Epigenetics

and its statistical methods

Hao Feng

Assistant Professor

Dept. of PQHS

JANUARY 18, 2010

Joe Klein:
The CIA's
Afghan Disaster

Yemen: The
New Center
Of Terror

Why the Recession
Hasn't Been Cool
To Teens

TIME

WHY YOUR DNA ISN'T YOUR DESTINY

The new science of epigenetics reveals how the choices you make can change your genes —and those of your kids

BY JOHN CLOUD

$4.95US $5.95CAN

03>

0 70989 10090 9

www.time.com

"The choice we make during our daily lives might ruin our short-term memory or make us fat or hasten death, but they won't affect our genes"

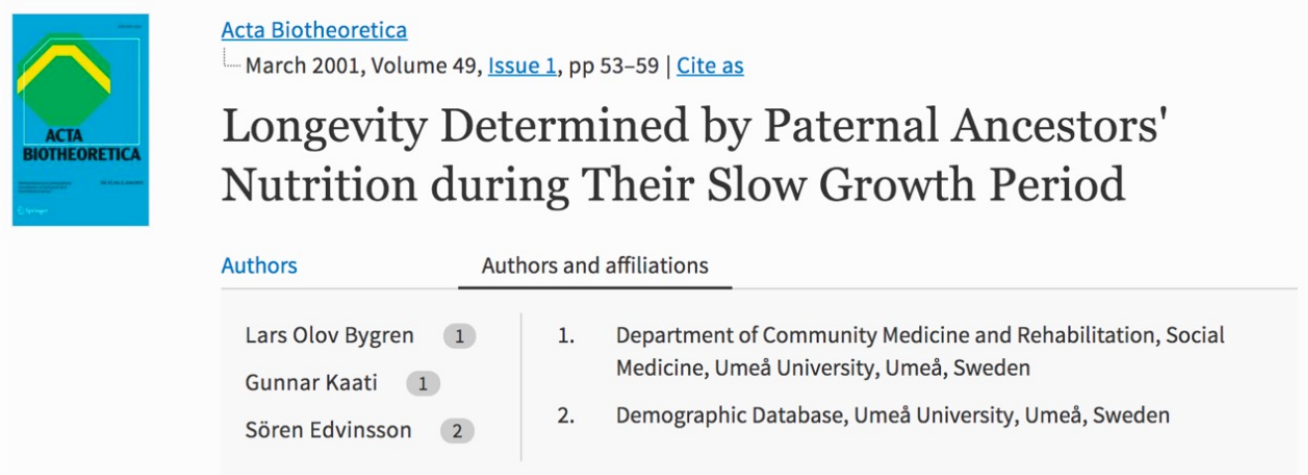
JANUARY 18, 2010
Joe Klein: The CIA's Afghan Disaster
Yemen: The New Center Of Terror
Why the Recession Hasn't Been Cool To Teens
TIME
WHY YOUR DNA ISN'T YOUR DESTINY
The new science of epigenetics reveals how the choices you make can change your genes—and those of your kids
BY JOHN CLOUD
$4.95US $5.95CAN
www.time.com

~~*"The choice we make during our daily lives might ruin our short-term memory or make us fat or hasten death, but they won't affect our genes"*~~

Environment → Epigenetics change

Lars Tunbjork / VU

Three generations: Dr. Lars Olov Bygren, with son Magnus and grandson Ludvig in Stockholm



Acta Biotheoretica
March 2001, Volume 49, Issue 1, pp 53–59 | Cite as

Longevity Determined by Paternal Ancestors' Nutrition during Their Slow Growth Period

Authors | Authors and affiliations

Lars Olov Bygren [1]
Gunnar Kaati [1]
Sören Edvinsson [2]

1. Department of Community Medicine and Rehabilitation, Social Medicine, Umeå University, Umeå, Sweden
2. Demographic Database, Umeå University, Umeå, Sweden

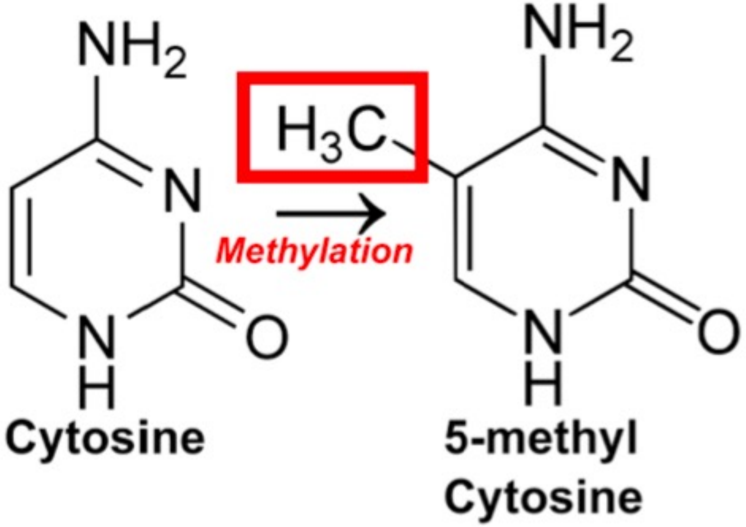
A single winter of overeating as a youngster

⬇

Could lead to shorter life expectancy for one's **grandchildren**

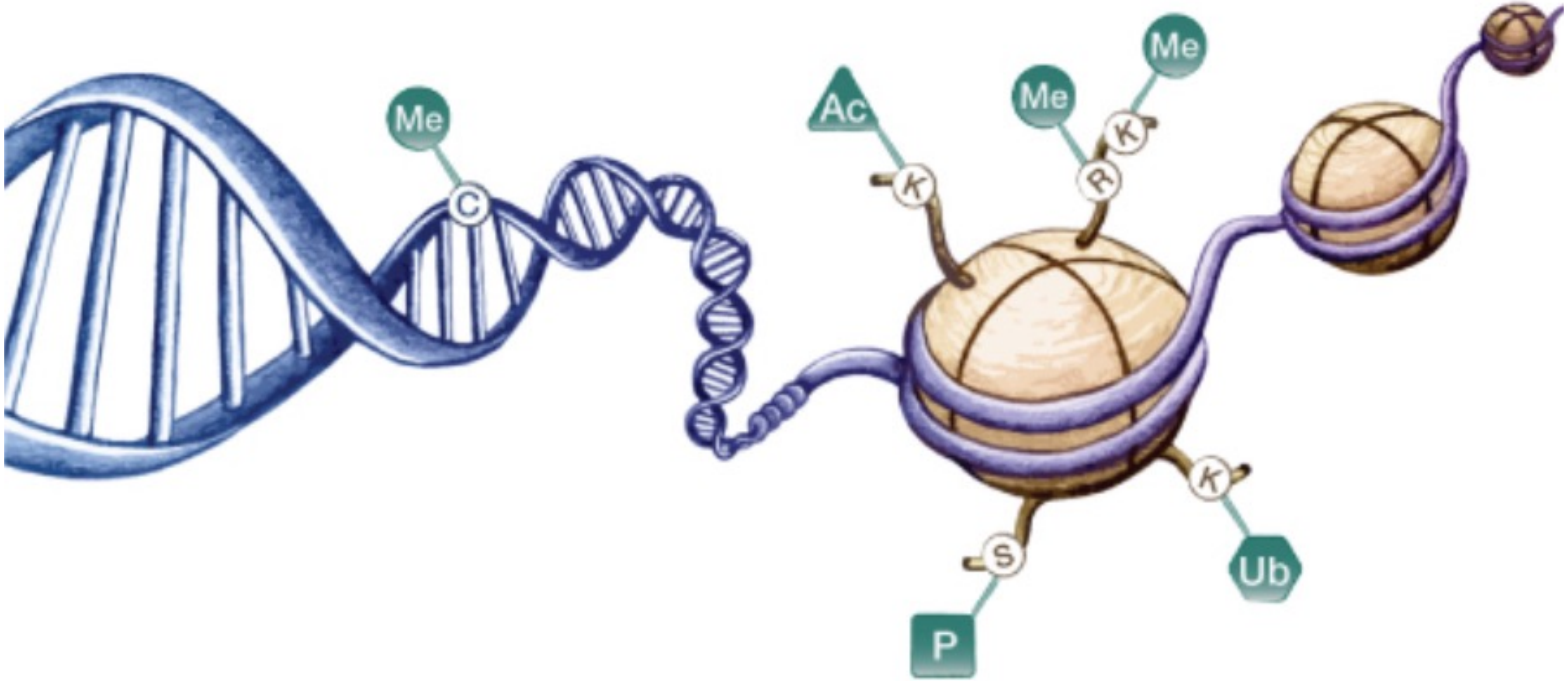
Epi- [*Greek*]: ‘on the top of’, ‘above’

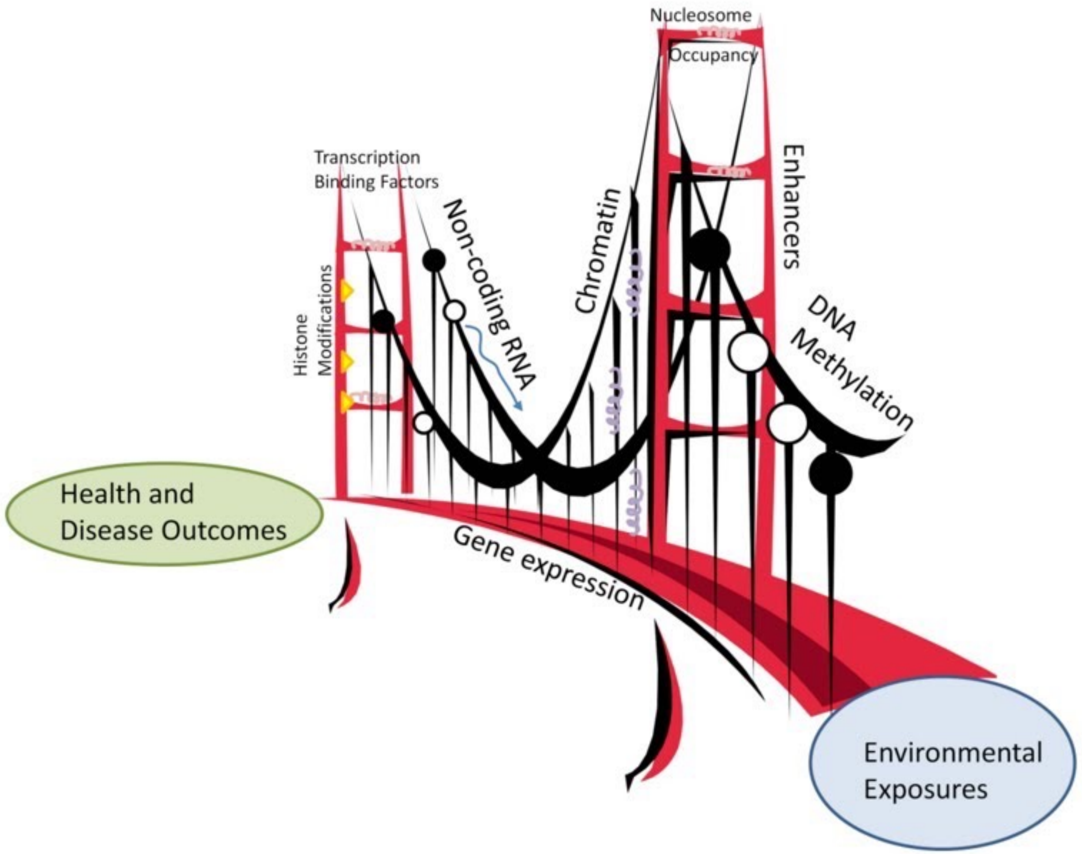
Epi- [*Greek*]: ‘on the top of’, ‘above’



Epigenetics: (heritable) changes on genetics that do NOT involve changes to the underlying DNA sequence.

Epigenetics: (heritable) changes on genetics that do NOT involve changes to the underlying DNA sequence.




Nucleosome
Occupancy
Transcription
Binding Factors
Non-coding RNA
Chromatin
Enhancers
Histone
Modifications
DNA
Methylation
Health and
Disease Outcomes
Gene expression
Environmental
Exposures

Epigenetics signals (1)

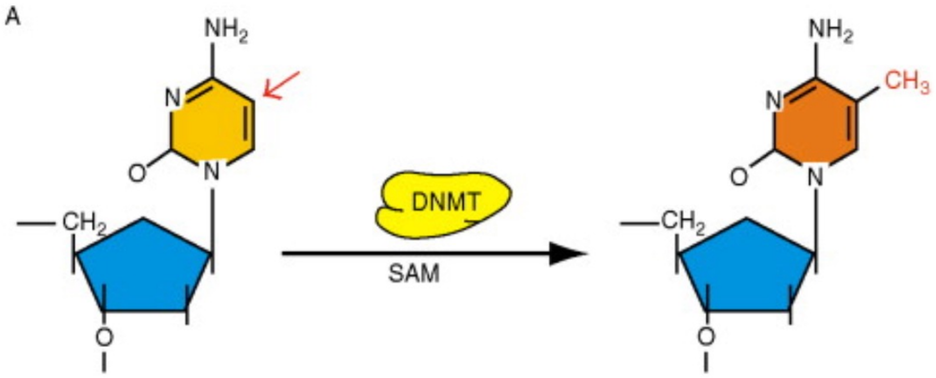
- DNA methylation
- Protein binding on DNA
- Histone modification
- Chromatin accessibility
- Nucleosome occupancy
- …

Epigenetics signals (1)

- DNA methylation
- Protein binding on DNA
- Histone modification
- Chromatin accessibility
- Nucleosome occupancy
- …

DNA Methylation

An epigenetic modification of the DNA sequence: adding a methyl group to the 5 position of cytosine (5mC)

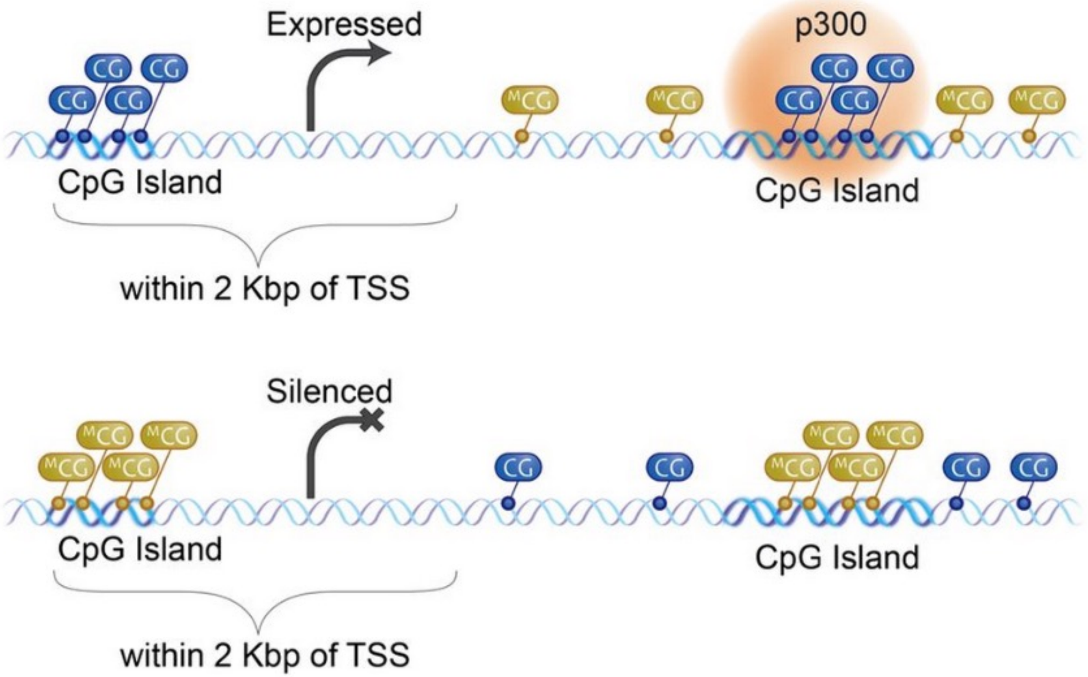


Primarily happens at **CpG sites** (C followed by a G), although non-CG methylation exists

Advances in Genetics Volume 70 2010 27 - 56

http://dx.doi.org/10.1016/B978-0-12-380866-0.60002-2

DNA Methylation



Varley K E et al. Genome Res. 2013;23:555-567

Methylation of CpG islands in/near promoter region of gene can silence gene expression

Function of DNA methylation

- Important in gene regulation
  - Methylation of promoter regions can suppress gene expression
- Plays crucial role in cell development
  - Heritable during cell division
  - Helps cells establish identity during cell/tissue differentiation
- Can be influenced by environment
  - Good candidate to mediate GxE interactions

Sequencing approaches for DNA methylation

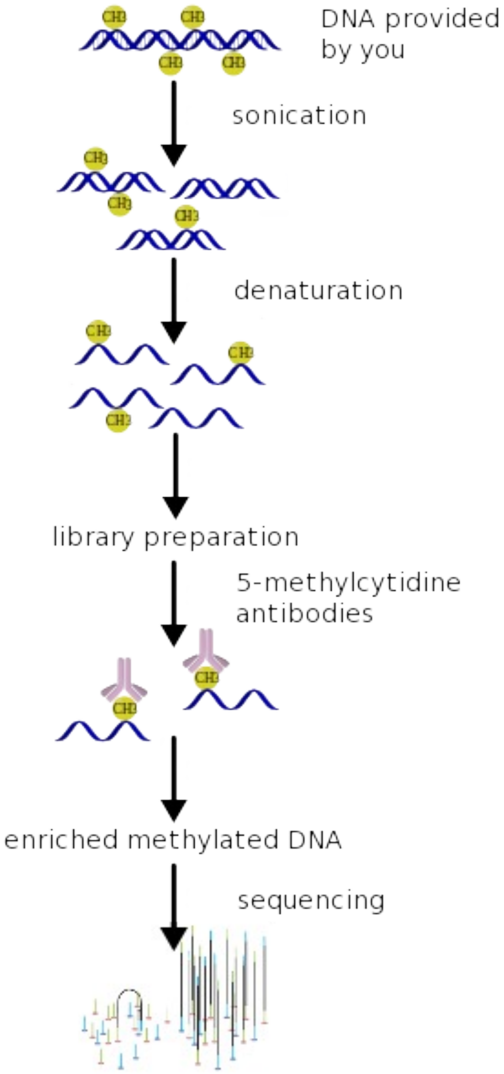
- Capture-based or enrichment-based sequencing
  - Use methyl-binding proteins or antibodies to capture methylated DNA fragments, then sequence fragments
  - **Resolution is low**: can typically quantify the amount of DNA methylation in 100-200 bp regions

Capture-based or enrichment-based sequencing

Two-Steps:

1. Capture of methylated DNA region
2. Sequencing

- MeDIP-seq (Methylated DNA ImmunoPrecipitation)[1]
  - – uses antibody against methylated DNA
  - – Assesses relative rather than absolute methylation levels
  - – MEDIPS[2] is a popular tool for analysis
- Other similar approaches: MBD-seq[3,] MIRA-seq[4], methylCap-seq[5,] MRE-seq[6]



[1] Weber et al. (2005) *Nat Genet;* [2] Chavez et al. (2010) *Gen Res;* [3] Serre et al. (2010) *NAR;* [4] Rauch et al. (2010) *Methods*; [5] Brinkman et al. (2010) *Methods;* [6] Maunakea et al. (2010) *Nature*

Sequencing approaches for DNA methylation

- Capture-based or enrichment-based sequencing
  - Use methyl-binding proteins or antibodies to capture methylated DNA fragments, then sequence fragments
  - **Resolution is low**: can typically quantify the amount of DNA methylation in 100-200 bp regions
- Bisulfite-conversion-based sequencing
  - Bisulfite treatment converts unmethylated C’s to T’s
  - Sequencing converted data gives single-bp resolution
  - Can measure methylation status of each CpG site
  - Until recently, not possible to distinguish 5mC from 5hmC
- Nowadays: bisulfite sequencing

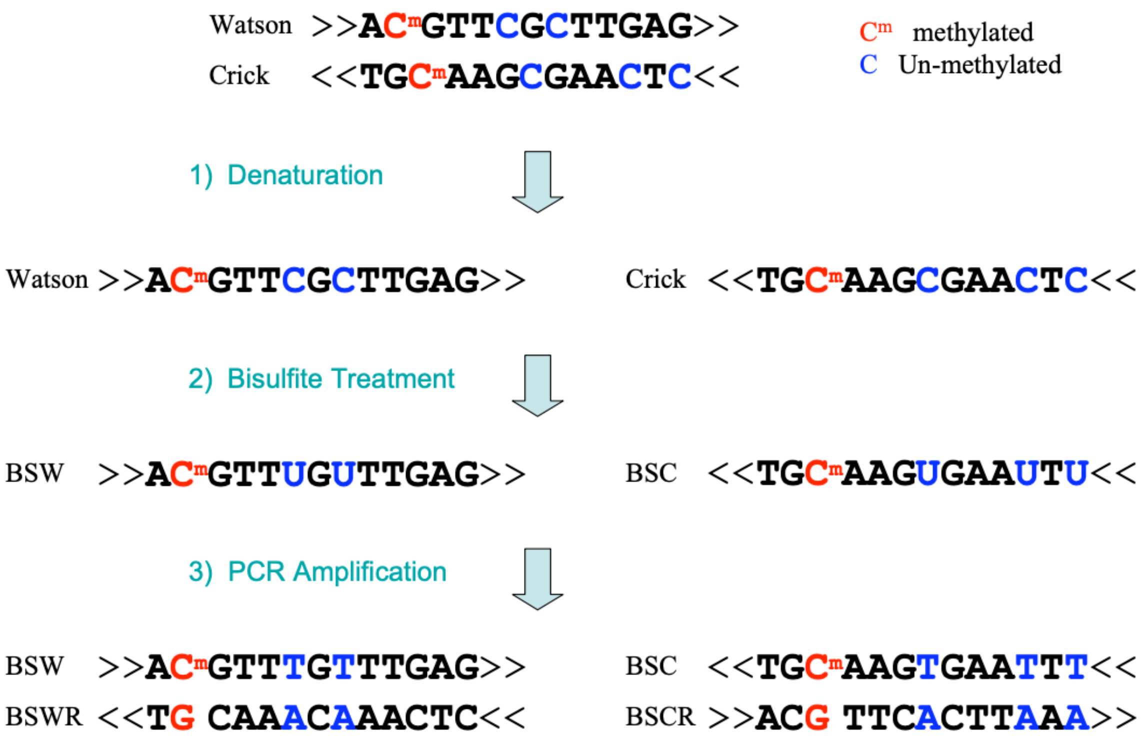
Sequencing approaches for DNA methylation

- Capture-based or enrichment-based sequencing
  - Use methyl-binding proteins or antibodies to capture methylated DNA fragments, then sequence fragments
  - **Resolution is low**: can typically quantify the amount of DNA methylation in 100-200 bp regions
- Bisulfite-conversion-based sequencing
  - Bisulfite treatment converts unmethylated C’s to T’s
  - Sequencing converted data gives single-bp resolution
  - Can measure methylation status of each CpG site
  - Until recently, not possible to distinguish 5mC from 5hmC
- Nowadays: bisulfite sequencing (BS-seq or WGBS)

Bisulfite sequencing (BS-seq)

- Technology in a nutshell:
  - Treat fragmented DNA with bisulfite
    - Unmethylated C will be converted to U, amplified as T $C \rightarrow T$
    - Methylated C will be protected and remain C $C^m \rightarrow C$
    - No change for other bases
  - Amplify the treated DNA
  - Sequence the DNA segments
  - Align sequence reads to genome

Bisulfite sequencing (BS-seq)



Watson >>AC^{m}GTTCGCTTGAG>>

Crick <<TGC^{m}AAGCGAACTC<<

C^m methylated

C Un-methylated

1) Denaturation

Watson >>AC^{m}GTTCGCTTGAG>>

Crick <<TGC^{m}AAGCGAACTC<<

2) Bisulfite Treatment

BSW >>AC^{m}GTTUGUTTGAG>>

BSC <<TGC^{m}AAGUGAAUTU<<

3) PCR Amplification

BSW >>AC^{m}GTTTGTTTGAG>>

BSWR <<TG CAAACAAACTC<<

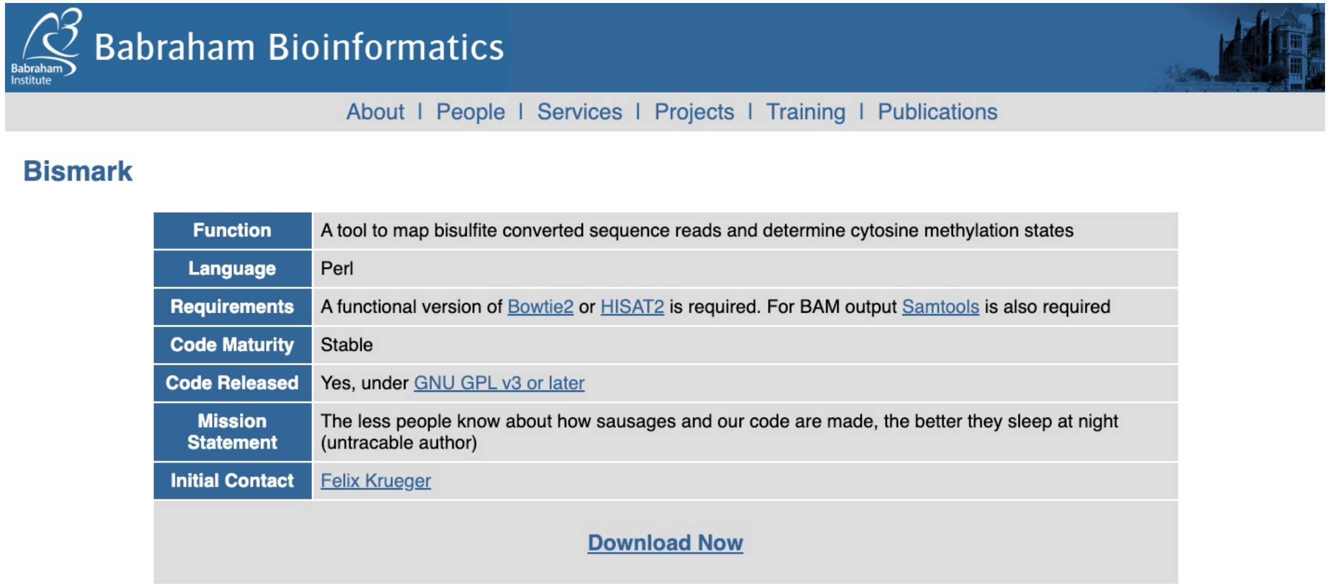
BSC <<TGC^{m}AAGTGAATTT<<

BSCR >>ACG TTCACTTAAA>>

Xi and Li (2009) *BMC Bioinformatics*

BS-seq alignment software

- Bismark
  - Faster than other programs
  - User-friendly in terms of extracting data, interfacing with other software

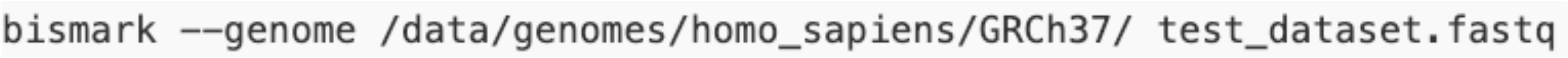

Babraham Institute
Babraham Bioinformatics
About | People | Services | Projects | Training | Publications

Bismark

Function	A tool to map bisulfite converted sequence reads and determine cytosine methylation states
Language	Perl
Requirements	A functional version of Bowtie2 or HISAT2 is required. For BAM output Samtools is also required
Code Maturity	Stable
Code Released	Yes, under GNU GPL v3 or later
Mission Statement	The less people know about how sausages and our code are made, the better they sleep at night (untracable author)
Initial Contact	Felix Krueger

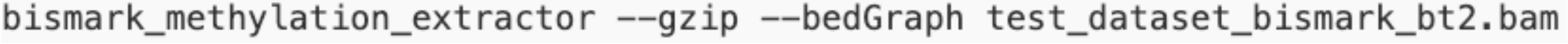
Bismark usage

1. Mapping



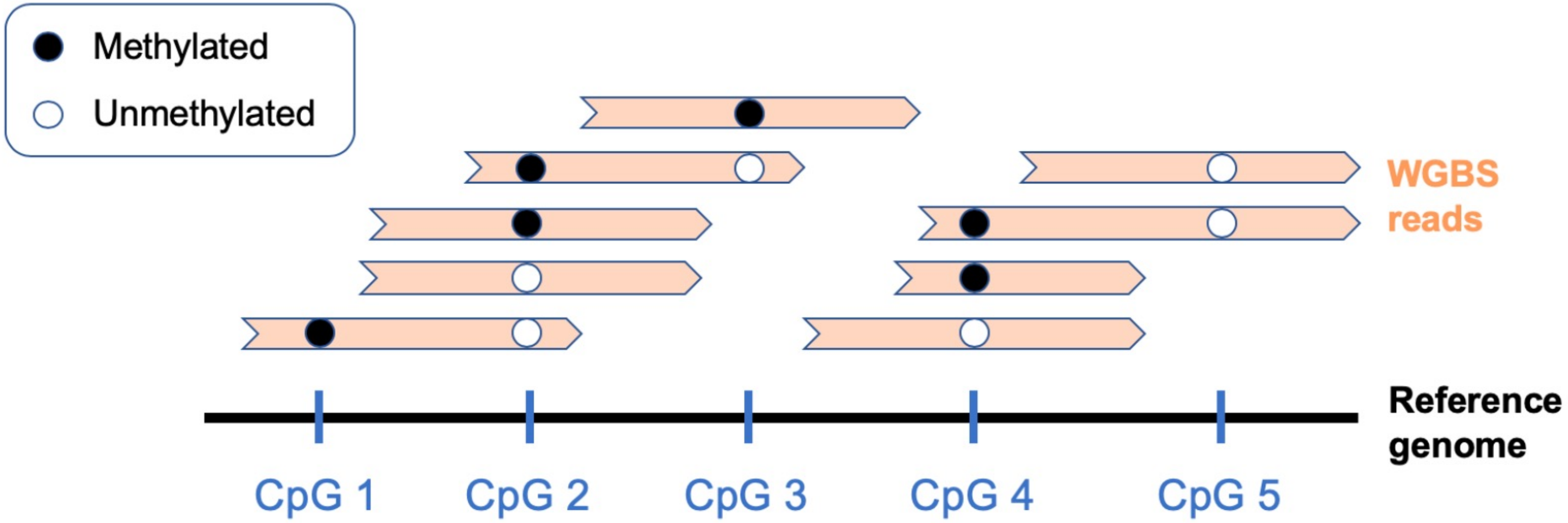
```
bismark --genome /data/genomes/homo_sapiens/GRCh37/ test_dataset.fastq
```

2. Methylation data extraction



```
bismark_methylation_extractor --gzip --bedGraph test_dataset_bismark_bt2.bam
```

BS-seq alignment summary



Methylated counts (X)	1	2	1	2	0
Coverage (N)	1	4	2	3	2
Methylation level (X/N)	1	0.5	0.5	0.67	0

WGBS data

BS-seq extracted data summary

- At each position, we have the total number of reads, and the methylated number of reads:

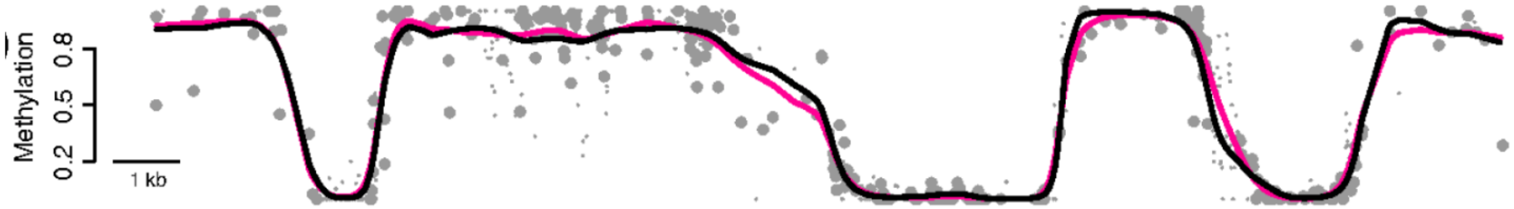
Position of CpG site		Total # reads	# methylated reads
chr1	3010874	22	18
chr1	3010894	31	27
chr1	3010922	12	10
chr1	3010957	7	6
chr1	3010971	6	6
chr1	3011025	7	5

Study design for BS-seq studies

- High costs → few samples typically analyzed
- Two common study designs
  - Analysis of a single sample:
    - Goal: observe methylation patterns across genome
    - Commonly done to **characterize methylome** for a particular cell type or species
  - Comparison of several samples:
    - Typical goal: compare methylation levels between groups
    - **Differential methylation analysis**
    - Compared with ChIP-seq and RNA-seq, methods are still in early stage, and are often *ad hoc*

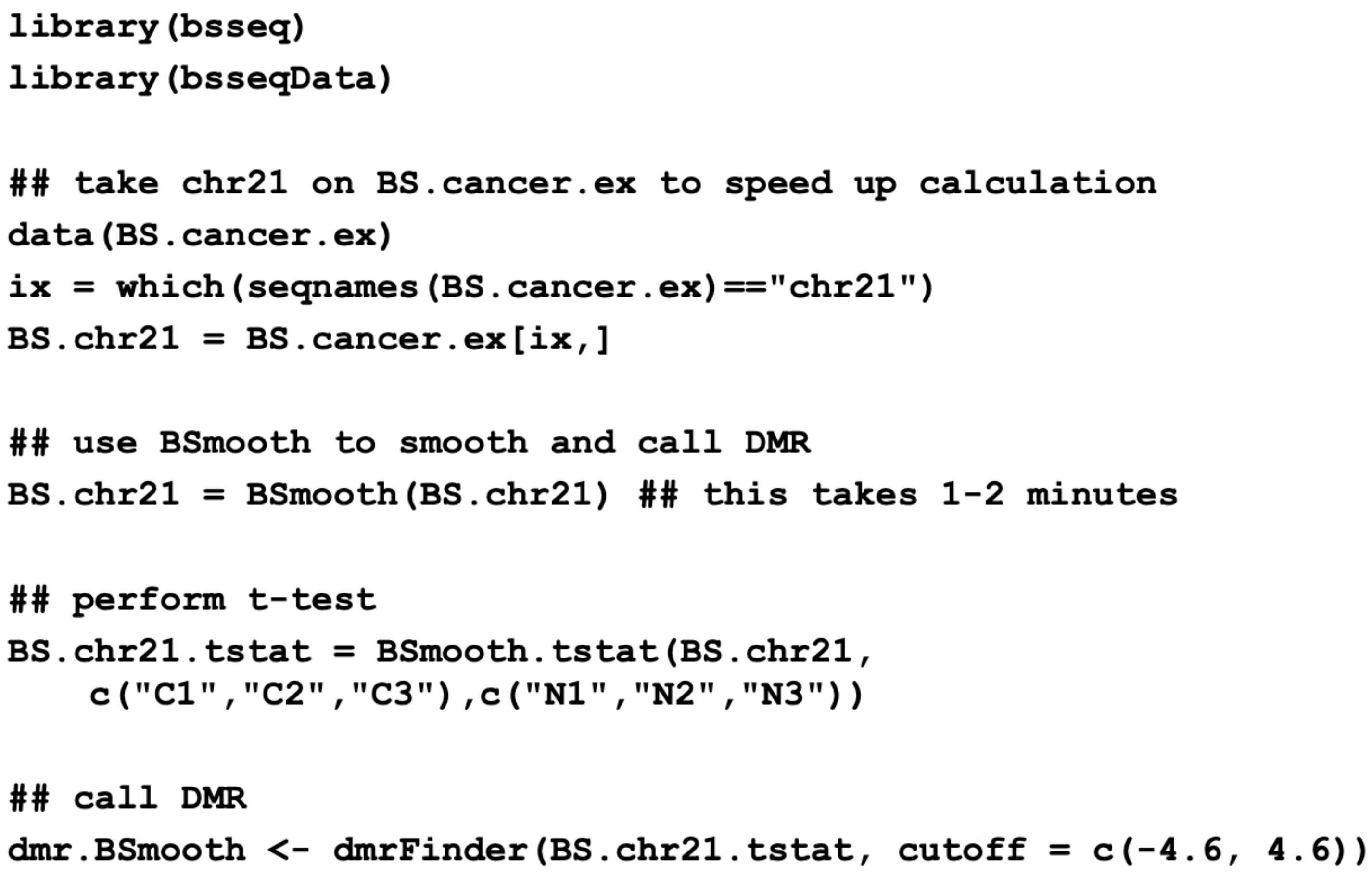
Single sample analysis: smoothing

- By borrowing information across sites, can achieve high precision even with low coverage
  - Pink line is from smoothing full 30x data
  - Black line is from smoothing 5x version of data
  - Correlation = .90 across entire dataset
  - Median absolute difference of .056



Hansen *et al.* 2012 ***Genome Biology***

Bioconductor package: bsseq



```
library(bsseq)
library(bsseqData)

## take chr21 on BS.cancer.ex to speed up calculation
data(BS.cancer.ex)
ix = which(seqnames(BS.cancer.ex)=="chr21")
BS.chr21 = BS.cancer.ex[ix,]

## use BSmooth to smooth and call DMR
BS.chr21 = BSmooth(BS.chr21) ## this takes 1-2 minutes

## perform t-test
BS.chr21.tstat = BSmooth.tstat(BS.chr21,
    c("C1","C2","C3"),c("N1","N2","N3"))

## call DMR
dmr.BSmooth <- dmrFinder(BS.chr21.tstat, cutoff = c(-4.6, 4.6))
```

Multiple sample analysis: differential methylation

- Goal: identify **differentially methylated regions** (DMRs) between groups.
  - BS-seq data from cancer patients
  - BS-seq data from healthy controls
  - Find the genomic regions that have methylation difference!!!

Multiple sample analysis: differential methylation

- If we have only one sample per group (no biological replicates), Fisher’s exact test is a natural choice
- Example: single CpG site sequenced for 2 samples
  - For tumor sample, 32/44 methylated reads
  - For normal sample, 8/12 methylated reads
- Can then perform Fisher’s exact test on the following table:
- OR = 1.33
- p = .73

	Methylated	Unmeth.	Total reads
Tumor	32	12	44
Normal	8	4	12
Total	40	16	56

Multiple sample analysis: differential methylation

Naïve t-test

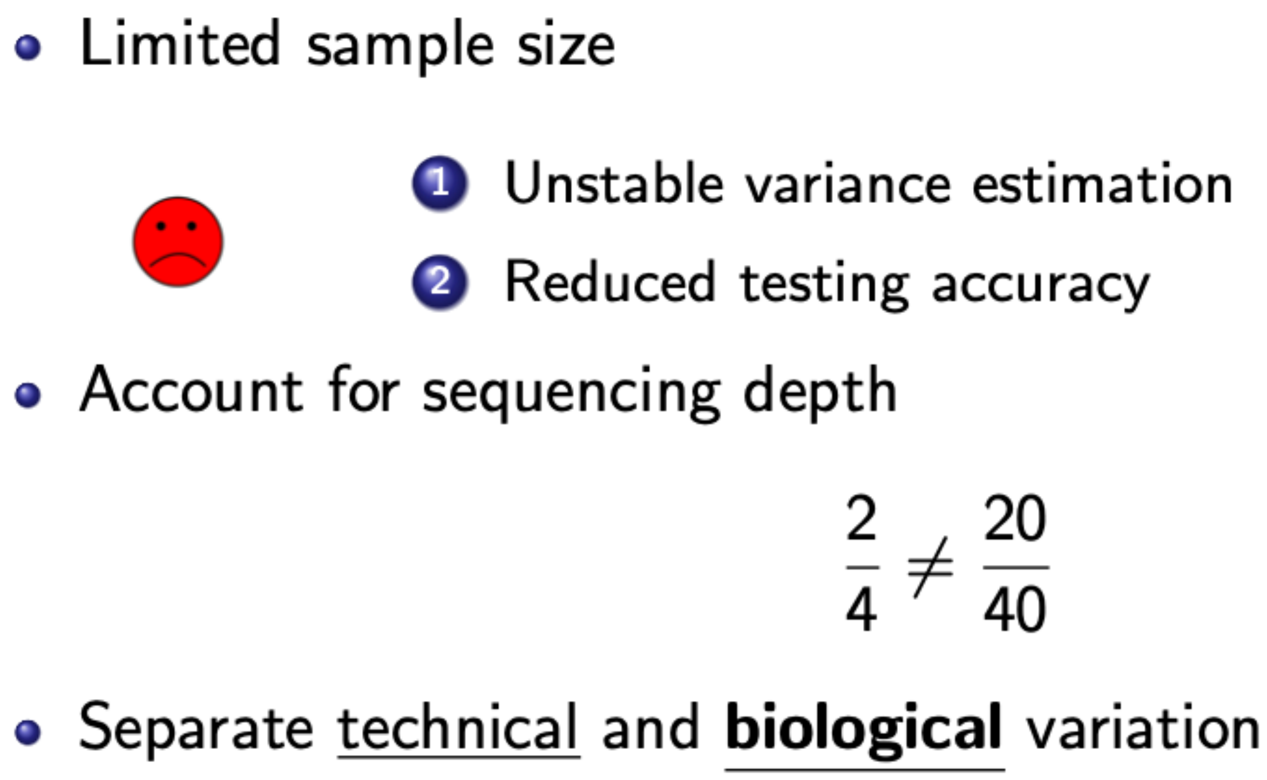
- Example: single CpG site sequenced for 4 samples
  - For 2 tumor samples, 32/44 and 4/10 methylated reads
  - For 2 normal samples, 8/12 and 12/34 methylated reads
- For t-test, compute a proportion for each sample
  - .727 and .400 for tumor samples
  - .667 and .353 for normal samples
- Difference in mean proportions = .563 - .510 = .053
- T-statistic = 0.2375
- p = .834

Multiple sample analysis: differential methylation

- Why Fisher’s and *t*-test are not good choices?

Multiple sample analysis: differential methylation

- Why Fisher’s and *t*-test are not good choices?



- Limited sample size

  ☹

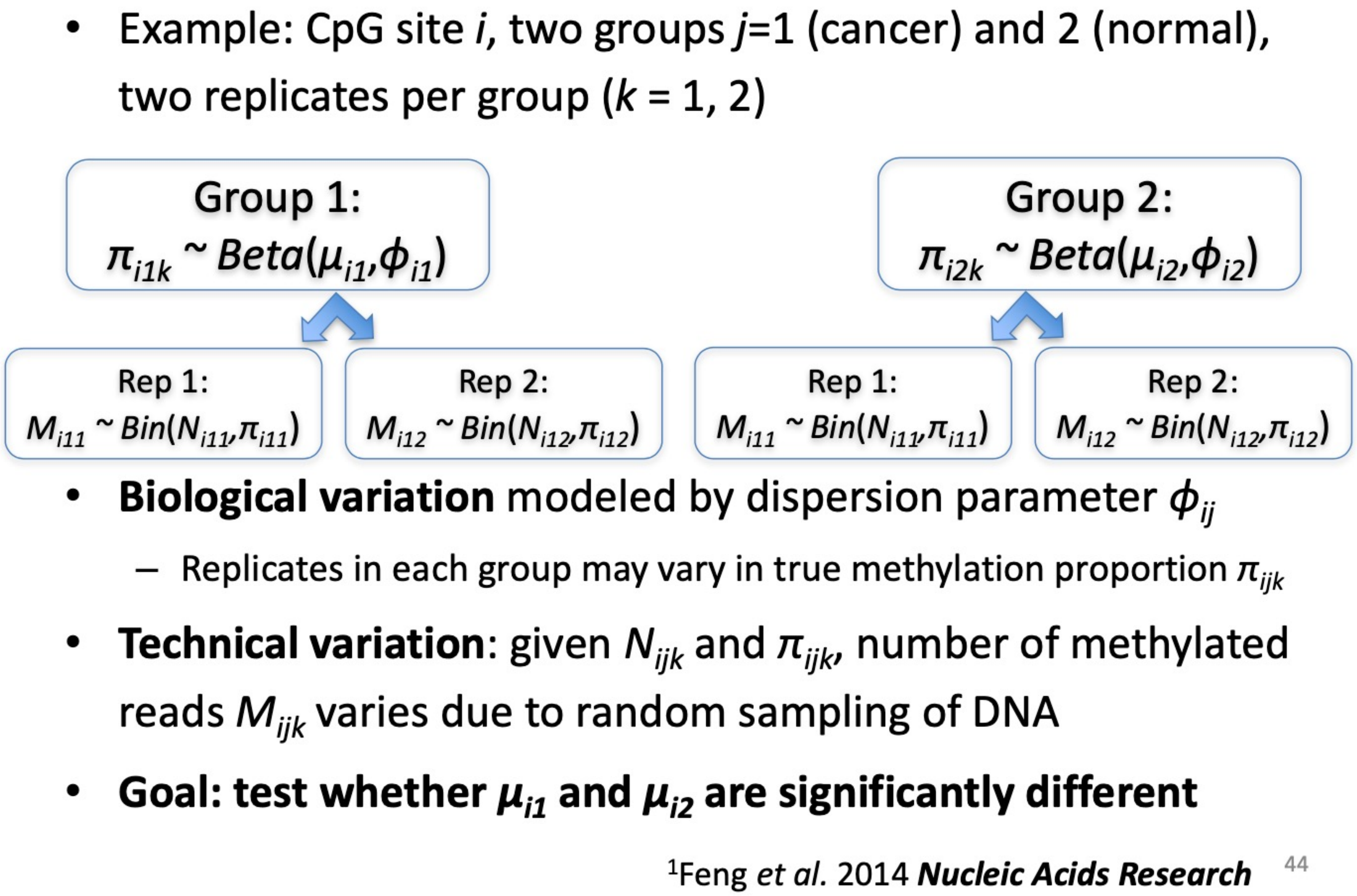
  1. Unstable variance estimation
  2. Reduced testing accuracy

- Account for sequencing depth

$$\frac{2}{4} \neq \frac{20}{40}$$

- Separate technical and **biological** variation

Beta-binomial hierarchical model



- Example: CpG site *i*, two groups *j*=1 (cancer) and 2 (normal), two replicates per group (*k* = 1, 2)

Group 1:
$\pi_{i1k} \sim Beta(\mu_{i1},\phi_{i1})$

Group 2:
$\pi_{i2k} \sim Beta(\mu_{i2},\phi_{i2})$

Rep 1:
$M_{i11} \sim Bin(N_{i11},\pi_{i11})$

Rep 2:
$M_{i12} \sim Bin(N_{i12},\pi_{i12})$

Rep 1:
$M_{i11} \sim Bin(N_{i11},\pi_{i11})$

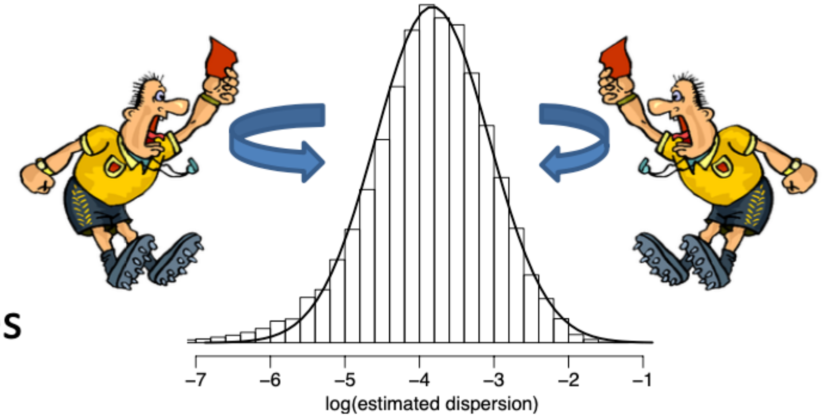
Rep 2:
$M_{i12} \sim Bin(N_{i12},\pi_{i12})$

- **Biological variation** modeled by dispersion parameter ϕ_{ij}
  - Replicates in each group may vary in true methylation proportion π_{ijk}
- **Technical variation**: given N_{ijk} and π_{ijk}, number of methylated reads M_{ijk} varies due to random sampling of DNA
- **Goal: test whether μ_{i1} and μ_{i2} are significantly different**

[1]Feng *et al.* 2014 ***Nucleic Acids Research***

Estimating dispersion parameter

- To obtain stable estimates of dispersion with few samples, we:
  - impose a log-normal prior on ϕ: $\phi_{ij} \sim \log normal\left(m_j, r_j^2\right)$
  - use information from all CpGs in the genome to estimate the parameters m_j and r_j^2
- Choice of log-normal prior was motivated by distribution of dispersion in bisulfite sequencing data
  - RRBS data from mouse embryogenesis study (Smith *et al.* 2012 ***Nature***)
  - Estimation robust to departure from log-normality
  - Prior provides a good "referee"
  - Encourages dispersion estimates to stay within bounds



[1] Feng *et al.* 2014 ***Nucleic Acids Research***

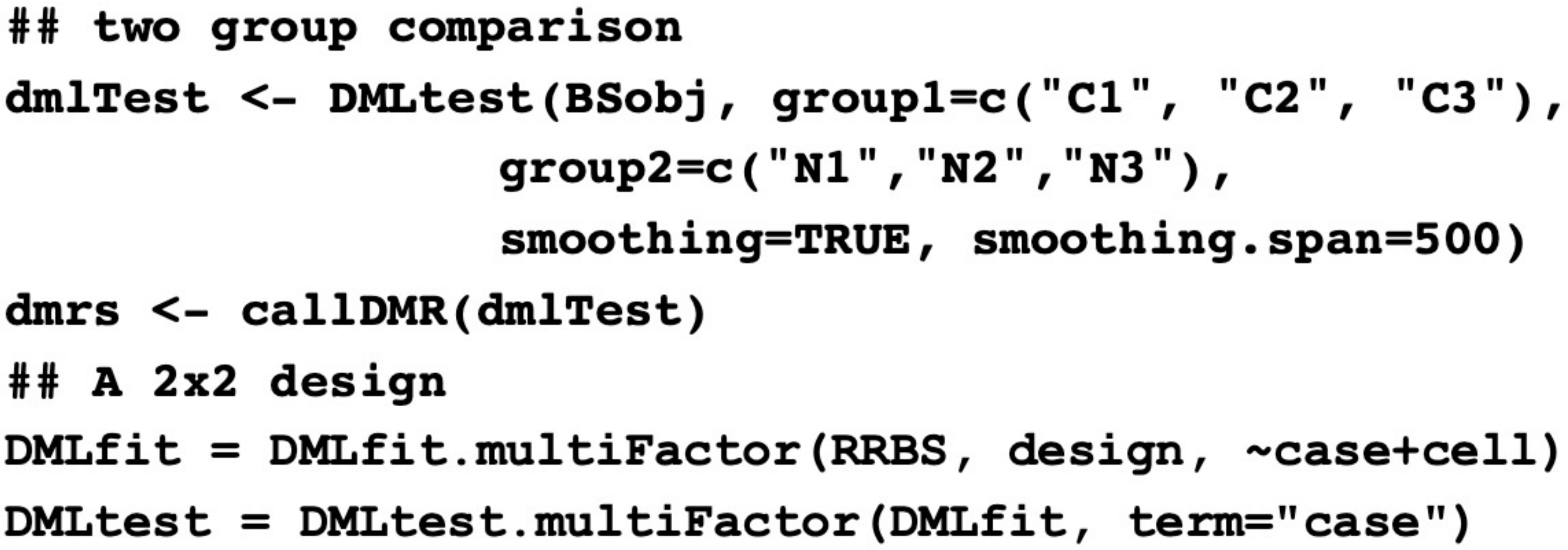
DMR identification

- DML: Differentially Methylated Loci
  - Test for differential methylation at each CpG site
- At site *i*, test: $H_0 : \mu_{i1} = \mu_{i2}$
- Basic algorithm:
  - Use naïve estimates of ϕ across genome to estimate prior
  - For each site *i*, estimate μ_{i1} and μ_{i2} as proportion of methylated reads for each group
  - Bayesian estimation of ϕ_{ij} based on data and prior
  - Plug in estimates of μ_{ij} and ϕ_{ij} to create Wald statistic of form $t_i = \frac{\hat{\mu}_{i1} - \hat{\mu}_{i2}}{\sqrt{Var\left(\hat{\mu}_{i1} - \hat{\mu}_{i2}\right)}}$

[1]Feng *et al.* 2014 ***Nucleic Acids Research***

Bioconductor package: DSS

- Input data object has the same format as `bsseq`.
- `DMLtest` performs Wald test at each CpG.
- `callDML/callDMR` calls DML or DMR.



```
## two group comparison
dmlTest <- DMLtest(BSobj, group1=c("C1", "C2", "C3"),
                 group2=c("N1","N2","N3"),
                 smoothing=TRUE, smoothing.span=500)
dmrs <- callDMR(dmlTest)
## A 2x2 design
DMLfit = DMLfit.multiFactor(RRBS, design, ~case+cell)
DMLtest = DMLtest.multiFactor(DMLfit, term="case")
```

DNA methylation summary

- Methylation plays important roles in many biological processes (stem cell generation, aging, caner, etc.)
- Analysis of BS-seq data presents unique challenges
    - Alignment of sequencing reads
    - Limited sample size + multiple testing
    - Splitting biological variability and technical variability
- Beta-binomial model is widely used

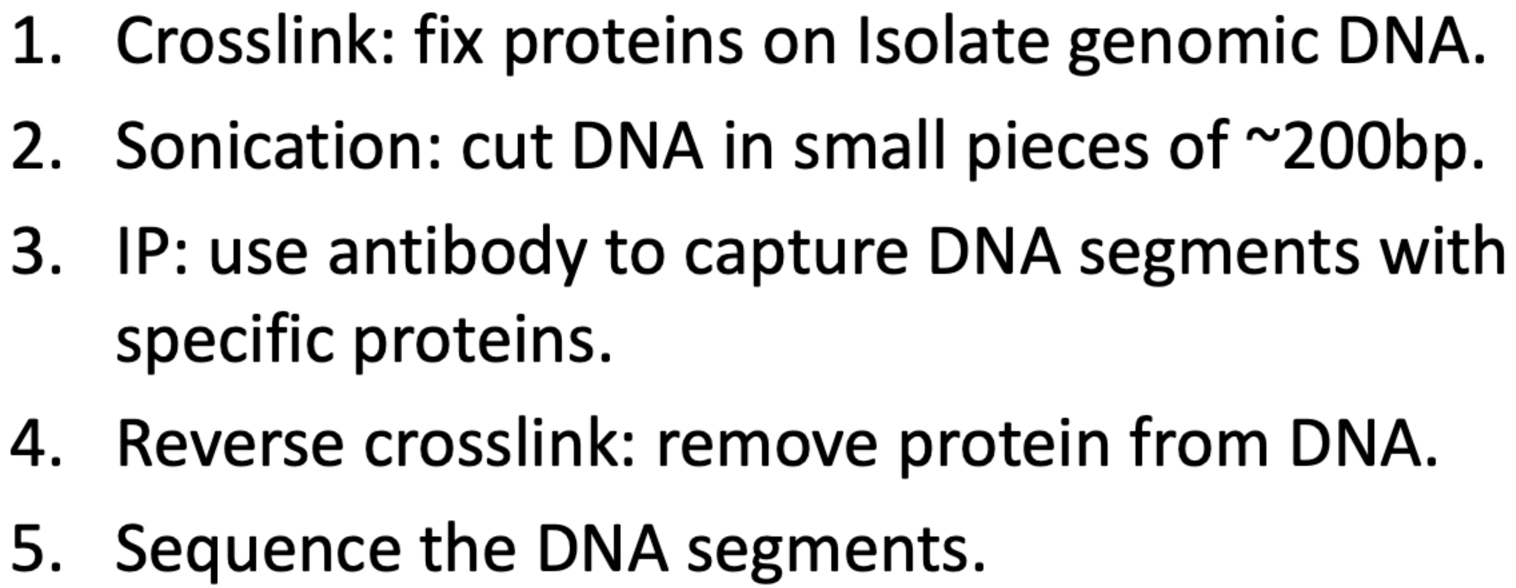
Epigenetics signals (2)

- DNA methylation
- Protein binding on DNA
- Histone modification
- Chromatin accessibility
- Nucleosome occupancy
- …

ChIP-seq: Chromatin ImmunoPrecipitation + sequencing

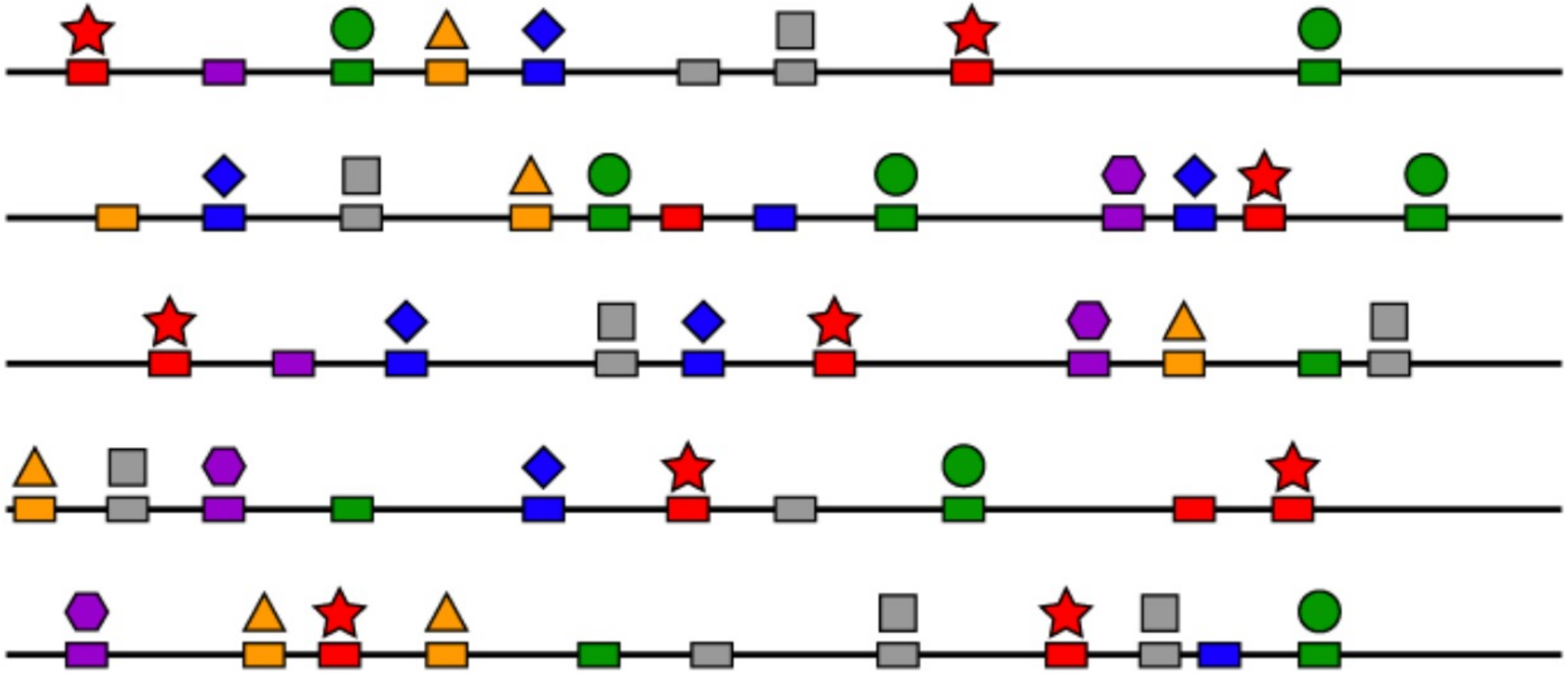
- Scientific motivation: measure specific biological modifications along the genome:
  - Detect binding sites of DNA-binding proteins (transcription factors, pol2, etc.) .
  - quantify strengths of chromatin modifications (e.g., histone modifications).

ChIP-seq experimental procedures



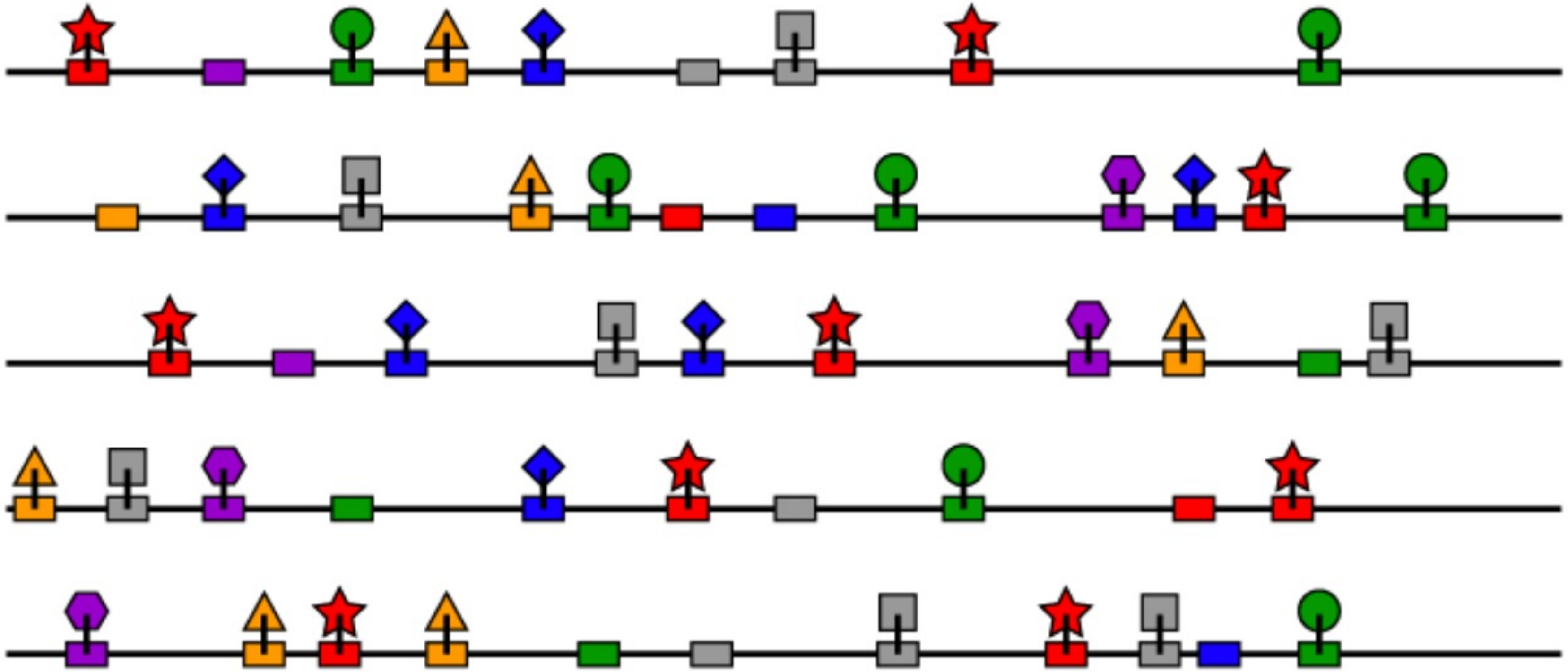
1. Crosslink: fix proteins on Isolate genomic DNA.
2. Sonication: cut DNA in small pieces of ~200bp.
3. IP: use antibody to capture DNA segments with specific proteins.
4. Reverse crosslink: remove protein from DNA.
5. Sequence the DNA segments.

DNA with proteins



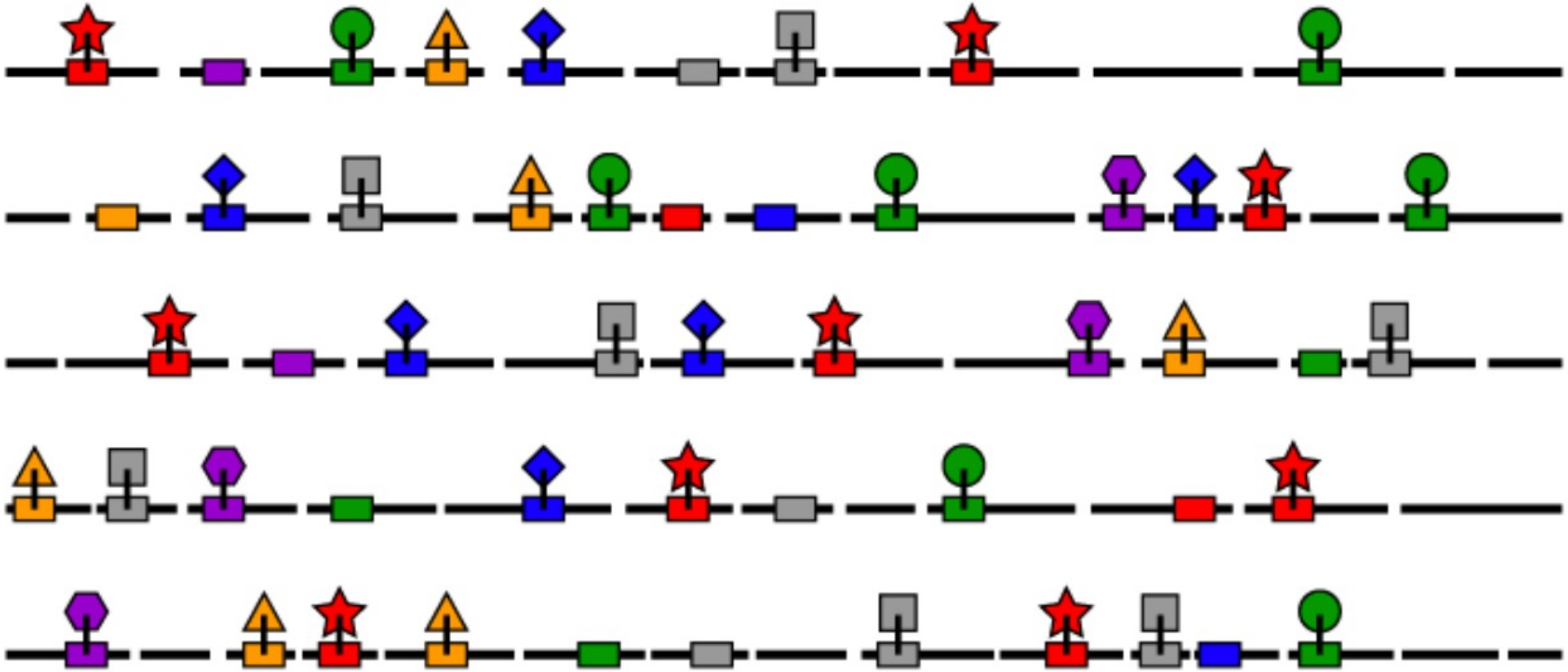
By Richard Bourgon at UC Berkley

Protein/DNA Crosslinking *in vivo*



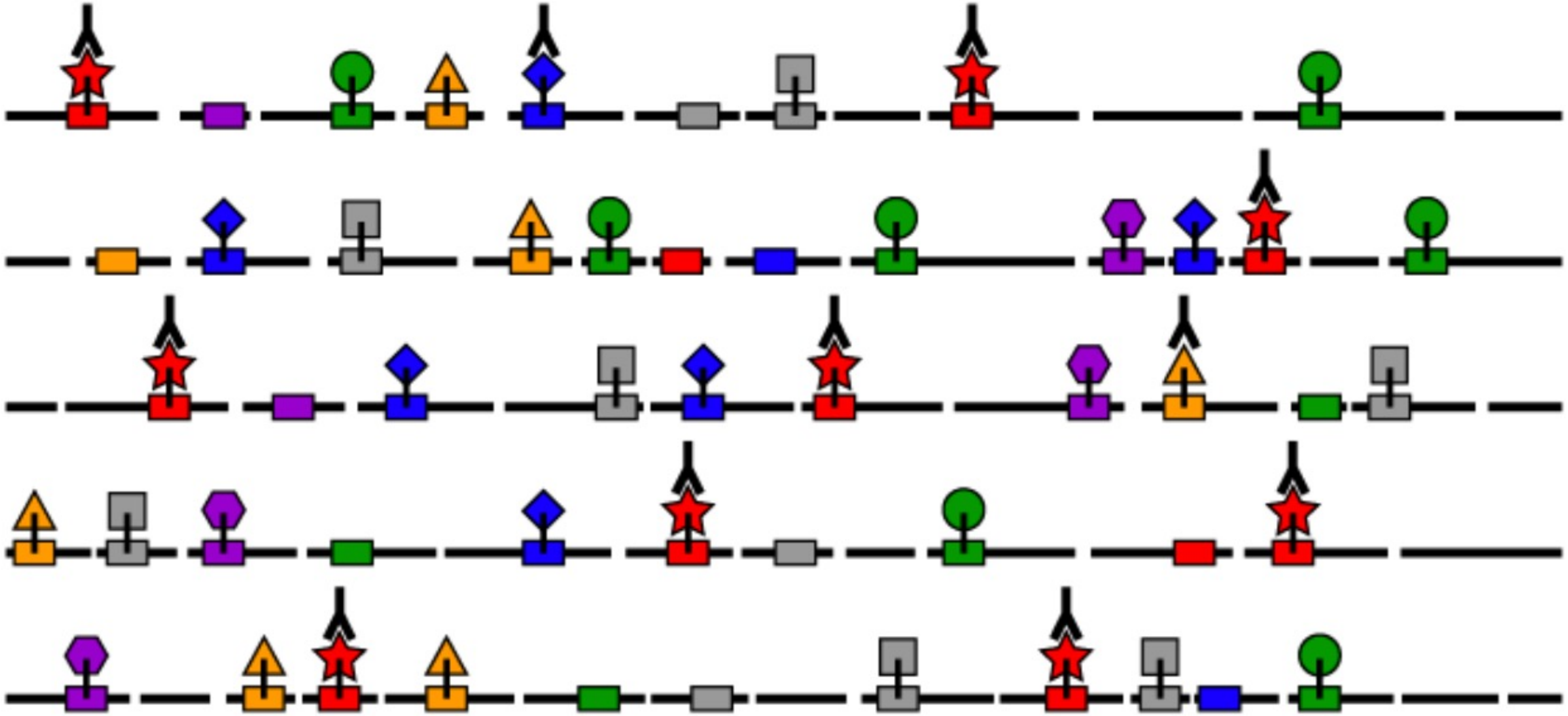
By Richard Bourgon at UC Berkley

Sonication (cut DNA into pieces)



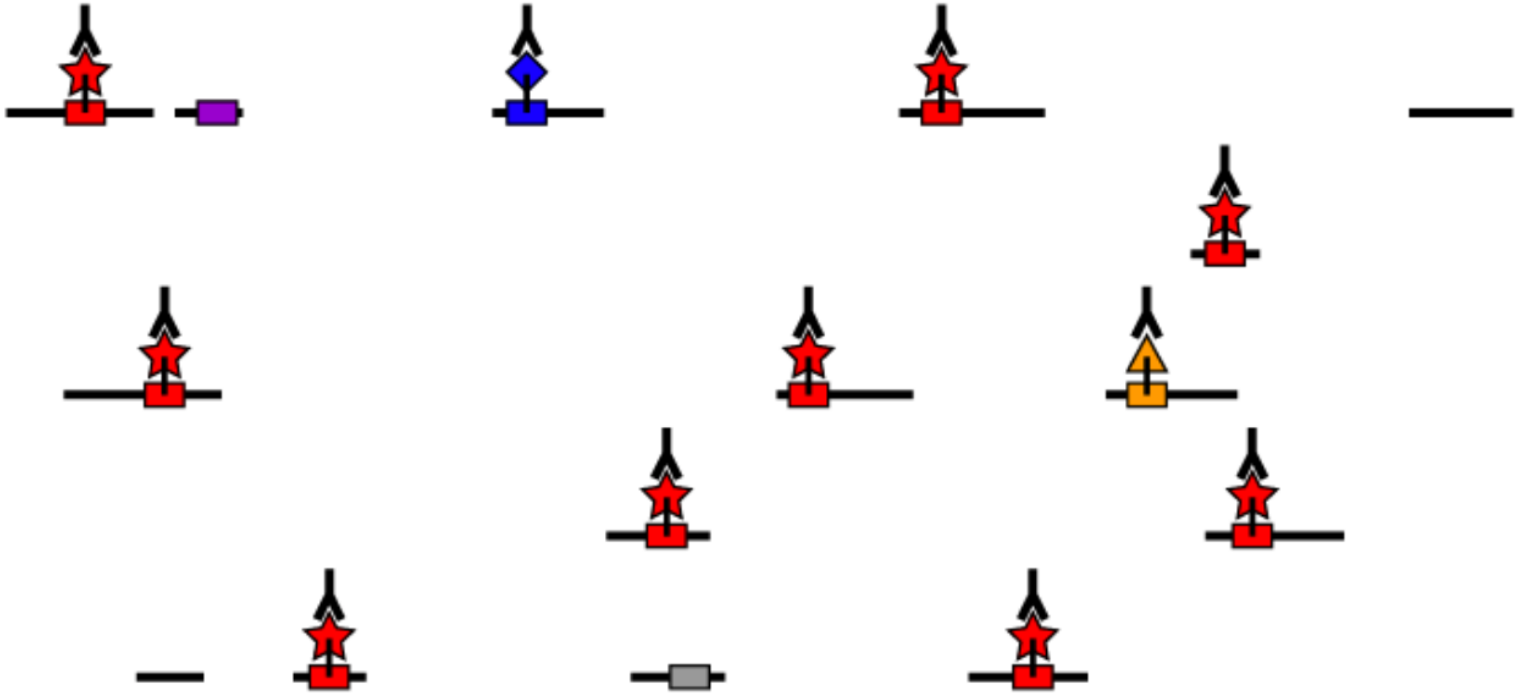
By Richard Bourgon at UC Berkley

Capture using specific antibody



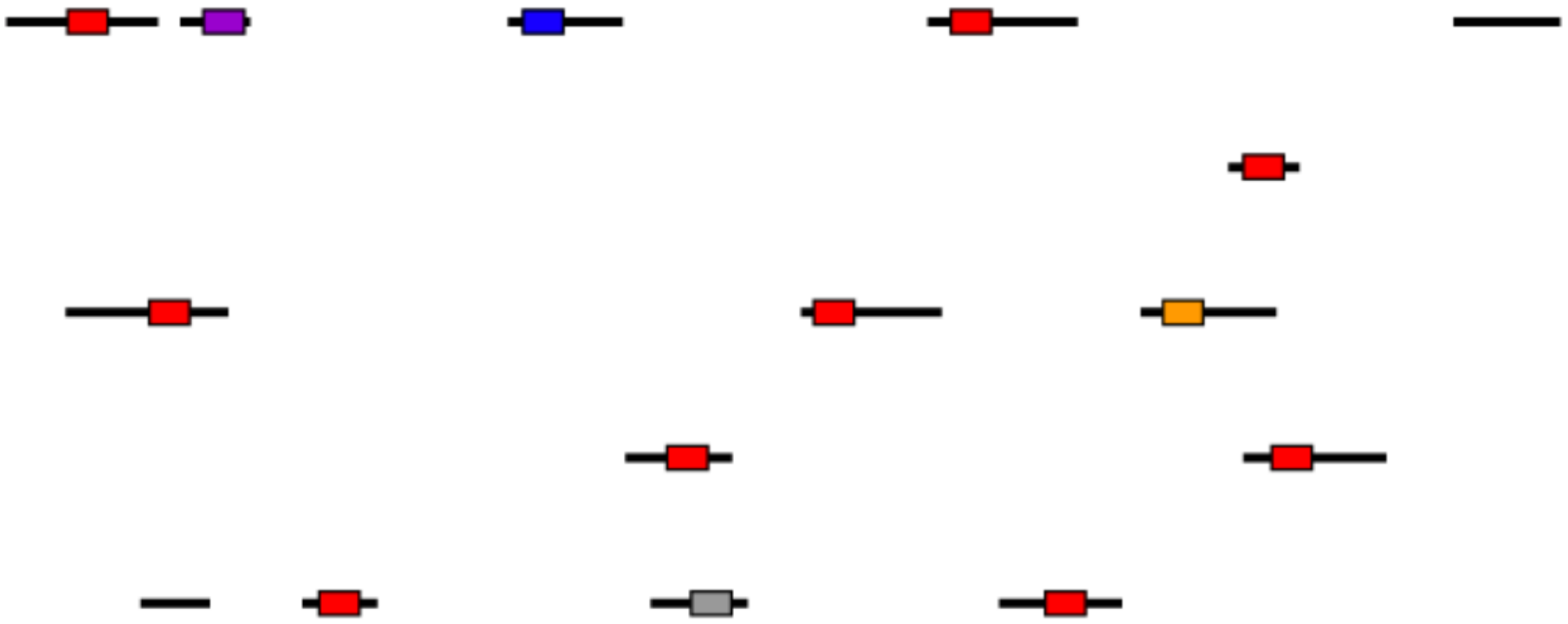
By Richard Bourgon at UC Berkley

Immunoprecipitation (IP)



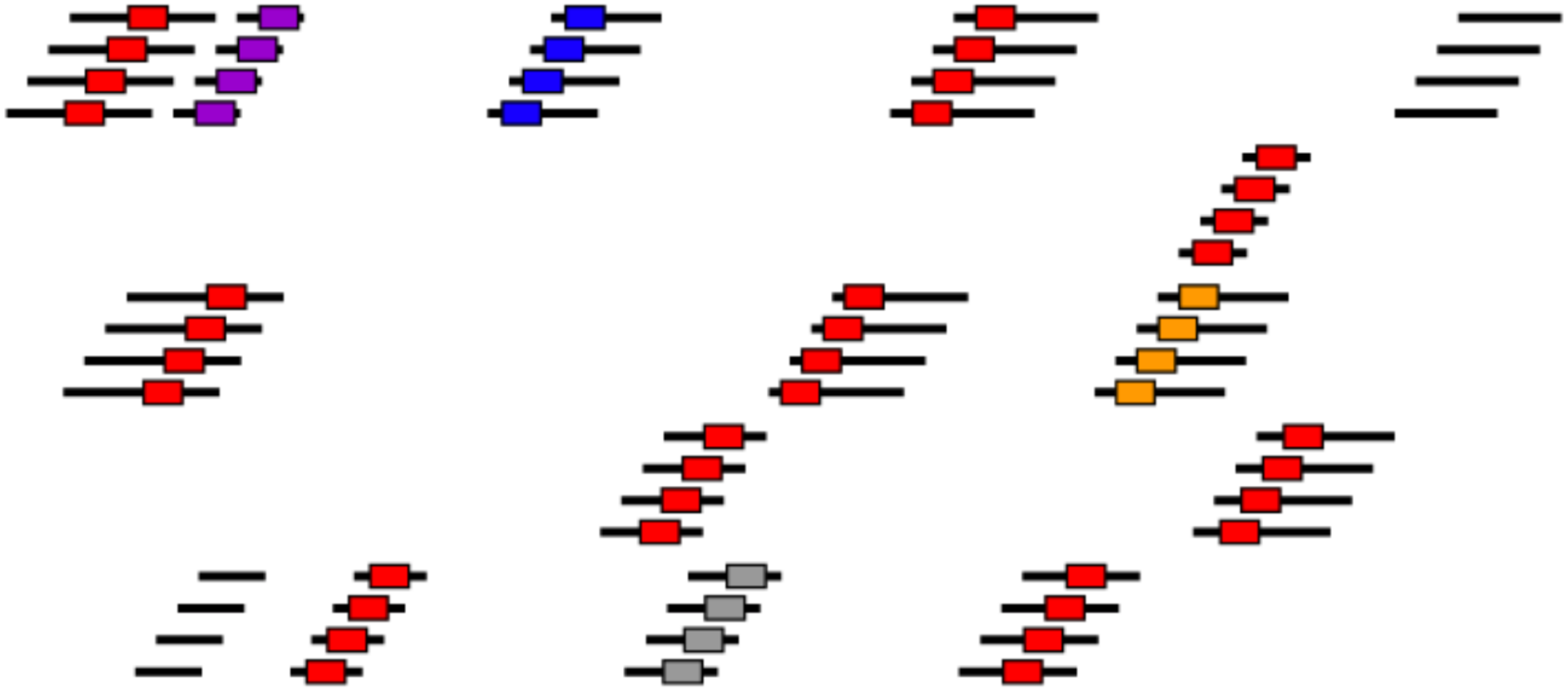
By Richard Bourgon at UC Berkley

Reverse Crosslink and DNA Purification



By Richard Bourgon at UC Berkley

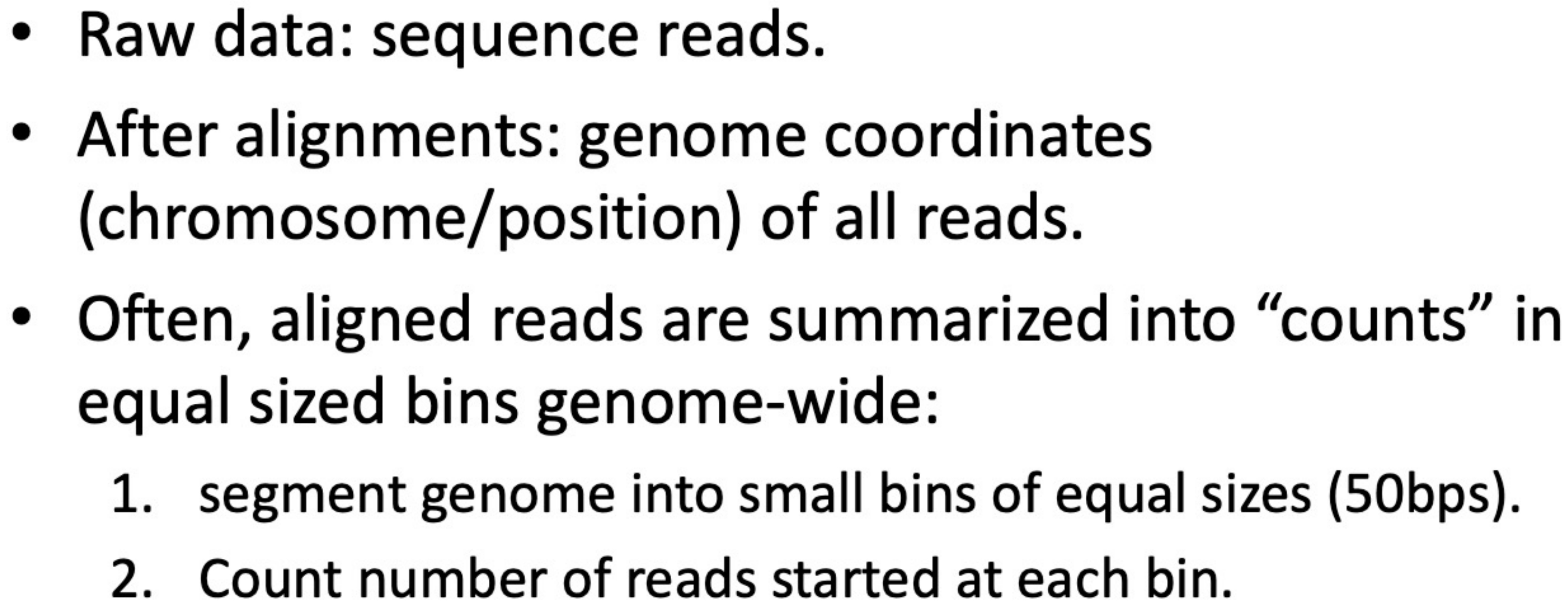
Amplification (PCR)



By Richard Bourgon at UC Berkley

Methods and software for ChIP-seq peak calling

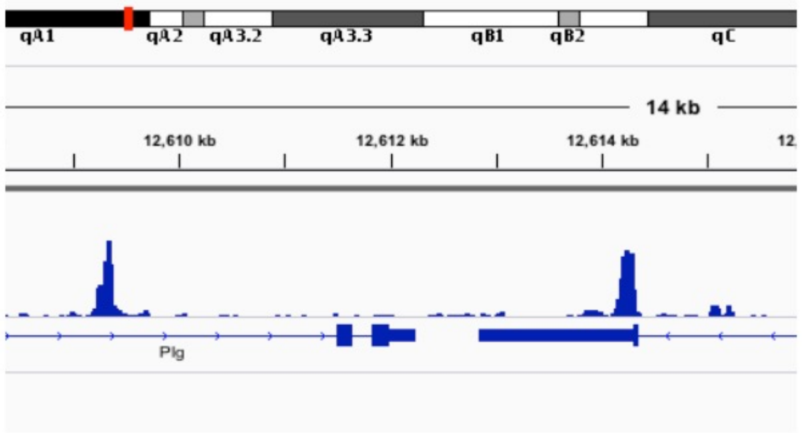
Data from ChIP-seq



- Raw data: sequence reads.
- After alignments: genome coordinates (chromosome/position) of all reads.
- Often, aligned reads are summarized into “counts” in equal sized bins genome-wide:
    1. segment genome into small bins of equal sizes (50bps).
    2. Count number of reads started at each bin.

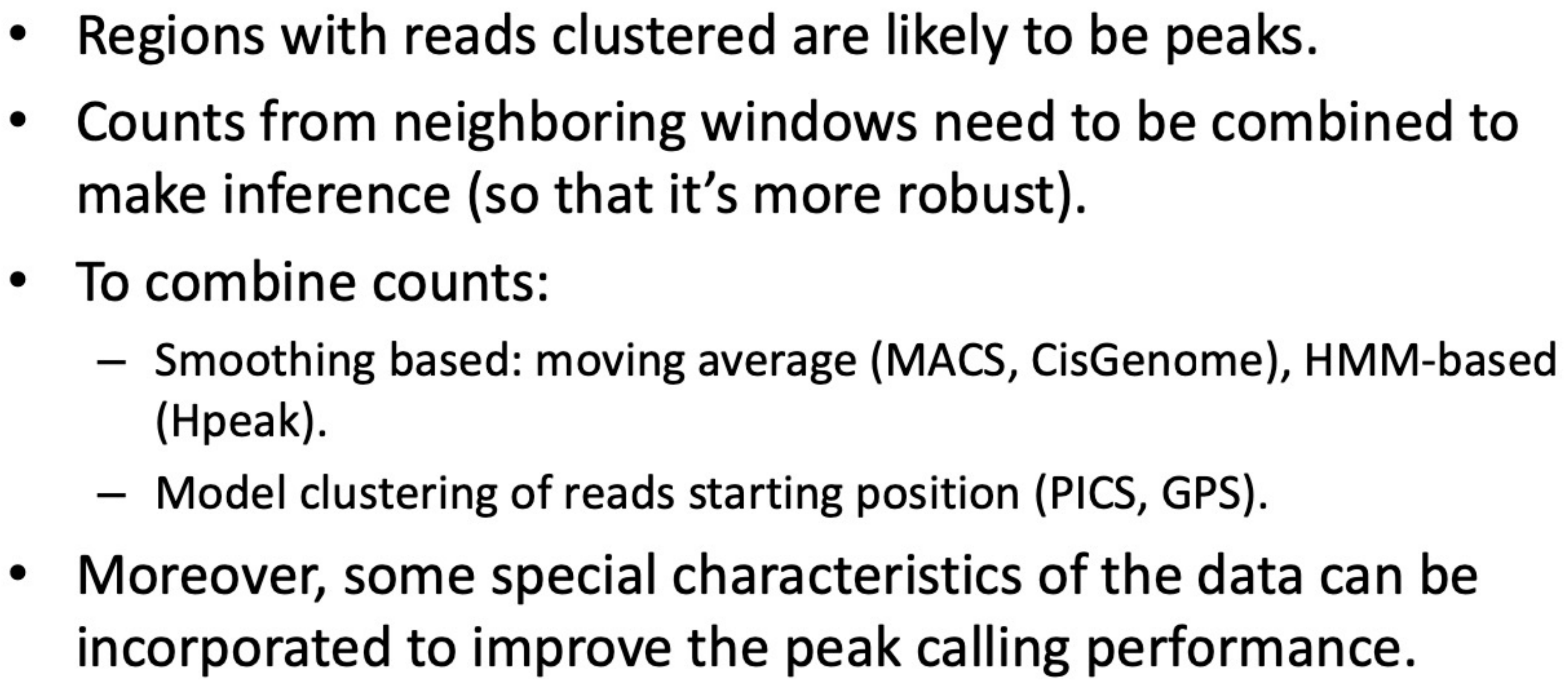
ChIP-seq ‘peak’ detection

- When plot the read counts against genome coordinates, the binding sites show a tall and pointy peak. So “peaks” are used to refer to protein binding or histone modification sites.



- Peak detection is the most fundamental problem in ChIP-seq data analysis.

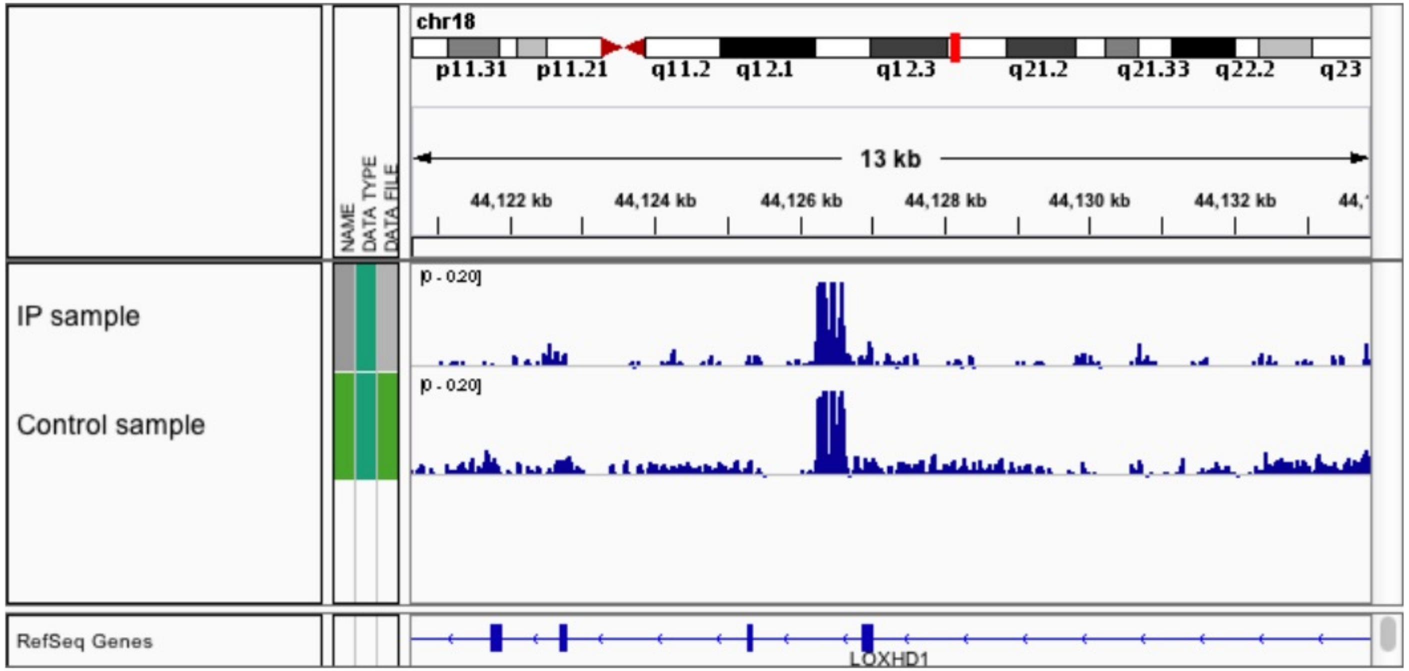
Simple ideas for peak detection



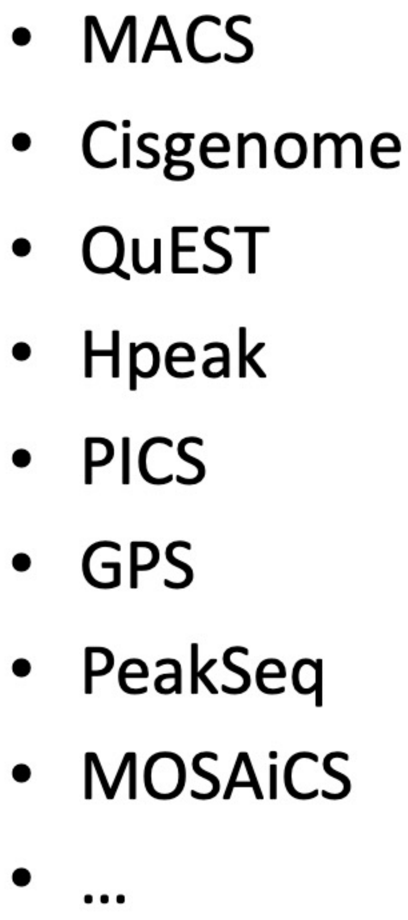
- Regions with reads clustered are likely to be peaks.
- Counts from neighboring windows need to be combined to make inference (so that it’s more robust).
- To combine counts:
  - Smoothing based: moving average (MACS, CisGenome), HMM-based (Hpeak).
  - Model clustering of reads starting position (PICS, GPS).
- Moreover, some special characteristics of the data can be incorporated to improve the peak calling performance.

Control sample is important

- A control sample is necessary for correcting many artifacts: DNA sequence dependent artifacts, chromatin structure, repetitive regions, etc.



Peak detection software



- MACS
- Cisgenome
- QuEST
- Hpeak
- PICS
- GPS
- PeakSeq
- MOSAiCS
- …

MACS (Model-based Analysis of ChIP-Seq) Zhang et al. 2008, *GB*

- Estimate shift size of reads *d* from the distance of two modes from + and – strands.
- Shift all reads toward 3' end by *d/2*.
- Use a dynamic Possion model to scan genome and score peaks. Counts in a window are assumed to following Poisson distribution with rate: $\lambda_{\text{local}} = \max(\lambda_{BG}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$
  - The dynamic rate capture the local fluctuation of counts.
- FDR estimates from sample swapping: flip the IP and control samples and call peaks. Number of peaks detected under each p-value cutoff will be used as null and used to compute FDR.

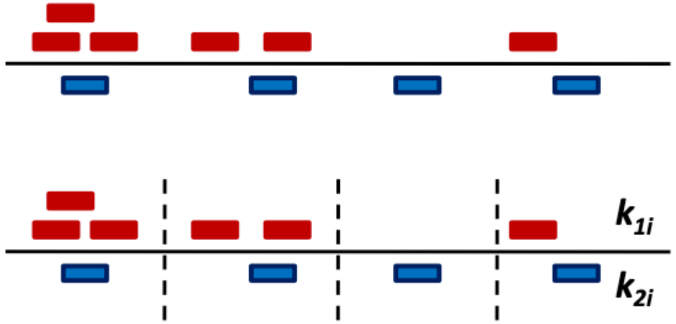
Using MACS

- http://liulab.dfci.harvard.edu/MACS/index.html
- Written in Python, runs in command line.
- Command:

```
macs14 -t sample.bed -c control.bed -n result
```

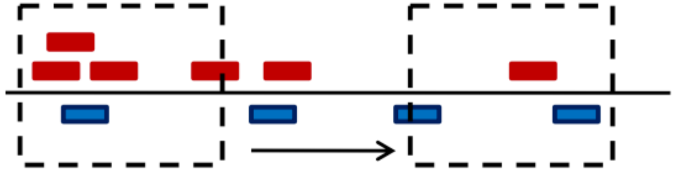
Cisgenome (Ji et al. 2008, *NBT*)

- Implemented with Windows GUI.
- Use a Binomial model to score peaks.



$n_i = k_{1i} + k_{2i}$

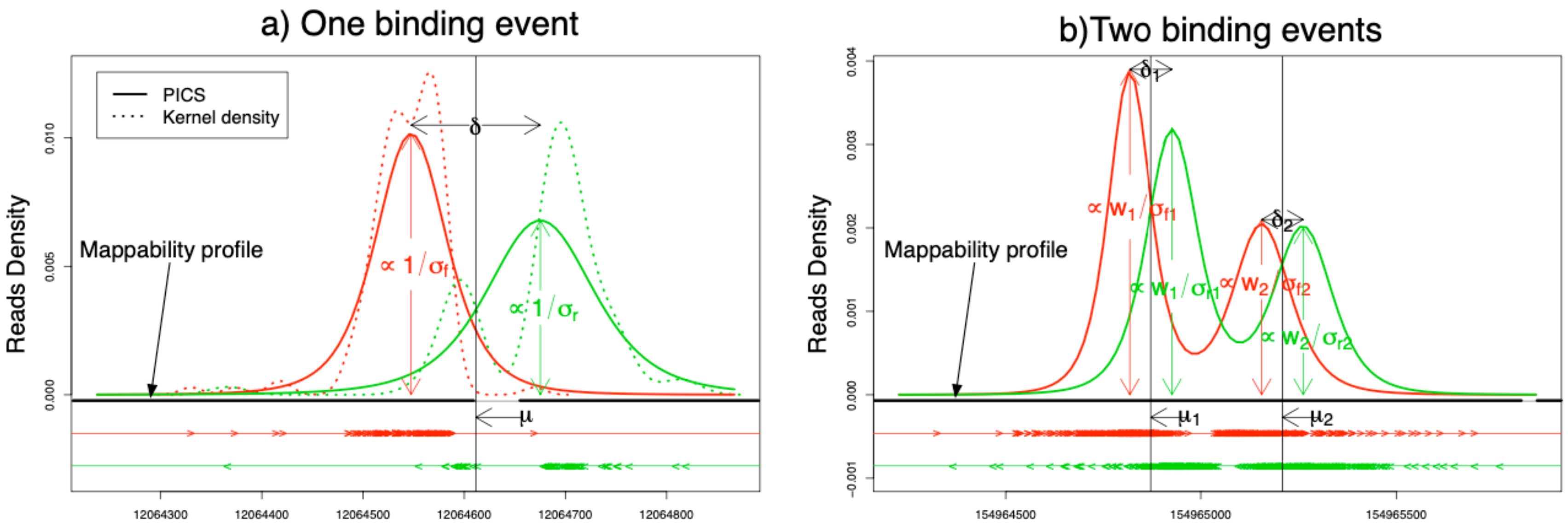
$k_{1i} \mid n_i \sim \text{Binom}(n_i, p_0)$



PICS: Probabilistic Inference for ChIP-seq (Zhang *et al.* 2010 *Biometrics*)

- Use shifted t-distributions to model peak shape.
- Can deal with the clustering of multiple peaks in a small region.
- A two step approach:
  - Roughly locate the candidate regions.
  - Fit the model at each candidate region and assign a score.
- EM algorithm for estimating parameters.
- Computationally very intensive.

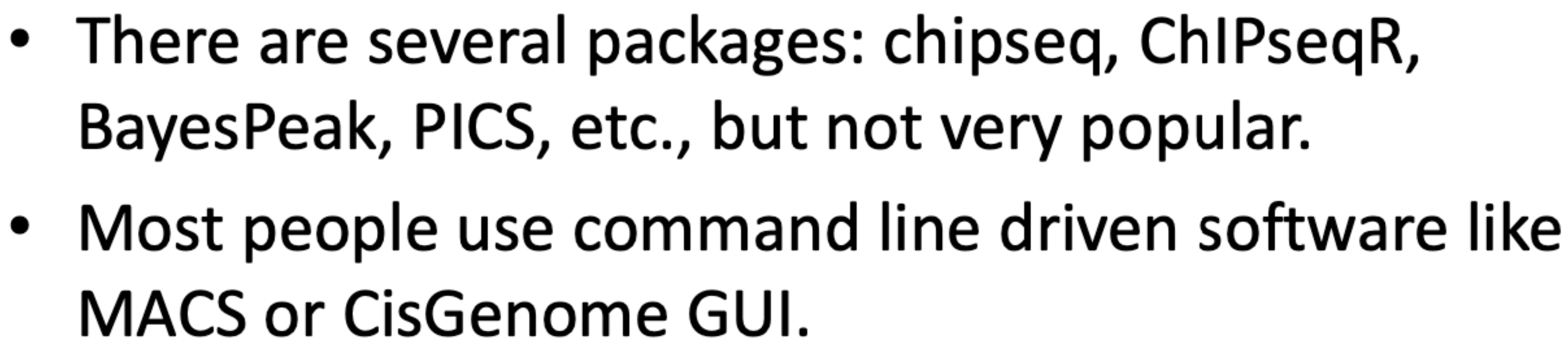
PICS



$$f_i \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{fk}, \sigma_{fk}^2\right) \overset{d}{=} g_f(f_i|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\sigma}_f)$$

$$r_j \sim \sum_{k=1}^{K} w_k t_4\left(\mu_{rk}, \sigma_{rk}^2\right) \overset{d}{=} g_r(r_j|\boldsymbol{w}, \boldsymbol{\mu}, \boldsymbol{\delta}, \boldsymbol{\sigma}_r)$$

Bioconductor packages for ChIP-seq

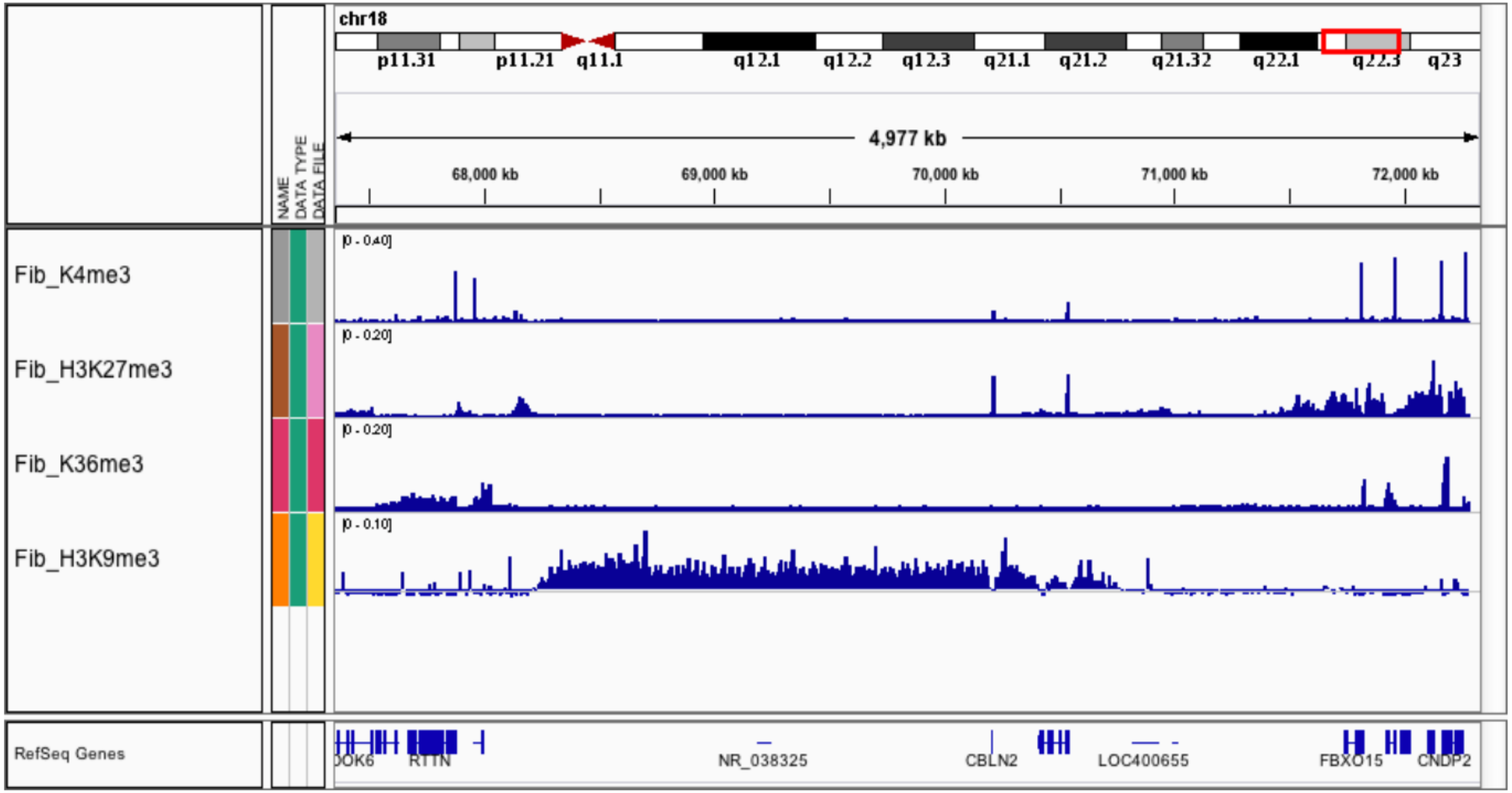


- There are several packages: chipseq, ChIPseqR, BayesPeak, PICS, etc., but not very popular.
- Most people use command line driven software like MACS or CisGenome GUI.

ChIP-seq for histone modification

- Histone modifications have various patterns.
  - Some are similar to protein binding data, e.g., with tall, sharp peaks: H3K4.
  - Some have wide (mega-bp) “blocks”: H3k9.
  - Some are variable, with both peaks and blocks: H3k27me3, H3k36me3.

Histone modification ChIP-seq data



Complications in histone peak/block calling

- Smoothing-based method:
  - Long block requires bigger smoothing span, which hurts boundary detection.
  - Data with mixed peak/block (K27me3, K36me3) requires varied span: adaptive fitting is computationally infeasible.
- HMM based method:
  - Tend to over fit. Sometimes need to manually specify transition matrix.

MACS2

- An updated version of MACS: https://github.com/taoliu/MACS/blob/master/README.rst.
- Has an option for broad peak calling, which uses post hoc approach to combine nearby peaks.
- Syntax:

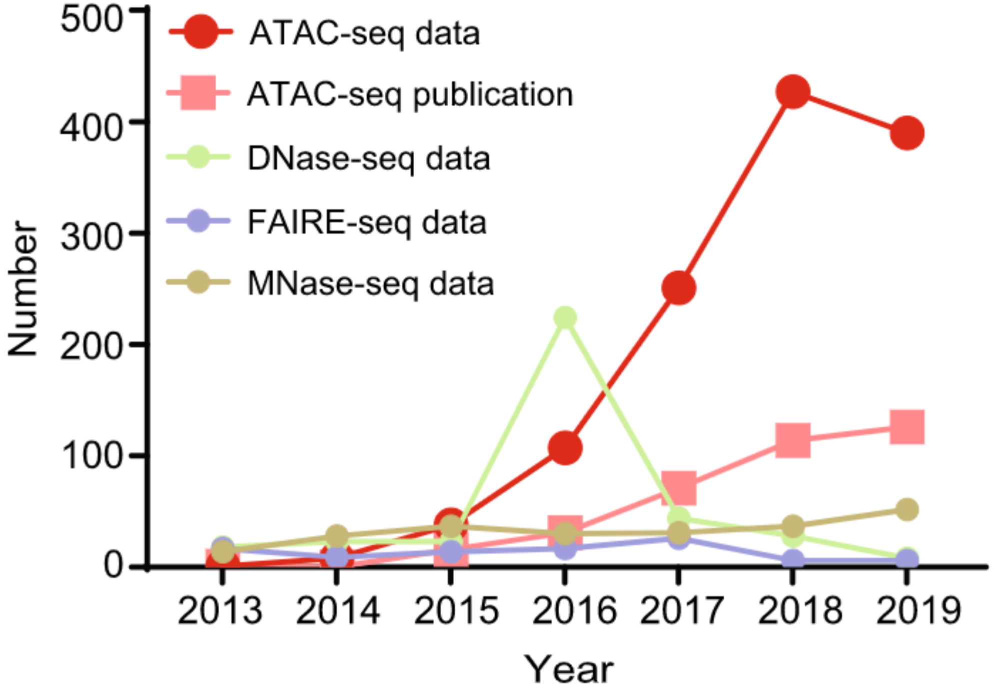
```
macs2 callpeak -t ChIP.bam -c Control.bam
--broad -g hs --broad-cutoff 0.1
```

Summary for ChIP-seq peak calling

- ChIP-seq detects protein binding and histone modification along the genome
- Detect regions with enriched reads
- Control sample is important
- Need to incorporate some special characteristics of the data to improve peak detection
- Calling long peaks is challenging
- Various software available

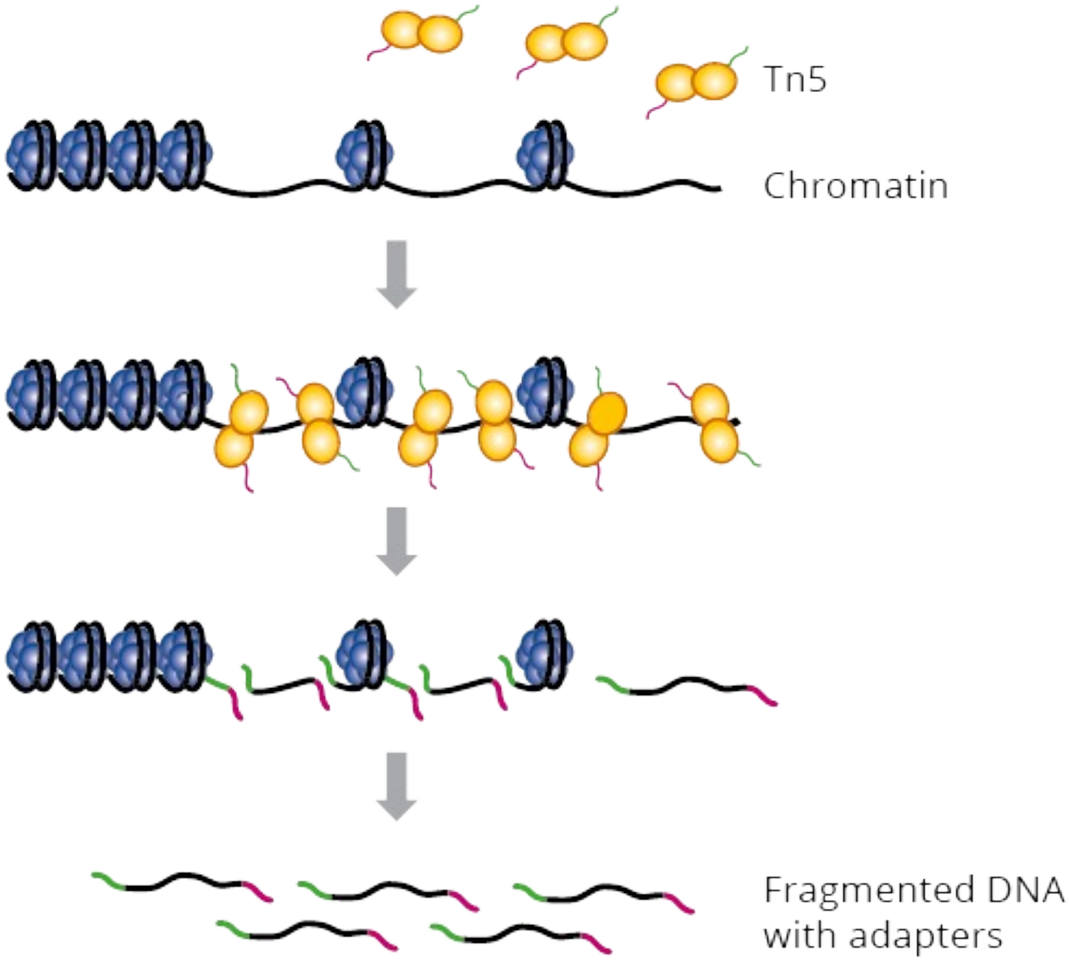
ATAC-seq

- ATAC-seq: **A**ssay for **T**ransposase-**A**ccessible **C**hromatin + sequencing
- Assess genome-wide chromatin accessibility
- Faster and more sensitive than old approach (DNase-seq, MNase-seq)



Yan et al. GB (2020) 21:22

ATAC-seq workflow

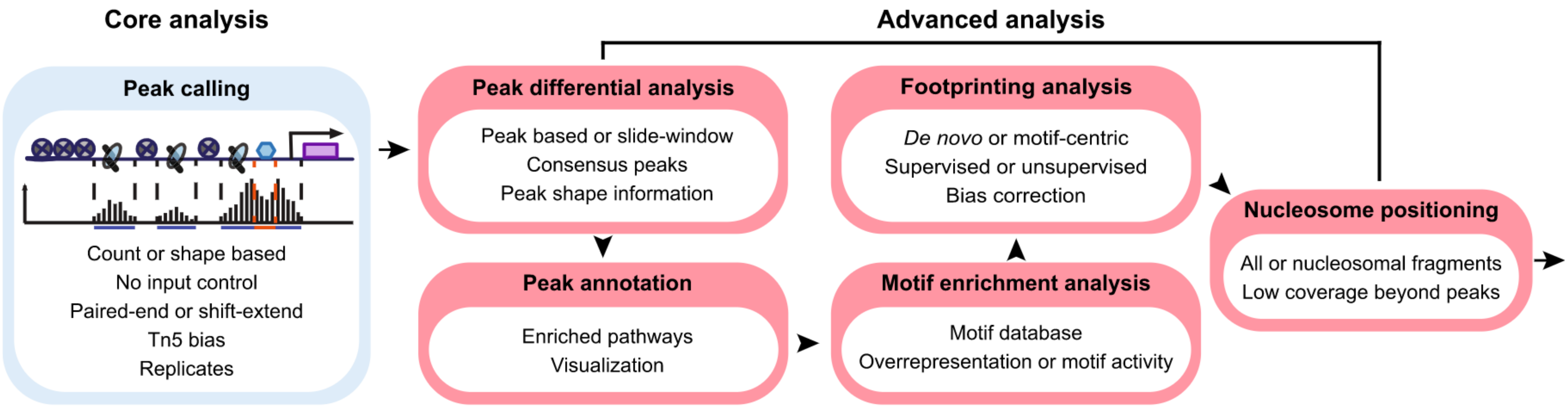


https://www.biostars.org/p/393678/

ATAC-seq data analysis: peak calling

- Can be adopted from ChIP-seq with the assumption that ATAC-seq peak patterns share the same properties
- Default software: MACS2
- A review is provided by Yan *et al.* on Genome Biology (2020)

ATAC-seq data analysis



Core analysis

Peak calling

Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates

Advanced analysis

Peak differential analysis

Peak based or slide-window
Consensus peaks
Peak shape information

Peak annotation

Enriched pathways
Visualization

Motif enrichment analysis

Motif database
Overrepresentation or motif activity

Footprinting analysis

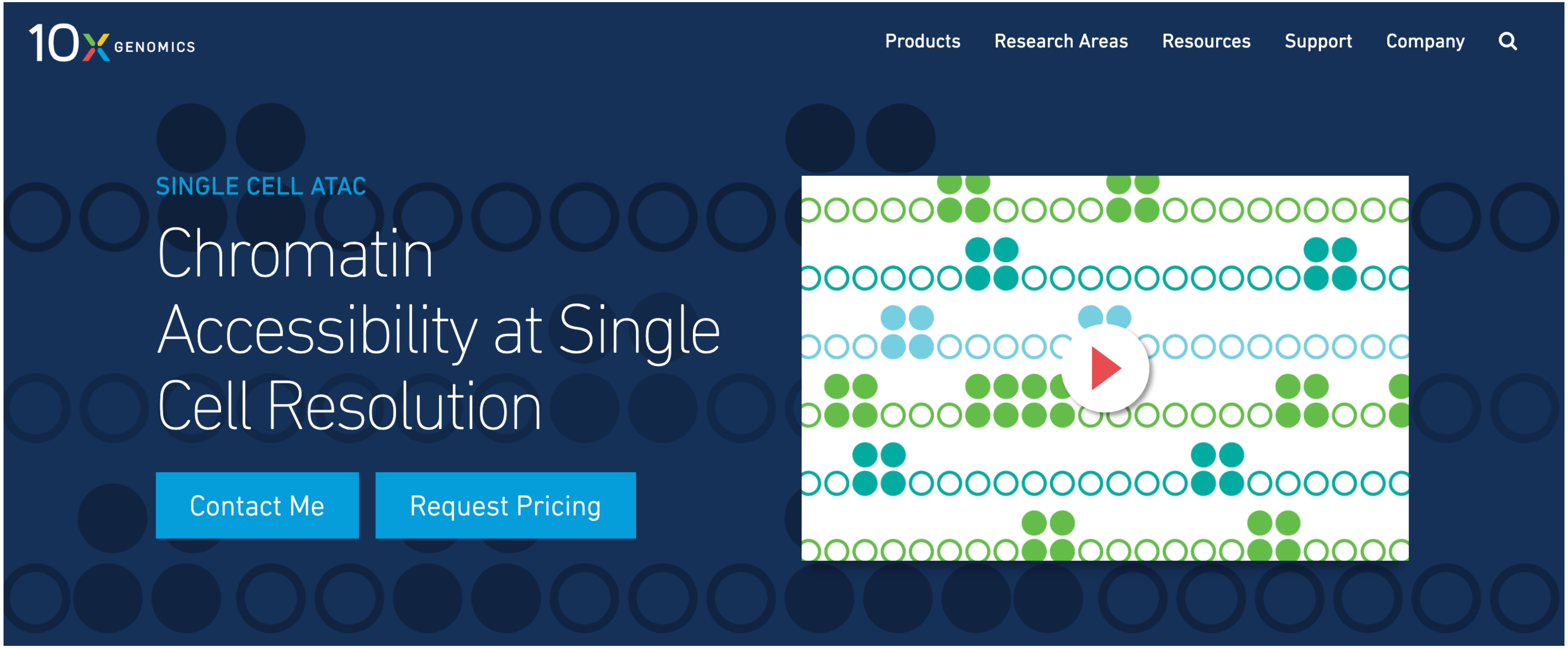
De novo or motif-centric
Supervised or unsupervised
Bias correction

Nucleosome positioning

All or nucleosomal fragments
Low coverage beyond peaks

Yan et al. GB (2020) 21:22

Single-cell ATAC-seq (scATAC-seq)



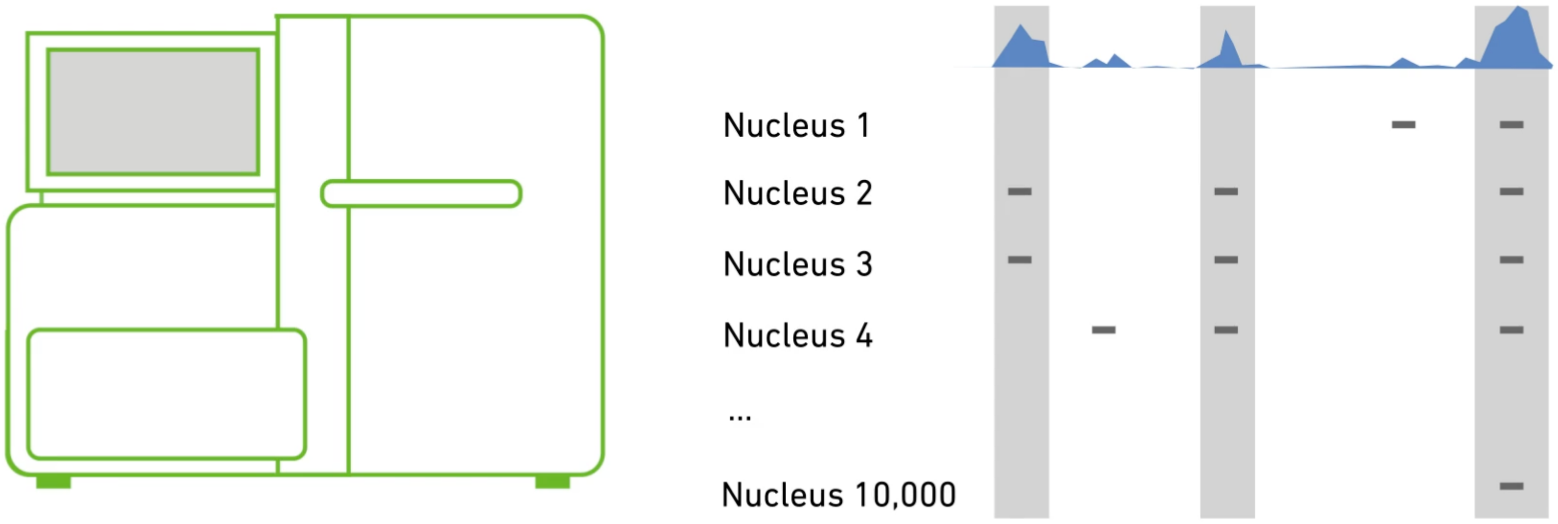
SINGLE CELL ATAC

Chromatin Accessibility at Single Cell Resolution

Source: 10X Genomics

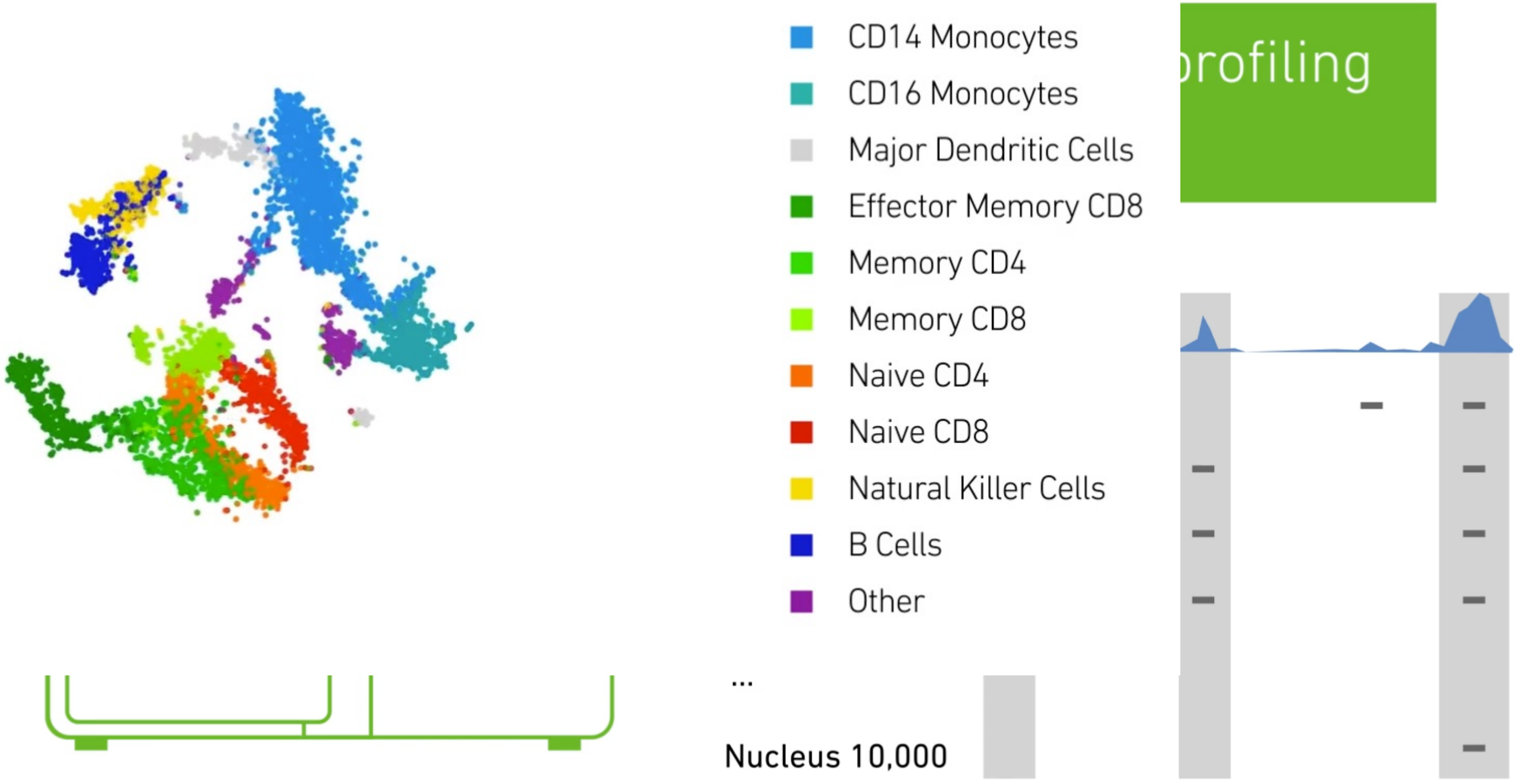
Single-cell ATAC-seq (scATAC-seq)

... enables open chromatin profiling of thousands of nuclei



Source: 10X Genomics

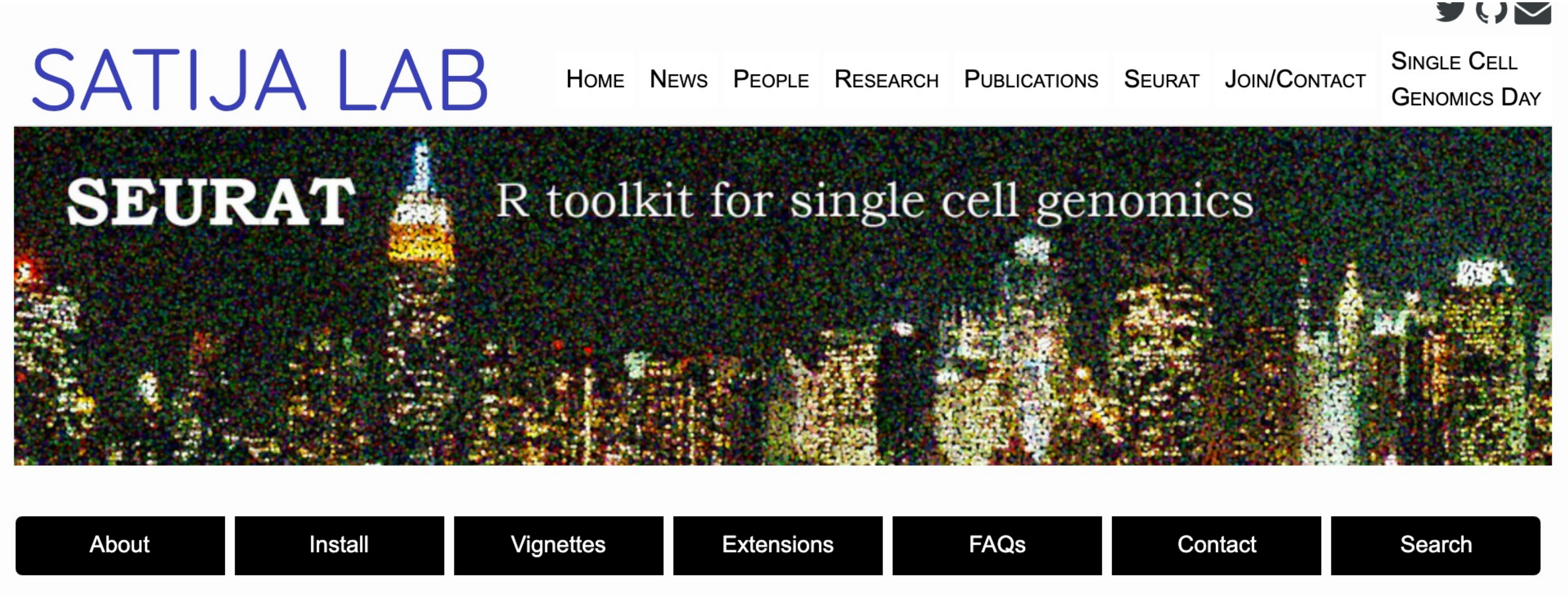
Single-cell ATAC-seq (scATAC-seq)



Source: 10X Genomics

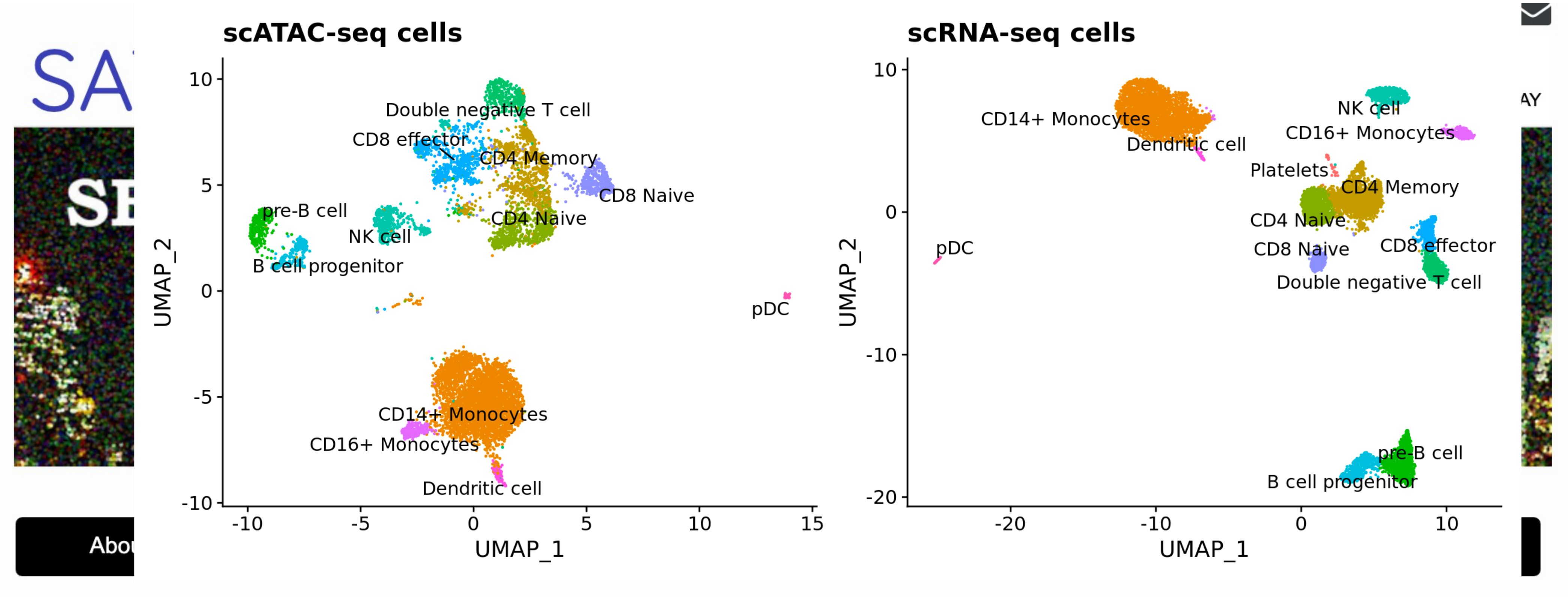
scATAC-seq data analysis

- Seurat (R, Bioconductor)



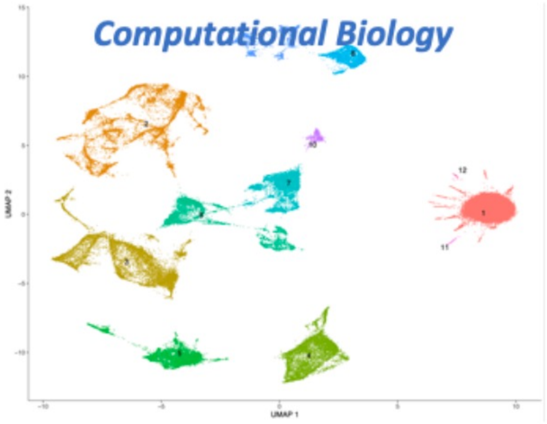
scATAC-seq data analysis

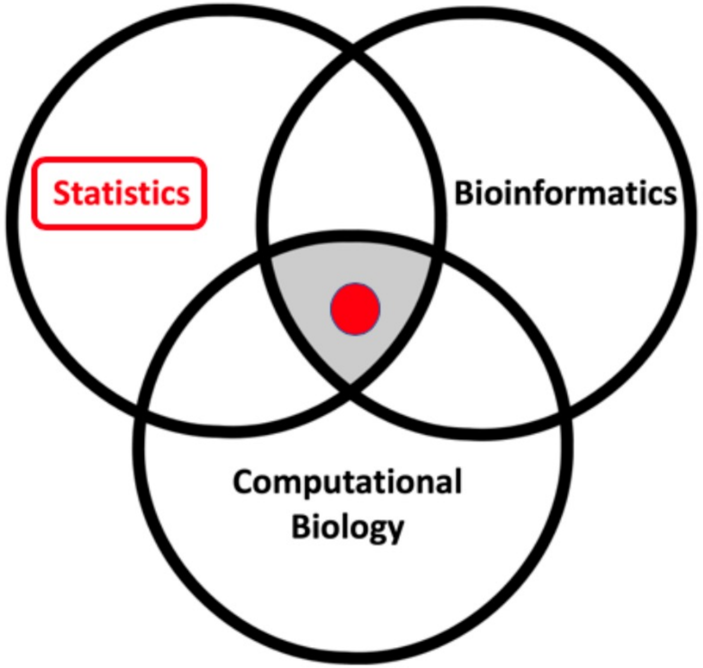
- Seurat (R, Bioconductor)



Other emerging methods

- scBS-seq: single-cell bisulfite sequencing
- NOME-seq: <u>**N**</u>ucleosome <u>**O**</u>ccupancy + <u>**ME**</u>thylation
- scNMT-seq: single-cell <u>**N**</u>ucleosome, <u>**M**</u>ethylation and <u>**T**</u>ranscription sequencing
- MeRIP-seq: mRNA epigenetics modifications (m6A)





✉ hfeng@case.edu

@HHarryFeng

https://hfenglab.org/

Internship positions in statistical bioinformatics are available!