Bioinformatics Practices

2 May 2016

Agenda Today

- Recap of various file format in bioinformatics
- Introduction of RGT (Regulatory Genomics Toolbox)
    - Website
    - Core modules
    - Practice
- Case study
    1. Peakcalling (Practice)
    2. Visualization by lineplot (Practice)
    3. Motif analysis (Practice)

Recap of file formats

FASTA, FASTQ	Sequences
	Alignment
SAM, BAM	Reads
	Peakcalling
BED, BIGBED	Regions
WIG, BIGWIG	Signals

?? FASTA,FASTQ,SAM,BAM,BED,WIG,BIGWIG ??

```
>AB000263 |acc=AB000263|Homo sapiens mRNA
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

?? FASTA,FASTQ,SAM,BAM,BED,WIG,BIGWIG ??

```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
hhhhhhhhhhghhghhhhhfhhhhhfffffe`ee[`X]b[d[ed`[Y[^Y
```

?? FASTA,FASTQ,SAM,BAM,BED,WIG,BIGWIG ??

```
chr7   127471196   127472363   Pos1   0   +
chr7   127472363   127473530   Pos2   0   +
chr7   127473530   127474697   Pos3   0   +
chr7   127474697   127475864   Pos4   0   +
chr7   127475864   127477031   Neg1   0   -
chr7   127477031   127478198   Neg2   0   -
```

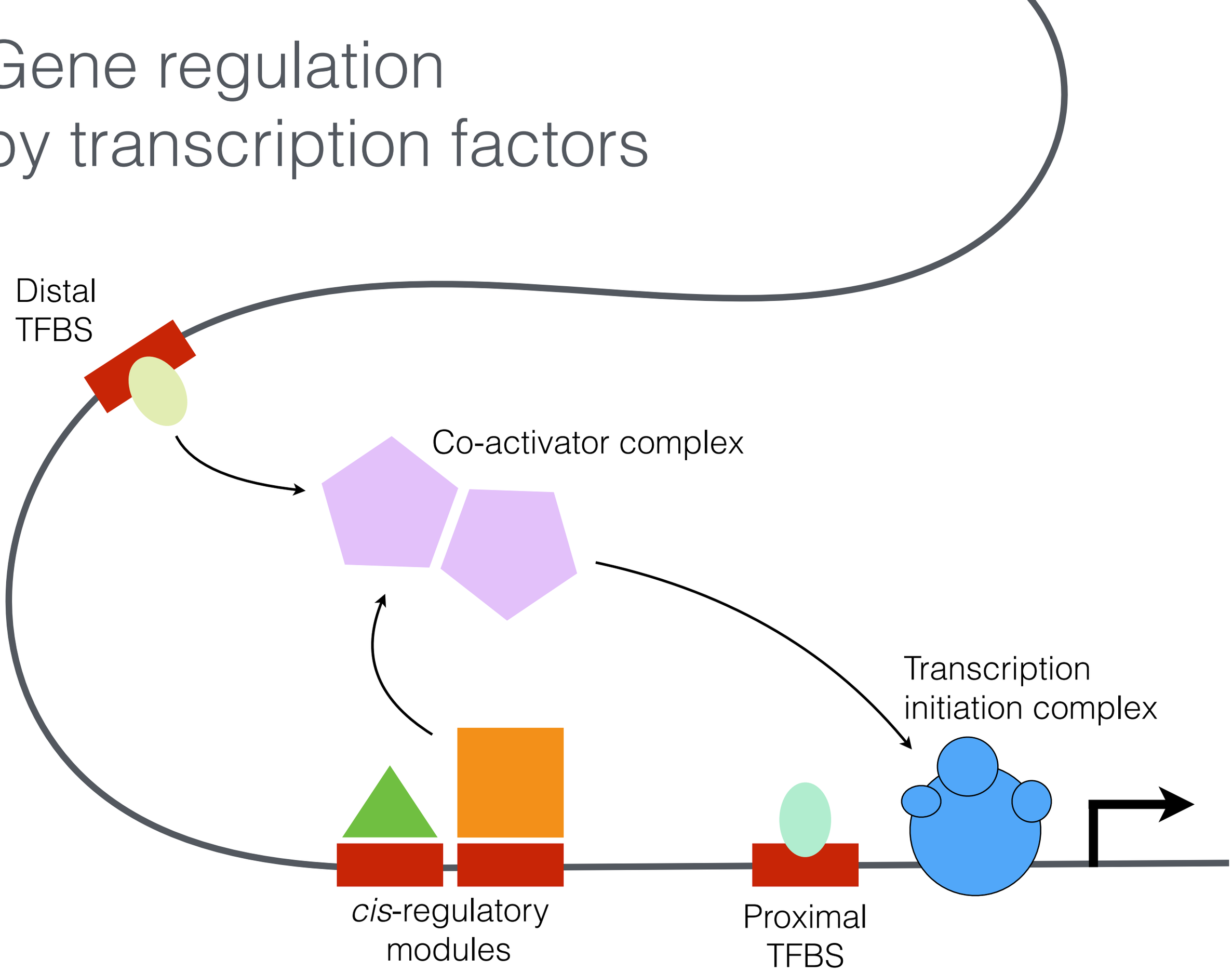
?? FASTA,FASTQ,SAM,BAM,BED,WIG,BIGWIG ??

```
variableStep chrom=chr2 span=5
300701   12.5
310500   15.0
```

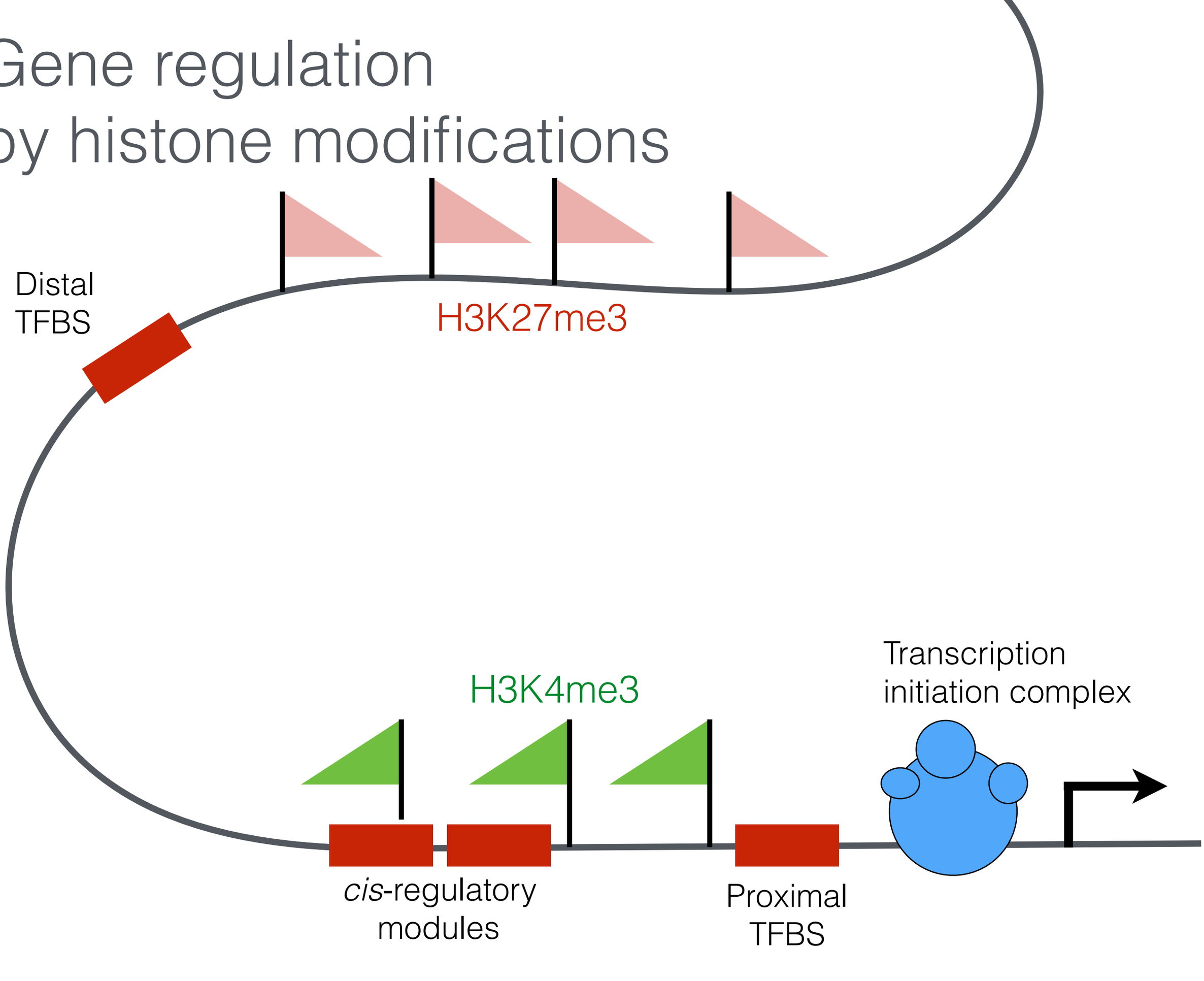
Introduction of RGT

Bioinformatic Lab 2016

Gene regulation by transcription factors



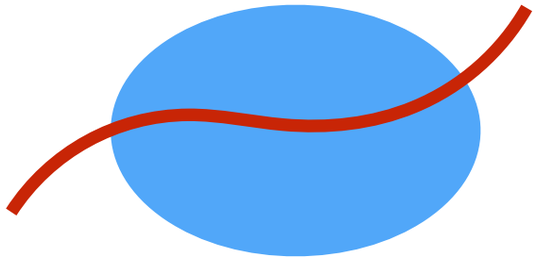
Gene regulation by histone modifications



Background: Protein-DNA interaction

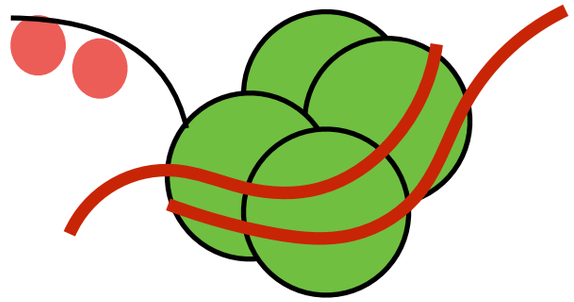
TF-TFBS

smaller regions



Histone modification

larger regions



ChIP-seq

chromatin immunoprecipitation *(ChIP)*
with massively parallel DNA sequencing *(seq)*

To identify the genome-wide locations of DNA binding proteins

Background: ChIP-seq data analysis

Acquired short sequences

reads

Alignment (reads mapping)

BAM

signal

Peak calling

BED

genomic region

genomic region set

Background

- Massive amounts of epigenetic data are produced by NGS techniques, such as ChIP-seq.

- The analysis of such data is mostly based on the manipulation of two common data structures:

1. **genomic signals**, which indicate the abundance of a ChIP-seq reads on genome; and

2. **genomic regions**, which represent candidate regions with the binding of a protein/histone with particular modifications to the DNA.

Histone code

H3K79me2 - Transcribed

H3K27ac - Active

H3K27me3 - Repressed

Transcription Factors

PU.1

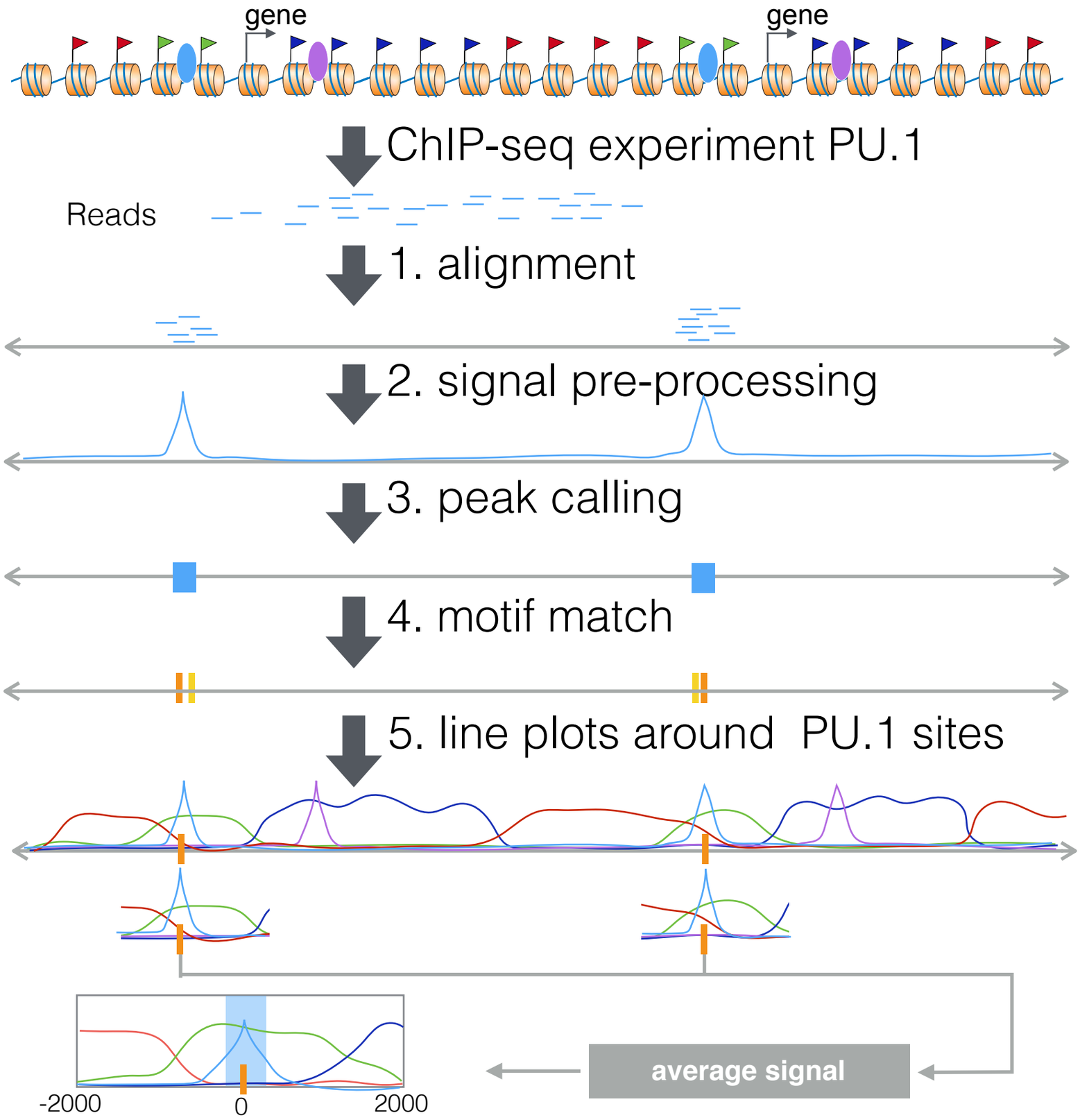
IRF4

Motif Database

PU.1

RUNX.1

…



low

medium

high

GenomicRegion

data: genomic range

set of

GenomicRegionSet

data: set of genomic range
file I/O: bed/bigBed
functions: range algebra

CoverageSet

data: genomic signals
file I/O: wig/bigWig
functions: soothing, frag. size,…

set of

MultiCoverageSet

data: sets of genomic signals
file I/O: wig/bigWig
functions: normalization

is

AnnotationSet

data: gene annotation
file I/O: gtf
functions: gene annotation

is

GenomicVariantSet

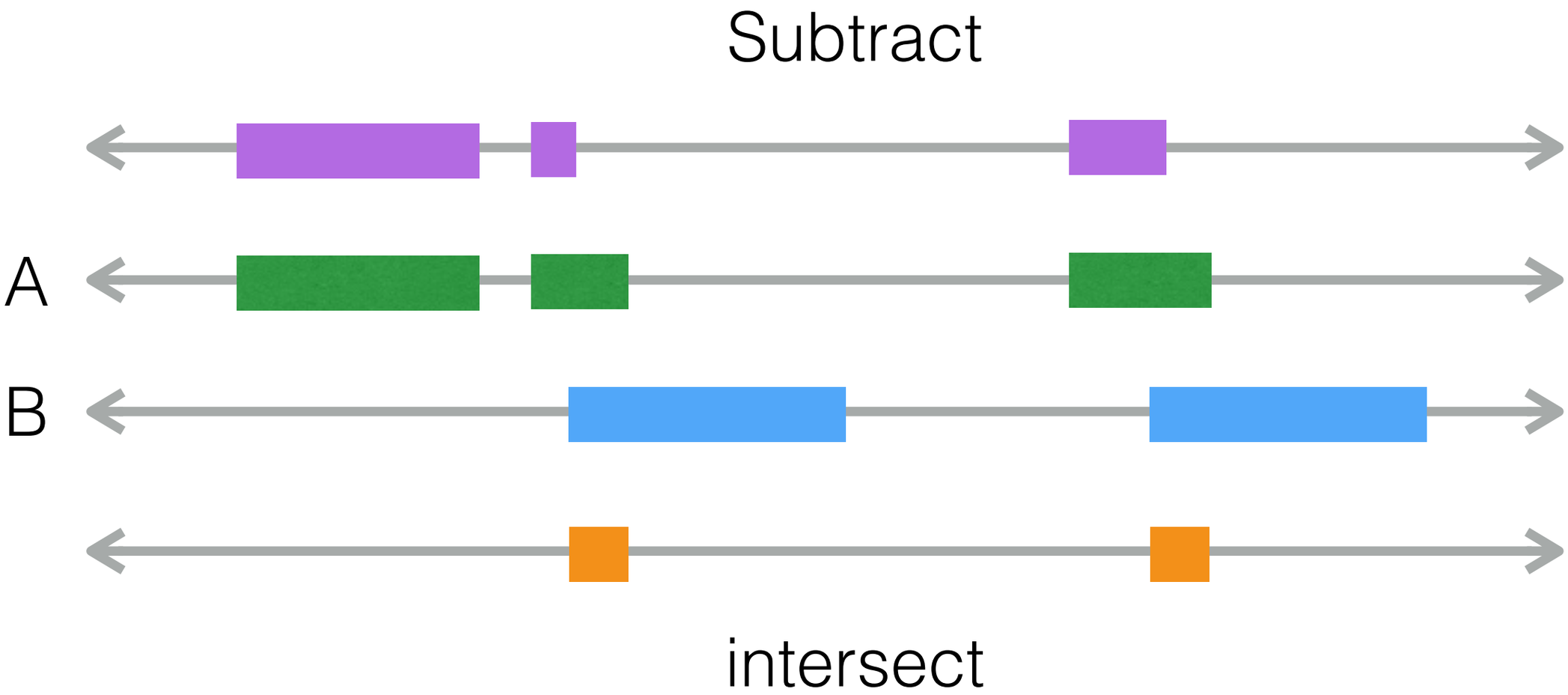
data: SNPs
file I/O: vcf

Core classes

- GenomicRegion
- GenomicRegionSet
- AnnotationSet
- GeneSet
- CoverageSet

GenomicRegion

GenomicRegionSet



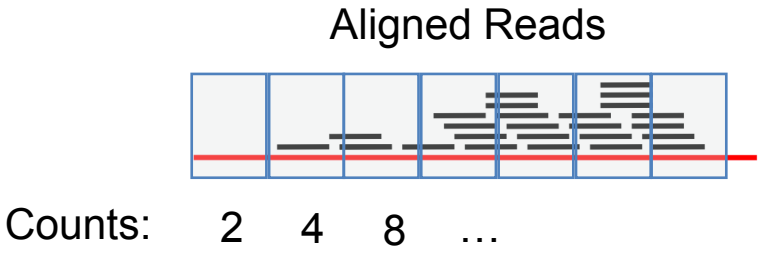
Let’s do it

Creating a Simple Peak Caller

Creating a simple peak caller

- Using RGT functions in Python.

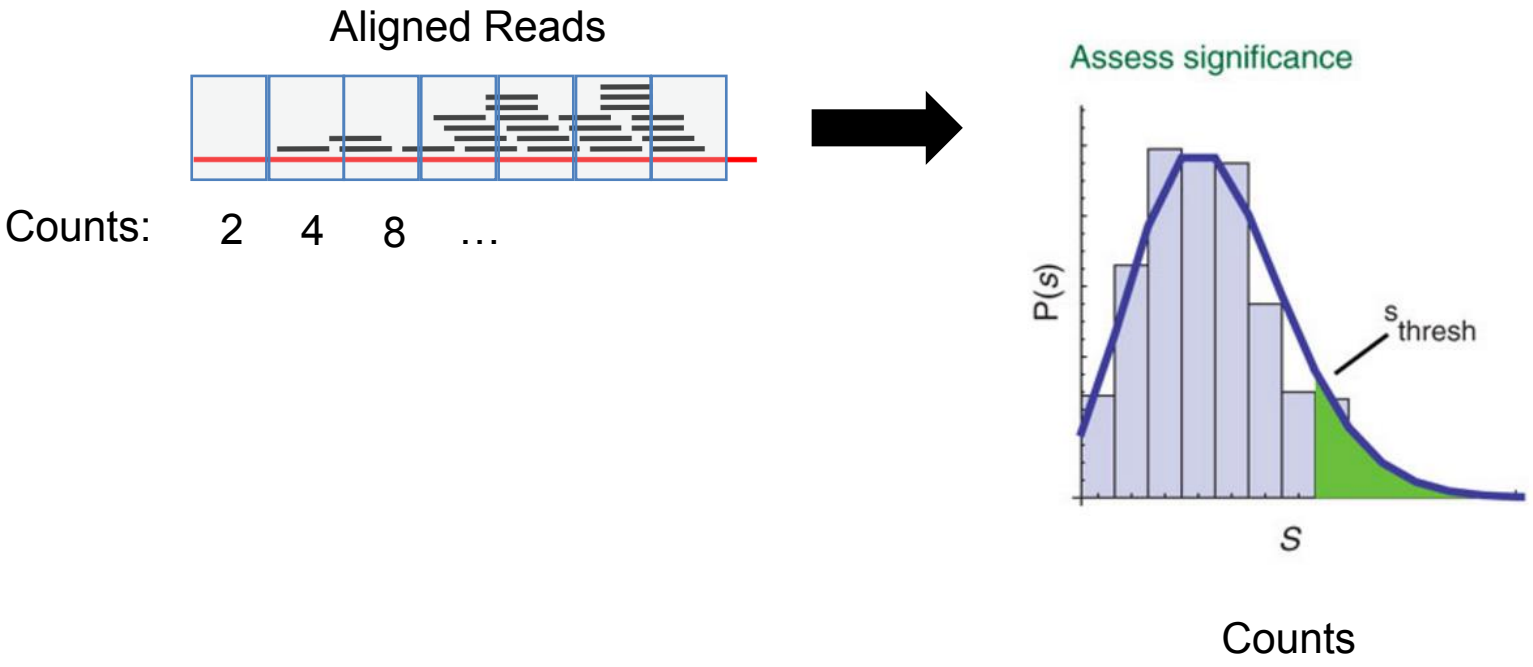
- Same basic idea of previous lectures.



Creating a simple peak caller

- Using RGT functions in Python.

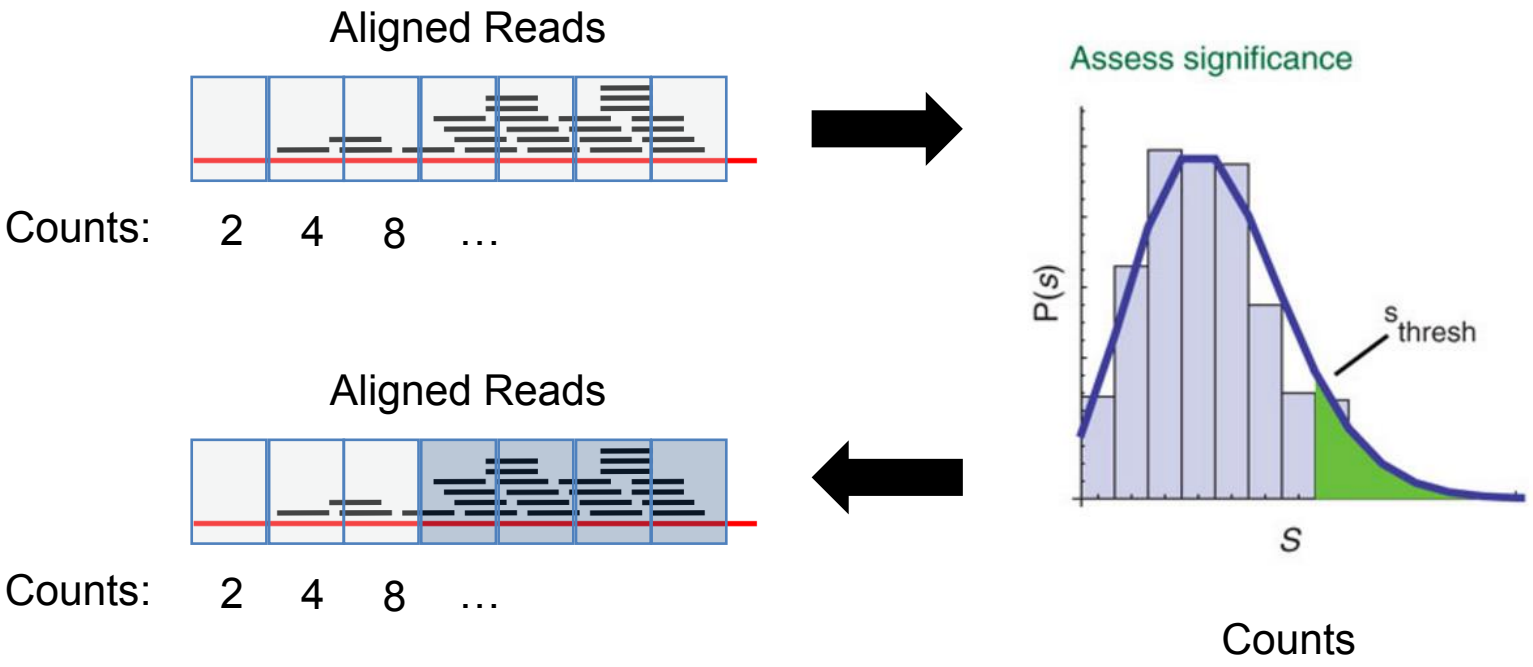
- Same basic idea of previous lectures.



Creating a simple peak caller

- Using RGT functions in Python.

- Same basic idea of previous lectures.



Our Peak Calling Pipeline

1. Normalize for CG content.

2. Normalize with input-DNA.

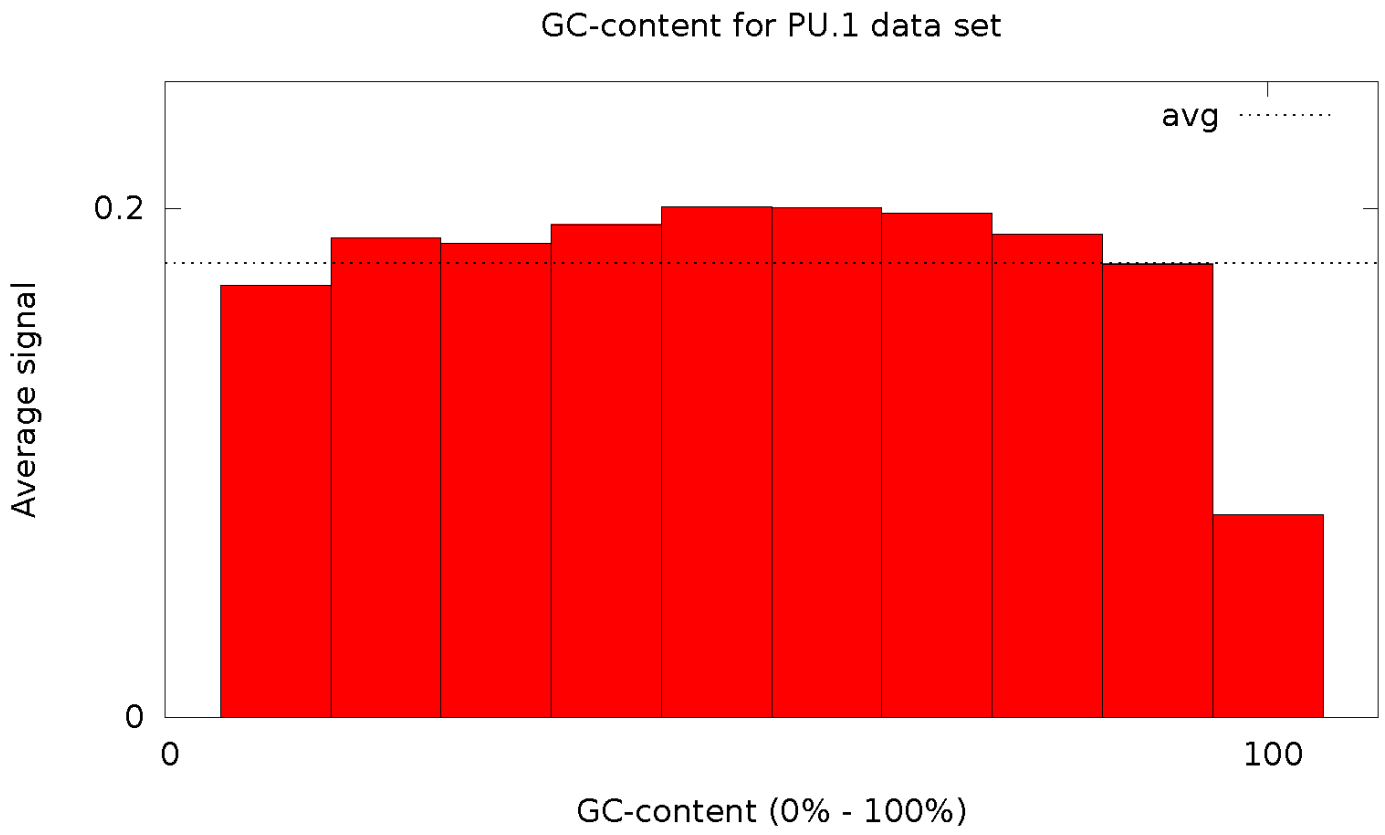
3. Use a binomial distribution to model read coverage.

4. Iterate over genomic bins performing binomial test.

5. Store the bins that pass the test.

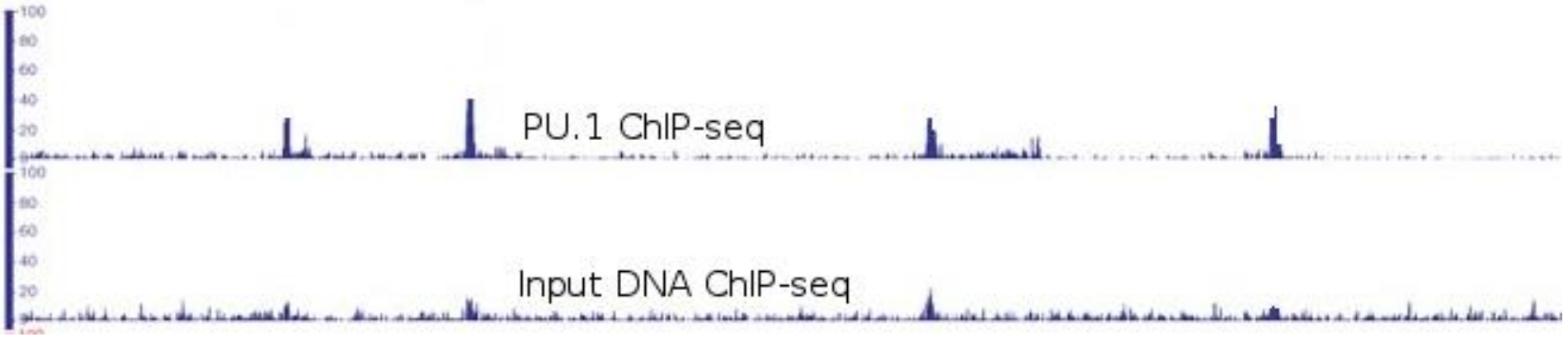
Normalize for CG content

- Coverage (signal intensity) varies given the frequency of C's and G's.



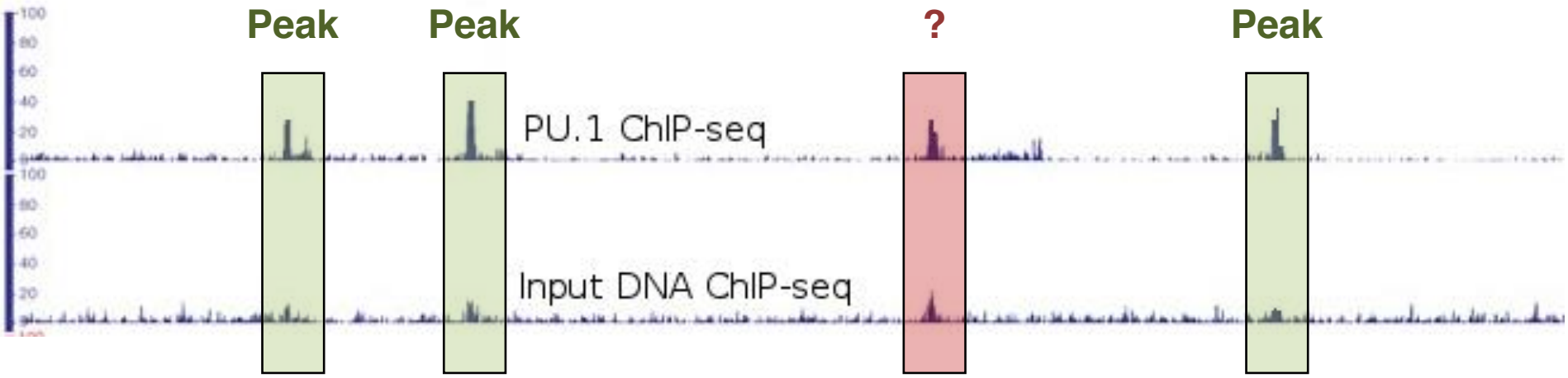
Normalize with Input DNA

- Input DNA ChIP-seq: A ChIP-seq experiment performed without the ChIP step.
- Many steps of the biological and computational process bias the signal's intensity within certain regions.



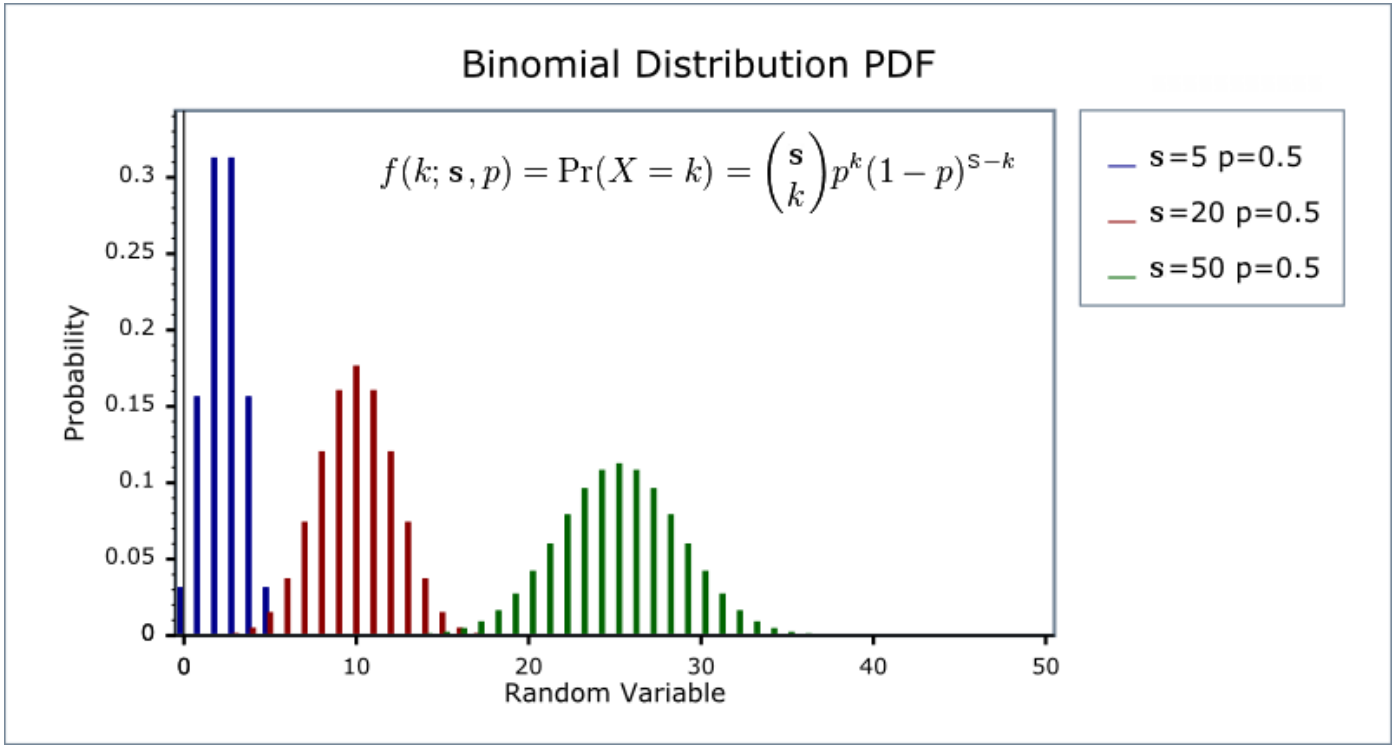
Normalize with Input DNA

- Input DNA ChIP-seq: A ChIP-seq experiment performed without the ChIP step.
- Many steps of the biological and computational process bias the signal's intensity within certain regions.



Model Distribution of Reads with Binomial

- Working assumption: ChIP-seq reads falling into a bin follow a Binomial distribution with parameters **s** and **p**.
- **s** = number of events = number of reads in the ChIP-seq library.
- **p** = probability of event = chance that a read falls into a bin.



Implementing peak caller

- Code in practices/2_peak_caller/peak_caller.py

- Execution:

```
cd practices/2_peak_caller/
python peak_caller.py
```

- Understanding the code

RGT-viz

Joseph Kuo

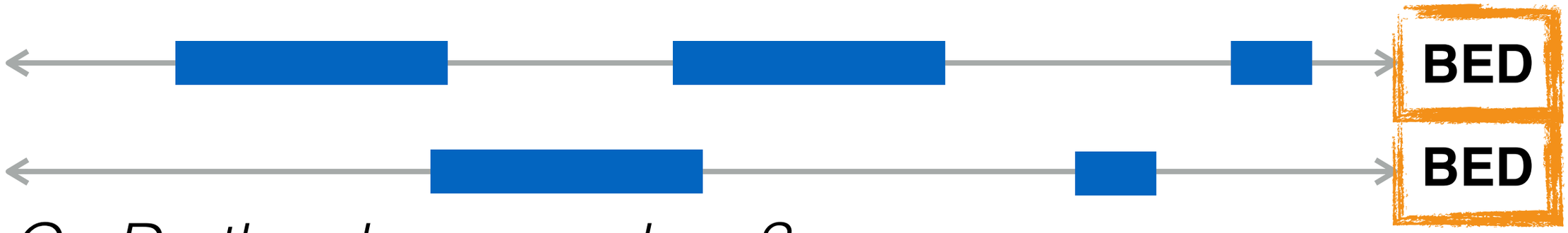
UNIKLINIK RWTH AACHEN

Computational Biology Group

IZKF AACHEN

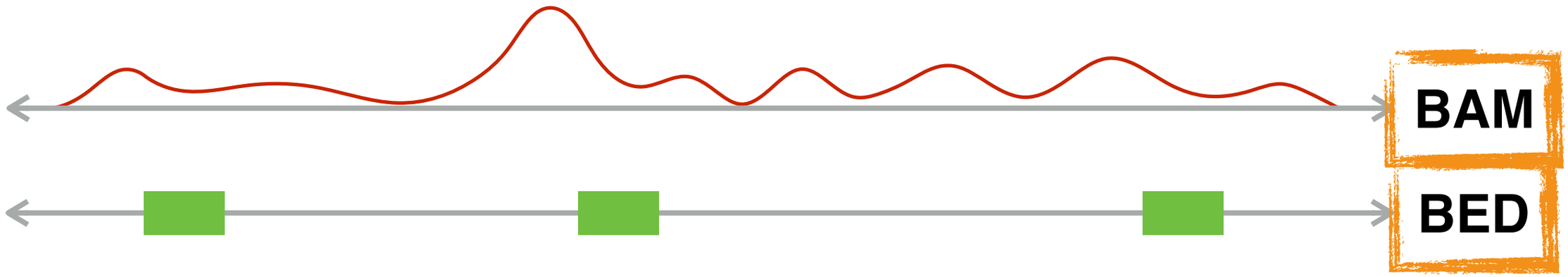
Objectives

1. **Association between two genomic region sets**



Q: Do they have overlaps?
These overlaps are due to chance or not?

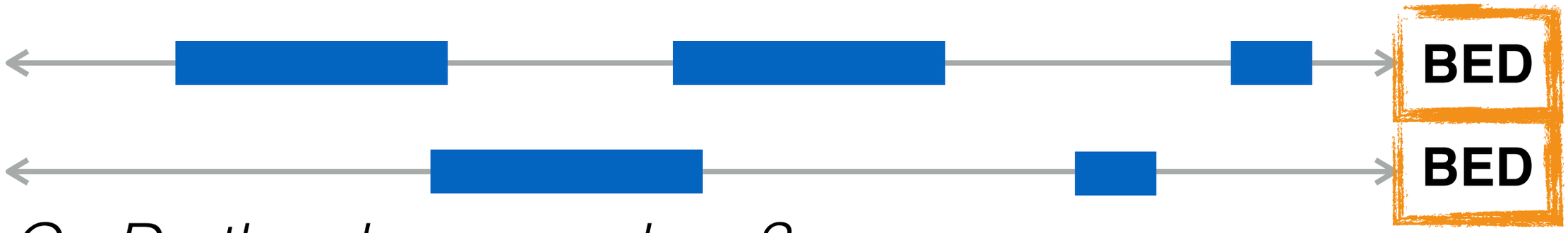
2. **Association between regions to signal**



Q: Is there stronger signal in the interested regions? How does it distribute?

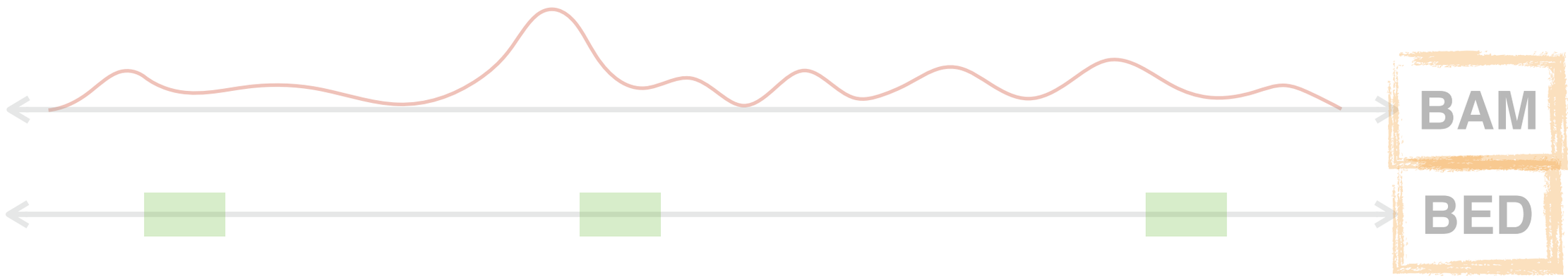
Objectives

1. Association between two genomic region sets



Q: Do they have overlaps?
These overlaps are due to chance or not?

2. Association between regions to signal

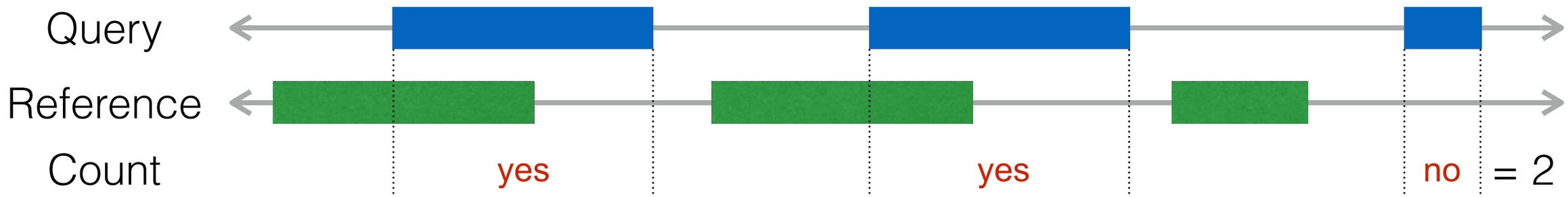


Q: Is there stronger signal in the interested regions? How does it distribute?

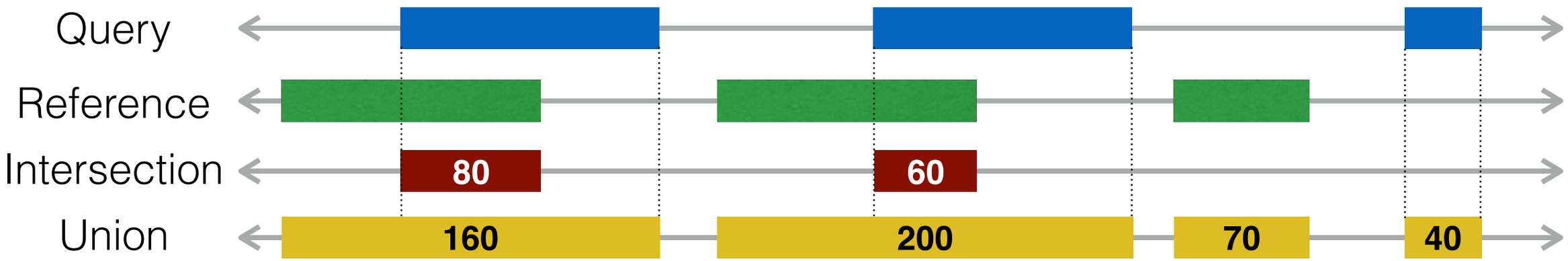
Evaluate association between regions

With overlap

1. Count the number of overlapped regions:



2. Measure the amount of overlapped regions:



The unit here is bps

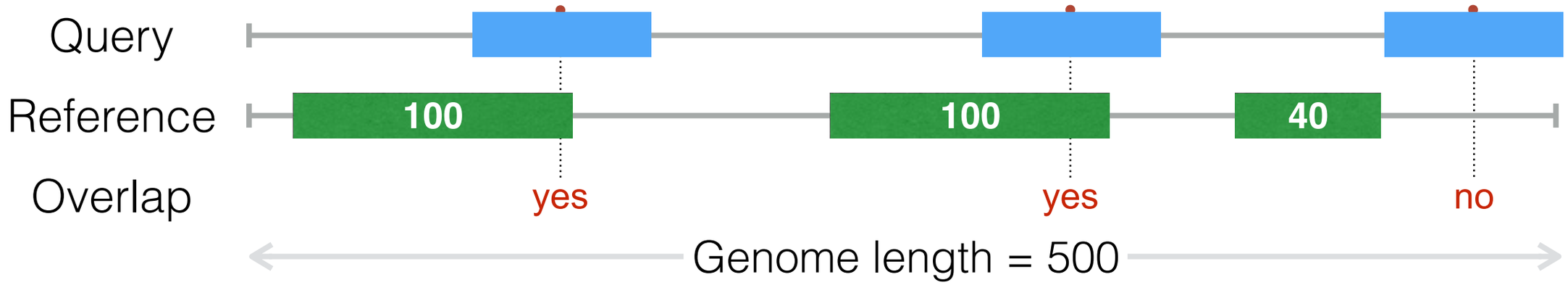
Tests for regions v.s. regions

	Measure object	Reference size	Query size
Projection test	number of overlaps	large	any
Intersection test	number of overlaps	any	any
Jaccard test	amount of overlap	not small	not small
Combinatorial test	number of overlap among different combinations	large	not small

Tests for regions v.s. regions

	Measure object	Reference size	Query size
Projection test	number of overlaps	large	any
Intersection test	number of overlaps	any	any
Jaccard test	amount of overlap	not small	not small
Combinatorial test	number of overlap among different combinations	large	not small

Projection test



$$\text{Background}(P) = \frac{\text{Coverage of reference}}{\text{Coverage of genome}} = \frac{100 + 100 + 40}{500} = 0.48$$

n = Number of regions in query = 3

x = Number of overlaps = 2

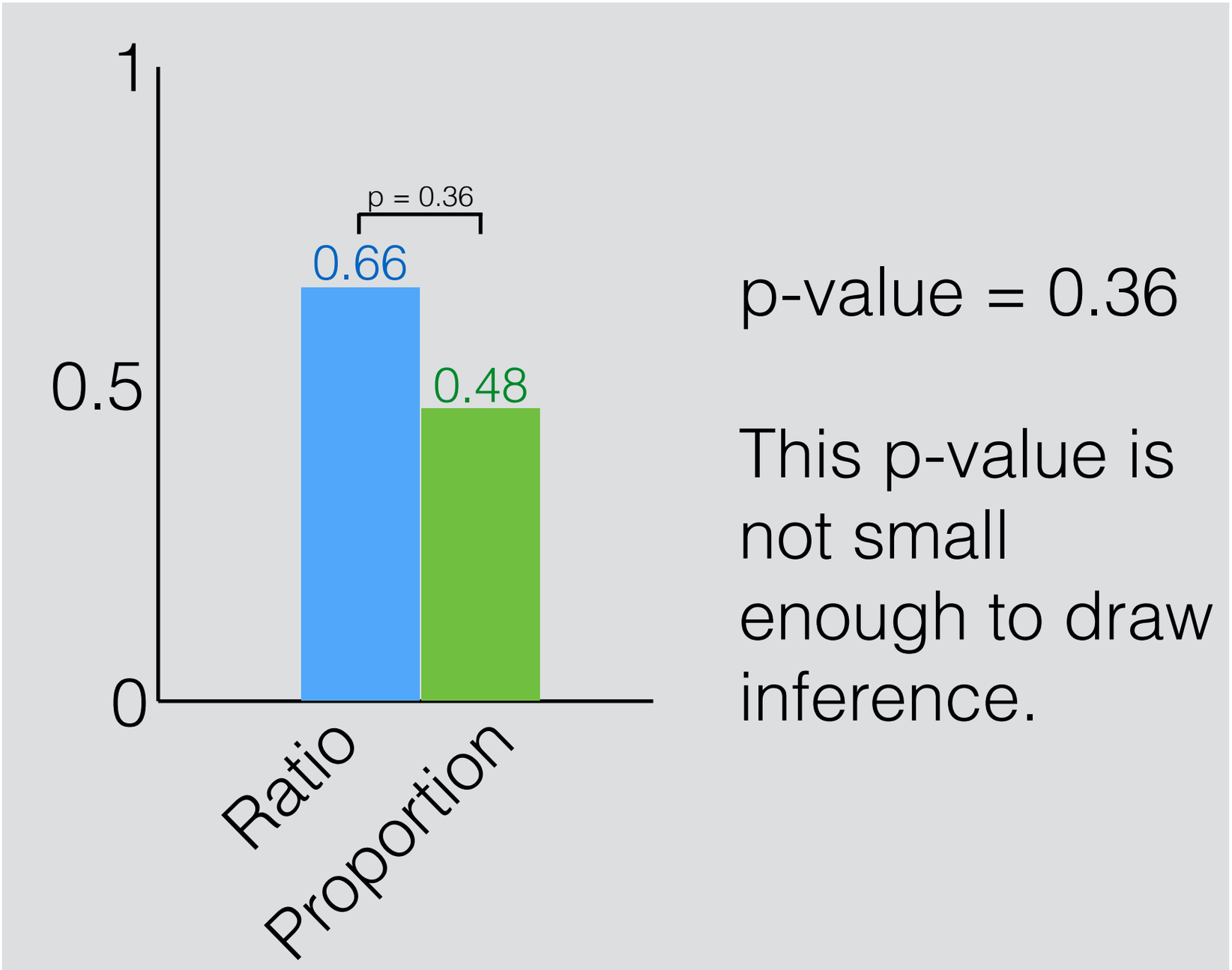
ratio(r) = x/n = 2/3

Using binomial test to evaluate their overlap is due to chance or not

$$\text{p-value} = \binom{n}{x} P^x (1-P)^{n-x}$$

p-value small: association exists
otherwise: overlap by chance

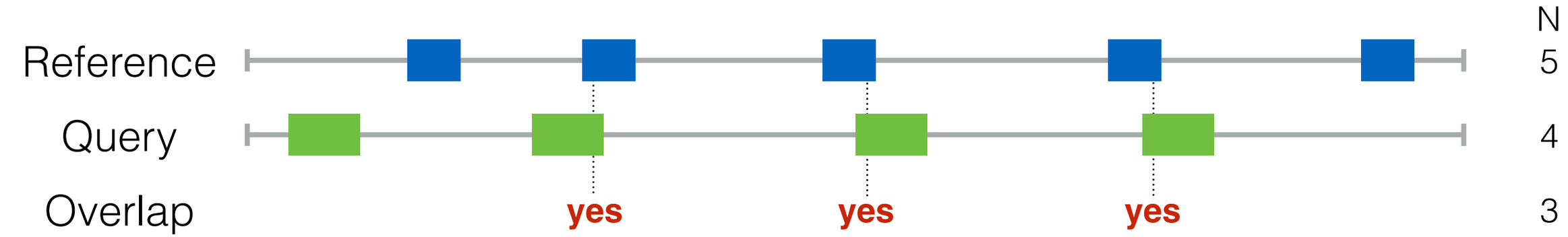
Projection test



Tests for regions v.s. regions

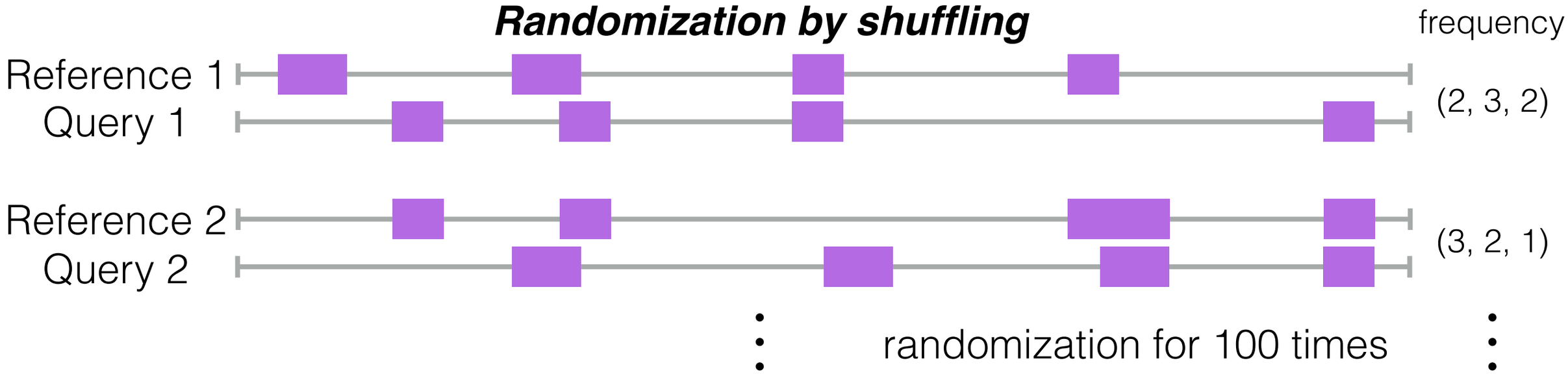
	Measure object	Reference size	Query size
Projection test	number of overlaps	large	any
Intersection test	number of overlaps	any	any
Jaccard test	amount of overlap	not small	not small
Combinatorial test	number of overlap among different combinations	large	not small

Intersection test



Observed frequency = (Overlaps, Only reference, Only query)
= (3, 5 - 3, 4 - 3)
= (3, 2, 1)

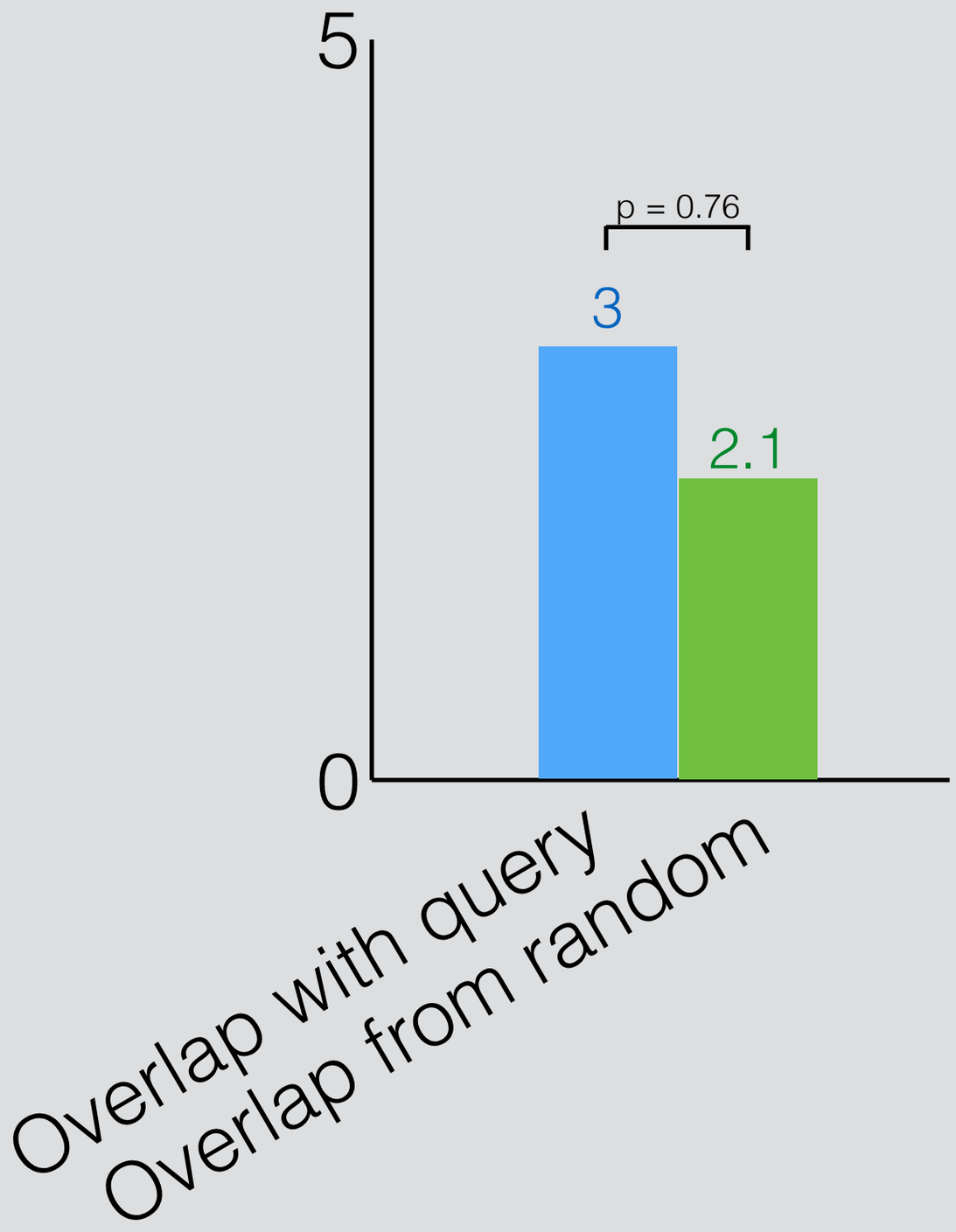
Randomization by shuffling



Expected frequency is the average of all frequencies from randomization

$$\chi^2 = \sum_{i=1}^{3} \frac{(F_i^{Ob} - F_{average,i}^{Ex})^2}{F_{average,i}^{Ex}}$$

Intersection test



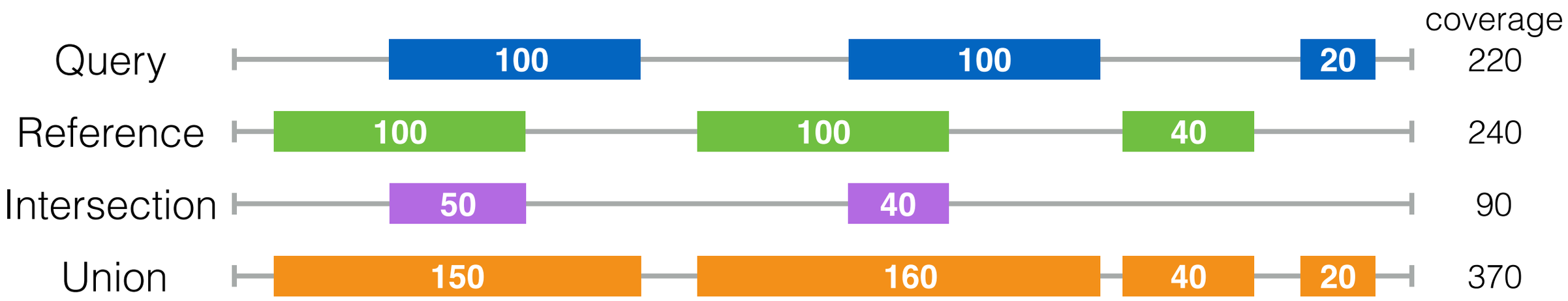
chi-square = 0.543
p-value = 0.76

This p-value is not small enough to draw inference.

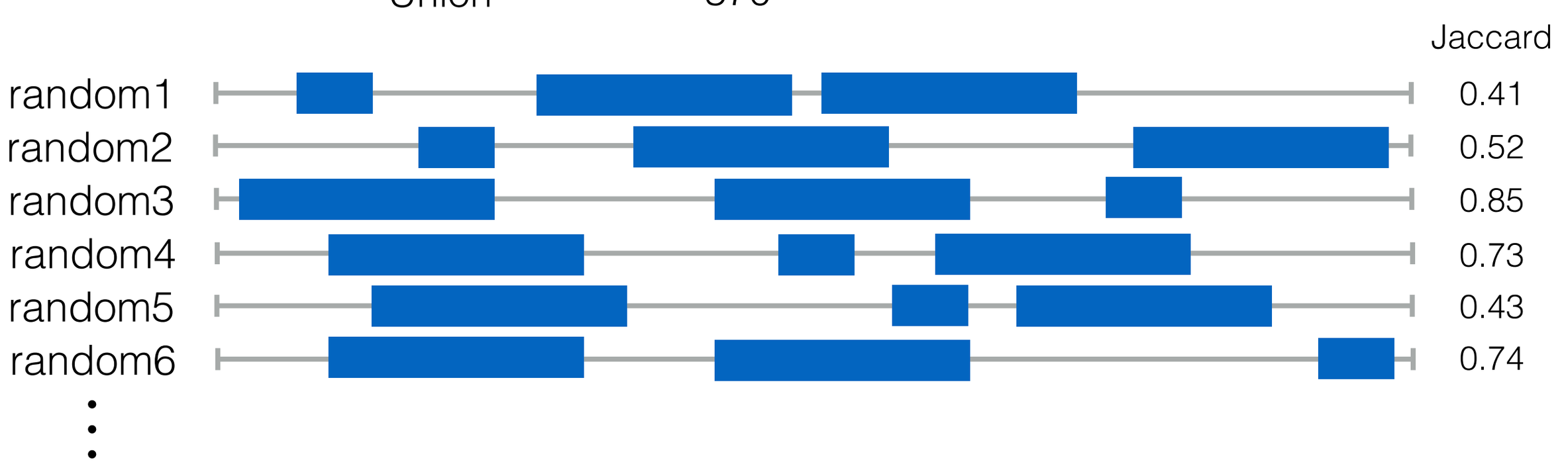
Tests for regions v.s. regions

	Measure object	Reference size	Query size
Projection test	number of overlaps	large	any
Intersection test	number of overlaps	any	any
Jaccard test	amount of overlap	not small	not small
Combinatorial test	number of overlap among different combinations	large	not small

Jaccard test

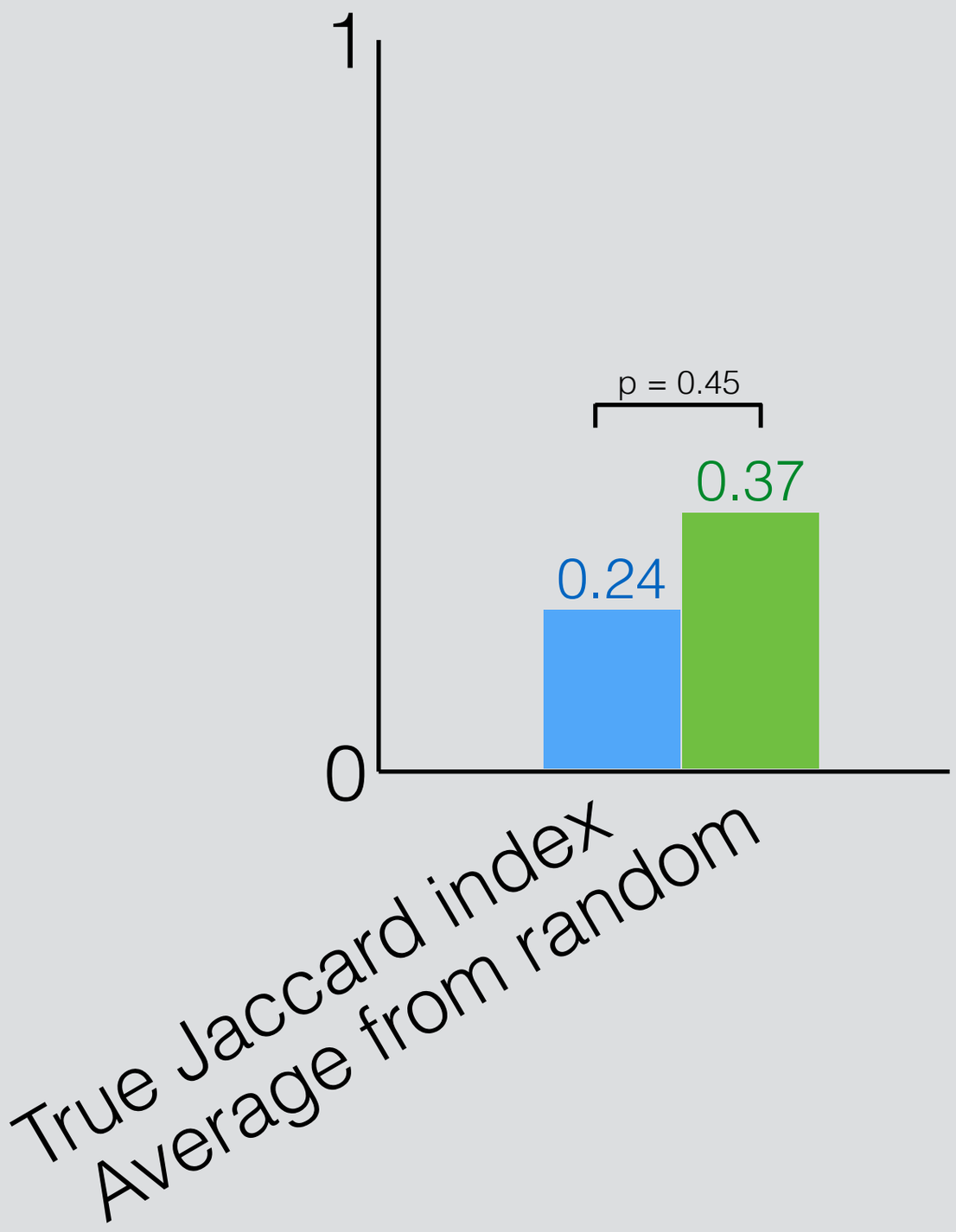


Jaccard index $= \frac{\text{Intersection}}{\text{Union}} = \frac{90}{370} = 0.24$



By comparing true Jaccard with random Jaccards,
the chance that true Jaccard is due to random is calculated.

Jaccard test



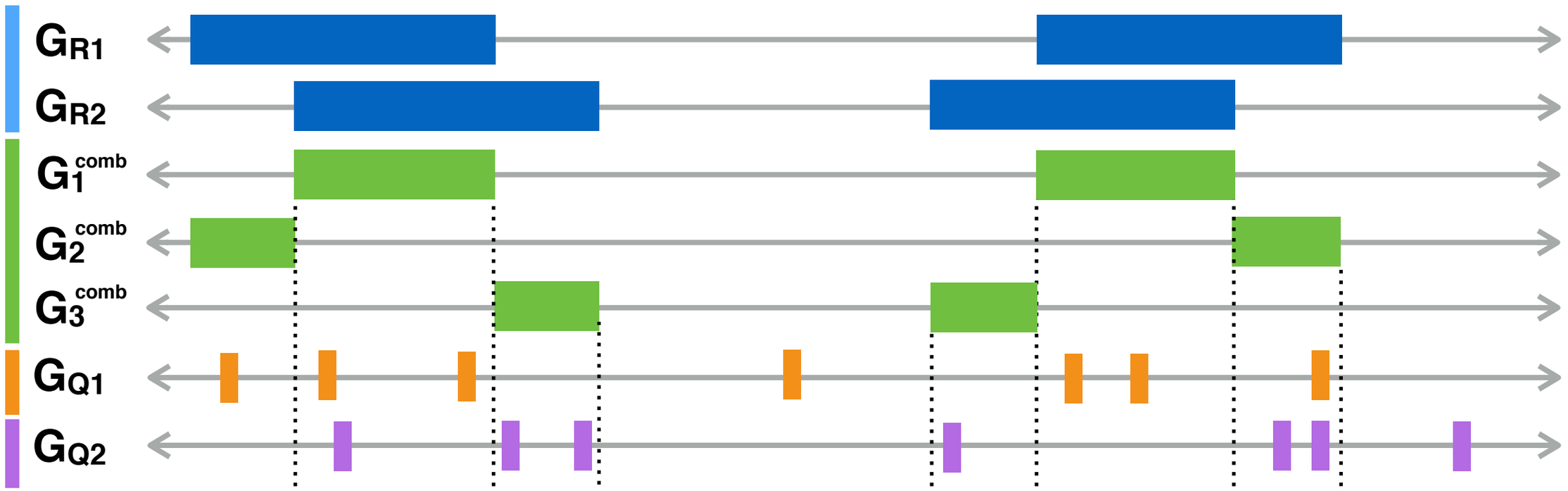
p-value = 0.45

This p-value is not small enough to draw inference.

Tests for regions v.s. regions

	Measure object	Reference size	Query size
Projection test	number of overlaps	large	any
Intersection test	number of overlaps	any	any
Jaccard test	amount of overlap	not small	not small
Combinatorial test	number of overlap among different combinations	large	not small

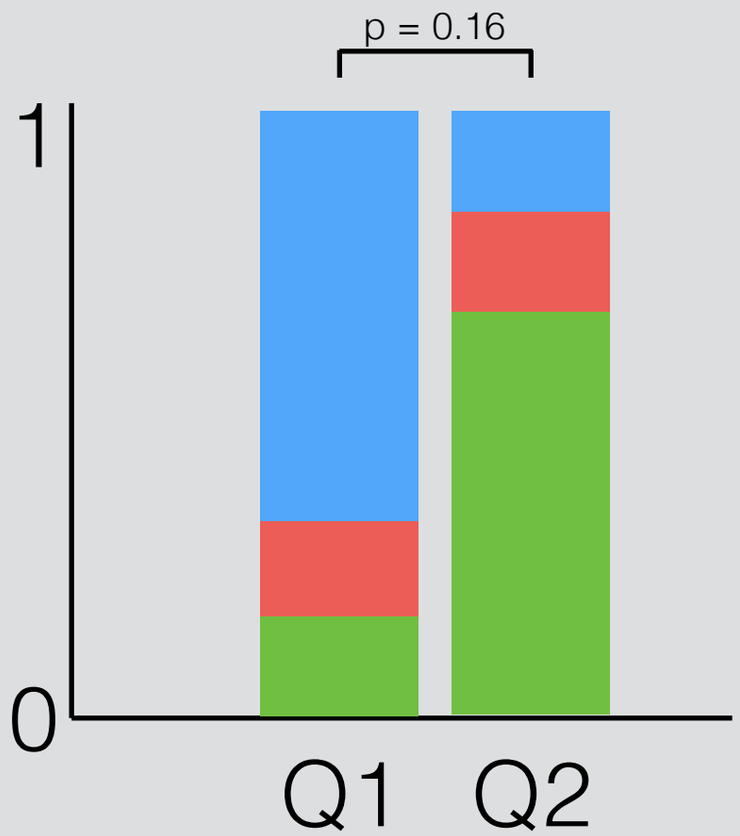
Combinatorial test



$$F_{Q1} = (F_1, F_2, F_3) = (4,1,1)$$
$$F_{Q2} = (F_1, F_2, F_3) = (1,1,4)$$

Using Chi-square test to test their difference

Combinatorial test

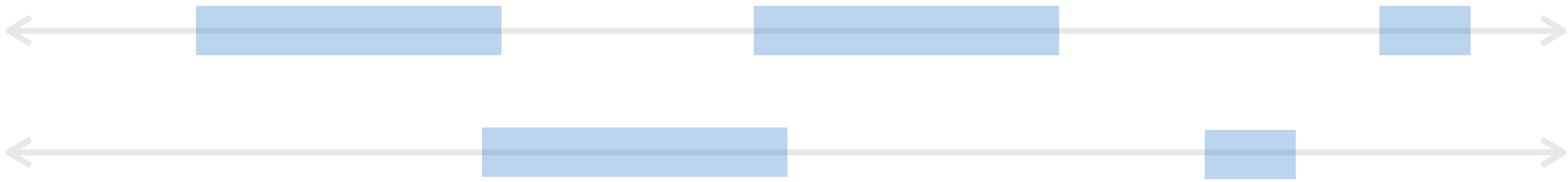


chi-square = 3.6
p-value = 0.16

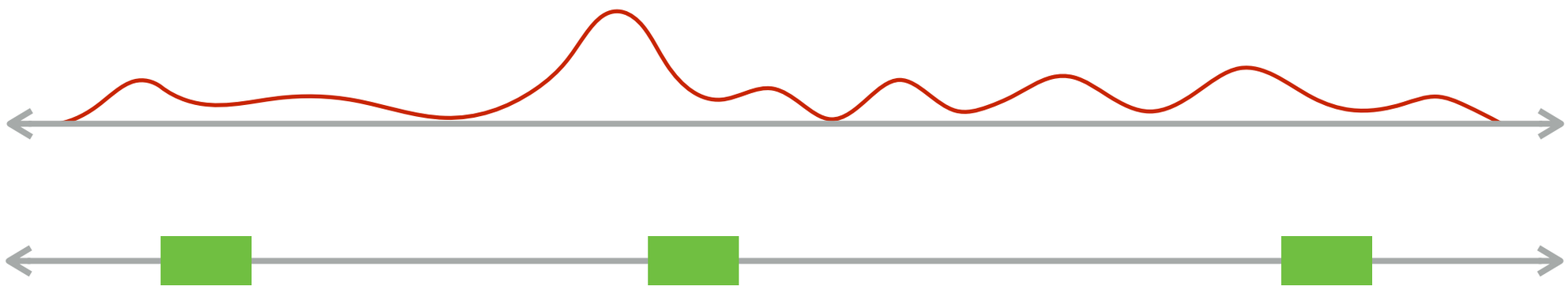
This p-value is not small enough to draw inference.

Objectives

1. Association between two genomic region sets



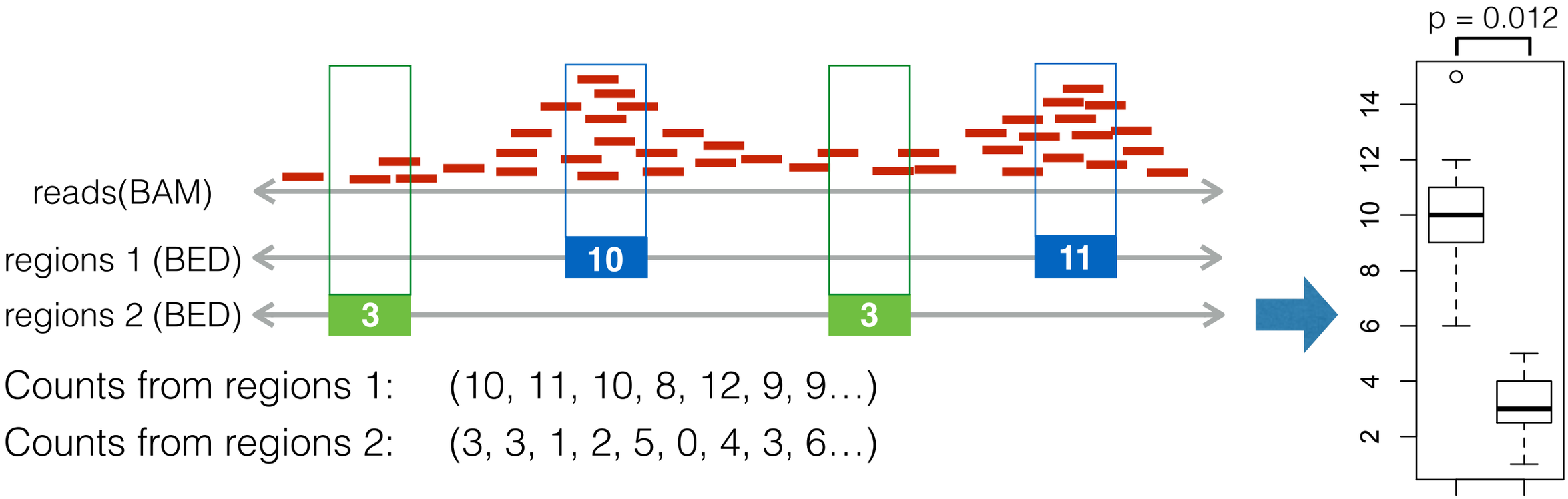
2. Association between regions to signal



Association between regions to signal

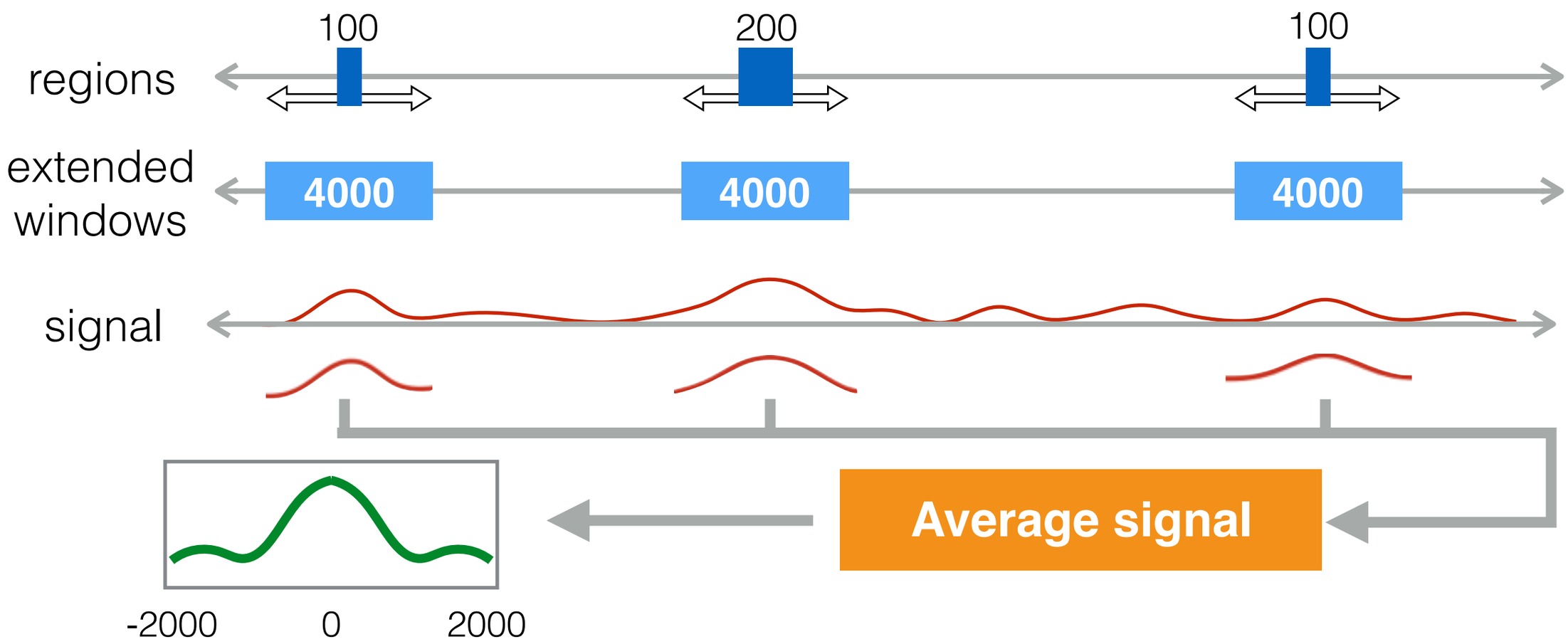
1. Boxplot
2. Lineplot
3. Heatmap

Boxplot (1/3)



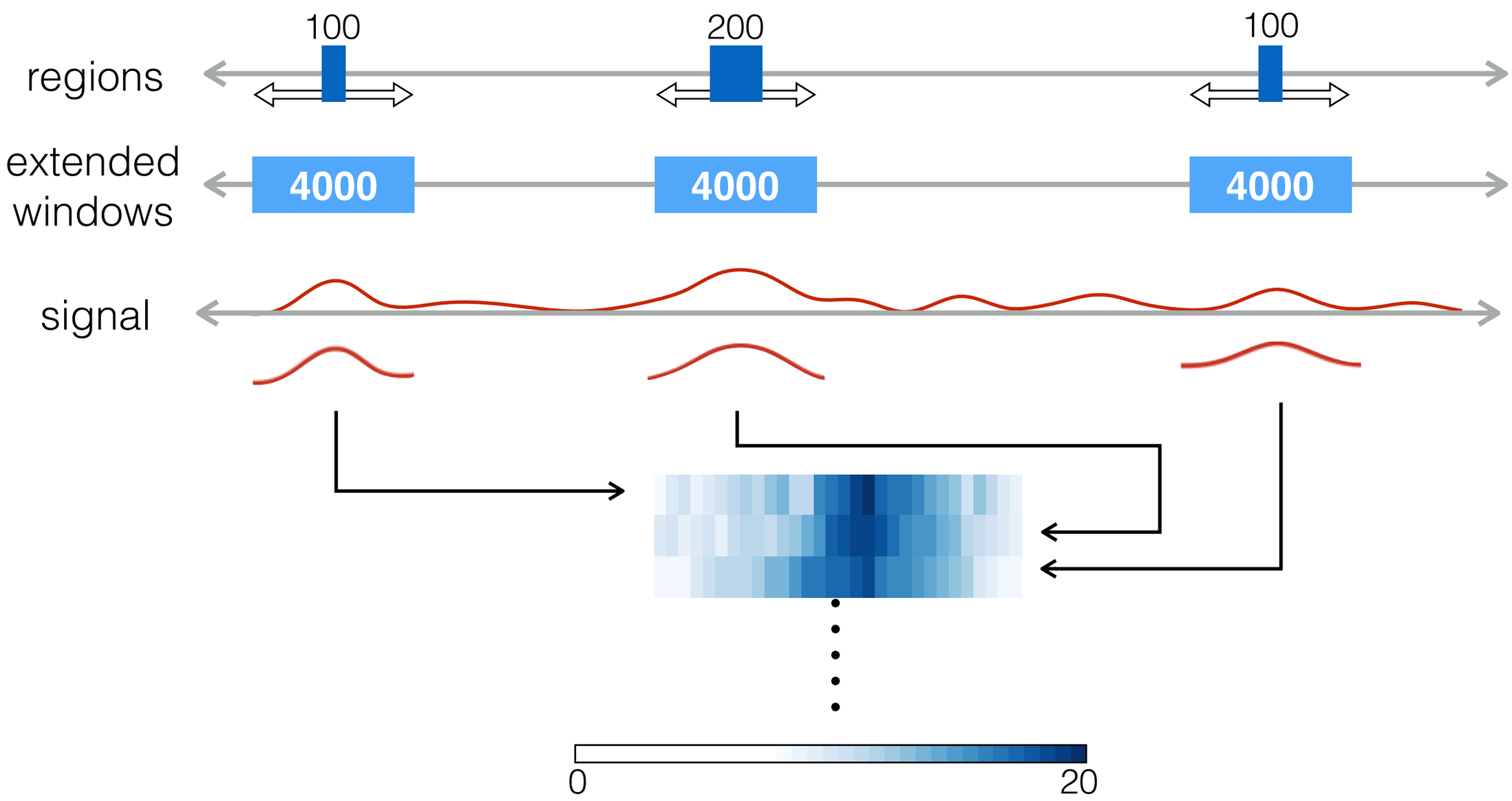
Obtain summarized result but lose detailed information.

Lineplot (2/3)



Display averaged spatial distribution of the reads on the given regions, but lose each single data point.

Heatmap (3/3)

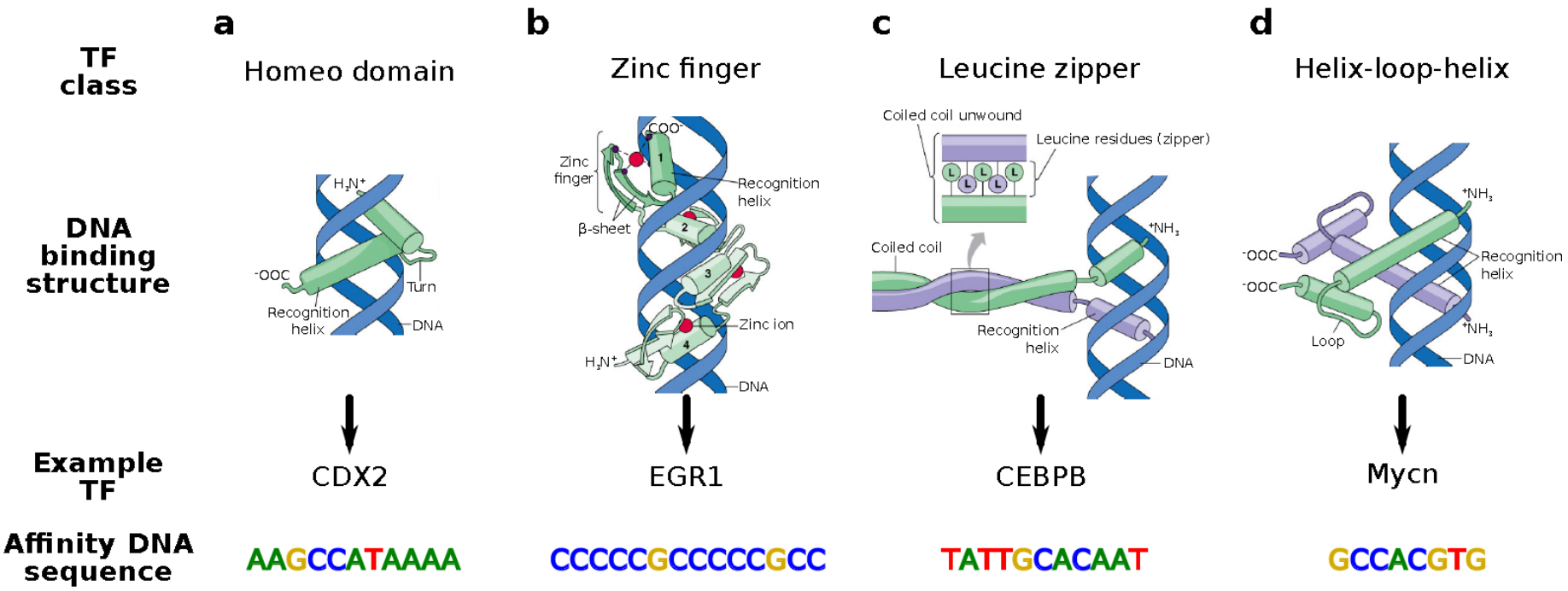


Display the signal by different colors and can preserve each single counting number.

Motif Analysis

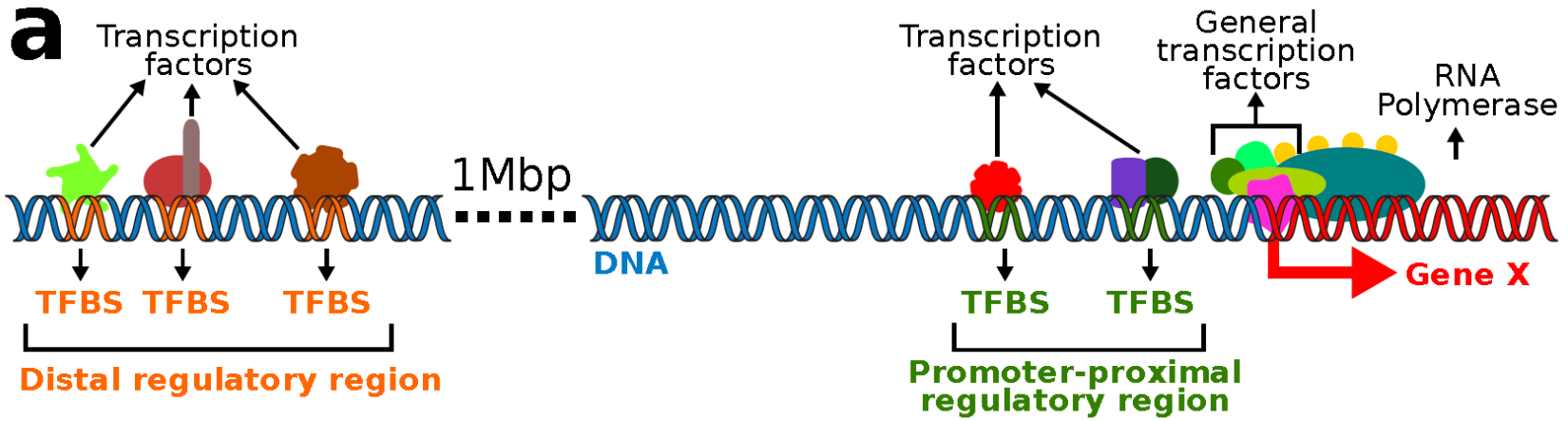
Motif Analysis

- DNA-binding proteins have sequence affinity.



Motif Analysis

- DNA-binding proteins work together (co-binding).



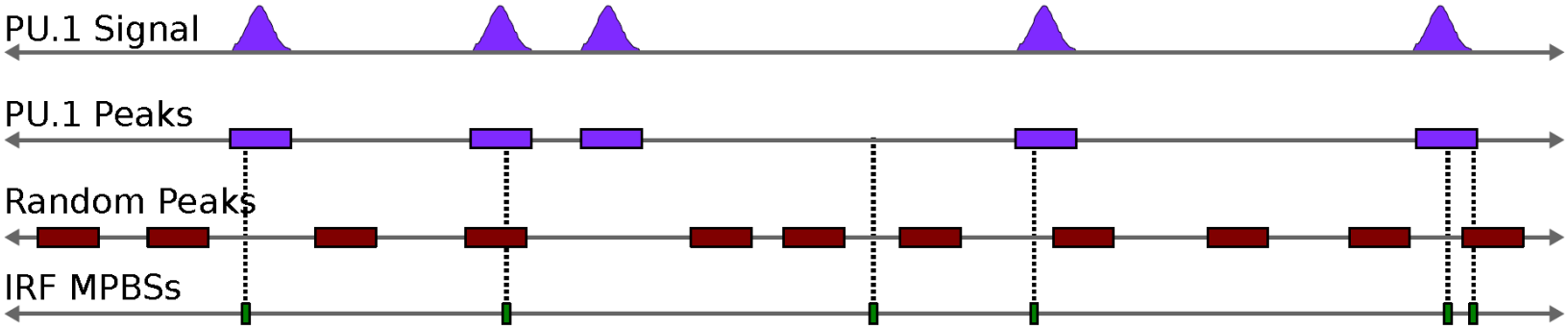
Motif Analysis

- The peaks called represent putative PU.1 binding sites.

- **Goal:** We want to check whether other transcription factors are enriched in these PU.1 binding regions.

- Use the motif enrichment analysis available in RGT.

Motif Enrichment Analysis

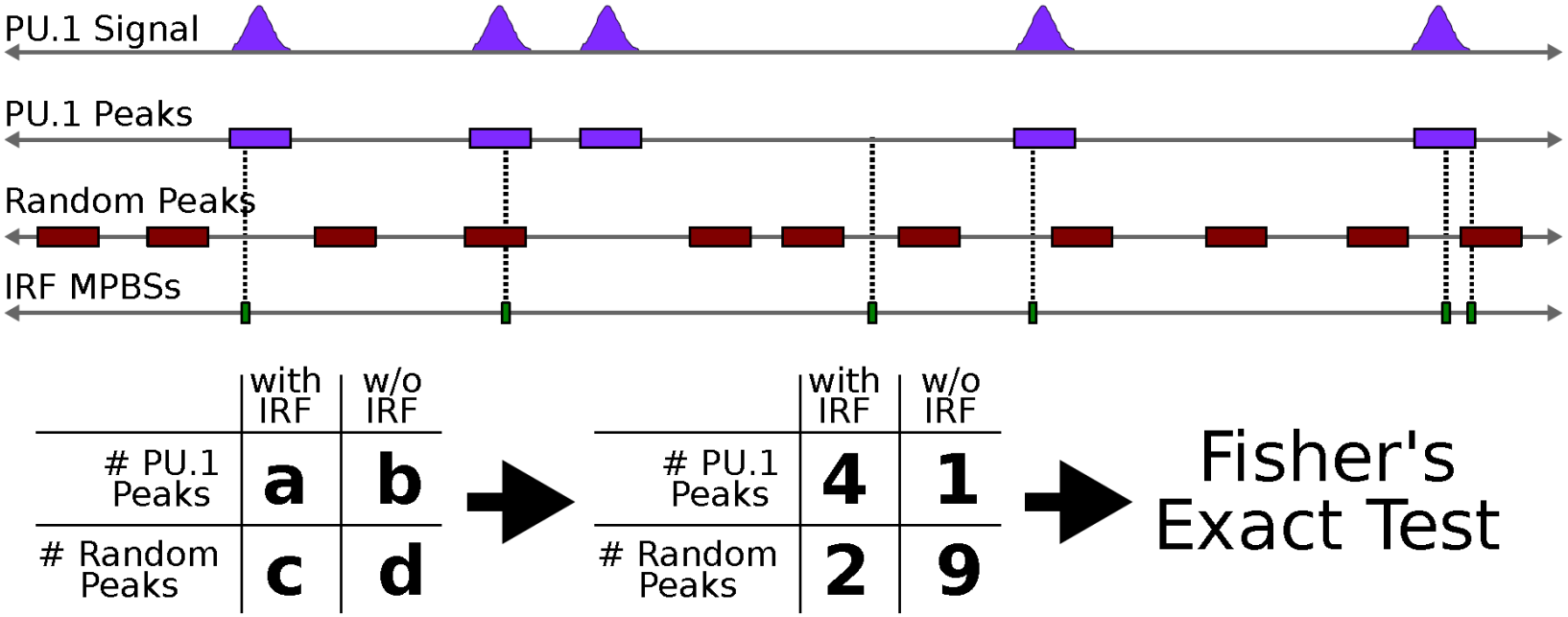


	with IRF	w/o IRF
# PU.1 Peaks	a	b
# Random Peaks	c	d

	with IRF	w/o IRF
# PU.1 Peaks	4	1
# Random Peaks	2	9

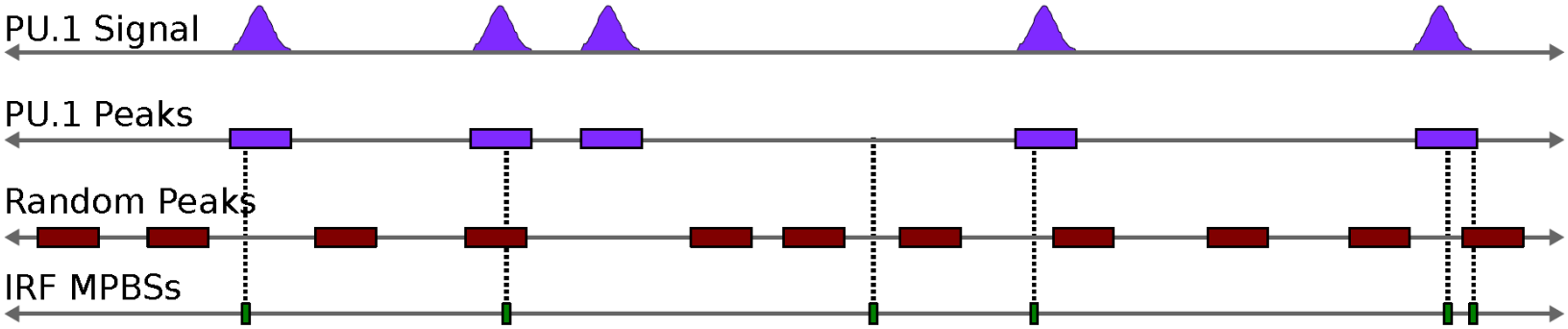
Fisher's Exact Test

Motif Enrichment Analysis



1. Find random regions.
2. Perform motif matching in the PU.1 peaks and random regions.
3. Count Fisher table.
4. Perform Fisher's Exact test.

Motif Enrichment Analysis



	with IRF	w/o IRF
# PU.1 Peaks	a	b
# Random Peaks	c	d

→

	with IRF	w/o IRF
# PU.1 Peaks	4	1
# Random Peaks	2	9

→ Fisher's Exact Test

1. Find random regions.
2. Perform motif matching in the PU.1 peaks and random regions.
3. Count Fisher table.
4. Perform Fisher's Exact test.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Performing Motif Enrichment Analysis

- Code in practices/4_motif_analysis/4_motif_analysis.py

- Execution:

```
cd practices/4_motif_analysis/
sh motif_analysis.sh
```

- Understanding the code:
    1. Motif matching
    2. Motif enrichment

Motif Enrichment Analysis Results