Bayesian multi-scale Poisson models for analyses of high-throughput sequencing data

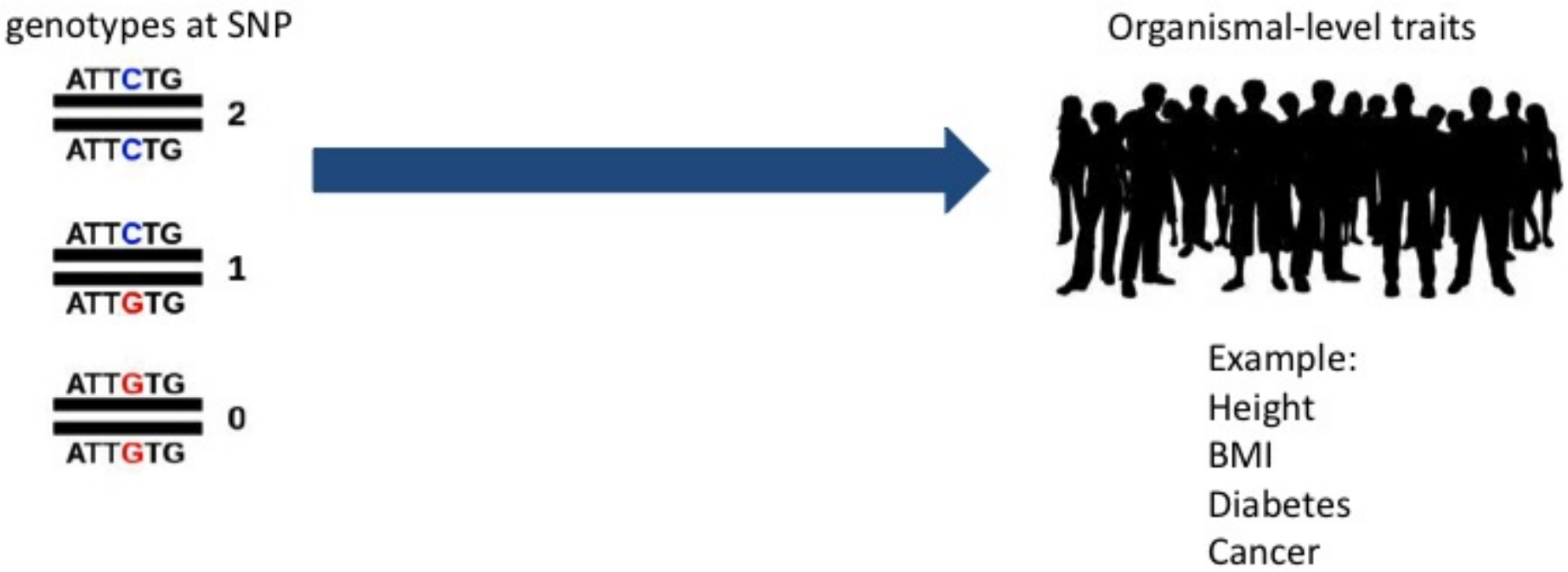
Heejung Shim

School of Mathematics and Statistics
Melbourne Integrative Genomics
University of Melbourne

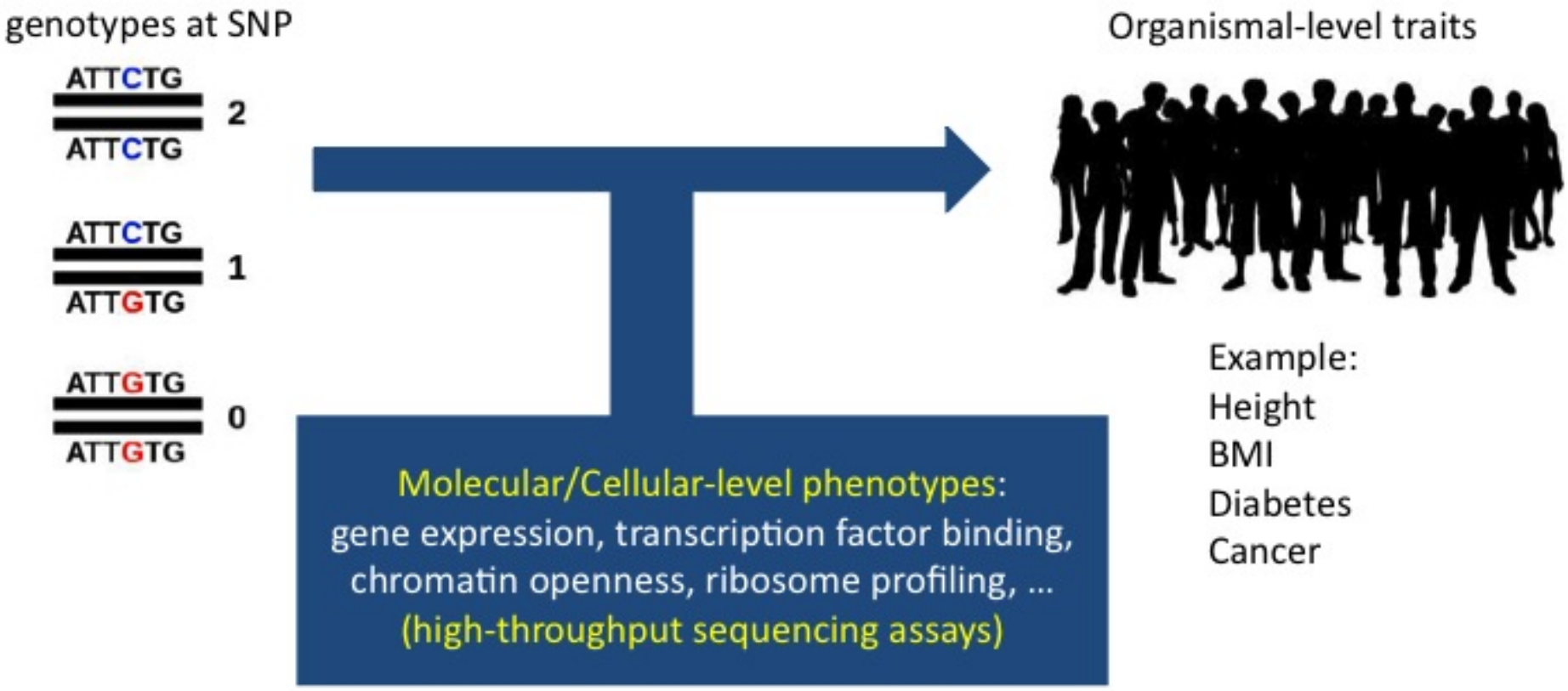
Joint work with Zhengrong Xing and Matthew Stephens

November 27, 2018

Genetic basis of complex traits



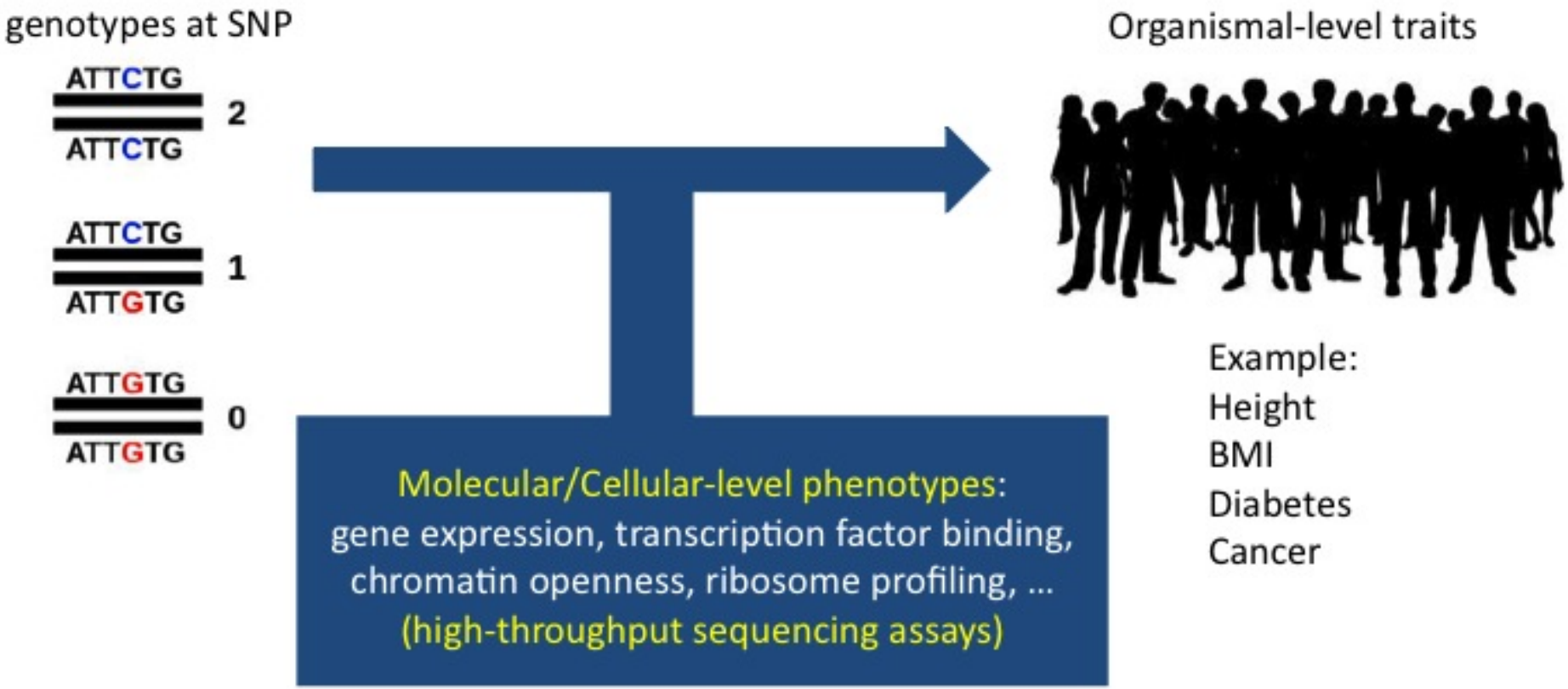
Genetic basis of complex traits



Genetic basis of complex traits

Problem of interest: detect/estimate differences in molecular-level phenotypes between multiple groups of samples using high-throughput sequencing data

- e.g., differential gene expression analysis, eQTLs analysis



High-throughput sequencing data: a very simple illustration

Example

RNA-seq : gene expression

DNase-seq and ATAC-seq :
chromatin accessibility

ChIP-seq : transcription factor binding

CAGE-seq : transcription start site usage

High-throughput sequencing data: a very simple illustration

Reads:

~ 20-70 bases for each read
~ 40M reads for each sample
(sequencing depth)

Sequencing Machine

Example
RNA-seq : gene expression
DNase-seq and ATAC-seq : chromatin accessibility
ChIP-seq : transcription factor binding
CAGE-seq : transcription start site usage

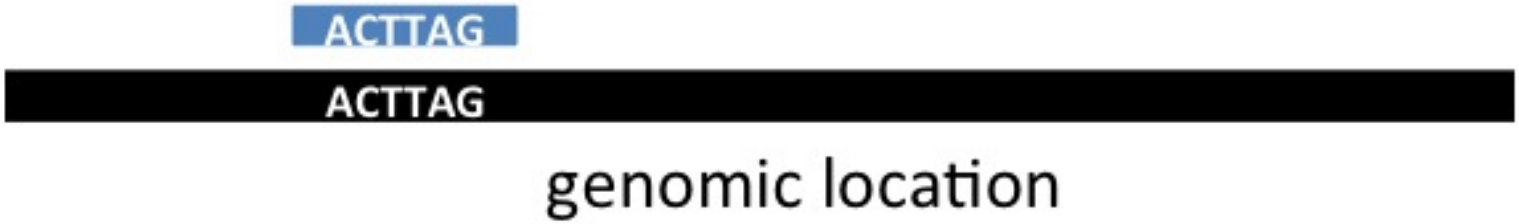
High-throughput sequencing data: a very simple illustration

Reads:

~ 20-70 bases for each read

~ 40M reads for each sample (sequencing depth)

Map reads:



High-throughput sequencing data: a very simple illustration

Reads:

~ 20-70 bases for each read
~ 40M reads for each sample (sequencing depth)

Map reads:

High-throughput sequencing data: a very simple illustration

Reads:

~ 20-70 bases for each read
~ 40M reads for each sample
(sequencing depth)

Map reads:

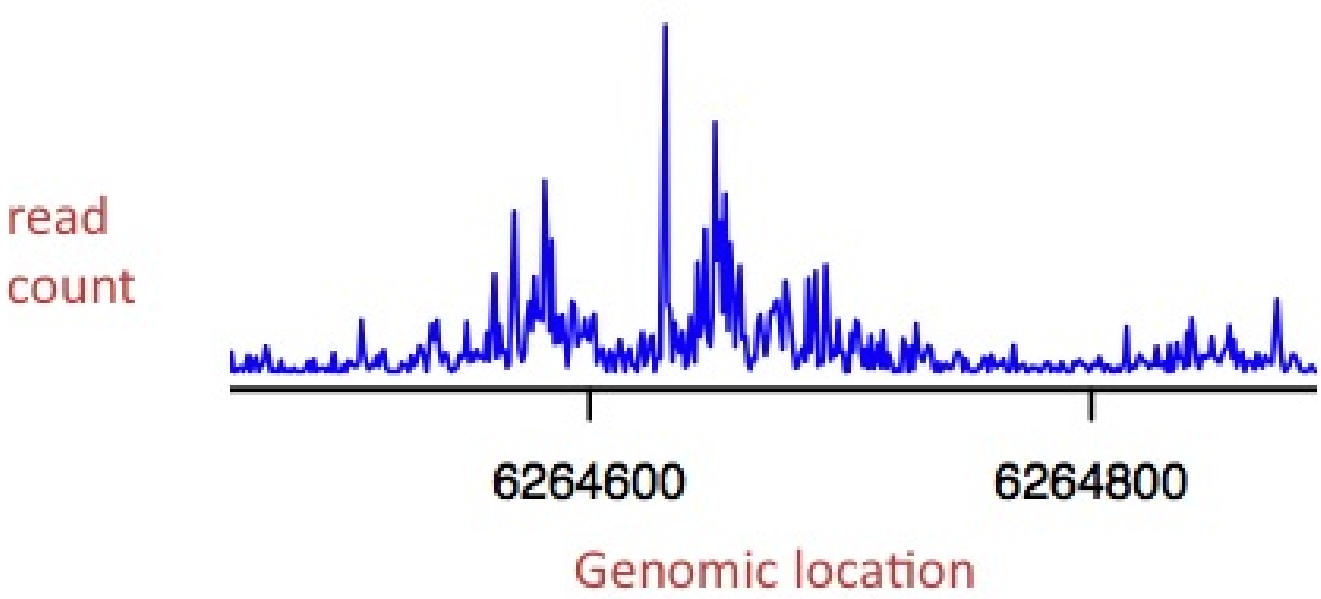
genomic location

Data: . . 0 1 1 0 2 1 0 0 1 1 1 0 1 1 0 0 0 0 0 1 1 3 0 1 0 . .

# of reads which start at each base ≈ level of trait at each base

High-throughput sequencing data

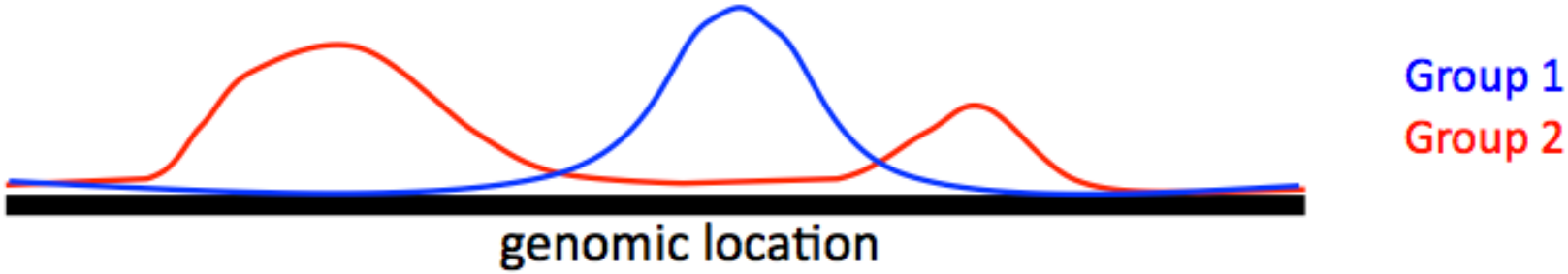
Common form: each sample consists of the number of reads which start at each base across the whole genome (3 billion bases for human genome).



- **Functional data** and **count data**.

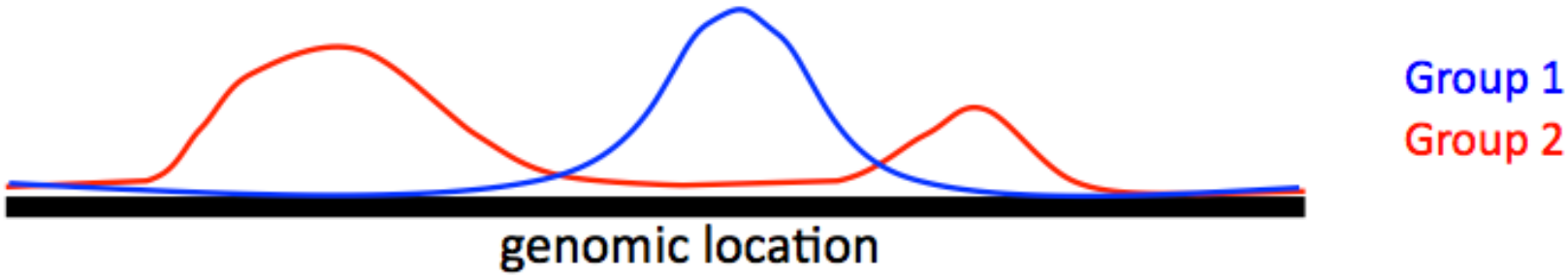
Problem of interest: detect/estimate differences in molecular-level phenotypes between multiple groups of samples using high-throughput sequencing data

Molecular-level phenotypes from base 1 to base T in a given region:



Problem of interest: detect/estimate differences in molecular-level phenotypes between multiple groups of samples using high-throughput sequencing data

Molecular-level phenotypes from base 1 to base T in a given region:



Data: multiple independent samples.

Sample id	High-throughput sequencing data from base 1 to base T				Group indicator
1	X_1^1	X_2^1	...	X_T^1	g^1
i	X_1^i	X_2^i	...	X_T^i	g^i
I	X_1^I	X_2^I	...	X_T^I	g^I

Typical approach: window based approach

- Define a window: genes or windows of fixed length (e.g, 100 bases)
- Use total number of reads mapped to the window.
  - Simple linear regression: Pickrell et al, 2010, Degner et al, 2012, ...
  - Negative Binomial model: edgeR (Robinson et al, 2010), DESeq2 (Love et al, 2014), ...

Typical approach: window based approach

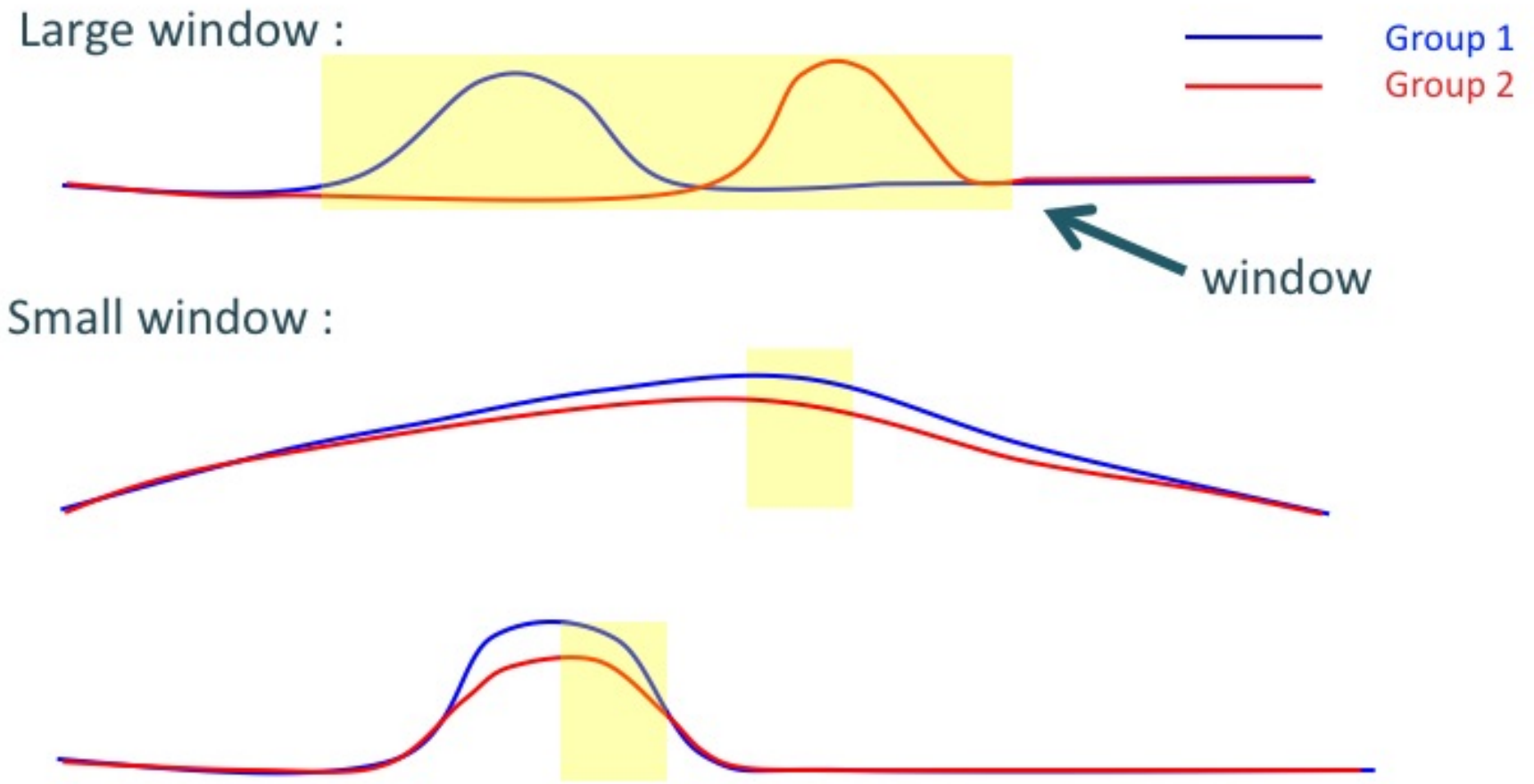
- Define a window: genes or windows of fixed length (e.g, 100 bases)
- Use total number of reads mapped to the window.
  - Simple linear regression: Pickrell et al, 2010, Degner et al, 2012, ...
  - Negative Binomial model: edgeR (Robinson et al, 2010), DESeq2 (Love et al, 2014), ...
- Fail to exploit high-resolution information.

Typical approach: window based approach

- Define a window: genes or windows of fixed length (e.g, 100 bases)
- Use total number of reads mapped to the window.
- Fail to exploit high-resolution information

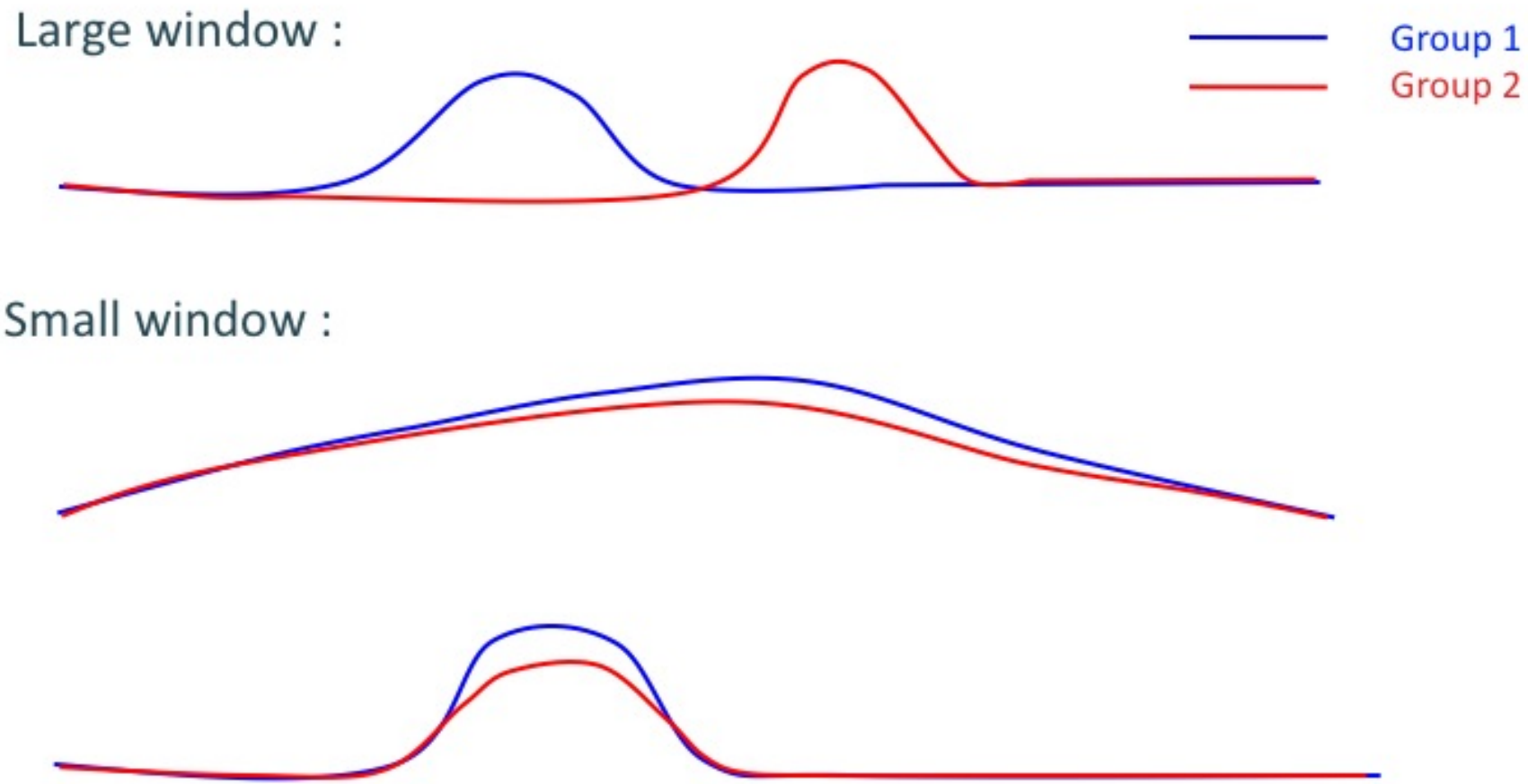
Typical approach: window based approach

- Define a window: genes or windows of fixed length (e.g, 100 bases)
- Use total number of reads mapped to the window.
- Fail to exploit high-resolution information



Typical approach: window based approach

- Define a window: genes or windows of fixed length (e.g, 100 bases)
- Use total number of reads mapped to the window.
- Fail to exploit high-resolution information

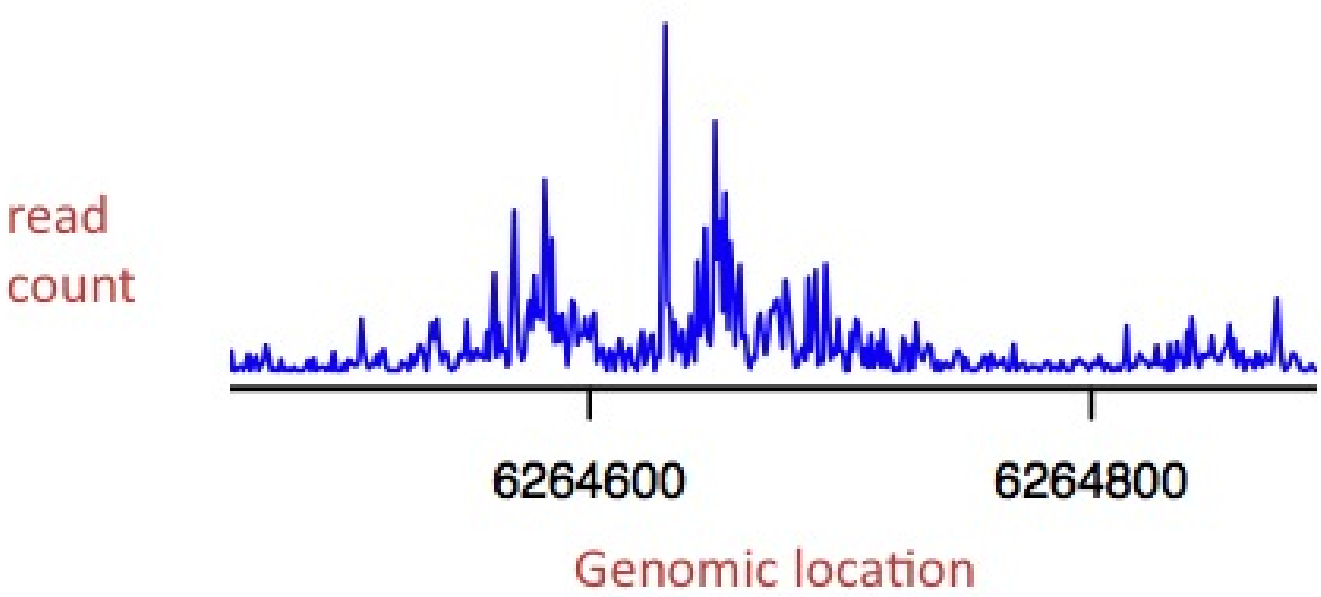


Multi-scale approaches use high-resolution information

Use high-resolution information?

no (window approach)	yes (multi-scale approach)
Pickrell et al, 2010, Degner et al, 2012, etc.	**Wavelet-based approach, Shim and Stephens, 2015**
DESeq2, edgeR, etc.	**Shim et al, In preparation**

Why multi-scale approaches?



- High-throughput sequencing data: very noisy measurements of an underlying molecular phenotype.
- The molecular phenotype is spatially structured and has a lot of local structure (inhomogeneous along the genome).

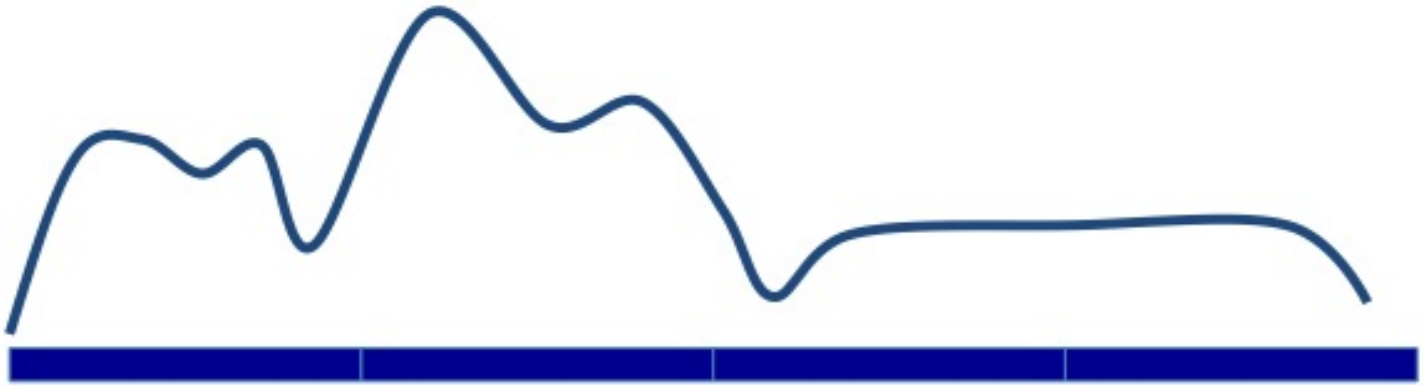
Key ideas: multi-scale transform

Molecular phenotype is spatially structured and has a lot of local structure.

$$\lambda_1, \quad \ldots \quad , \lambda_T$$

Key ideas: multi-scale transform

Molecular phenotype is spatially structured and has a lot of local structure.

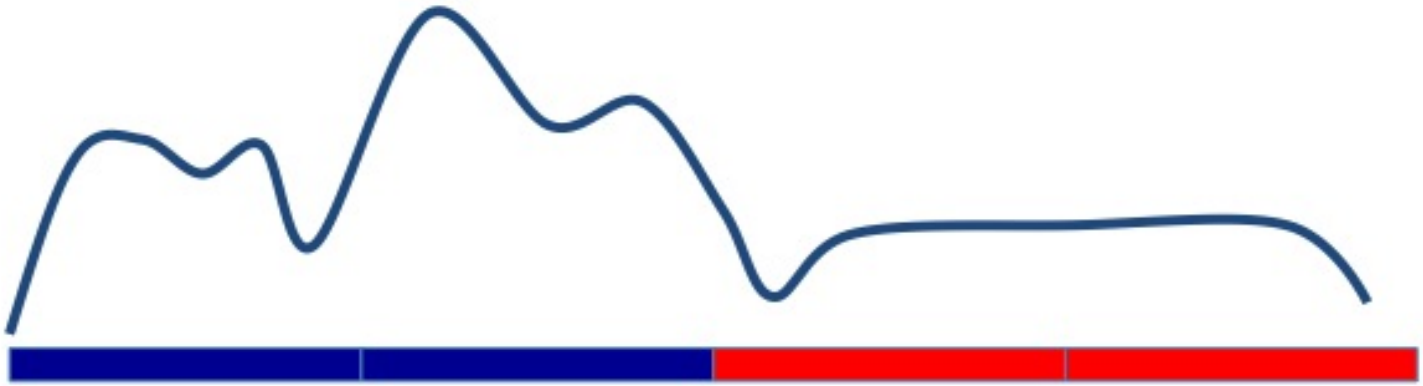


$$\lambda_1, \quad \ldots \quad , \lambda_T$$

$$\theta_{\mathbf{01}} = \sum_{i=1}^{T} \lambda_i$$

Key ideas: multi-scale transform

Molecular phenotype is spatially structured and has a lot of local structure.



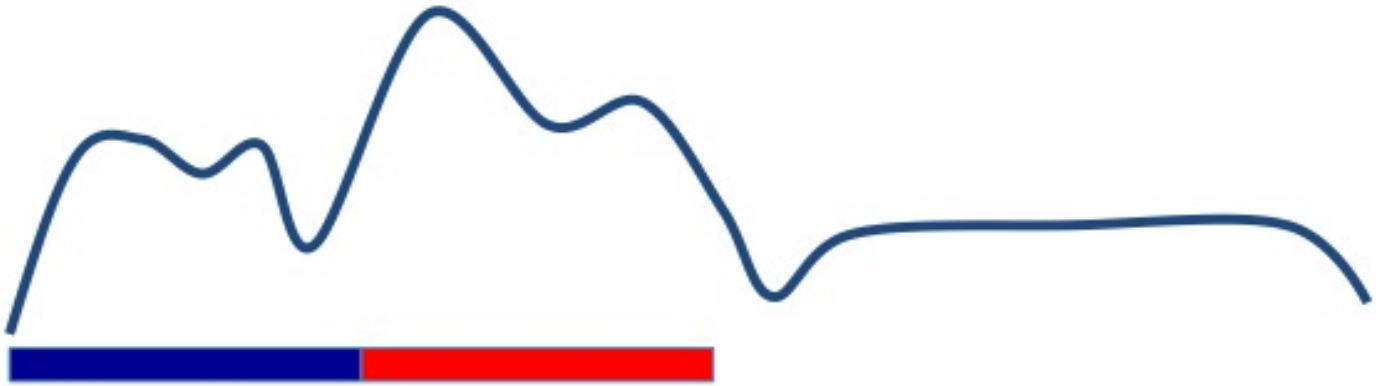
$$\lambda_1, \quad \dots \quad , \lambda_T$$

$$\theta_{\mathbf{01}} = \sum_{i=1}^{T} \lambda_i$$

$$\theta_{\mathbf{11}} = \sum_{i=1}^{T/2} \lambda_i - \sum_{i=T/2+1}^{T} \lambda_i$$

Key ideas: multi-scale transform

Molecular phenotype is spatially structured and has a lot of local structure.



$$\lambda_1, \quad \dots \quad , \lambda_T$$

$$\theta_{\mathbf{01}} = \sum_{i=1}^{T} \lambda_i$$

$$\theta_{\mathbf{11}} = \sum_{i=1}^{T/2} \lambda_i - \sum_{i=T/2+1}^{T} \lambda_i$$

$$\theta_{\mathbf{21}} = \sum_{i=1}^{T/4} \lambda_i - \sum_{i=T/4+1}^{T/2} \lambda_i$$

Key ideas: multi-scale transform

Molecular phenotype is spatially structured and has a lot of local structure.

$$\lambda_1, \quad \ldots \quad , \lambda_T$$

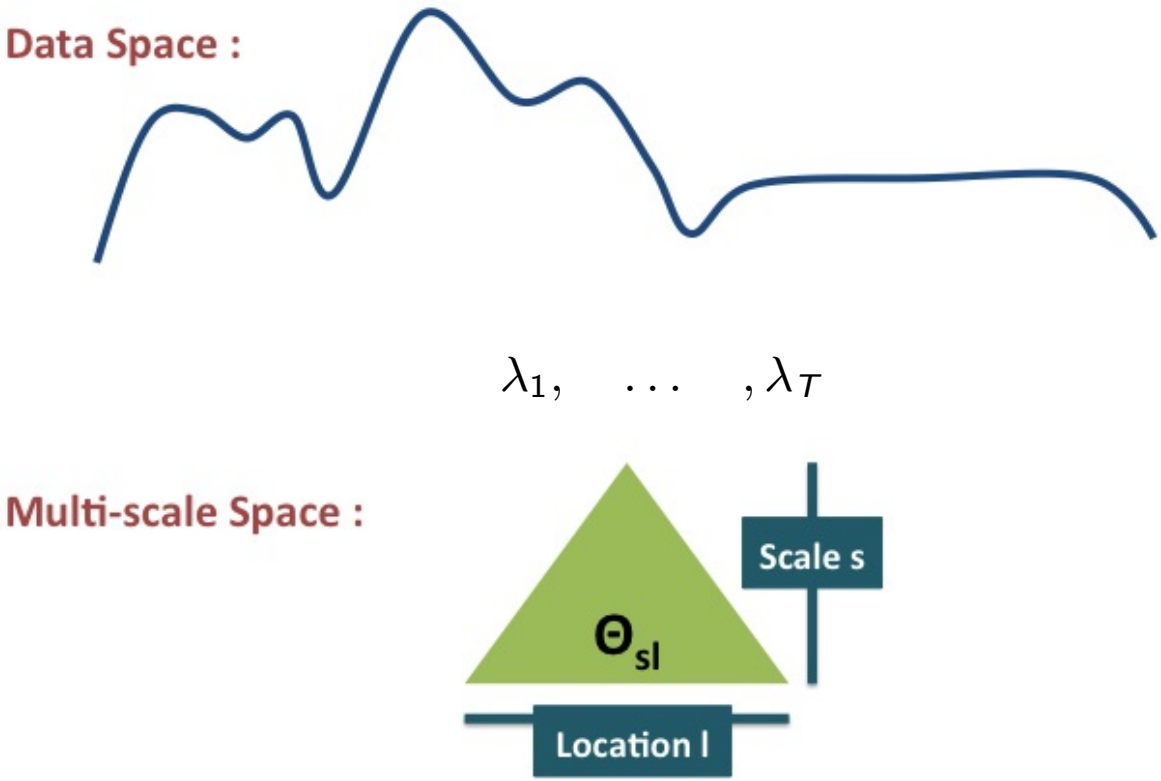
$$\theta_{\mathbf{01}} = \sum_{i=1}^{T} \lambda_i$$

$$\theta_{\mathbf{11}} = \sum_{i=1}^{T/2} \lambda_i - \sum_{i=T/2+1}^{T} \lambda_i$$

$$\theta_{\mathbf{21}} = \sum_{i=1}^{T/4} \lambda_i - \sum_{i=T/4+1}^{T/2} \lambda_i \qquad \theta_{\mathbf{22}} = \sum_{i=T/2+1}^{3T/4} \lambda_i - \sum_{i=3T/4+1}^{T} \lambda_i$$

Key ideas: multi-scale transform

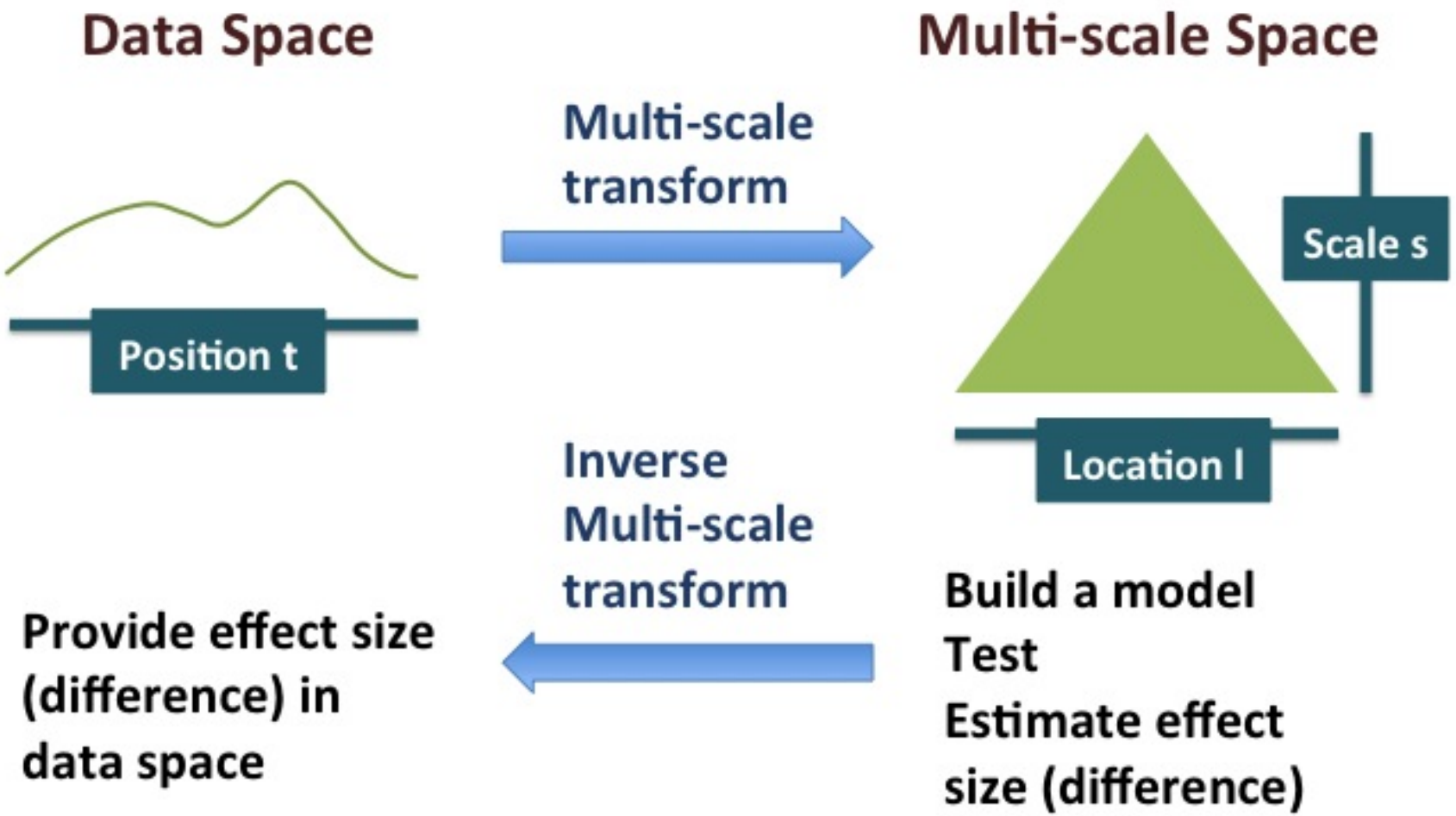
Molecular phenotype is spatially structured and has a lot of local structure.



The multi-scale transform is 1-1, but important advantages:

- Spatial structure in λ implies sparsity in θ.
- Easy to capture local structure.

Multi-scale approaches



Multi-scale approaches use high-resolution information

		Use high-resolution information? No (window approach)	Use high-resolution information? Yes (multi-scale approach)
Model count data?	No	Pickrell et al, 2010, Degner et al, 2012, etc.	Wavelet-based approach Shim and Stephens, 2015
	Yes	DESeq2, edgeR, etc	Shim et al, In preparation

Wavelet-based ("normal") multi-scale approach

- Software **WaveQTL** : https://github.com/heejungshim/WaveQTL
- We demonstrated that WaveQTL has more power than simpler window-based approaches (sample size: 70).
- Potential limitations in application to small sample sizes or low read count data.

Multi-scale approaches use high-resolution information

		Use high-resolution information? No (window approach)	Use high-resolution information? Yes (multi-scale approach)
Model count data?	No	Pickrell et al, 2010, Degner et al, 2012, etc.	Wavelet-based approach Shim and Stephens, 2015
	Yes	DESeq2, edgeR, etc	Shim et al, In preparation

Multi-scale model for count data

- Model the count nature of the sequencing data directly.
- Software **multiseq** : https://github.com/stephenslab/multiseq

multiseq: multi-scale method to model
multiple samples of functional count data with a covariate.

multi-scale model for inhomogeneous Poisson processes: one sample

Kolaczyk 1999 and Timmermann and Nowak 1999

$$\mathrm{P}(X_1, \ldots, X_T \mid \lambda_1, \ldots, \lambda_T) = \prod_t Pois(X_t \mid \lambda_t) \quad T = 2^J$$

multi-scale model for inhomogeneous Poisson processes: one sample

Kolaczyk 1999 and Timmermann and Nowak 1999

$$P(X_1, \ldots, X_T \mid \lambda_1, \ldots, \lambda_T) = \prod_t Pois(X_t \mid \lambda_t) \quad T = 2^J$$

$$= P(X_1, \ldots, X_T \mid \mu_0, p_{11}, p_{21}, p_{22}, \ldots, p_{J1}, \ldots, p_{J,2^{J-1}})$$

$$\mu_0 : \lambda_1 + \ldots + \lambda_T$$

$$p_{11}: \frac{\lambda_1+\ldots+\lambda_{T/2}}{\lambda_1+\ldots+\lambda_T}$$

$$p_{21}: \frac{\lambda_1+\ldots+\lambda_{T/4}}{\lambda_1+\ldots+\lambda_{T/2}} \qquad p_{22}: \frac{\lambda_{T/2+1}+\ldots+\lambda_{3T/4}}{\lambda_{T/2+1}+\ldots+\lambda_T} \qquad \ldots$$

multi-scale model for inhomogeneous Poisson processes: one sample

Kolaczyk 1999 and Timmermann and Nowak 1999

$$\mathrm{P}(X_1, \ldots, X_T \mid \lambda_1, \ldots, \lambda_T) = \prod_t Pois(X_t \mid \lambda_t) \quad T = 2^J$$

$$= \mathrm{P}(X_1, \ldots, X_T \mid \mu_0, p_{11}, p_{21}, p_{22}, \ldots, p_{J1}, \ldots, p_{J,2^{J-1}})$$

$$\mu_0 : \lambda_1 + \ldots + \lambda_T$$

$$p_{11}: \frac{\lambda_1+\ldots+\lambda_{T/2}}{\lambda_1+\ldots+\lambda_T}$$

$$p_{21}: \frac{\lambda_1+\ldots+\lambda_{T/4}}{\lambda_1+\ldots+\lambda_{T/2}} \qquad p_{22}: \frac{\lambda_{T/2+1}+\ldots+\lambda_{3T/4}}{\lambda_{T/2+1}+\ldots+\lambda_T} \qquad \ldots$$

$$= Pois(\sum_{t=1}^{T} X_t \mid \mu_0) Binomial(\sum_{t=1}^{T/2} X_t \mid \sum_{t=1}^{T} X_t, p_{11})$$

$$Binomial(\sum_{t=1}^{T/4} X_t \mid \sum_{t=1}^{T/2} X_t, p_{21}) Binomial(\sum_{t=T/2+1}^{3T/4} X_t \mid \sum_{t=T/2+1}^{T} X_t, p_{22}), \ldots$$

multi-scale model for inhomogeneous Poisson processes: one sample

Kolaczyk 1999 and Timmermann and Nowak 1999

$$\mathrm{P}(X_1, \ldots, X_T \mid \lambda_1, \ldots, \lambda_T) = \prod_t Pois(X_t \mid \lambda_t) \quad T = 2^J$$

$$= \mathrm{P}(X_1, \ldots, X_T \mid \mu_0, p_{11}, p_{21}, p_{22}, \ldots, p_{J1}, \ldots, p_{J,2^{J-1}})$$

$$\mu_0 : \lambda_1 + \ldots + \lambda_T$$

$$p_{11}: \frac{\lambda_1+\ldots+\lambda_{T/2}}{\lambda_1+\ldots+\lambda_T}$$

$$p_{21}: \frac{\lambda_1+\ldots+\lambda_{T/4}}{\lambda_1+\ldots+\lambda_{T/2}} \qquad p_{22}: \frac{\lambda_{T/2+1}+\ldots+\lambda_{3T/4}}{\lambda_{T/2+1}+\ldots+\lambda_T} \qquad \ldots$$

$$= Pois(\sum_{t=1}^{T} X_t \mid \mu_0) Binomial(\sum_{t=1}^{T/2} X_t \mid \sum_{t=1}^{T} X_t, p_{11})$$

$$Binomial(\sum_{t=1}^{T/4} X_t \mid \sum_{t=1}^{T/2} X_t, p_{21}) Binomial(\sum_{t=T/2+1}^{3T/4} X_t \mid \sum_{t=T/2+1}^{T} X_t, p_{22}), \ldots$$

g^i: a group indicator of sample i

multiseq: extension of window based methods

model at the zeroth scale (i.e., s = 0):

$$\sum_{t=1}^{T} X_t^i \sim Pois(\mu_0^i), \quad \text{where} \quad \mu_0^i = \sum_{t=1}^{T} \lambda_t^i$$

- Model additional variation across multiple samples.
- Negative binomial models considered in window based methods.

multiseq: model multiple samples with covariate

For each scale s and location l, we model the potential association by

$$logit(p_{sl}^{i}) = \alpha_{sl} + \beta_{sl} g^{i} + \epsilon_{sl}^{i},$$

where g^{i} is a group indicator of sample i.

multiseq: model multiple samples with covariate

For each scale s and location l, we model the potential association by

$$logit(p_{sl}^i) = \alpha_{sl} + \beta_{sl} g^i + \epsilon_{sl}^i,$$

where g^i is a group indicator of sample i.

We place the following prior on β_{sl}:

$$\begin{aligned} \beta_{sl} &\sim \gamma_{sl} N(0, \tau_{sl}^2) + (1 - \gamma_{sl})\delta_0, \\ \gamma_{sl} &\sim \text{Bernoulli}(\pi_s). \end{aligned}$$

where δ_0 is a point mass at zero.

multiseq: model multiple samples with covariate

For each scale s and location l, we model the potential association by

$$logit(p_{sl}^i) = \alpha_{sl} + \beta_{sl} g^i + \epsilon_{sl}^i,$$

where g^i is a group indicator of sample i.

We place the following prior on β_{sl}:

$$\begin{aligned} \beta_{sl} &\sim \gamma_{sl} N(0, \tau_{sl}^2) + (1 - \gamma_{sl})\delta_0, \\ \gamma_{sl} &\sim \text{Bernoulli}(\pi_s). \end{aligned}$$

where δ_0 is a point mass at zero.

In practice:

- Approximate the likelihood by a Normal likelihood.
- Prior: a mixture of a point mass at zero and multiple normal distributions with known variances (the ashr "Adaptive SHrinkage" package, Stephens, 2016).

multiseq: model multiple samples with covariate

For each scale s and location l, we model the potential association by

$$logit(p^i_{sl}) = \alpha_{sl} + \beta_{sl} g^i + \epsilon^i_{sl},$$

where g^i is a group indicator of sample i.

We place the following prior on β_{sl}:

$$\begin{aligned} \beta_{sl} &\sim \gamma_{sl} N(0, \tau^2_{sl}) + (1 - \gamma_{sl})\delta_0, \\ \gamma_{sl} &\sim \text{Bernoulli}(\pi_s). \end{aligned}$$

where δ_0 is a point mass at zero.

To detect difference:

- Posterior joint alternative probability : $1 - \text{P}(\gamma_{sl} = 0 \ \ \forall s, l | X)$

multiseq: model multiple samples with covariate

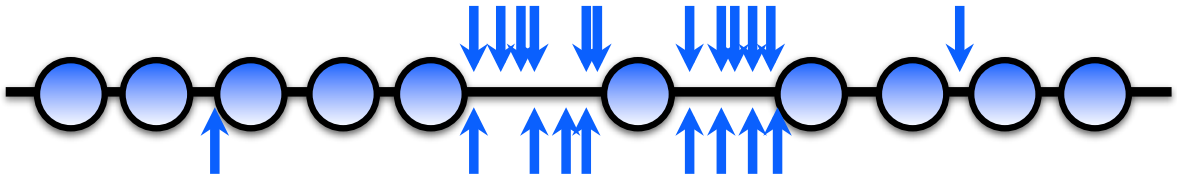
To explain observed difference/association:

- $P(\beta_{sl} \mid X)$: a mixture of a point mass at zero and normal distributions.
- Provide posterior mean and variance on difference (log scale) in the data space (approximation by using Taylor expansion).
- Other types of posterior inference (e.g. pointwise credible intervals): sampling procedure.

Analysis of ATAC-seq data from Luca/Pique-Regi group

ATAC-seq measures chromatin accessibility

- Tn5 transposase cuts DNA more often in regions that are accessible.

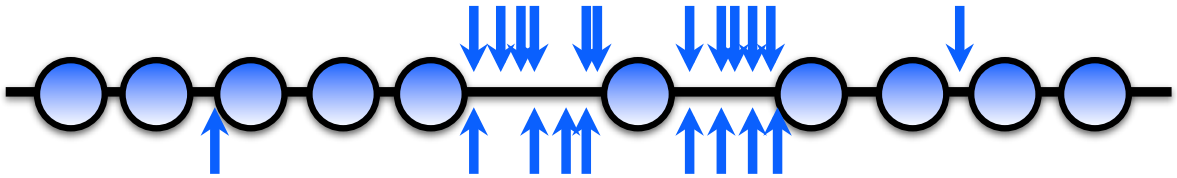


- Higher ATAC-seq read count corresponds to higher chromatin accessibility at each base.

Analysis of ATAC-seq data from Luca/Pique-Regi group

ATAC-seq measures chromatin accessibility

- Tn5 transposase cuts DNA more often in regions that are accessible.



- Higher ATAC-seq read count corresponds to higher chromatin accessibility at each base.
- Chromatin accessibility is related to functional elements of the genome (e.g., transcription factor binding sites).

Analysis of ATAC-seq data from Luca/Pique-Regi group

- Data
  - ATAC-seq data in Copper-treated and control samples (3 vs 3)
- Question: detect regions with difference in chromatic accessibility between two groups

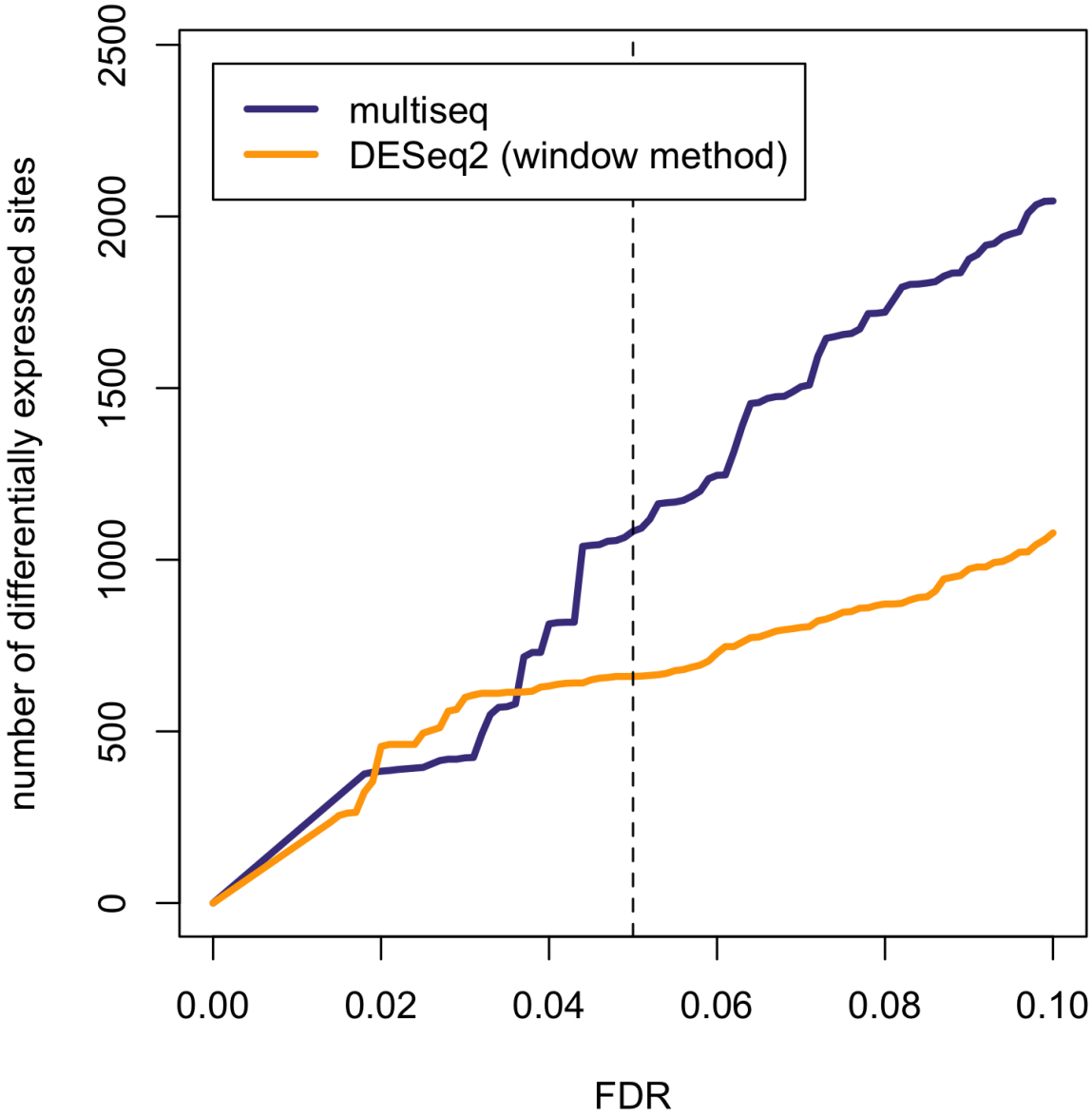
Analysis of ATAC-seq data from Luca/Pique-Regi group

- Data
  - ATAC-seq data in Copper-treated and control samples (3 vs 3)
- Question: detect regions with difference in chromatic accessibility between two groups
- Analysis
  - 237K 1024bp ($\approx$ 1kb) regions
  - For each region, test statistic
    - multiseq: posterior joint alternative probability
    - window methods (1024 bp as window size): p-value from DESeq2

Analysis of ATAC-seq data from Luca/Pique-Regi group

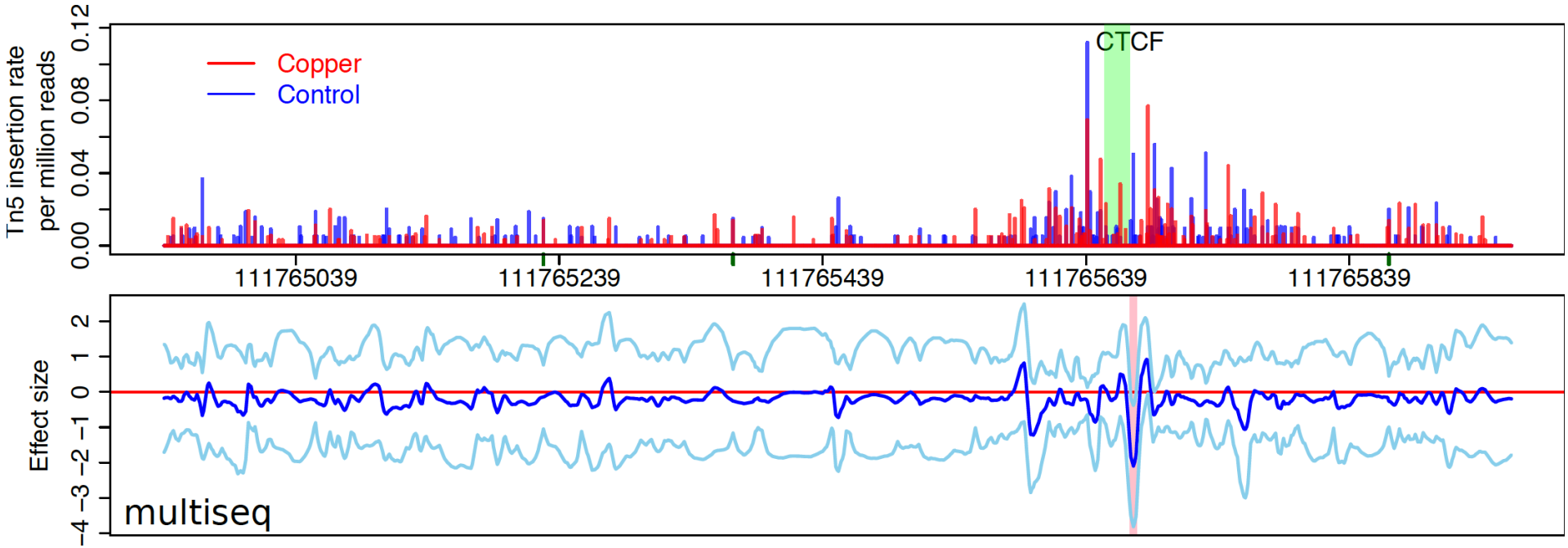
- Data
  - ATAC-seq data in Copper-treated and control samples (3 vs 3)
- Question: detect regions with difference in chromatic accessibility between two groups
- Analysis
  - 237K 1024bp ($\approx$ 1kb) regions
  - For each region, test statistic
    - multiseq: posterior joint alternative probability
    - window methods (1024 bp as window size): p-value from DESeq2
  - p-value from an empirical null distribution of test statistic.
  - FDR (the qvalue package, Storey 2002, 2003) for each method computed using p values from 237K 1kb regions.

multiseq has better power than a window approach



Differential chromatin accessibility found only by multiseq

chr1:111764939-111765962



Further information:

- More results (e.g. simulation studies, comparison with WaveQTL) can be shared after the talk.
- Shim et al, in preparation
- Software multiseq : https://github.com/stephenslab/multiseq

Summary

- Presented multi-scale methods (multiseq)
  - model the count nature of the sequencing data directly.
  - model multiple samples of functional count data with a covariate.
- Demonstrated that
  - multi-scale methods outperform window-based methods.

Discussion

- Putting dependent (Markov Tree) priors on differences in multi-scale space (Crouse et al., 1998, Ma and Soriano, 2018)

Acknowledgments

University of Chicago

- Matthew Stephens
- Zhengrong Xing

Wayne State University

- Roger Pique-Regi
- Francesca Luca