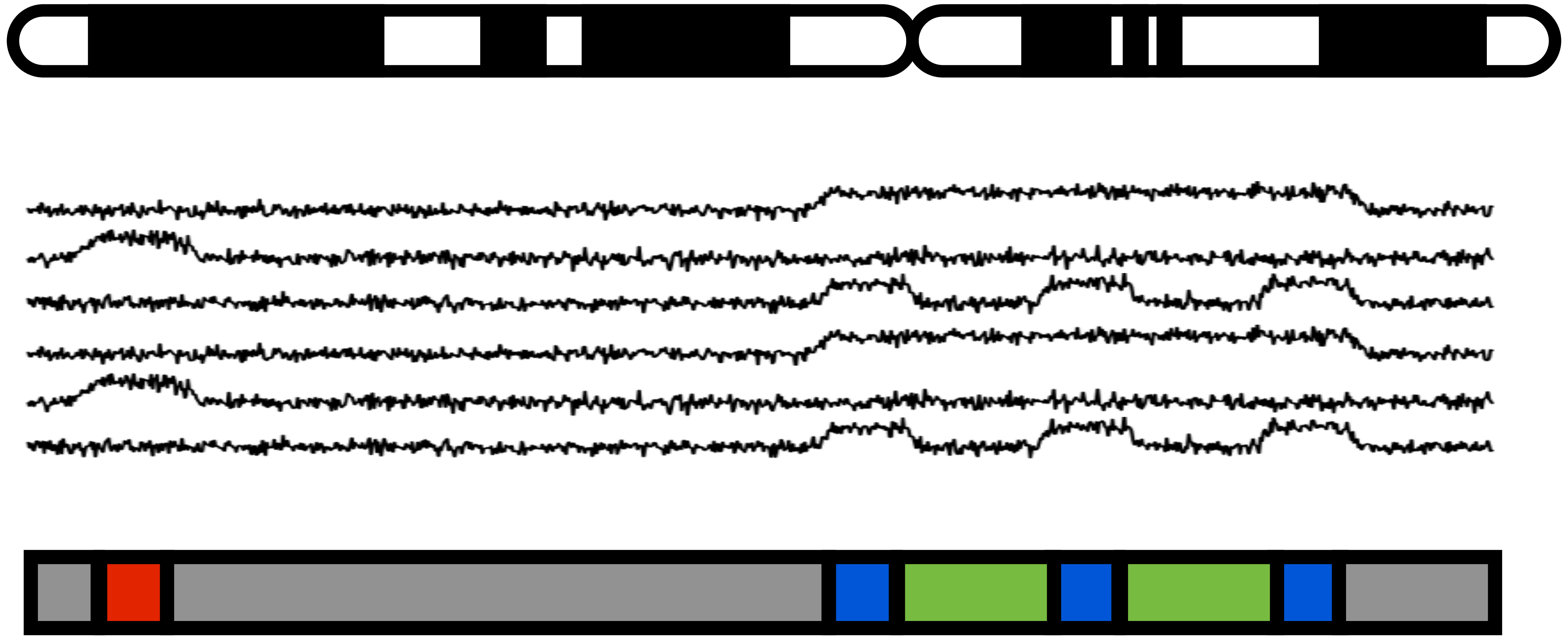
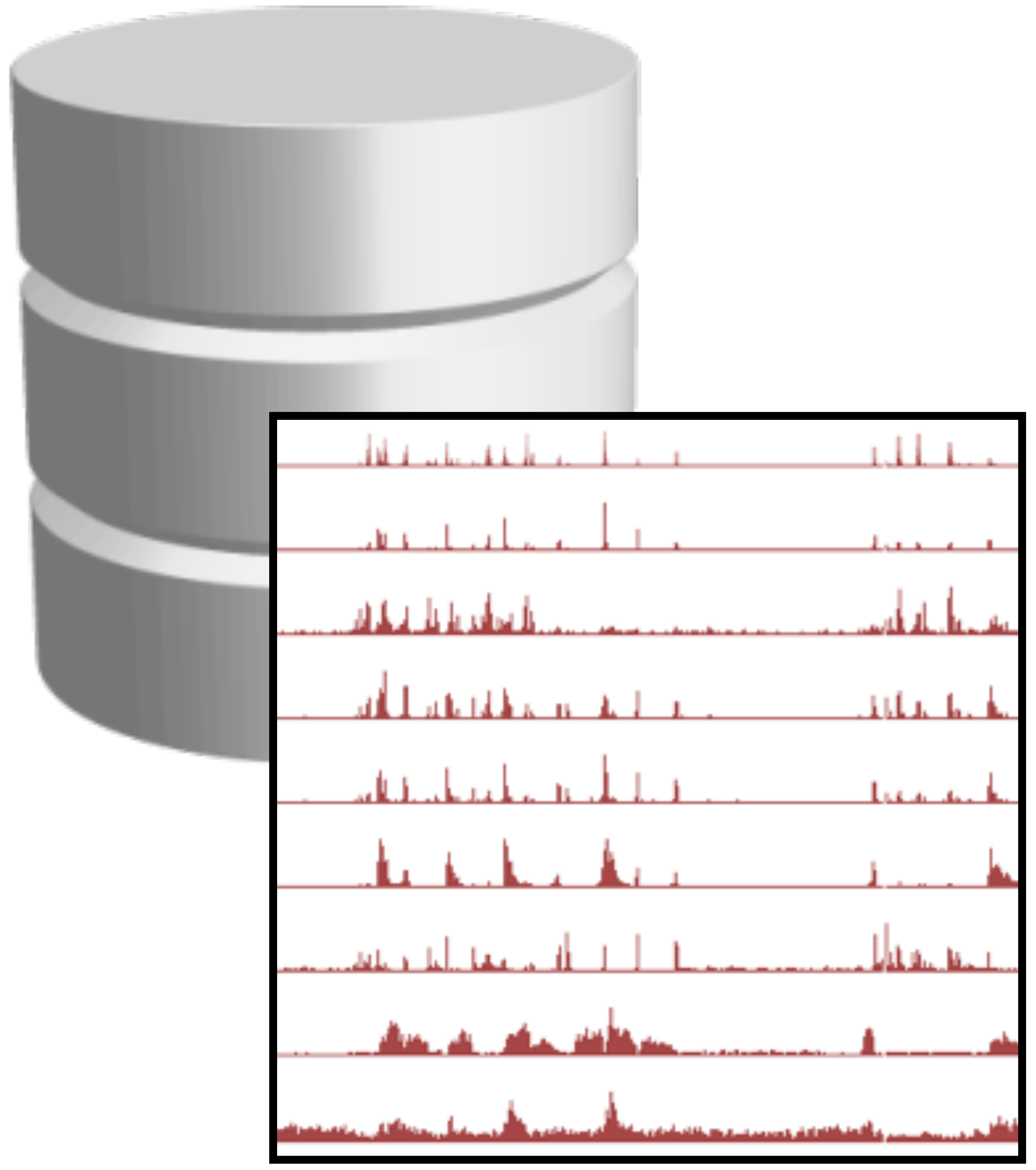
Storing and analyzing genomics signal data sets with Genomedata, Segway and Segtools



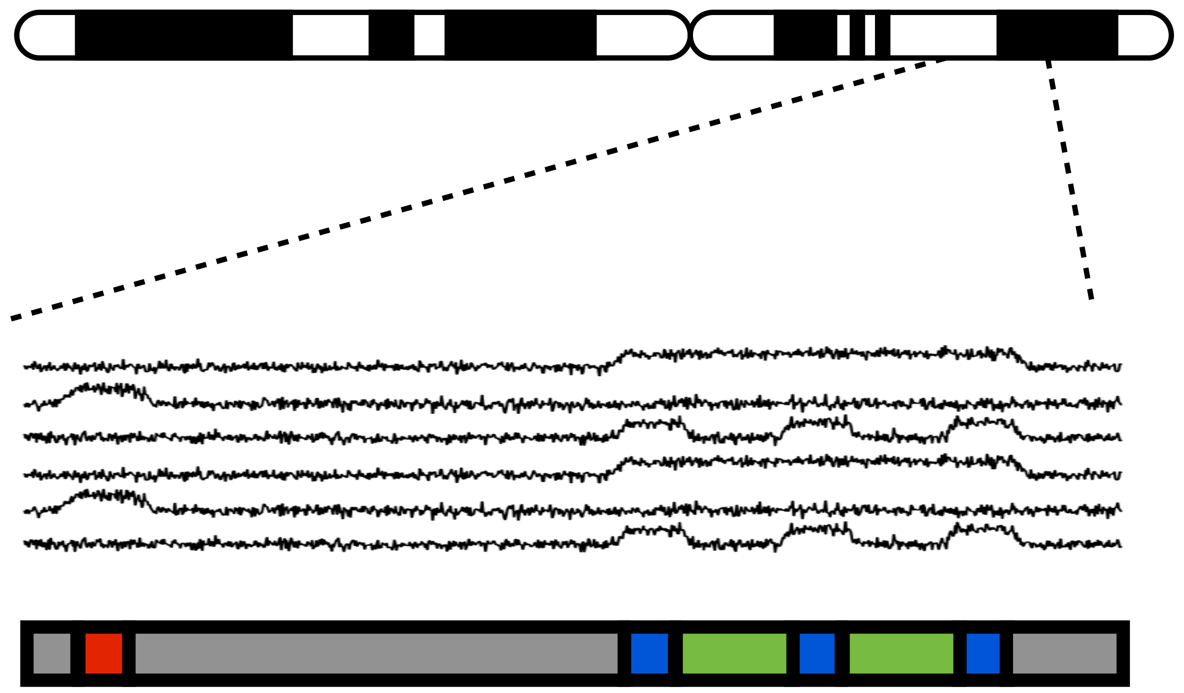
Max Libbrecht
Postdoc, Bill Noble's group
University of Washington, Seattle

Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics signal data sets

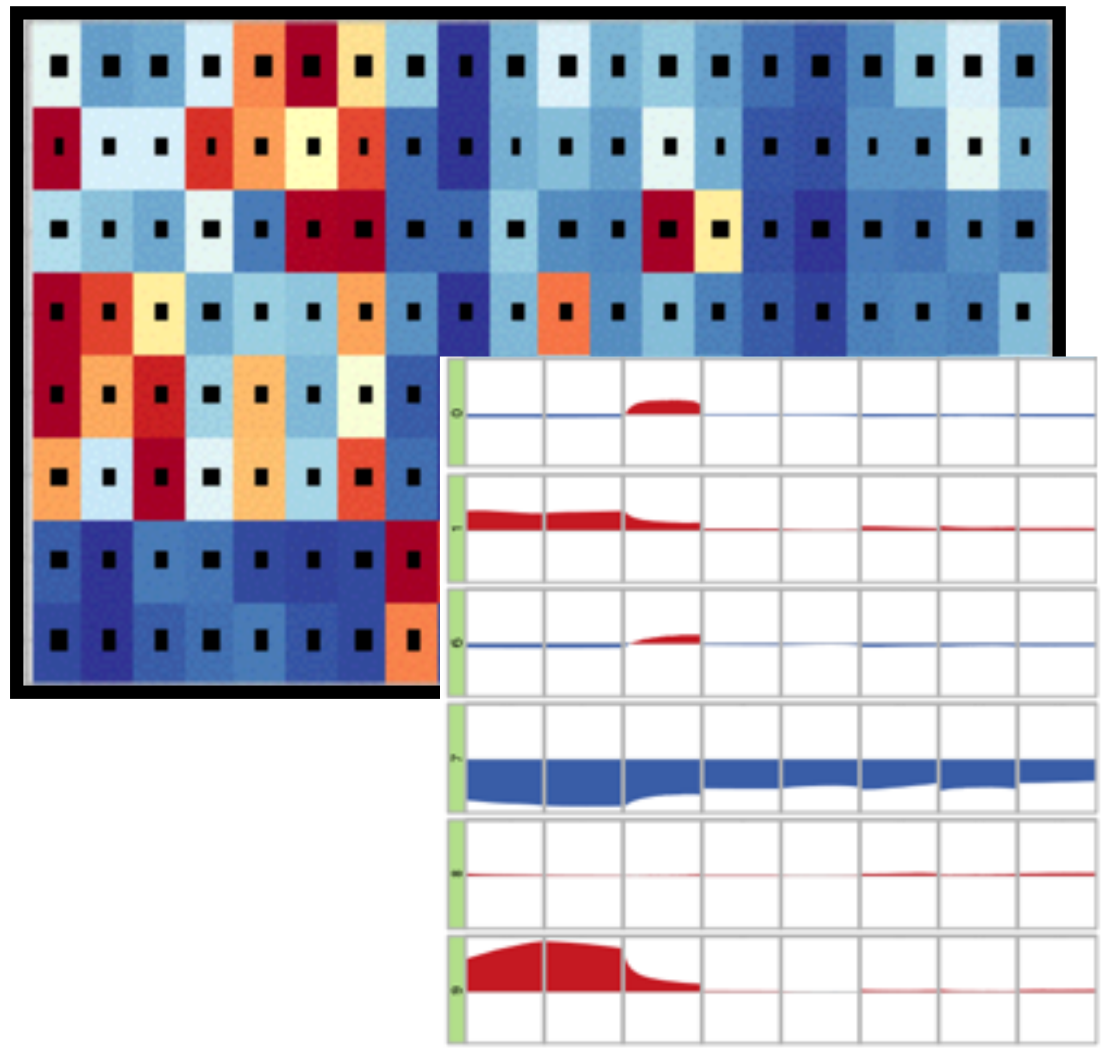
Genomedata



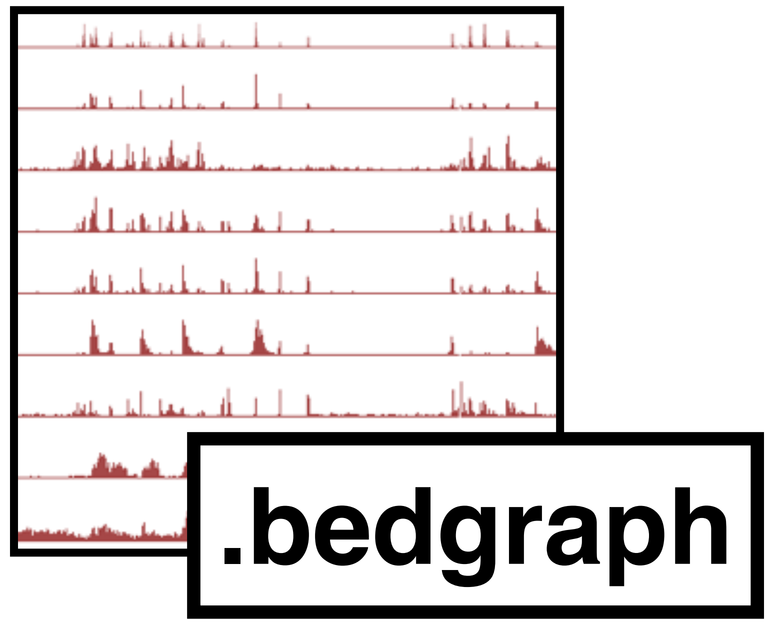
Segway



Segtools

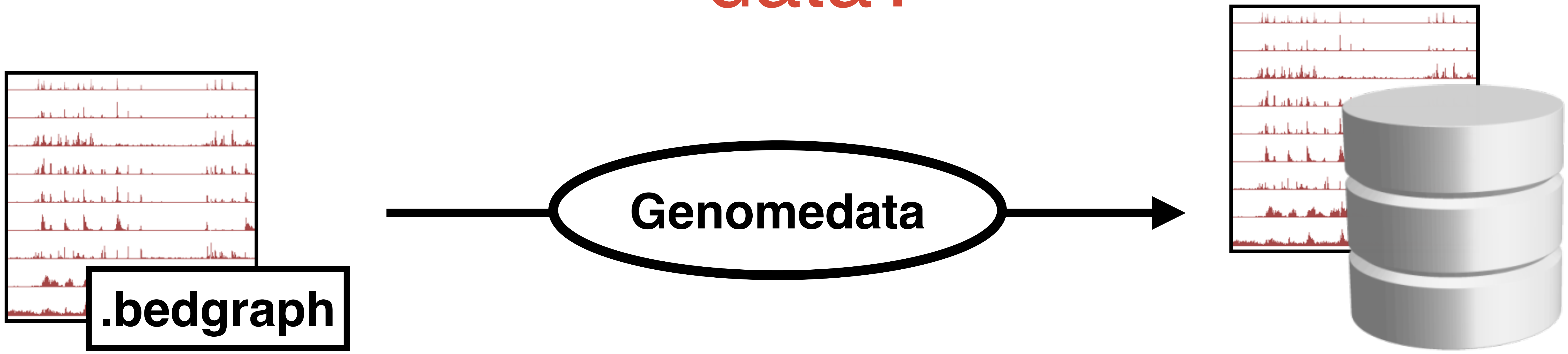


How can we understand a new collection of genomics data?



Signal data:
bedgraph format

How can we understand a new collection of genomics data?



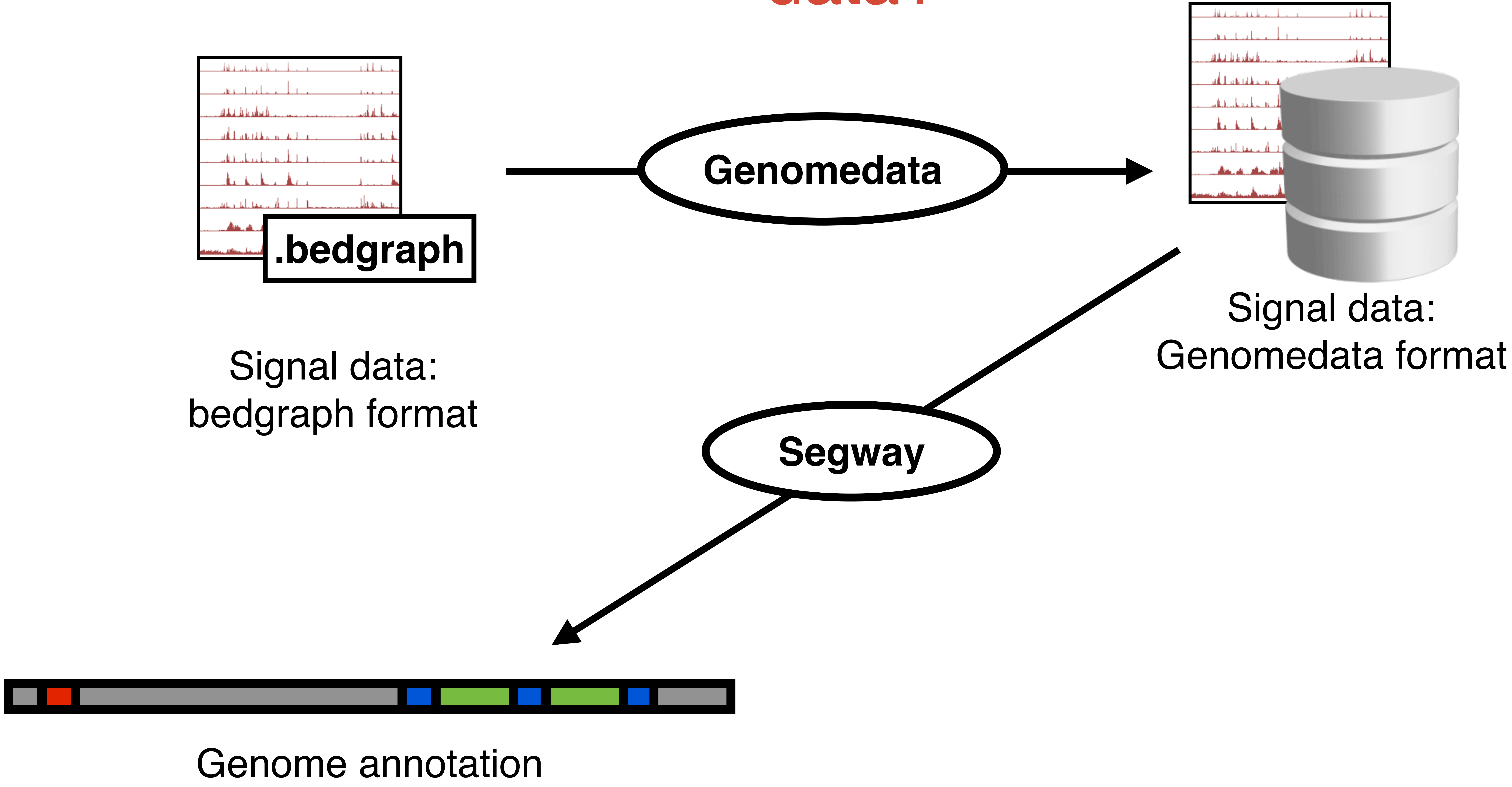
.bedgraph

Genomedata

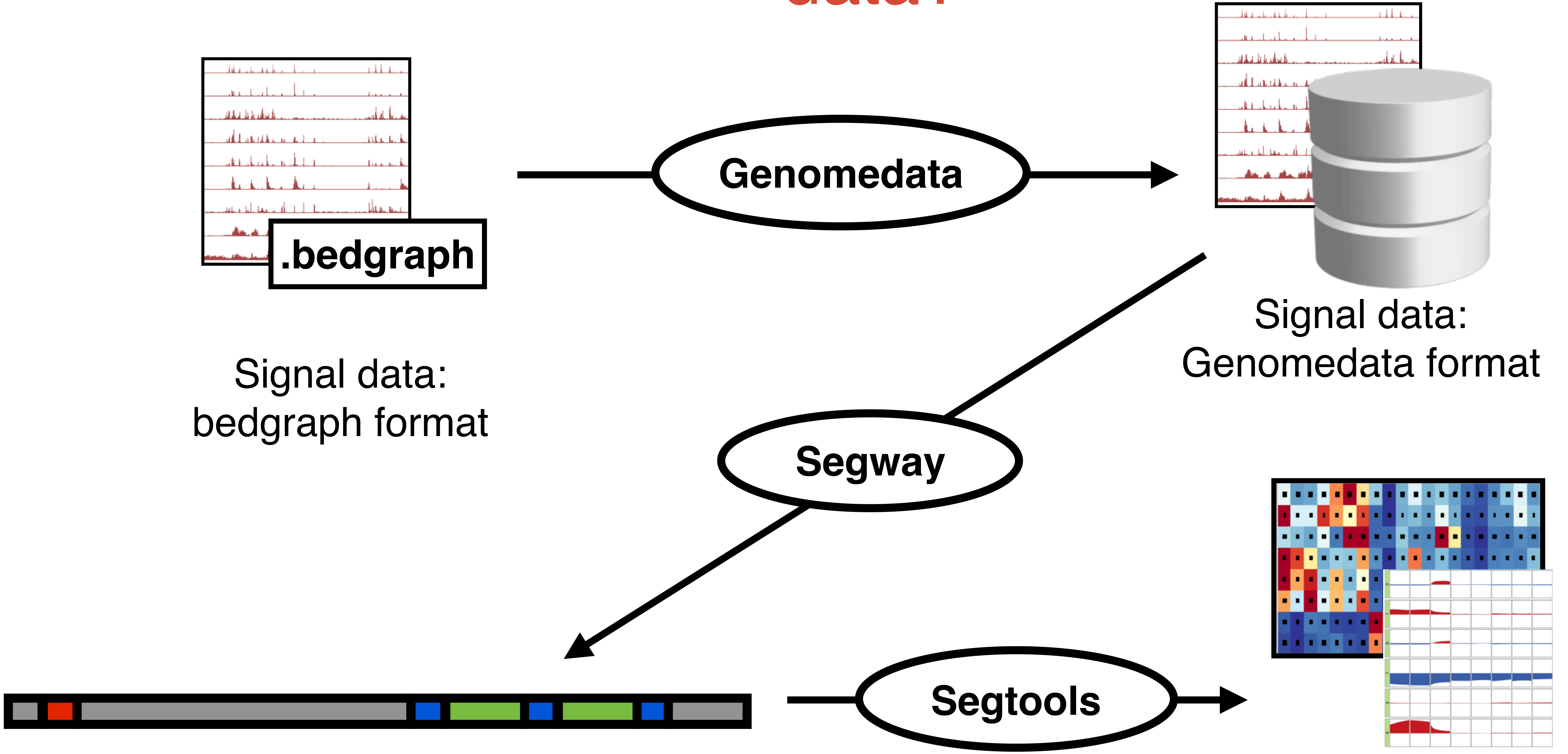
Signal data:
bedgraph format

Signal data:
Genomedata format

How can we understand a new collection of genomics data?



How can we understand a new collection of genomics data?



.bedgraph

Signal data:
bedgraph format

Genomedata

Signal data:
Genomedata format

Segway

Genome annotation

Segtools

Summary plots

Platform and installation

Genomedata, Segway and Segtools are supported on Linux

To install:

```
# Ubuntu/Debian:
sudo apt-get install libhdf5-serial-dev hdf5-tools
# CentOS/RHEL/Fedora:
sudo yum -y install hdf5 hdf5-devel
# OpenSUSE:
sudo zypper in hdf5 hdf5-devel libhdf5

wget http://melodi.ee.washington.edu/downloads/gmtk/gmtk-1.4.4.tar.gz
tar xf gmtk-1.4.4.tar.gz
cd gmtk-1.4.4
./configure
make
make install

pip install numpy
pip install numexpr
pip install cython
pip install genomedata
pip install segway
pip install segtools
```

Documentation and more information

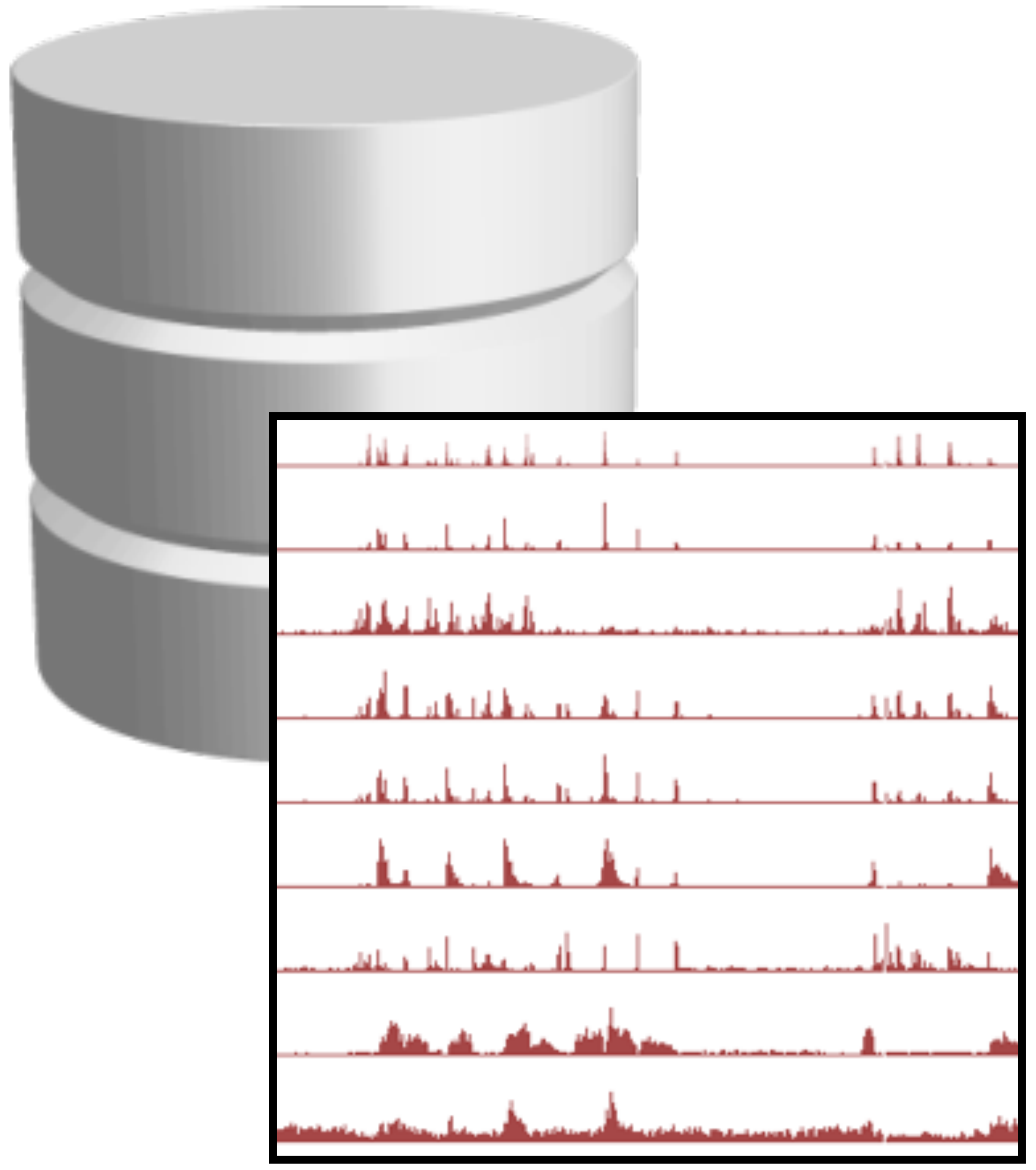
Genomedata: https://www.pmgenomics.ca/hoffmanlab/proj/genomedata/

Segway: https://www.pmgenomics.ca/hoffmanlab/proj/segway/

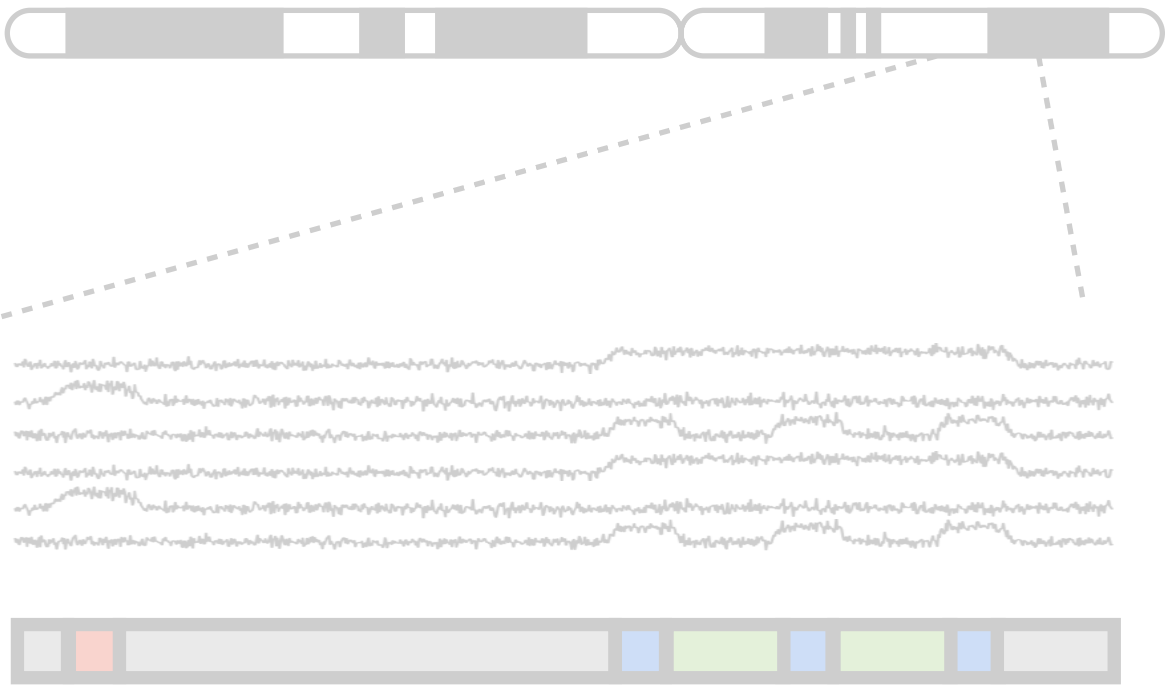
Segtools: https://www.pmgenomics.ca/hoffmanlab/proj/segtools

Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics signal data sets

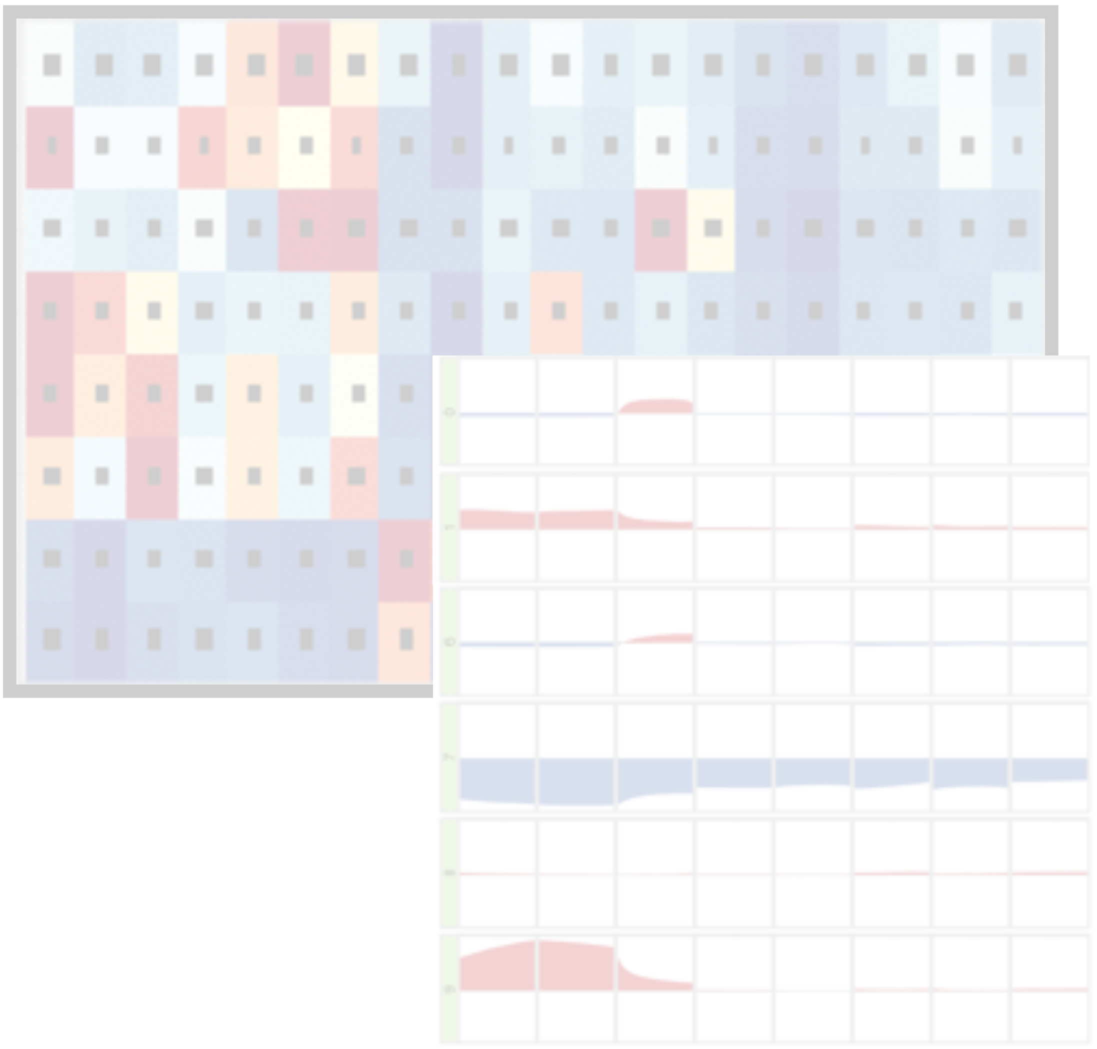
Genomedata



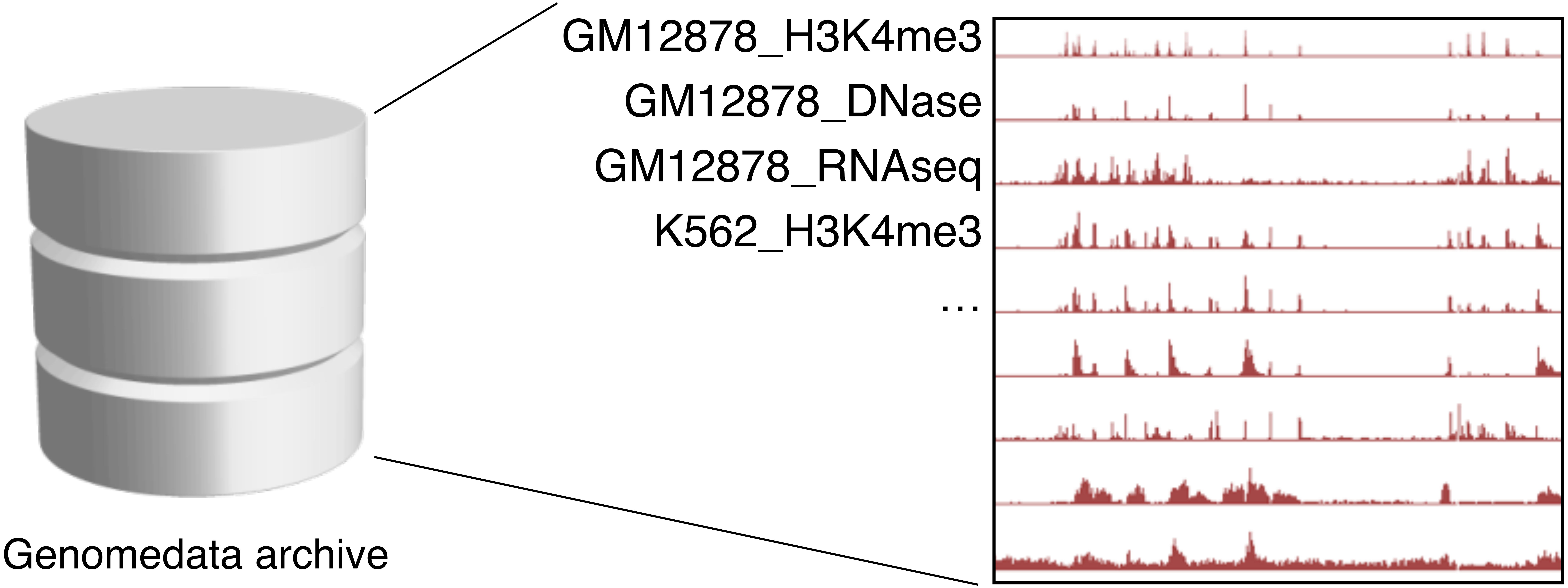
Segway



Segtools



Genomedata stores a collection of genomics tracks



GM12878_H3K4me3

GM12878_DNase

GM12878_RNAseq

K562_H3K4me3

…

Genomedata archive
(HDF5)

Key feature: **random access**

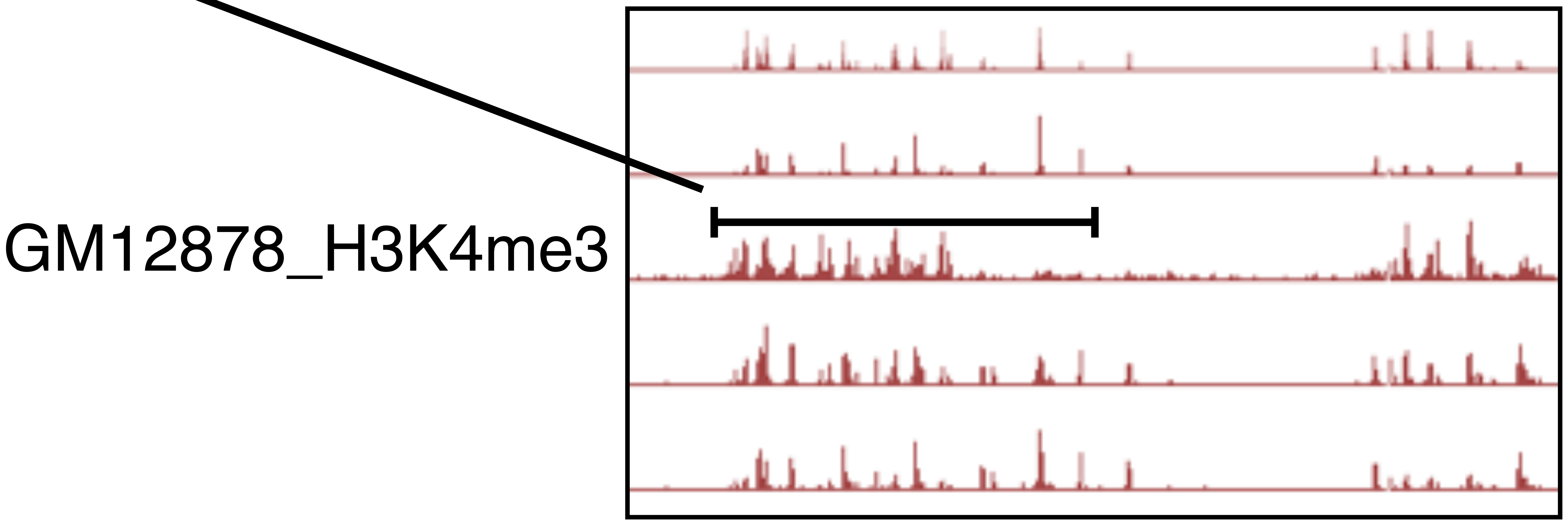
Loading data into genomedata

http://hgdownload.cse.ucsc.edu/downloads.html

```
$ genomedata-load-assembly data.genomedata hg19.fa

# For each track:
$ genomedata-open-data data.genomedata GM12878_H3K4me3

$ genomedata-load-data data.genomedata GM12878_H3K4me3
    < GM12878_H3K4me3.bedgraph

$ genomedata-close-data data.genomedata
```

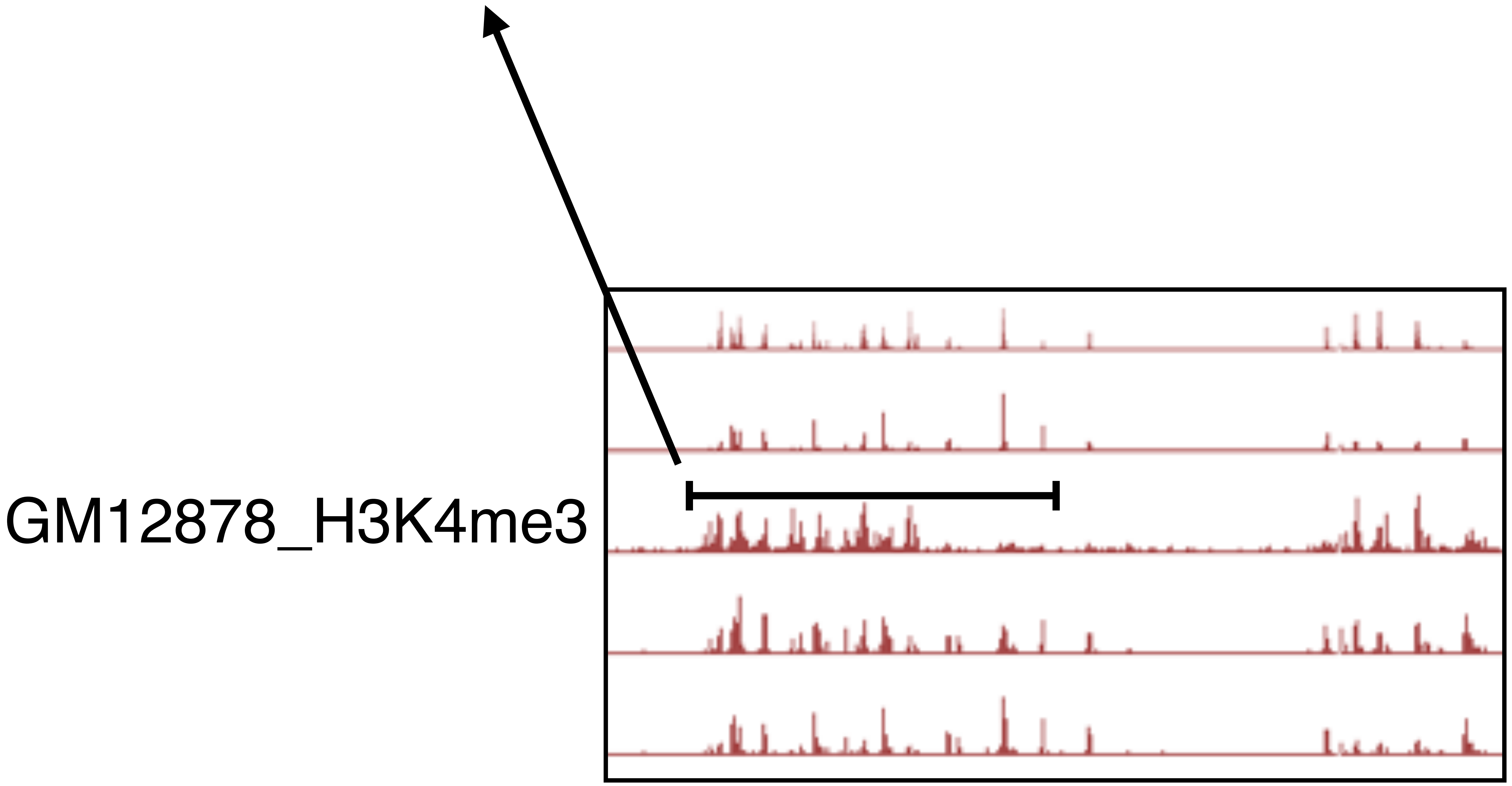
Accessing data (command line)

```
$ genomedata-query data.genomedata GM12878_H3K4me3 chr1 1000000 1000100
fixedStep  chrom=chr1 start=1000000
16.8
17.9
14.0
1.2
...
```



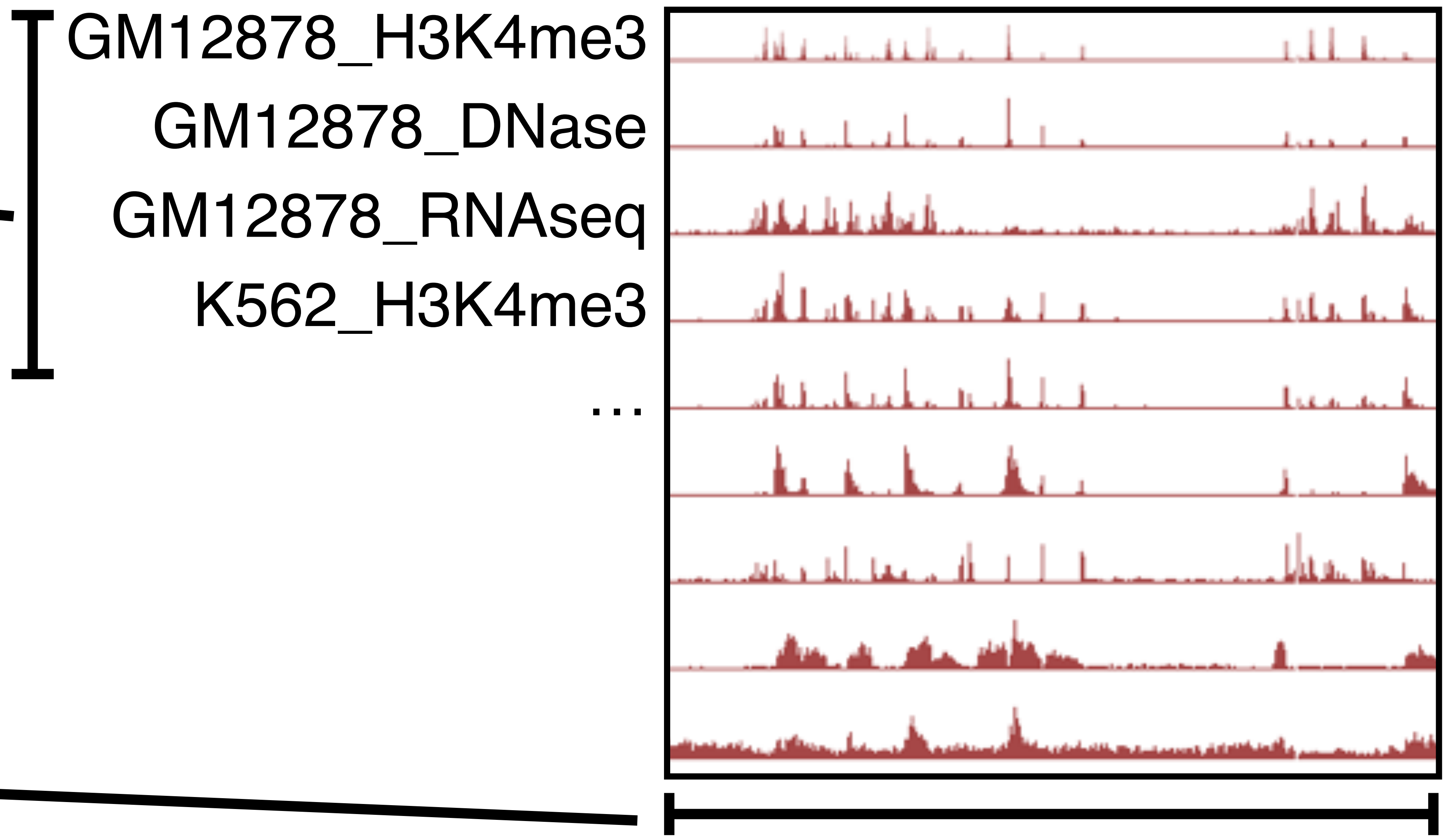
Accessing data (Python)

```
>>> import genomedata
>>> g = genomedata.Genome(“data.genomedata”)
>>> g[“chr1”][1000000:1000100, “GM12878_H3K4me3”]
array([ 16.8, 17.9, 14.0, 1.2, ...], dtype=float32)
```



Information about a genomedata archive (command line)

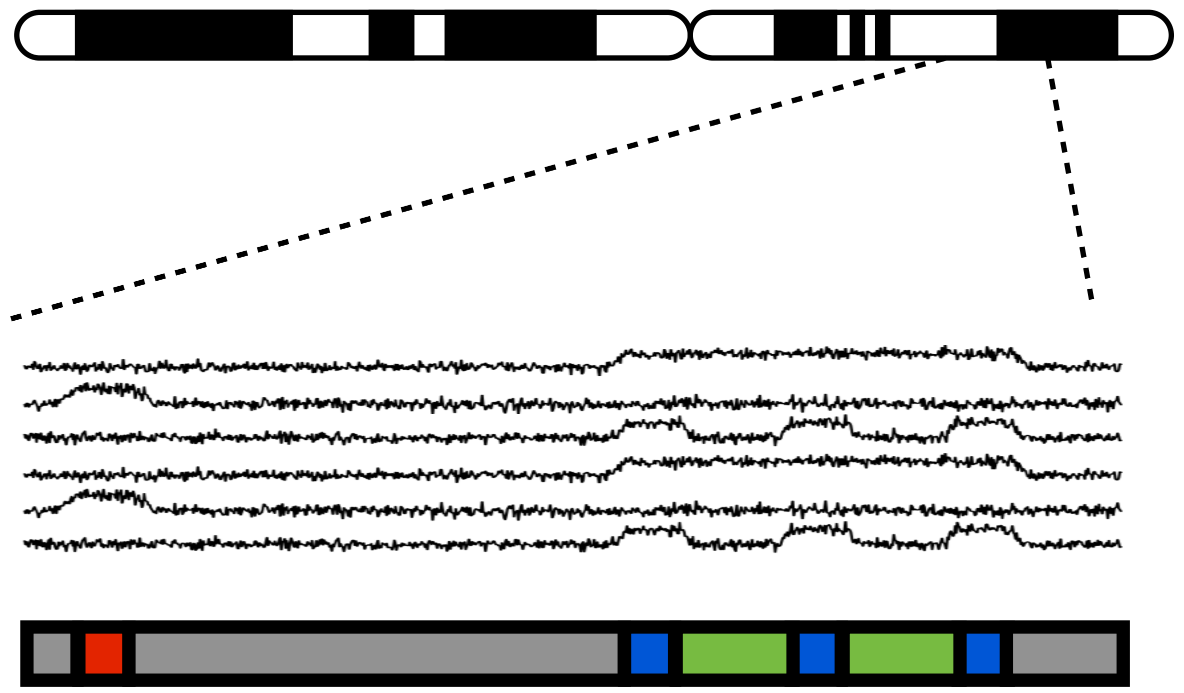
```
$ genomedata-info tracknames data.genomedata
GM12878_H3K4me3
GM12878_DNase
GM12878_RNAseq
K562_H3K4me3
...

$ genomedata-info contigs data.genomedata
chr1  0  249250621
chr2  0  243199373
...
```



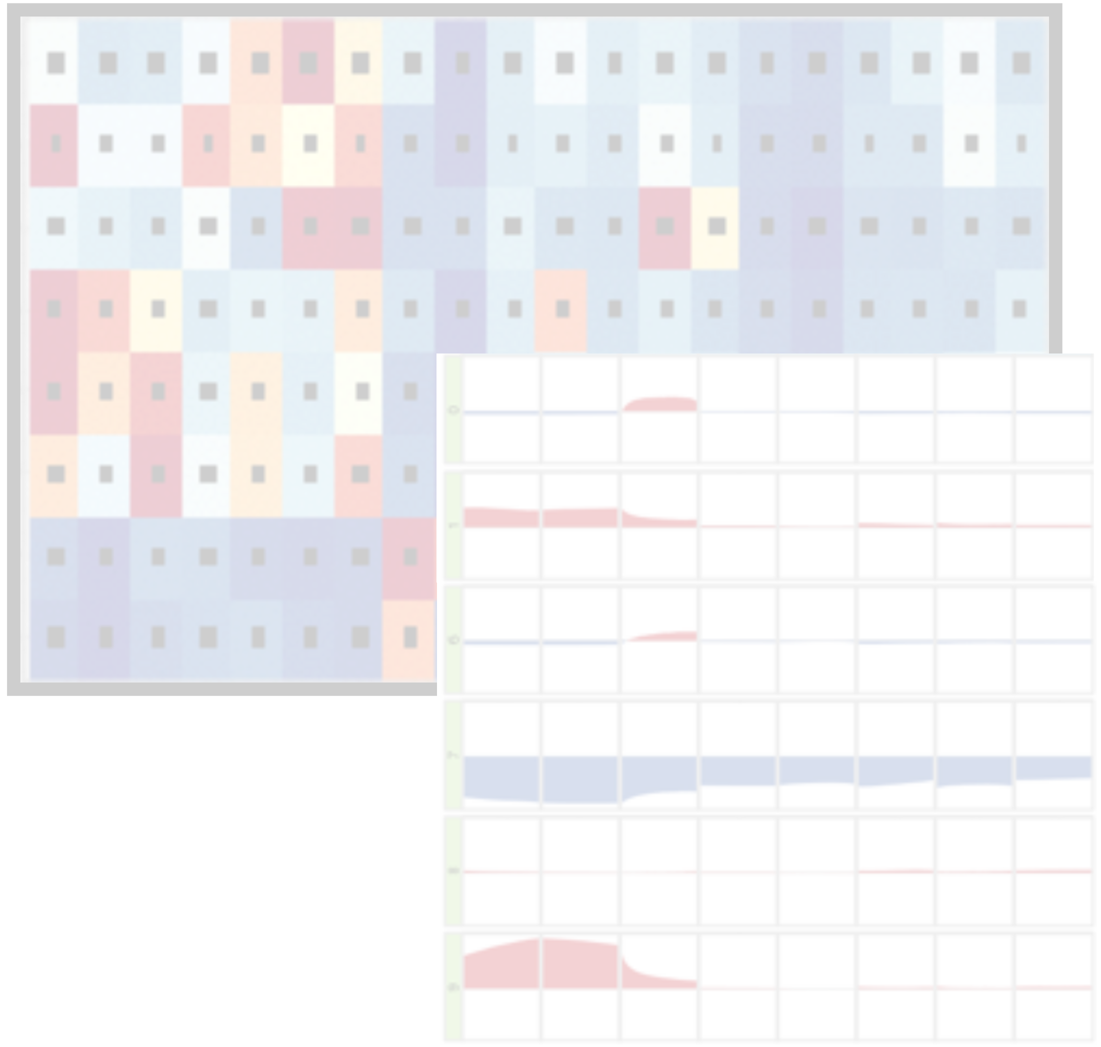
Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics signal data sets

Genomedata

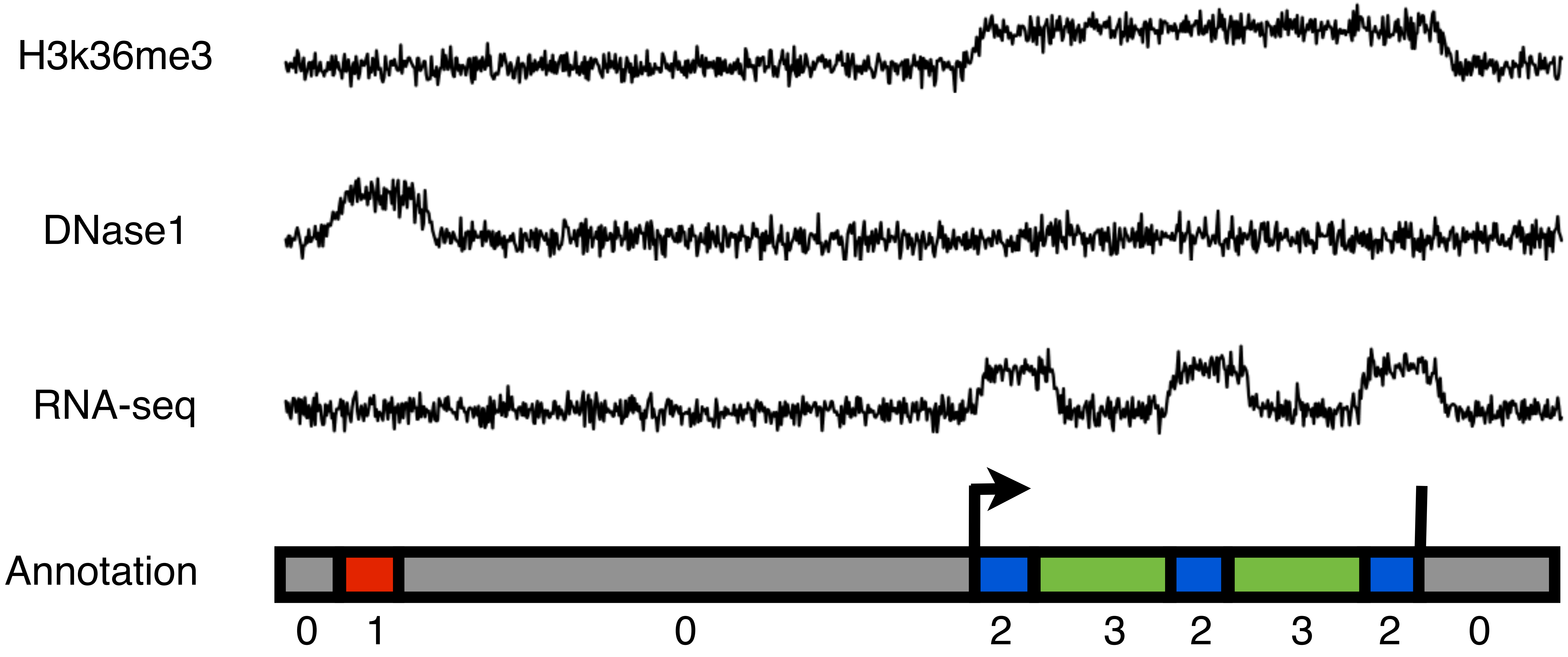
Segway



Segtools



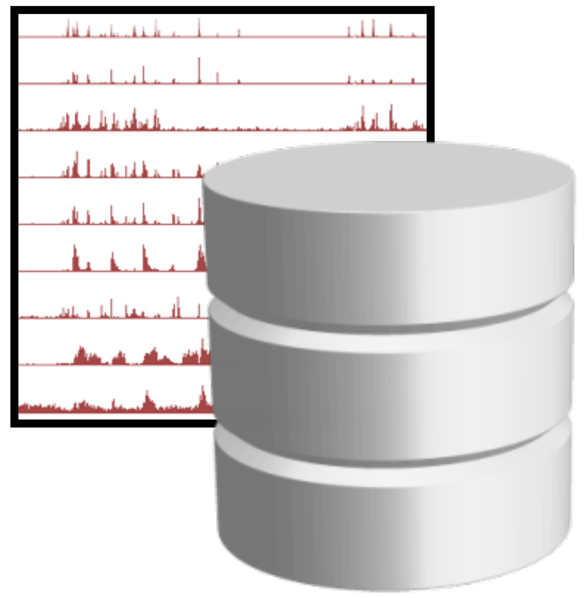
Semi-automated genome annotation algorithms partition and label the genome on the basis of functional genomics tracks



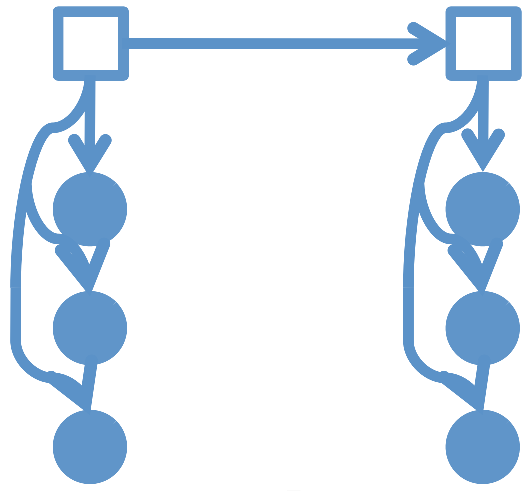
Human interpretation: 1 = “Enhancer”, 2 = “Exon”, …

HMMSeg: Day et al. *Bioinformatics*, 2007
ChromHMM: Ernst, J. and Kellis, M. *Nature Biotechnology*, 2010
Segway: Hoffman, M et al. *Nature Methods*, 2012

Running Segway



```
segway train data.genomedata traindir
```



```
segway identify data.genomedata traindir identifydir
```

```
output: identifydir/segway.bed.gz
chr1    0     150    5
chr1    150   700    2
...
```

Using a compute cluster

Segway supports distributed computing using **Grid Engine** and **Platform LSF**.

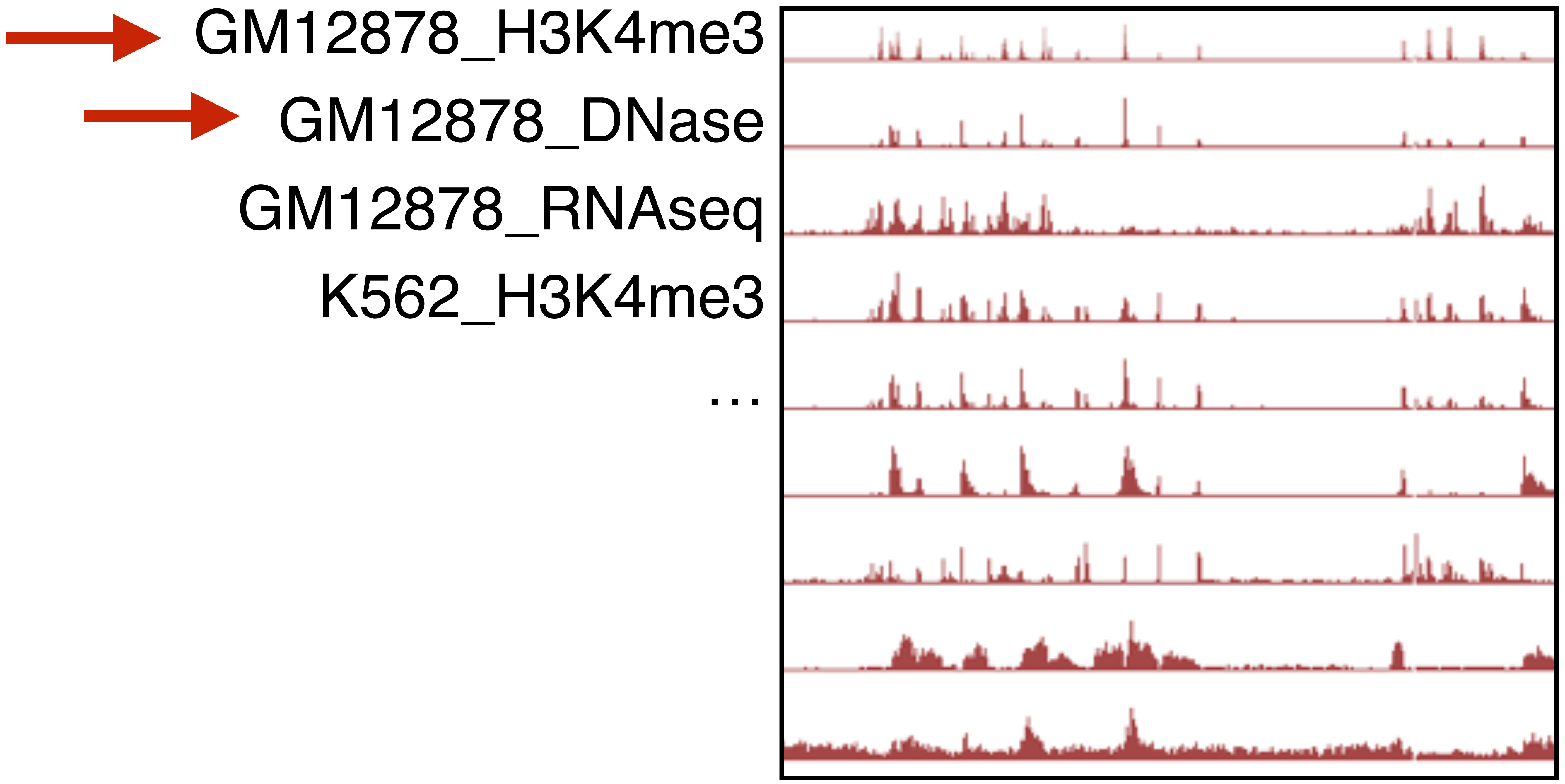
To run Segway without a cluster, set

```
$ export SEGWAY_CLUSTER=local
```

Input tracks

Input tracks

```
--track=GM12878_H3K27ac --track=GM12878_H3K4me3
OR
--tracks-from=tracks.txt

tracks.txt:
GM12878_H3K27ac
GM12878_DNase
```



GM12878_H3K4me3

GM12878_DNase

GM12878_RNAseq

K562_H3K4me3

…

Input coordinates

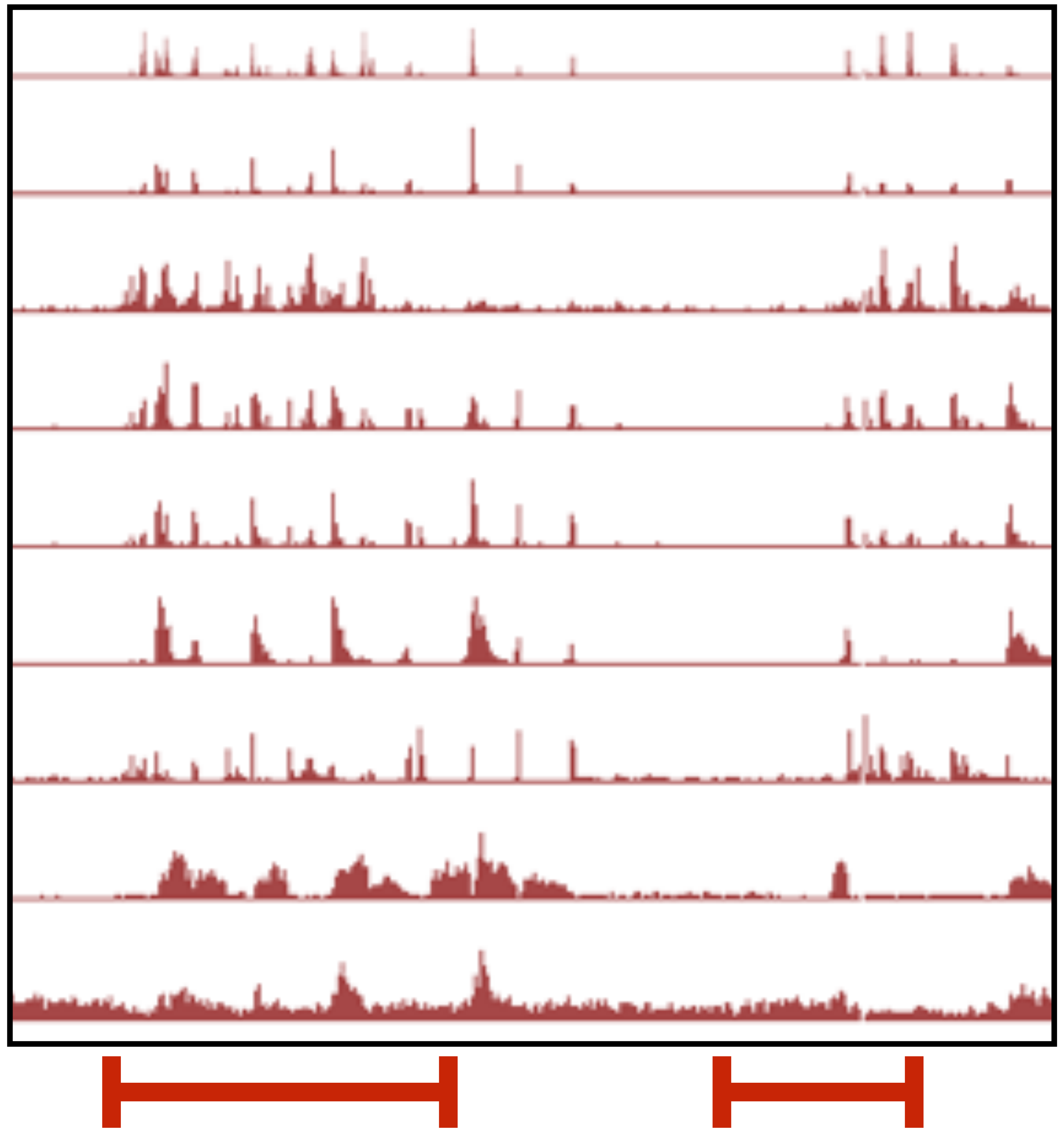
Genome coordinates

```
--include-coords=coords.bed

coords.bed:
chr1     151158060    151658060
chr10     55483812     55983812

--exclude-coords=blacklist.bed
```

Training minibatch size

```
--minibatch-fraction=0.01
```

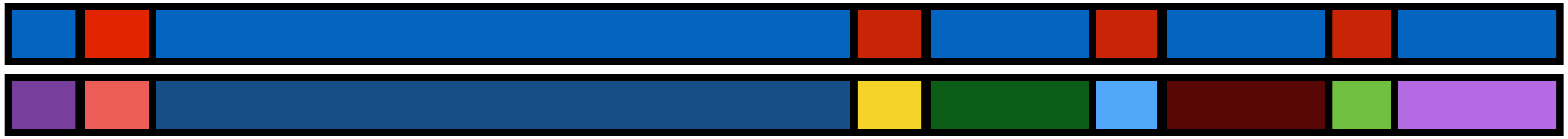
https://sites.google.com/site/anshulkundaje/projects/blacklists



Training parameters

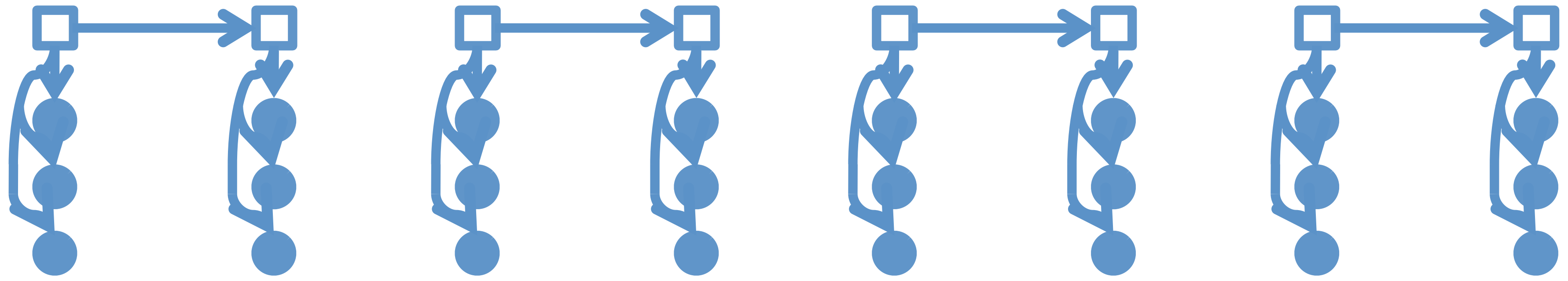
Number of annotation labels

```
--num-labels=25 (Recommended: 4 - 50)
```



Number of EM intializations

```
--num-instances=10 (Recommended: 10)
```



Maximum number of EM training iterations

```
--max-train-rounds=100 (Recommended: 100)
```

Controlling segment lengths

Downsampling resolution

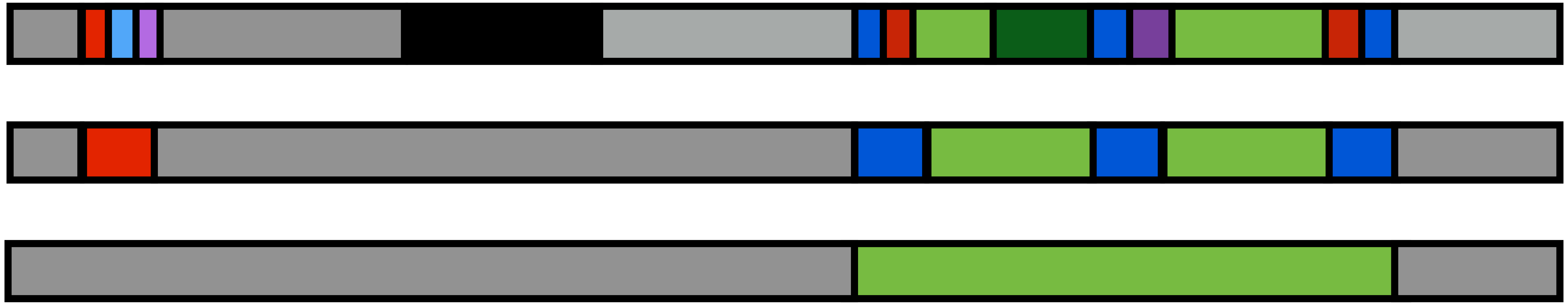
```
--resolution=10 (Recommended: 1 - 10,000)
```

Long segments prior

```
--prior-strength=1.0 (Recommended: 0 - 10+)
```

Weight on transition part of the model

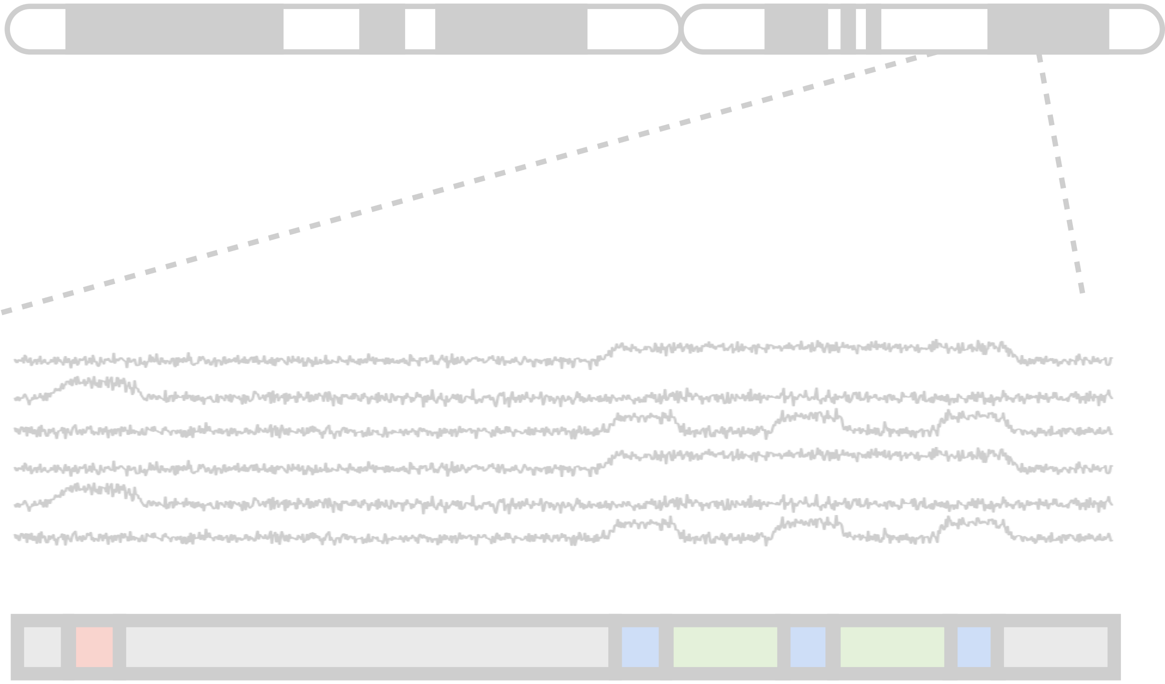
```
--segtransition-weight-scale=10 (Recommended: ≈ number of tracks)
```



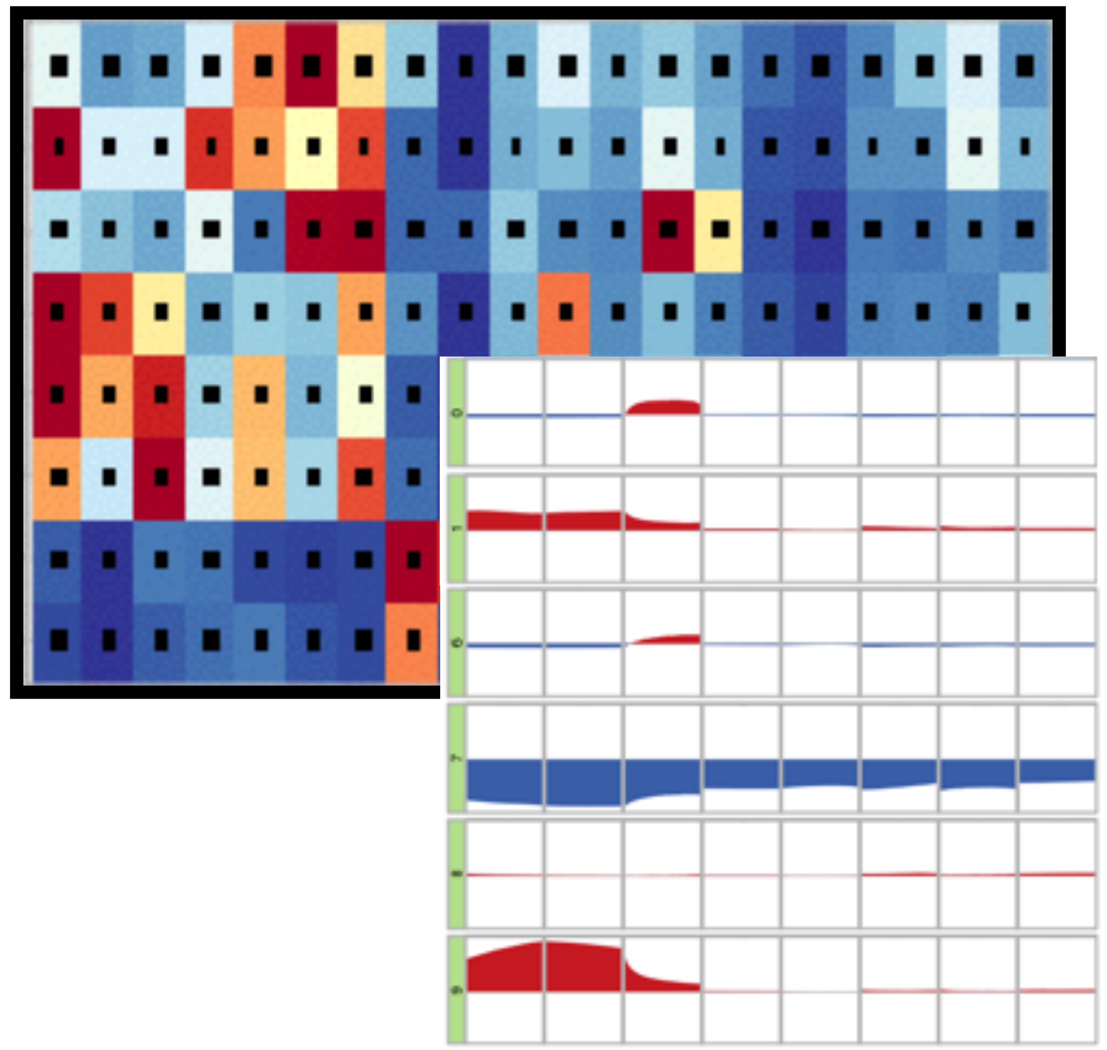
Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics signal data sets

Genomedata

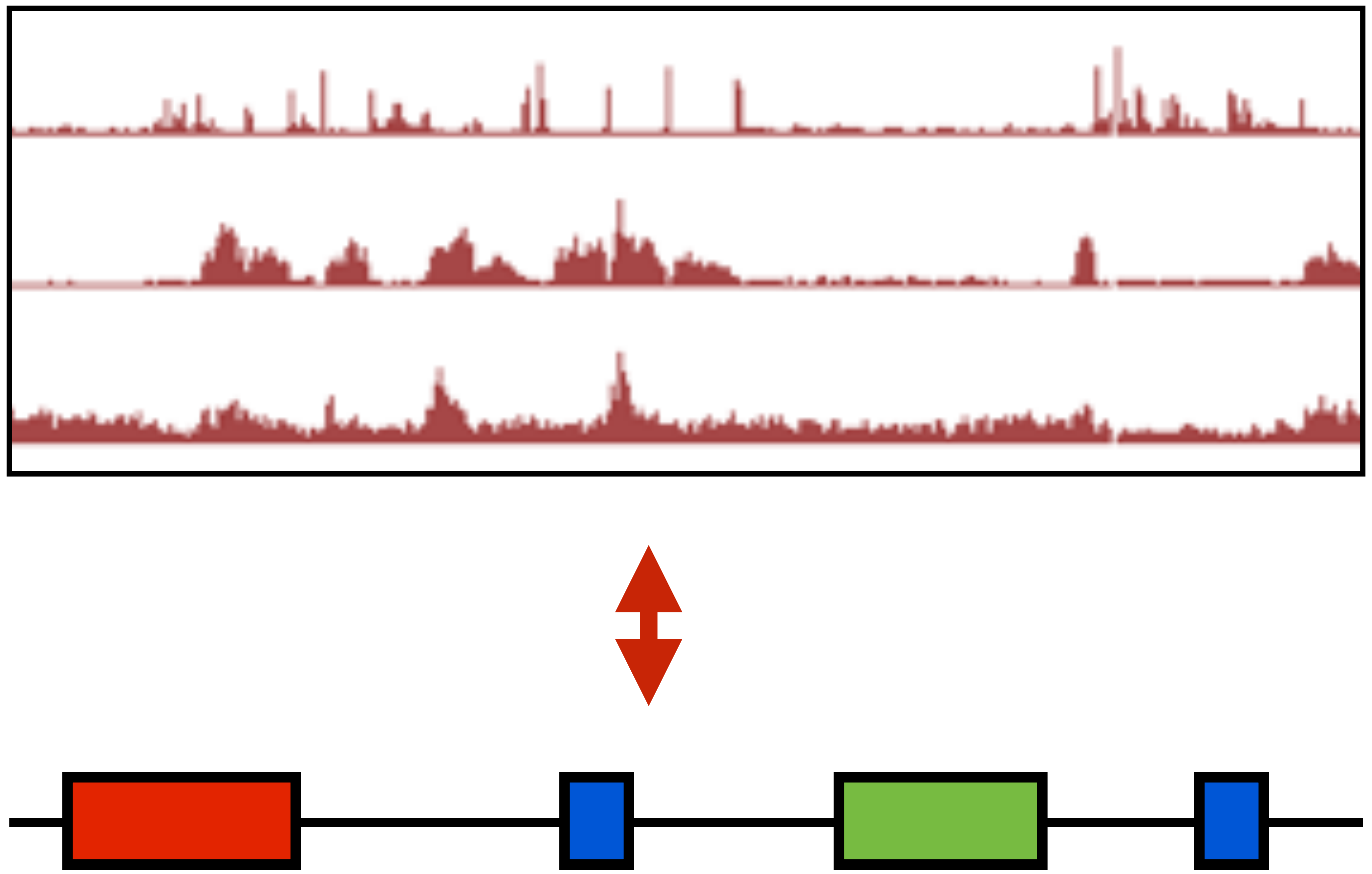
Segway

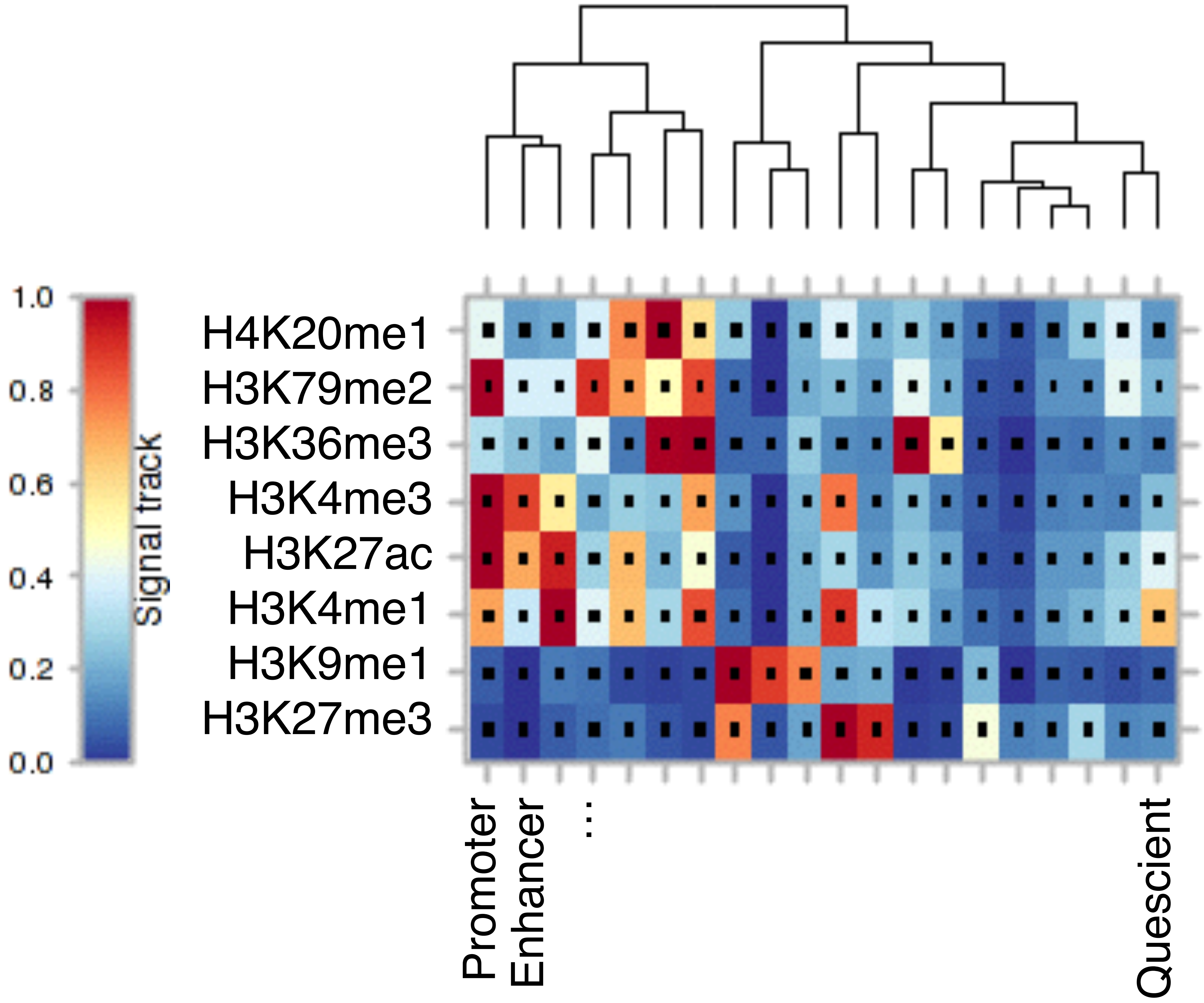


Segtools



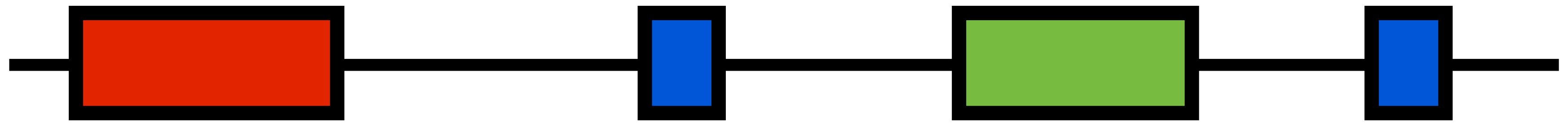
segtools-signal-distribution measures relationships between annotation labels and signal tracks

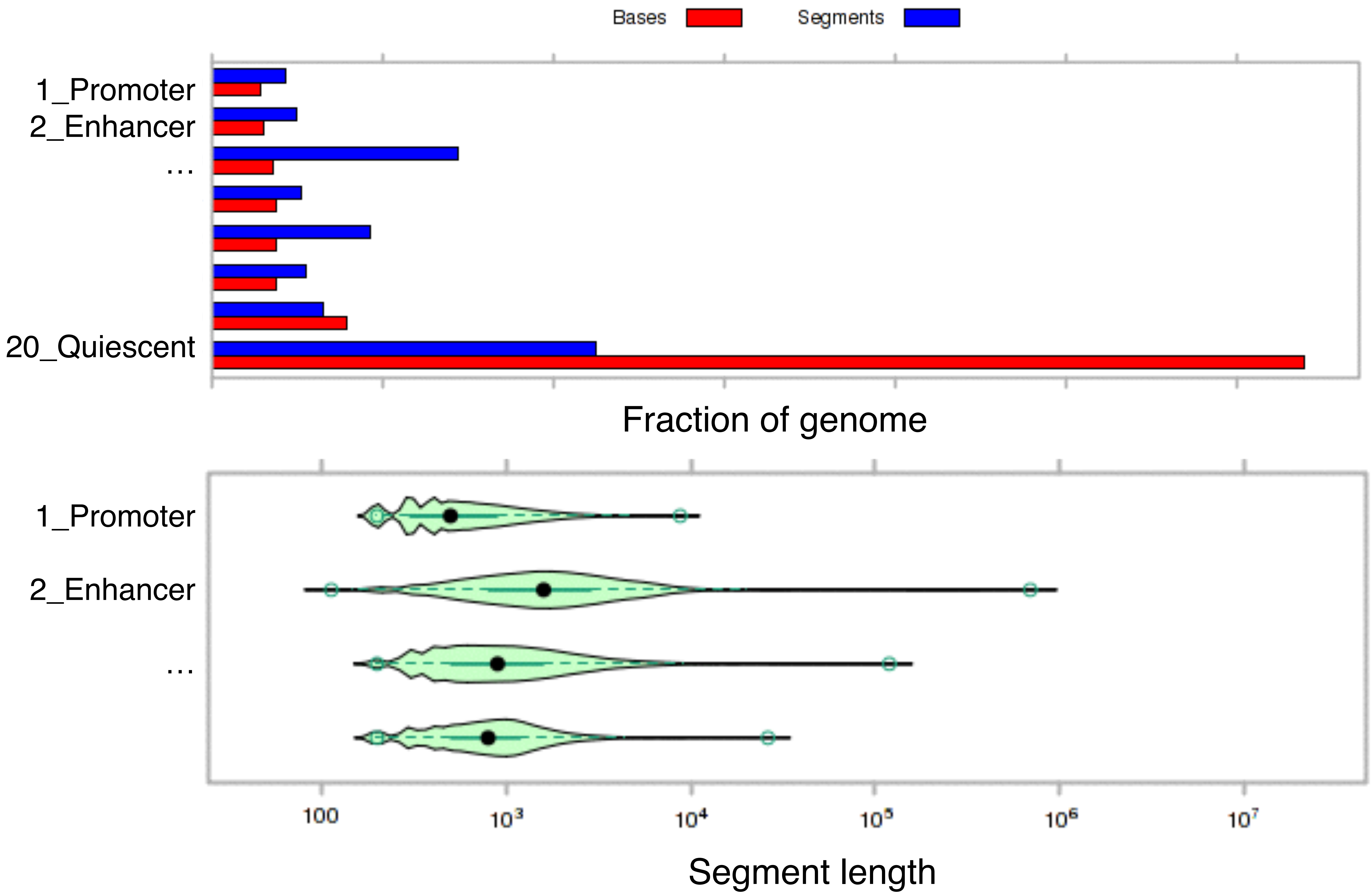




segtools-length-distribution measures segment lengths genome coverage

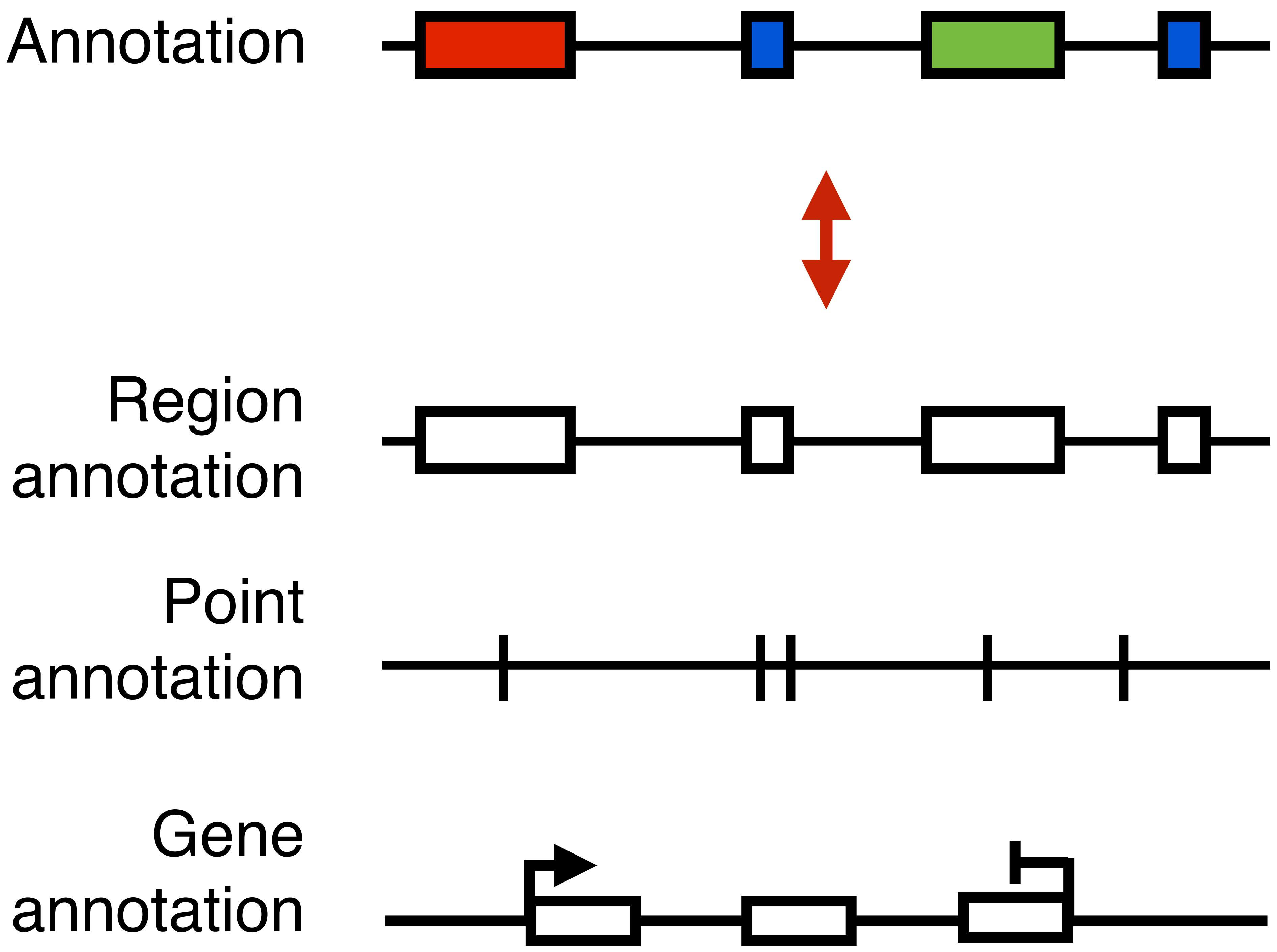
```
segtools-length-distribution segway.bed.gz
```

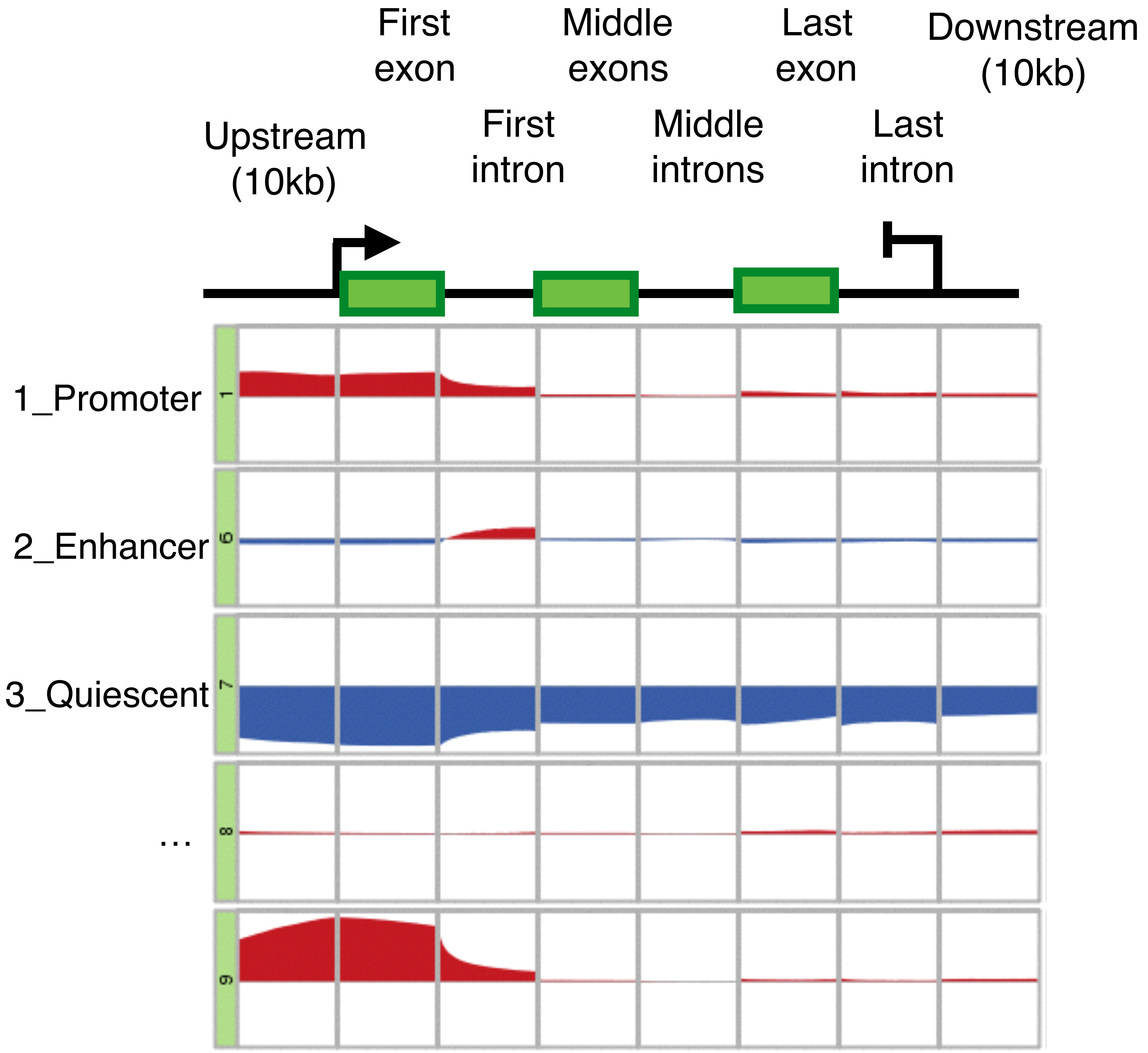




segtools-aggregation measures associations with other genome annotations

```
segtools-aggregation --normalize --mode=gene segway.bed.gz gencode.gff
```





Thank you