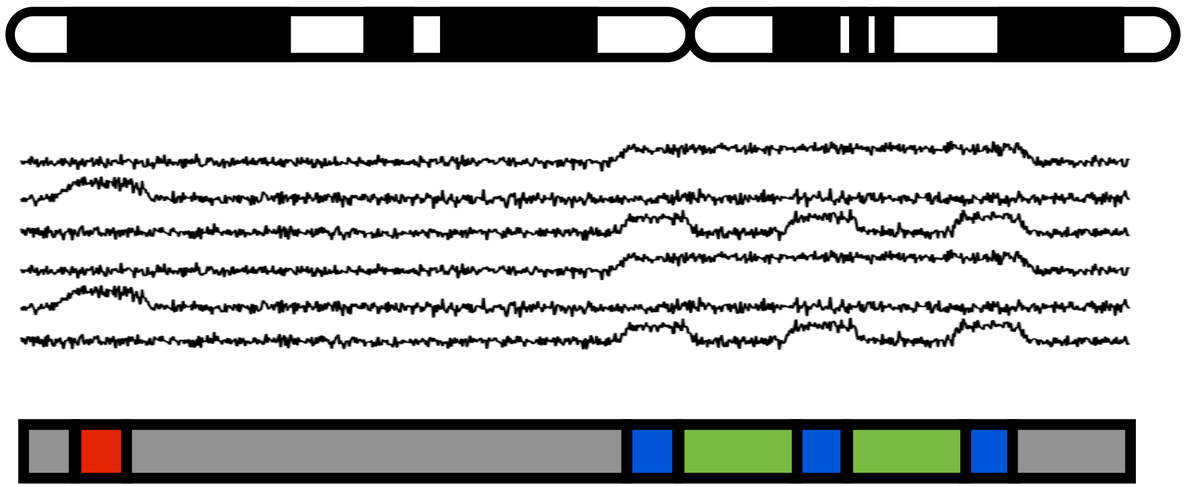
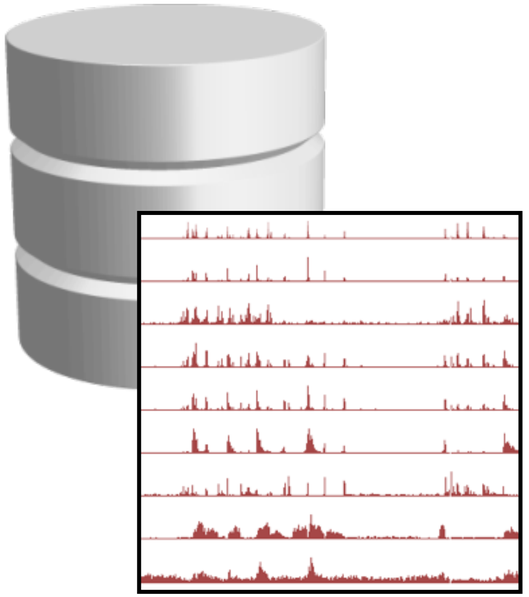
Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics data sets



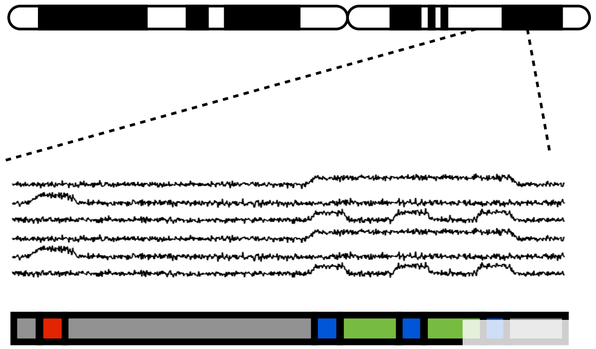
Max Libbrecht

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

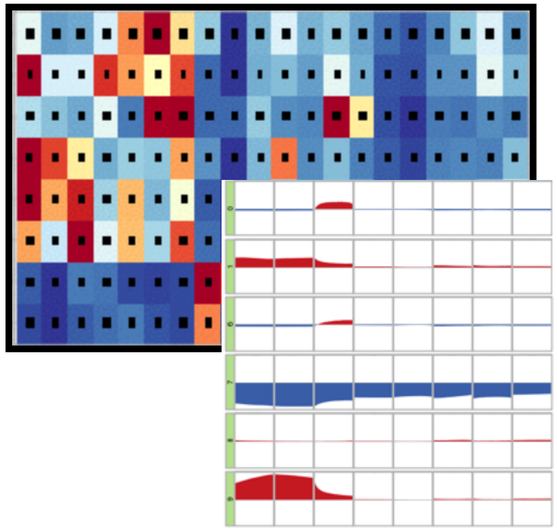
Genomedata



Segway

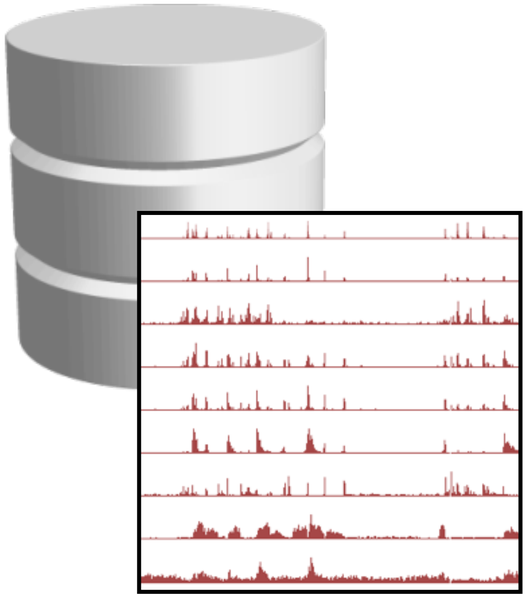


Segtools

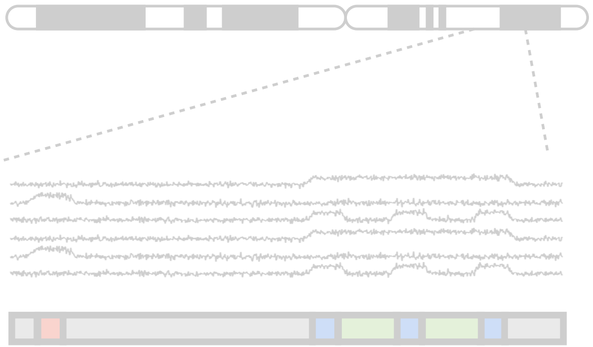


Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics data sets

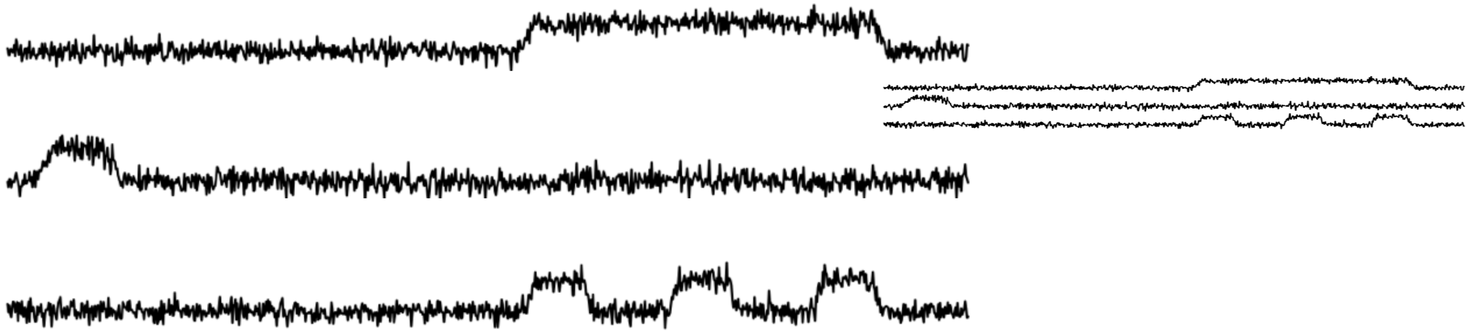
Genomedata



Segway



Segtools



Installing Genomedata

```
# HDF5
# 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

# Pytables
pip install numpy
pip install numexpr
pip install cython

# Genomedata
pip install genomedata
```

Loading data into genomedata

```
genomedata-load-assembly --sizes my_genomedata hg19.sizes
genomedata-open-data my_genomedata my_trackname
zcat input.bedgraph.gz | genomedata-load-data my_genomedata my_trackname
genomedata-close-data my_genomedata
```

```
hg19.sizes:
chr1  249250621
chr2  243199373
chr3  198022430
chr4  191154276
```

Accessing data: command line

```
$ genomedata-query my_genomedata my_trackname chr1 1000000 1000100
fixedStep  chrom=chr1 start=1000000
0.0
0.0
0.0
0.0
0.0
0.0
0.0
...
```

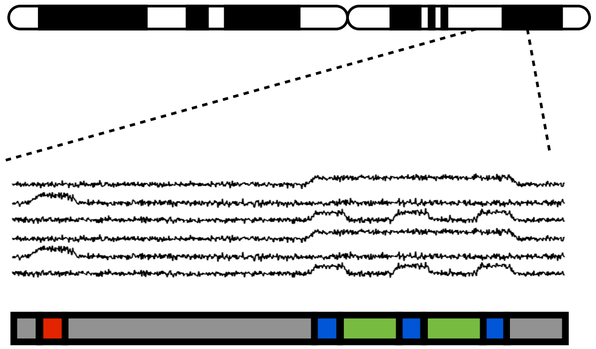
Accessing data: Python

```
>>> import genomedata
>>> g = genomedata.Genome(“my_genomedata”)
>>> g[“chr1”][1000000:1000100, “my_trackname”]
array([ 17.89999962,  17.89999962,  17.89999962,  17.89999962,
        17.89999962,  17.89999962,  17.89999962,  17.89999962,
        17.89999962,  17.89999962], dtype=float32)
```

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

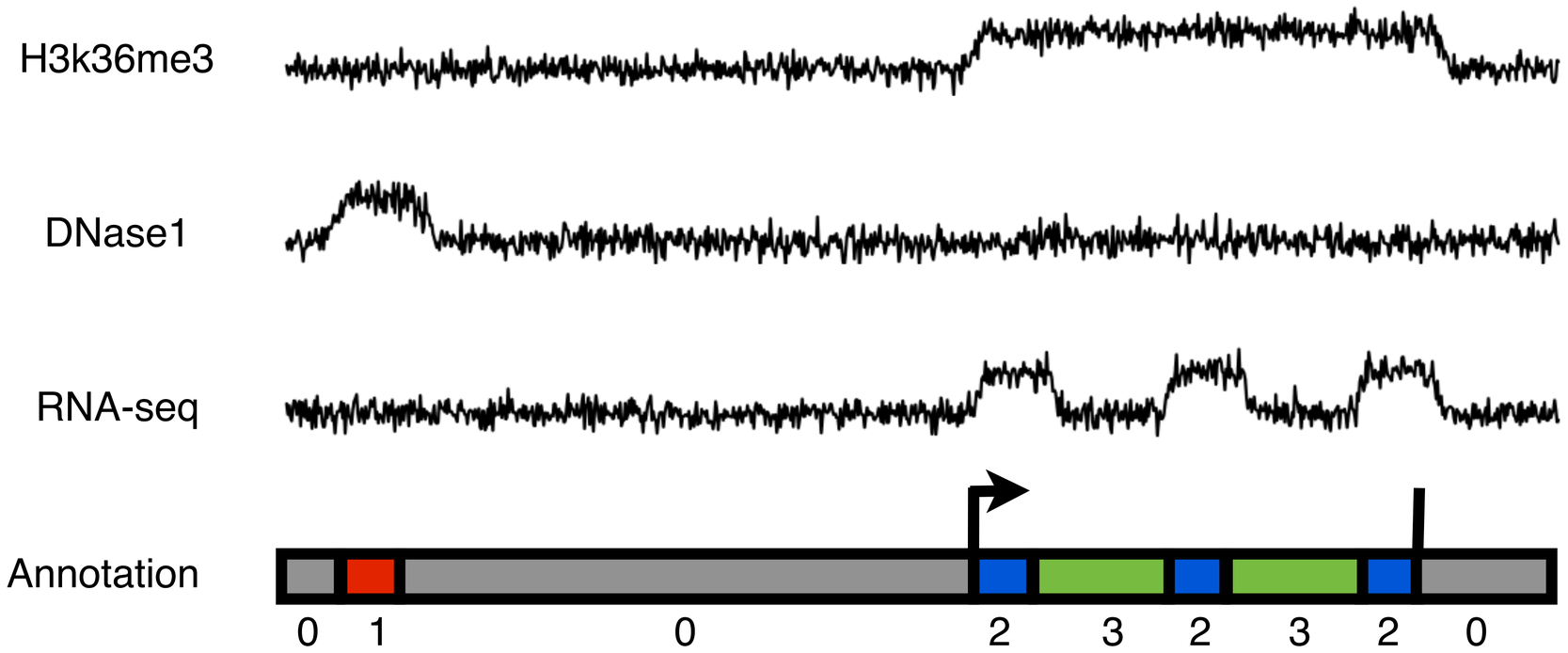
Genomedata

Segway



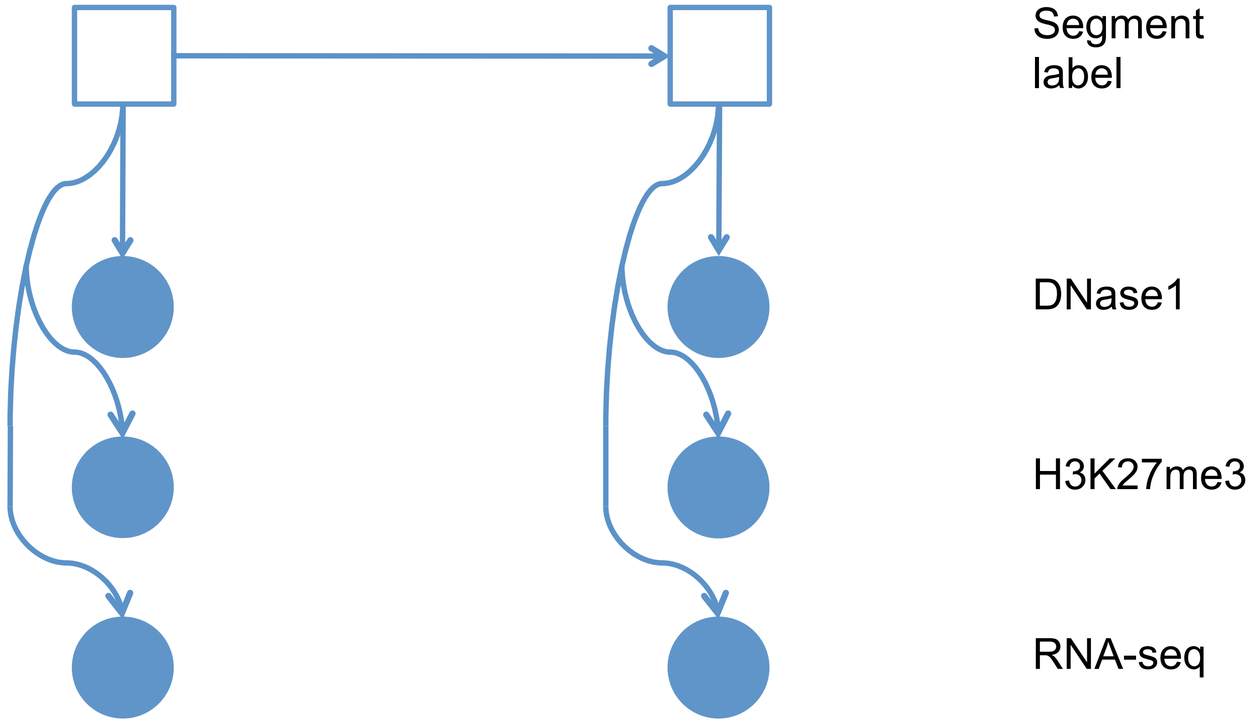
Segtools

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



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

Semi-automated genome annotation algorithms use dynamic Bayesian network models



hidden random variable

observed random variable

Installing Segway

```
# GMTK
wget http://melodi.ee.washington.edu/downloads/gmtk/gmtk-1.4.0.tar.gz
tar -xzvf gmtk-1.4.0.tar.gz
./configure
make
make install
cd ..

# Segway
pip install segway
```

Running Segway

```
segway train my_genomedata my_traindir
segway identify my_genomedata my_traindir my_identifydir

output: my_identifydir/segway.bed.gz
```

Model parameters

Number of annotation labels

```
--num-labels=25
```

Number of EM intializations

```
--num-instances=10
```

Maximum number of EM training iterations

```
--max-train-rounds=100
```

Input data

Input tracks

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

tracks.txt:
GM12878_H3K27ac
GM12878_H3K4me3
```

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
```

Controlling segment lengths

Downsampling resolution
`--resolution=10`

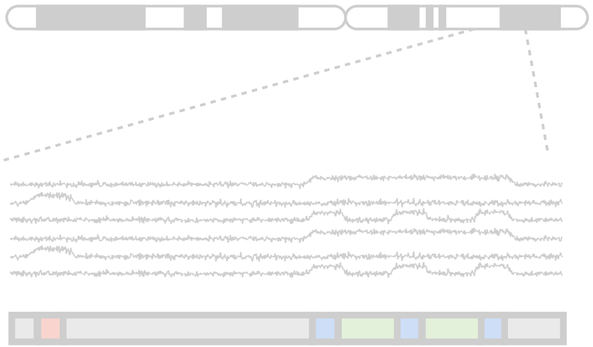
Long segments prior
`--prior-strength=1.0`

Weight on transition part of the model
`--segtransition-weight-scale=10`

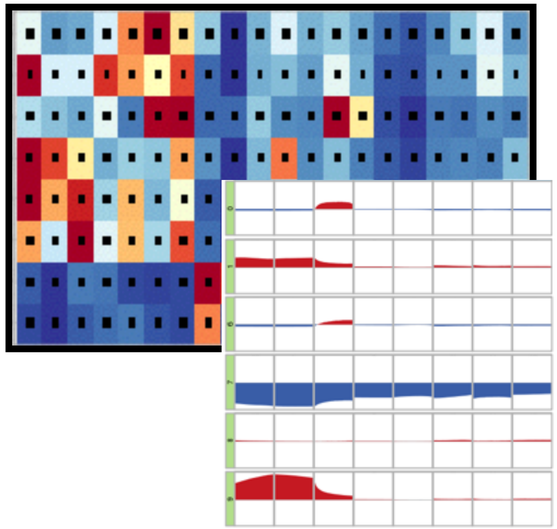
Genomedata, Segway and Segtools: How to use the Segway pipeline to store and analyze genomics data sets

Genomedata

Segway



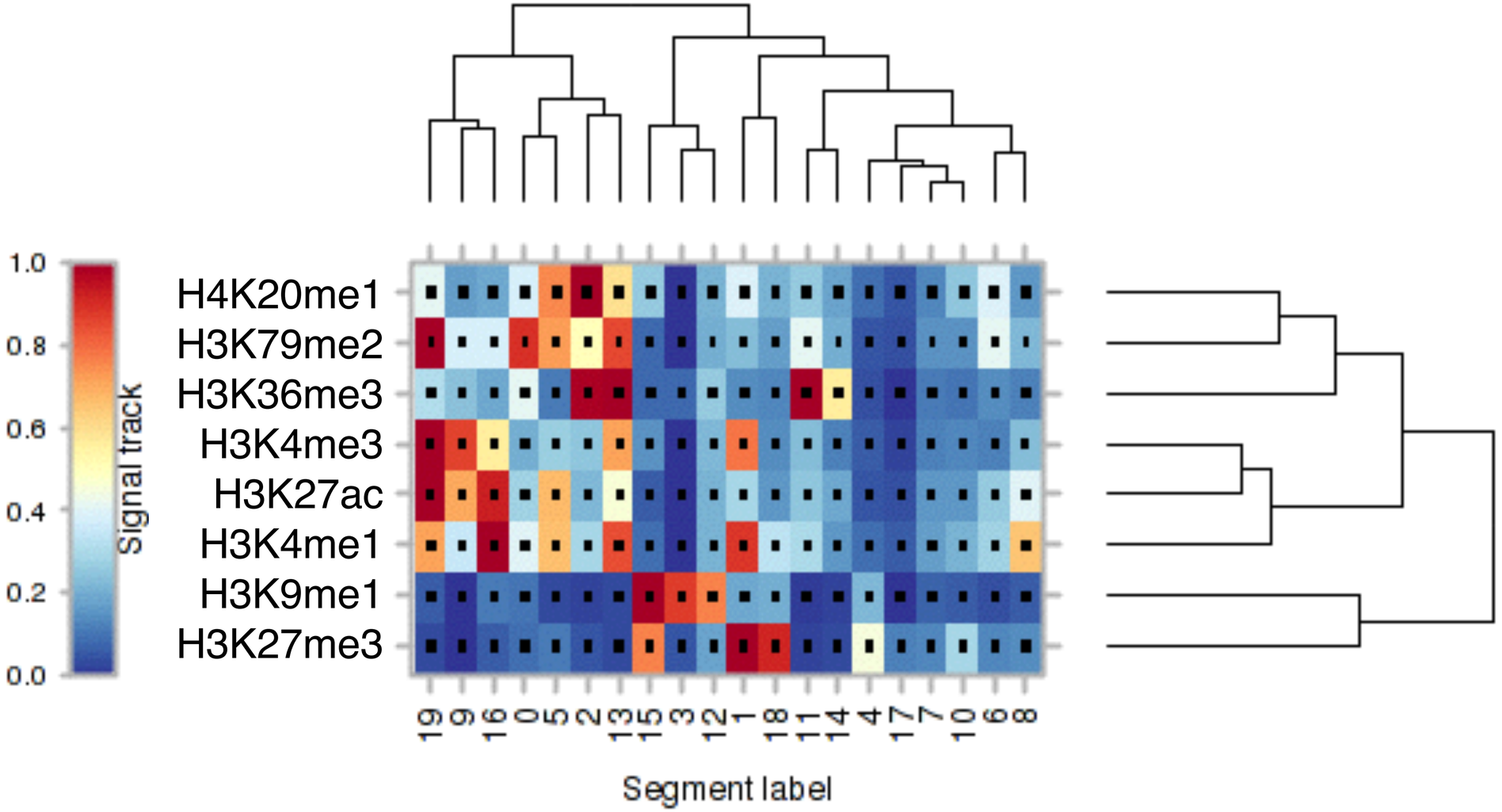
Segtools



Installing Segtools

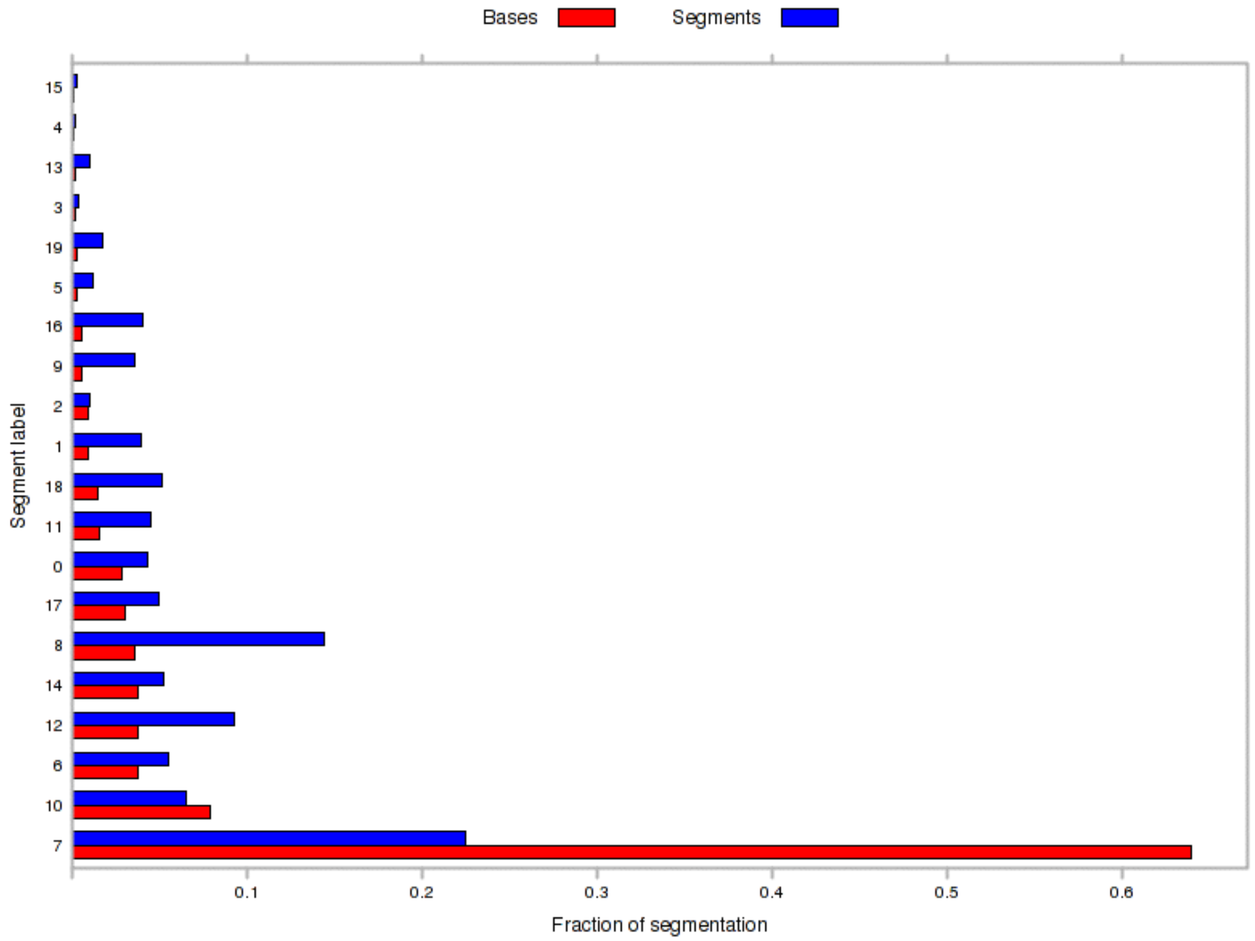
```
pip install segtools
```

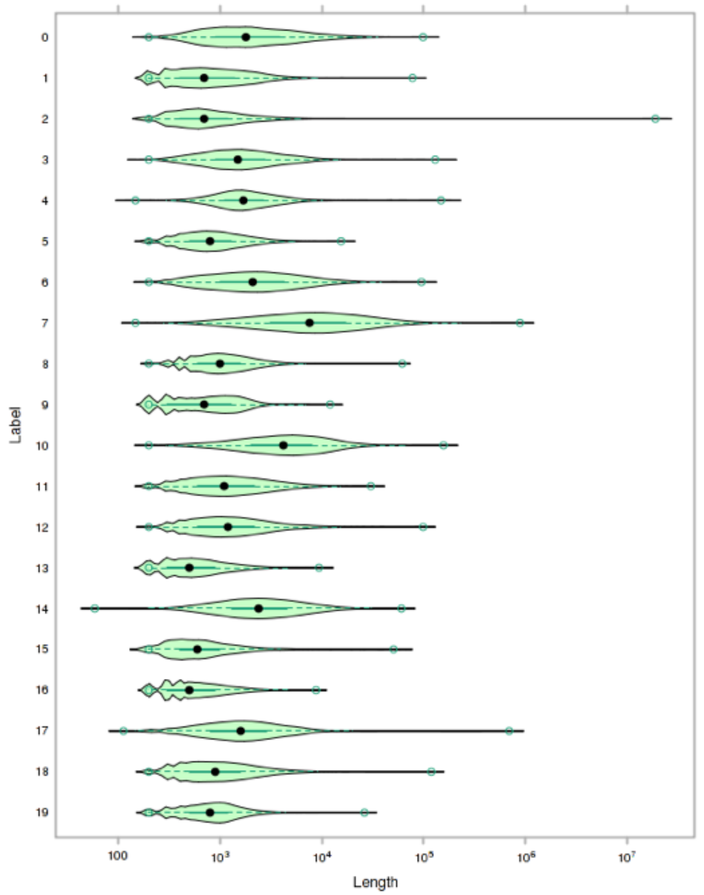
segtools-signal-distribution measures relationships between annotation labels and signal tracks



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
```