Statistical analysis of ChIP-seq data

Sündüz Keleş

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Feb 9, 2012

Basic principles of gene expression

- Each cell contains a complete copy of the organism's genome (the same hardware!).
- Cells are of many different types and states. **E.g.** skin, blood, and nerve cells, cancerous cells, etc.
- What makes the cells different?
- Each cell utilizes only a subset of the whole set of genes. Differential gene expression, i.e., when, where, and how much each gene is expressed.
- The mechanism that controls gene expression is called the *regulation of gene expression*.

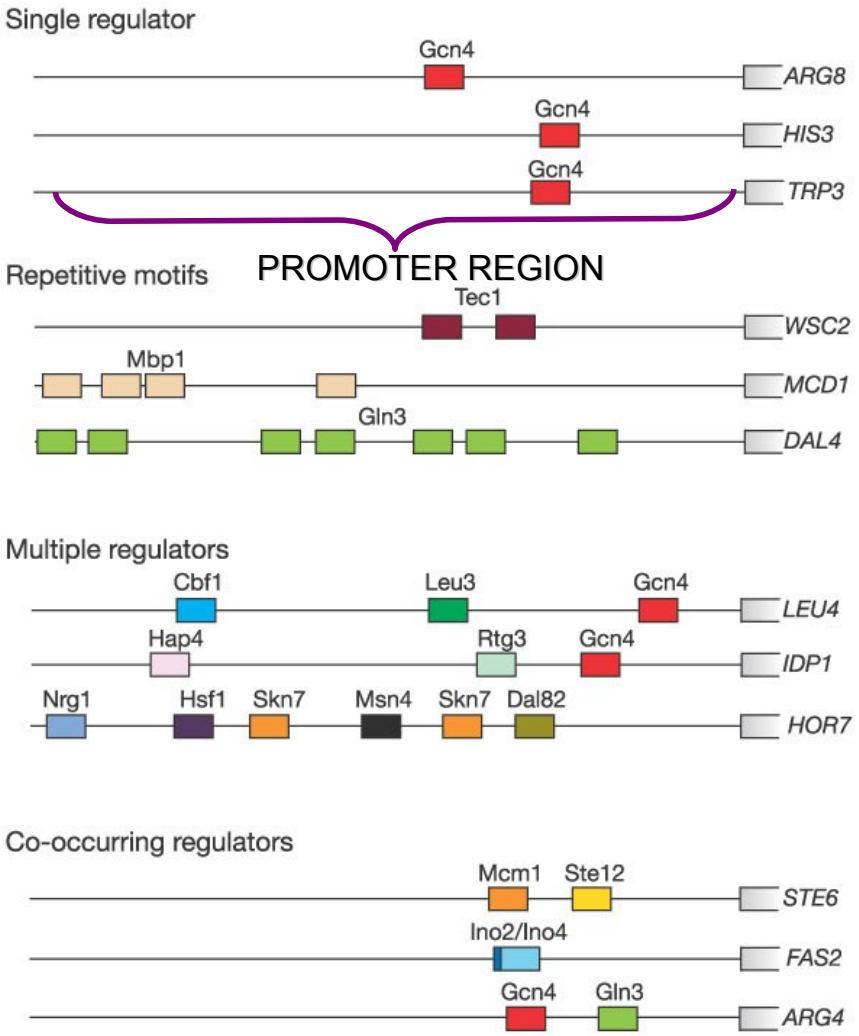
Stages of regulation of gene expression

- during chromatin modifications (DNA packaging),
- during transcription control,
- splicing,
- transport and translation control.

Transcriptional control: most common way of regulation; occurs during the transcription phase when the DNA is transcribed into RNA.
Basic elements of transcriptional control:

- Transcription factors,
- DNA binding sites (regulatory motifs), enhancers,
- Promoters.

Complexity of eukaryotic transcriptional regulation



TGA-T-

[A/T]CGCG[A/T]

TCACGTG

CCG----CGG
GGC----GCC

Harbison et al. Nature (2004)

Data collection for the motif finding problem

Through microarray technology:

- Genomewide gene expression data: Relative abundance of gene expression in different cell types is measured.
- ChIP-chip data: Genome-wide maps of DNA and protein interactions (Ren, B. et al. (2000) & Iyer, V.R. et al. (2001), Simon, I. (2001), Lieb, J. (2001), Lee et al. (2002).) [a.k.a. ChIP-chip data]

Through comparative genomics: Multiple genome sequences from related species.

Through high throughput sequencing: Genome-wide maps of DNA and protein interactions by ChIP-Seq experiments.

Data collection for motif finding problem

- Based on expression or multiple species data, we extract 500-1000bps upstream of the transcription start sites (TSS).
- ChIP-chip/seq data generates specific coordinates of binding which may not be restricted to upstream of the TSS.

High throughput ChIP assay (ChIP-seq): Chromatin immunoprecipitation combined with high throughput sequencing

- ChIP: Chromatin immunoprecipitation (*in vivo*).
- seq: High throughput sequencing (mainly Illumina for us).
- ChIP-seq: ChIP followed by high throughput sequencing.

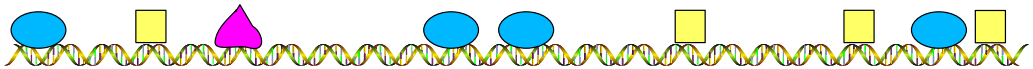
ChIP assay

Target protein =

ChIP assay

Target protein =

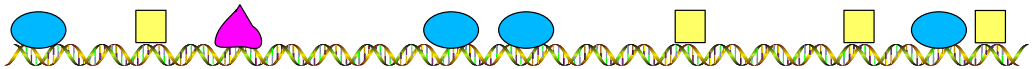
1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.



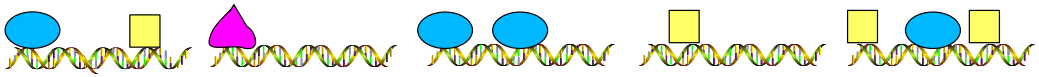
ChIP assay

Target protein = 

1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.



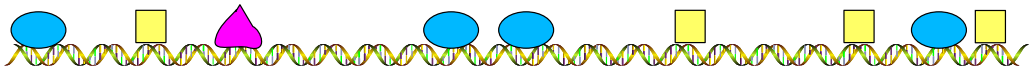
2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).



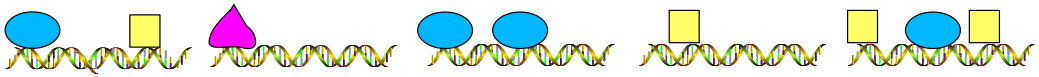
ChIP assay

Target protein =

1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.

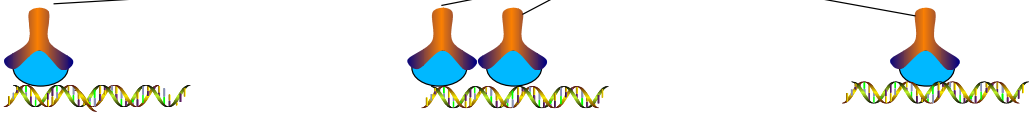


2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).



3. Immunoprecipitate using a target protein-specific antibody.

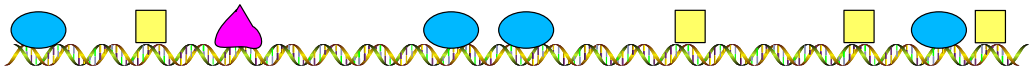
Selectively binds to the target protein.



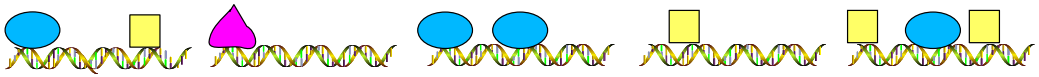
ChIP assay

Target protein =

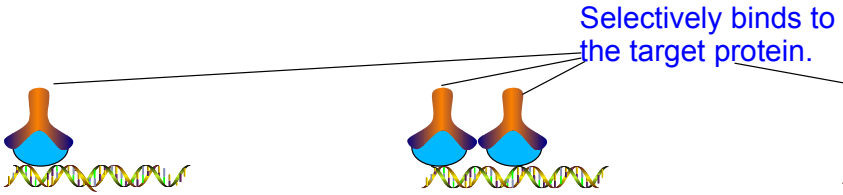
1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.



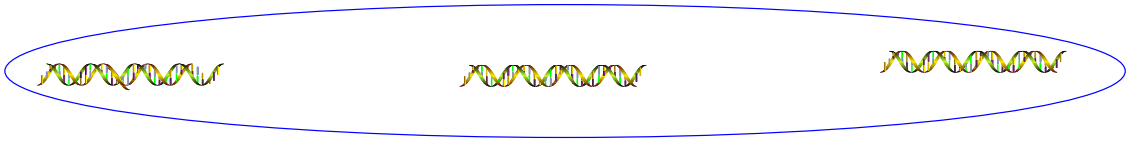
2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).



3. Immunoprecipitate using a target protein-specific antibody.



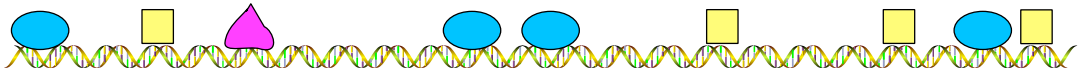
4. Reverse the cross-links and purify DNA.



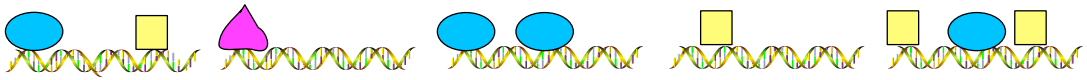
ChIP assay

Target protein =

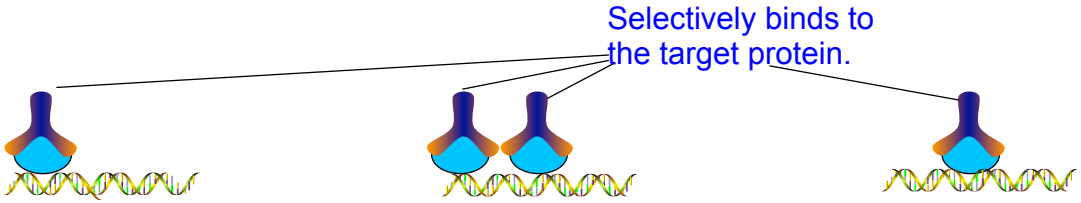
1. Crosslink DNA and protein *in vivo* by exposing cells to formaldehyde.



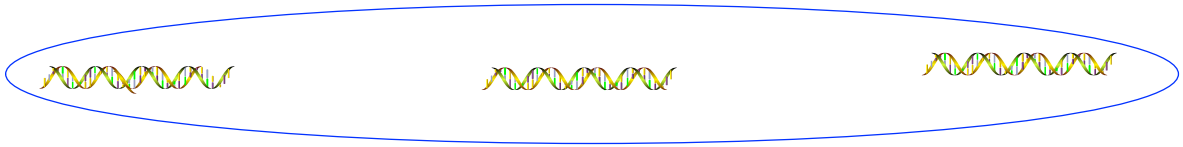
2. Extract the chromatin from cells and fragment by sonication (~ 500-1000 bps).



3. Immunoprecipitate using a target protein-specific antibody.



4. Reverse the cross-links and purify DNA.



5. Find the identity of the isolated DNA fragments.

Variations: MNase-seq for nucleosome occupancy

In high throughput experiments experiments measuring nucleosome occupancy, an enzyme called "Micrococcal nuclease" is used to digest nucleosome free regions instead of sonication + immunoprecipitation.

Traditional ChIP assay

- The identity of the DNA fragments isolated in complex with the protein of interest can then be determined by polymerase chain reaction (PCR) using primers specific for the DNA regions that the protein in question is hypothesized to bind.
- One experiment per hypothesized region.
- Identify the identity of all the immunoprecipitated regions? Sequencing (ChIP-Seq) (previously with ChIP-chip).

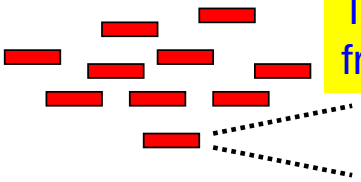
ChIP-Seq data generation

Next Gen Seq Tech

ChIP-Seq data generation

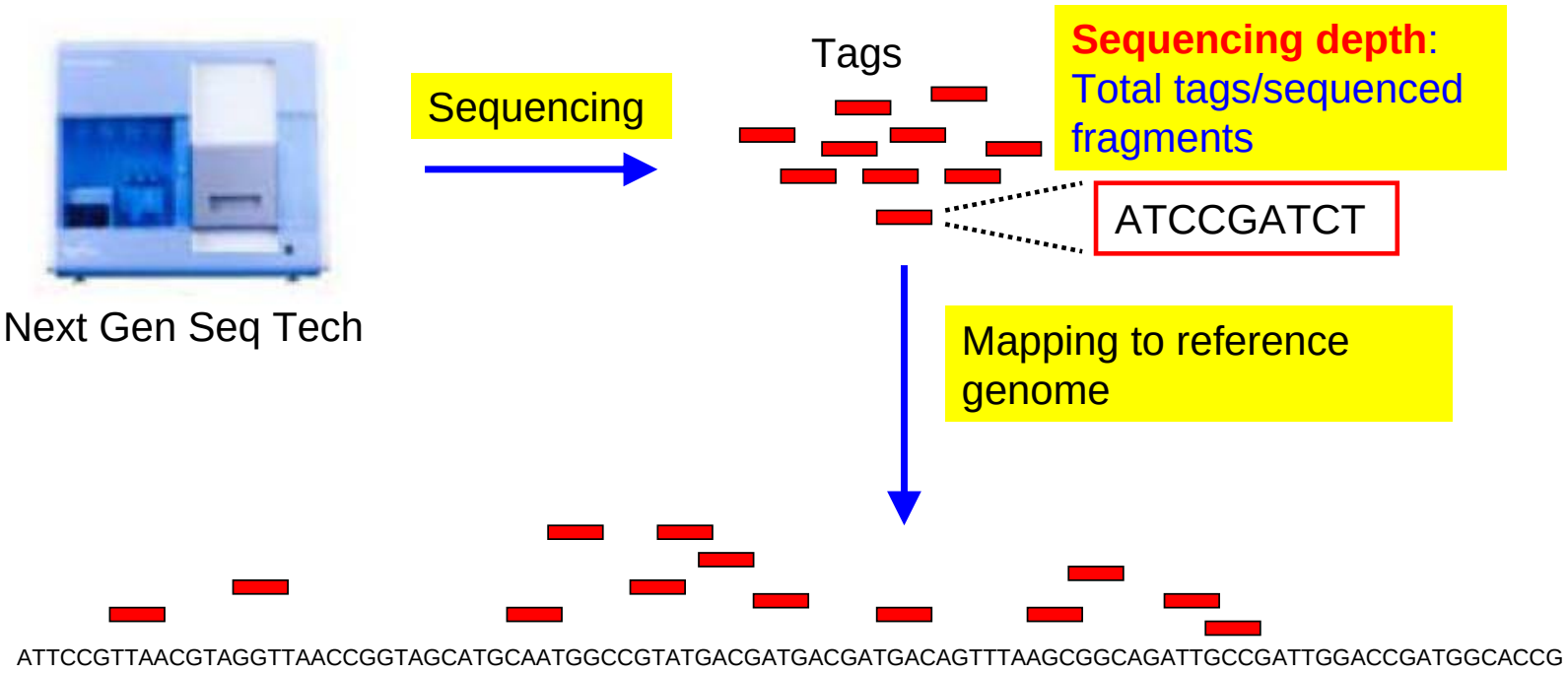
Next Gen Seq Tech

Sequencing

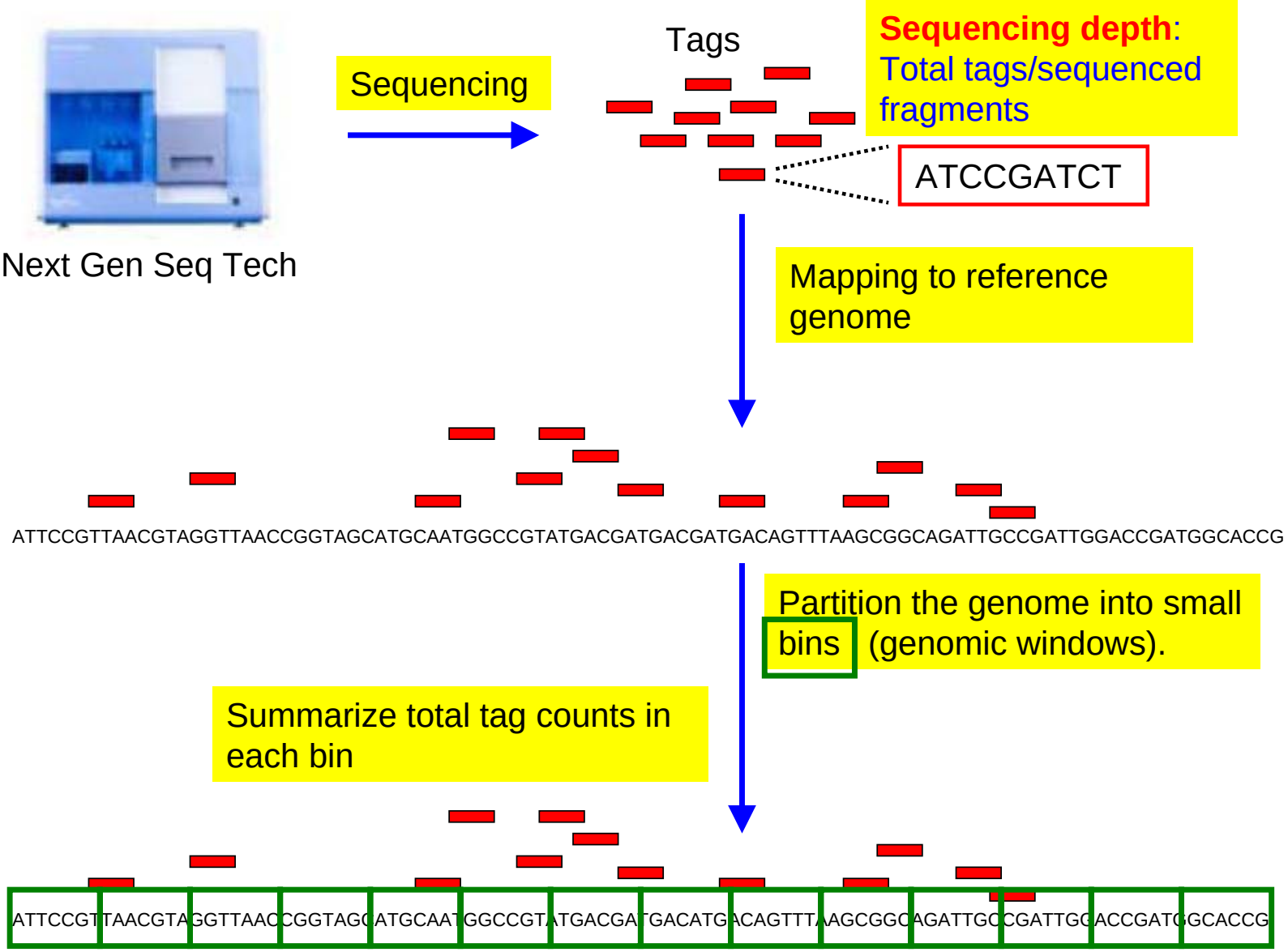


Sequencing depth: Total tags/sequenced fragments

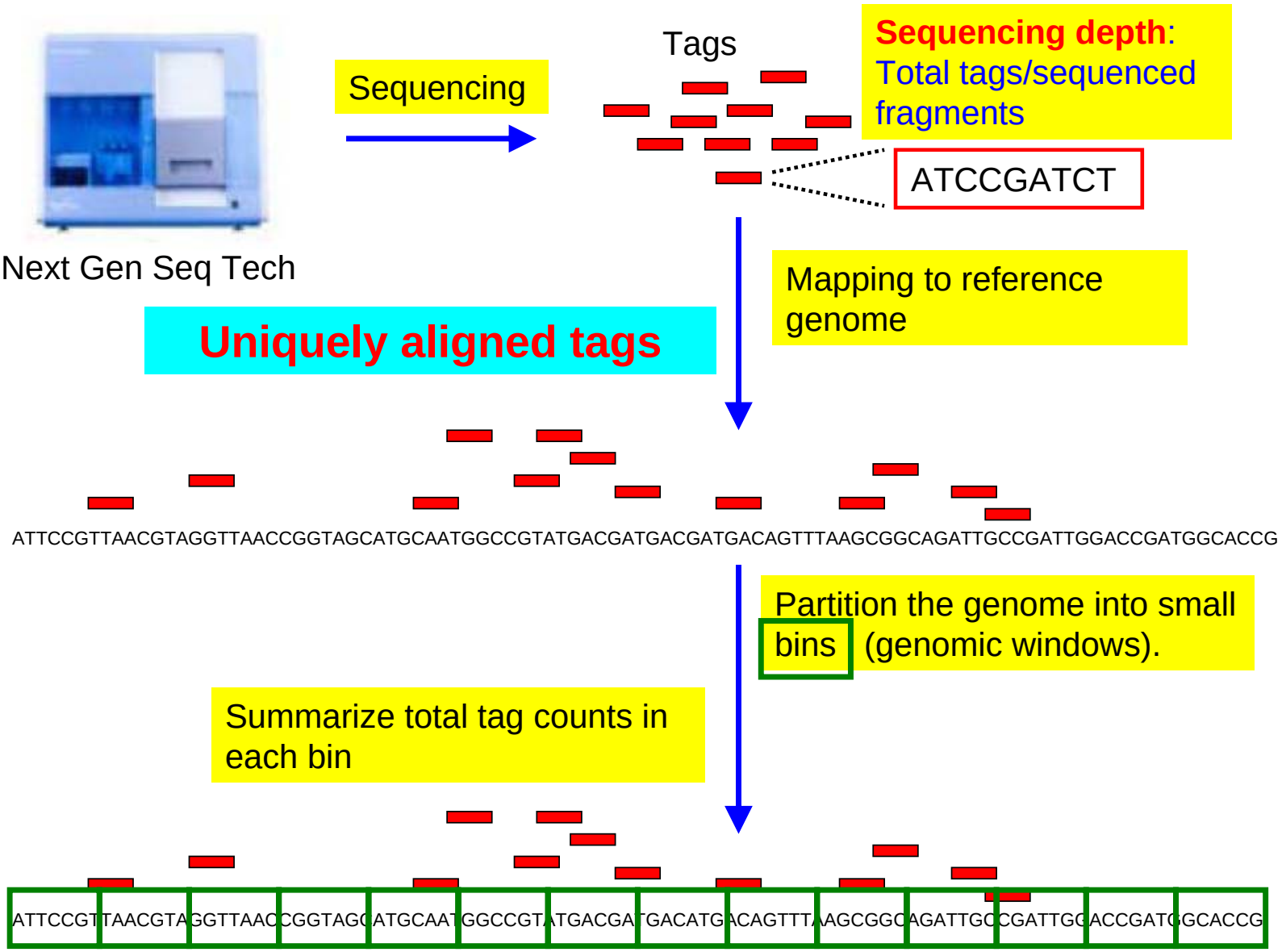
ChIP-Seq data generation



ChIP-Seq data generation



ChIP-Seq data generation



FASTQ format

```
@HWUSI-EAS1789_0000:5:1:1049:9966#GGCTAN/1
CAGAAGTGCATCAAACATGATTTAGAGCTTGTTTAT
+HWUSI-EAS1789_0000:5:1:1049:9966#GGCTAN/1
‘ddadc^aYc\b\Ybc^d^^cdd\cdddccaccc^d
```

- The first line begins with an @ symbol and is followed by the sequence name.

  `@lane:tile:x_coordinate_on_tile:y_coordinate_on_tile:quality_filter`

- The second line contains the base call (in this case for each of 36 nucleotides).
- The third line begins with a + symbol and may (or may not) repeat the sequence name.
- The fourth line contains a symbol that measures the quality score for the corresponding base call as listed on the second line. There should be one symbol for each base call. The symbol on the fourth line uses an ASCII character (American Standard Code for Information Interchange) to encode the quality score.

Mapping reads to reference genome: Tools

- Eland
- Bowtie
- MAQ
- ...

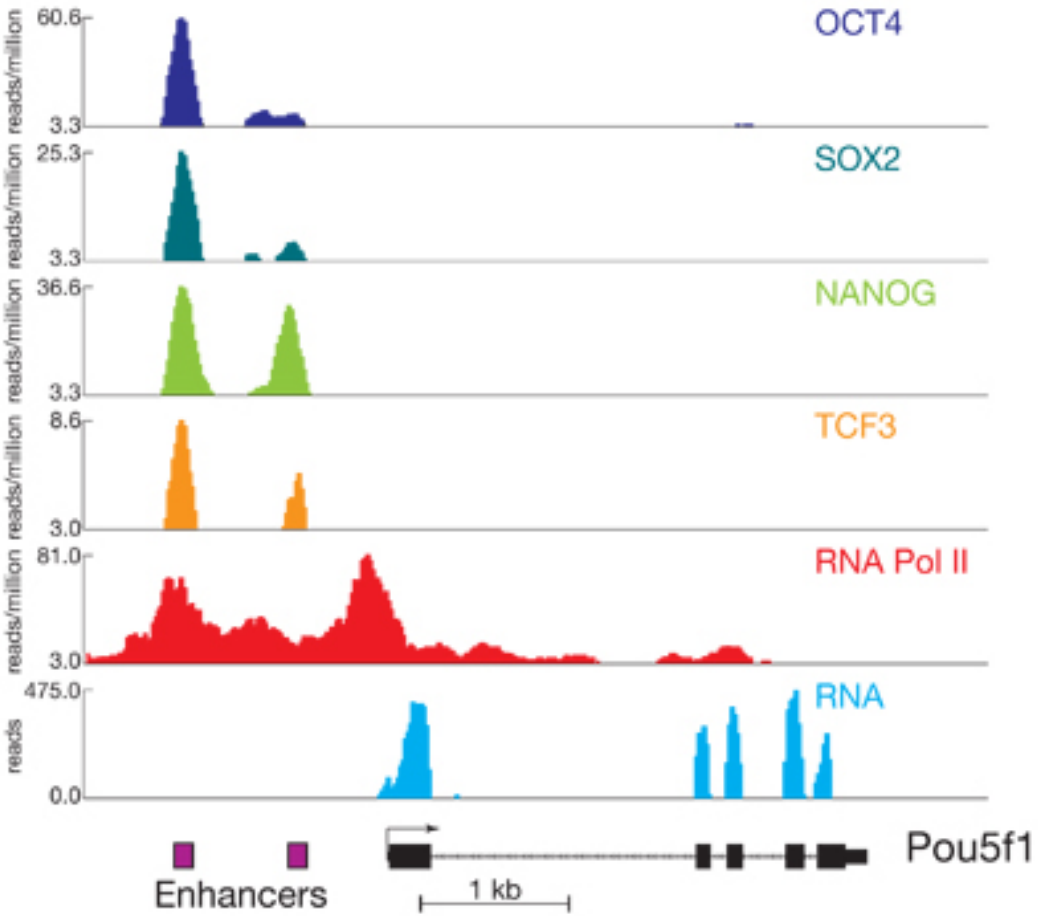
Mapping reads to reference genome

From an `eland_extended` output:

```
>HWUSI-EAS1789_0000:6:1:14464:6867#TTAGG./1 ACTGGTAGTCTGACTGTACATTGAAACATTCCTTAA
1:0:0 chr8.fa:87764854F36
>HWUSI-EAS1789_0000:6:120:18004:6530#TTAGGC/1 AAGTCTGCTCTGTGTAAAGGATCGTTCGACTCTGTG
0:2:0 chr1.fa:121185632R27A8,121186651R27A8
>HWUSI-EAS1789_0000:6:120:17588:21429#TTAGGC/1 GAATCTGAAAGTGGATATTTGGATAGCTTTGCGGAT
0:2:32 chr2.fa:91638898F31A4,chr9.fa:66565311F31A4
```

`chr9.fa:66565311F31A4` denotes a match with 1 mismatch to position 66565311 in the forward strand of chr 9. 31+A+4 = 36 bps.
As a result of mapping, we obtain the "observations/measurements" that we can do statistical inference on.

What does the aligned data look like?



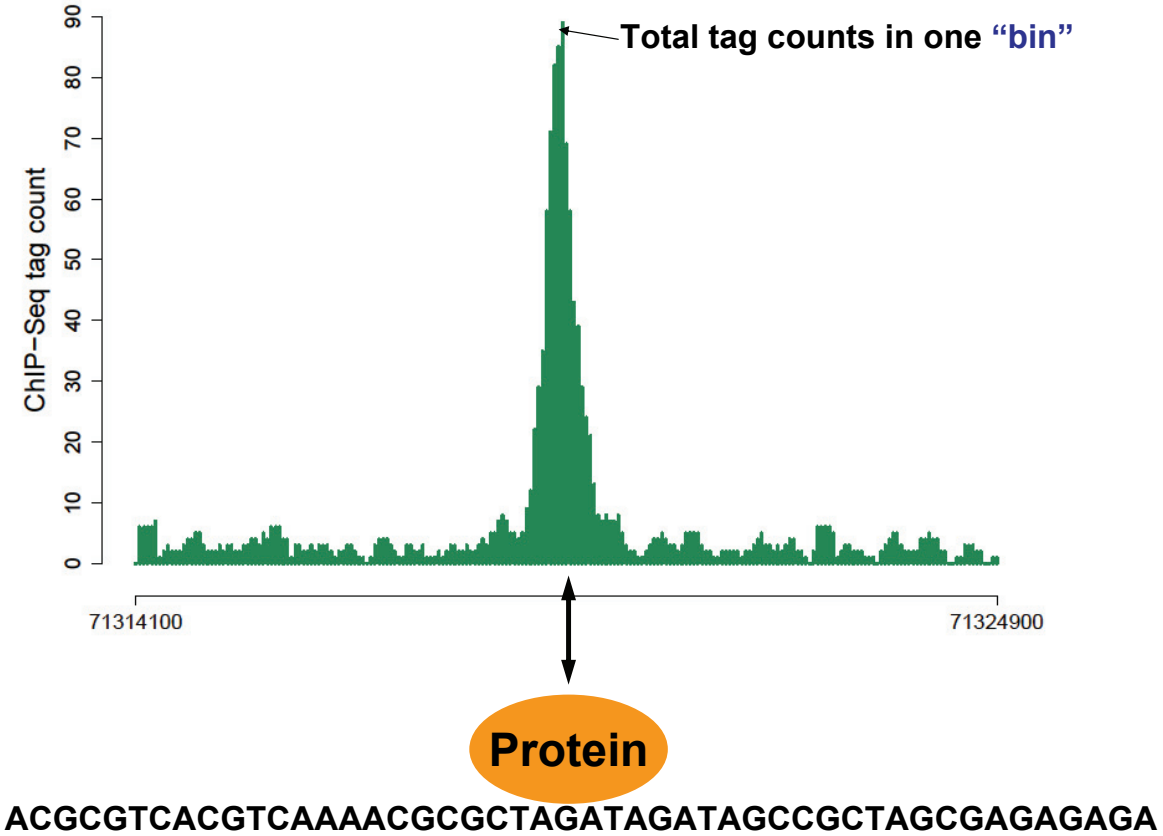
Practical challenges of dealing with Next Gen data

- Useful to know some scripting language, e.g. perl, python.
- Storage of the data is a big problem.
  - An aligned read file with 95 million reads is around 15-20GB. This is typically "one sample" for us.
  - Then for each treatment sample we typically have a control sample.
  - If we are going beyond identifying peaks, e.g., differential binding etc, we have at least 2 reps per treatment. A simple study adds up to 8 samples (or to 200 GB).
  - The good news is that, often, we can build statistical inference on data extracted from the aligned files.
  - In our work, we call these bin files. We partition the genome into non-overlapping intervals of 200 bps and count the number of reads falling into each interval – reduces size to 200MB.

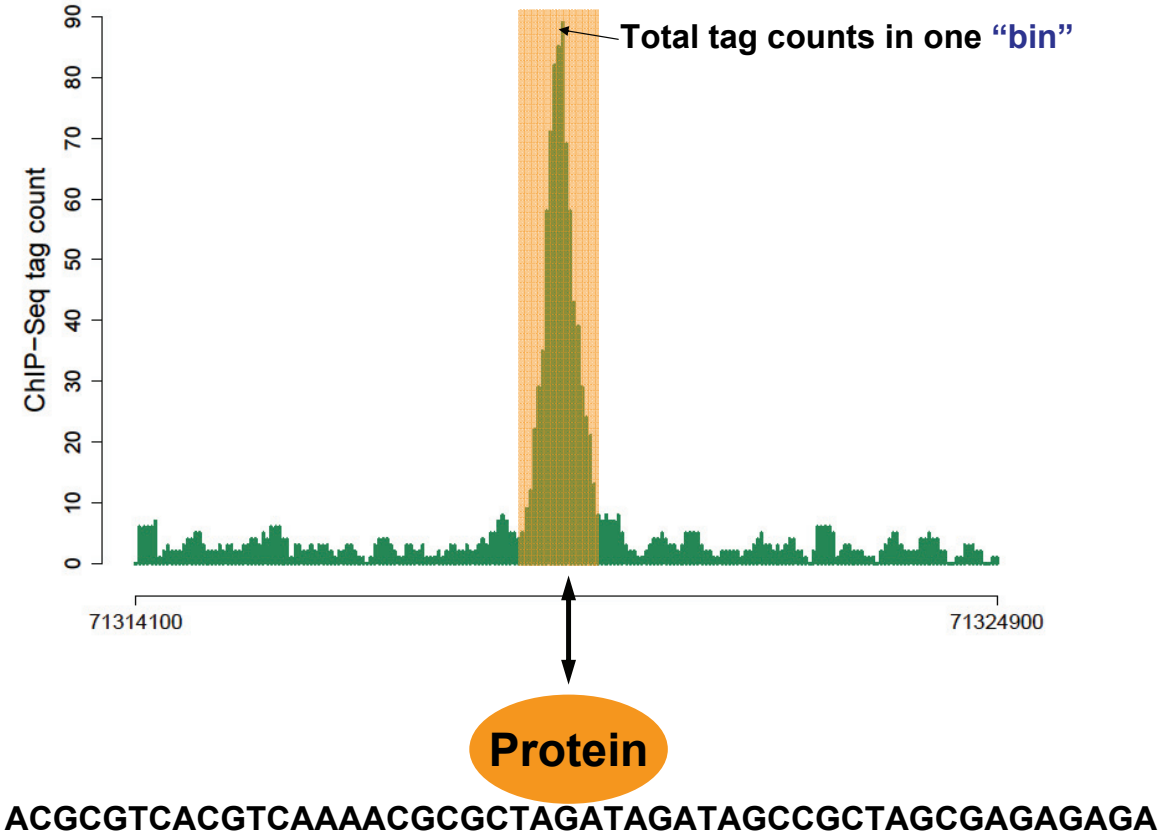
```
chr5 152001400 8
chr5 152001600 12
chr5 152001800 20
chr5 152002000 20
chr5 152002200 13
chr5 152002400 6
```

ChIP-Seq data structure

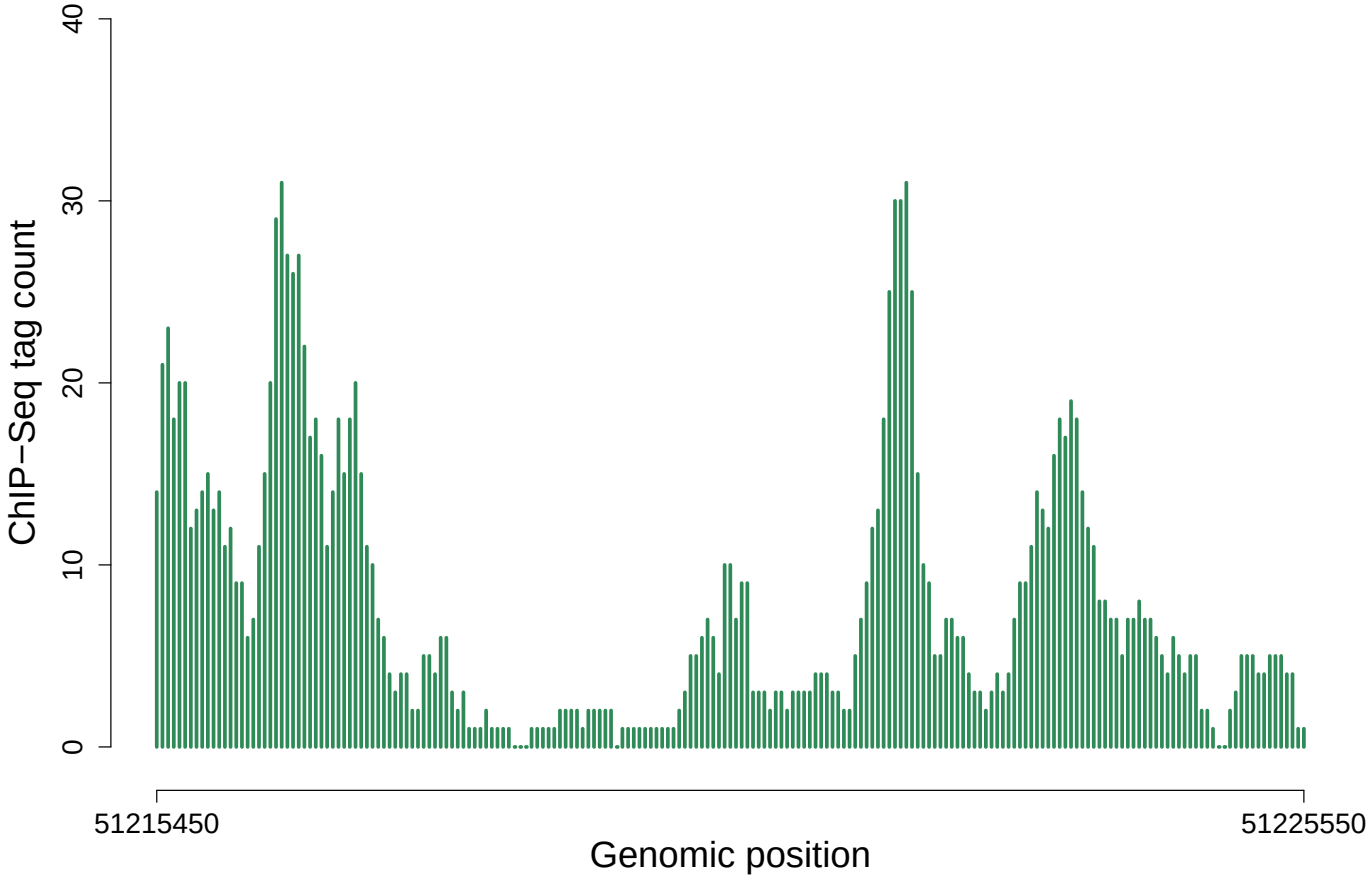
ChIP-Seq data structure



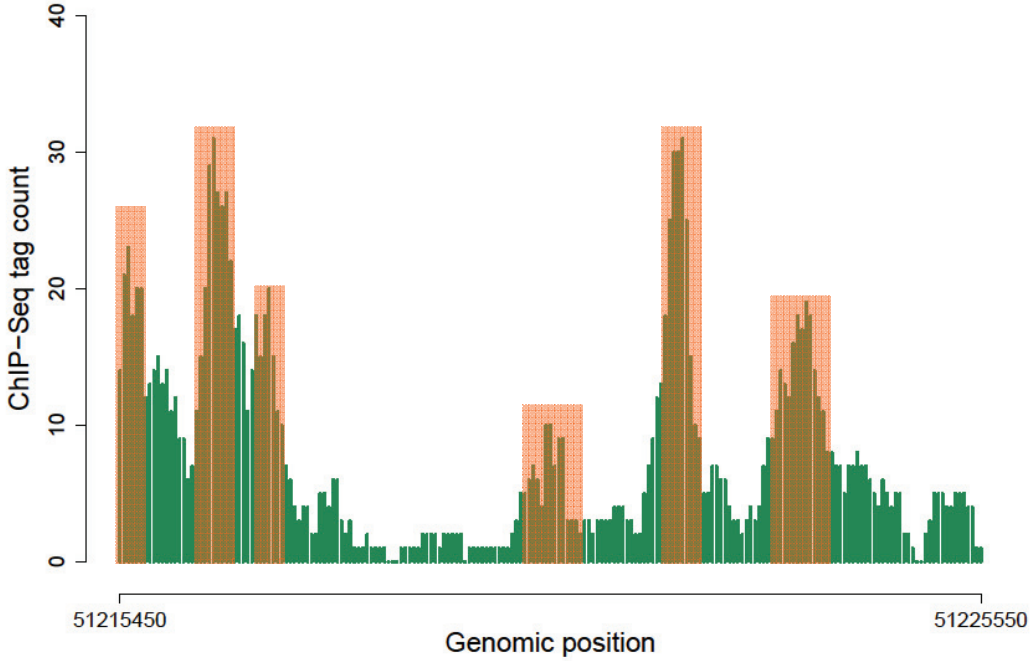
ChIP-Seq data structure



ChIP-Seq data structure



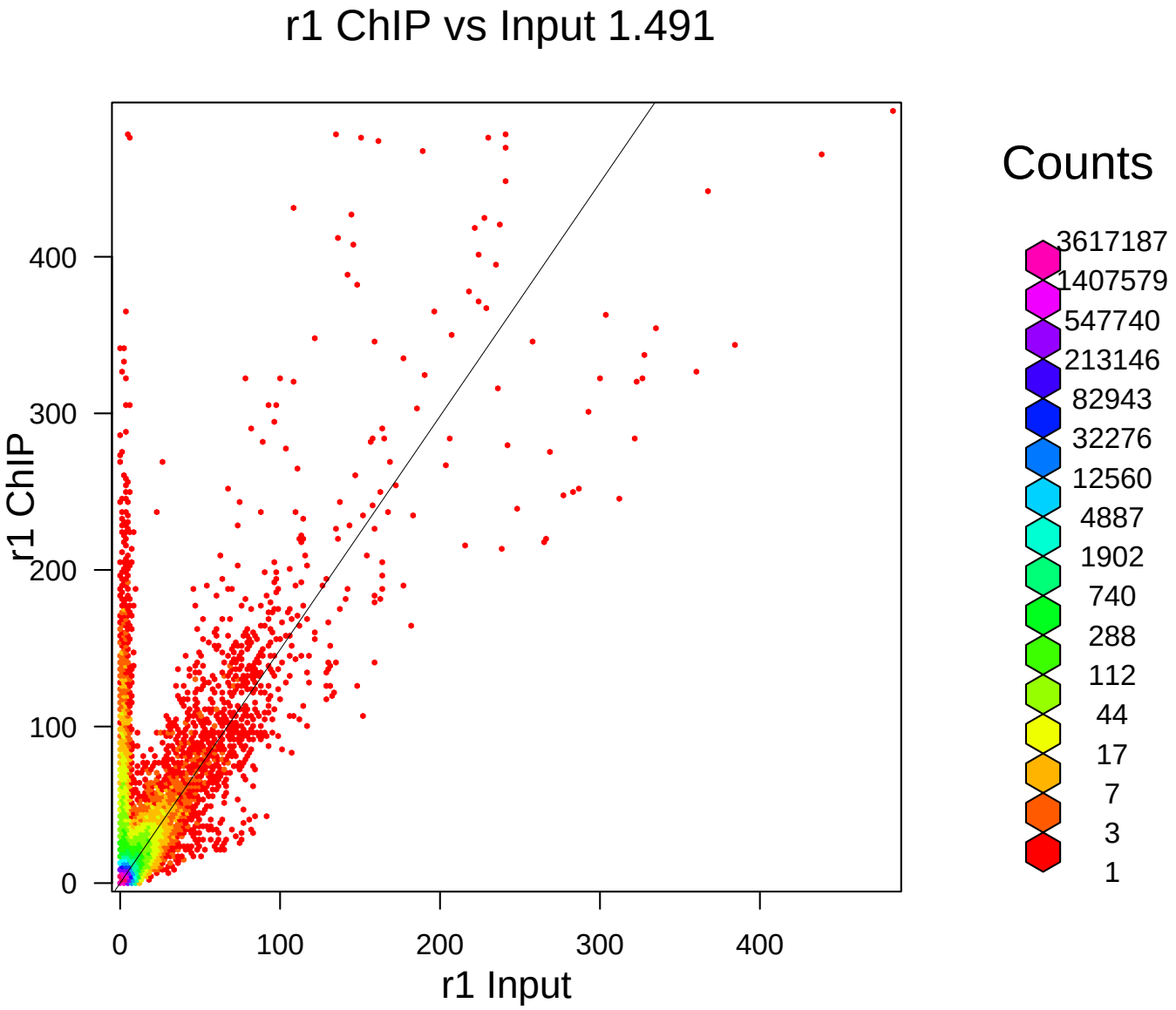
ChIP-Seq data structure



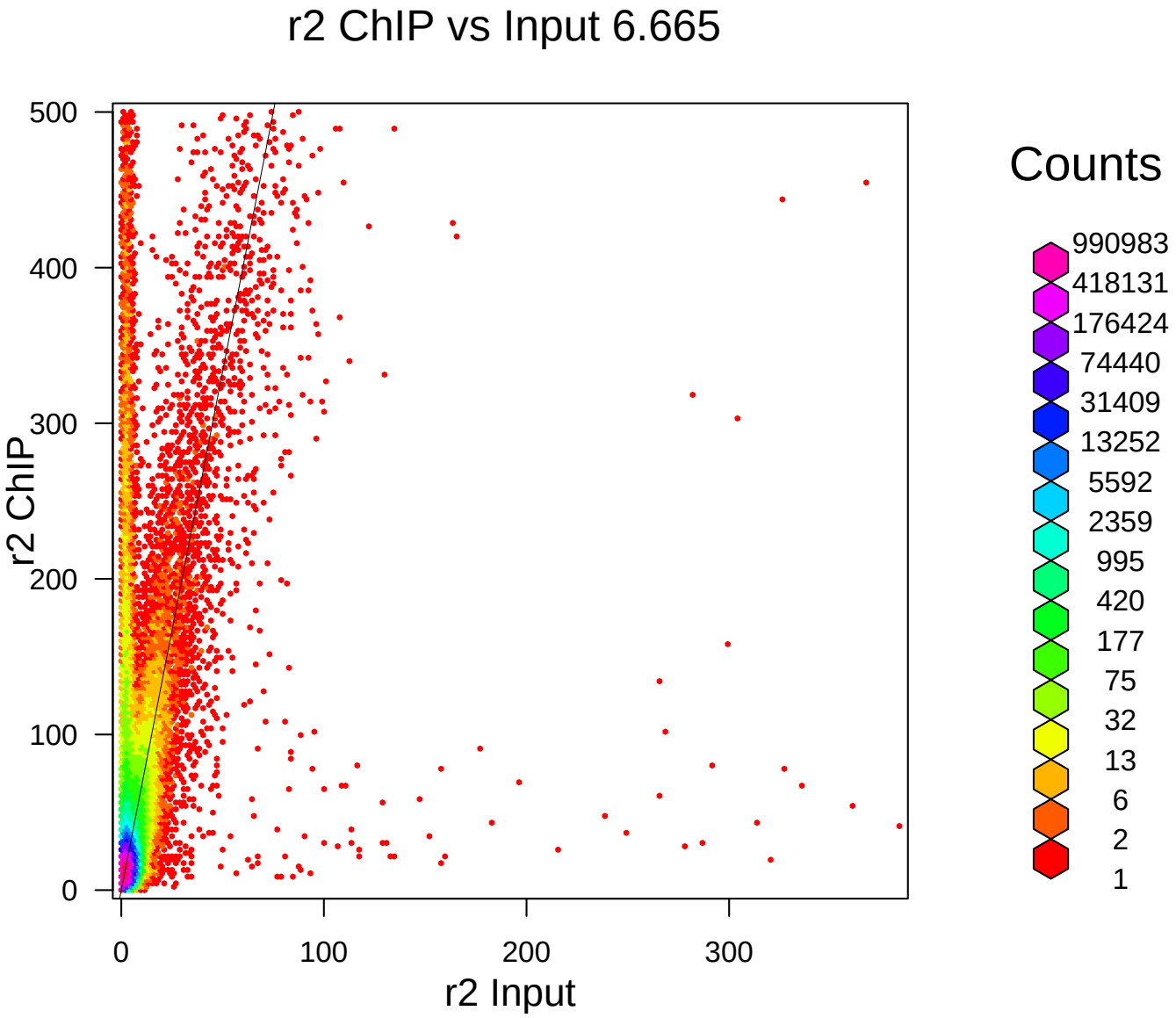
Which of these are real peaks?

How would the data look like under the null distribution? $\Longrightarrow$ Same threshold for all the peaks? Same null distribution along the genome?

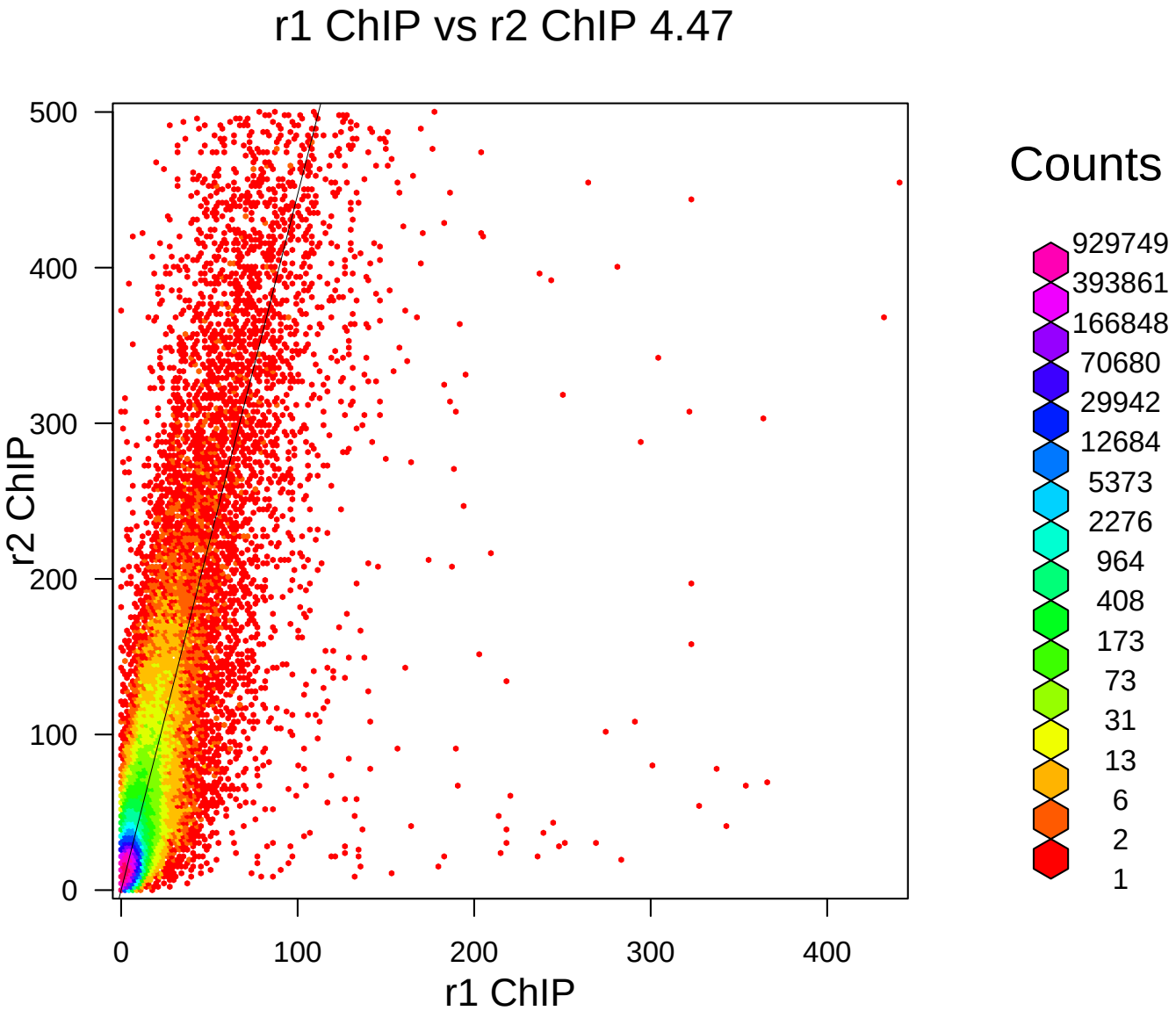
A key plot for ChIP-seq data



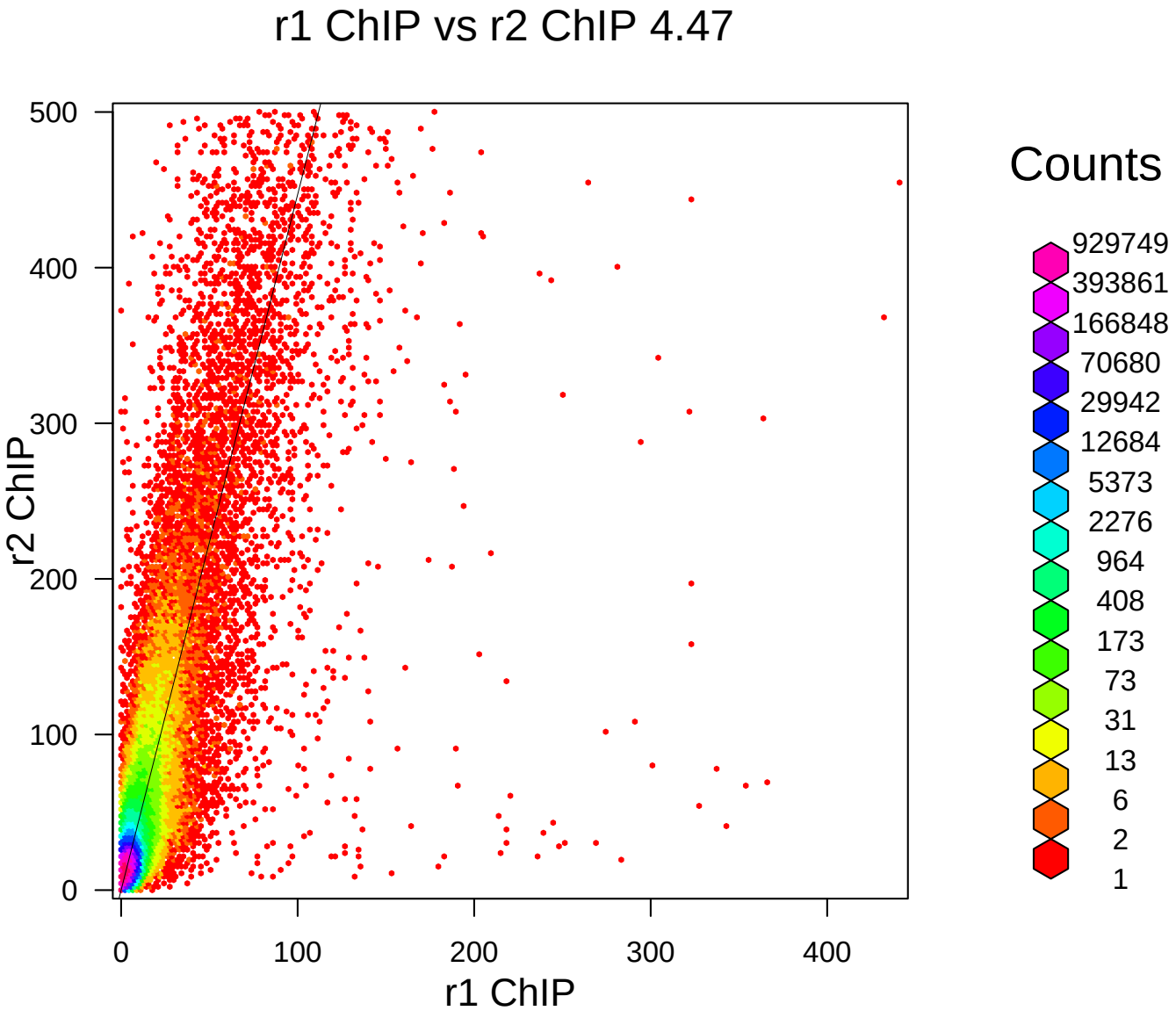
A key plot for ChIP-seq data



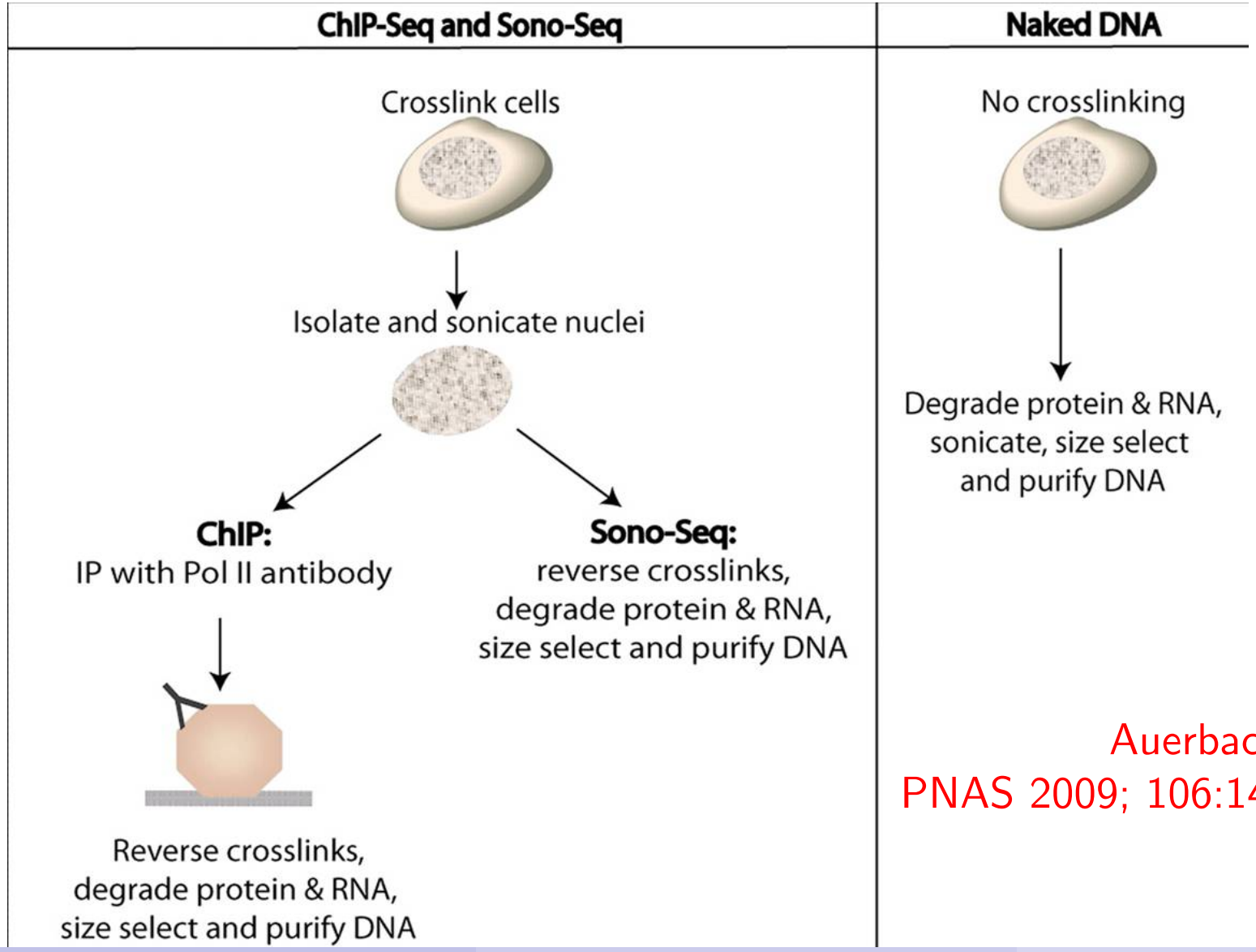
A key plot for ChIP-seq data



A key plot for ChIP-seq data

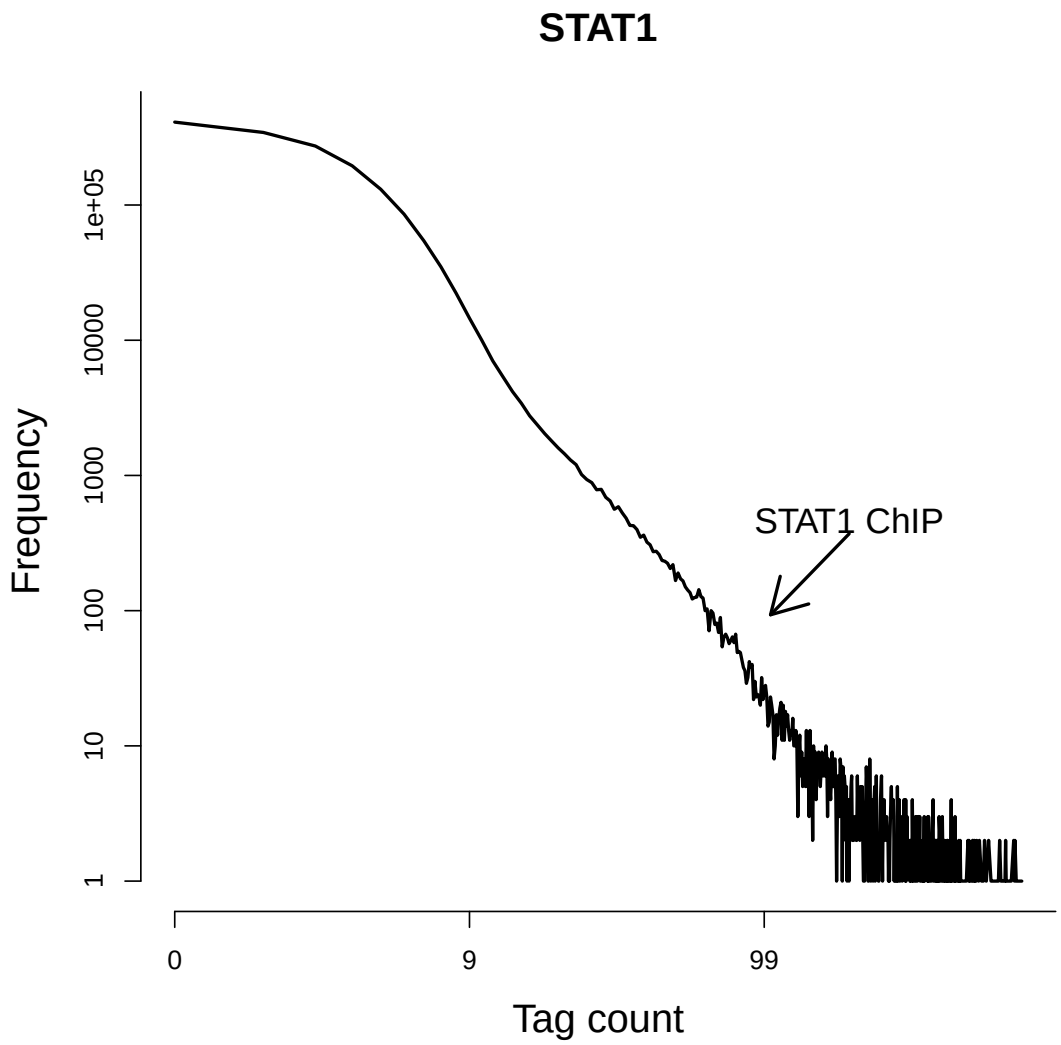


ChIP-Seq vs. Control experiments: Sono-Seq (Input-Seq), Naked-DNA-Seq

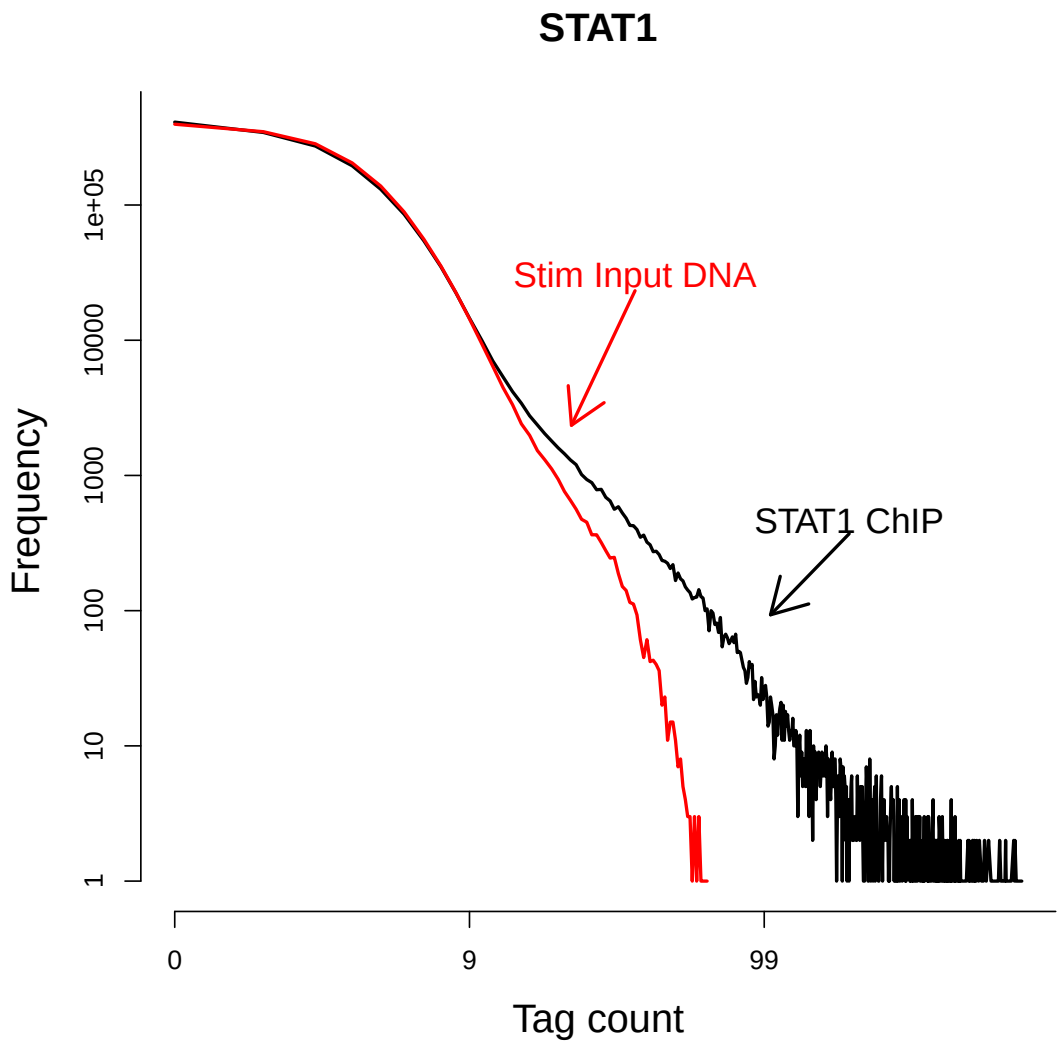


Auerbach et al.
PNAS 2009; 106:14926-14931.

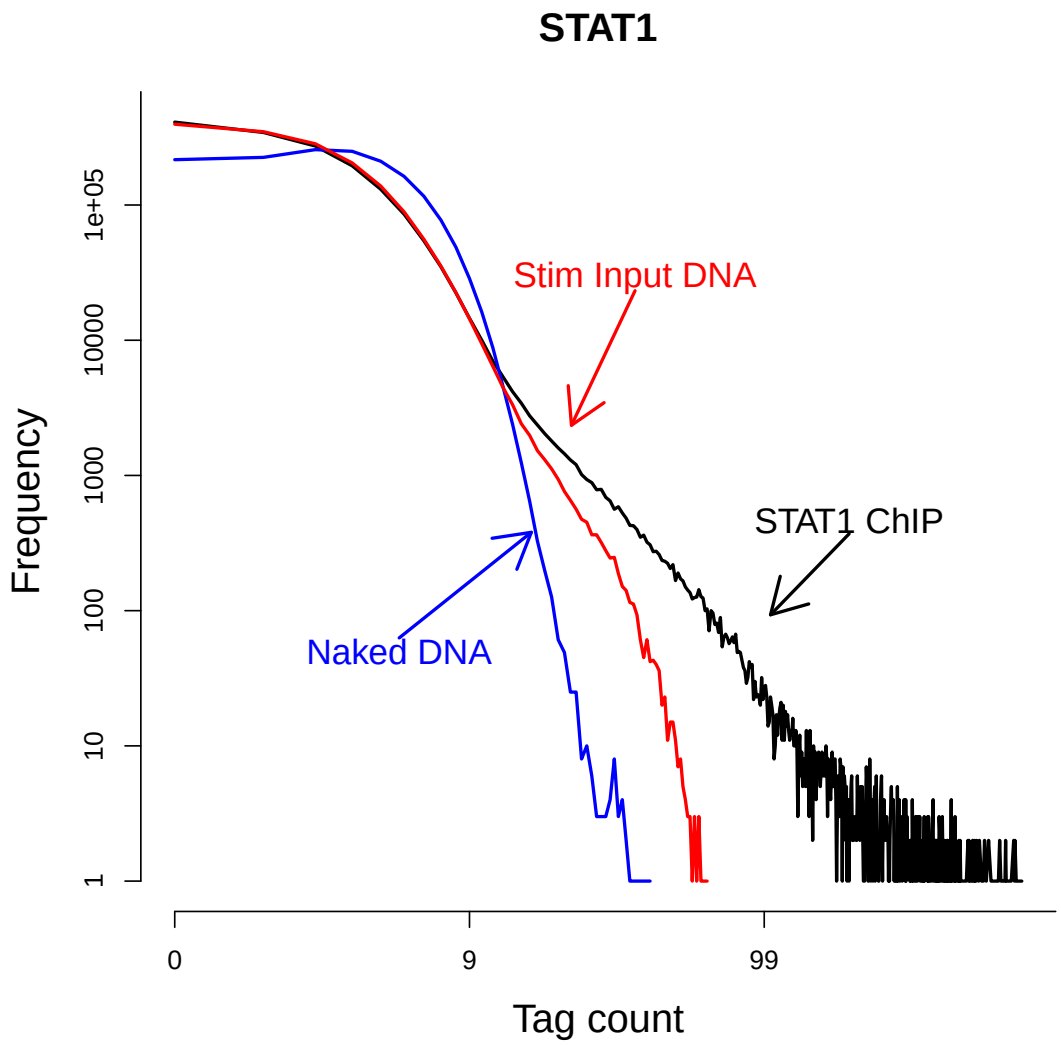
ChIP-Seq, Sono-Seq, Naked-DNA-Seq



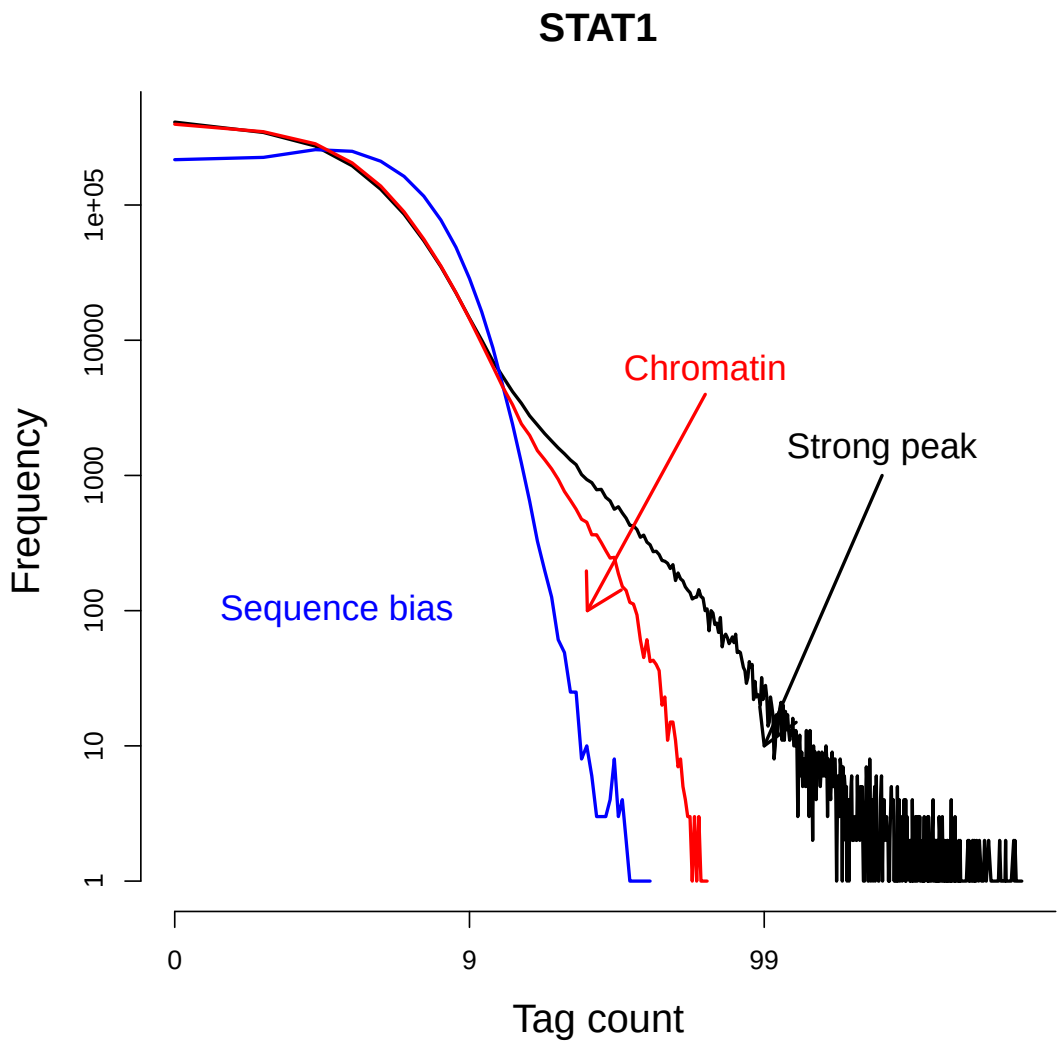
ChIP-Seq, Sono-Seq, Naked-DNA-Seq



ChIP-Seq, Sono-Seq, Naked-DNA-Seq



ChIP-Seq, Sono-Seq, Naked-DNA-Seq



Mappability and *GC* biases in ChIP-Seq data

- Mappability bias: due to retaining only uniquely aligning tags. Rozowsky *et al.* (2009).

  79.6% of the human genome is uniquely mappable using 30bp tags. (91.1% for 75bp tags).

- GC bias: Dohm *et al.* (2008), Vega *et al.* (2009).

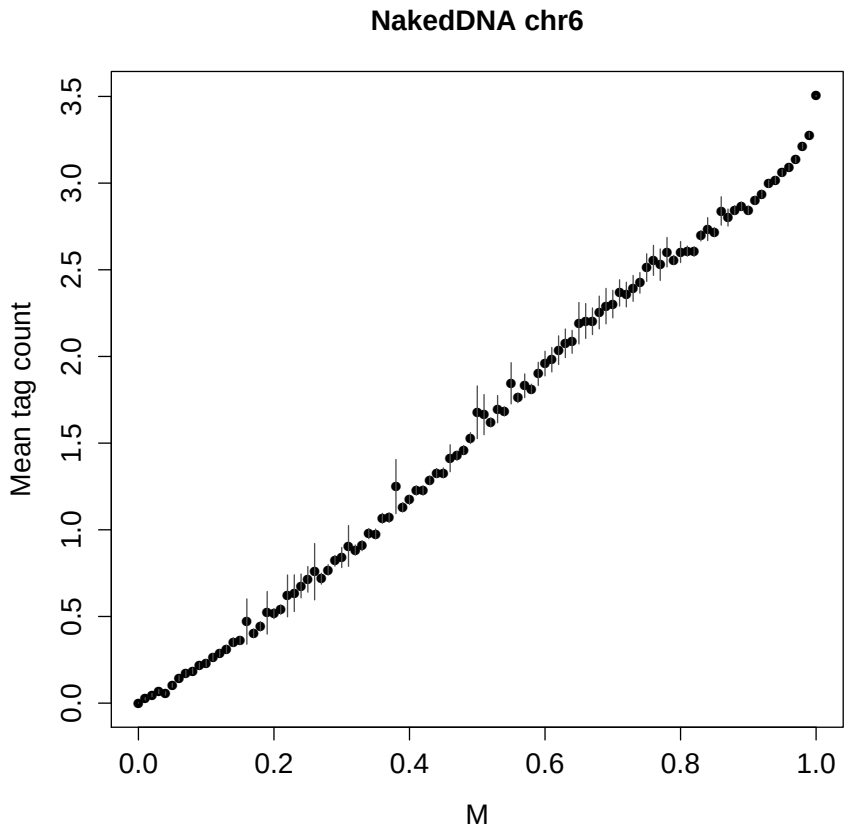
> Background for ChIP-Seq data is not uniform $\Longrightarrow$ need for region/location specific cut-offs for calling peaks.

> Naked DNA sequencing data provides an excellent platform to <u>investigate</u> and <u>model</u> these effects.

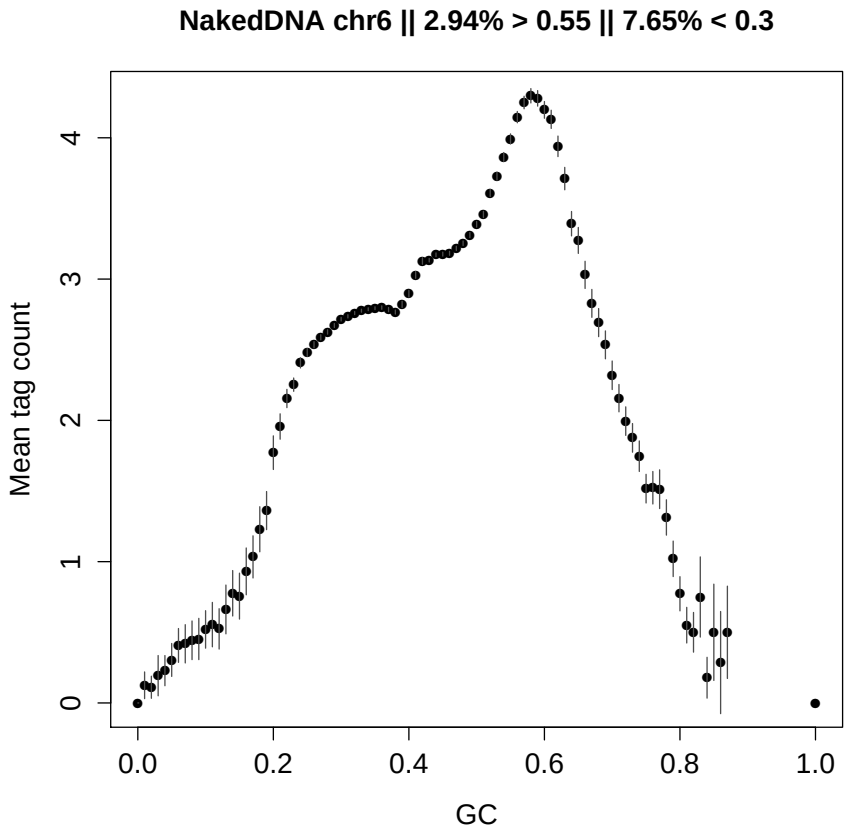
Sequence bias in ChIP-Seq data: Mean tag count vs Mappability

(bin-level) and GC content: HeLa S3 Naked-DNA-Seq (GSE14022)

Mappability



GC

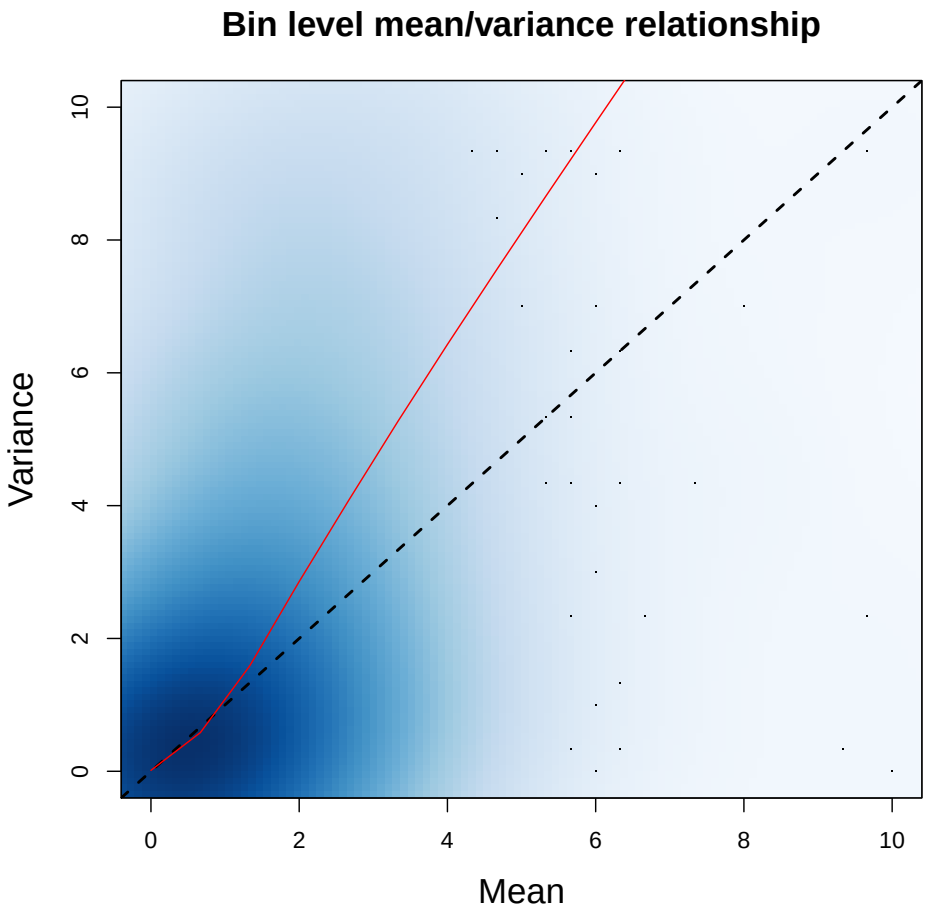


M-GC

Same relationships hold for data from mouse. Mouse

Background/Null model for one-sample ChIP-Seq data

- Existing methods:
  - Poisson distribution.
  - Negative binomial distribution (CisGenome, Ji et al. (2008)).
- Excess zeroes and over-dispersion.
- Mappability and GC bias $\Longrightarrow$ bin specific distributions.



Poisson vs. Neative Binomial Distributions

$$\begin{aligned} Y &\sim \text{Poisson}(\lambda), \\ &\Longrightarrow E[Y] = \lambda, \text{var}[Y] = \lambda. \end{aligned}$$

a: shape, b:scale.

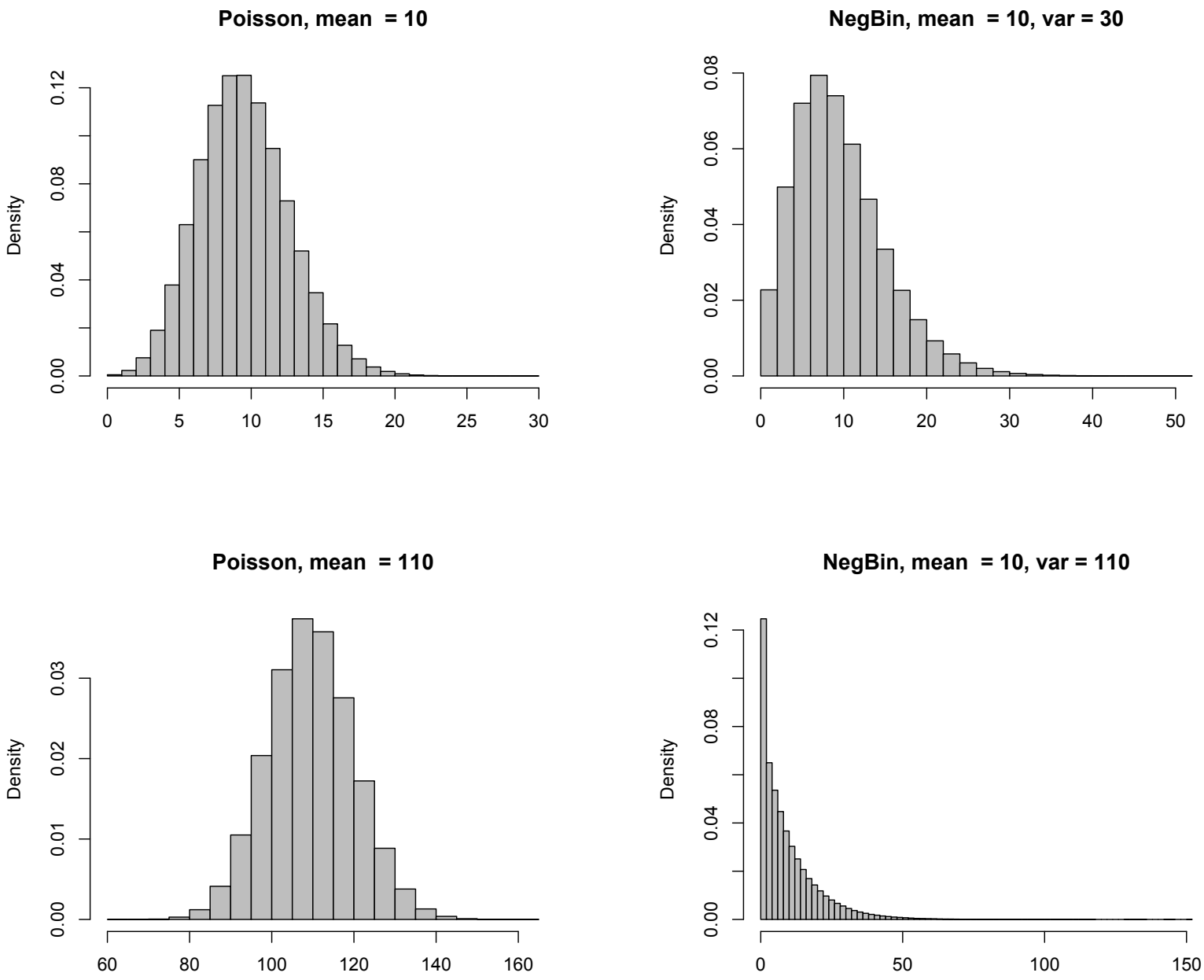
$$\begin{aligned} Y &= \text{NegBin}(a, b) \\ &\Longrightarrow E[Y] = a/b, \text{var}[Y] = a(1+b)/b^2. \end{aligned}$$

Alternative parametrization (in `R`)

$$\begin{aligned} Y &= \text{NegBin}(\rho, \mu) \\ &\Longrightarrow E[Y] = \mu, \text{var}[Y] = \mu + \mu^2/\rho, \end{aligned}$$

where $\rho = a$, $\mu = a/b$.

$1/\rho$ is referred to as the "dispersion" parameter.


Poisson, mean = 10
Density
0.00
0.04
0.08
0.12
0
5
10
15
20
25
30
NegBin, mean = 10, var = 30
Density
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
50
Poisson, mean = 110
Density
0.00
0.01
0.02
0.03
60
80
100
120
140
160
NegBin, mean = 10, var = 110
Density
0.00
0.04
0.08
0.12
0
50
100
150

Background/Null model for one sample ChIP-Seq data

Y_j : observed tag counts for bin j.
N_j : background tag counts for the bin.
M_j : average mappability score.
GC_j : average GC content.

Non-homogeneous background

- $Y_j \sim N_j$
- $N_j | \mu_j \sim g(\mu_j)$

Candidate models for $g(\mu_j)$

1. $g(\mu_j) \sim Po(\mu_j)$ (Poi Reg)
2. $g(\mu_j) \sim NegBin(a, a/\mu_j)$ (NegBin Reg)

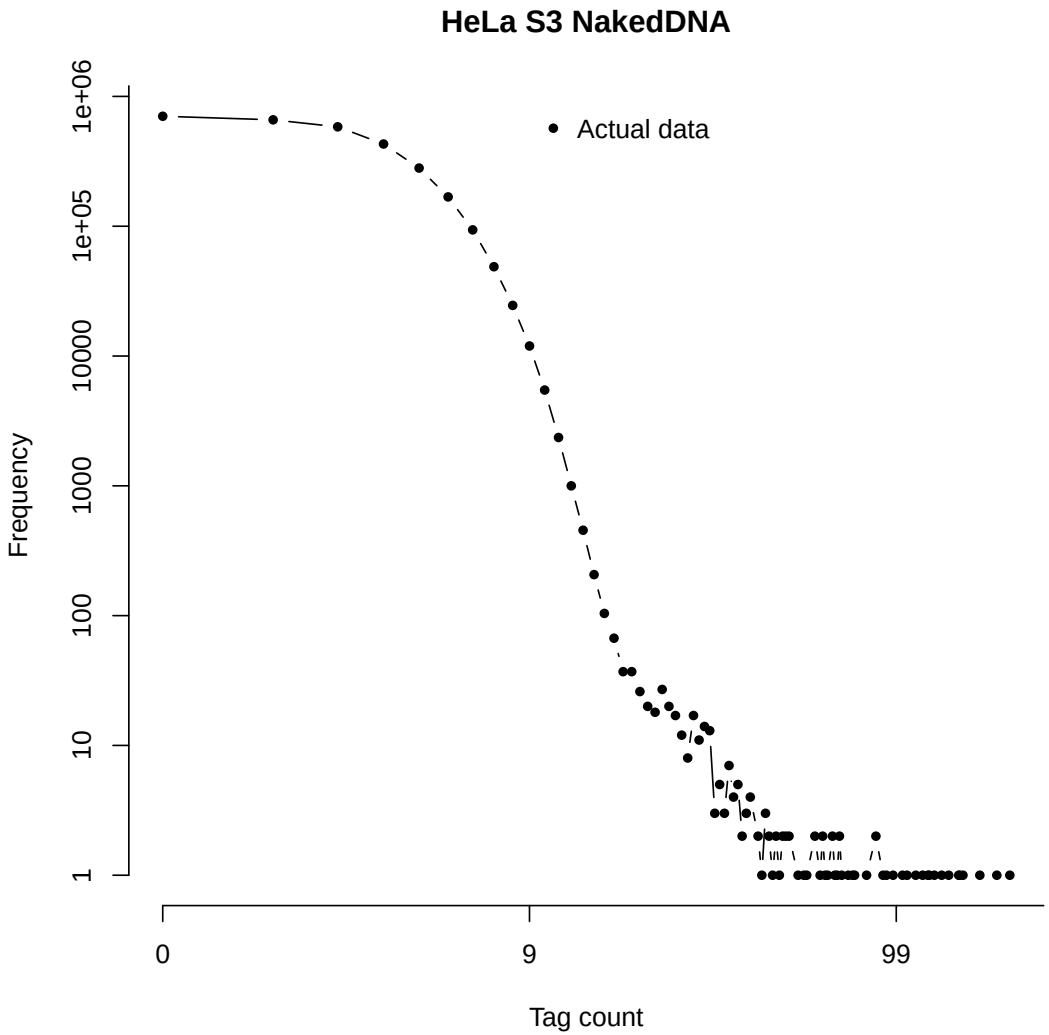
Candidate models for μ_j

1. $\mu_j = \exp(\beta_0)$
2. $\mu_j = \exp(\beta_0 + \beta_M \log_2(M_j + 1))$
3. $\mu_j = \exp(\beta_0 + \beta_{GC} GC_j)$
4. $\mu_j = \exp(\beta_0 + \beta_M \log_2(M_j + 1) + \beta_{GC} GC_j)$
5. $\mu_j = \exp(\beta_0 + \beta_{GC} Sp(GC_j))$
6. $\mu_j = \exp(\beta_0 + \beta_M \log_2(M_j + 1) + \beta_{GC} Sp(GC_j))$

CisGenome (Ji et al. (2009))

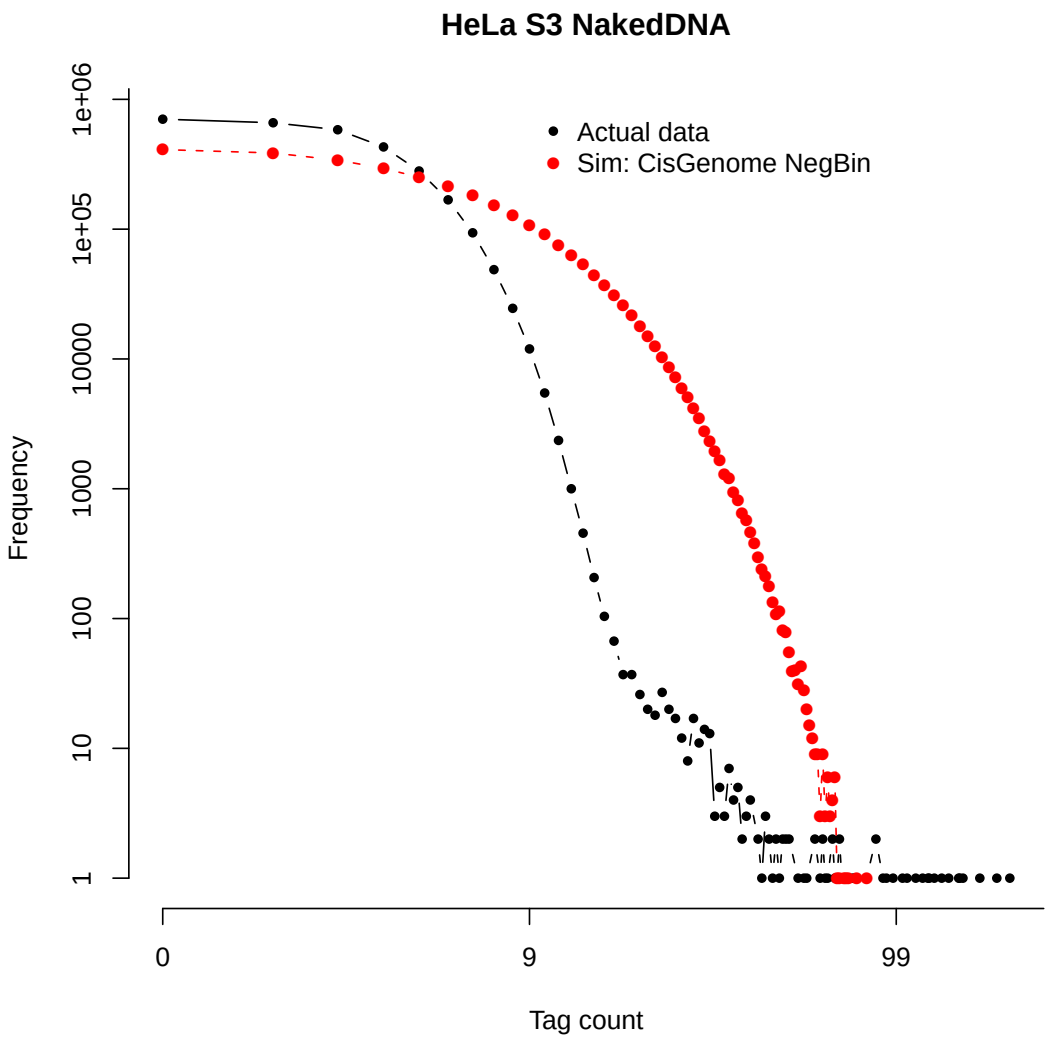
- $Y_j \sim NegBin(a, b)$

Goodness of Fit



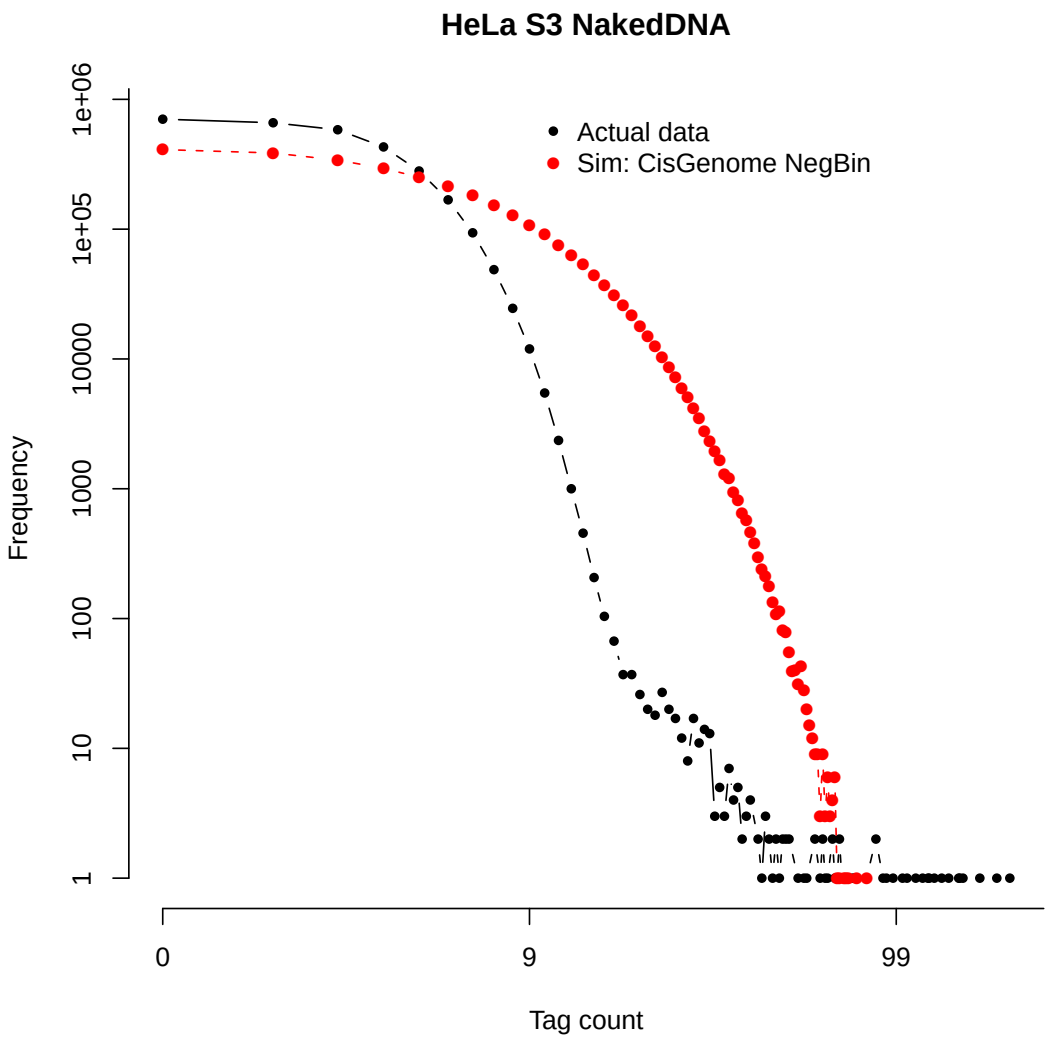
Black line: Actual data

Goodness of Fit: CisGenome NegBin



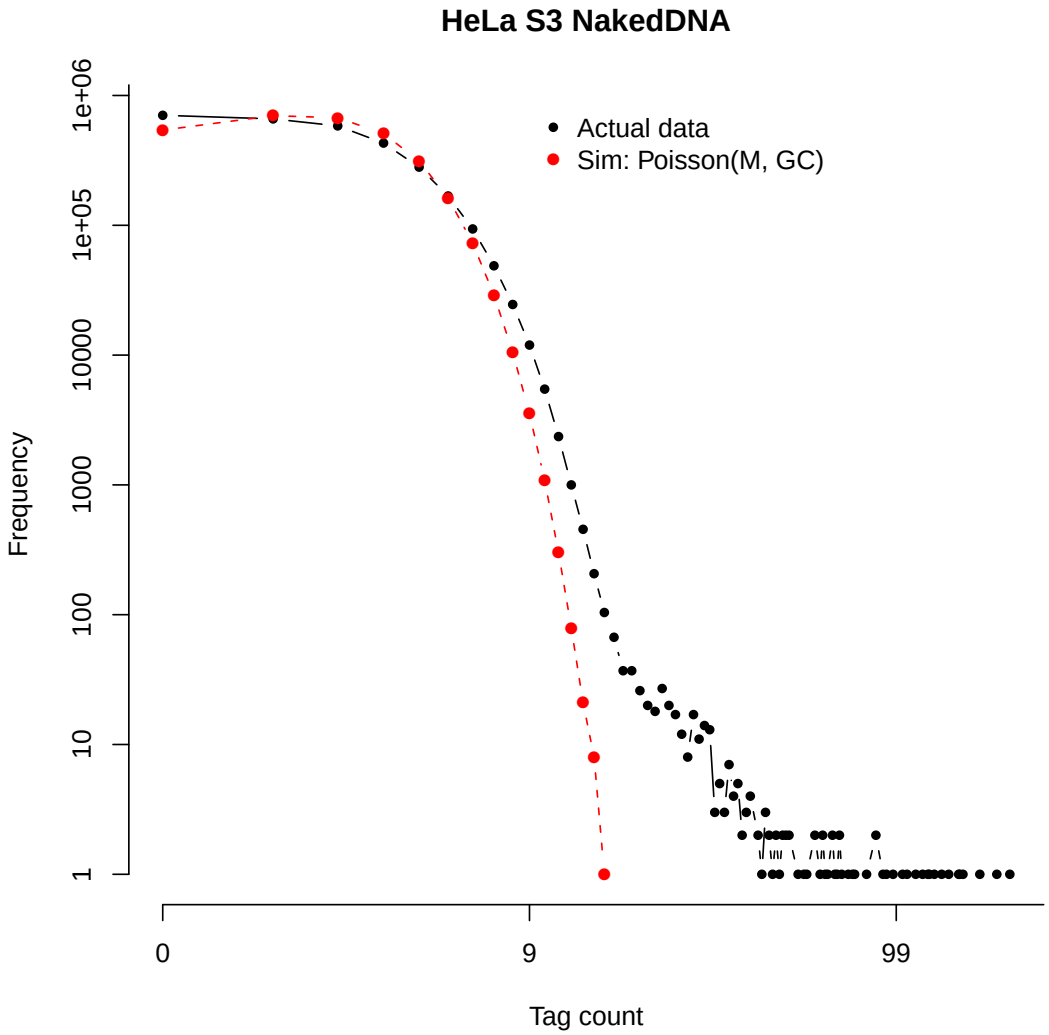
Black line: Actual data Red line: Simulated data from the fitted CisGenome

Goodness of Fit: CisGenome NegBin



Over-estimated background!

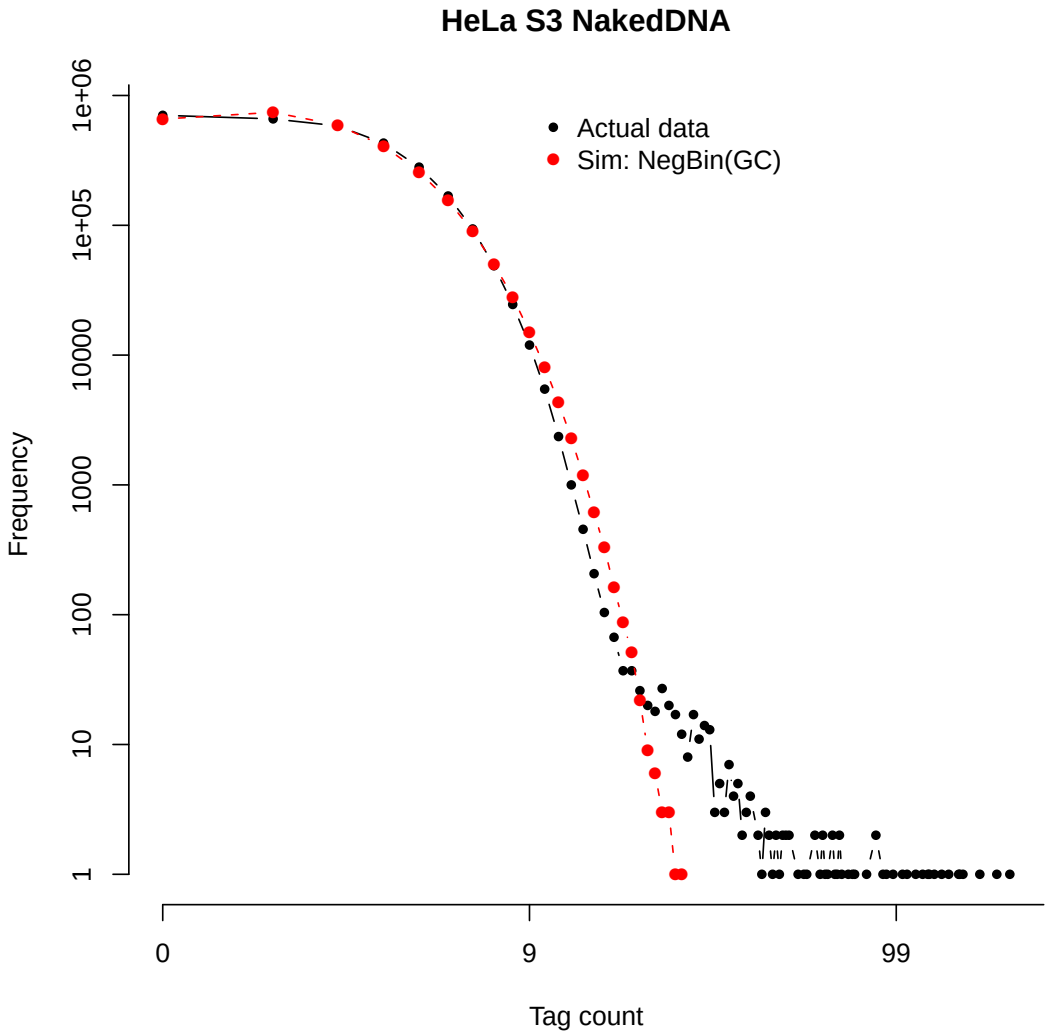
Goodness of Fit: Poisson Reg (M, GC)



Black line: Actual data

Red line: Simulated data from the fitted Poisson Reg(M, GC)

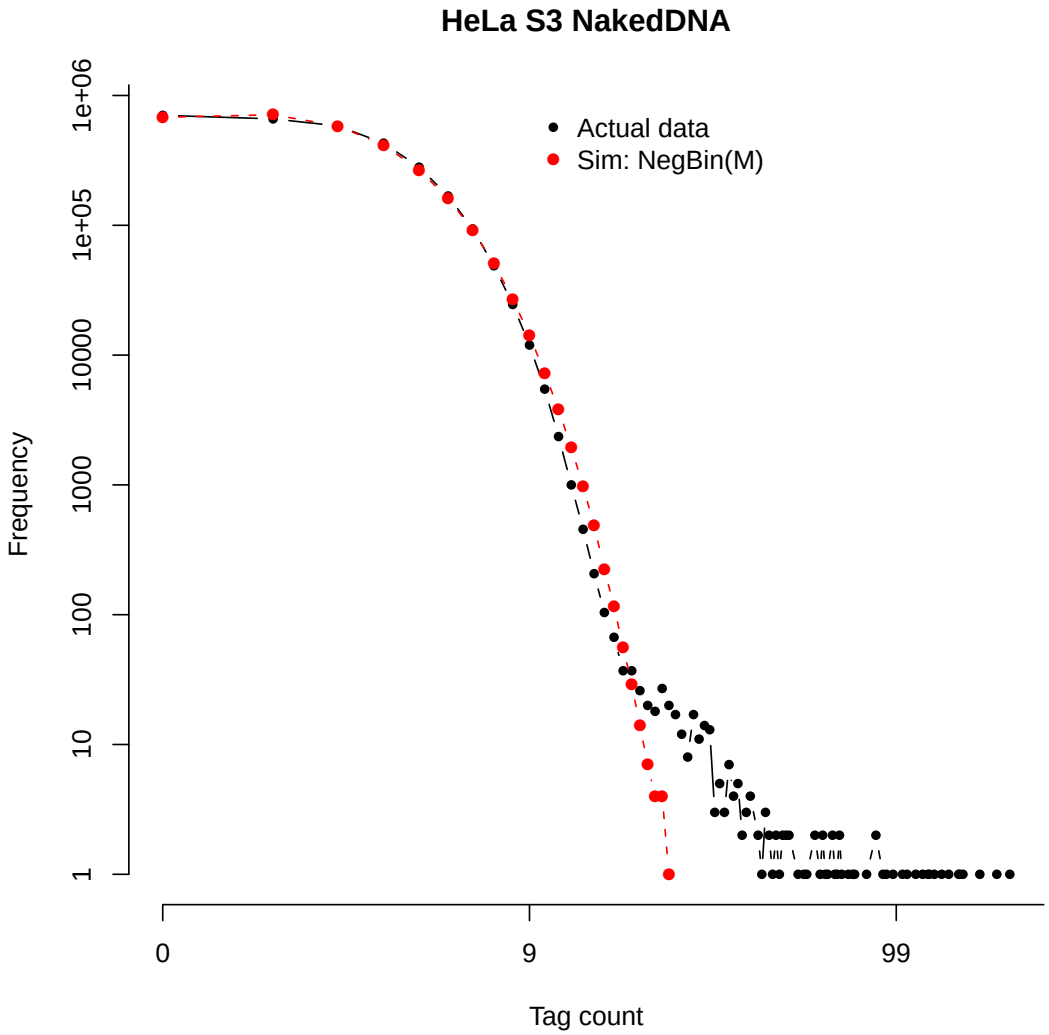
Goodness of Fit: NegBin Reg (GC)



Black line: Actual data

Red line: Simulated data from the fitted NegBin Reg (GC)

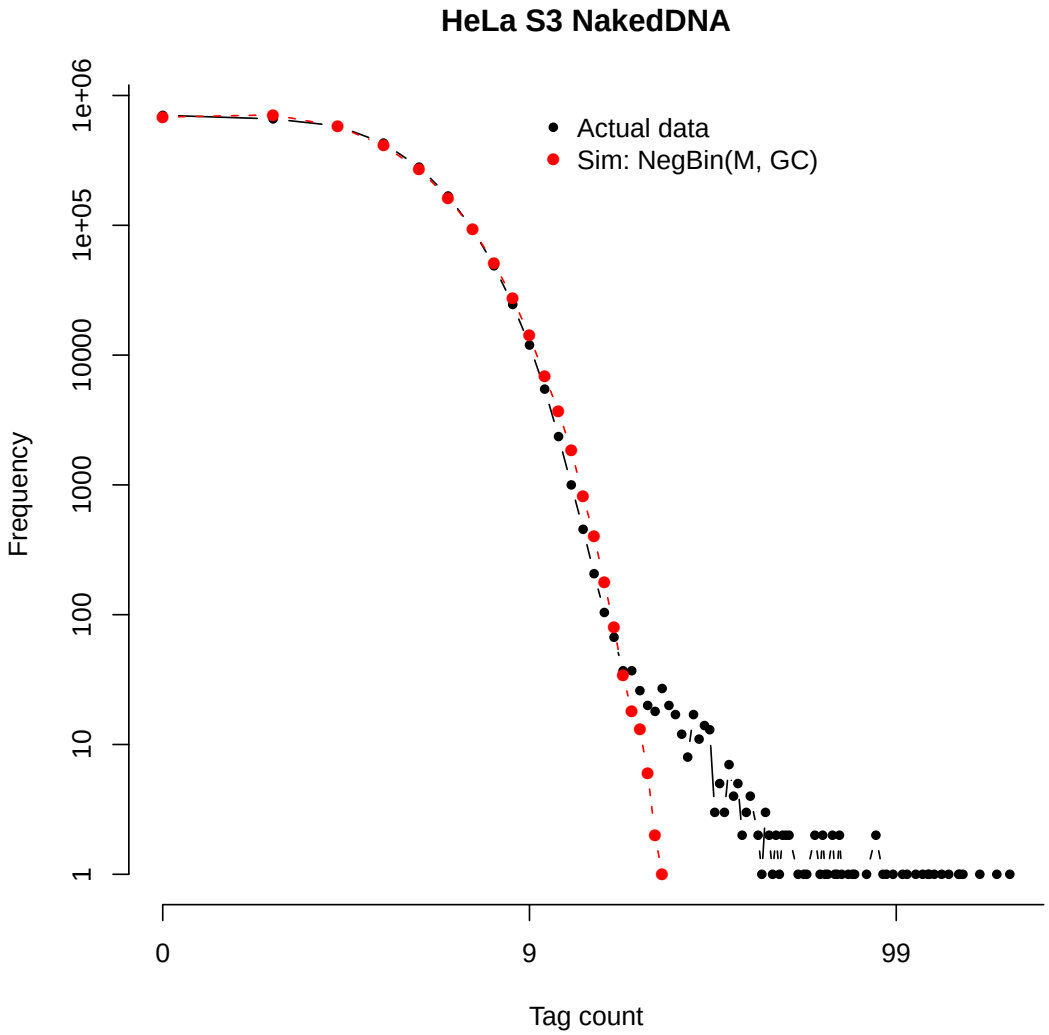
Goodness of Fit: NegBin Reg (M)



Black line: Actual data

Red line: Simulated data from the fitted NegBin Reg (M)

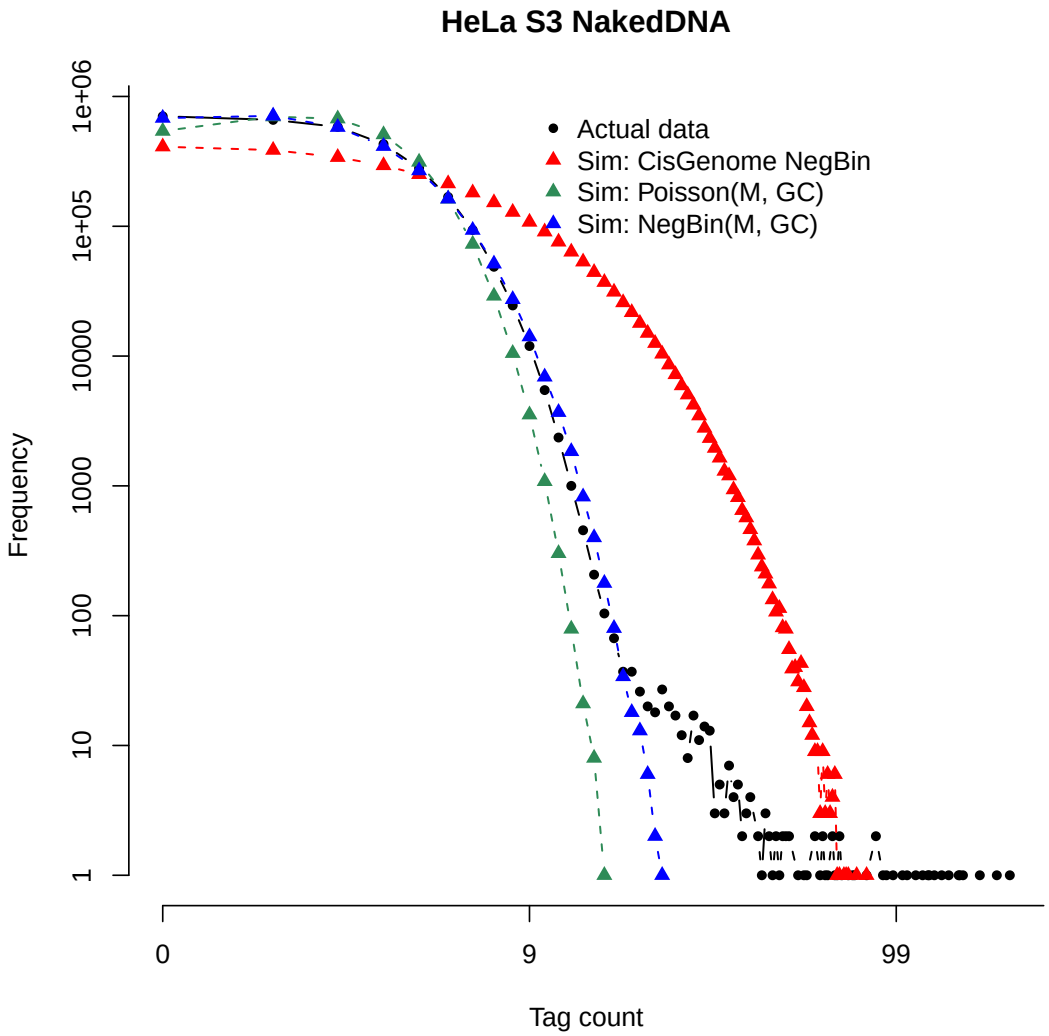
Goodness of Fit: NegBin Reg (M, GC)



Black line: Actual data

Red line: Simulated data from the fitted NegBin Reg (M, GC)

Goodness of Fit



Bayesian Information Criterion (BIC) for model selection (smaller better):

None$> GC > M > M + GC > M + Sp(GC)$.

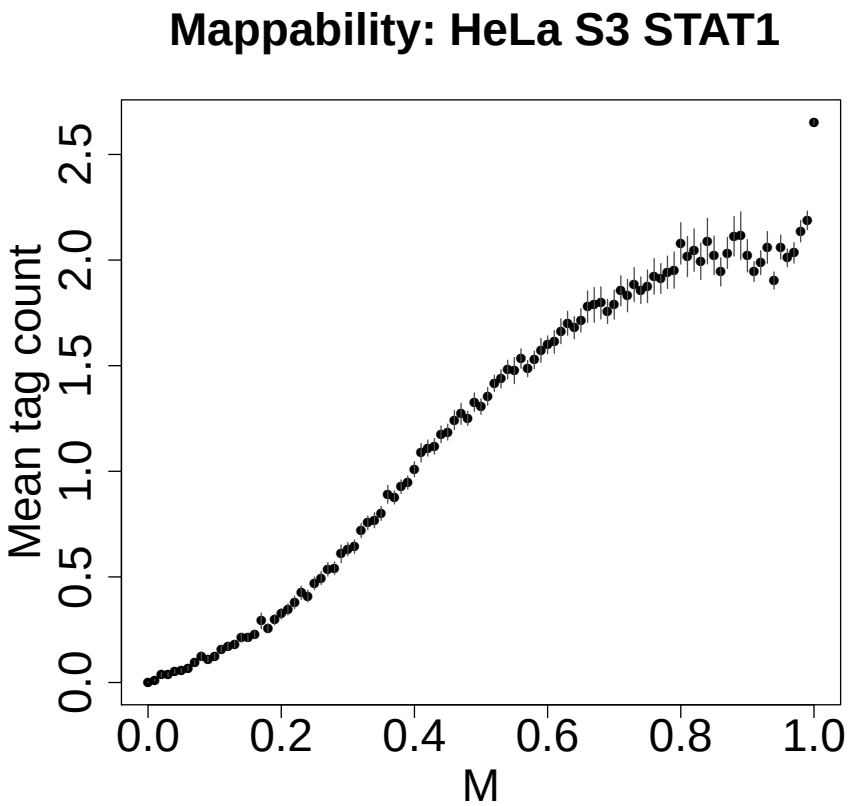
Mixture model for one-sample ChIP-Seq data

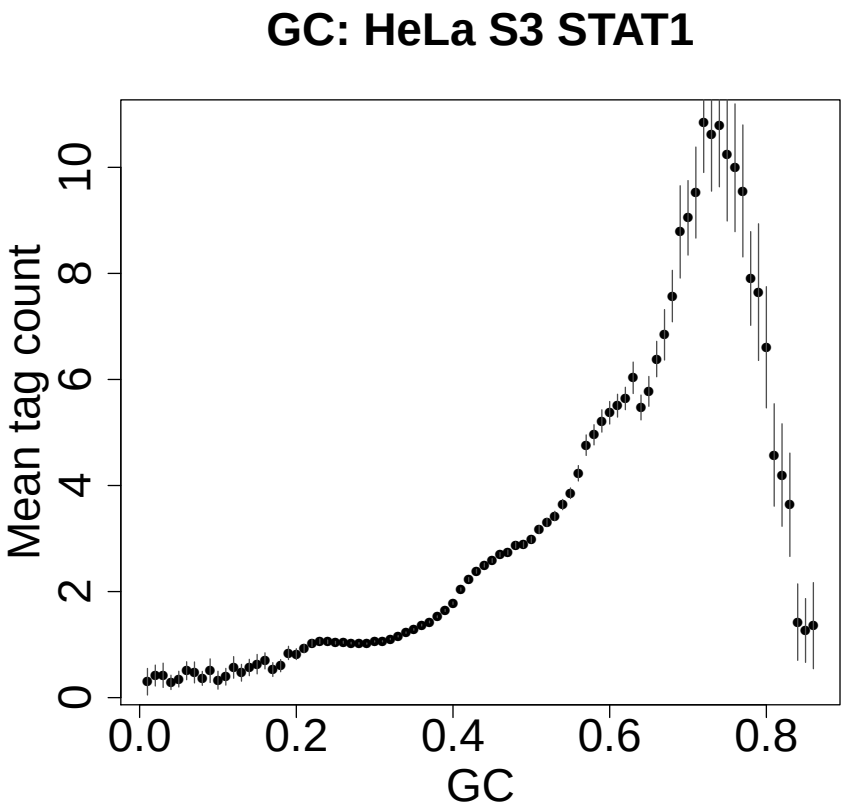
- Z_j: unknown/latent state, $Z_j = 1(0)$ if bound (unbound).
- $Y_j | Z_j = 0 \sim N_j$ NegBin Reg(M_j, GC_j)
- $Y_j | Z_j = 1 \sim N_j + S_j$
- S_j: protein-binding signal
  1. $S_j \sim NegBin(b_1, c_1)$ (1-component)
  2. $S_j \sim p_1 NegBin(b_1, c_1) + (1 - p_1) NegBin(b_2, c_2)$ (2-component)

Estimate unknown parameters with maximum likelihood method using the EM algorithm.

Case Study 1: STAT1 ChIP-Seq data

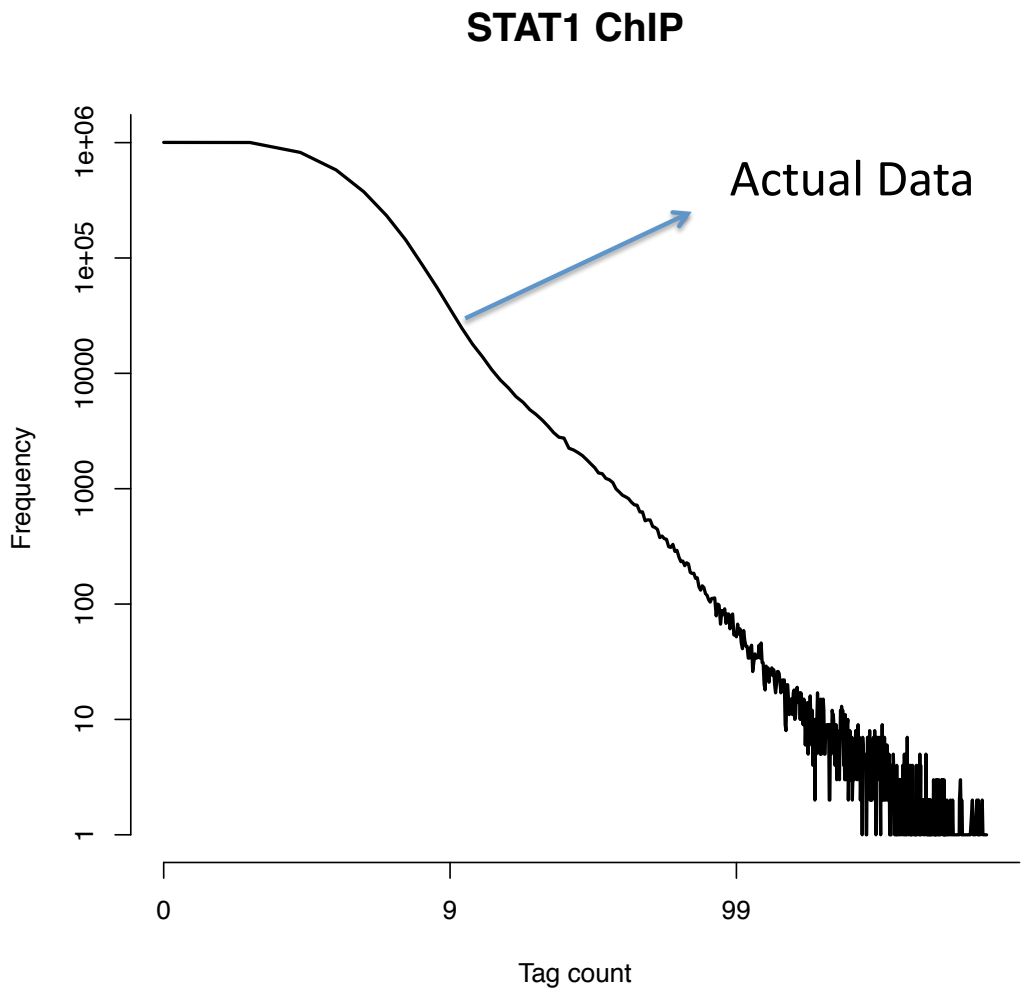
- STAT1 in IFN-γ-stimulated HeLa S3 ChIP-Seq data (GSE12782).
- 6 lanes of Illumina sequencing data, 23 million mapped reads.



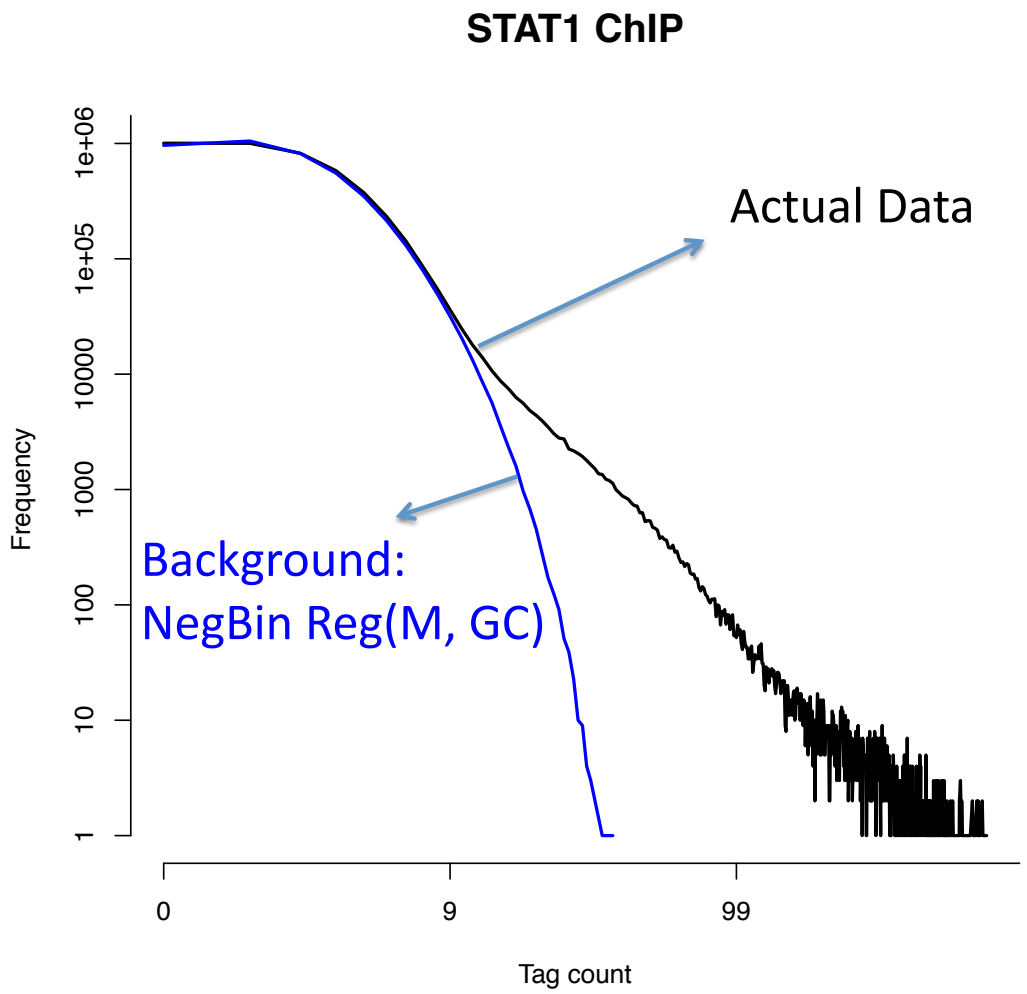


Case Study 2: GATA1 Sono-Seq

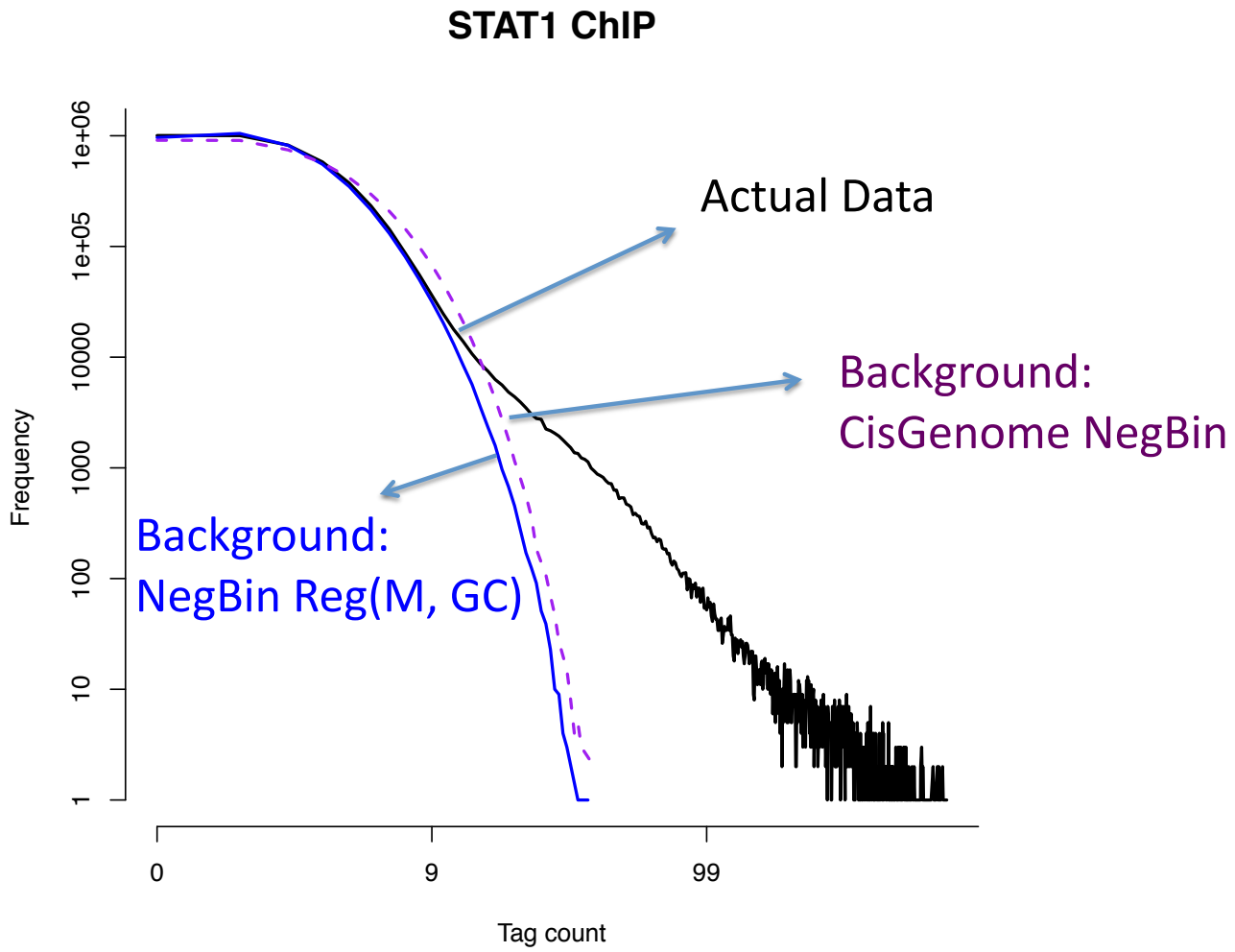
Case Study 1: STAT1 ChIP-Seq data



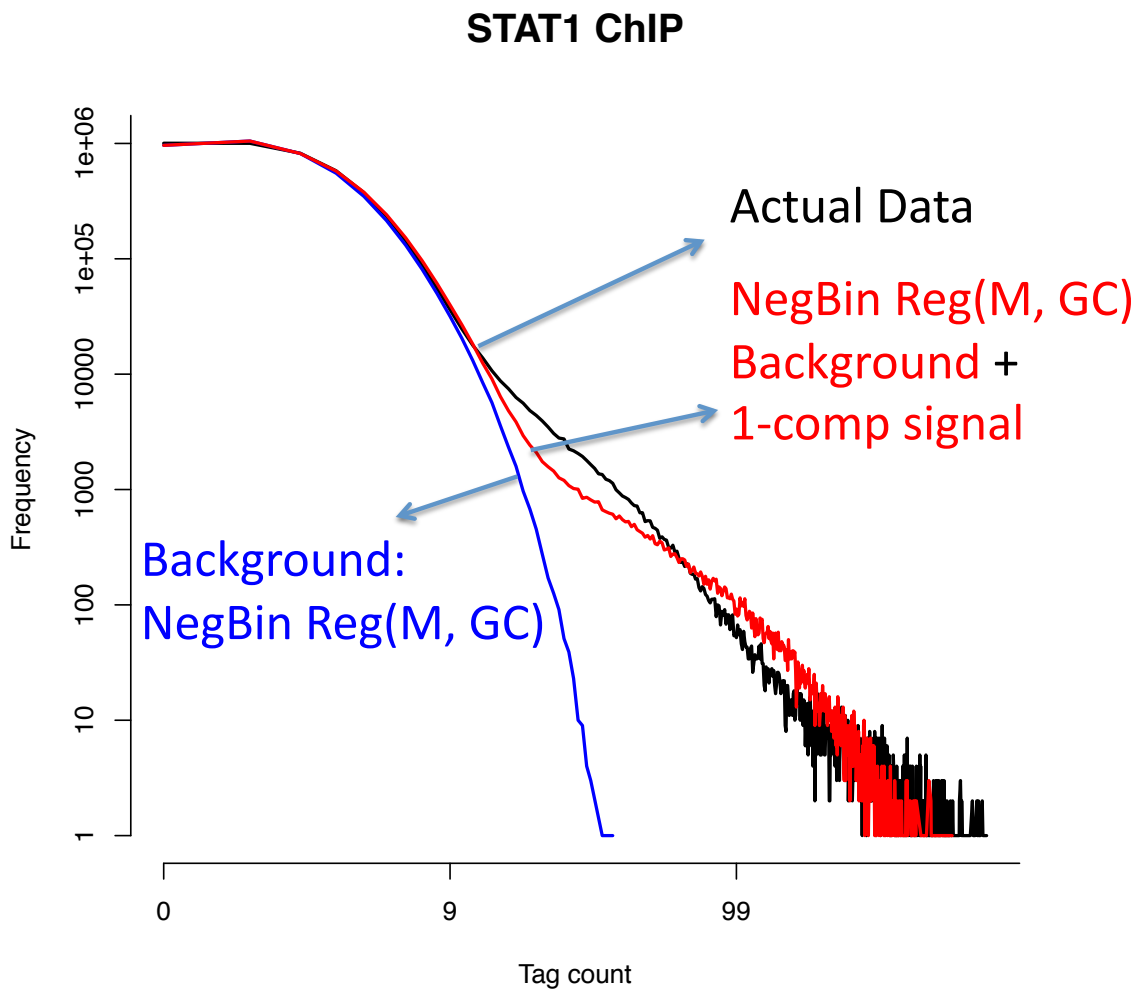
Case Study 1: STAT1 ChIP-Seq data



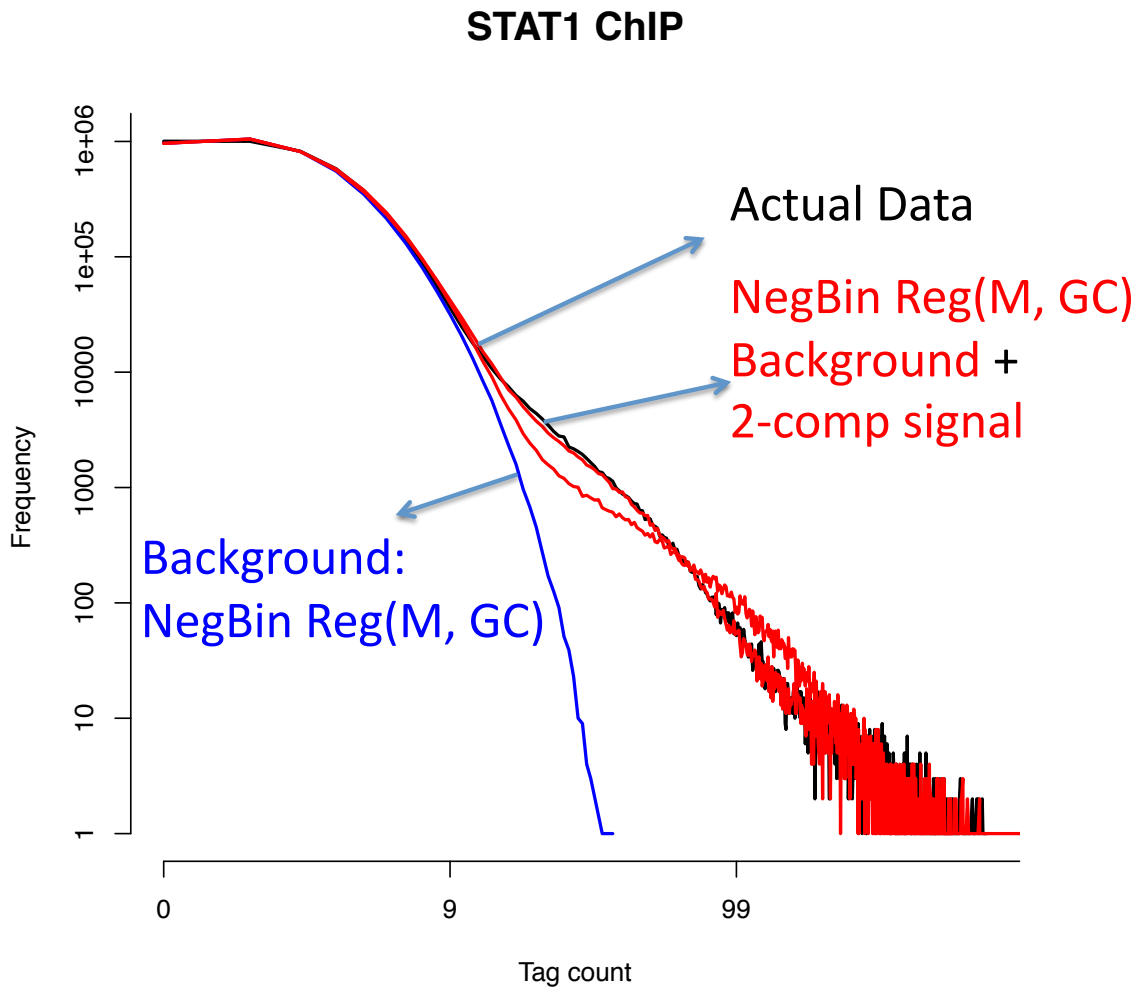
Case Study 1: STAT1 ChIP-Seq data



Case Study 1: STAT1 ChIP-Seq data

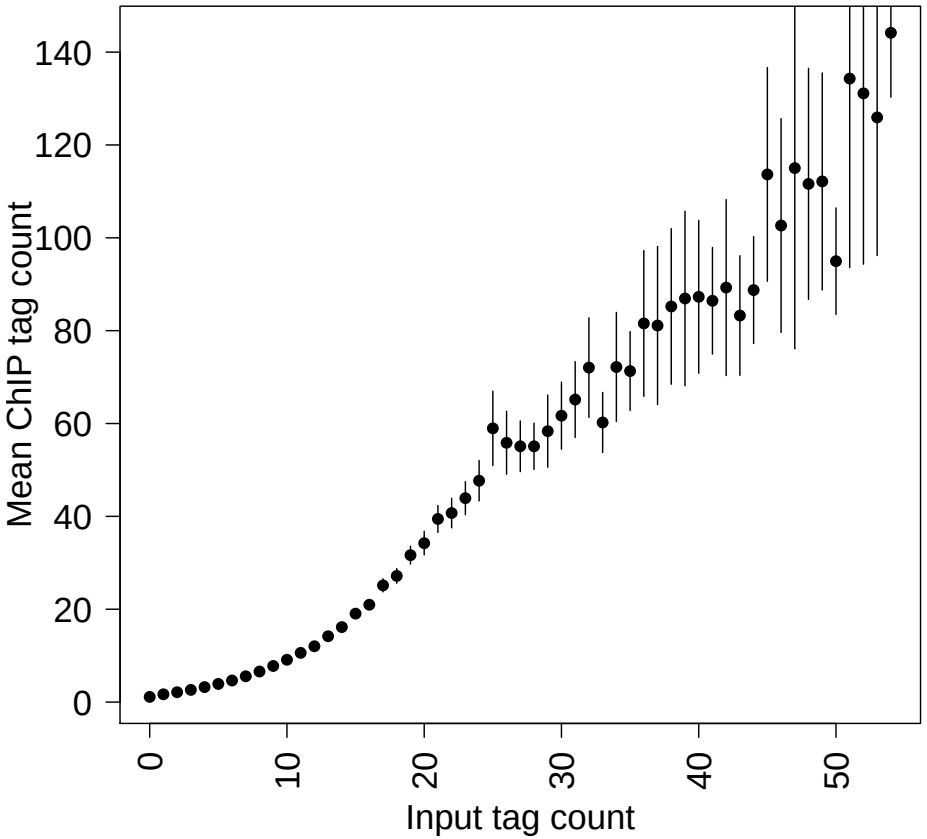


Case Study 1: STAT1 ChIP-Seq data

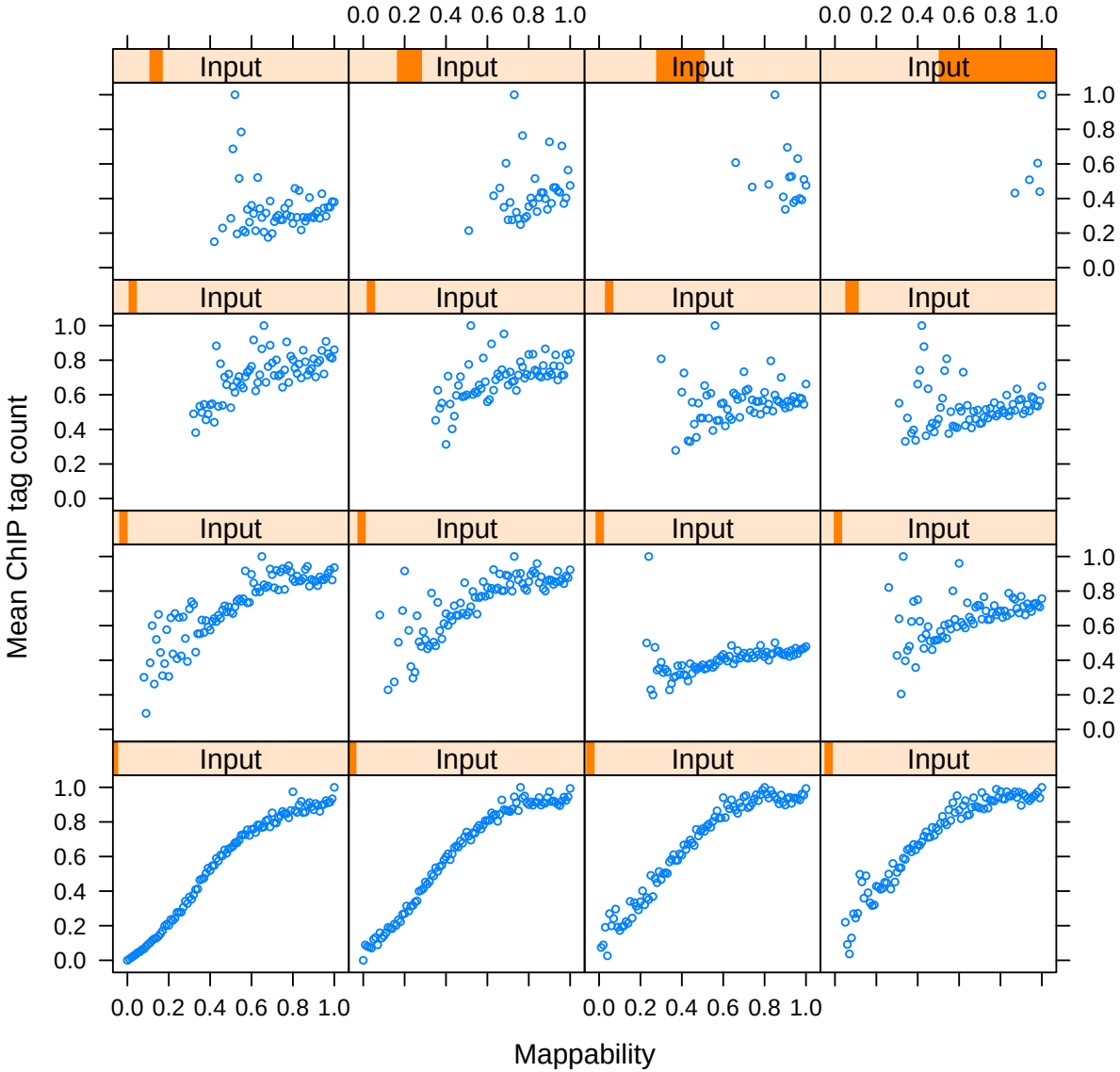


MOSAiCS with Input-Seq: Two-sample analysis

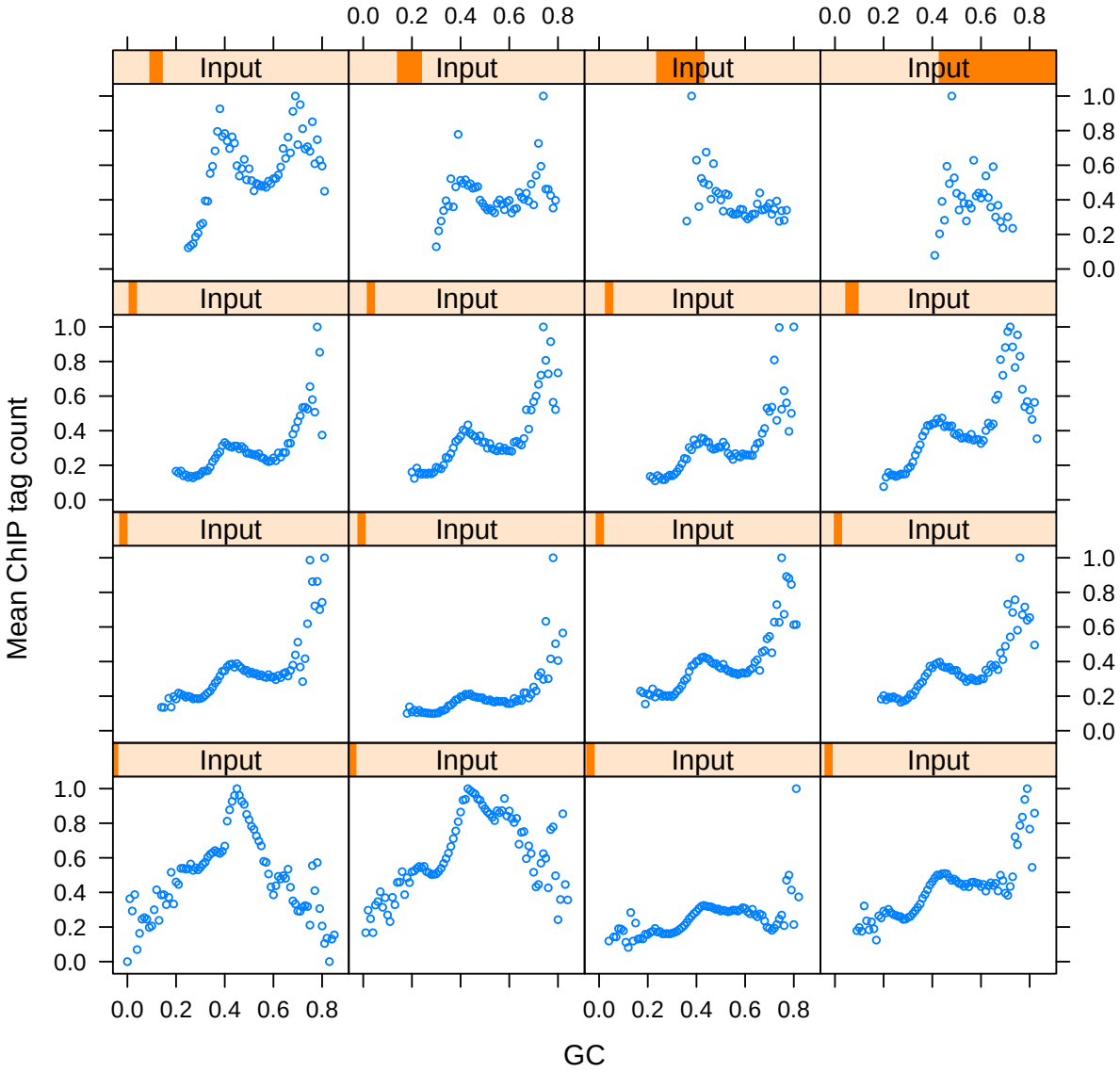
Y_j: tag count from ChIP-Seq;
X_j: tag count from Input-Seq.



ChIP-Seq vs Input-Seq: Does Input-Seq account for all the M and GC bias?



ChIP-Seq vs Input-Seq: Does Input-Seq account for all the M and GC bias?



More

MOSAiCS two-sample background model

$$
\begin{aligned}
Y_j \mid Z_j = 0, X_j, M_j, GC_j &\sim NegBin(a, a/\mu_j) \\
\mu_j &= \exp(\beta_0 + f(M_j, GC_j, X_j))
\end{aligned}
$$

where

$$
\begin{aligned}
f(M, GC, X) &= I(X \leq c)\left[\beta_M \log_2(M+1) + \beta_C Sp(GC) + \beta_X^1 X^d\right] \\
&\quad + I(X > c)\beta_X^2 X^d
\end{aligned}
$$

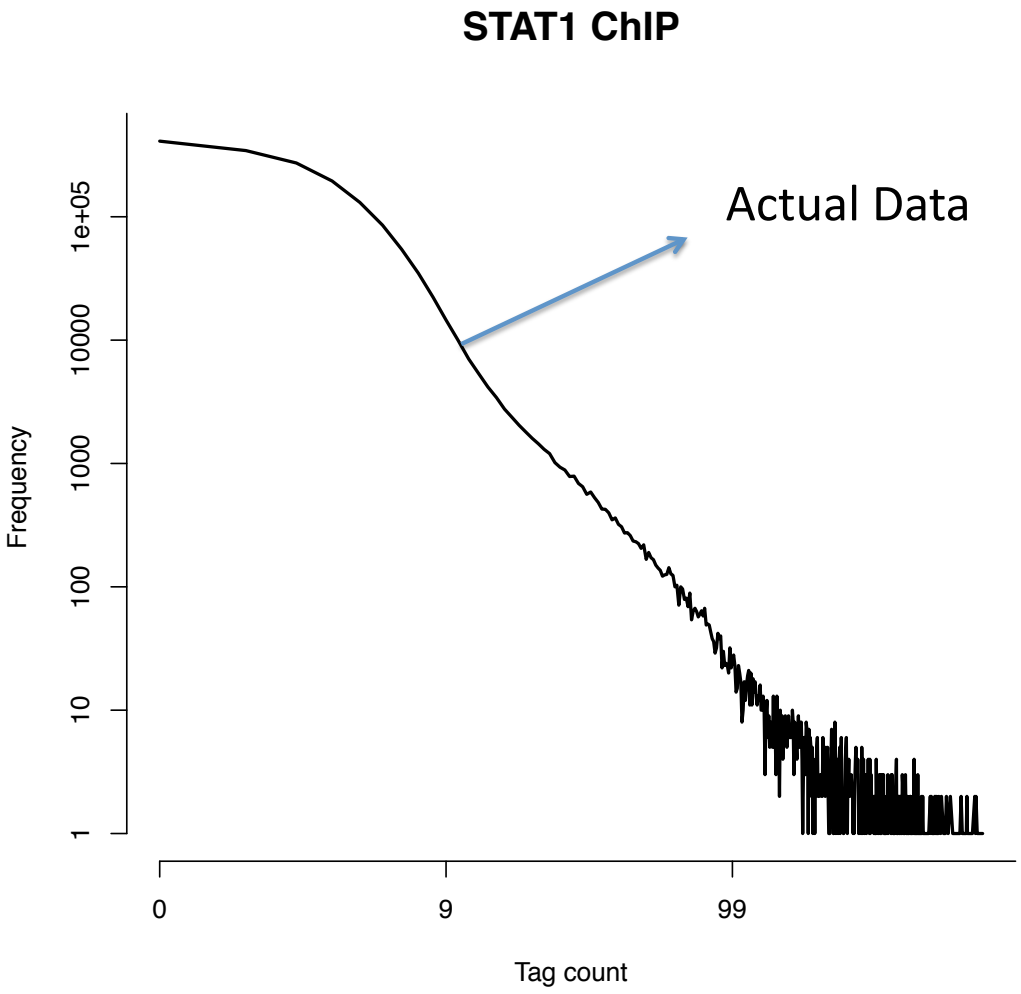
MOSAiCS two-sample background model

$$
\begin{aligned}
Y_j \mid Z_j = 0, X_j, M_j, GC_j &\sim NegBin(a, a/\mu_j) \\
\mu_j &= \exp(\beta_0 + f(M_j, GC_j, X_j))
\end{aligned}
$$

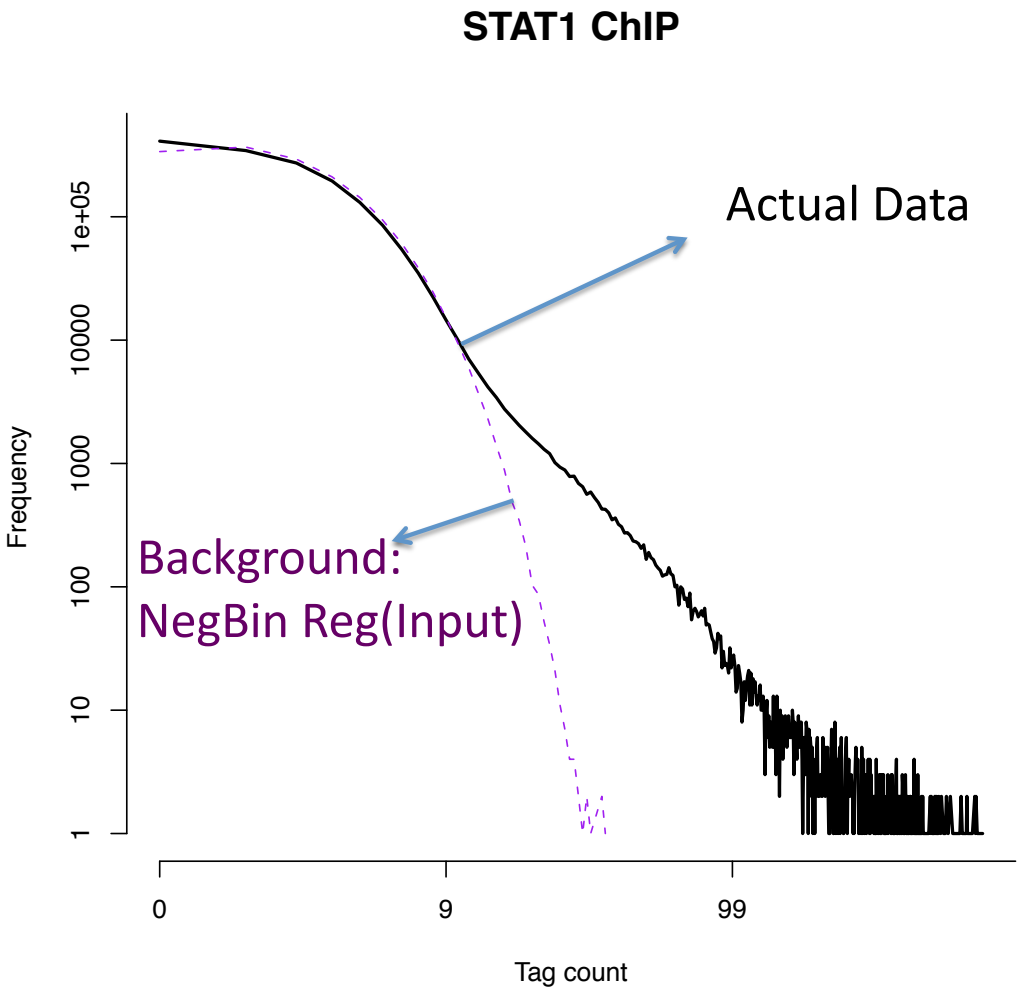
where

$$
\begin{aligned}
f(M, GC, X) &= I(X \leq c)\left[\beta_M \log_2(M+1) + \beta_C Sp(GC) + \beta_X^1 X^d\right] \\
&\quad + I(X > c)\beta_X^2 X^d
\end{aligned}
$$

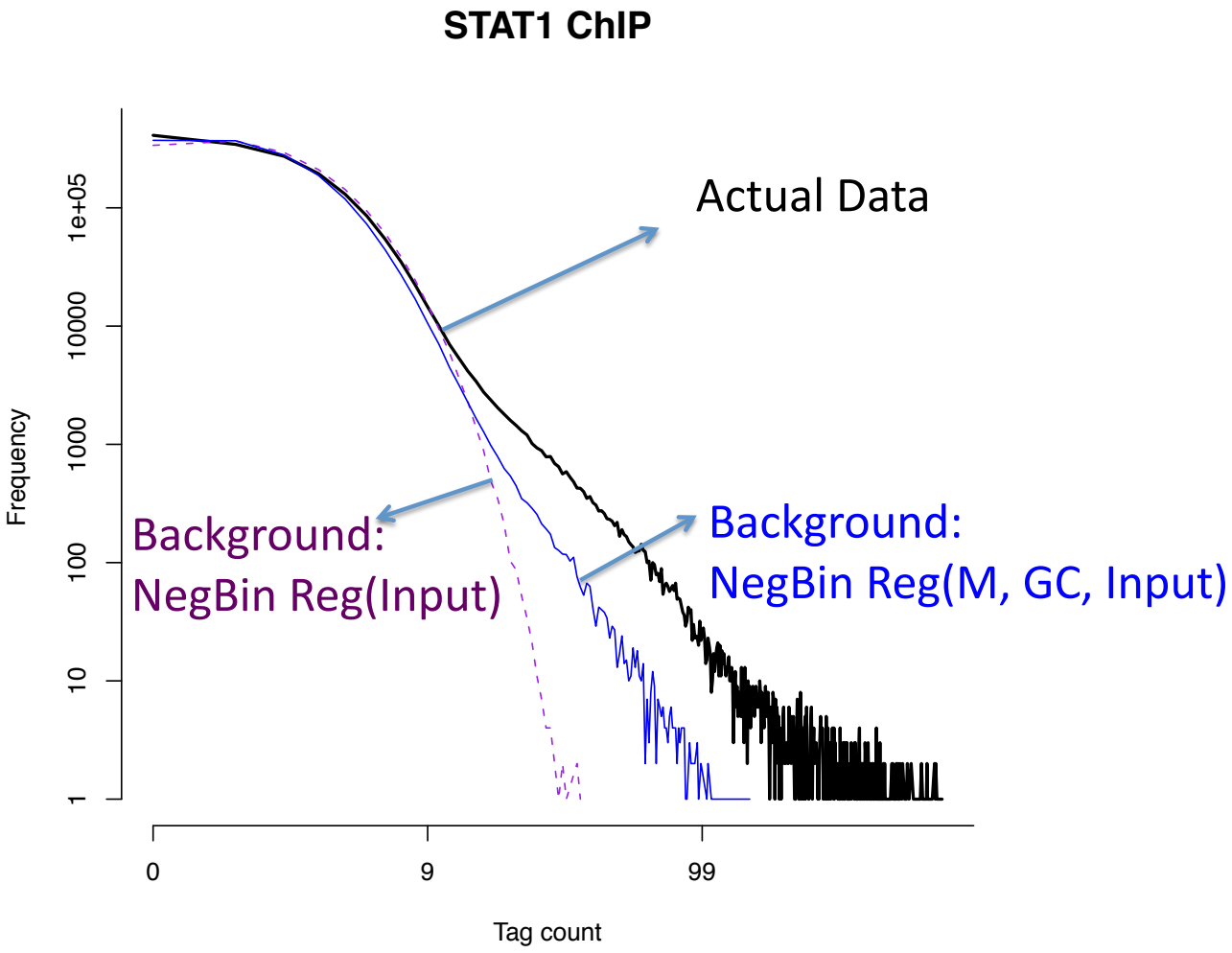
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq



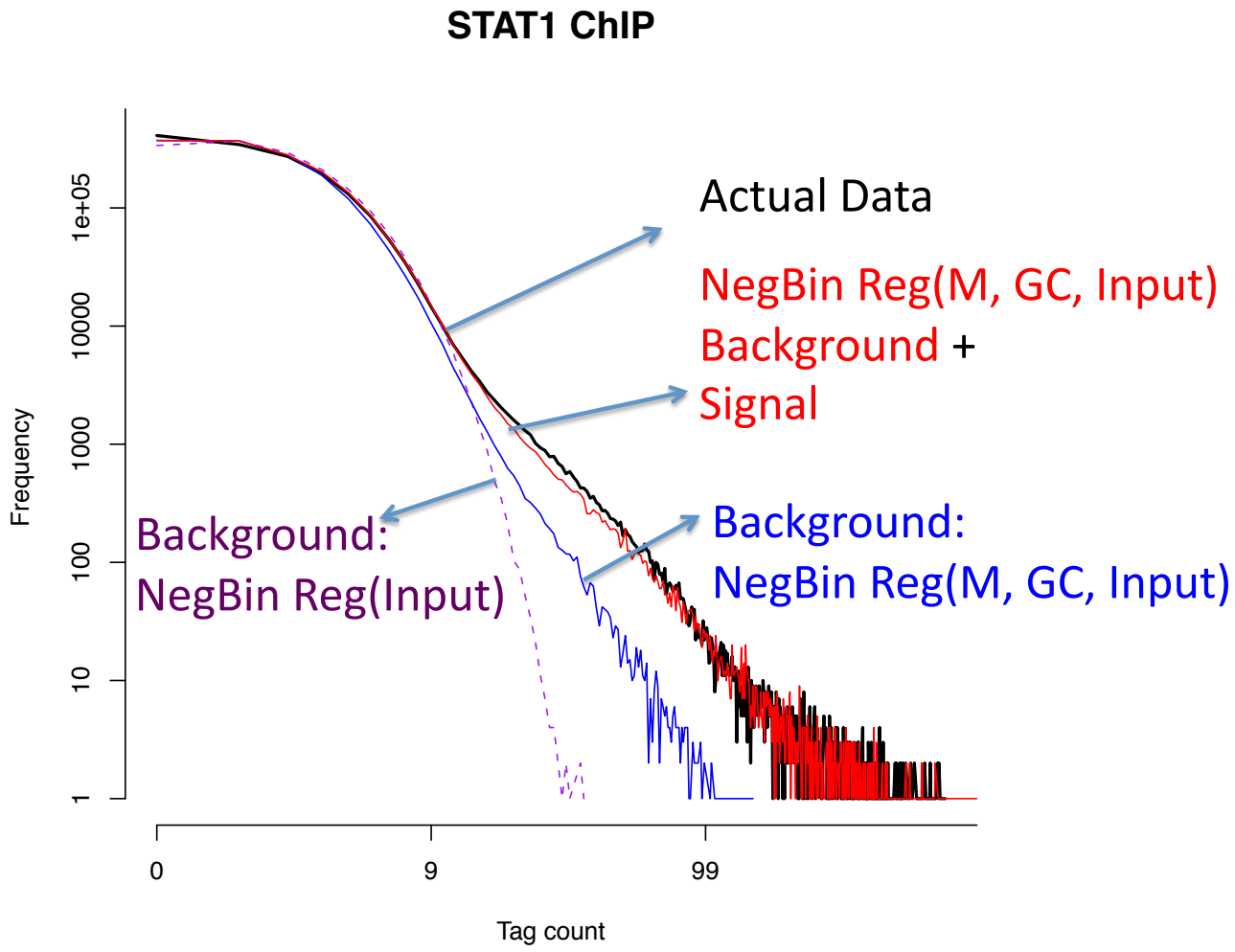
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq



MOSAiCS two-sample model GOF for STAT1 ChIP-Seq

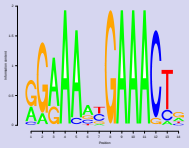


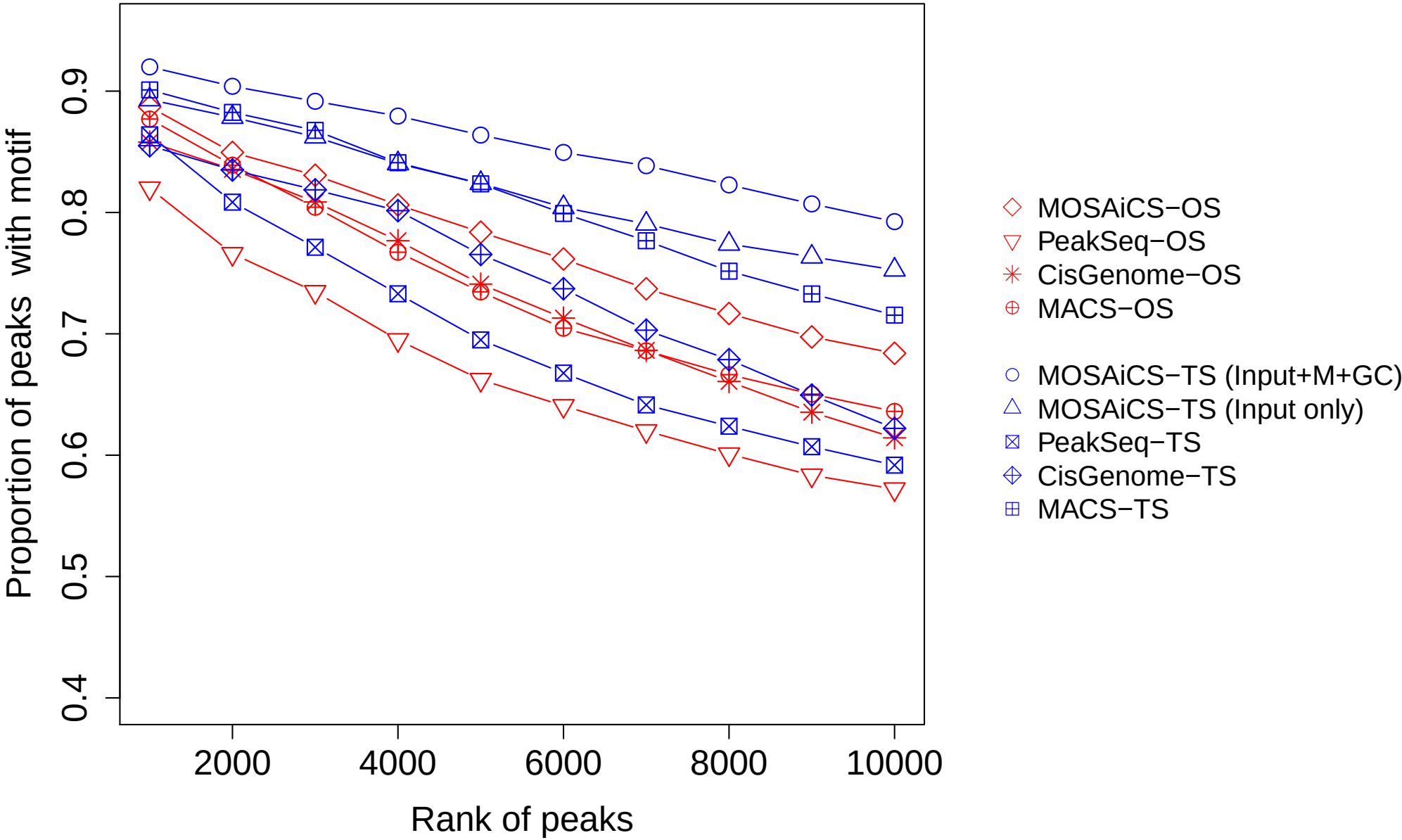
MOSAiCS two-sample model GOF for STAT1 ChIP-Seq



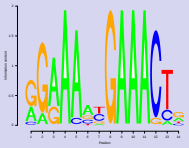
▸ MOSAiCS two-sample for GATA1

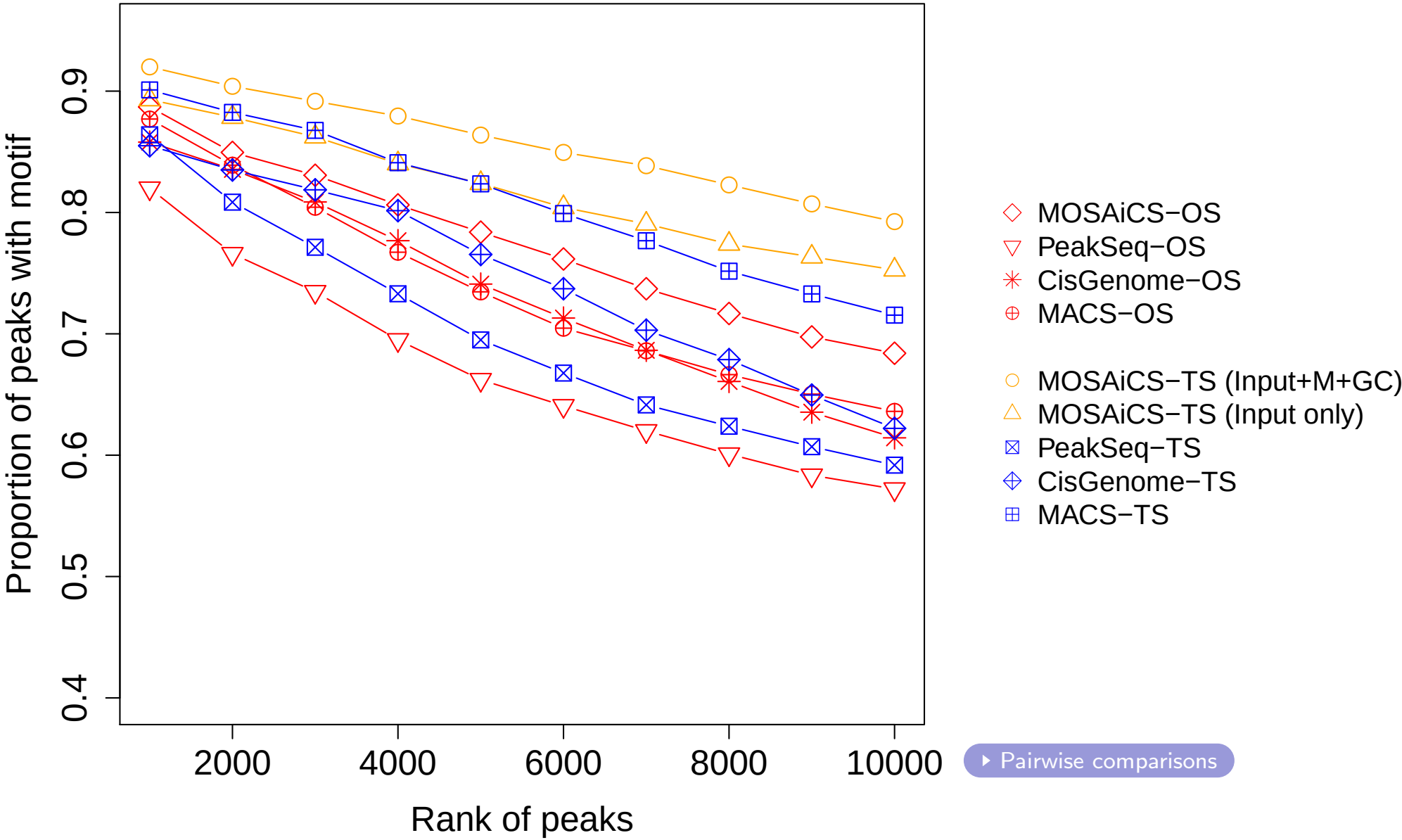
STAT1 motif scans





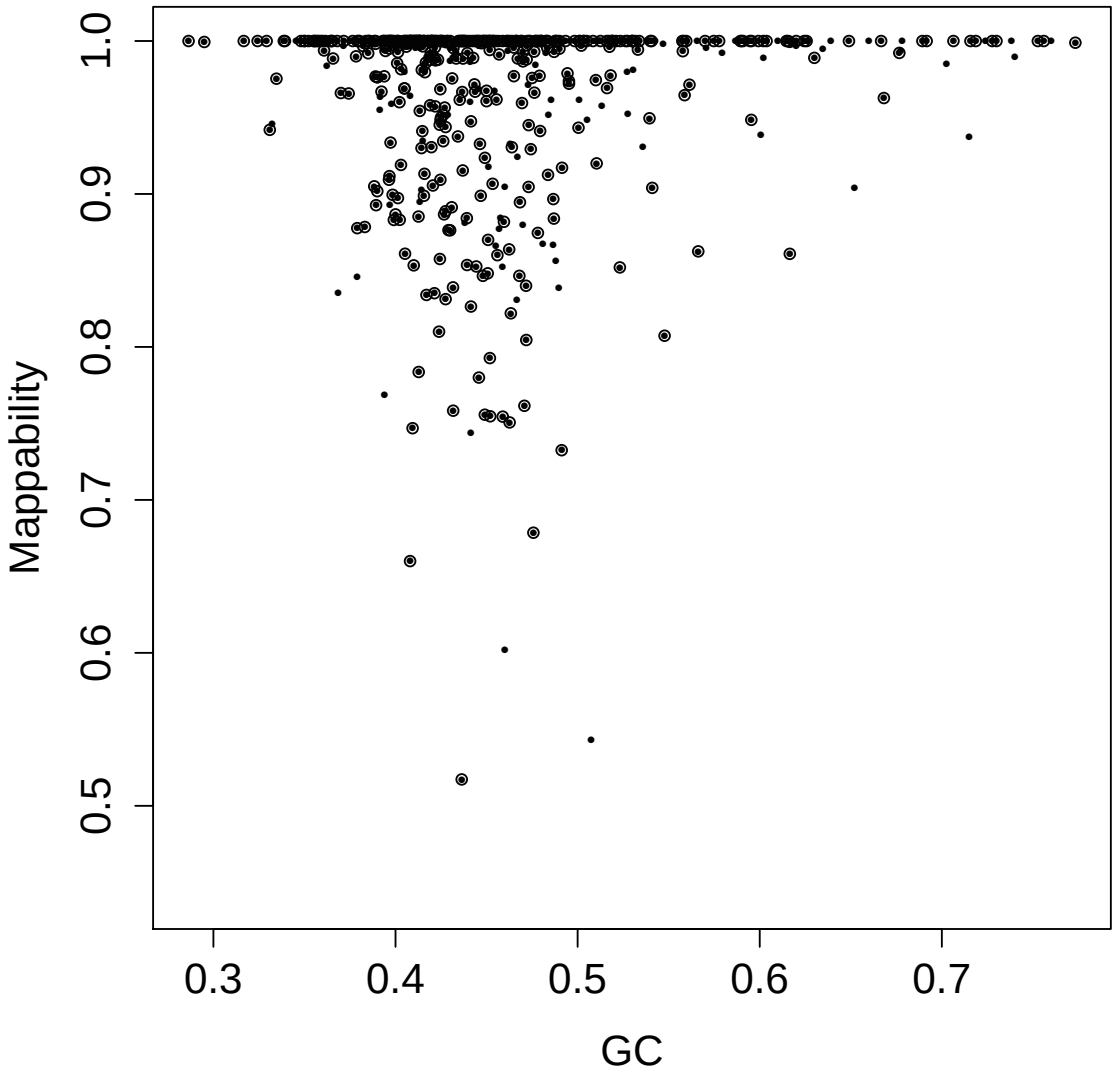
STAT1 motif scans



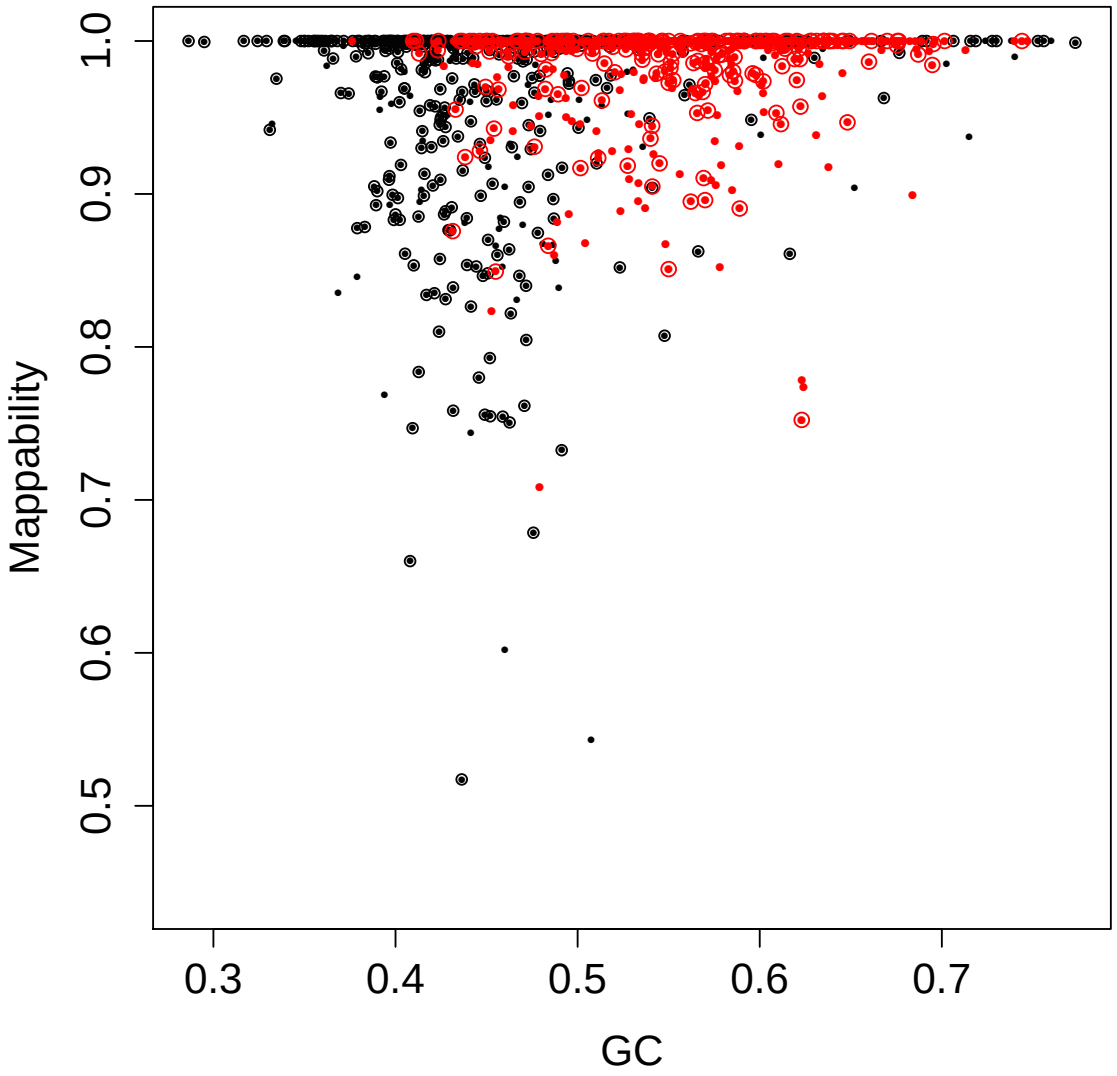


Pairwise comparisons

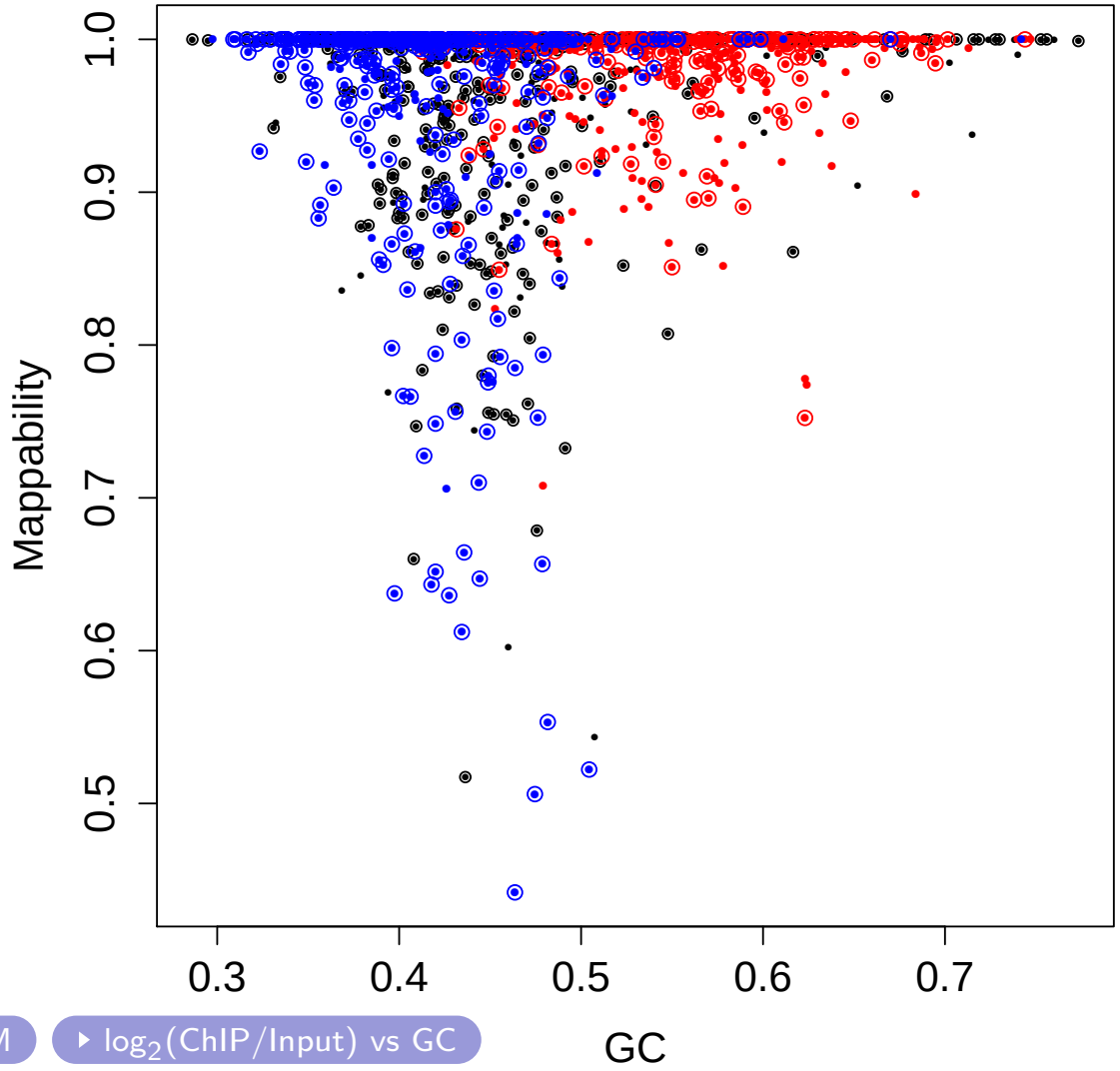
MOSAiCS vs. PeakSeq: Mappability vs GC of Common Peaks



MOSAiCS vs. PeakSeq: Mappability vs GC of PeakSeq only Peaks

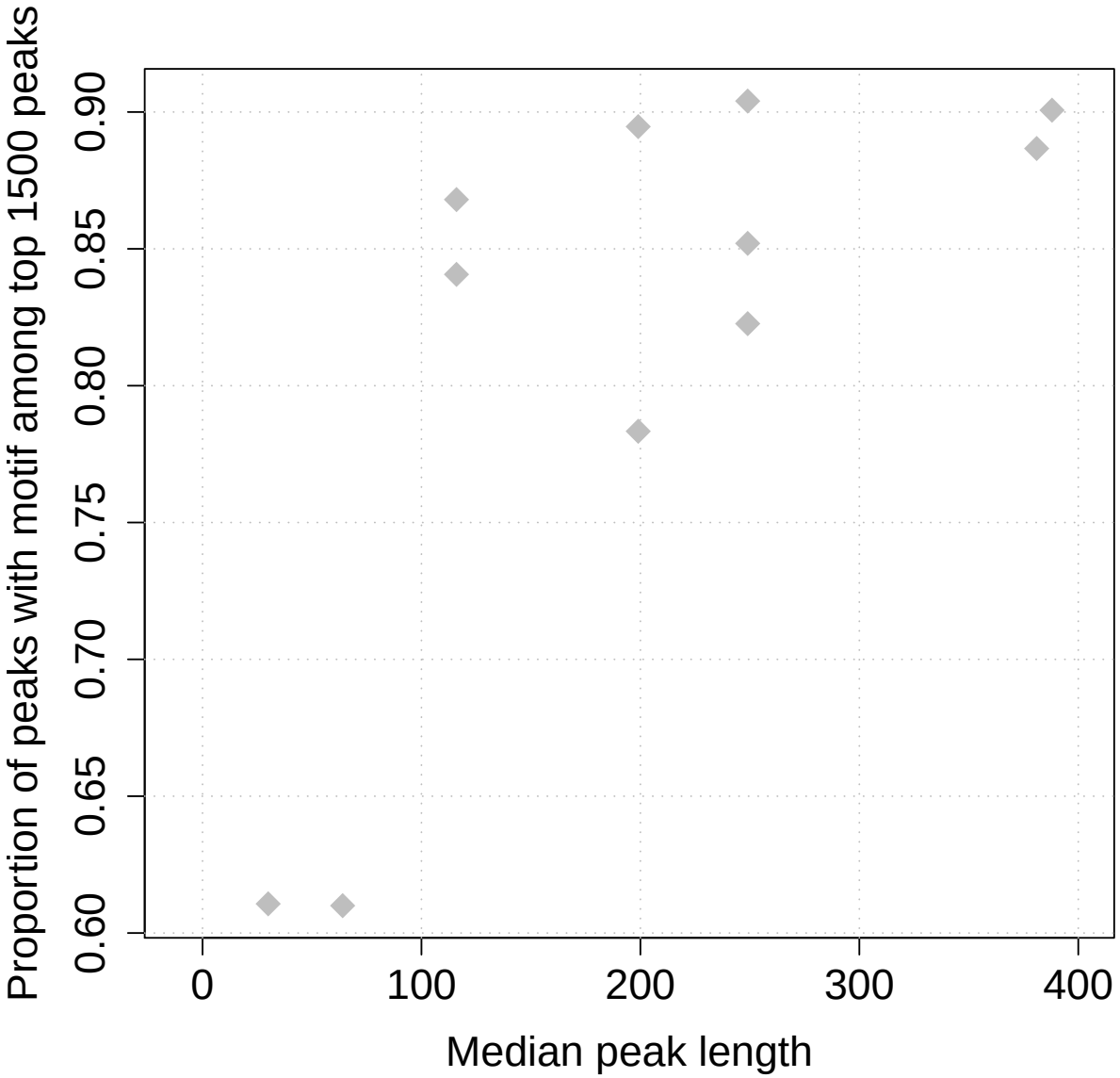


MOSAiCS vs. PeakSeq: Mappability vs GC of MOSAiCS only Peaks

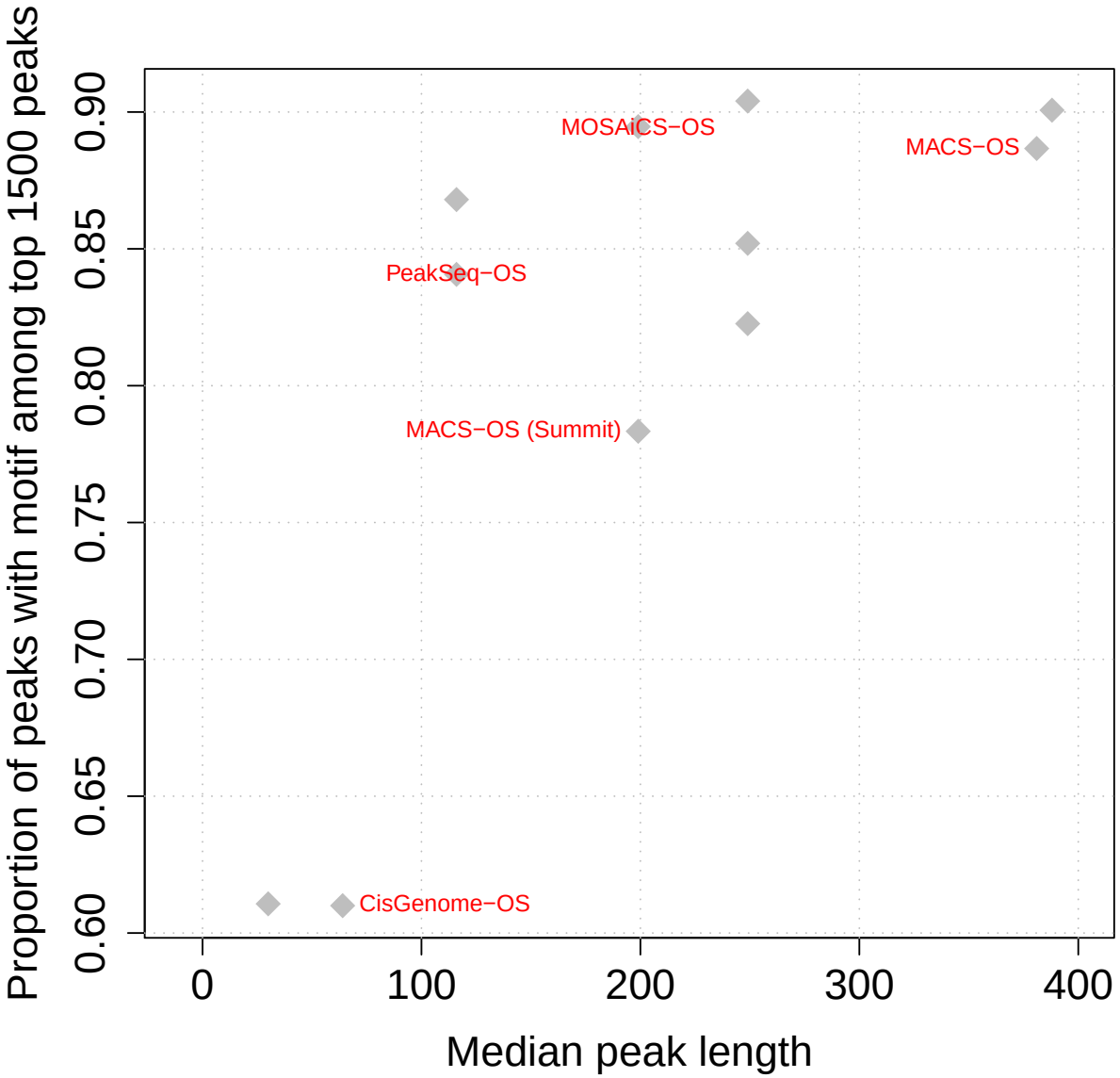


▸ $\log_2$(ChIP/Input) vs M ▸ $\log_2$(ChIP/Input) vs GC

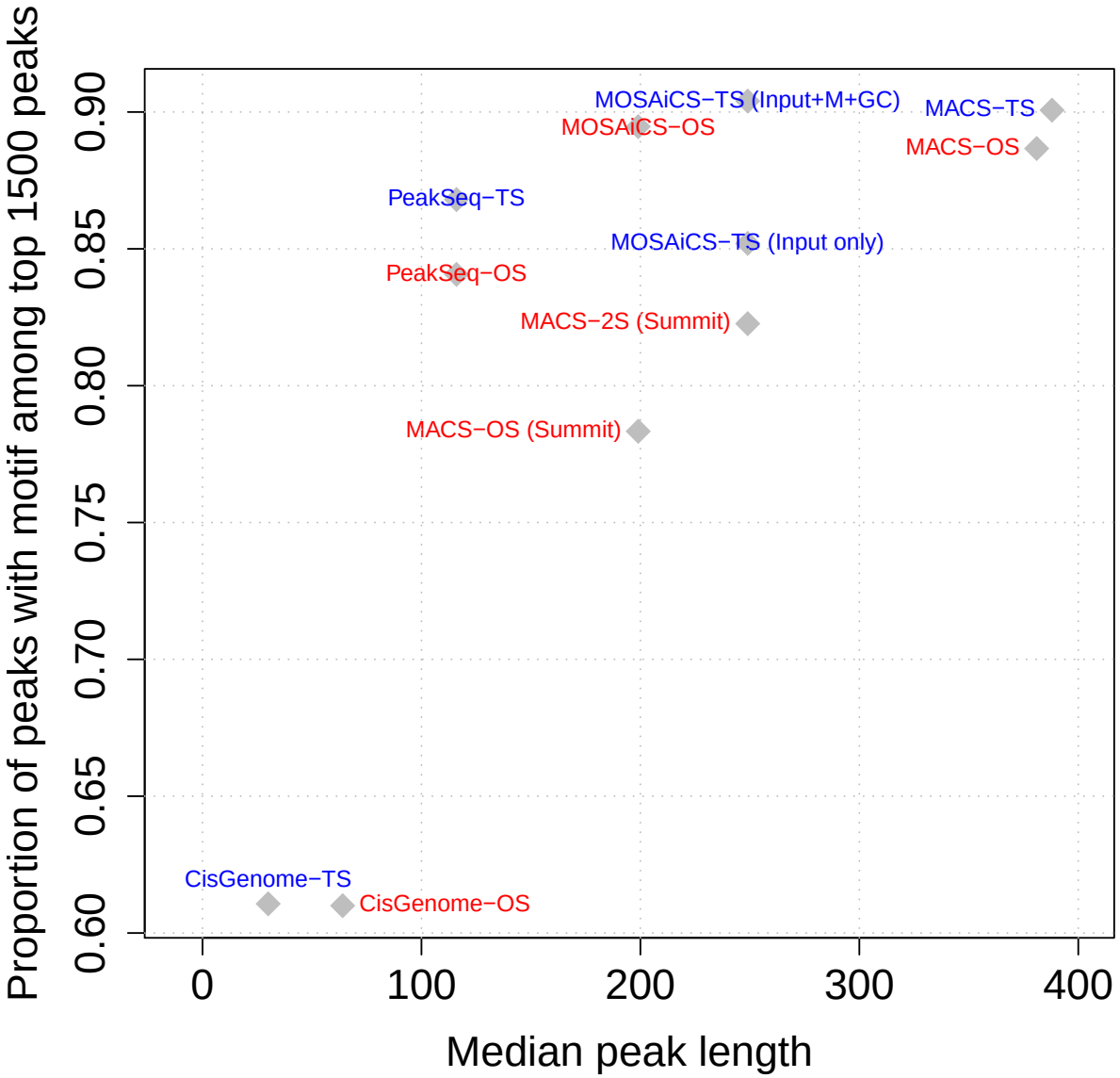
GATA1 motif scans: WGATAA



GATA1 motif scans: WGATAA



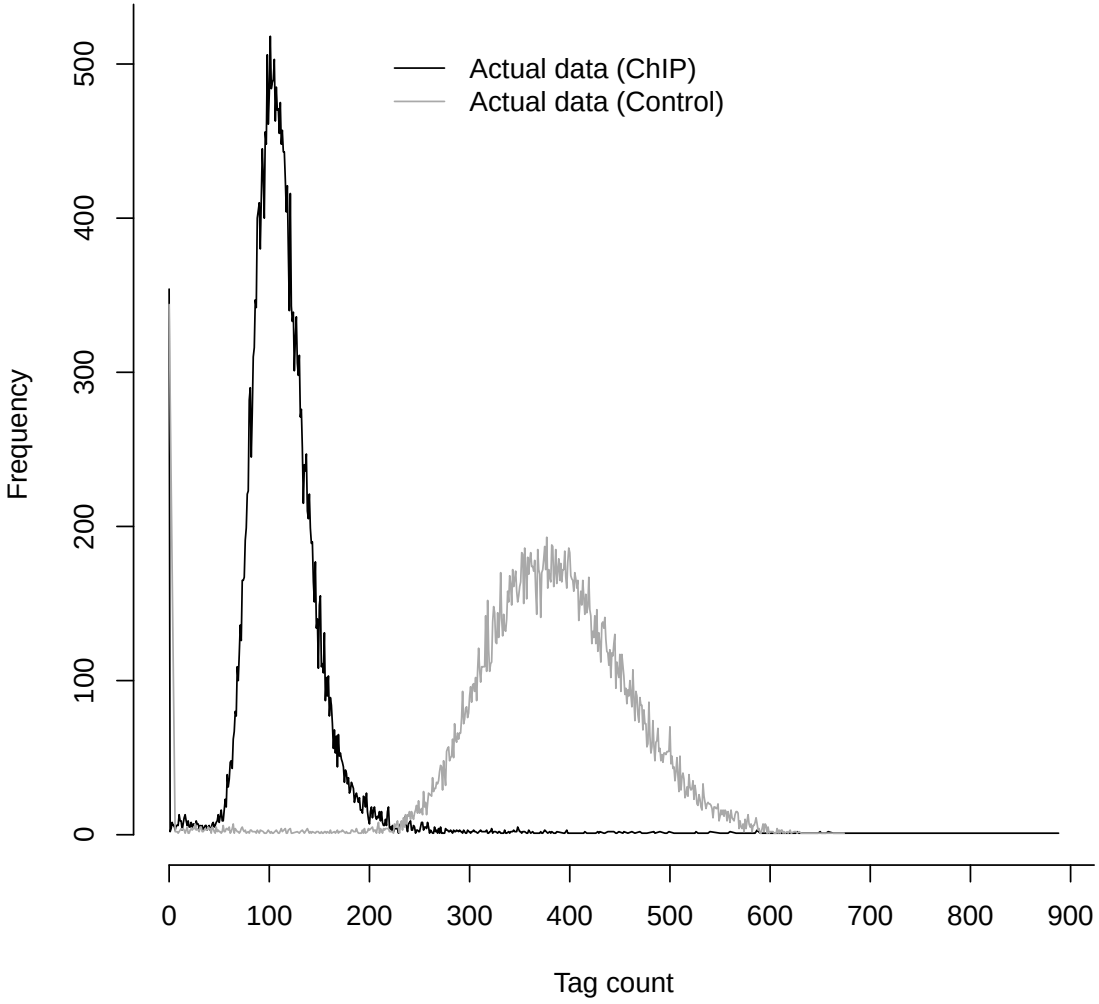
GATA1 motif scans: WGATAA



If the sample is deeply sequenced

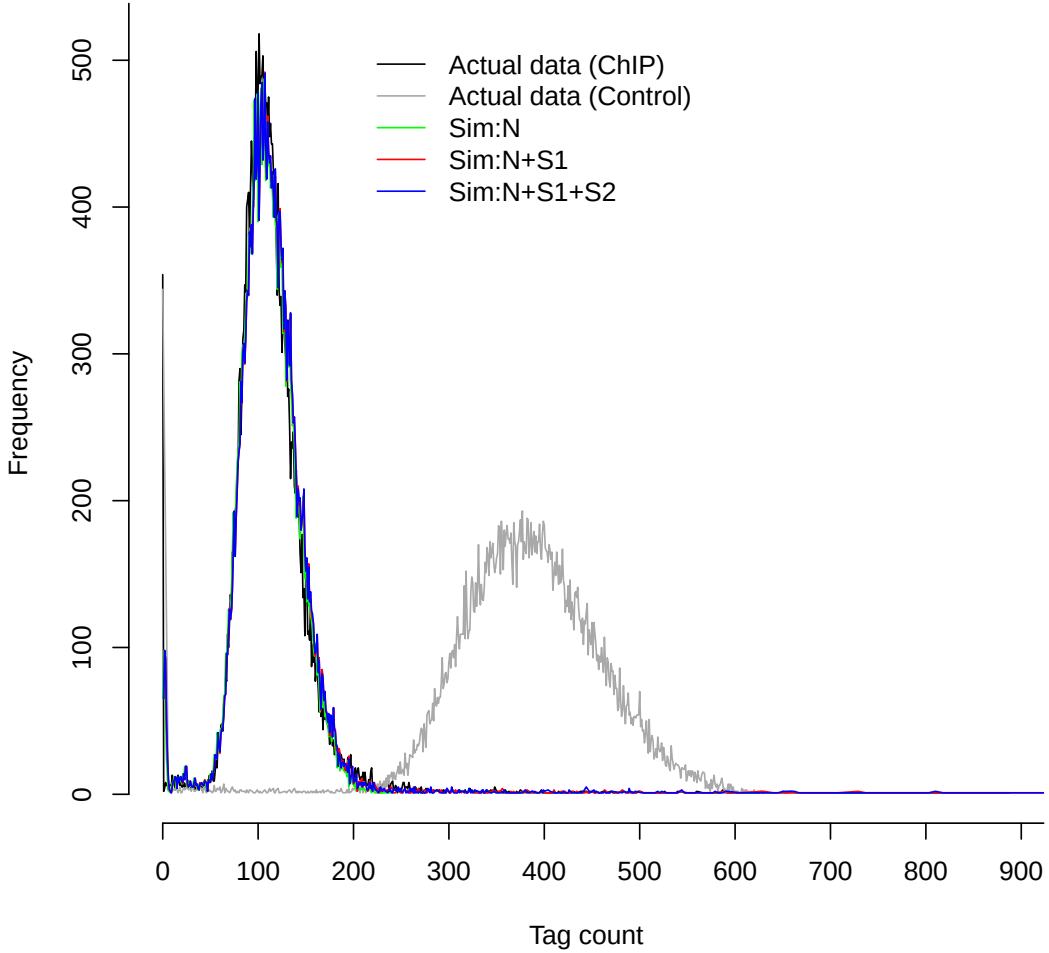
E.g.,

- *E. coli* ChIP-seq - one lane from the Illumina GA-II.
- Higher eukaryote ChIP-seq sample on Illumina Hi-seq.



MOSAiCS on the FNR ChIP-seq data from the Kiley Lab @ UW Madison

▸ GOF on log scale

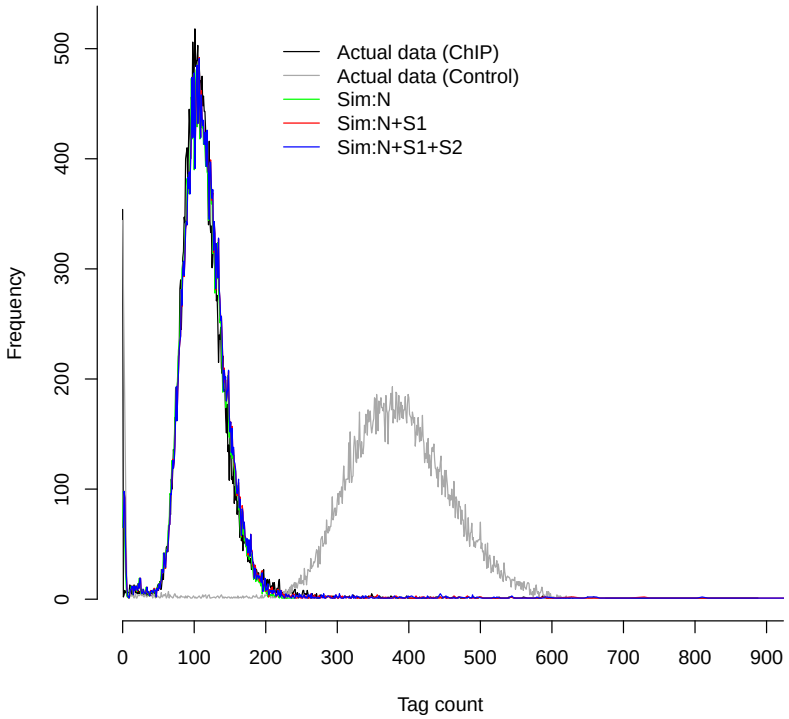


MOSAiCS on the FNR ChIP-seq data from the Kiley Lab @ UW Madison

▸ GOF on log scale

If the sample is deeply sequenced,

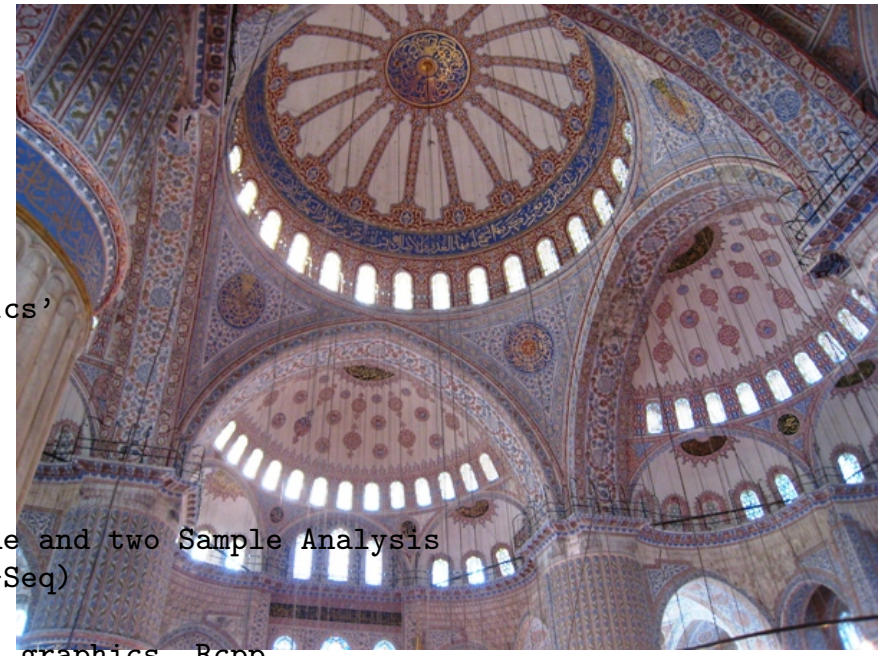
Input-only model fits well.



Software implementation: R package mosaics

Available through

- Bioconductor http://www.bioconductor.org/packages/2.9/bioc/html/mosaics.html
- Galaxy http://toolshed.g2.bx.psu.edu/.



```
> library(mosaics)
> library(help = mosaics)
                Information on package 'mosaics'

Description:

Package:               mosaics
Type:                  Package
Title:                 MOSAiCS (MOdel-based one and two Sample Analysis
                       and Inference for ChIP-Seq)
Version:               1.2.2
Depends:               R (>= 2.11.1), methods, graphics, Rcpp
Imports:               MASS, splines, lattice, IRanges
Suggests:              mosaicsExample, multicore
LinkingTo:             Rcpp
SystemRequirements:    Perl
Date:                  2012-01-10
Author:                Dongjun Chung, Pei Fen Kuan, Sunduz Keles
Maintainer:            Dongjun Chung <chungdon@stat.wisc.edu>
Description:           This package provides functions for fitting
```

R package: mosaics

ChIP file has 95 million reads.

Input file has 14 million reads.
mosaics runs in about 2 hrs using a single CPU.

```
library(mosaics)
 mosaicsRunAll(
 chipDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/",
 chipFileName="GSM746584_tal1_ter119_r2_mapped.txt",
 chipFileFormat="bowtie",
 controlDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/",
 controlFileName="GSM746580_input_ter119_mapped.txt",
 controlFileFormat="bowtie",
 binfileDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/",
 peakDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/peak/",
 peakFileName="tal1_ter119_peak_list_r2.txt",
 peakFileFormat="txt",
 reportSummary=TRUE,
 summaryDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
 summaryFileName="mosaics_summary_tal1_ter119_r2.txt",
 reportExploratory=FALSE,
 exploratoryDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
 exploratoryFileName="mosaics_exploratory_tal1_ter119_r2.pdf",
 reportGOF=TRUE,
 gofDir="/scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/reports/",
 gofFileName="mosaics_GOF_tal1_ter119_r2.pdf",
 byChr=FALSE,
 FDR=0.05,
 fragLen=200,
 binSize=200,
 capping=0,
 analysisType="IO",
```

R package: mosaics

```
d=0.25,
signalModel="BIC",
maxgap=200,
minsize=50,
thres=40,
nCore=20)
```

R package: mosaics

```
Info: constructing bin-level files...
------------------------------------------------------------
Info: setting summary
------------------------------------------------------------
Directory of aligned read file: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/
Name of aligned read file: GSM746584_tal1_ter119_r2_mapped.txt
Aligned read file format: Bowtie default
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Construct bin-level files by chromosome? N
Fragment length: 200
Bin size: 200
------------------------------------------------------------
------------------------------------------------------------
Info: setting summary
------------------------------------------------------------
Directory of aligned read file: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/
Name of aligned read file: GSM746580_input_ter119_mapped.txt
Aligned read file format: Bowtie default
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Construct bin-level files by chromosome? N
Fragment length: 200
Bin size: 200
------------------------------------------------------------
Info: reading the aligned read file and processing it into bin-level files...
Info: reading the aligned read file and processing it into bin-level files...
Info: done!
```

R package: mosaics

```
-----------------------------------------------------------
Info: processing summary
-----------------------------------------------------------
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Processed bin-level file: GSM746580_input_ter119_mapped.txt_fragL200_bin200.txt
-----------------------------------------------------------
Info: done!
-----------------------------------------------------------
Info: processing summary
-----------------------------------------------------------
Directory of processed bin-level files: /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/bin/
Processed bin-level file: GSM746584_tal1_ter119_r2_mapped.txt_fragL200_bin200.txt
-----------------------------------------------------------
```

R package: mosaics

```
Info: analyzing bin-level files...
Info: fitting MOSAiCS model & call peaks...
Info: reading and preprocessing bin-level data...
Info: data contains more than one chromosome.
Info: done!

Info: background estimation method is determined based on data.
Info: background estimation based on robust method of moment
Info: two-sample analysis (Input only).
Info: use adaptive griding.
Info: fitting background model...
Info: done!
Info: fitting one-signal-component model...
Info: fitting two-signal-component model...
Info: calculating BIC of fitted models...
Info: done!
Info: use two-signal-component model.
Info: calculating posterior probabilities...
Info: calling peaks...
Info: done!
Info: writing the peak list...
Info: peak file was exported in TXT format:
Info: file name = tal1_ter119_peak_list_r2.txt
Info: directory = /scratch/ChIPSeqDesign/HardisonData/tal1_ter119_r2/Results/peak/
Info: generating reports...

> proc.time()
    user   system  elapsed
7154.502  114.432 6986.742
```

Run time $\approx$ 2 hours.

R package: mosaics - Main output

- A peak list with columns: `chrID peakStart peakStop peakSize aveP minP aveChipCount maxChipCount aveInputCount aveInputCountScaled aveLog2Ratio`.
- Exploratory plots. Mean read count vs. Input, vs. Mappability, vs. GC content.
- Goodness-of-fit plots.

Some other statistical problems regarding ChIP-seq data

- Utilizing multi-reads, i.e., read mapping to multiple locations on the reference genome.

OPEN ACCESS Freely available online PLOS COMPUTATIONAL BIOLOGY

Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data

Dongjun Chung[1,2], Pei Fen Kuan[3], Bo Li[4], Rajendran Sanalkumar[5], Kun Liang[1,2], Emery H. Bresnick[5], Colin Dewey[2,4], Sündüz Keles[1,2]*

1 Department of Statistics, University of Wisconsin, Madison, Wisconsin, United States of America, **2** Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, Wisconsin, United States of America, **3** Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, United States of America, **4** Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin, United States of America, **5** Wisconsin Institutes for Medical Research, UW Carbone Cancer Center, Department of Cell and Regenerative Biology, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin, United States of America

- Differential binding.

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 28 no. 1 2012, pages 121–122*
doi:10.1093/bioinformatics/btr605

Genome analysis Advance Access publication November 3, 2011

Detecting differential binding of transcription factors with ChIP-seq

Kun Liang[1,2,*] and Sündüz Keleş[1,2]

[1]Department of Statistics and [2]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53706, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: Increasing number of ChIP-seq experiments are

Summary

Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

A Statistical Framework for the Analysis of ChIP-Seq Data

Pei Fen KUAN, Dongjun CHUNG, Guangjin PAN, James A. THOMSON, Ron STEWART, and Sündüz KELEŞ

Chromatin immunoprecipitation followed by sequencing (ChIP-Seq) has revolutionalized experiments for genome-wide profiling of DNA-binding proteins, histone modifications, and nucleosome occupancy. As the cost of sequencing is decreasing, many researchers are switching from microarray-based technologies (ChIP-chip) to ChIP-Seq for genome-wide study of transcriptional regulation. Despite its increasing and well-deserved popularity, there is little work that investigates and accounts for sources of biases in the ChIP-Seq technology. These biases typically arise from both the standard preprocessing protocol and the underlying DNA sequence of the generated data.

We study data from a naked DNA sequencing experiment, which sequences noncross-linked DNA after deproteinizing and shearing, to understand factors affecting background distribution of data generated in a ChIP-Seq experiment. We introduce a background model that accounts for apparent sources of biases such as mappability and GC content and develop a flexible mixture model named MOSAiCS for detecting peaks in both one- and two-sample analyses of ChIP-Seq data. We illustrate that our model fits observed ChIP-Seq data well and further demonstrate advantages of MOSAiCS over commonly used tools for ChIP-Seq data analysis with several case studies. This article has supplementary material online.

KEY WORDS: GC content; Mappability; Mixture model; Negative binomial regression; Next generation sequencing.

MOSAiCS:

- implements a model-based approach for ChIP-seq data analysis;
- available as a `R` package through Bioconductor `http://www.bioconductor.org/` and a Galaxy tool from the Galaxy tool shed `toolshed.g2.bx.psu.edu/`;
- provides basic pre-processing functions for ChIP-seq data;
- provides plotting functions.

Acknowledgements

Dongjun Chung (Department of Statistics, UW Madison).

Pei Fen Kuan (Department of Biostatistics, UNC Chapel Hill).

Kun Liang (Department of Statistics, UW Madison).

Xin Zeng (Department of Statistics, UW Madison).

Chen Zuo (Department of Statistics, UW Madison).

Colin Dewey (Department of Biostatistics and Medical Informatics, UW Madison); Bo Li (Department of Computer Science, UW Madison).

Emery Bresnick; Sanal Kumar (Department of Cell and Regenerative Biology, UW Madison) .

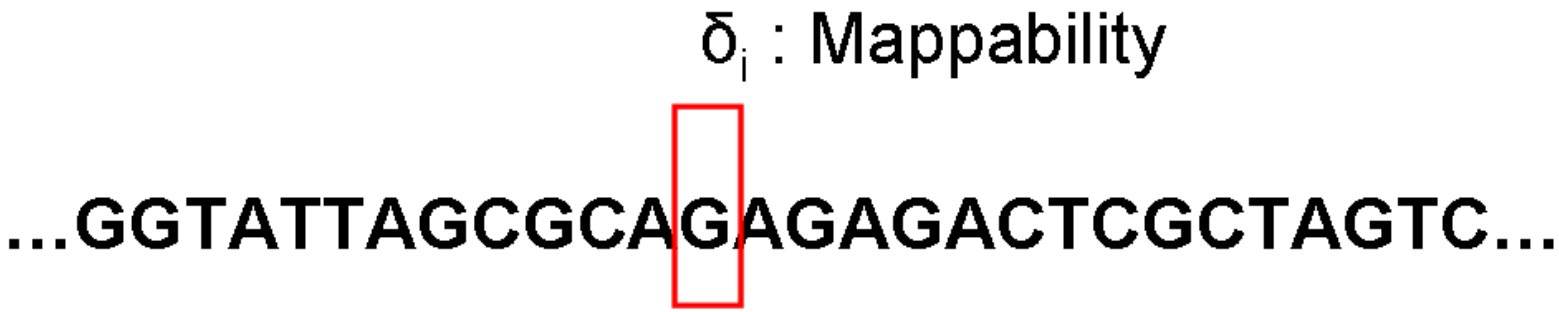
Peggy Farnham (Departments of Biochemistry & Molecular Biology, Keck School of Medicine, USC).

Qiang Chang (Cellular and Molecular Neurosciences Core, Waisman Center, UW Madison).

NIH, NSF.

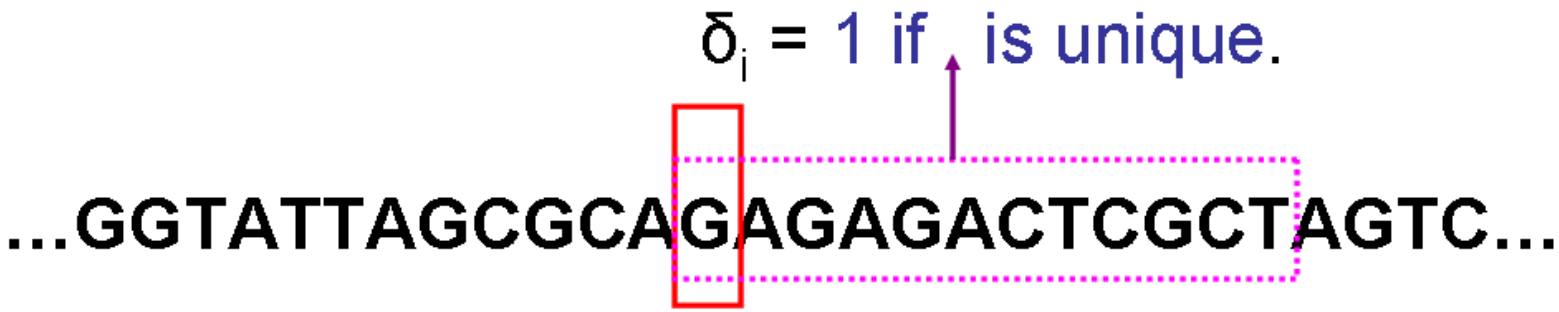
Sequence bias in ChIP-Seq data

- Mappability bias. Original definition by Rozowsky *et al.* (2009).



Sequence bias in ChIP-Seq data

- Mappability bias. Original definition by (Rozowsky *et al.* (2009)).

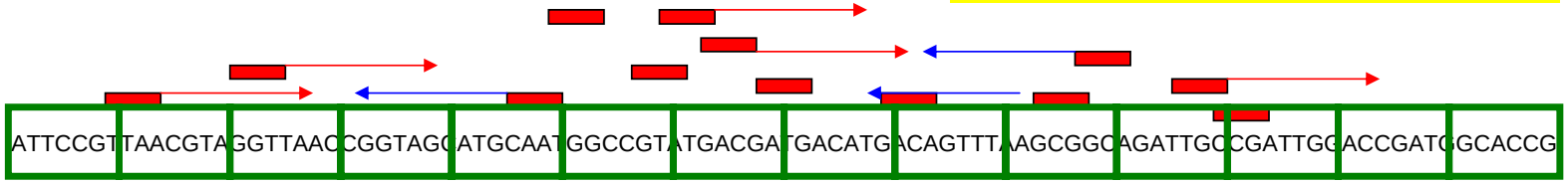


Tag/Read extension

Extending each tag by expected fragment length and strand

Partition the genome into small bins (genomic windows).

Summarize total tag counts in each bin



Sequence bias in ChIP-Seq data

- Mappability. Consider tags originating from nearby nucleotides.

k: read length

$$m_i = [1/(2L)] \left(\sum \delta_{j,\ j = \{i-L+1, \ldots i\}} + \sum \delta_{j,\ j = \{i-k+1, \ldots, i+L-k\}} \right)$$

Sequence bias in ChIP-Seq data

- Mappability. Bin level.

Mappability for j^{th} bin
$M_j = \text{mean}(m_i)$ i^{th} bp in the bin.

...GGTATTAGCGCAGAGAGACTCGCTAGTC...