Next-generation sequencing for epigenetics studies

Jung Kyoon Choi, PhD

Assistant Professor, KAIST, Korea
Senior Research Scientist, Genome Institute of Singapore

Overview

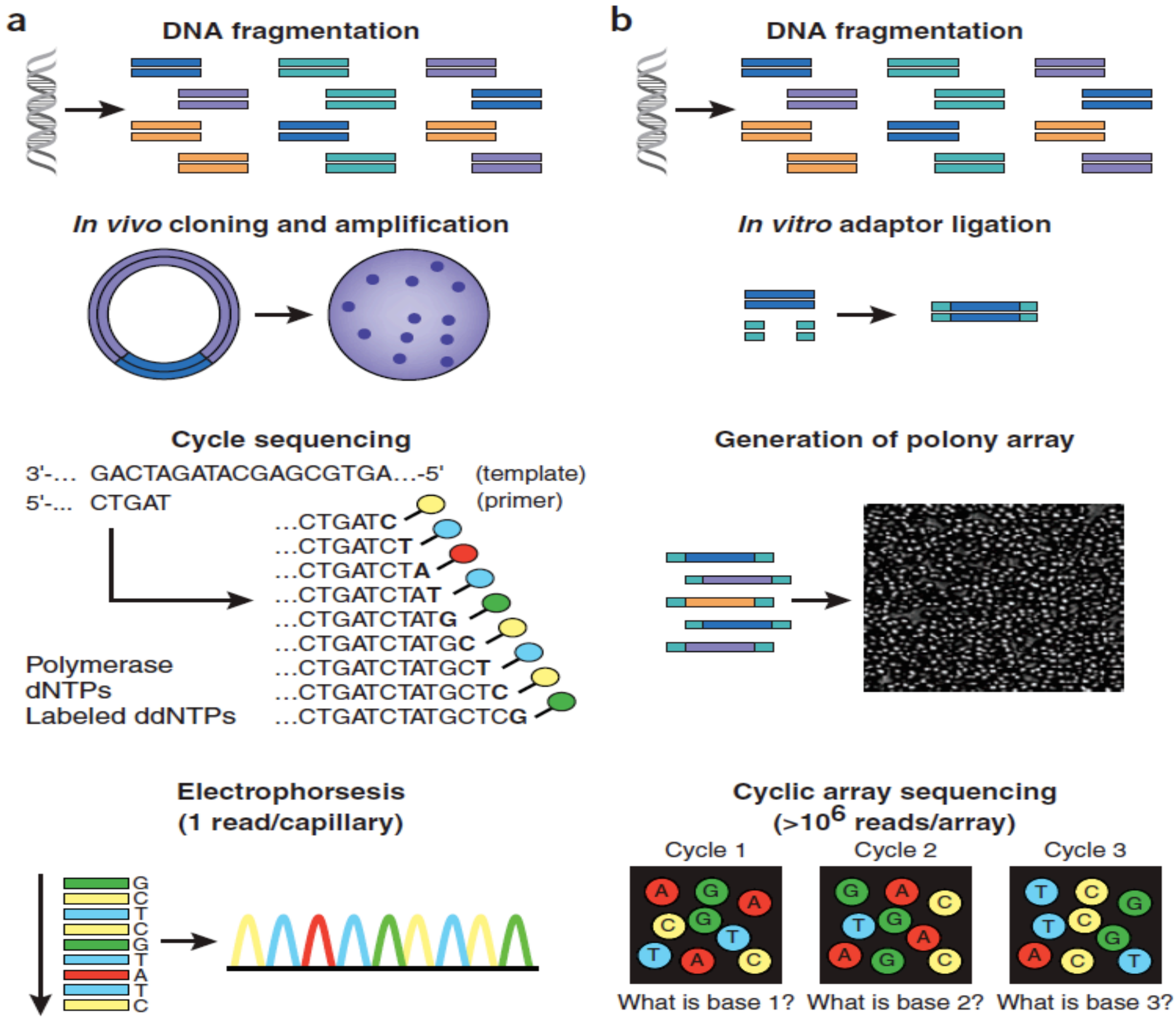
- Next-generation sequencing (NGS)
- What is epigenetics?
- Experimental techniques for epigenomics
- Data analysis and visualization

Overview

- Next-generation sequencing (NGS)
- What is epigenetics?
- Experimental techniques for epigenomics
- Data analysis and visualization

- Sanger sequencing
  - long read (~500 bp)
  - self-assembly by overlapping
  - *de novo* sequencing

- Next-generation sequencing
  - high throughput (>1 Gb per run)
  - mapping to a reference genome
  - resequencing : variant detection or counting

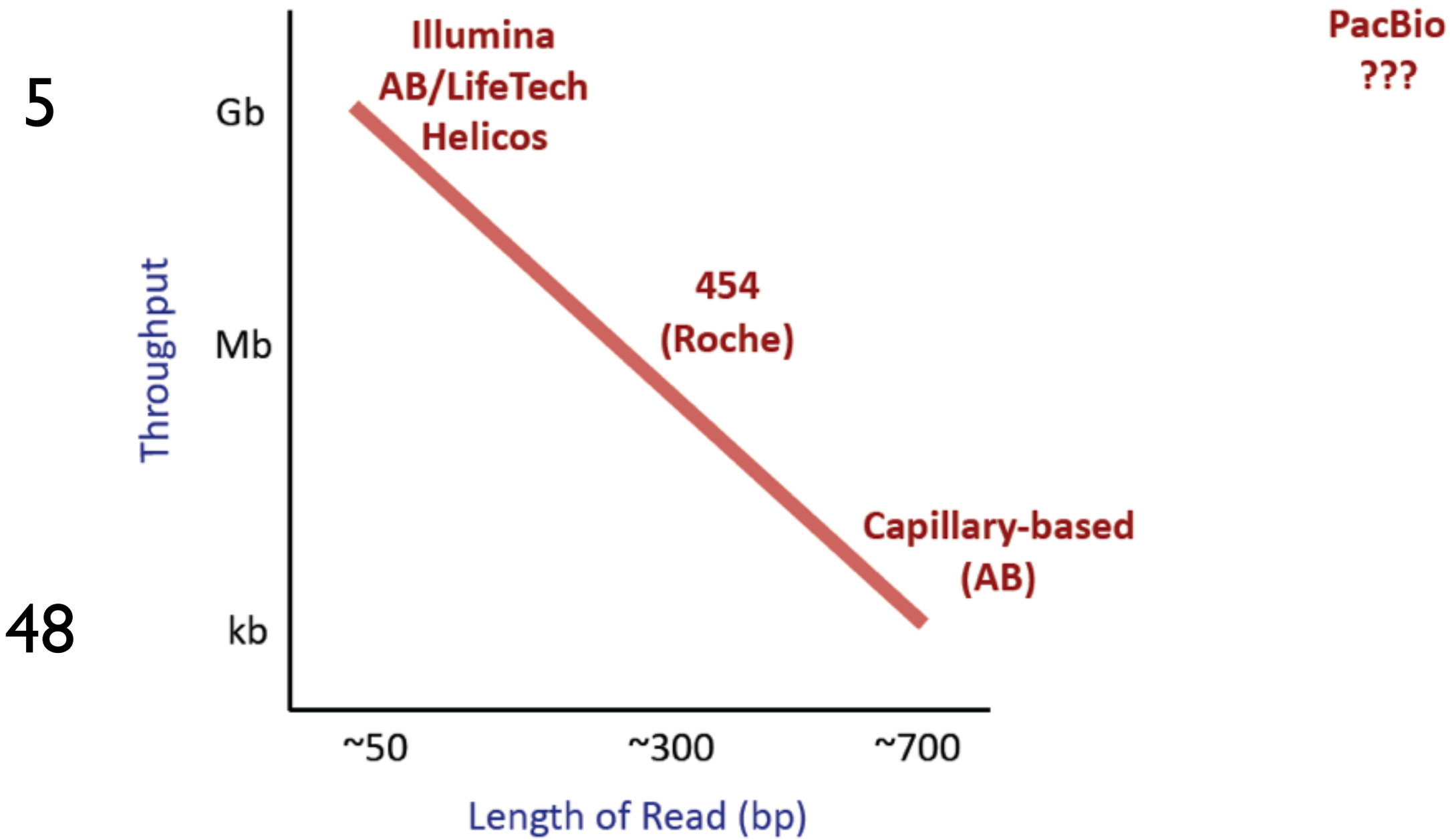

a
DNA fragmentation
In vivo cloning and amplification
Cycle sequencing
3'-... GACTAGATACGAGCGTGA...-5' (template)
5'-... CTGAT (primer)
...CTGATC
...CTGATCT
...CTGATCTA
...CTGATCTAT
...CTGATCTATG
...CTGATCTATGC
...CTGATCTATGCT
...CTGATCTATGCTC
...CTGATCTATGCTCG
Polymerase
dNTPs
Labeled ddNTPs
Electrophorsesis
(1 read/capillary)
G
C
T
C
G
T
A
T
C
b
DNA fragmentation
In vitro adaptor ligation
Generation of polony array
Cyclic array sequencing
(>10^6 reads/array)
Cycle 1
Cycle 2
Cycle 3
What is base 1?
What is base 2?
What is base 3?

Synonyms

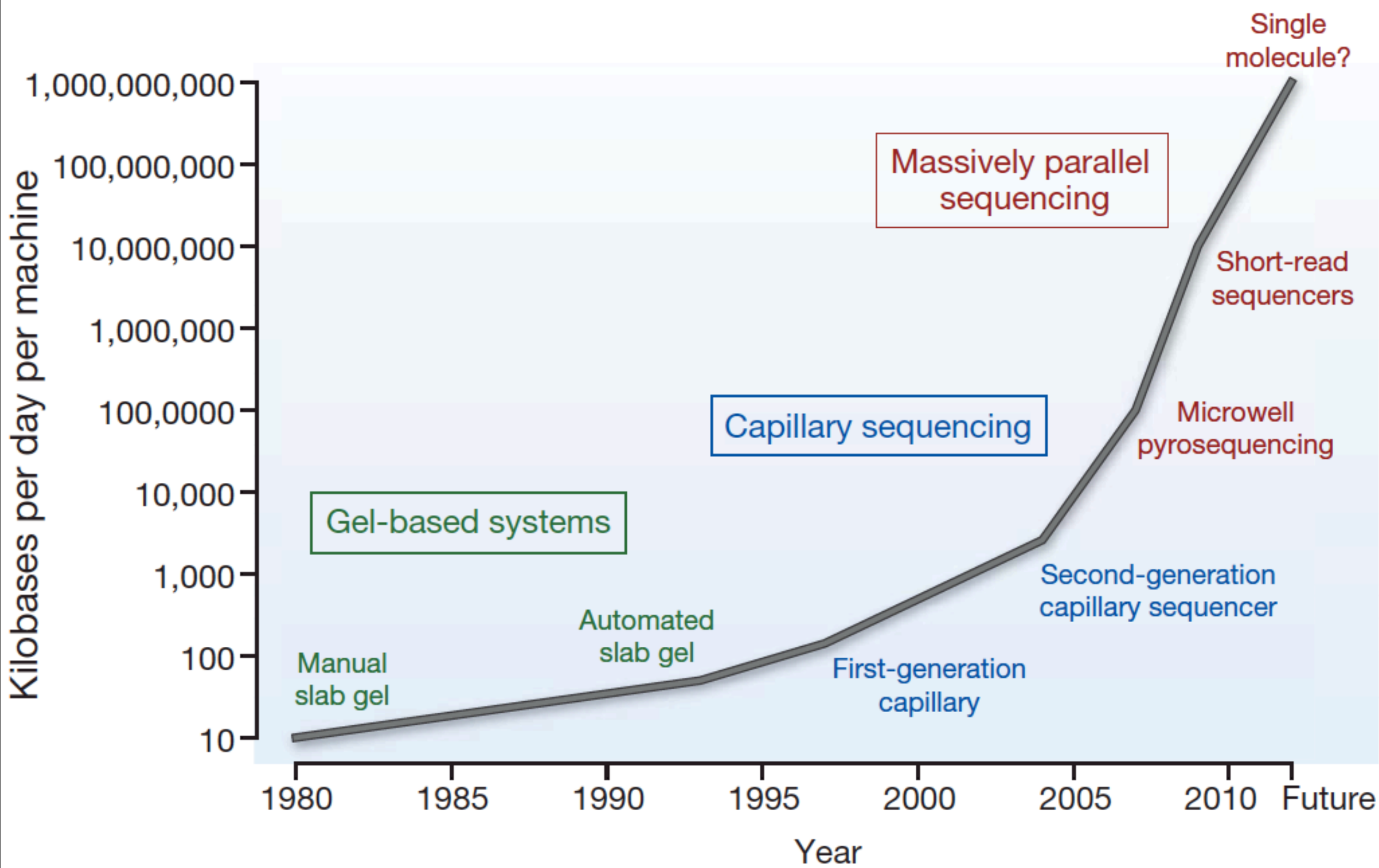
- NGS
  - Next-generation sequencing
  - New-generation sequencing
- Massively parallel sequencing
- Deep sequencing

Vendor	Chemistry	Machine
Illumina	Solexa™	Genome Analyzer HiSeq2000
Applied Biosystems	SOLiD™	SOLiD system
Roche (454 Life Sciences)	pyrosequencing	GS FLX 454
Helicos BioSciences	tSMS™	HeliScope
Pacific Biosciences	SMRT™	PacBio RS

Trade-offs with Newer Sequencing Technologies

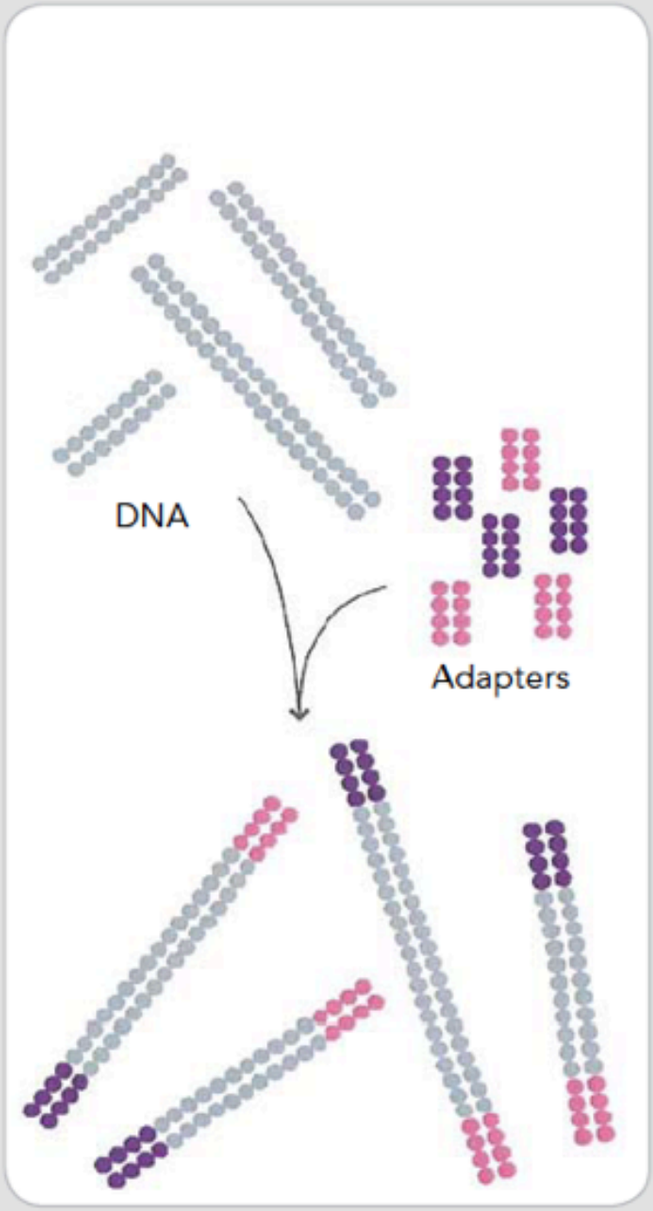


$$\text{Throughput} = \frac{\text{Amount of Sequence Generated}}{\text{Unit of Time or Cost}}$$


Single molecule?
Massively parallel sequencing
Short-read sequencers
Microwell pyrosequencing
Capillary sequencing
Gel-based systems
Second-generation capillary sequencer
Automated slab gel
Manual slab gel
First-generation capillary
1,000,000,000
100,000,000
10,000,000
1,000,000
100,0000
10,000
1,000
100
10
Kilobases per day per machine
1980
1985
1990
1995
2000
2005
2010
Future
Year

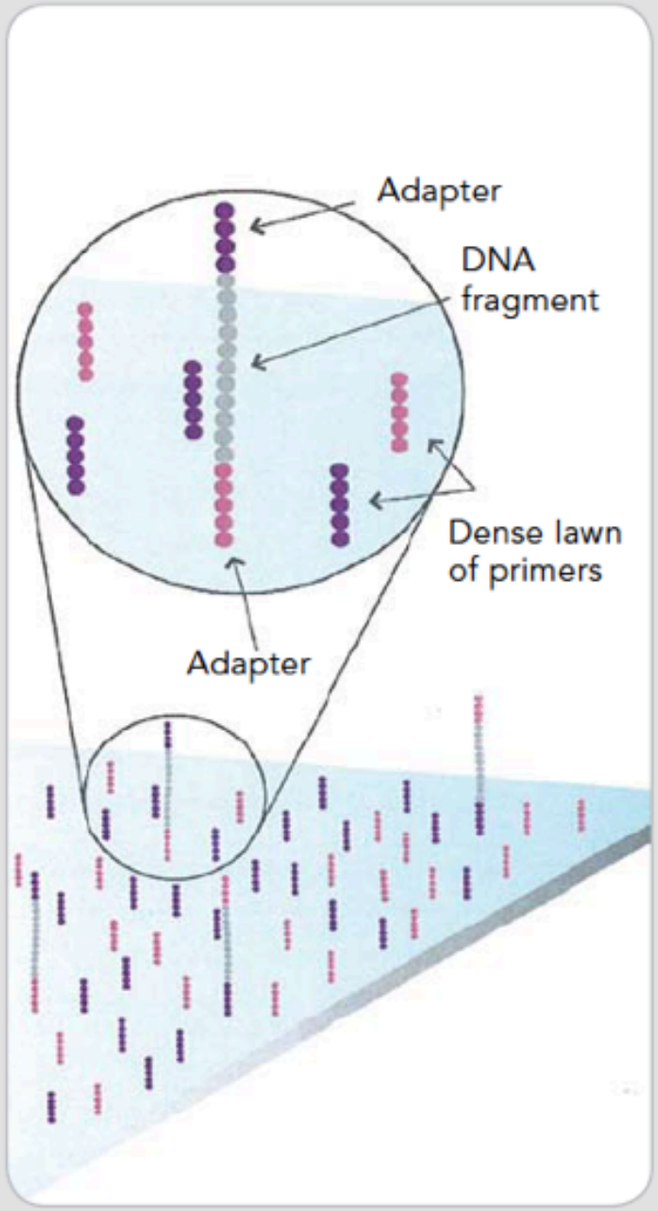
Illumina/Solexa Sequencing

1. PREPARE GENOMIC DNA SAMPLE



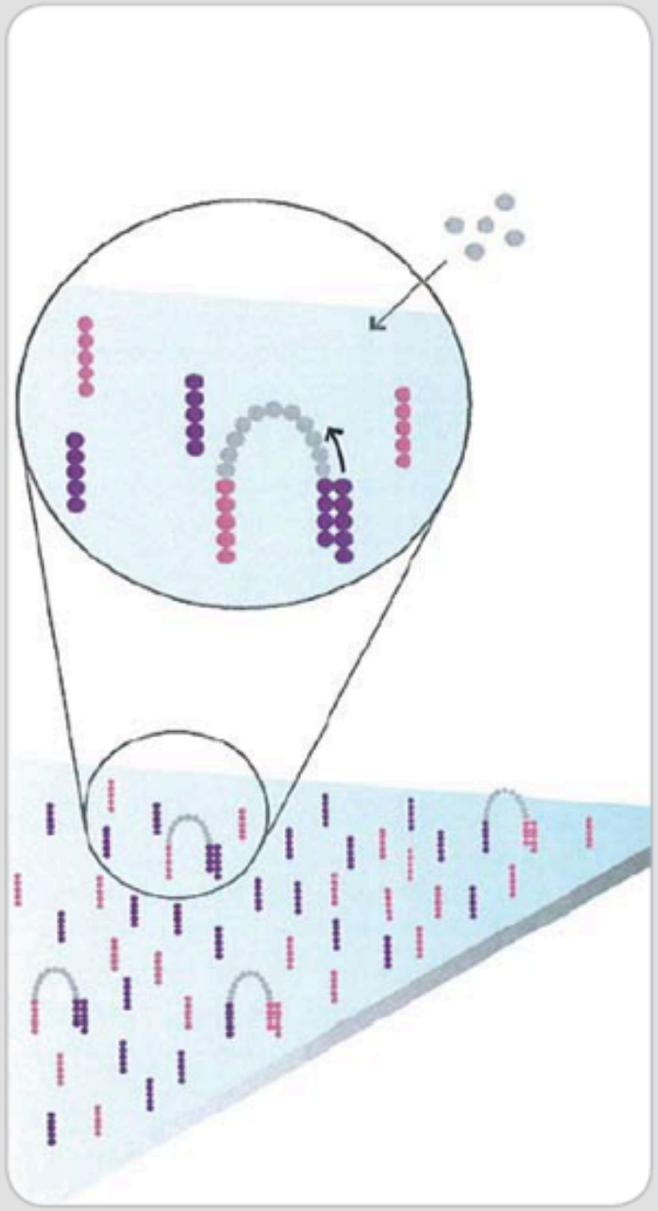
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

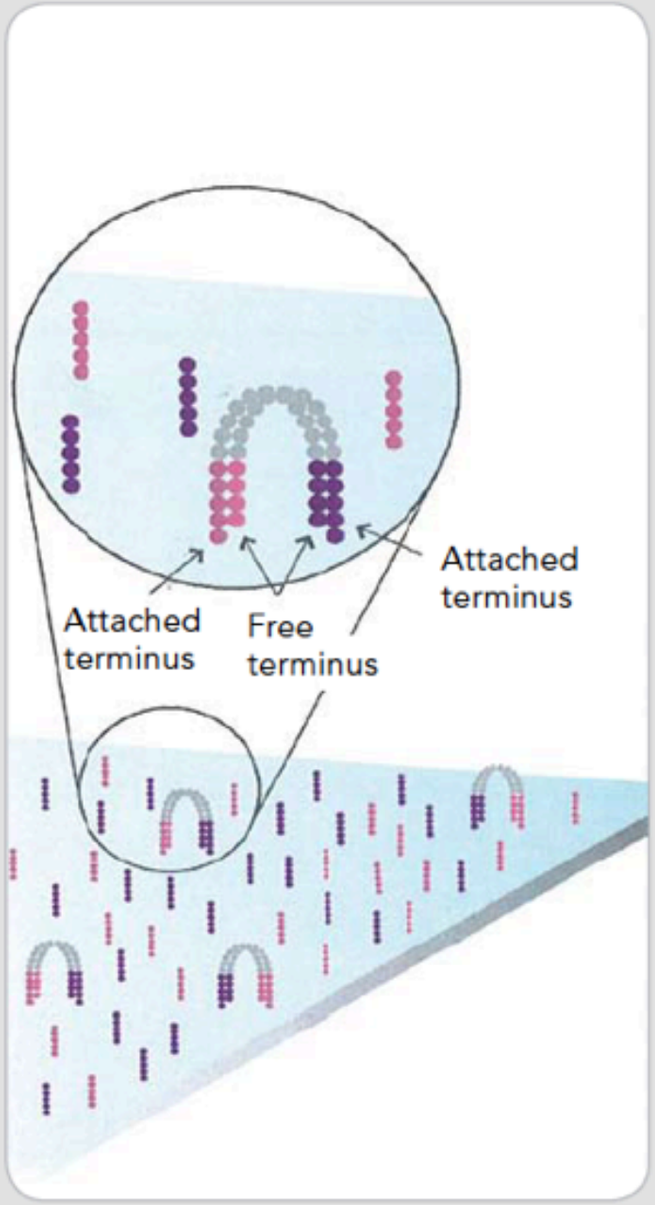
3. BRIDGE AMPLIFICATION



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

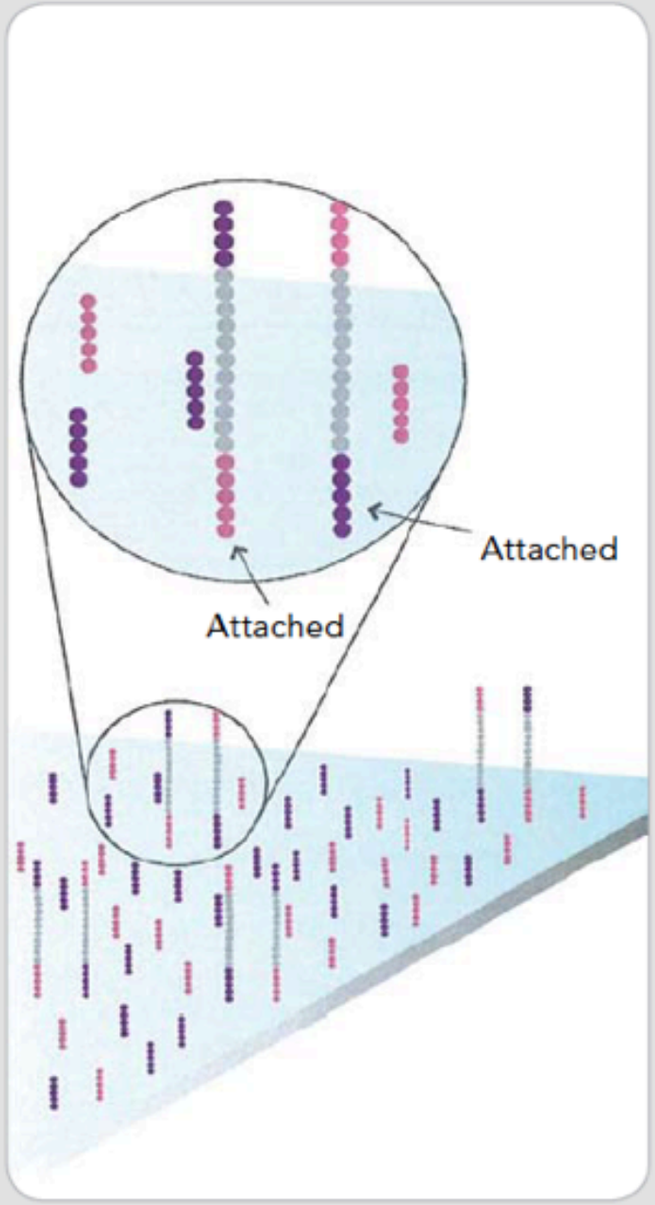
Illumina/Solexa Sequencing

4. FRAGMENTS BECOME DOUBLE-STRANDED



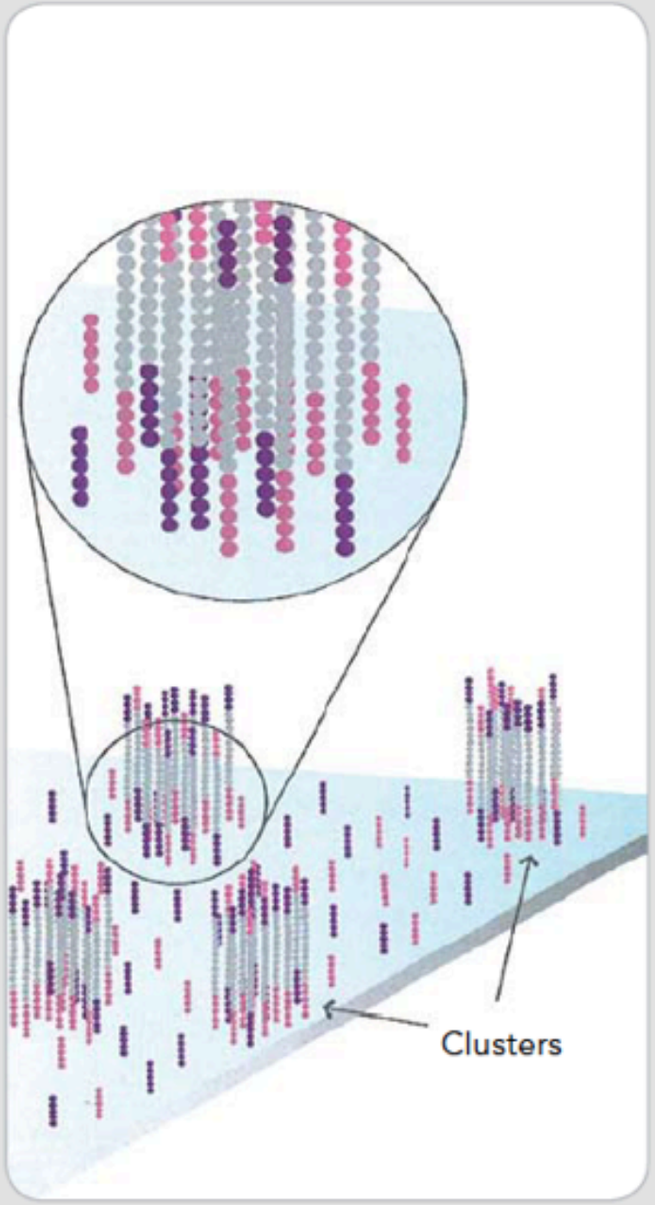
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



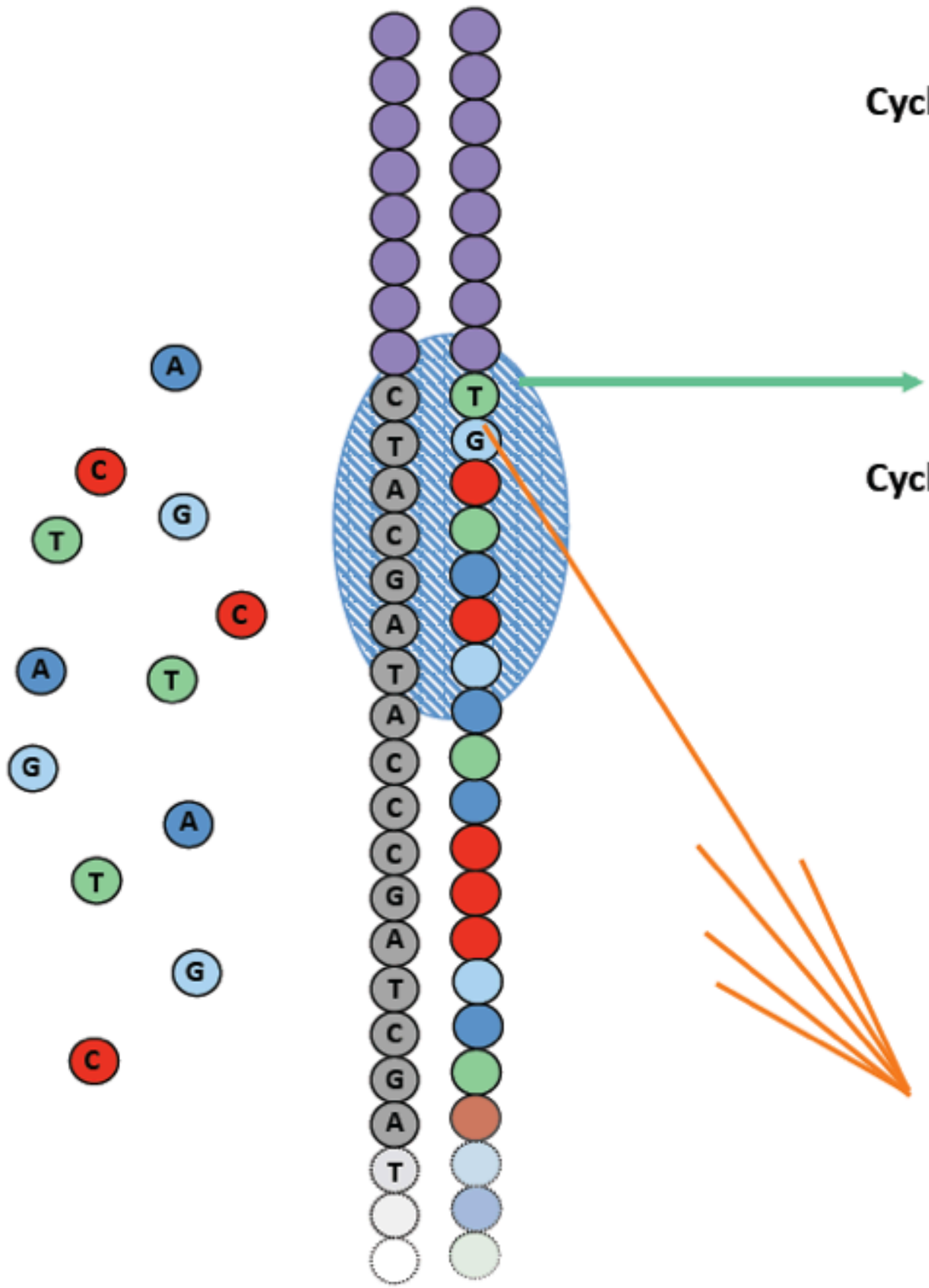
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Sequencing By Synthesis (SBS)



Cycle 1: Add sequencing reagents

First base incorporated

Remove unincorporated bases

Detect signal

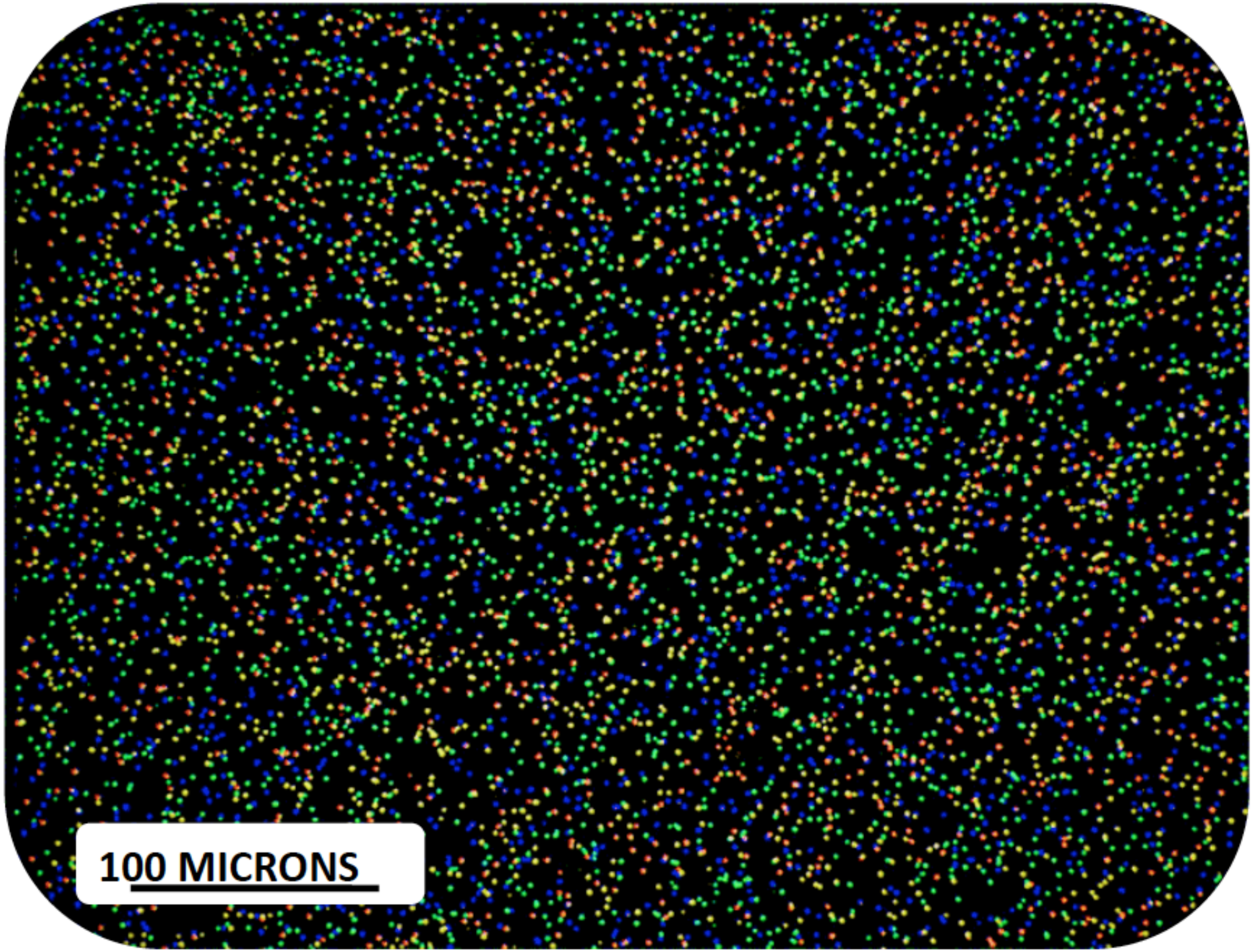
Cycle 2-n: Add sequencing reagents and repeat

All four labeled nucleotides in one reaction

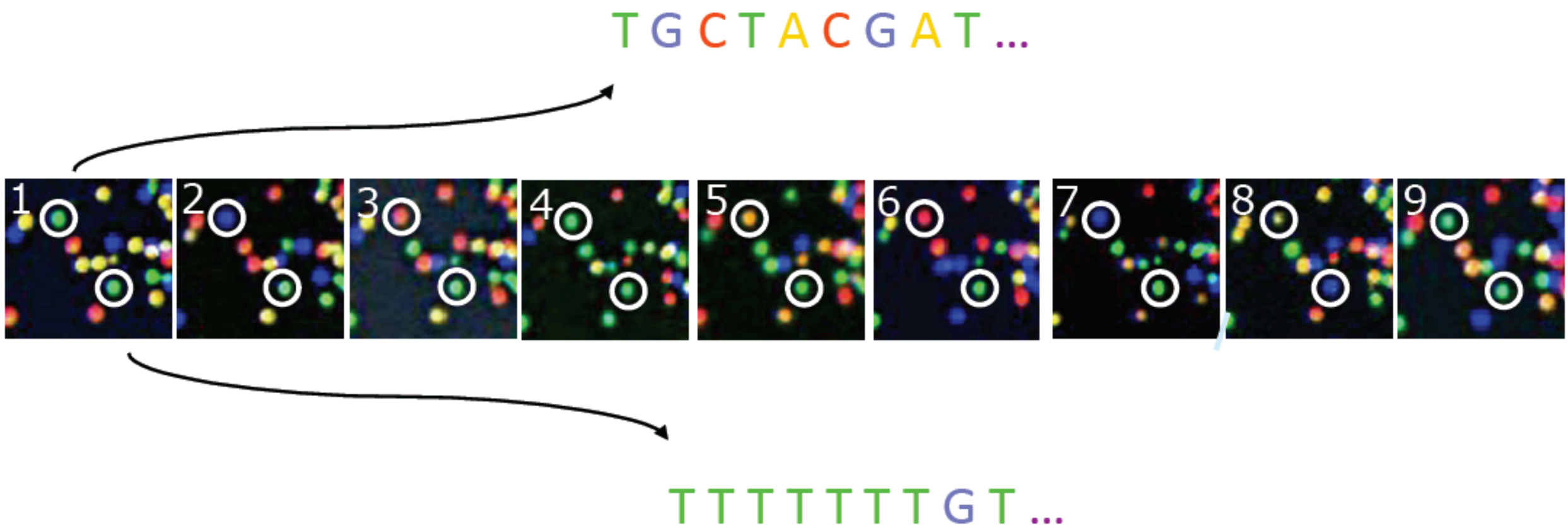
Base-by-base sequencing

No problems with homopolymer repeats

Pseudo-color Enhanced Image



Base Calling from Raw Data

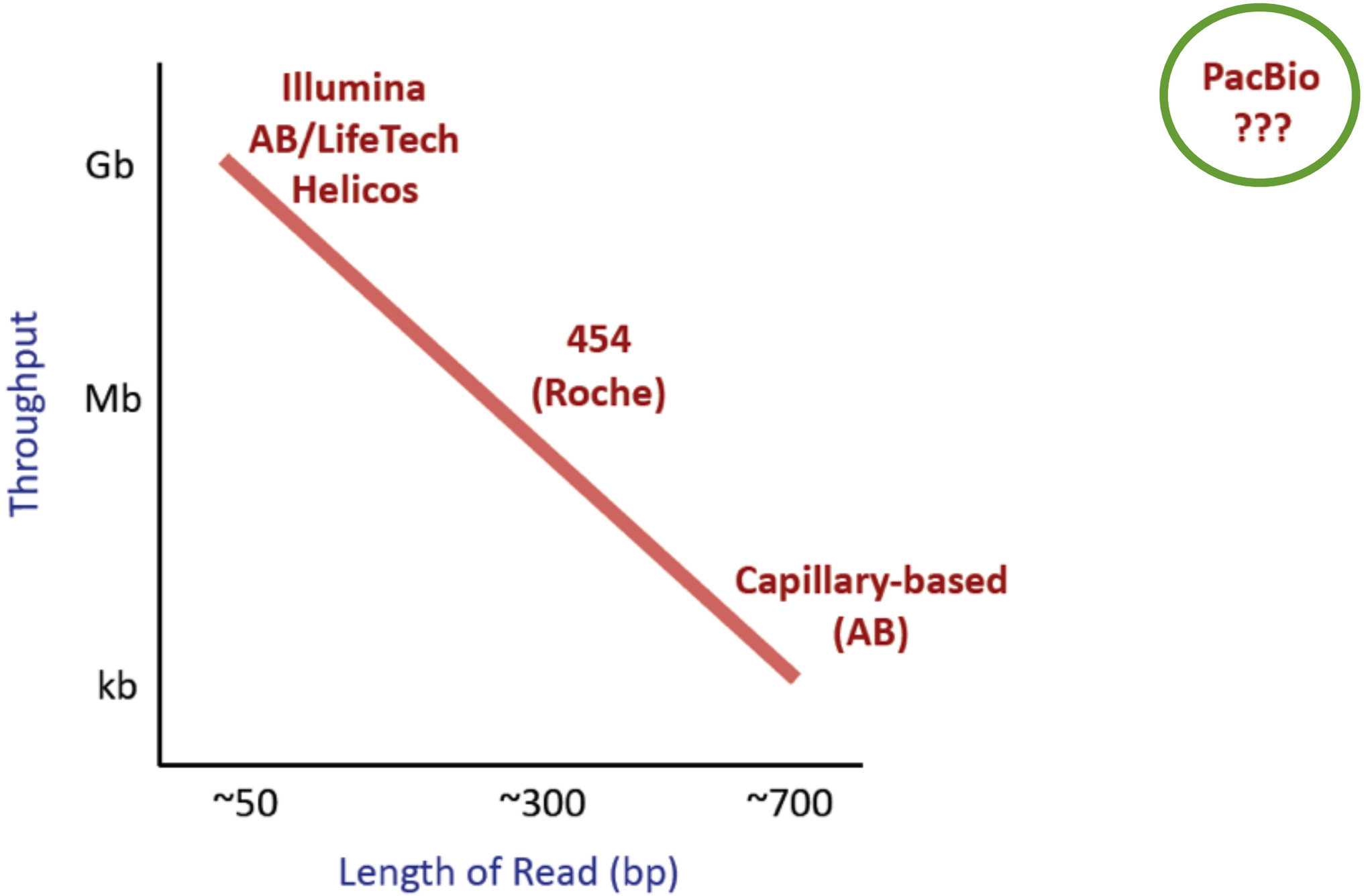


The identity of each base of a cluster is read off from sequential images.

HiSeq2000

- **Same chemistry**
- **Runs 2 flowcells at the same time**
  - Imaging one flowcell – chemistry on the other
- **Flowcells are bigger**
  - More surface area can be scanned
  - Focuses on top and bottom of flowcell
- **Improvements to hardware**
  - Better lasers, cameras, etc.

Trade-offs with Newer Sequencing Technologies



$$\text{Throughput} = \frac{\text{Amount of Sequence Generated}}{\text{Unit of Time or Cost}}$$

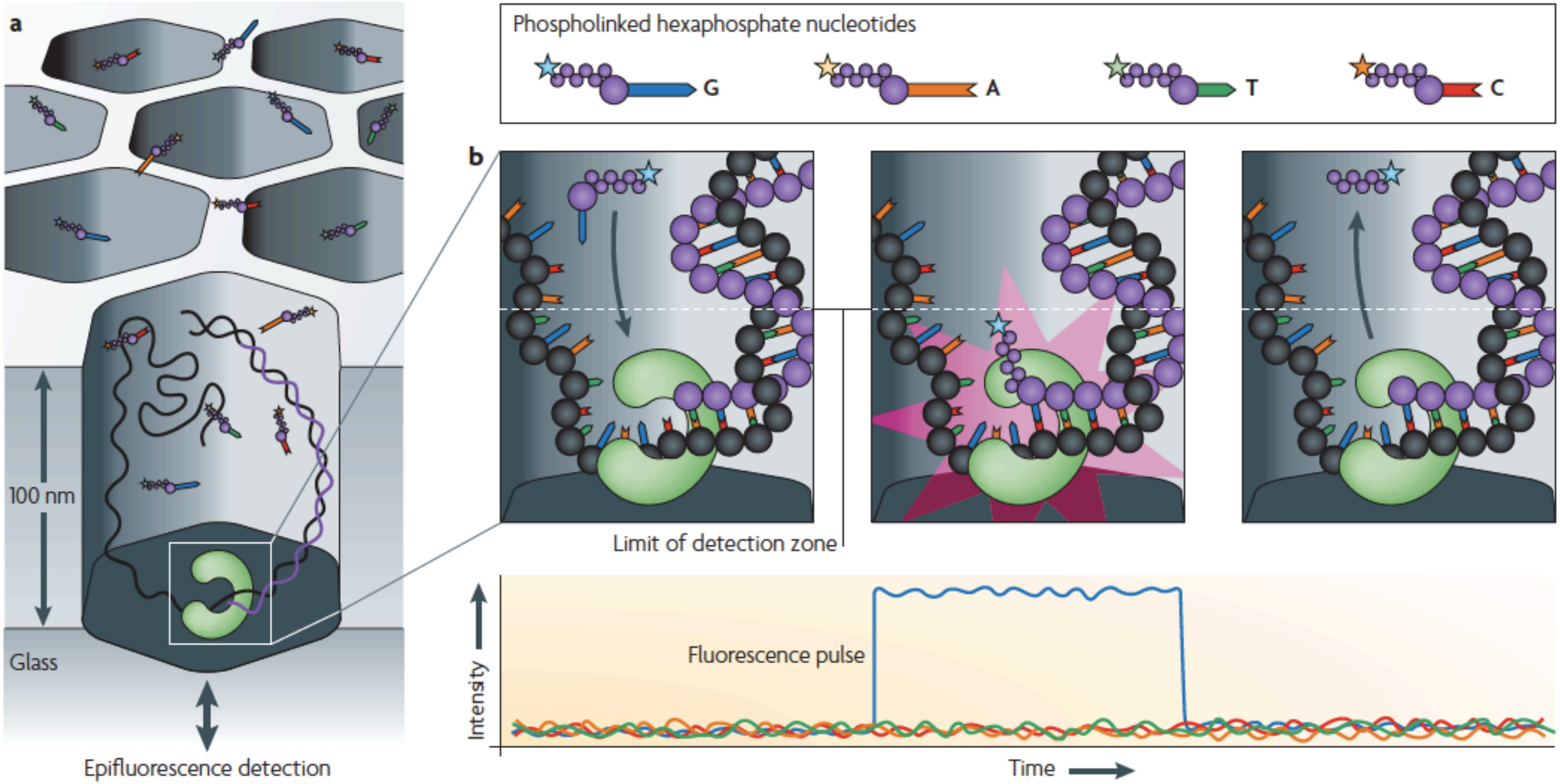
SMRT™

PACBIO *RS*

Pacific Biosciences is proud to introduce the revolutionary third generation DNA sequencing system: the PacBio *RS*. Our system incorporates novel, single molecule sequencing techniques and advanced analytics to reveal more biology in real time. We call this SMRT™ (Single Molecule Real Time) technology.

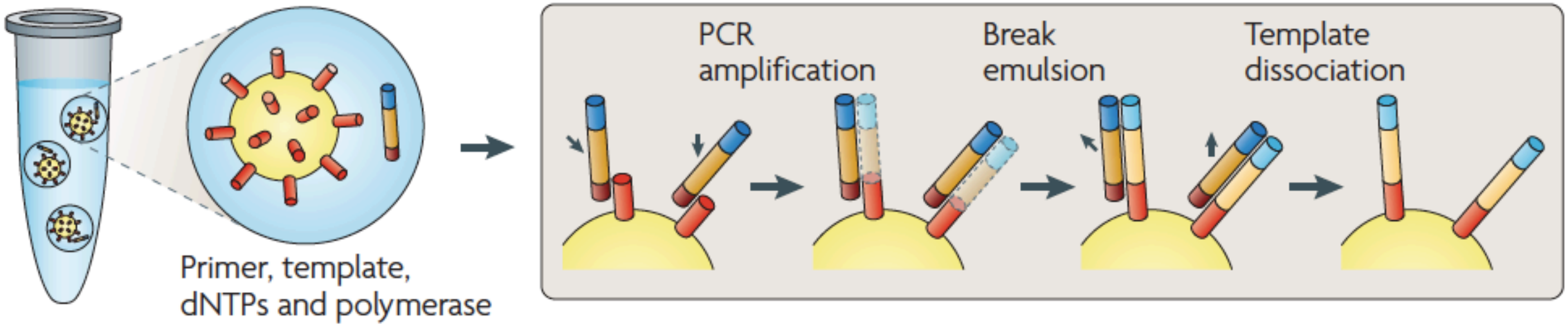
How does this technology enable longer reads?

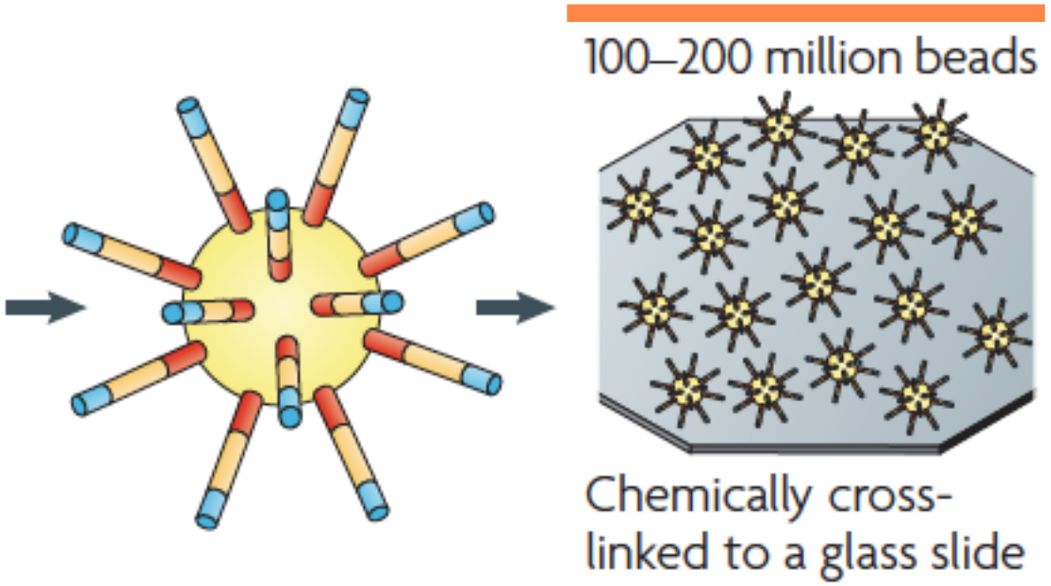
Pacific Biosciences — Real-time sequencing



signal chemistry is similar to pyrosequencing but
- no amplification
- four different fluorescent labels used instead of pyrophosphate detection, which enables RT signal detection

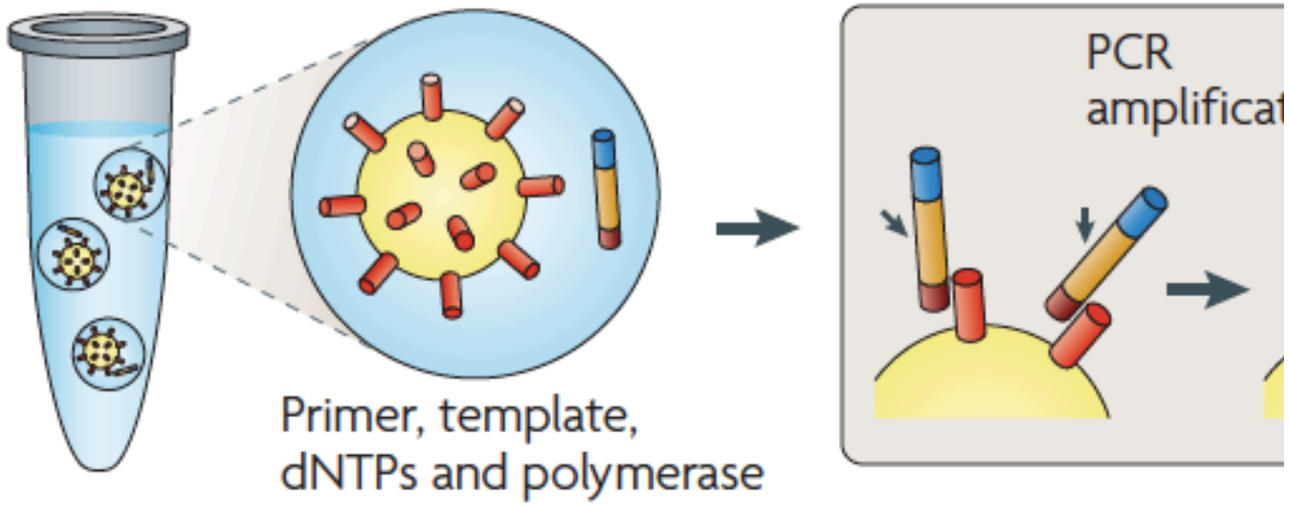
a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thousands of copies occurs in microreactors in an emulsion

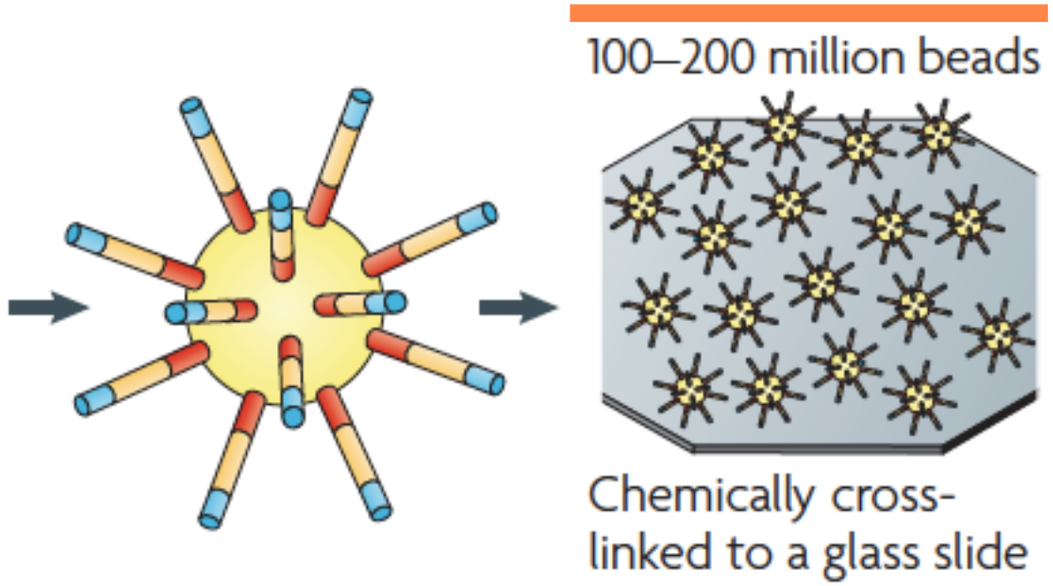




SOLiD

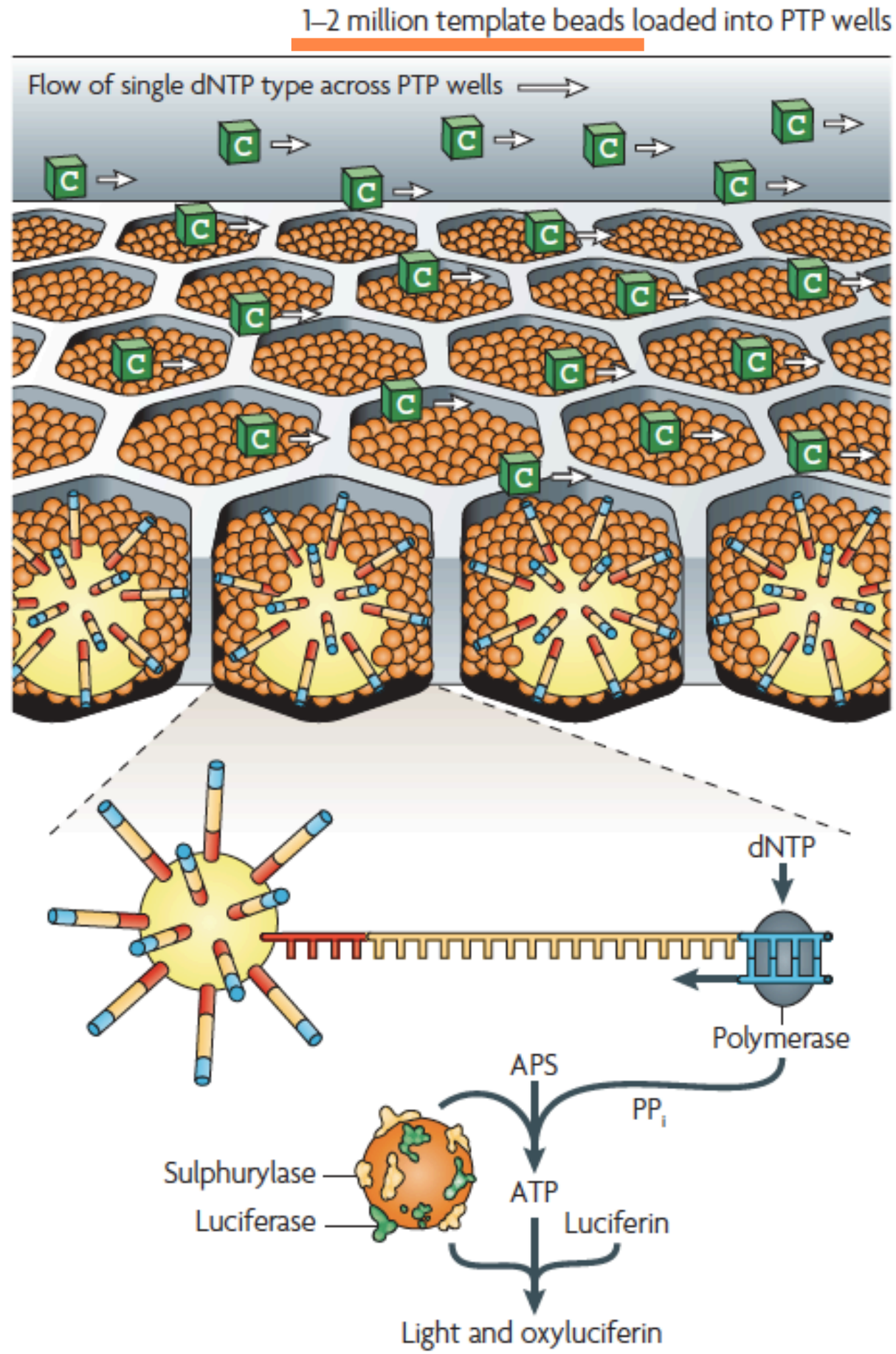
a Roche/454, Life/APG, Polonator
Emulsion PCR
One DNA molecule per bead. Clonal amplification to thou



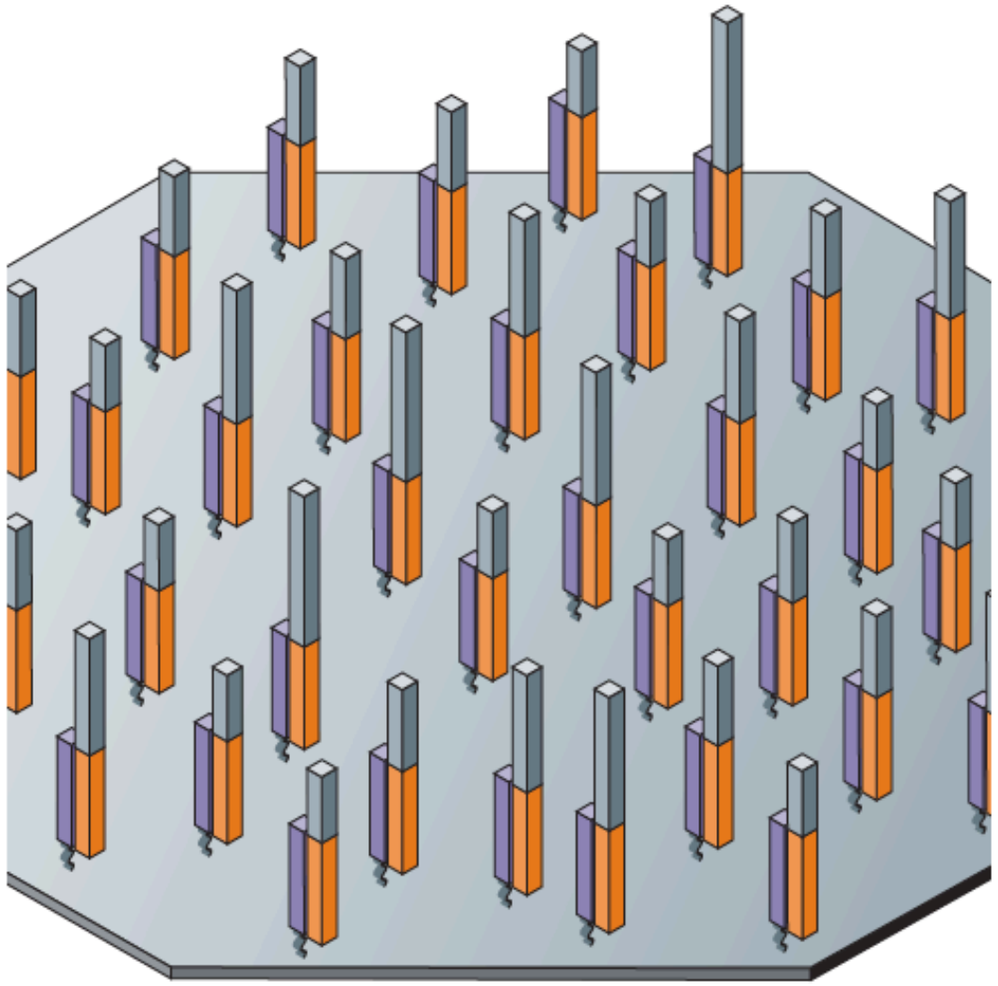


SOLiD

Roche/454 — Pyrosequencing

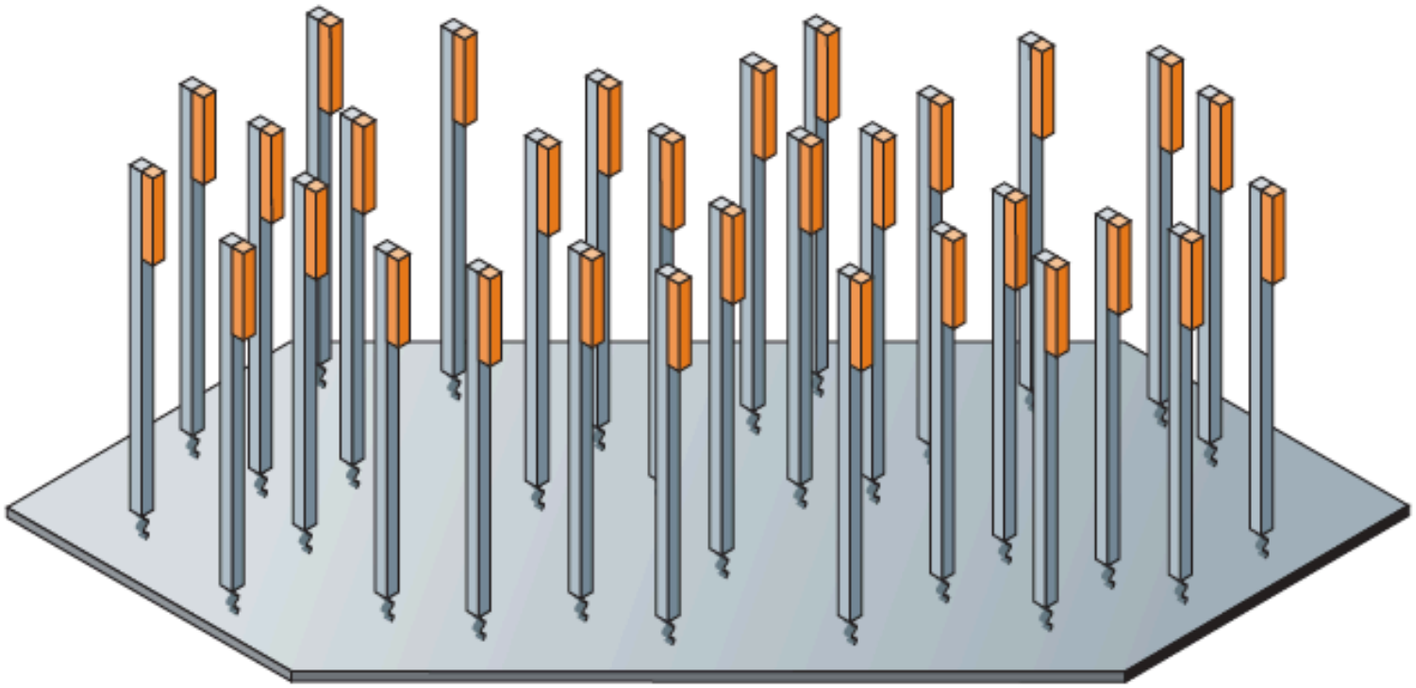


c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized



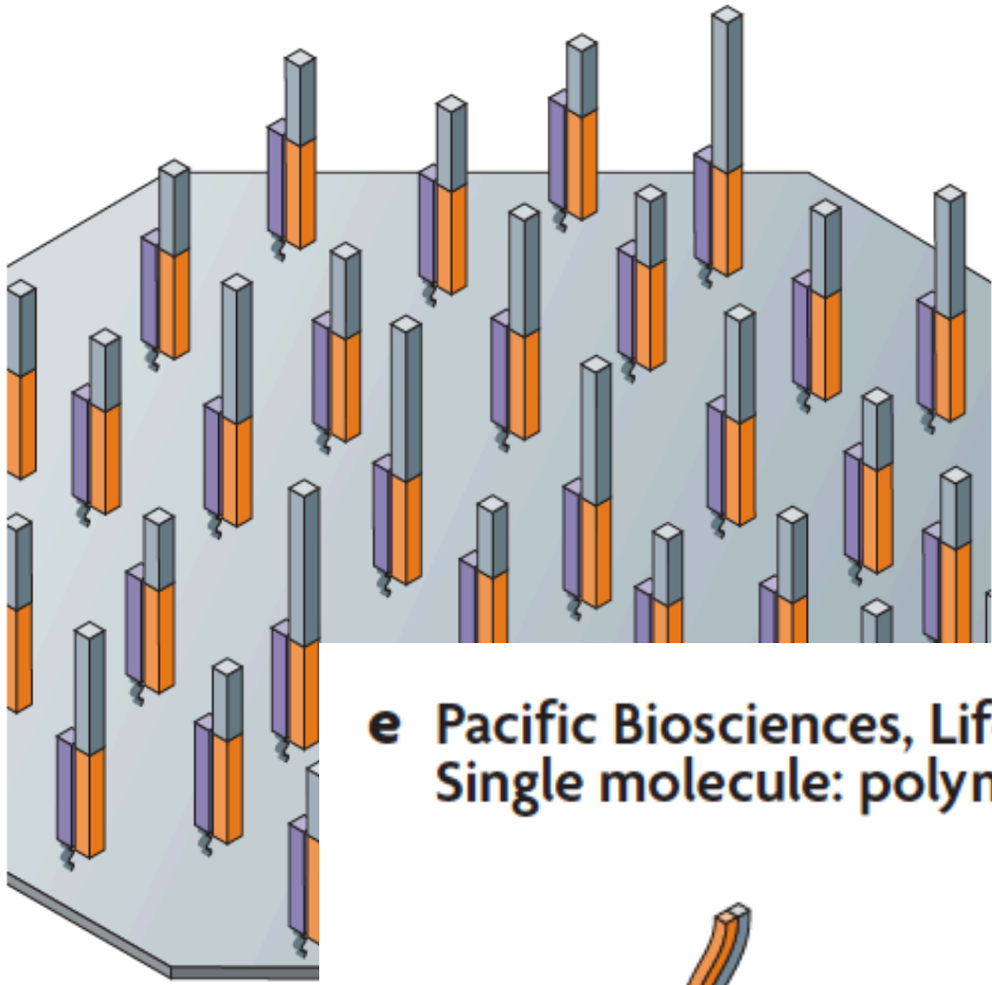
Billions of primed, single-molecule templates

d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized



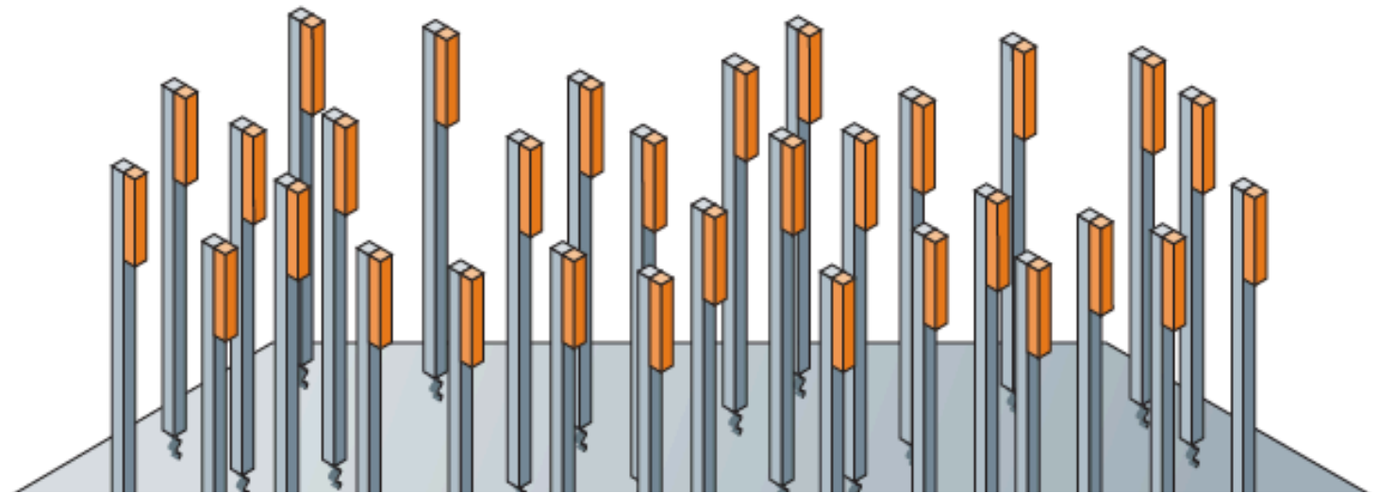
Billions of primed, single-molecule templates

c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized

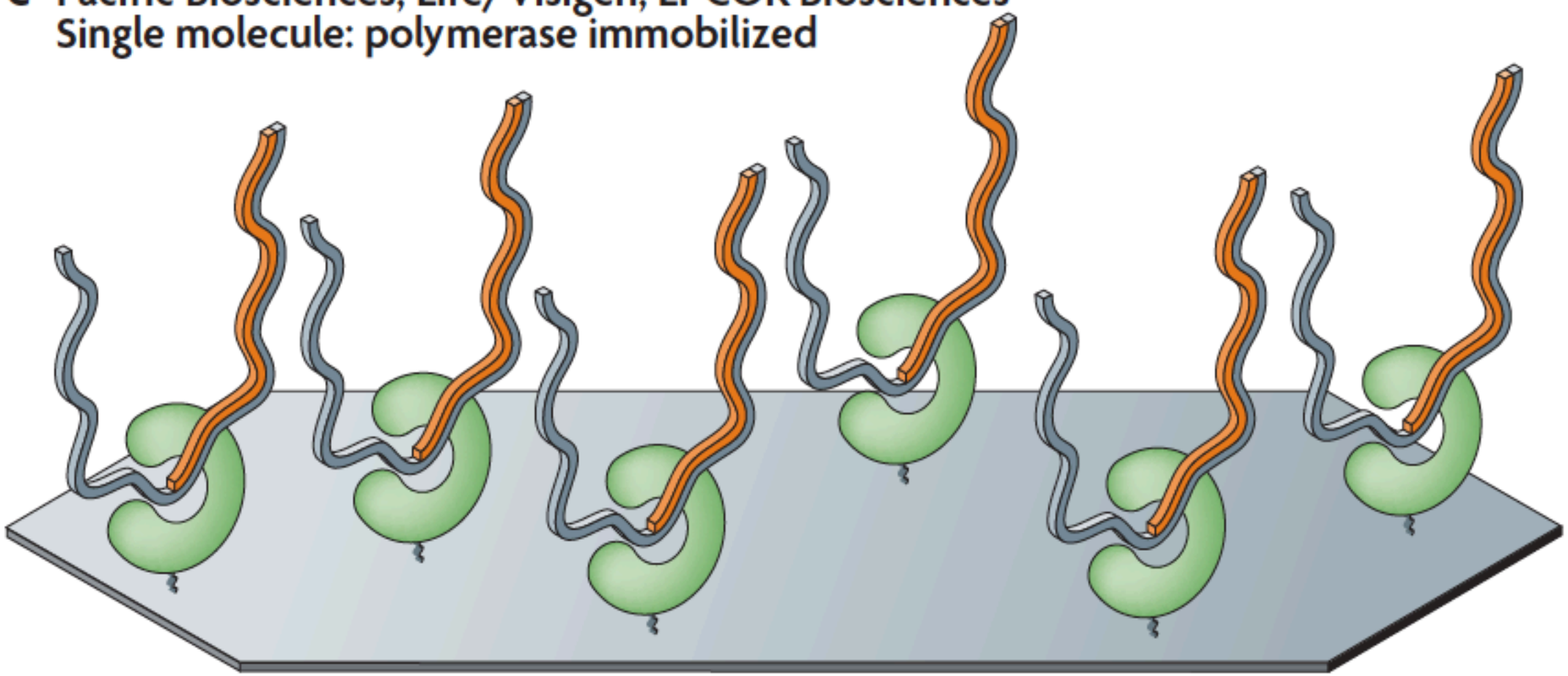


Billions of prir

d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized

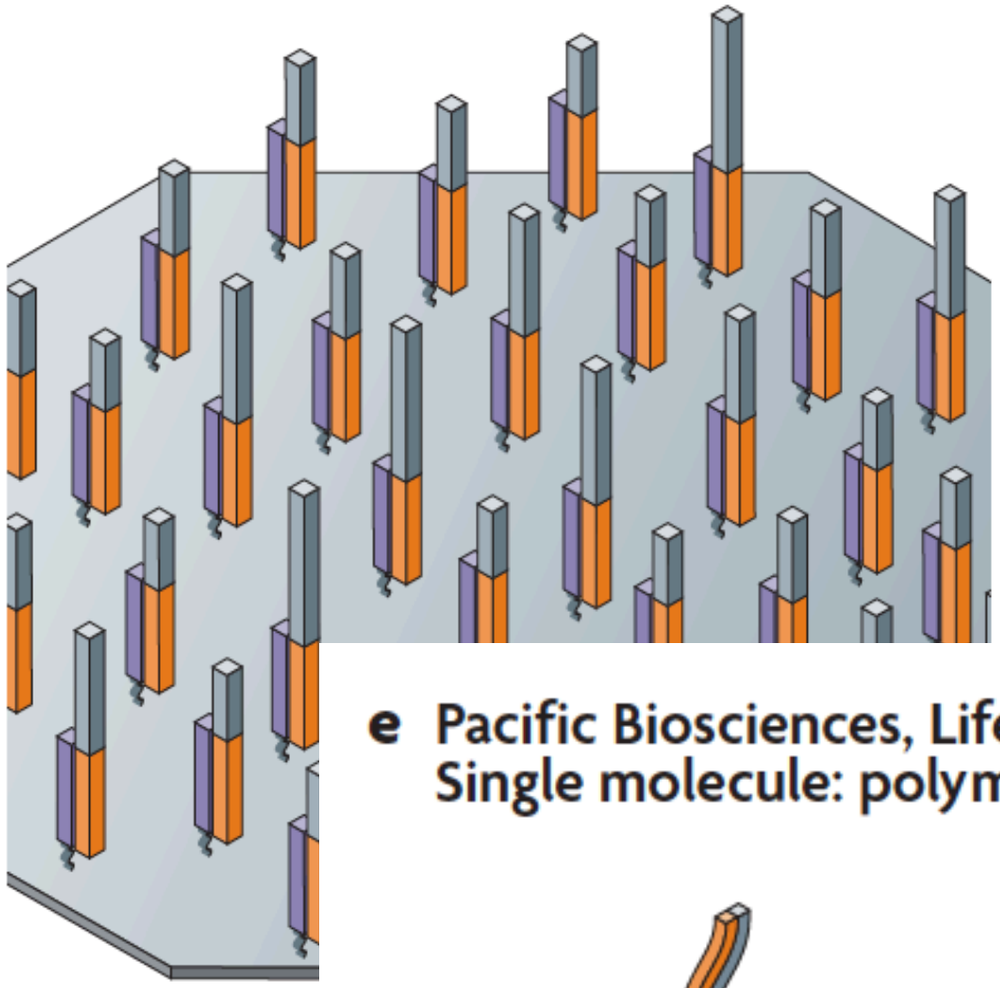


e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized



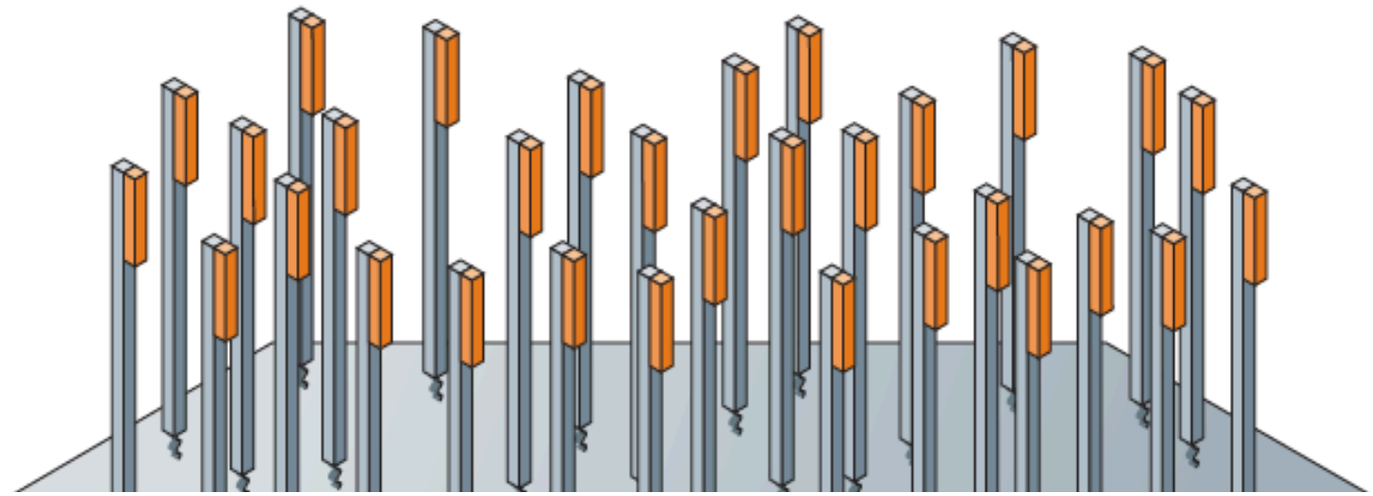
Thousands of primed, single-molecule templates

c Helicos BioSciences: one-pass sequencing
Single molecule: primer immobilized

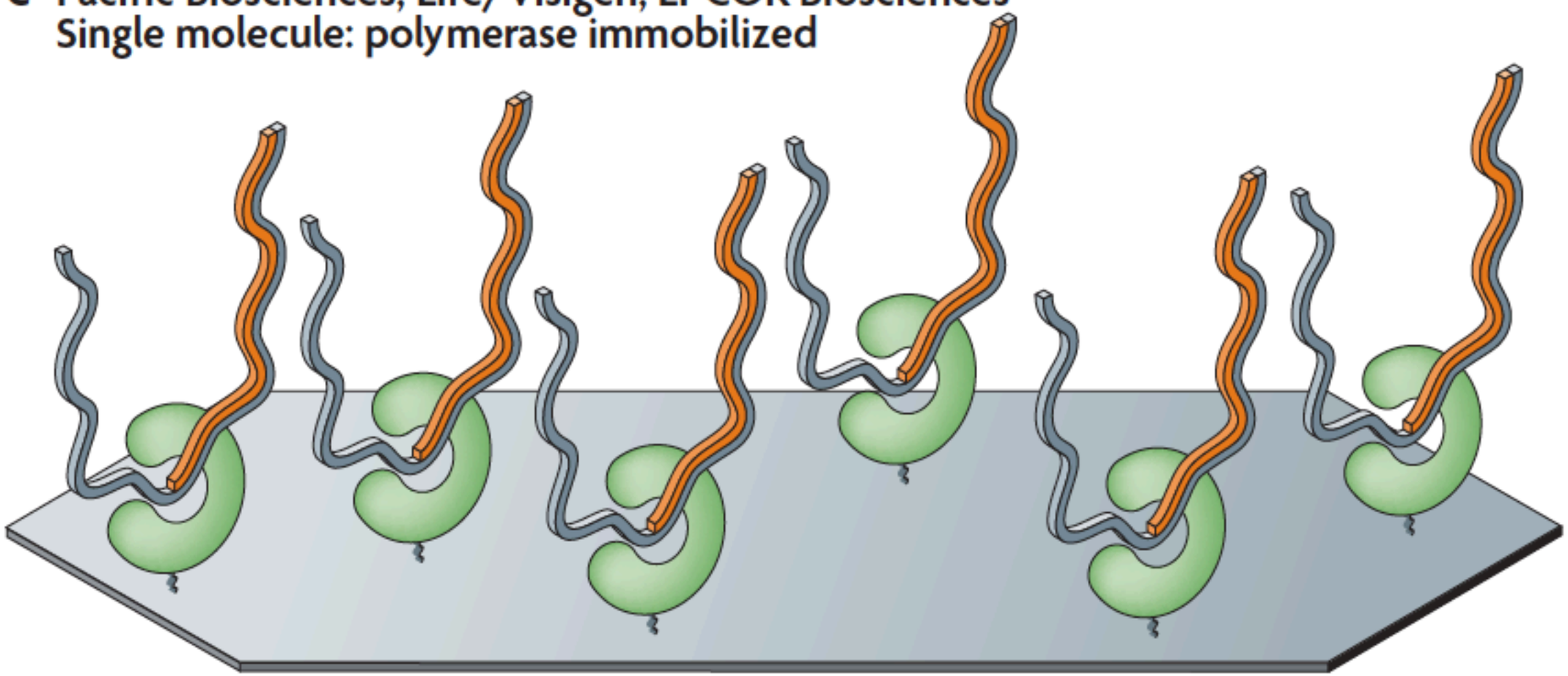


Billions of prir

d Helicos BioSciences: two-pass sequencing
Single molecule: template immobilized



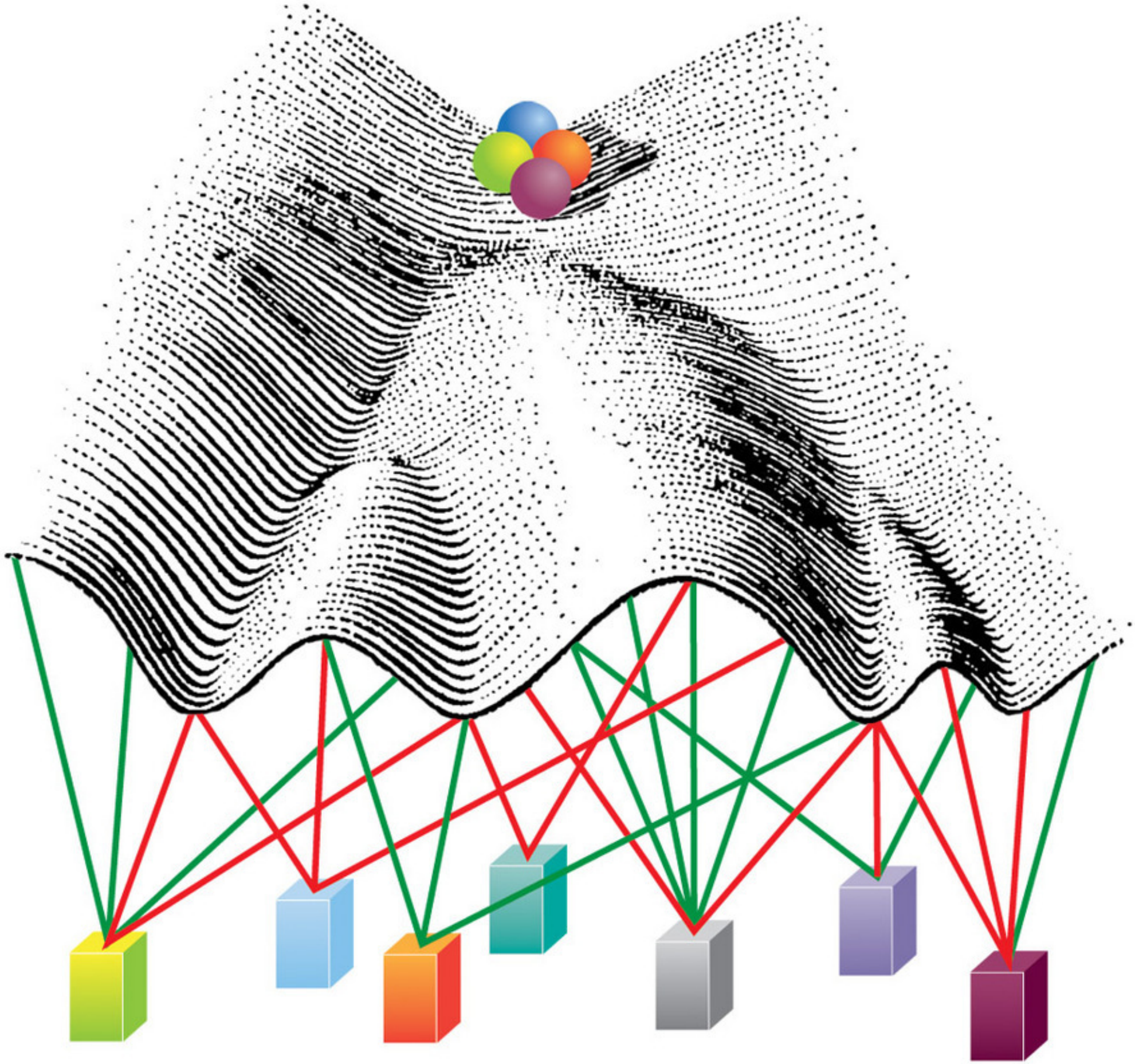
e Pacific Biosciences, Life/Visigen, LI-COR Biosciences
Single molecule: polymerase immobilized



Thousands of primed, single-molecule templates

Overview

- Next-generation sequencing (NGS)
- What is epigenetics?
- Experimental techniques for epigenomics
- Data analysis and visualization



“The difference between genetics and epigenetics can probably be compared to the difference between writing and reading a book. Once a book is written, the text (the genes or DNA: stored information) will be the same in all the copies distributed to the interested audience. However, each individual reader of a given book may interpret the story slightly differently, with varying emotions and projections as they continue to unfold the chapters. In a very similar manner, epigenetics would allow different interpretations of a fixed template (the book or genetic code) and result in different read-outs, dependent upon the variable conditions under which this template is interrogated.”

Epigenetics : mechanisms that regulate gene expression (behavior or function) without influencing genetic codes

Gene expression is initiated by TF binding to promoters or enhancers

TF binding is determined by surrounding structure of DNA

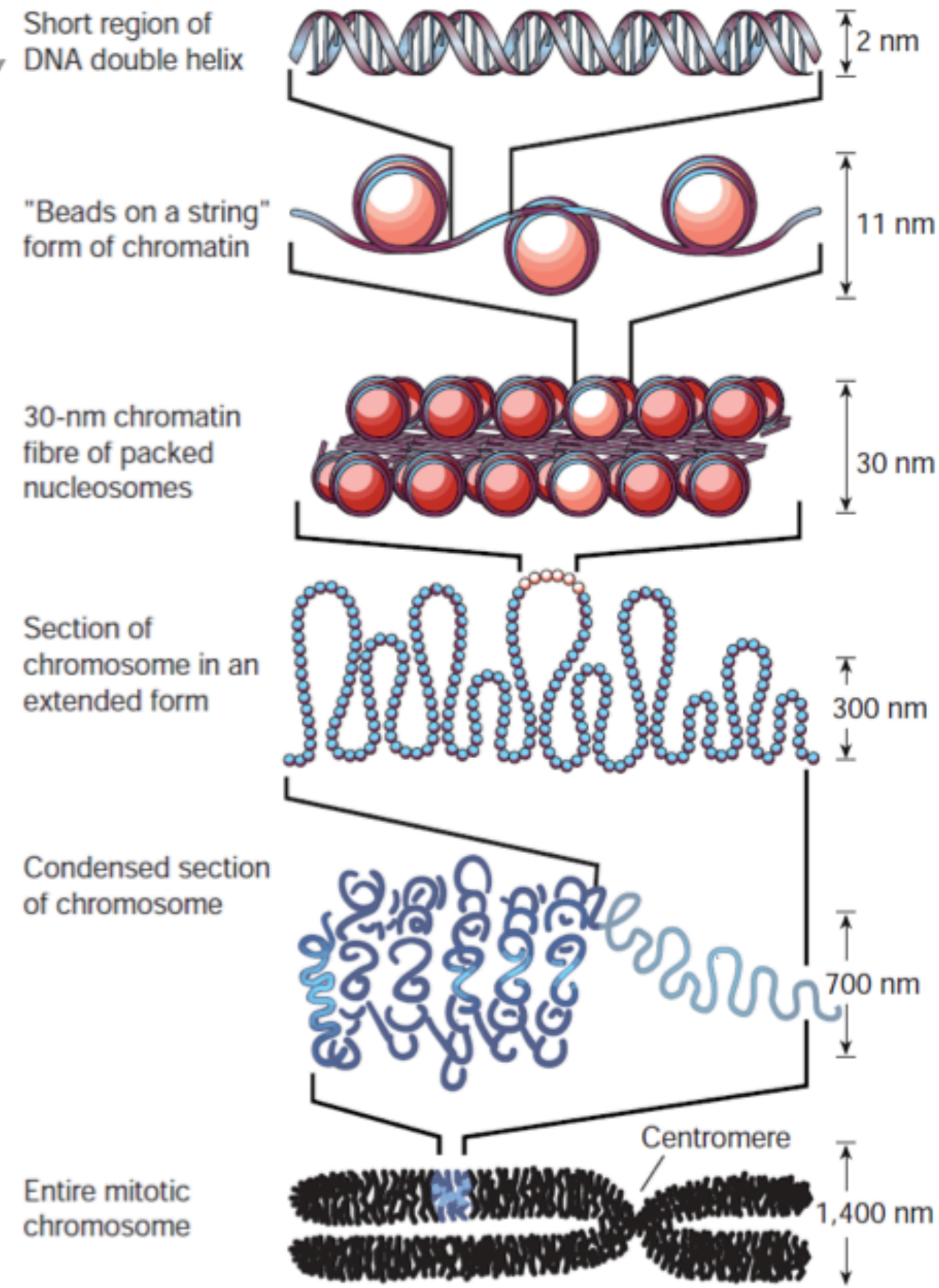
Surrounding structure of DNA is controlled by *epigenetic* mechanisms

DNA methylation

Nucleosome positioning and remodeling

Histone modifications

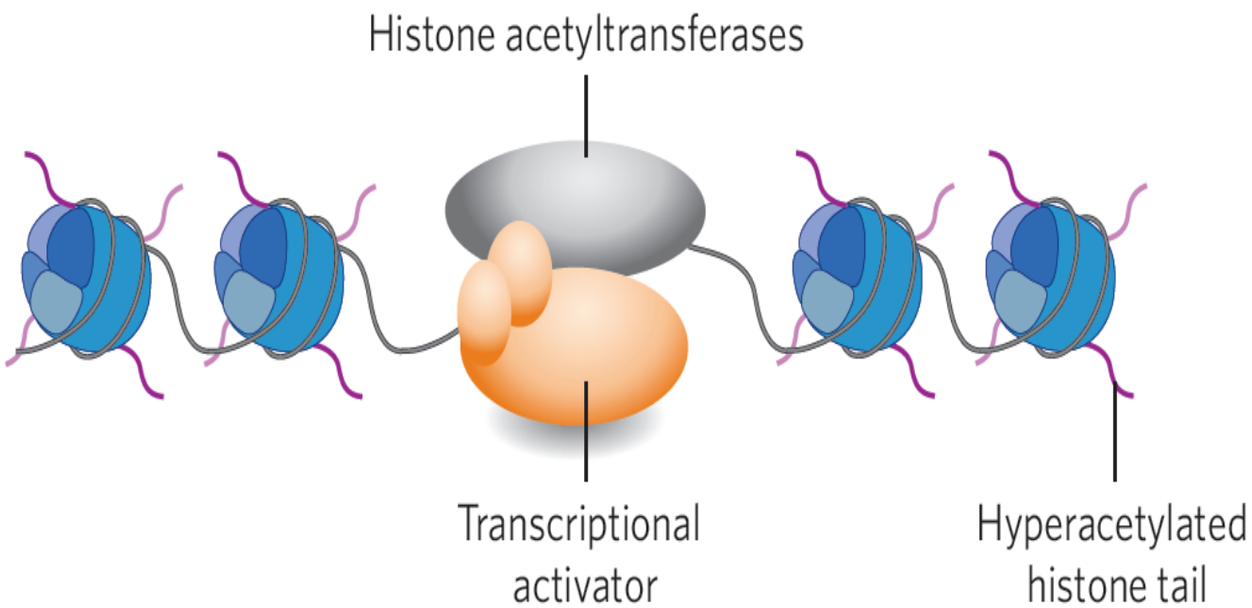
Higher-order chromatin structure



Nature 421:448 (2003)

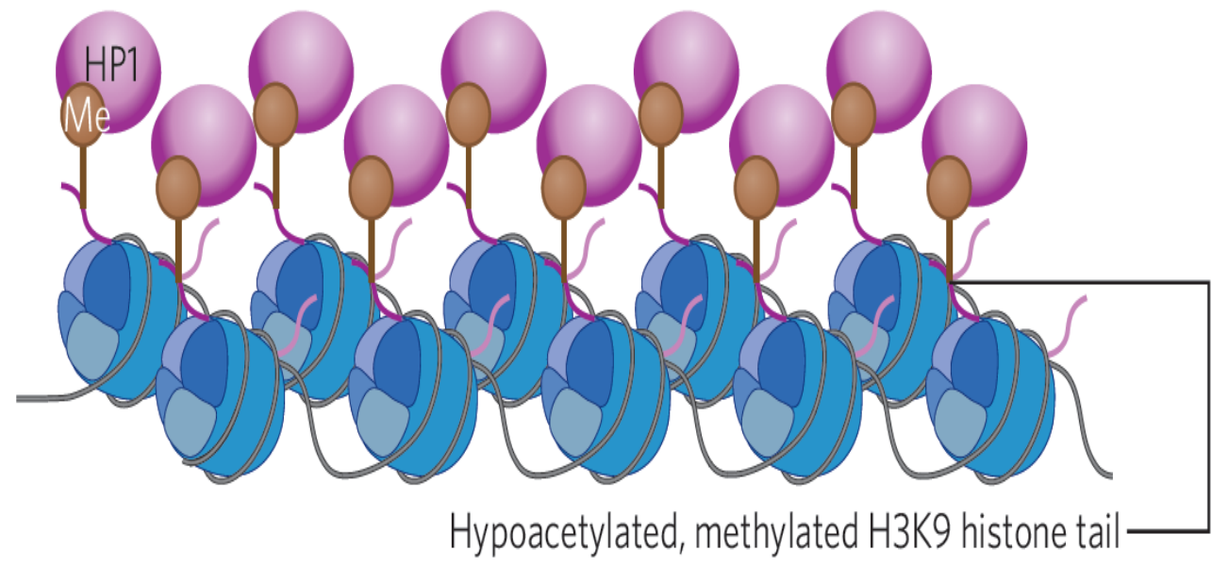
DNA methylation

- Open chromatin
  - Sparse nucleosome positioning
  - Histone acetylations & activating histone methylations



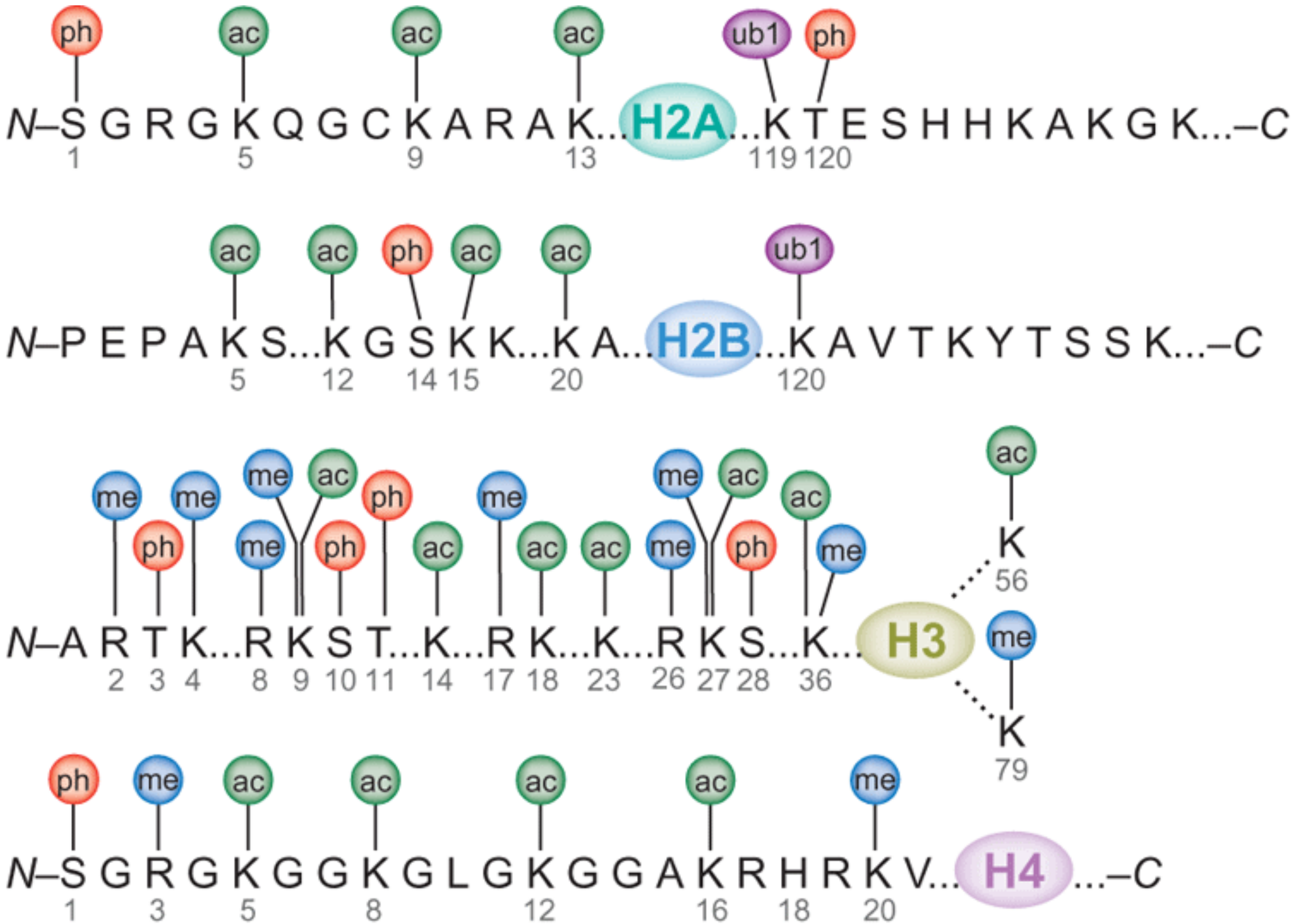
- Less condensed
- At chromosome arms
- Contains unique sequences
- Gene-rich

- Close chromatin
  - Compact nucleosome positioning
  - Lack of histone acetylations & repressing histone methylations

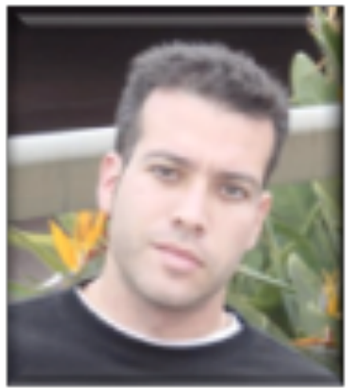


- Highly condensed
- At centromeres and telomeres
- Contains repetitious sequences
- Gene-poor

Histone code



Nucleosome code



Dr. Eran Segal

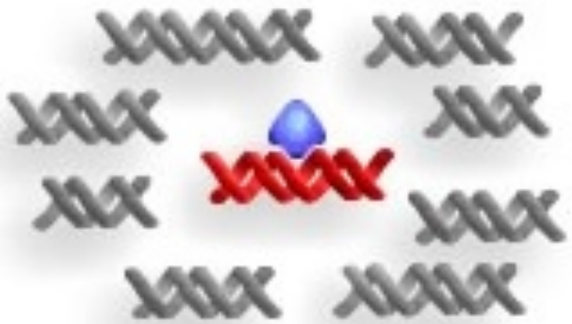
- hot paper
  - Nature cover, News & Views, ESI hot paper, NY Times, Nature Review Genetics, Faculty of 1000 review, Codes and Enigmas, about the authors
- DNA sequence encodes not only protein production but also its own physical packaging
- Nucleosome positioning, part of epigenetic information, is governed by genetic codes

Overview

- Next-generation sequencing (NGS)
- What is epigenetics?
- Experimental techniques for epigenomics
- Data analysis and visualization

Chromatin
Immuno
Precipitation

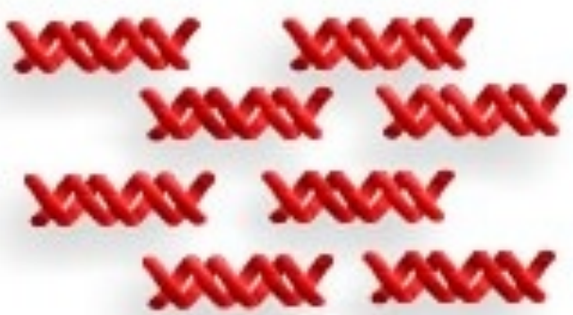
DNA-binding proteins are crosslinked to DNA with formaldehyde in vivo.



Isolate the chromatin. Shear DNA along with bound proteins into small fragments.

chromatin = the complex of DNA and protein

Bind antibodies specific to the DNA-binding protein to isolate the complex by precipitation. Reverse the cross-linking to release the DNA and digest the proteins.



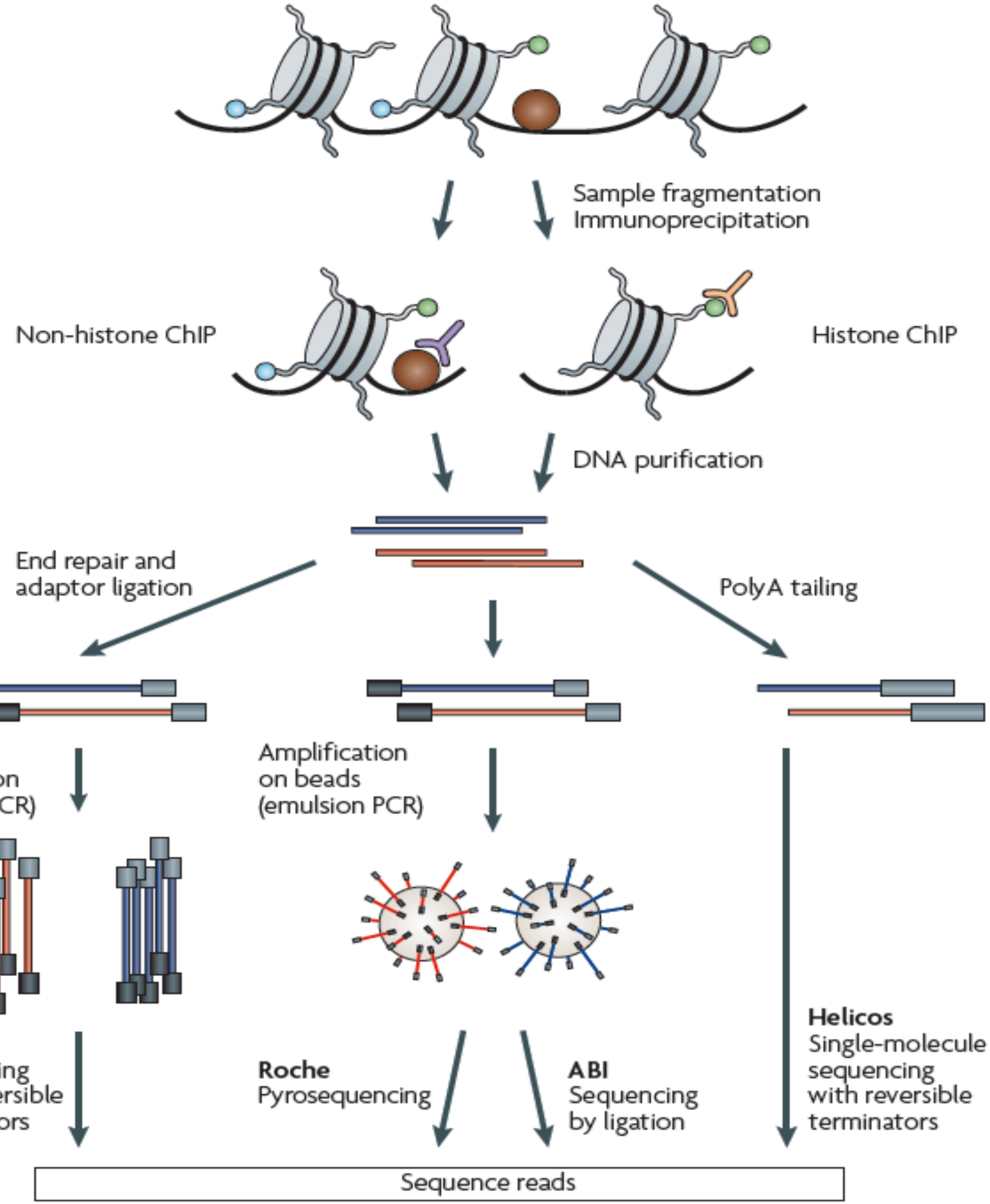
Use PCR to amplify specific DNA sequences to see if they were precipitated with the antibody.

By analyzing the bound DNA, we can know…

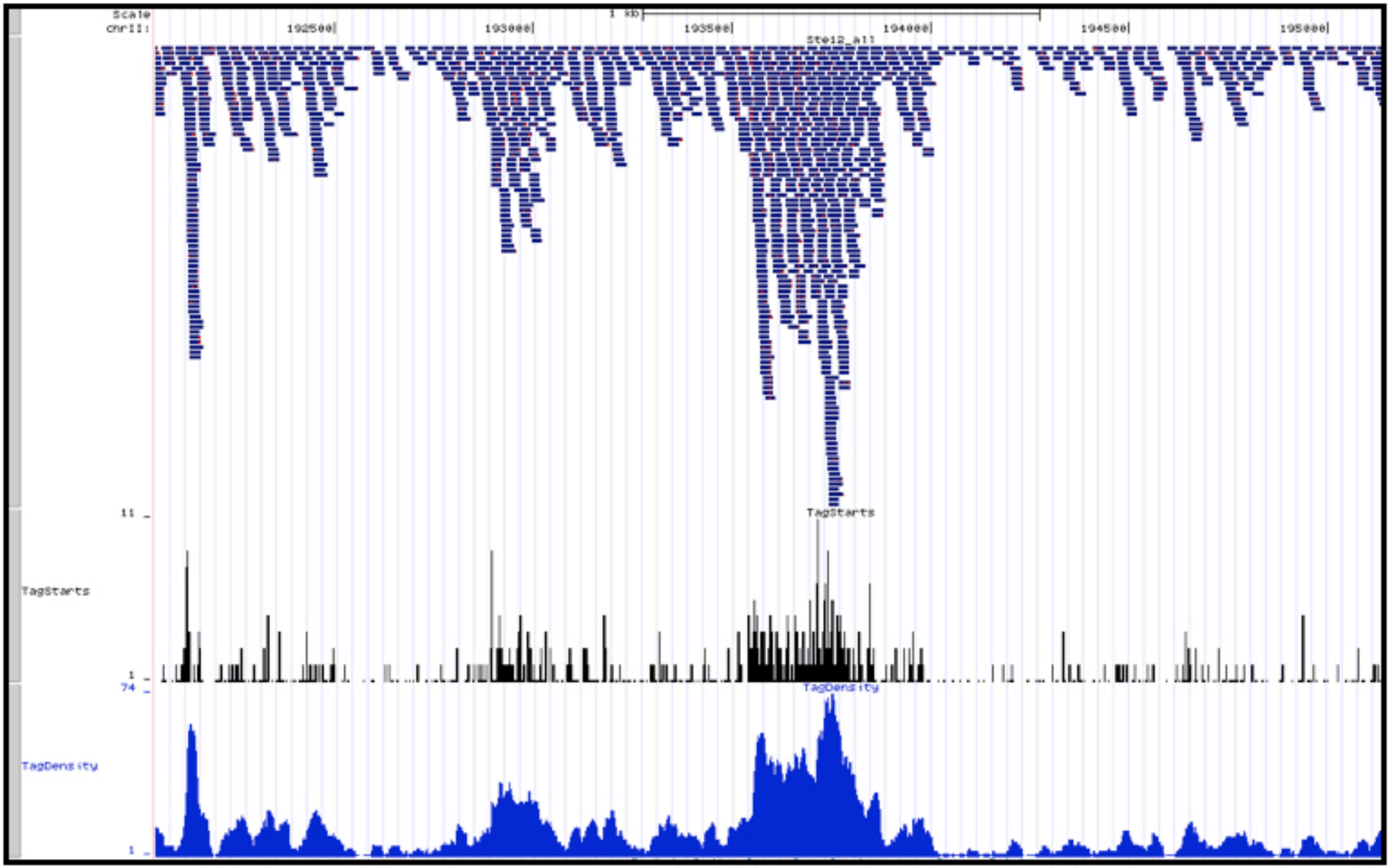
1. DNA location >
   genomic location of the protein of question

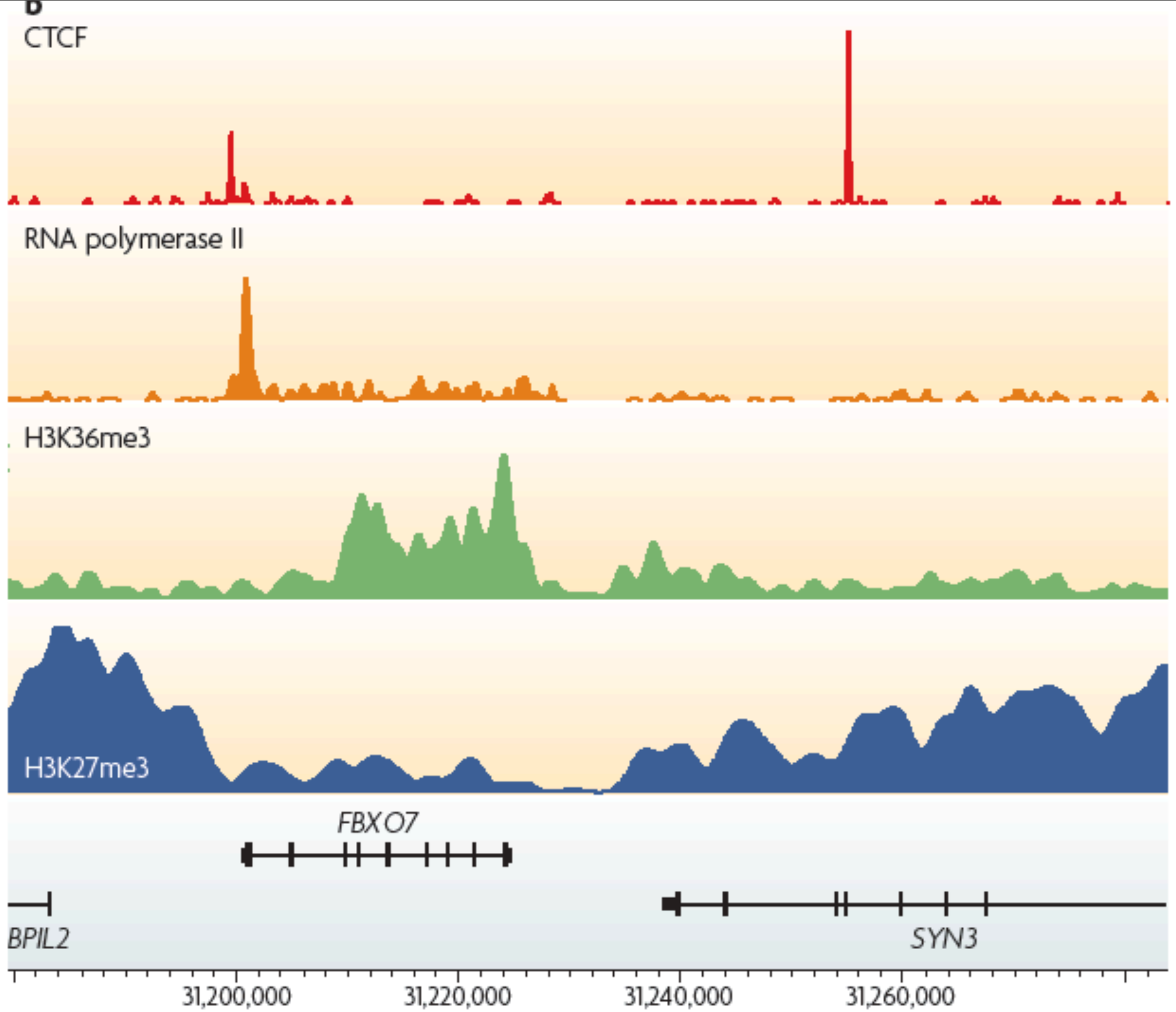
2. DNA counts >
   affinity of protein-DNA interaction

ChIP-seq workflow



tag counting




b
CTCF
RNA polymerase II
H3K36me3
H3K27me3
FBXO7
BPIL2
SYN3
31,200,000
31,220,000
31,240,000
31,260,000

Experiment techniques for NGS

- DNA methylation
- Histone modification
- Nucleosome positioning
- Open chromatin for regulatory function

DNA methylation

- Bisulphite-seq
  - conversion of unmethylated C to T
- MeDIP-seq
  - methylated DNA immunoprecipitation
- MBD-seq
  - methyl-binding-domain protein

Watson >> A C^m G T T C T C C A G T C >>

Crick << T G C^m A A G A G G T C A G <<

Bisulfite conversion

BSW >> AC^mGTTTTTTAGTT >>

BSC << TGC^mAAGAGGTTAG <<

PCR

BSW >> AC^mGTTTTTTAGTT >>

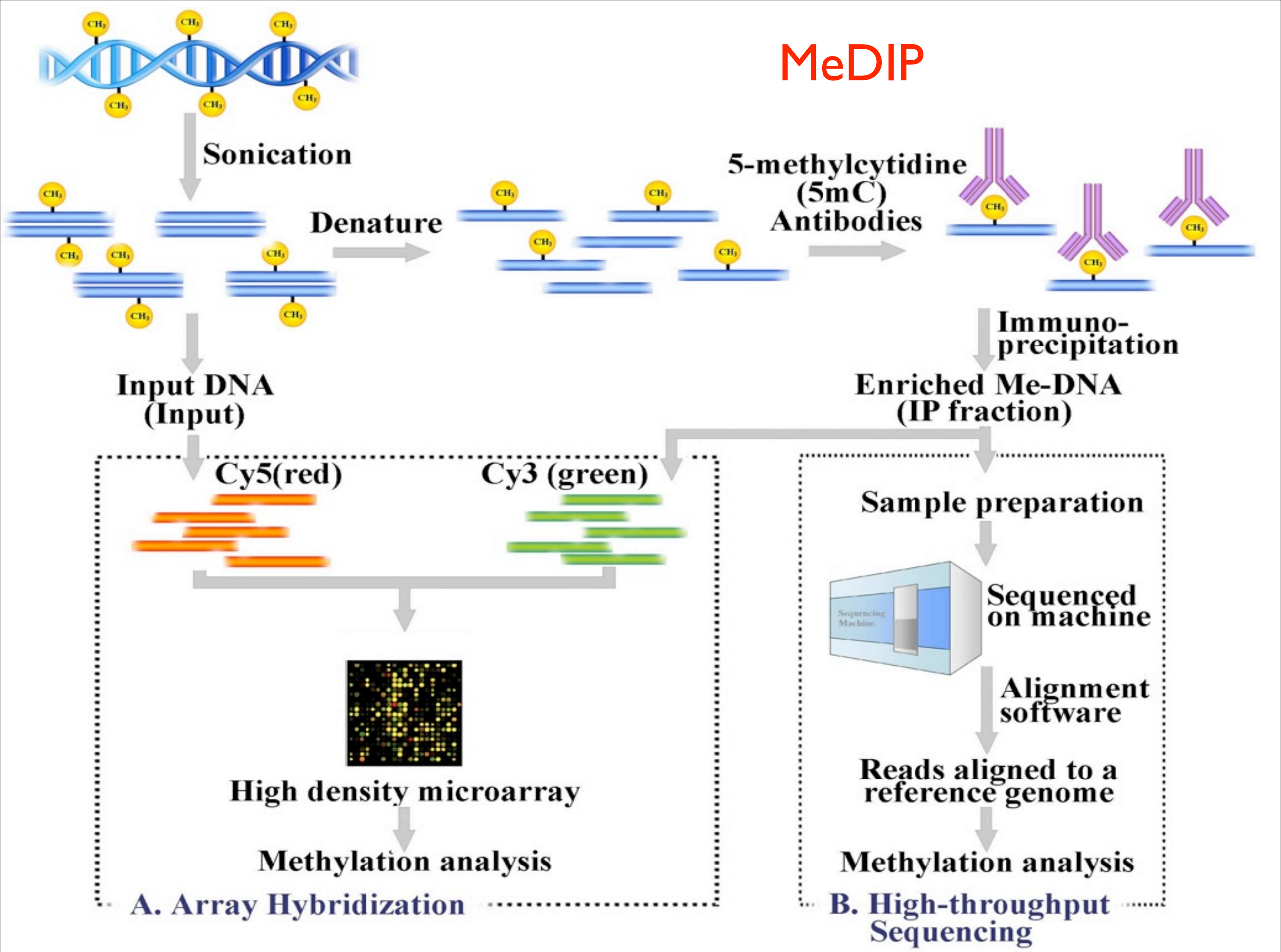
BSWR << TG CAAAAAATCAA <<

BSCR >> ACG TTCTCCAAGA >>

BSC << TGC^mAAGAGGTTAG <<

```
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAATGGTTAAGGTTGGTTTGTGTTAGAACGGTGTGC
TTTGGATAGTTTGTTTATTTATTGTAATGGTTAAGGTTGGTTTGTGTTGGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAATGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGT
TTTGGATAGTTTGTTTATTTATTGTAATGGTTAAGGTTGGTTTGTGTCAGAACGGTGTGT
TTTGGATAGTTTGTTTATTTATTGTAATGGTTAAGGTTGGTTTGTGTTAGAACGGTGTGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGCAACGGTTAAGGTTGGTTTGTGTTATAACGGTGCGC
TTTGGACAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTCGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTTGG--AAGTTGTTTATTTATTT--ACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGG--CGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTCGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAATGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGTGTGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGCGCGC
TTTGGATAGTTTGTTTATTTATTGTAACGGTTAAGGTTGGTTTGTGTTAGAACGGCGCGC
```

MeDIP


Sonication
Denature
5-methylcytidine
(5mC)
Antibodies
Immuno-
precipitation
Input DNA
(Input)
Enriched Me-DNA
(IP fraction)
Cy5(red)
Cy3 (green)
High density microarray
Methylation analysis
A. Array Hybridization
Sample preparation
Sequencing
Machine
Sequenced
on machine
Alignment
software
Reads aligned to a
reference genome
Methylation analysis
B. High-throughput
Sequencing

Histone modification

- ChIP-seq
  - antibodies against various histone modifications or modifying enzymes

5 HATs & 4 HDACs

Cell

Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes

nature genetics

39 histone acetylations and methylations

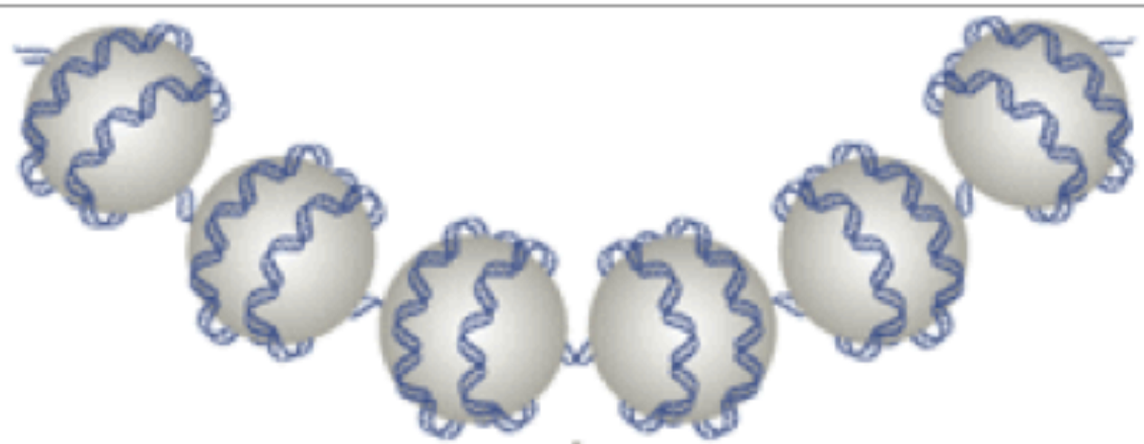
Combinatorial patterns of histone acetylations and methylations in the human genome

Nucleosome positioning or open chromatin

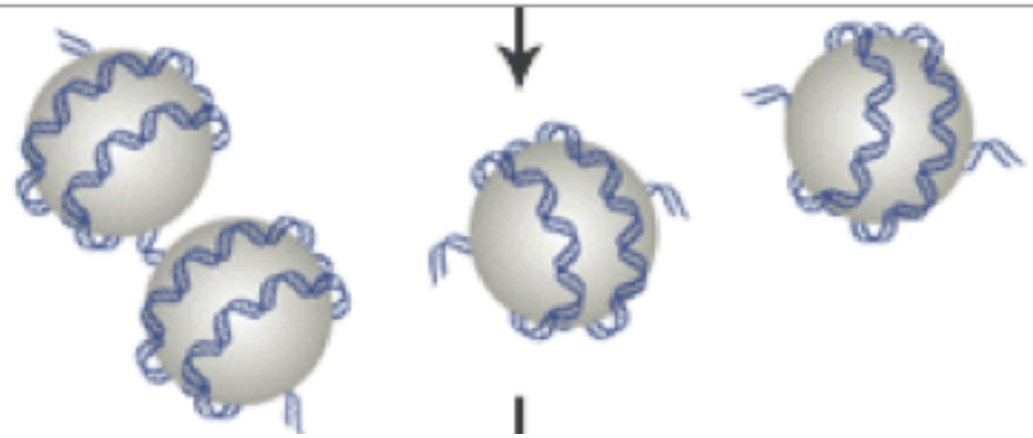
- ChIP-seq
  - antibodies against histones
- MNase-seq
  - capture nucleosomal DNA
- DNase-seq
  - capture nucleosome-free regions

MNase digestion

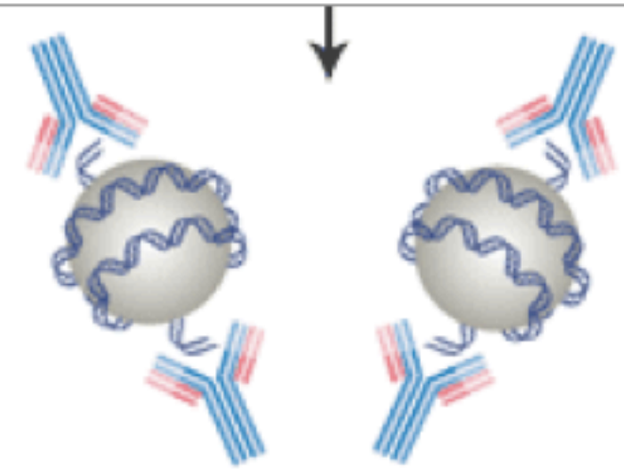
Cells are fixed with formaldehyde to cross-link histone and non-histone proteins to DNA.



Chromatin is digested with Micrococcal Nuclease into 150–900 bp DNA/protein fragments.

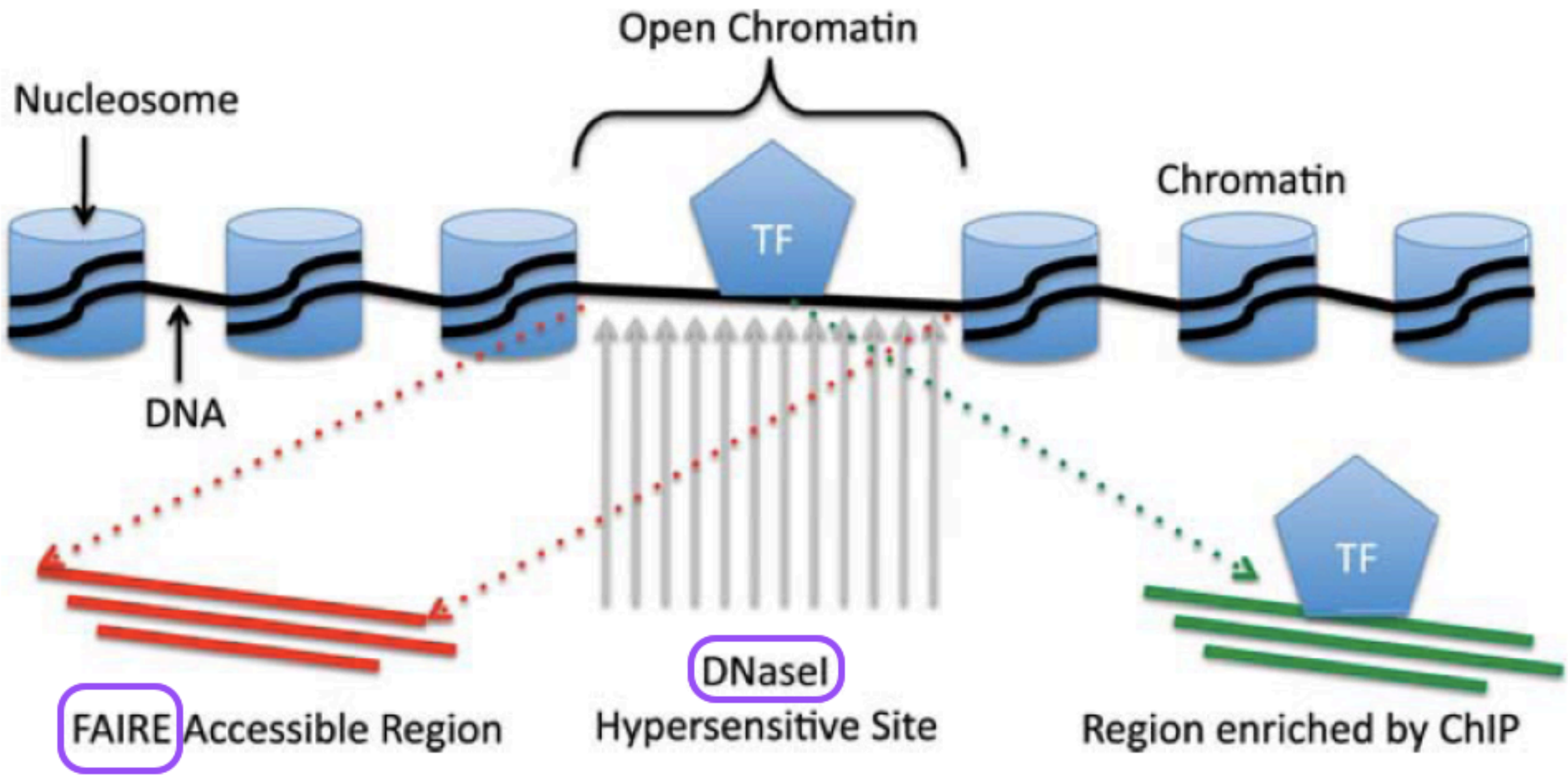


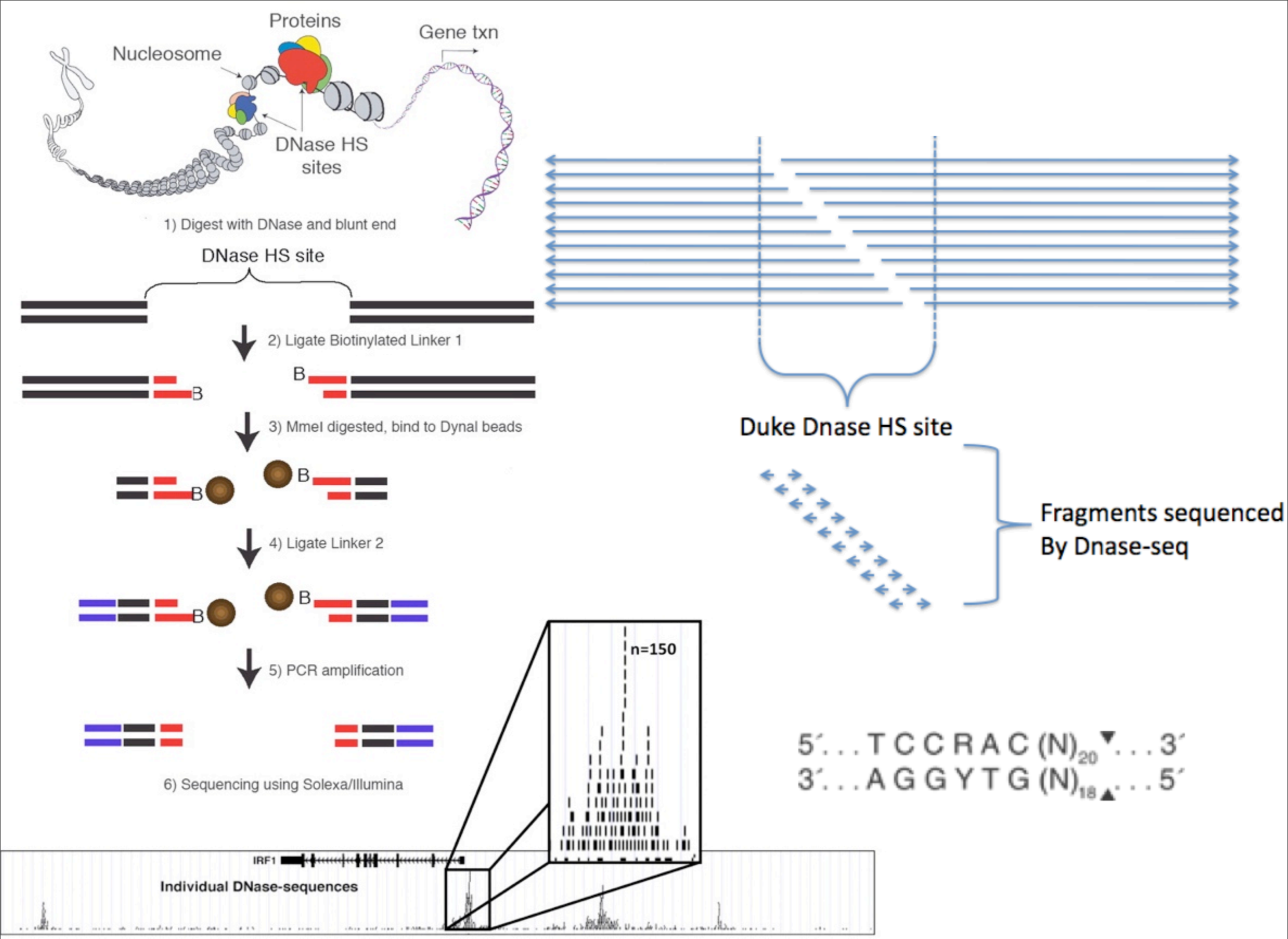
Antibodies specific to histone or non-histone proteins are added and the complex co-precipitates and is captured by Protein G Agarose or Protein G magnetic beads.

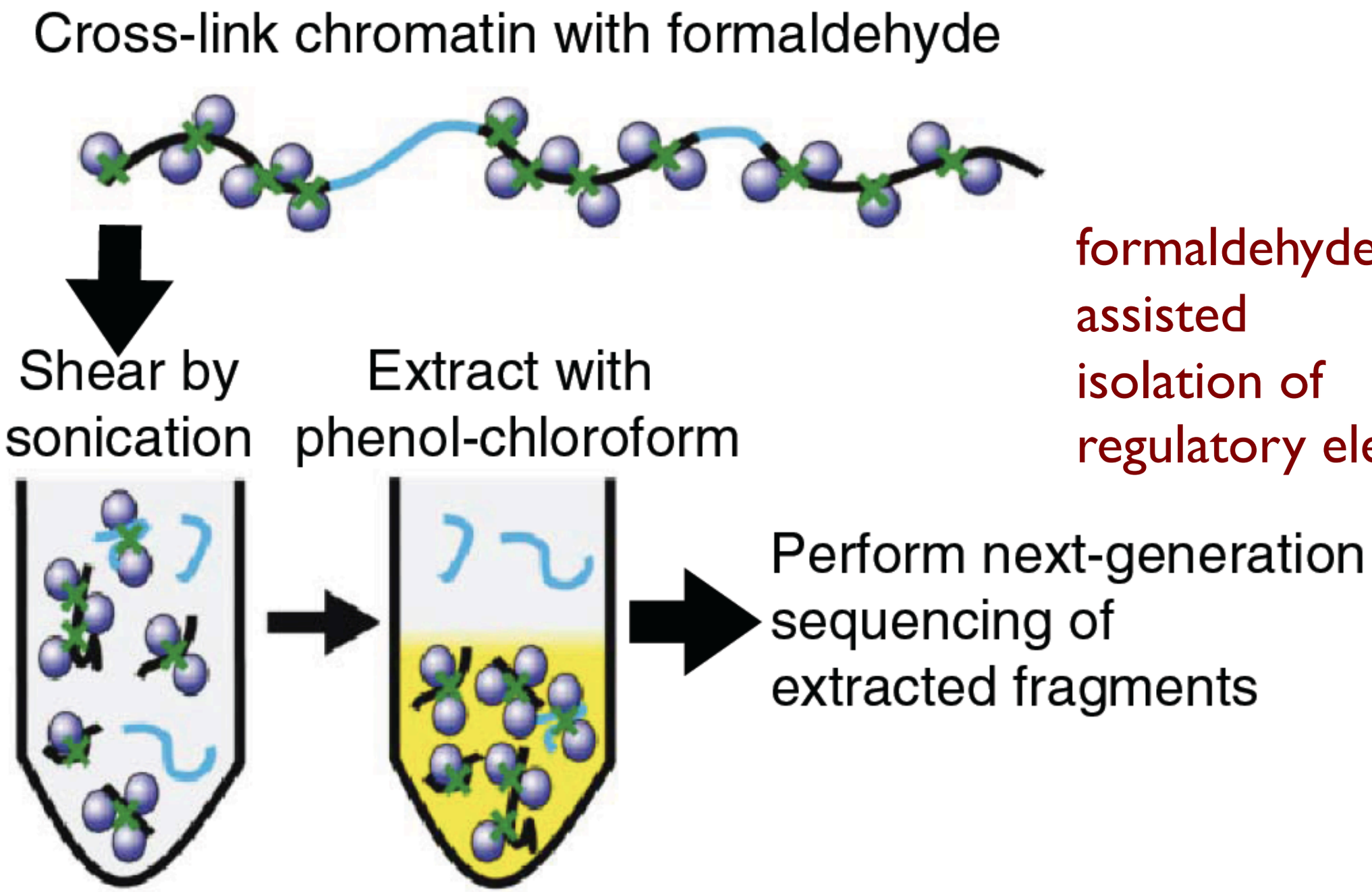


Cross-links are reversed, and DNA is purified and ready for analysis.

Profiling regulatory elements




Proteins
Nucleosome
Gene txn
DNase HS sites
1) Digest with DNase and blunt end
DNase HS site
2) Ligate Biotinylated Linker 1
B
3) MmeI digested, bind to Dynal beads
4) Ligate Linker 2
5) PCR amplification
6) Sequencing using Solexa/Illumina
Duke Dnase HS site
Fragments sequenced By Dnase-seq
n=150
5′...TCCRAC(N)20▼...3′
3′...AGGYTG(N)18▲...5′
IRF1
Individual DNase-sequences



formaldehyde-assisted isolation of regulatory elements

Overview

- Next-generation sequencing (NGS)
- What is epigenetics?
- Experimental techniques for epigenomics
- Data analysis and visualization

Tools

- BEDTools
- Peak finding
  - MACS, Peak Finder, CCAAT, FindPeaks
- HOMER
- Galaxy / GeneTrack

Tools

- BEDTools
- Peak finding
  - MACS, Peak Finder, CCAAT, FindPeaks
- HOMER
- Galaxy / GeneTrack

BED format

(Browser extensible data)

Required fields **Optional fields**

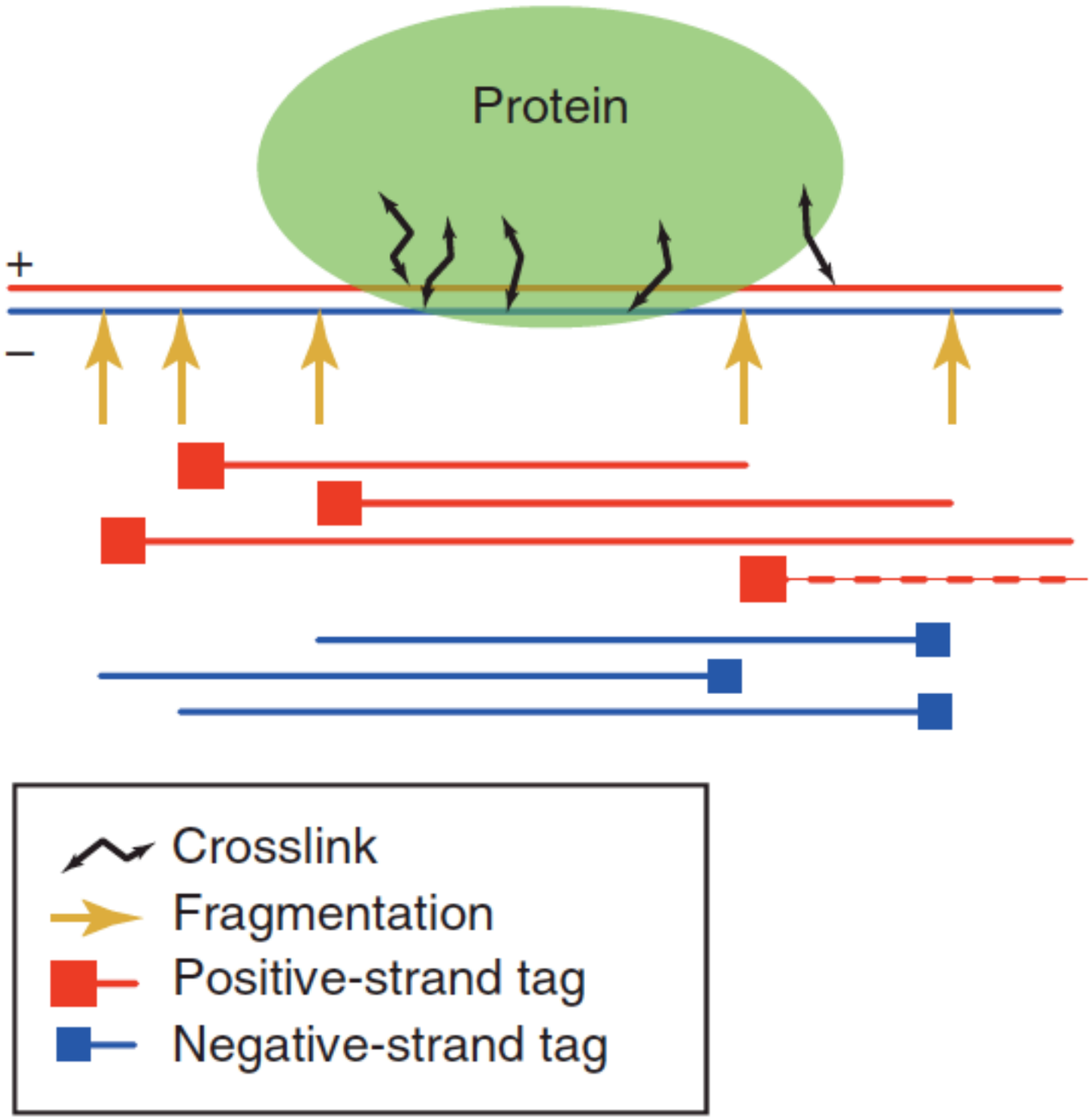
chr7	127471196	127472363	Pos1	0	+
chr7	127472363	127473530	Pos2	0	+
chr7	127473530	127474697	Pos3	0	+
chr7	127474697	127475864	Pos4	0	+
chr7	127475864	127477031	Neg1	0	-
chr7	127477031	127478198	Neg2	0	-
chr7	127478198	127479365	Neg3	0	-
chr7	127479365	127480532	Pos5	0	+
chr7	127480532	127481699	Neg4	0	-

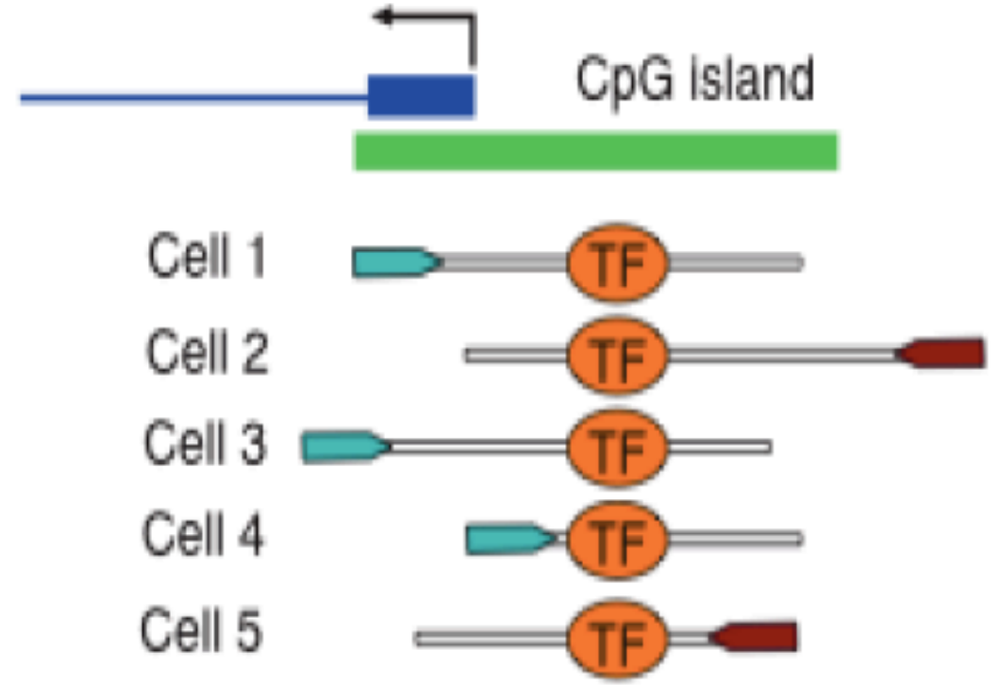
BEDTools

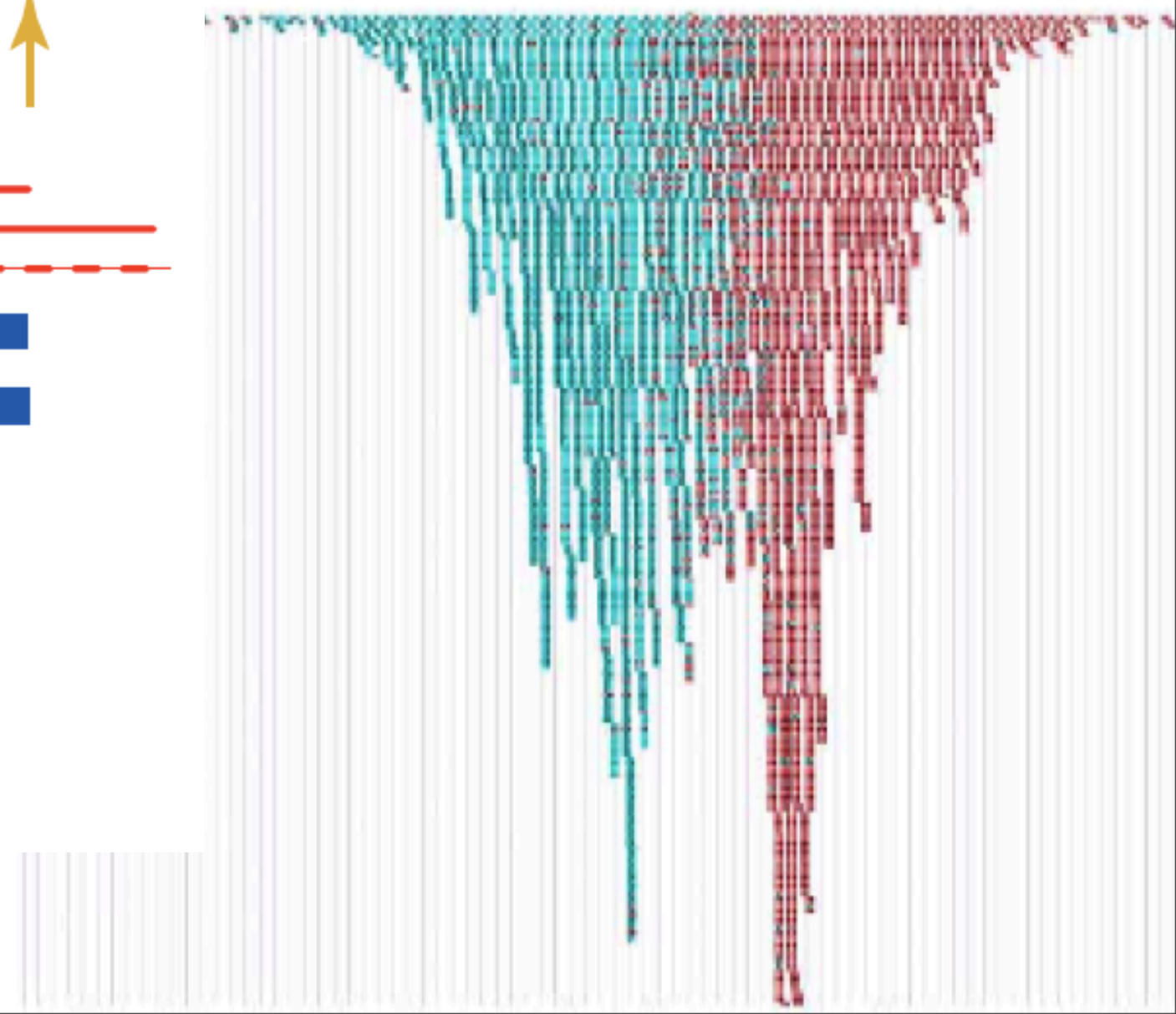
Utility	Description
intersectBed	Returns overlapping features between two BED/GFF files. *Supports BAM format as input and output.*
pairToBed	Returns overlaps between a BEDPE file and a regular BED/GFF file. *Supports BAM format as input and output.*
pairToPair	Returns overlaps between two BEDPE files.
bamToBed	Converts BAM alignments to BED and BEDPE formats. *Supports BAM format as input.*
windowBed	Returns overlapping features between two BED/GFF files within a "window".
closestBed	Returns the closest feature to each entry in a BED/GFF file.
subtractBed	Removes the portion of an interval that is overlapped by another feature.
mergeBed	Merges overlapping features into a single feature.
coverageBed	Summarizes the depth and breadth of coverage of features in one BED/GFF file (e.g., aligned reads) relative to another (e.g., user-defined windows).
genomeCoverageBed	Histogram or a "per base" report of genome coverage.
fastaFromBed	Creates FASTA sequences from BED/GFF intervals.
maskFastaFromBed	Masks a FASTA file based upon BED/GFF coordinates.
shuffleBed	Permutes the locations of features within a genome.
slopBed	Adjusts features by a requested number of base pairs.
sortBed	Sorts BED/GFF files in useful ways.
linksBed	Creates an HTML links from a BED/GFF file.
complementBed	Returns intervals not spanned by features in a BED/GFF file.

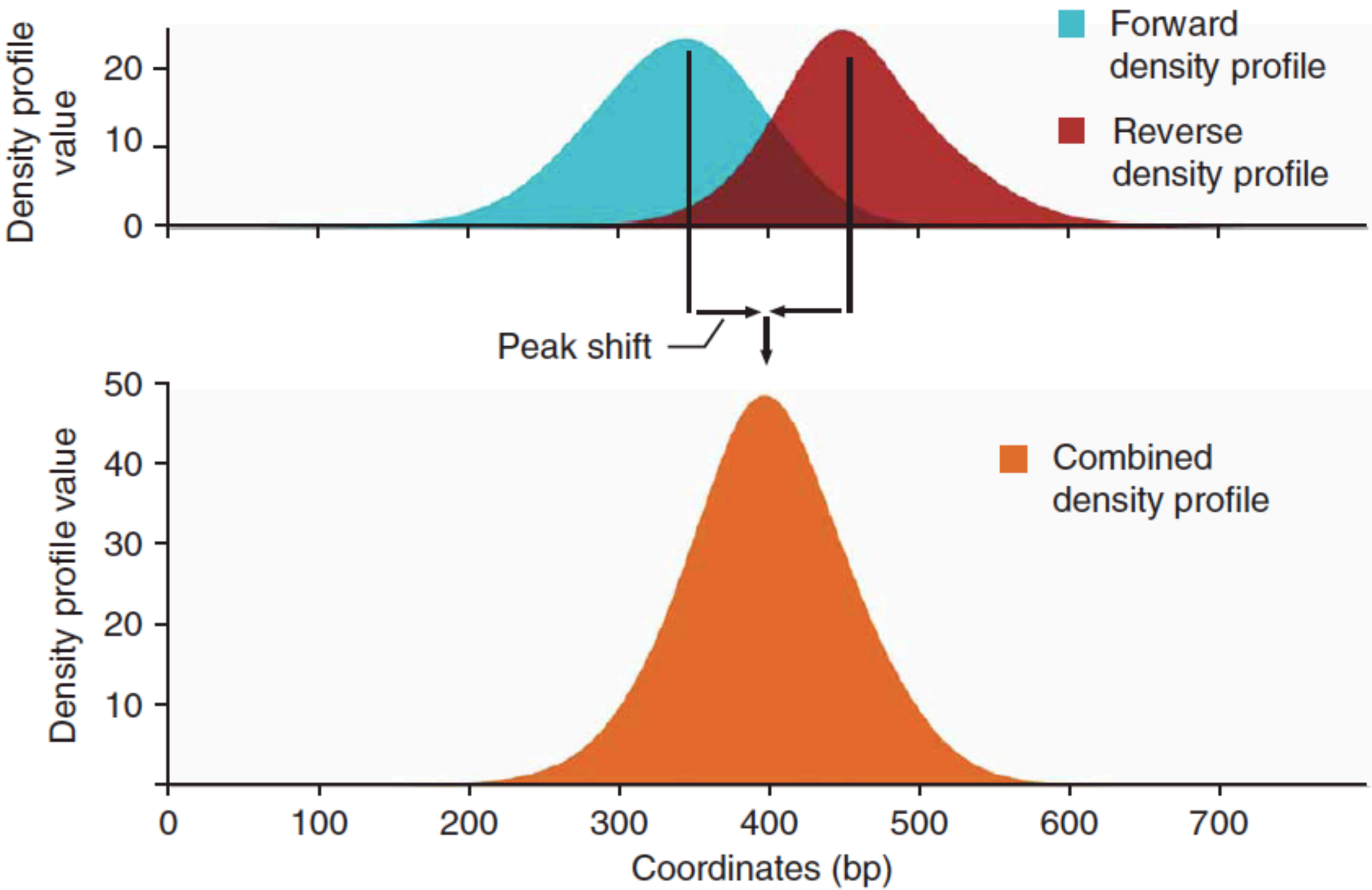
Tools

- BEDTools
- Peak finding
  - MACS, Peak Finder, CCAAT, FindPeaks
- HOMER
- Galaxy / GeneTrack


Protein
+
−
Crosslink
Fragmentation
Positive-strand tag
Negative-strand tag


CpG Island
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
TF
TF
TF
TF
TF



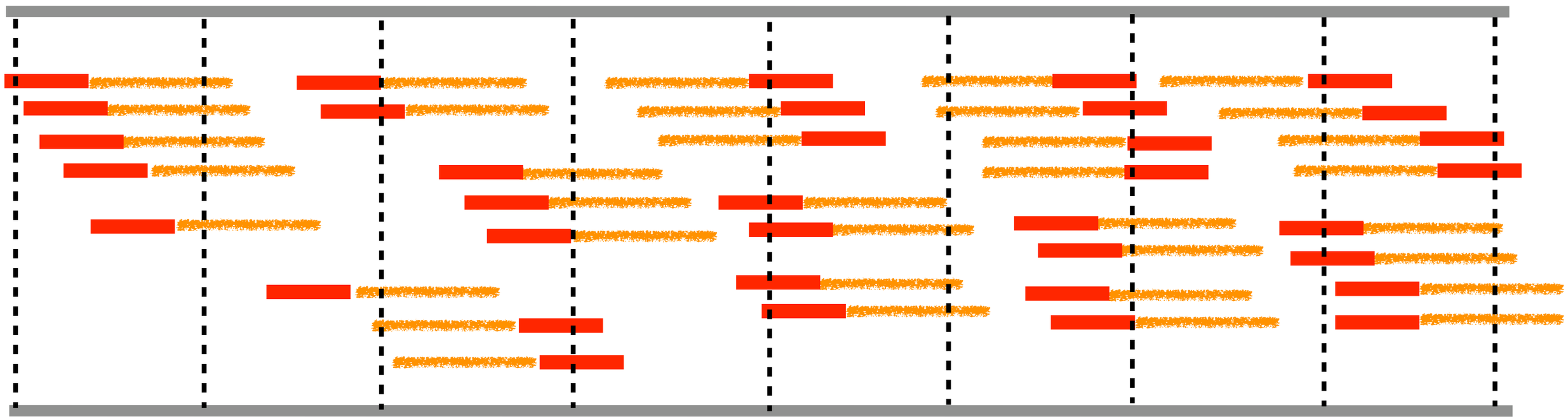

Forward density profile
Reverse density profile
Density profile value
20
10
0
Peak shift
Combined density profile
Density profile value
50
40
30
20
10
0
100
200
300
400
500
600
700
Coordinates (bp)

Tag extension to the library size

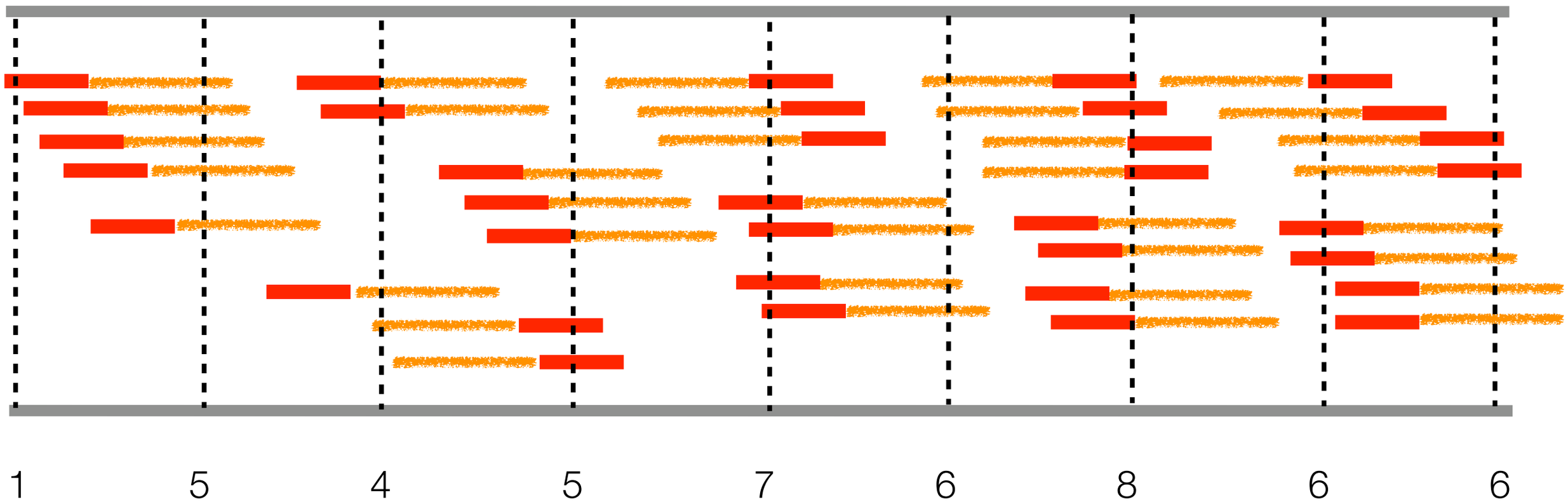
Tag extension to the library size

Tag extension to the library size

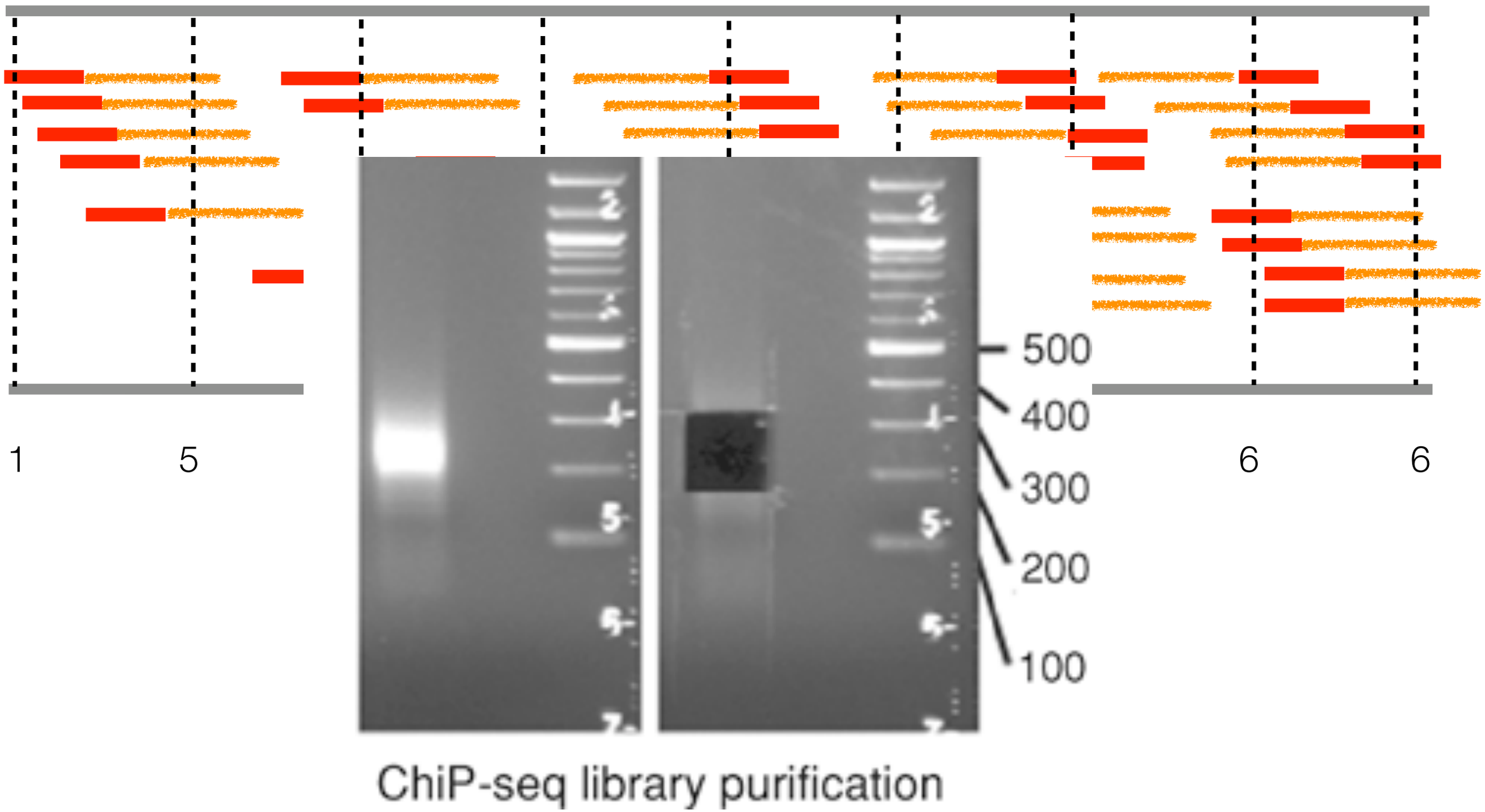
Tag extension to the library size



Tag extension to the library size



Tag extension to the library size



WIG (wiggle) format

```
variableStep  chrom=chrN  [span=windowSize]
chromStartA  dataValueA
chromStartB  dataValueB
... etc ...  ... etc ...
```

```
variableStep chrom=chr2
300701 12.5
300702 12.5
300703 12.5
300704 12.5
300705 12.5
```

```
variableStep chrom=chr2 span=5
300701 12.5
```

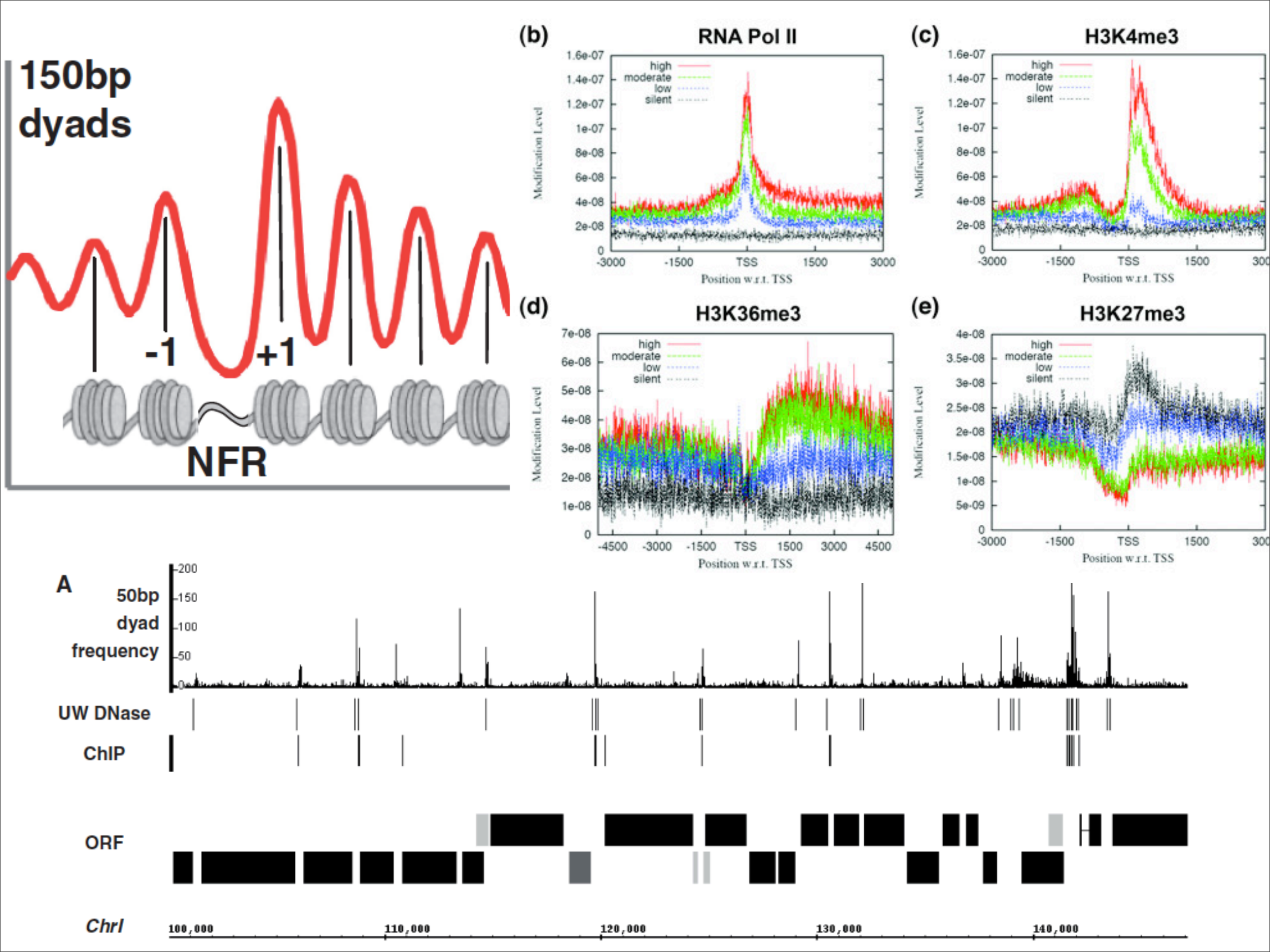
```
fixedStep chrom=chrN start=position step=stepInterval [span=windowSize]
dataValue1
dataValue2
... etc ...
```

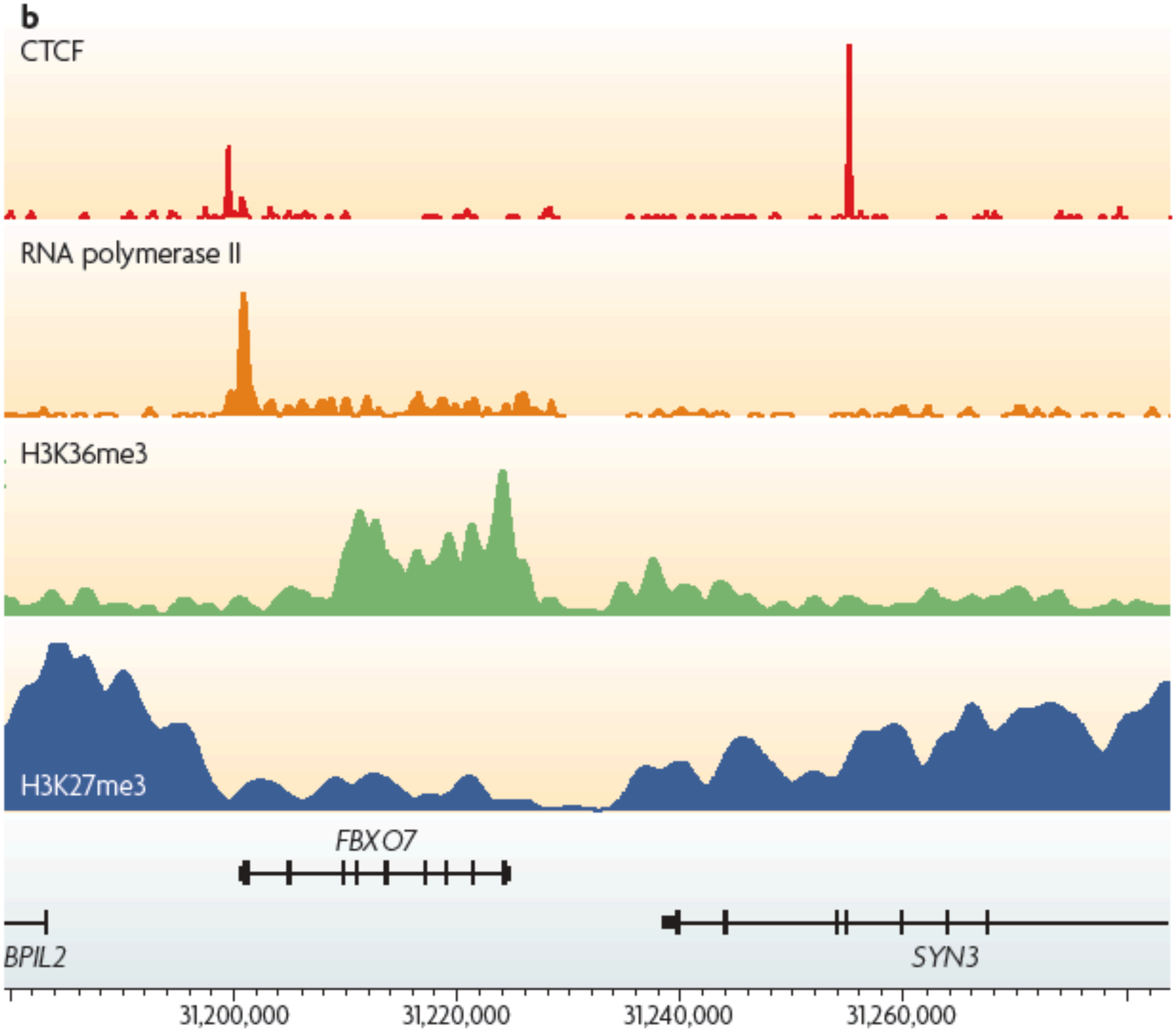
```
fixedStep chrom=chr3 start=400601 step=100
11
22
33
```

displays the values 11, 22, and 33 as single-base regions on chromosome 3 at positions 400601, 400701, and 400801, respectively. Adding span=5 to the declaration line:

```
fixedStep chrom=chr3 start=400601 step=100 span=5
11
22
33
```

causes the values 11, 22, and 33 to be displayed as 5-base regions on chromosome 3 at positions 400601-400605, 400701-400705, and 400801-400805, respectively.


150bp
dyads
-1
+1
NFR
(b)
RNA Pol II
(c)
H3K4me3
(d)
H3K36me3
(e)
H3K27me3
high
moderate
low
silent
Modification Level
Position w.r.t. TSS
TSS
A
50bp
dyad
frequency
UW DNase
ChIP
ORF
ChrI
100,000
110,000
120,000
130,000
140,000


b
CTCF
RNA polymerase II
H3K36me3
H3K27me3
FBXO7
BPIL2
SYN3
31,200,000
31,220,000
31,240,000
31,260,000

Tools

- BEDTools
- Peak finding
  - MACS, Peak Finder, CCAAT, FindPeaks
- HOMER
- Galaxy / GeneTrack

http://biowhat.ucsd.edu/homer

HOMER(v2.6, 10-22-2010)

Software for motif discovery and ChIP-Seq analysis

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and ChIP-Seq analysis. It is a collection of command line programs for unix-style operating systems written in mostly perl and c++. Homer was primarily written as a *de novo* motif discovery algorithm that is well suited for finding 8-12 bp motifs in large scale genomics data.

News

(10-11-2010) Some people have been having trouble after updating - I think I have the configure script fixed such that it won't happen in the future.

(09-01-2010) UCSC has updated their software to do more rigorous error checking - as a result, old UCSC files made with HOMER may not work. The new version of the software fixes this problem.

(09-01-2010) This version was a little rushed because of the UCSC issue - there will likely be an update fixing a boatload of problems in the near future.

Program Download

configureHomer.pl script v2.6 (10-22-10) - use for downloading and updating HOMER program and associated data

Instructions: Download configureHomer.pl (right click and select "save link as") and place in a directory dedicated for HOMER (such as homer/).
Run the script by typing "perl configureHomer.pl" - consult the links below for more information. Additional software and configuration will be required the first time you install HOMER (see Installation).

To upgrade, change your directory to where you installed HOMER, and type:

perl configureHomer.pl -update
- or -
perl configureHomer.pl -install homer (this is good for forcing the software to reinstall - preferred if you think there is something wrong)

If something appears to be wrong, redownload the "configureHomer.pl" script and use "perl configureHomer.pl -install homer" - this fixes a majority of issues.

Hardware Requirements (recommended): 2+ Gb memory (4-8+ Gb), 10+ Gb Hard Drive space (50+ Gb)
Software Requirements: Unix compatible OS (or cygwin), perl, gcc, make, wget, ghostscript, weblogo, blat (see documentation)

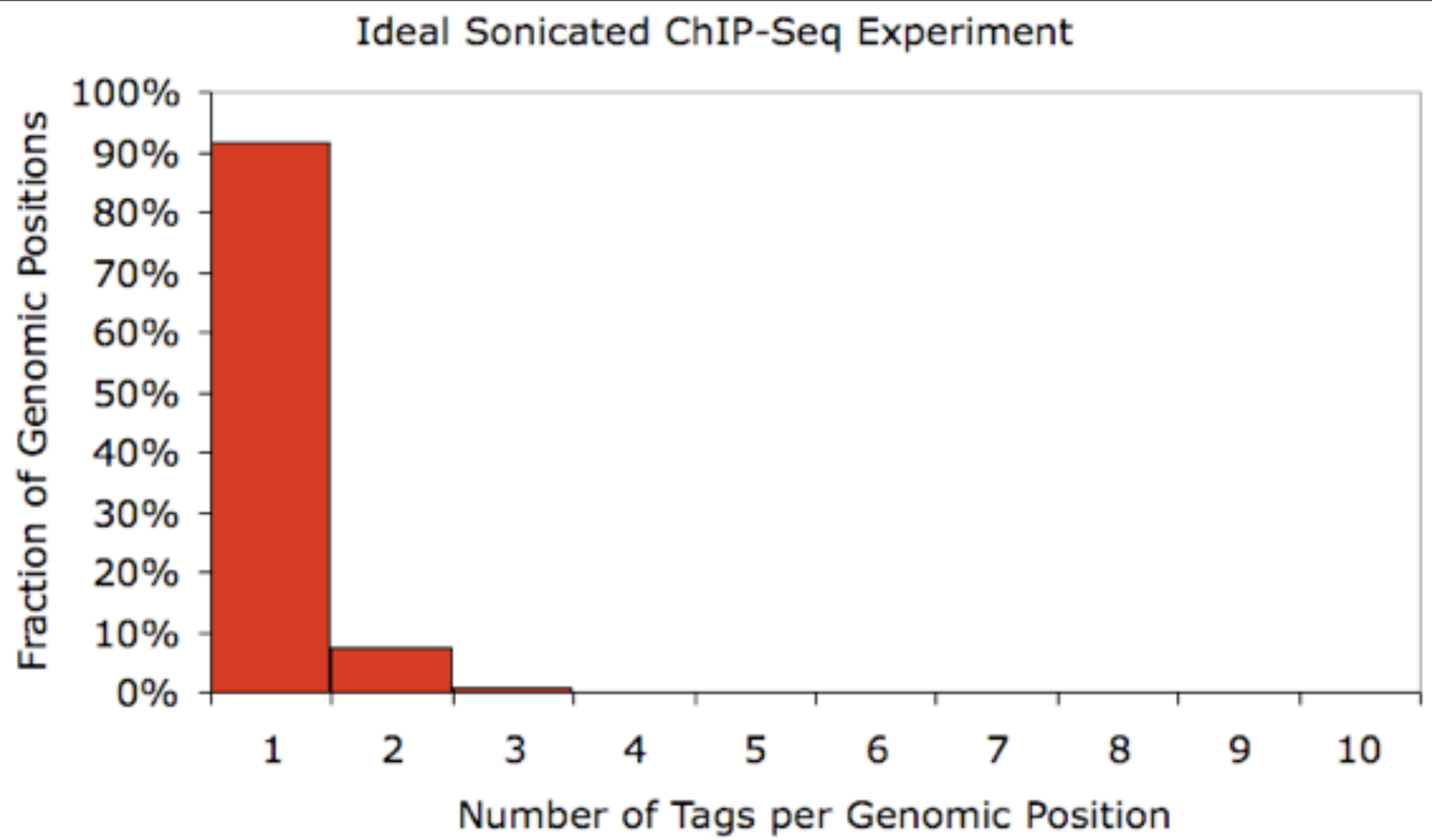
Standard ChIP-Seq analysis with HOMER:

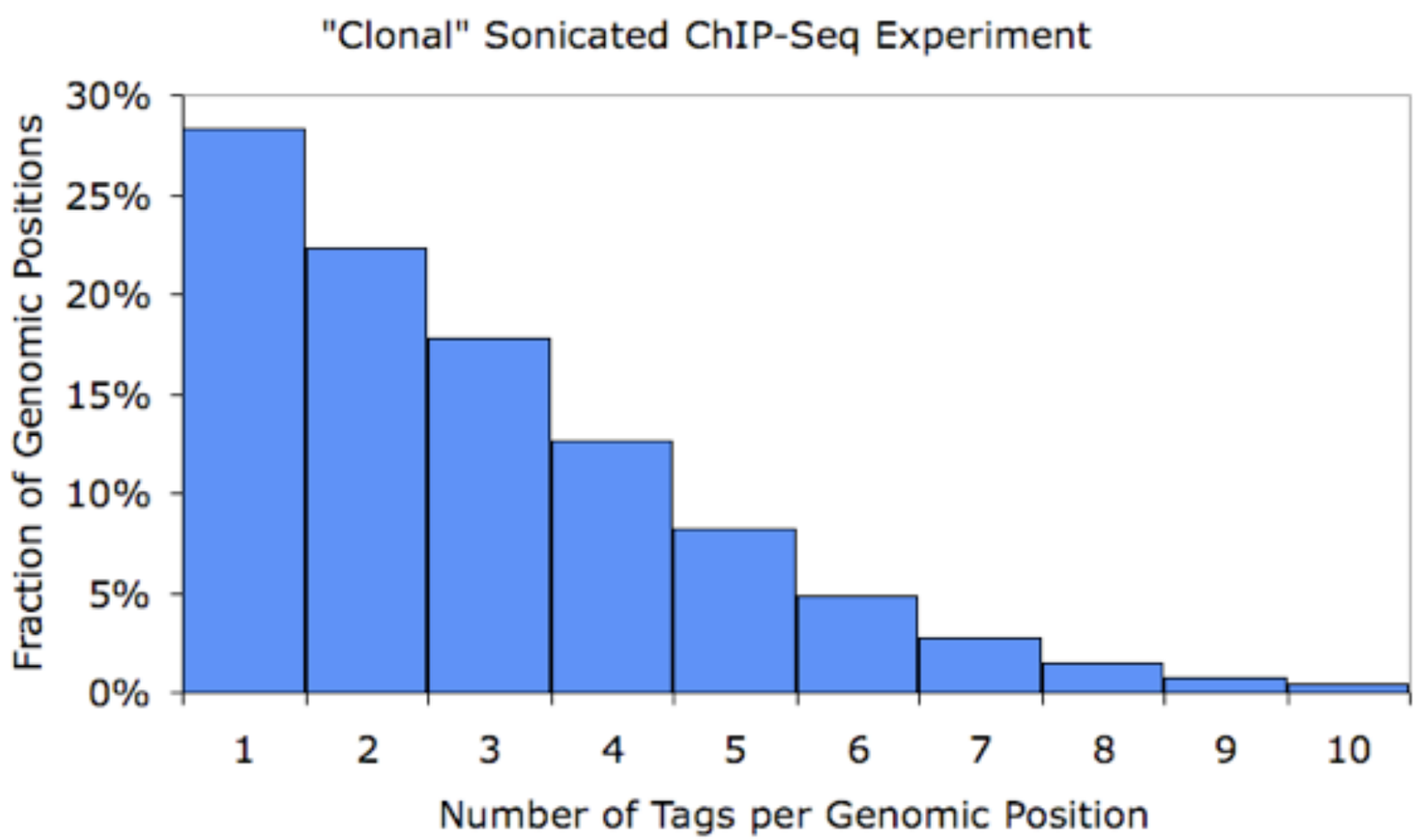
1. Creating a "Tag Directory" from aligned sequences
2. Basic quality control (sequence bias, fragment length estimation)
3. Creating files to view your data in the UCSC Genome Browser
4. Finding Peaks (ChIP-enriched regions) in the genome
5. Finding enriched motifs in ChIP-Seq peaks
6. Annotating Peaks (and cross referencing other experiments and motifs)

   Automating standard ChIP-Seq analysis with analyzeChIP-Seq.pl

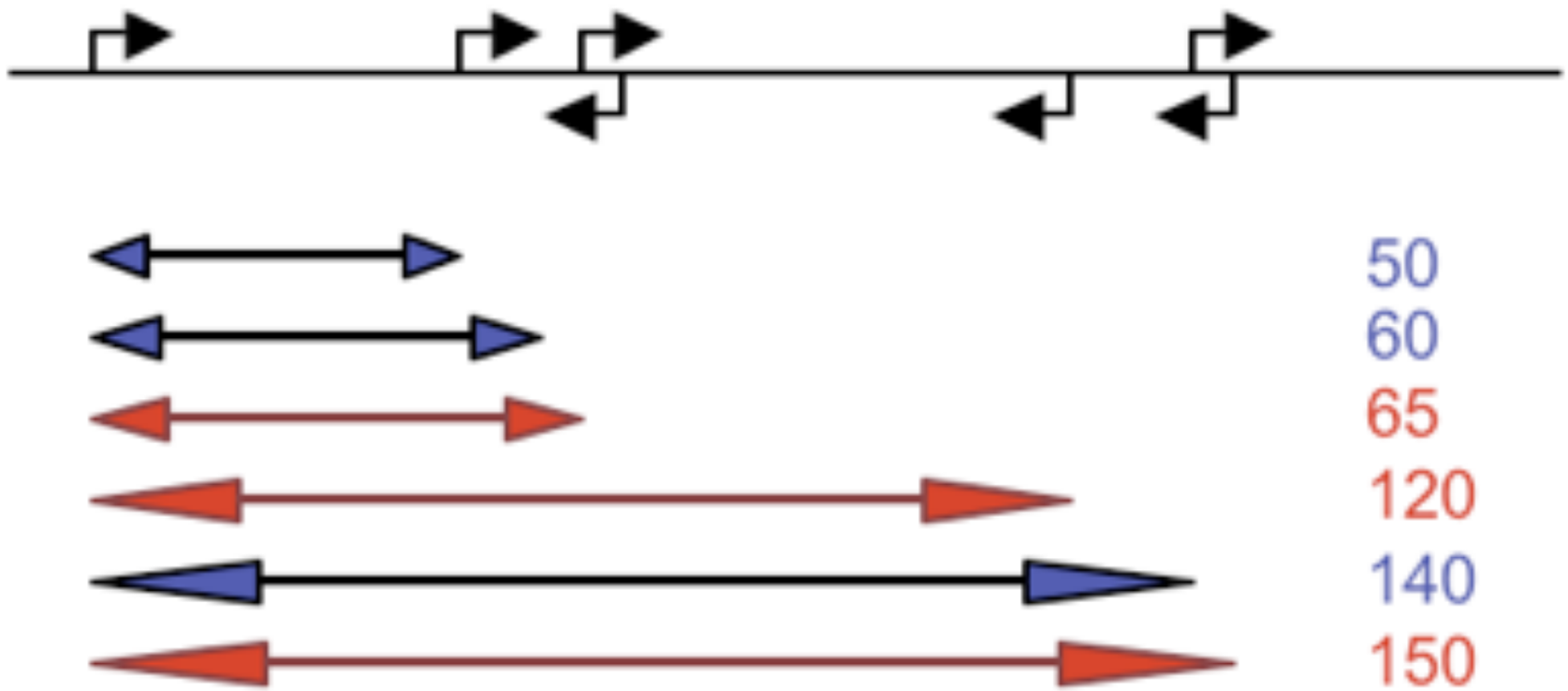
Advanced ChIP-Seq Analysis with HOMER:

- Finding overlapping or differentially bound peaks
- Creating histograms with sequencing data
- Creating heatmaps with sequencing data
- Re-centering peaks on motifs

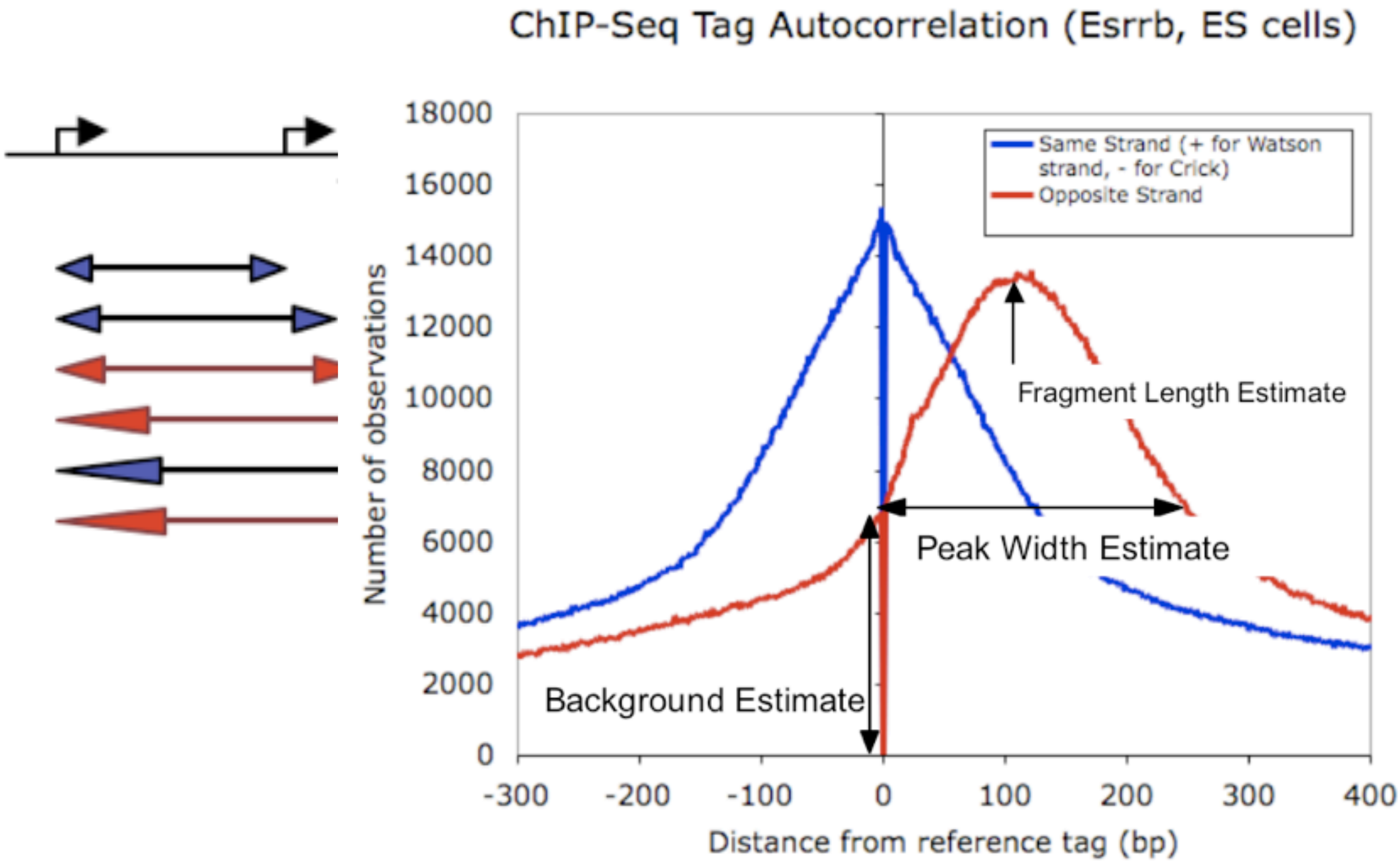

Ideal Sonicated ChIP-Seq Experiment
Fraction of Genomic Positions
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1
2
3
4
5
6
7
8
9
10
Number of Tags per Genomic Position

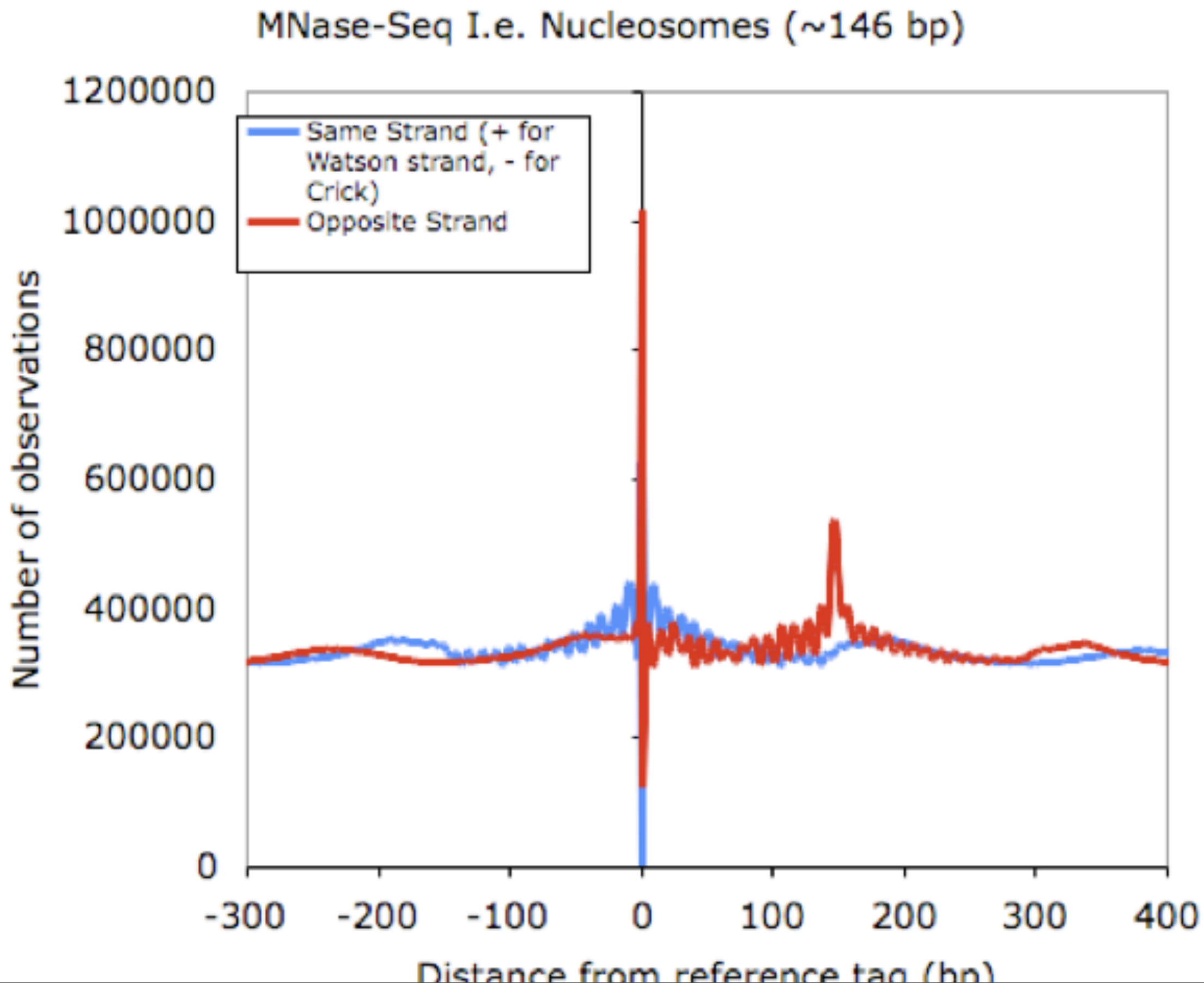

"Clonal" Sonicated ChIP-Seq Experiment
Fraction of Genomic Positions
30%
25%
20%
15%
10%
5%
0%
1
2
3
4
5
6
7
8
9
10
Number of Tags per Genomic Position

Tag Autocorrelation Schematic

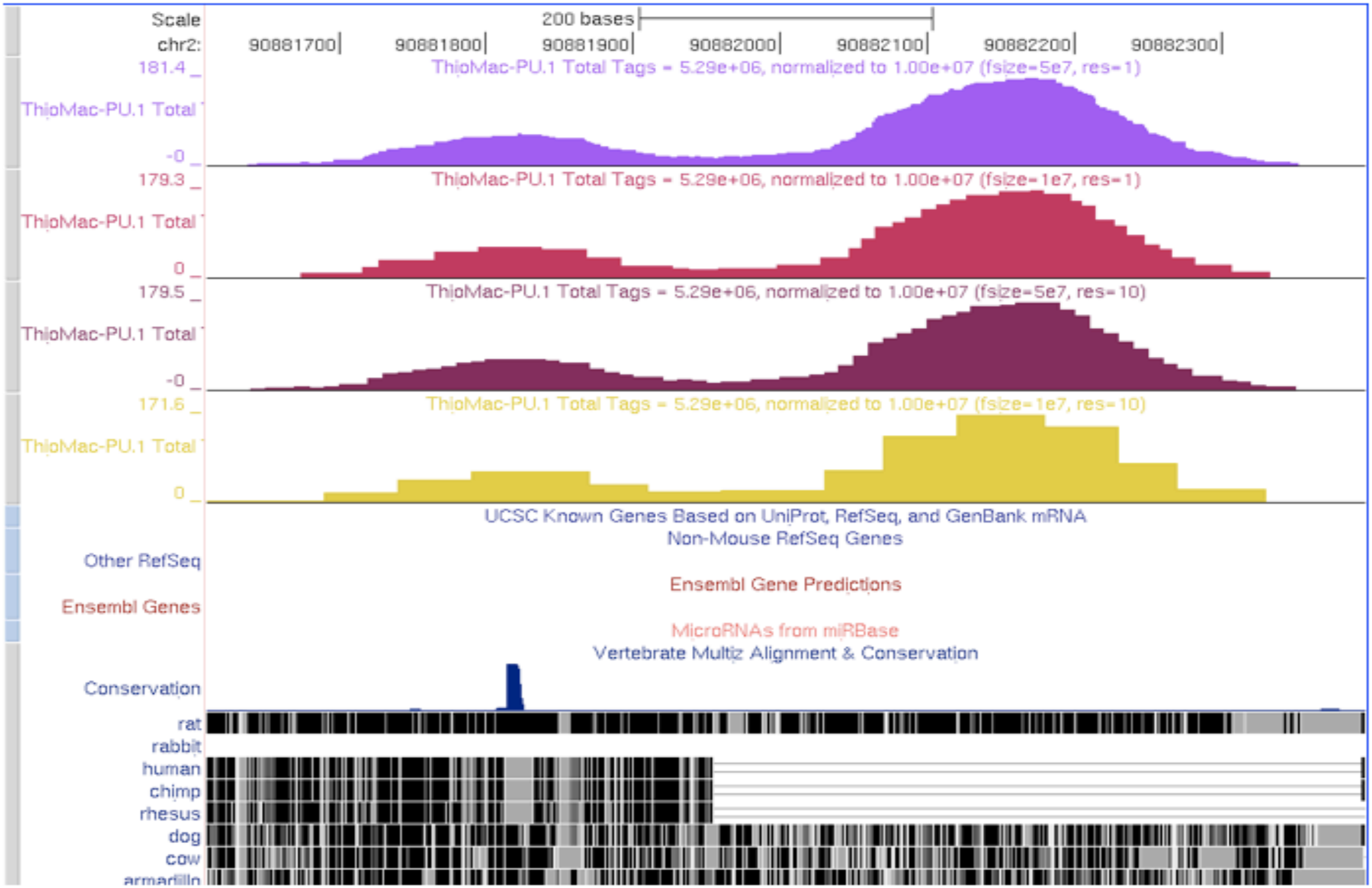


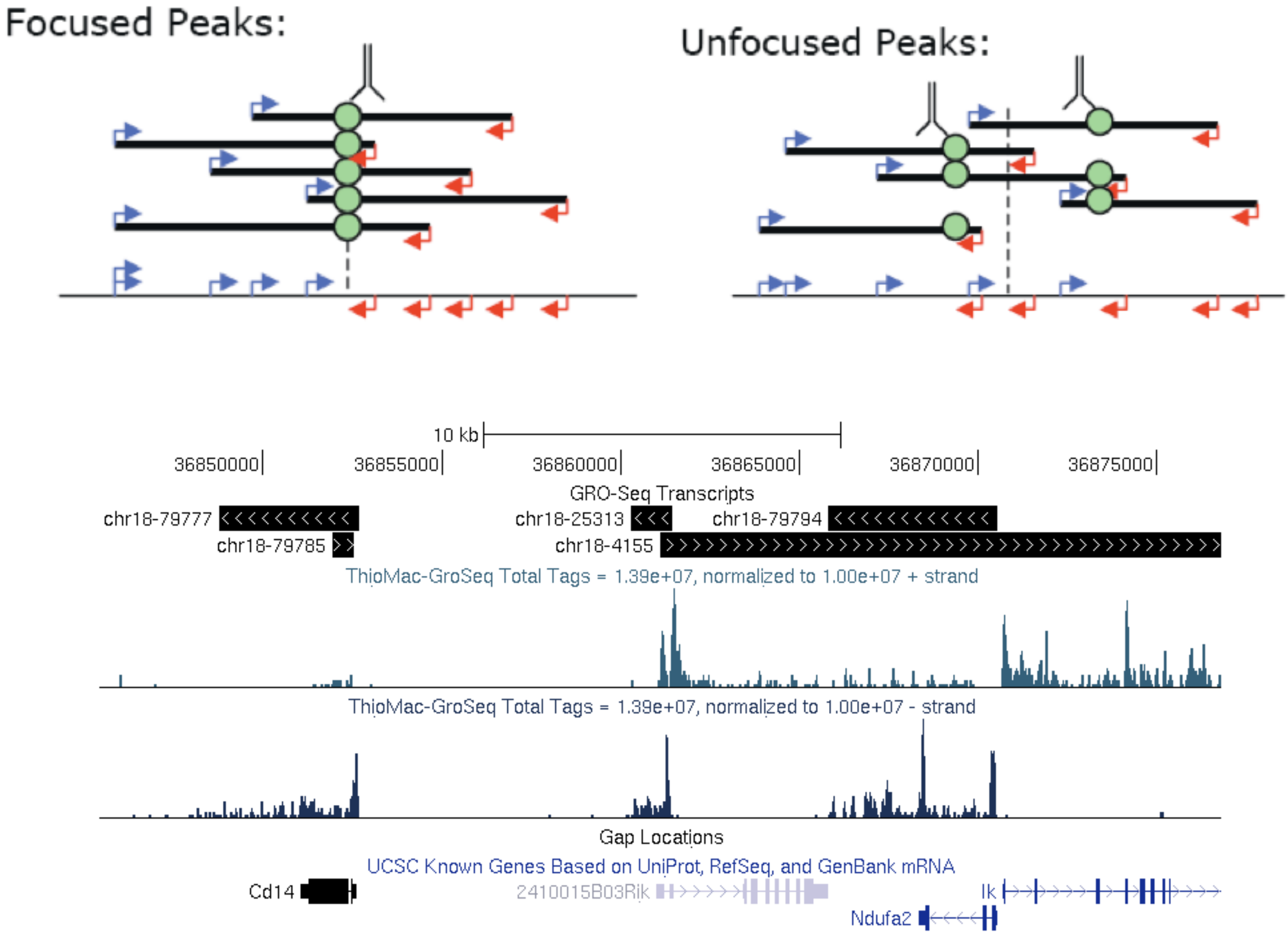
Tag Autocorrelation Schematic

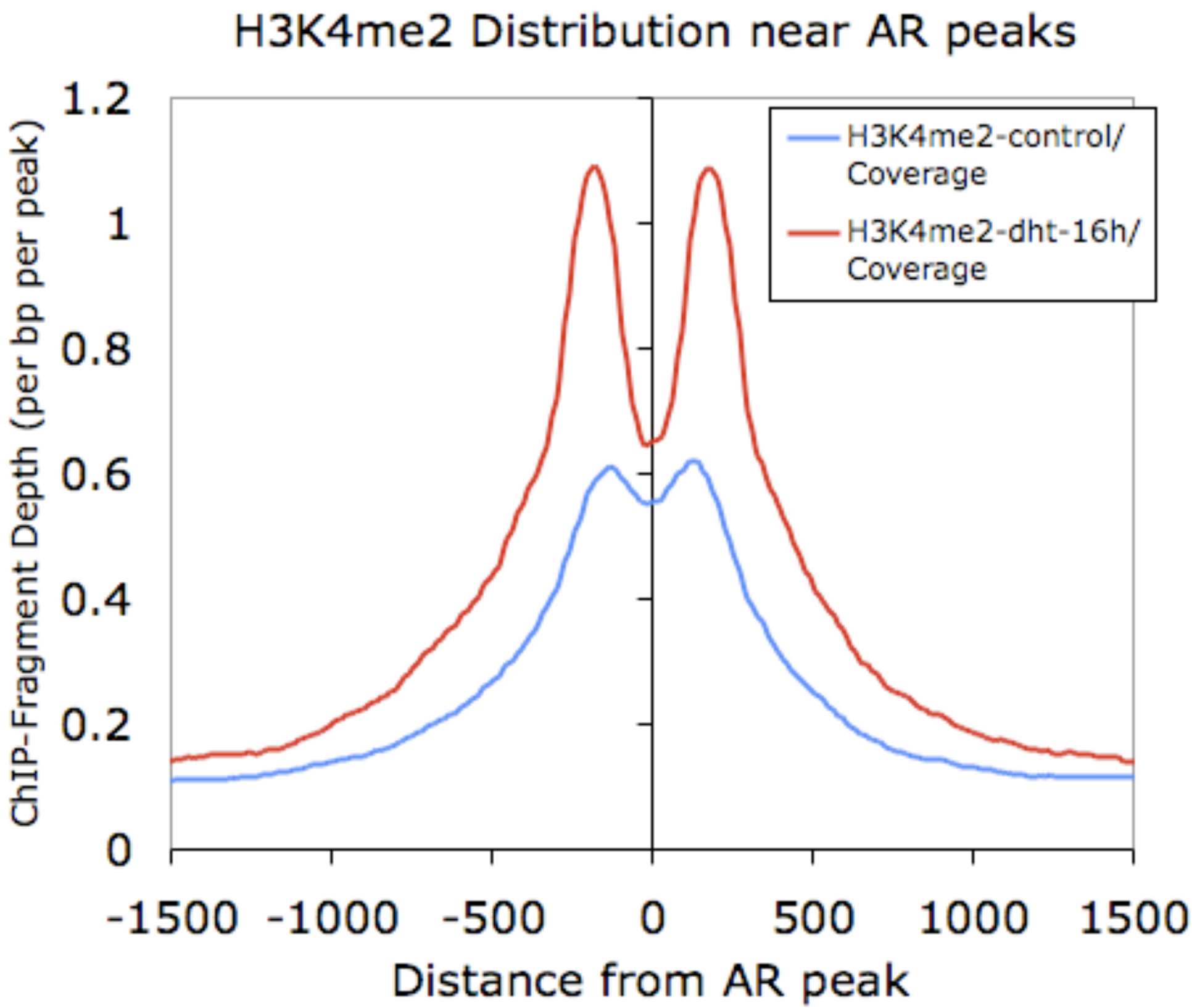


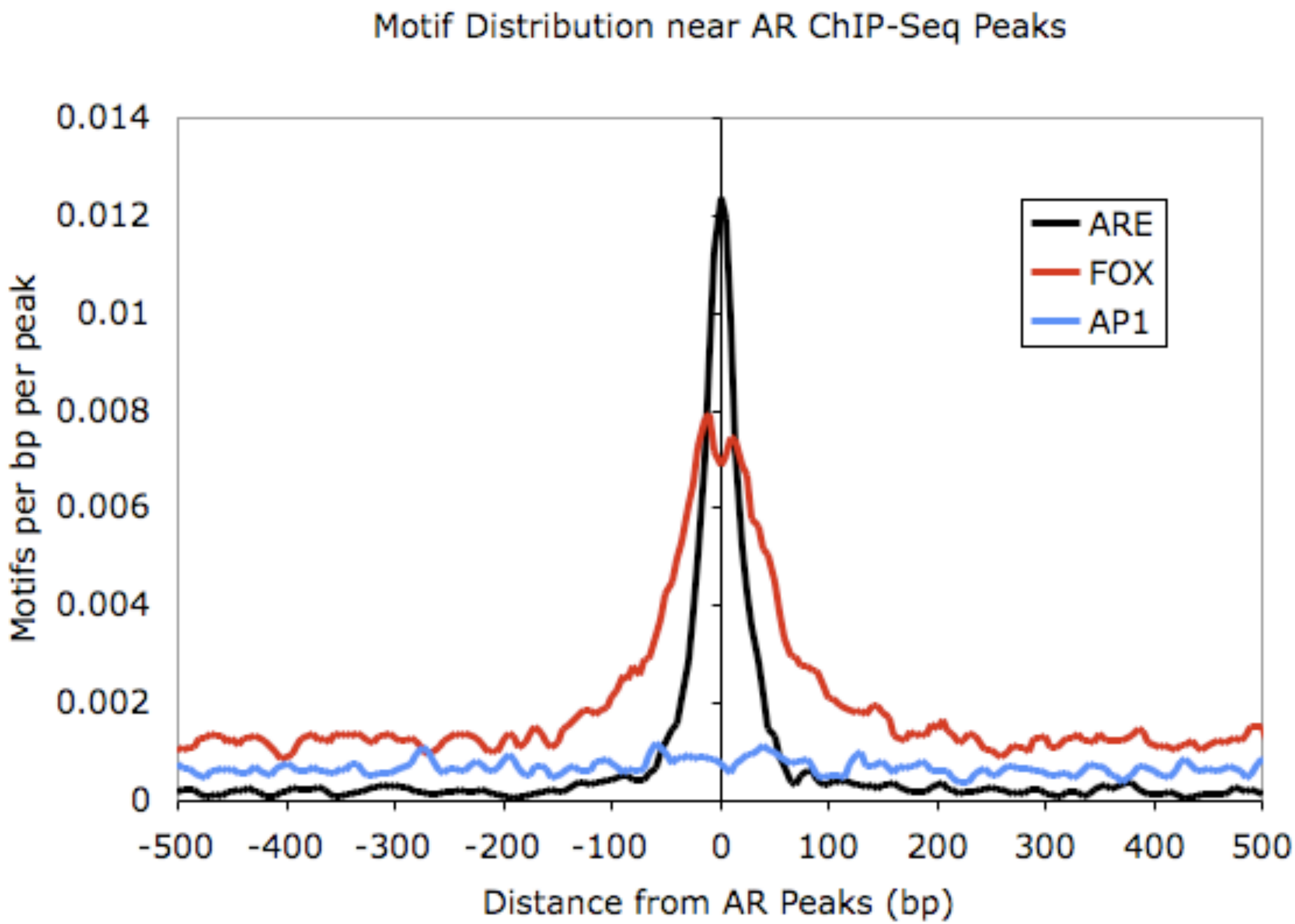

MNase-Seq I.e. Nucleosomes (~146 bp)
Same Strand (+ for Watson strand, - for Crick)
Opposite Strand
Number of observations
1200000
1000000
800000
600000
400000
200000
0
-300
-200
-100
0
100
200
300
400
Distance from reference tag (bp)

Visualization of bedGraph




Focused Peaks:
Unfocused Peaks:
10 kb
36850000
36855000
36860000
36865000
36870000
36875000
GRO-Seq Transcripts
chr18-79777
chr18-25313
chr18-79794
chr18-79785
chr18-4155
ThioMac-GroSeq Total Tags = 1.39e+07, normalized to 1.00e+07 + strand
ThioMac-GroSeq Total Tags = 1.39e+07, normalized to 1.00e+07 - strand
Gap Locations
UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA
Cd14
2410015B03Rik
Ik
Ndufa2


H3K4me2 Distribution near AR peaks
ChIP-Fragment Depth (per bp per peak)
1.2
1
0.8
0.6
0.4
0.2
0
-1500
-1000
-500
0
500
1000
1500
Distance from AR peak
H3K4me2-control/ Coverage
H3K4me2-dht-16h/ Coverage


Motif Distribution near AR ChIP-Seq Peaks
Motifs per bp per peak
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
-500
-400
-300
-200
-100
0
100
200
300
400
500
Distance from AR Peaks (bp)
ARE
FOX
AP1

Tools

- BEDTools
- Peak finding
  - MACS, Peak Finder, CCAAT, FindPeaks
- HOMER
- Galaxy / GeneTrack

Thank you