ChIP-Seq Data Analysis

Maciej Łapiński, IIMCB, Warsaw

NGSchool 2016

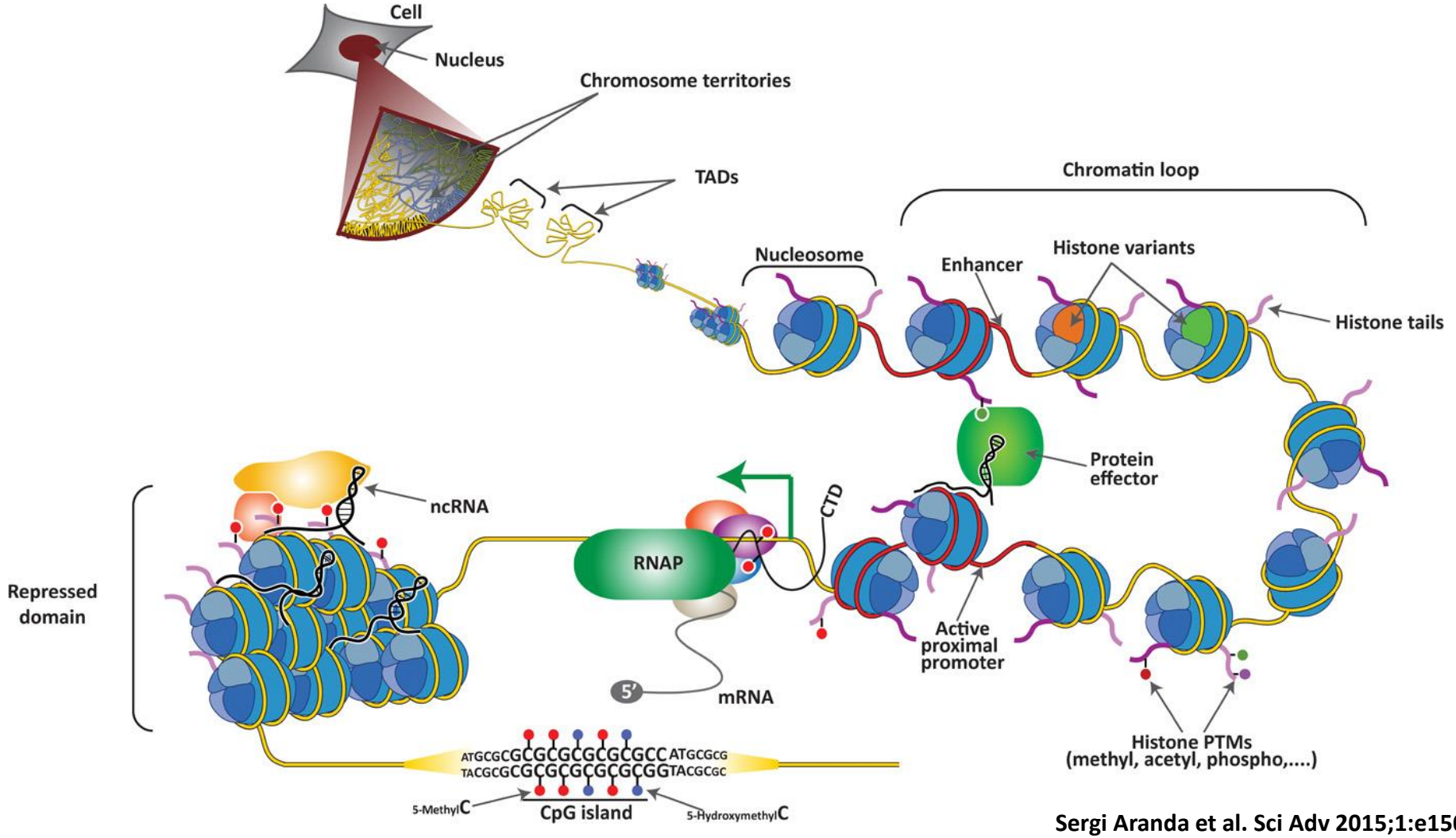
Dolný Smokovec, Slovakia

Walther Flemming
(21 April 1843 – 4 August 1905)

„Therefore, we will designate as chromatin that substance, in the nucleus, which upon treatment with dyes known as nuclear stains does absorb the dye.”

Zellsubstanz, Kern und Zelltheilung (1882)

Deciphering transcriptional regulatory network



Sergi Aranda et al. Sci Adv 2015;1:e1500737

Chromatin immunoprecipitation followed by massively paralel sequencing

Technology that can identify, in an unbiased manner, all DNA segments in the genome physically associated with a specific DNA-binding protein.

ChIP-Seq- Applications

Map the chromosomal locations of:

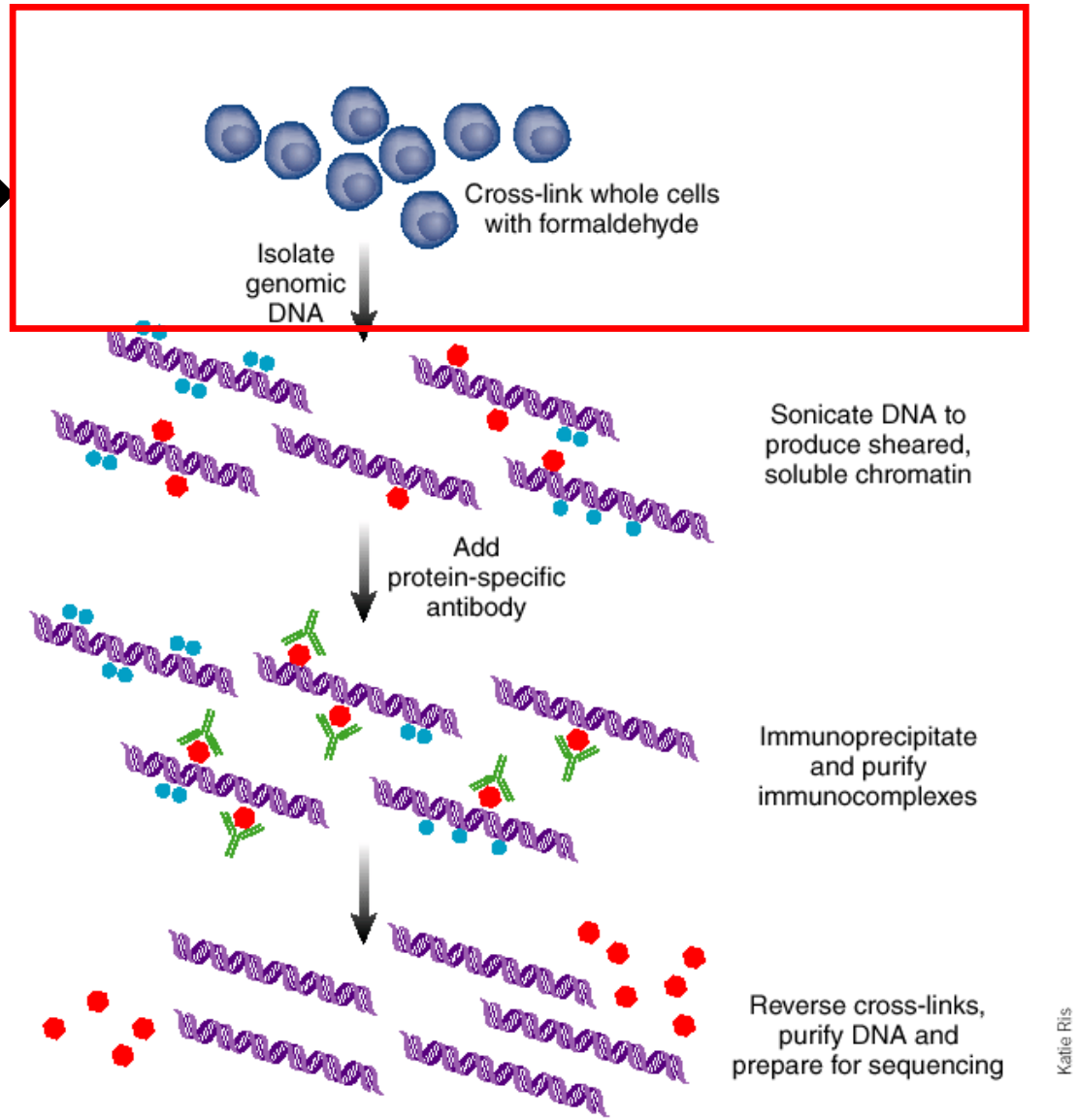
- transcription factors,
- nucleosomes,
- histone modifications,
- chromatin remodeling enzymes,
- chaperones,
- polymerases

ChIP-Seq- Applications

- Identification of precise regulatory sites across the genome
- Computation of recognized DNA sequence motifs
- Determination of downstream targets of transcription factors
- Clustering of multiple regulatory proteins at specific genomic positions
- Determination of chromatin states and DNA modifications

Methodology

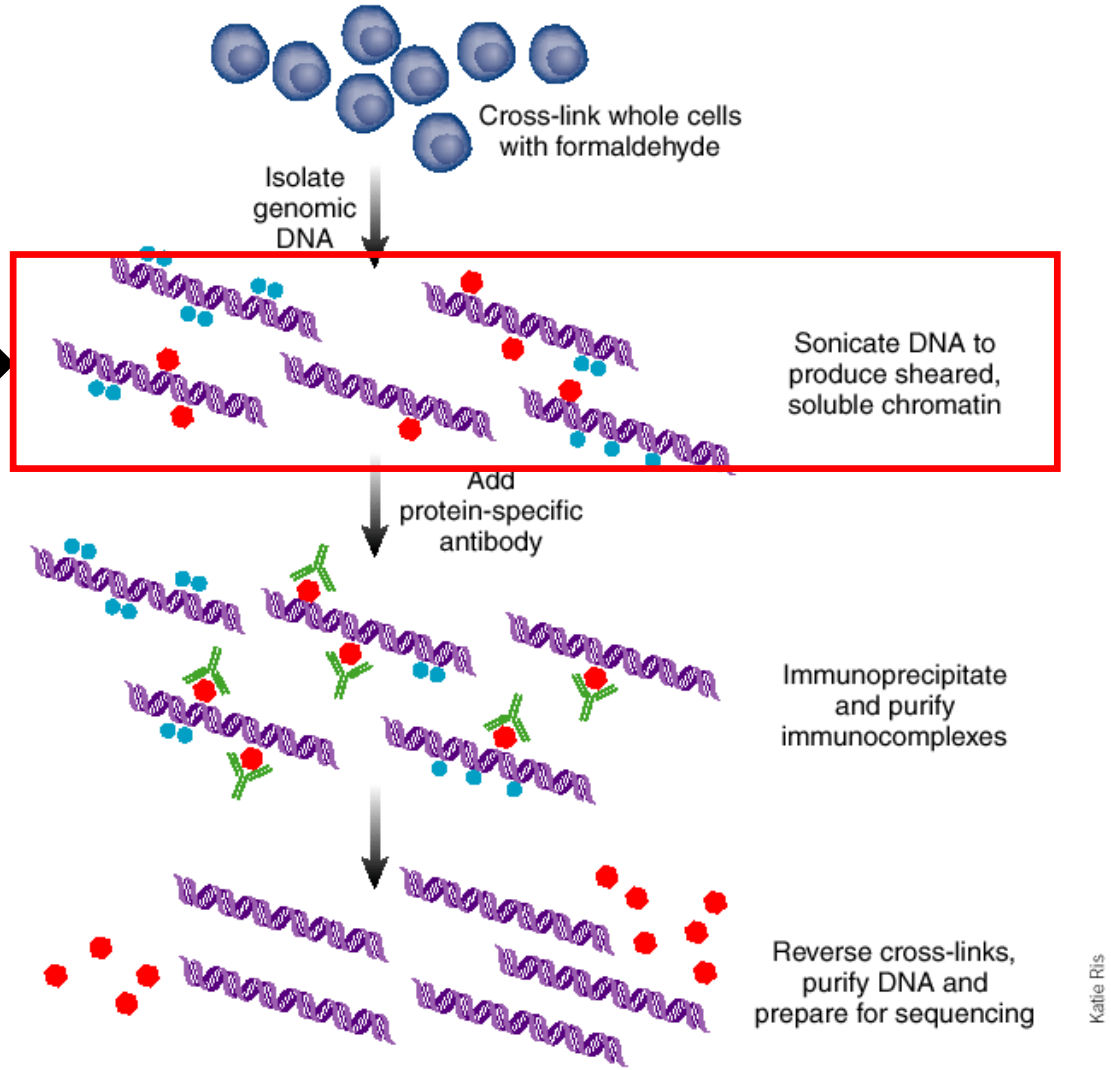
Formaldehyde crosslinking



Mardis E R; Nature Methods 4, 613 - 614 (2007)

Methodology

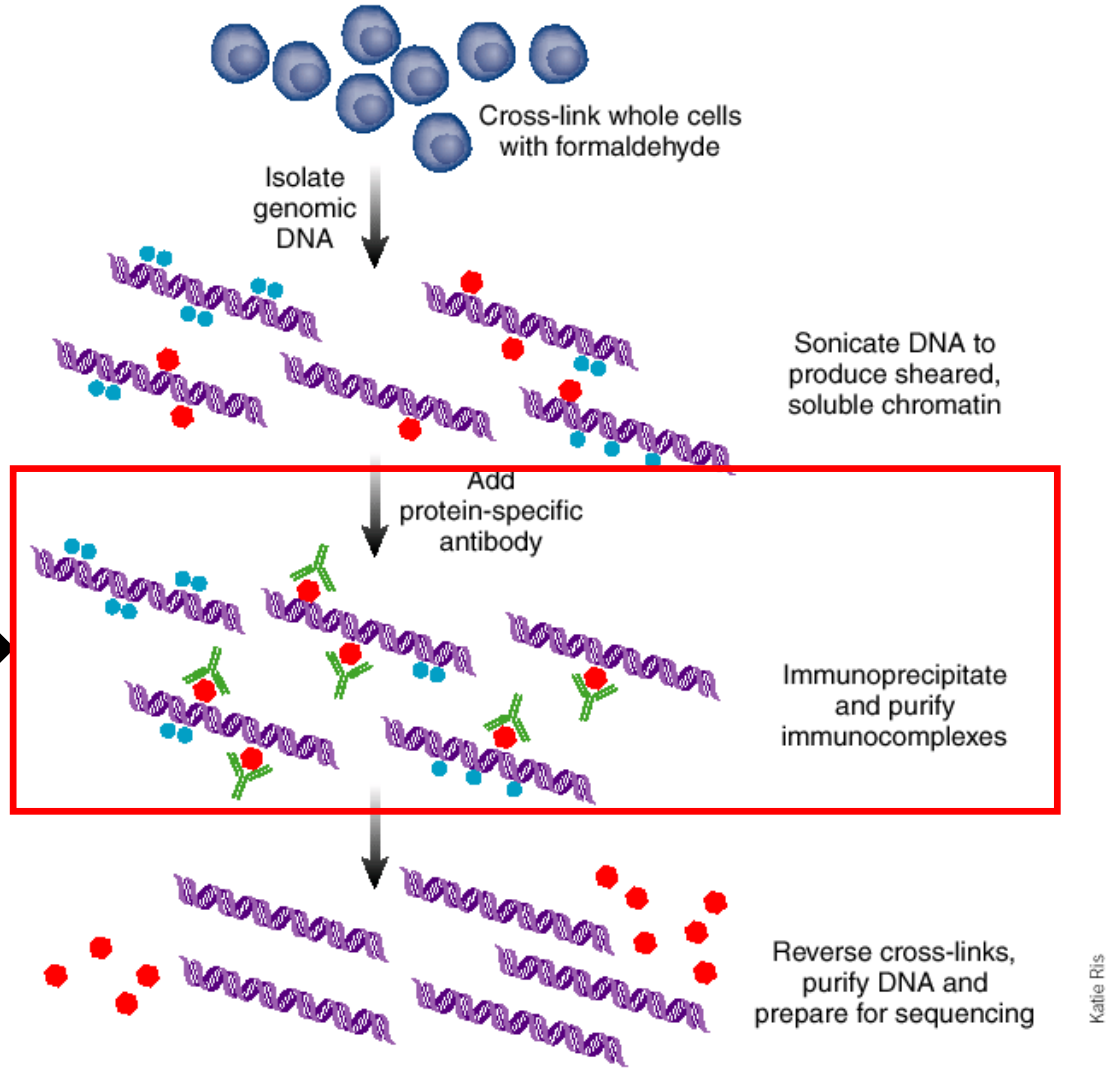
Chromatin shearing



Mardis E R; Nature Methods 4, 613 - 614 (2007)

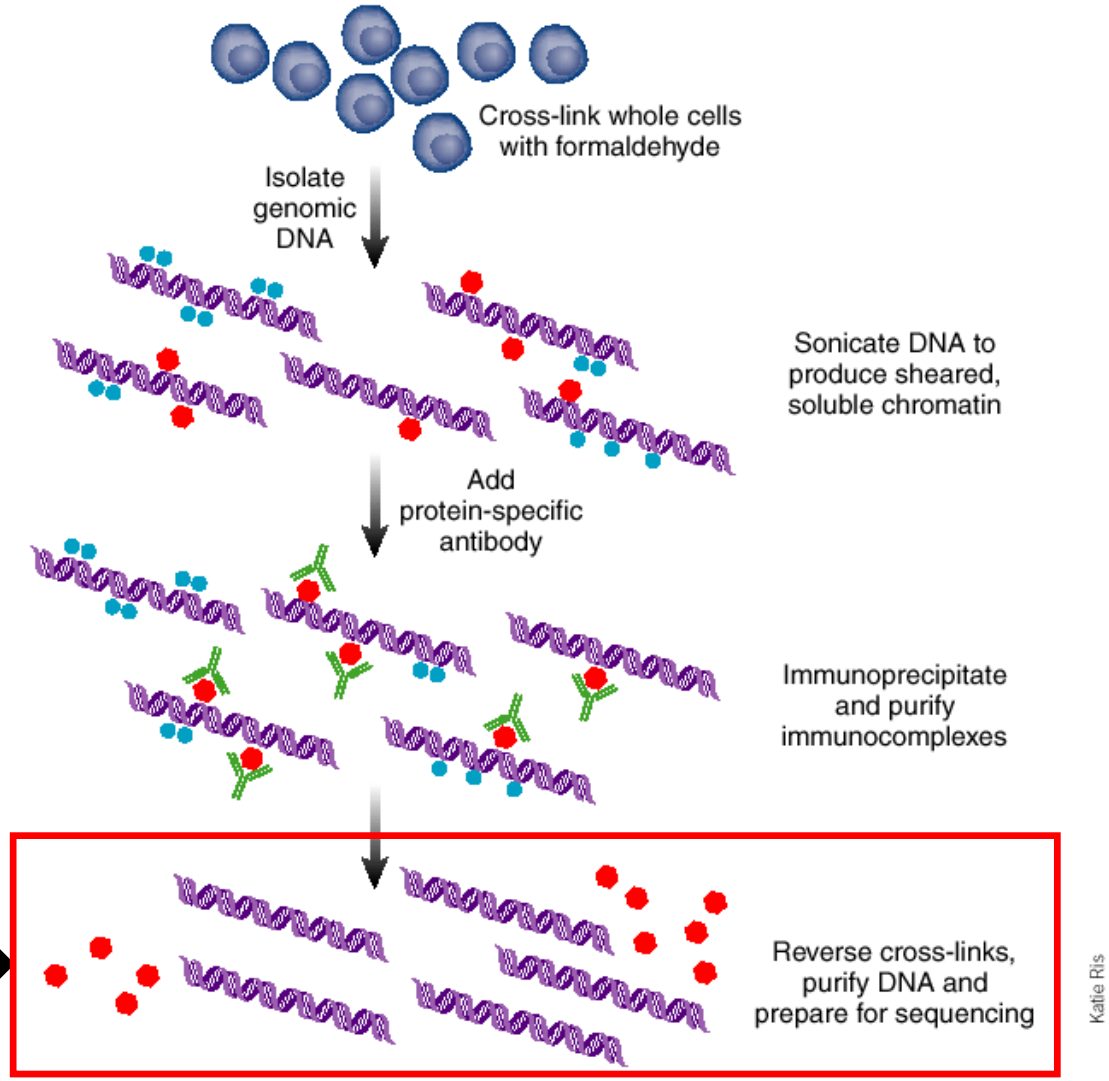
Methodology

Immunoprecipitation and purification



Mardis E R; Nature Methods 4, 613 - 614 (2007)

Methodology



Library preparation

Mardis E R; Nature Methods 4, 613 - 614 (2007)

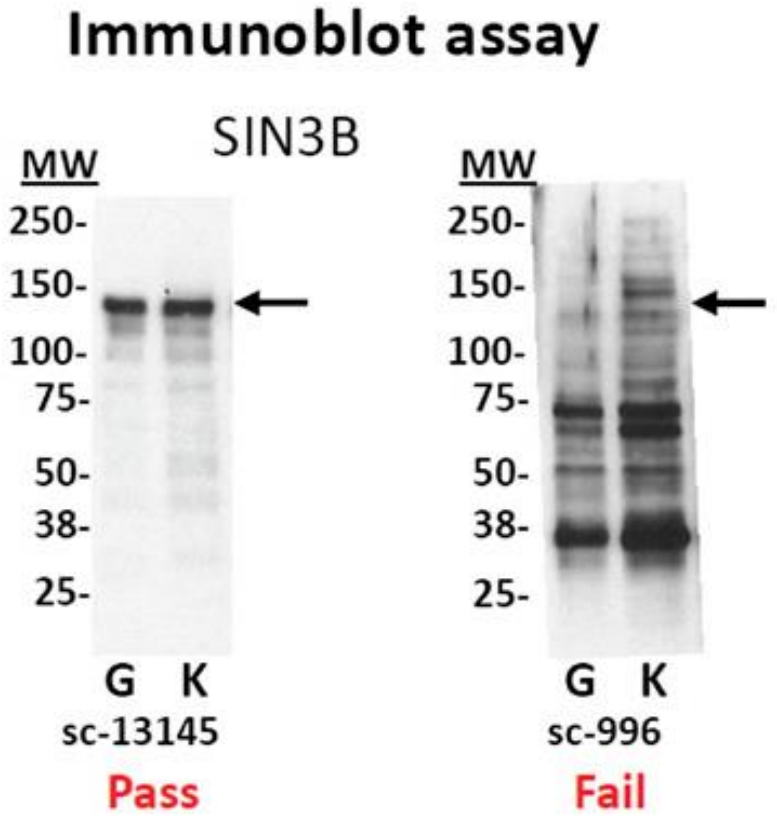
The importance of controls

Antibody and immunoprecipitation specificity

Antibody flaws:

- Poor reactivity against a chosen target protein
- Cross-reactivity with other DNA-binding proteins

ENCODE standard: primary reactive band contains >**50%** of the signal observed on the blot.



Landt SG et al. **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. Genome Res. 2012 Sep;22(9):1813-31. doi:10.1101/gr.136184.111

The importance of controls

Control sample

Critical for the experiment due to:

- Non-uniform chromatin shearing- regions of open chromatin are preferentially represented
- Repetitive regions with tendency to accumulate more sequencing tags
- Platform-specific efficiency bias (eg. GC-bias)

The importance of controls

Control sample

Different approaches:

1. „Input” DNA control- cross-linked and fragmented under the same conditions as the immunoprecipitaded sample
2. „Mock” IP control- using antibody against non-nuclear antigen

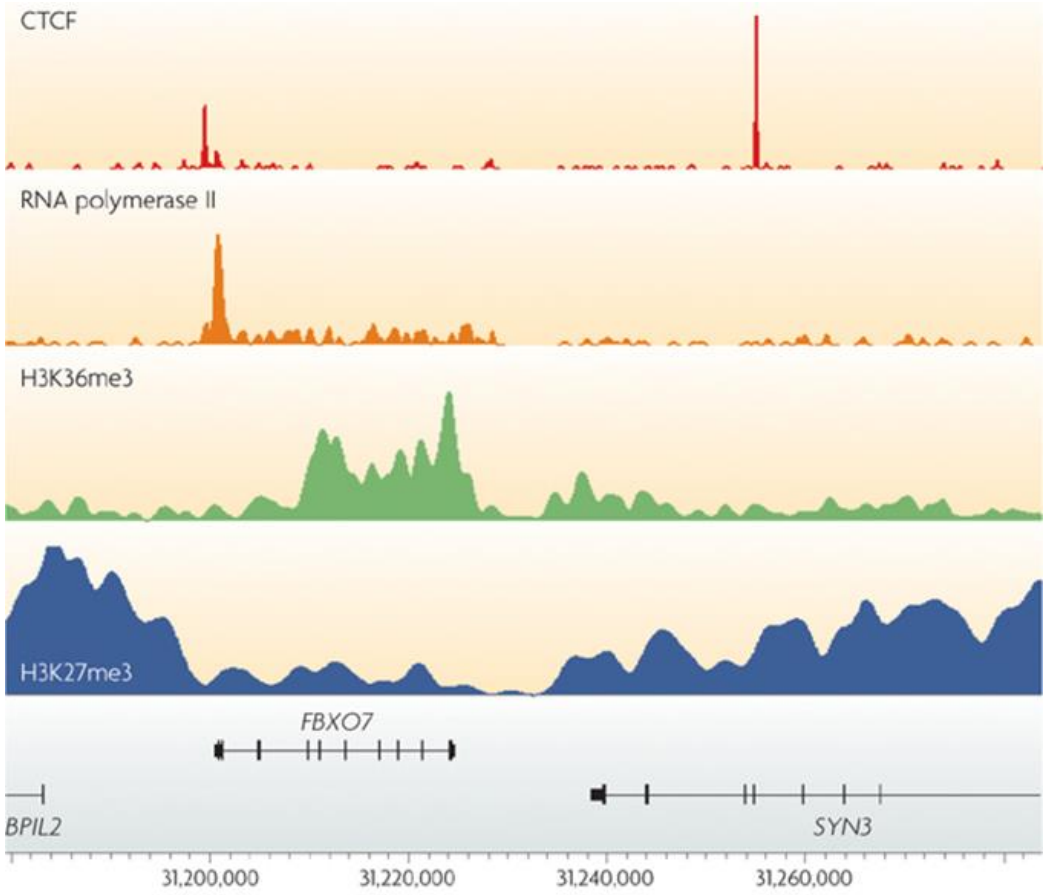
Experimental design

Distinct modes of interaction of DNA-binding proteins:

- *Point source* factors: sequence specific TFs, cofactors, transcription start site or enhancer-associated histone marks

- *Broad source* factors: chromatin marks and chromatin proteins associated with transcriptional elongation or repression

- *Mixed-source* factors



Park PJ, Nature Reviews Genetics 10, 669-680 (2009)

Experimental design

Sequencing depth

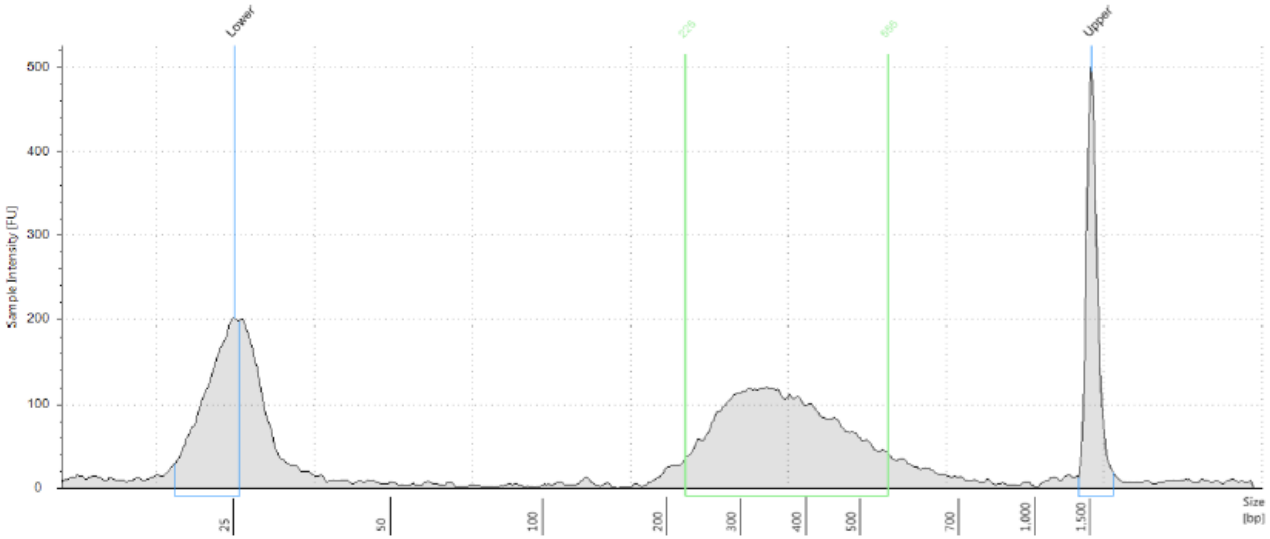
- *Point source*: mammals: mammals: at least 20 million reads per factor in two replicates, ENCODE./worms and flies: minimum of 2 million uniquely mapped reads per replicate, ENCODE/
- *Broad source*: mammals: up to 60 million reads / worms and flies: 5 million uniquely mapped reads per replicate, ENCODE/

Nakato R, Shirahige K**. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation**. Brief Bioinform. 2016 Mar 15

Experimental design

Read Length

36-100 bases



Nieścierowicz K, Unpublished

Library Type

Single-end

(if particularly interested in improving the signal strength in repetitive regions paired-end reads recommended)

ChIP-Seq analysis workflow

Quality Control

Pre-processing

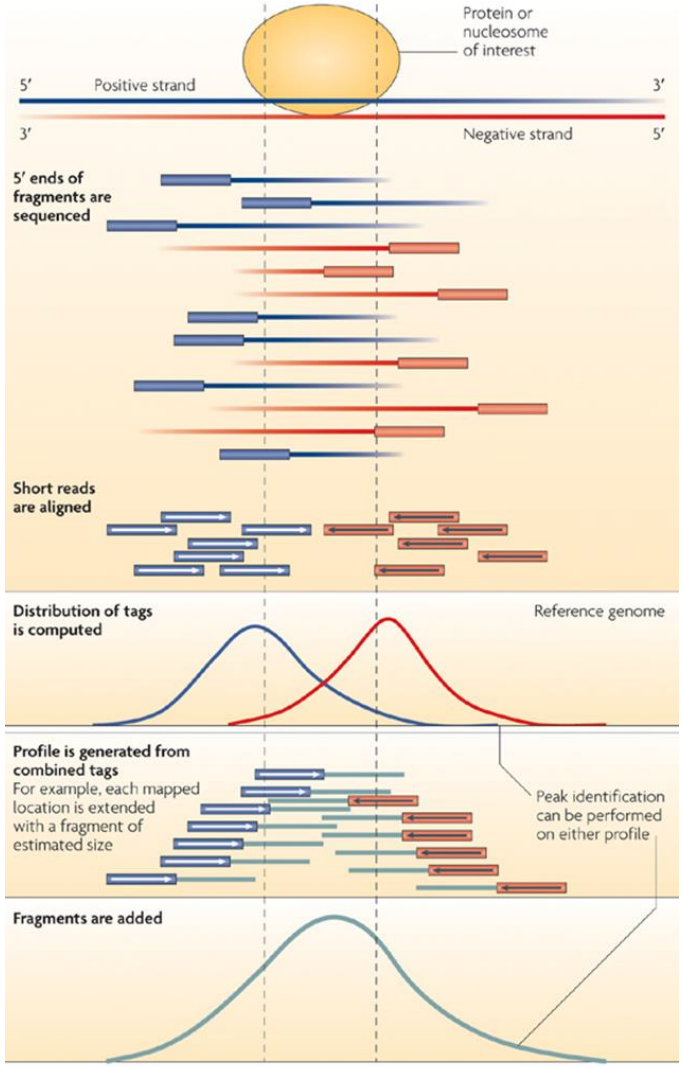
- filter out erroneous and low-quality reads
- **FastQC**, **cutadapt**, trimmomatic

Mapping

- Short read alignment- 70% uniquely mapped normal, <50% might be a concern
- Multi mapping reads will be ignored by most peak-calling programs
- **BWA**, Bowtie, SOAP, or MAQ

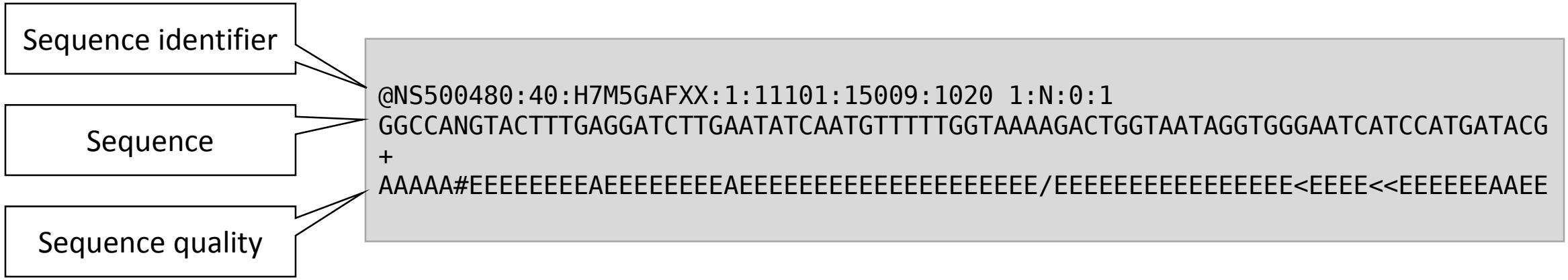
Peak finding

- identifying candidate peaks, testing for statistical significance
- **MACS2**, GEM, MUSIC, BCP, ZINBA, SPP…

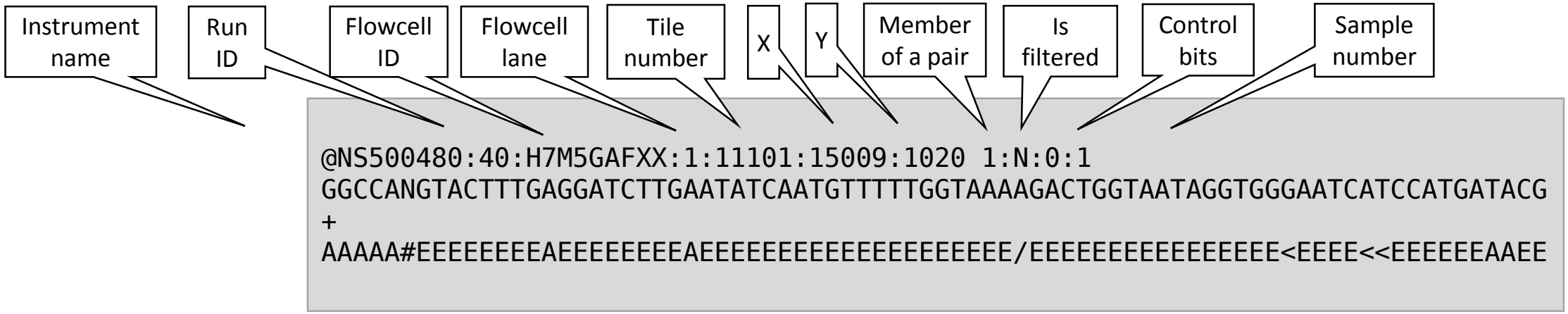


Park PJ, Nature Reviews Genetics 10, 669-680 (2009) Nature Reviews | Genetics

File formats: FASTQ



File formats: FASTQ



Quality:

$$Q_{\mathrm{phred}} = -10 \log_{10} e$$

e- the estimated probability of an incorrect base

ASCII encoding:

Sanger/Illumina 1.8+: characters Q_{phred} + 33

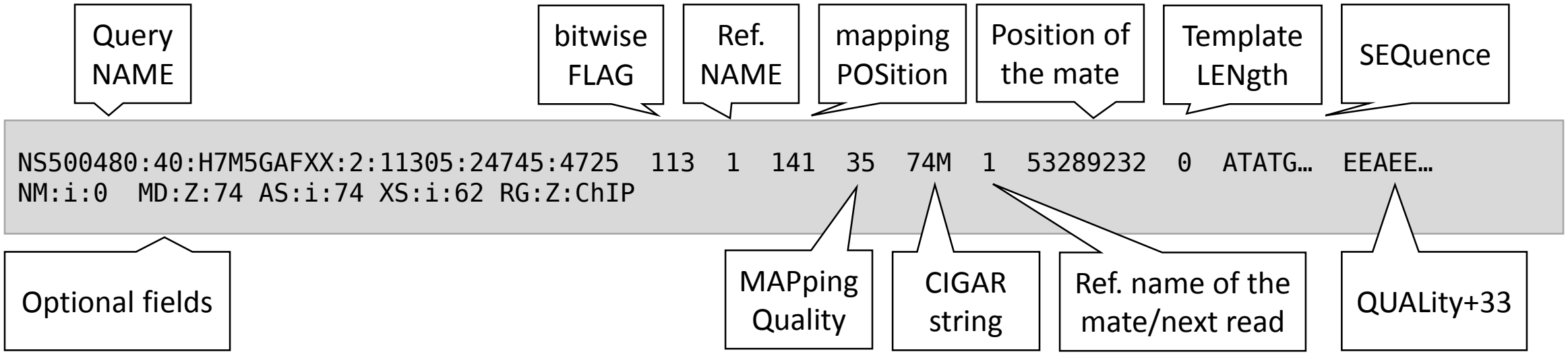
Solexa/Illumina 1.3-1.7: Q_{phred} + 64

File formats: SAM

Header section

```
@HD     VN:1.3  SO:coordinate
@SQ     SN:1    LN:58871917
@RG     ID:ChIP   SM:ChIP
@PG     ID:bwa  PN:bwa  VN:0.7.13-r1126
```

Alignments section



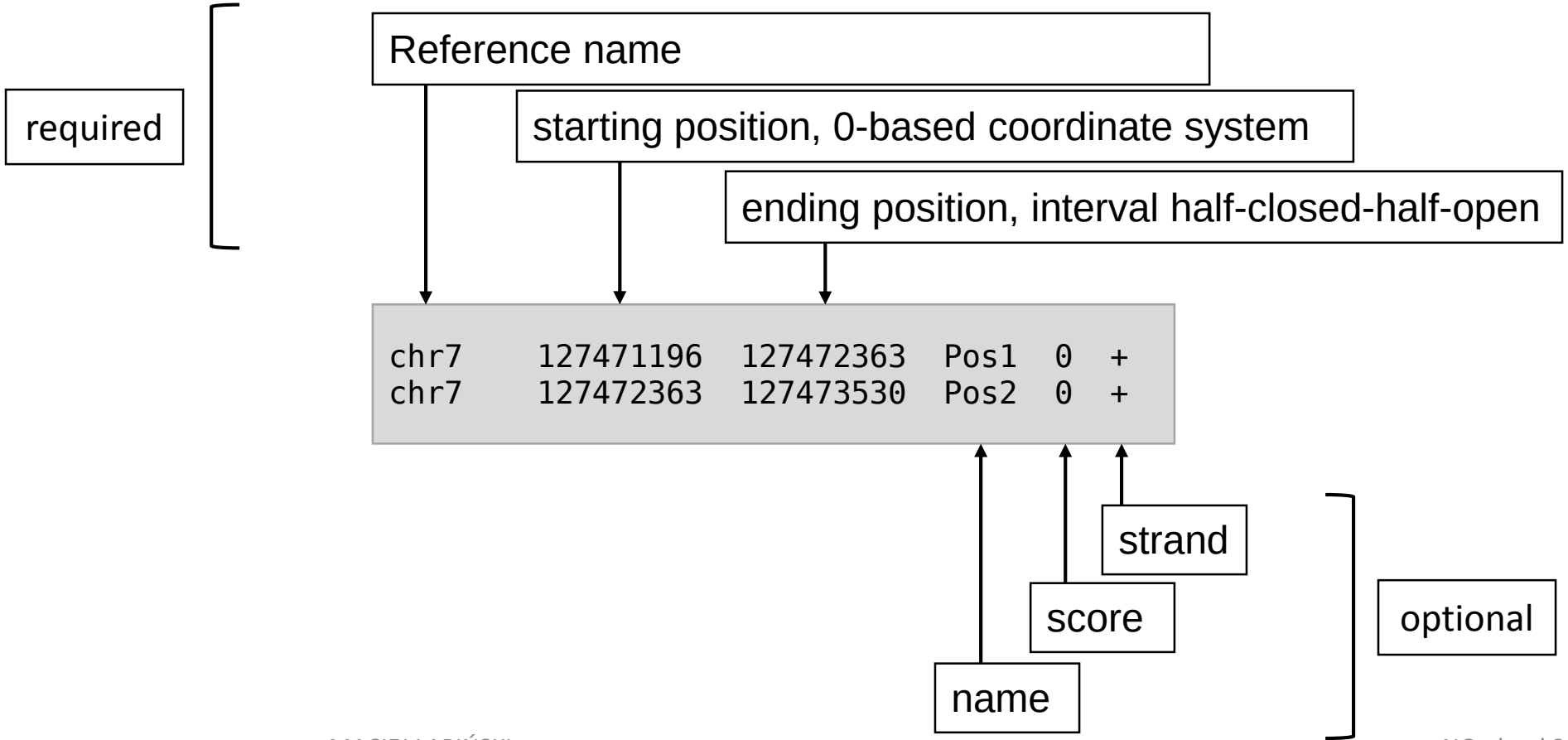
File formats: BAM

Compressed, binary version of sam.

- Space efficiency: BGZF compression- block compression on top of standard gzip file format
- Efficient random access for the indexed queries

File formats: BED

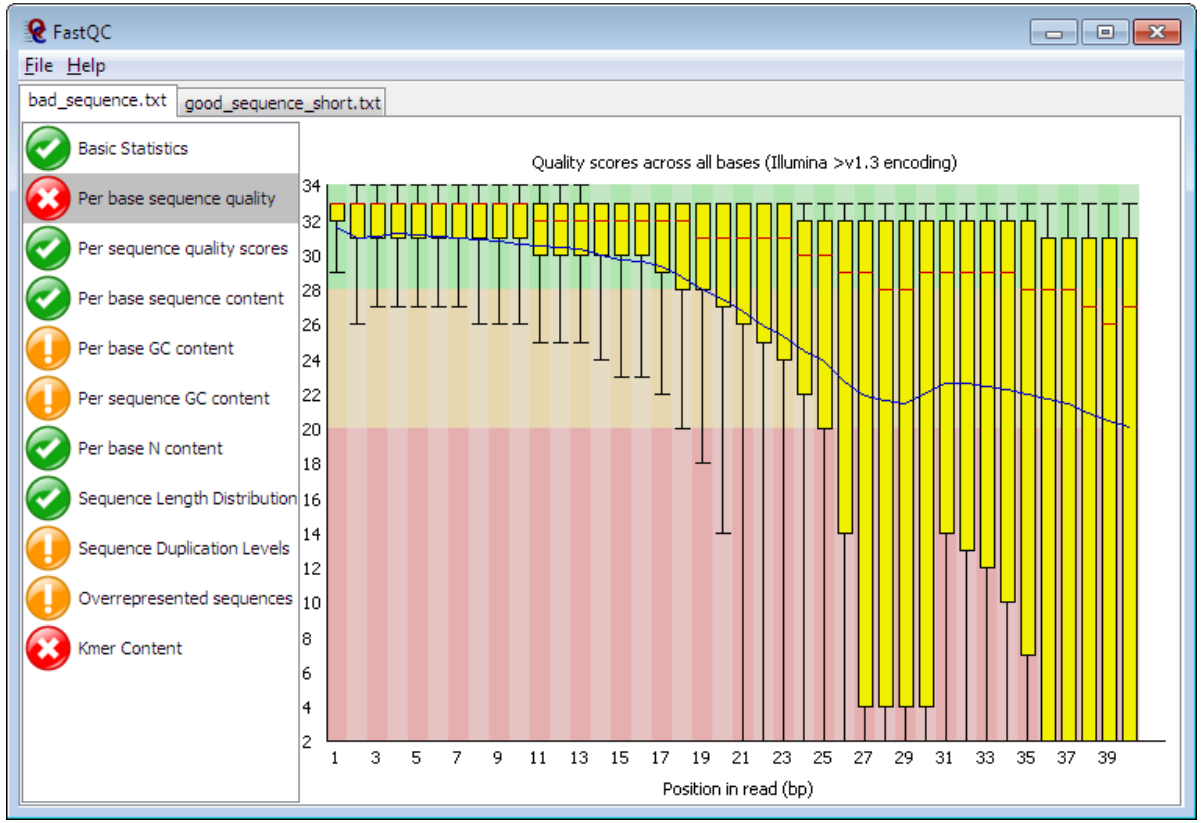
Browser Extensible Data format



Sequencing quality metrics

FastQC- quality control of raw sequence data

cutadapt- adaper and quality trimming, read filtering



Short read alignment

Maping low-divergent sequences against a large reference genome.

- Burrows-Wheeler Alignment tool (BWA)
- String matching using Burrows–Wheeler Transform (BWT)
- Same principle: SOAPv2, Bowtie
- Inexact matching algorithm
- Capable of gapped alignment of single reads

Li H. and Durbin R. (2009) **Fast and accurate short read alignment with Burrows-Wheeler transform**. Bioinformatics, 25, 1754-1760.

Resource

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia

Stephen G. Landt,[1,26] Georgi K. Marinov,[2,26] Anshul Kundaje,[3,26] Pouya Kheradpour,[4] **(...)**

Genome Res. 2012 Sep

Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation

Ryuichiro Nakato and Katsuhiko Shirahige

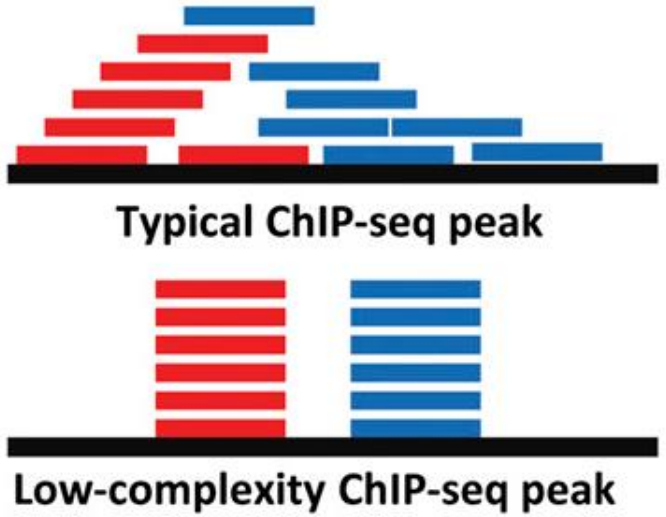
Brief Bioinform. 2016 Mar 15

Library complexity

- Measured as the fraction of nonredundant mapped reads (**NRF**) in a data set

$$NRF = \frac{N_{nonred}}{N_{all}}$$

- NRF decreases with sequencing depth
- ENCODE: NRF ≥0.8 for 10M uniquely mapped reads

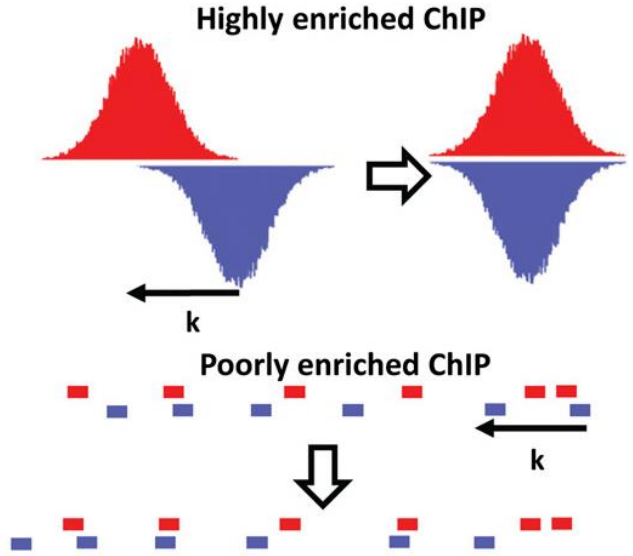


Cross-correlation analysis

Highquality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site.

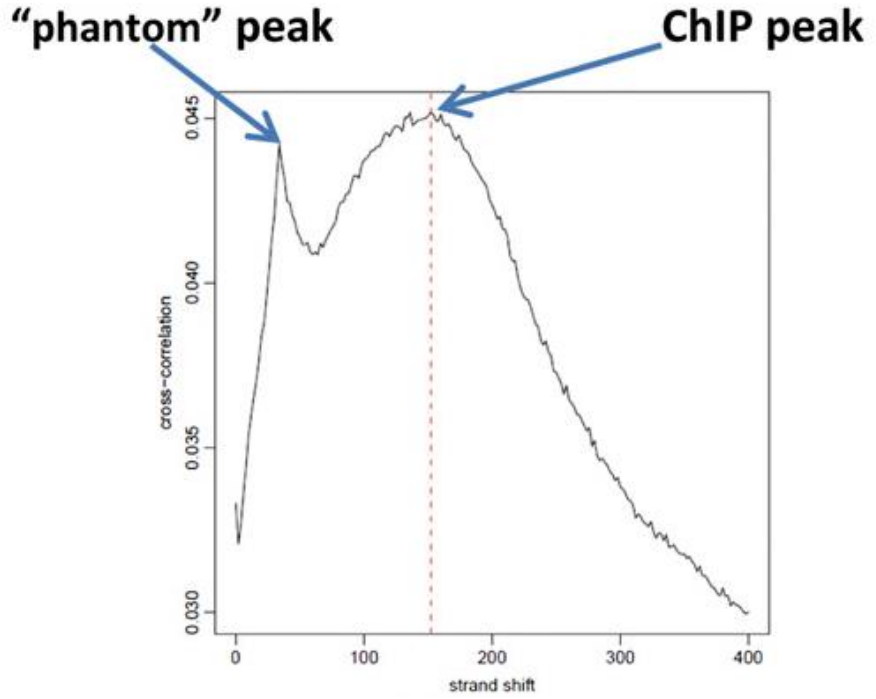
Sequence tags are positioned at a distance from the binding site center that depends on the fragment size distribution

Lack of pattern of shifted stranded tag densities



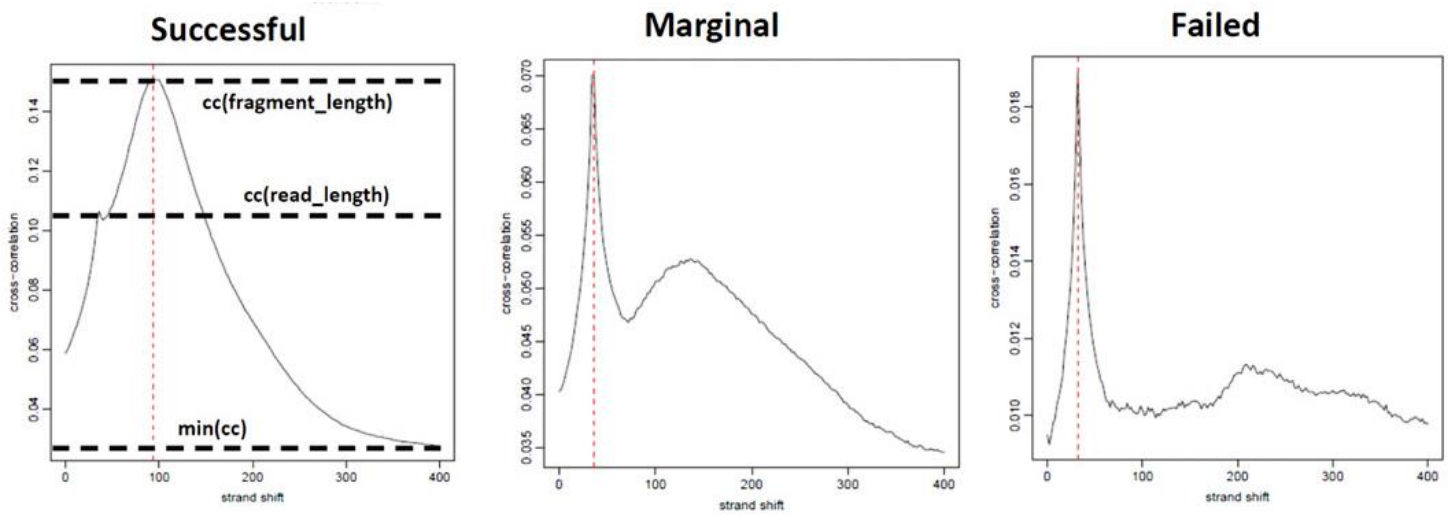
Cross-correlation analysis

Correlation between genome-wide stranded tag densities is computed as the Pearson linear correlation between the Crick strand and the Watson strand after shifting Watson by *k* base pairs.



Typically this produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length (‘‘phantom’’ peak).

Cross-correlation analysis



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

Normalized strand coefficient (**NSC**)- the normalized ratio between the fragment-length crosscorrelation peak and the background cross-correlation

Relative strand correlation (**RSC**)- the ratio between the fragment length peak and the read-length peak

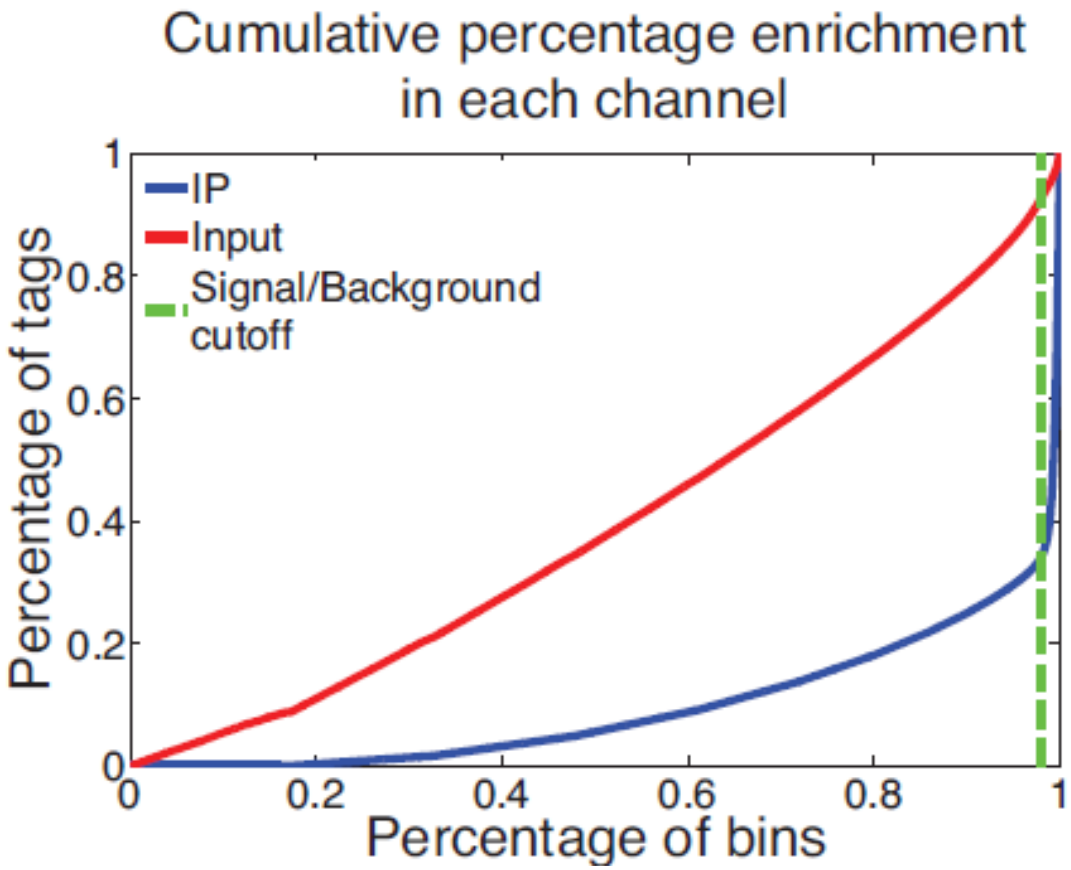
ENCODE: NSC >**1.05** and RSC >**0.8** for *point source* TFs

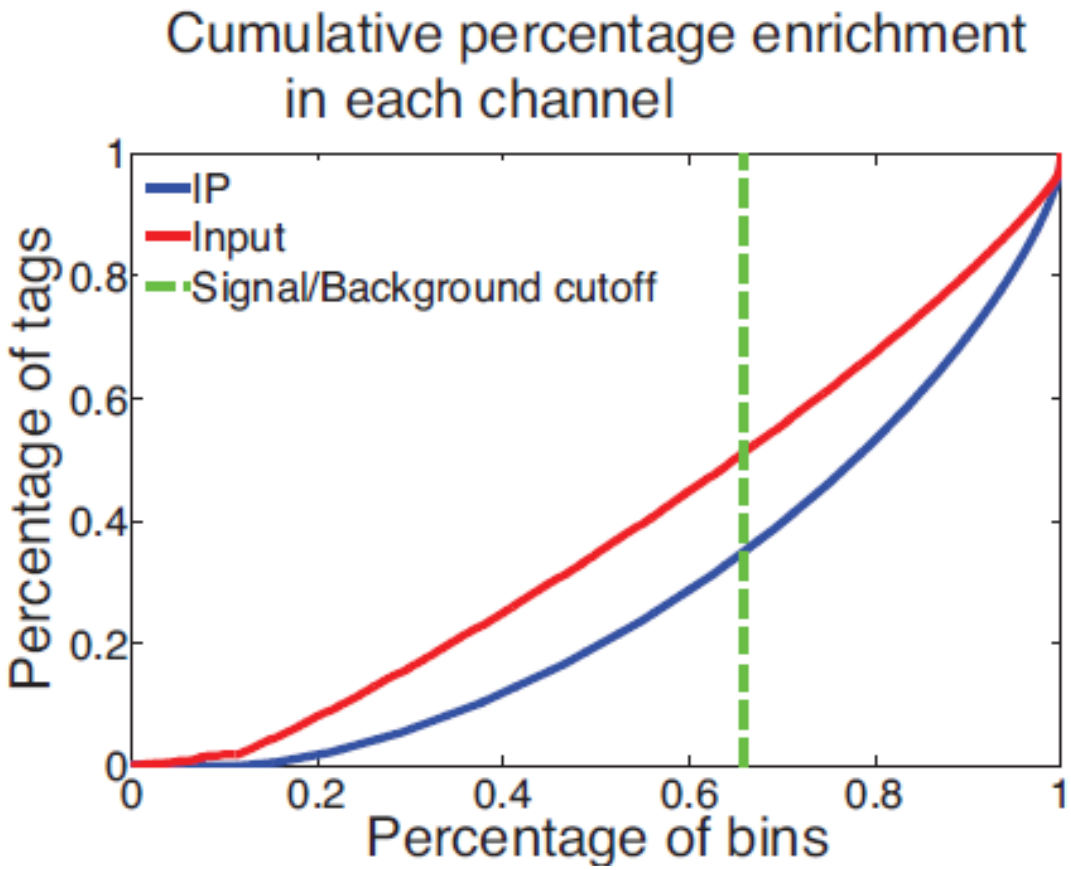
Signal extraction scaling

1. Partition the genome into non-overlaping windows
2. Count the number of alignments that fall within a given window
3. Order the windows based on the tag count
4. Plot the percentage of tags that fall into the given percentage of ordered windows

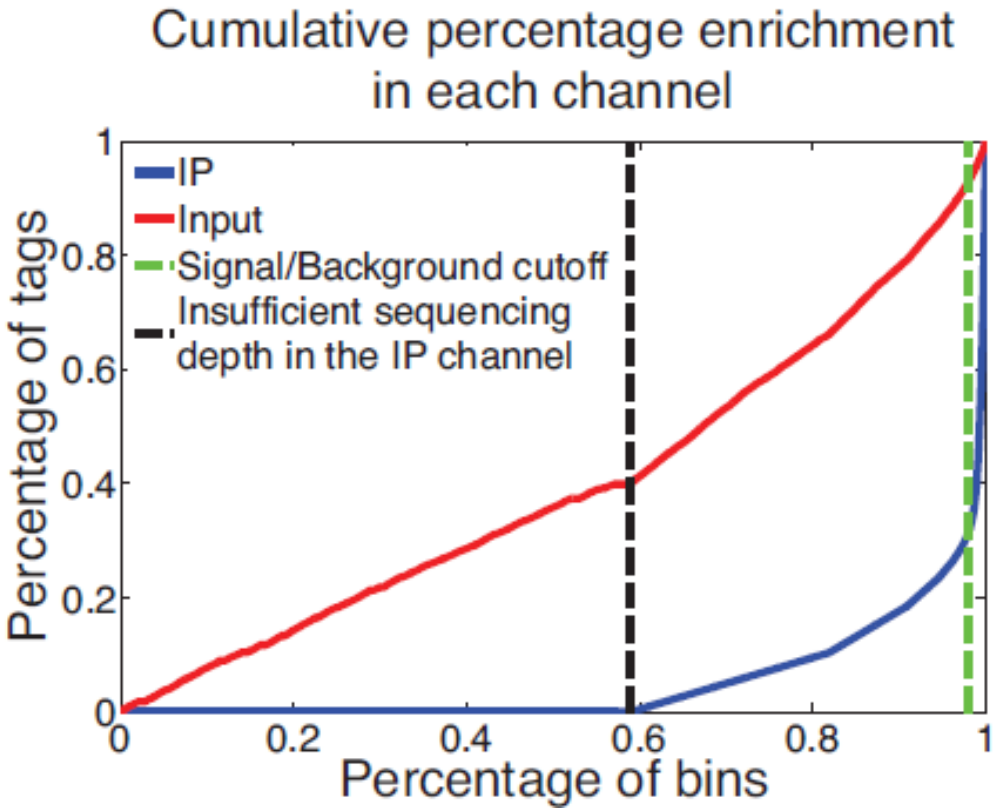
Diaz A, Nellore A, Song JS. **CHANCE: comprehensive software for quality control and validation of ChIP-seq data**. Genome Biology. 2012;13(10):R98. doi:10.1186/gb-2012-13-10-r98.

Signal extraction scaling

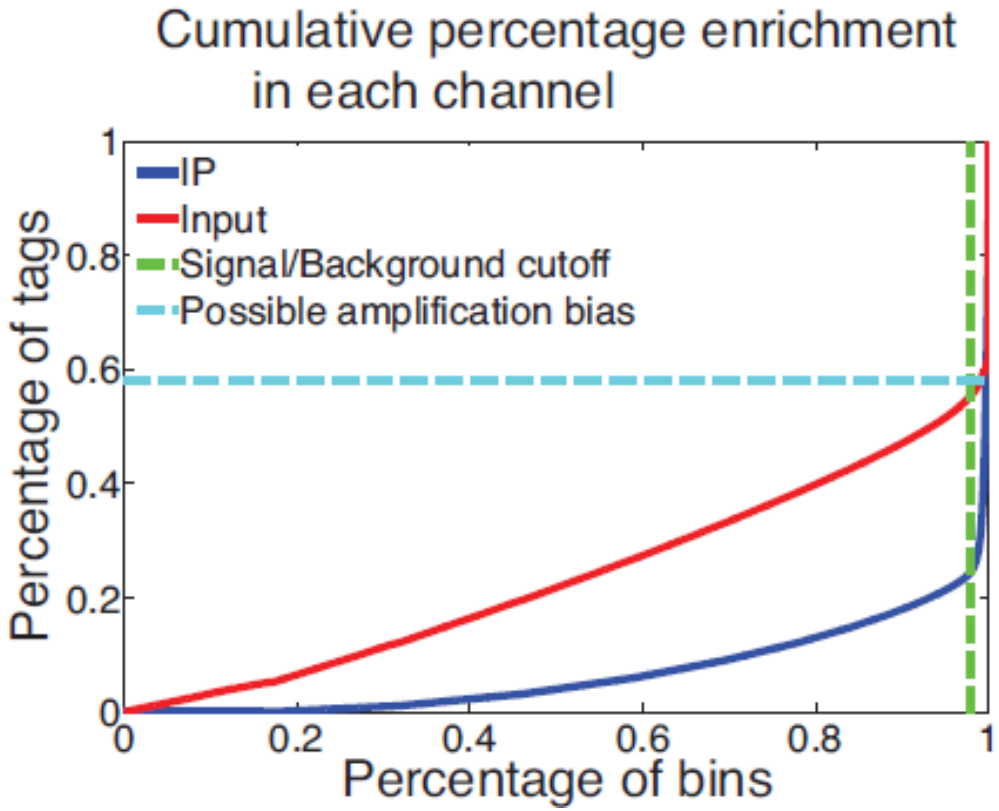




Signal extraction scaling



Insufficient sequencing depth



Amplification bias

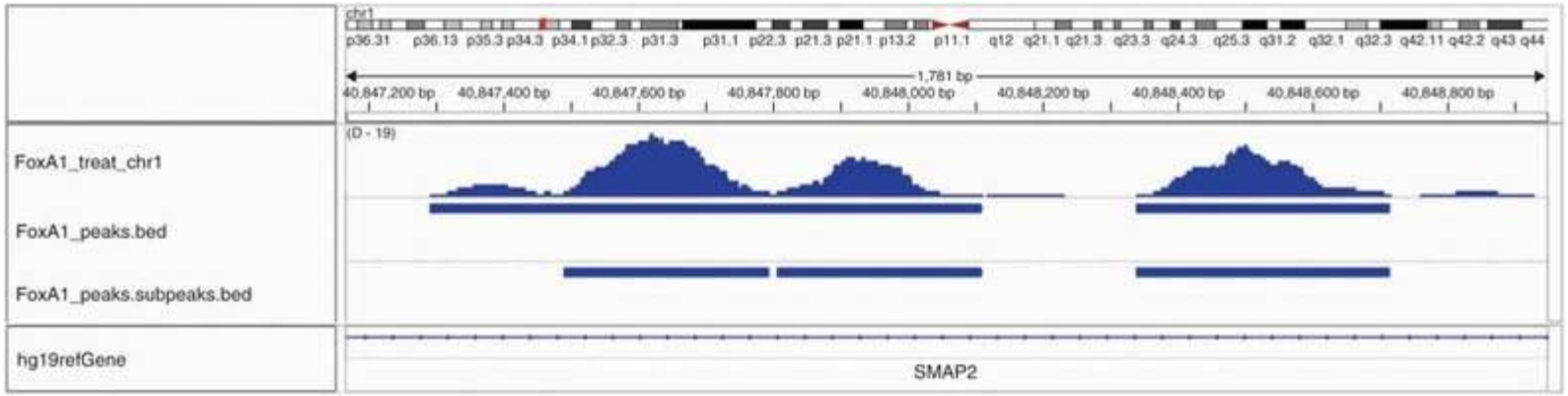
Peak calling

create a profile

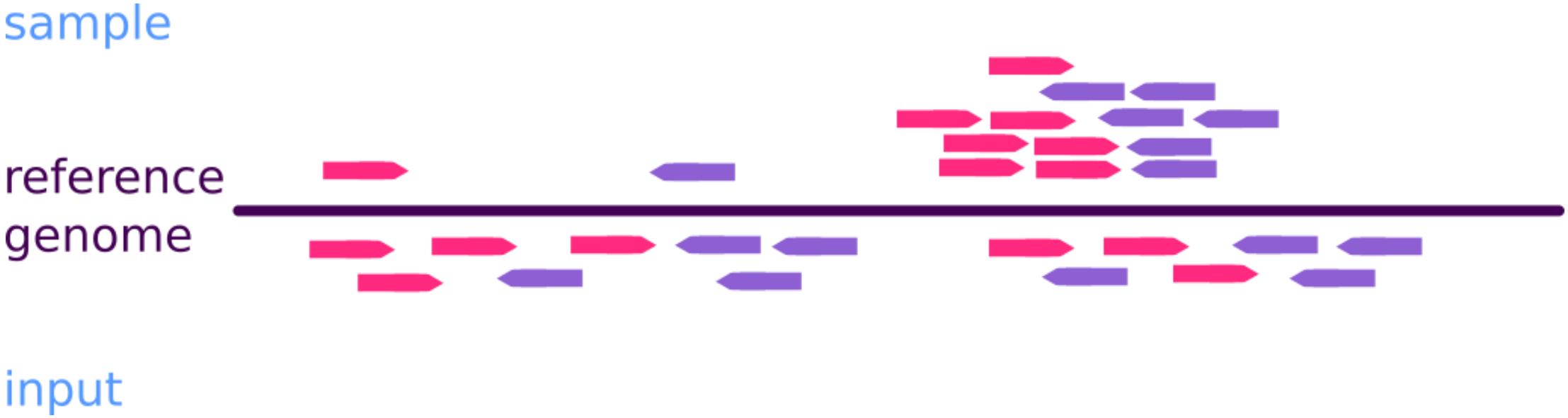
select candidate sites

calculate significance (p-value)

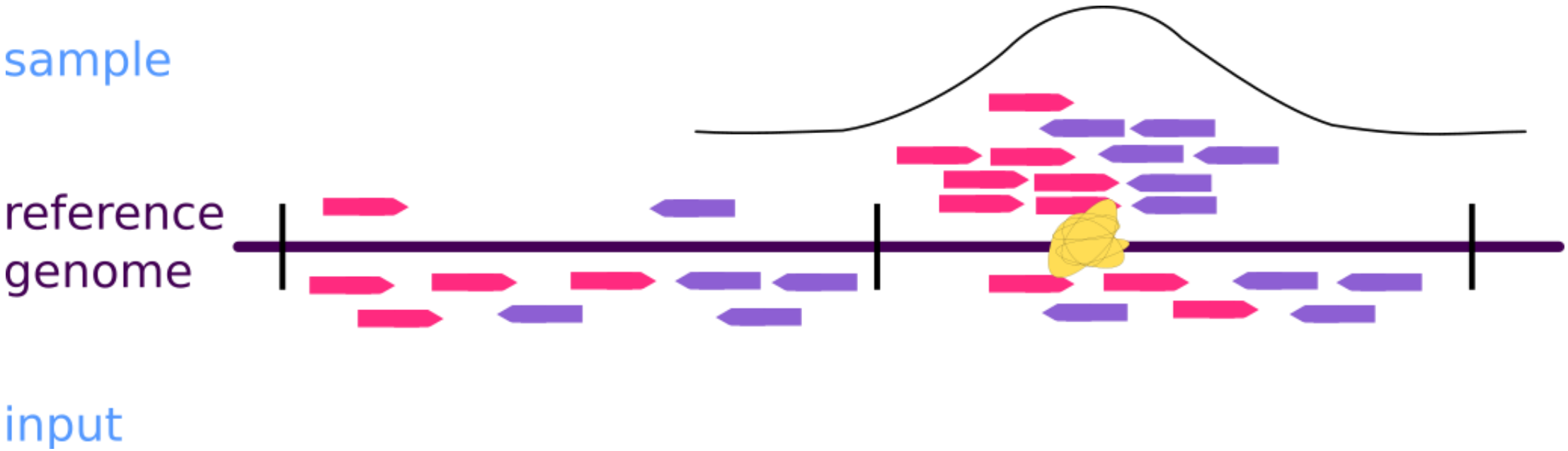
determine cut-off threshold (ie FDR)



Peak calling



Peak calling



Peak calling

Selecting the best peak caller for a given application:

- *Point source* factors: **GEM**, MACS2, BCP, SPP, ZINBA…
- *Broad source factors*: **BCP, ZINBA**, **MUSIC**, BroadPeak, MACS2, ZINBA, SPP…

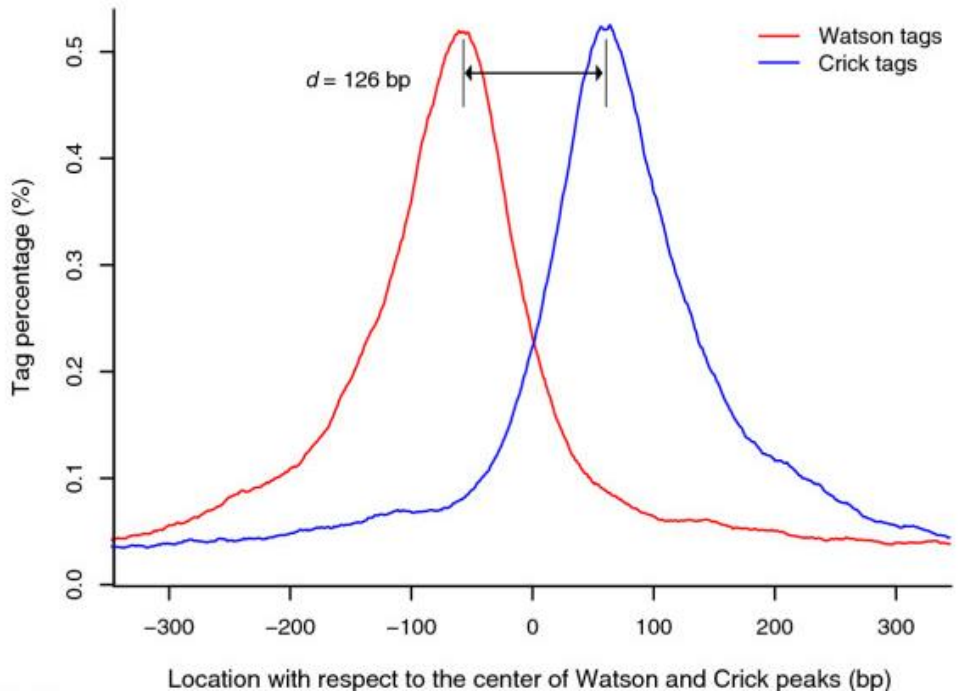
Reuben Thomas, Sean Thomas, Alisha K. Holloway, and Katherine S. Pollard **Features that define the best ChIP-seq peak calling algorithms**. Brief Bioinform 2016 : bbw035v1-bbw035.

MACS2

Decide the fragment length ***d***

Empirical modeling of '*d*' and tag shifting by d/2 to putative protein-DNA interaction site.

'*d*' is used to extend the ChIP-sample and compute ChIP coverage



MACS2

Build local bias track from control

Tag distribution along the genome is modeled by a Poisson distribution, described by the average numer of events:

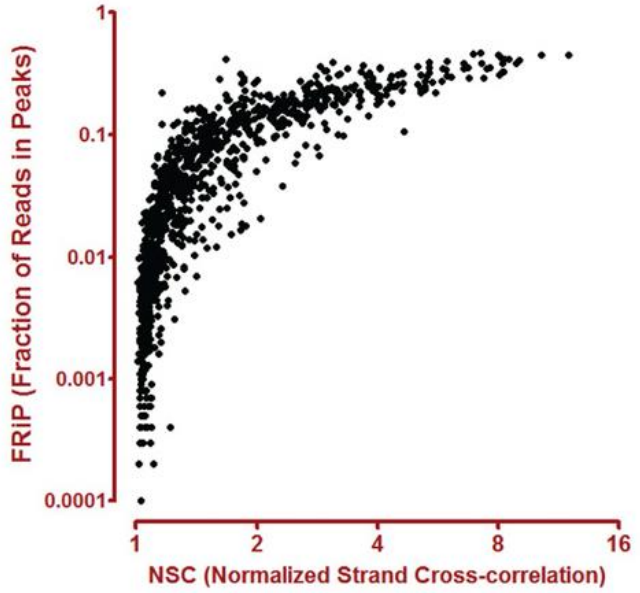
$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k}),$$

where λ_{1k}, λ_{5k} and λ_{10k} are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample

Global ChIP enrichment (FRiP)

- Fraction of Reads in Peaks
- Positive and linear correlaton with the number of called regions
- ENCODE: **> 1%**, most successful point-source factor- FRiP values of 0.2–0.5 and NSC/RSC values of 5–12.

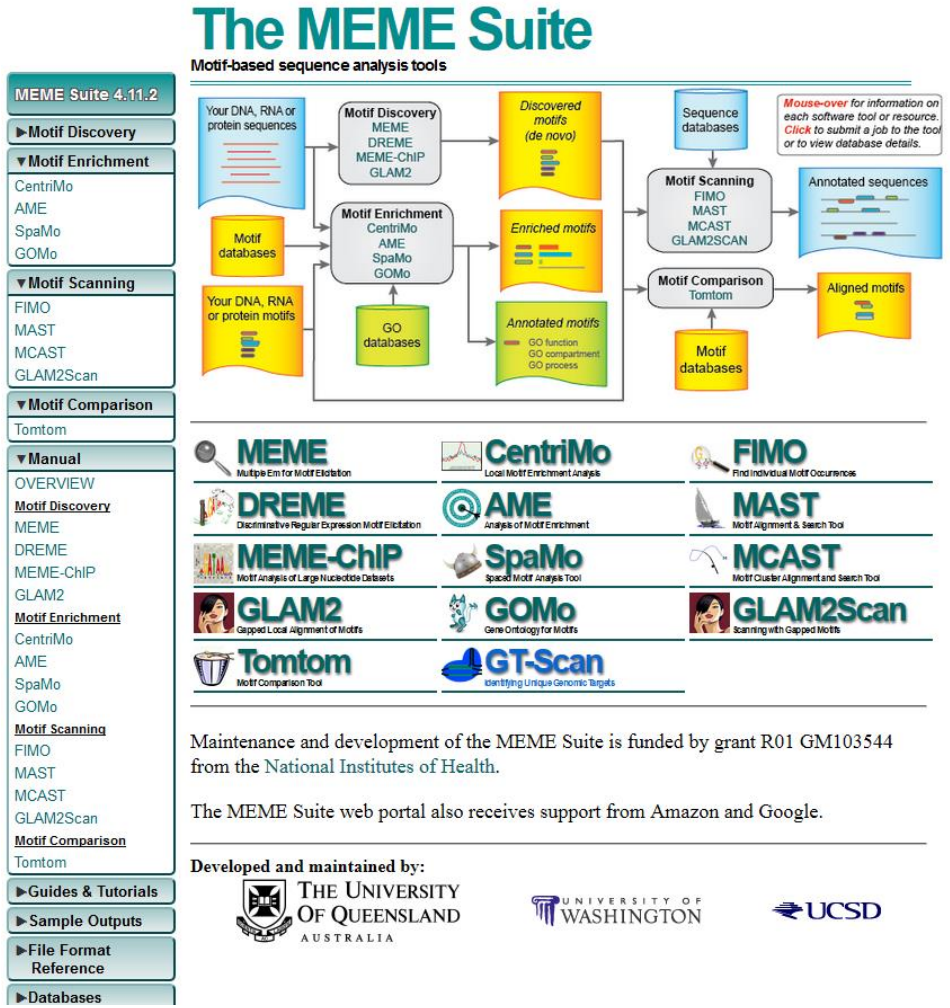
Correlatio between FRiP an NSC for 1052 human ChiP-Seq experiments



MEME- novel motifs discovery

- Discover novel sequence motifs
- Scan your sequences with a given motif
- Analyse the similarity to known motifs

http://meme-suite.org



Gene set enrichment analysis

- **GREAT**- annotate non-coding regions, http://bejerano.stanford.edu/great/public/html/

- **DAVID**- functional annotation, https://david.ncifcrf.gov/

Thank you!