

Statistical Methods for Epigenetic

Yongjin Park (slide credit: Keegan Korthauer, Manolis Kellis)

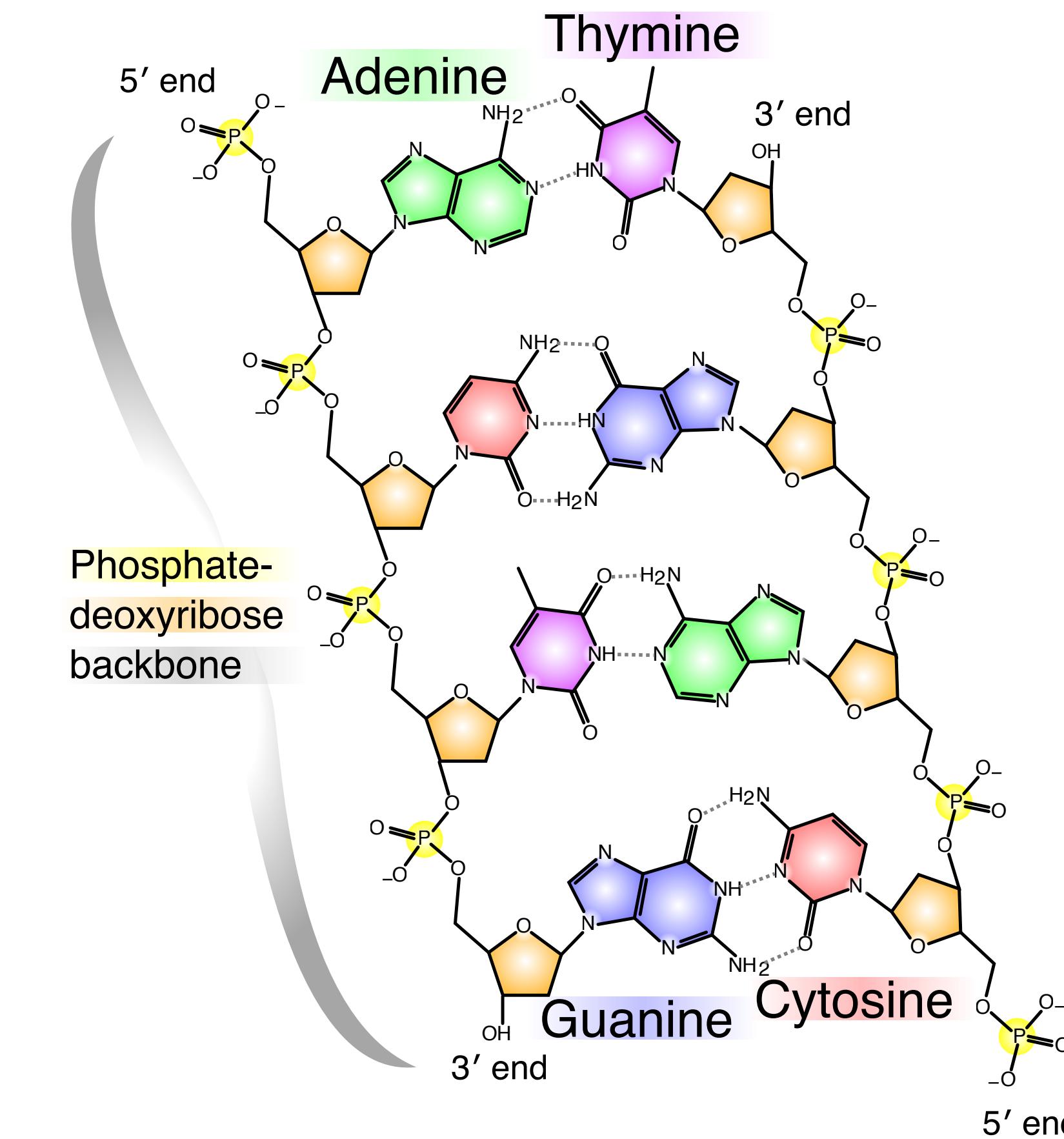
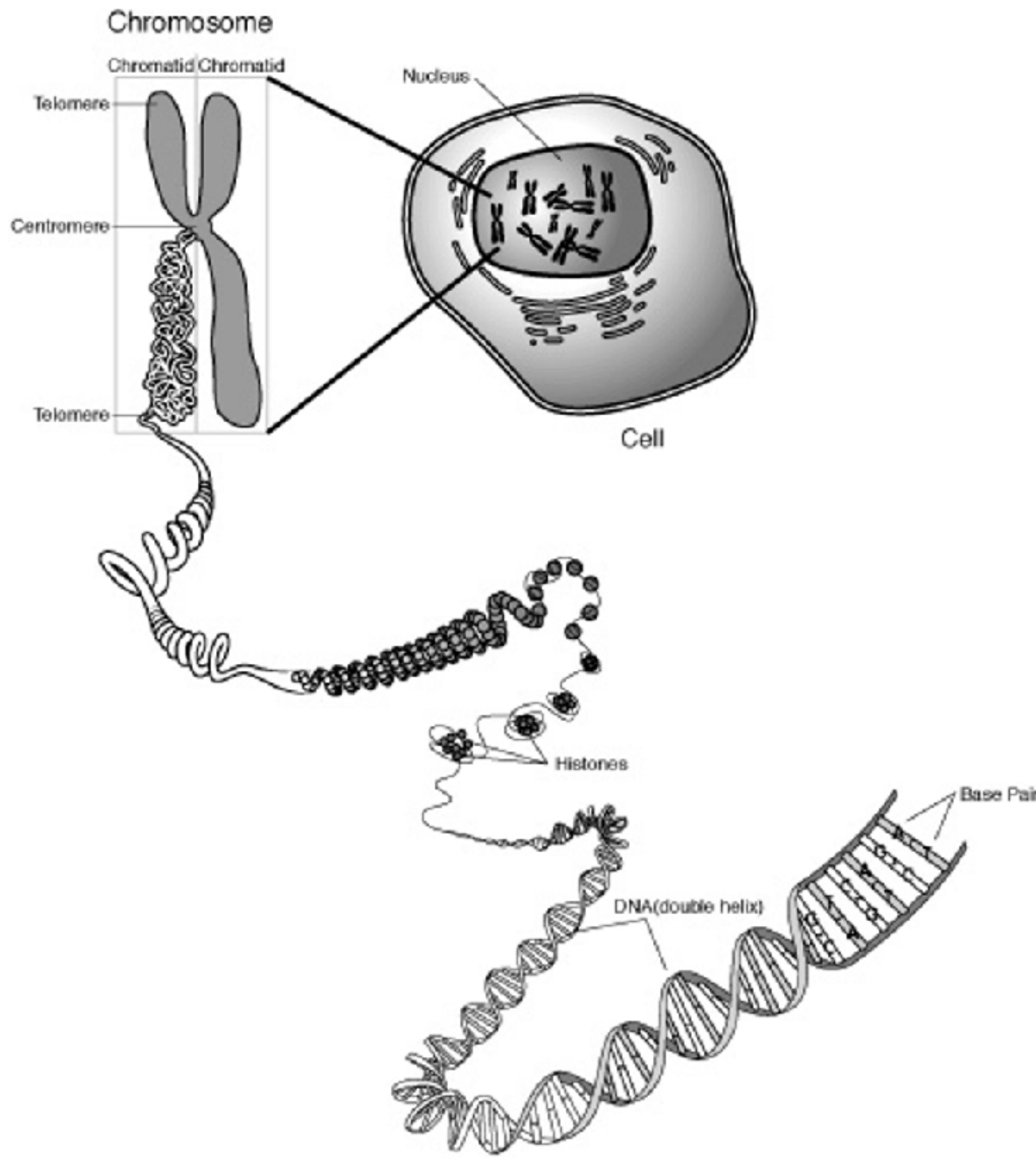
Announcement

- **If you want full marks on the analysis assignment, please write down your interpretations.**
- **(We want you to be a scientist)**

Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

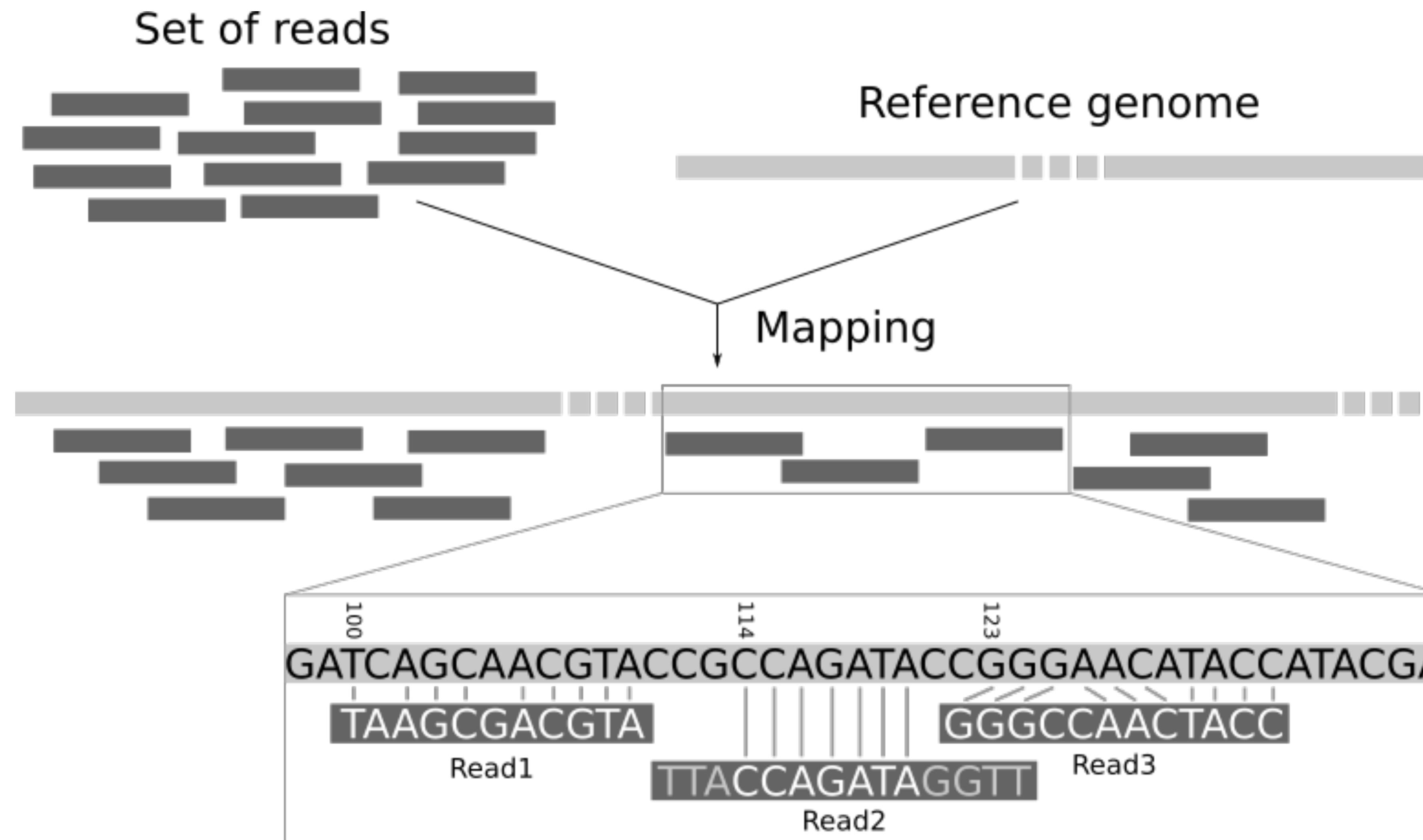
A reference genome can serve as an efficient computing tool (a template for matching)



Given reference, we can look up locations

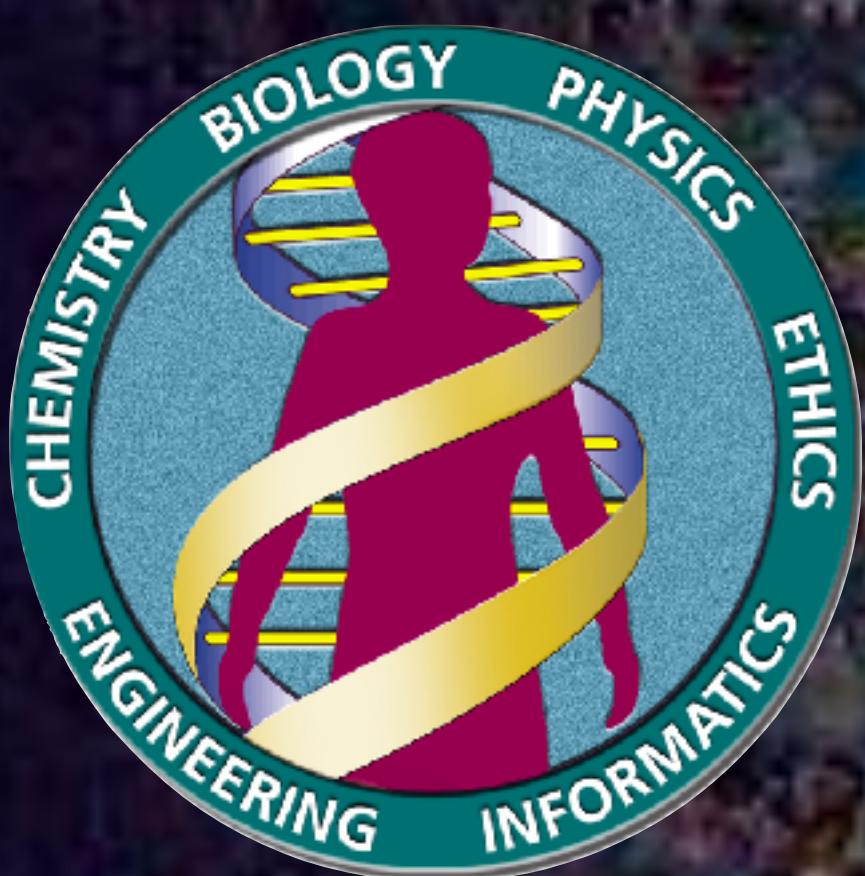
Where is
CCTTCTGTG
TCGGA

Short fragments/reads to reference genome



Human Genome Project

A reference DNA sequence for
3.2B basepairs x diploid for each individual

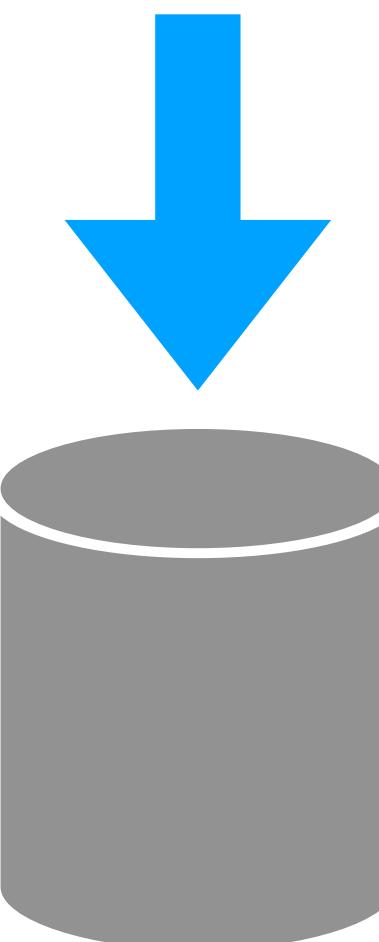


<https://www.genome.gov/human-genome-project>

How do we align/map/match a short read to known genomic DNAs?

Computational Method

```
TTATATTGAATTTCAAAAAATTCTTACTTTTT  
TTTGATGGACGCAAAGAAGTTAATAAT  
CATATTACATGGCATTACCACCATACATA  
TCCATATCTAACCTTACTTATATGTTGTGGA  
AATGTAAAGAGGCCCCATTATCTTAGCCTAA  
AAAAACCTTCTTTGGAACCTTCAGTAAT  
ACGCTTAAC TGCTCATTGCTATATTGAAGT
```



Index reference
genome for a quick lookup
(e.g., BWT, Bloom filter)

GCCCCATTATCTT?



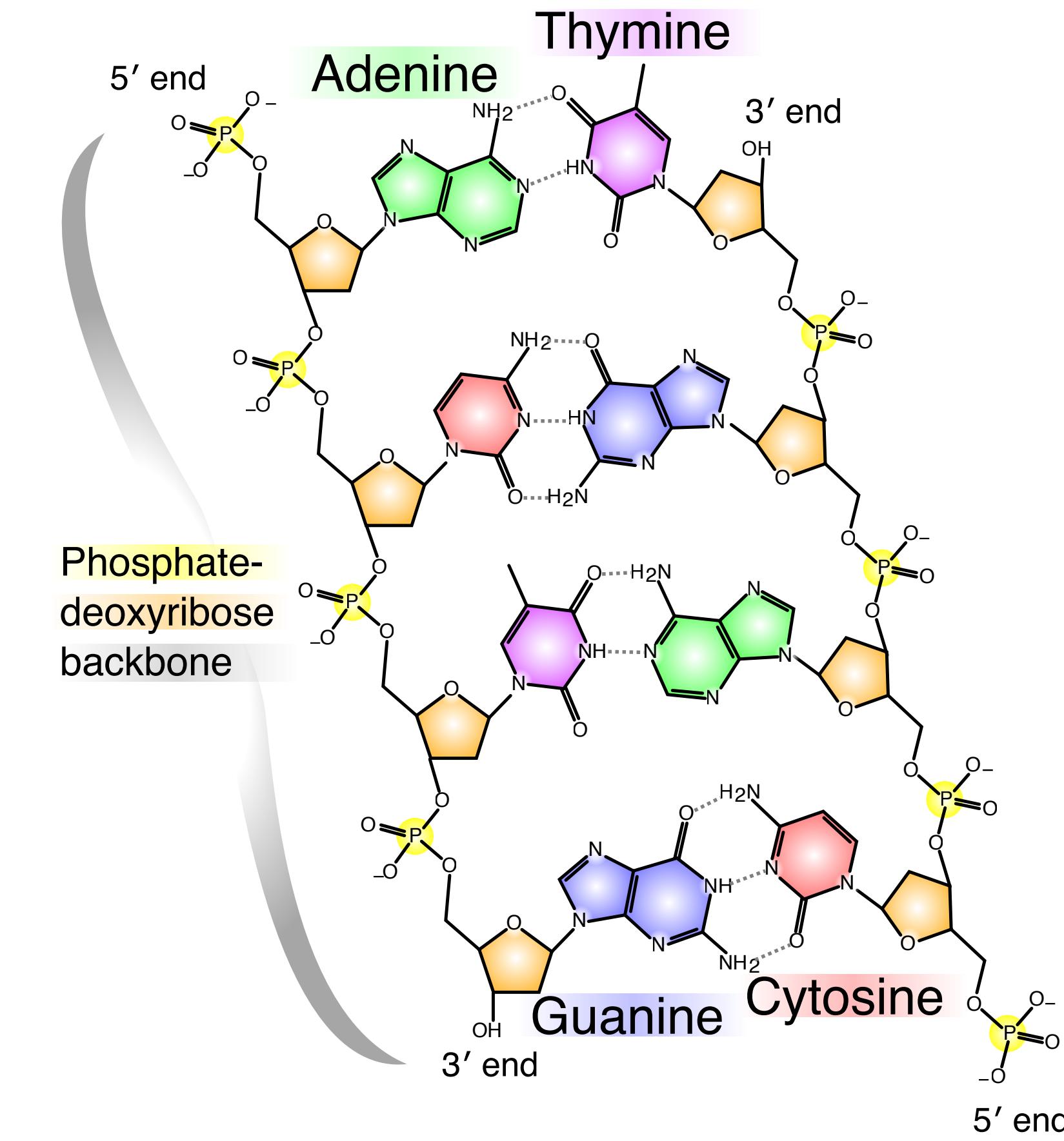
vs.

GCCTTTTATCTT?



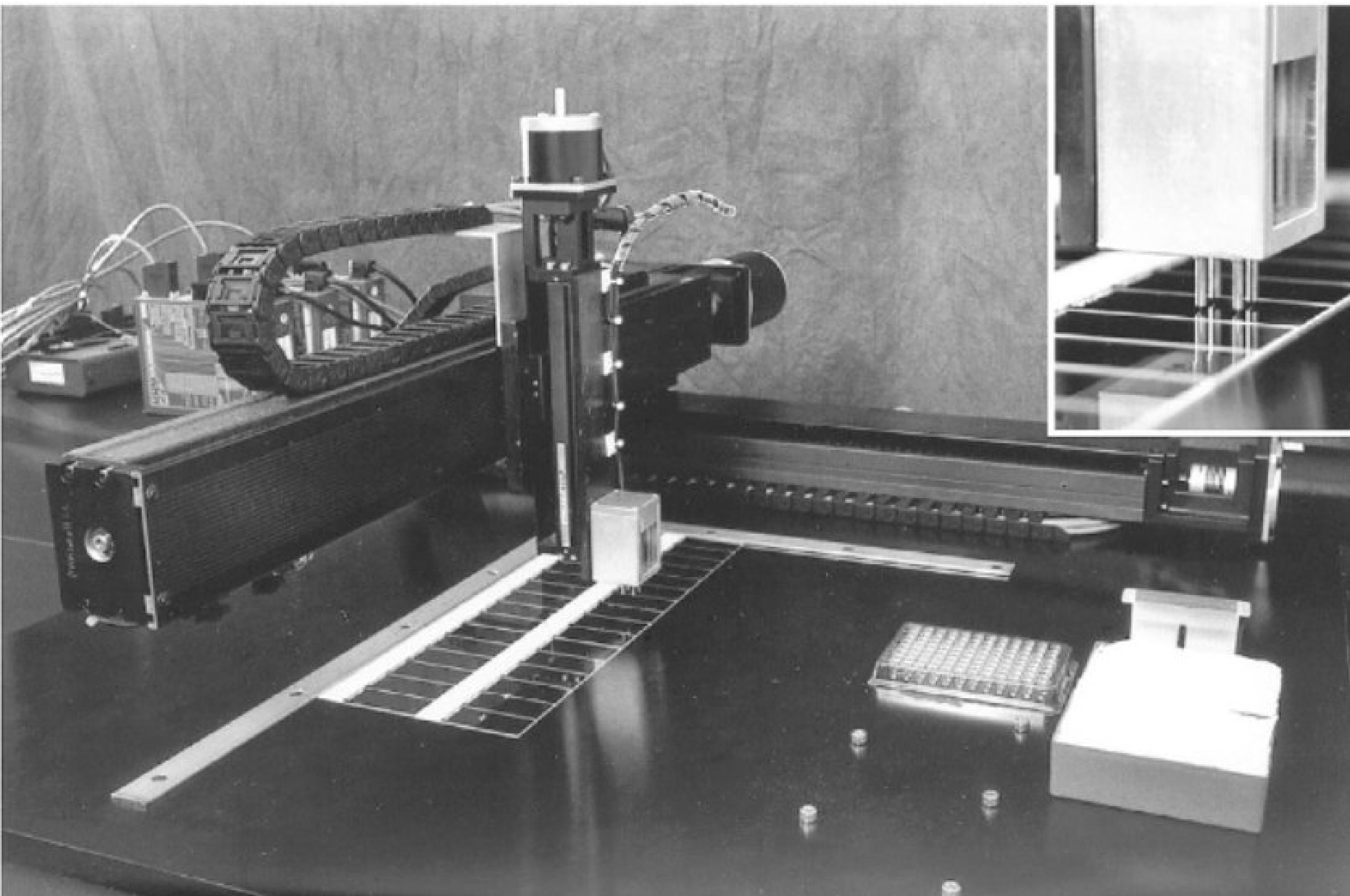
Can we find this short read?
If so, where?

Biochemical Method



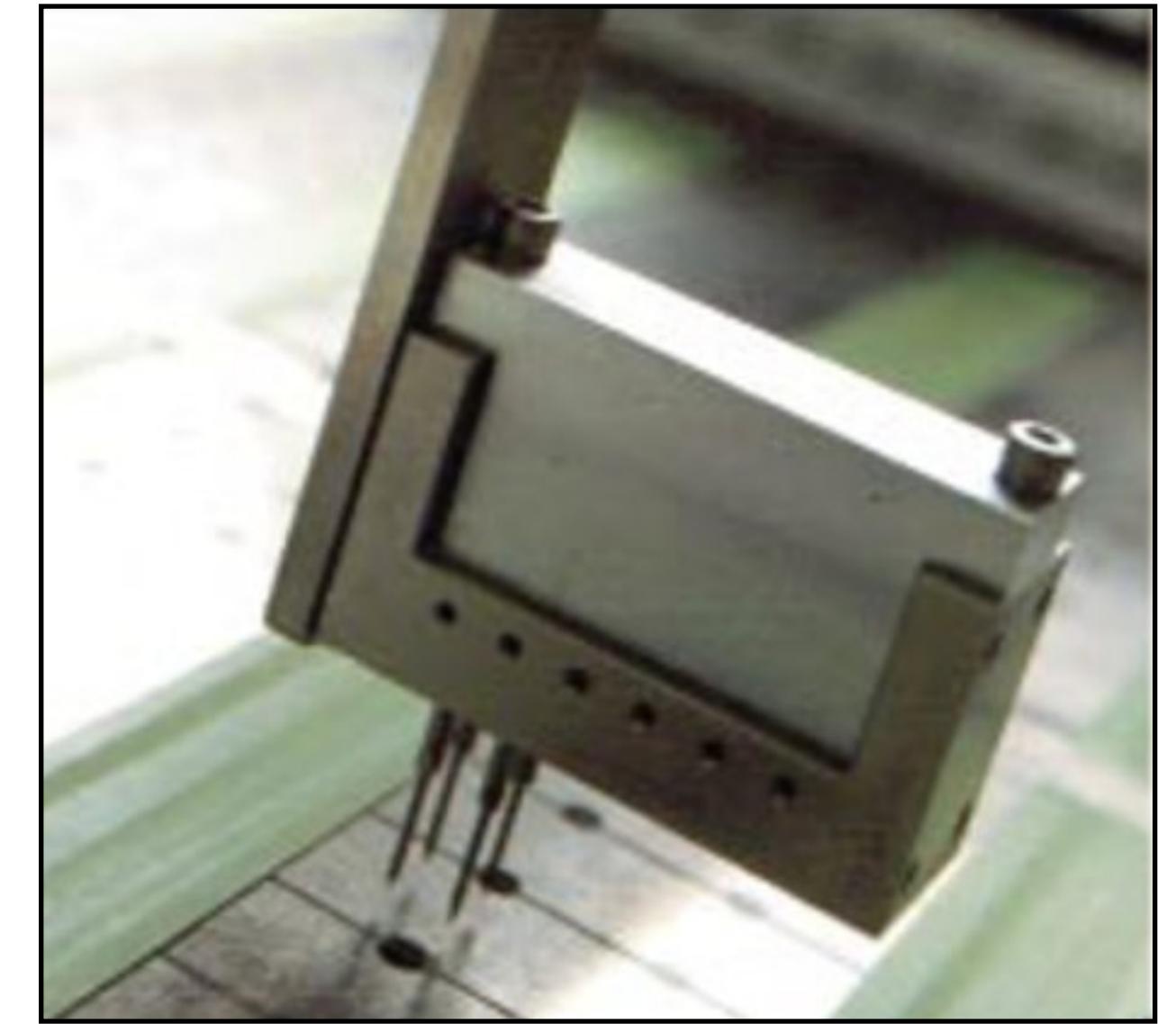
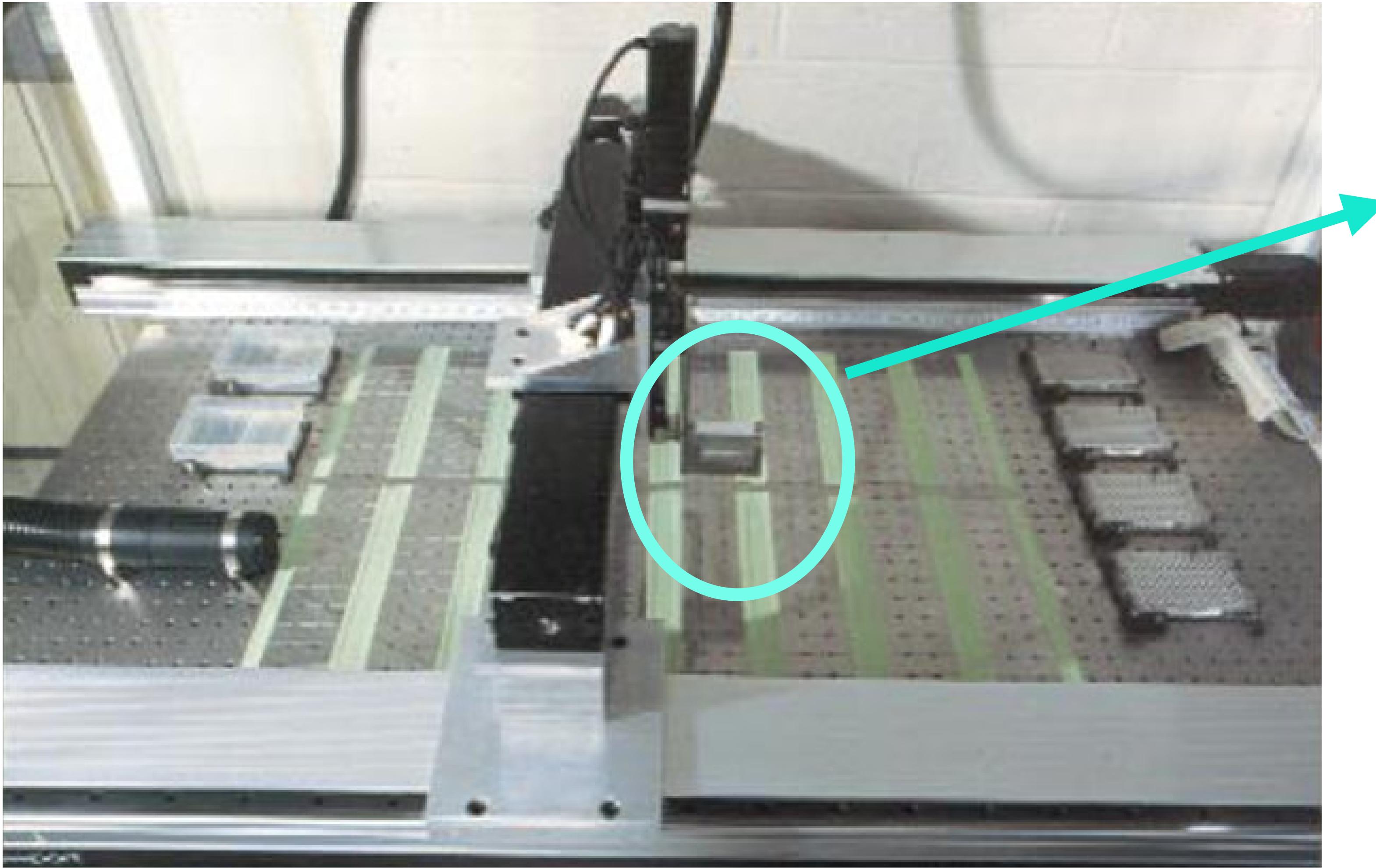
A long time ago...

Array technology (Pat Brown's pipetting machine)



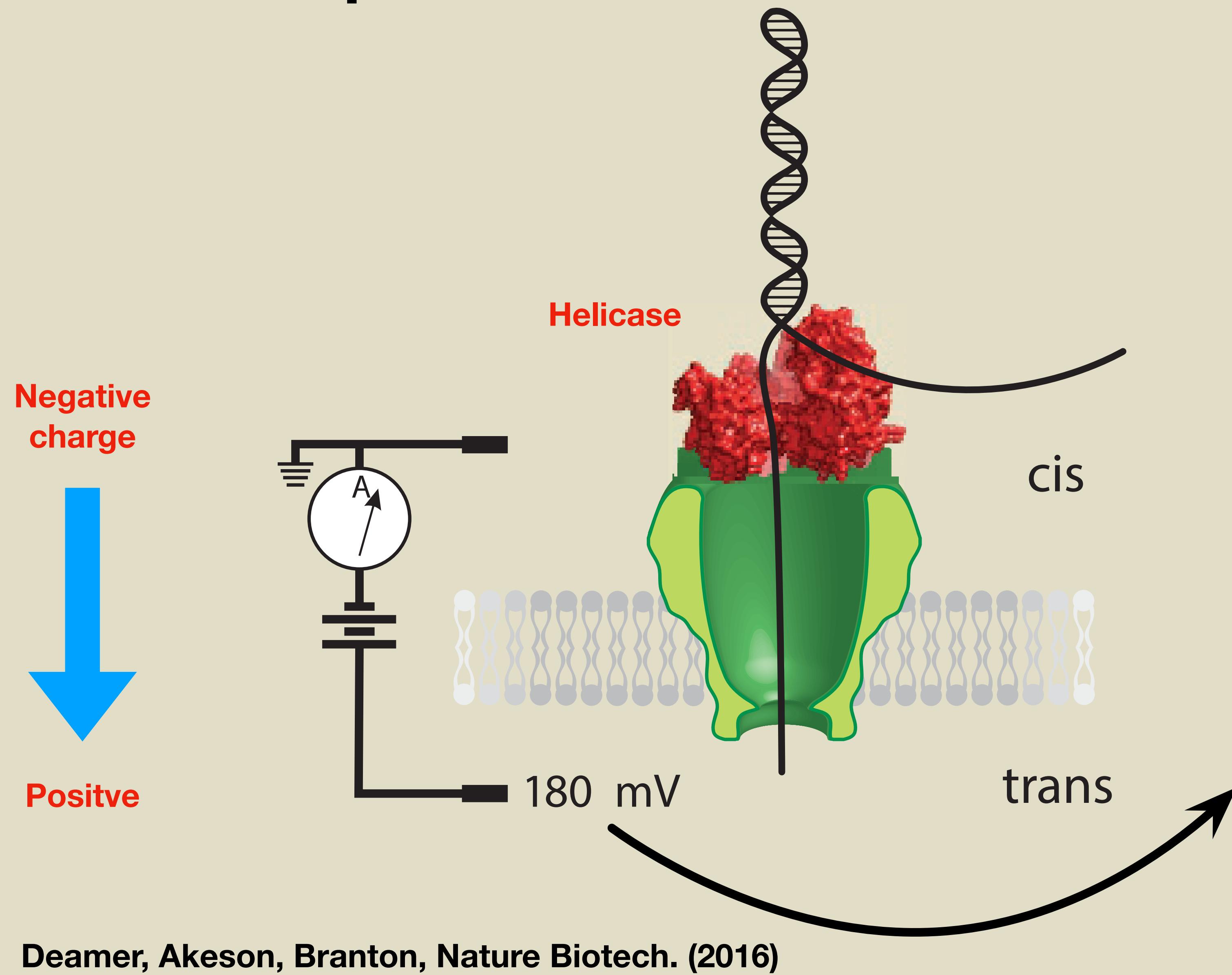
Cheung et al. (1999)

Pat Brown's pipetting machine



Cheung et al. (1999)

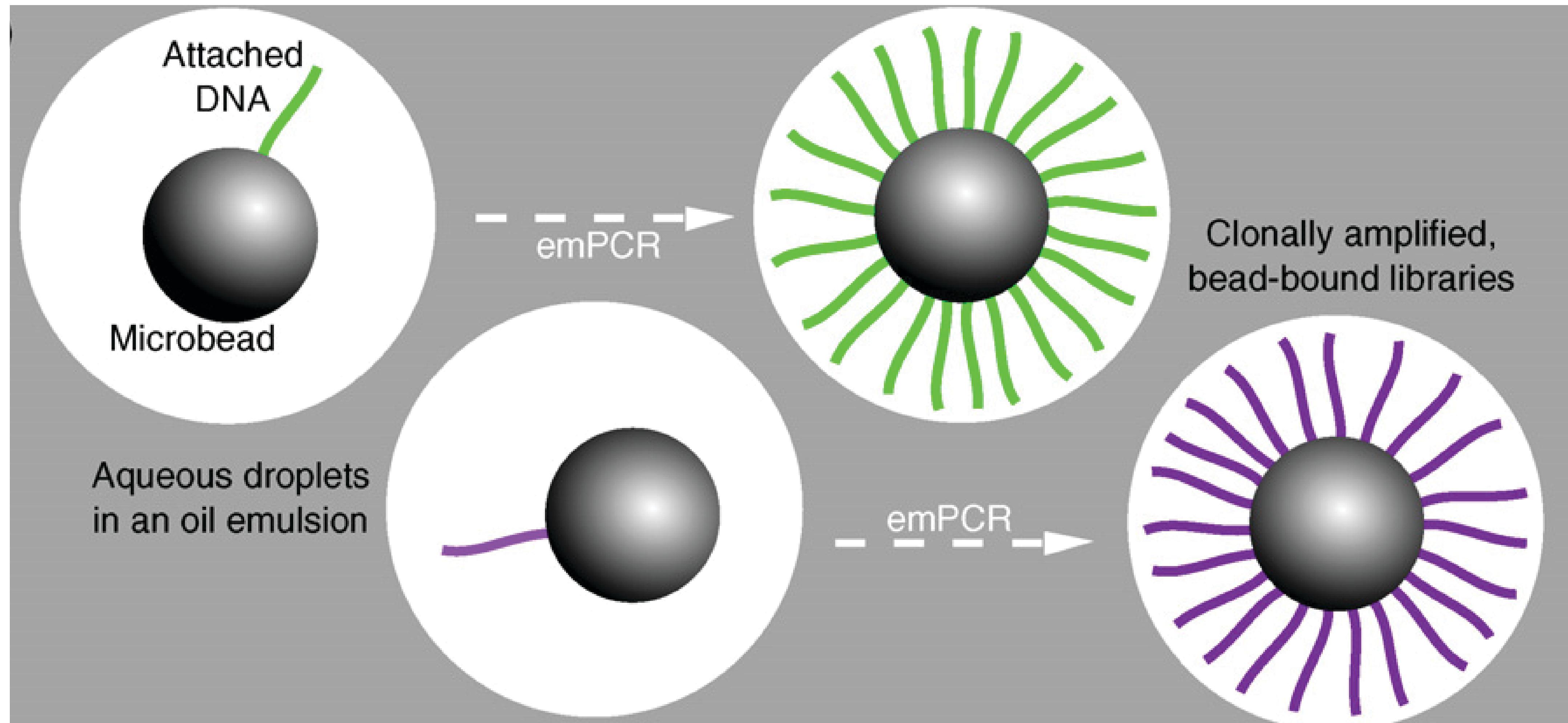
Nanopore: Yet another method for sequencing



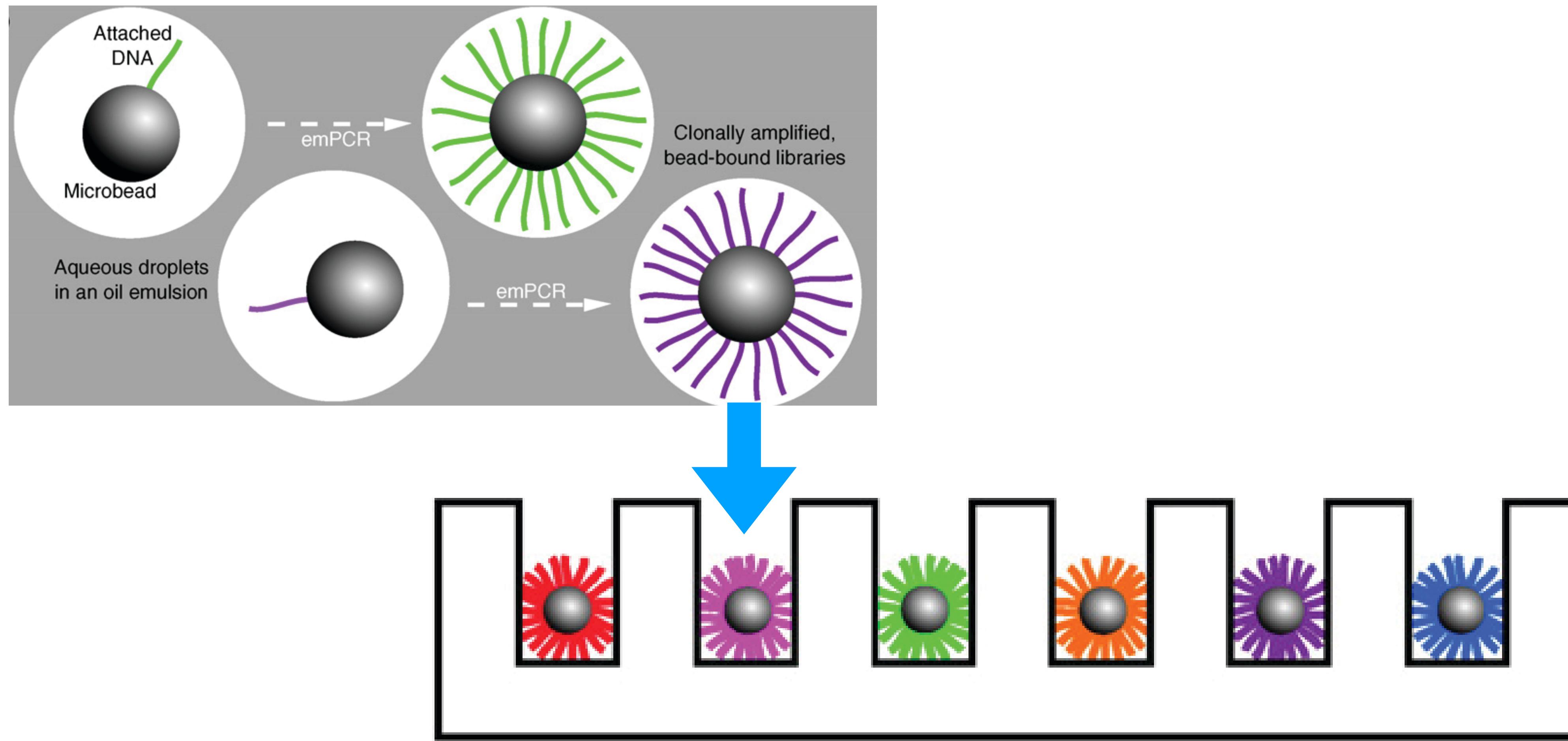
Deamer, Akeson, Branton, Nature Biotech. (2016)

- The ionic conductivity is sensitive to the presence of the nucleotide.
- Characteristic "squiggle"

Micro bead-based applications followed by massively parallel short read sequencing

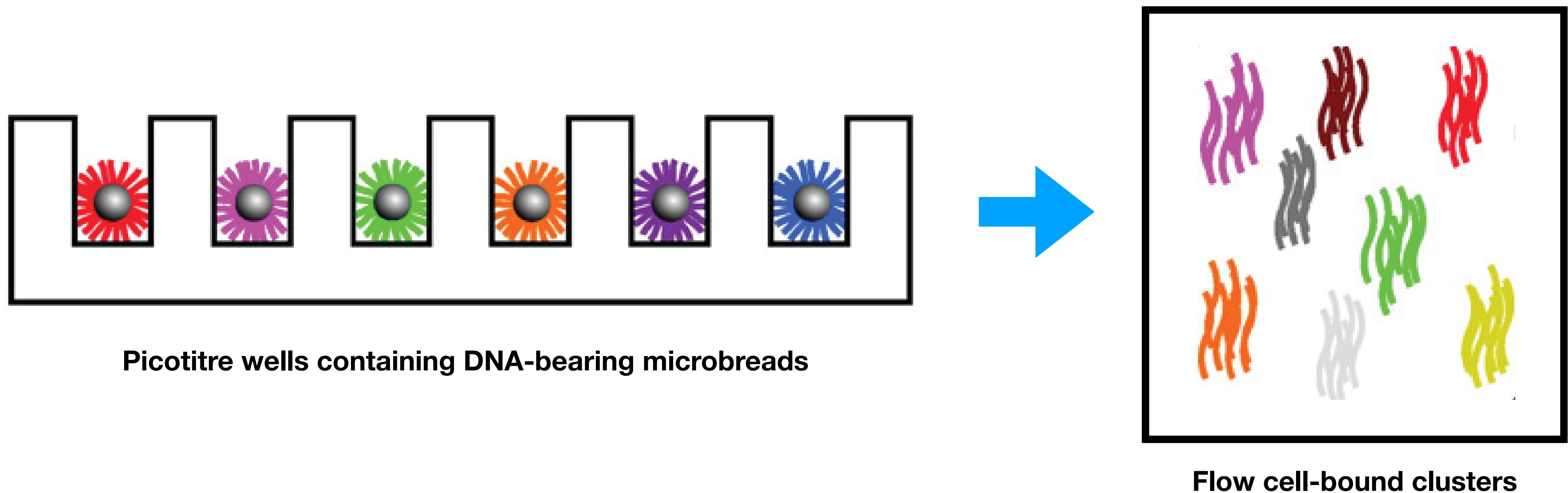


Micro bead-based applications followed by massively parallel short read sequencing

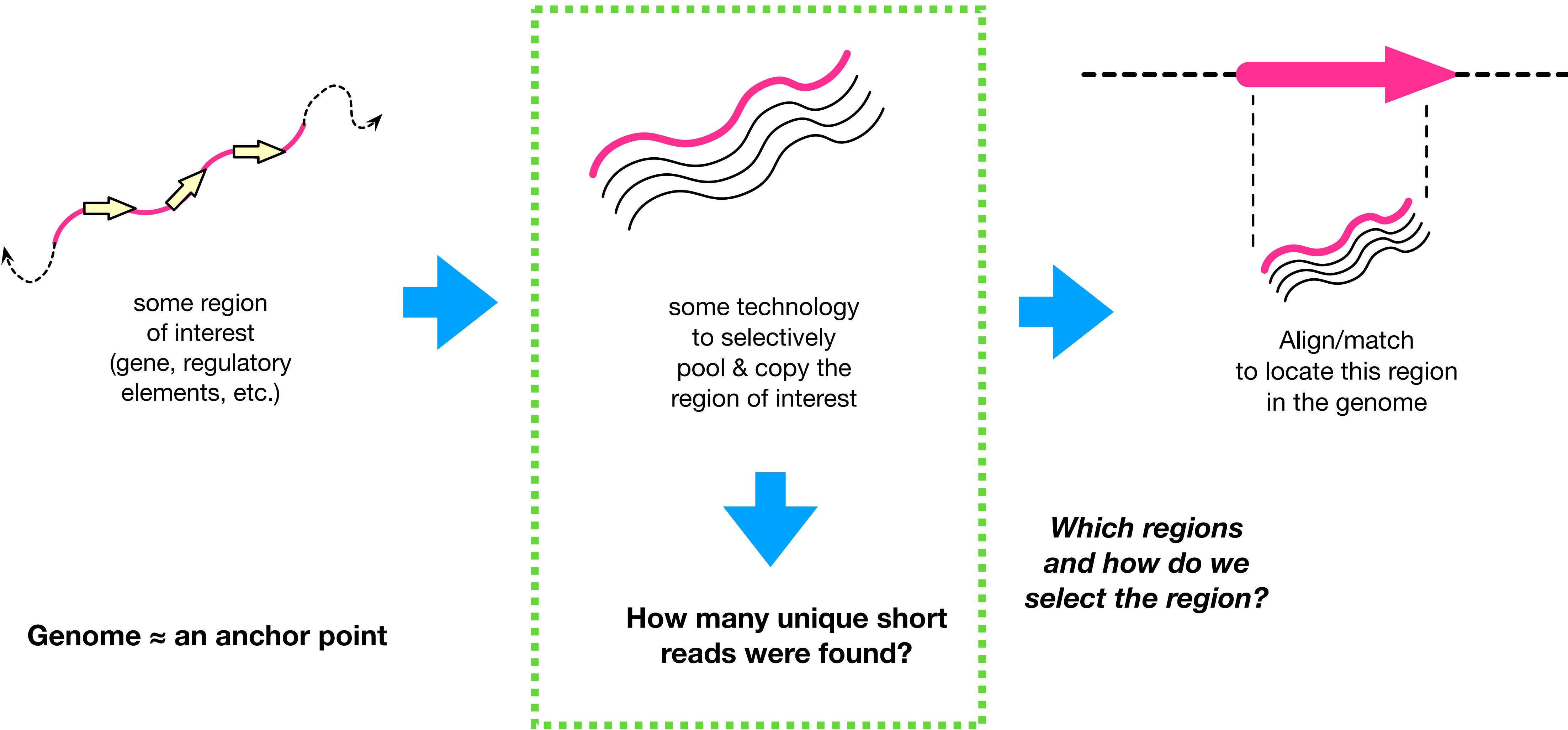


Picotitre wells containing DNA-bearing microbeads

Micro bead-based applications followed by massively parallel short read sequencing



The basic idea of High-throughput methods: selection followed by sequencing



Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

Epigenetic modifications

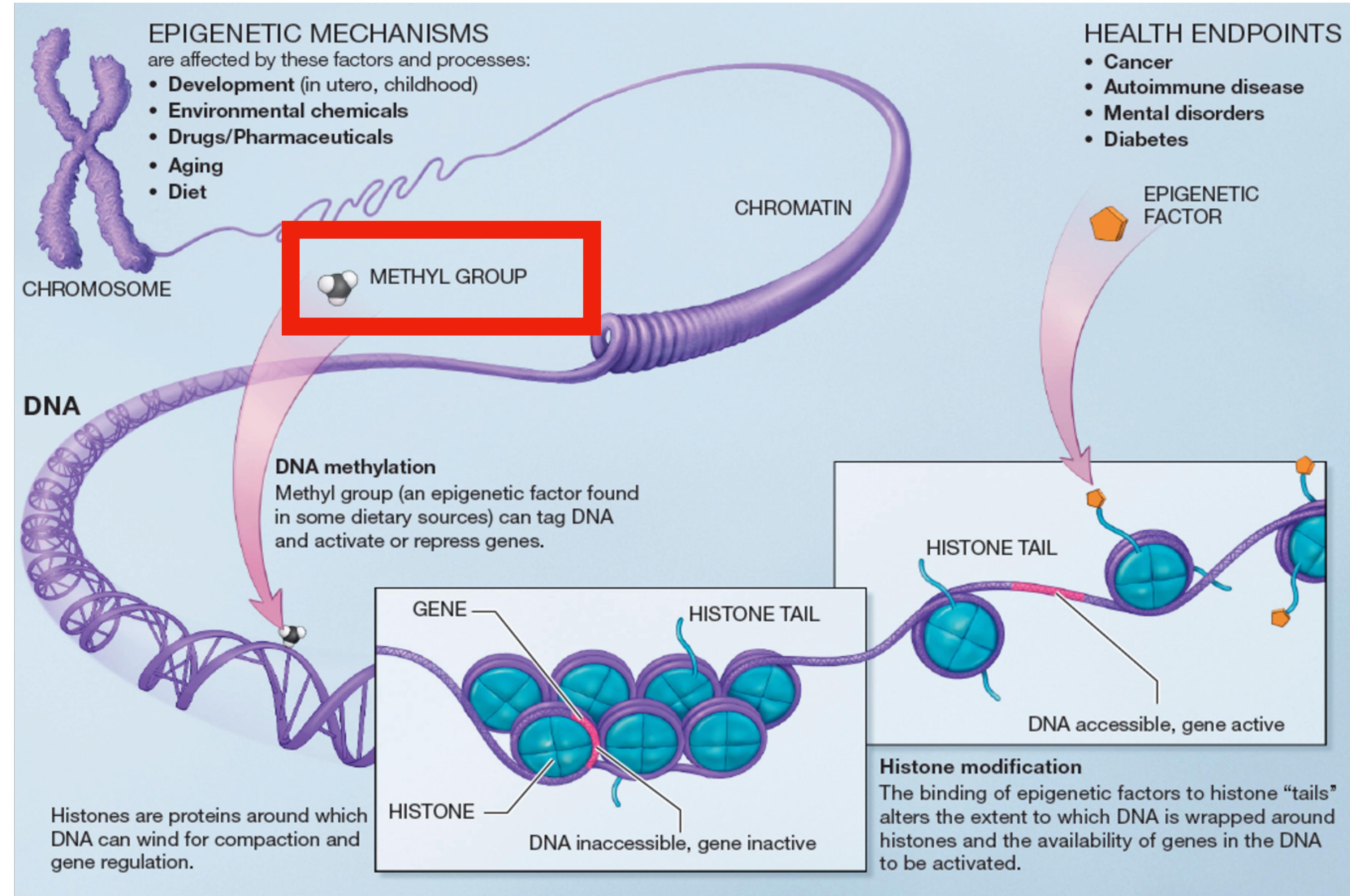
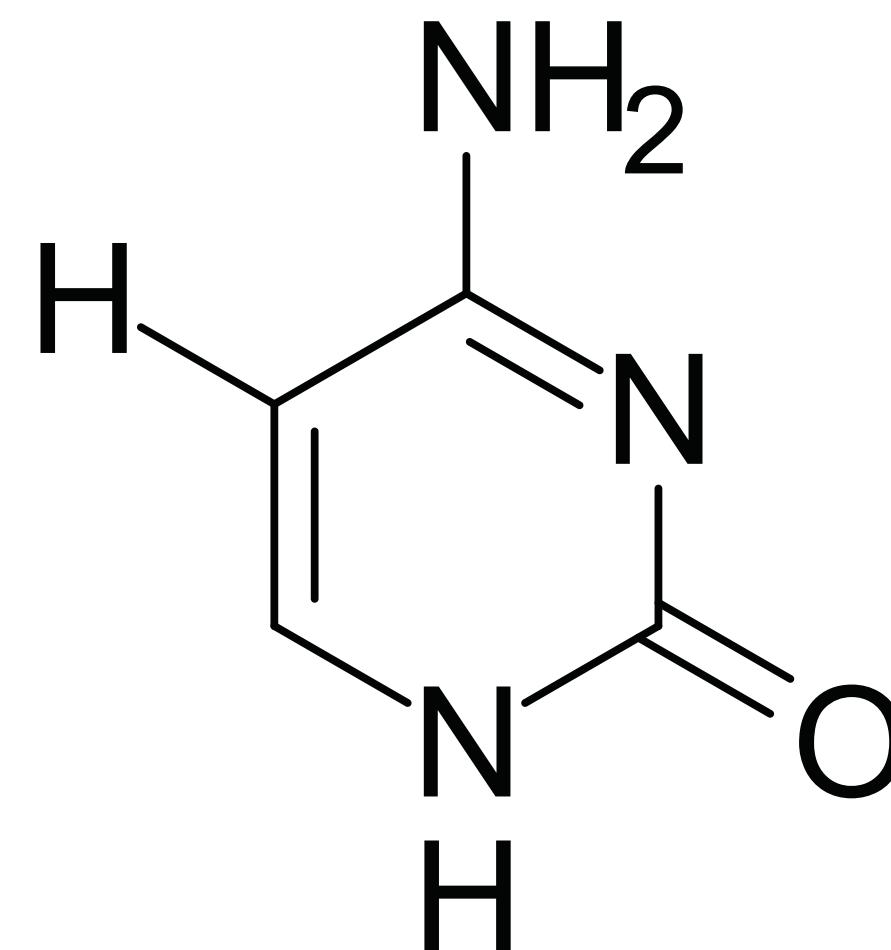


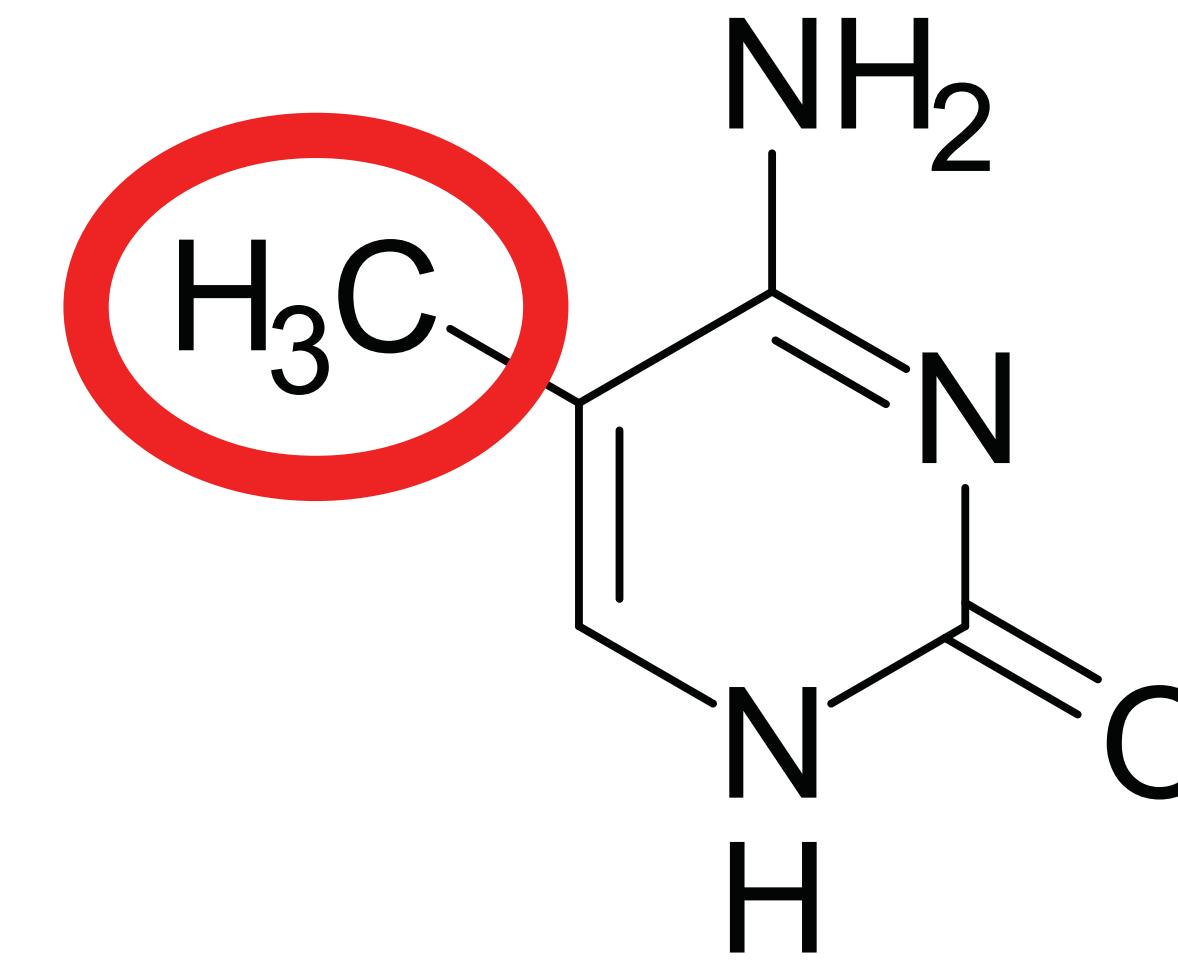
Image source: <http://nihroadmap.nih.gov/epigenomics/>

Histones are proteins around which DNA can wind for compaction and gene regulation.

DNA methylation modifies the biochemical properties of a particular base

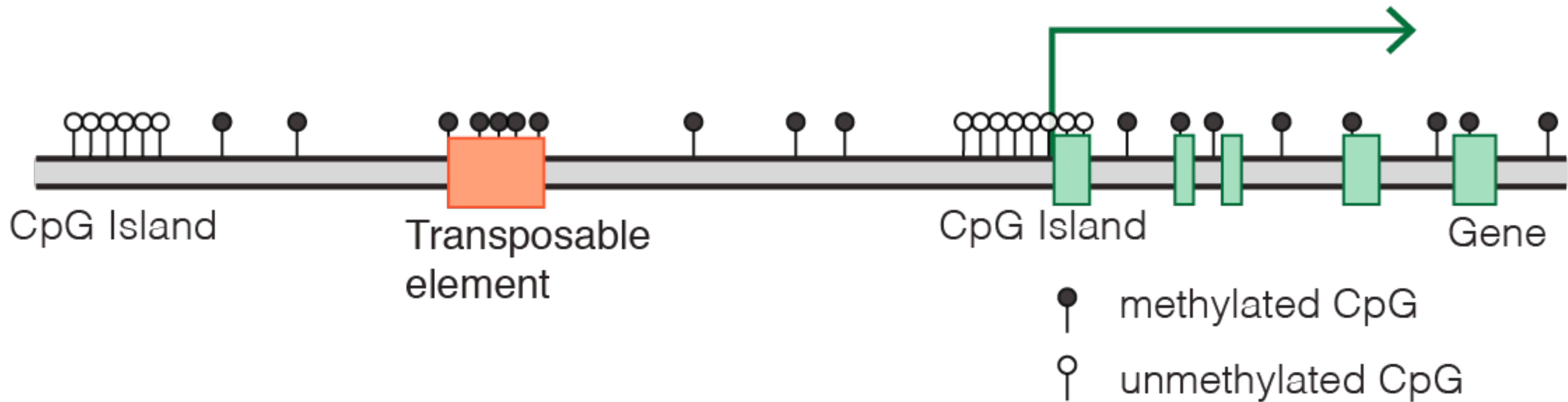


cytosine

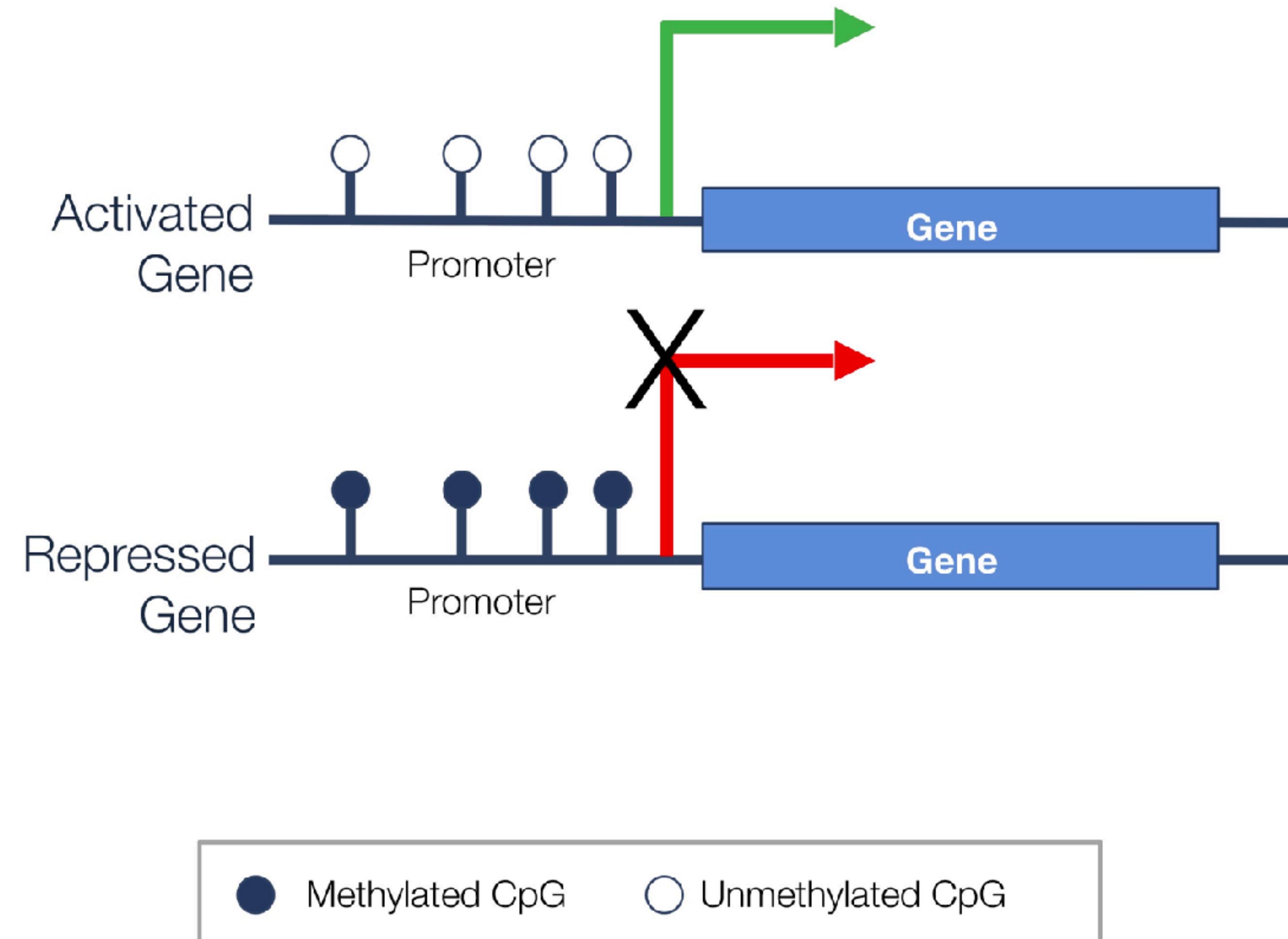


methylated
cytosine

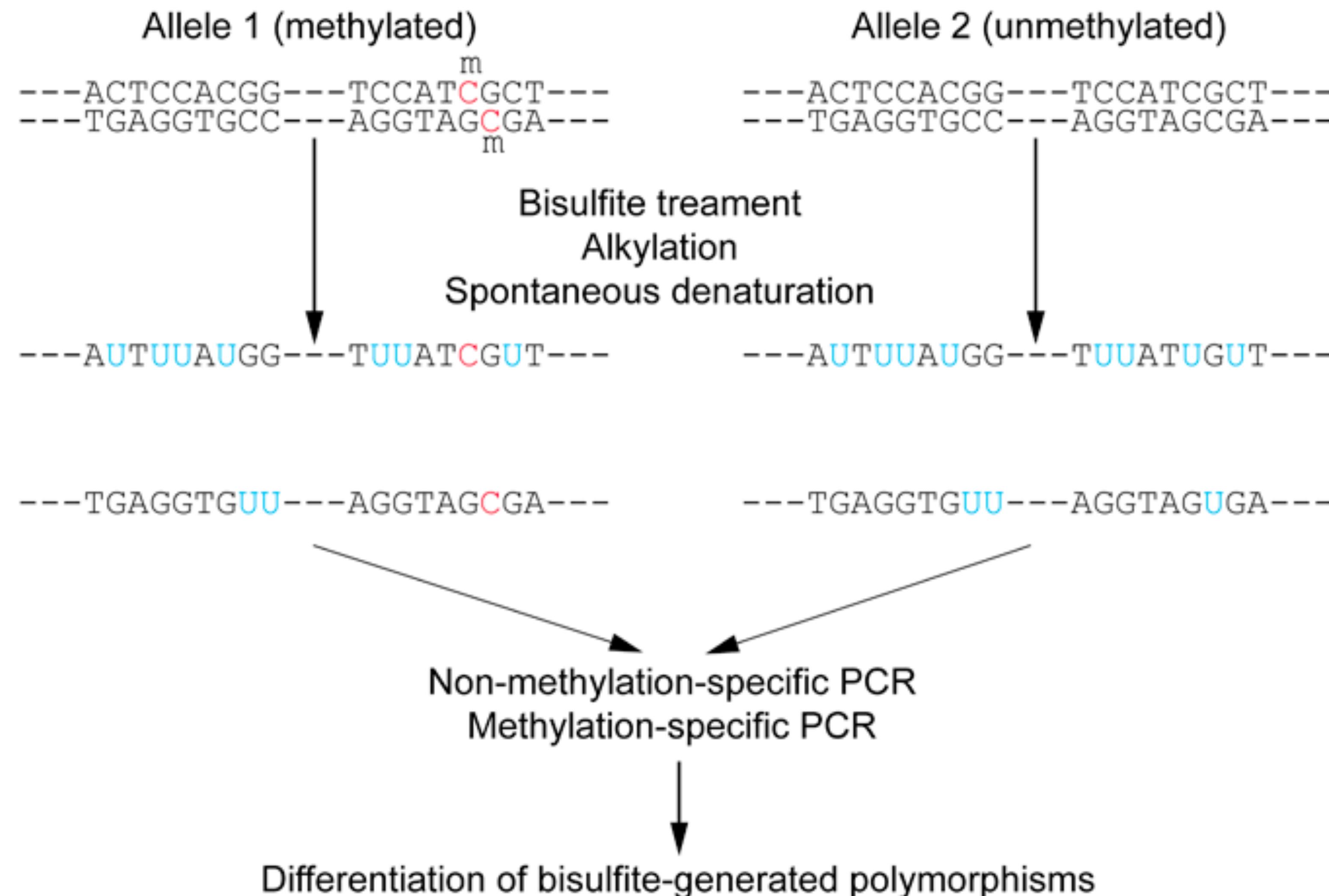
Typical mammalian DNA methylation landscape



Regulatory functions of DNA methylation

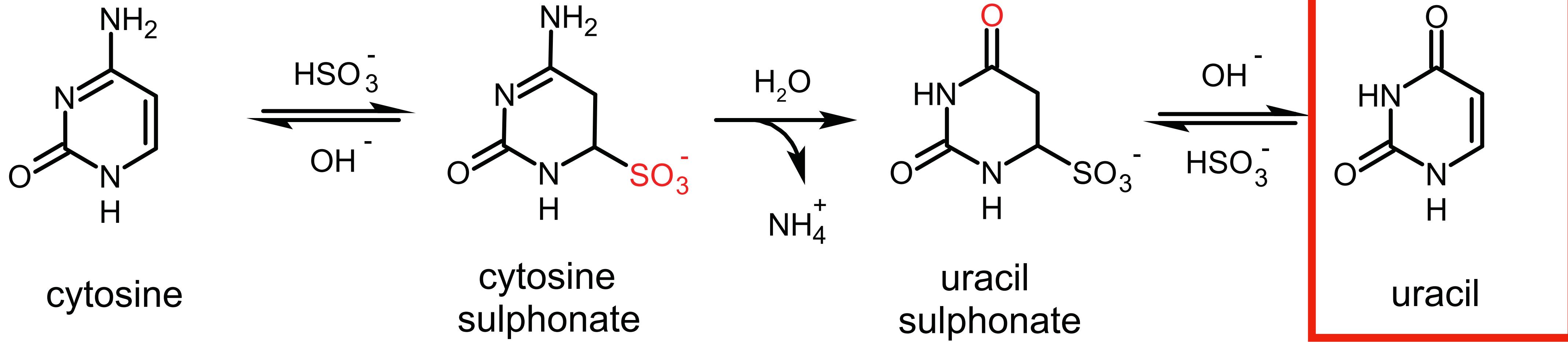


How do we know if a certain "C" is methylated?

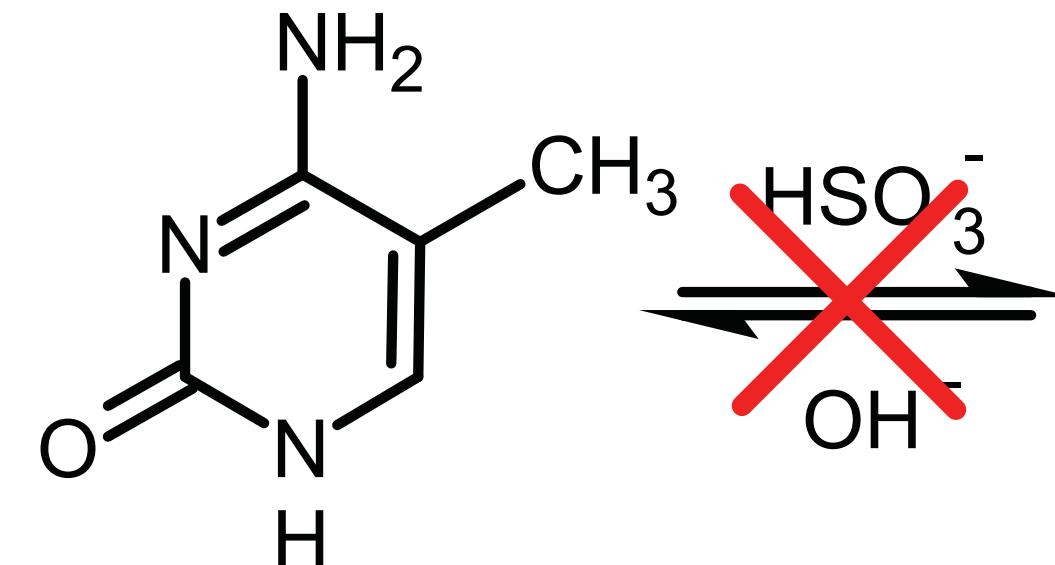


Bisulfite conversion: "unmethylated" C U

For unmethylated/unprotected Cytosine

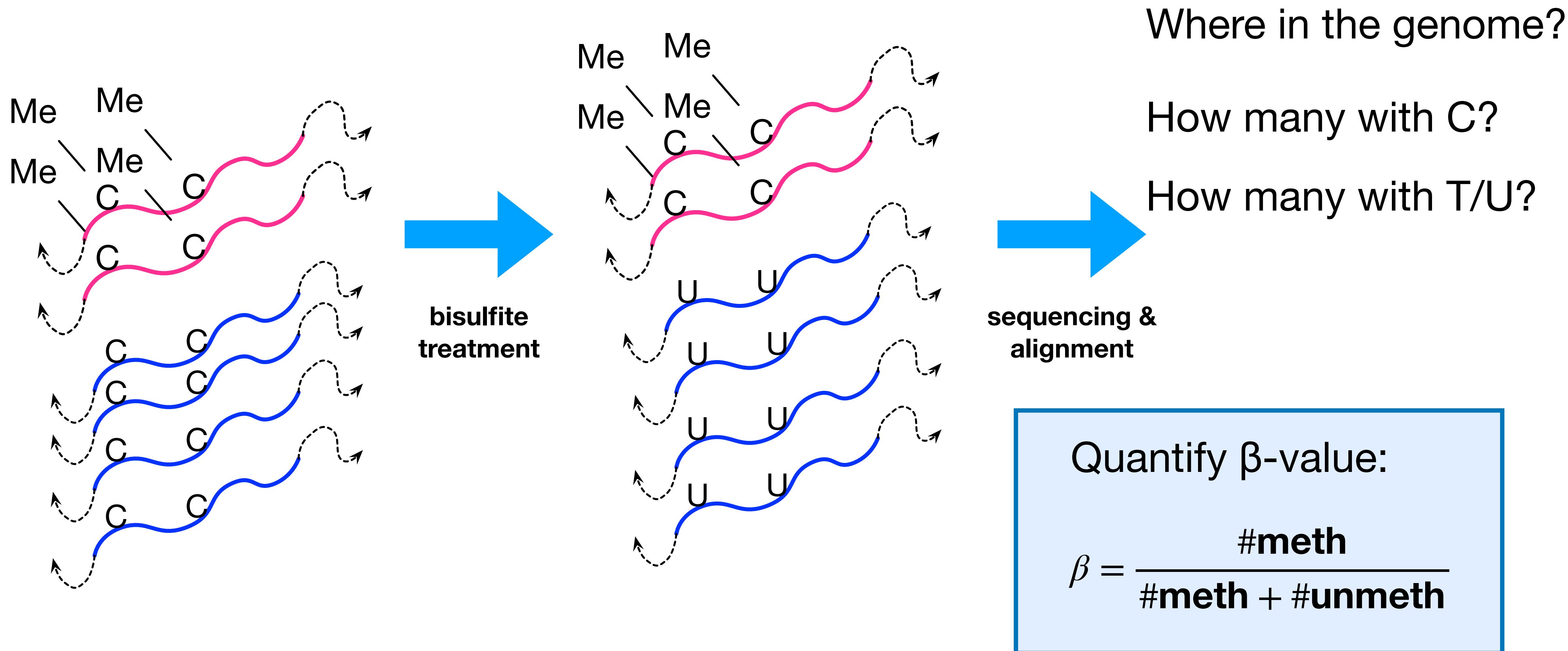


For methylated Cytosine

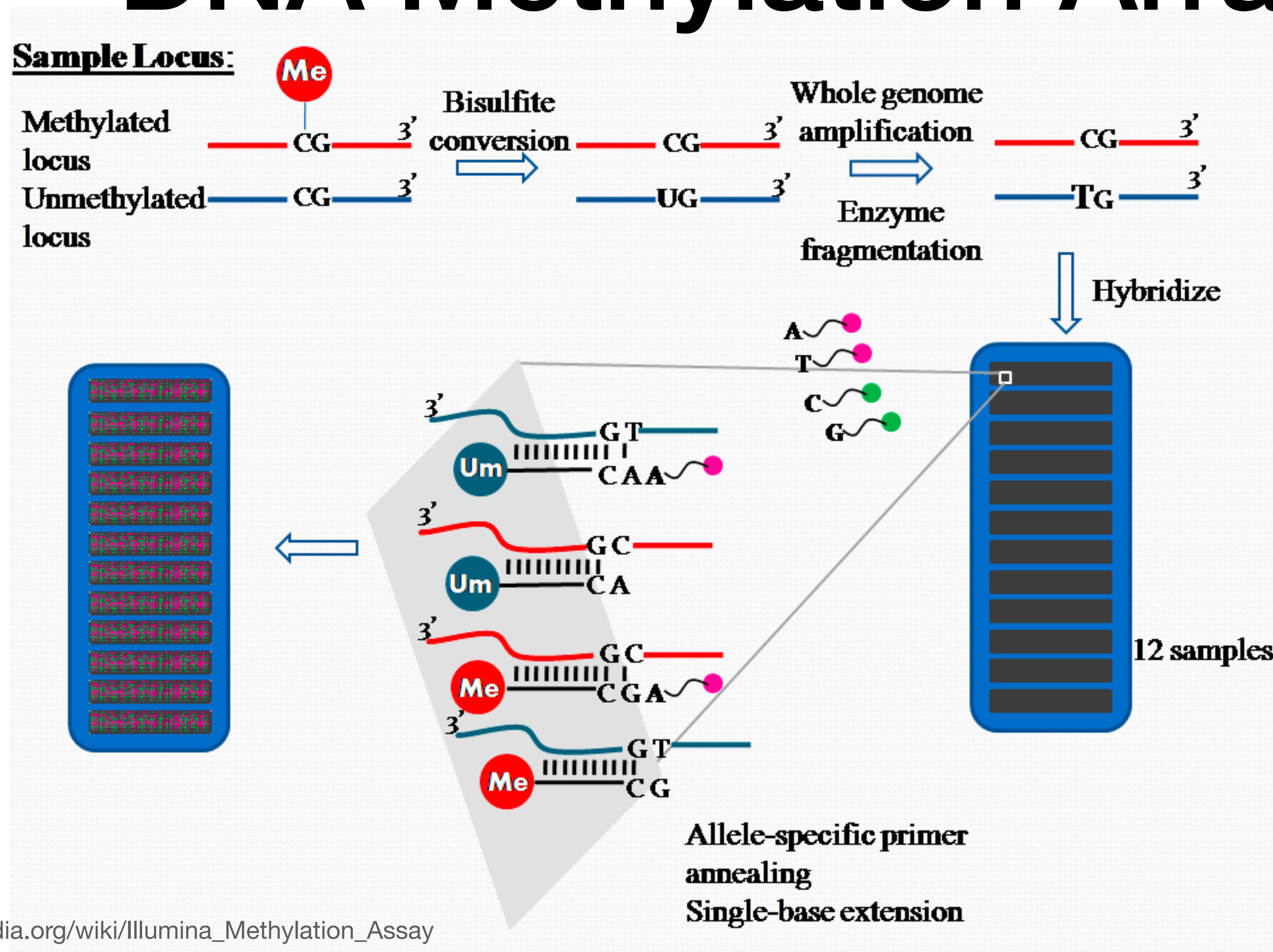


5-methylcytosine

Bisulfite sequencing/array: Distinguish DNAs methylated vs. un-methylated



DNA Methylation Arrays



Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

Beta values and M values

Need some way to summarize the two relative measures for each probe:

1. methylated intensity
2. unmethylated intensity

Beta (β) values

- ▶ Combine two measurements for each probe, and measure level of methylation:

$$\beta = \frac{\text{methylated intensity}}{(\text{methylated intensity} + \text{unmeth intensity})}$$

- ▶ Range from 0 to 1
- ▶ Easily interpretable as a methylation proportion
- ▶ Best used for visualization

M-values

- ▶ Transformed β values:

$$M = \log_2 \left(\frac{\text{methylated intensity}}{\text{unmethylated intensity}} \right) = \log_2 \left(\frac{\beta}{1 - \beta} \right)$$

- ▶ This type of transformation is referred to as the *logit*
 - ▶ maps a proportion onto real line
- ▶ alleviate heteroscedasticity of β values (recall that Binomial(n,p) variance = $np(1-p)$)

Du et al. (2010)

Methylation array analysis

- ▶ Analysis strategy: linear regression on M-values
- ▶ R Packages:
 - ▶ minfi for preprocessing and normalization
 - ▶ limma for individual cytosine analysis
 - ▶ bumphunter or dmrcate for region-level analysis
 - ▶ Workflow that combines these: `methylationArrayAnalysis` - explore in Seminar 7¹

¹Note that this seminar requires installation of a couple of packages that download data and can take some time - please try run the installation step before class time!

A working Example - DNA Methylation Arrays in sorted T cells

“Genome-wide DNA methylation analysis identifies hypomethylated genes regulated by FOXP3 in human regulatory T cells” from Zhang et al. (2013)

- ▶ GEO accession GSE49667
- ▶ Here we use filtered and normalized Bioc objects that were created by this workflow, with details described in Seminar 7

```
mSet <- readRDS("data/mSet.rds")      # see Seminar 7
targets <- readRDS("data/targets.rds") #
```



```
## class: GenomicRatioSet
## dim: 439918 10
## metadata(0):
## assays(2): M CN
## rownames(439918): cg13869341 cg24669183 ... cg08265308 cg14273923
## rowData names(0):
## colnames(10): naive.1 rTreg.2 ... act_naive.9 act_rTreg.10
## colData names(13): Sample_Name Sample_Well ... yMed predictedSex
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
## Preprocessing
```

Metadata values

```
##   Sample_Name Sample_Well Sample_Source Sample_Group Sample_Label
## 1           1         A1       M28      naive      naive
## 2           2         B1       M28      rTreg      rTreg
## 3           3         C1       M28  act_naive  act_naive
## 4           4         D1       M29      naive      naive
## 5           5         E1       M29  act_naive  act_naive
## 6           6         F1       M29  act_rTreg  act_rTreg

##
## act_naive act_rTreg      naive      rTreg
##      3        2          3        2
```

Beta values

```
## [1] 439918      10

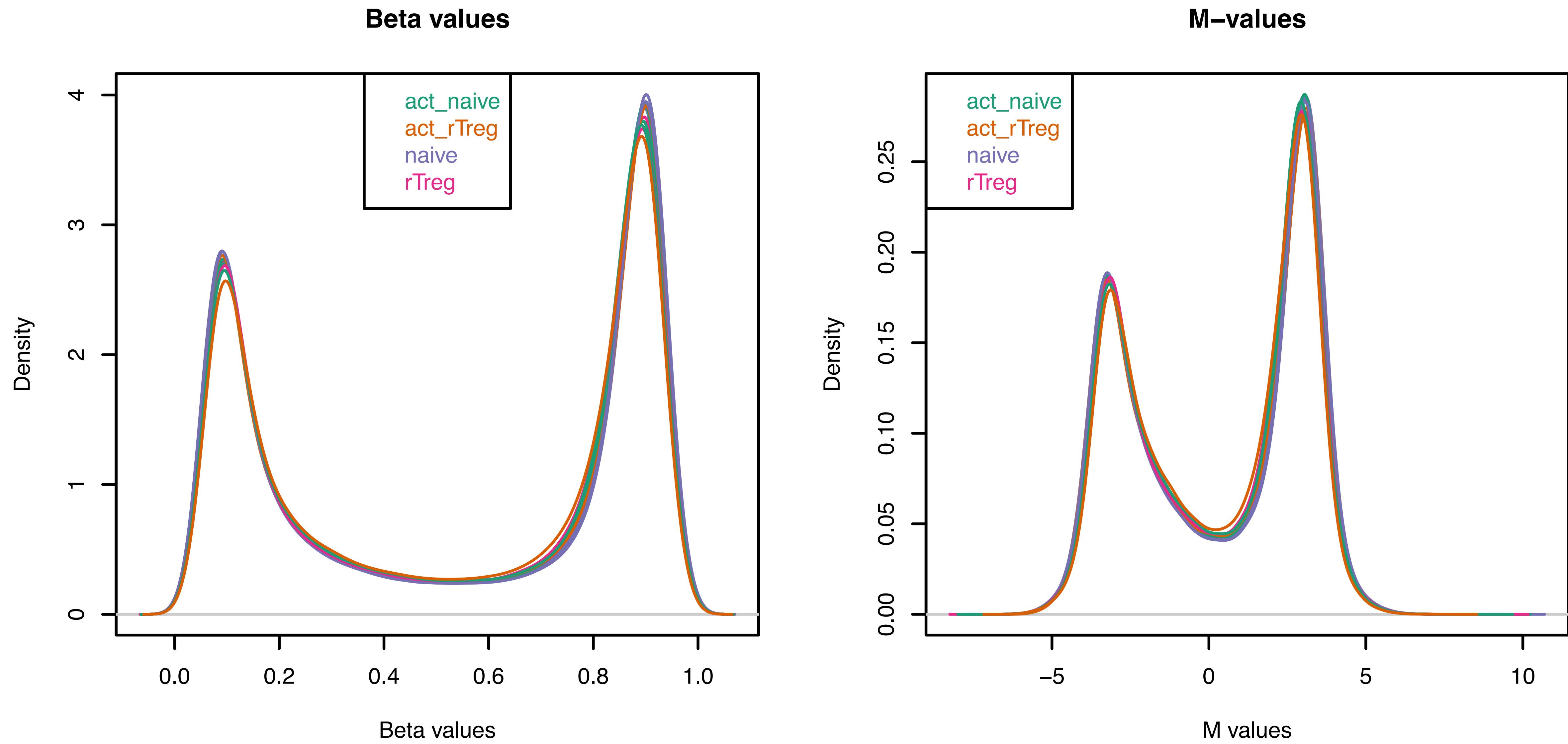
##          naive.1    rTreg.2 act_naive.3    naive.4 act_naive.5 act_rTreg.6
## cg13869341 0.84267937 0.85118462  0.8177504 0.82987650 0.81186174 0.8090798
## cg24669183 0.81812908 0.82489238  0.8293297 0.75610281 0.81967323 0.8187838
## cg15560884 0.77219626 0.74903910  0.7516263 0.77417882 0.77266205 0.7721528
## cg01014490 0.08098986 0.06590459  0.0233755 0.04127262 0.04842397 0.0644404
## cg17505339 0.89439216 0.93822870  0.9471357 0.90520570 0.92641305 0.9286016
## cg11954957 0.74495496 0.79008516  0.7681146 0.84450764 0.75431167 0.8116911
##          naive.7    rTreg.8 act_naive.9 act_rTreg.10
## cg13869341 0.8891851 0.88537940 0.90916748 0.88334231
## cg24669183 0.7903763 0.85304116 0.80930568 0.80979554
## cg15560884 0.7658623 0.75909061 0.78099397 0.78569274
## cg01014490 0.0245281 0.02832358 0.07740468 0.04640659
## cg17505339 0.8889361 0.87205348 0.90099782 0.93508348
## cg11954957 0.7832207 0.84929777 0.84719430 0.83350220
```

M values

```
## [1] 439918      10

##          naive.1    rTreg.2 act_naive.3    naive.4 act_naive.5 act_rTreg.6
## cg13869341  2.421276  2.515948   2.165745  2.286314   2.109441  2.083313
## cg24669183  2.169414  2.235964   2.280734  1.632309   2.184435  2.175771
## cg15560884  1.761176  1.577578   1.597503  1.777486   1.764999  1.760819
## cg01014490 -3.504268 -3.825119  -5.384735 -4.537864  -4.296526 -3.859792
## cg17505339  3.082191  3.924931   4.163206  3.255373   3.654134  3.701096
## cg11954957  1.546401  1.912204   1.727910  2.441267   1.618331  2.107829
##          naive.7    rTreg.8 act_naive.9 act_rTreg.10
## cg13869341  3.004332  2.949430   3.323265  2.920691
## cg24669183  1.914738  2.537203   2.085423  2.090007
## cg15560884  1.709728  1.655782   1.834341  1.874285
## cg01014490 -5.313593 -5.100400  -3.575205 -4.360973
## cg17505339  3.000690  2.768876   3.185991  3.848438
## cg11954957  1.853192  2.494570   2.470994  2.323683
```

Comparing beta and M values



Differentially methylated cytosines (DMC)

Which cytosines/probes are differentially methylated between naive & activated naive T cells?

```
## this is the factor of interest
cellType <- factor(targets$Sample_Group)

## this is the individual effect that we need to account for
individual <- factor(targets$Sample_Source)

## use the above to create a design matrix
design <- model.matrix(~ cellType + individual,
                       data = targets)

design
```

| | (Intercept) | cellTypeact_rTreg | cellTypenaive | cellTyperTreg | individualM29 |
|------|-------------|-------------------|---------------|---------------|---------------|
| ## 1 | 1 | 0 | 1 | 0 | 0 |
| ## 2 | 1 | 0 | 0 | 1 | 0 |
| ## 3 | 1 | 0 | 0 | 0 | 0 |
| ## 4 | 1 | 0 | 1 | 0 | 1 |
| ## 5 | 1 | 0 | 0 | 0 | 1 |
| ## 6 | 1 | 1 | 0 | 0 | 1 |
| ## 7 | 1 | 0 | 1 | 0 | 0 |
| ## 8 | 1 | 0 | 0 | 1 | 0 |
| ## 9 | 1 | 0 | 0 | 0 | 0 |

Differential methylation: limma on M-values

Here, the reference cell type is act_naive cells

```
## fit the linear model (one for each probe)
mfit <- lmFit(mVals, design)
cfit <- eBayes(mfit)

## look at top DM CpGs between naive and ref (act_naive)
topTable(cfit, coef="cellTypenaive")
```

| | logFC | AveExpr | t | P.Value | adj.P.Val | B |
|---------------|-----------|-------------|-----------|--------------|-------------|----------|
| ## cg15459165 | -3.892074 | -1.47895212 | -22.09541 | 1.993958e-08 | 0.002793543 | 6.810420 |
| ## cg19704755 | -4.139885 | 0.09100138 | -20.35779 | 3.787335e-08 | 0.002793543 | 6.594795 |
| ## cg13531460 | 3.484448 | -0.63132888 | 20.23020 | 3.978221e-08 | 0.002793543 | 6.577256 |
| ## cg17137500 | -3.007052 | -2.47635924 | -20.21569 | 4.000601e-08 | 0.002793543 | 6.575245 |
| ## cg17048073 | -3.650550 | -0.87247048 | -20.04737 | 4.270889e-08 | 0.002793543 | 6.551674 |
| ## cg03899643 | 4.603864 | 0.02760746 | 19.86354 | 4.589787e-08 | 0.002793543 | 6.525407 |
| ## cg19827923 | 3.316270 | 0.15932891 | 19.82438 | 4.661126e-08 | 0.002793543 | 6.519739 |
| ## cg26800893 | -3.080184 | -2.39958457 | -19.60718 | 5.080116e-08 | 0.002793543 | 6.487839 |
| ## cg14688905 | -3.119261 | -1.31909419 | -18.94774 | 6.635412e-08 | 0.003243375 | 6.385901 |
| ## cg21837189 | -3.717982 | -1.08764112 | -18.00052 | 9.895908e-08 | 0.003769603 | 6.224845 |

Summary for all coefficients

Significant tests (BH-adjusted p-value less than 0.05) of the null hypotheses: each coefficient equal to zero

```
##          (Intercept) cellTypeact_rTreg cellTypenaive cellTyperTreg individualM29
## Down      184747             919        400         827           1367
## NotSig    14791            438440     439291       438570        437499
## Up       240380             559        227         521           1052
##          individualM30
## Down      1245
## NotSig   437716
## Up       957
```

Note that the first column tests whether the reference group (act_naive) has an M value of zero (not usually of interest)

Interpretation of parameters and effect sizes

- ▶ To interpret parameter estimates from M-values on β scale, need to back-transform:
 - ▶ Recall: $M = \log_2\left(\frac{\text{methylated intensity}}{\text{unmethylated intensity}}\right) = \log_2(\beta/(1 - \beta))$
 - ▶ Solving for beta gives $\beta = 2^M / (1 + 2^M)$

```
# fitted mean M-value for cg15459165 in ref group:  
(M_act_naive <- cfit$coefficients["cg15459165", "(Intercept)"])
```

```
## [1] 0.4121952  
(M_naive <- M_act_naive + cfit$coefficients["cg15459165", "cellTypenaive"])
```

```
## [1] -3.479879  
# back transform to beta values  
inv_logit <- function(x){ 2^x / (1+2^x) }  
inv_logit(M_act_naive)
```

```
## [1] 0.570946  
inv_logit(M_naive)  
  
## [1] 0.08225703
```

Summary of array DMC analysis

- ▶ **Just like for limma analysis of gene expression microarray:** Can use additive models, models with interactions, continuous and categorical variables
 - ▶ **interpretation of hypothesis tests, contrasts** with `makeContrasts` (for other comparisons not represented by coefficients in our design matrix): follows as before
 - ▶ **interpretation of parameters/effect sizes:** keep in mind that M-values are *transformed* proportions

Summary of array DMC analysis

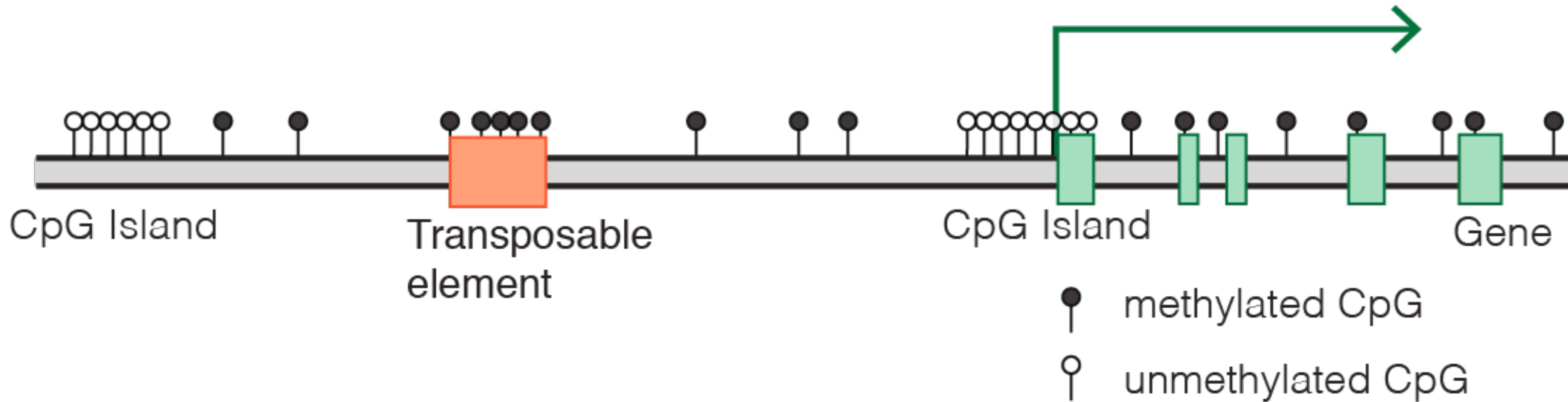
- ▶ Just like for limma analysis of gene expression microarray: Can use additive models, models with interactions, continuous and categorical variables
 - ▶ interpretation of hypothesis tests, contrasts with makeContrasts (for other comparisons not represented by coefficients in our design matrix): follows as before
 - ▶ interpretation of parameters/effect sizes: keep in mind that M-values are *transformed* proportions
- ▶ Annotation packages (e.g. IlluminaHumanMethylation450kanno.ilmn12.hg19) allow for adding information about each CpG
 - ▶ chromosome and position
 - ▶ strand
 - ▶ whether in CpG Island

Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

Isn't DNA methylation analysis the same as RNA?

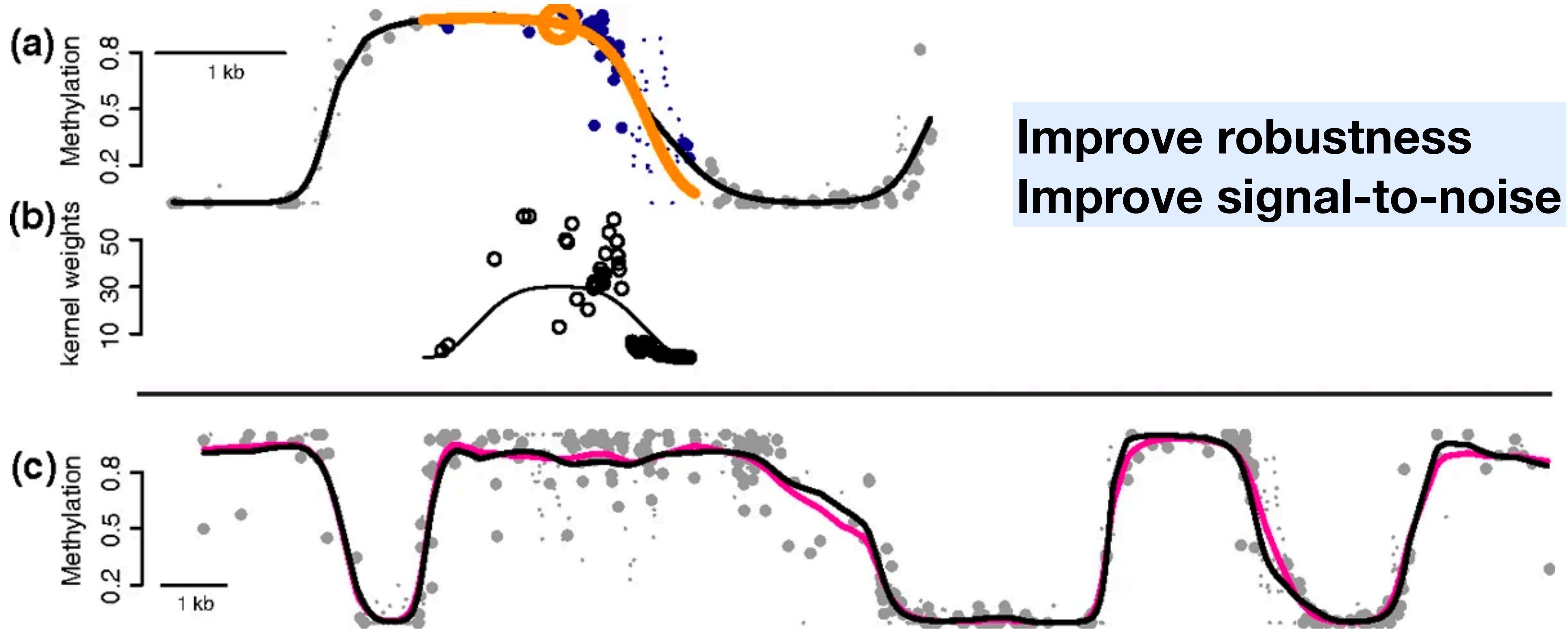
No, so many CpGs exist in the vicinity of a gene



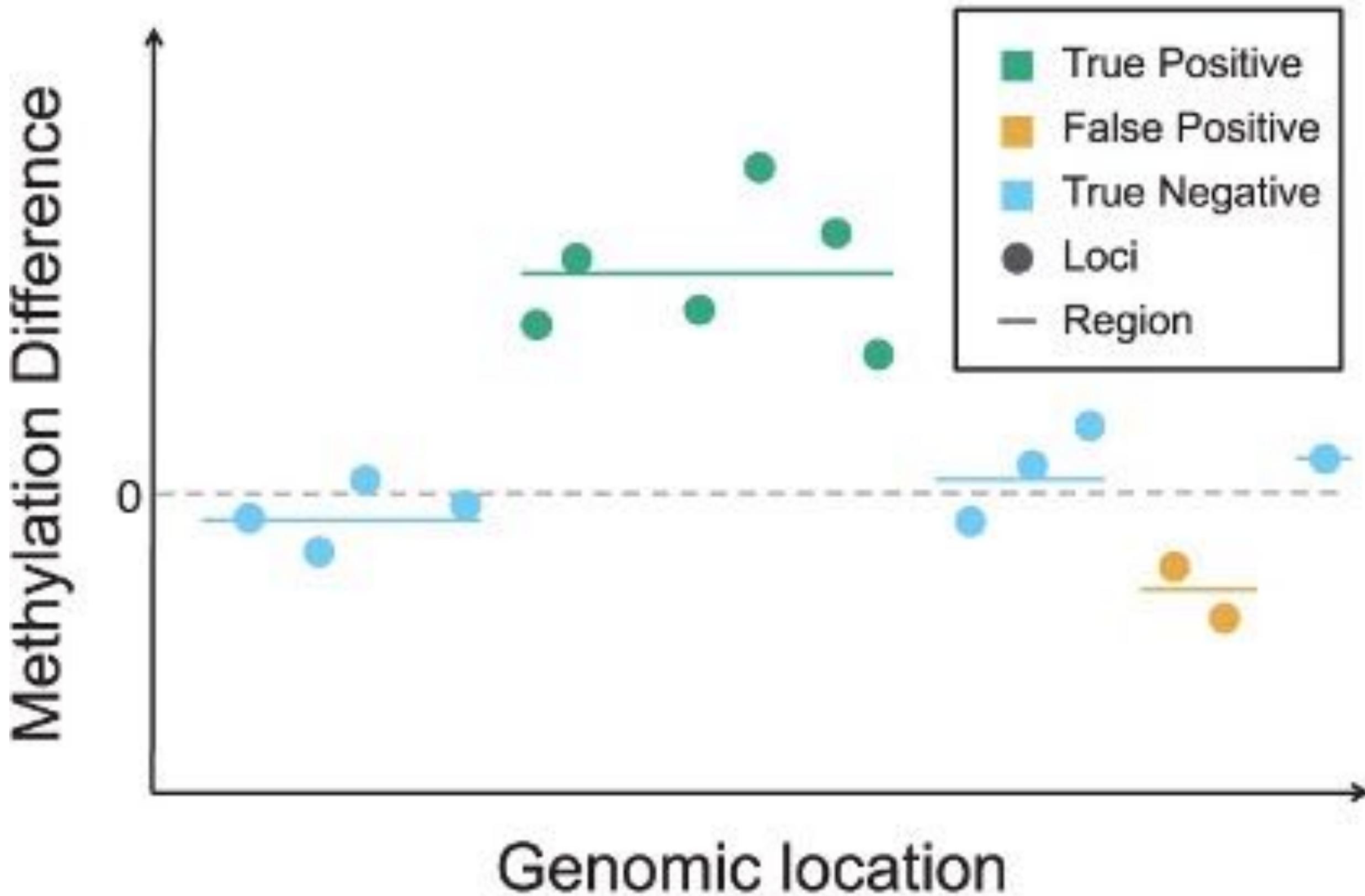
Who are your neighbours?

https://en.wikipedia.org/wiki/DNA_methylation

Why DMR: borrow information across genome



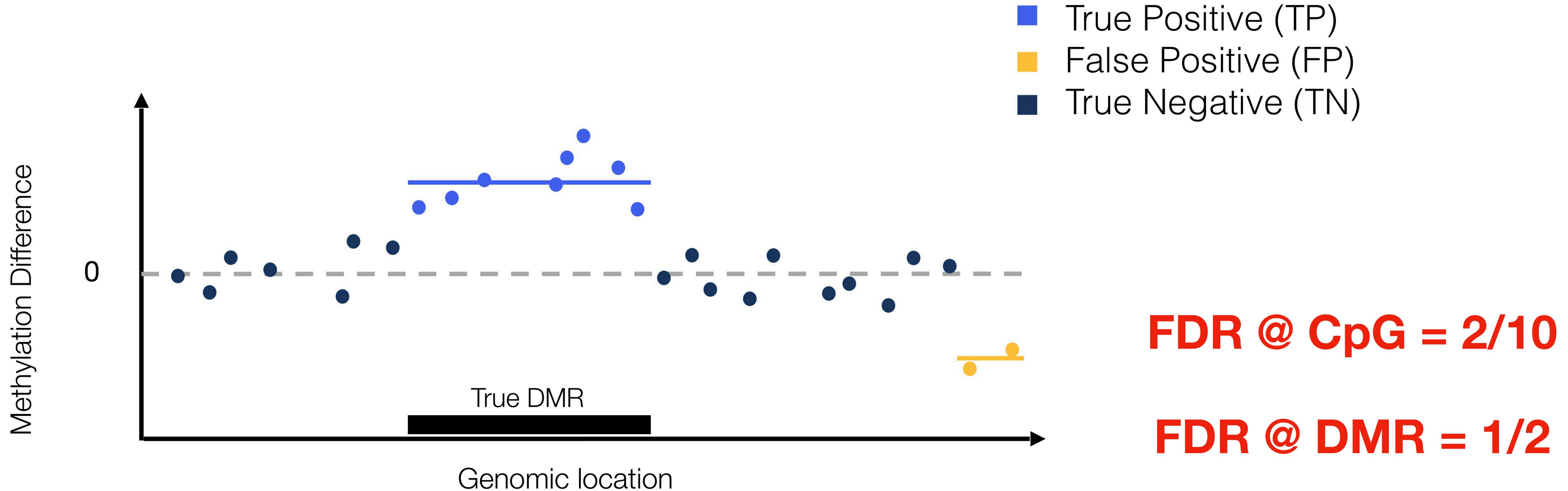
CpG-level analysis may be too optimistic



What is the fundamental assumption of multiple hypothesis correction?

Can you treat each CpG and the corresponding hypothesis test independently?

Why DMR: correct false discovery calibration



$$\text{False Discovery Rate (FDR)} = E \left[\frac{FP}{FP + TP} \right]$$

Korthauer et al. (2019)

First of all, how do we find the regions?

Published by Oxford University Press on behalf of the International Epidemiological Association
© The Author 2012; all rights reserved.

International Journal of Epidemiology 2012;41:200–209
doi:10.1093/ije/dyr238

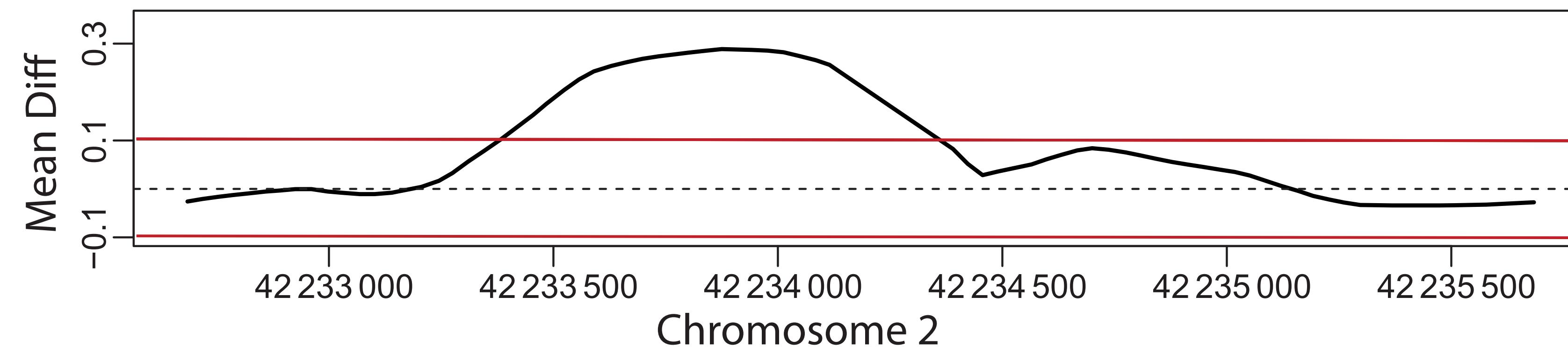
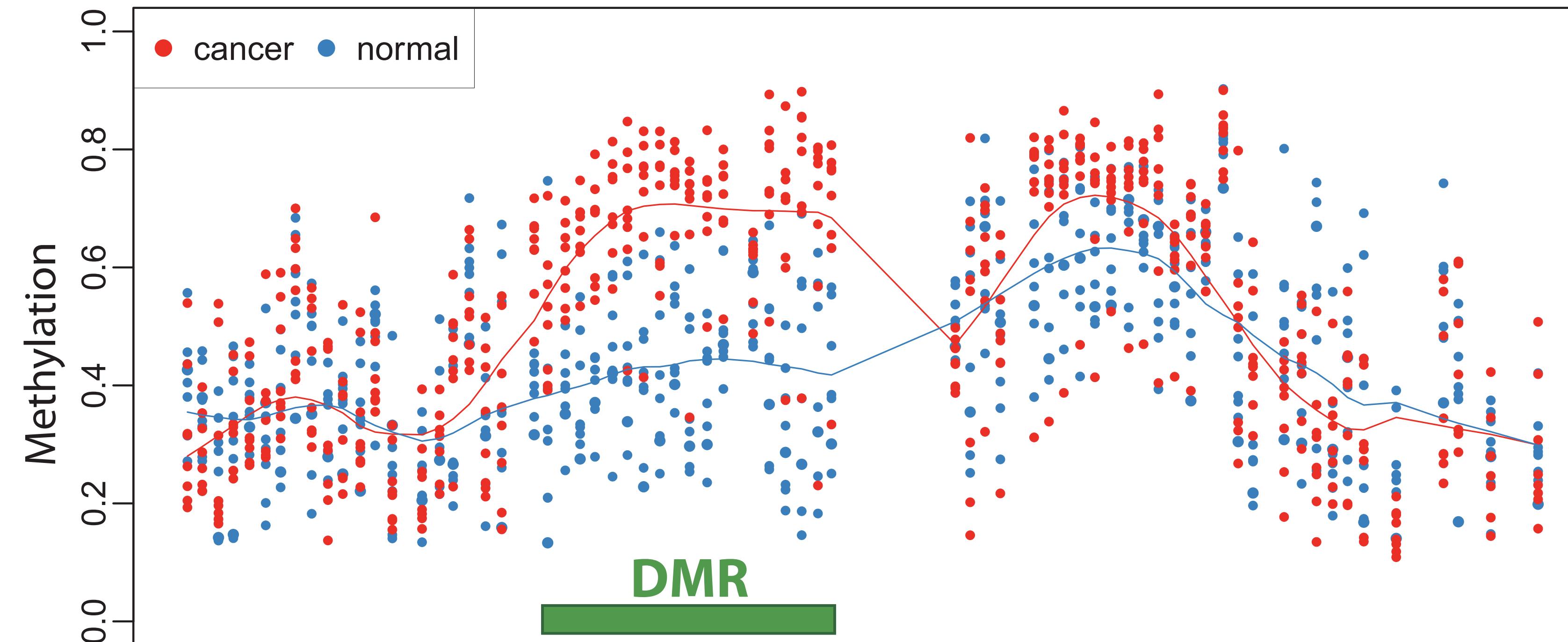
Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies

**Andrew E Jaffe,^{1,2,3} Peter Murakami,³ Hwajin Lee,³ Jeffrey T Leek,¹ M Daniele Fallin,^{1,2,3,4}
Andrew P Feinberg^{1,3,4} and Rafael A Irizarry^{1,3*}**

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, ²Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, ³Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, MD, USA and ⁴Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

*Corresponding author. Department of Biostatistics, 615 N. Wolfe St E3620, Baltimore, MD 21205. E-mail: rafa@jhu.edu

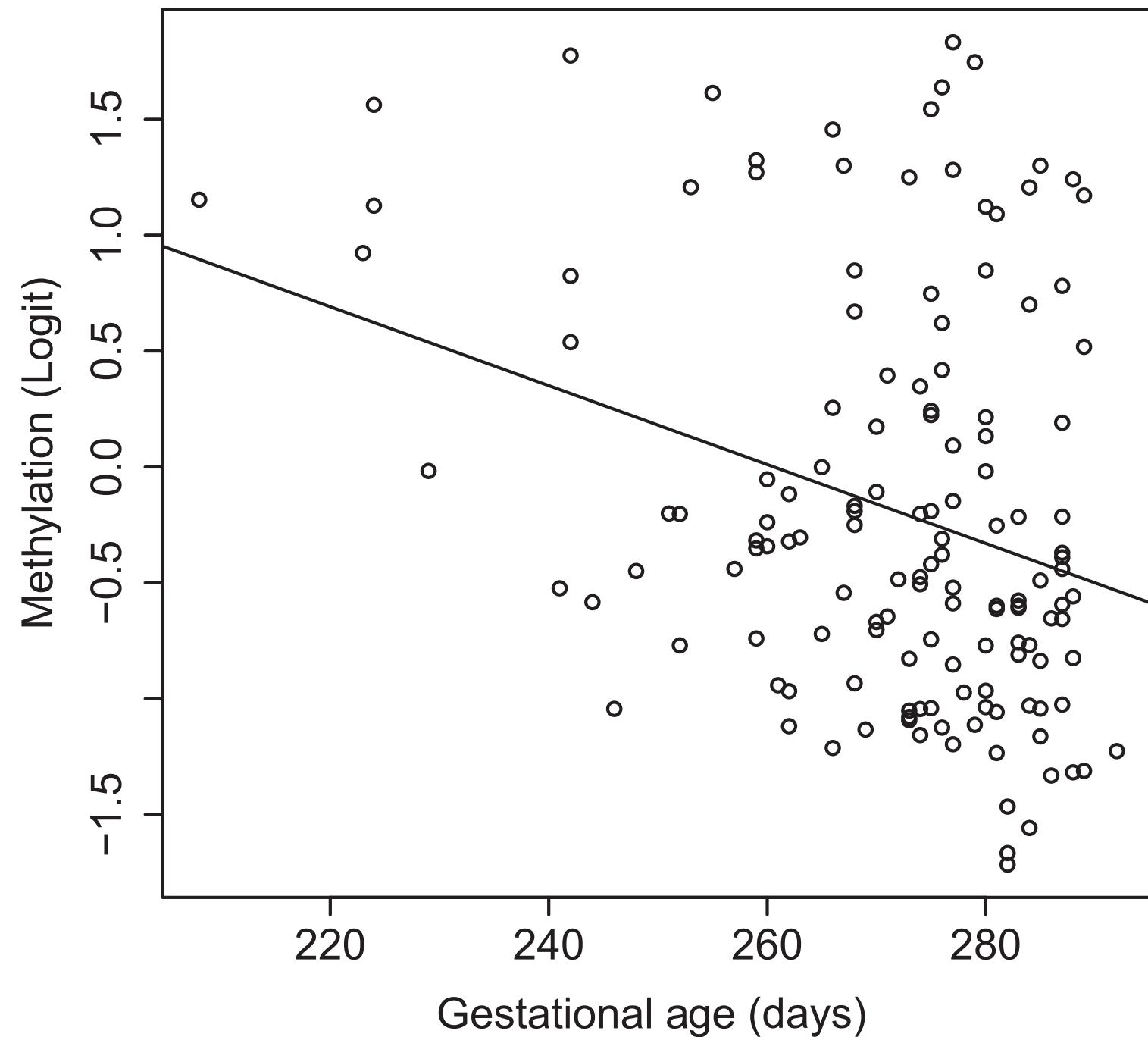
Goal: Identify differentially methylated regions



Jaffe et al. (2012)

DMR bump hunter

Treat each CpG independently
Perform association, thus get β

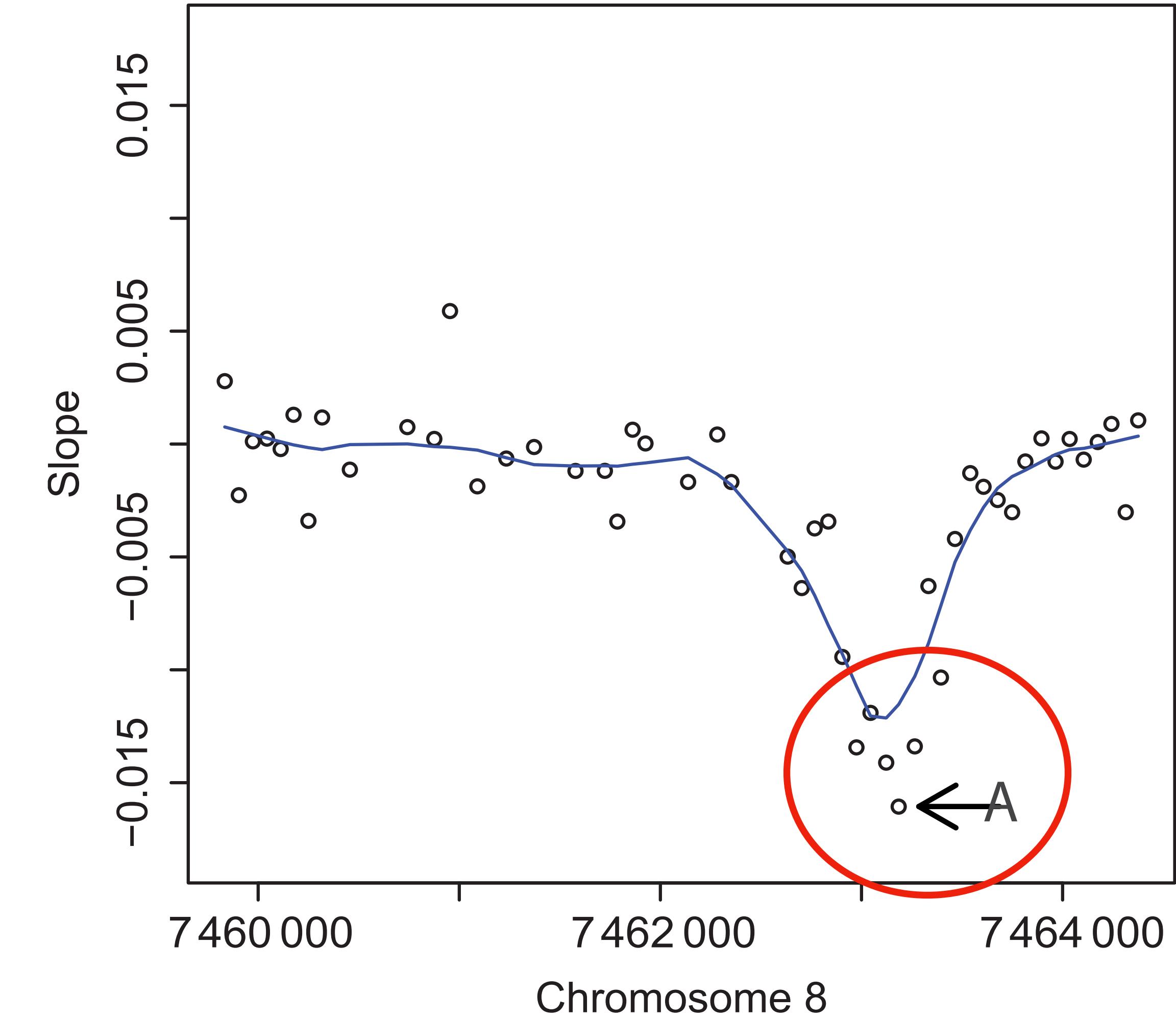
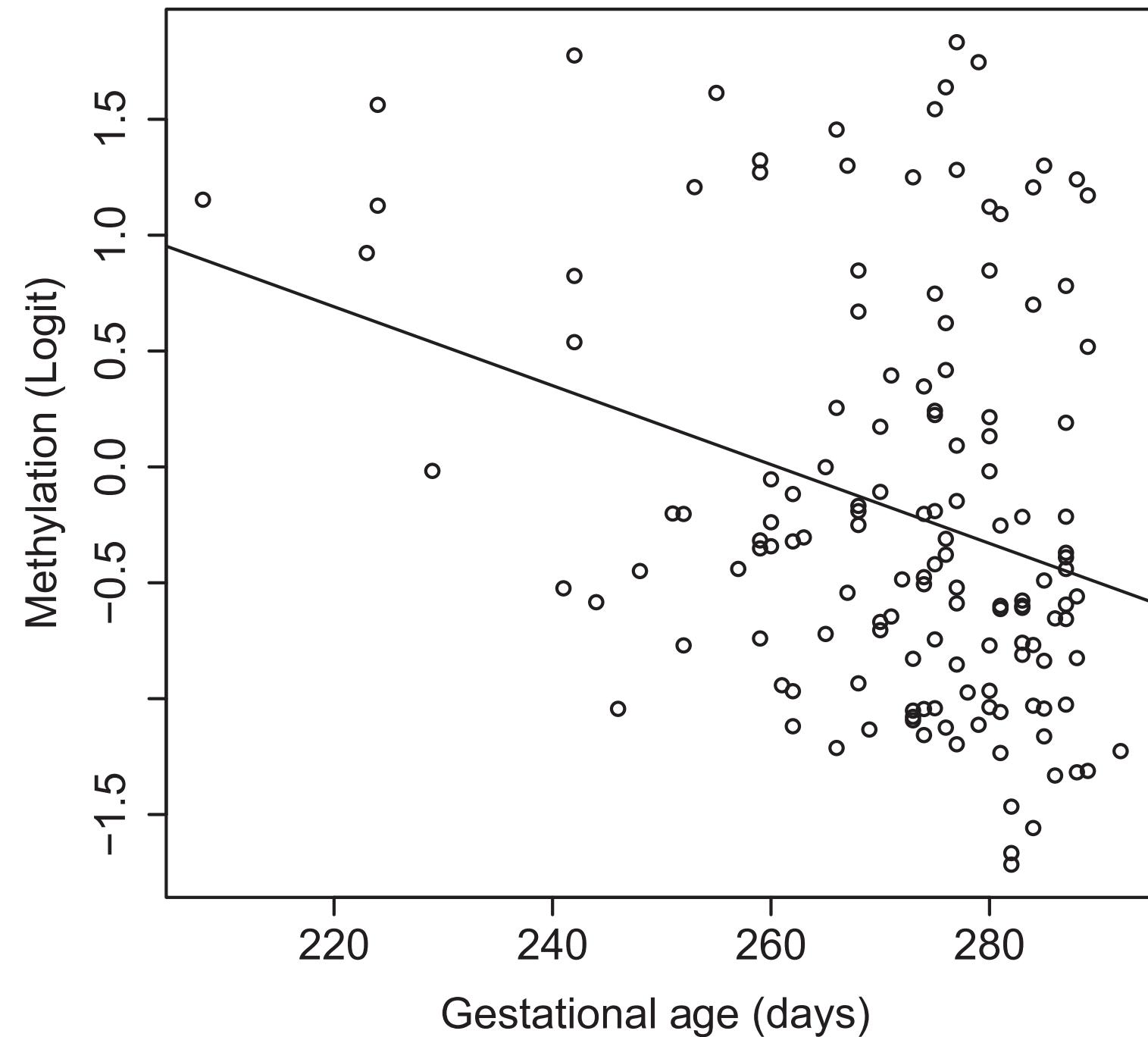


$$M_{ig} \sim \boxed{\beta_g} \times X_i + \dots$$

Jaffe et al. (2012)

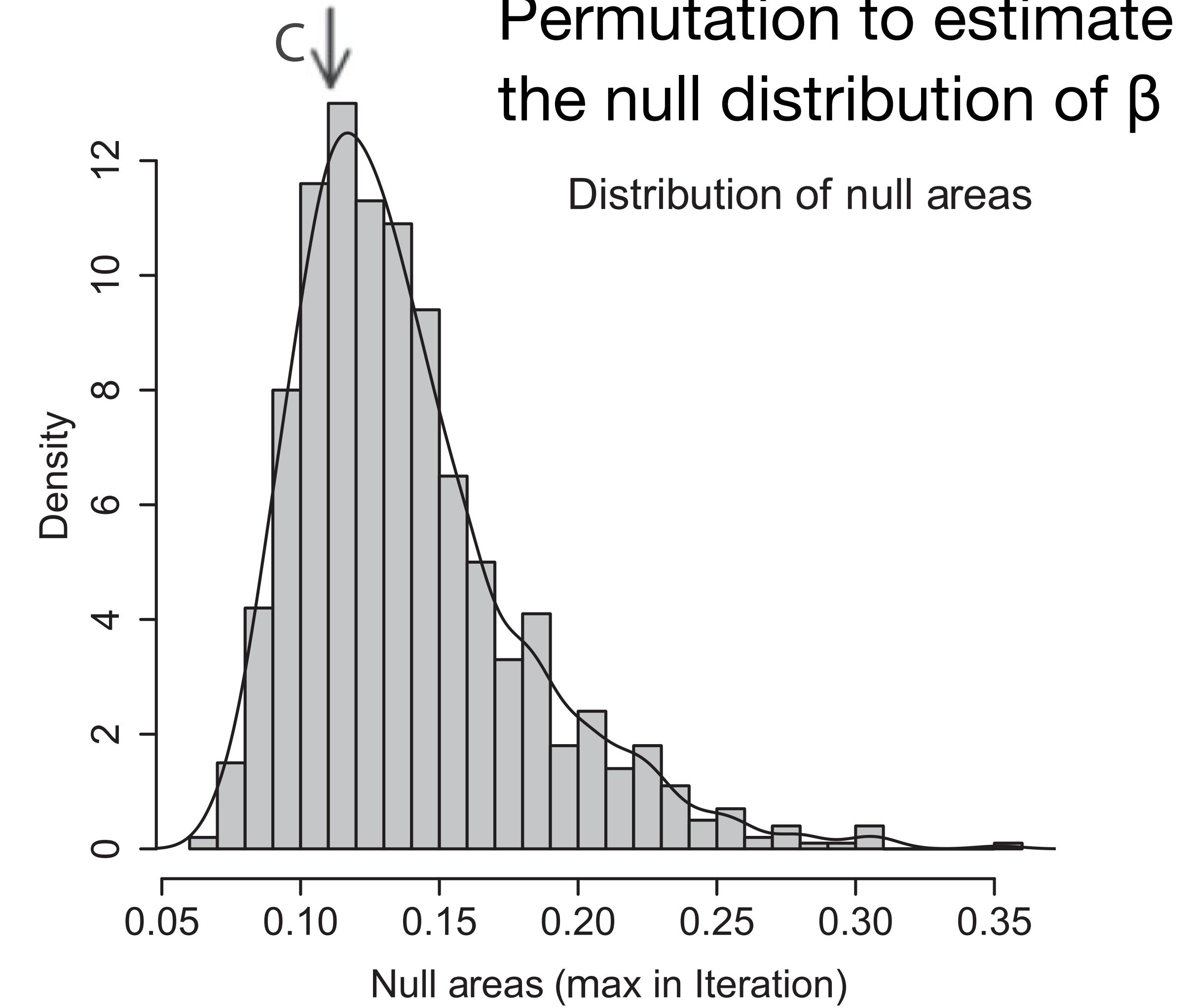
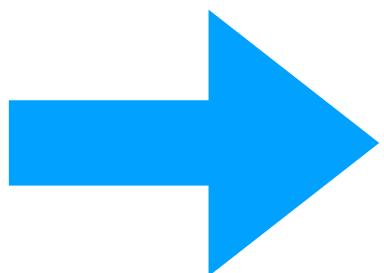
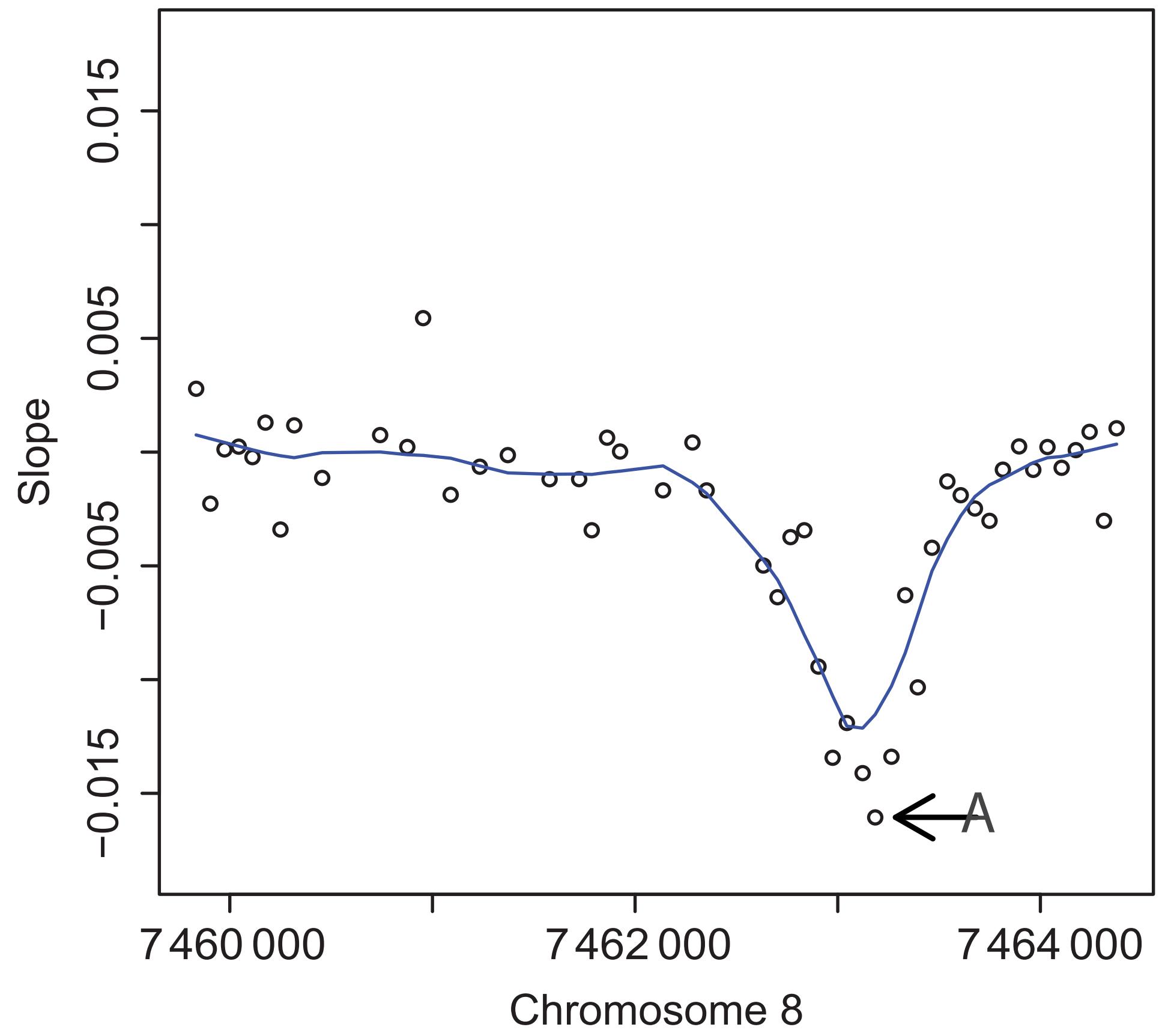
DMR bump hunter

Treat each CpG independently
Perform association, thus get β



Jaffe et al. (2012)

DMR bump hunter

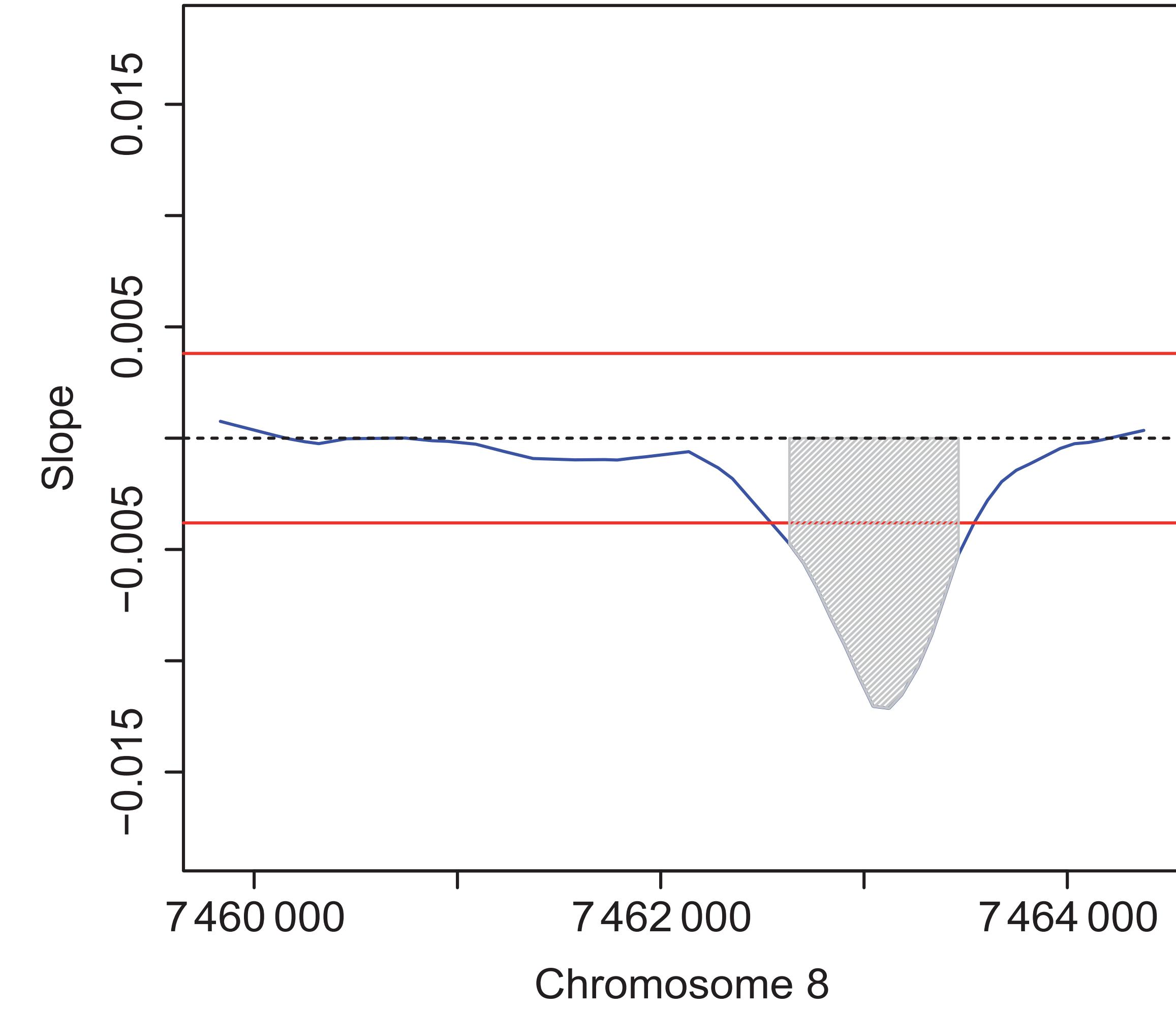
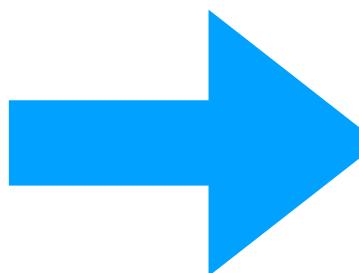
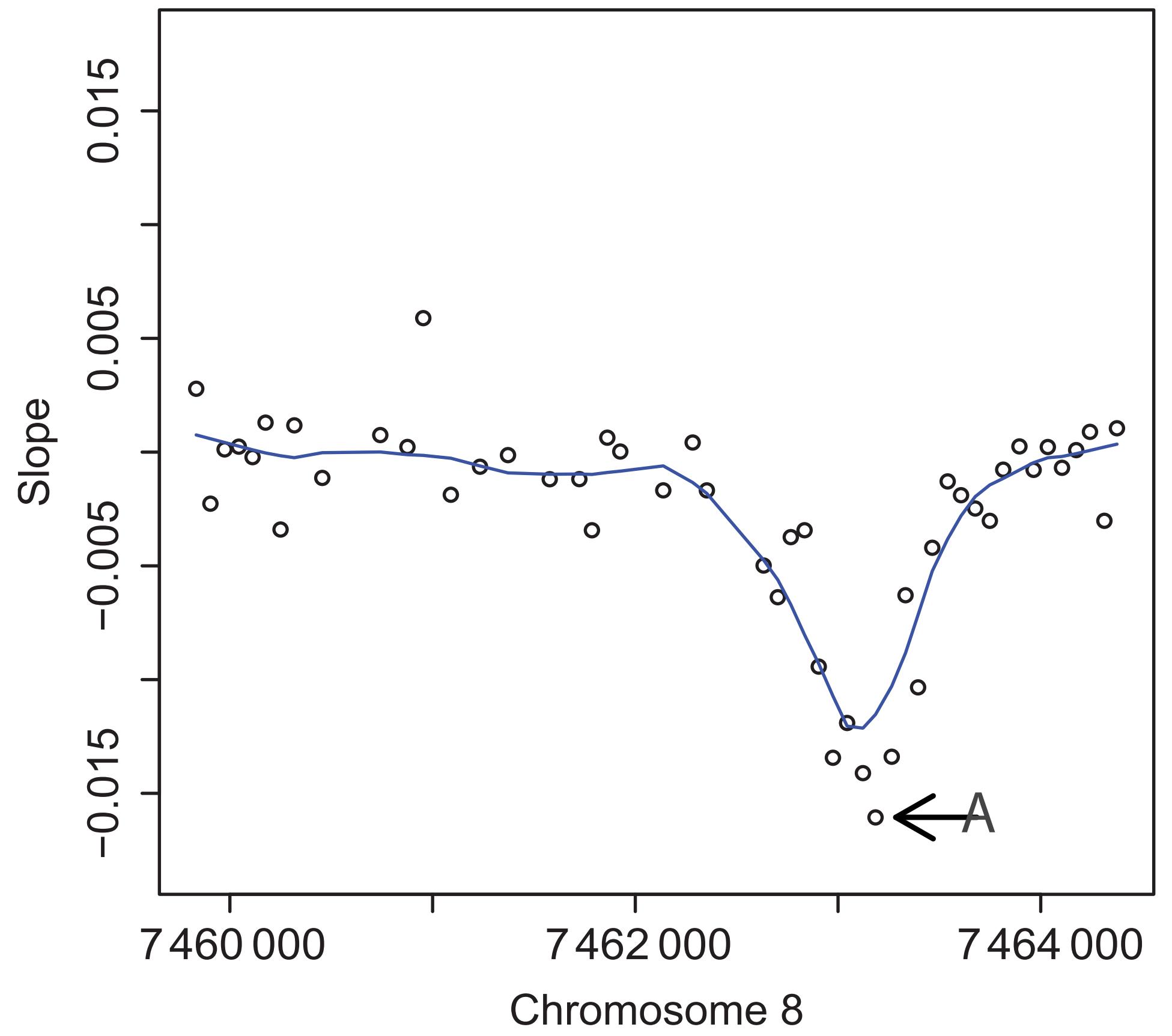


Permutation to estimate
the null distribution of β

Distribution of null areas

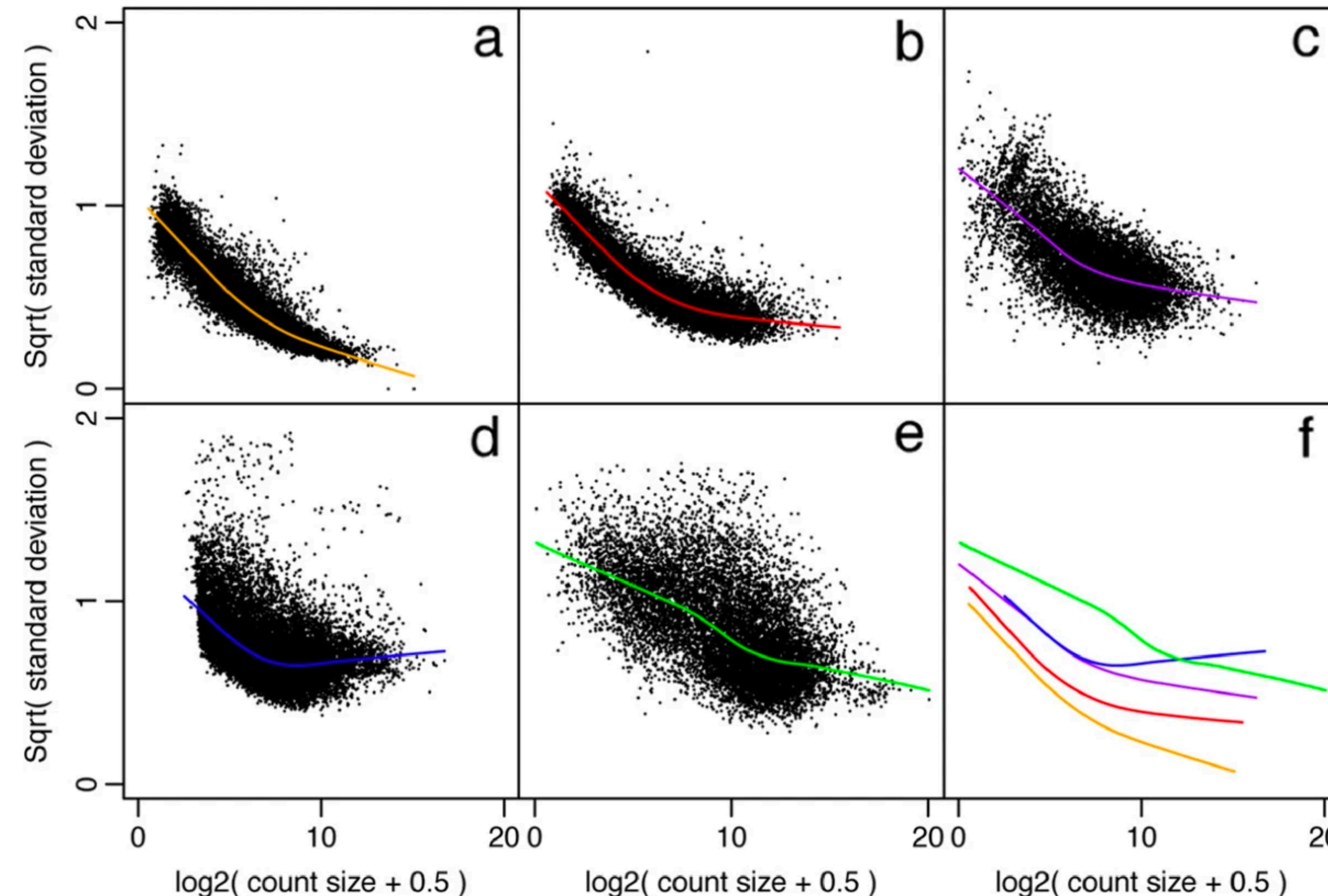
Jaffe et al. (2012)

DMR bump hunter

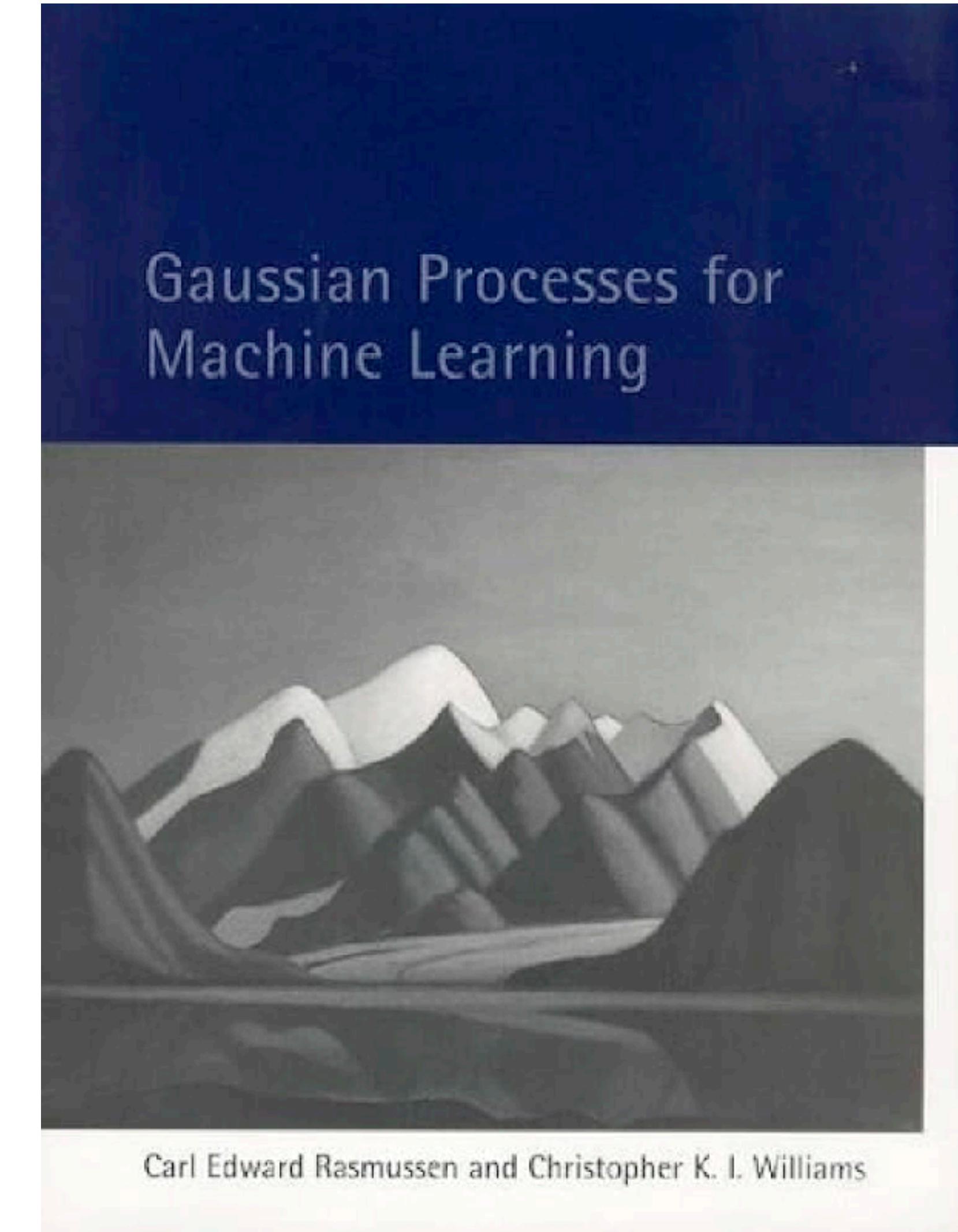
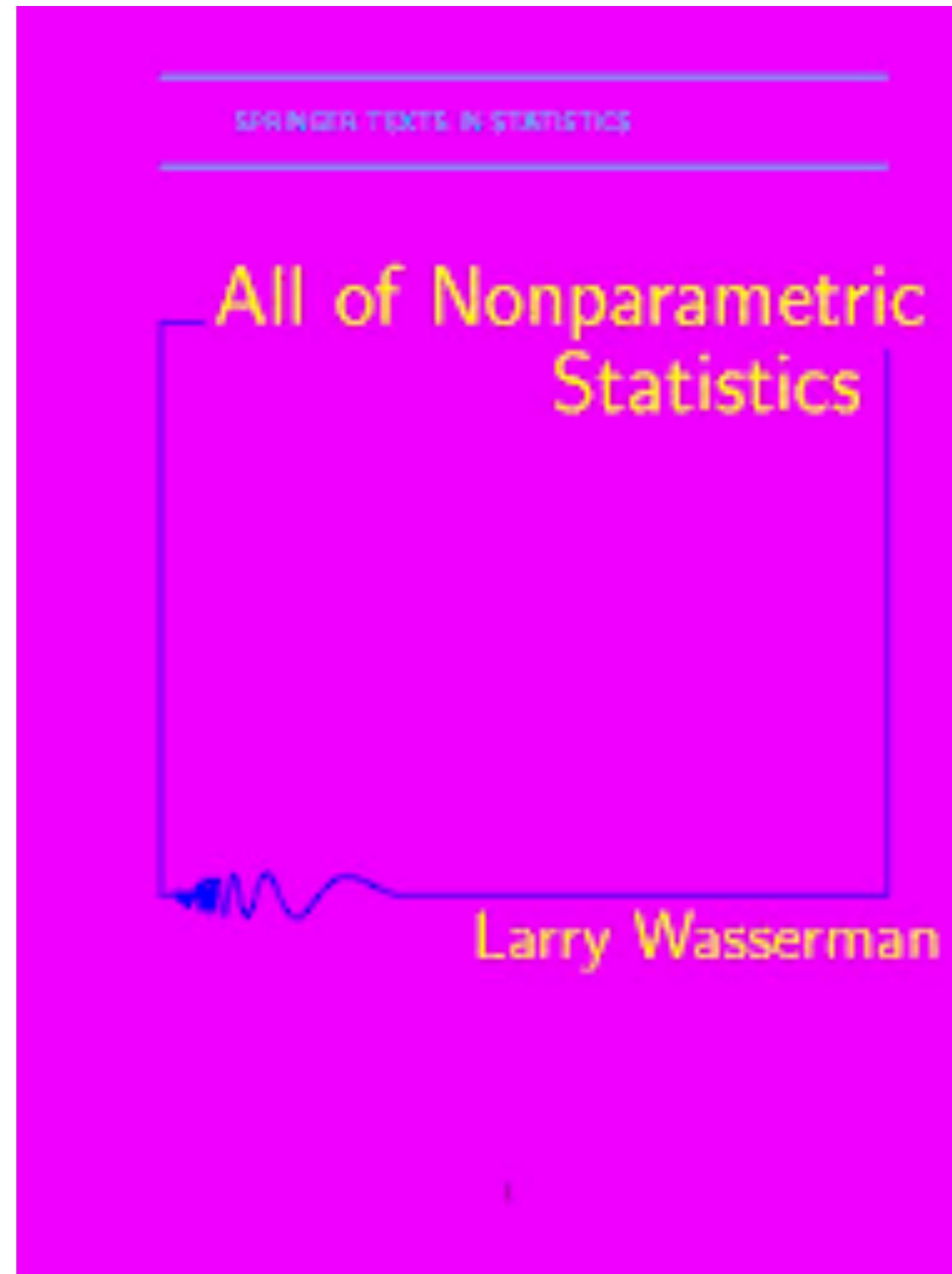


Jaffe et al. (2012)

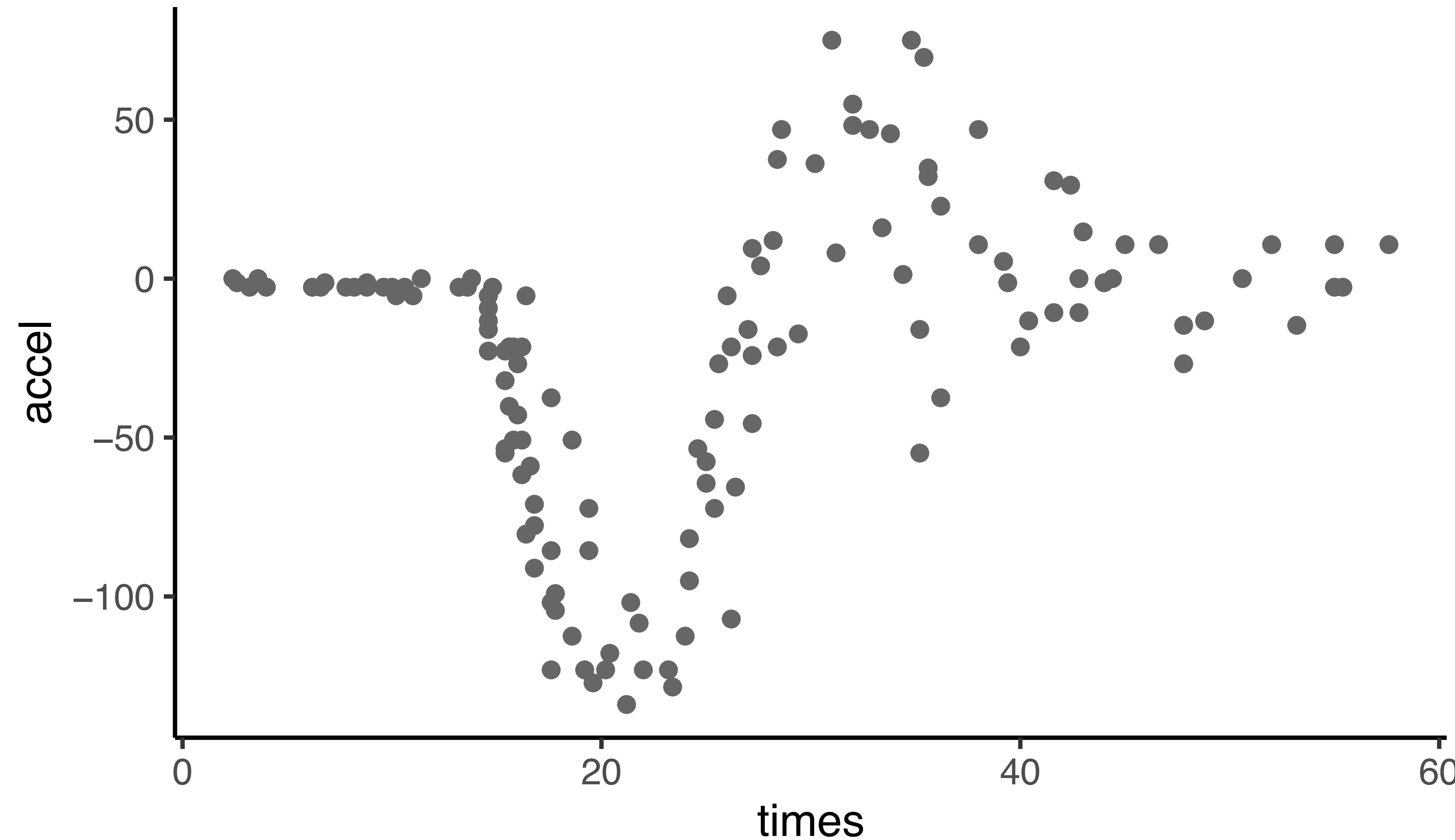
How can we model the wiggly patterns in 1D?



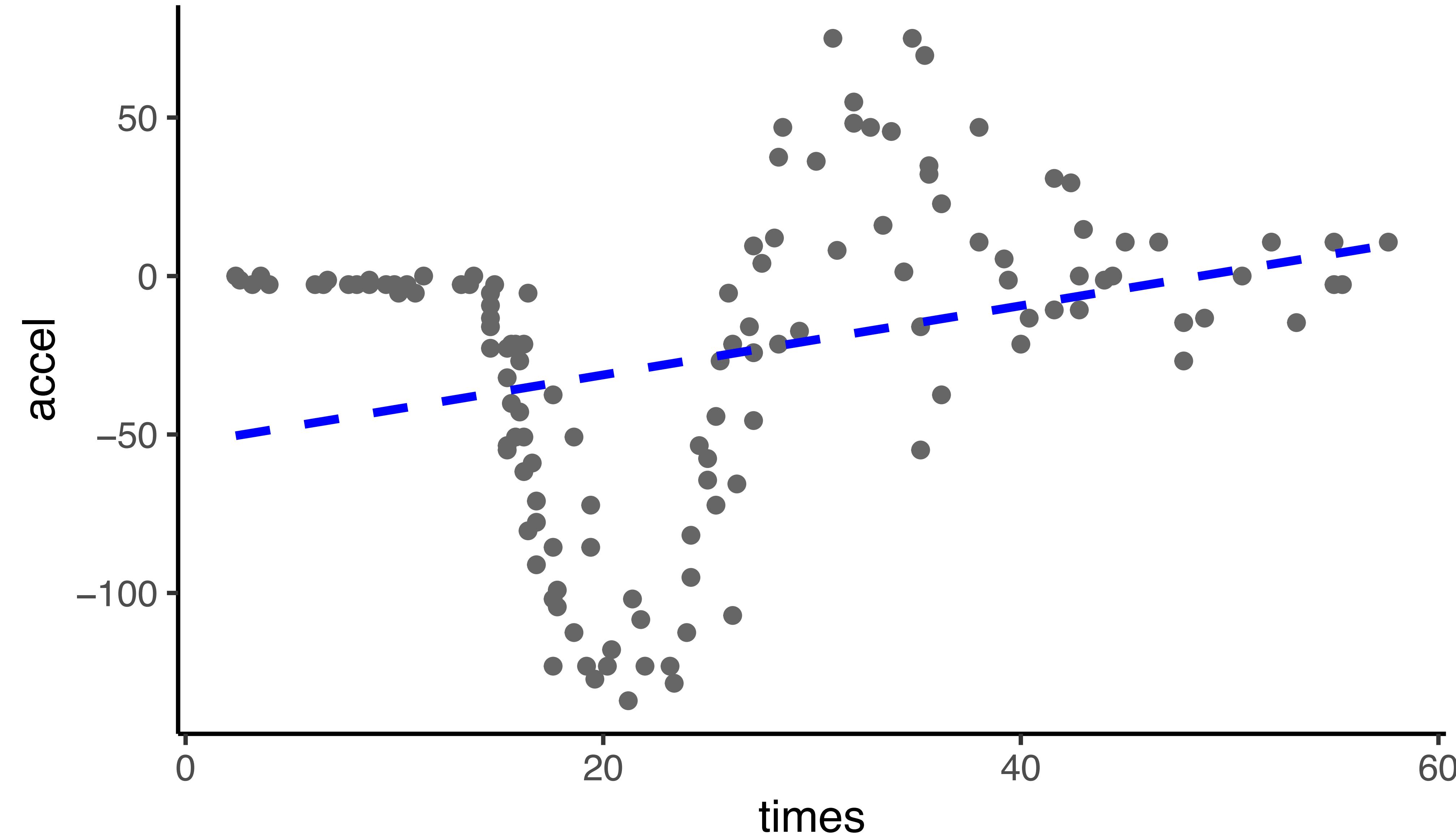
In case you want to know more:



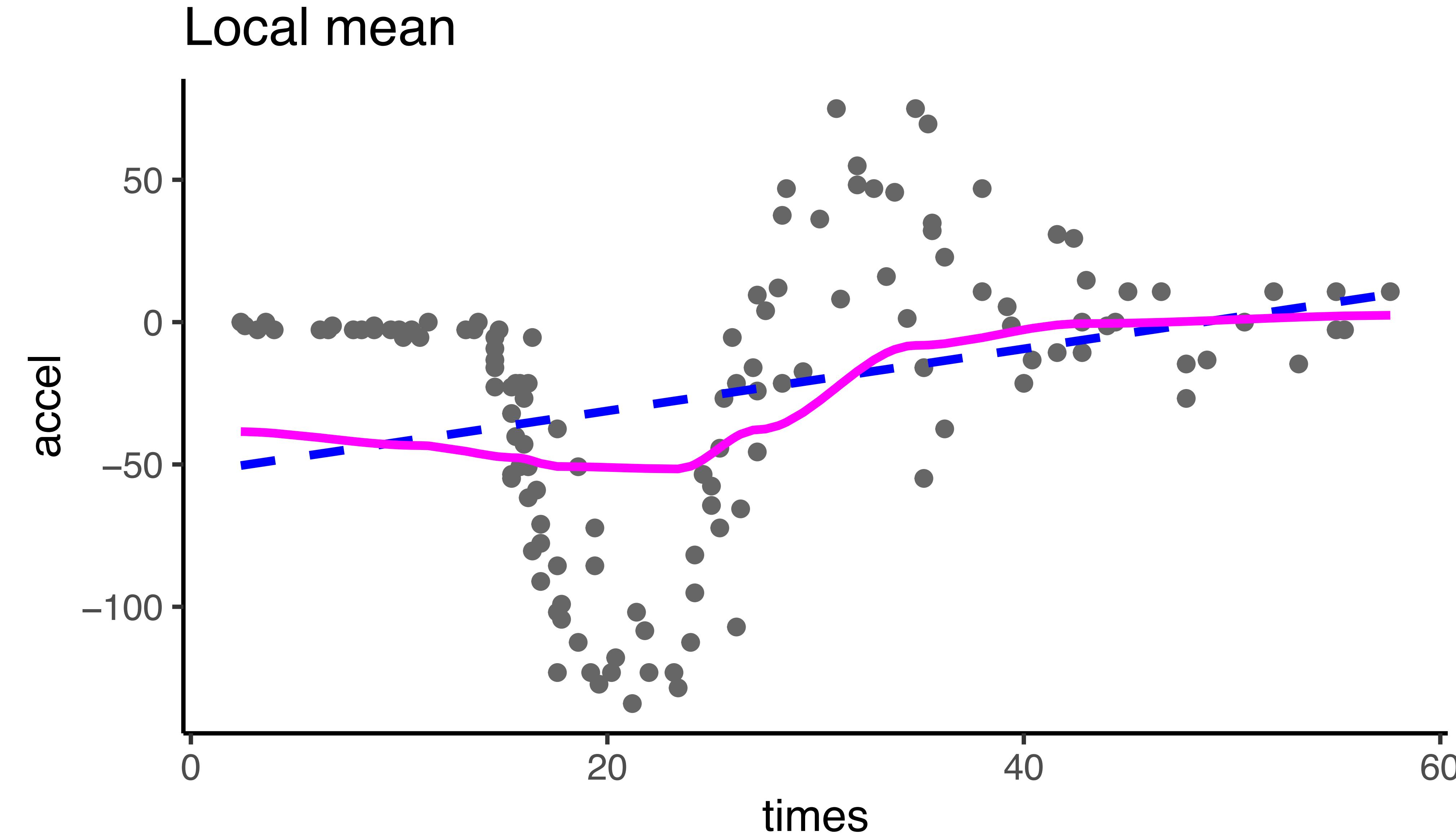
There are many things that a linear model cannot capture



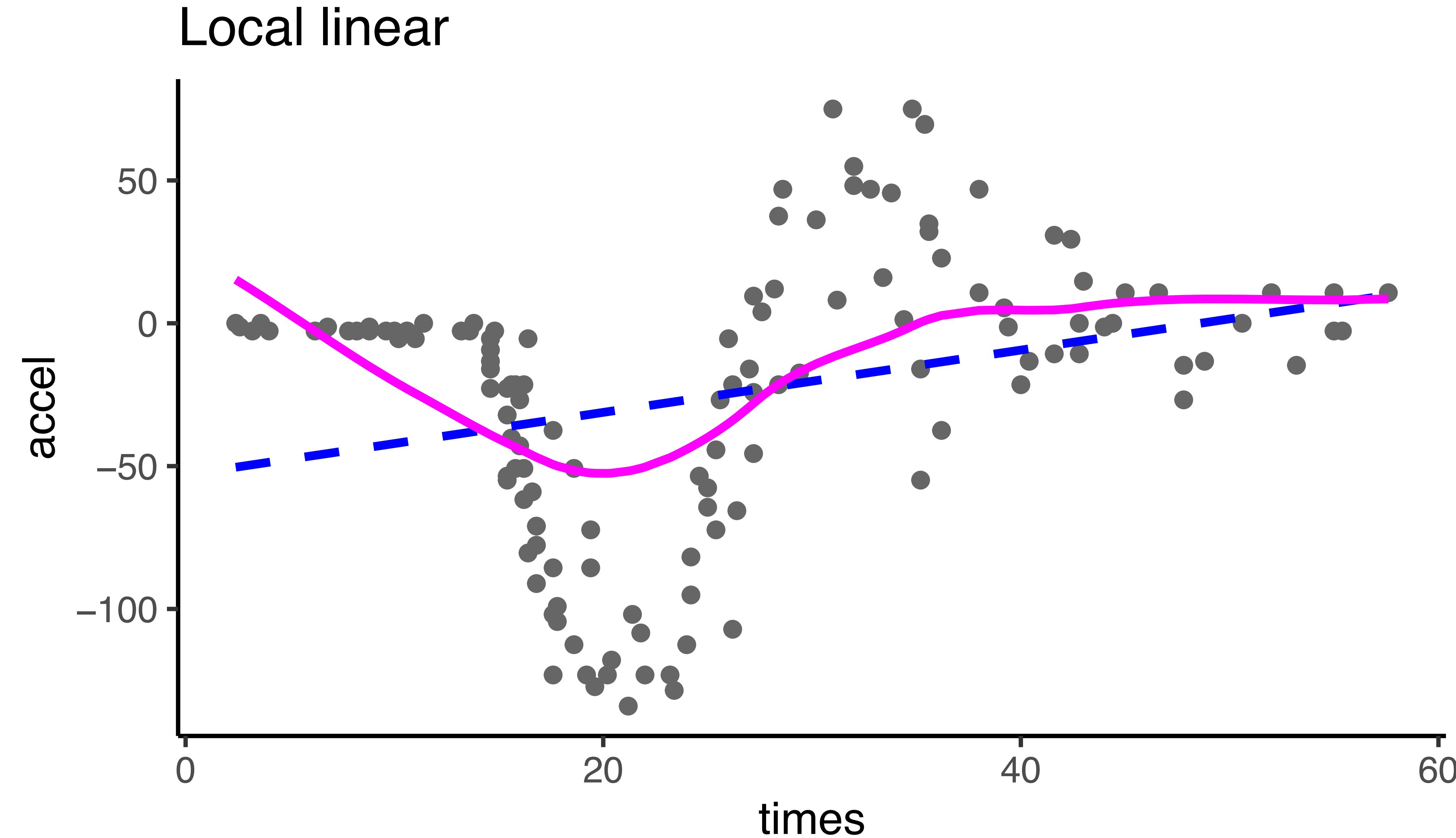
There are many things that a linear model cannot capture



Let's think locally ... local polynomial regression loess

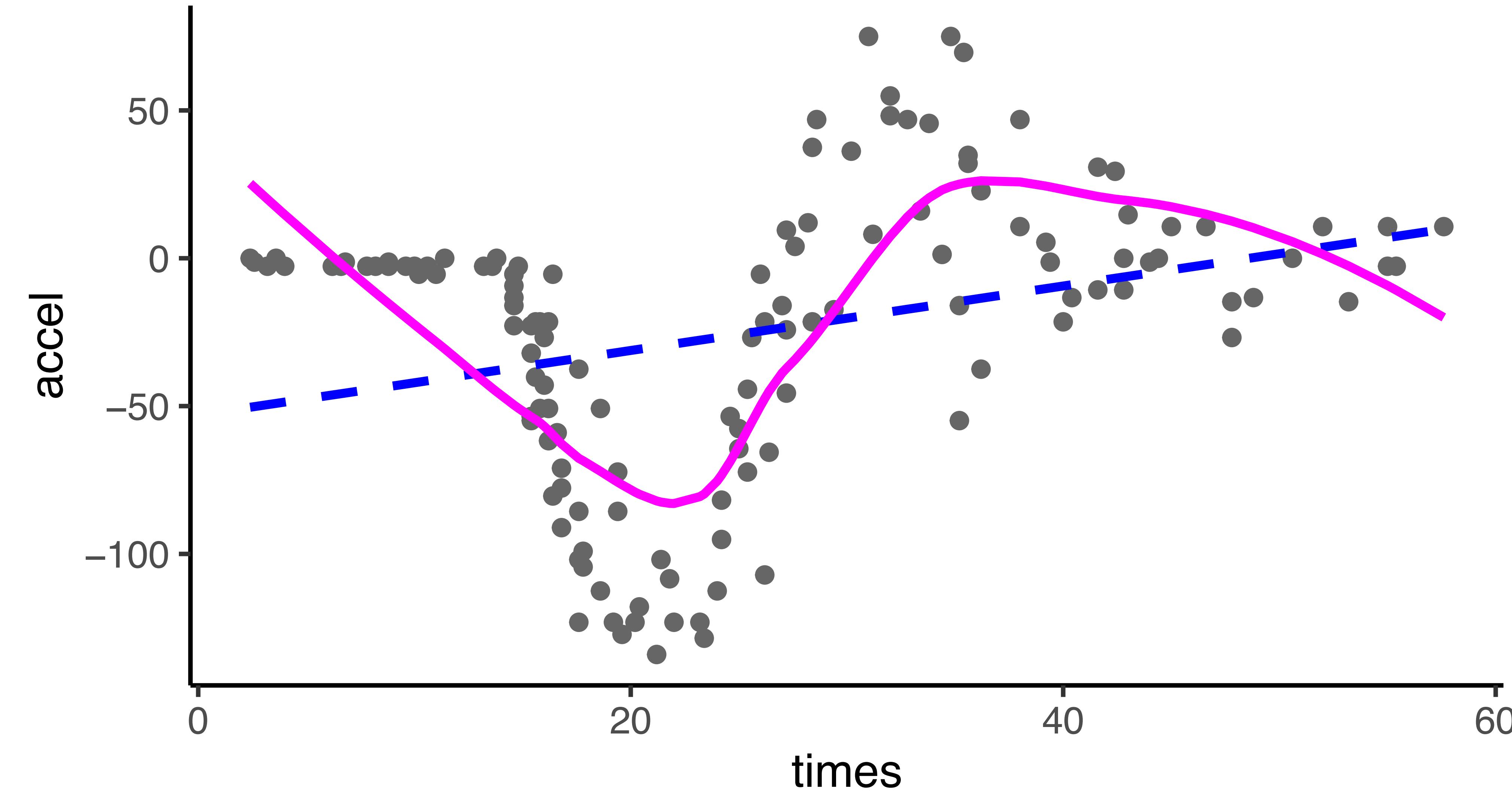


Let's think locally ... local polynomial regression loess



Let's think locally ... local polynomial regression loess

Local quadratic



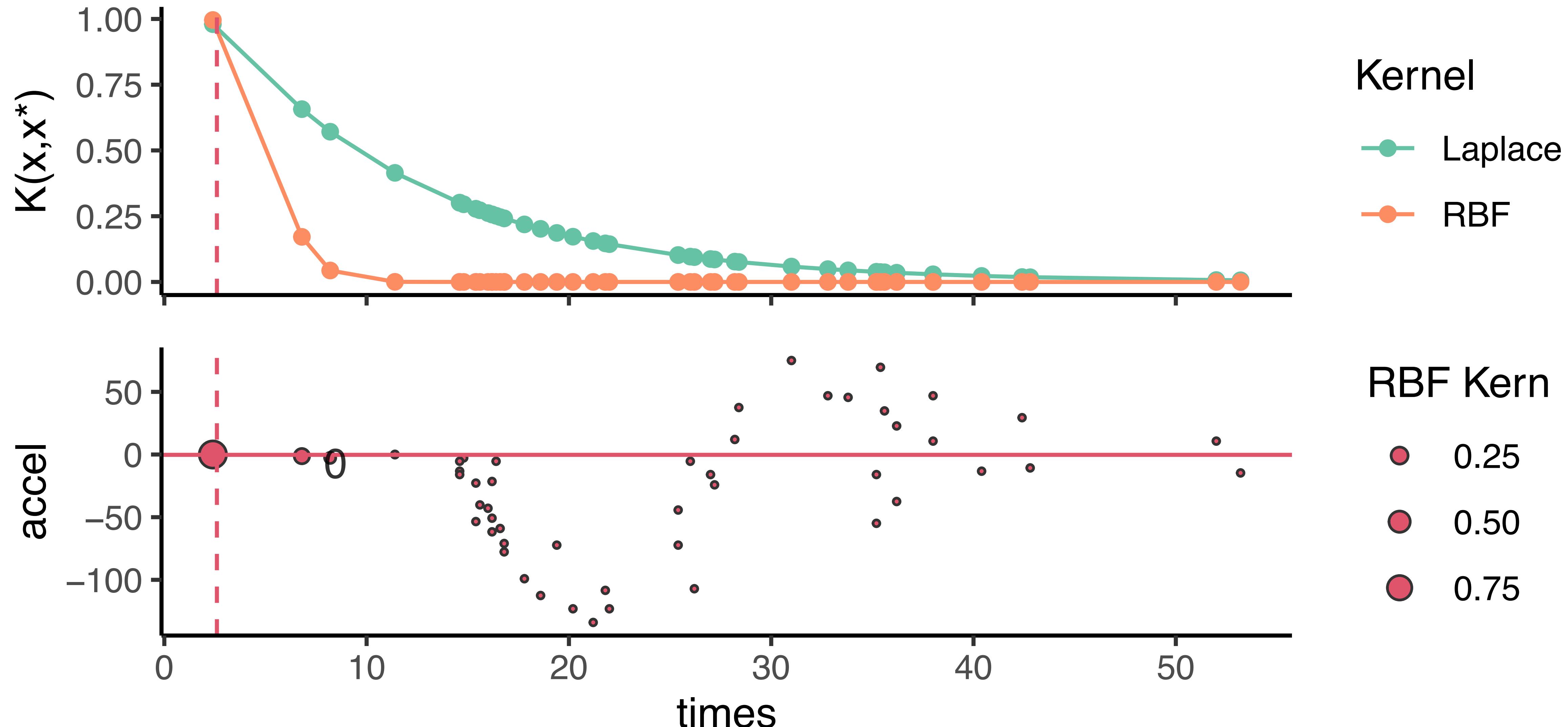
Nadaraya-Watson (kernelized local) estimator

```
nn <- nrow(mcycle)
train.idx <- sample(nn, 50)
x.test <-
  mcycle[-train.idx, 1, drop = FALSE] %>%
  as.matrix
y.test <-
  mcycle[-train.idx, 2, drop = FALSE] %>%
  as.matrix
x.train <-
  mcycle[train.idx, 1, drop = FALSE] %>%
  as.matrix
y.train <-
  mcycle[train.idx, 2, drop = FALSE] %>%
  as.matrix
```

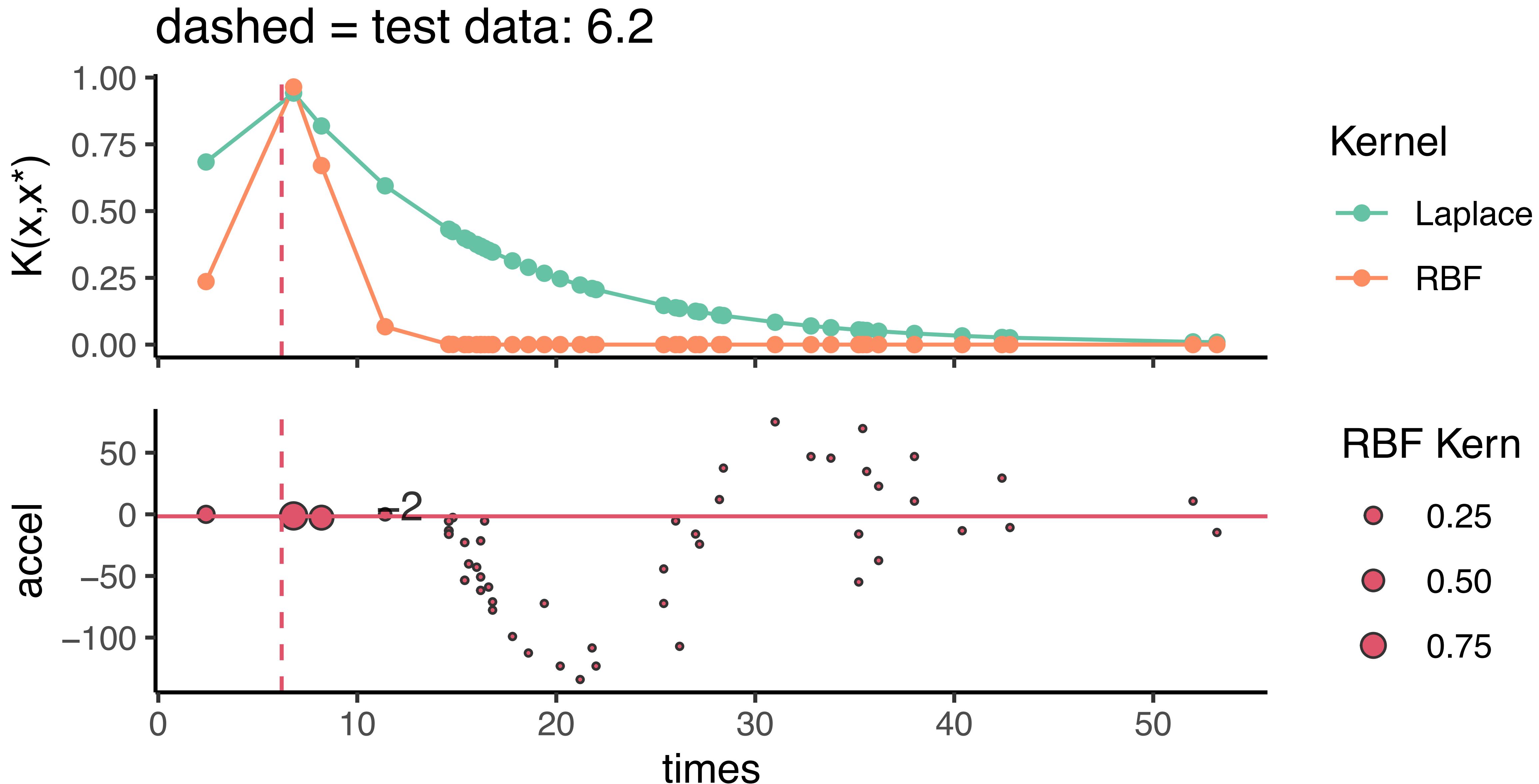
```
#' @param x.test testing data points
#' @param X training X
#' @param Y training Y
#' @param h bandwidth h
nw.rbf <- function(x.test, X, Y, h) {
  rbf <- kernlab::rbfdot(sigma=h)
  K <- kernlab::kernelMatrix(rbf, x.test, X)
  W <- K / rowSums(K)
  W %*% as.matrix(Y)
}
```

Basic idea: borrow neighbours' Y values in training data

dashed = test data: 2.6

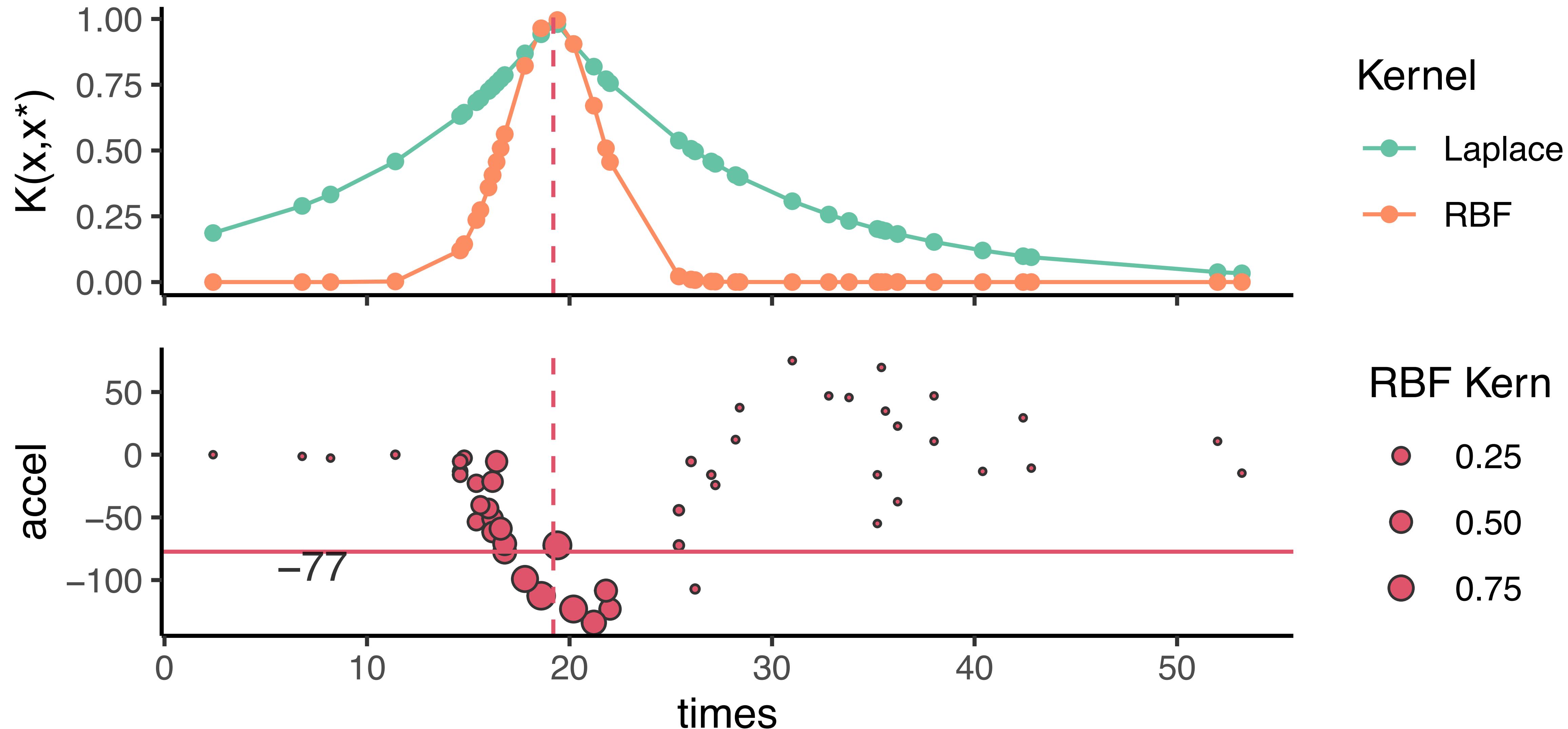


Basic idea: borrow neighbours' Y values in training data



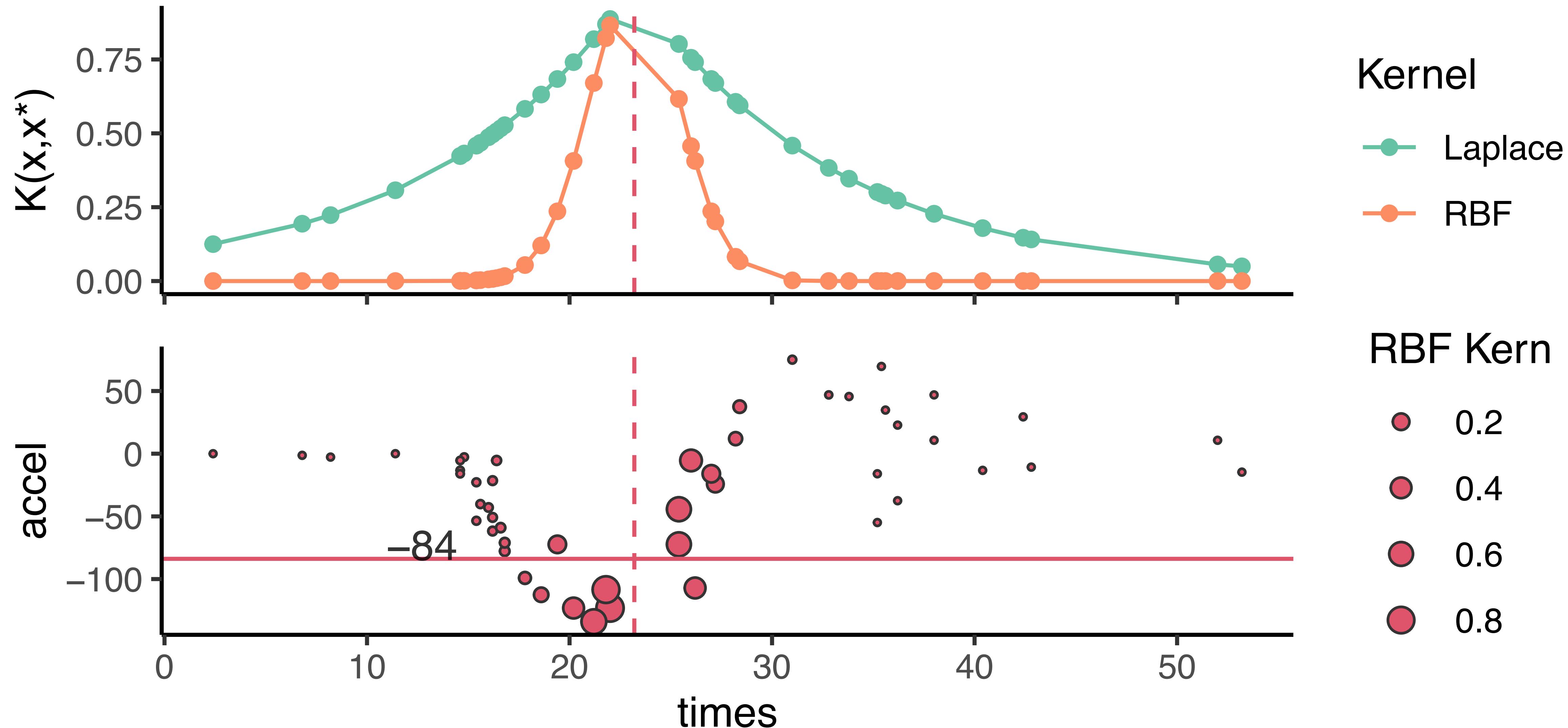
Basic idea: borrow neighbours' Y values in training data

dashed = test data: 19.2



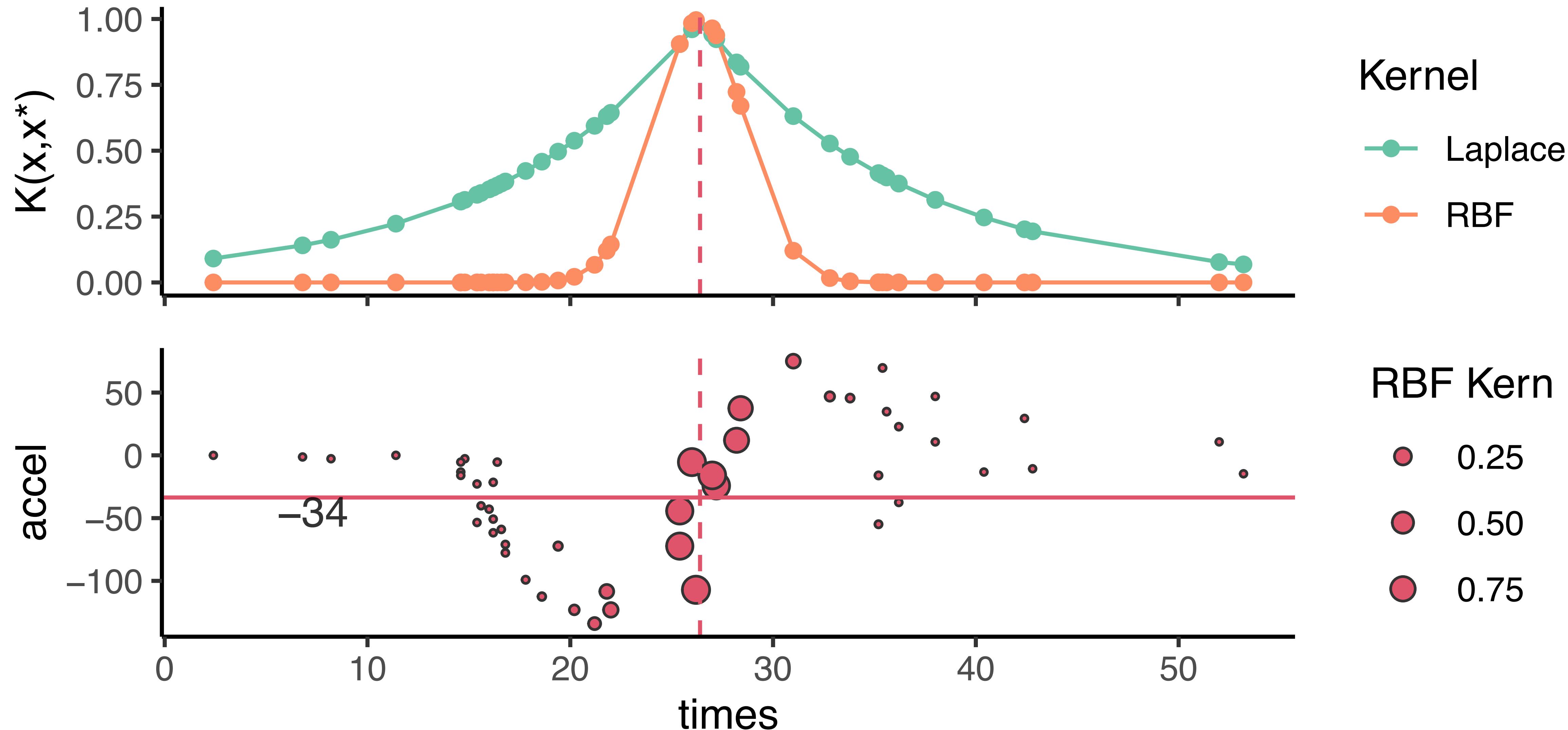
Basic idea: borrow neighbours' Y values in training data

dashed = test data: 23.2



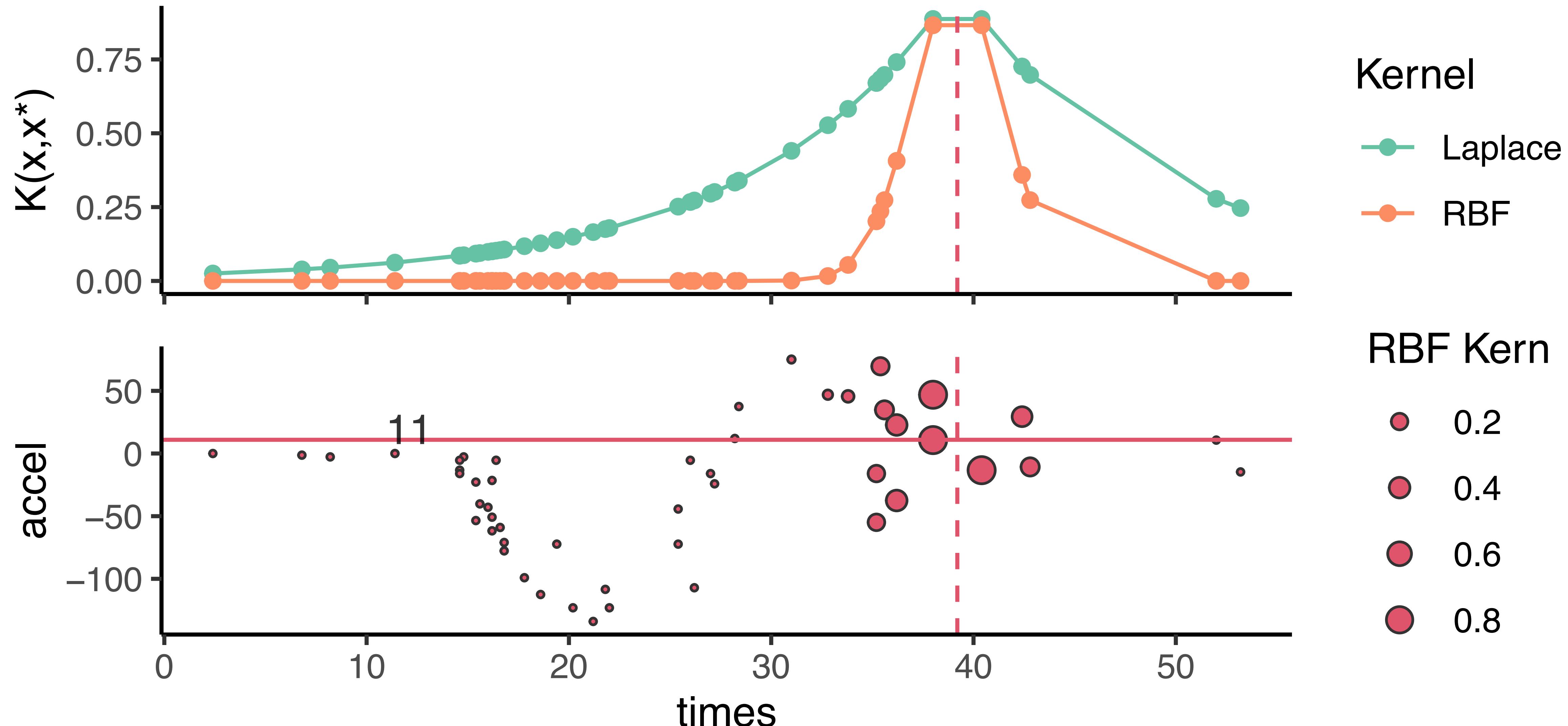
Basic idea: borrow neighbours' Y values in training data

dashed = test data: 26.4

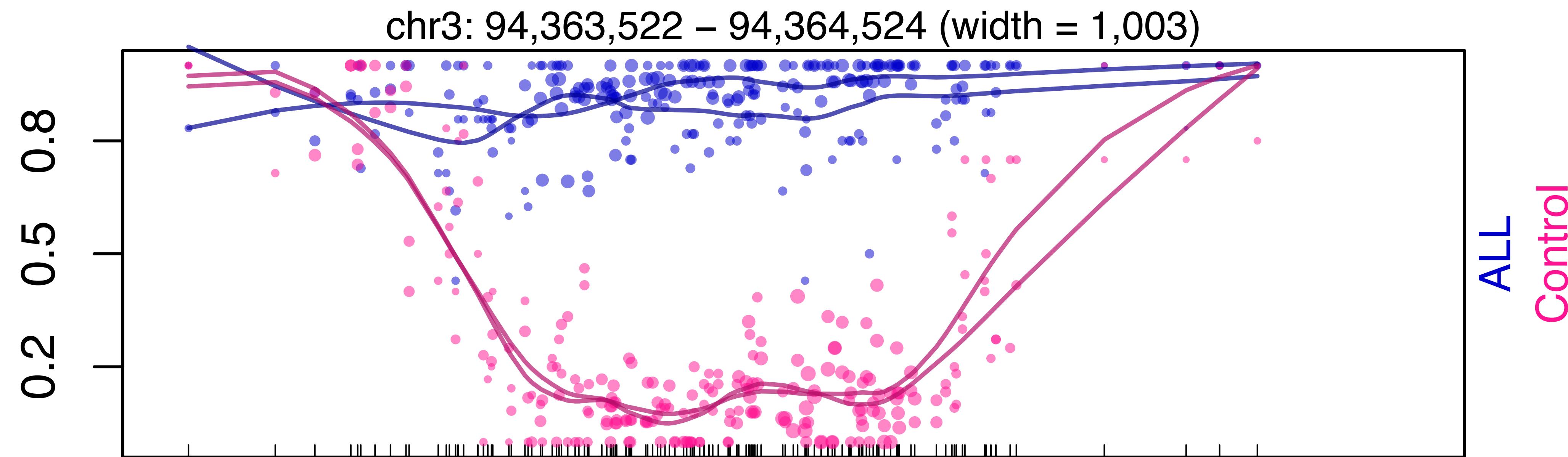


Basic idea: borrow neighbours' Y values in training data

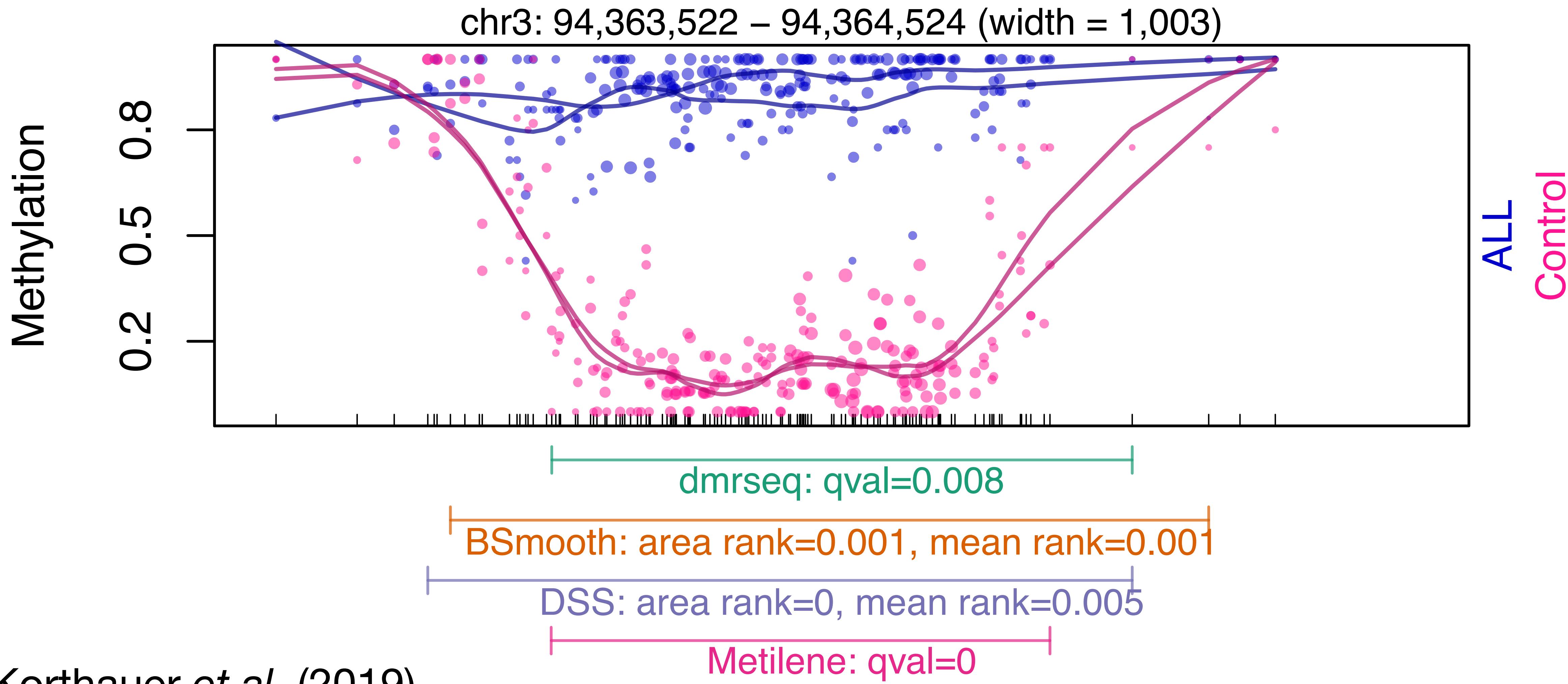
dashed = test data: 39.2



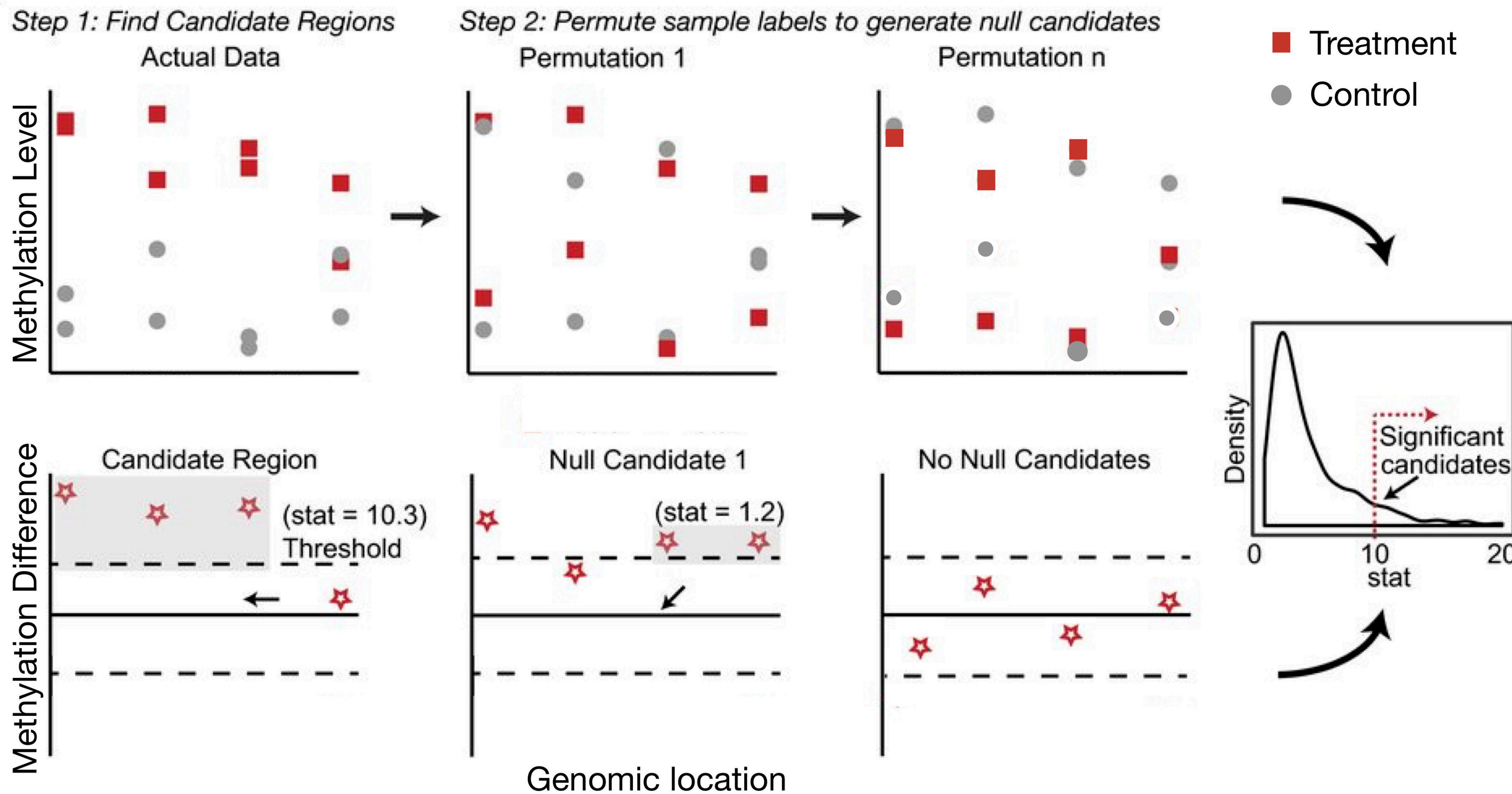
Calibrating DMR considering FDR



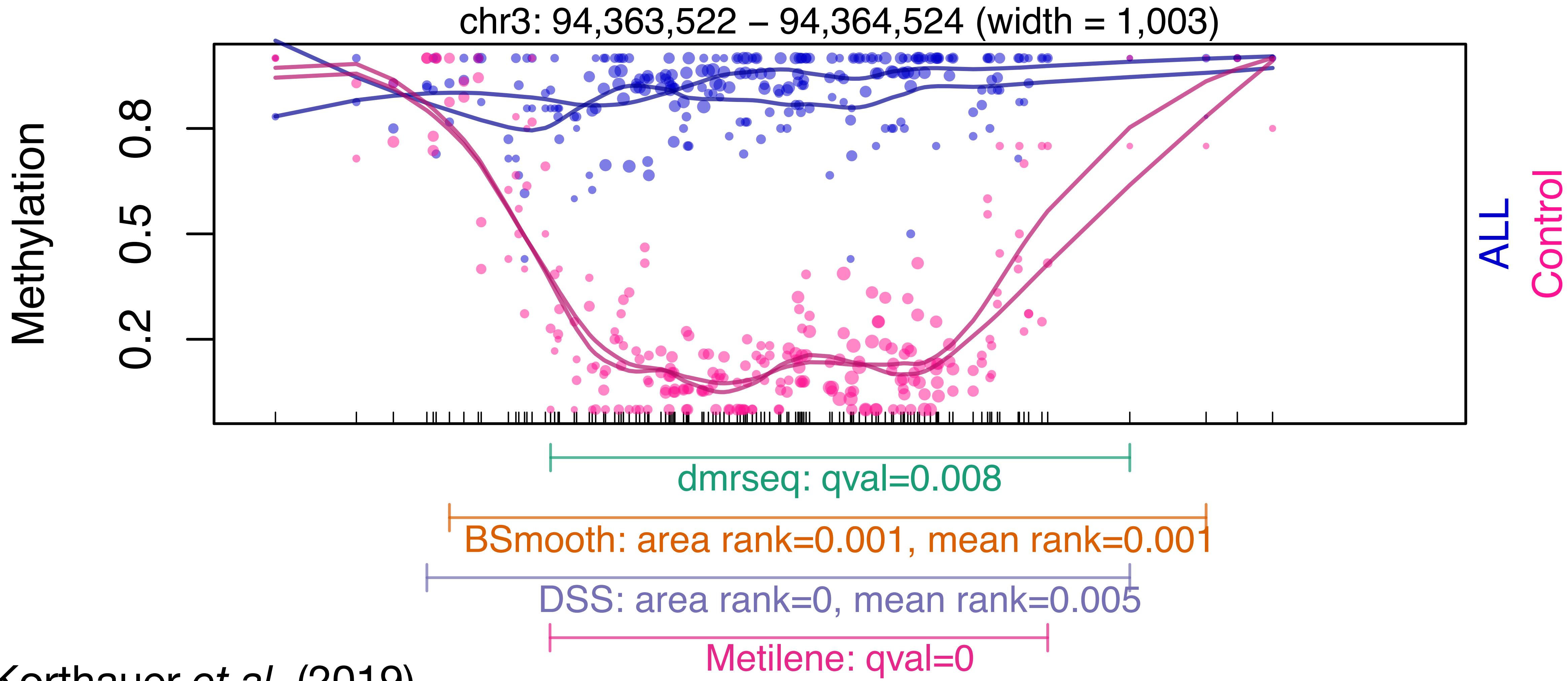
dmrseq improves statistical power with tight FDR clibration



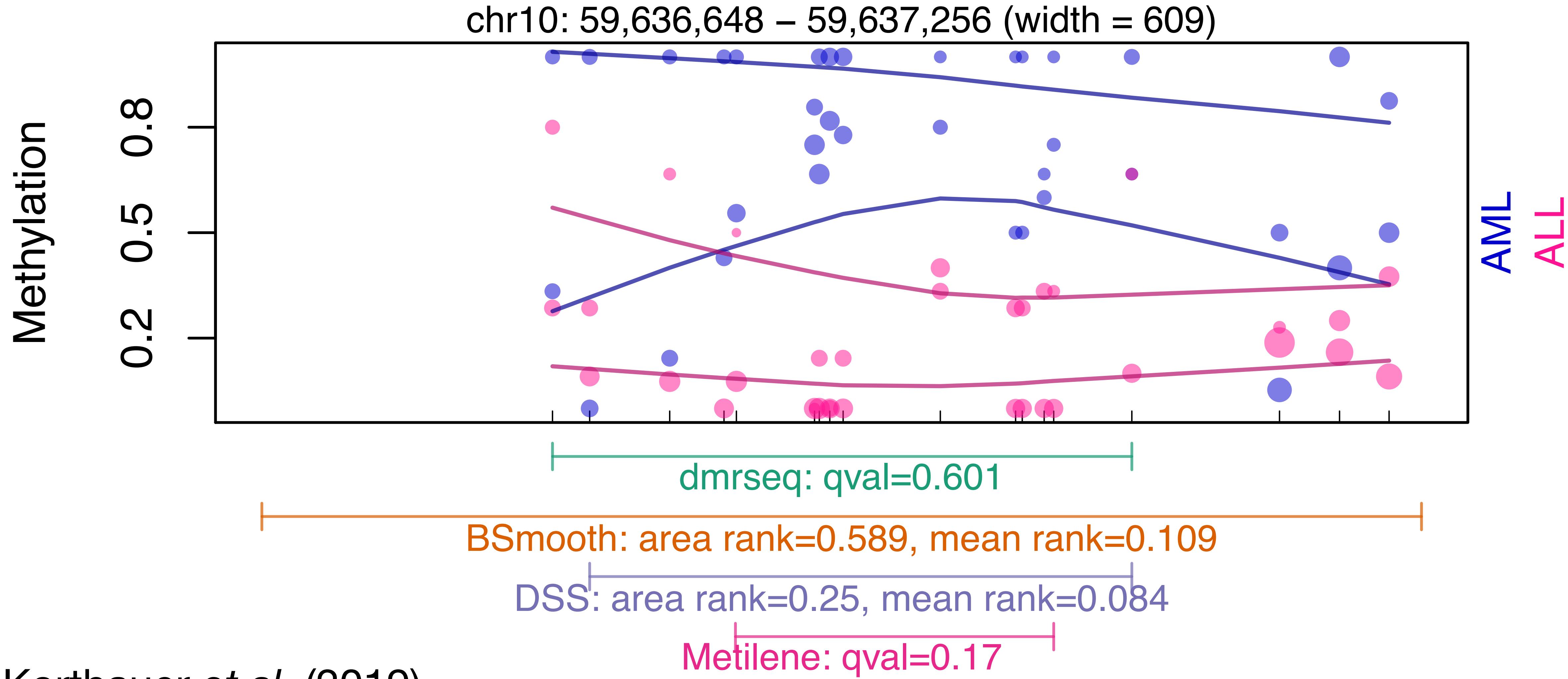
How can we tighten up FDR correction?



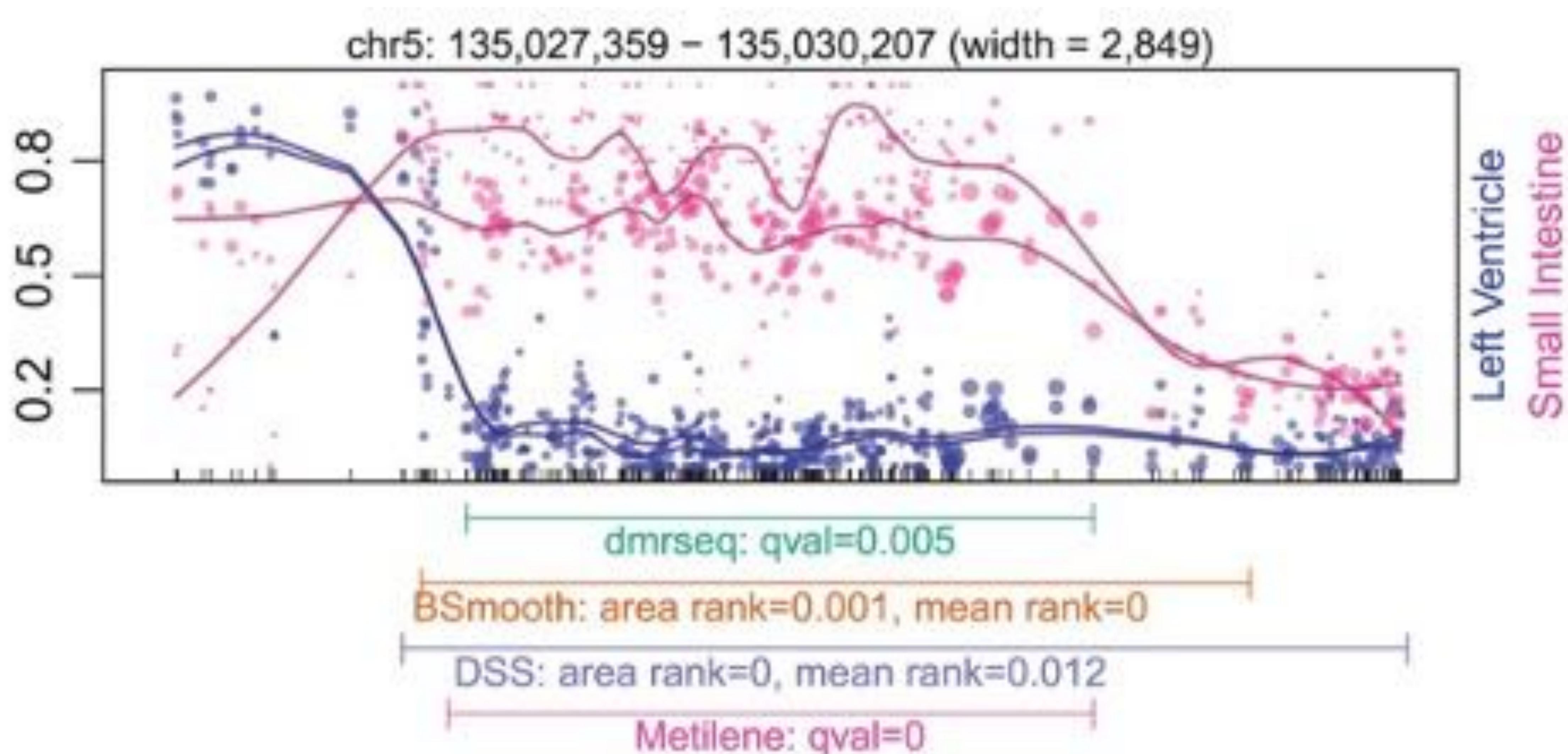
Do we see a DMR?



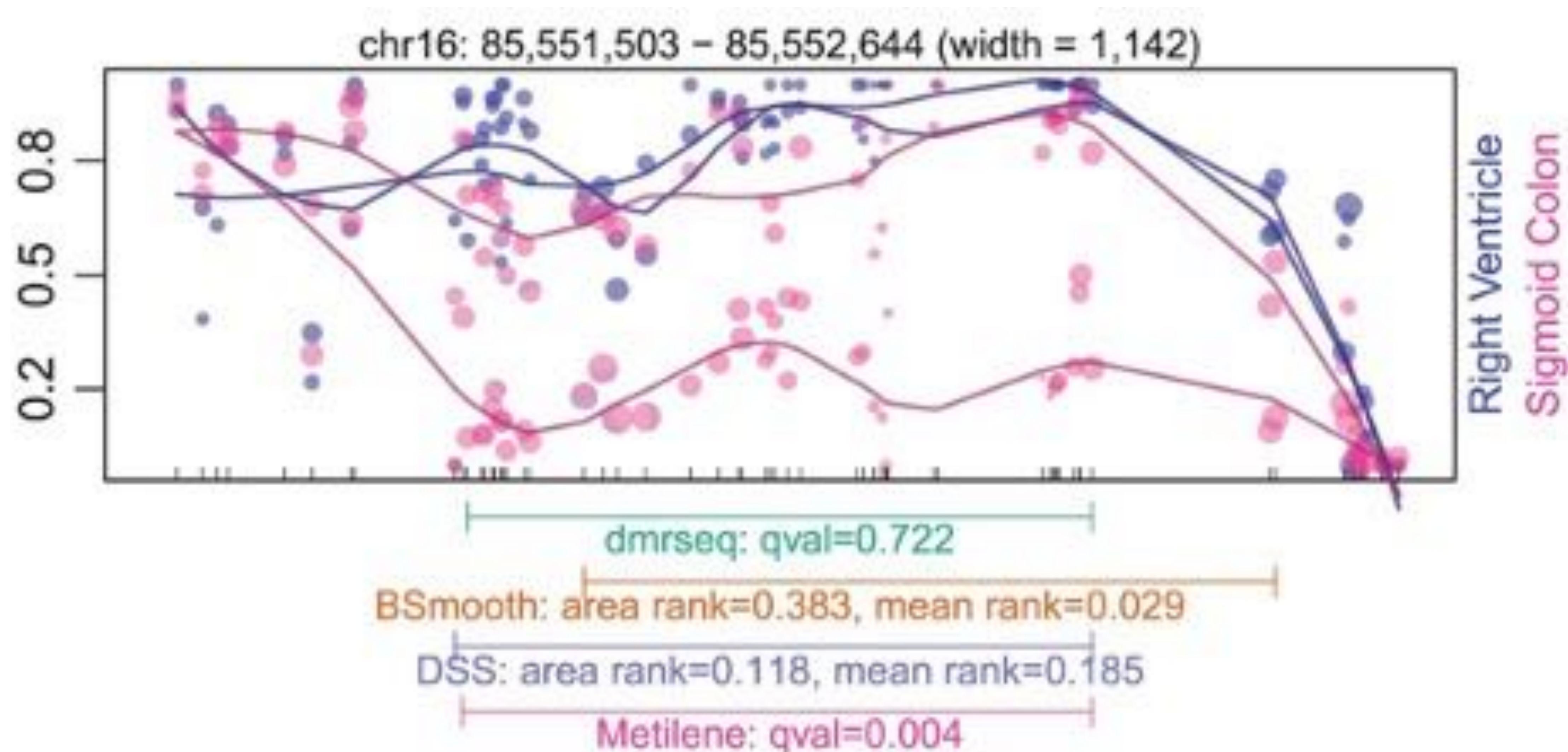
Do we see a DMR?



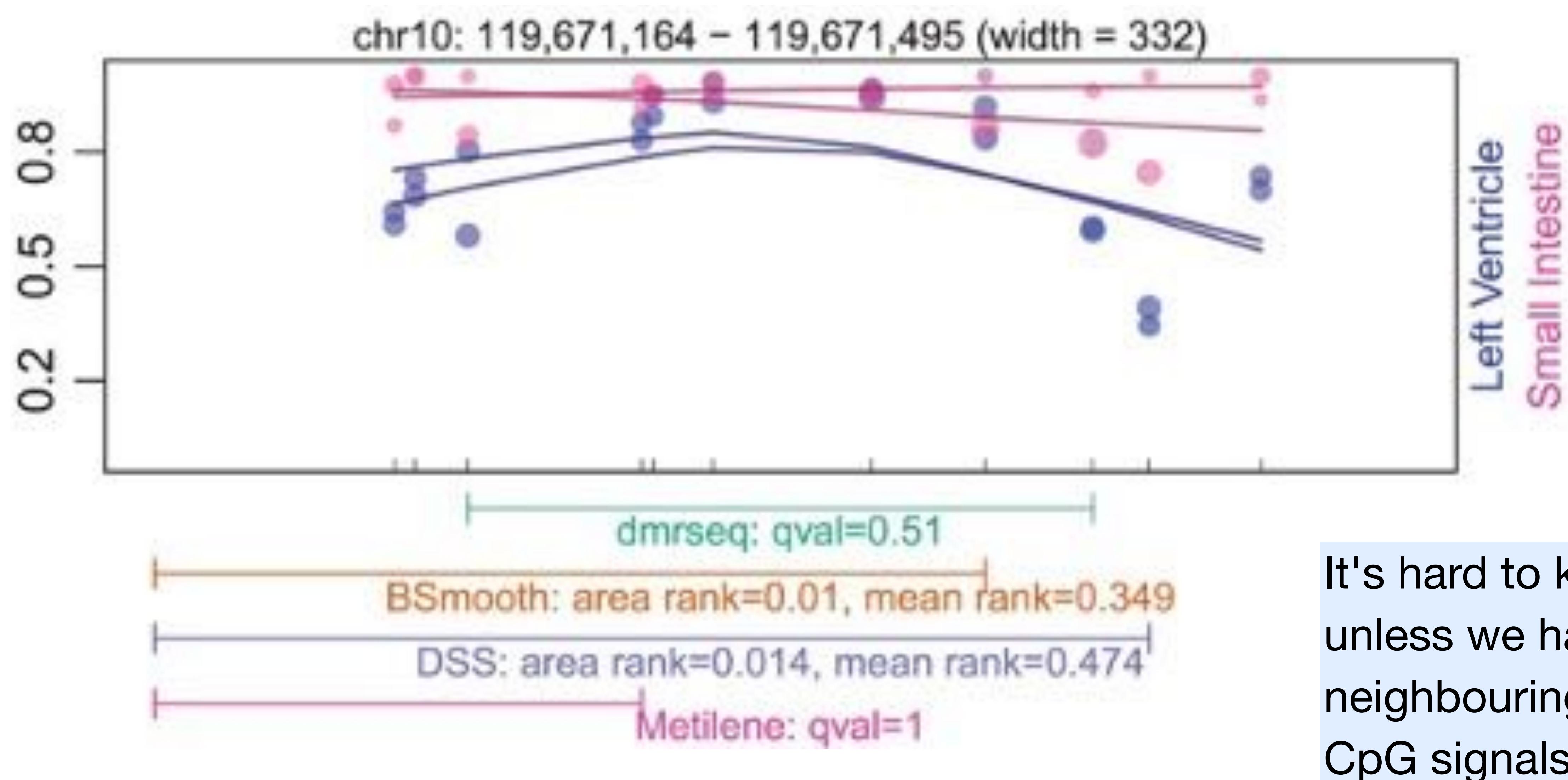
An obvious case where most methods work fine



Some start breaking down in a harder case



Heuristic approaches report wrong



Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

Epigenetic modifications

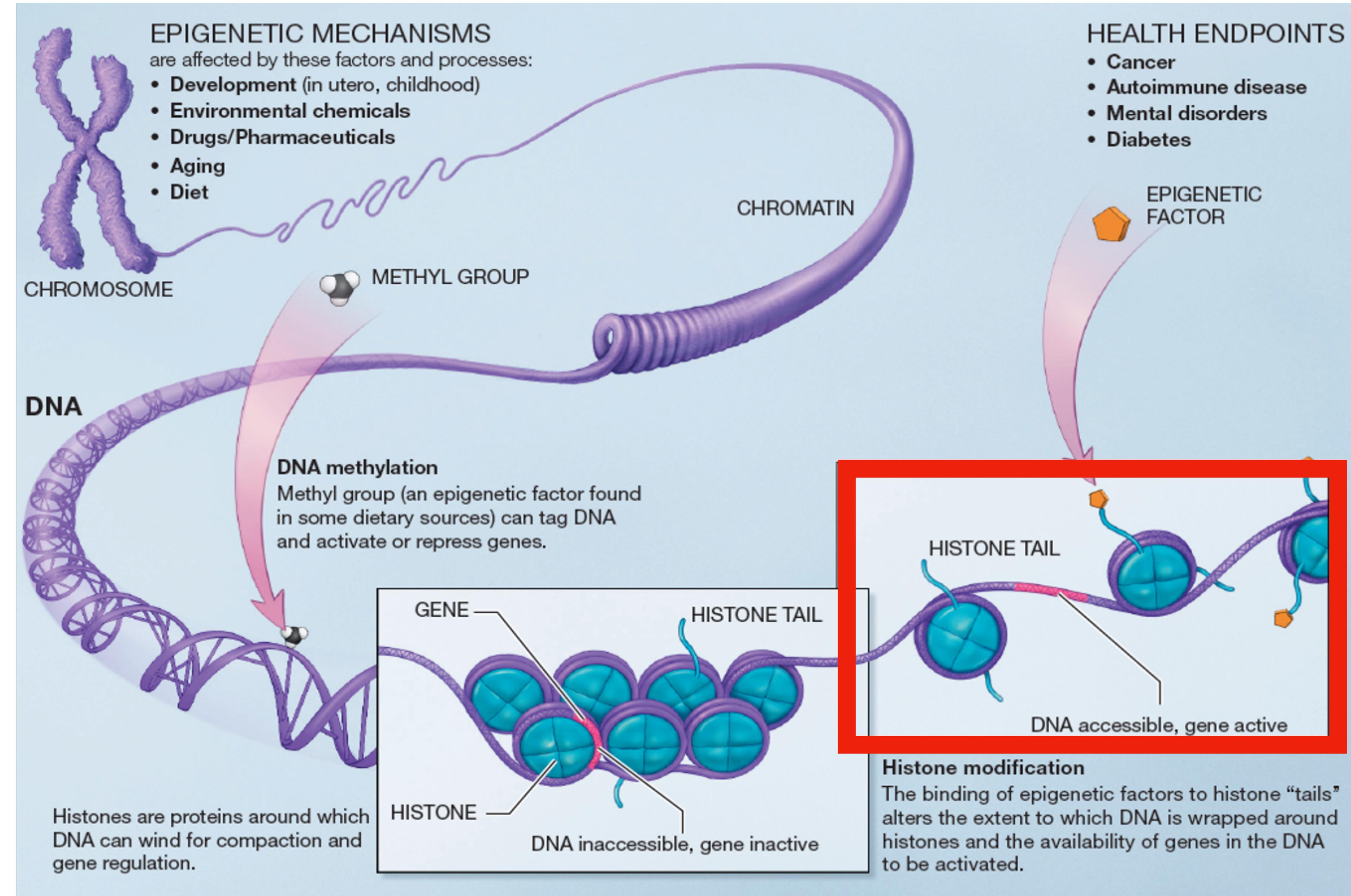
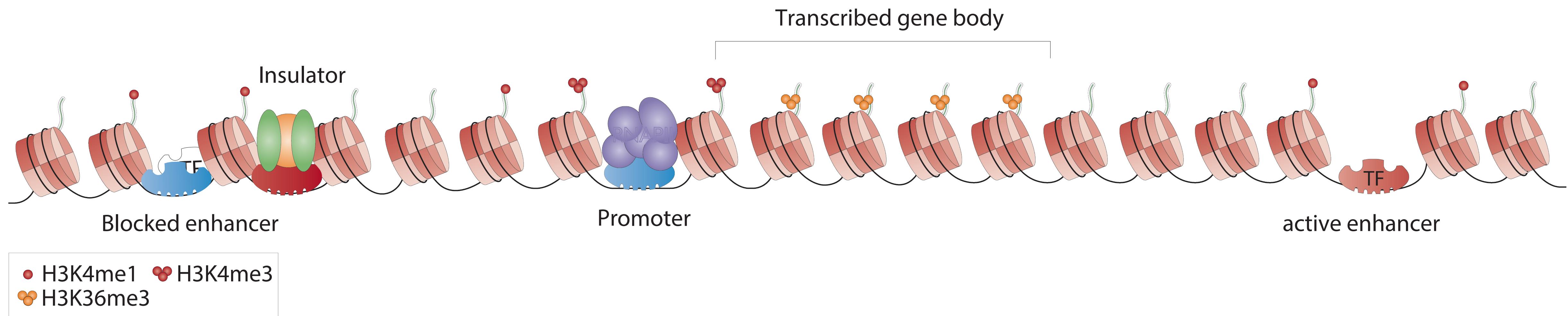
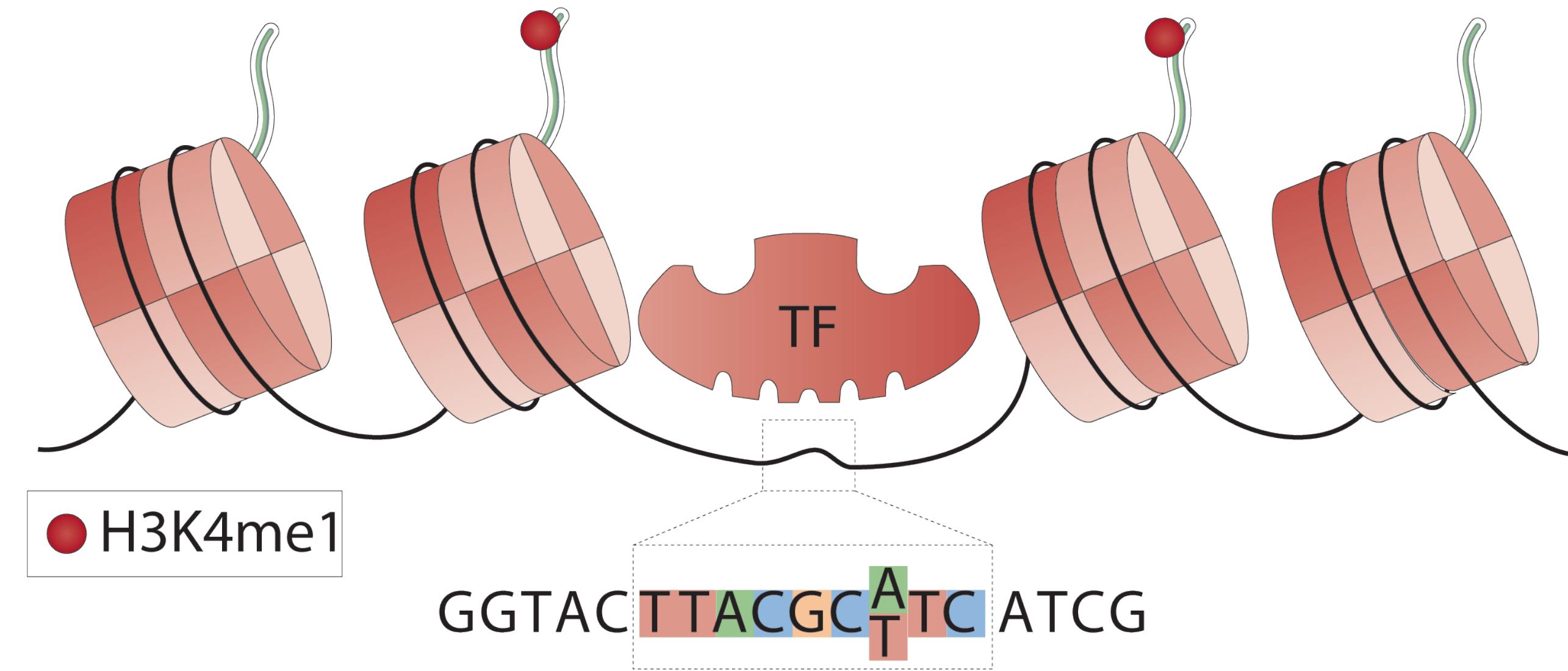


Image source: <http://nihroadmap.nih.gov/epigenomics/>

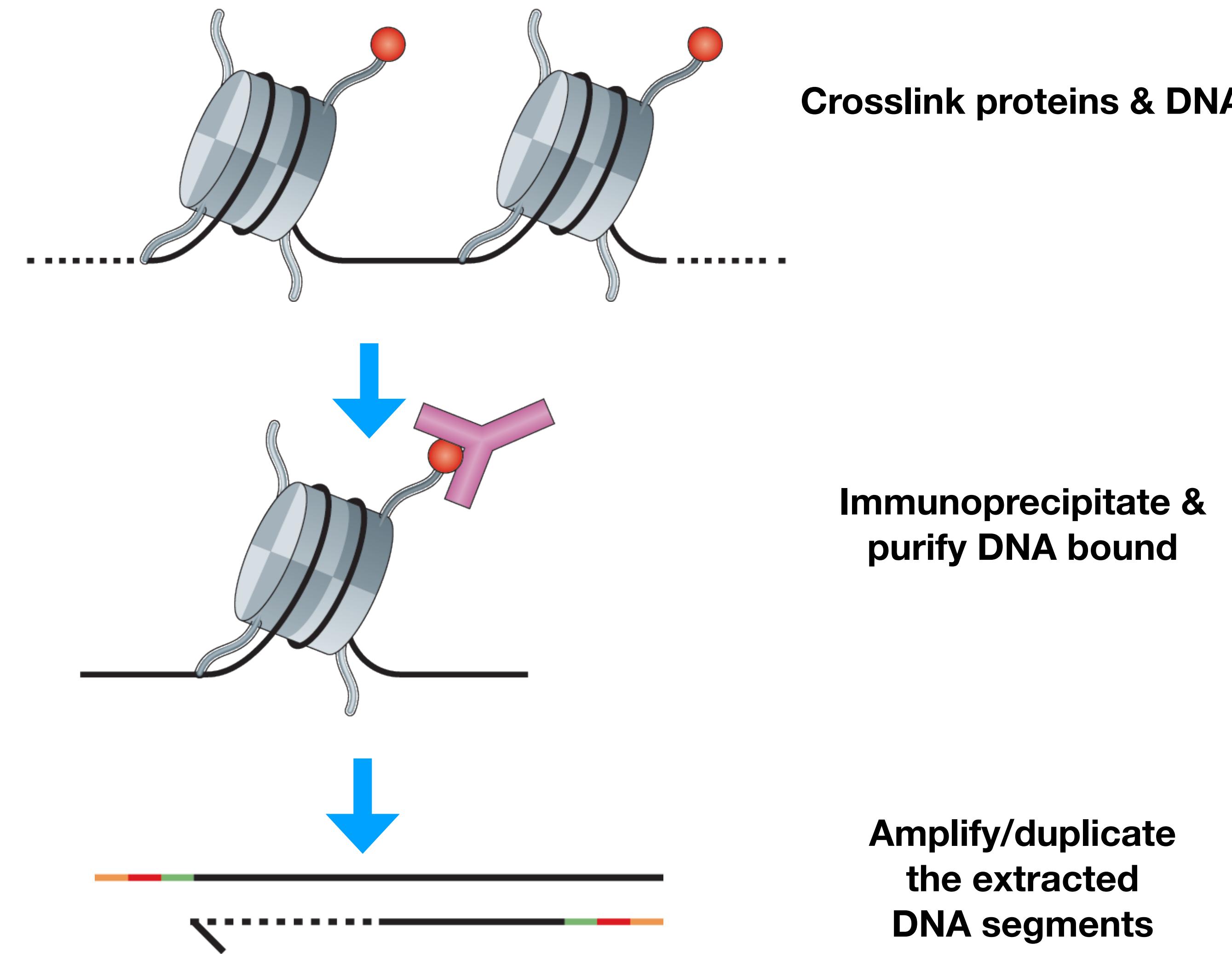
Different histone code marks different types of regulatory elements



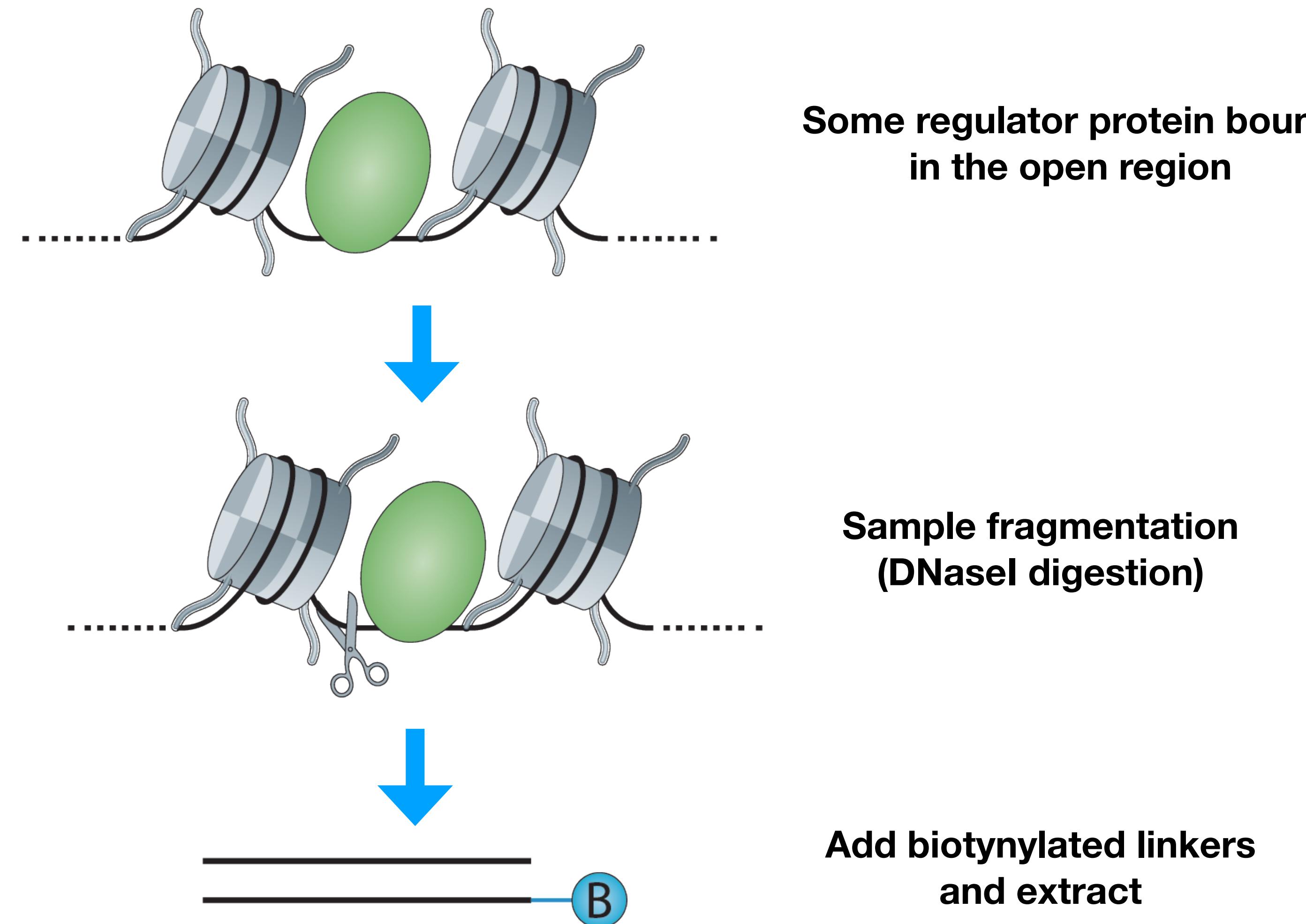
One of the goals is to understand the logic of non-coding DNA sequences



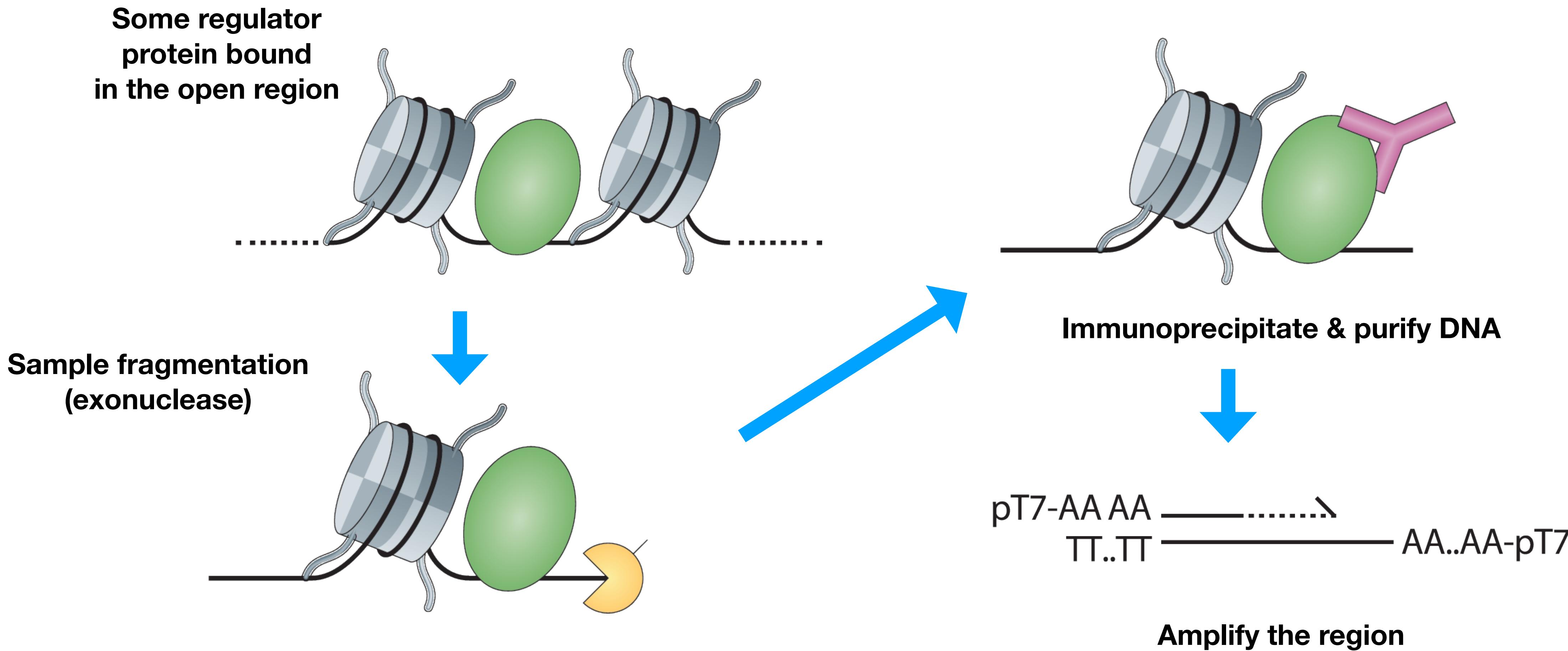
ChIP-seq: Quantifying histone modifications



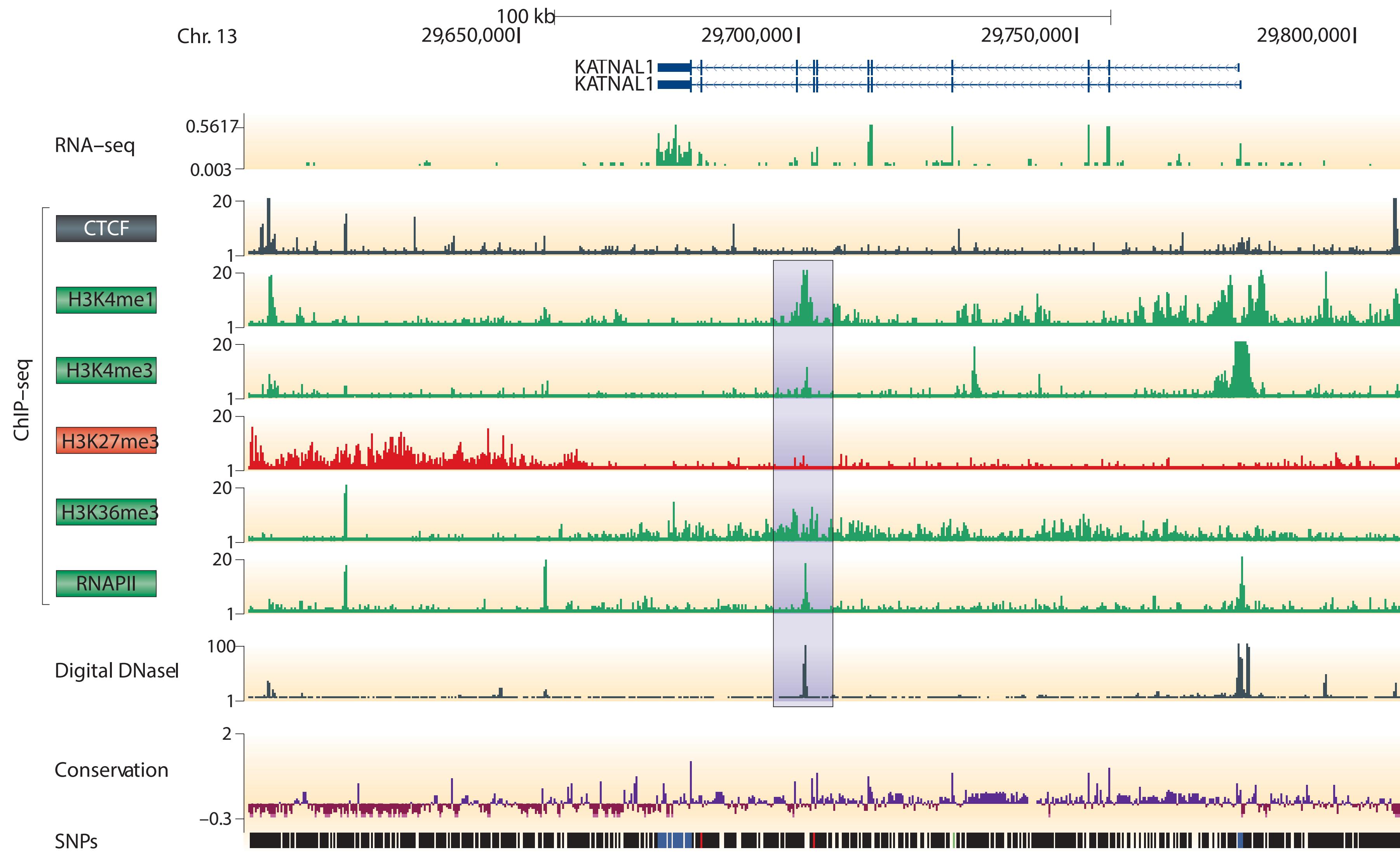
DNase-seq: Quantifying DNA accessibility



ChIP-seq: Quantifying regions bound by a particular protein



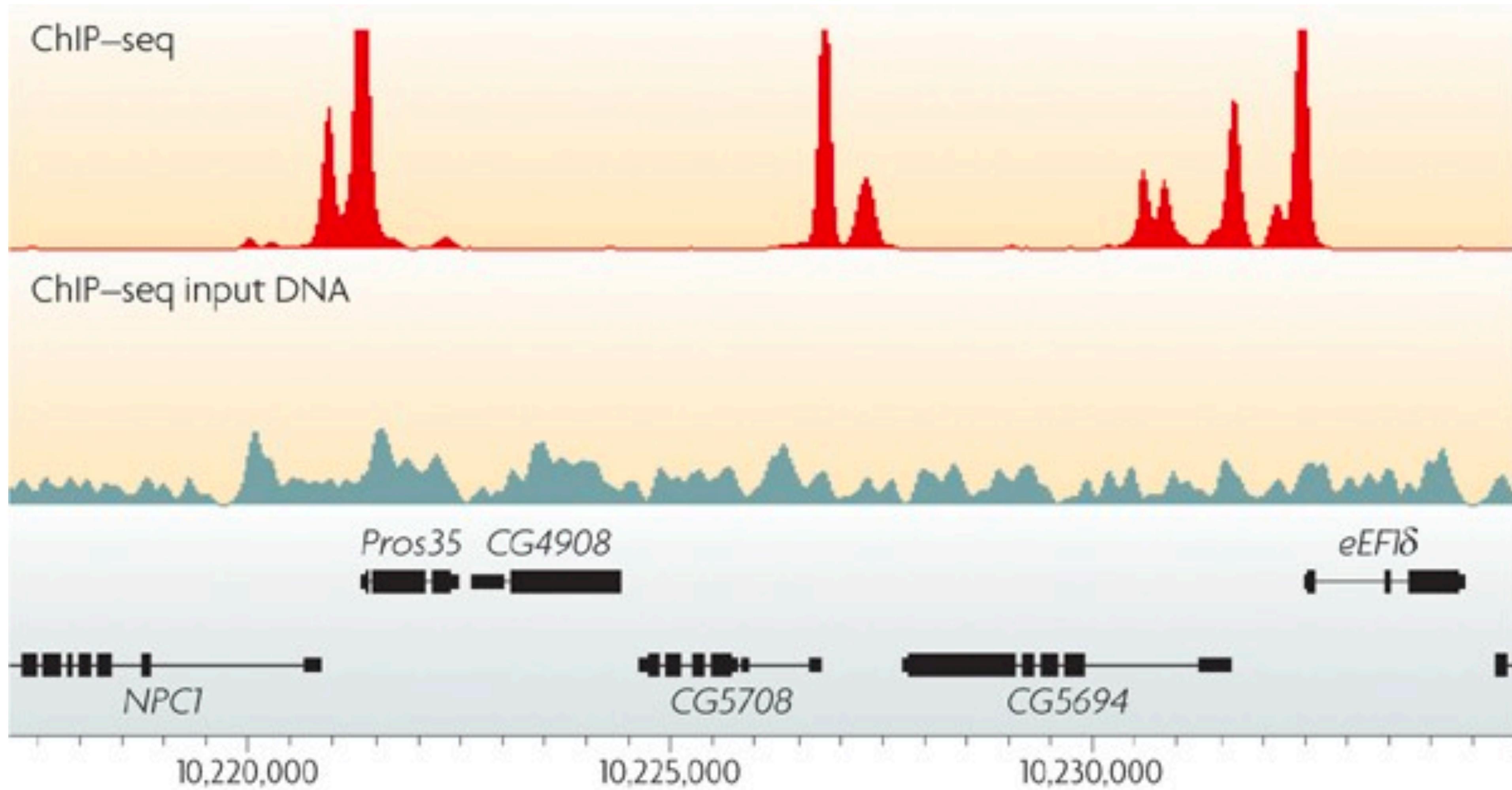
An example of multiple ChIP-seq tracks



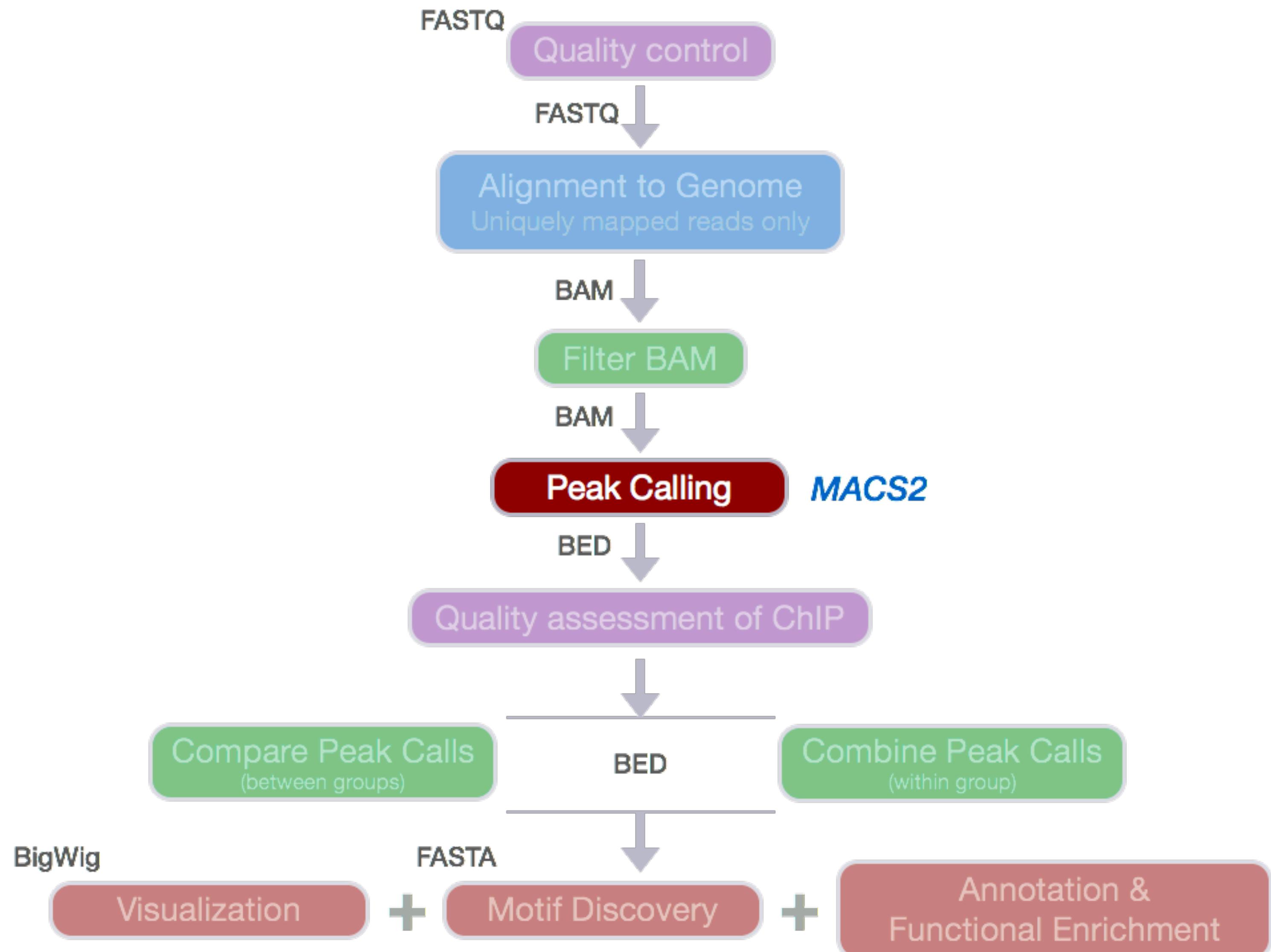
Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

How do we know which regions are "peaks?"



ChIP-seq analysis pipeline



Method | [Open access](#) | Published: 17 September 2008

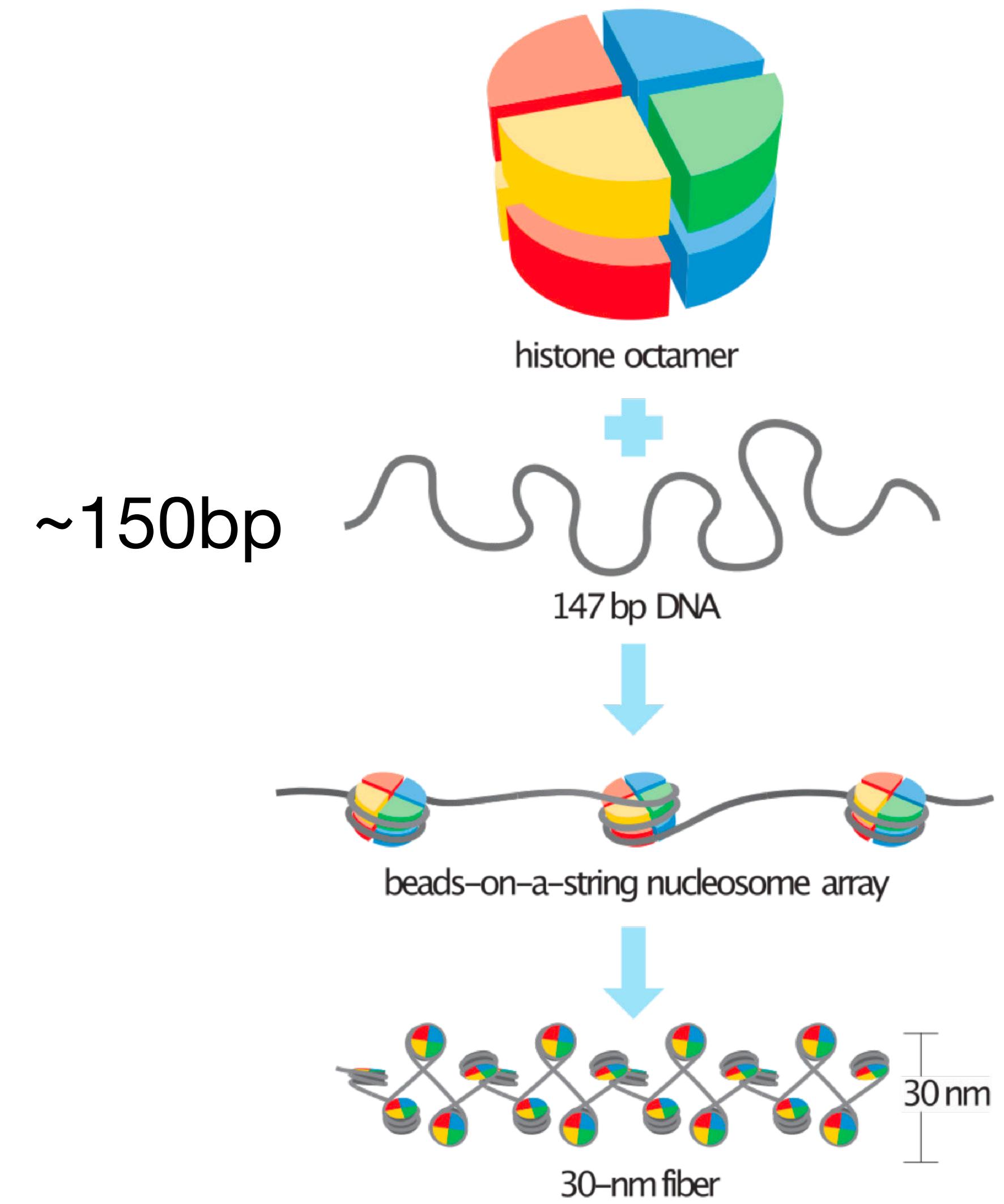
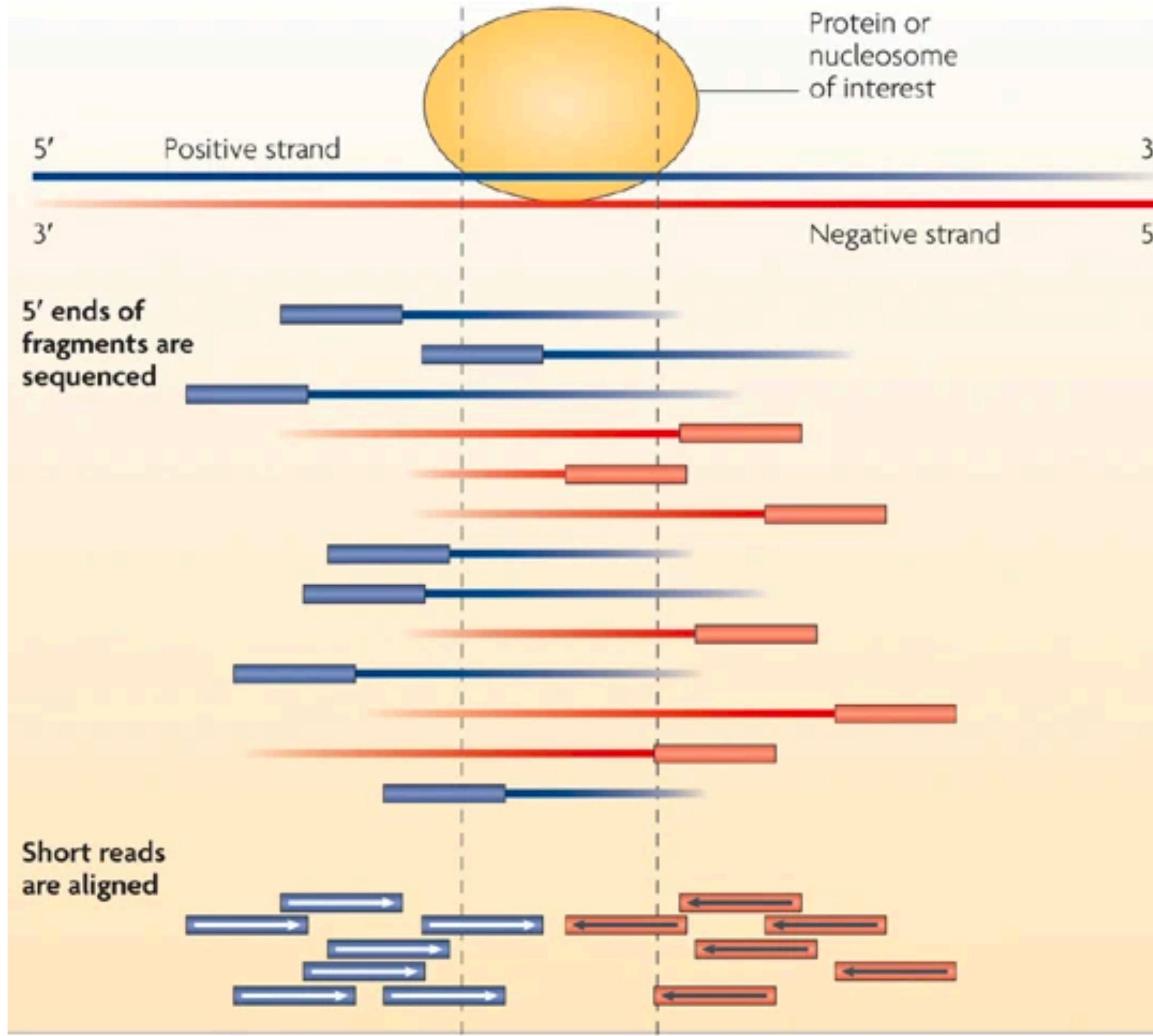
Model-based Analysis of ChIP-Seq (MACS)

[Yong Zhang](#), [Tao Liu](#), [Clifford A Meyer](#), [Jérôme Eeckhoute](#), [David S Johnson](#), [Bradley E Bernstein](#), [Chad Nusbaum](#), [Richard M Myers](#), [Myles Brown](#), [Wei Li](#)✉ & [X Shirley Liu](#)✉

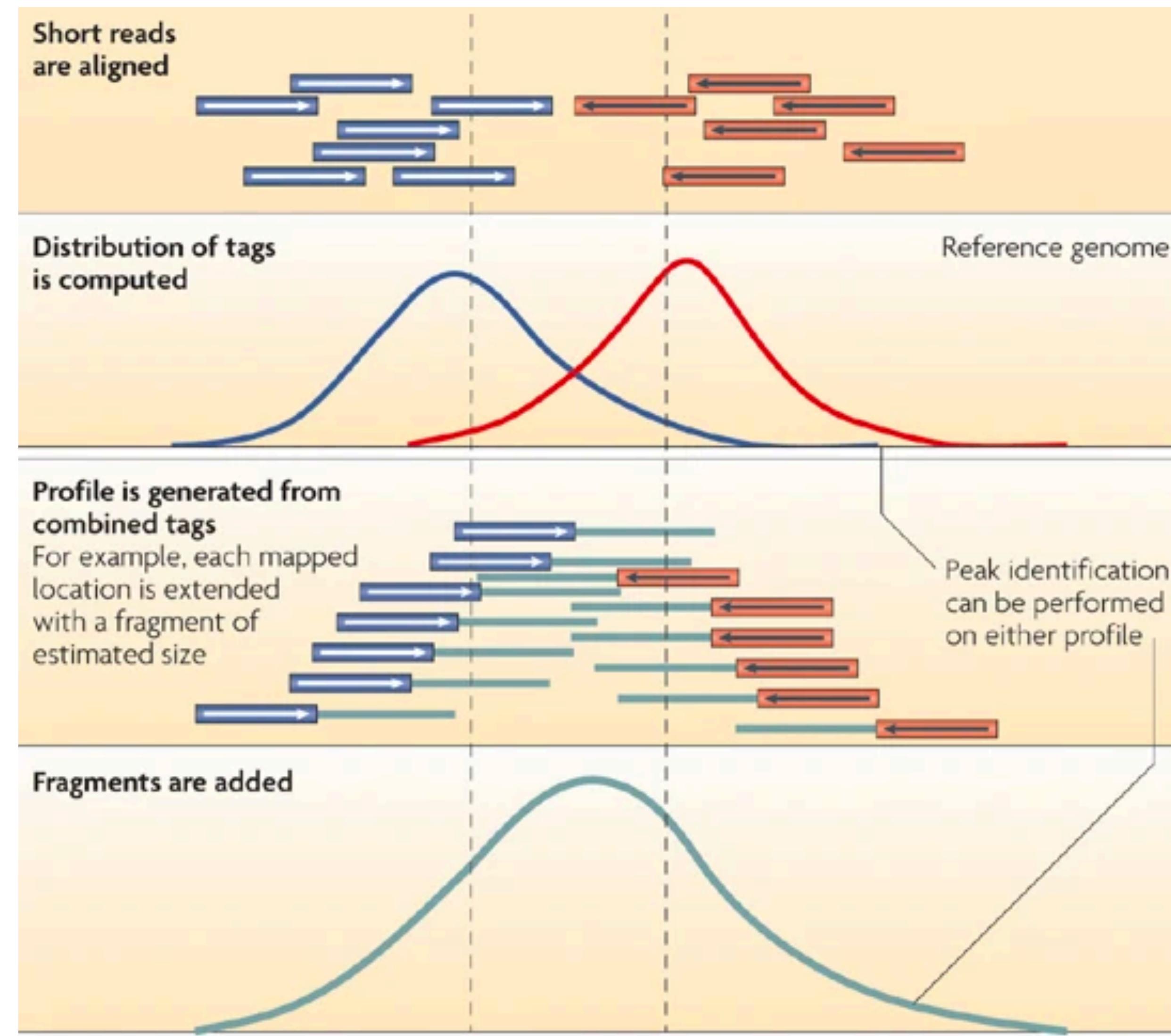
[Genome Biology](#) **9**, Article number: R137 (2008) | [Cite this article](#)

196k Accesses | **9294** Citations | **62** Altmetric | [Metrics](#)

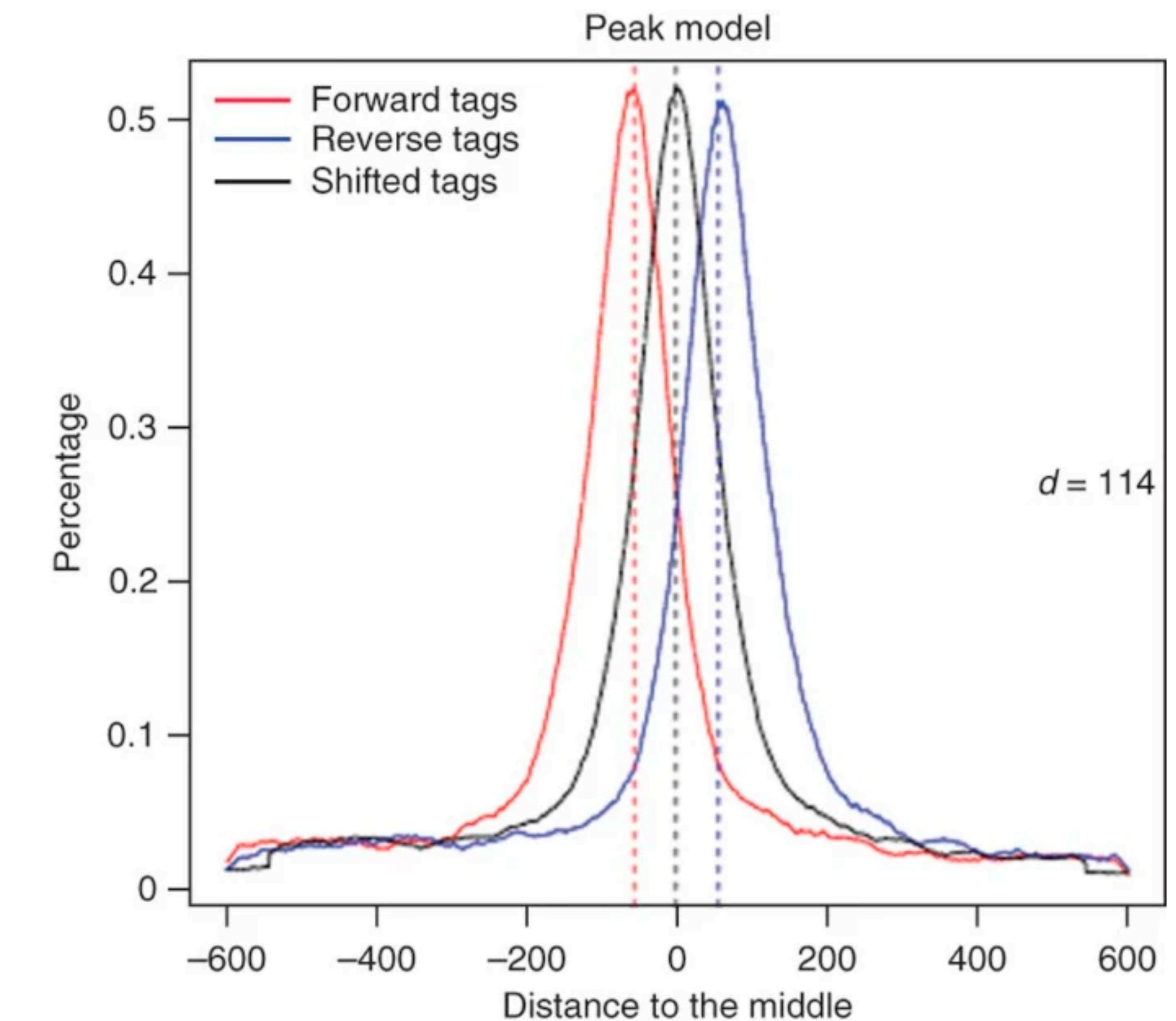
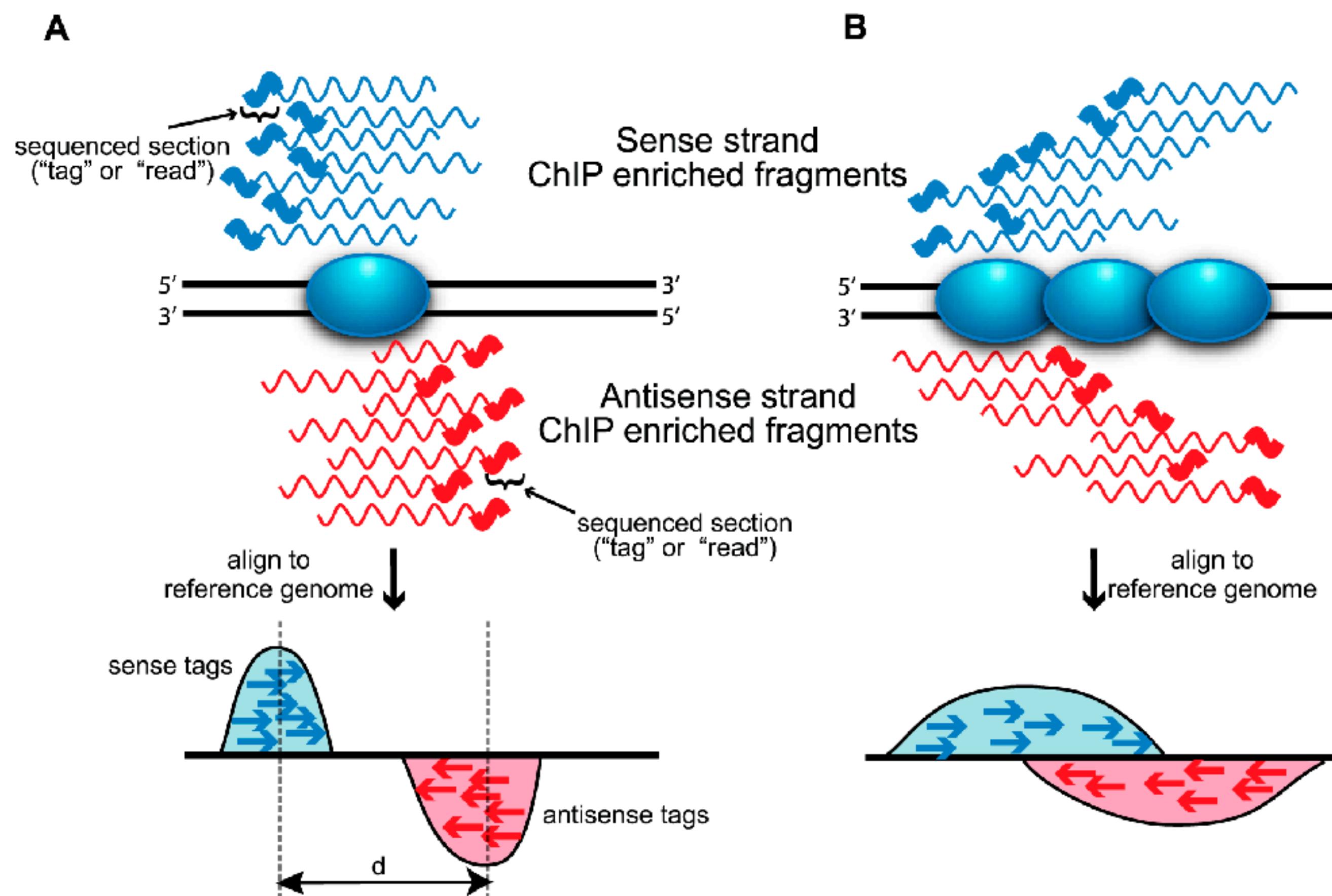
Why not using the same type of bump hunter?



Paired-end, sense, anti-sense seq = more accurate nucleosome positioning



MACS: shift reads toward 3' end



MACS: multiple Poisson p-values

Poisson p-value thresholds

- Read count model: Locally-adjusted' Poisson distribution
- $P(\text{count} = x | \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$
- $\lambda_{\text{local}} = \max(\lambda_{\text{BG}}, [\lambda_{1k},] \lambda_{5k}, \lambda_{10k})$ estimated from control data
- Poisson p -value = $P(\text{count} \geq x)$
- q -value : Multiple hypothesis correction

Peaks: Genomic locations that pass a user-defined p -value
(e.g. 1e-5) or q -value (e.g. 0.01) threshold



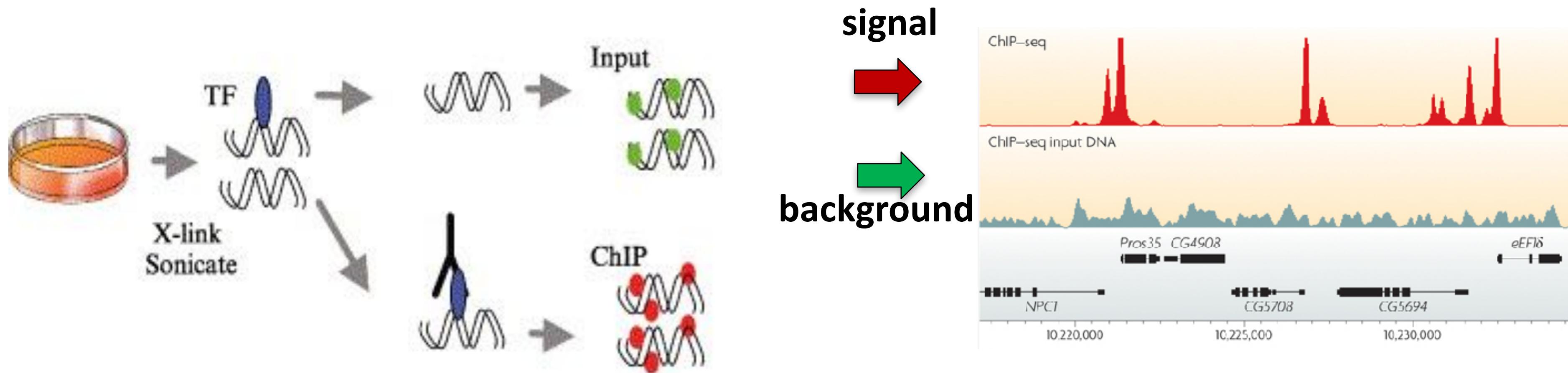
MACS: How do we get the null distribution?

Empirical False discovery rates

- Swap ChIP and input-DNA tracks
- Recompute p -values
- At each p -value, eFDR = Number of control peaks / Number of ChIP peaks
- Use an FDR threshold to call peaks



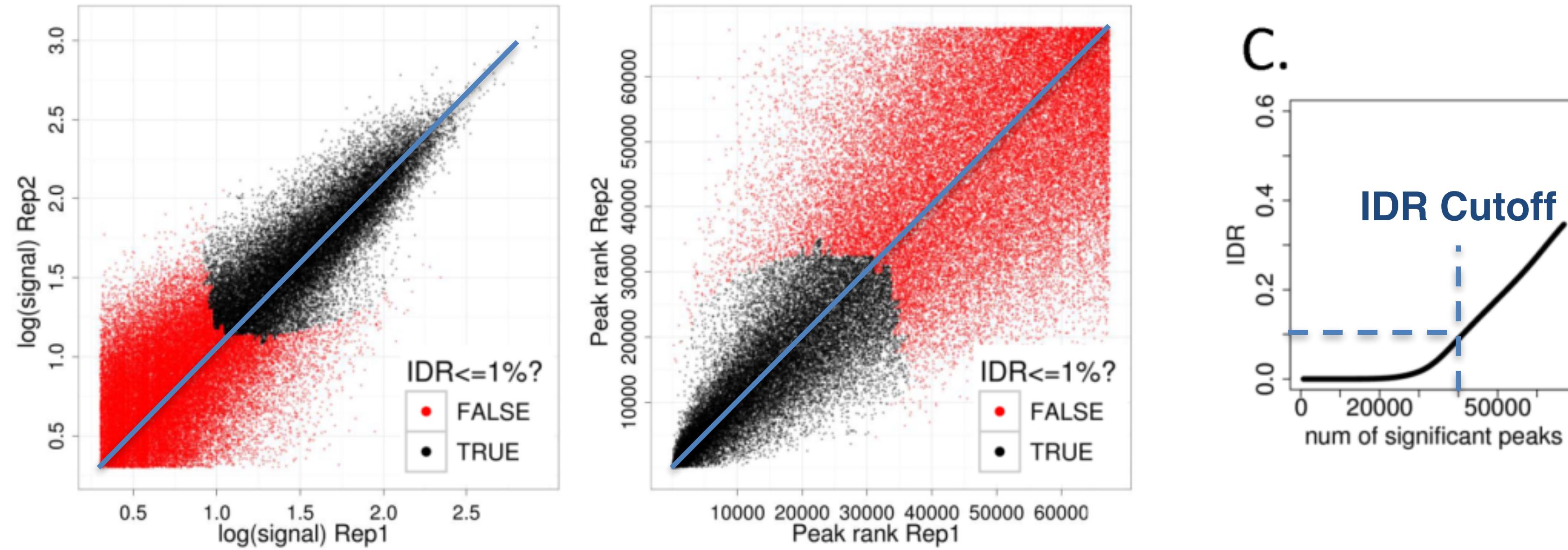
Q/C: Use of input DNA as control dataset



- **Challenge:**
 - Even without antibody: Reads are not uniformly scattered
- **Sources of bias in input dataset scatter:**
 - Non-uniform fragmentation of the genome
 - Open chromatin fragmented more easily than closed regions
 - Repetitive sequences over-collapsed in the assembled genome.
- **How to control for these biases:**
 - Remove portion of DNA sample before ChIP step
 - Carry out control experiment without an antibody (input DNA)
 - Fragment input DNA, sequence reads, map, use as background



How do we know peaks are good enough? Measure Irreproducibility!



- Key idea: True peaks will be highly ranked in both replicates
 - Keep going down rank list, until ranks are no longer correlated
 - This cutoff could be different for the two replicates
 - The actual peaks included may differ between replicates
- Adaptively learn optimal peak calling threshold
 - FDR threshold of 10% → 10% of peaks are false (widely used)
 - IDR threshold of 10% → 10% of peaks are not reproducible



The IDR model: A two component mixture model

- Looking only at ranks means that the marginals are uniform, so all the information is encoded in the joint distribution.
- Model the joint distribution of ranks as though it came from a two component Gaussian mixture model:

$$(x, y) \sim pN(\mu, \mu, \sigma, \sigma, \rho) + (1 - p)N(0, 0, 1, 1, 0)$$



- 2D Copula mixture

Read & use IDR if interested in

The Annals of Applied Statistics
2011, Vol. 5, No. 3, 1752–1779
DOI: 10.1214/11-AOAS466
© Institute of Mathematical Statistics, 2011

MEASURING REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS¹

BY QUNHUA LI, JAMES B. BROWN, HAIYAN HUANG AND PETER J. BICKEL

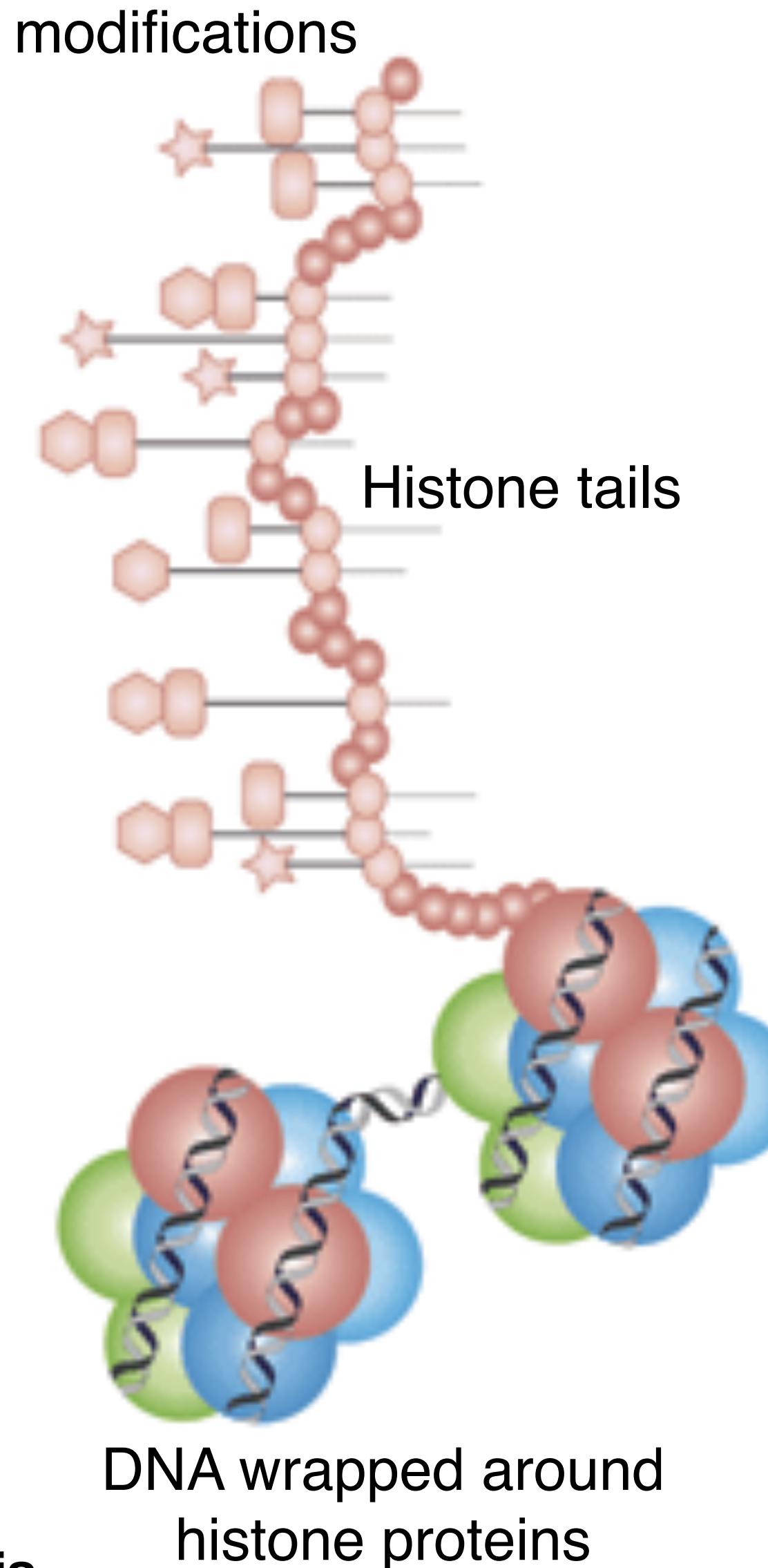
University of California at Berkeley

idr: Irreproducible Discovery Rate

This is a package for estimating the copula mixture model and plotting correspondence curves. Details are in "Measuring reproducibility of high-throughput experiments" (2011), Annals of Applied Statistics, Vol. 5, No. 3, 1752-1779, by Li, Brown, Huang, and Bickel.

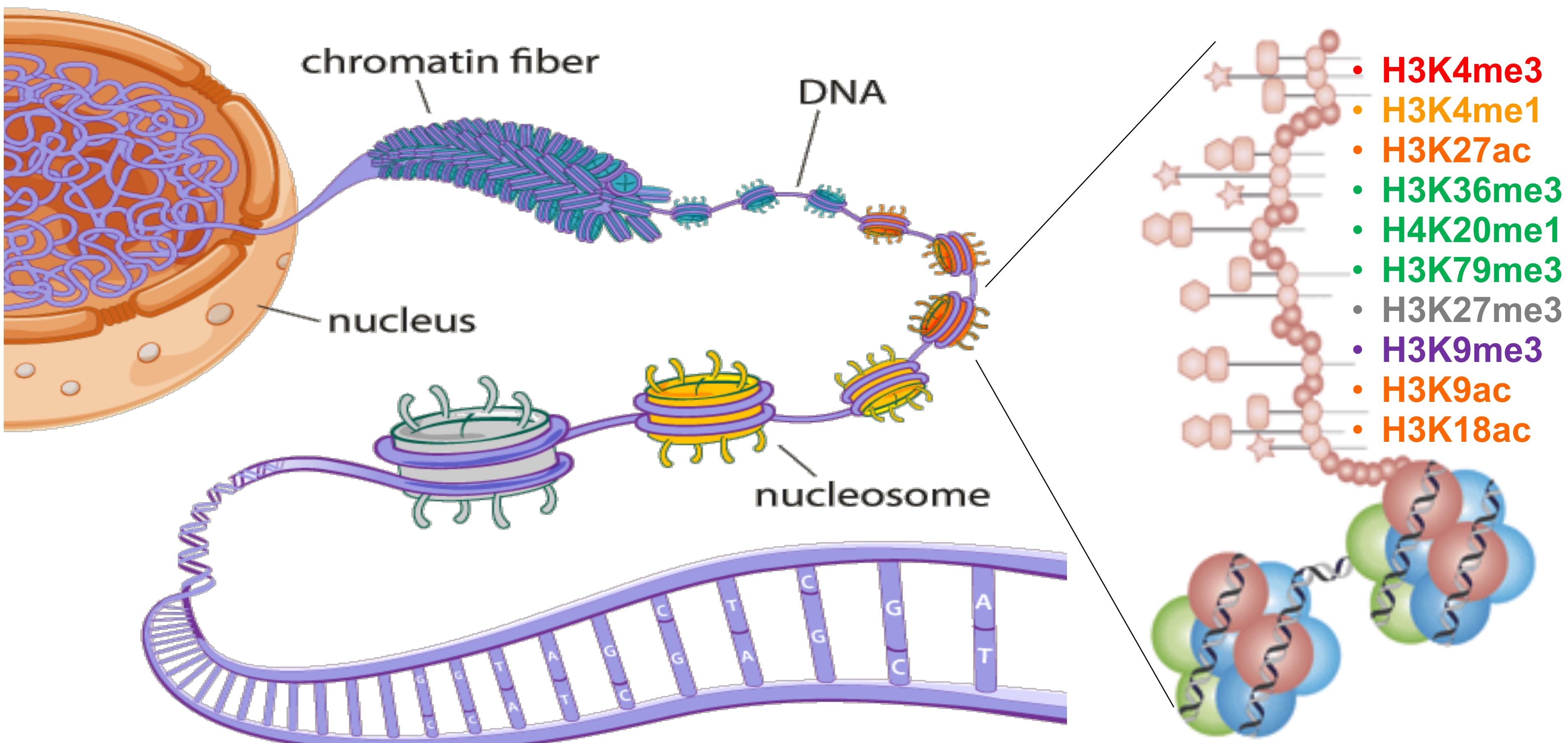
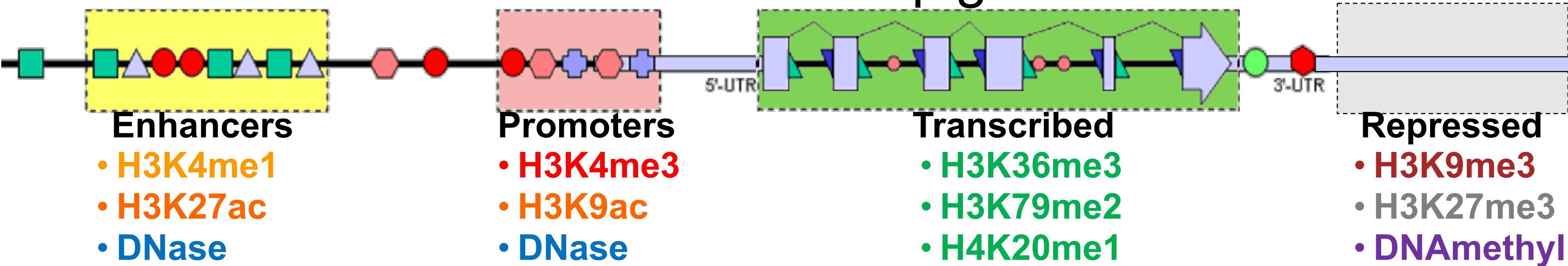
Version: 1.3
Imports: stats
Published: 2022-06-21
Author: Qunhua Li
Maintainer: Qunhua Li <qunhua.li at psu.edu>
License: [GPL-2](#) | [GPL-3](#) [expanded from: GPL (≥ 2.0)]
NeedsCompilation: no
CRAN checks: [idr results](#)

100s of histone tail modifications



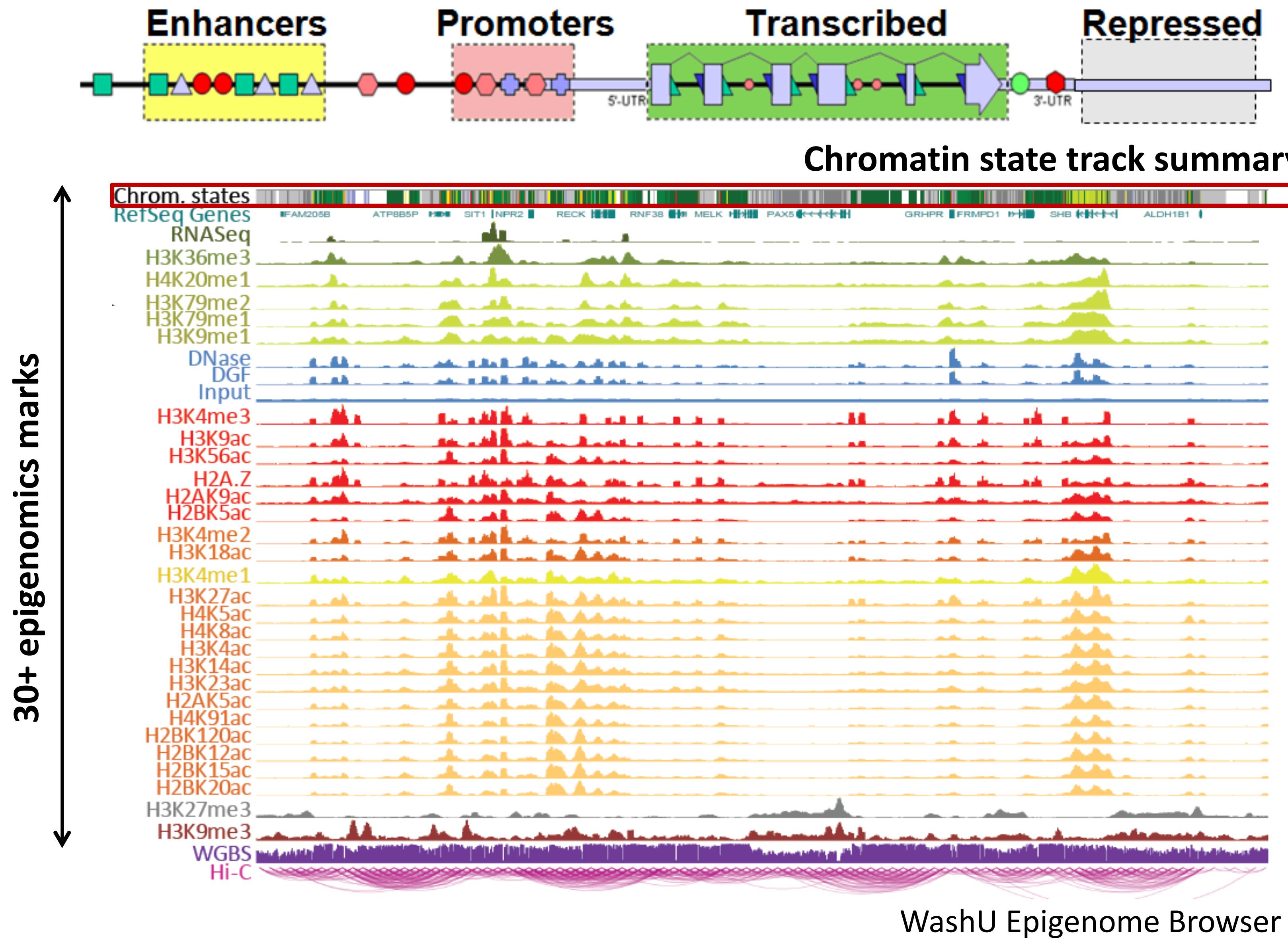
- 100+ different histone modifications
 - Histone protein → H3/H4/H2A/H2B
 - AA residue → Lysine4(K4)/K36...
 - Chemical modification → Met/Pho/Ubi
 - Number → Me-Me-Me(me3)
 - Shorthand: H3K4me3, H2BK5ac
- In addition:
 - DNA modifications
 - Methyl-C in CpG / Methyl-Adenosine
 - Nucleosome positioning
 - DNA accessibility
- The constant struggle of gene regulation
 - TF/histone/nucleo/GFs/Chrom compete⁹⁵

Combinations of marks encode epigenomic state

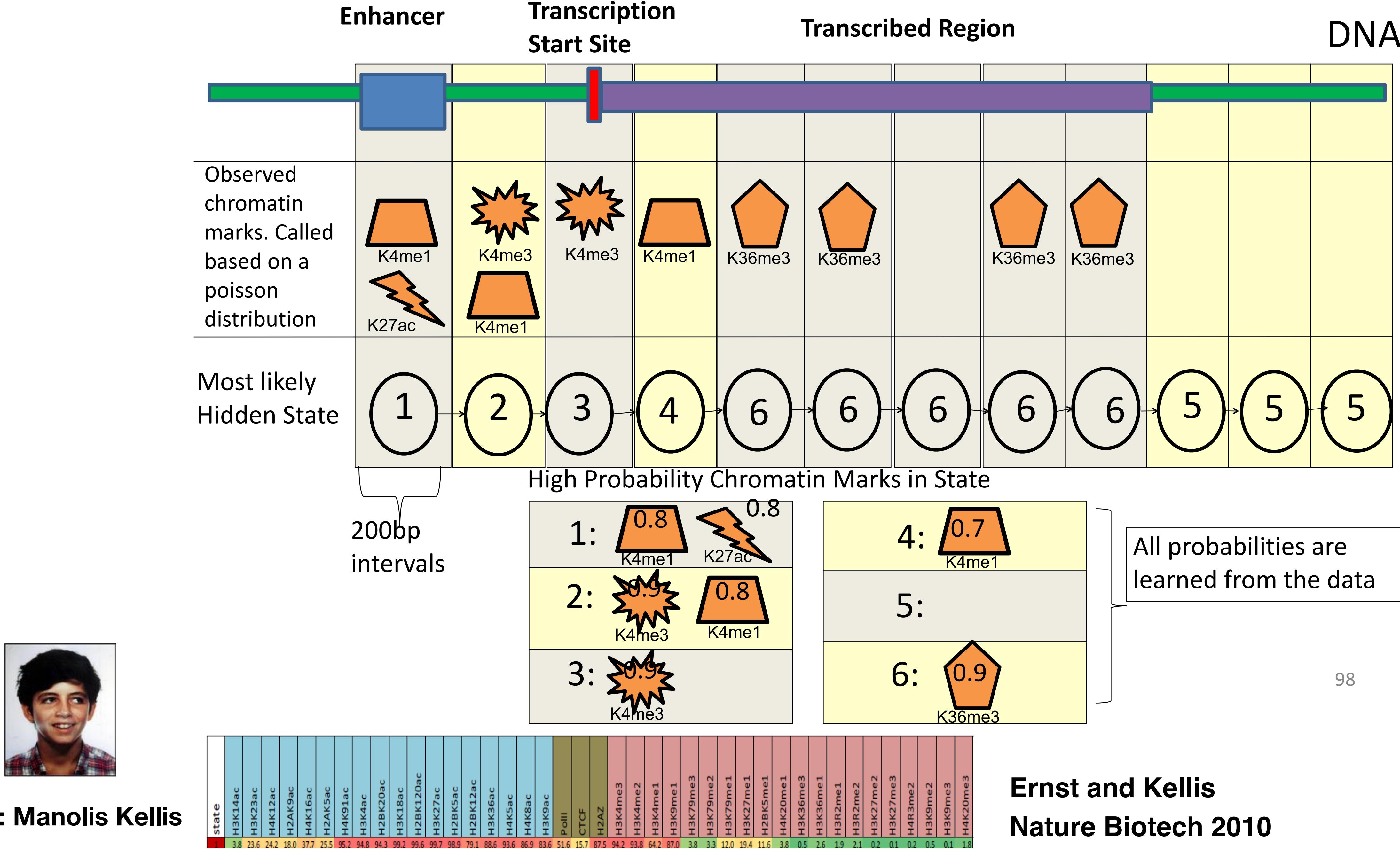


- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

Summarize multiple marks into chromatin states



Multivariate HMM for Chromatin States



98

| state | H3K14ac | H3K23ac | H4K12ac | H2AK9ac | H4K16ac | H2AK5ac | H4K91ac | H3K4ac | H2BK20ac | H3K27ac | H2BK12ac | H3K36ac | H2BK5ac | H4K5ac | H3K9ac | POLI | CTCF | H2AZ | H3K4me3 | H3K4me2 | H3K4me1 | H3K9me1 | H3K79me3 | H3K79me2 | H3K27me1 | H3K5me1 | H4K20me1 | H3K36me3 | H3K36me1 | H3R2me1 | H3R2me2 | H3K27me2 | H3K27me3 | H4R3me2 | H3K9me2 | H3K9me3 | H4K20me3 | | | | |
|-------|---------|---------|---------|---------|---------|---------|---------|--------|----------|---------|----------|---------|---------|--------|--------|------|------|------|---------|---------|---------|---------|----------|----------|----------|---------|----------|----------|----------|---------|---------|----------|----------|---------|---------|---------|----------|-----|-----|-----|-----|
| 1 | 3.8 | 23.6 | 24.2 | 18.0 | 37.7 | 25.5 | 95.2 | 94.8 | 94.3 | 99.2 | 99.6 | 99.7 | 98.9 | 79.1 | 88.6 | 93.6 | 86.9 | 83.6 | 51.6 | 15.7 | 87.5 | 94.2 | 93.8 | 64.2 | 87.0 | 3.8 | 3.3 | 12.0 | 19.4 | 11.6 | 3.8 | 0.5 | 2.6 | 1.9 | 2.1 | 0.2 | 0.1 | 0.2 | 0.5 | 0.1 | 1.8 |

Ernst and Kellis
Nature Biotech 2010

Statistical Methods for Epigenomics

- **Review: a set-up for epigenomics profiling**
 - Importance of Reference Genome
- **DNA methylation--basics**
 - Why do we investigate DNA methylation?
 - Bisulfite conversion: methyl-CpG tagging
- **Statistical methods for DNA methylation analysis**
 - A method treating each CpG as a variant
 - A method treating aggregating signals across genome
- **A brief overview of other ChIP-seq analysis**
 - Technology and biology
 - Peak calling
 - A step forward (too big for one person's project)

