ChIP-seq data analysis

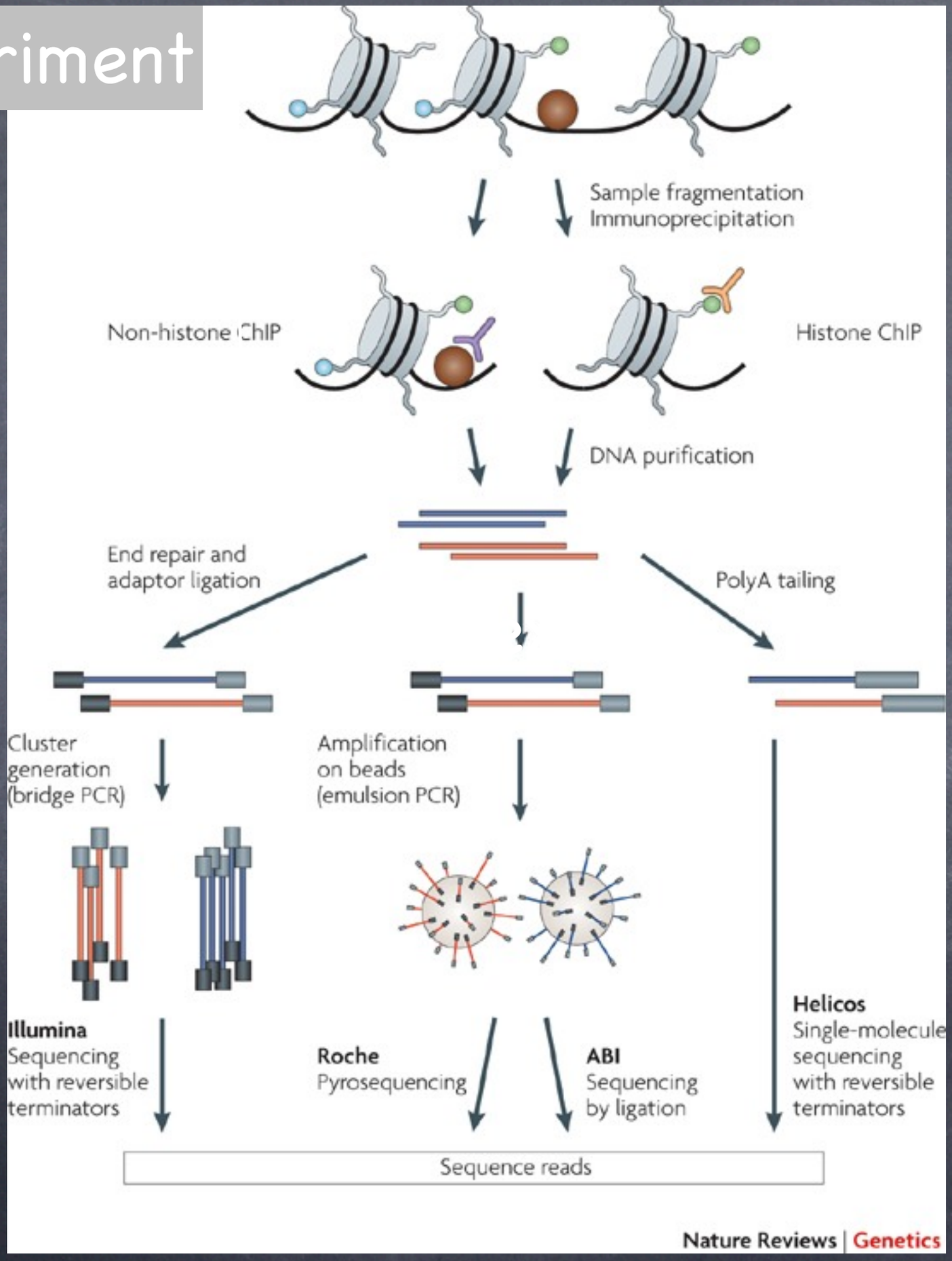
with SWEMBL

08. 06. 2010 --- Petra Schwalie

overview

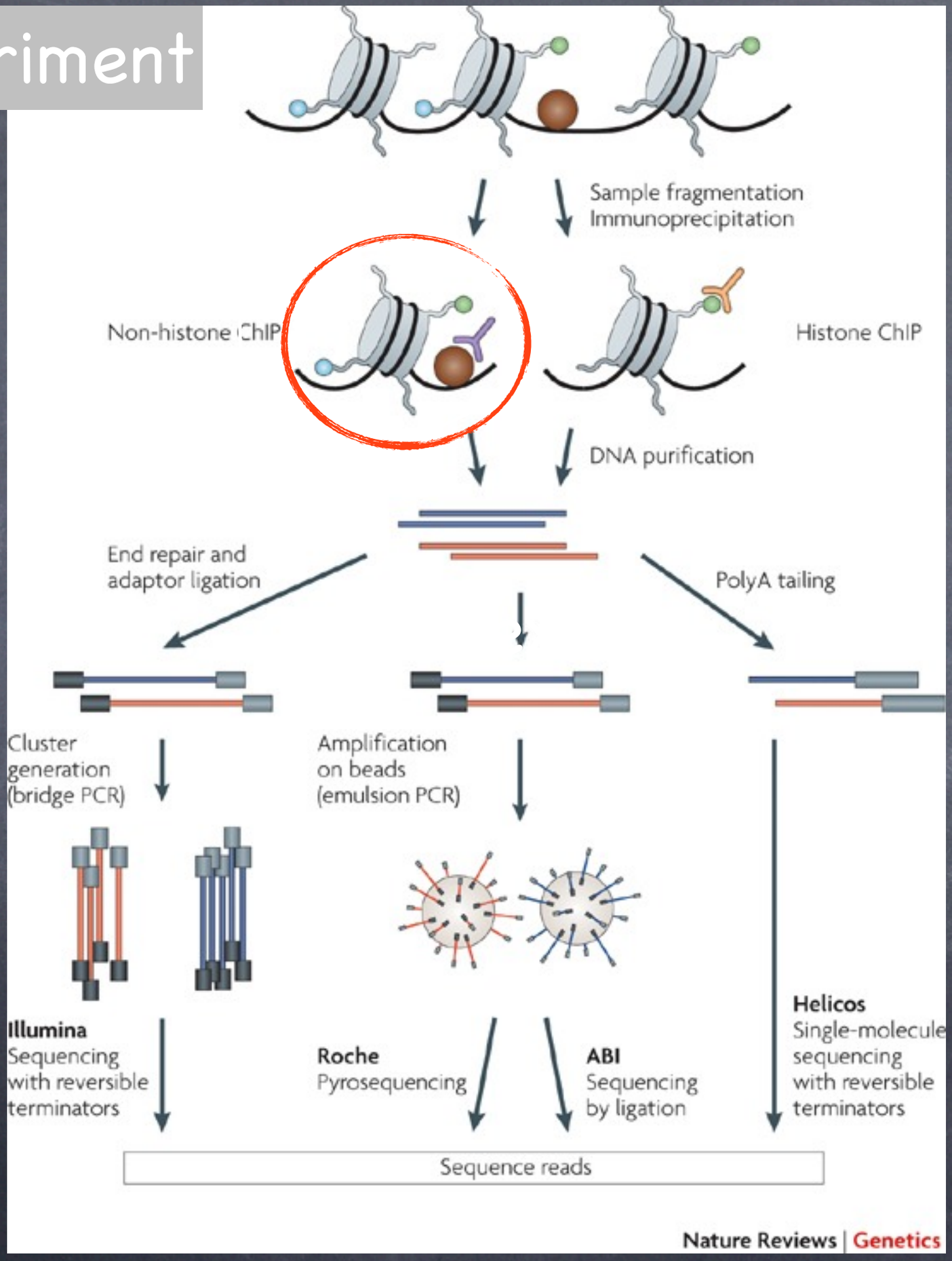
- the experiment
- the data (QC, filtering, aligning, storing)
- peak-calling with SWEMBL
- downstream analysis

The experiment



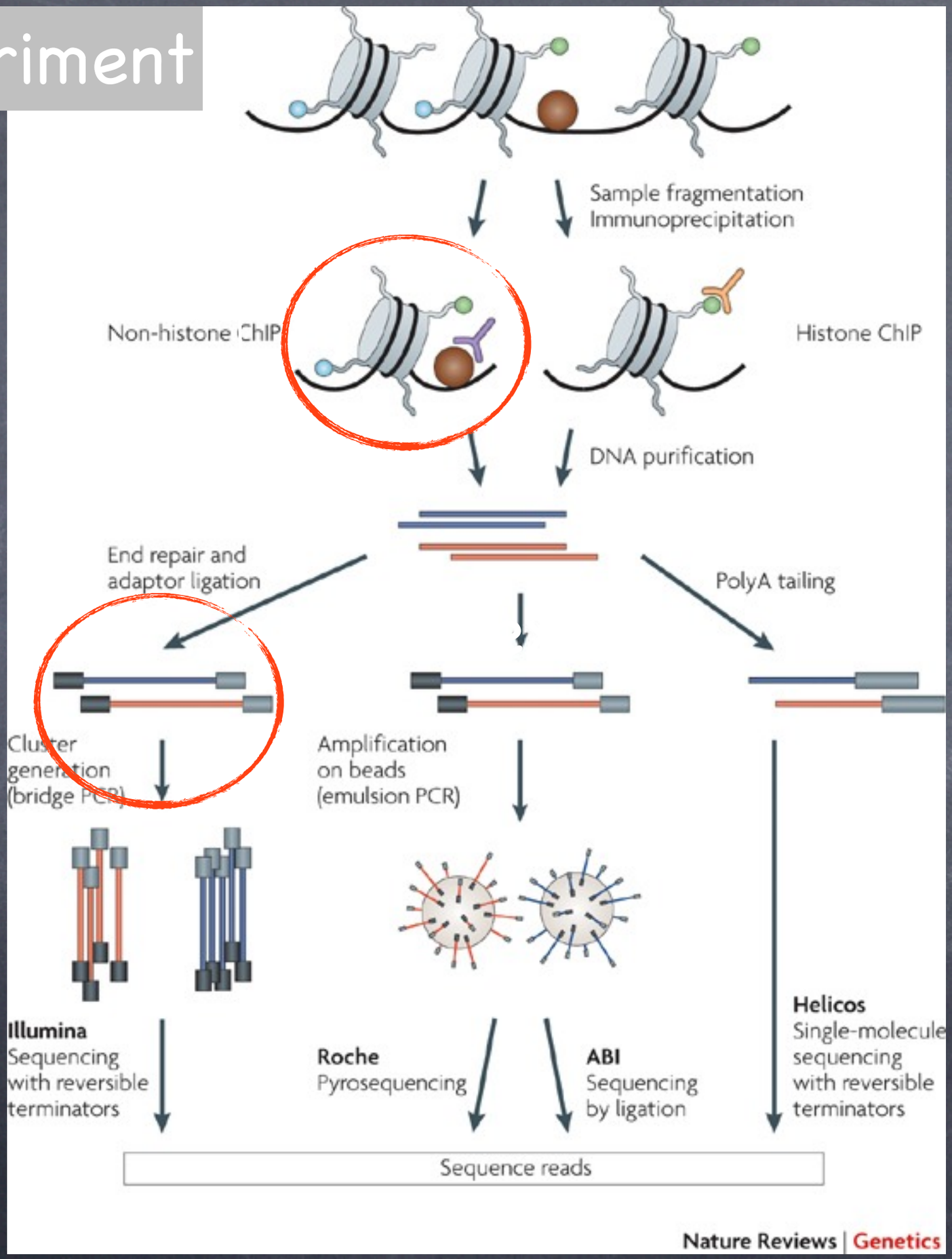
Park

The experiment



Park

The experiment



Park

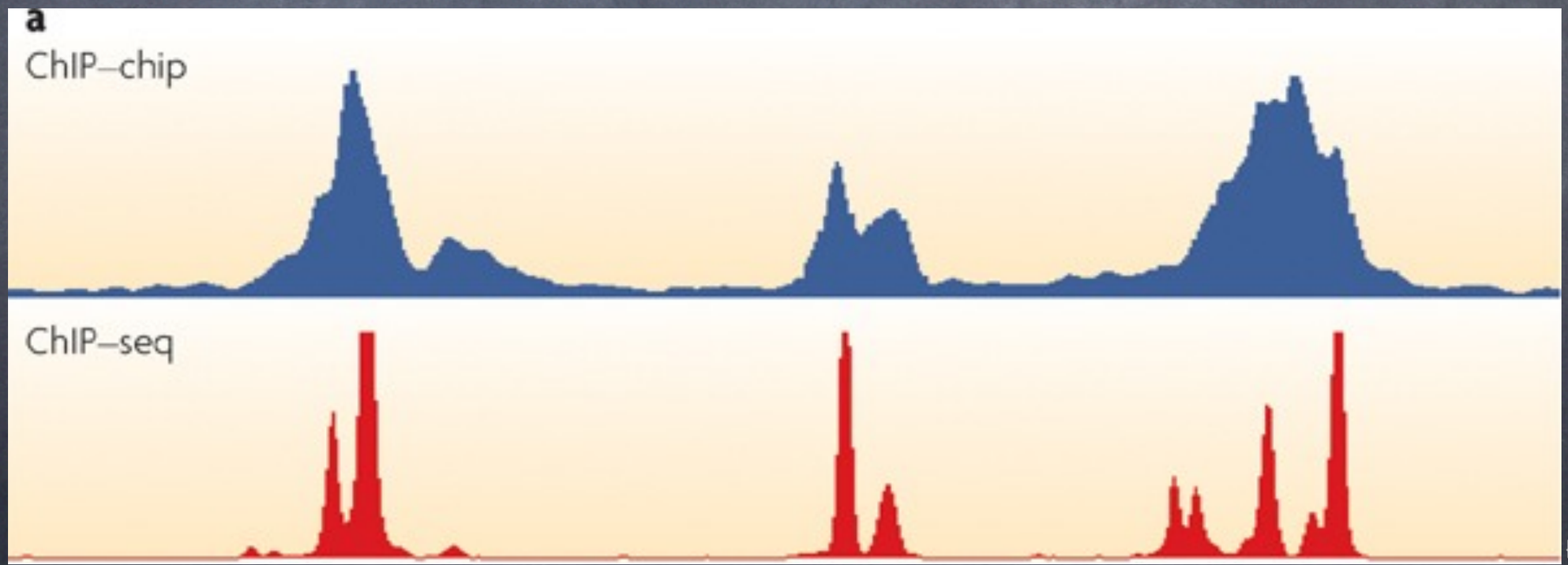
From chip to seq

- greater coverage (tiling resolution, repetitive regions)
- less noise (no probe-specific behavior, dye biases, less PCR)
- less input material
- lower cost

Park

From chip to seq

- greater coverage (tiling resolution, repetitive regions)



Park

Challenges and biases

Biases

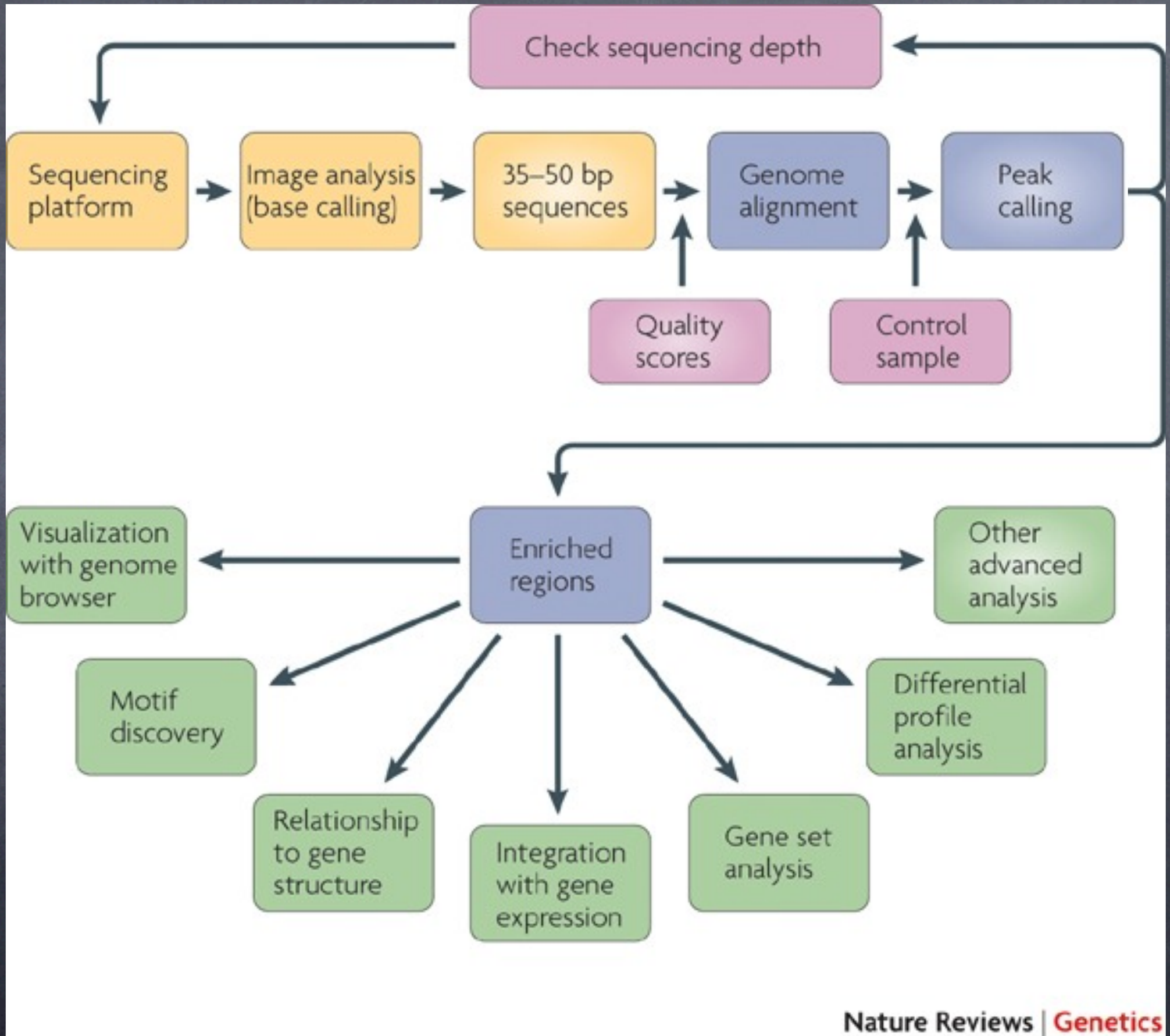
- Sequencing errors
- GC-rich content (library preparation and amplification)
- multiple hits/position

Challenges

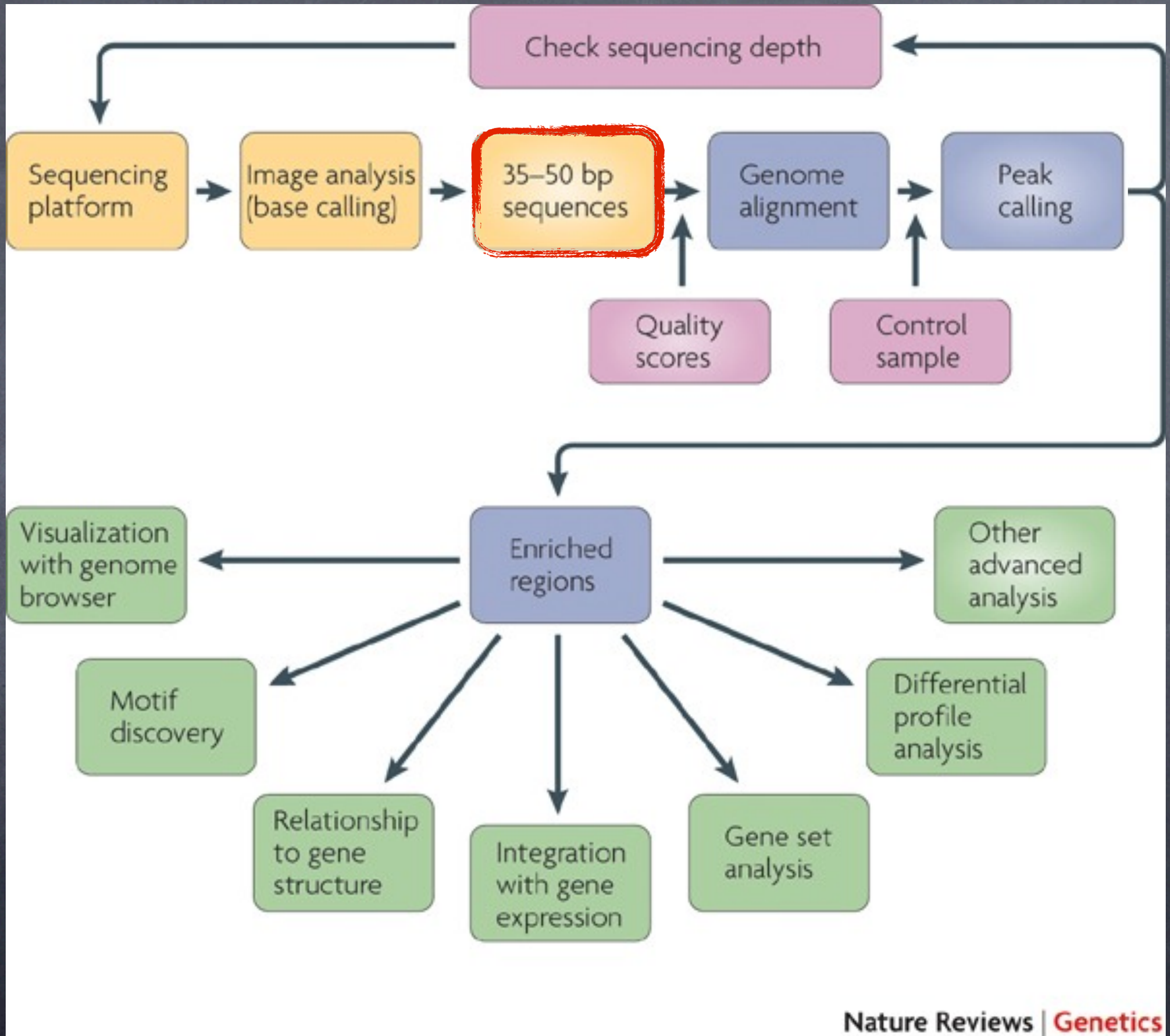
- Filtering & alignment
- Background tag distribution
- Required sequencing depth
- Protein binding positions

Antibody!

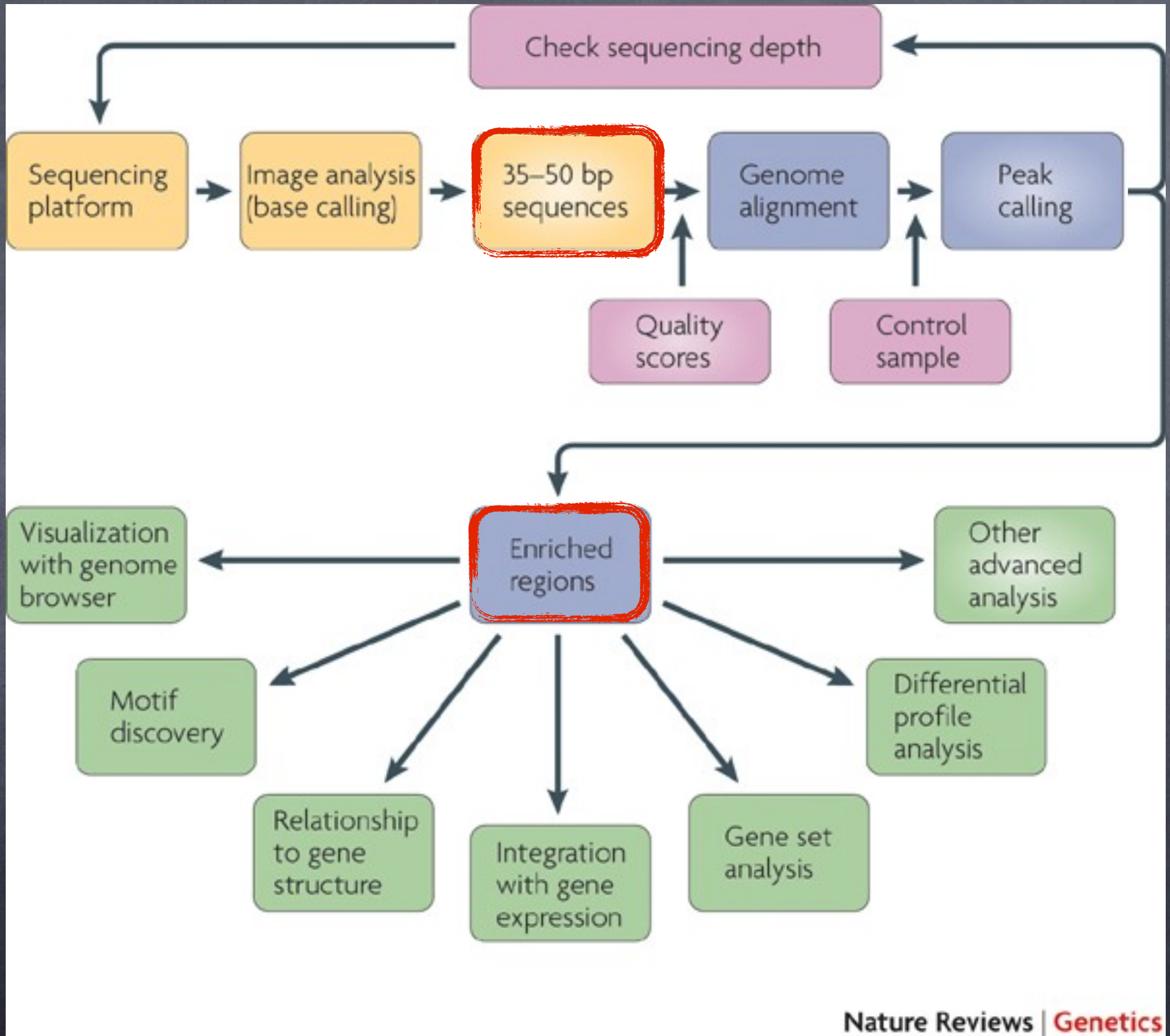
- good enrichment = good antibody (highly specific)
- ChIP grade antibodies, validation!


Check sequencing depth
Sequencing platform
Image analysis (base calling)
35–50 bp sequences
Genome alignment
Peak calling
Quality scores
Control sample
Visualization with genome browser
Enriched regions
Other advanced analysis
Motif discovery
Relationship to gene structure
Integration with gene expression
Gene set analysis
Differential profile analysis
Nature Reviews | Genetics

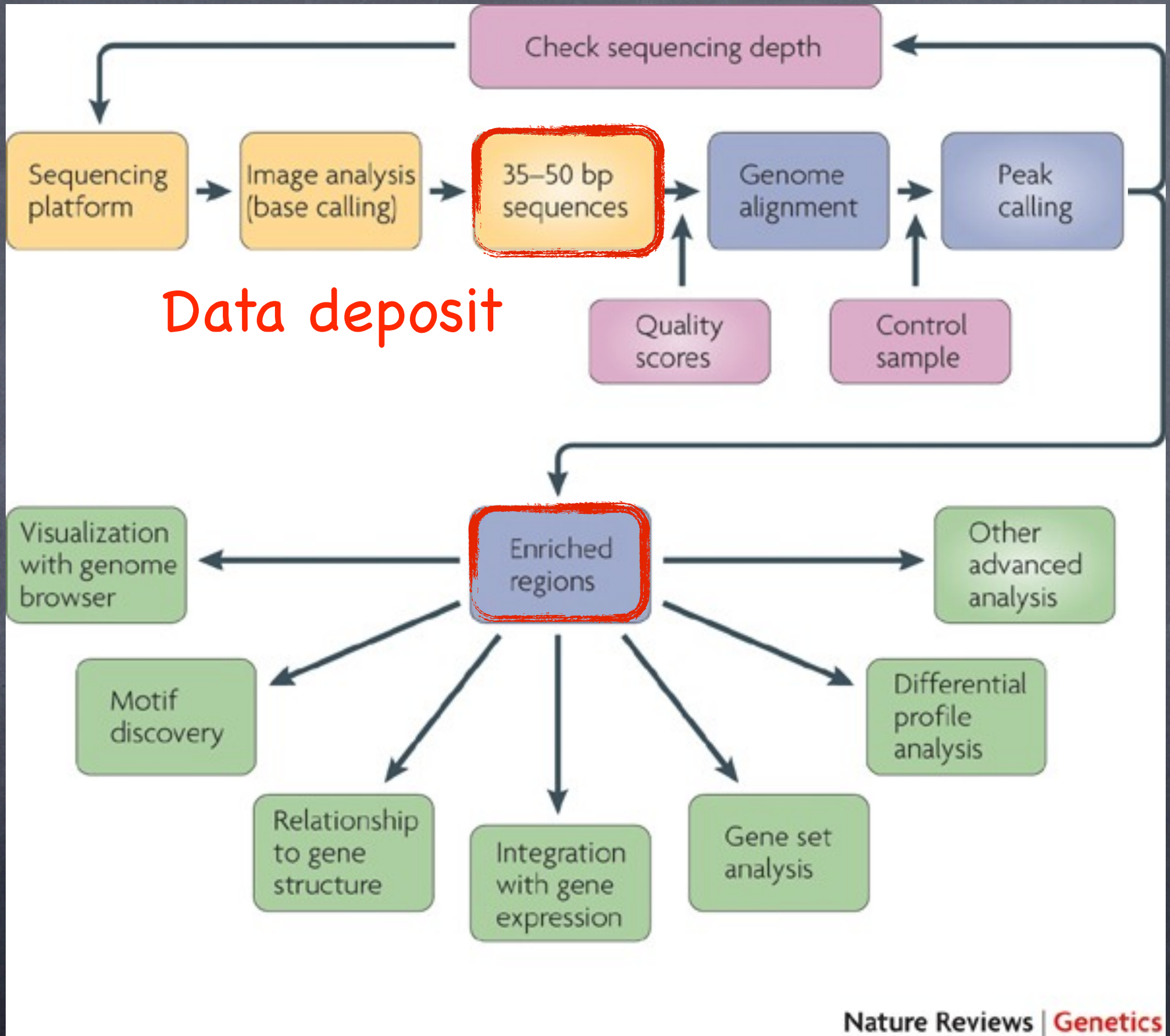
Park

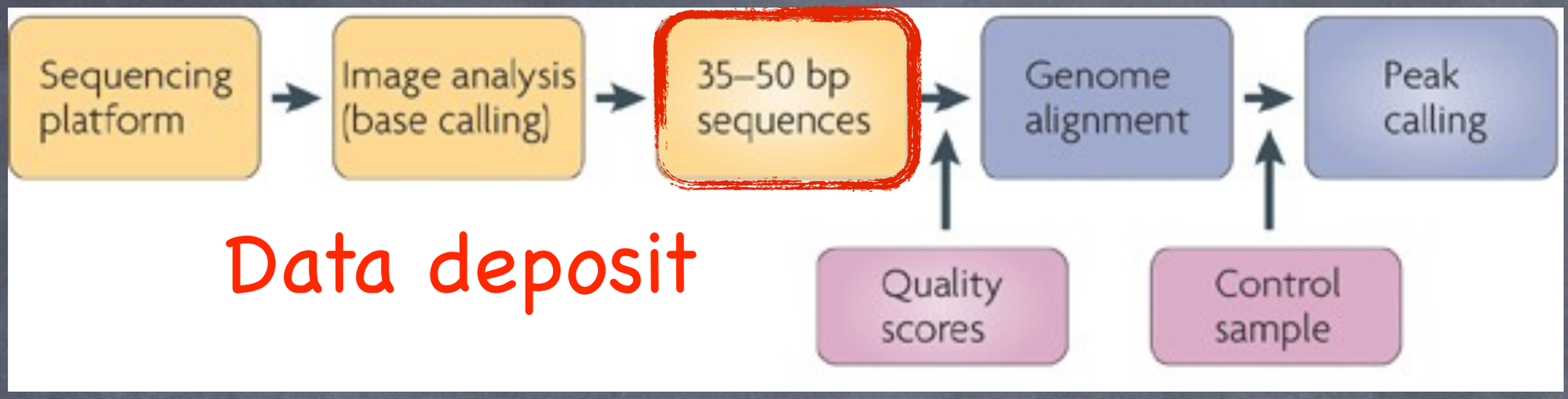

Check sequencing depth
Sequencing platform
Image analysis (base calling)
35–50 bp sequences
Genome alignment
Peak calling
Quality scores
Control sample
Visualization with genome browser
Enriched regions
Other advanced analysis
Motif discovery
Relationship to gene structure
Integration with gene expression
Gene set analysis
Differential profile analysis
Nature Reviews | Genetics

Park


Check sequencing depth
Sequencing platform
Image analysis (base calling)
35–50 bp sequences
Genome alignment
Peak calling
Quality scores
Control sample
Visualization with genome browser
Enriched regions
Other advanced analysis
Motif discovery
Relationship to gene structure
Integration with gene expression
Gene set analysis
Differential profile analysis
Nature Reviews | Genetics

Park


Check sequencing depth
Sequencing platform
Image analysis (base calling)
35–50 bp sequences
Genome alignment
Peak calling
Data deposit
Quality scores
Control sample
Park
Visualization with genome browser
Enriched regions
Other advanced analysis
Motif discovery
Relationship to gene structure
Integration with gene expression
Gene set analysis
Differential profile analysis
Nature Reviews | Genetics



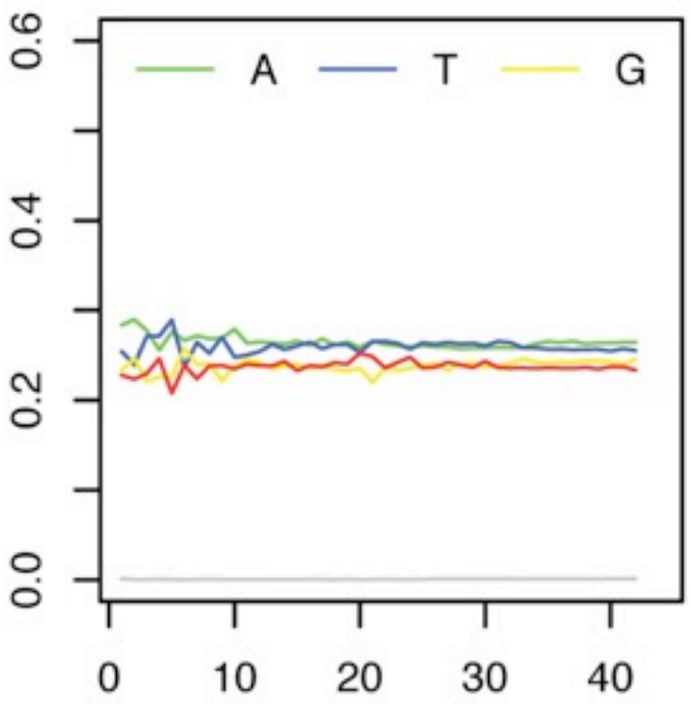
Data deposit

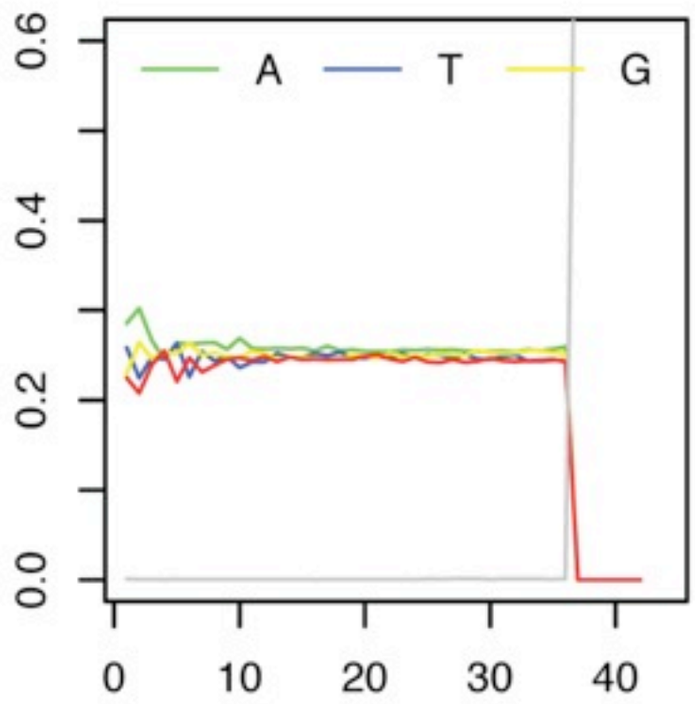
Park

Initial QC

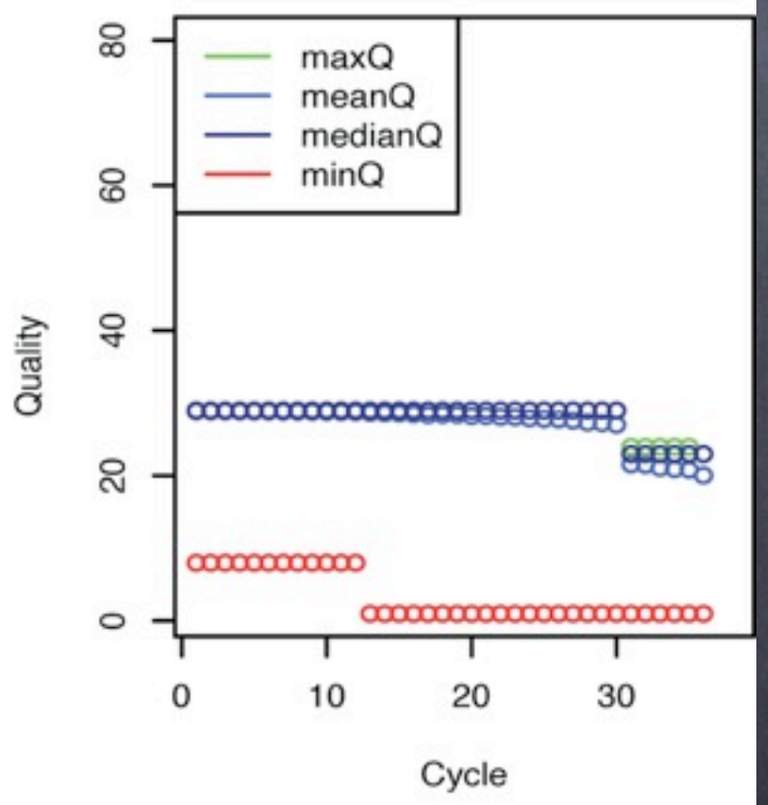
(fastq level)

Base frequency plots



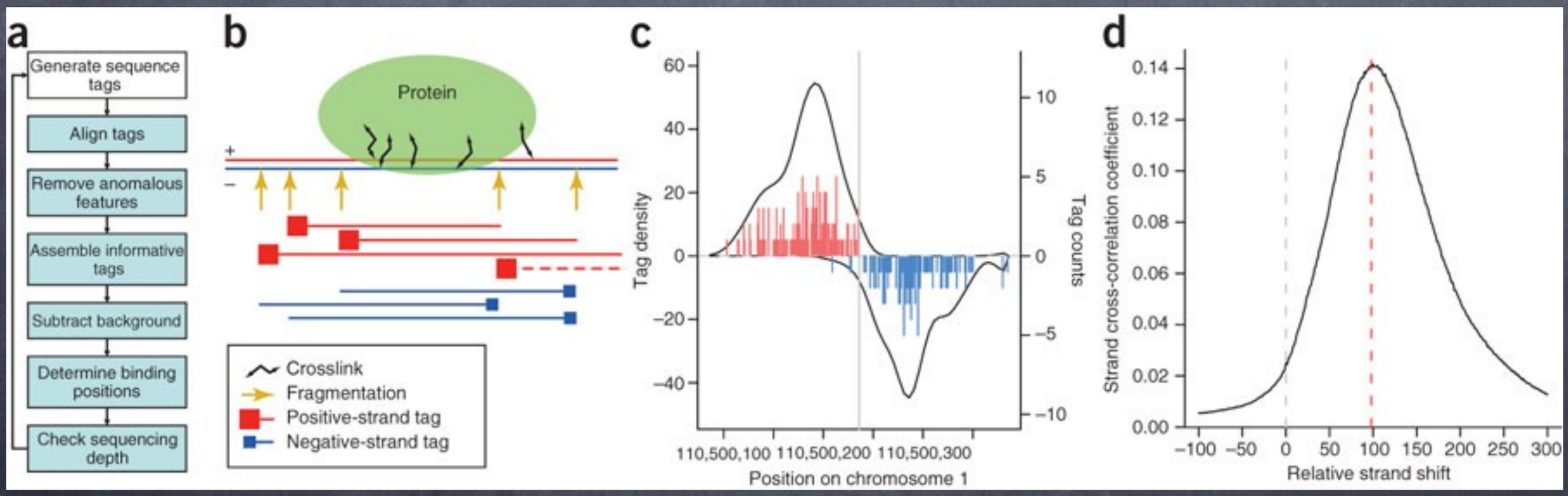


Quality score



=> reject lanes? trim reads?

Alignment

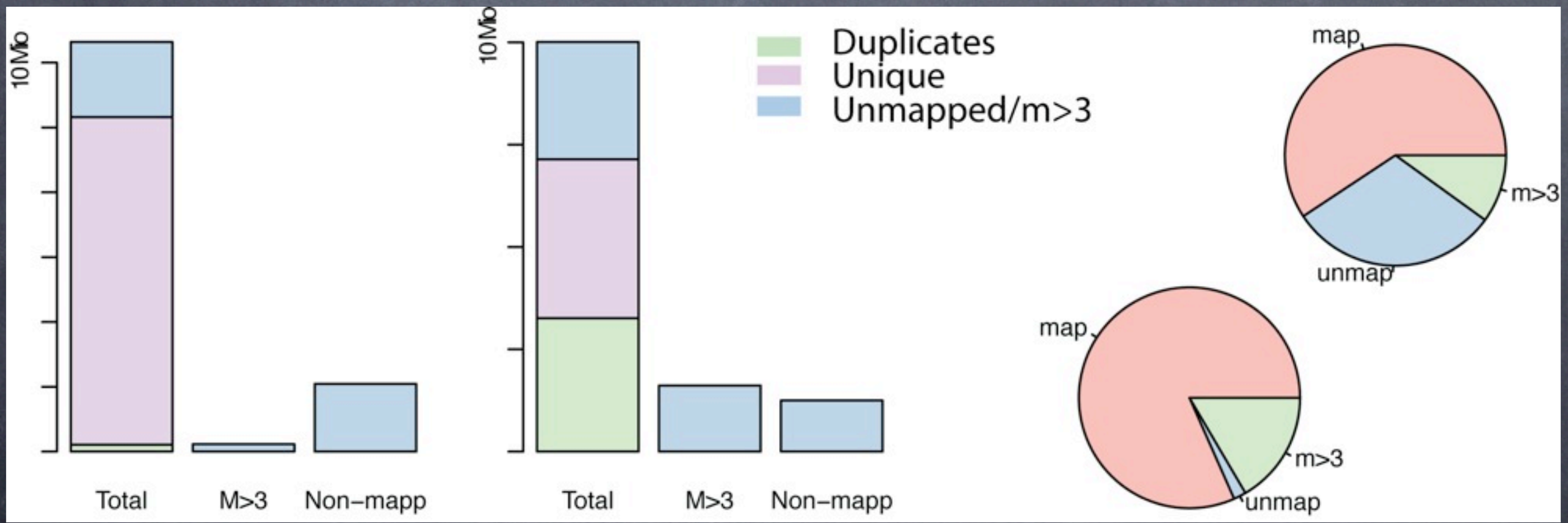


Kharchenko et al., Nature Biotechnology

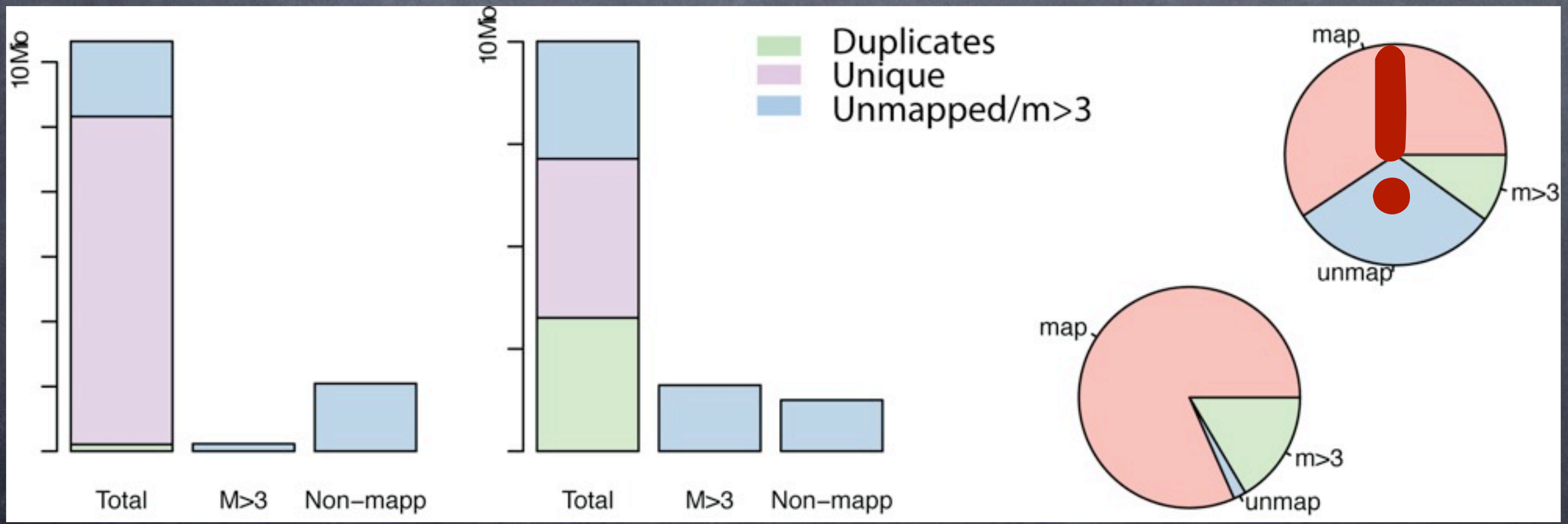
Alignment

- poorly aligned tags
- alignable genomes (repeat content)
- aligner of choice (typically BWA, Bowtie), parameters
- filtering for uniquely mapping sites/unique reads

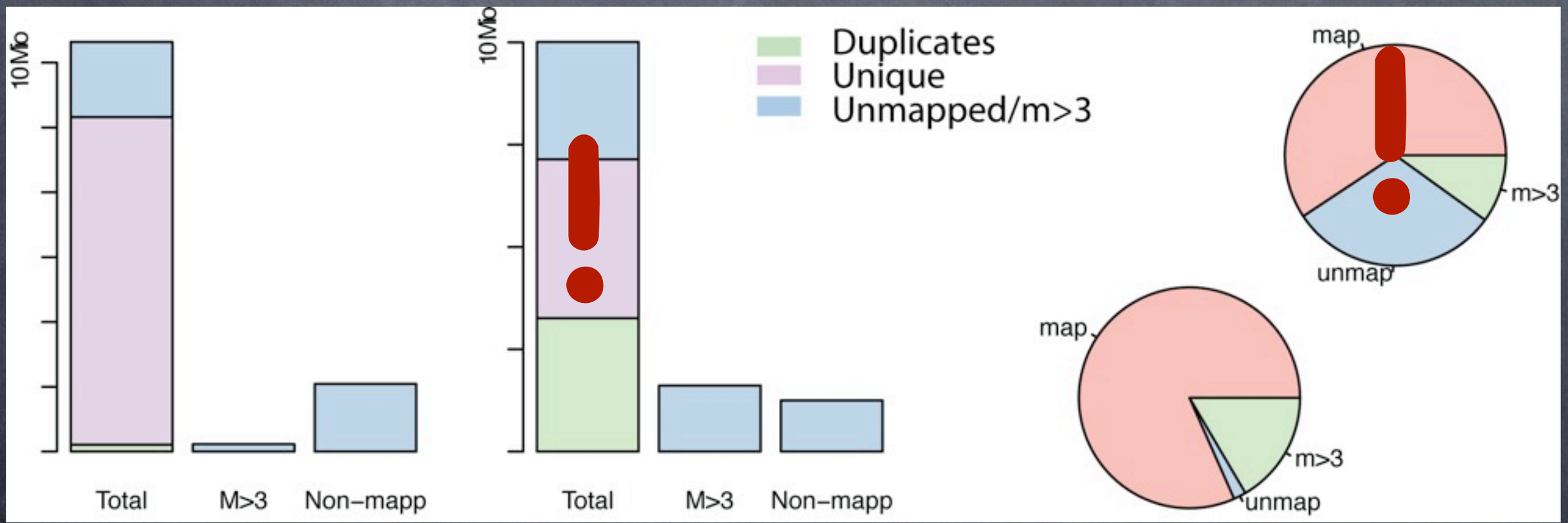
Alignment


10Mio
Total
M>3
Non−mapp
10Mio
Total
M>3
Non−mapp
Duplicates
Unique
Unmapped/m>3
map
m>3
unmap
map
m>3
unmap

Alignment


10Mio
Total
M>3
Non−mapp
10Mio
Total
M>3
Non−mapp
Duplicates
Unique
Unmapped/m>3
map
m>3
unmap
map
m>3
unmap

Alignment


10Mio
Total
M>3
Non−mapp
10Mio
Total
M>3
Non−mapp
Duplicates
Unique
Unmapped/m>3
map
m>3
unmap
map
m>3
unmap

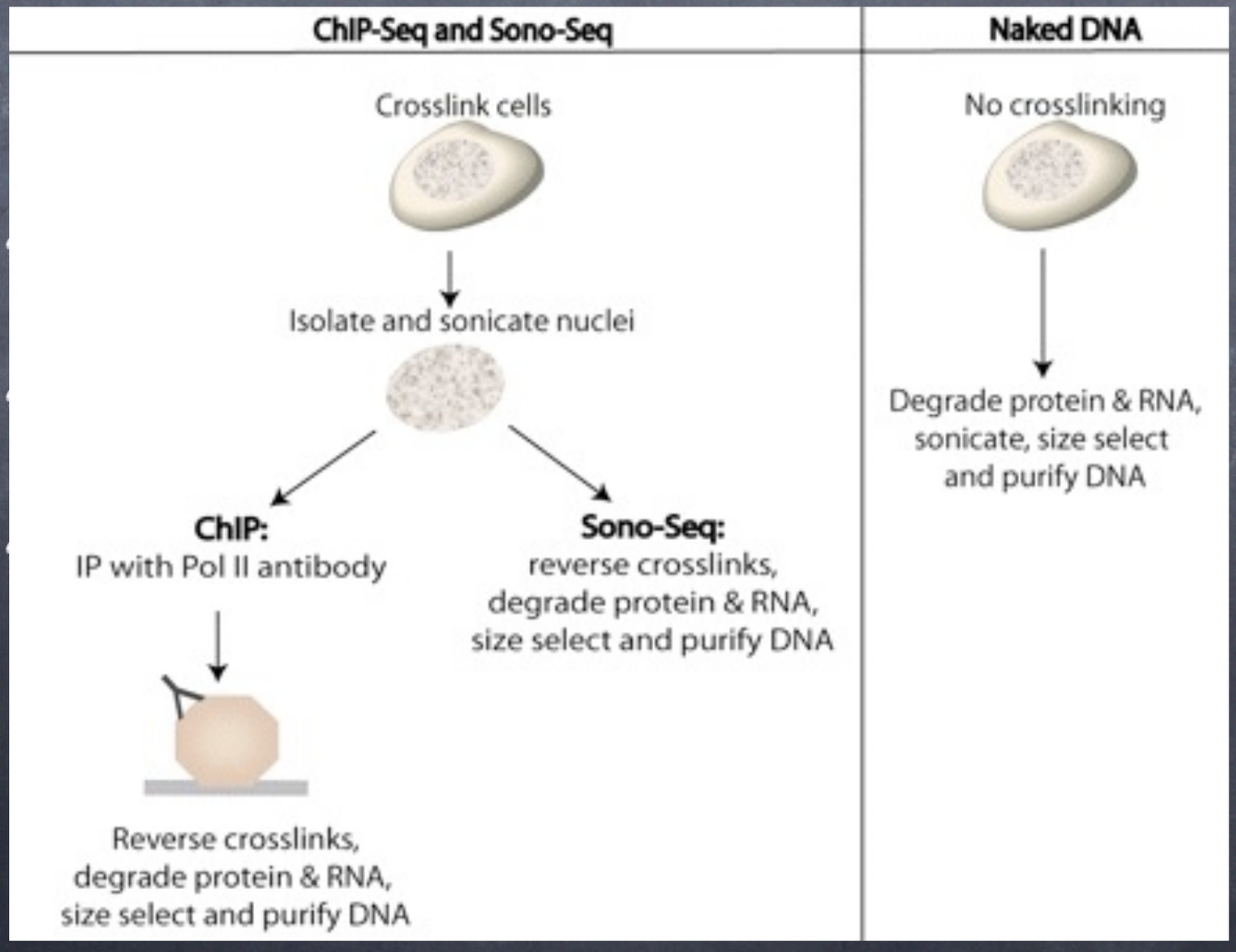
Control

need for an INPUT - background tag distribution non-uniformly around genome, several types of anomalies:

- non-uniform DNA shearing
- repetitive regions
- GC biases

Control

need for an INPUT - background tag distribution non-uniformly around genome, several types of anomalies:



Control

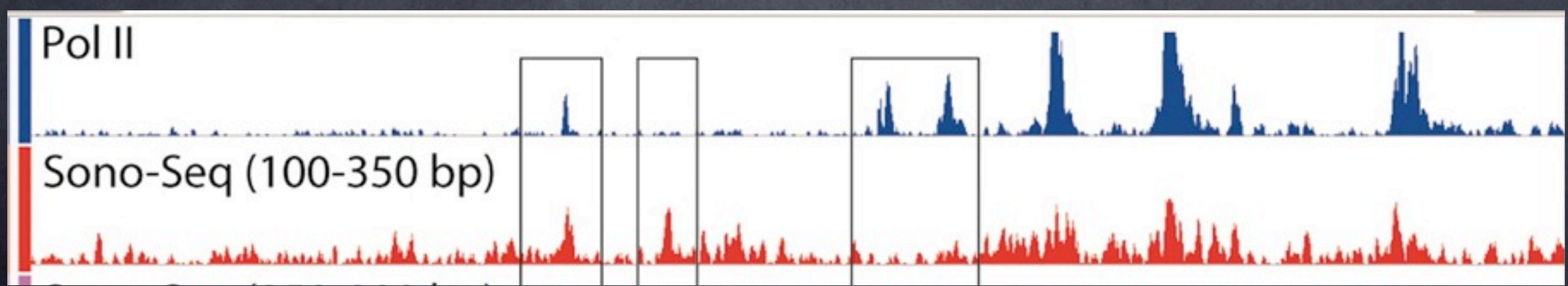
need for an INPUT - background tag distribution non-uniformly around genome, several types of anomalies:

- non-uniform DNA shearing
- repetitive regions
- GC biases

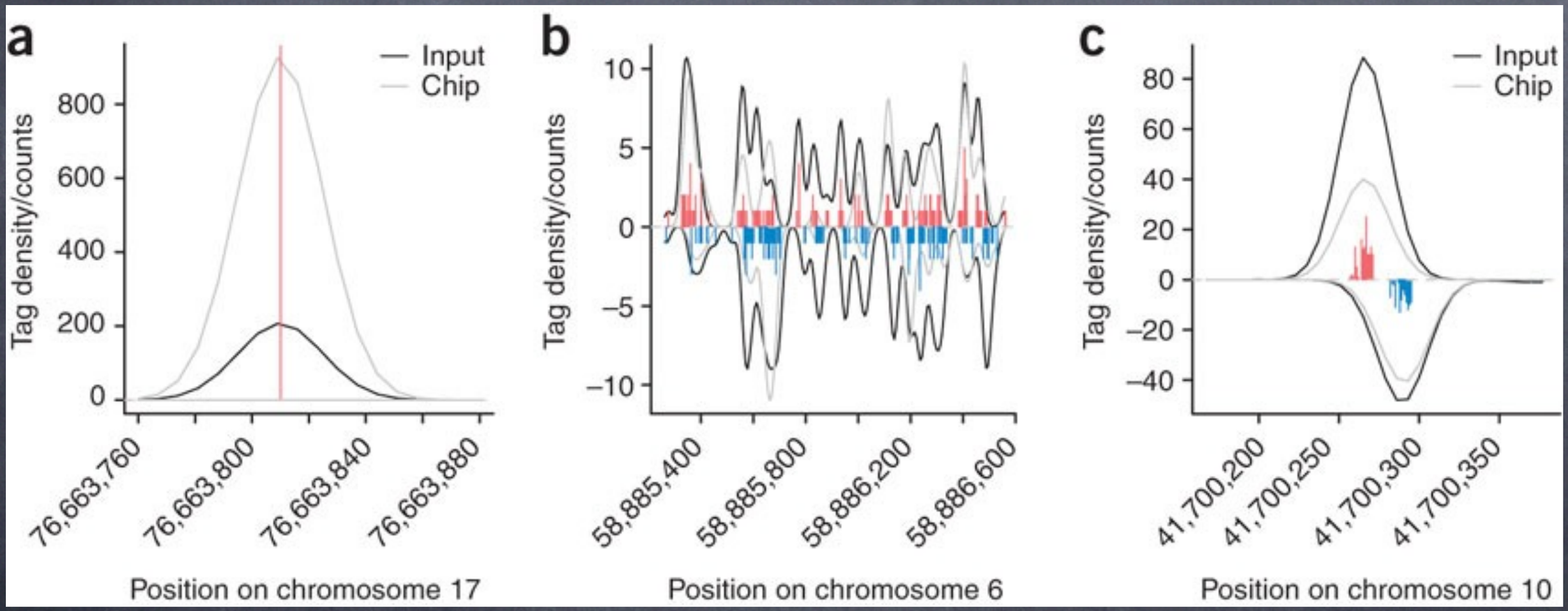
Control

need for an INPUT - background tag distribution non-uniformly around genome, several types of anomalies:

- non-uniform DNA shearing
- repetitive regions
- GC biases

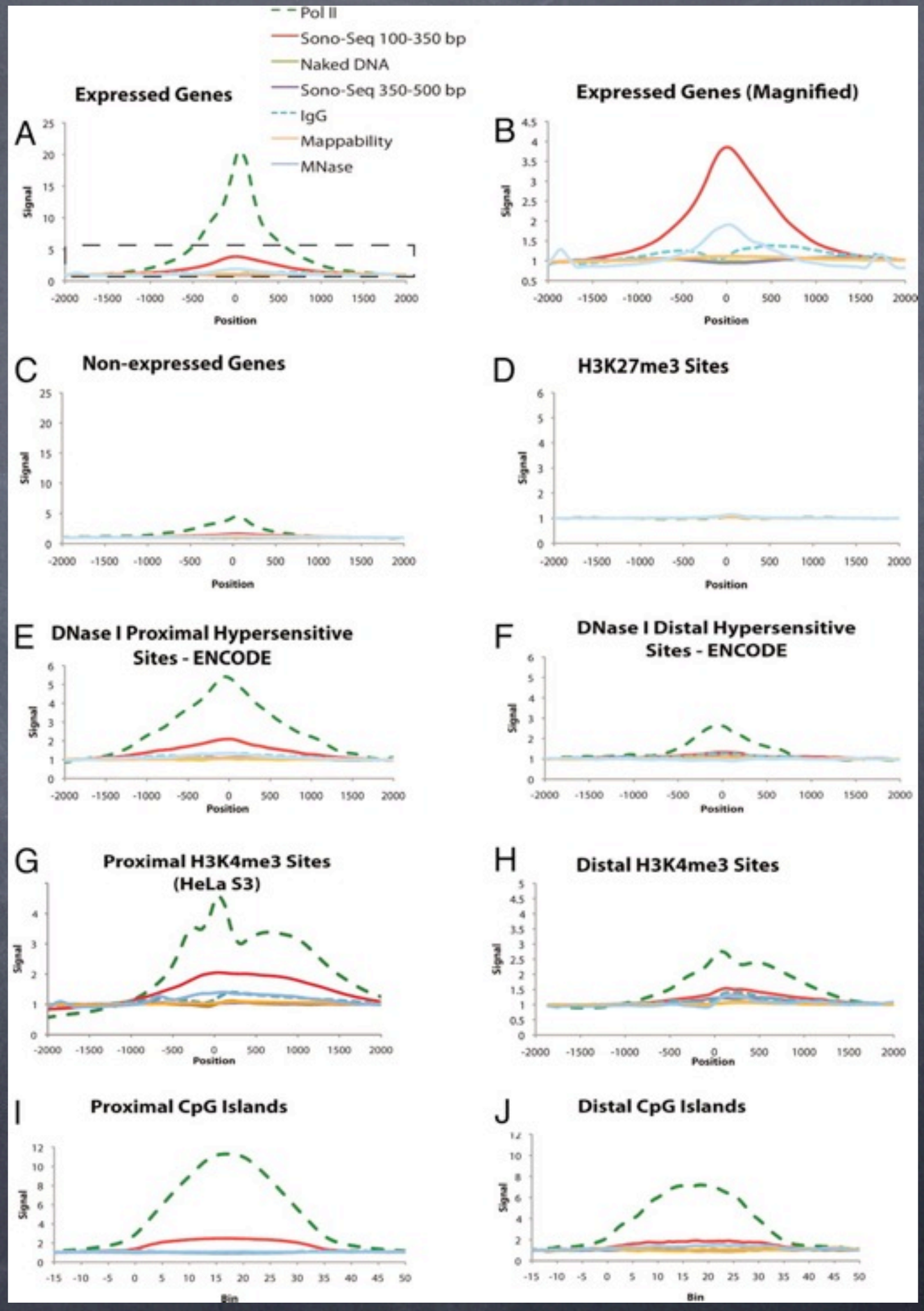


Input tag distribution



Auerbach et al, PNAS

Sono-seq


Pol II
Sono-Seq 100-350 bp
Naked DNA
Sono-Seq 350-500 bp
IgG
Mappability
MNase
A
Expressed Genes
B
Expressed Genes (Magnified)
C
Non-expressed Genes
D
H3K27me3 Sites
E
DNase I Proximal Hypersensitive Sites - ENCODE
F
DNase I Distal Hypersensitive Sites - ENCODE
G
Proximal H3K4me3 Sites (HeLa S3)
H
Distal H3K4me3 Sites
I
Proximal CpG Islands
J
Distal CpG Islands
Signal
Position
Bin

Use of replicates

Peak calling on the pool

Intersection of replicates

Intersection of replicates and pool

- Testing on mouse data (several factors, 3 mouse strains): highest peak reproducibility for pooled approach
- replicate pooling before peak-calling

Read depth

- how many reads are needed?
- saturation

Multiplexing

- no. of reads/lane increasing --> ability to sequence multiple samples in one lane important for cost effectiveness
- barcoding of samples during preparation

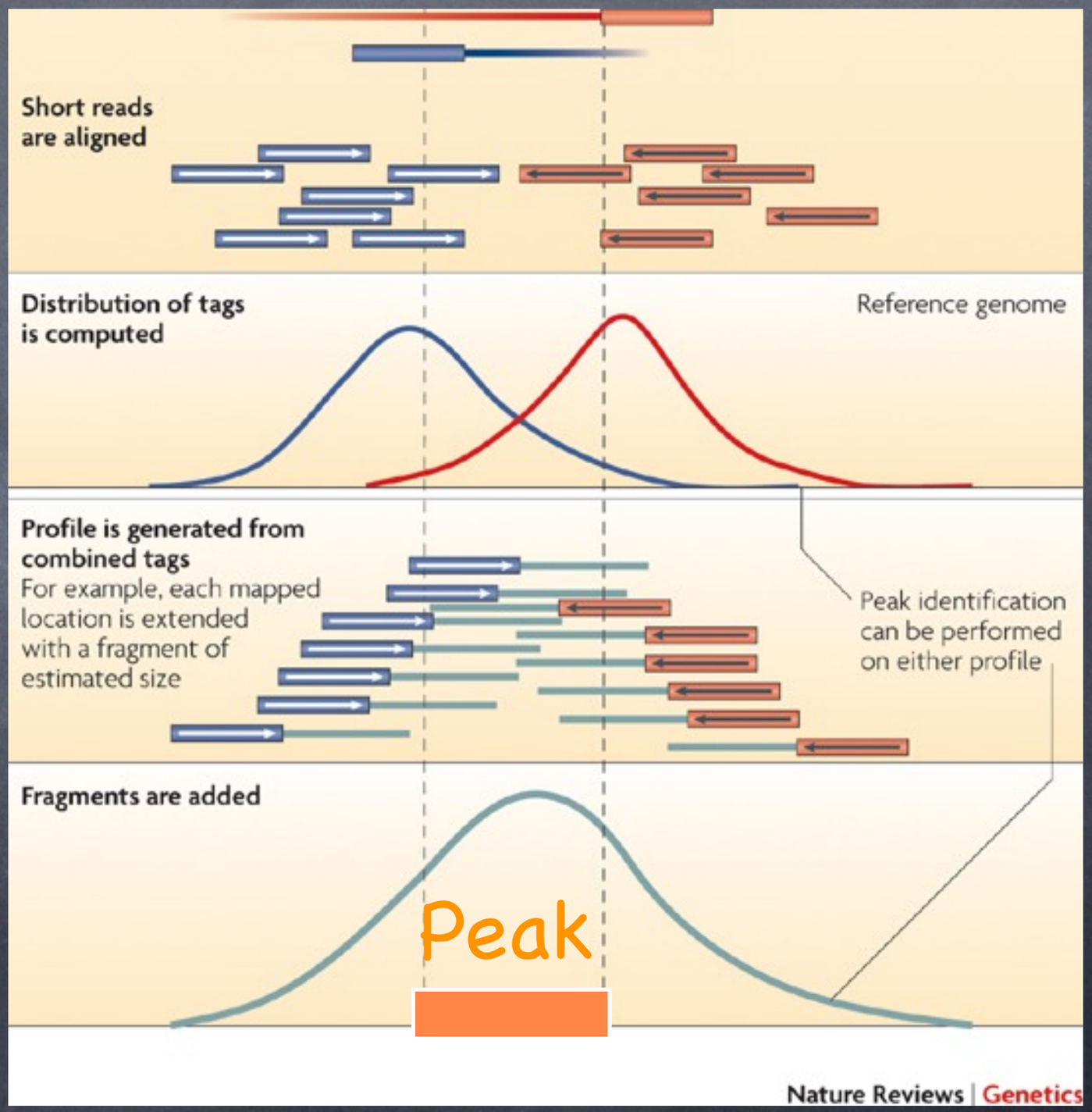
Peak-calling

pehaps add pic from review;
mention how many methods there a
popular ones, that will be presented i
of the talk

enrichment over input
minimum tag density
reads directionality

MACS, FindPeaks,
PeakSeq, BayesPeak,
Useq, ...

run time
memory
compilation
flexibility



SWEMBL

Steven Wilder, EMBL

- fast and stable
- precise localization of the peak summit
- no assumptions about the shape of the peak
- automatically optimizes parameters for different no. of factor and input reads
- no linear dependance between read and peak number

SWEMBL

- simple function that can be rapidly calculated at every position in the genome
- uses read directionality

$F_0 = 0$ at start of chromosome

$F_n = \max(F_{n-1} + count_n - penalty_n, 0)$

- value 0: starting a new region
- peak called from start of the region to max F_n

SWEMBL

- Penalty function: distance from previous aligned read, reads from control DNA sample, read base quality, read uniqueness, GC content, sequence features
- extension: copy number variants

add swembl output screenshot and explain

SWEMBL I/O

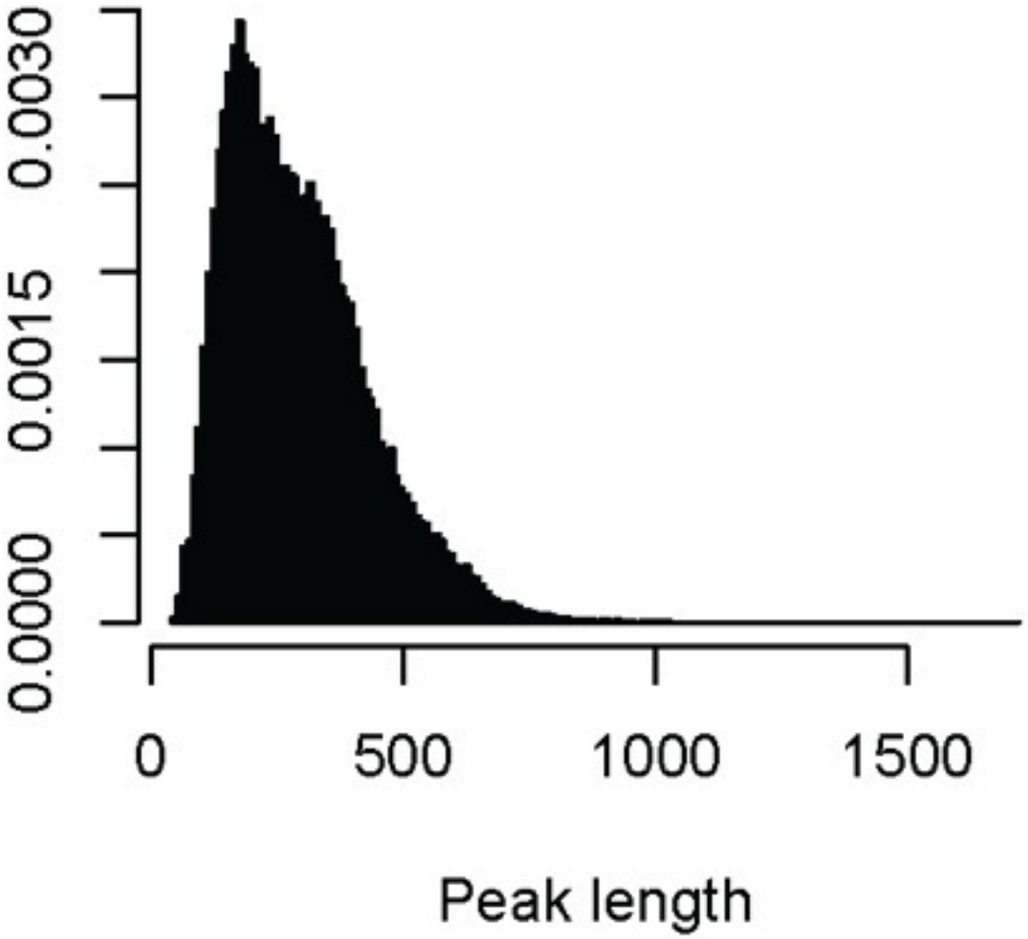
- supports sam, bed, ... input; automatically estimates parameters based on read counts

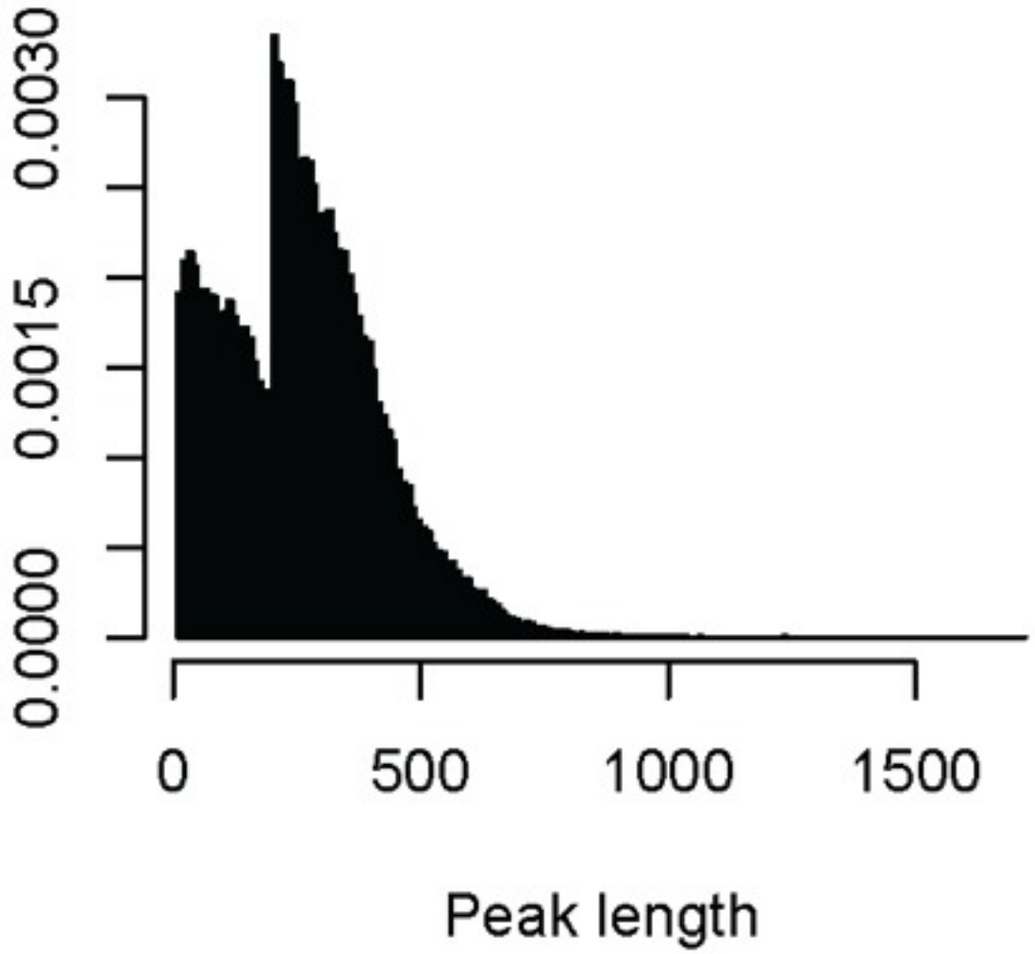
```
$ SWEMBL  -R 0.005 -S -i Aligned.factor.sam
      -r Aligned.input.sam -o Peaks.sw3
```

- output in SWEMBL tab-delimited format, includes peak score, max.coverage and summit

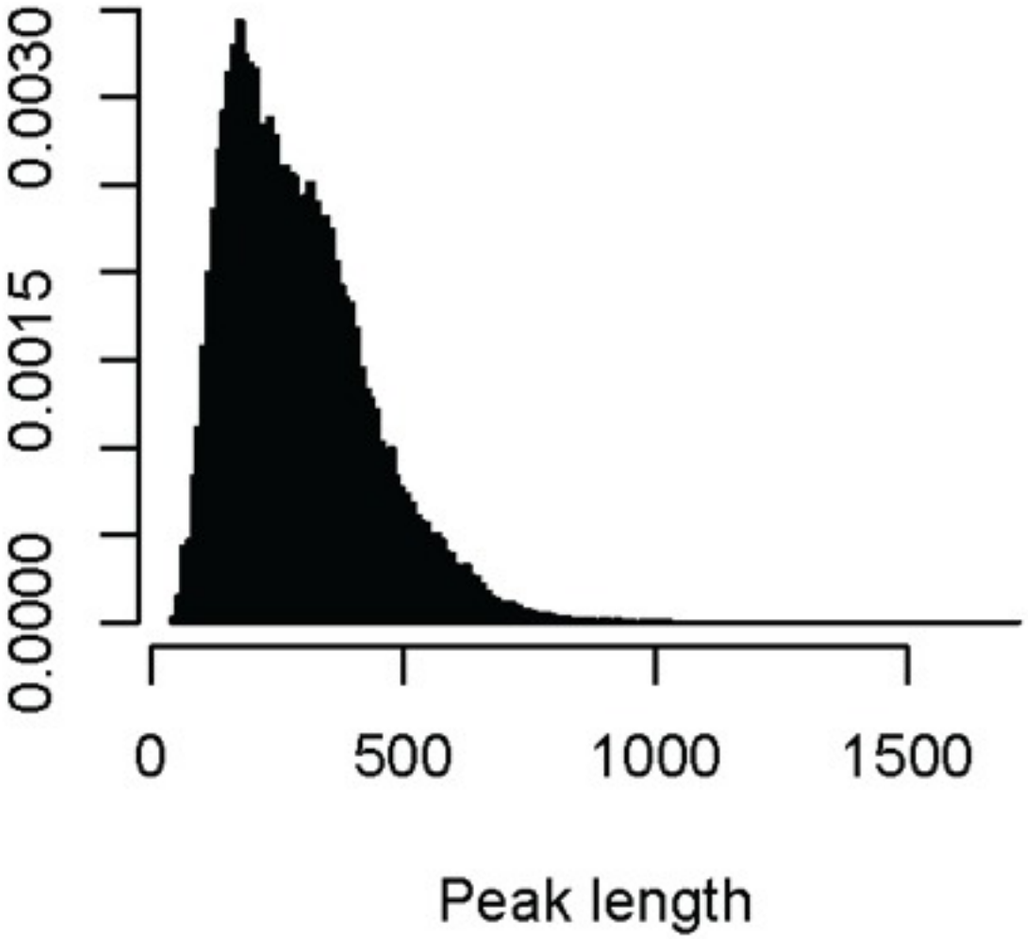
Region	Start pos.	End pos.	Count	Length
Unique pos.	Score	Ref. count	Max. Coverage	Summit

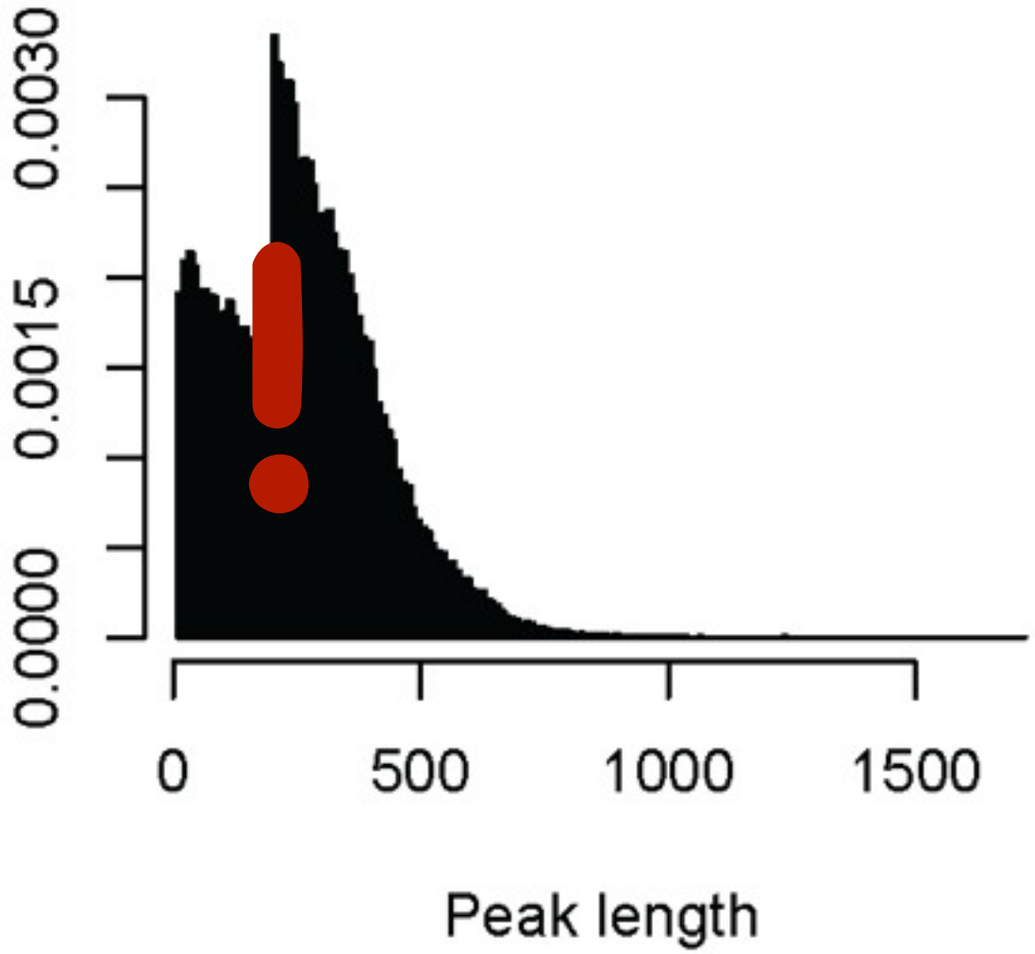
peak width?


0.0030
0.0015
0.0000
0
500
1000
1500
Peak length

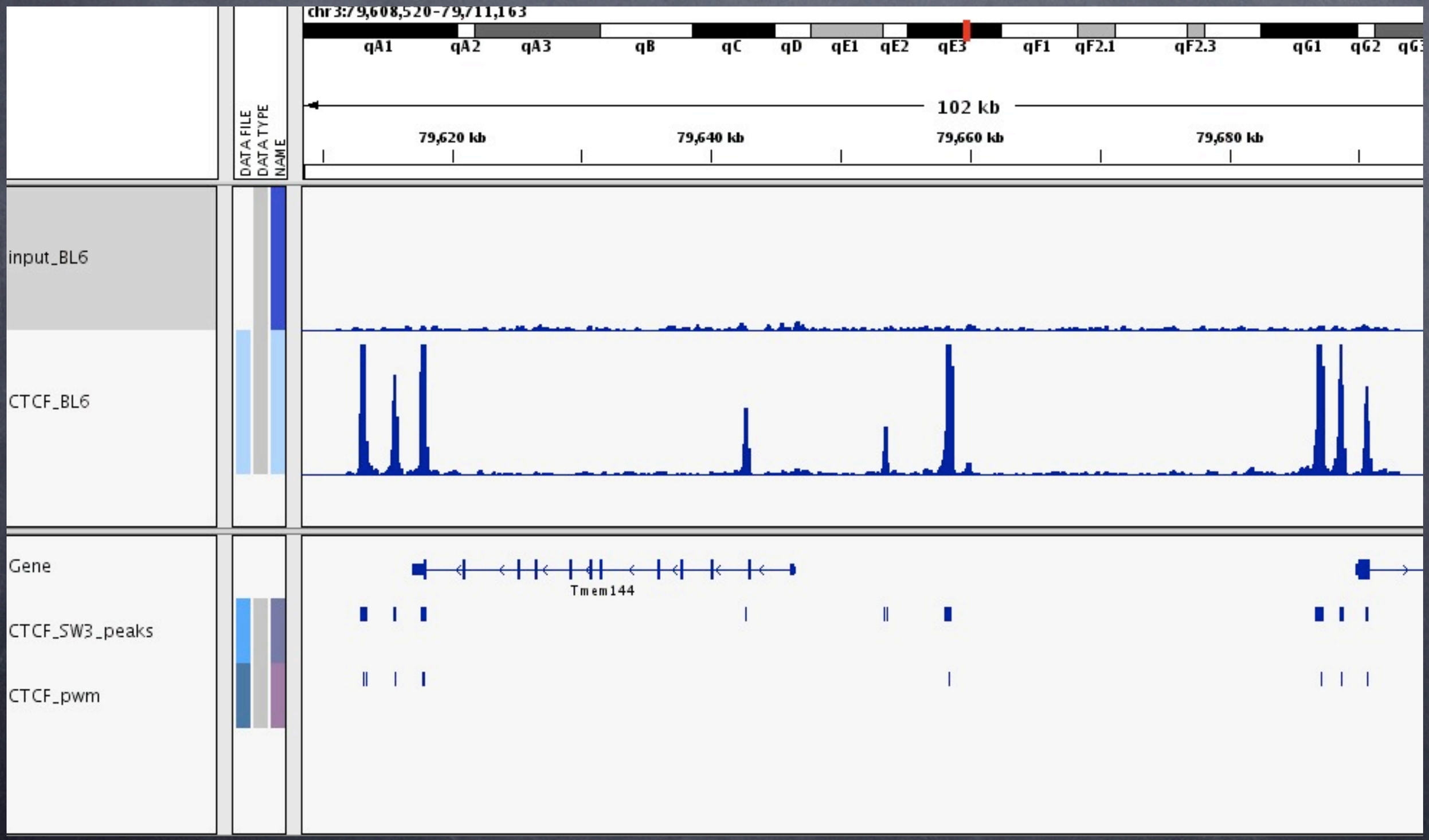

0.0030
0.0015
0.0000
0
500
1000
1500
Peak length

peak width?

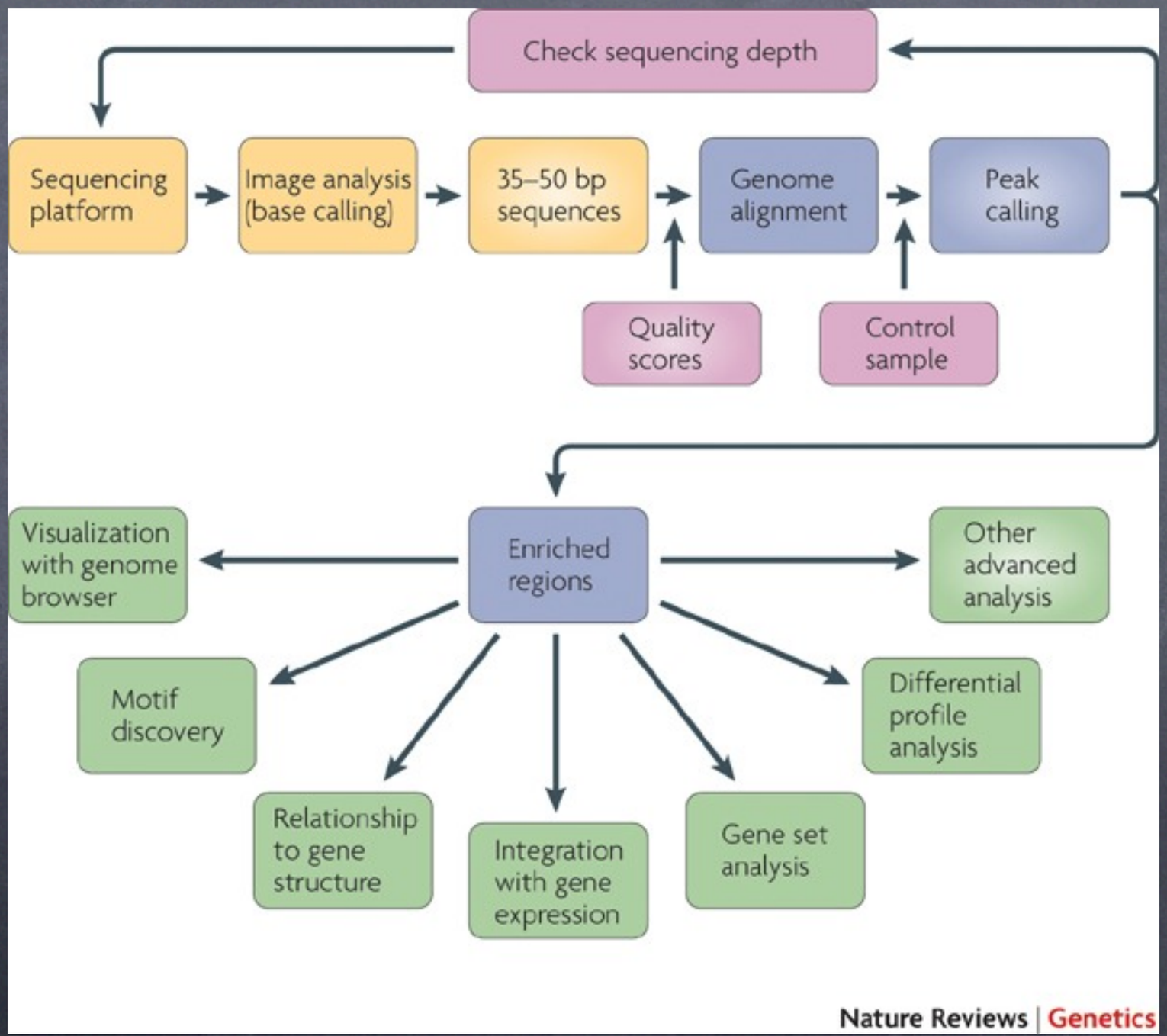

0.0030
0.0015
0.0000
0
500
1000
1500
Peak length


0.0030
0.0015
0.0000
0
500
1000
1500
Peak length

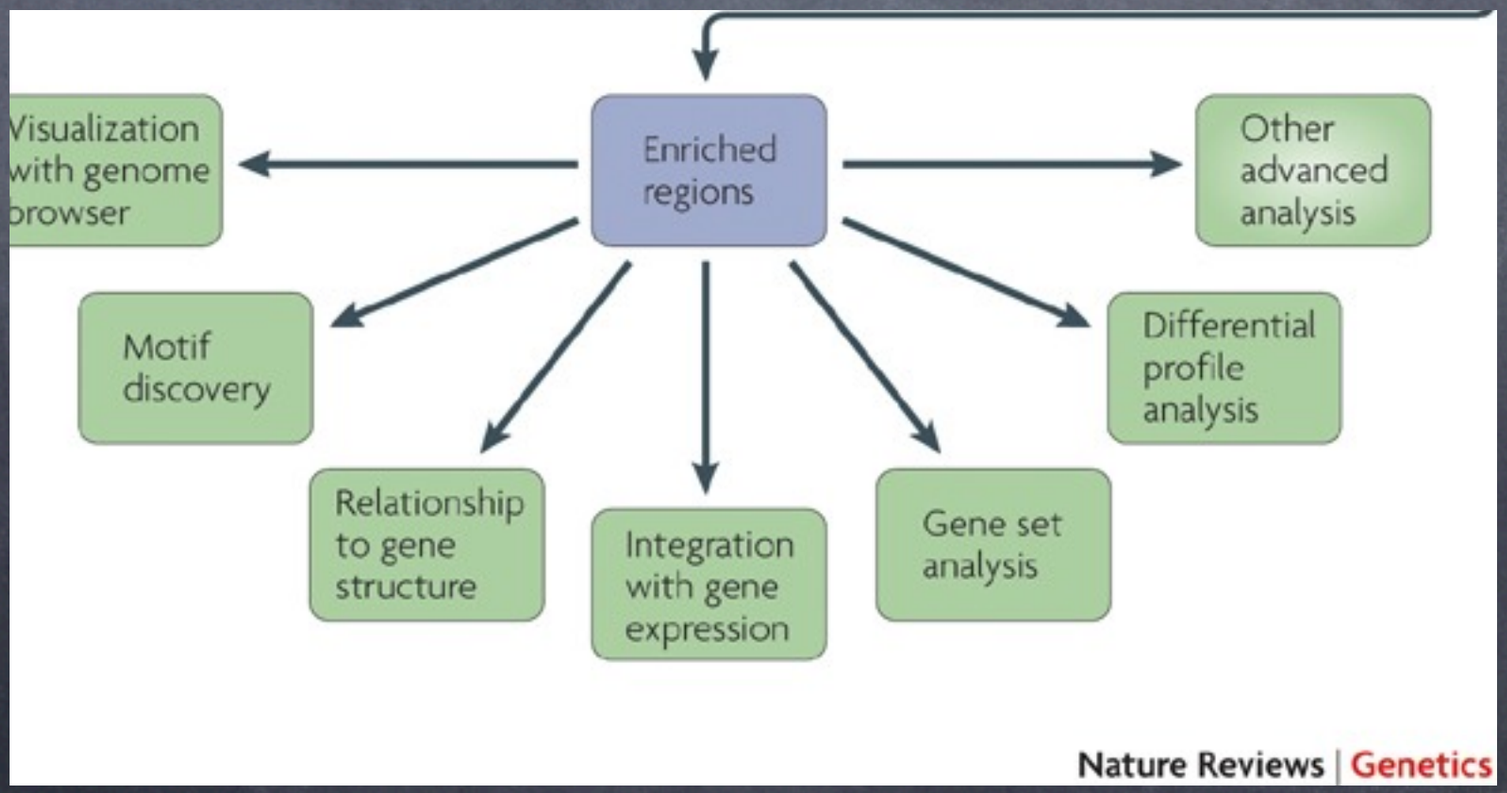
SWEMBL


chr3:79,608,520-79,711,163
qA1
qA2
qA3
qB
qC
qD
qE1
qE2
qE3
qF1
qF2.1
qF2.3
qG1
qG2
102 kb
79,620 kb
79,640 kb
79,660 kb
79,680 kb
DATA FILE
DATA TYPE
NAME
input_BL6
CTCF_BL6
Gene
Tmem144
CTCF_SW3_peaks
CTCF_pwm

Downstream analysis



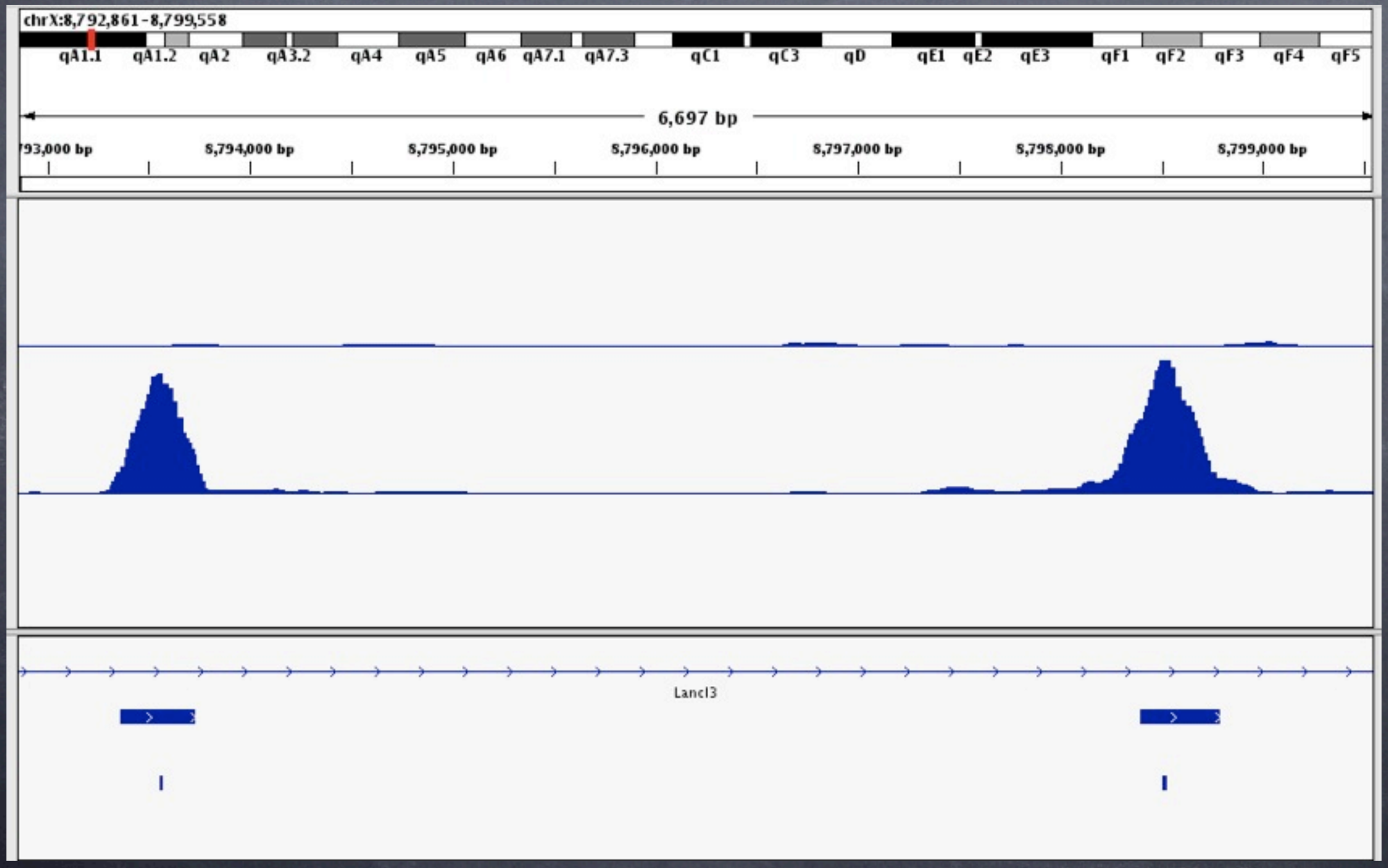
Downstream analysis



Visualisation

- formats: sam, bam, wig, bedgraph
- toolboxes: Samtools (C), Picard (Java), Bio-SamTools (Perl), Pysam (Python), igvtools
- Browsers: USCS Genome Browser, Ensembl, IGB, IGV, ...

Visualisation: IGV



Motif discovery

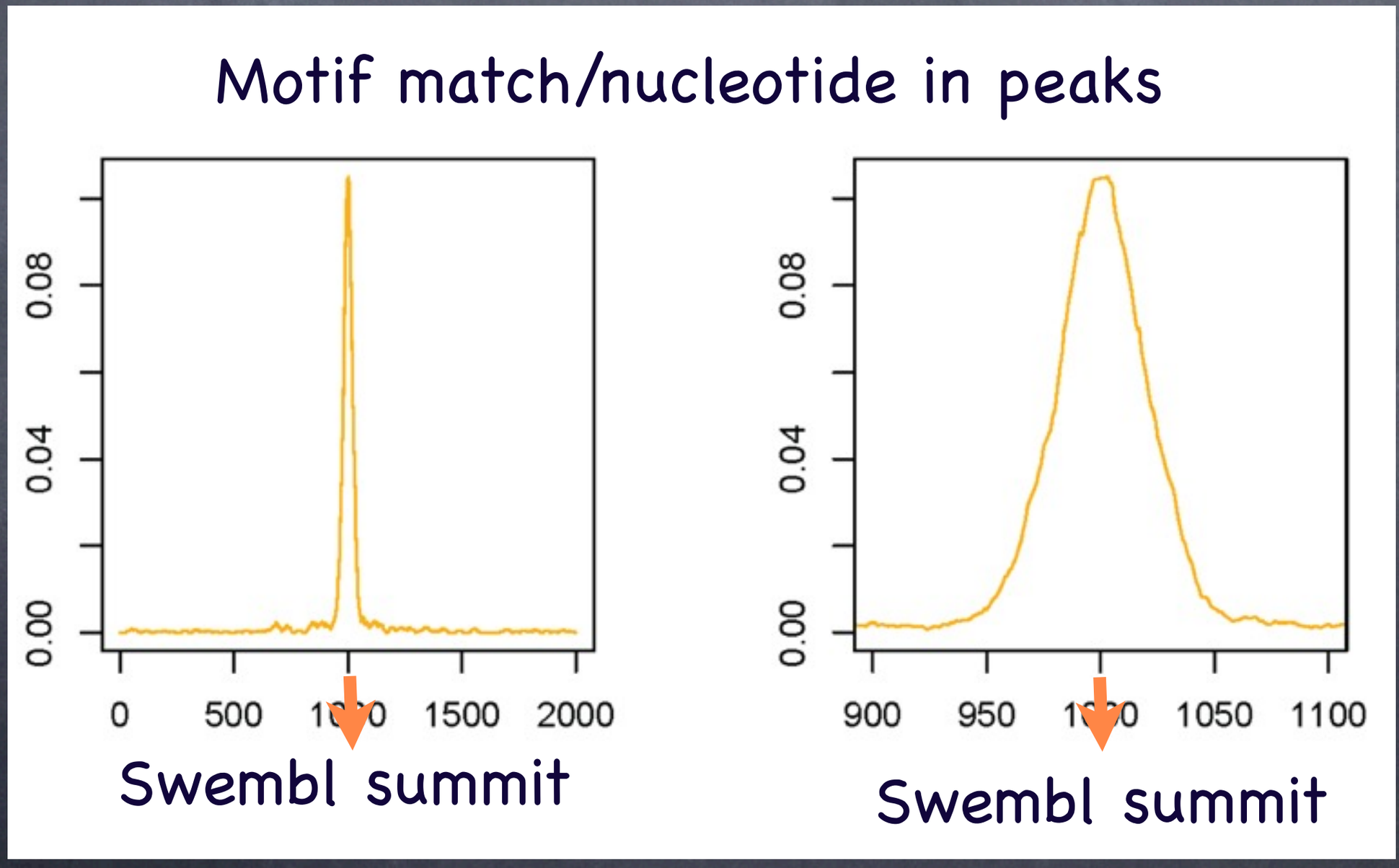
- motif scan, de novo motif discovery
- popular tools like MEME, Nmica, Weeder, AlignACE, ... but also new ones especially designed for ChIP-seq
- high ChIP resolution: centering of the motif under the peak summit, high % of sites that have the motif

iMotifs

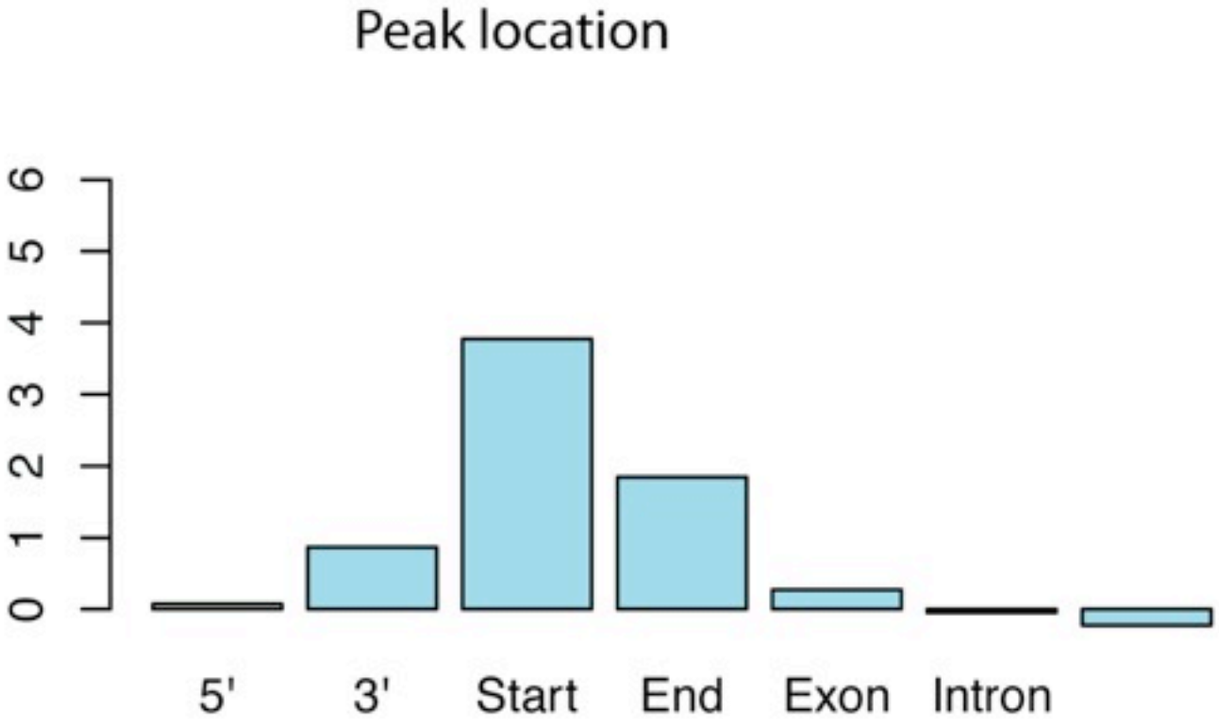
(Matias Piipari, Sanger)

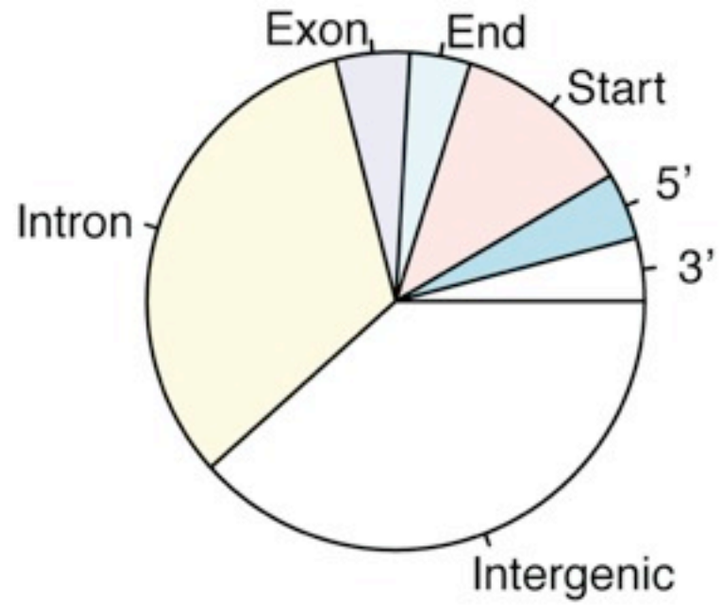
iMotifs

Motif scan


Motif match/nucleotide in peaks
0.08
0.04
0.00
0
500
1500
2000
Swembl summit
0.08
0.04
0.00
900
950
1050
1100
Swembl summit

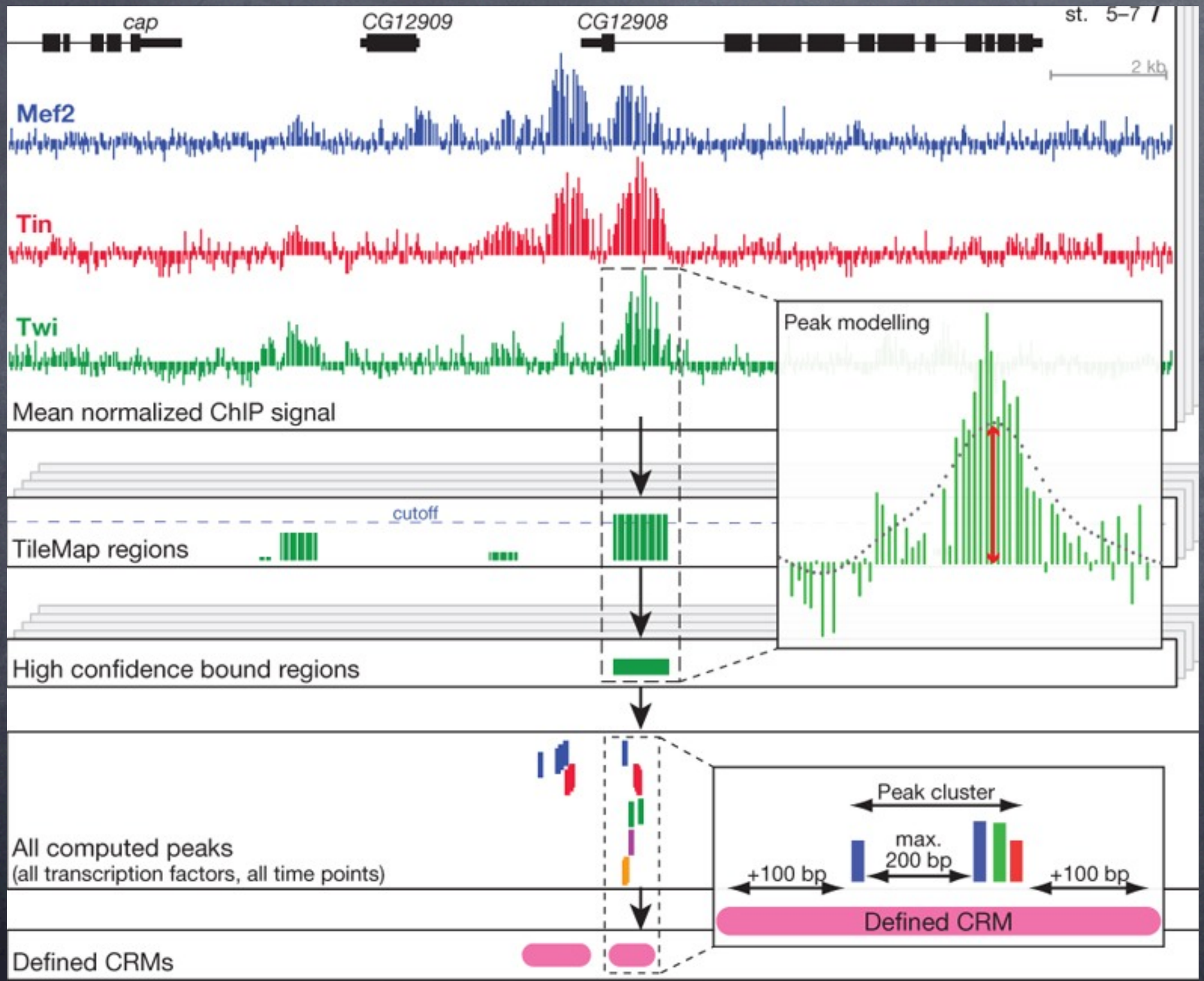
Peak location





- gene proximal location of peaks?
- integration of expression information: peaks in proximity of genes belonging to a group
- gene ontology, GSEA

Co-occurrence, modules

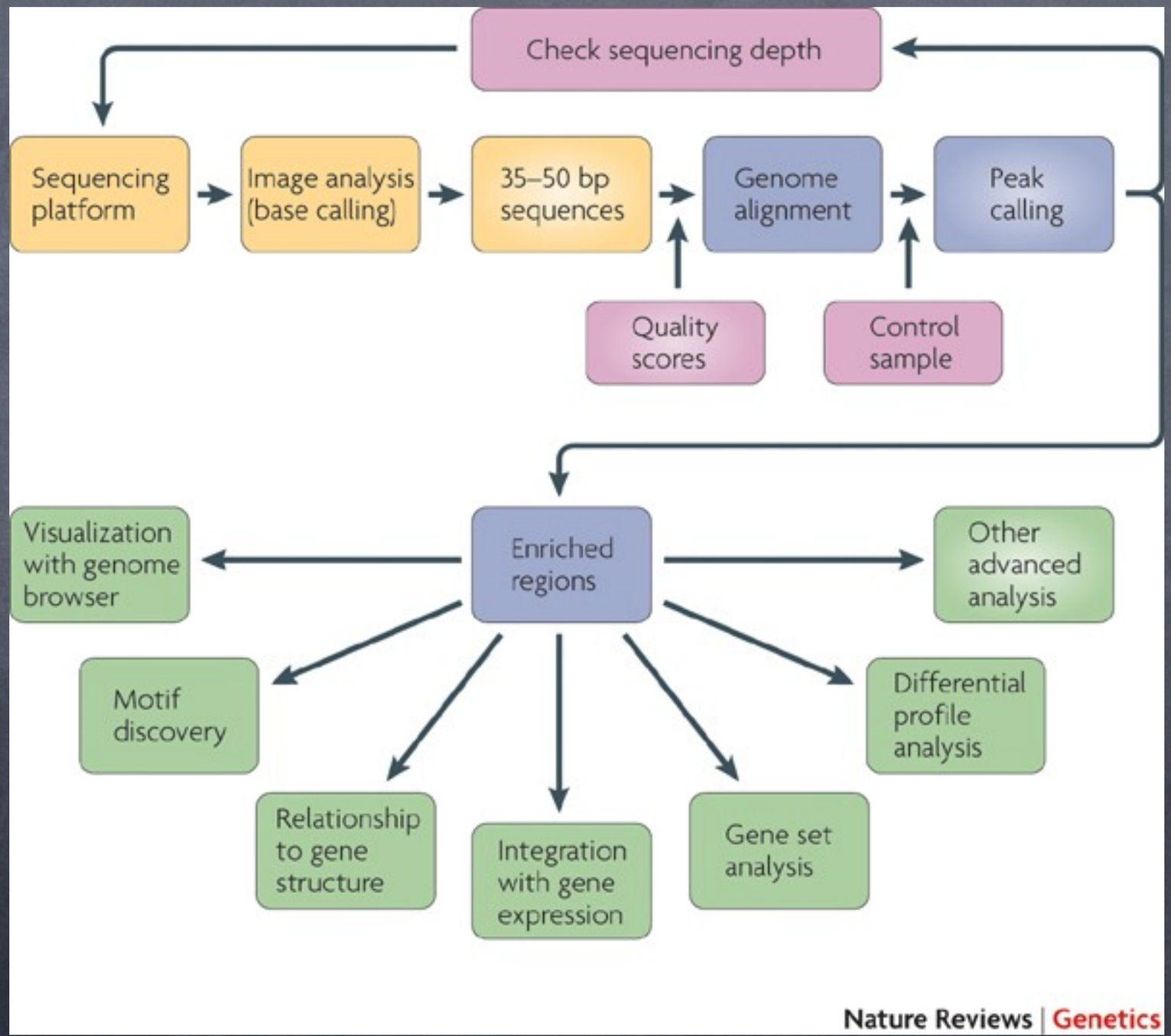


Zinzen et. al, Nature

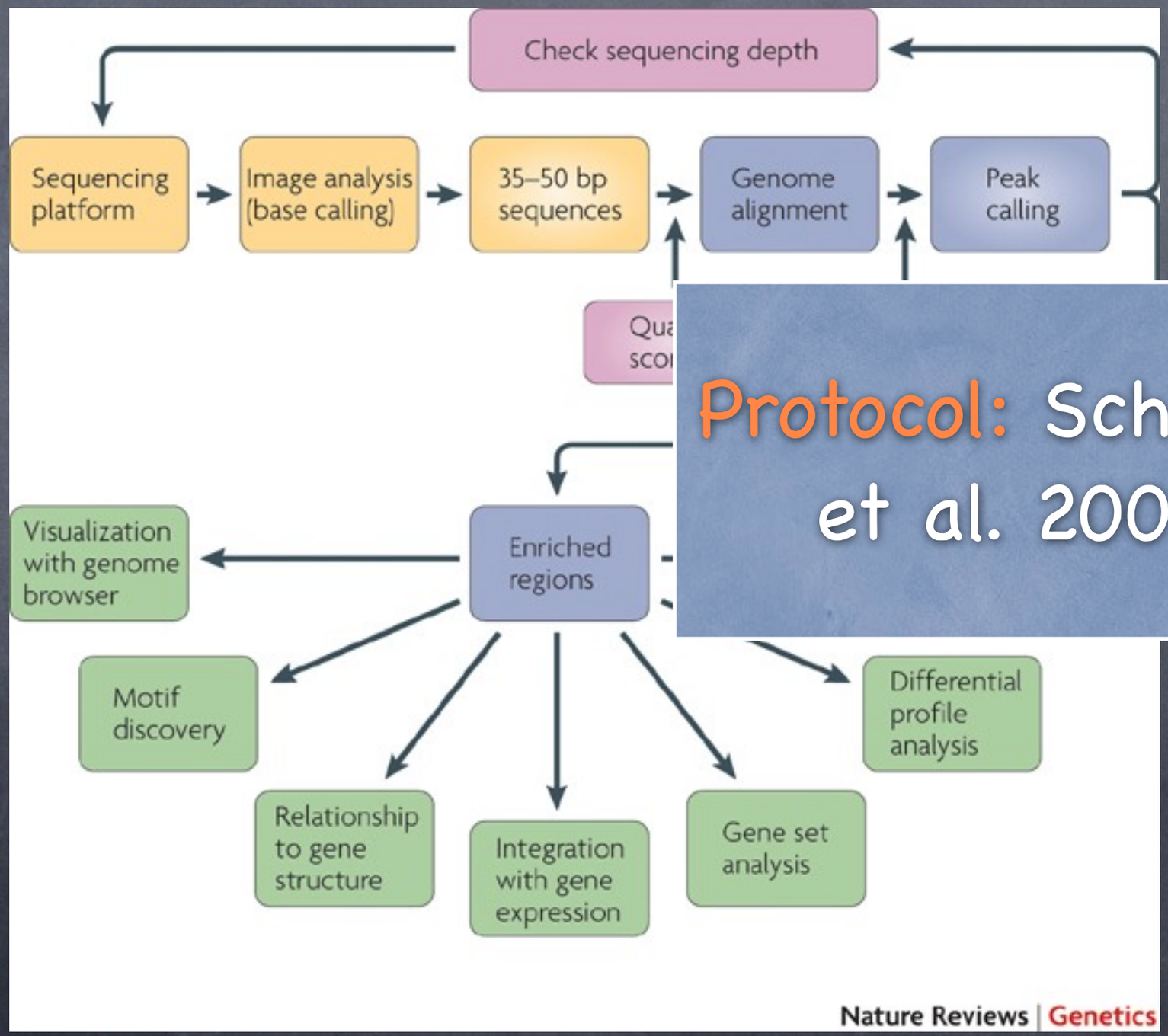
Further analysis

- Additional info: chromatin state
- Differential binding: conditions, cell types
- Evolutionary aspect: binding versus sequence conservation
- Integration with expression information: eQTLs in TFBS

summary

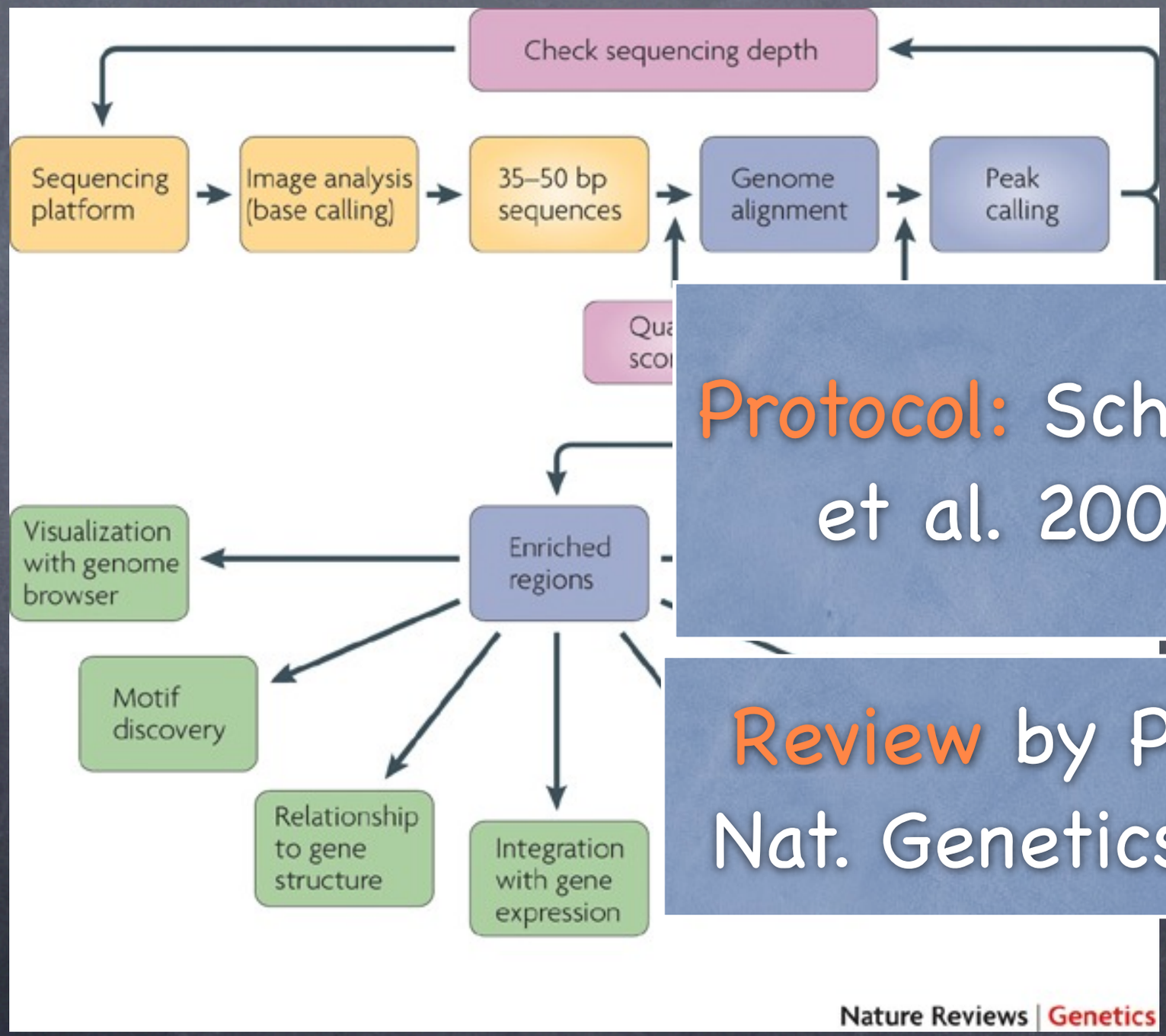

Check sequencing depth
Sequencing platform
Image analysis (base calling)
35–50 bp sequences
Genome alignment
Peak calling
Quality scores
Control sample
Visualization with genome browser
Enriched regions
Other advanced analysis
Motif discovery
Differential profile analysis
Relationship to gene structure
Integration with gene expression
Gene set analysis
Nature Reviews | Genetics

summary



Protocol: Schmidt et al. 2009

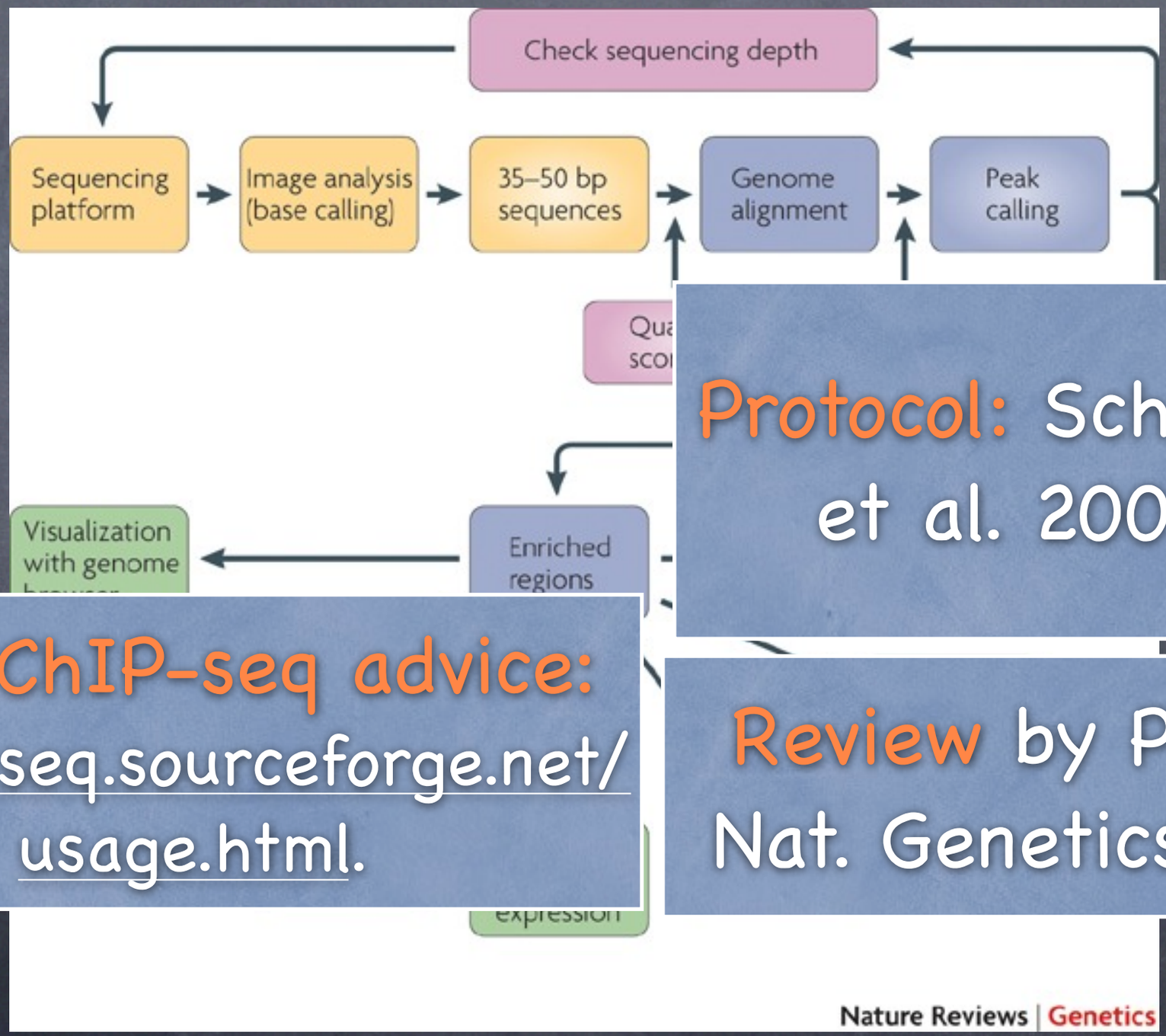
summary



Protocol: Schmidt et al. 2009

Review by Park Nat. Genetics 09

summary

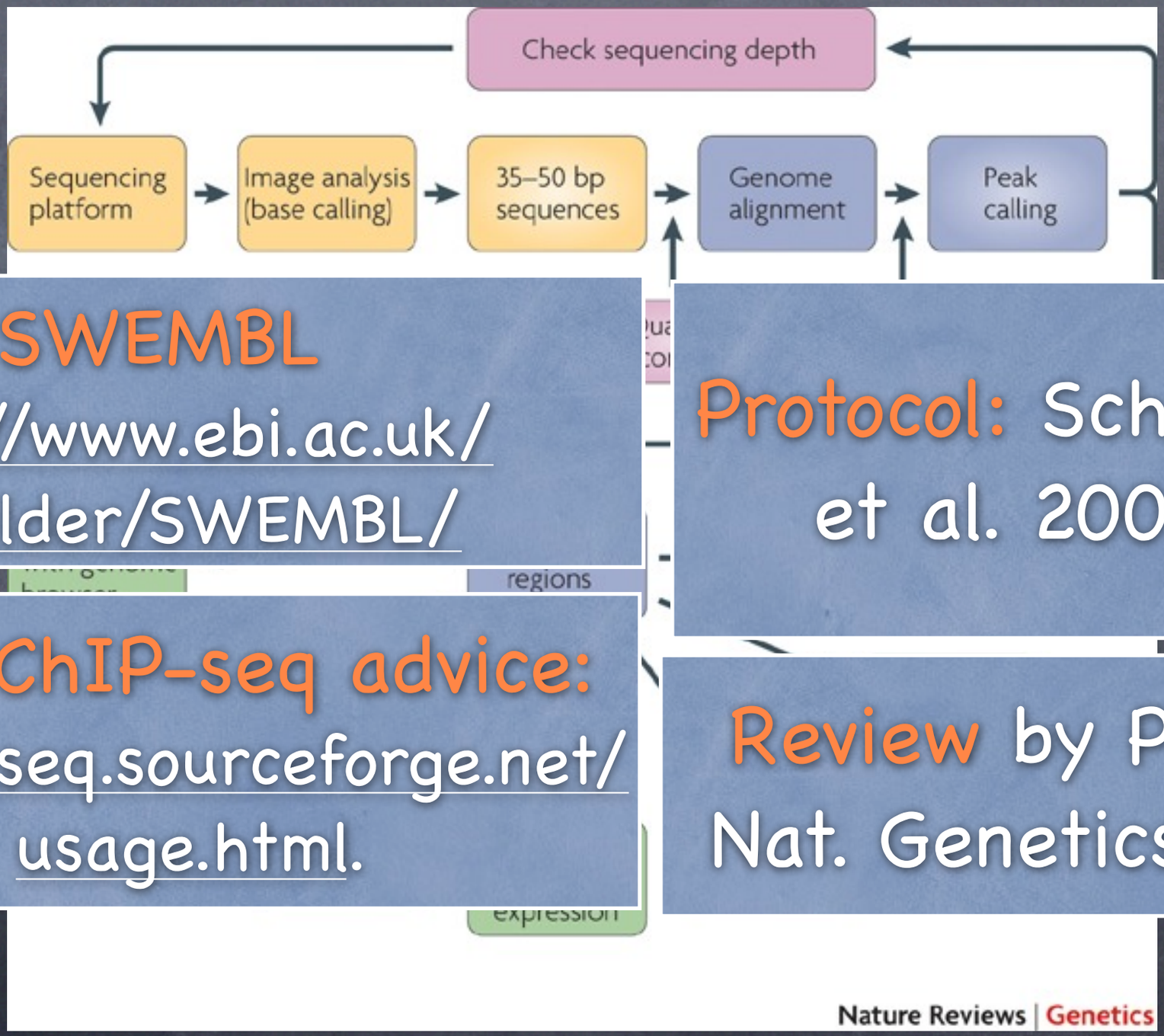


Protocol: Schmidt et al. 2009

Useq ChIP-seq advice: http://useq.sourceforge.net/usage.html.

Review by Park Nat. Genetics 09

summary



SWEMBL
http://www.ebi.ac.uk/~swilder/SWEMBL/

Protocol: Schmidt et al. 2009

Useq ChIP-seq advice:
http://useq.sourceforge.net/usage.html.

Review by Park Nat. Genetics 09

summary

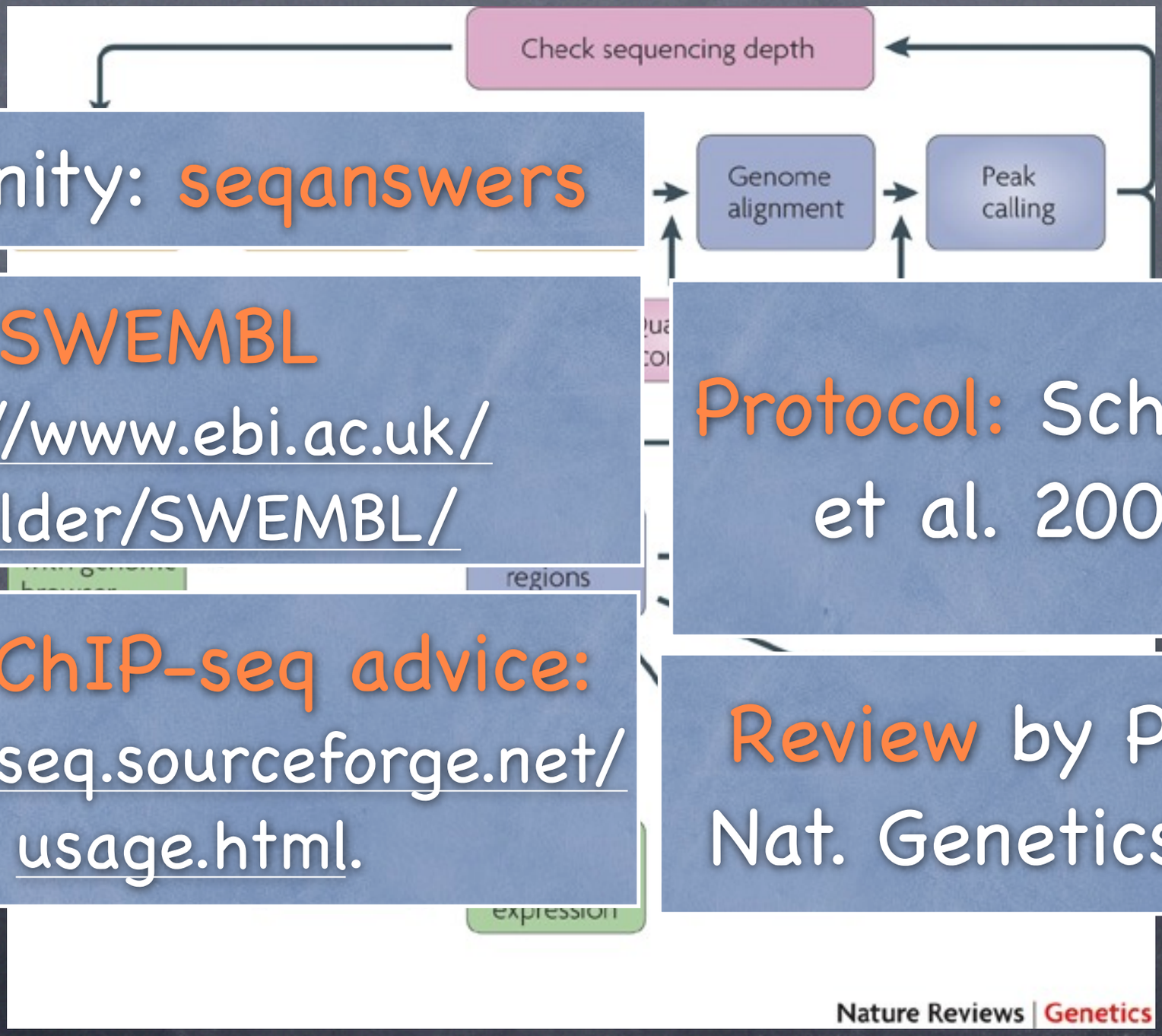
Community: seqanswers

SWEMBL
http://www.ebi.ac.uk/~swilder/SWEMBL/

Protocol: Schmidt et al. 2009

Useq ChIP-seq advice:
http://useq.sourceforge.net/usage.html.

Review by Park Nat. Genetics 09



Acknowledgments

- Steven Wilder,
  David Thybert,
  Benoit Ballester,
  Paul Flicek
  (EMBL-EBI, Hinxton, UK)

- Dominic Schmidt,
  Duncan Odom
  (CRI Cambridge, UK)