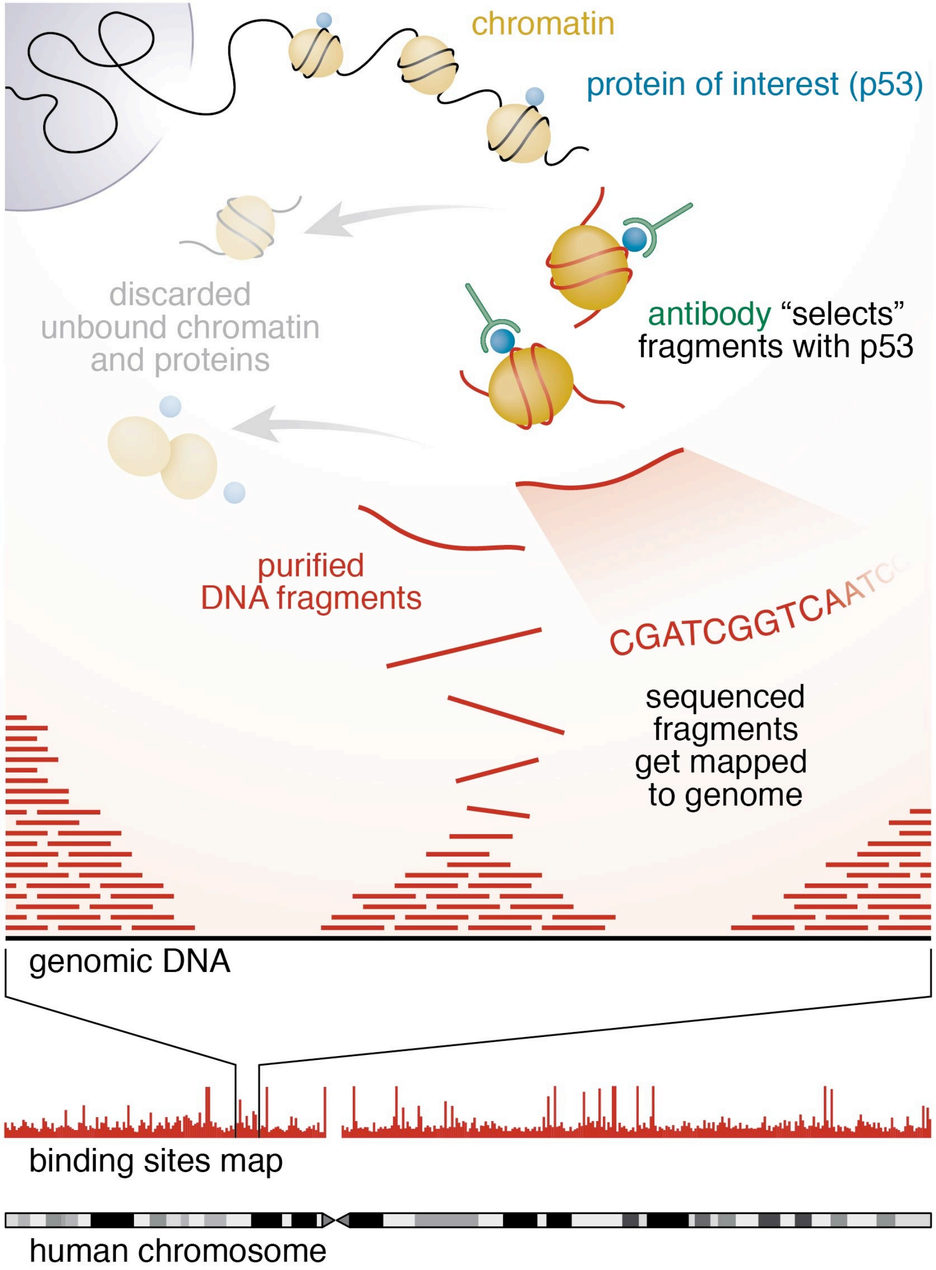
PLATEFORME
BIOPUCES & SÉQUENÇAGE
INSTITUT DE GÉNÉTIQUE ET DE BIOLOGIE MOLÉCULAIRE ET CELLULAIRE

ChIPseq: library preparation and data analysis

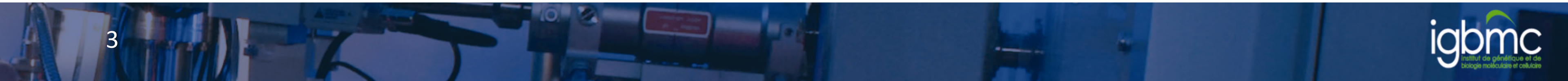
Stéphanie Le Gras

igbmc
institut de génétique et de biologie moléculaire et cellulaire

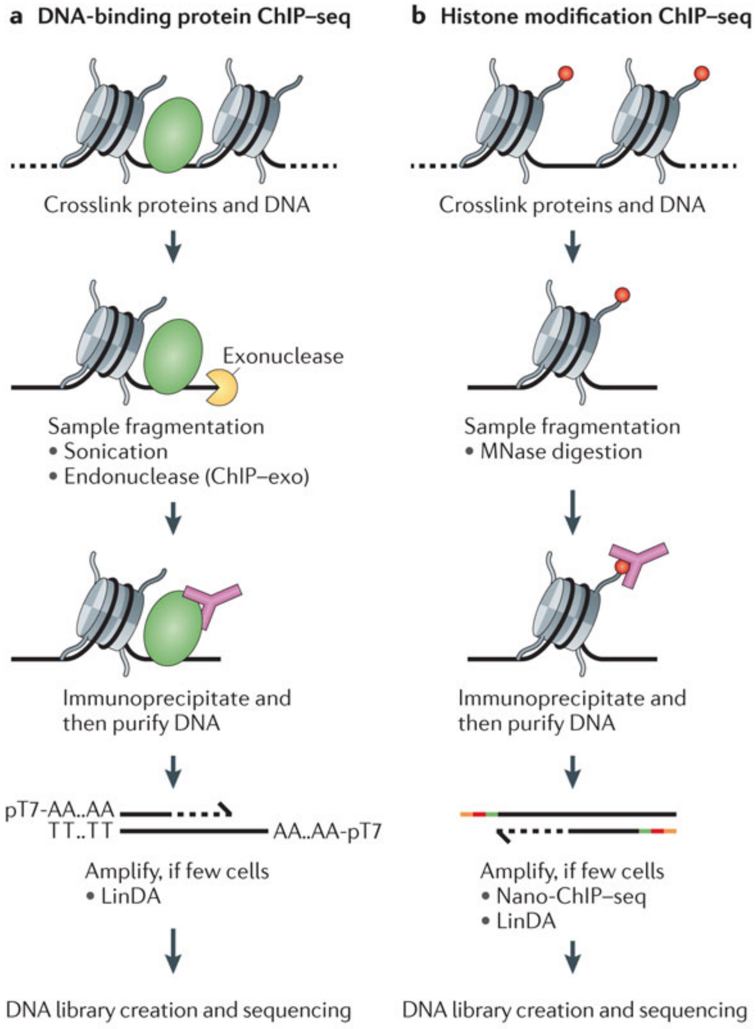
ChIP-seq


chromatin
protein of interest (p53)
discarded unbound chromatin and proteins
antibody "selects" fragments with p53
purified DNA fragments
CGATCGGTCAATCC
sequenced fragments get mapped to genome
genomic DNA
binding sites map
human chromosome

CHIP AND LIBRARY PREP CONSIDERATIONS



Chromatin ImmunoPrecipitation


a DNA-binding protein ChIP–seq
b Histone modification ChIP–seq
Crosslink proteins and DNA
Crosslink proteins and DNA
Exonuclease
Sample fragmentation
• Sonication
• Endonuclease (ChIP–exo)
Sample fragmentation
• MNase digestion
Immunoprecipitate and then purify DNA
Immunoprecipitate and then purify DNA
pT7-AA..AA
TT..TT
AA..AA-pT7
Amplify, if few cells
• LinDA
Amplify, if few cells
• Nano-ChIP–seq
• LinDA
DNA library creation and sequencing
DNA library creation and sequencing

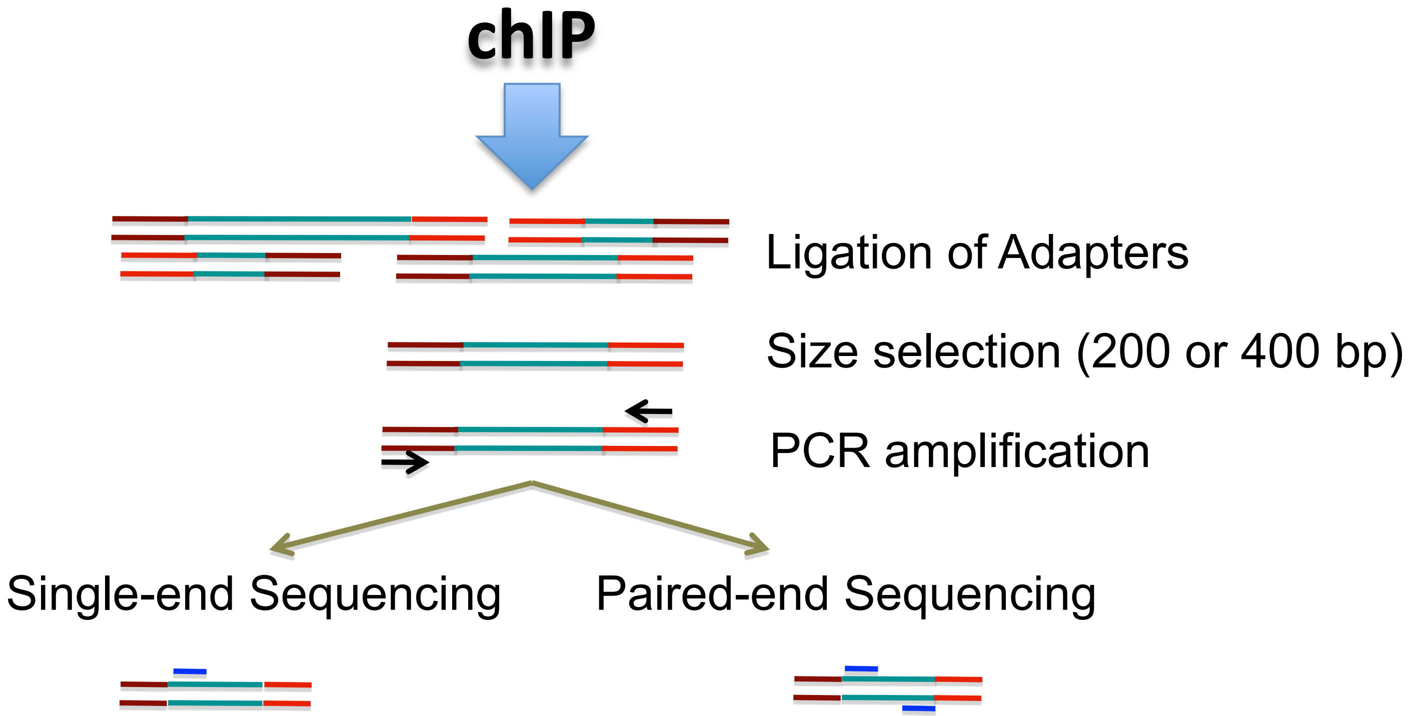
Nature Reviews Genetics 13, 840-852 (December 2012) doi:10.1038/nrg3306

Considerations on chIP

- Antibody
  - Antibody quality varies, even between independently prepared lots of the same antibody (Egelhofer, T. A. *et al*. 2011)
  - Multiple histone modifications can alter the efficacy of certain antibodies (Fuchs, S. M. et al, 2011)
- Number of cells
  - large numbers of cells (~10 million) are required for a ChIP experiment (limitation for small organisms)
    - Nano-ChIP–seq (Adli et al, 2011)
    - LinDA (Shankaranarayanan et al, 2011)
- Shearing of DNA (Mnase I, sonication, Covaris) : trying to narrow down the size distribution of DNA fragments
- **Complexity in DNA fragments**

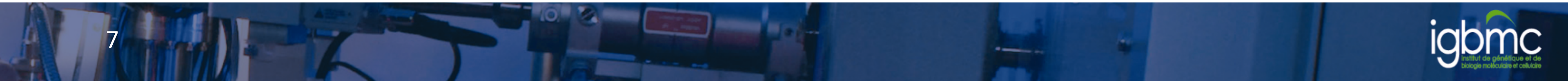
Library prep

- Step between chIP and sequencing
- The goal is to prepare DNA for the sequencing
- 5-10 ng of sheared DNA



Library prep

- PCR amplification : to increase amount of starting DNA.
  - Number of PCR as low as possible
  - PCR free protocols.



Sequencing

- Sequencer : Illumina HiSeq 2500
- No. of reads per run, per sample :
  - 1st run on the GAIIx : 10-20 millions of reads per lane
  - (HiSeq 2500) 4 samples per lane :~50-70 millions per sample
- Length of DNA fragment : ~200bp
- No. of cycle per run : 50

Single end or paired end?

- Single end (most of the time)
- Paired-end sequencing
  - Improve identification of duplicated reads
  - Better estimation of the fragment size distribution
  - Increases the efficiency of mapping to **repeat regions**
  - The price!

Sequencing depth

- Consider the depth needed
- For human genomes, 20 million uniquely mapped read sequences are suggested for point-source peaks, or 40 million for broad-source peaks.
- For fly genome: 8 million reads
- For worm genome: 10 million reads

Controls

- Used mostly to filter out false positives (high level noise)
  - Idea: potential false positive will be enriched in both treatment and control.
- 3 types of control are commonly used :
  - Input DNA : a portion of DNA sample removed prior to IP
  - DNA from non specific IP : DNA obtained from IP with an antibody not known to be involved in DNA binding or chromatin modification such as IgG
  - Mock IP DNA : DNA obtained from IP without antibodies

Controls

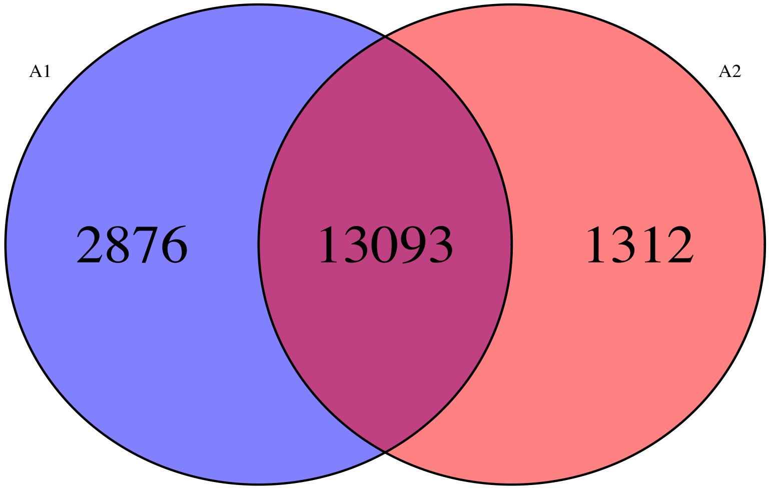
- Using Input DNA as a control corrects for biases due to :
  - Variable solubility of different regions
  - Shearing of DNA
  - Amplification
- Choice of control is extremely important
- A control will fail to filter out false positives if its enrichment profile is very different from the enrichment profile of false positive regions in the treatment sample

Replicates

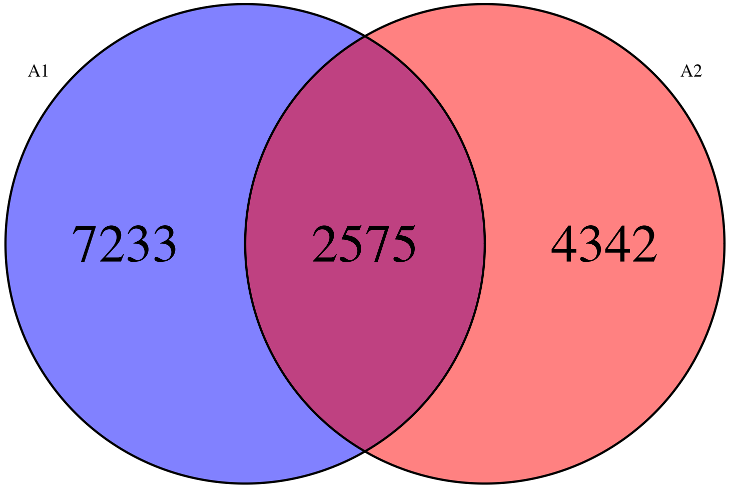
- A minimum of two replicates should be carried out per experiment.
- Each replicate should be a biological rather than a technical replicate; that is, it represents an independent cell culture, embryo pool or tissue sample.
- For two replicates, either 80% of the top 40% of identified targets in one replicate must be among the targets in the second replicate; alternatively, 75% of target lists must be in common between both replicates.

Biological replicates

- H3K9ac (sharp peak)

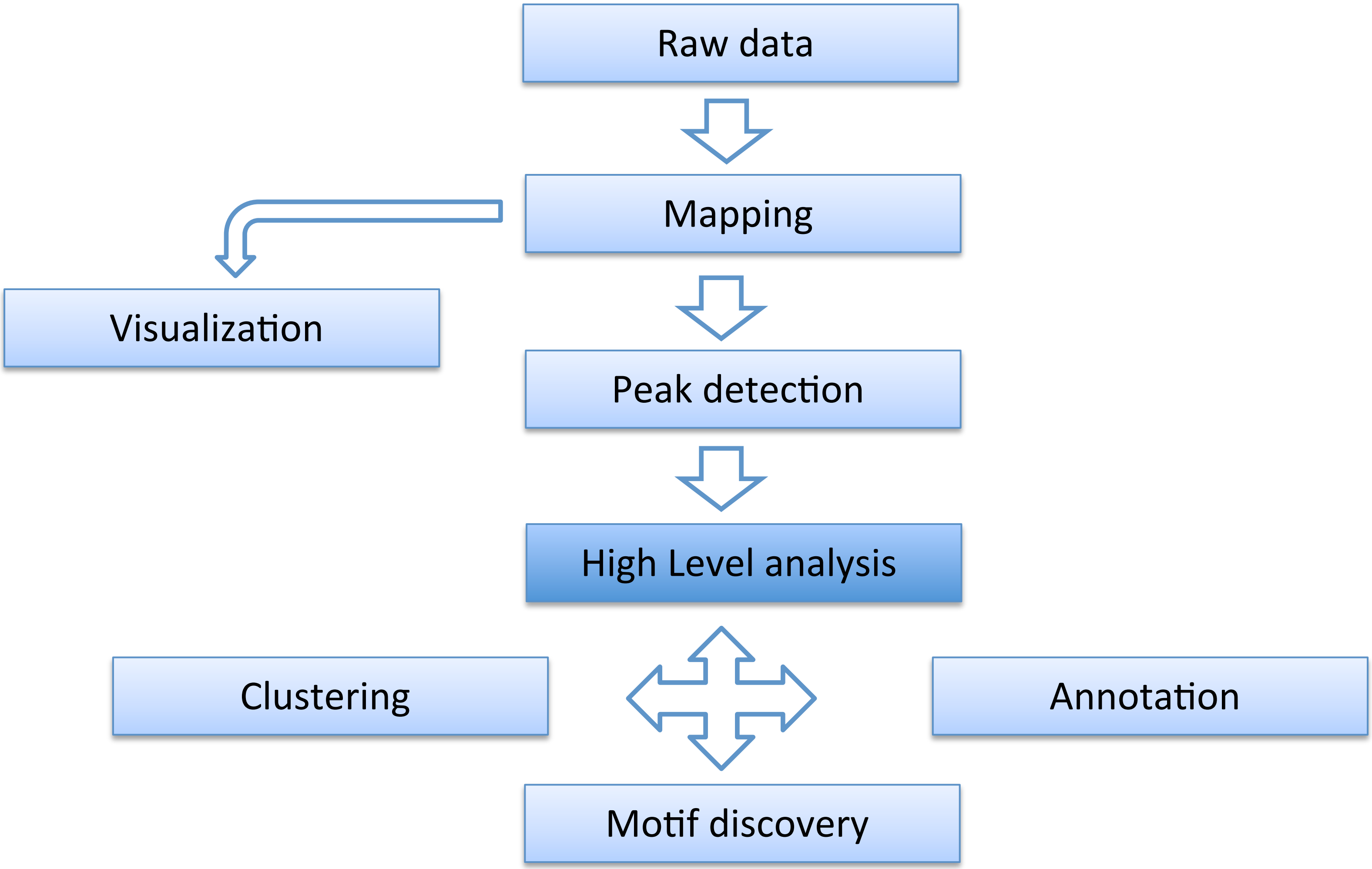


- H4K16ac (broad enrichment)



DATA ANALYSIS

Guidelines



Raw data

Mapping

Visualization

Peak detection

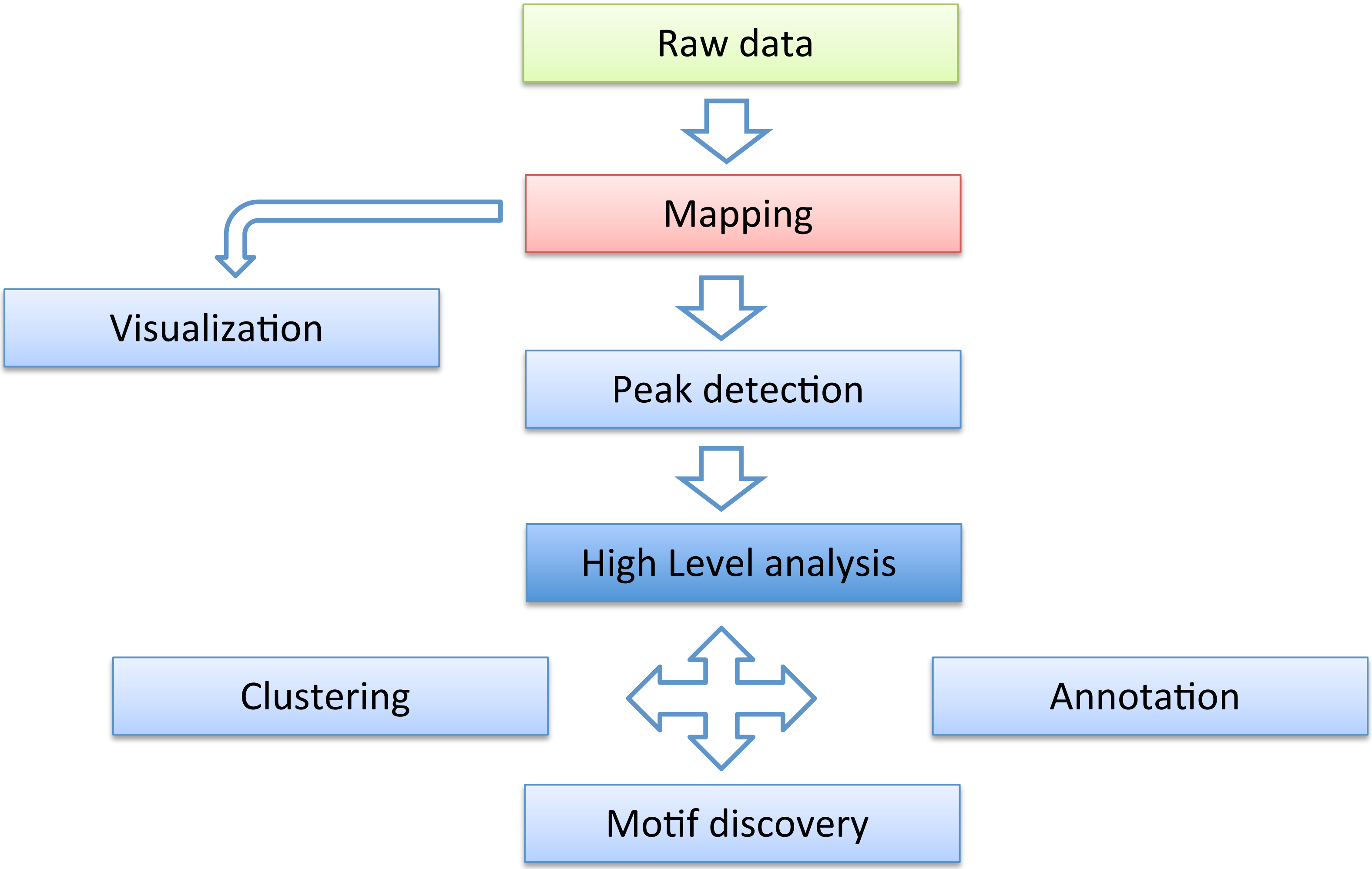
High Level analysis

Clustering

Annotation

Motif discovery

Guidelines



Raw data

Mapping

Visualization

Peak detection

High Level analysis

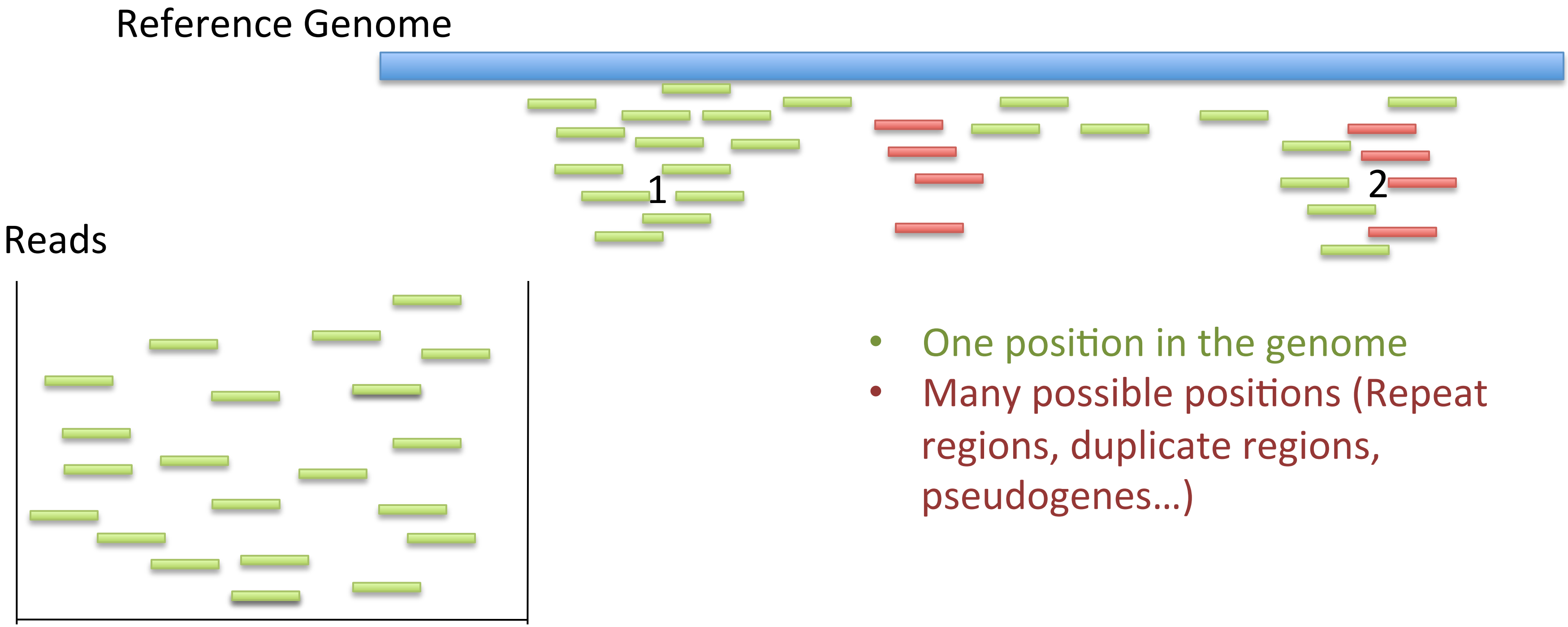
Clustering

Annotation

Motif discovery

Mapping

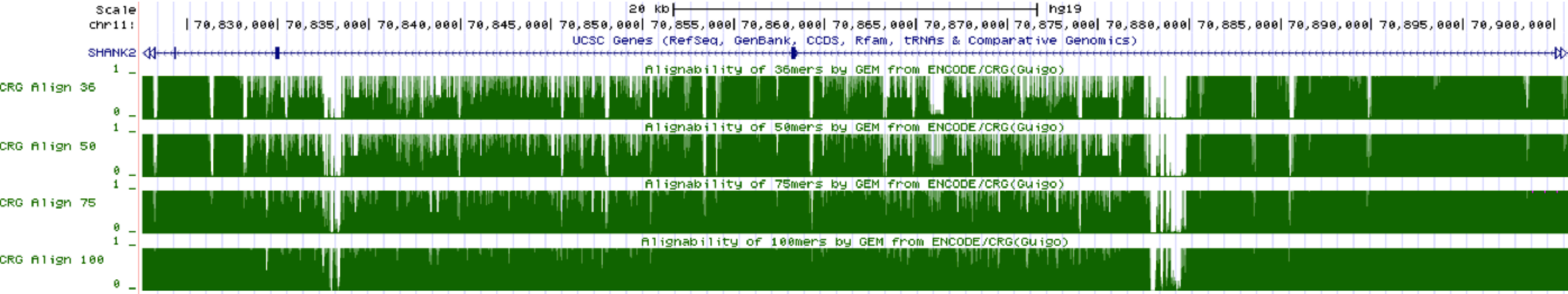
- Find out the position of the reads within the genome



- One position in the genome
- Many possible positions (Repeat regions, duplicate regions, pseudogenes…)

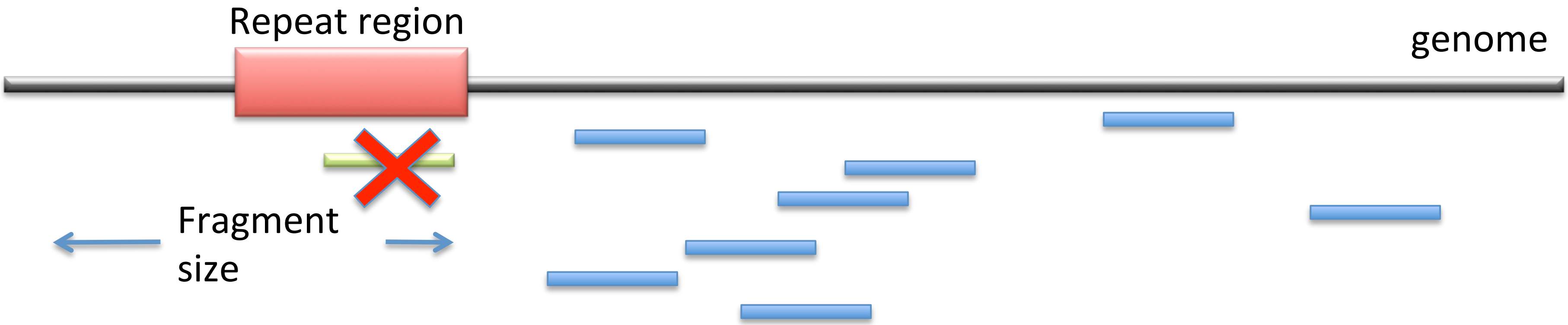
Tricky regions

- Low complexity regions (homopolymers)
- Repeat regions (pseudogenes, …)
  - – Mappability
    - depend on the read length
    - Best if paired end reads


Scale
20 kb
hg19
chr11:
70,830,000
70,835,000
70,840,000
70,845,000
70,850,000
70,855,000
70,860,000
70,865,000
70,870,000
70,875,000
70,880,000
70,885,000
70,890,000
70,895,000
70,900,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
SHANK2
1
Alignability of 36mers by GEM from ENCODE/CRG(Guigo)
CRG Align 36
0
1
Alignability of 50mers by GEM from ENCODE/CRG(Guigo)
CRG Align 50
0
1
Alignability of 75mers by GEM from ENCODE/CRG(Guigo)
CRG Align 75
0
1
Alignability of 100mers by GEM from ENCODE/CRG(Guigo)
CRG Align 100
0

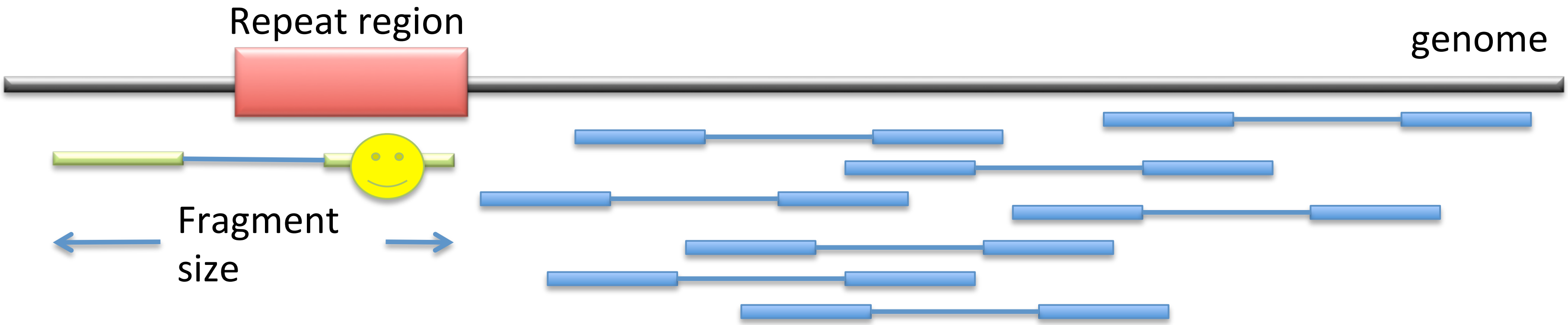
Repeat regions

- Usually not kept for downstream analysis (source of bias)

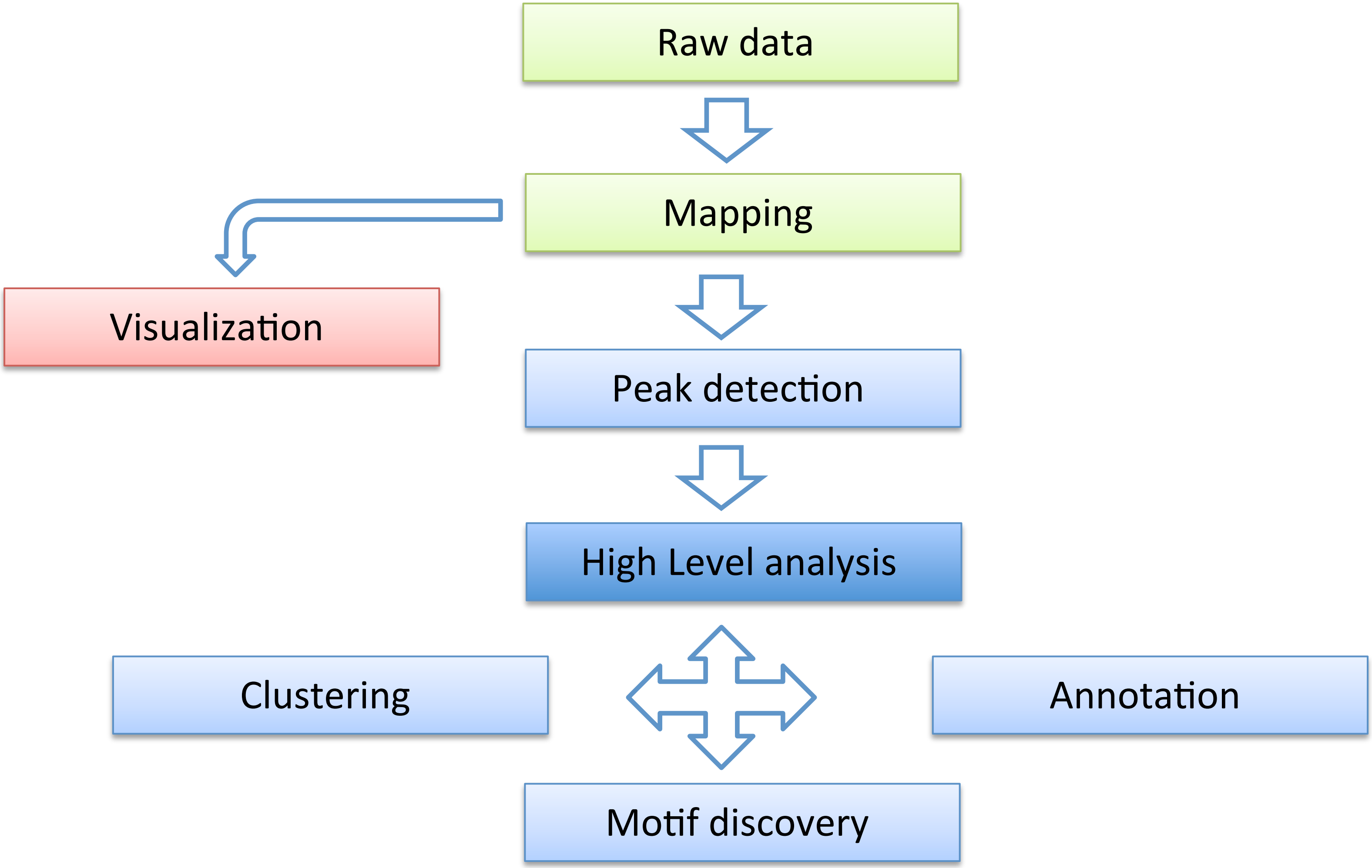


Repeat regions

- Usually not kept for downstream analysis (source of bias)



Guidelines



Raw data

Mapping

Visualization

Peak detection

High Level analysis

Clustering

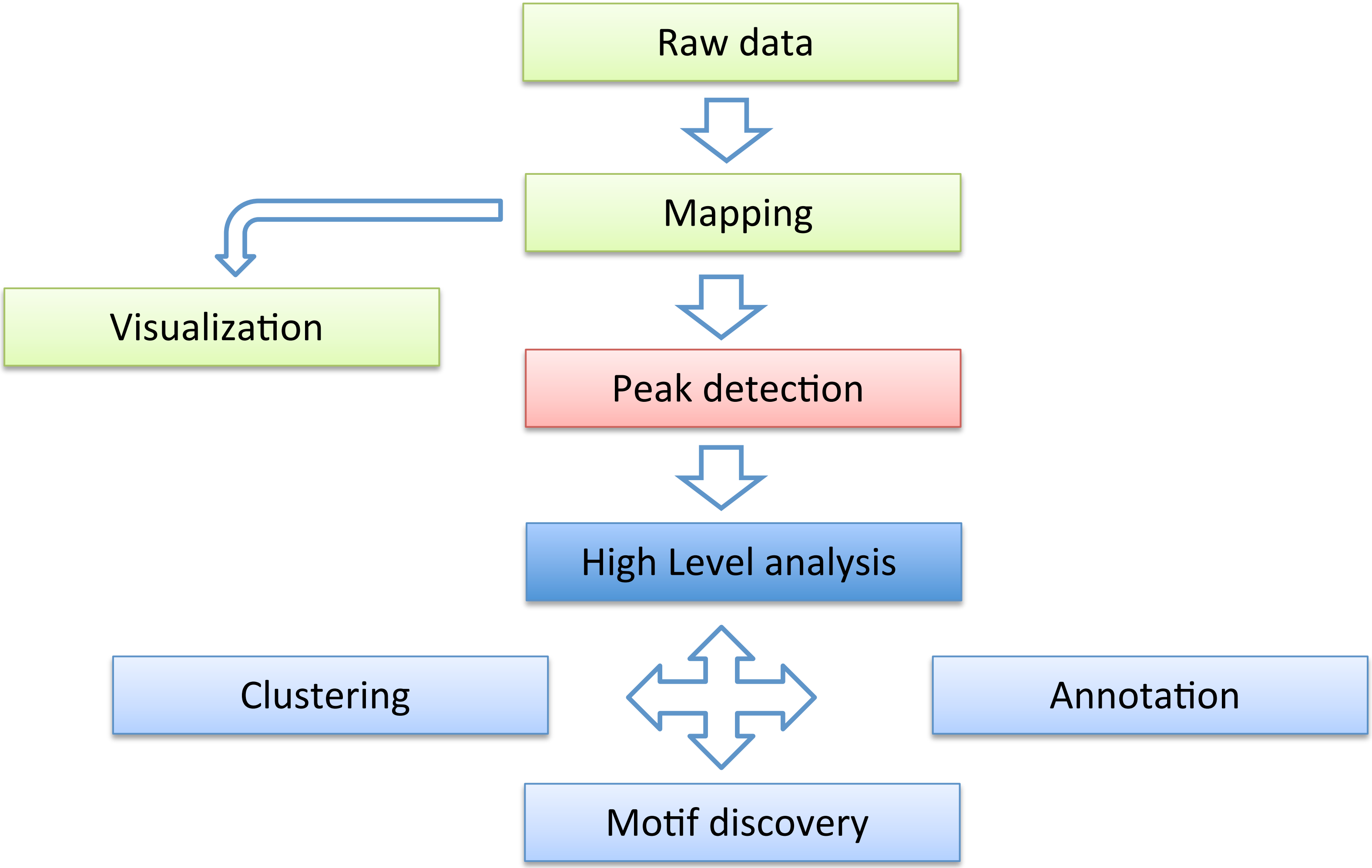
Annotation

Motif discovery

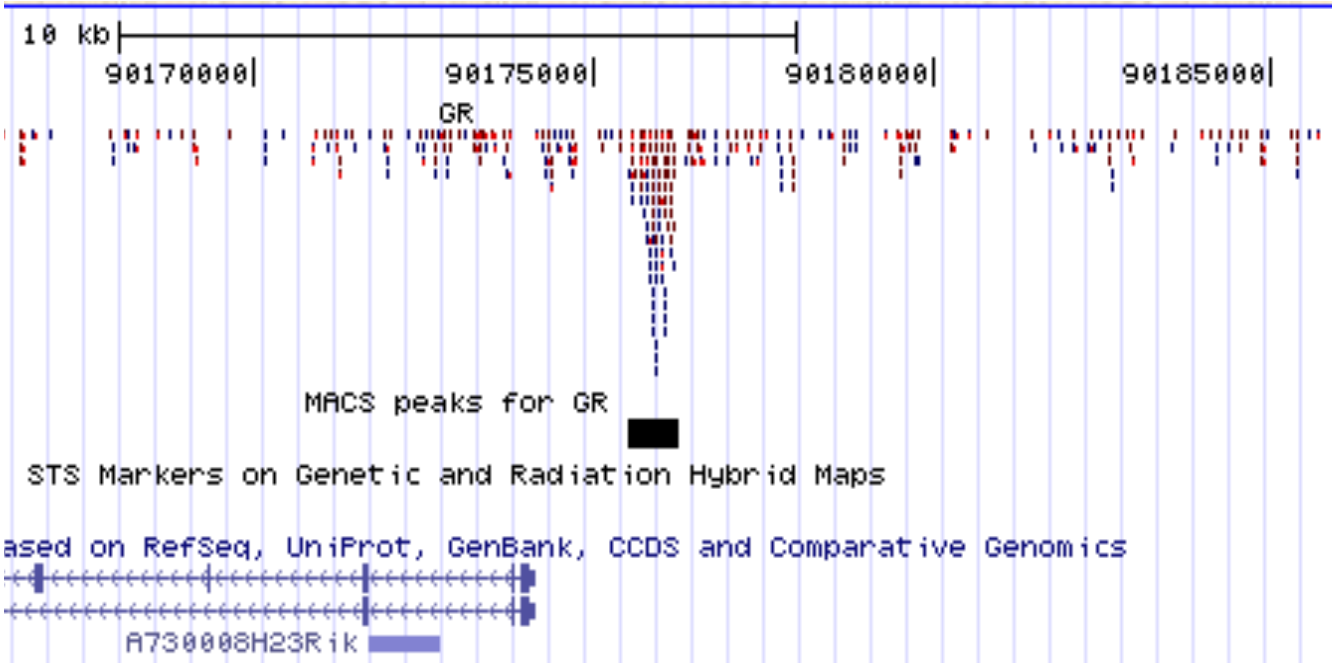
Visualisation

- http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=slegras&hgS_otherUserSessionName=Mitf%20data

Guidelines


Raw data
Mapping
Visualization
Peak detection
High Level analysis
Clustering
Annotation
Motif discovery

Peaks

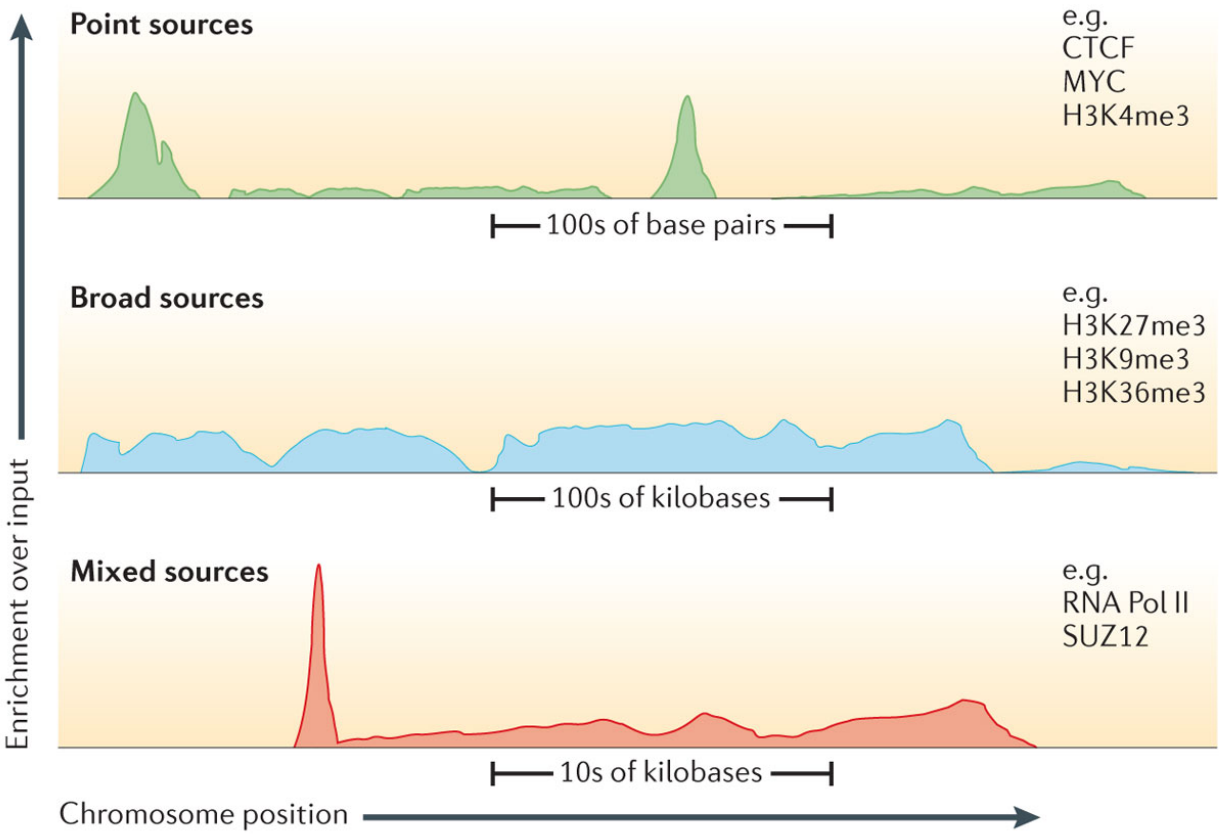


10 kb
90170000|
90175000|
90180000|
90185000|
GR
MACS peaks for GR
STS Markers on Genetic and Radiation Hybrid Maps
ased on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
A730008H23Rik

Majority (60-90%?) are `background' (Pepke et al., 2009)

Not as bad as it sounds { 40% of reads distributed over 99.9% of the genome, vs 60% over 0.1%.

Binding profiles

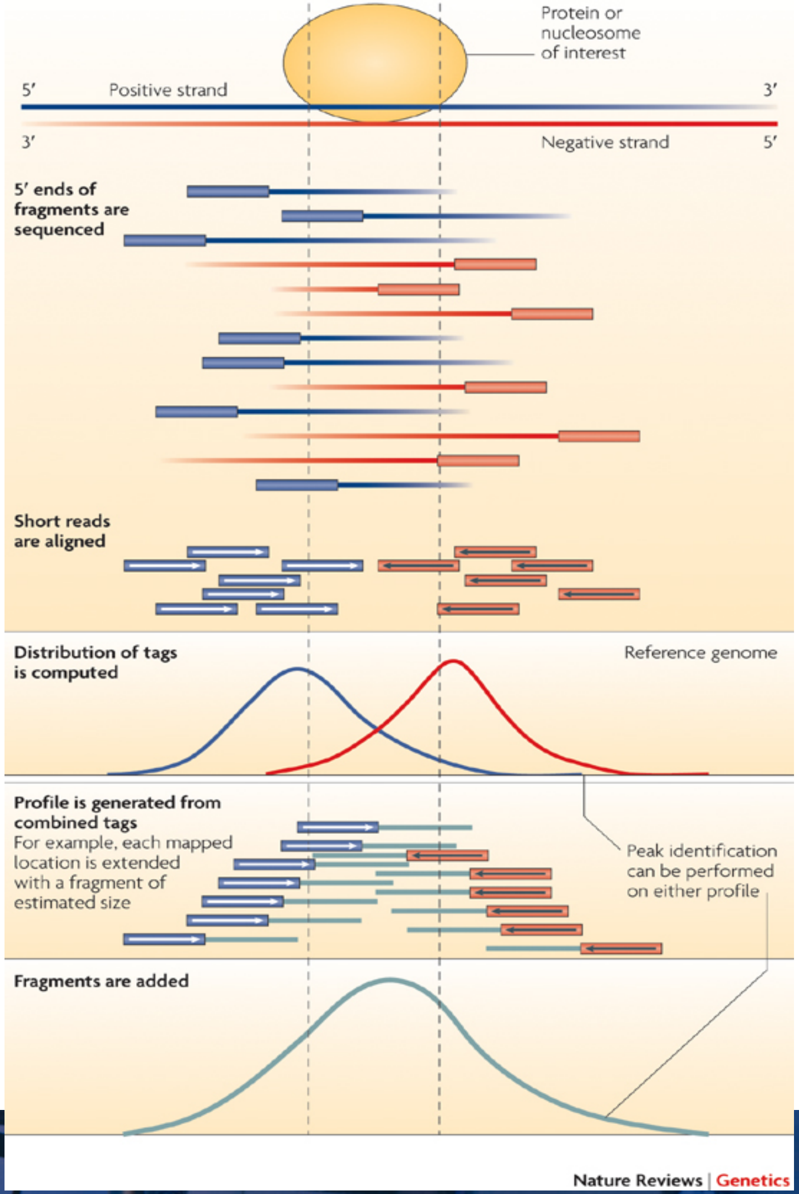


Nature Reviews | Genetics

Peak detection

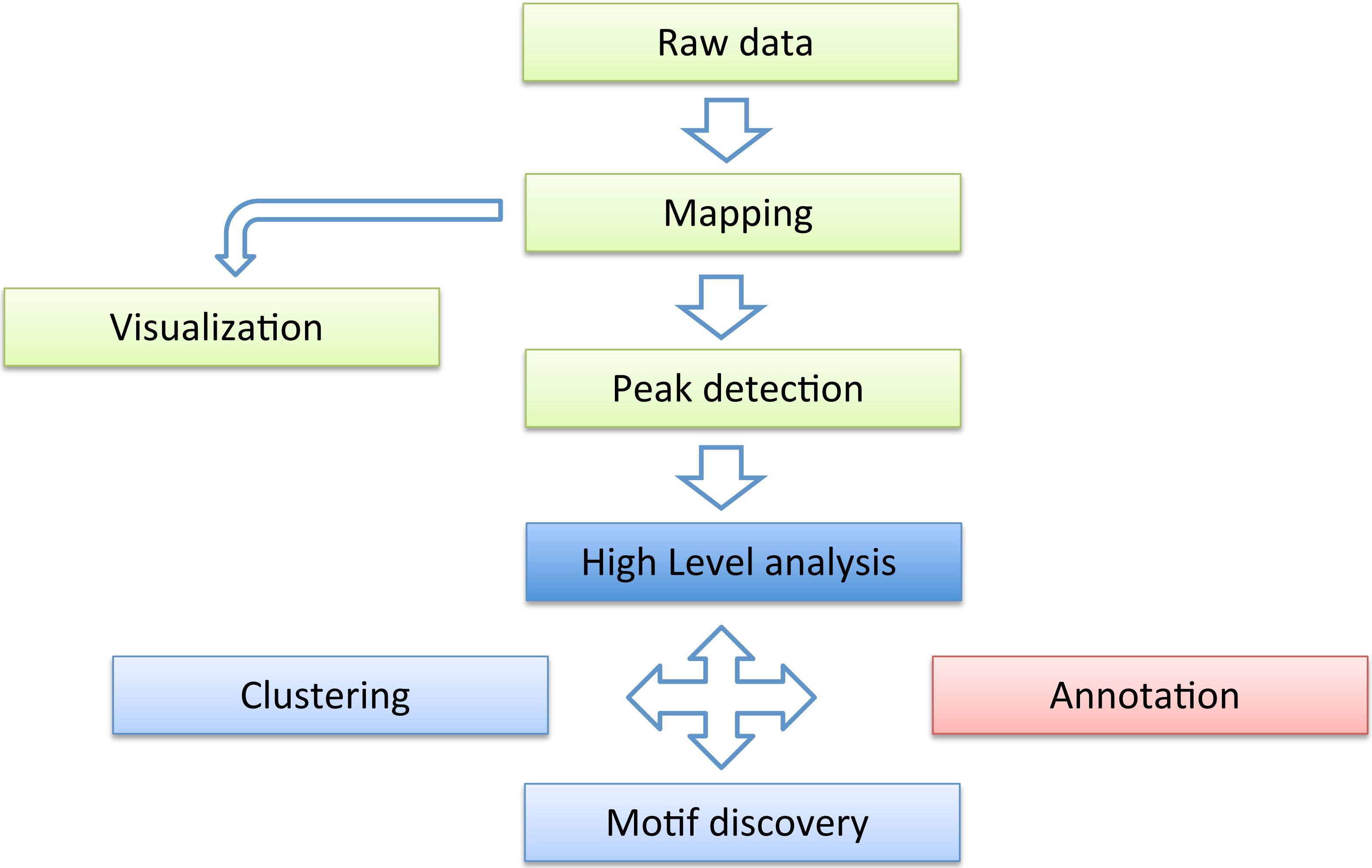
- Discover interaction sites from aligned reads
- Idea: loci with lots of reads/fragments = signal site
- Loci with lots of reads could also be due to
  - Sequencing biases
  - Chromatin biases
  - PCR biases/artefacts
  - Biases/artefacts of unknown origin
  - So need to separate signal from noise
- Need to use an input to correct for the biases (Expect that the biaises are similar in input and in IP)

Peak detection


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Nature Reviews | Genetics

Guidelines


Raw data
Mapping
Visualization
Peak detection
High Level analysis
Clustering
Annotation
Motif discovery

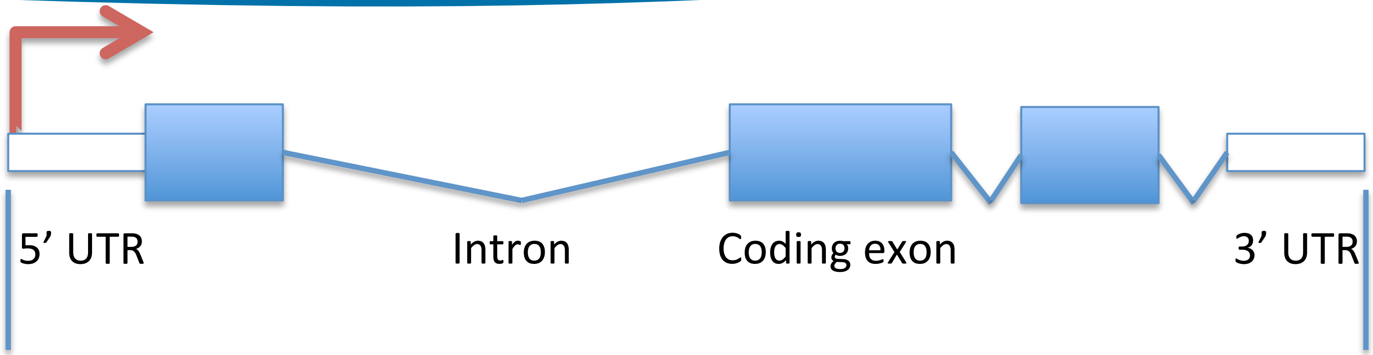
Peak annotation

- Goal: assigning a peak to one ore many genome features
- Always be careful on the database used to annotate the peaks (either RefSeq or Ensembl)
- Many tools exist (GPAT, CEAS, CisGenome, Homer…)

Peak annotation (Homer)

- Default behaviour is to use RefSeq annotations
- Works in two parts:
  - Determines the distance to the nearest TSS and assigns the peak to that gene
  - Determines the genomic annotation of the region occupied by the center of the peak/region

Peak annotation (Homer)



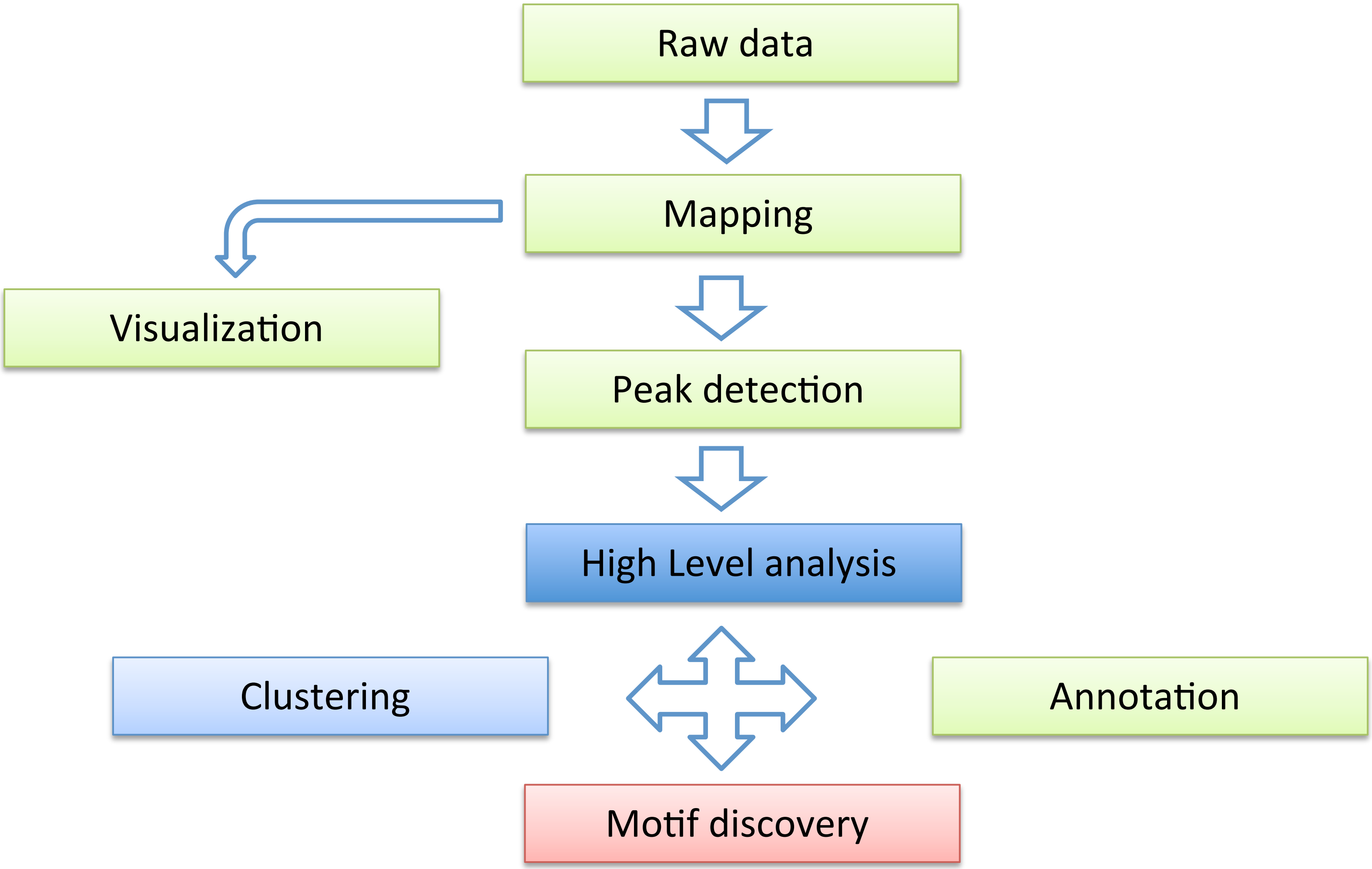
TSS (Transcription start site)

TTS (transcription termination site)

- Rank:

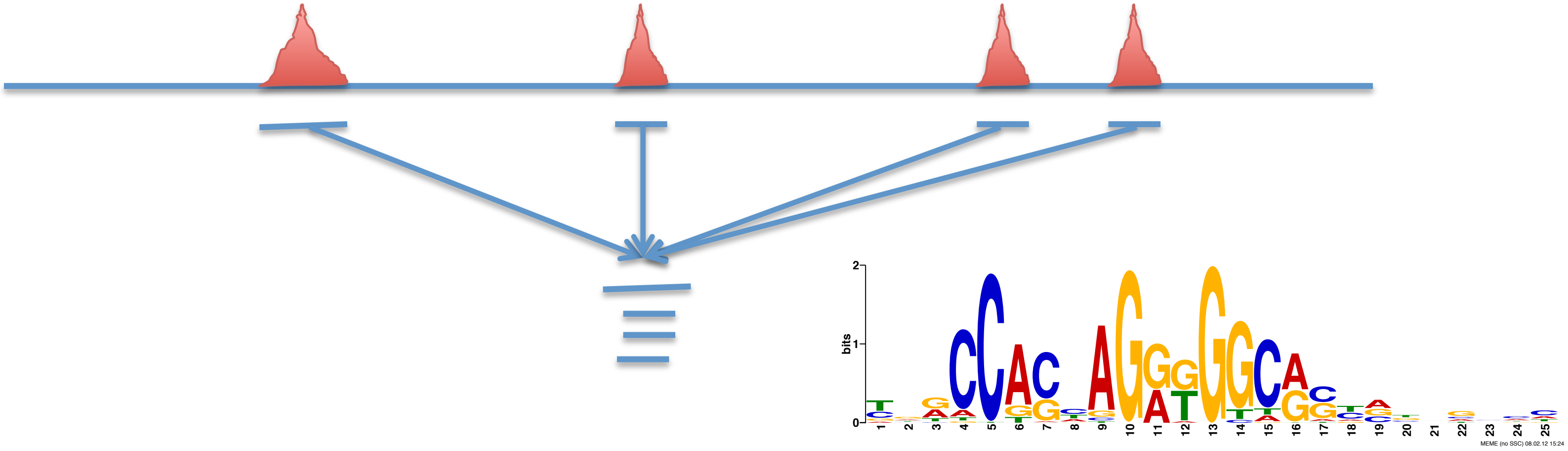
1. TSS (by default defined from -1kb to +100bp)
2. TTS (by default defined from -100 bp to +1kb)
3. CDS Exons
4. 5' UTR Exons
5. 3' UTR Exons
6. **CpG Islands
7. **Repeats
8. Introns
9. Intergenic

Guidelines


Raw data
Mapping
Visualization
Peak detection
High Level analysis
Clustering
Annotation
Motif discovery

Motif discovery

- Sequence to which the protein of interest may be bound
- Searching for enriched nucleotide sequence (i.e motif) within peak sequence.

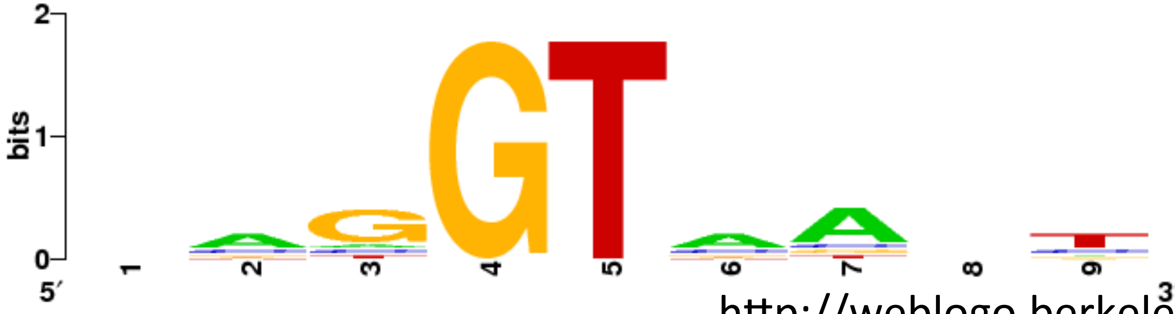


- De novo motif searching
- Motif searching based on motif databases (JASPAR)

PWM

- **position weight matrix (PWM)**, also known as a **position-specific weight matrix (PSWM)** or **position-specific scoring matrix (PSSM)**

$$M = \begin{array}{c} A \\ C \\ G \\ T \end{array} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

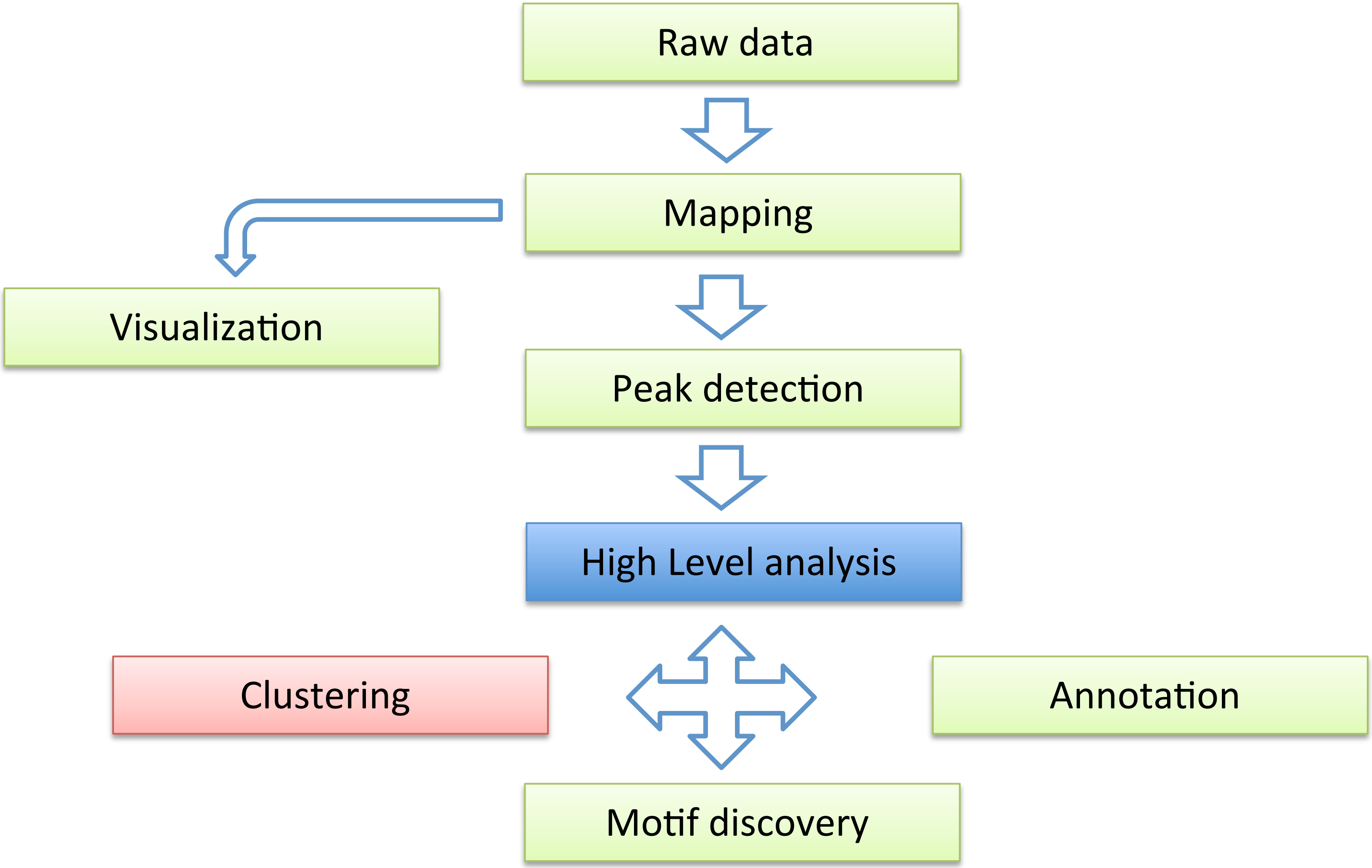


http://weblogo.berkeley.edu/logo.cgi

Known motif searching

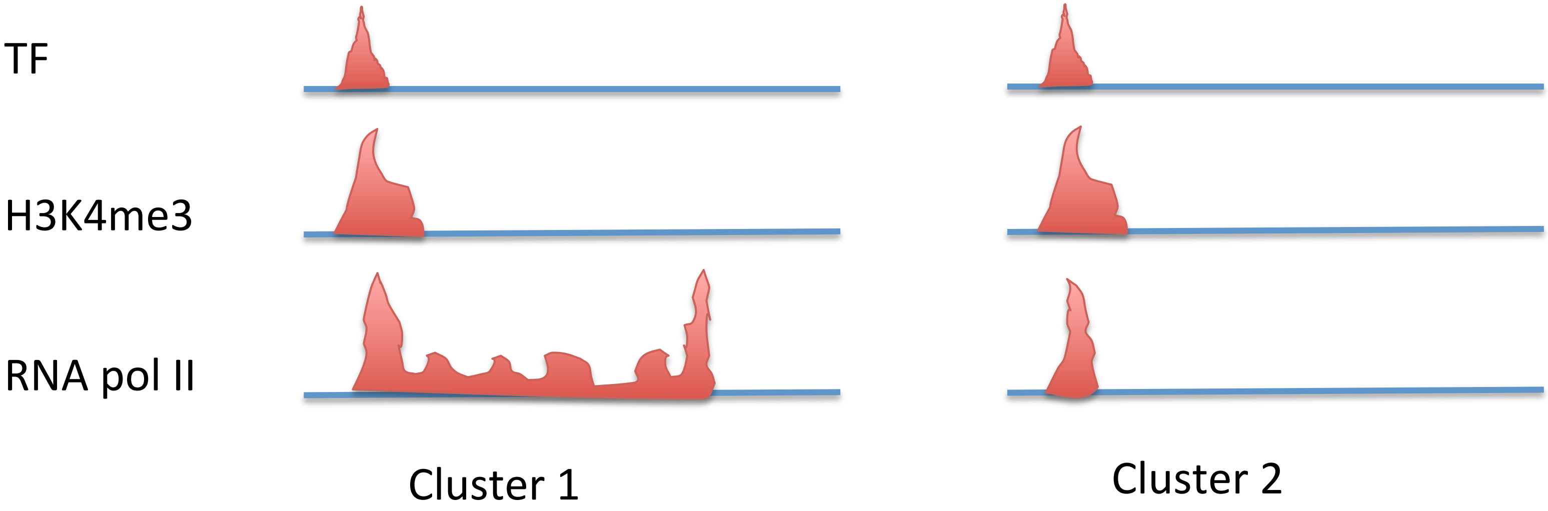
- Charles E. Grant, Timothy L. Bailey, and William Stafford Noble, "FIMO: Scanning for occurrences of a given motif", *Bioinformatics* 27(7):1017–1018, 2011
- Scan nucleotide sequences of interest for PWMs.
- JASPAR, Transfac databases

Guidelines

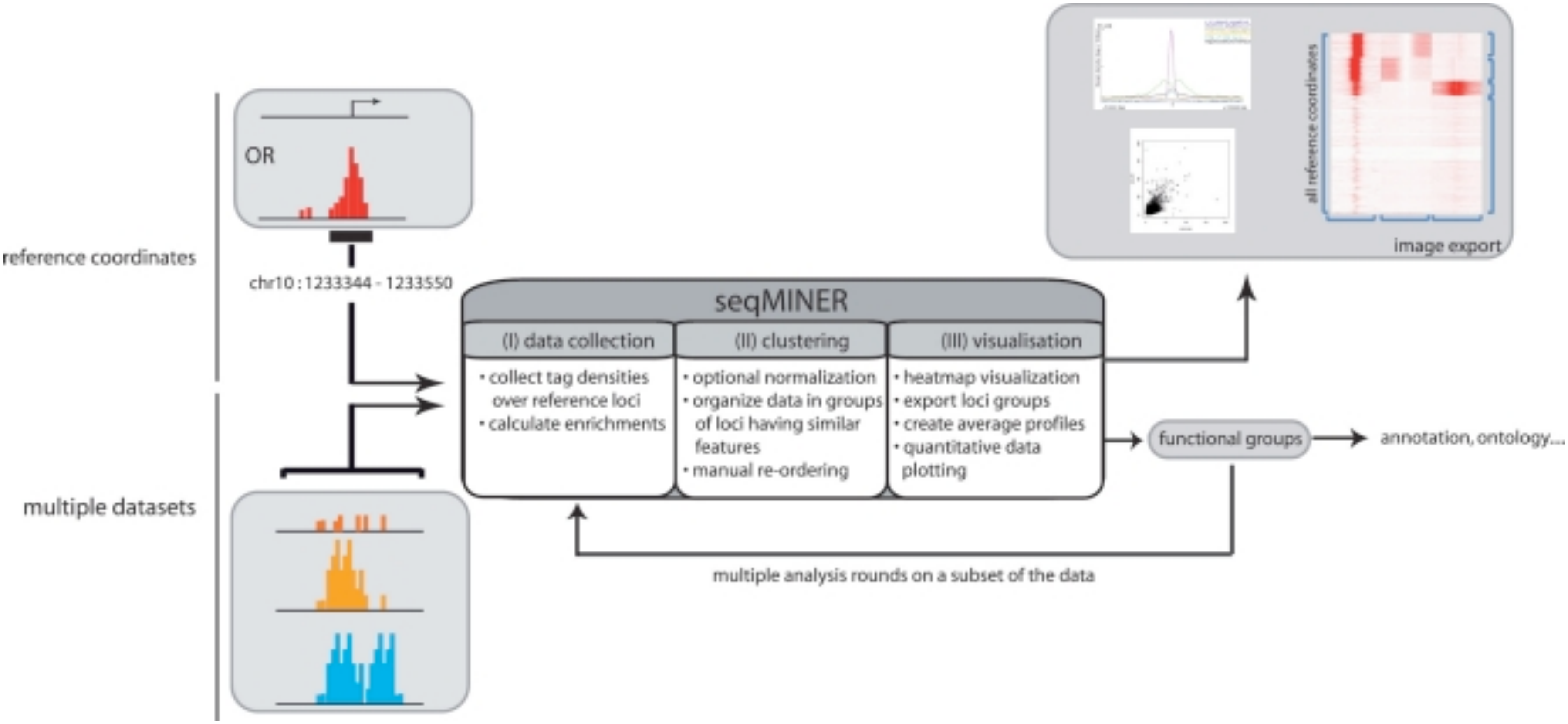

Raw data
Mapping
Visualization
Peak detection
High Level analysis
Clustering
Annotation
Motif discovery

Clustering

- Comparing several datasets all together at defined genomic loci
- Group together genomic regions with similar binding profiles

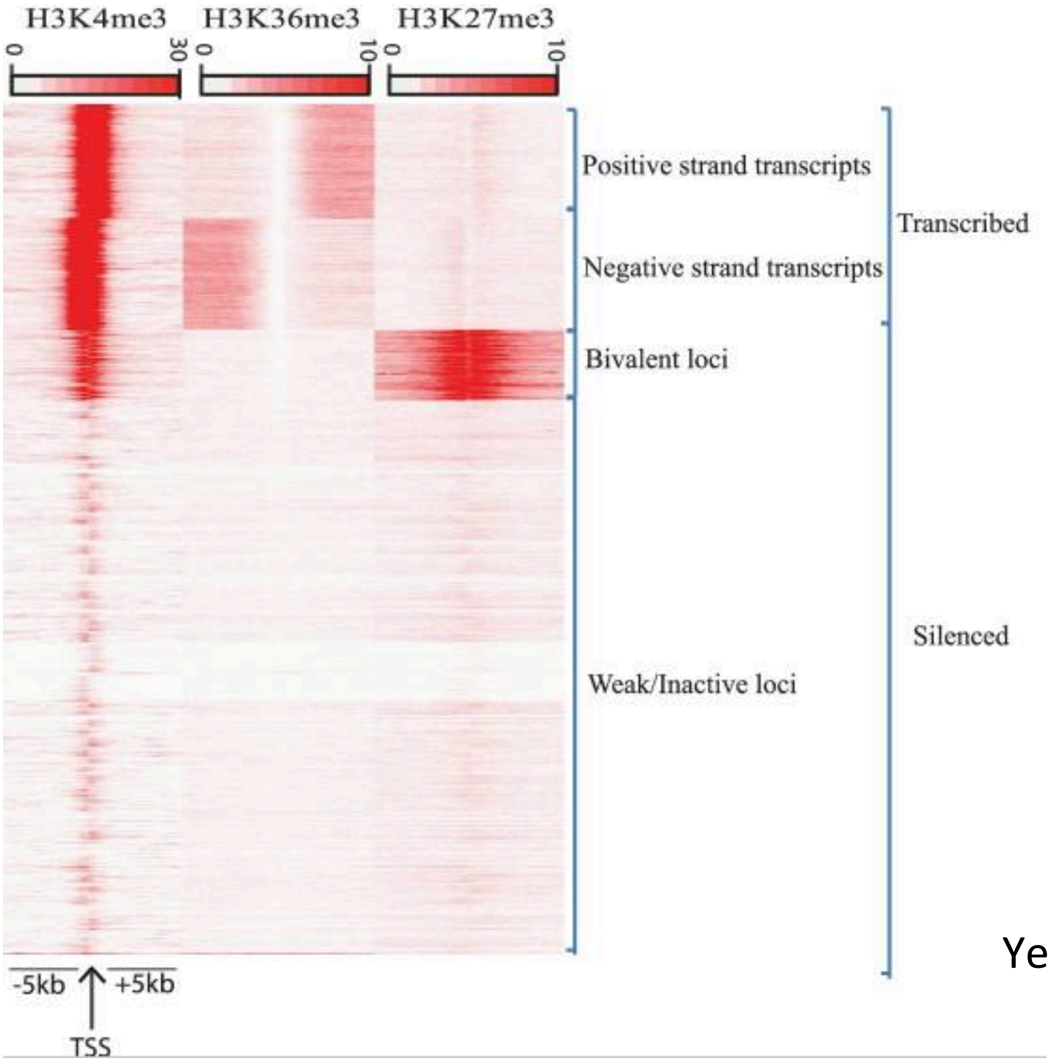


Clustering: SeqMINER



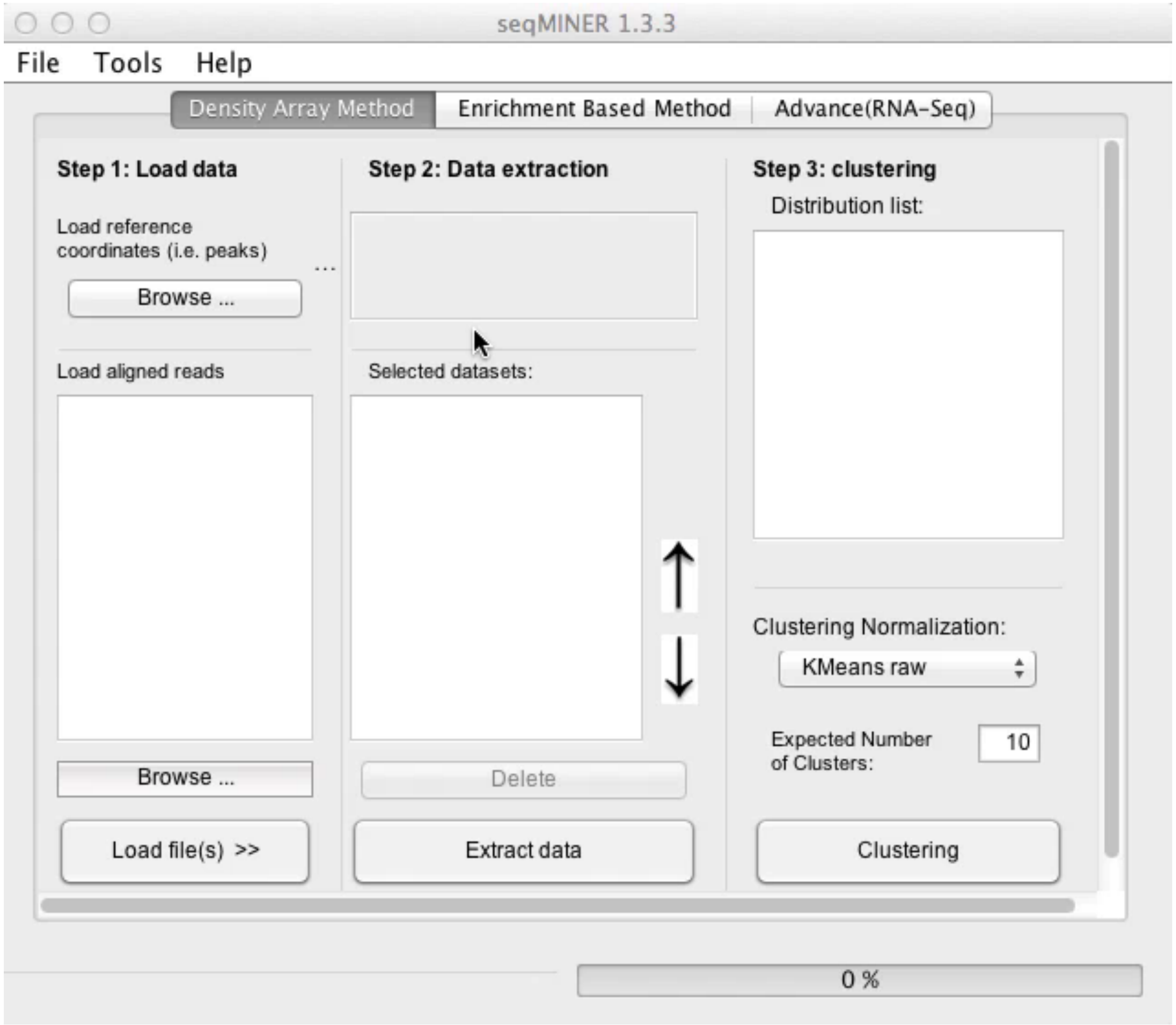
Ye et al, 2011, NAR.

Clustering: SeqMINER



Ye et al, 2011, NAR.

Example



seqMINER 1.3.3

File Tools Help

Density Array Method | Enrichment Based Method | Advance(RNA-Seq)

Step 1: Load data

Load reference coordinates (i.e. peaks)

Browse ...

Load aligned reads

Browse ...

Load file(s) >>

Step 2: Data extraction

Selected datasets:

Delete

Extract data

Step 3: clustering

Distribution list:

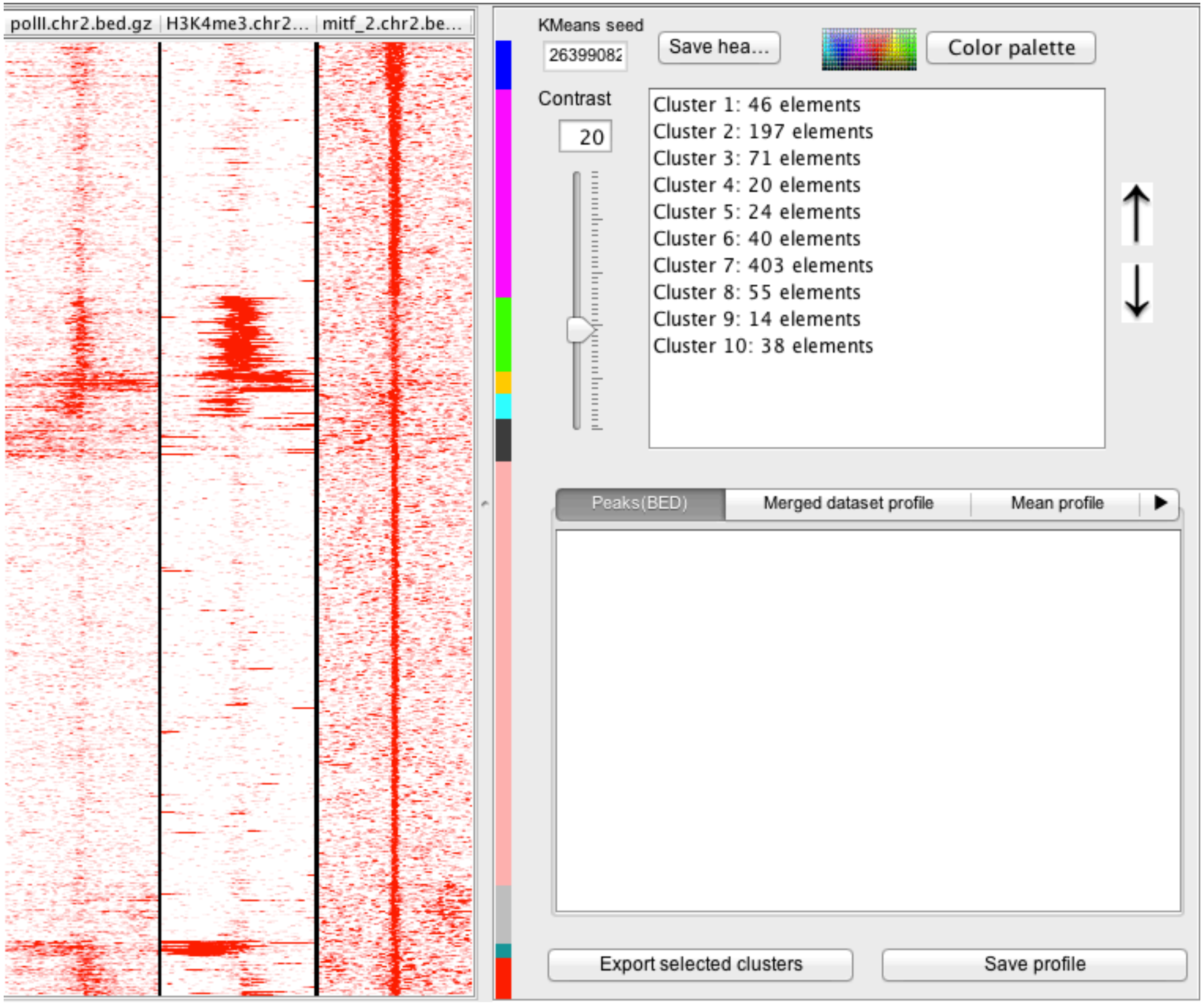
Clustering Normalization: KMeans raw

Expected Number of Clusters: 10

Clustering

0 %

Example



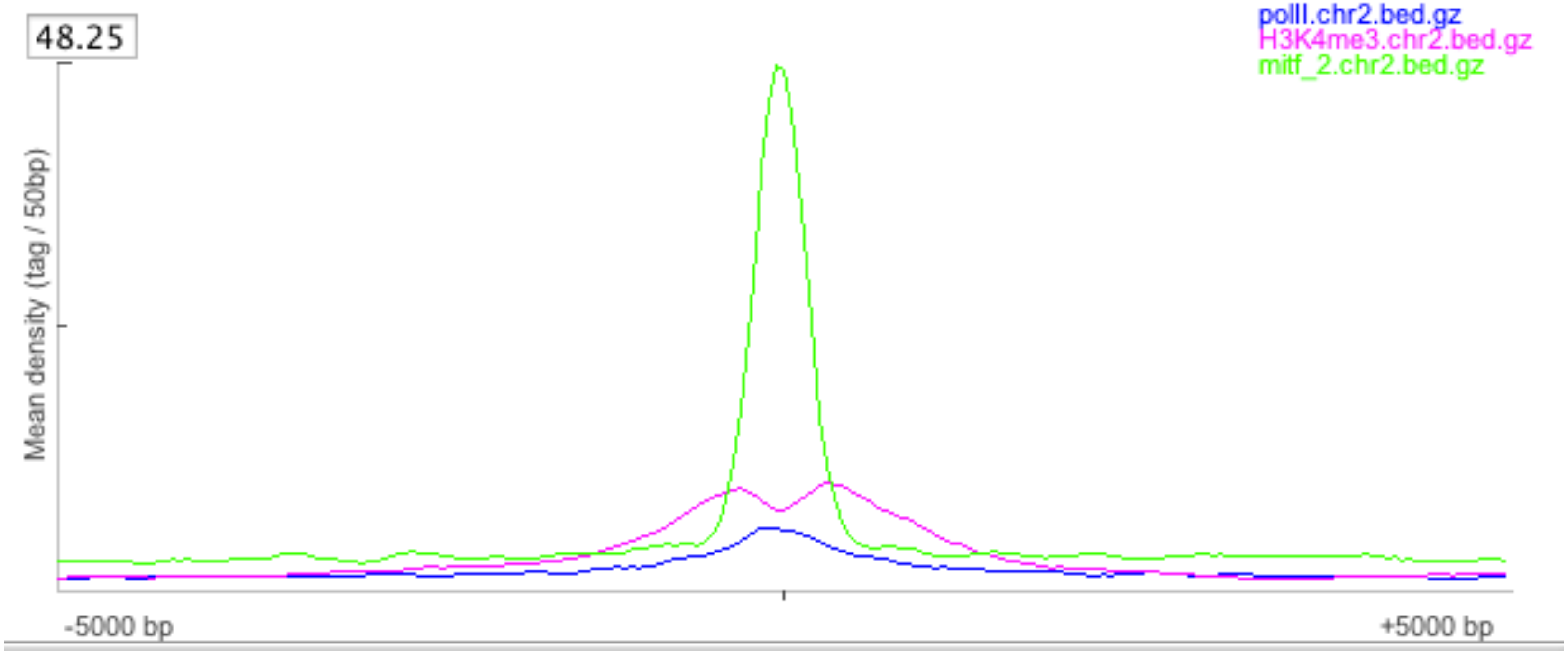
KMeans seed: 26399082

Contrast: 20

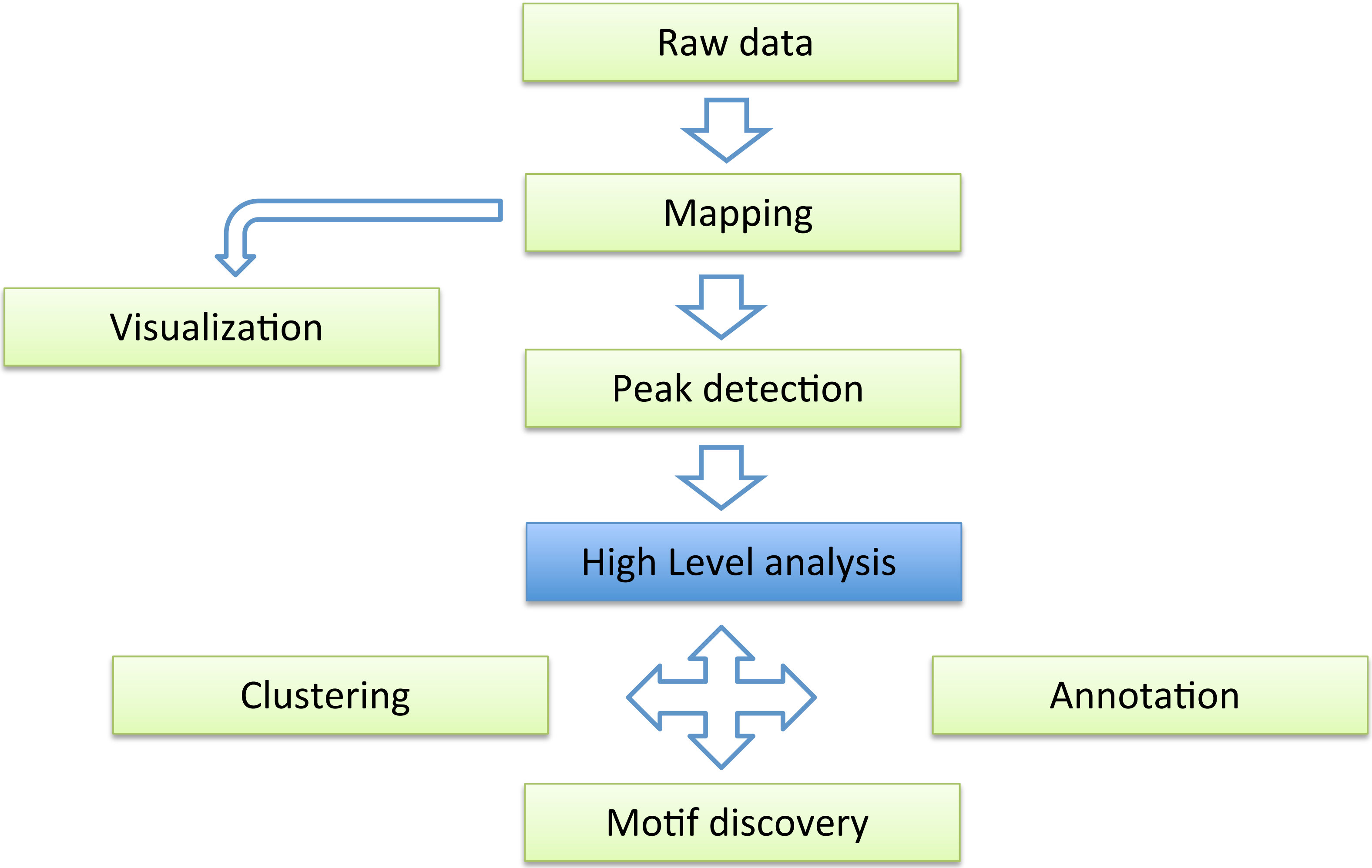
Cluster 1: 46 elements
Cluster 2: 197 elements
Cluster 3: 71 elements
Cluster 4: 20 elements
Cluster 5: 24 elements
Cluster 6: 40 elements
Cluster 7: 403 elements
Cluster 8: 55 elements
Cluster 9: 14 elements
Cluster 10: 38 elements

Example

Mean profile



Guidelines



Raw data

Mapping

Visualization

Peak detection

High Level analysis

Clustering

Annotation

Motif discovery