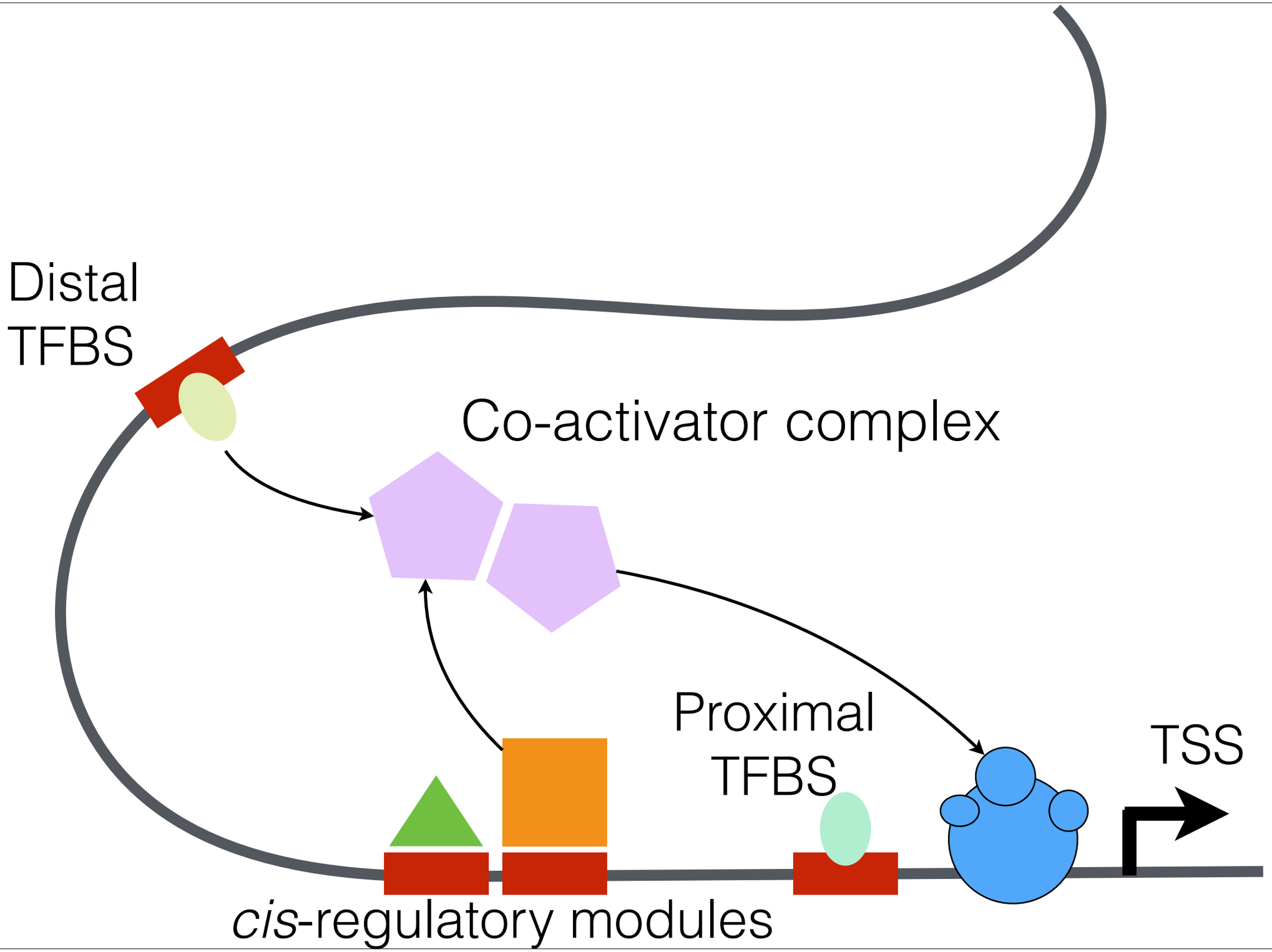
Bioinformatics Lab: Introduction of RGT

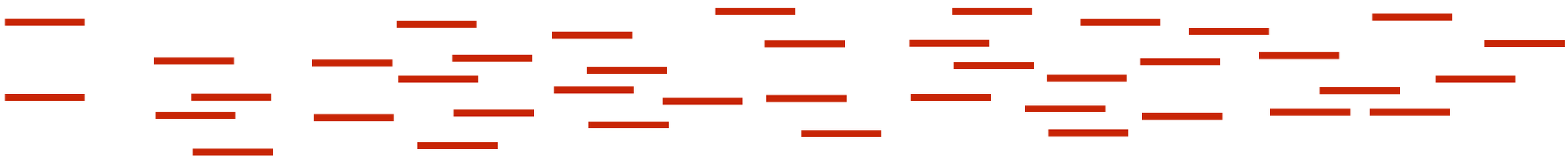
Ivan Gesteira Costa & Zhijian Li
Institute for Computational Genomics

Gene regulation by transcription factors



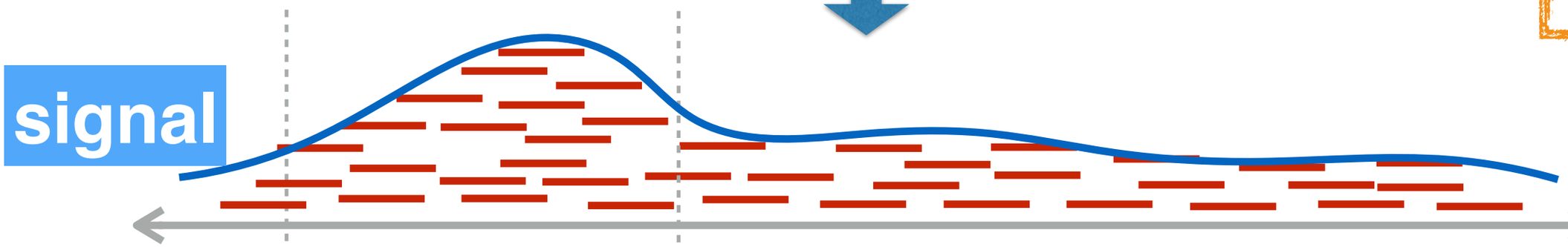
Example: ChIP-seq data analysis

Acquired short sequences



Alignment (reads mapping)

BW



Peak calling

BED

genomic region

genomic region set

Background of RGT

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

Background of RGT

- 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:

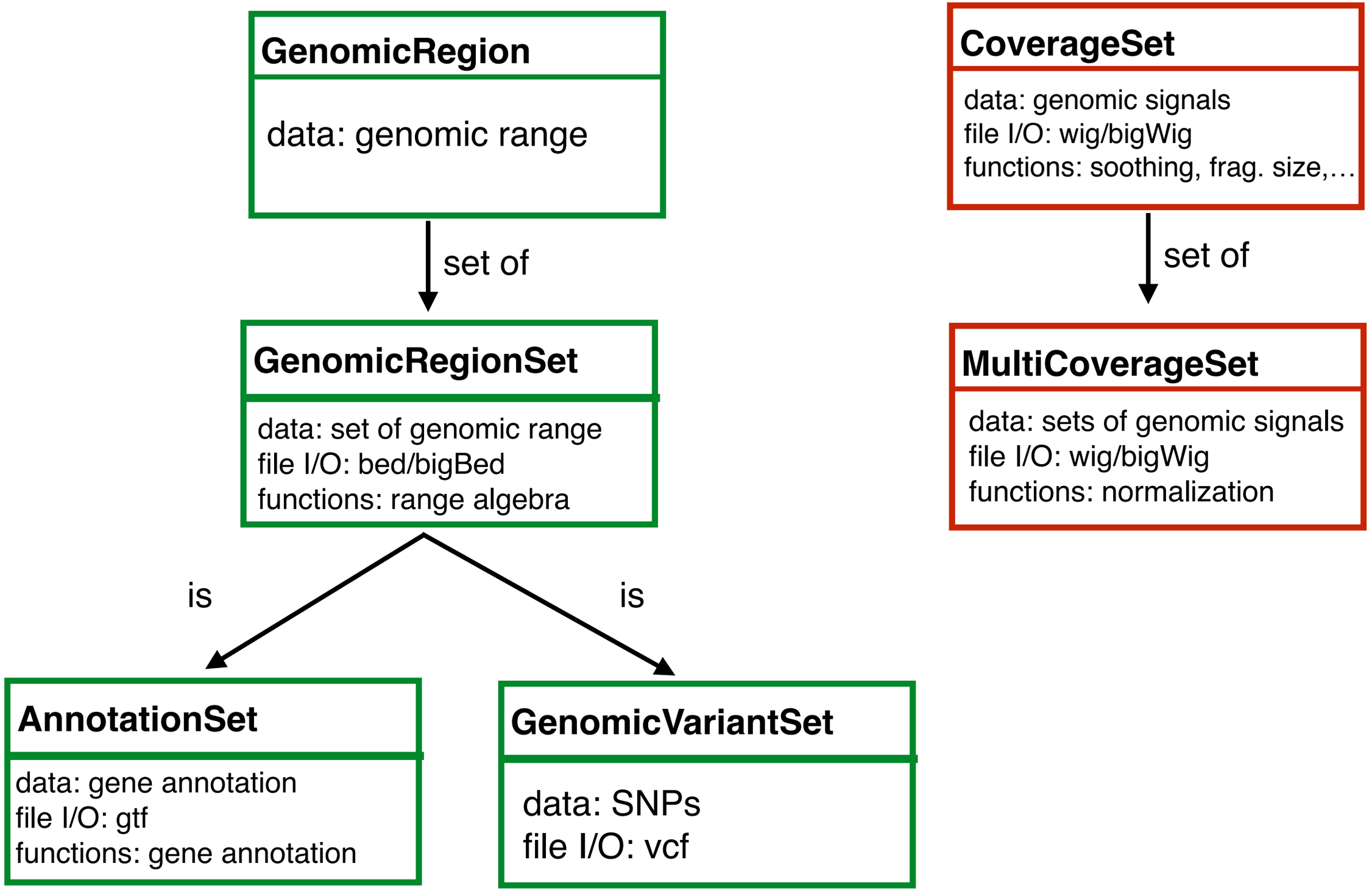
- **genomic signals**, which indicate the abundance of a ChIP-seq read on genome;

- **genomic regions**, which represent the regions that we are interested in

Core classes of RGT

- GenomicRegion
- GenomicRegionSet
- AnnotationSet
- GeneSet
- CoverageSet

Core classes of RGT



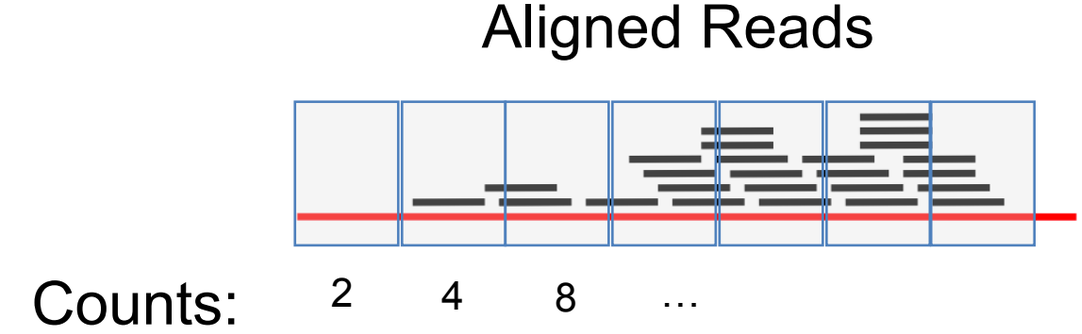
More Information about RGT

- http://www.regulatory-genomics.org/
- https://github.com/CostaLab/reg-gen

Create a Simple Peak Caller

- Using RGT functions in Python.

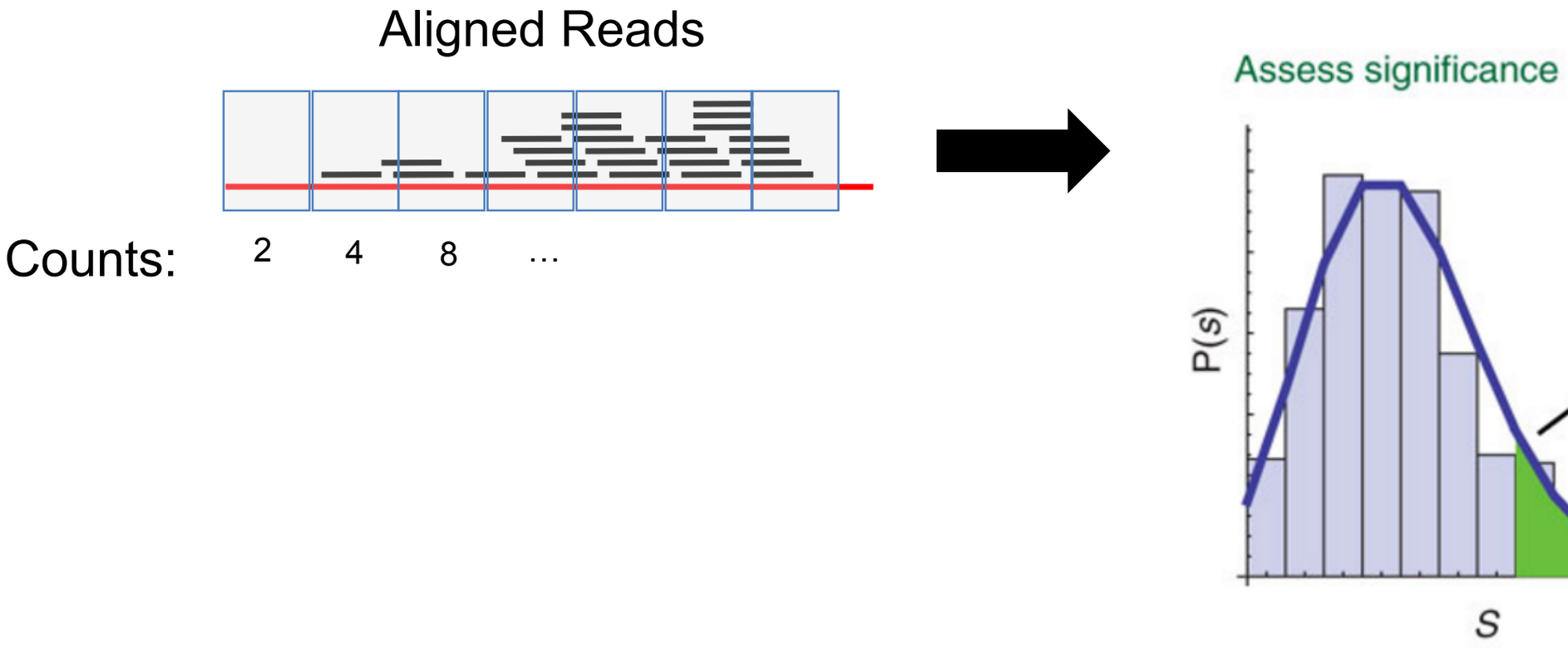
- Same basic idea of previous lectures.



Create a Simple Peak Caller

- Using RGT functions in Python.

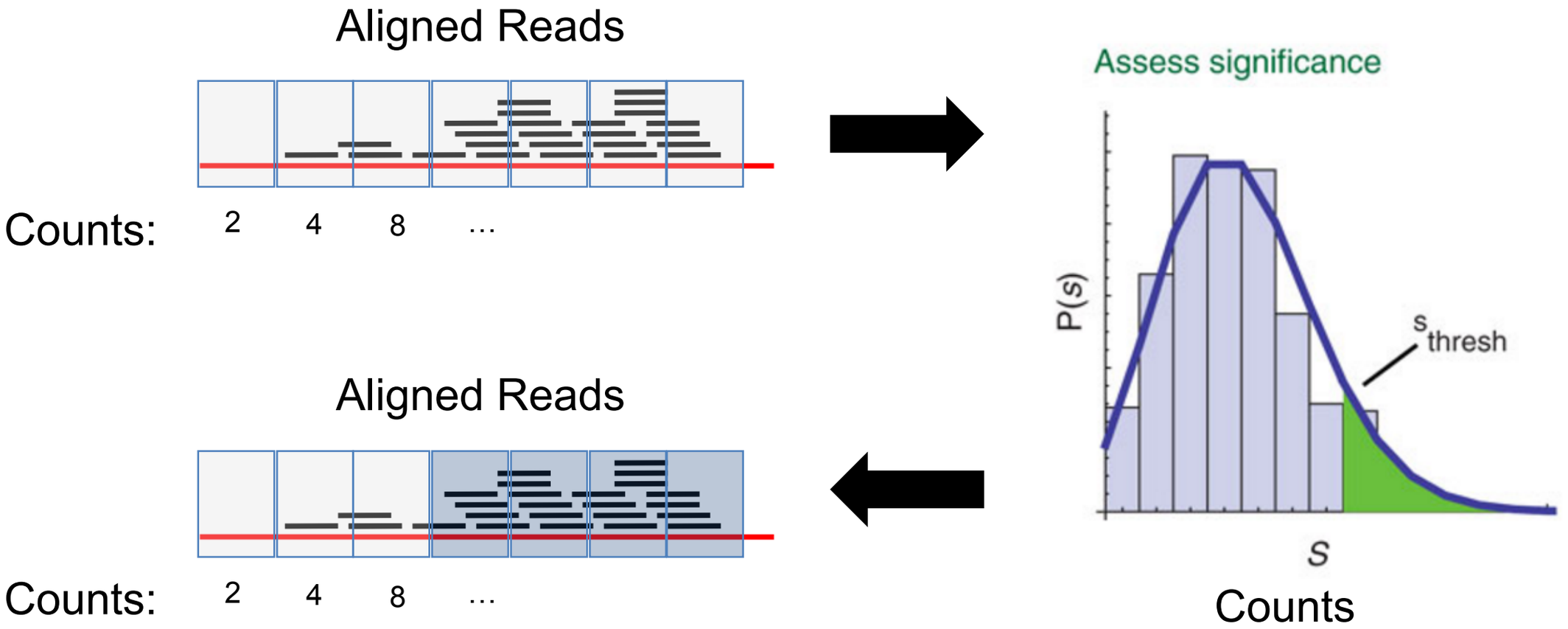
- Same basic idea of previous lectures.



Create a Simple Peak Caller

- Using RGT functions in Python.

- Same basic idea of previous lectures.

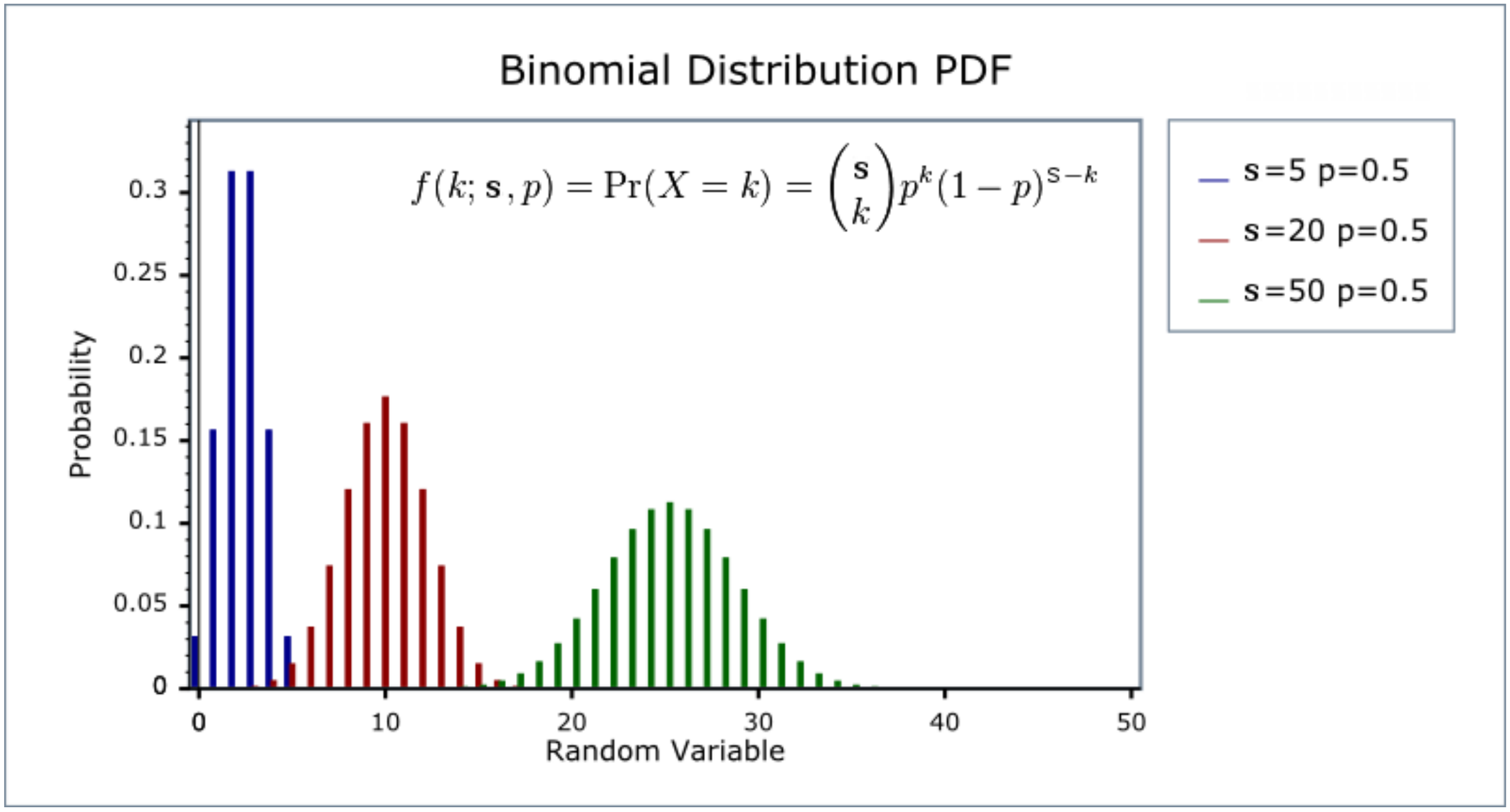


Model Distribution of Reads with Binomial

- 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.

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.



Our Peak Calling pipeline

1. Count number of reads for each bin.

2. Use a binomial distribution to model read coverage.

3. Iterate over genomic bins performing binomial test.

4. Store the bins that pass the test.

Let’s implement our peak caller