

# Selected Topics in ChIP-seq data analysis

Tamás Schauer

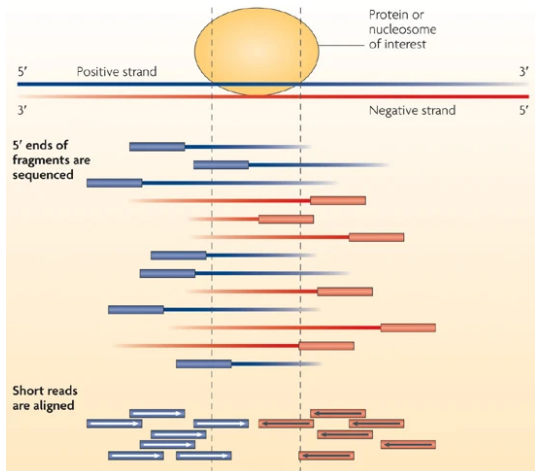
03.03.2021

# Overview

- ▶ Introduction
- ▶ ChIP-seq Coverage
- ▶ Normalization Methods
- ▶ Peak Overlaps
- ▶ Statistical Analysis

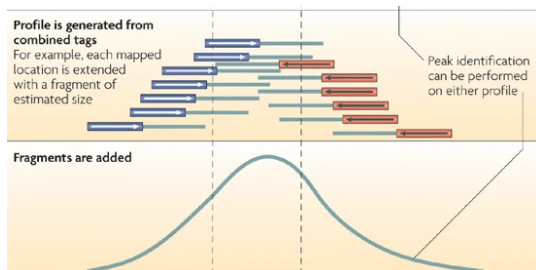
Source: [https://github.com/tschauer/ChIPseq\\_Talk](https://github.com/tschauer/ChIPseq_Talk)

# Introduction



Peter J. Park, 2009

# Introduction



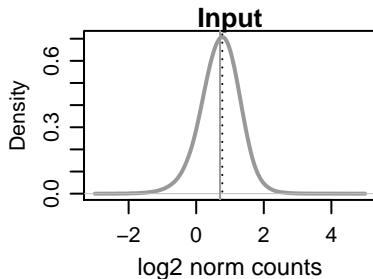
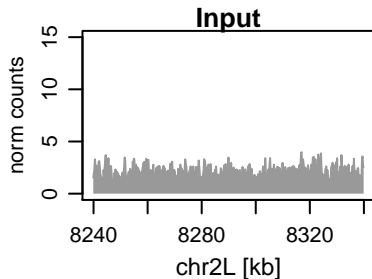
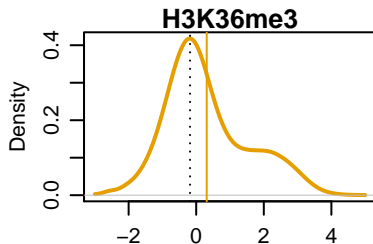
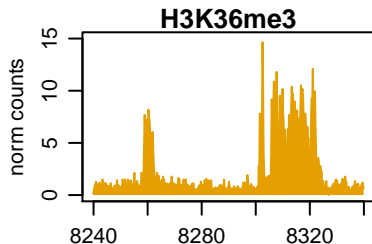
Nature Reviews | Genetics

Peter J. Park, 2009

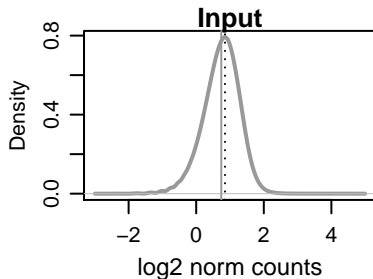
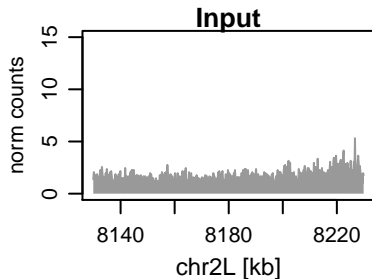
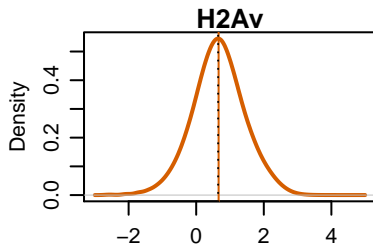
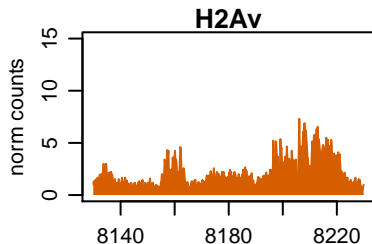
# Overview

- ▶ Introduction
- ▶ **ChIP-seq Coverage**
- ▶ Normalization Methods
- ▶ Peak Overlaps
- ▶ Statistical Analysis

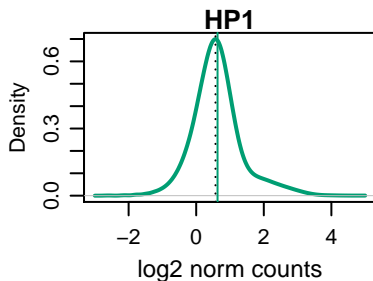
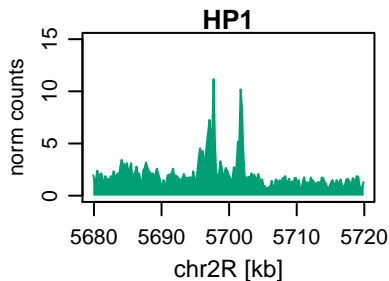
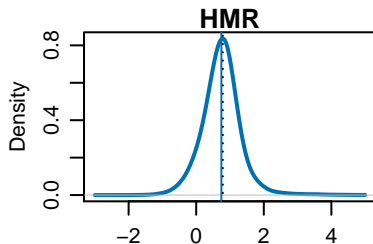
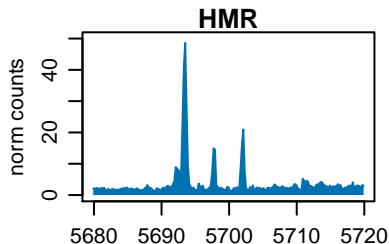
# ChIP Coverage



# ChIP Coverage



# ChIP Coverage





# Overview

- ▶ Introduction
- ▶ ChIP-seq Coverage
- ▶ **Normalization Methods**
- ▶ Peak Overlaps
- ▶ Statistical Analysis

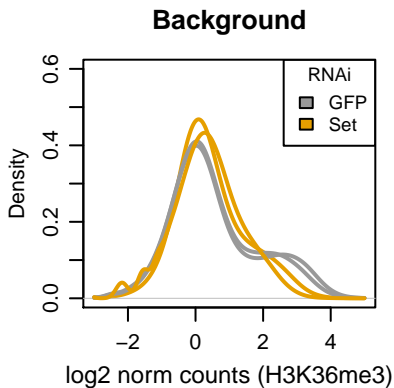
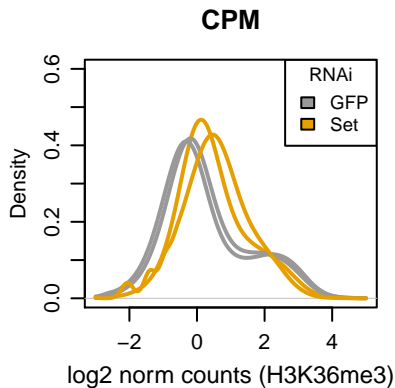
# ChIP-seq Normalization

- ▶ **Sub-sampling**: random selection of certain number of reads
- ▶ **Counts Per Million (CPM)**: divide by the total number reads
- ▶ **Background**: remove compositional bias
- ▶ **Spike-In**: add constant amount of foreign chromatin

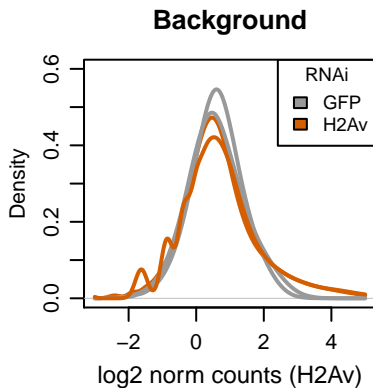
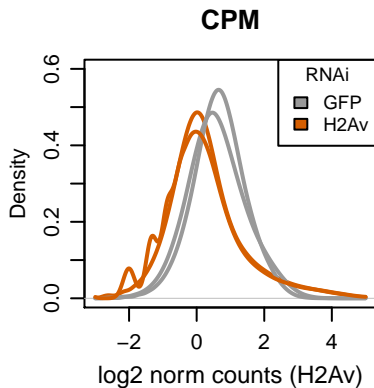
# Background Normalization

- ▶ use large bins (10 kb)
- ▶ TMM - trimmed mean of M-values
- ▶ trim away extreme values

# Background Normalization



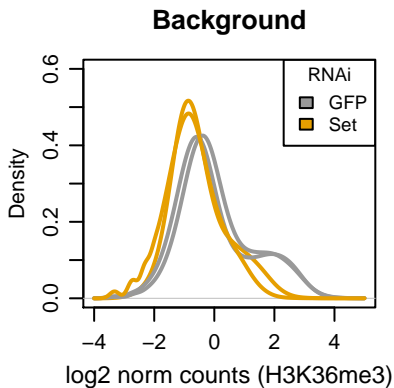
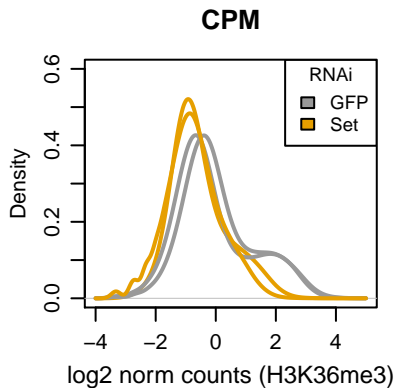
# Background Normalization



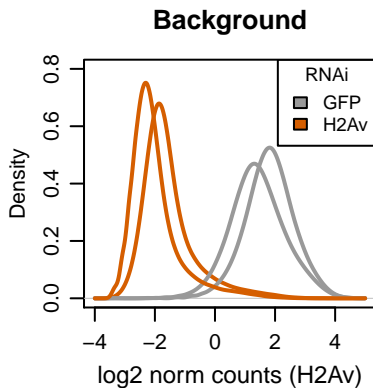
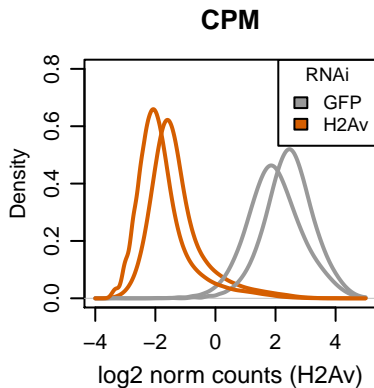
# Spike-In Normalization

- ▶ Spike-In chromatin: synthetic or different species
- ▶ Cell number and chromatin amounts have to be constant!
- ▶ Apply CPM or BG normalization on Spike-In reads

# Spike-In Normalization



# Spike-In Normalization





# Spike-In Normalization

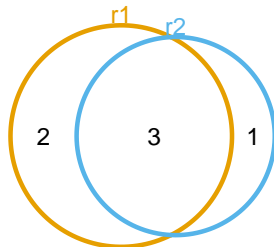
- ▶ When to use Spike-Ins?
  - ▶ global effect
  - ▶ effect has to be larger than variability
  - ▶ more replicates might be required

# Overview

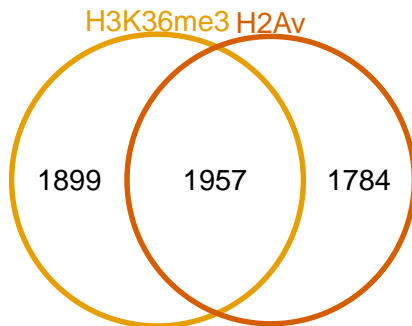
- ▶ Introduction
- ▶ ChIP-seq Coverage
- ▶ Normalization Methods
- ▶ **Peak Overlaps**
- ▶ Statistical Analysis

# Peak Overlaps

## ► overlap counting rules



# Peak Overlaps



- ▶ reviewers question: is this significant?

# Peak Overlaps

- ▶ 2x2 contingency table

H2Av / H3K36me3	Yes	No
Yes	1957	1784
No	1899	NA

- ▶ what should be the number of unbound regions?
- ▶ use gene-based approach (unbound genes)?
- ▶ formula?

$$n = \frac{GenomeSize * (Fraction_{coding} + Fraction_{regulatory})}{(2 * PeakWidth)}$$

# Peak Overlaps

H2Av / H3K36me3	Yes	No
Yes	1957	1784
No	1899	8966

Odds Ratio 95% CI = 4.77 - 5.62

Fisher's exact test p-value < 2.2e-16

- ▶ What is wrong here?

## Peak Overlaps

H2Av / H3K36me3	Yes	No
Yes	1957	1784
No	1899	2000

Odds Ratio 95% CI = 1.06 - 1.27

Fisher's exact test p-value < 1.7e-03

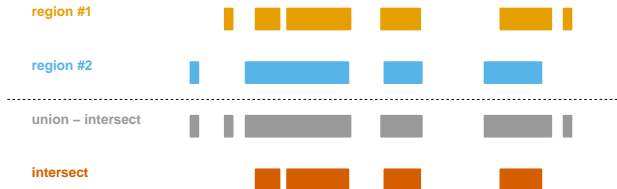
- What is wrong here?

# Peak Overlaps

- ▶ Fisher's exact test
  - ▶ hard to interpret such p-values
  - ▶ p-value is highly dependent on 'N'
  - ▶ main problem: 'N' is number of peaks
  - ▶ peaks are likely not independent
  - ▶ no information about replicates
- ▶ NOT recommended



# Peak Overlaps



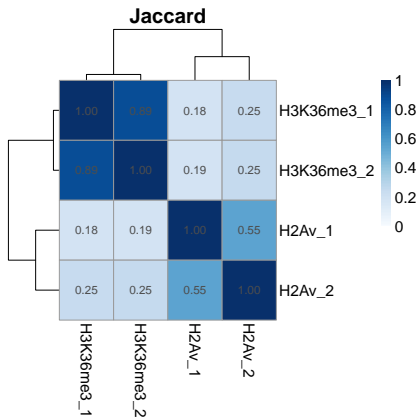
## ► Jaccard Similarity Index

$$Jaccard = \frac{Length_{intersect}}{Length_{union} - Length_{intersect}}$$

- Value: 0 - 1
- Example: 0.69

# Peak Overlaps

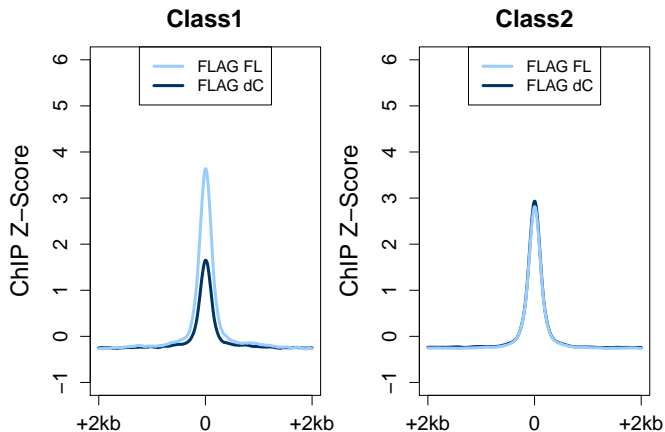
## ► H3K36me3 vs H2Av



# Overview

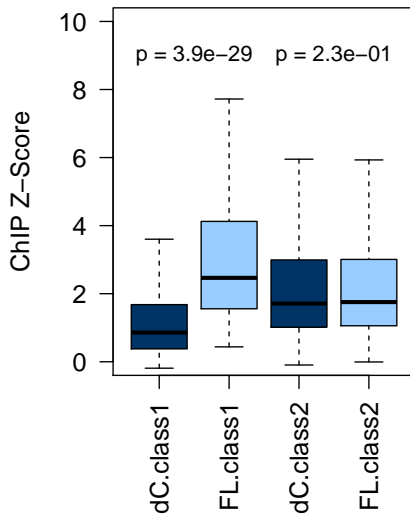
- ▶ Introduction
- ▶ ChIP-seq Coverage
- ▶ Normalization Methods
- ▶ Peak Overlaps
- ▶ **Statistical Analysis**

# Statistical Analysis



► reviewers question: is this significant?

# Statistical Analysis



► What is wrong here?

- ▶ Wilcoxon rank sum test

- ▶ What is N?

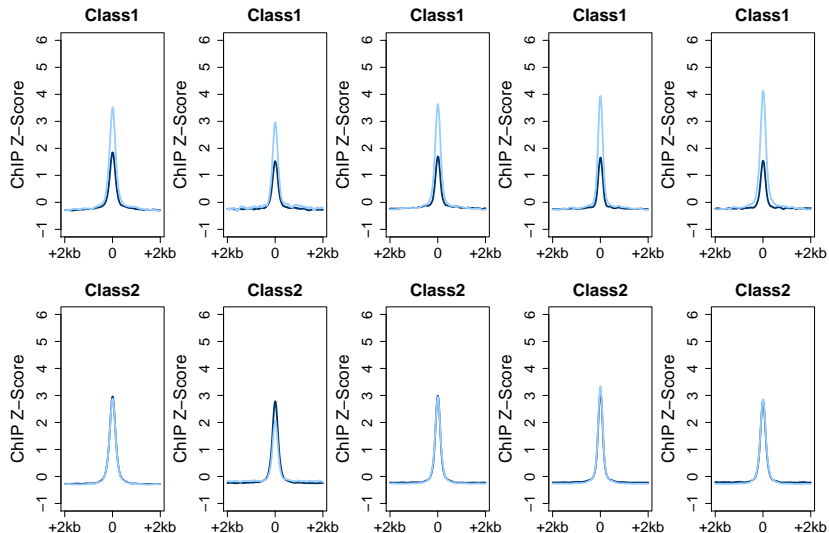
$$N_{class1} = 230, N_{class2} = 2067$$

- ▶ peaks are likely not independent

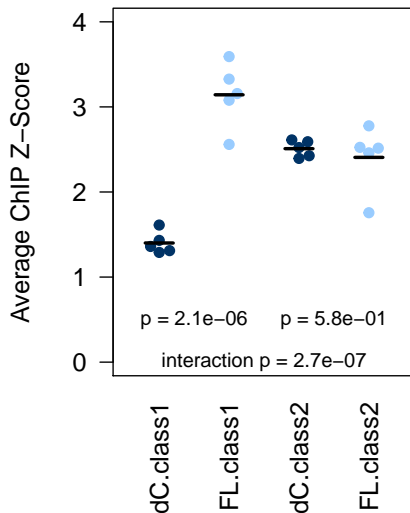
- ▶ no replicate information!

- ▶ NOT recommended

# Statistical Analysis



# Statistical Analysis



- linear mixed effect model



# Overview

- ▶ Introduction
- ▶ ChIP-seq Coverage
- ▶ Normalization Methods
- ▶ Peak Overlaps
- ▶ Statistical Analysis

# Acknowledgements

- ▶ **BMC, Bioinformatics**

- ▶ Tobias Straub

- ▶ **BMC, Molecular Biology**

- ▶ Alessandro Scacchetti

- ▶ Andrea Lukacs

- ▶ Catherine Regnard

