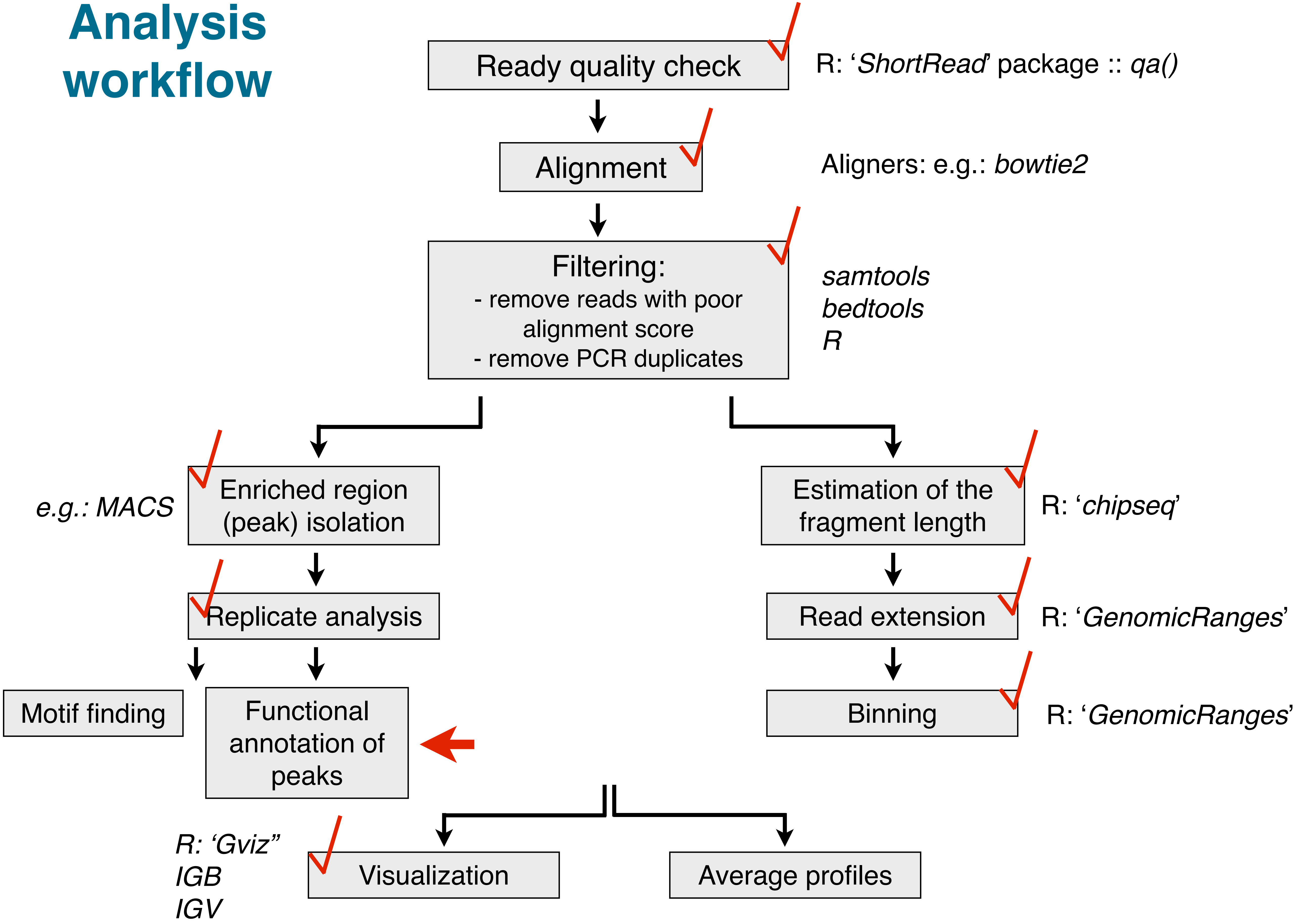
.bed

.bedGraph

.wig

.bigWig

Analysis workflow



Ready quality check — R: '*ShortRead*' package :: *qa()*

Alignment — Aligners: e.g.: *bowtie2*

Filtering:
- remove reads with poor alignment score
- remove PCR duplicates

samtools
bedtools
R

e.g.: MACS — Enriched region (peak) isolation

Replicate analysis

Motif finding

Functional annotation of peaks

Estimation of the fragment length — R: '*chipseq*'

Read extension — R: '*GenomicRanges*'

Binning — R: '*GenomicRanges*'

R: 'Gviz"
IGB
IGV — Visualization

Average profiles

Peak analysis

Frequently asked questions include:

- Localization of peaks with respect to functional elements in the genome (promoters, gene body, introns, transcription termination sites, intergenic regions etc.)
- Co-ocurrence between enriched regions
- The distribution of signal at the peaks

ChIPpeakAnno - provides functions performing peak annotation to promoters etc.

biomaRt - easy access to data bases including gene annotation, sequence conservation, sequence retrieval etc.

GenomicRanges - fast comparison between genomic intervals:

findOverlaps()

countOverlaps()

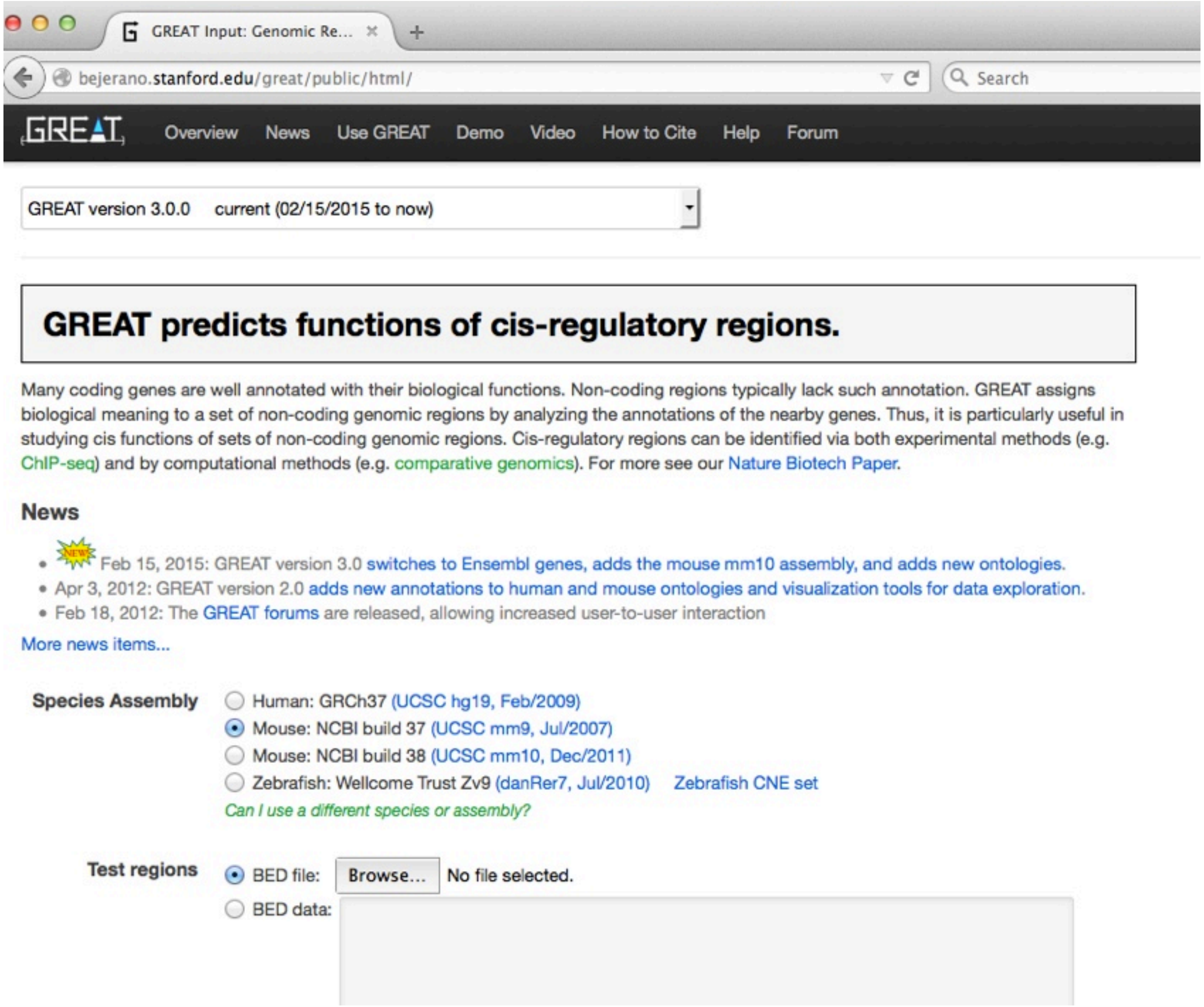
nearest()

Easy peak annotation to pre-established or new genomic features, cross-comparisons between peak locations and any kind of imaginable analysis

VennDiagram - visualization of two or multi-sample overlaps

Rcade - integrates ChIP-seq analysis with differential expression

Peak analysis - GREAT tool



GREAT Input: Genomic Re...
bejerano.stanford.edu/great/public/html/
Search
GREAT
Overview
News
Use GREAT
Demo
Video
How to Cite
Help
Forum

GREAT version 3.0.0 current (02/15/2015 to now)

GREAT predicts functions of cis-regulatory regions.

Many coding genes are well annotated with their biological functions. Non-coding regions typically lack such annotation. GREAT assigns biological meaning to a set of non-coding genomic regions by analyzing the annotations of the nearby genes. Thus, it is particularly useful in studying cis functions of sets of non-coding genomic regions. Cis-regulatory regions can be identified via both experimental methods (e.g. ChIP-seq) and by computational methods (e.g. comparative genomics). For more see our Nature Biotech Paper.

News

- NEW Feb 15, 2015: GREAT version 3.0 switches to Ensembl genes, adds the mouse mm10 assembly, and adds new ontologies.
- Apr 3, 2012: GREAT version 2.0 adds new annotations to human and mouse ontologies and visualization tools for data exploration.
- Feb 18, 2012: The GREAT forums are released, allowing increased user-to-user interaction

More news items...

Species Assembly

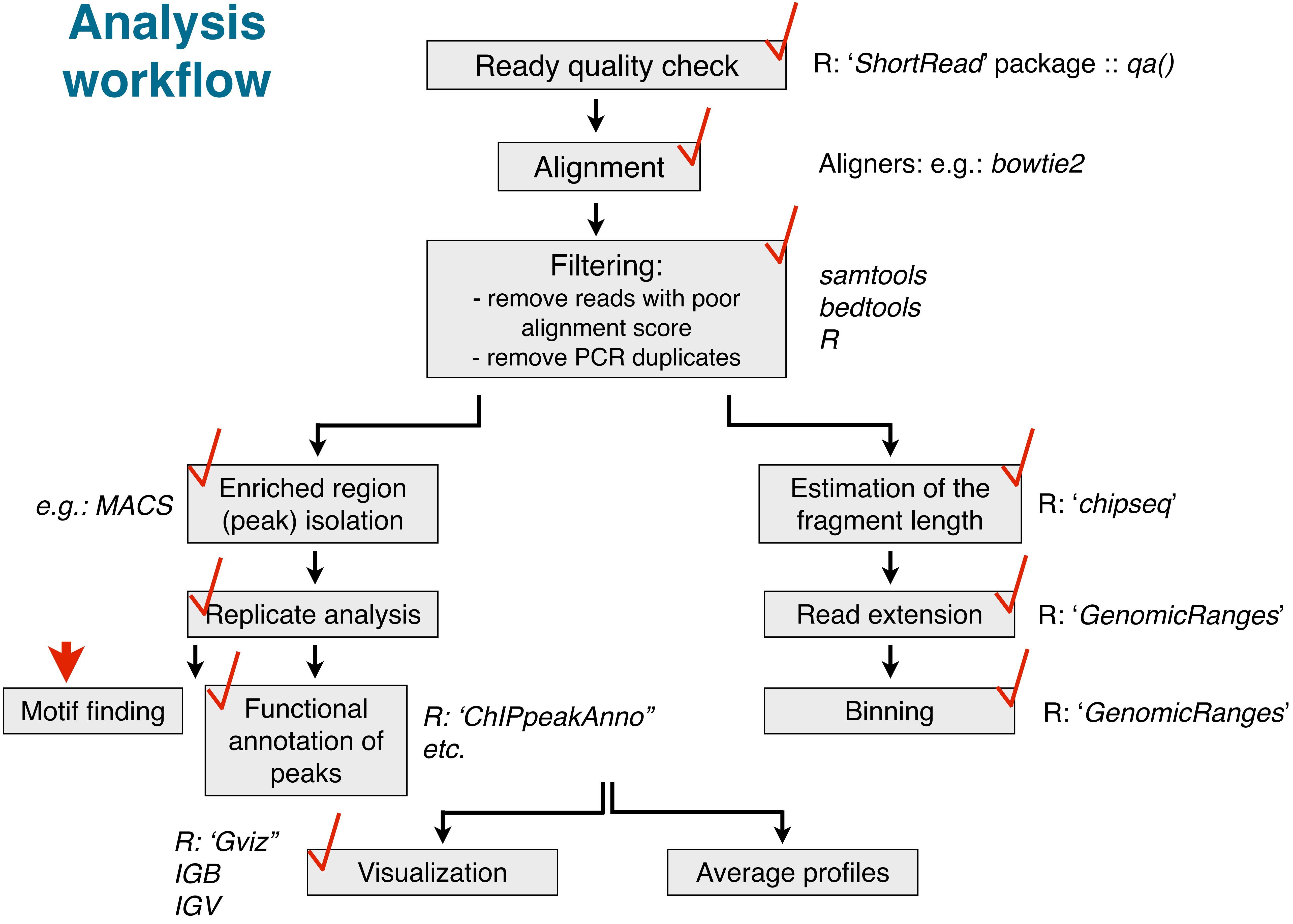
○ Human: GRCh37 (UCSC hg19, Feb/2009)
● Mouse: NCBI build 37 (UCSC mm9, Jul/2007)
○ Mouse: NCBI build 38 (UCSC mm10, Dec/2011)
○ Zebrafish: Wellcome Trust Zv9 (danRer7, Jul/2010) Zebrafish CNE set

Can I use a different species or assembly?

Test regions

● BED file: Browse... No file selected.
○ BED data:

Analysis workflow



Ready quality check ✓ — R: ‘*ShortRead*’ package :: *qa()*

Alignment ✓ — Aligners: e.g.: *bowtie2*

Filering: ✓
- remove reads with poor alignment score
- remove PCR duplicates

samtools
bedtools
R

e.g.: MACS — Enriched region (peak) isolation ✓

Replicate analysis ✓

Motif finding

Functional annotation of peaks ✓ — *R: ‘ChIPpeakAnno” etc.*

Estimation of the fragment length ✓ — R: ‘*chipseq*’

Read extension ✓ — R: ‘*GenomicRanges*’

Binning ✓ — R: ‘*GenomicRanges*’

R: ‘Gviz”
IGB
IGV

Visualization ✓

Average profiles

Peak analysis - motifs

MEME - provides functions performing motif discovery

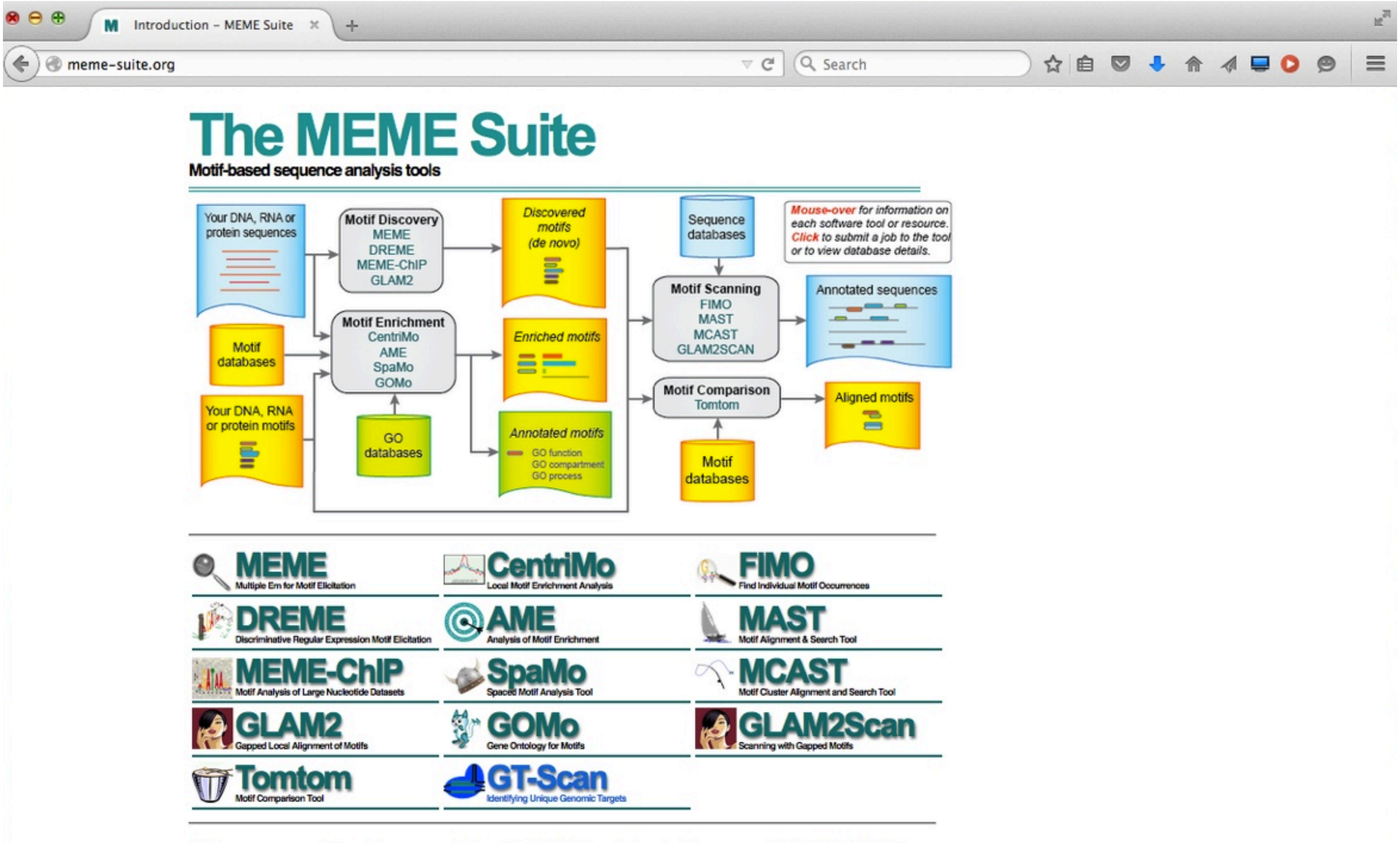
RSAT - complete suite for motif finding

↕

Position Weight Matrix (PWM) - describes the probability of each nucleotide at each position of a motif

JASPAR/TRANSFAC - data bases of PWM

R: MotifDb, FIMO and others



The MEME Suite

Motif-based sequence analysis tools

Mouse-over for information on each software tool or resource. Click to submit a job to the tool or to view database details.

MEME Multiple Em for Motif Elicitation

CentriMo Local Motif Enrichment Analysis

FIMO Find Individual Motif Occurrences

DREME Discriminative Regular Expression Motif Elicitation

AME Analysis of Motif Enrichment

MAST Motif Alignment & Search Tool

MEME-ChIP Motif Analysis of Large Nucleotide Datasets

SpaMo Spaced Motif Analysis Tool

MCAST Motif Cluster Alignment and Search Tool

GLAM2 Gapped Local Alignment of Motifs

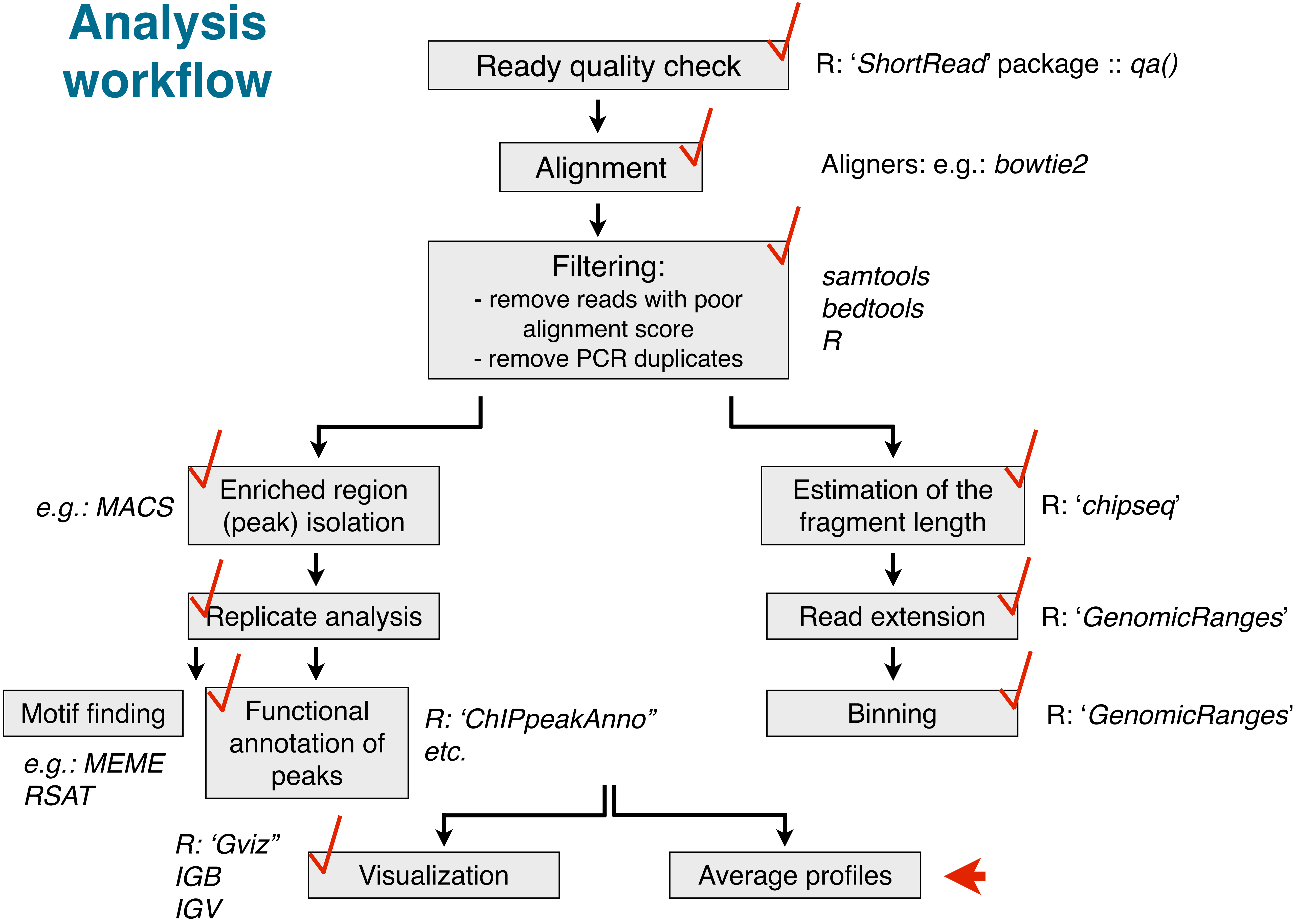
GOMo Gene Ontology for Motifs

GLAM2Scan Scanning with Gapped Motifs

Tomtom Motif Comparison Tool

GT-Scan Identifying Unique Genomic Targets

Analysis workflow



Ready quality check ✓ — R: '*ShortRead*' package :: *qa()*

Alignment ✓ — Aligners: e.g.: *bowtie2*

Filtering: ✓
- remove reads with poor alignment score
- remove PCR duplicates

samtools
bedtools
R

e.g.: MACS — Enriched region (peak) isolation ✓

Replicate analysis ✓

Motif finding

e.g.: MEME
RSAT

Functional annotation of peaks ✓ — *R: 'ChIPpeakAnno" etc.*

Estimation of the fragment length ✓ — R: '*chipseq*'

Read extension ✓ — R: '*GenomicRanges*'

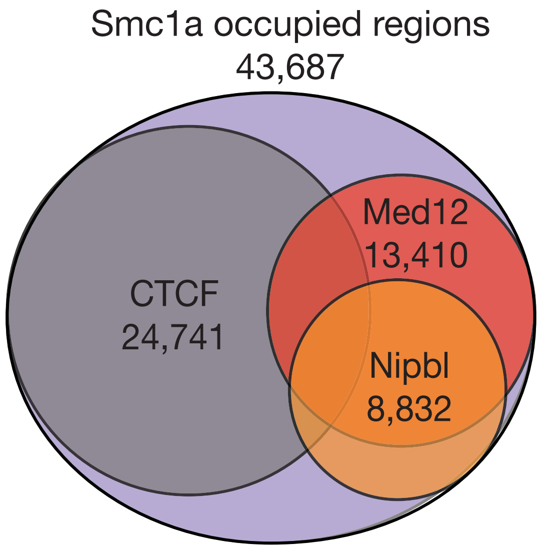
Binning ✓ — R: '*GenomicRanges*'

R: 'Gviz"
IGB
IGV

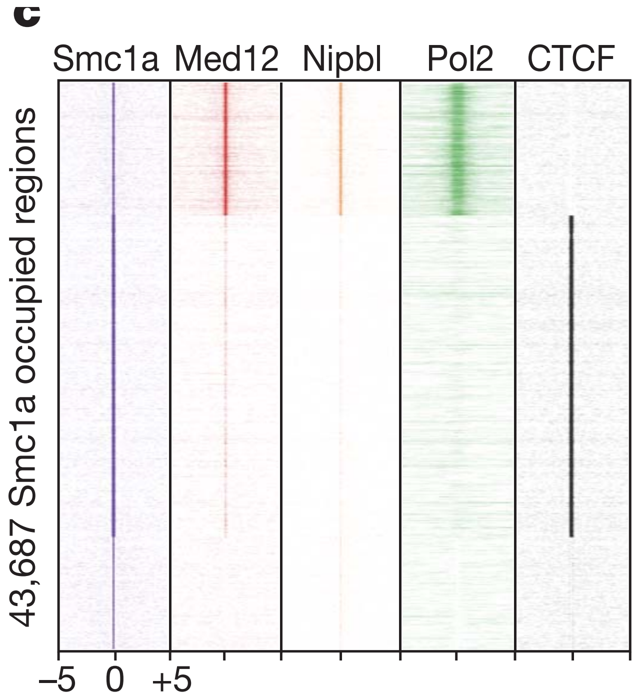
Visualization ✓

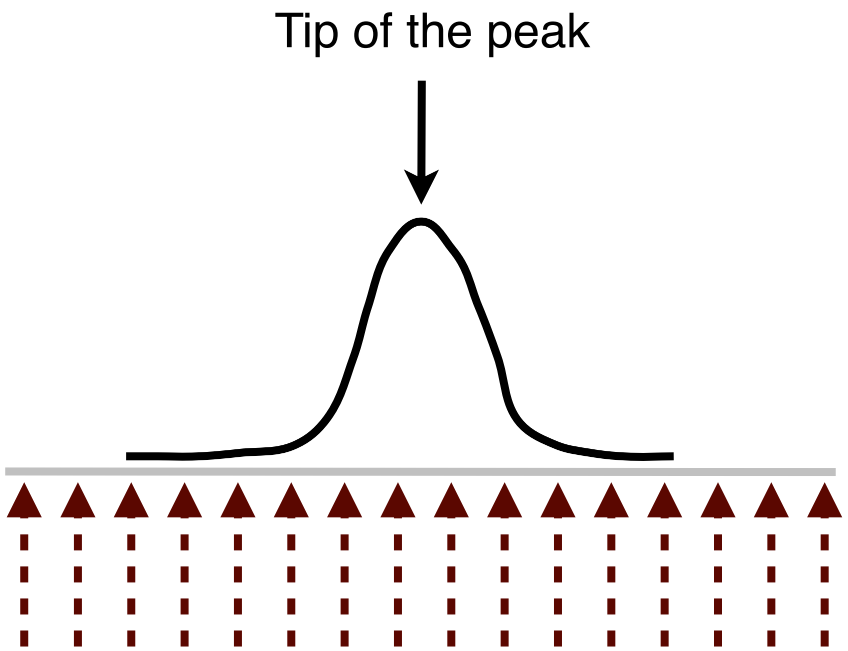
Average profiles

Co-enrichment and signal distribution analysis



Kagey 2010

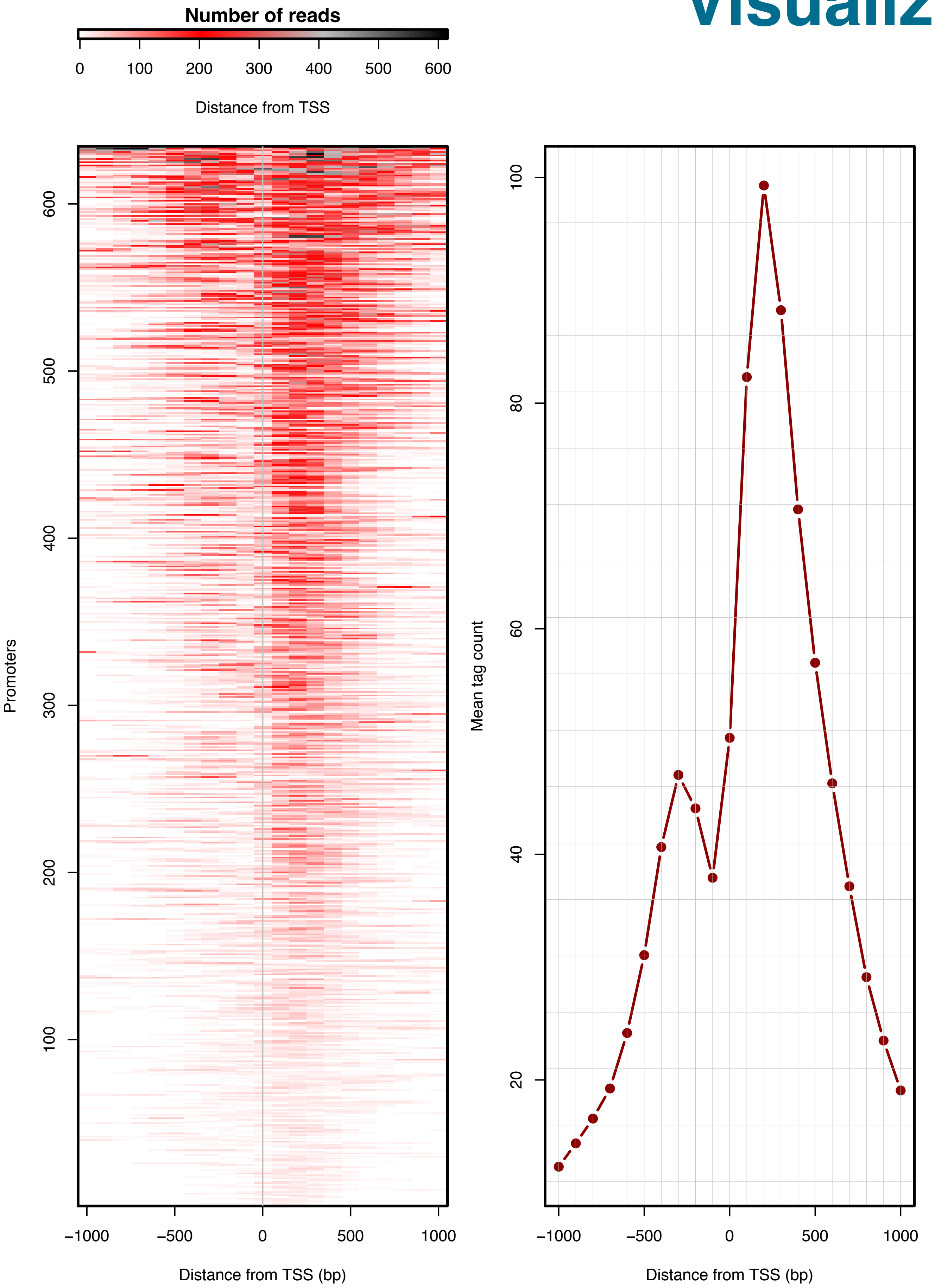




Region divided in to tiles

Count how many fragments fall into each tile

Visualization



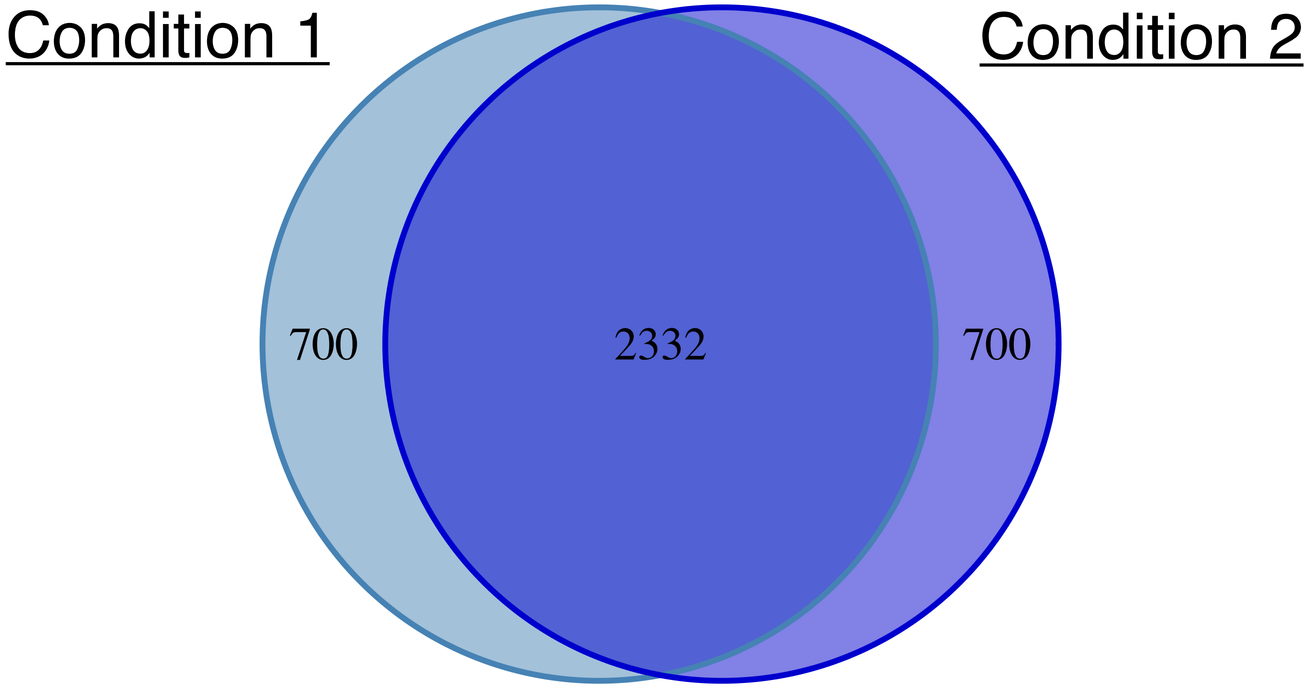
Heatmaps of signal enrichment at

- promoters
- loci enriched with factors of interest

We will see an example of such an analysis using R package ***GenomicRanges***

A nice alternative: ***HT-Seq*** (python)

Comparative peak analysis



Threshold issues affecting all qualitative analyses

Comparative peak analysis

DiffBind

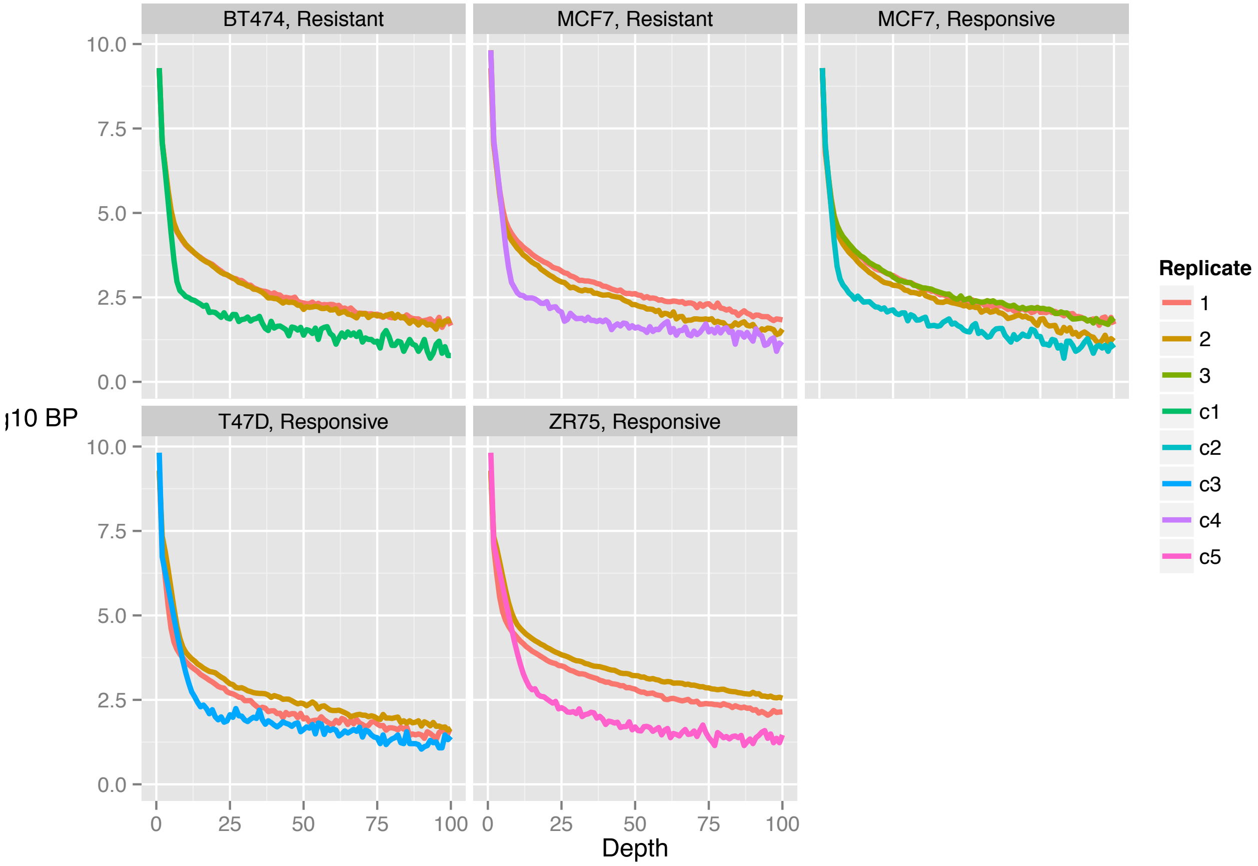
1. Count reads in peaks in all the replicates and conditions
2. Perform *edgeR* or *DESeq2* analysis - *dba.analyze()*
3. Provides various plotting functions

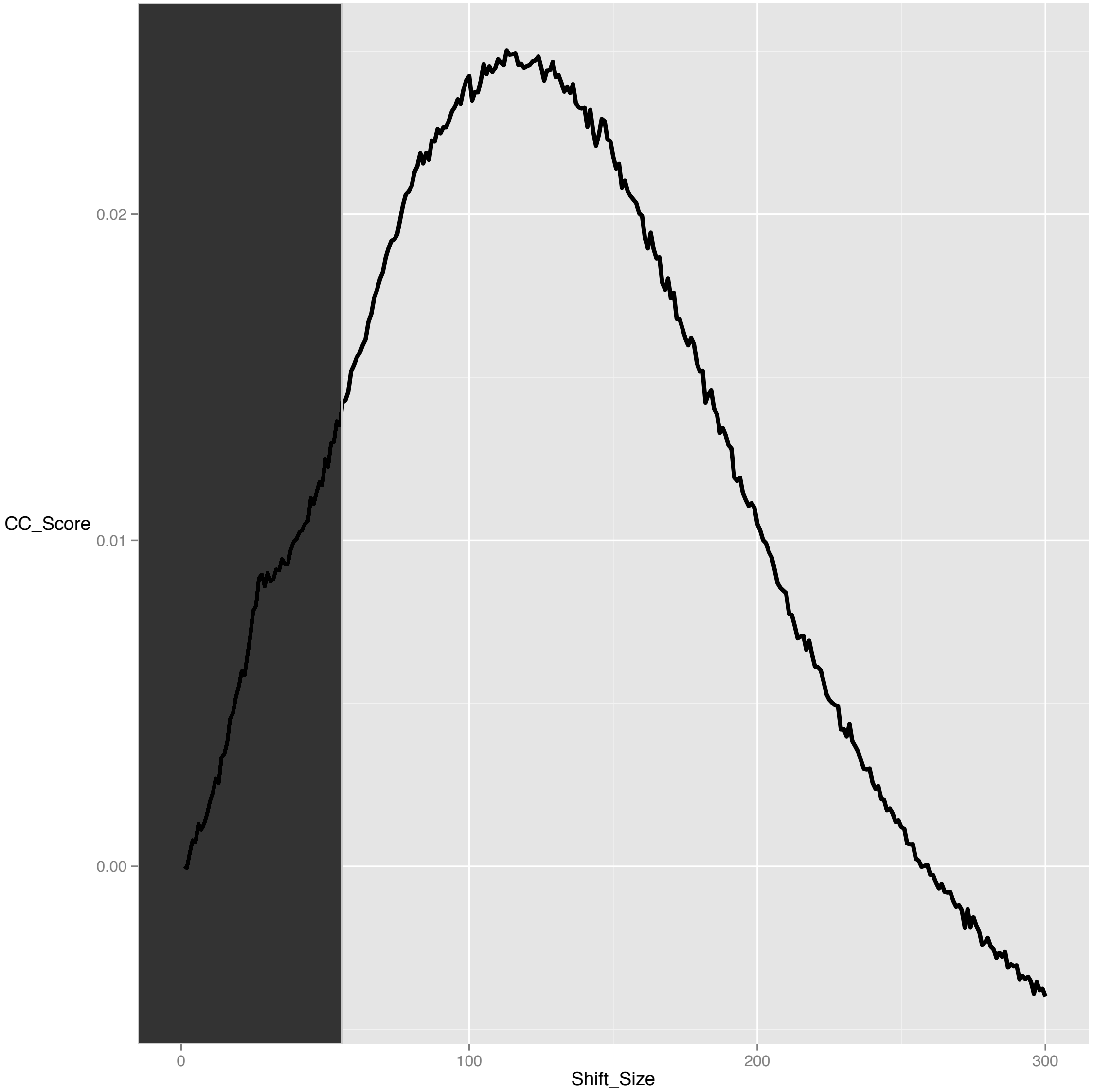
MMDiff

1. Count reads in peaks in all the replicates and conditions
2. Performs *DESeq* normalisation
3. Compares peak shapes using kernel based statistical tests

ChIPQC package for quality control checks and quantitative analysis of peak strengths

1. Plotting coverage histograms for peaks
2. Cross-coverage analysis in the function of shift sizes
3. Plotting peak profiles
4. Sample clustering





References (I)

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics 31 (2): 166-169

Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Madrigal P, Taslim C, Zhang J. 2013. Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol 9: 5–12.

Barski A, Cuddapah S, Cui K, Roh T, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell 129: 823–837.

Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, Wong E, Orlov YL, Zhang W, Jiang J, et al. 2008. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell 133: 1106–1117.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. PNAS 107.

Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. Nat Protoc 7: 1728–1740.

Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res 36: 5221–5231.

References (II)

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS, et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. Nature 467: 430–435.

Kharchenko P V, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26: 1351–1359.

Landt S, Marinov G. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 1813–1831.

Li Q, Brown B, Huang H, Bickel P. 2010. IDR analysis 101 Measuring consistency between replicates in high-throughput experiments. 1–7.

Lun ATL, Smyth GK. 2014. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: Controlling error rates correctly. Nucleic Acids Res 42: 1–11.

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, Van Helden J. 2012. RSAT peak-motifs: Motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 40.

Zhang Y, Liu T, Meyer C a, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.