



# Next generation genomics

- \* alignment file formats: CIGAR strings, SAM
- \* peak finding
- \* RNASeq mapping

## \* alignment file formats

**How do we communicate the results of an alignment experiment?**

universal formats

driven by large projects

and large genomics centres

e.g. the 1000 human genomes initiative



# \* alignment file formats: CIGAR strings

## CIGAR FORMAT

Op	Description
M	alignment match (can be a sequence match or mismatch)
I	insertion to the reference
D	deletion from the reference
N	skipped region from the reference
S	soft clipping (clipped sequences present in SEQ)
H	hard clipping (clipped sequences NOT present in SEQ)
P	padding (silent deletion from padded reference)
=	sequence match
X	sequence mismatch

## \* alignment file formats: CIGAR strings

describing the EDIT path to transform one sequence into another

Read 1

**AGCTGCTTTGCA**

->

Read 2

**AGCT-CT**A**TGCA**

3M1D2M1S4M

**\* alignment file formats: SAM format**

# Sequence Alignment/Map format

**The SAM Format Specification (v1.3-r882)**

The SAM Format Specification Working Group

December 11, 2010

<http://samtools.sourceforge.net/SAM-1.3.pdf>

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
Coord      12345678901234   5678901234567890123456789012345
ref        AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1    TTAGATAAAGGATA*CTG
+r002      aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004      ATAGCT.....TCAGC
-r003      ttagctTAGGC
-r001/2    CAGCGCCAT
```

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5H6M * 0 0 AGCTAA * NM:i:1
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 16 ref 29 30 6H5M * 0 0 TAGGC * NM:i:0
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

**@ = header rows**

can [SHOULD] include experimental metadata such as program used, version, parameters, researcher, data source, date, etcetera

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
1      2      3      4  5  6              7  8  9  10              11  12
r001   83   ref 37 30 9M              = 7  -39 CAGCGCCAT          *          *
```

### COLUMNS

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]f1,255g	Query template NAME
2	FLAG	Int	[0,216-1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>--][!-~]*	Reference sequence NAME
4	POS	Int	[0,229-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,28-1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>--][!-~]*	Ref. name of the mate/next fragment
8	PNEXT	Int	[0,229-1]	Position of the mate/next fragment
9	TLEN	Int	[-229+1,229-1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	fragment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
1      2      3      4 5 6      7 8 9 10      11      12
r001  83   ref 37 30 9M      = 7 -39 CAGCGCCAT      *      
```

### COLUMNS

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]f1,255g	Query template NAME
2	FLAG	Int	[0,216-1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>--][!-~]*	Reference sequence NAME
4	POS	Int	[0,229-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,28-1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>--][!-~]*	Ref. name of the mate/next fragment
8	PNEXT	Int	[0,229-1]	Position of the mate/next fragment
9	TLEN	Int	[-229+1,229-1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	fragment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
1      2      3      4  5  6      7  8  9  10      11      12
r001  83   ref 37 30 9M      =  7  -39  *      *      
```

### COLUMNS

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]f1,255g	Query template NAME
2	FLAG	Int	[0,216-1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>--][!-~]*	Reference sequence NAME
4	POS	Int	[0,229-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,28-1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>--][!-~]*	Ref. name of the mate/next fragment
8	PNEXT	Int	[0,229-1]	Position of the mate/next fragment
9	TLEN	Int	[-229+1,229-1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	fragment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

### COLUMNS

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]f1,255g	Query template NAME
2	FLAG	Int	[0,216-1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>--][!-~]*	Reference sequence NAME
4	POS	Int	[0,229-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,28-1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>--][!-~]*	Ref. name of the mate/next fragment
8	PNEXT	Int	[0,229-1]	Position of the mate/next fragment
9	TLEN	Int	[-229+1,229-1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	fragment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

```
r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT *
```

Bit	Description
0x1	template having multiple fragments in sequencing
0x2	each fragment properly aligned according to the aligner
0x4	fragment unmapped
0x8	next fragment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next fragment in the template being reversed
0x40	the first fragment in the template
0x80	the last fragment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate

# \* alignment file formats: SAM format

## The SAM Format Specification (v1.3-r882)

1	2	3	4	5	6	7	8	9	10	11	12
r003	0	ref	9	30	5H6M	*	0	0	AGCTAA	*	NM:i:1
r003	16	ref	29	30	6H5M	*	0	0	TAGGC	*	NM:i:0

## OPTIONAL COLUMNS

TAG:TYPE:VALUE

# \* alignment file formats: SAM format

## TAG:TYPE:VALUE

Tag	Type	Description
X?	?	Reserved fields for end users (together with Y? and Z?)
AM	i	The smallest template-independent mapping quality of fragments in the rest
AS	i	Alignment score generated by aligner
BQ	Z	Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, $BAQ_i = Q_i - (BQ_i - 64)$ where $Q_i$ is the i-th base quality.
CM	i	Edit distance between the color sequence and the color reference (see also NM)
CQ	Z	Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.
CS	Z	Color read sequence on the original strand of the read. The primer base must be included.
E2	Z	The 2nd most likely base calls. Same encoding and same length as QUAL.
FI	i	The index of fragment in the template.
FS	Z	Fragment suffix.
LB	Z	Library. Value to be consistent with the header RG-LB tag if @RG is present.
H0	i	Number of perfect hits
H1	i	Number of 1-difference hits (see also NM)
H2	i	Number of 2-difference hits
HI	i	Query hit index, indicating the alignment record is the i-th one stored in SAM
IH	i	Number of stored alignments in SAM that contains the query in the current record
MD	Z	String for mismatching positions. Regex : $[0-9]+(([\text{ACGTN}] \^[ACGTN]+)[0-9]+)^* 1$
MQ	i	Mapping quality of the mate/next fragment
NH	i	Number of reported alignments that contains the query in the current record
NM	i	Edit distance to the reference, including ambiguous bases but excluding clipping
OQ	Z	Original base quality (usually before recalibration). Same encoding as QUAL.
OP	i	Original mapping position (usually before realignment)
OC	Z	Original CIGAR (usually before realignment)
PG	Z	Program. Value matches the header PG-ID tag if @PG is present.
PQ	i	Phred likelihood of the template, conditional on both the mapping being correct
PU	Z	Platform unit. Value to be consistent with the header RG-PU tag if @RG is present.
Q2	Z	Phred quality of the mate/next fragment. Same encoding as QUAL.
R2	Z	Sequence of the mate/next fragment in the template.
RG	Z	Read group. Value matches the header RG-ID tag if @RG is present in the header.
SM	i	Template-independent mapping quality
TC	i	The number of fragments in the template.
U2	Z	Phred probability of the 2nd call being wrong conditional on the best being wrong. The same encoding as QUAL.
UQ	i	Phred likelihood of the fragment, conditional on the mapping being correct

## \* alignment file formats: BAM format

**BAM is binary SAM,**  
compressed in the BGZF format.

- BGZF is block compression implemented on top of the standard gzip file format

BAM files are indexed by reference sequence block

- indexing aims to achieve fast retrieval of alignments overlapping a specified region without going through the whole alignment.



# Next generation genomics

- \* alignment file formats: CIGAR strings, SAM
- \* peak finding
- \* RNASeq mapping

## \* peak finding

- many next-gen experiments are **not** attempting to cover the whole genome evenly
- aiming to identify **functionally significant** regions of the genome in reduced-representation subsets
- using technical enrichment *or*
- biological enrichment

# \* reduced representation technologies

targeted/capture resequencing

ChIP-Seq

*chromatin immunoprecipitation*

ccc- or 3C-Seq

*chromosome conformation capture*

RNA-IP-Seq

*immunoprecipitated RNA*

RNA-Seq

*sequencing the transcriptome directly*

deepSAGE

*serial analysis of gene expression using NG*

RAD-Seq

*restriction-site associated DNA*

MeDIP-Seq

*methylated DNA*

BS-Seq

*(bisulphite sequencing) methylated DNA*

## \* bisulphite sequencing

- DNA can be methylated on C residues
- bisulphite treatment converts non-methylated C to U
- direct sequencing of bisulphite treated methylated DNA results in reads where meC is read as T

## \* bisulphite sequencing

- map using 3 reference sequences
  - no changes
  - all C converted to T
  - all G converted to A

# \* reduced representation technologies

targeted/capture resequencing

ChIP-Seq

*chromatin immunoprecipitation*

ccc- or 3C-Seq

*chromosome conformation capture*

RNA-IP-Seq

*immunoprecipitated RNA*

RNA-Seq

*sequencing the transcriptome directly*

deepSAGE

*serial analysis of gene expression using NG*

RAD-Seq

*restriction-site associated DNA*

MeDIP-Seq

*methylated DNA*

## \* peak finding

- these technologies are ENRICHMENT based
- expect to find **over-representation** of targets compared to bulk genome
- specificity depends on reagents (e.g. IP) or biological signal-noise ration (e.g. RNA-Seq)

## \* peak finding

- 1 *Map reads*
- 2 Signal profiling
- 3 Measure or Estimate background
- 4 Call peaks
- 5 Filter artifacts
- 6 Assess significance of remaining peaks

[Journal home](#) > [Supplement](#) > [Review](#) > [Full Text](#)**Journal content**[Journal home](#)[Advance online publication](#)[Current issue](#)[Archive](#)[Focuses and Supplements](#)[Methagora](#)[Method of the Year 2010](#)[Press releases](#)**Journal information**[Guide to authors](#)[Online submission](#)[Subscribe](#)[New Subscription](#)**REVIEW***Nature Methods* **6**, S22 - S32 (2009)

doi:10.1038/nmeth.1371

**Computation for ChIP-seq and RNA-seq studies**Shirley Pepke<sup>1</sup>, Barbara Wold<sup>2</sup> & Ali Mortazavi<sup>2</sup>

**Genome-wide measurements of protein-DNA interactions and transcriptomes are increasingly done by deep DNA sequencing methods (ChIP-seq and RNA-seq). The power and richness of these counting-based measurements comes at the cost of routinely handling tens to hundreds of millions of reads. Whereas early adopters necessarily developed their own custom computer code to analyze the first ChIP-seq and RNA-seq datasets, a new generation of more sophisticated algorithms and software tools are emerging to assist in the analysis phase of these projects. Here we describe the multilayered analyses of ChIP-seq and RNA-seq datasets, discuss the software packages currently available to perform tasks at each layer and describe some upcoming challenges and features for future analysis tools. We also discuss how software choices and uses are affected by specific aspects of the underlying biology and data structure, including genome size, positional clustering of transcription factor binding sites, transcript discovery and expression quantification.**

**Apply for your free  
subscription to  
Nature Methods**[Subscribe](#)**This issue**[Table of contents](#)[Previous article](#)**Article tools**[Download PDF](#)[Send to a friend](#)[Export citation](#)[Export references](#)[Rights and permissions](#)[Order commercial reprints](#)[Save this link](#)

Information extraction

ChIP-seq

RNA-seq  
quantification

RNA-seq  
discovery

Integrate

RNA-seq, ChIP-seq and external data

Analyze

Associated  
genes

Differential  
expression

Novel splice  
isoforms

Motif finding

Expression  
levels

Novel gene  
models

Aggregate  
and identify

Binding sources

Novel  
transfrags

Enriched  
regions

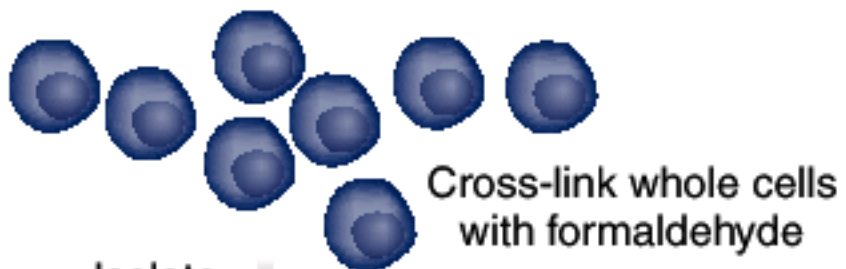
Density on known  
exons

*De novo*  
transcript  
assembly

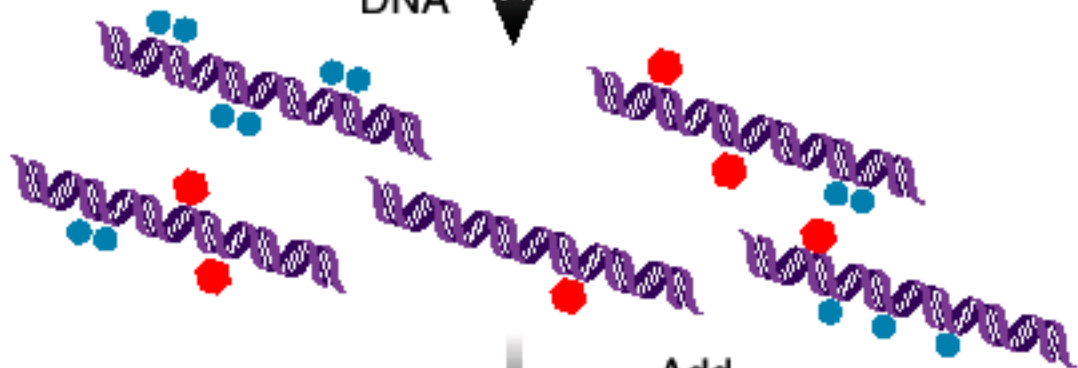
Maps read

Splice-crossing reads

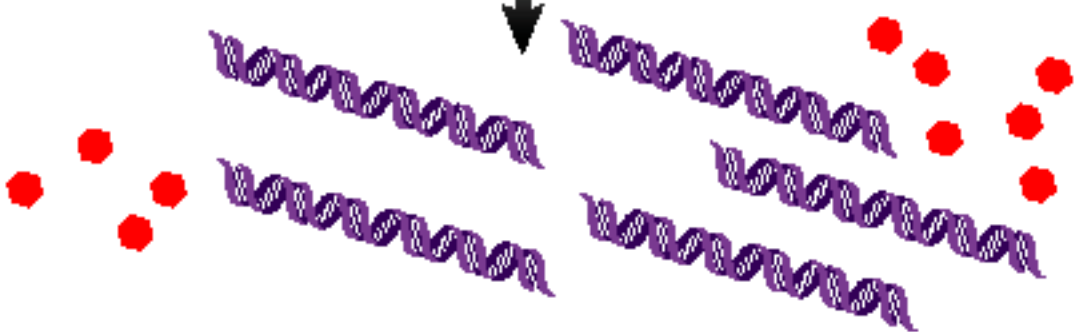
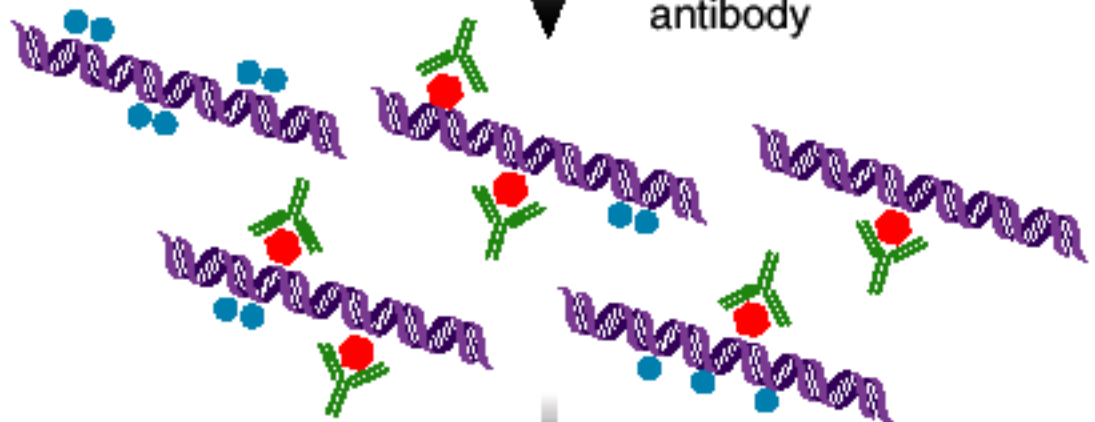
Contiguous reads

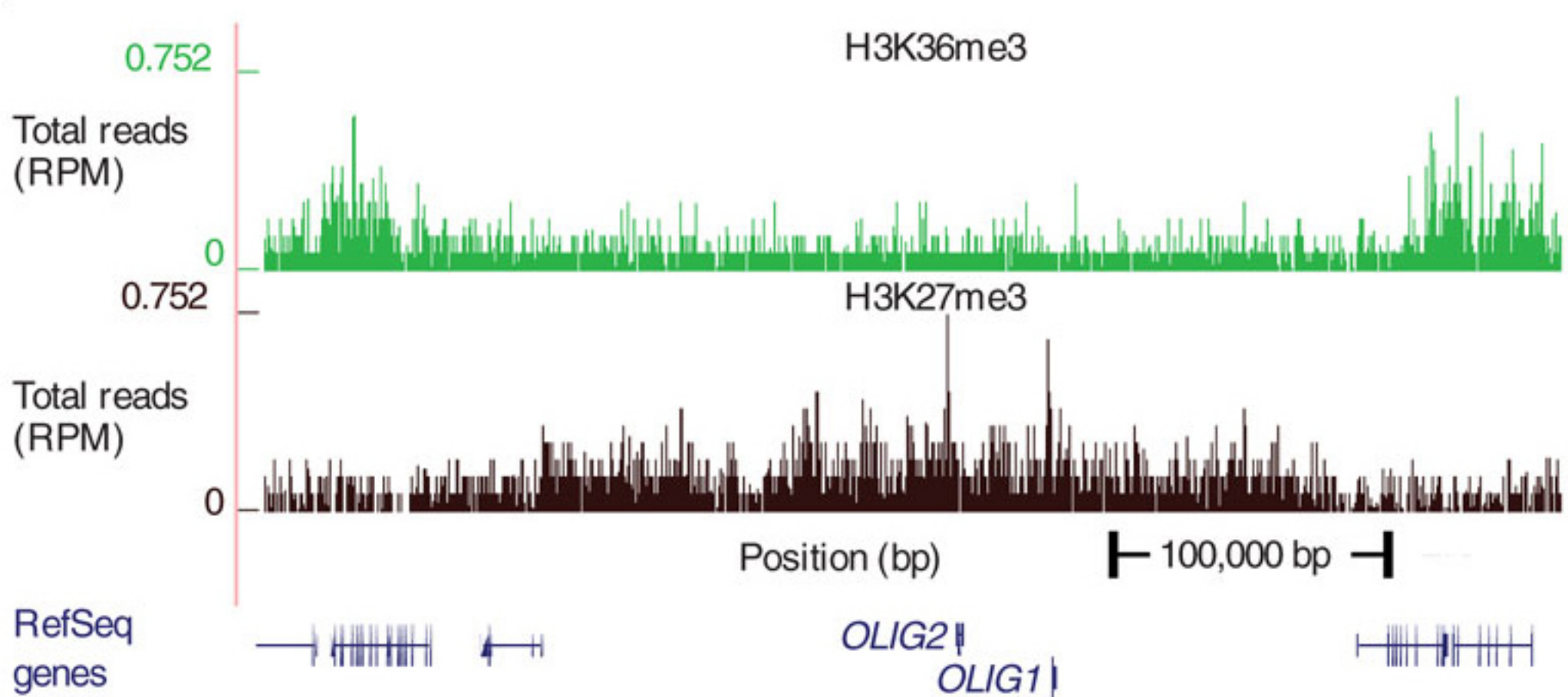


Isolate genomic DNA



Add protein-specific antibody





# RNA polymerase II

Watson (+) reads  
minus Crick (-)  
reads (RPM)

10.63

-7.24

16.9

Total reads  
(RPM)

0

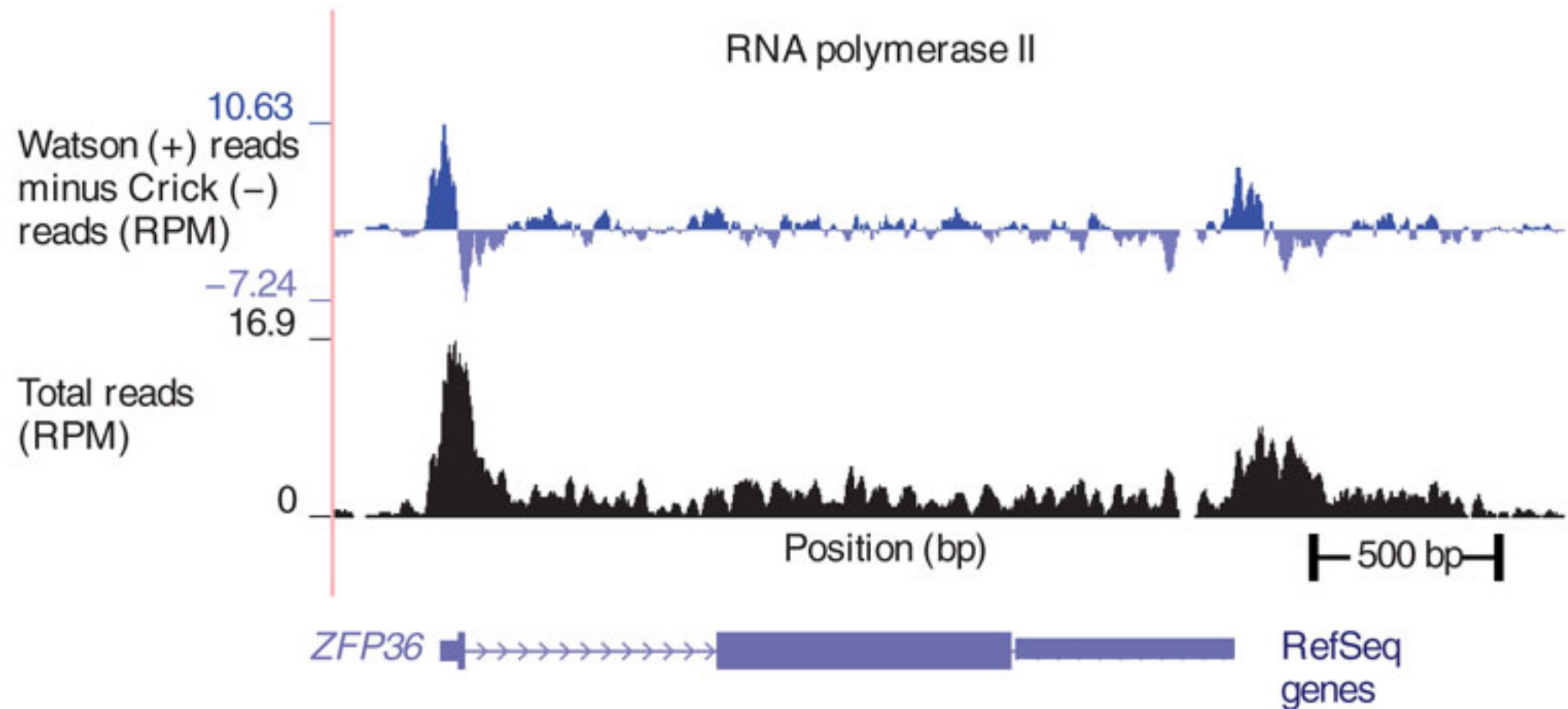
Position (bp)

500 bp

*ZFP36*



RefSeq  
genes



Watson (+) reads  
minus Crick (-)  
reads (RPM)

5.4  
0  
-5.3766  
8.3653

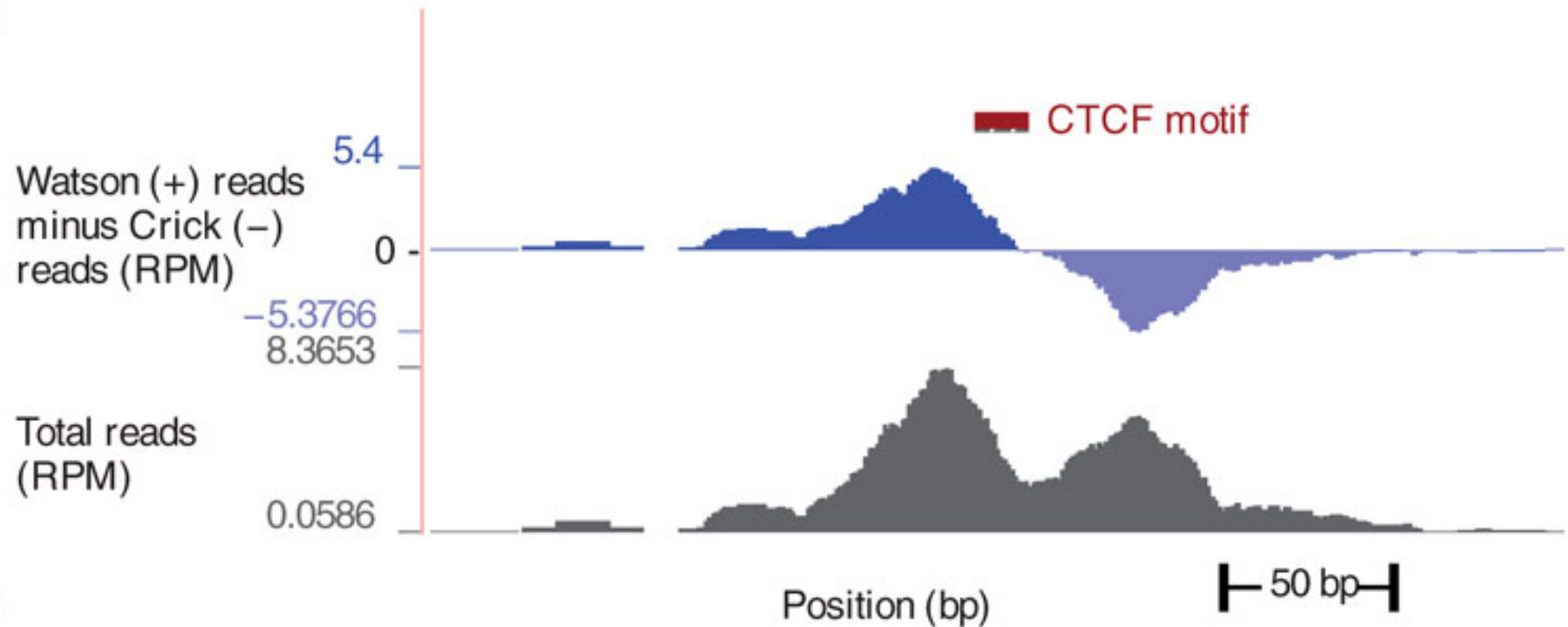
CTCF motif

Total reads  
(RPM)

0.0586

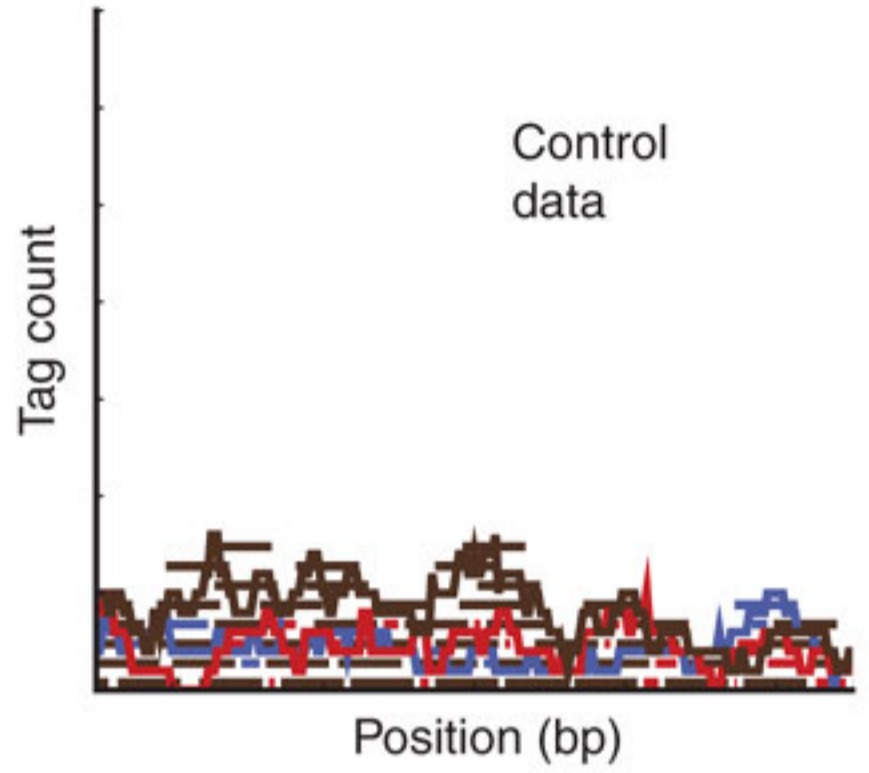
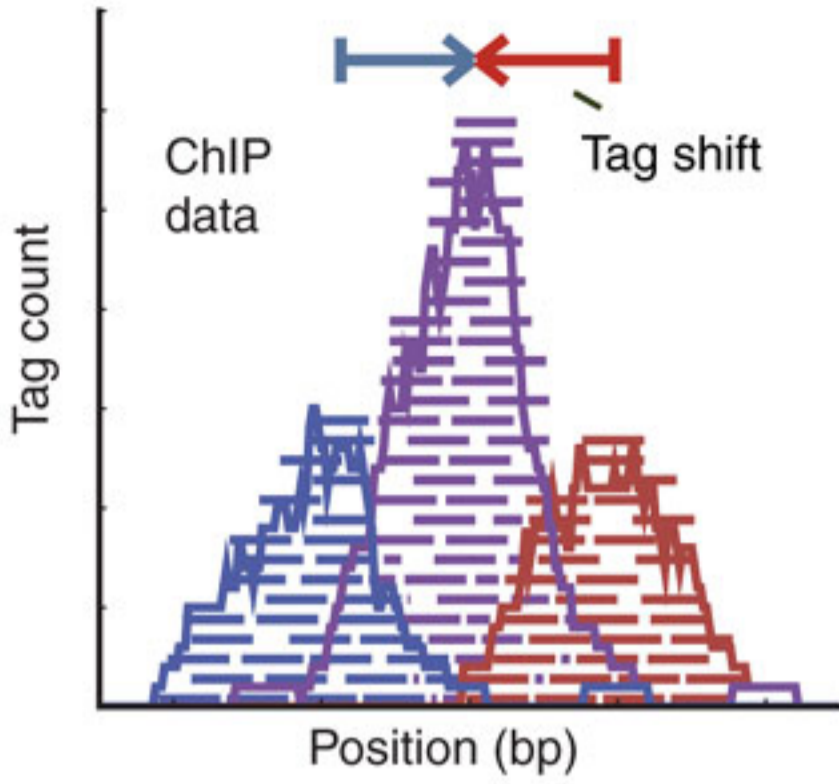
Position (bp)

50 bp

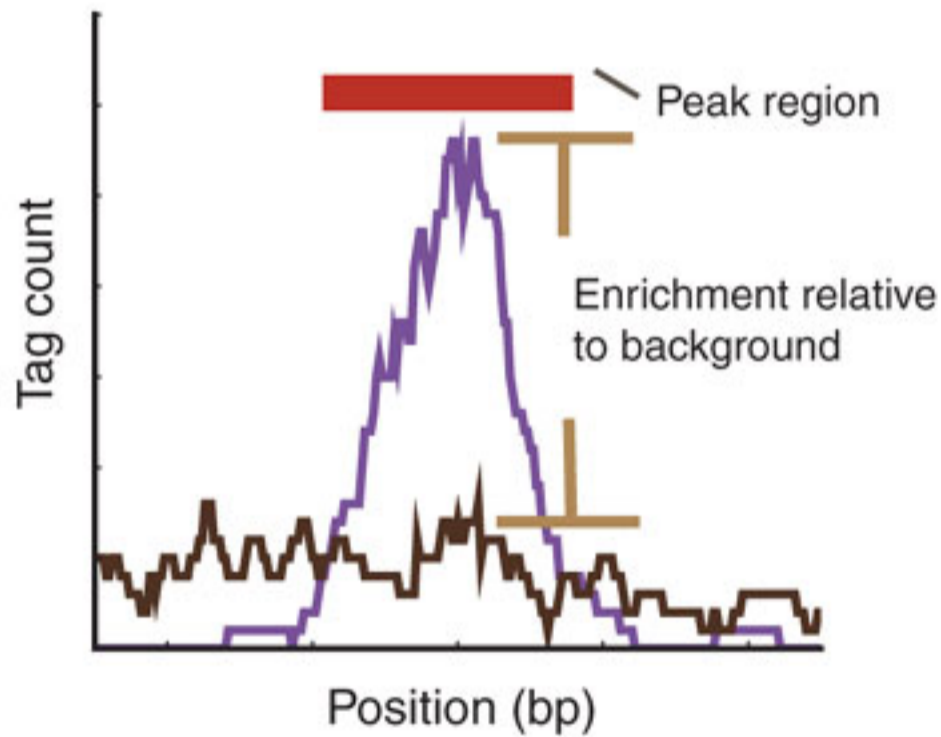


Generate signal profile along each chromosome

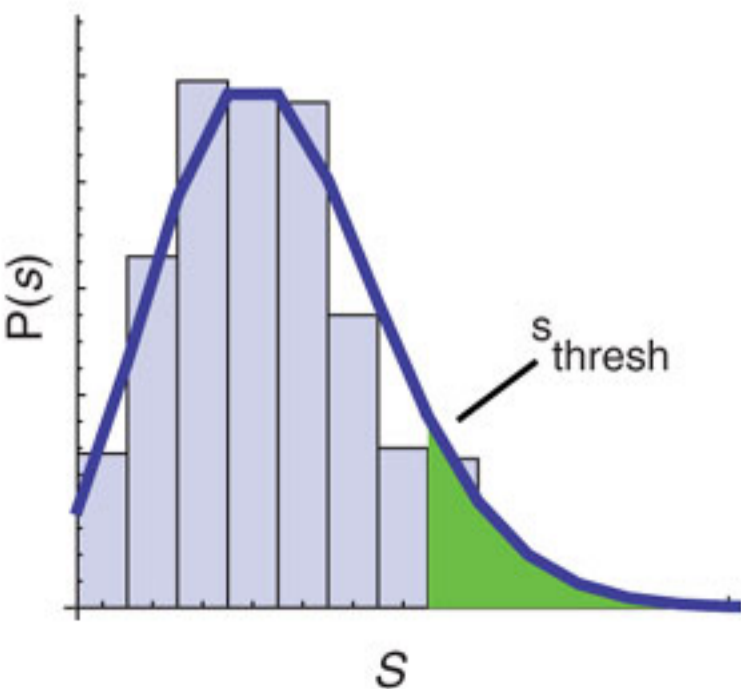
Define background (model or data)



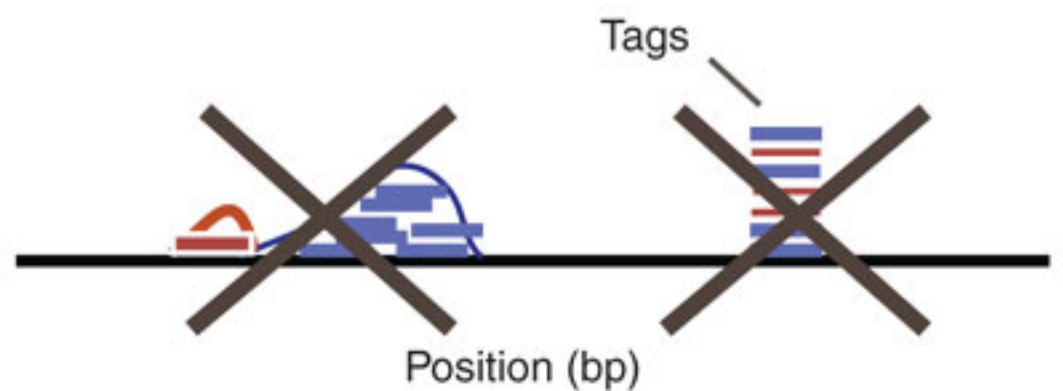
Identify peaks in ChIP signal

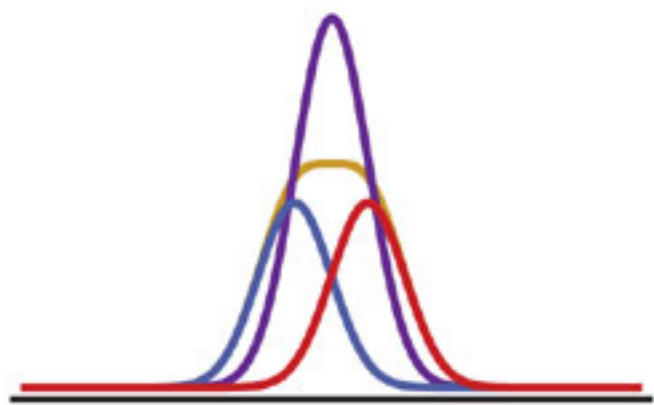
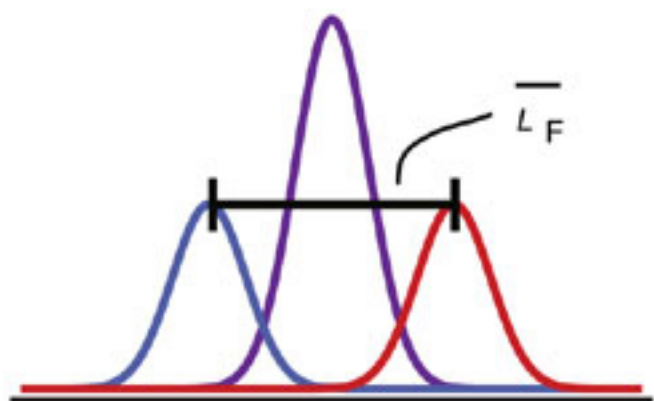


Assess significance

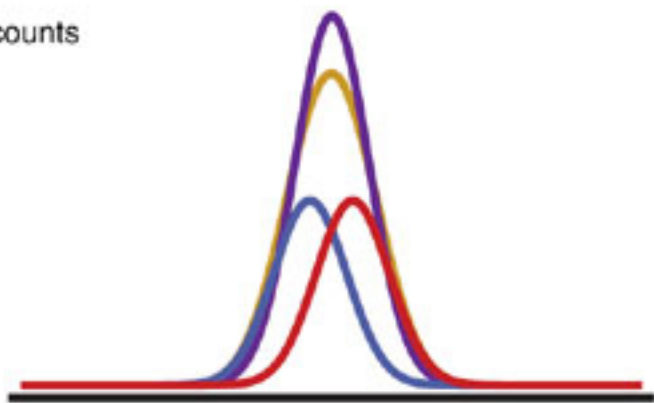


Filter artifacts



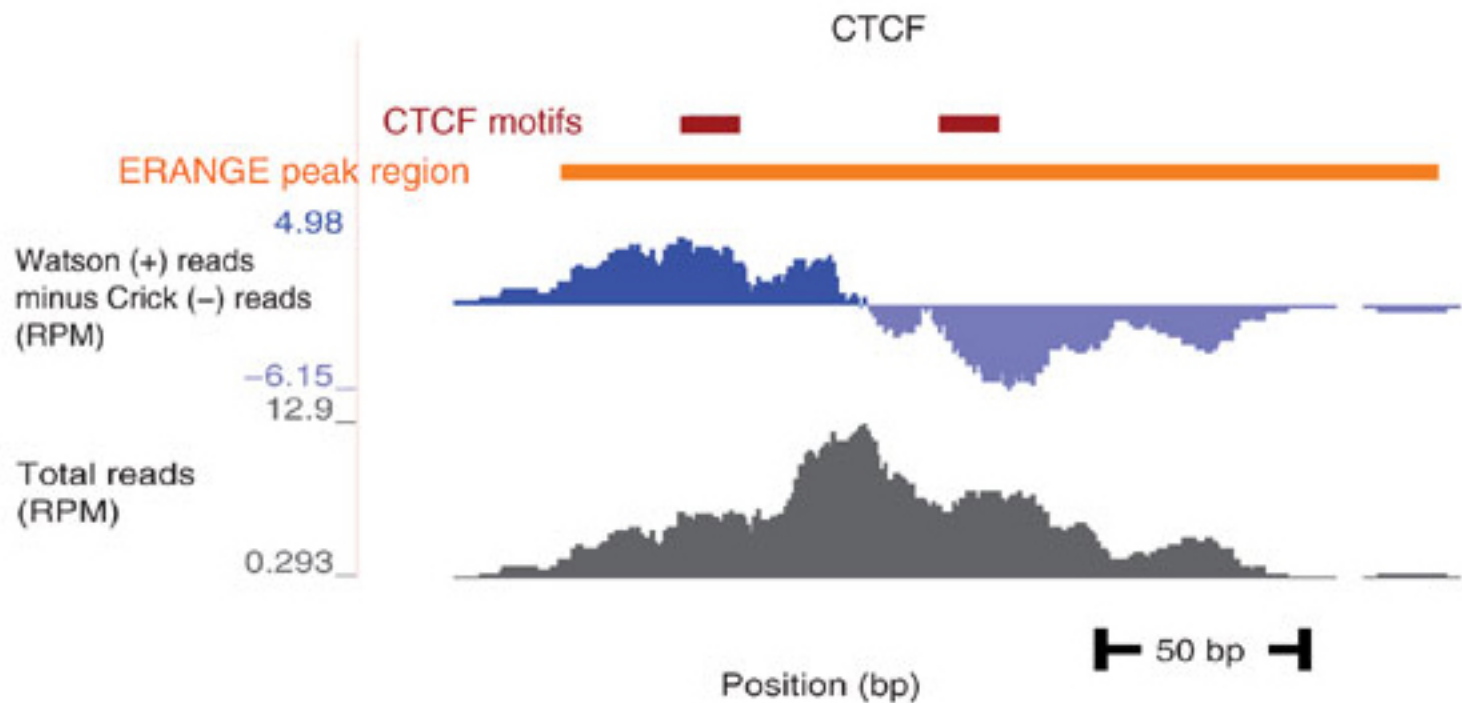


Tag counts



Position (bp)

- █ Forward strand tags
- █ Reverse strand tags
- █  $f + r$
- █  $f + r$ , shifted by  $\overline{L}_F / 2$





# \* reduced representation technologies

targeted/capture resequencing

ChIP-Seq

*chromatin immunoprecipitation*

ccc- or 3C-Seq

*chromosome conformation capture*

RNA-IP-Seq

*immunoprecipitated RNA*

RNA-Seq

*sequencing the transcriptome directly*

deepSAGE

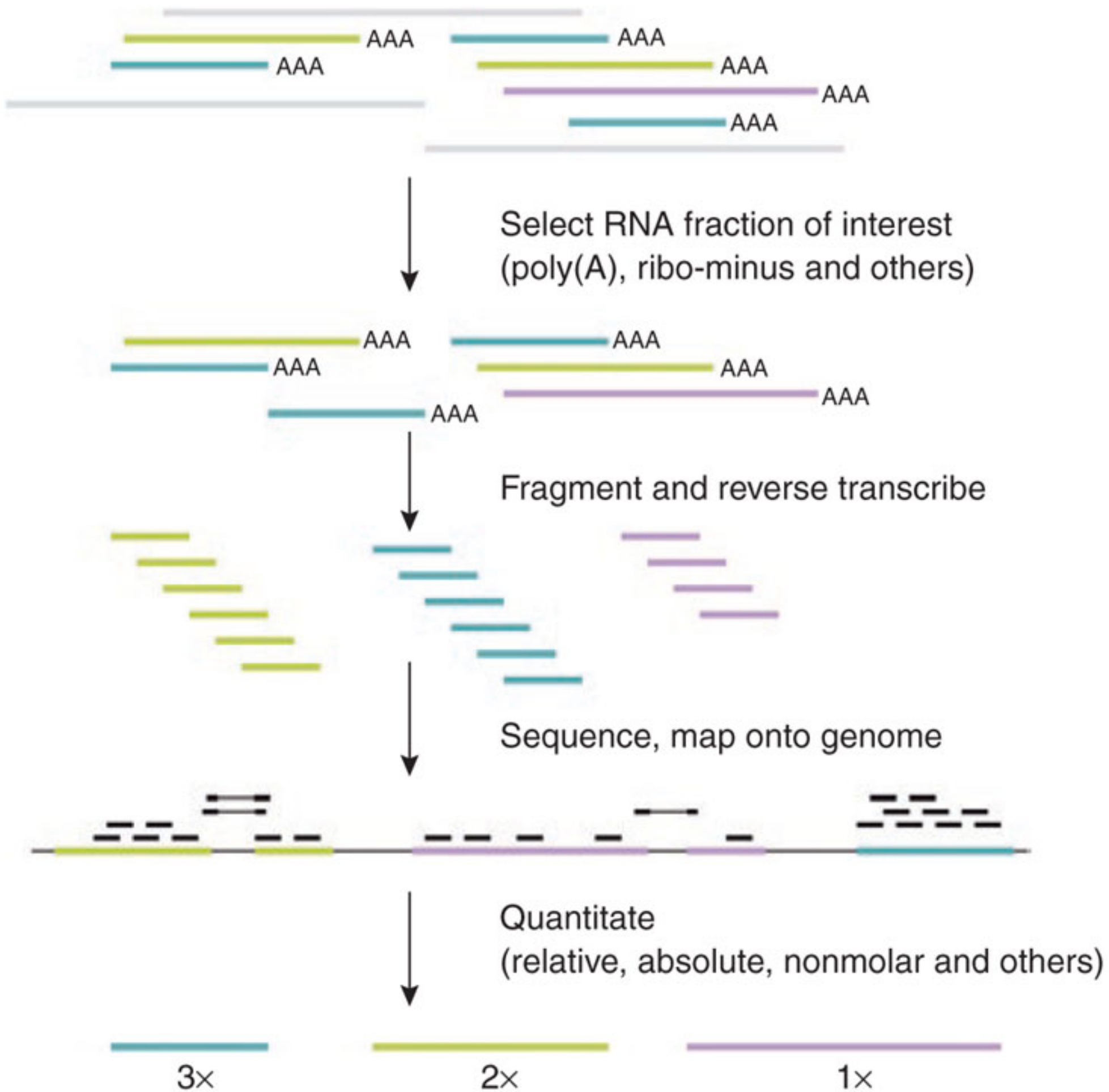
*serial analysis of gene expression using NG*

RAD-Seq

*restriction-site associated DNA*

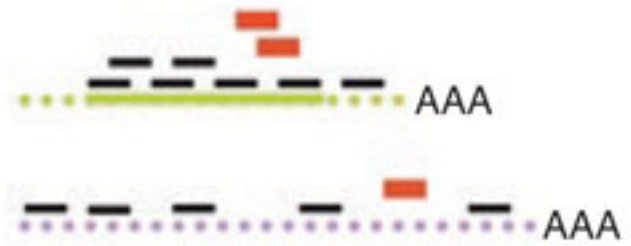
MeDIP-Seq

*methylated DNA*



**a***De novo* assembly of the transcriptome

Highly expressed gene  
 Lowly expressed gene



Read coverage must be high enough to build EST contigs (solid bar)

**b**

## Map onto the genome



Read mapper must support splitting reads to record splices

**c**

## Map onto the genome and splice junctions



Splice junction sequences from either annotations or inferred

