Nonparametric inference
for functional and translational genomics

Ben Brown

Statistics, UC Berkeley

Part 1

A general model of feature co-association

The Genome Structure Correction (GSC)

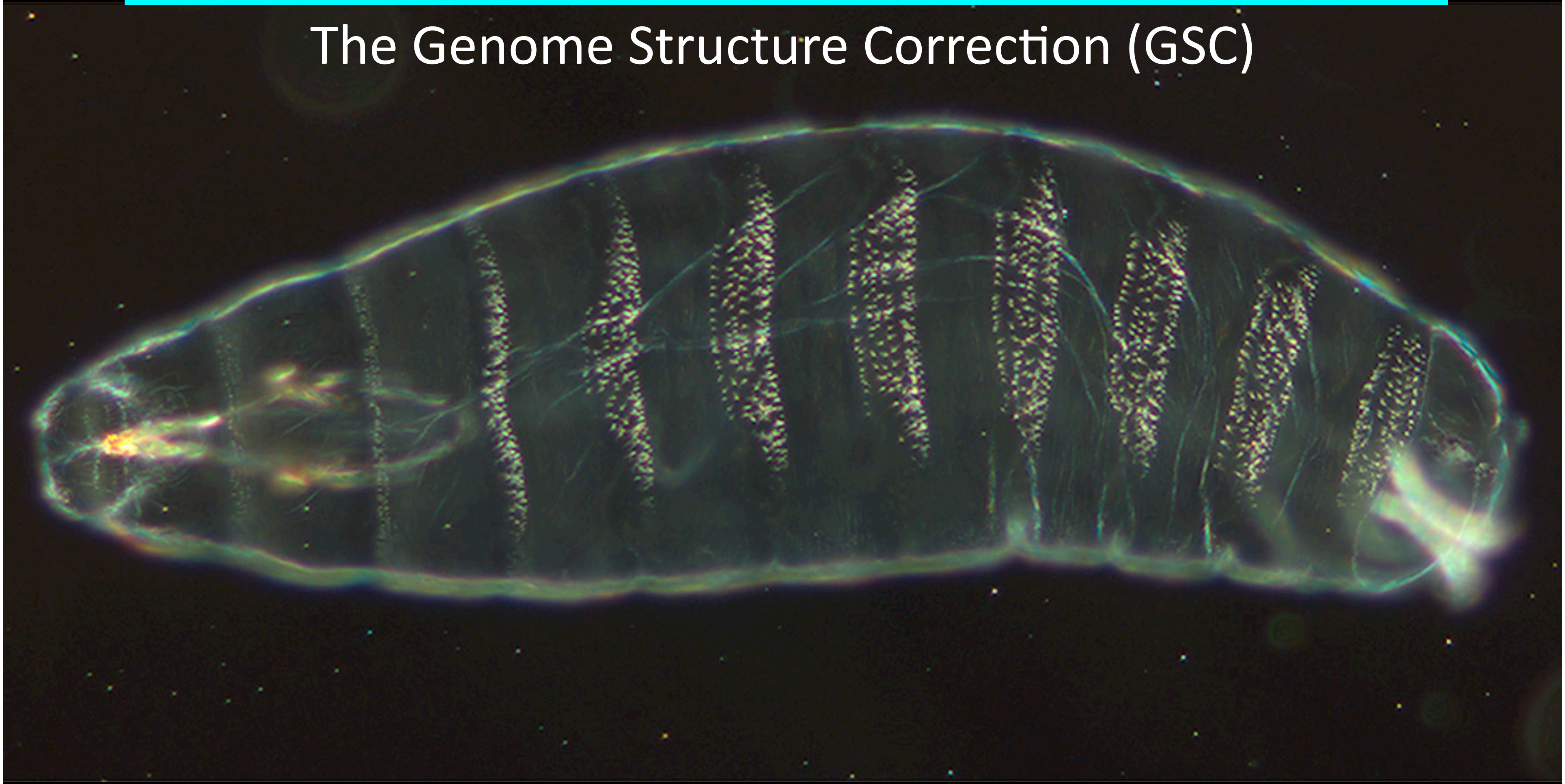
Part 2

Beyond heuristics

A generic statistical tool for the analysis of *-seq assays

A general model of feature co-association

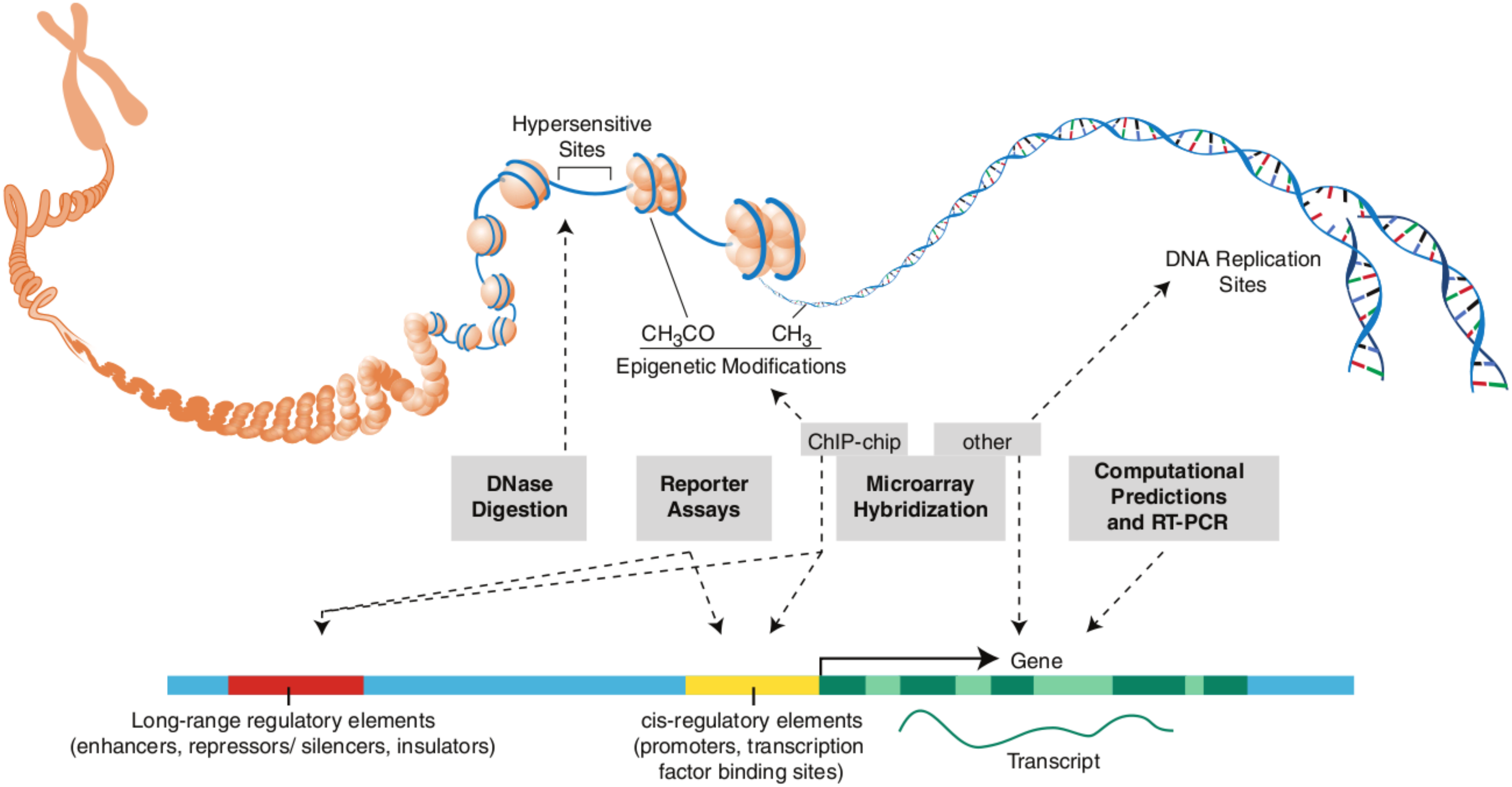
The Genome Structure Correction (GSC)



Dr. Ben Brown,

Statistics, UC Berkeley

The ENCODE Project



Hypersensitive Sites
CH_3CO
CH_3
Epigenetic Modifications
DNA Replication Sites
ChIP-chip
other
DNase Digestion
Reporter Assays
Microarray Hybridization
Computational Predictions and RT-PCR
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

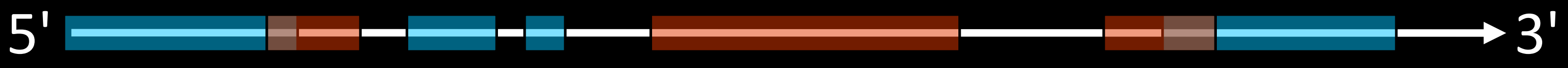
The ENCODE Project Consortium. 2004. Science 22: 306 (5696).

Feature Overlap

- Do a pair of features overlap more, or less than “expected at random”?

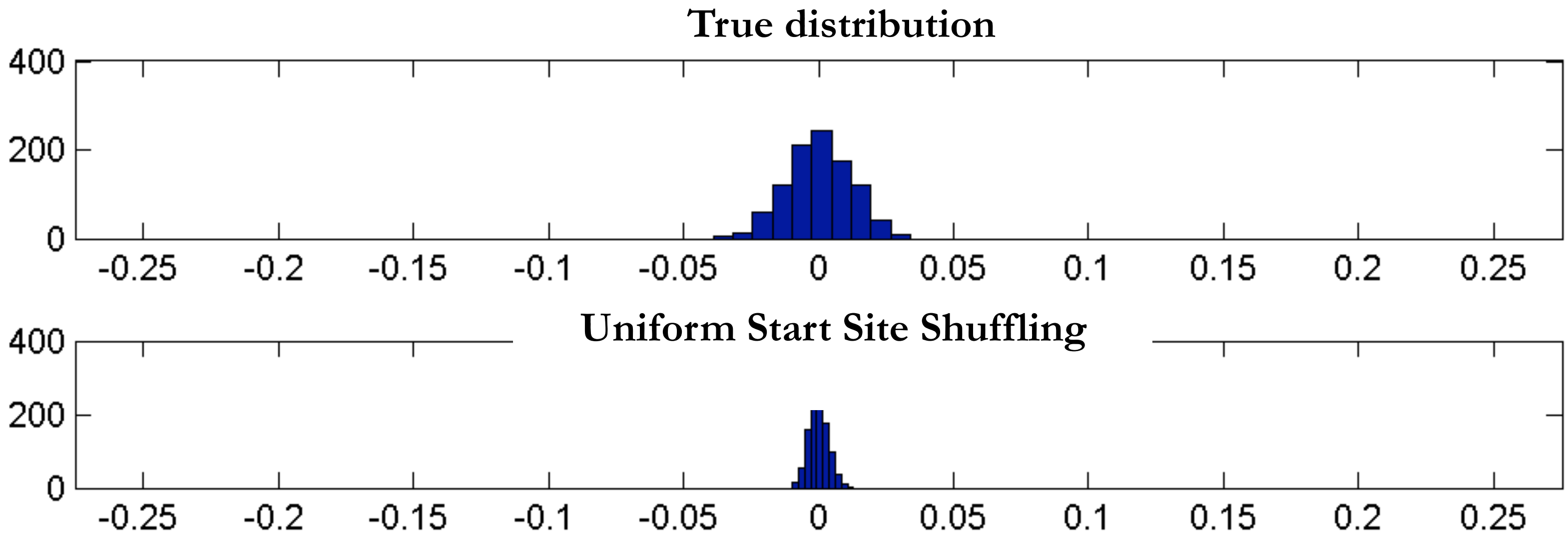
🟦 →Transcription Fragments

🟫 → Conserved sequence



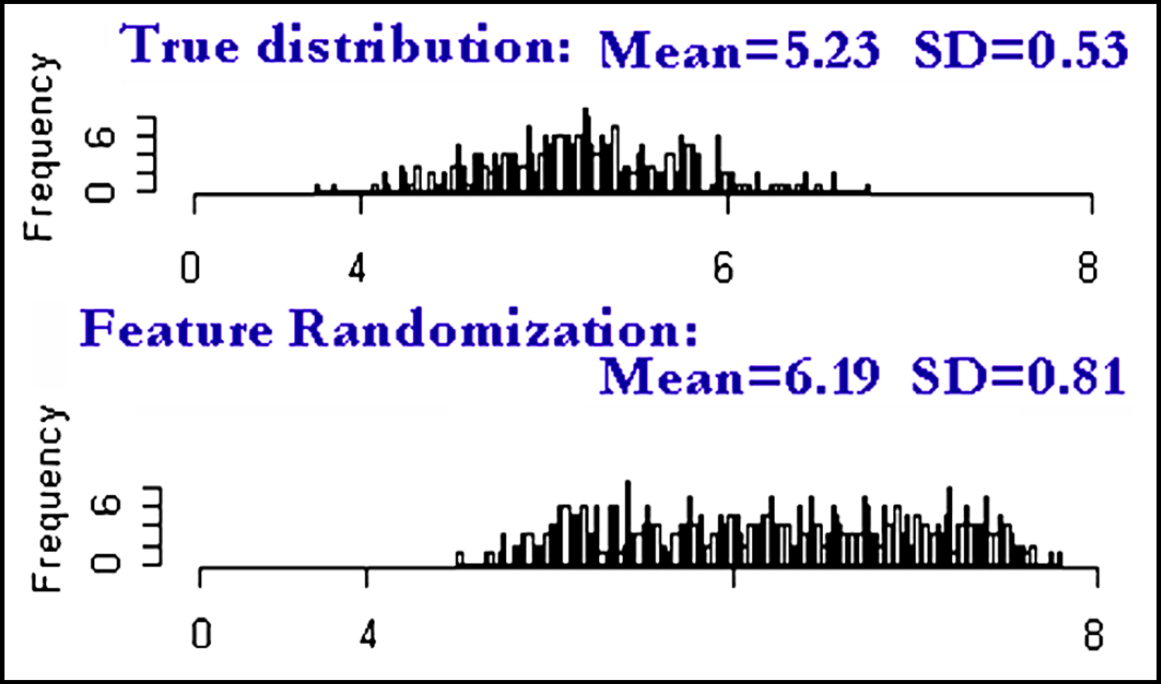
Naïve methods

- Uniform feature start site shuffling
  - – Big assumption: feature inter-arrival distances are Poisson, i.e. no big clumping, clustering, or underlying structure



Naïve methods

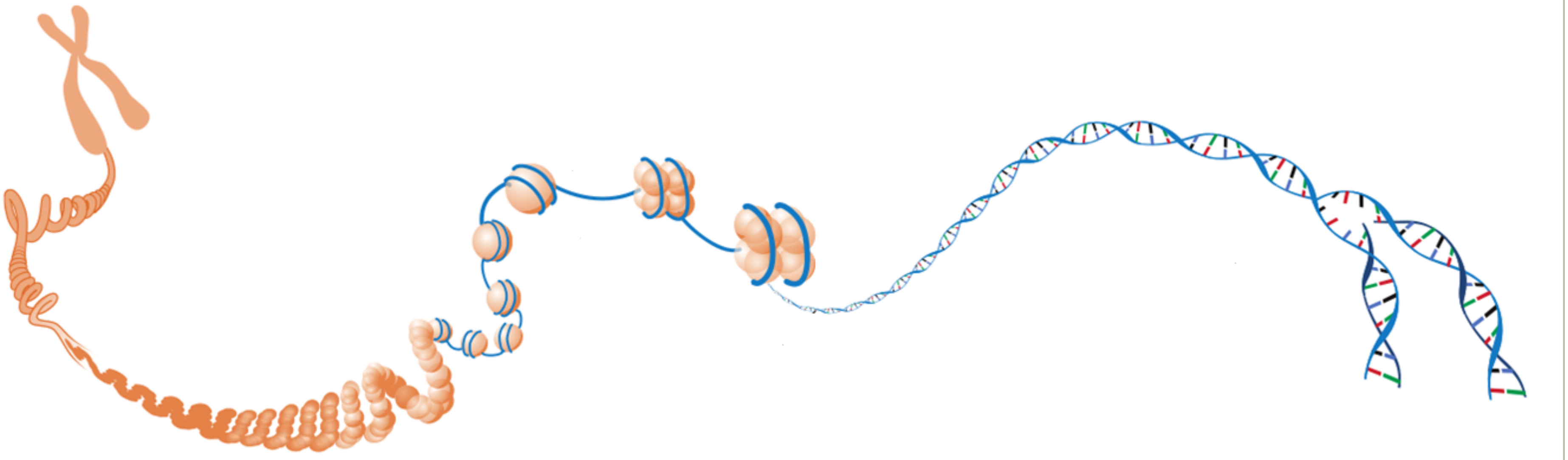
- Shuffle one feature, keep the other fixed
  - Not a consistent estimator



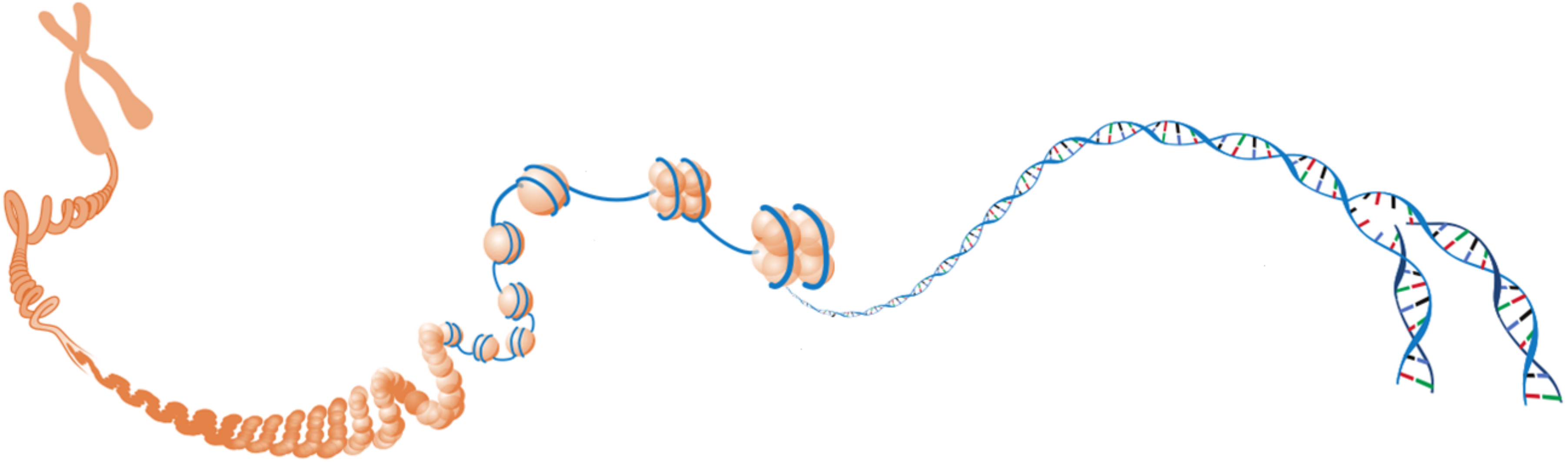
Requirements

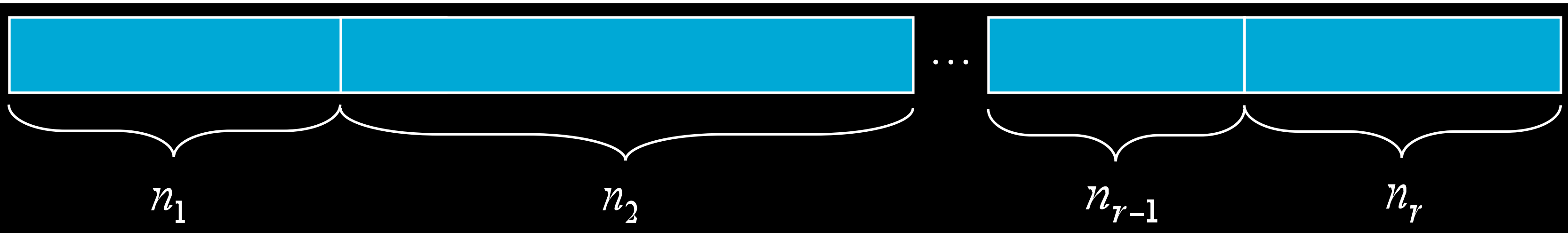
- **General:** any other consistent estimator is a special case (submodel)
- **Self diagnostic:** if assumptions aren't met, it shows up during analysis
- **Conservative:** any p-value is assured to be greater than or equal to the "true" p-value

Toward a model



“Segmented Stationarity”



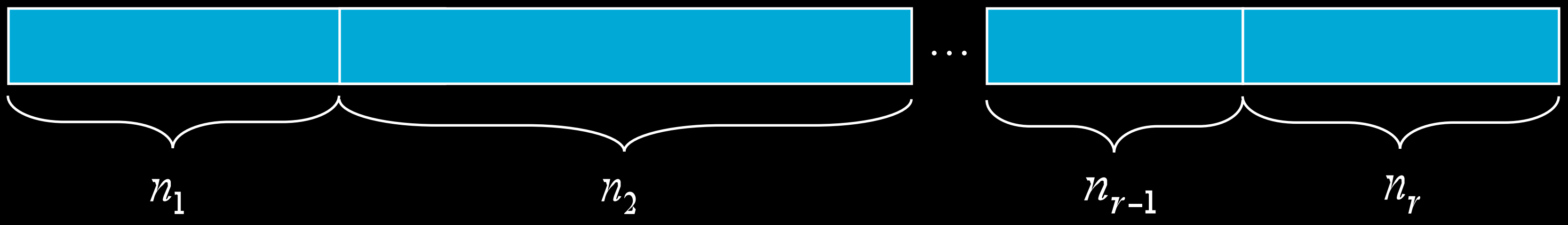


“Segmented Stationarity”

Let X_i = base at position i, $i=1,\ldots,n$
such that for each $k=1,\ldots,r$, $\{X_{k_j} : 1 \le j \le n_k\}$ is:

1) Stationary (homogeneity within blocks)
2) Mixing (bases at distant positions are nearly independent)
3) And, $r << n$

$$(X_1,\ldots,X_n) = (X_{1_1},\ldots,X_{1_{n_1}},\ldots,X_{r_1},\ldots,X_{r_{n_r}}), \qquad n = n_1 + \ldots + n_r$$



The GSC := “Segmented Stationarity”

1) Stationary (homogeneity within blocks)
2) Mixing (bases at distant positions are nearly independent)
3) And, $r << n$

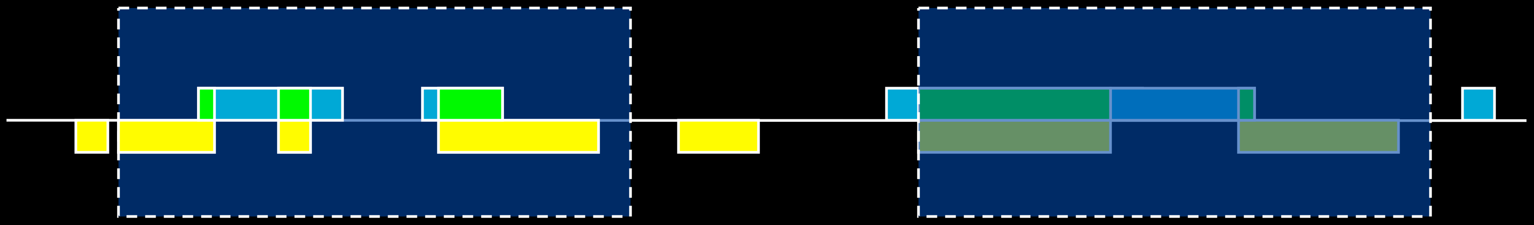
✓ **General:** any other consistent estimator is a special case (submodel)

✓ **Self diagnostic:** if assumptions aren’t met, it shows up during analysis

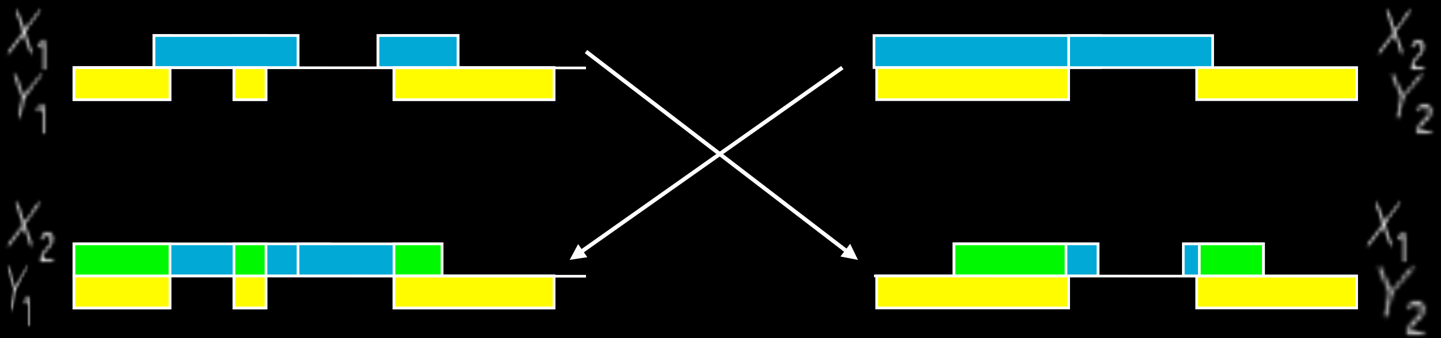
✓ **Conservative:** any p-value is assured to be greater than or equal to the “true” p-value

Testing Independence

Observed Sequence (Feature 1 = 🟦, Feature 2 = 🟨):



Sample two blocks of equal length.



Calculate overlap in the blocks after swapping = $(X_2)(Y_1)+(X_1)(Y_2)$
Align Feature 1 of first block with Feature 2 of second block, and vice versa.

Statistic is: $(X_2)(Y_1)+(X_1)(Y_2)$, properly normalized and set to mean 0.
Under the null hypothesis of independence, this should be Gaussian.

Inference under Segmented Stationarity

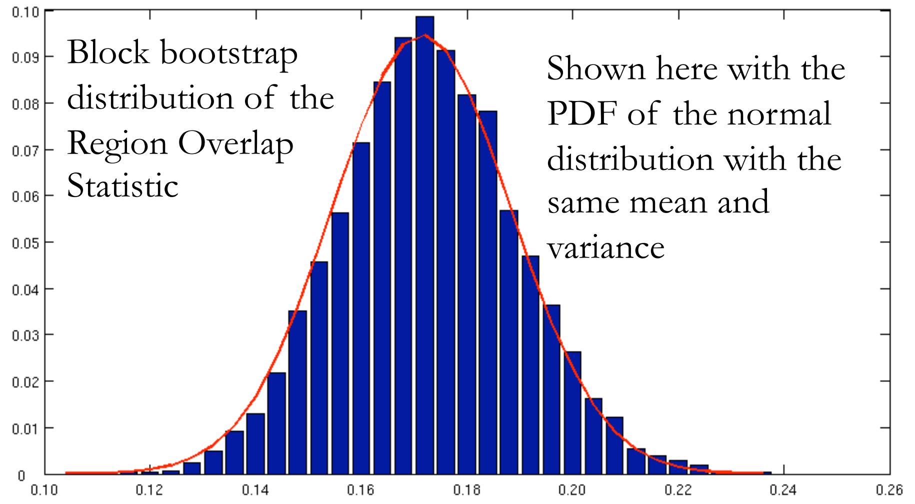
Many genomic statistics are function of one or more sums of the form:

$$S = \sum_{i=1}^{n} g(U_i)$$

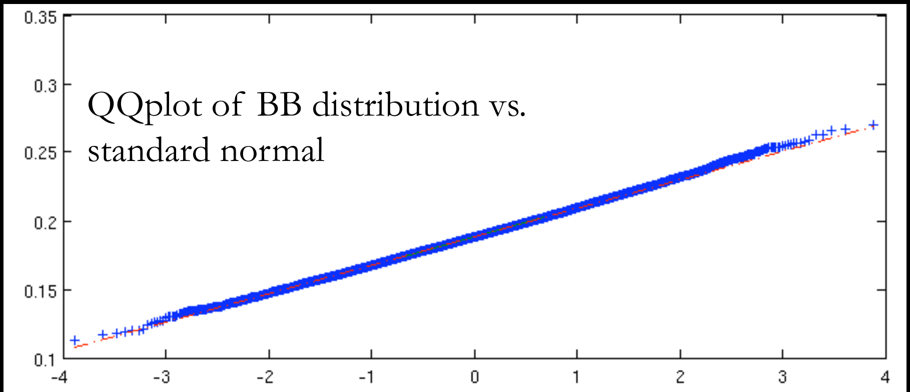
e.g. $g(X_k)$ is 1 or 0 depending on the presence or absence of a feature or features

Under segmented stationarity,
these distributions are asymptotically Gaussian
and can be estimated from the data

How Gaussian?


Block bootstrap distribution of the Region Overlap Statistic

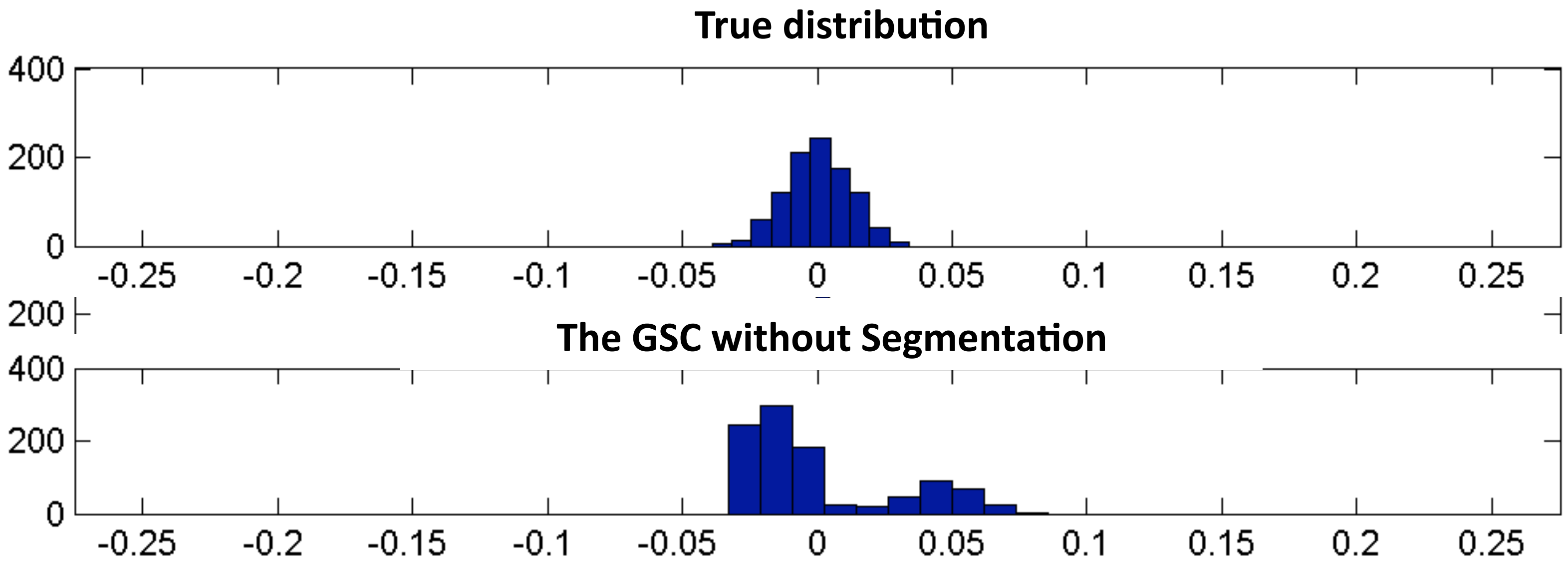
Shown here with the PDF of the normal distribution with the same mean and variance


QQplot of BB distribution vs. standard normal

All ENCODE Pilot biochemically active elements

vs

All ENCODE Pilot conserved regions



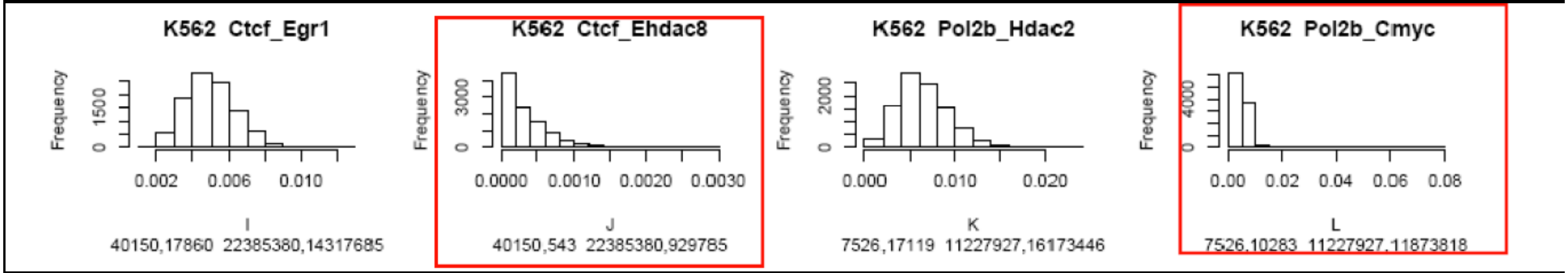
True distribution

The GSC without Segmentation

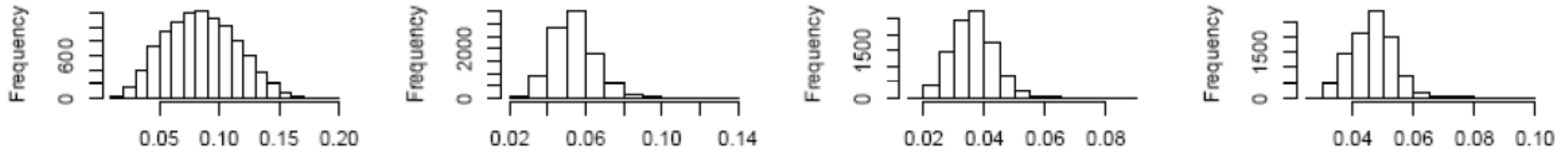
The effects of segmentation on real data

(Kevin and Nitin are amazing)

Unsegmented



Segmented



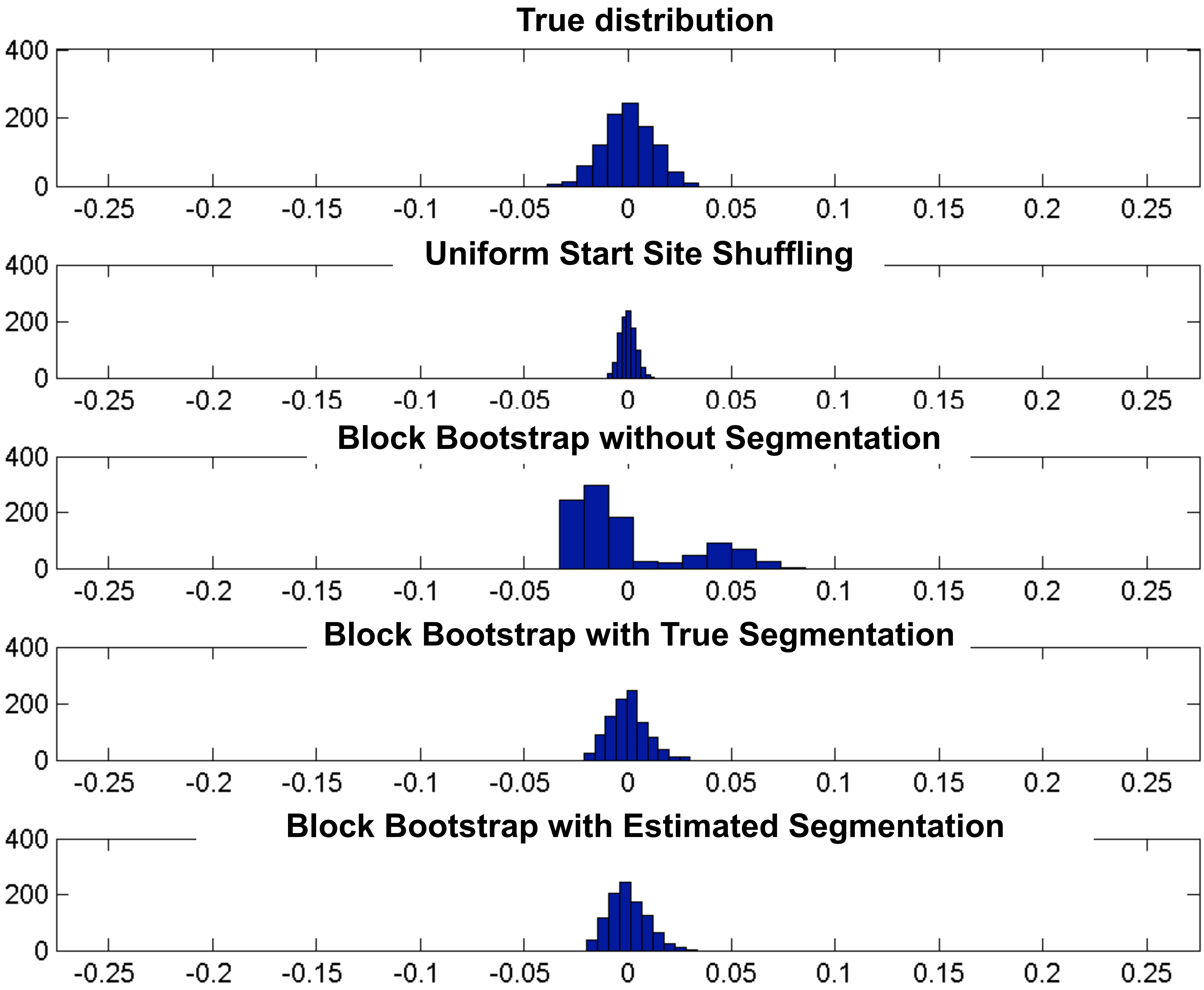
Dyadic Segmentation

- Define,

$$M(j) = \frac{j}{n}\left(1 - \frac{j}{n}\right)\Delta_j^2$$

$$\Delta_j \equiv Ave\{X_i : 1 \le i \le j\} - Ave\{X_i : j+1 \le i \le n\}$$

- Find j_{max} maximizing *M(j)* creating intervals I_{left} and I_{right}
- If length of both intervals falls below a stopping criterion, stop
- Else, repeat process for I_{left} and/or I_{right}, whichever are longer than stopping criterion, with redefined *M(j)*


True distribution

Uniform Start Site Shuffling

Block Bootstrap without Segmentation

Block Bootstrap with True Segmentation

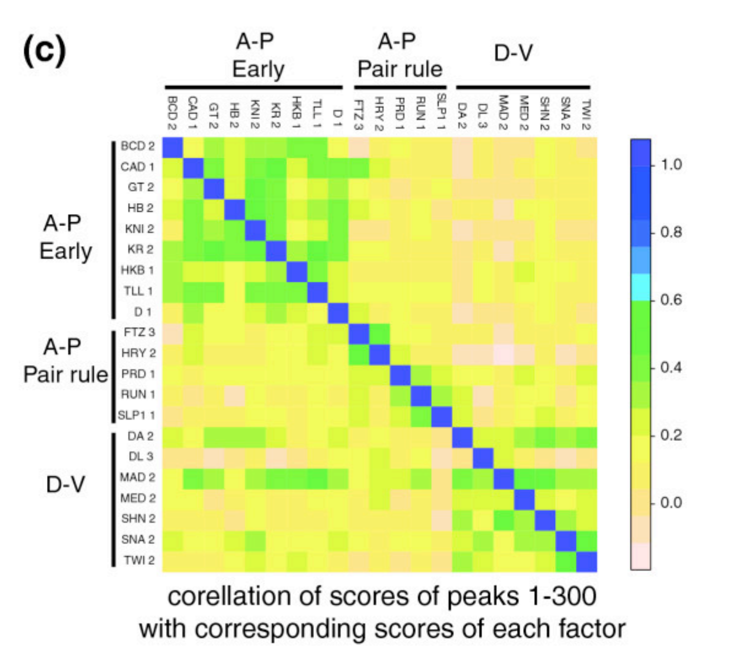
Block Bootstrap with Estimated Segmentation

Nancy's Dyadic Theorem

(Nancy is magnificent)

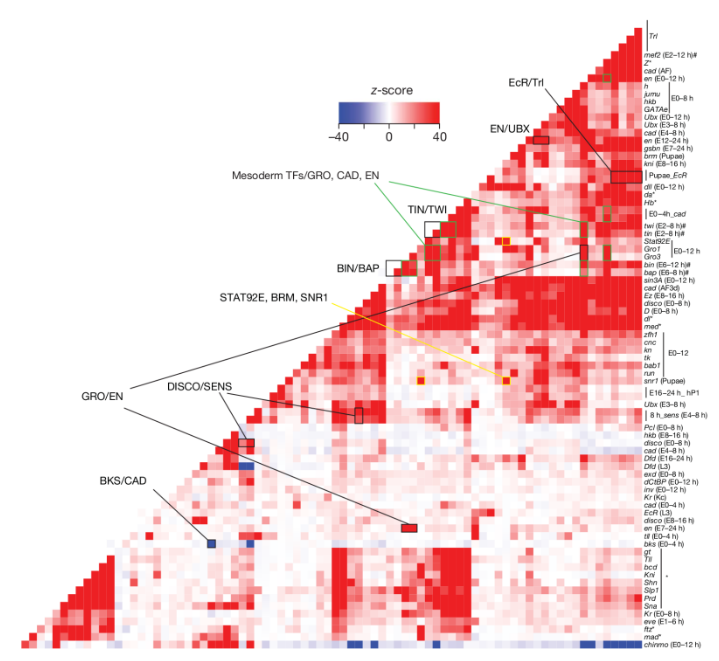
$$\hat{\sigma} = \sigma + \sum \gamma \left(\mu_i - \mu_j\right)^2$$

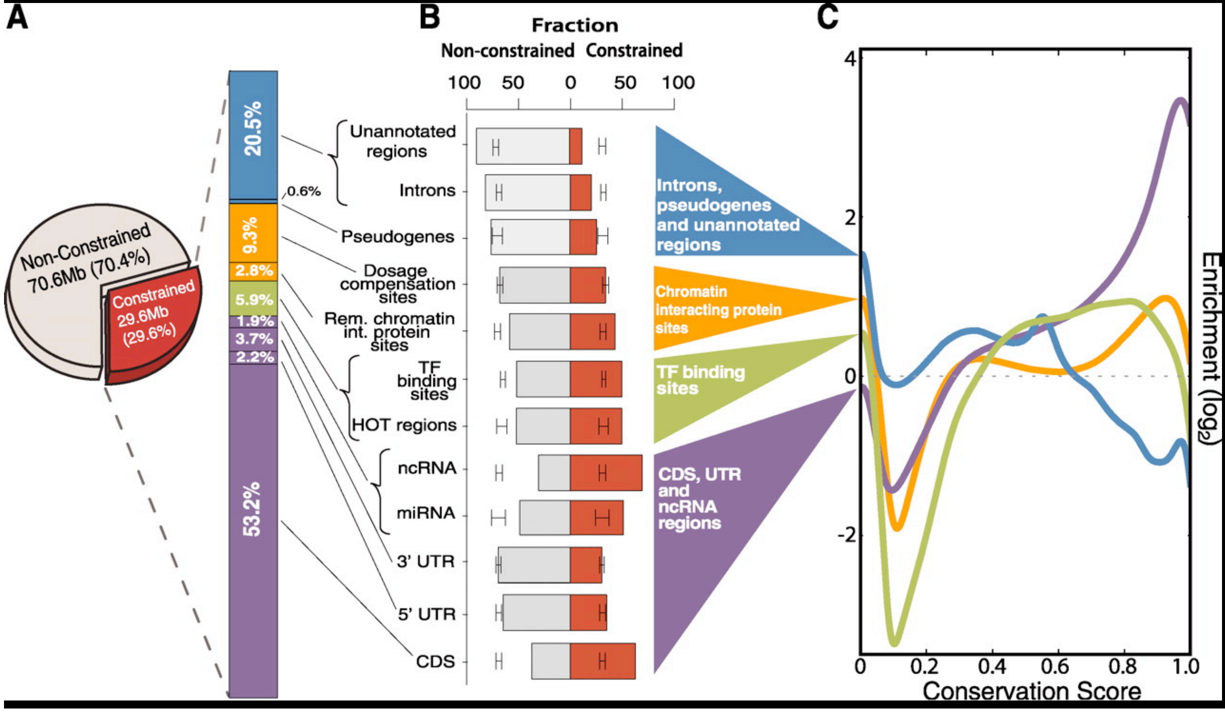
MacArthur et al, 2009, Genome Biology



corellation of scores of peaks 1-300 with corresponding scores of each factor

Negre et al, 2011, Naure

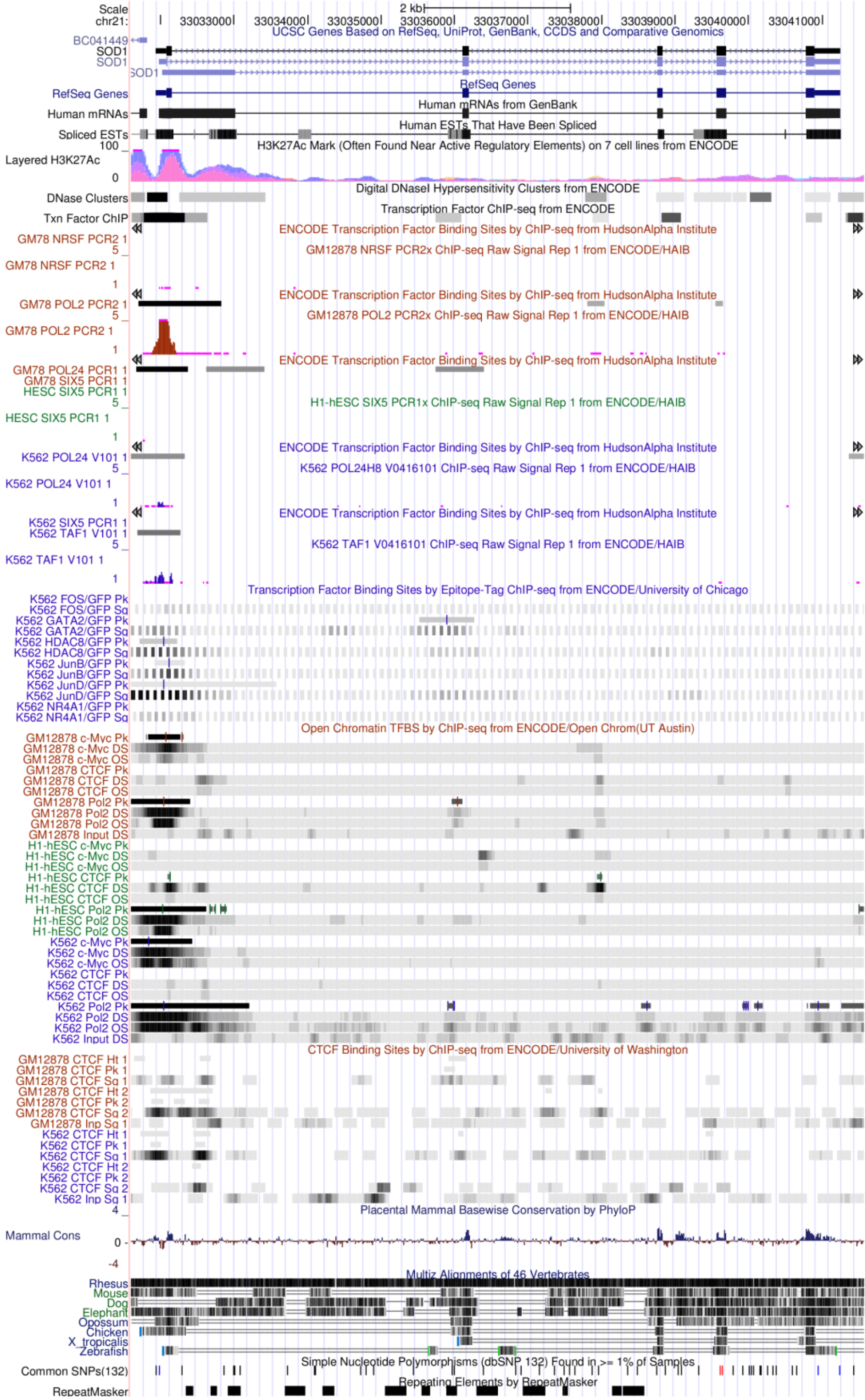


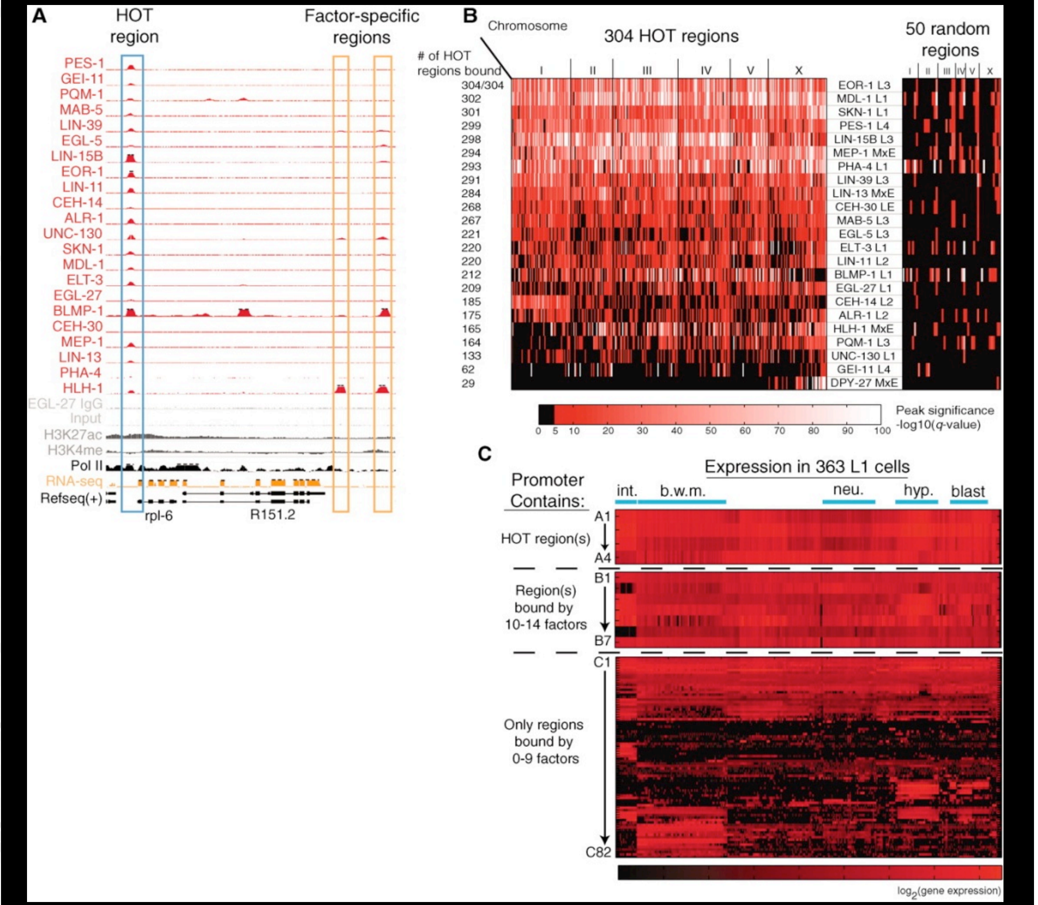


Applied in high impact papers...

But only by the usual suspects!

Gerstein et al, 2011, Science





M B Gerstein et al. Science 2010;330:1775-1787

The Team

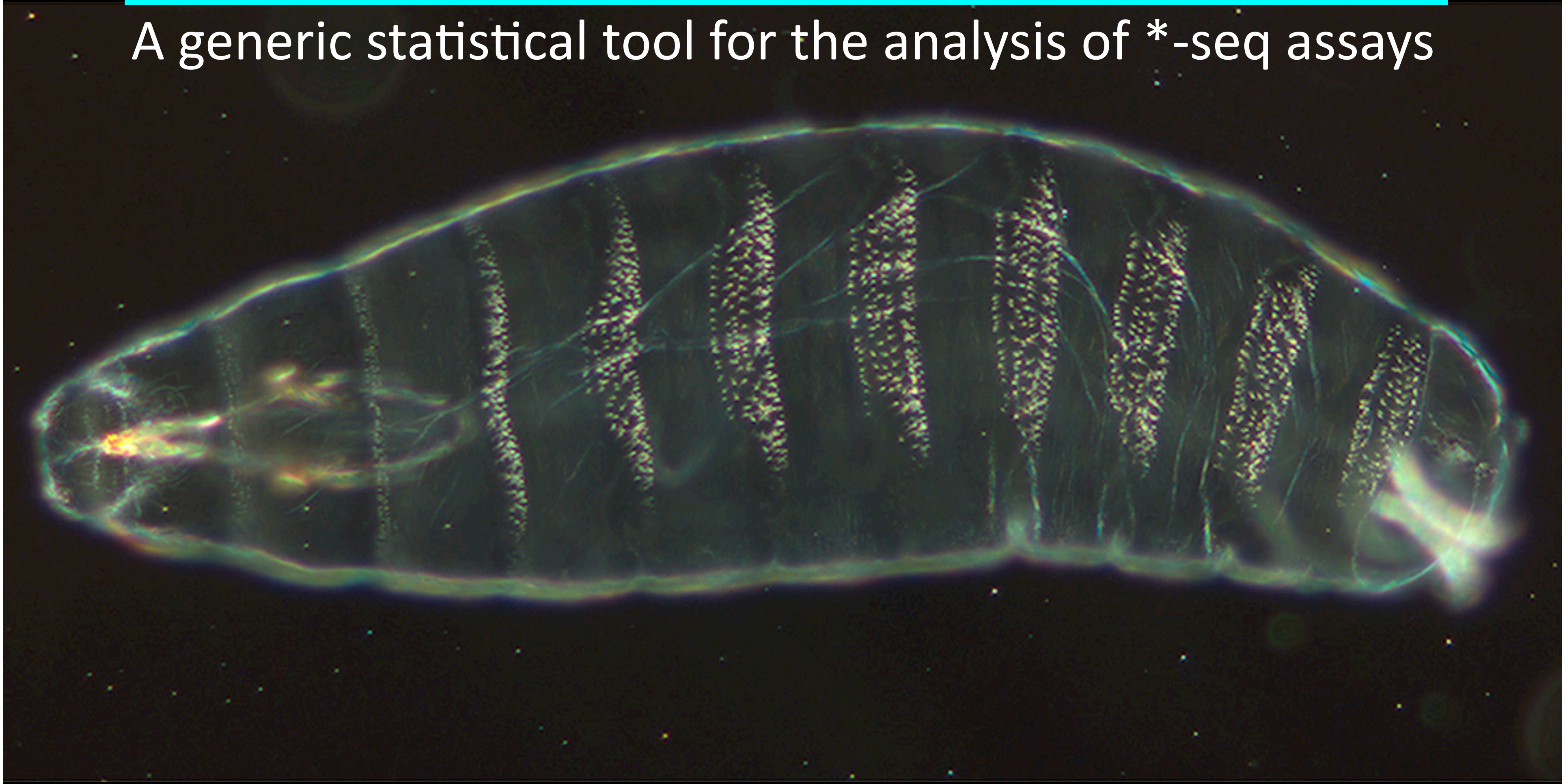
- Nancy Zhang
- Haiyan Huang
- Nathan Boley
- Peter Bickel

and now:

- Jasmine Mu
- Kevin Yip
- Joel Rozowsky
- and Mark Gerstein

Beyond heuristics

A generic statistical tool for the analysis of *-seq assays



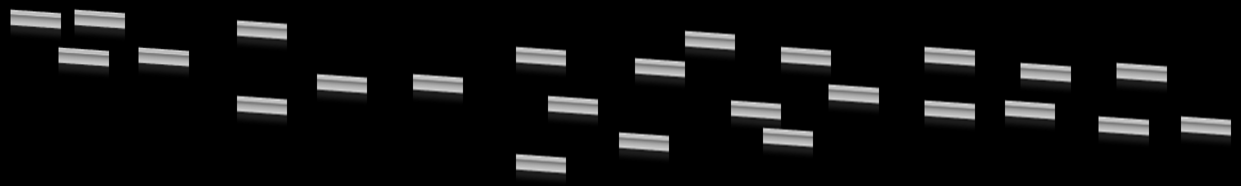
Dr. Ben Brown,

Statistics, UC Berkeley

Definition: *-seq assay

Input:
DNA or RNA

Selection:
interrogate the
input

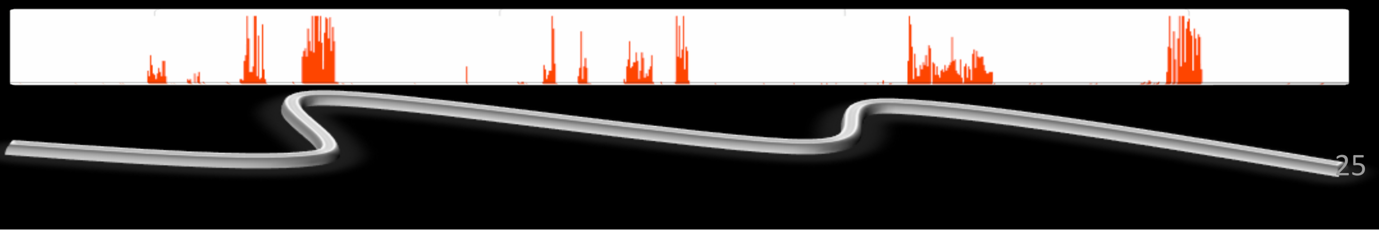


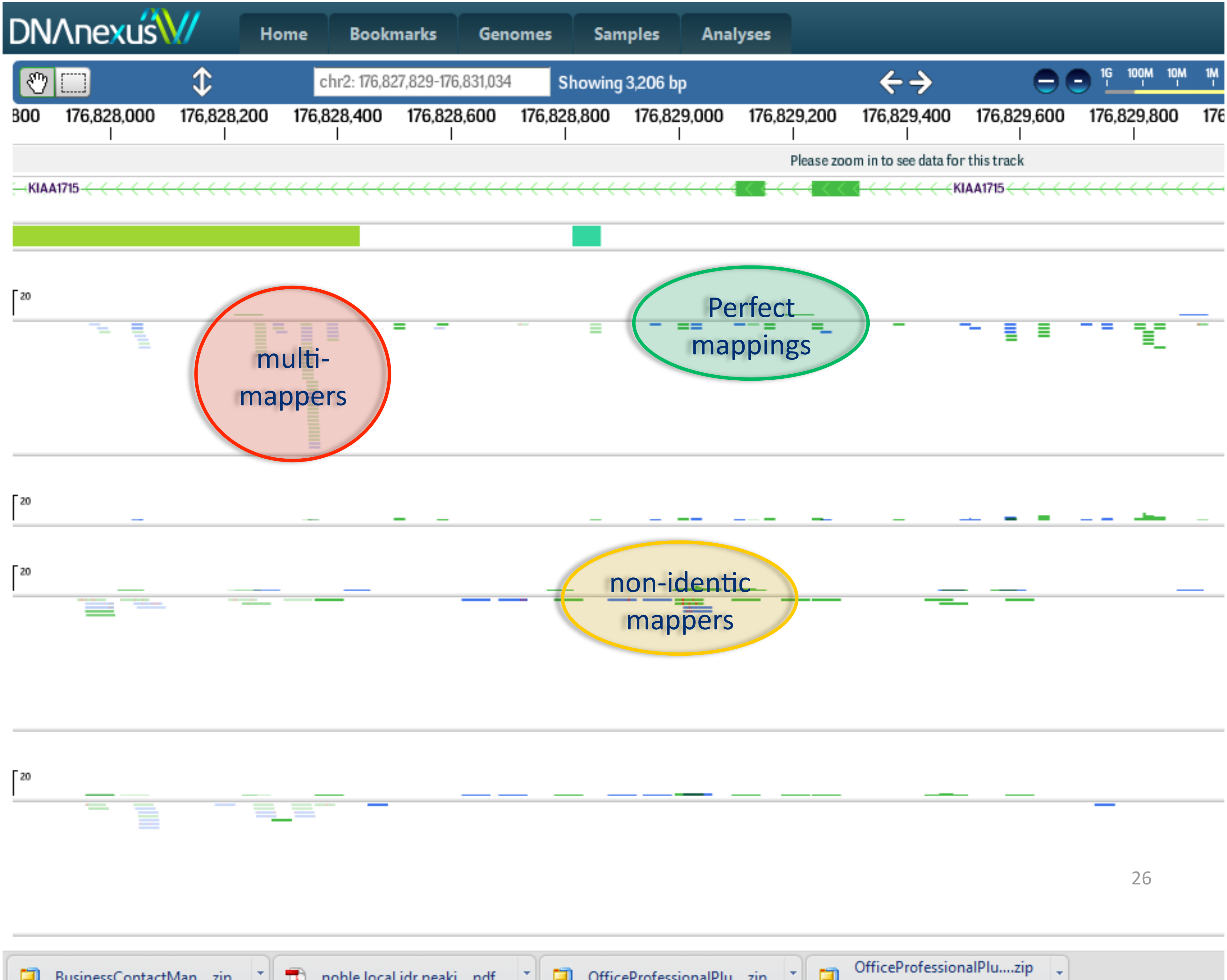
Output:
sequence reads

```
AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
hhHAJhha;hhhhhhhhhhhhhhhhhhhhhhhhhh
```

Map reads back
to Input

Interpret the
mapping





chr2: 176,827,829-176,831,034

Showing 3,206 bp

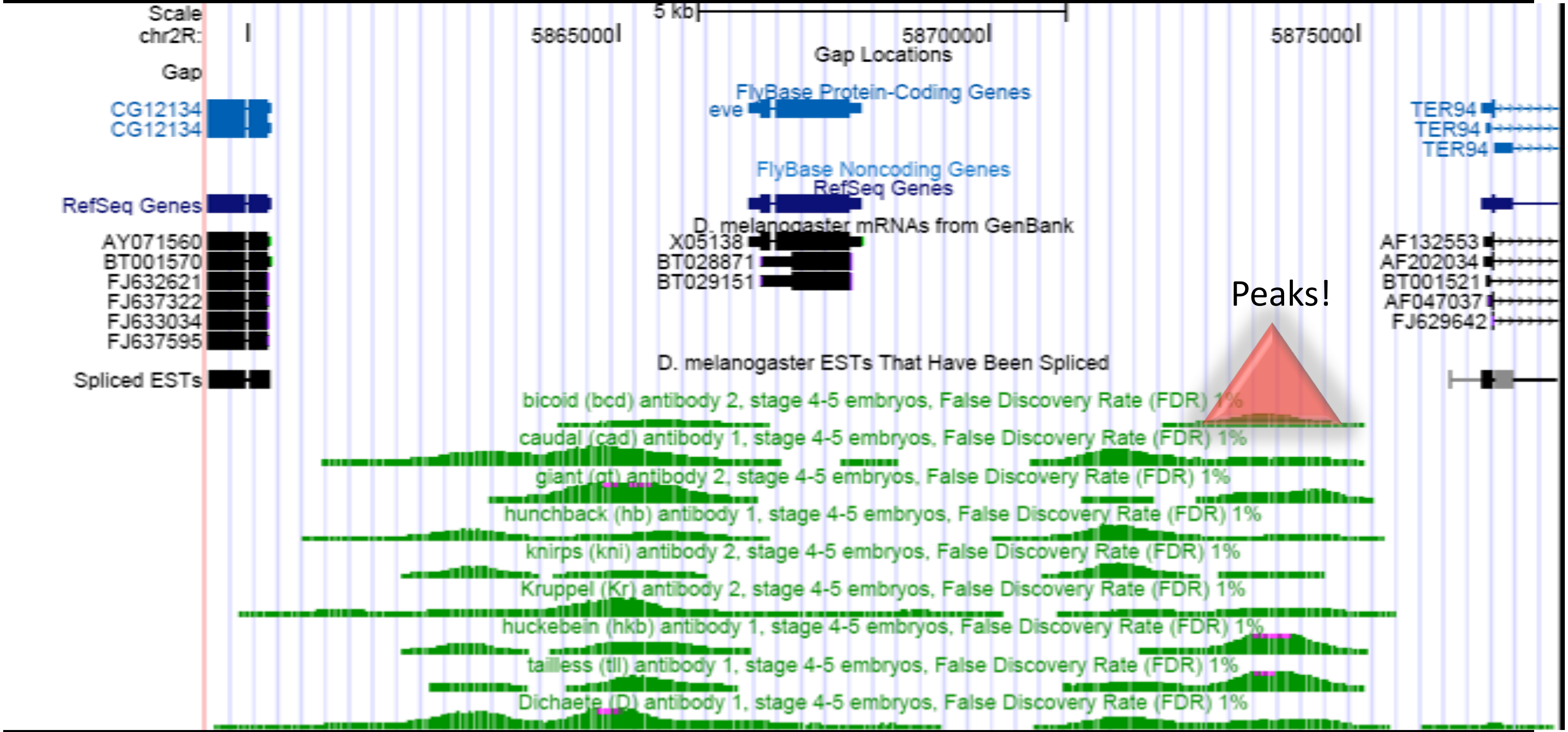
Please zoom in to see data for this track

multi-mappers

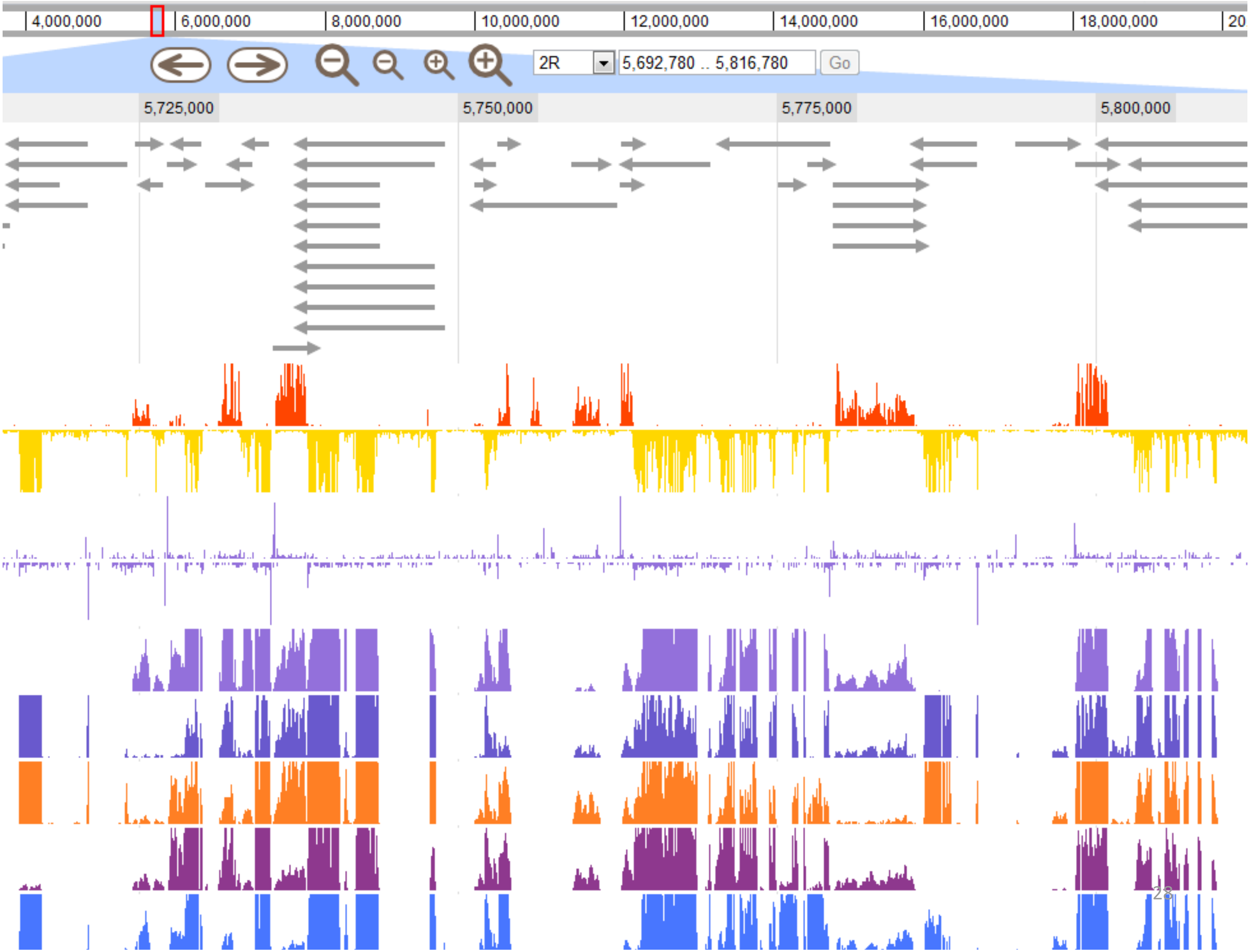
Perfect mappings

non-identic mappers

Complex statistics are computed on the mapped trace


Scale
chr2R:
5 kb
5865000
5870000
5875000
Gap
Gap Locations
FlyBase Protein-Coding Genes
CG12134
CG12134
eve
TER94
TER94
TER94
FlyBase Noncoding Genes
RefSeq Genes
RefSeq Genes
D. melanogaster mRNAs from GenBank
AY071560
BT001570
FJ632621
FJ637322
FJ633034
FJ637595
X05138
BT028871
BT029151
AF132553
AF202034
BT001521
AF047037
FJ629642
D. melanogaster ESTs That Have Been Spliced
Spliced ESTs
bicoid (bcd) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
caudal (cad) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
giant (gt) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
hunchback (hb) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
knirps (kni) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
Kruppel (Kr) antibody 2, stage 4-5 embryos, False Discovery Rate (FDR) 1%
huckebein (hkb) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
tailless (tll) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%
Dichaete (D) antibody 1, stage 4-5 embryos, False Discovery Rate (FDR) 1%

Peaks!



2R

5,692,780 .. 5,816,780

Go

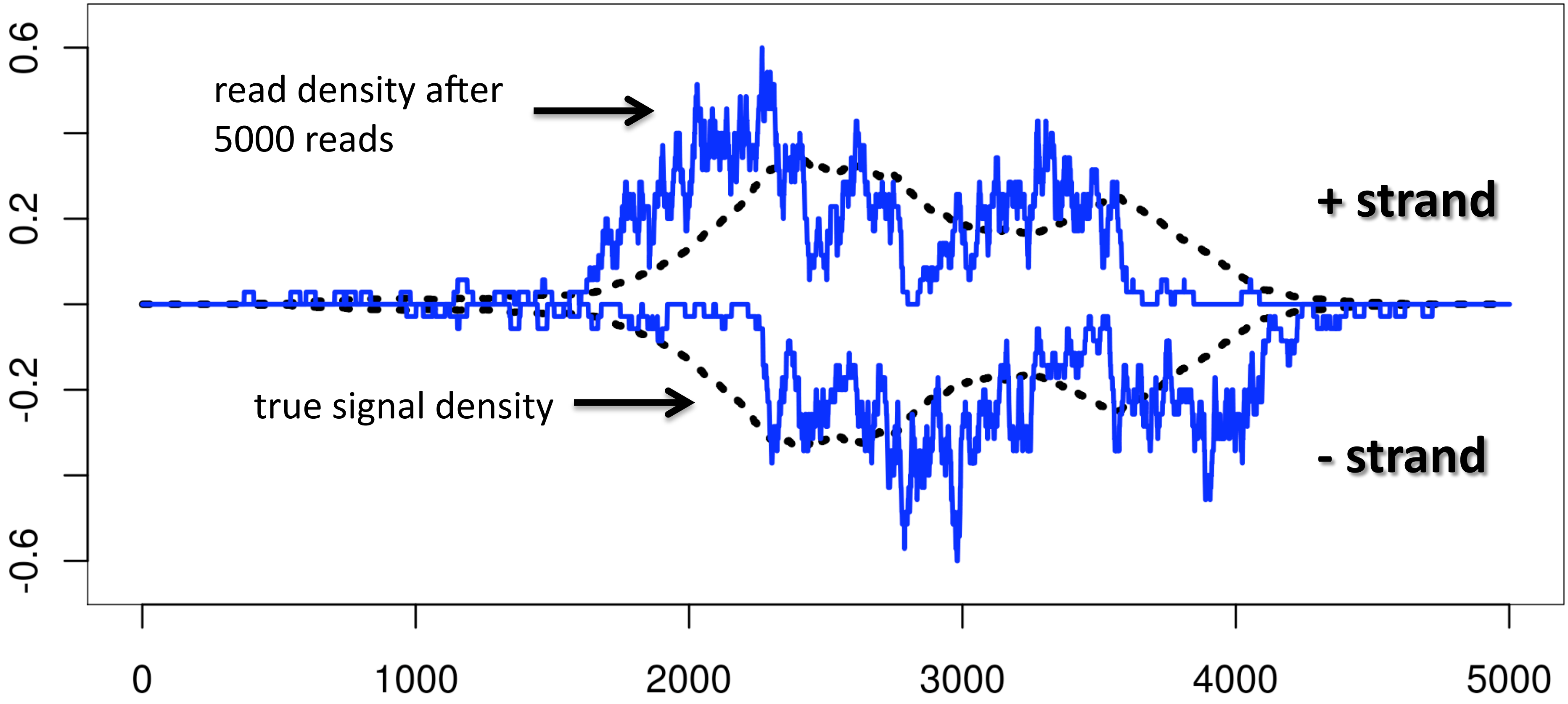
toward a generic statistical framework

- **Candidate Mapping**
  - Exhaustive: find every candidate mapping above some probability threshold
  - Correct: accurately estimate read quality scores
- **Parameter Estimation**
  - Formalize assay specific knowledge
- **Mapping Variance**
  - Find all "likely" mappings
  - Put confidence on estimated parameters
  - Estimate variance for a wide class of statistics
- **Variation**
  - Map to non-isogenic genomes
  - Dynamically infer SNPs/variation

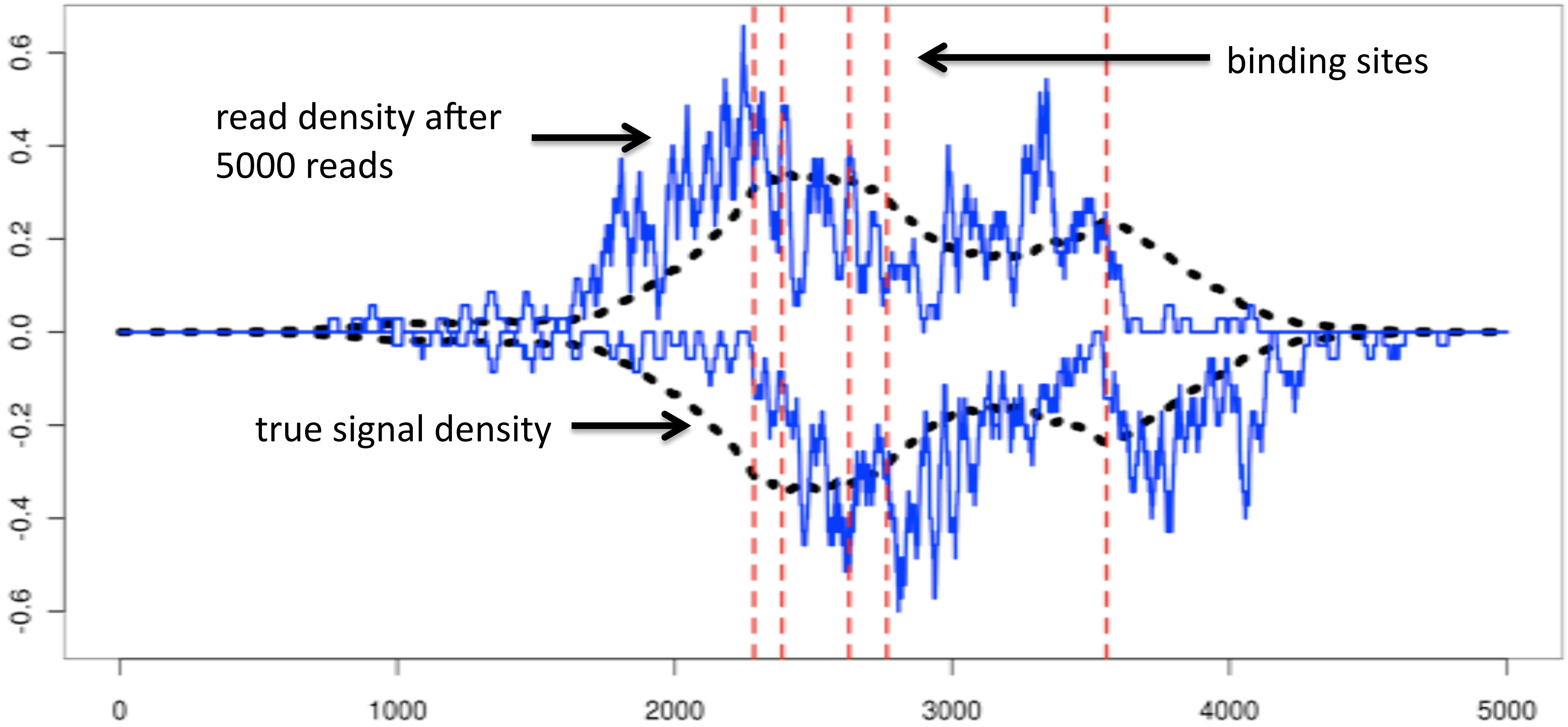
Statmap

- **Candidate Mapping**
  - Exhaustive: find every candidate mapping above some probability threshold
  - Correct: accurately re-estimate base-calling error rates
- **Parameter Estimation**
  - Formalize assay specific knowledge
- **Mapping Variance**
  - Find ~all "likely" mappings
  - Put confidence on estimated parameters
  - Estimate variance for a wide class of statistics
- **Variation**
  - Map to non-isogenic genomes
  - Dynamically infer SNPs/variation

Illustrative simulation



Illustrative simulation

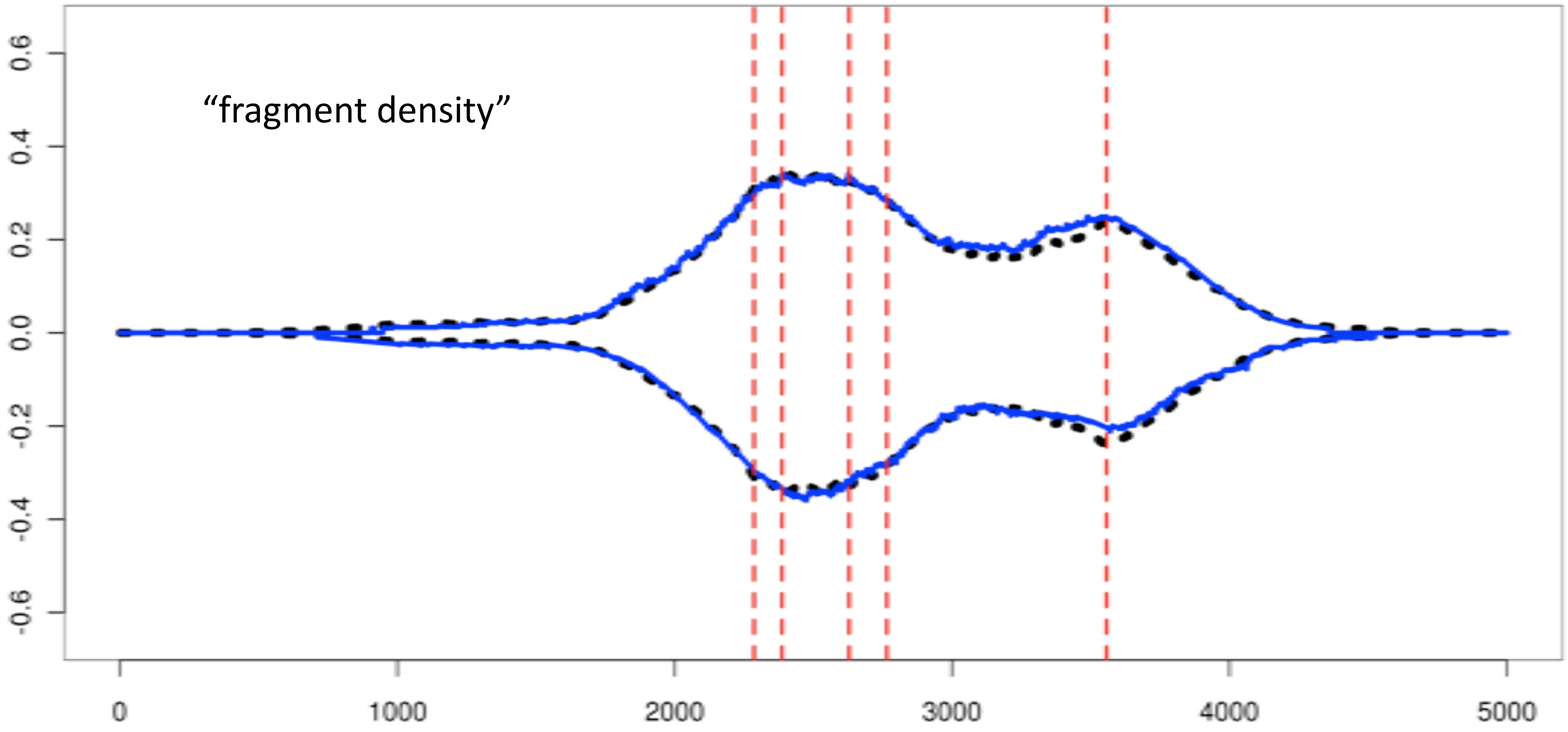


binding sites

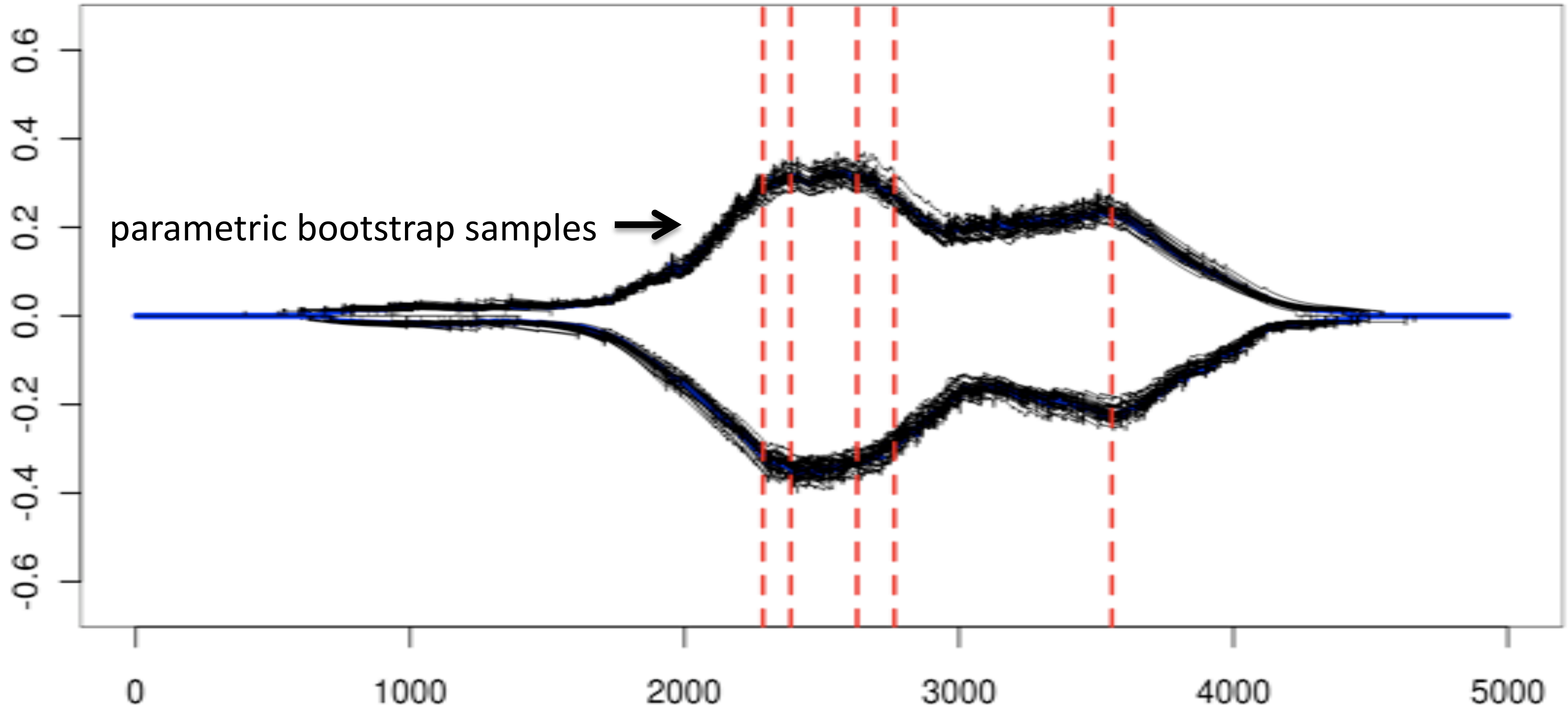
read density after 5000 reads

true signal density

parameter estimation

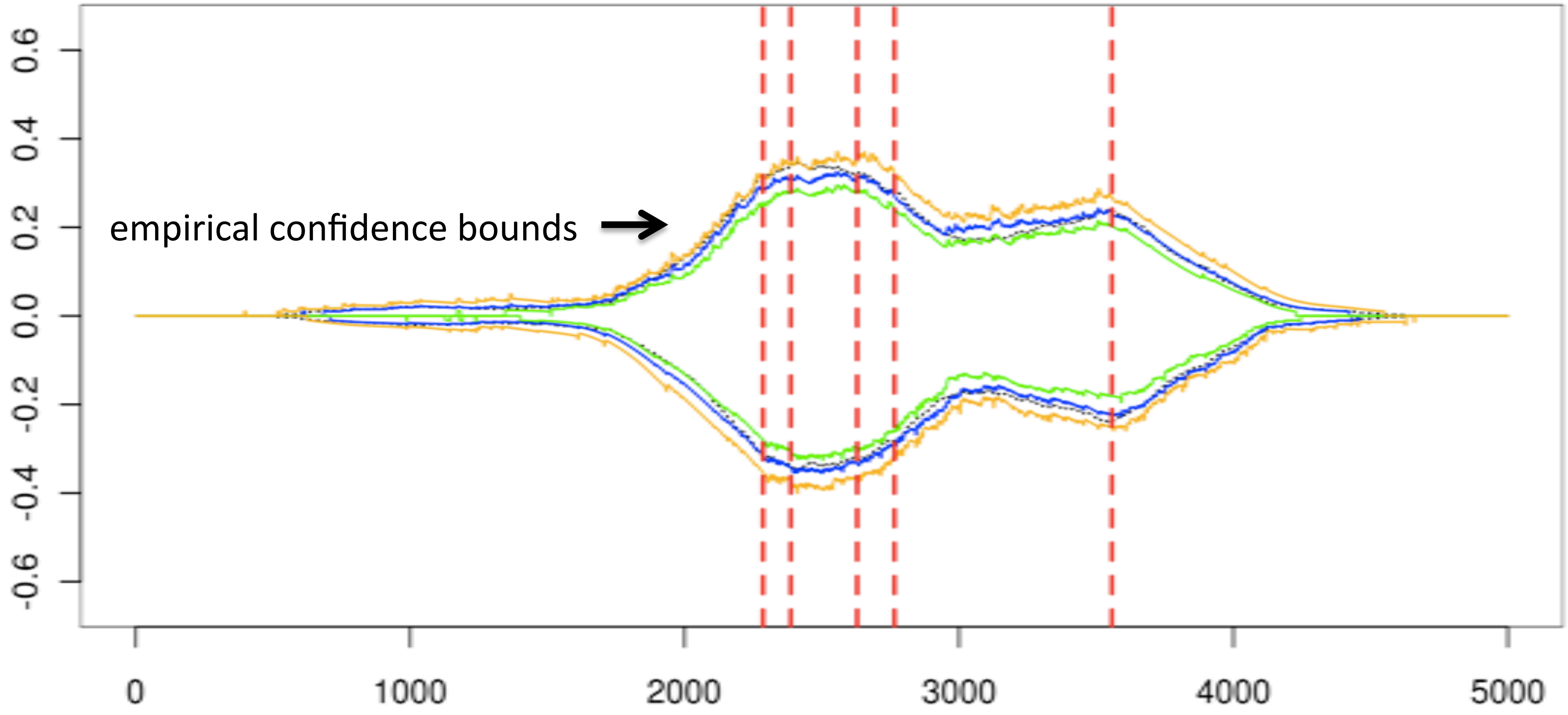


variance estimation



parametric bootstrap samples

confidence bounds



empirical confidence bounds

What the bootstrap buys us

- Place confidence bounds on fragment coverage.
- More generally:

evaluate the variance of any statistic that is a function of the mapped read density by computing the statistic over all bootstrap samples

Sampling Variance

...and what it doesn't

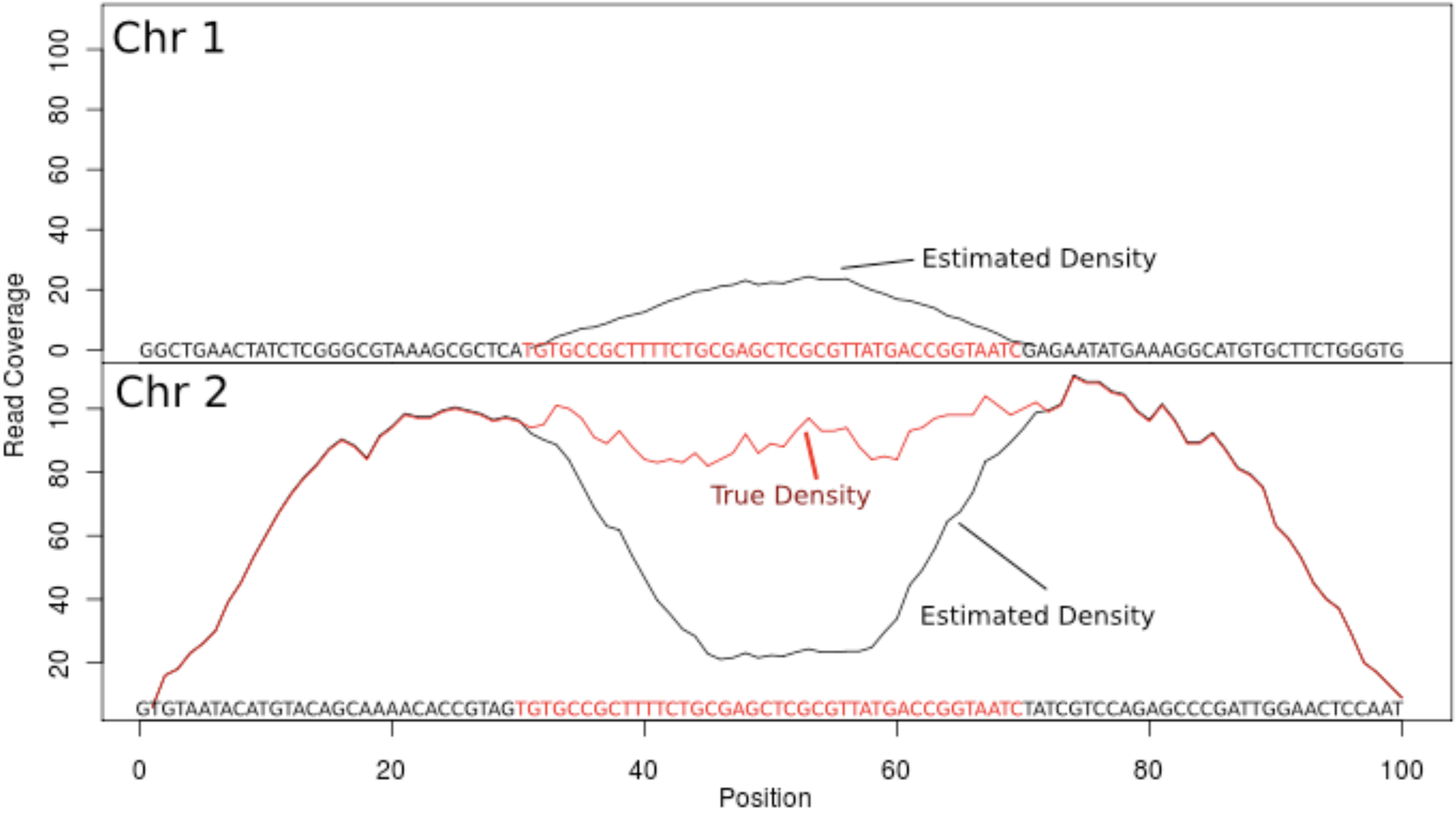
- Estimates **conditional** on the marginal read density

when the estimated read density deviates from the truth the bootstrap estimates will be poor

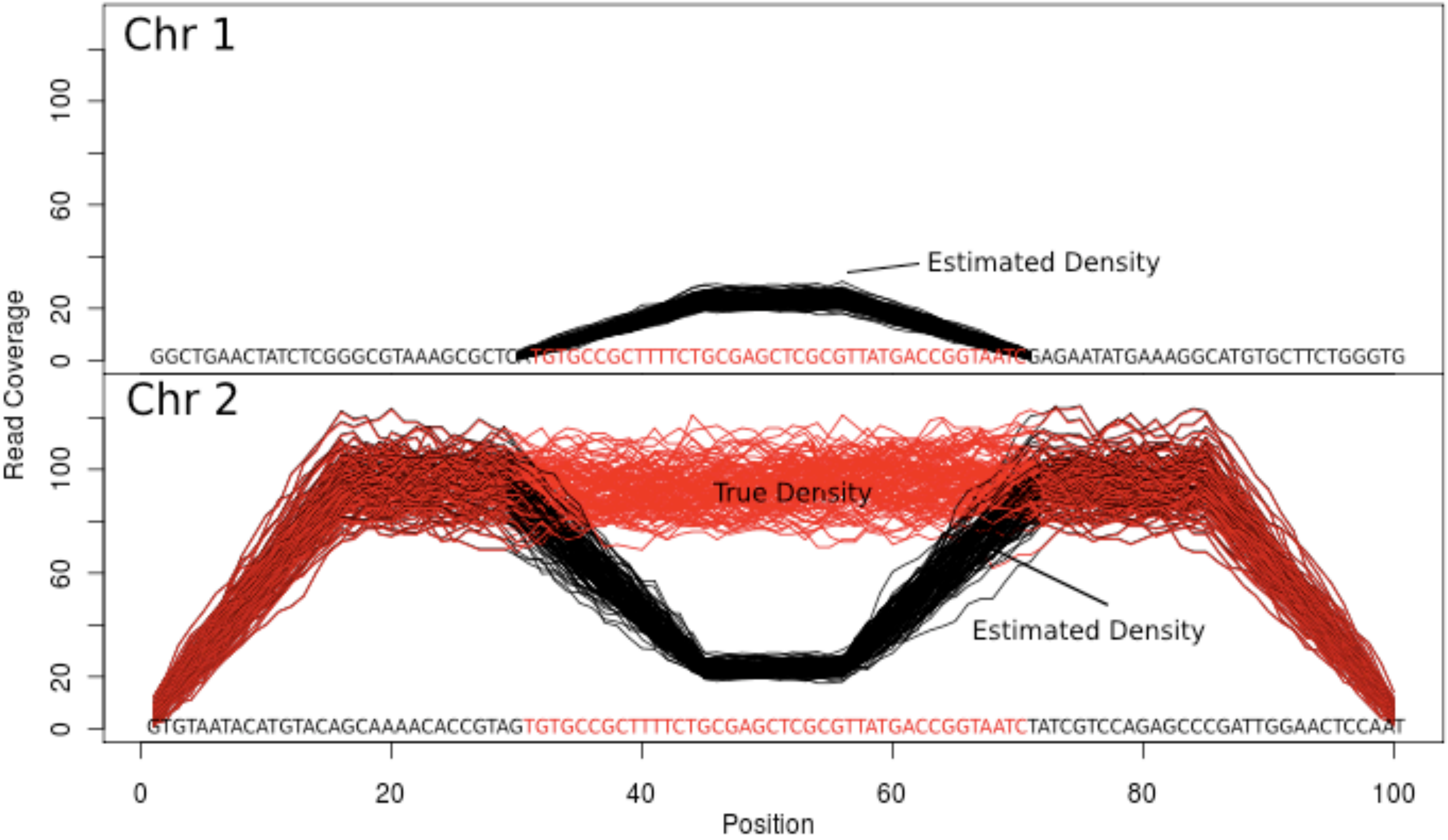
- When there are multiple 'equally valid' interpretations of an assay (mappings)

Analytical Variance

Analytical variance



The bootstrap is condition on the marginal read density

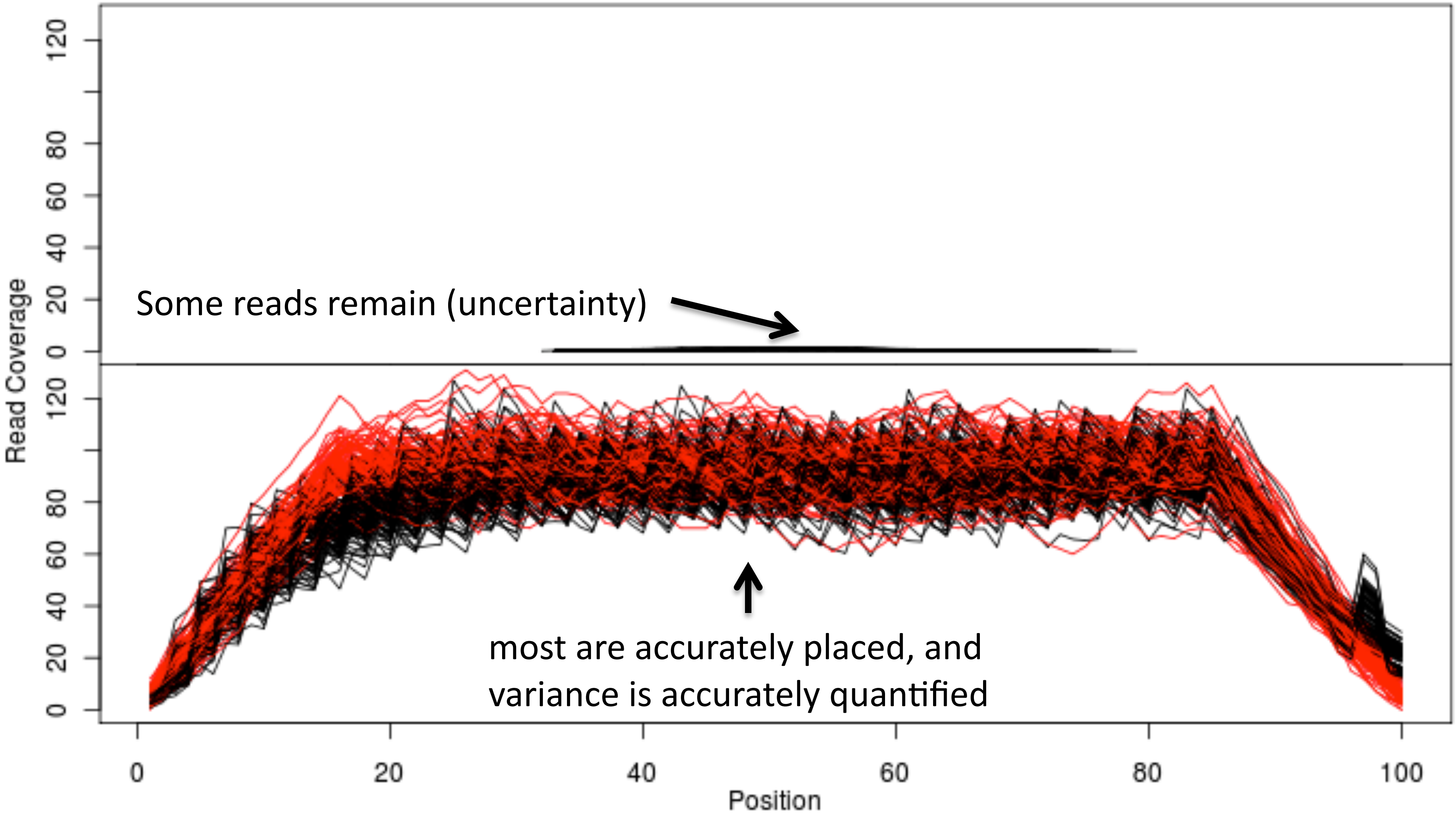


Beyond the marginal read density

- Assay specific knowledge

empowers us to consider dependence between reads

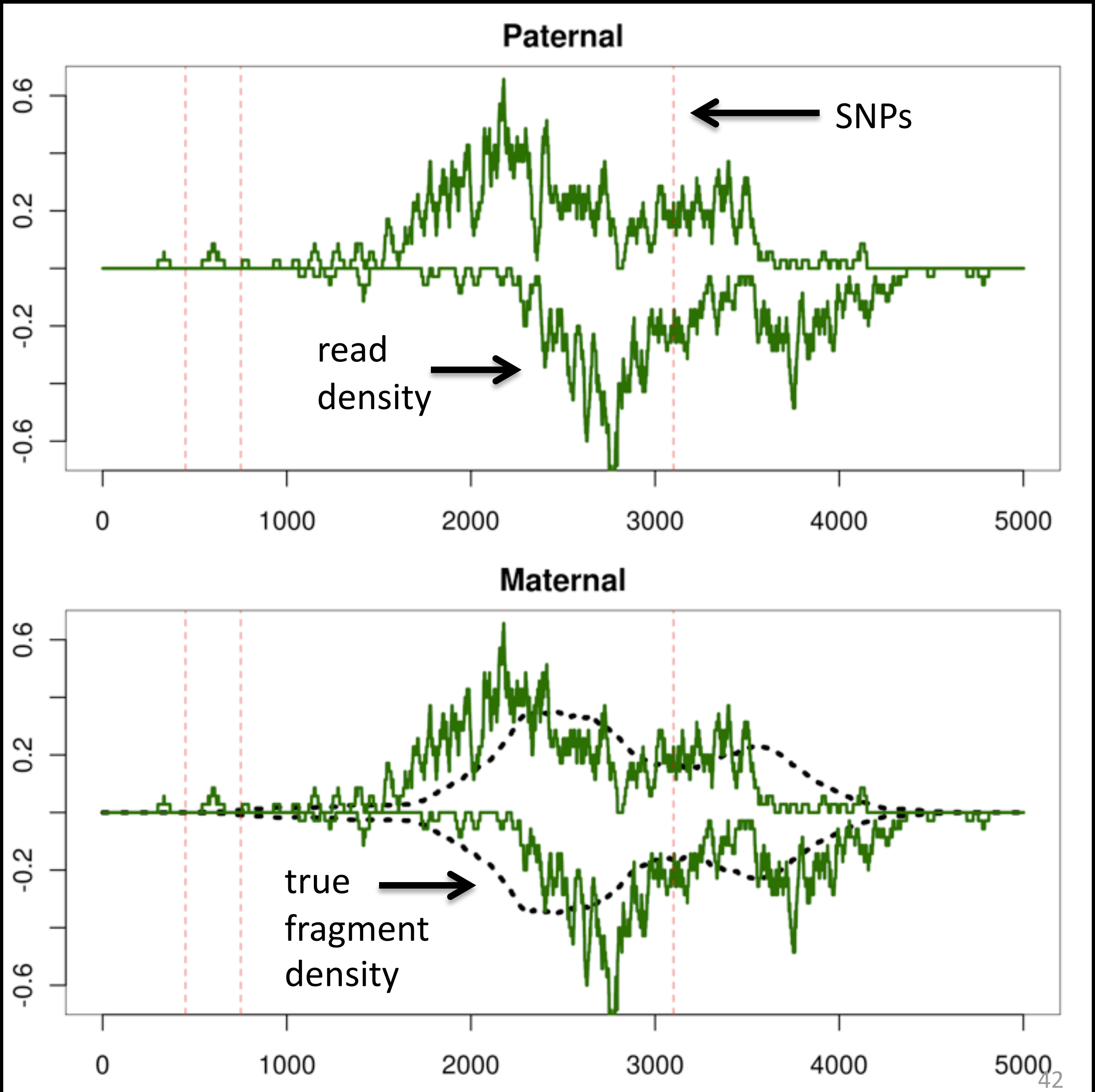
Assay specific kernels



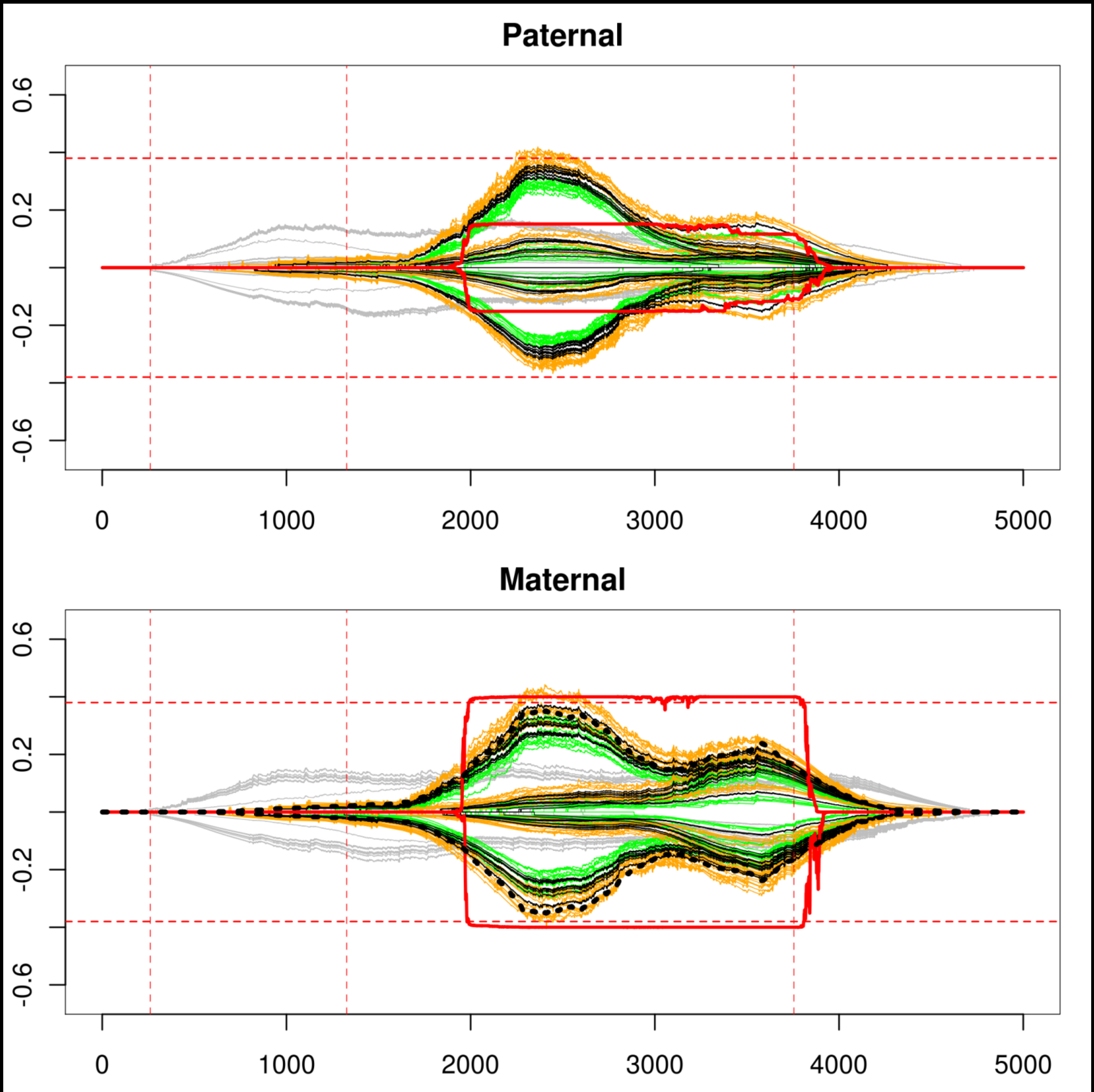
Some reads remain (uncertainty)

most are accurately placed, and variance is accurately quantified

ChIP-seq in a non-isogenic background



All reads came from the maternal chromosome



Paternal

Maternal

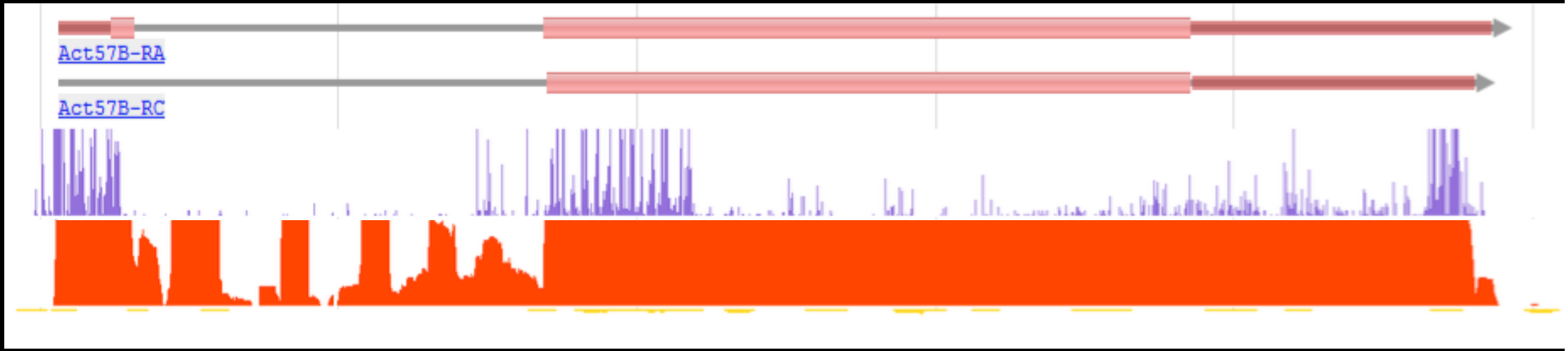
Confidence for any statistic:
the local bootstrap + a search
heuristic for likely mappings

CAGE: complex signal

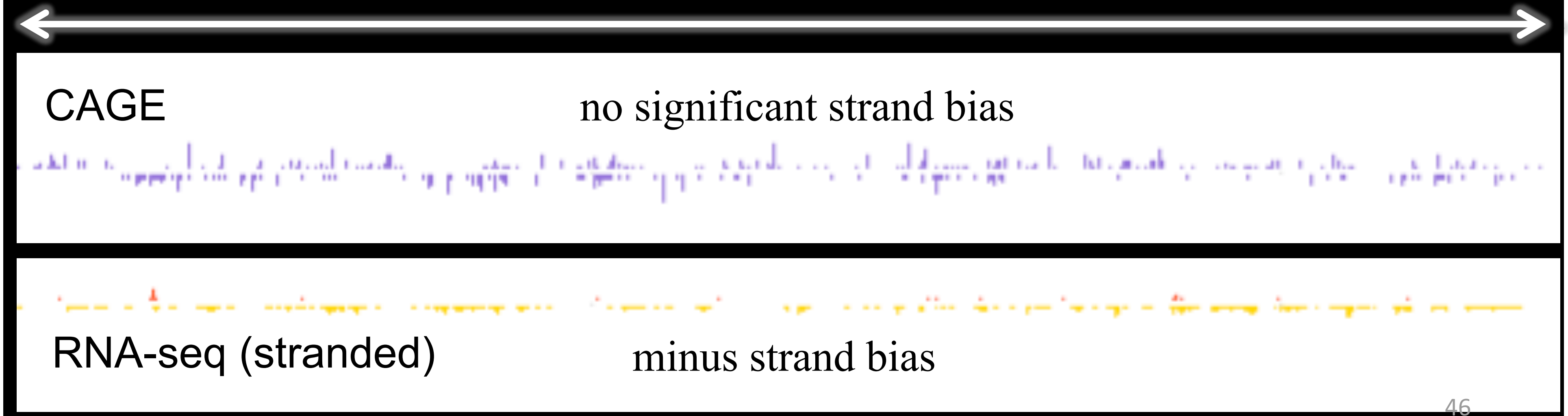
FlyBase

CAGE

RNA-seq (stranded)



50Kb of intergenic space



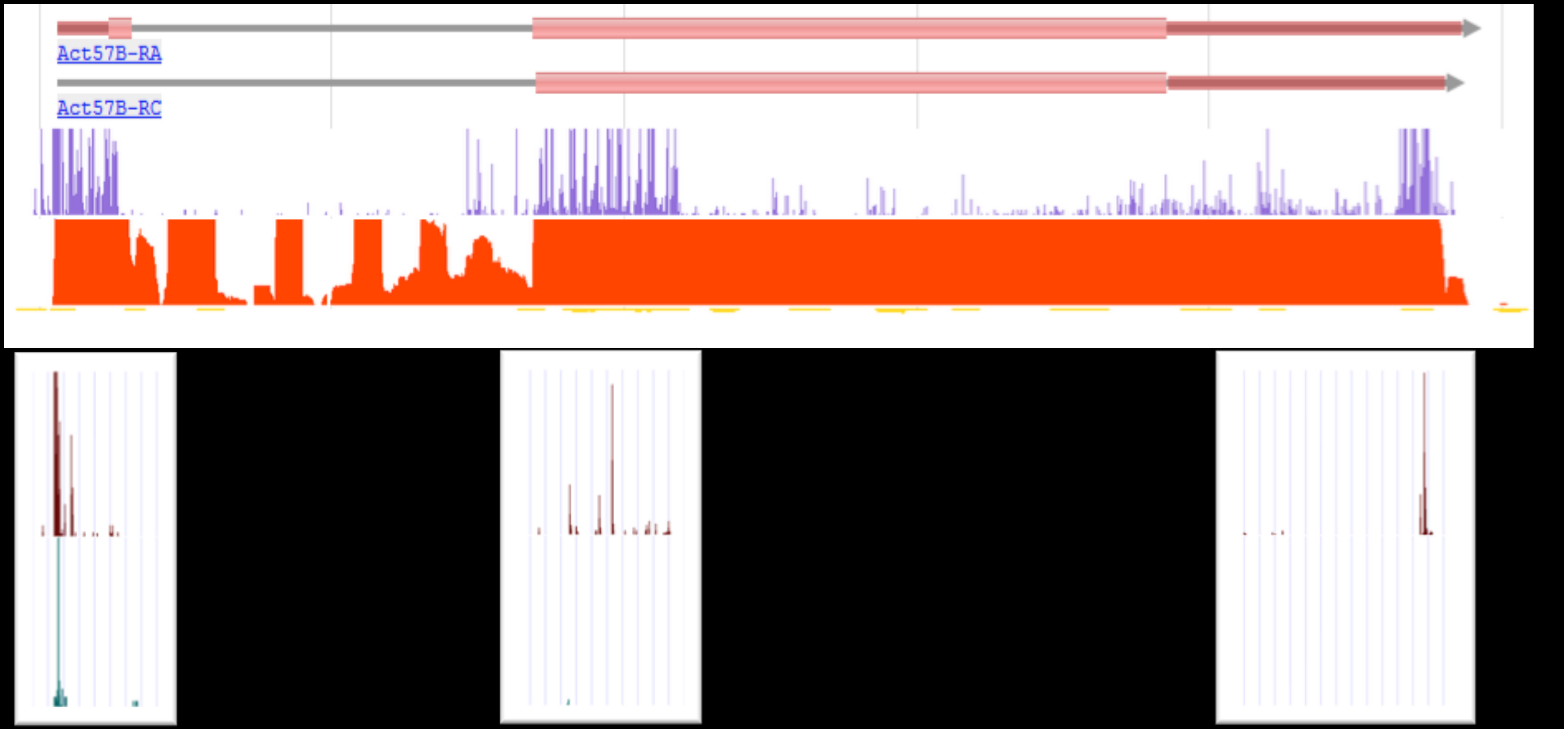
CAGE

no significant strand bias

RNA-seq (stranded)

minus strand bias

Statmapping CAGE



FlyBase

CAGE

RNA-seq (stranded)

CAGE peaks

RACE

At most 22,000
not 120,000

- Failing to account for variance and background in CAGE has had consequences:

Number of active promoters in *D. melanogaster* embryo has been over estimated, consistently, in the literature at least 5 fold

A (not so) new mindset:
after spending $10k+ on your assay, spend $50 to reliably interpret it on the Statmap implementation on the Amazon EC2 cluster

The team

- Nathan Boley
- Peter Bickel

- Special thanks to Ewan Birney and Anshul Kundaje for many enlightening conversations

Questions?

Key concepts

g: a specific location in the genome

$\Pr[g]$: the frequency with which oligos originating at g are sequenced

$\Pr[r|g]$: prob of observing read r given that g was sequenced

$\Pr[g|r] = \frac{\Pr[r|g]\Pr[g]}{\sum_{g'} \Pr[r|g']\Pr[g']}$: assumes all reads came from the genome

EM,

M step: $\Pr_{OLD}[g|r] = \frac{\Pr_{OLD}[g] \sum_{j=1}^{J} \pi_j^{OLD} f_j(g|r)}{\sum_{g\prime} \sum_k \pi_k^{OLD} f_k(g\prime|r)}$ assay specific kernel

E step: $\Pr_{NEW}[g] = \sum_r \Pr_{OLD}[g|r] \Pr[r]$

A likelihood function for any mapping:

$$lhd[\bar{r}, g] = \prod_r \sum_g \Pr[r|g] \Pr[g]$$