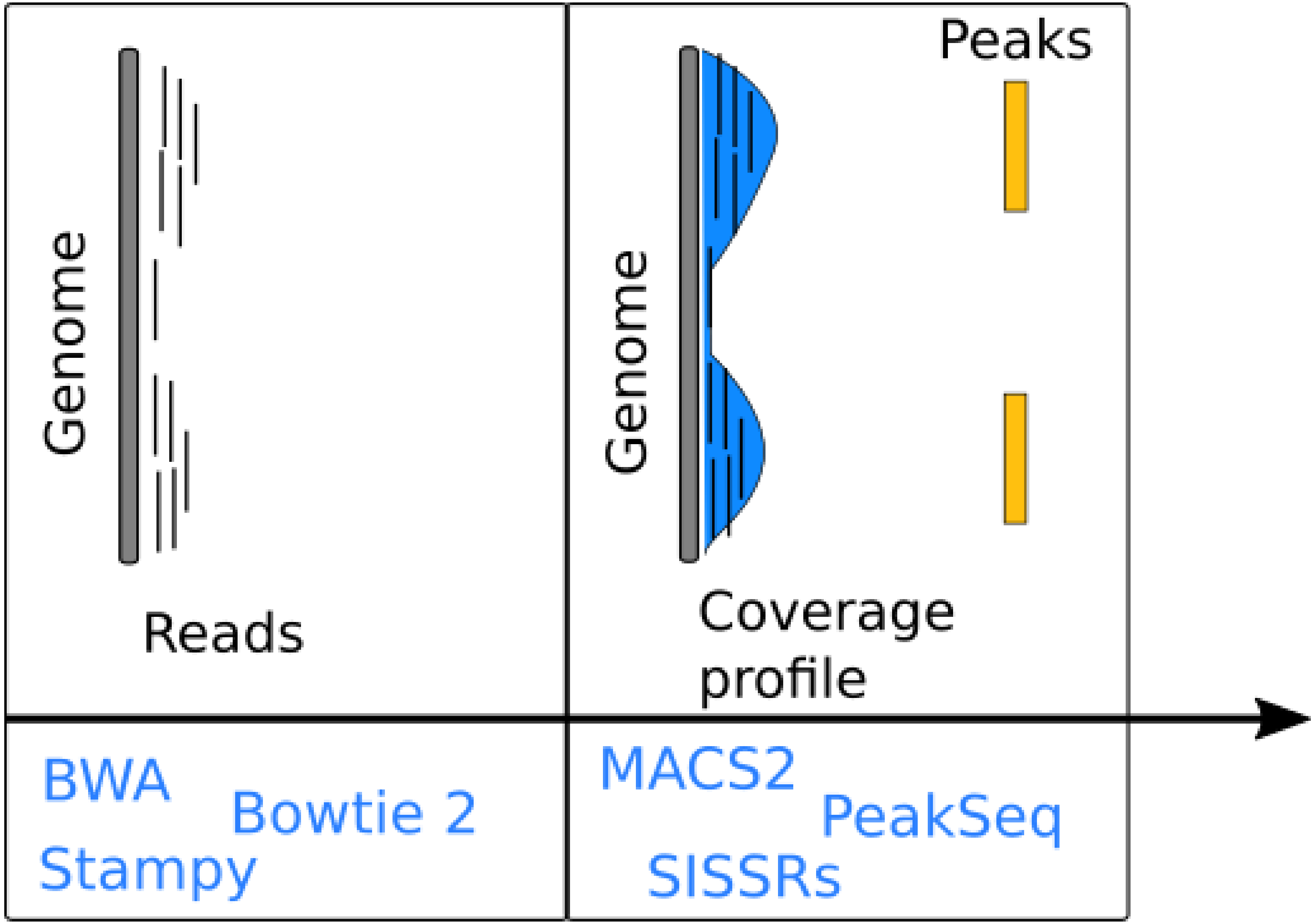
Importing data

CHIP-SEQ WITH BIOCONDUCTOR IN R

Peter Humburg
Statistician, Macquarie University


Genome
Reads
Genome
Peaks
Coverage profile
BWA
Bowtie 2
Stampy
MACS2
PeakSeq
SISSRs

Handling sequence reads

- Usually stored in **B**inary Sequence **A**lignment/**M**ap (BAM) format files.
- BAM record fields:
  - Read name: `SRR1782620.7265769`
  - Binary flag: `0`
  - Reference sequence name and position of alignment: `chr20 29803915`
  - Mapping quality: `0`
  - CIGAR string (alignment summary): `51M`
  - Reference sequence and position of paired read (not used here): `0 0`
  - Read sequence: `AATGAAATGGAA` ...
  - Read quality (ASCII encoded): `CCCFFFFFHHHH` ...

Importing mapped reads into R

- Use `Rsamtools` package to interact with BAM files.
- `Rsamtools` provides functions for indexing, reading, filtering and writing of BAM files.

Use `readGAlignments` to import mapped reads.

```r
library(GenomicAlignments)
reads <- readGAlignments(bam_file)
```

Returns `GAlignments` object.

Importing selected regions

- Use `BamViews` to define regions of interest.

```r
library(GenomicRanges)
library(Rsamtools)
ranges <- GRanges(...)
views <- BamViews(bam_file, bamRanges=ranges)
```

- Then import reads as before.

```r
reads <- readGAlignments(views)
```

The `BamViews` function supports multiple BAM files.

Importing peak calls

Use `import.bed` to load peak calls from a BED file.

```r
library(rtracklayer)
peaks <- import.bed(peak_bed, genome="hg19")
```

Use `peaks` to define views into the BAM files.

```r
bams <- BamViews(bam_file, bamRanges=peaks)
reads <- readGAlignments(bams)
```

Let's practice!

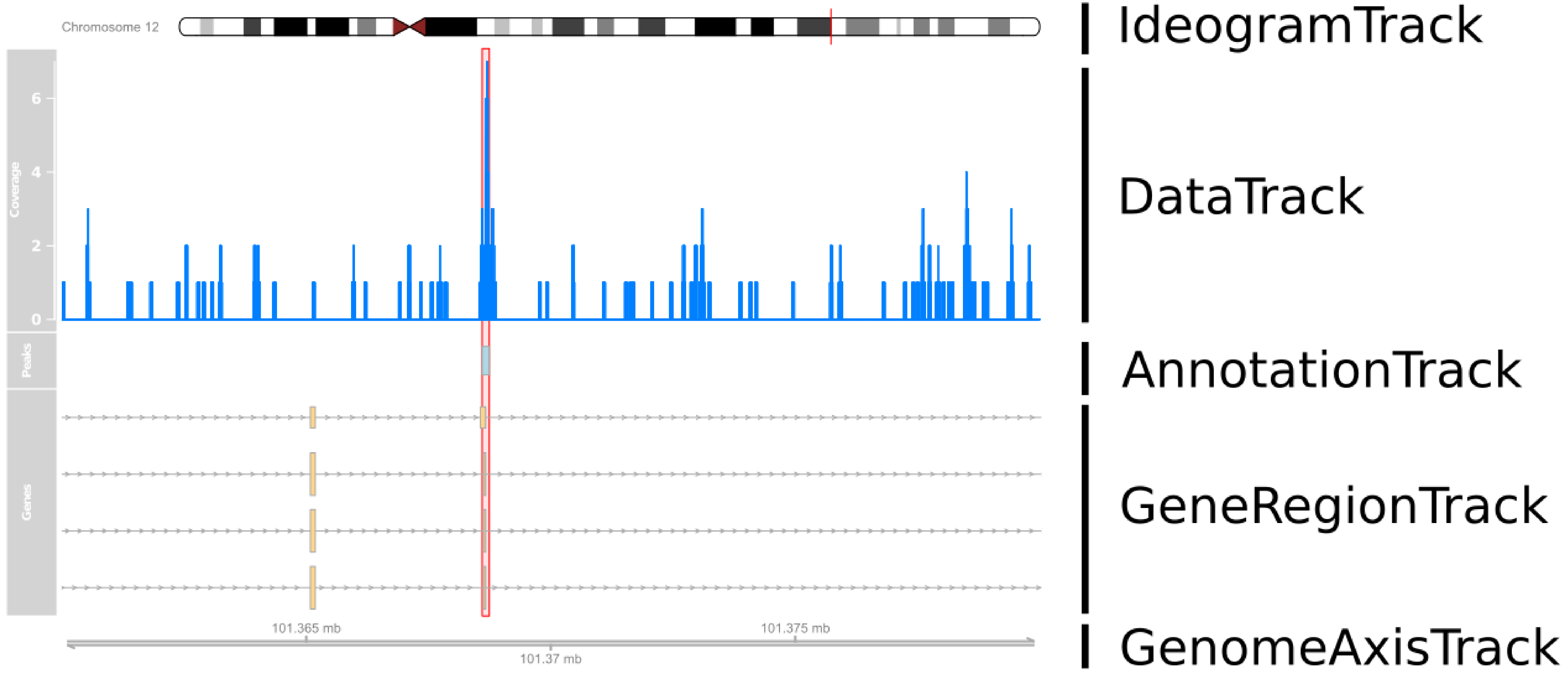
CHIP-SEQ WITH BIOCONDUCTOR IN R

Taking a closer look at peaks

CHIP-SEQ WITH BIOCONDUCTOR IN R

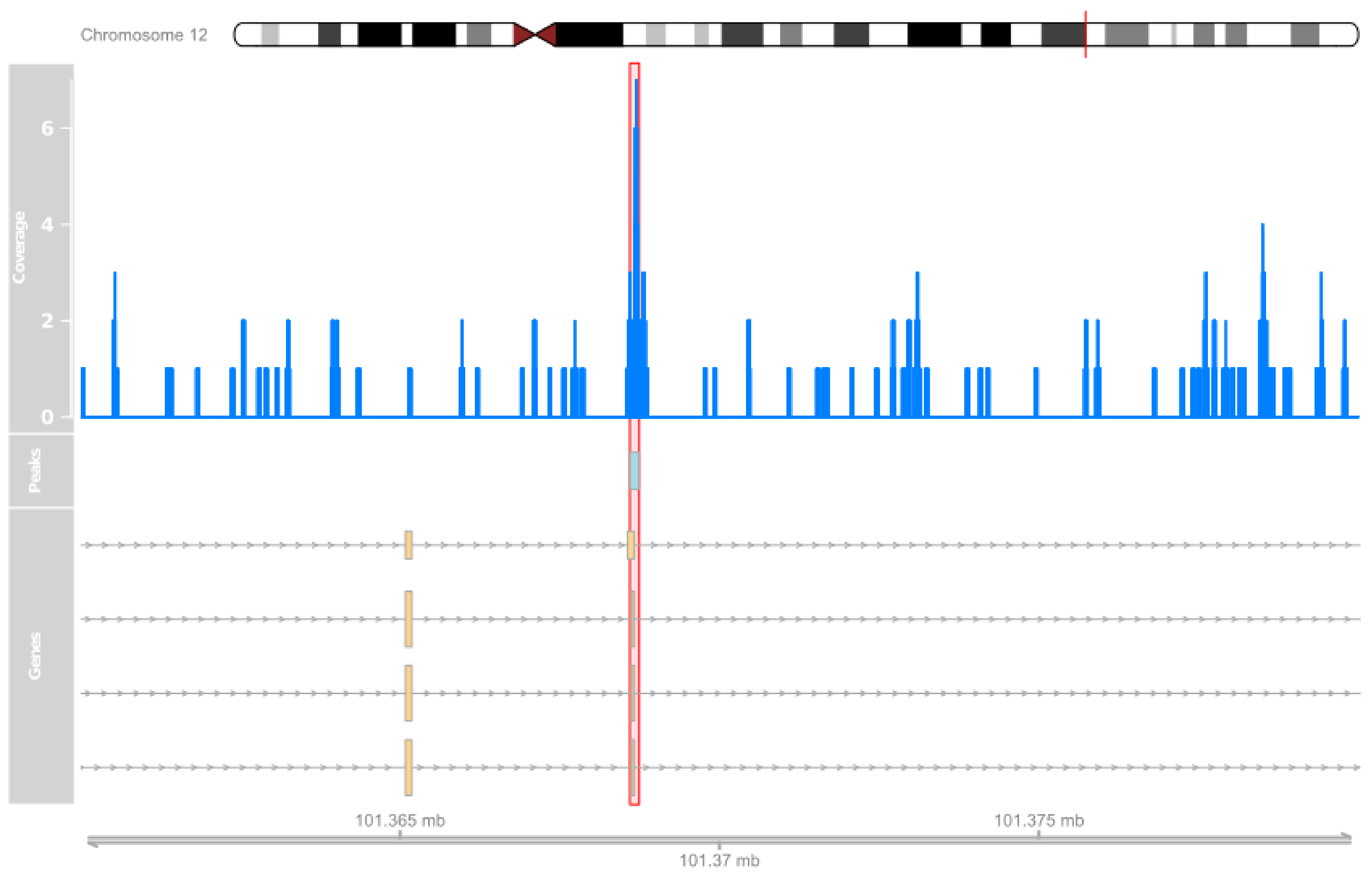
Peter Humburg
Statistician, Macquarie University

Using Gviz



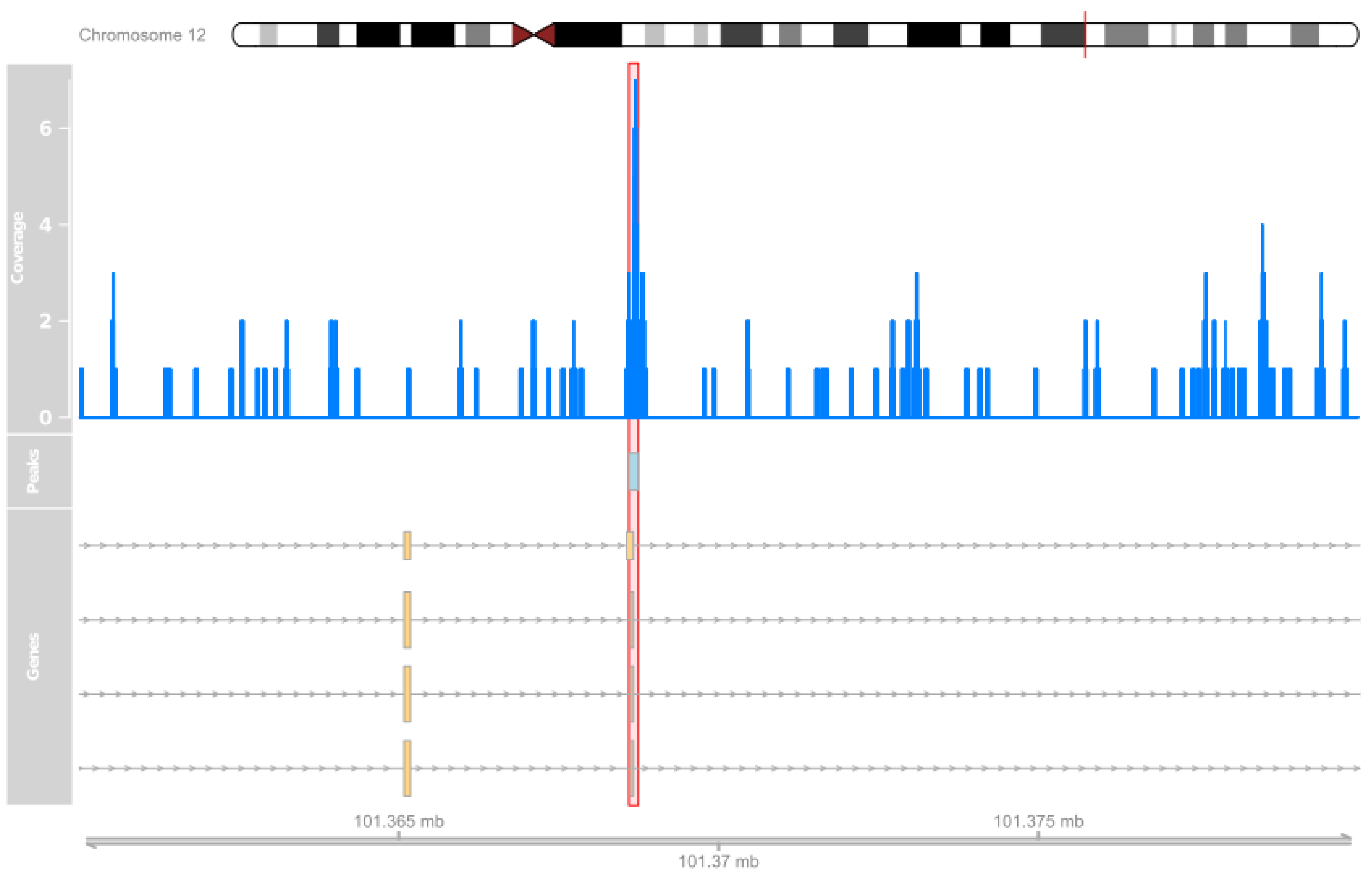
Chromosome 12
Coverage
6
4
2
0
Peaks
Genes
101.365 mb
101.37 mb
101.375 mb
IdeogramTrack
DataTrack
AnnotationTrack
GeneRegionTrack
GenomeAxisTrack

What are we plotting?



Ideogram

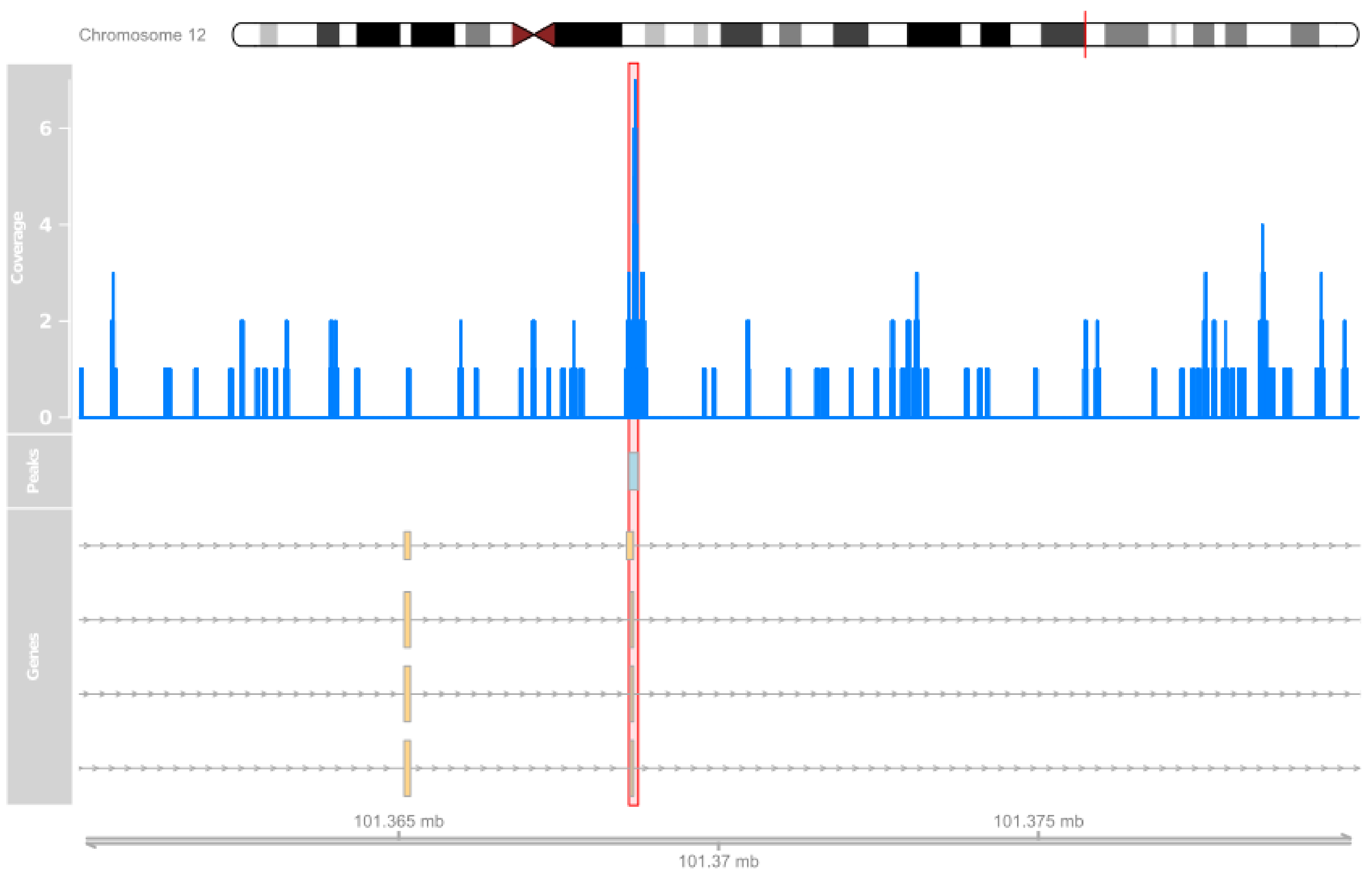
What are we plotting?



Ideogram

Coverage

What are we plotting?



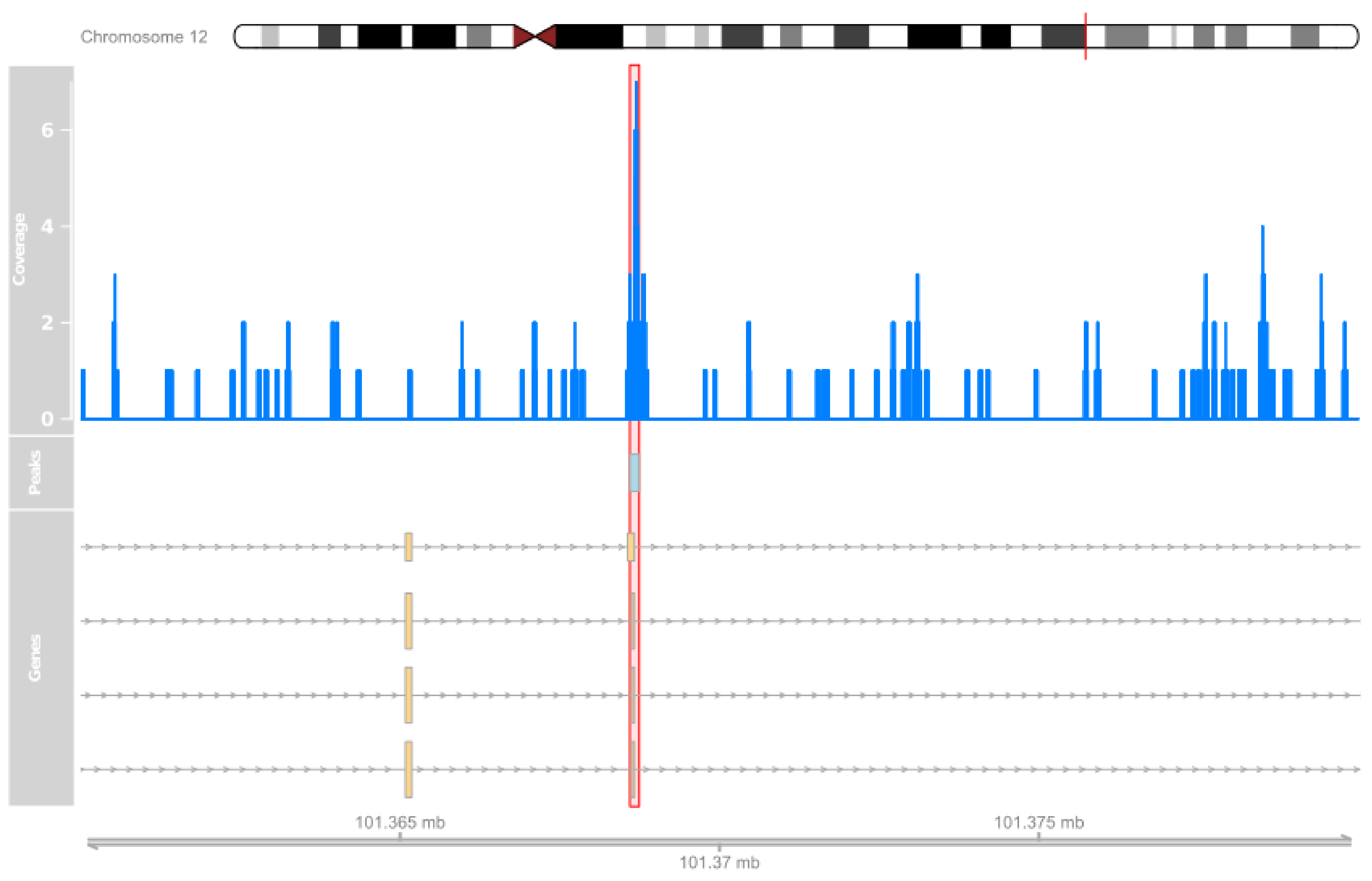
Ideogram

Coverage

Peaks

Genes

What are we plotting?



Ideogram

Coverage

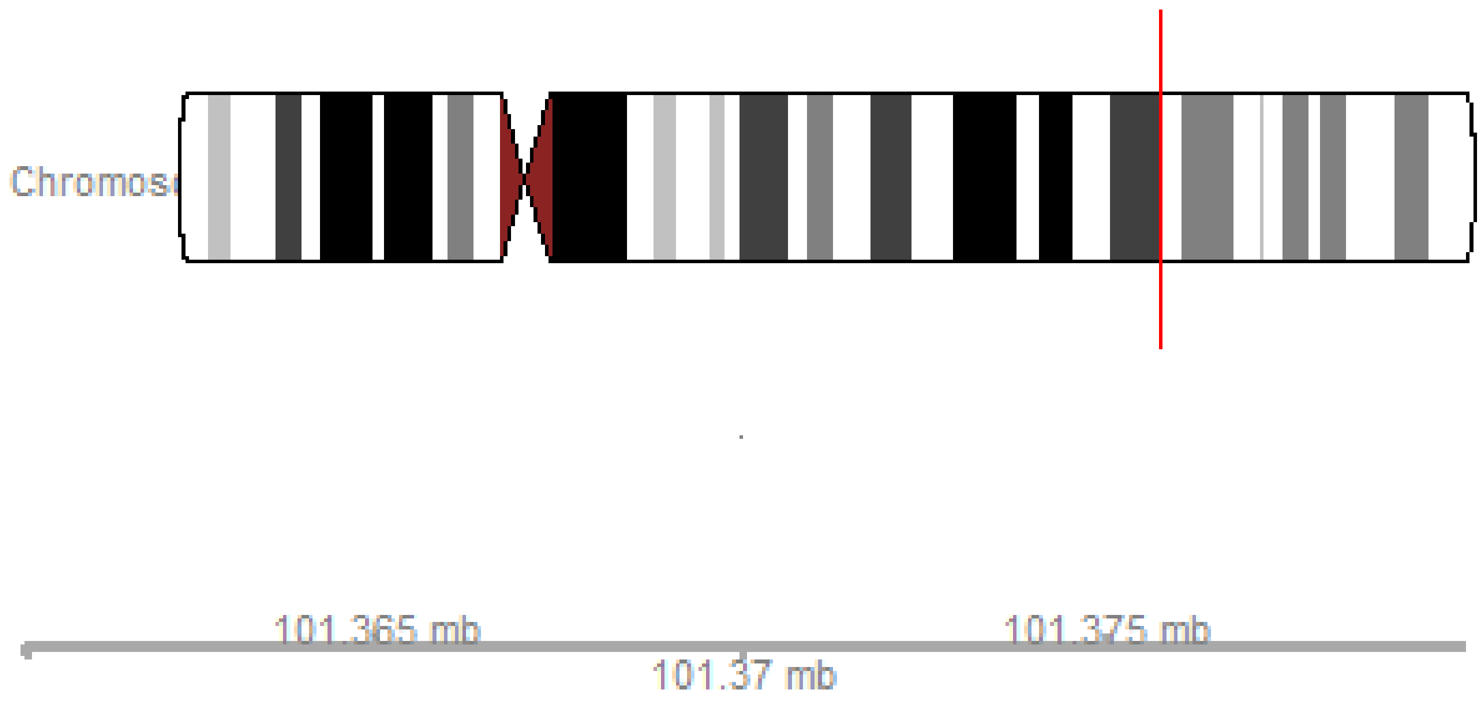
Peaks

Genes

Chromosome position

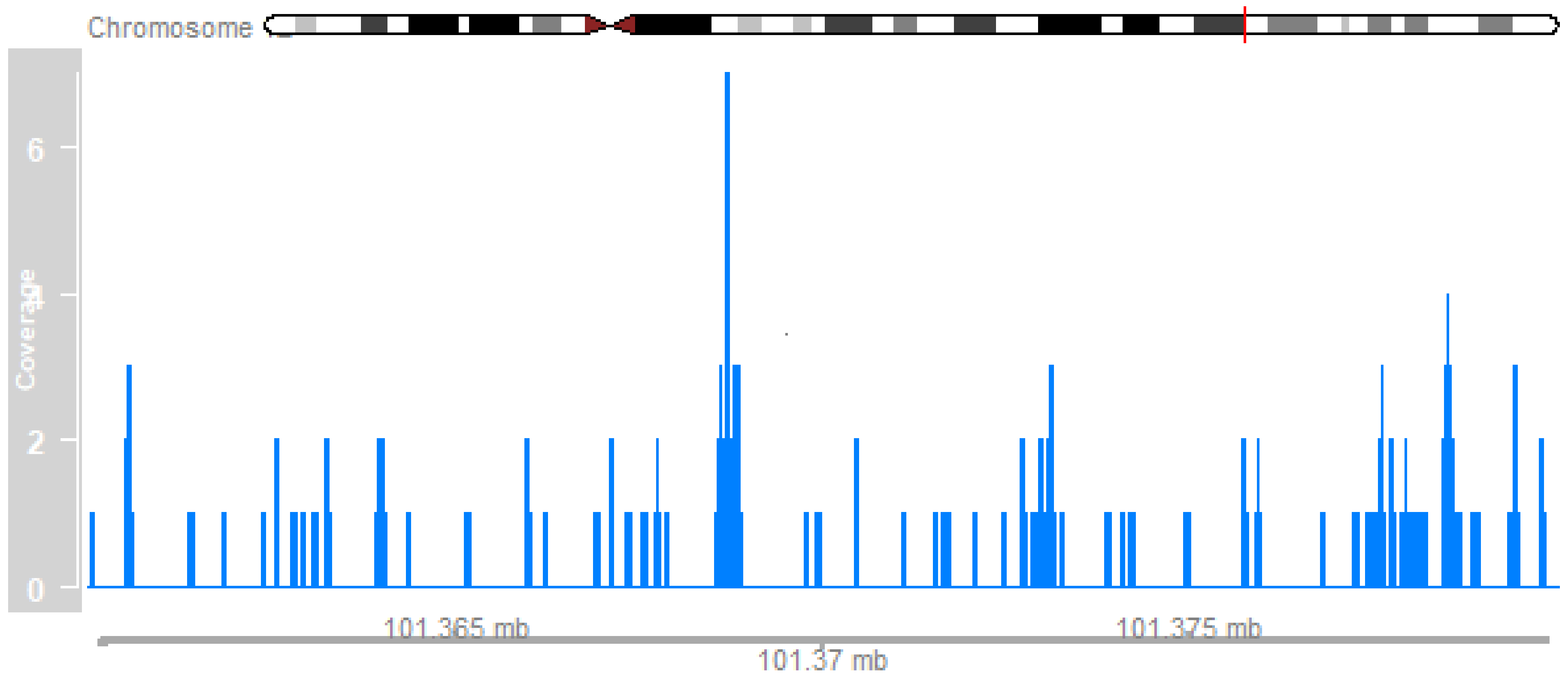
Setting-up coordinates

```
library(Gviz)
ideogram <- IdeogramTrack("chr12",
                          "hg19")
axis <- GenomeAxisTrack()
plotTracks(list(ideogram, axis),
           from=101360000,
           to=101380000)
```



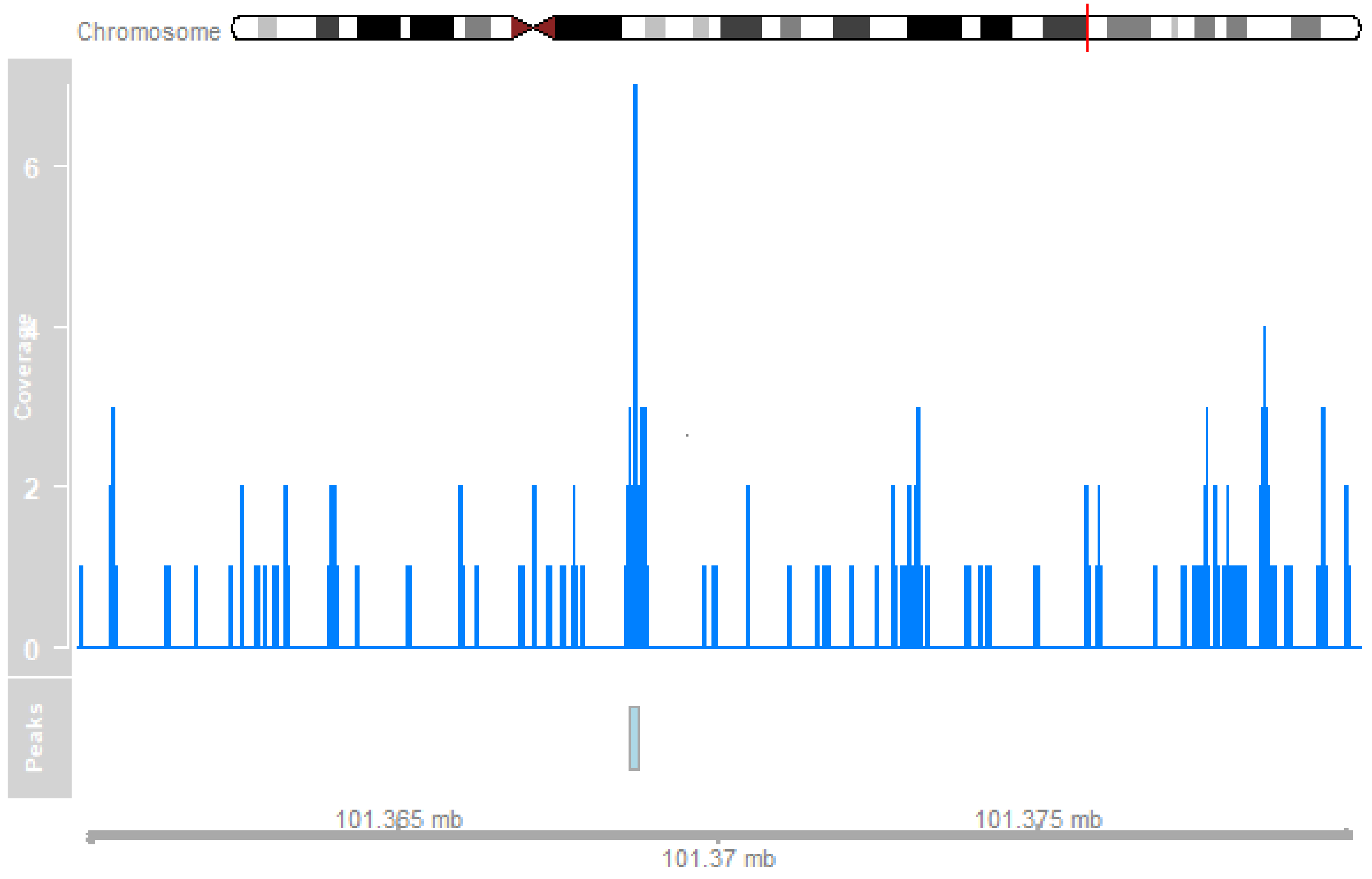
Adding Data

```
cover_track <- DataTrack(cover_ranges,window=100000,type='h',name="Coverage")
plotTracks(list(ideogram, cover_track, axis), from=101360000, to=101380000)
```



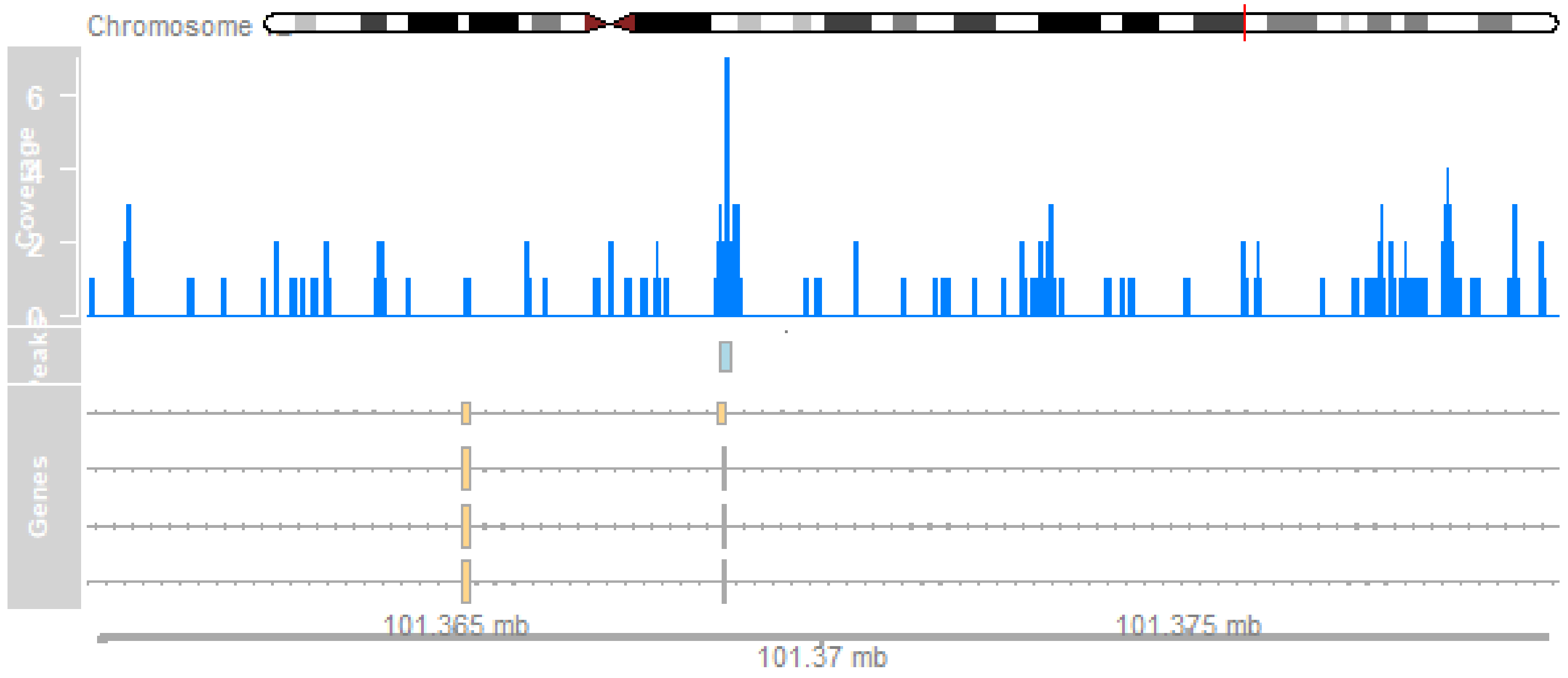
Adding Annotations

```
peak_track <- AnnotationTrack(peaks, name="Peaks")
plotTracks(list(ideogram, cover_track, peak_track, axis),
           from=101360000, to=101380000)
```



Gene Annotations

```r
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
tx <- GeneRegionTrack(TxDb.Hsapiens.UCSC.hg19.knownGene, chromosome="chr12",
                      start=101360000, end=101380000, name="Genes")
plotTracks(list(ideogram, cover_track, peak_track, tx, axis),
           from=101360000, to=101380000)
```



Let's practice!

CHIP-SEQ WITH BIOCONDUCTOR IN R

Cleaning ChIP-seq data

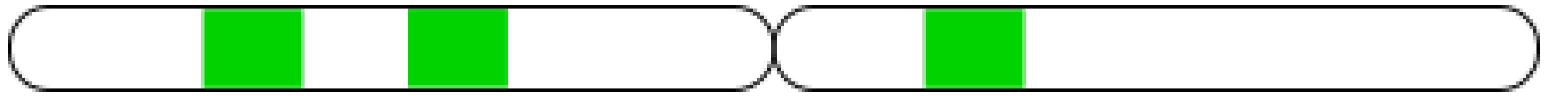
CHIP-SEQ WITH BIOCONDUCTOR IN R

Peter Humburg
Statistician, Macquarie University

Common Problems

Incorrectly mapped reads may produce false peaks.

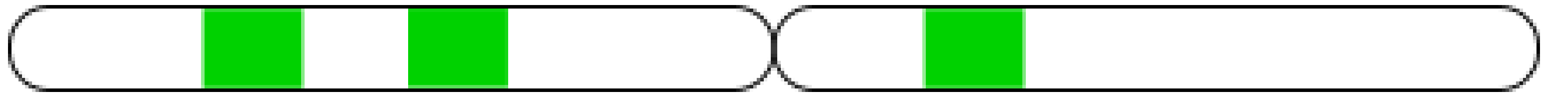
- Genomic repeats.



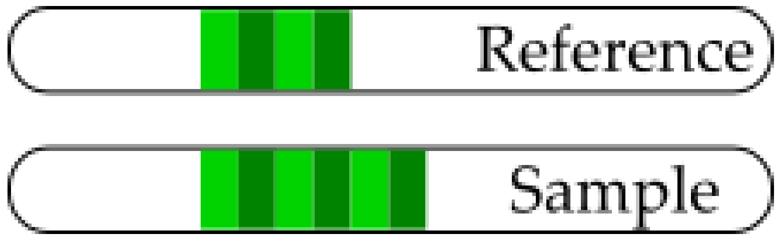
Common Problems

Incorrectly mapped reads may produce false peaks.

- Genomic repeats.



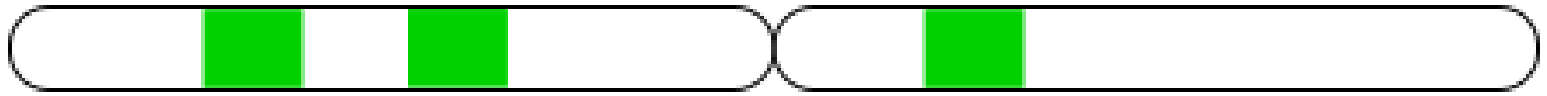
- Incomplete reference sequence.



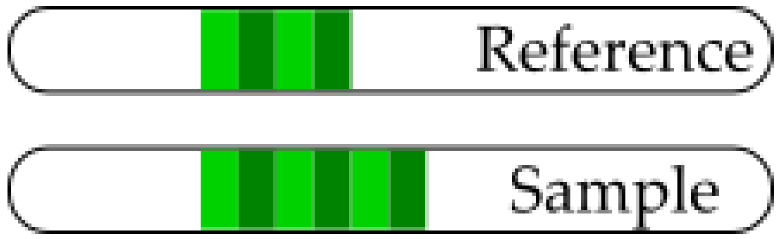
Common Problems

Incorrectly mapped reads may produce false peaks.

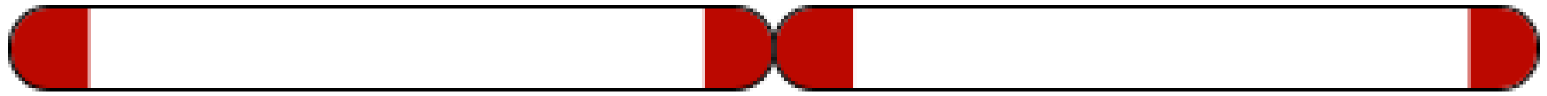
- Genomic repeats.

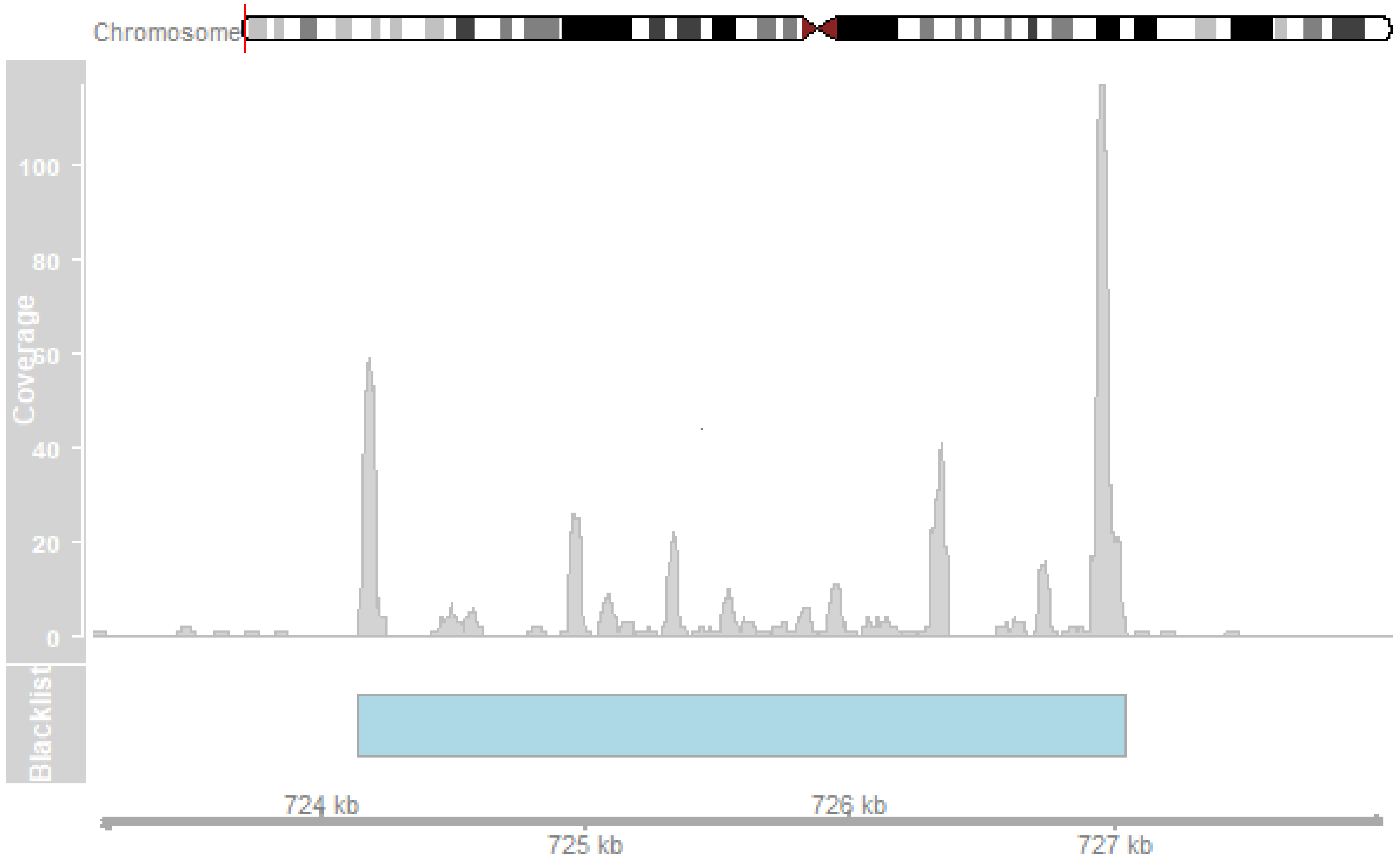


- Incomplete reference sequence.



- Low complexity regions.



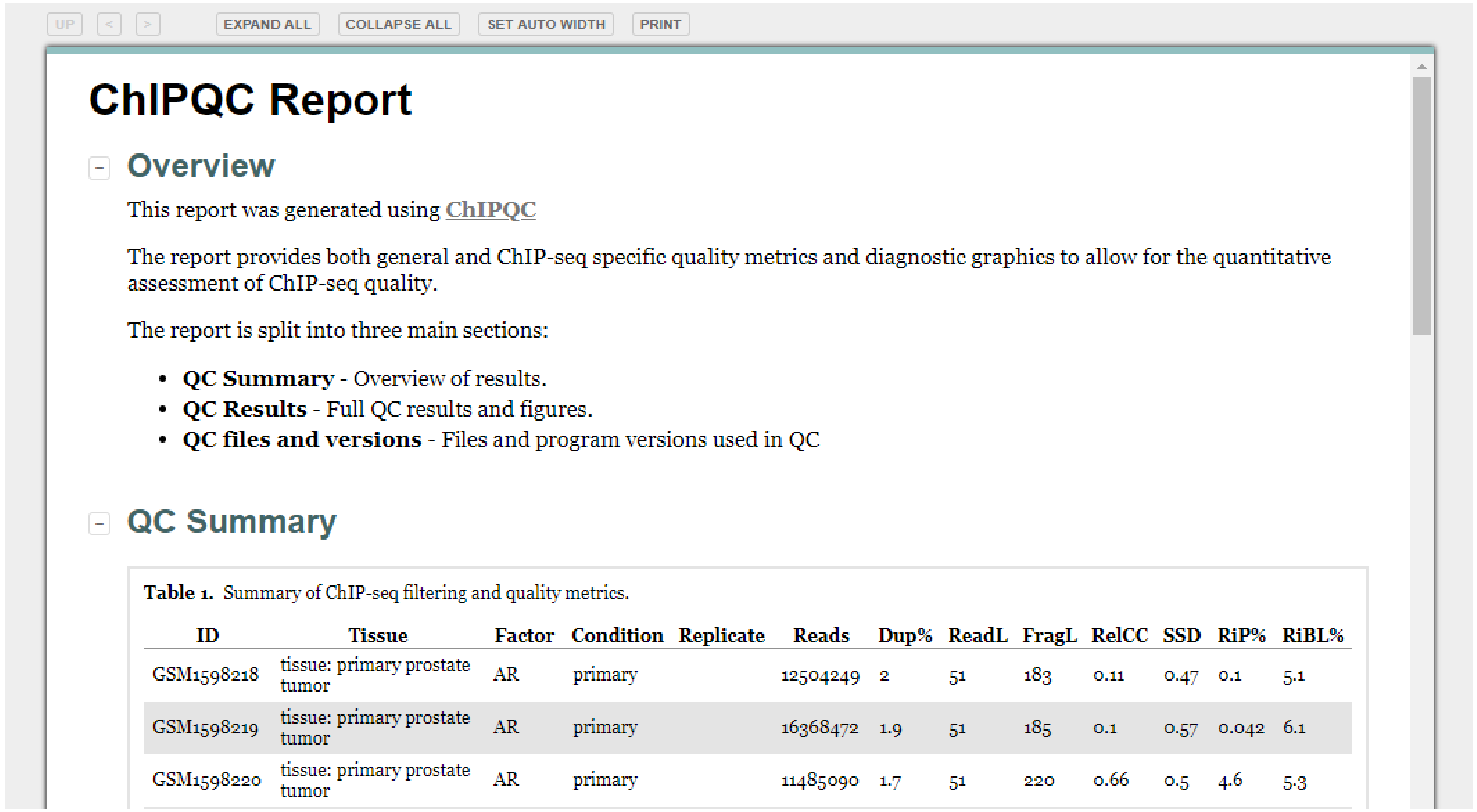

Chromosome
Coverage
100
80
60
40
20
0
Blacklist
724 kb
725 kb
726 kb
727 kb

Amplification Bias

- DNA fragments extracted from cells are copied multiple times prior to sequencing.
- Not all fragments produce the same number of copies.
- Multiple copies of the same fragment may be sequenced.
- A single DNA fragment may inflate coverage and lead to incorrect peak calls.

Quality Control Reports

```r
library(ChIPQC)
qc_report <- ChIPQC(experiment="sample_info.csv", annotation="hg19")
ChIPQCreport(qc_report)
```



ChIPQC Report

Overview

This report was generated using ChIPQC

The report provides both general and ChIP-seq specific quality metrics and diagnostic graphics to allow for the quantitative assessment of ChIP-seq quality.

The report is split into three main sections:

- **QC Summary** - Overview of results.
- **QC Results** - Full QC results and figures.
- **QC files and versions** - Files and program versions used in QC

QC Summary

Table 1. Summary of ChIP-seq filtering and quality metrics.

ID	Tissue	Factor	Condition	Replicate	Reads	Dup%	ReadL	FragL	RelCC	SSD	RiP%	RiBL%
GSM1598218	tissue: primary prostate tumor	AR	primary		12504249	2	51	183	0.11	0.47	0.1	5.1
GSM1598219	tissue: primary prostate tumor	AR	primary		16368472	1.9	51	185	0.1	0.57	0.042	6.1
GSM1598220	tissue: primary prostate tumor	AR	primary		11485090	1.7	51	220	0.66	0.5	4.6	5.3

Preparing input files

SampleID	Factor	Condition	Tissue	Treatment	bamReads	Peaks	PeakCaller
S1	AR	primary	primary prostate tumor	gleason score: 3+4=7	S1.bam	S1.bed	macs
S2	AR	primary	primary prostate tumor	gleason score: 3+4=7	S2.bam	S2.bed	macs
...	...	...	...	...	...	...	...

Cleaning the Data

- **Remove duplicate reads.**
- Remove reads with multiple hits.
- Remove reads with low mapping quality.
- Remove peaks in blacklisted regions.

Cleaning the Data

- Remove duplicate reads.
- **Remove reads with multiple hits.**
- **Remove reads with low mapping quality.**
- Remove peaks in blacklisted regions.

Cleaning the Data

- Remove duplicate reads.
- Remove reads with multiple hits.
- Remove reads with low mapping quality.
- **Remove peaks in blacklisted regions.**
    - Blacklisted regions are available from the ENCODE project.

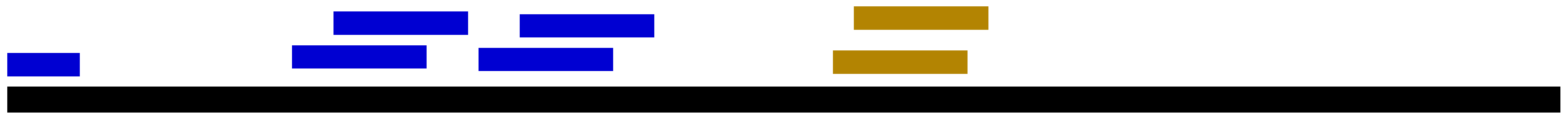
Let's practice!

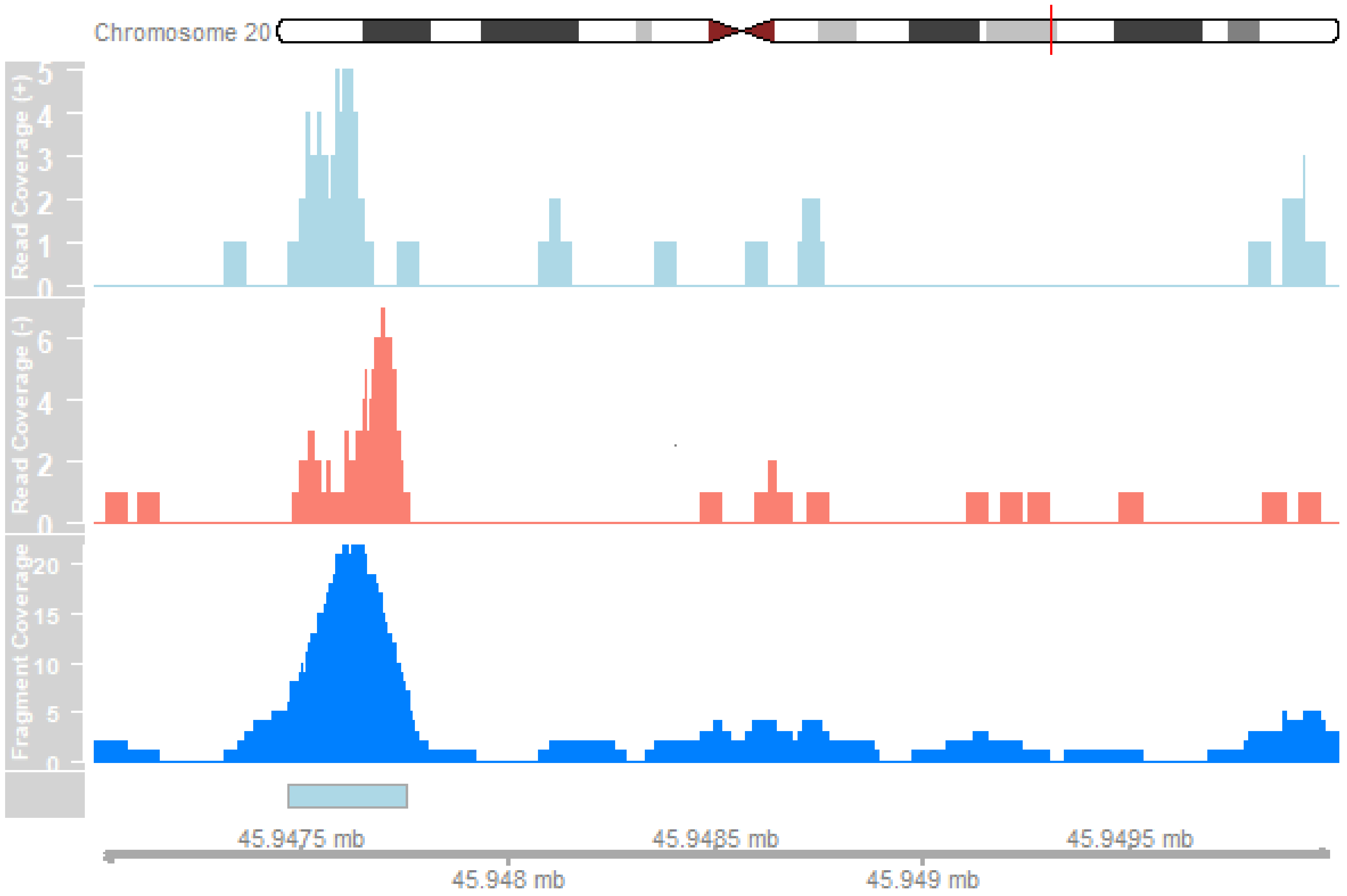
CHIP-SEQ WITH BIOCONDUCTOR IN R

Assessing enrichment

CHIP-SEQ WITH BIOCONDUCTOR IN R

Peter Humburg
Statistician, Macquarie University




Chromosome 20
Read Coverage (+)
5
4
3
2
1
0
Read Coverage (-)
6
4
2
0
Fragment Coverage
20
15
10
5
0
45.9475 mb
45.9485 mb
45.9495 mb
45.948 mb
45.949 mb

Extending reads

Load the data:

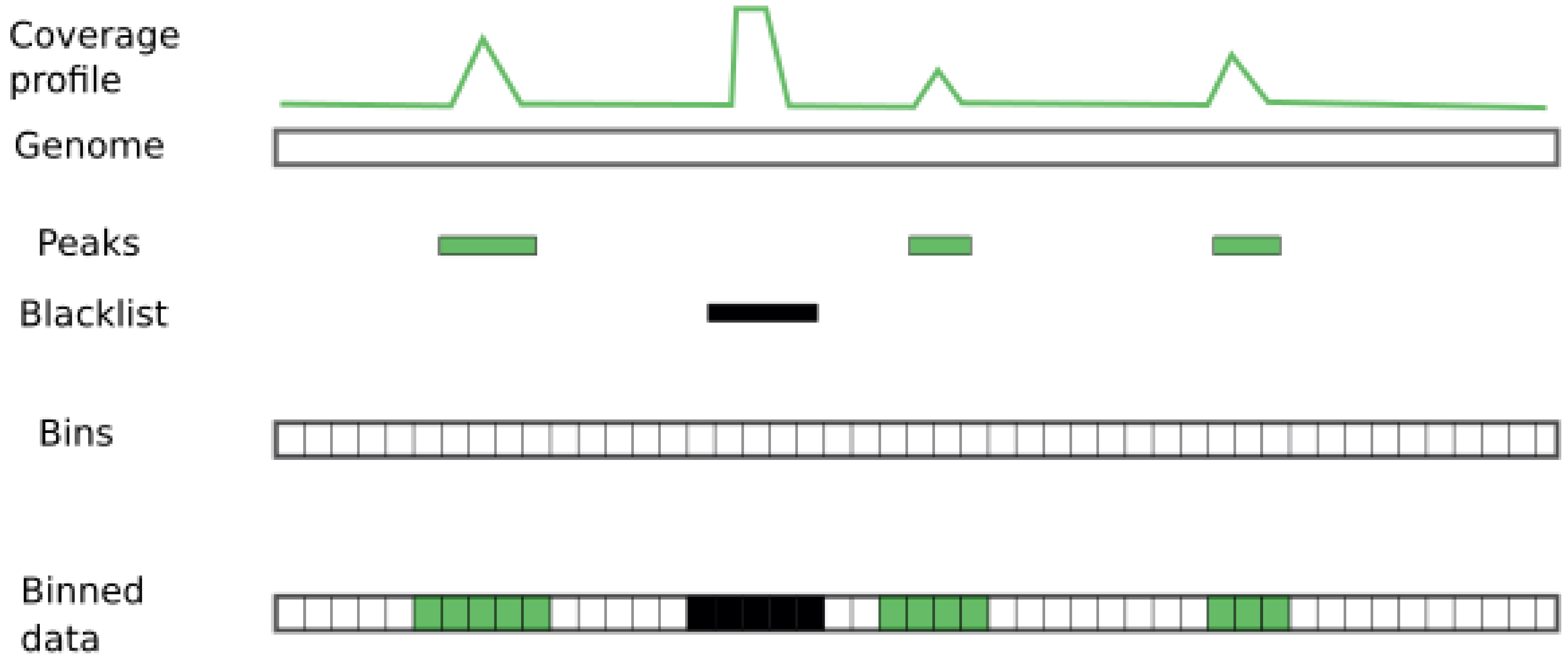
```
reads <- readGAlignments(bam)
reads_gr <- granges(reads[[1]])
```

Obtain average fragment length:

```
frag_length <- fragmentlength(qc_report)["GSM1598218"]
```

Extend reads and compute coverage:

```
reads_ext <- resize(reads_gr, width=frag_length)
cover_ext <- coverage(reads_ext)
```


Coverage profile
Genome
Peaks
Blacklist
Bins
Binned data

Coverage for peaks

Create 200 bp bins along the genome.

```
bins <- tileGenome(seqinfo(reads), tilewidth=200,
                   cut.last.tile.in.chrom=TRUE)
```

Find all bins overlapping peaks.

```
peak_bins_overlap <- findOverlaps(bins, peaks)
peak_bins <- bins[from(peak_bins_overlap), ]
```

Count the number of reads overlapping each peak bin.

```
peak_bins$score <- countOverlaps(peak_bins, reads)
```

Binned coverage function

```r
count_bins <- function(reads, target, bins){
  # Find all bins overlapping peaks
  overlap <- from(findOverlaps(bins, target))
  target_bins <- bins[overlap, ]

  # Count the number of reads overlapping each peak bin
  target_bins$score <- countOverlaps(target_bins, reads)
  target_bins
}
```

Coverage for blacklisted regions

```
peak_bins <- count_bins(reads_ext, peaks, bins)
bl_bins <- count_bins(reads_ext, blacklist.hg19, bins)
```

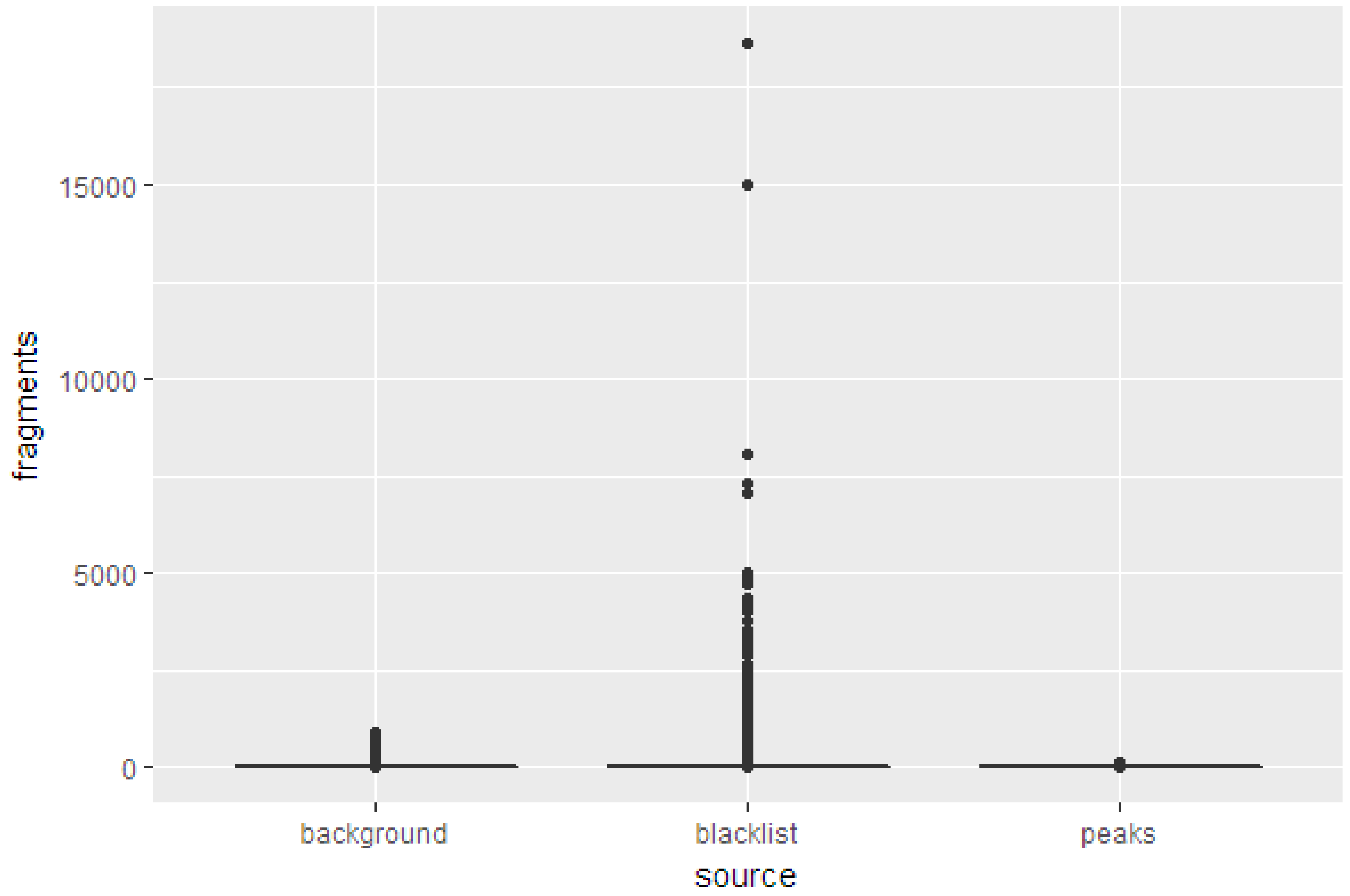
Background coverage

Remove all bins already accounted for.

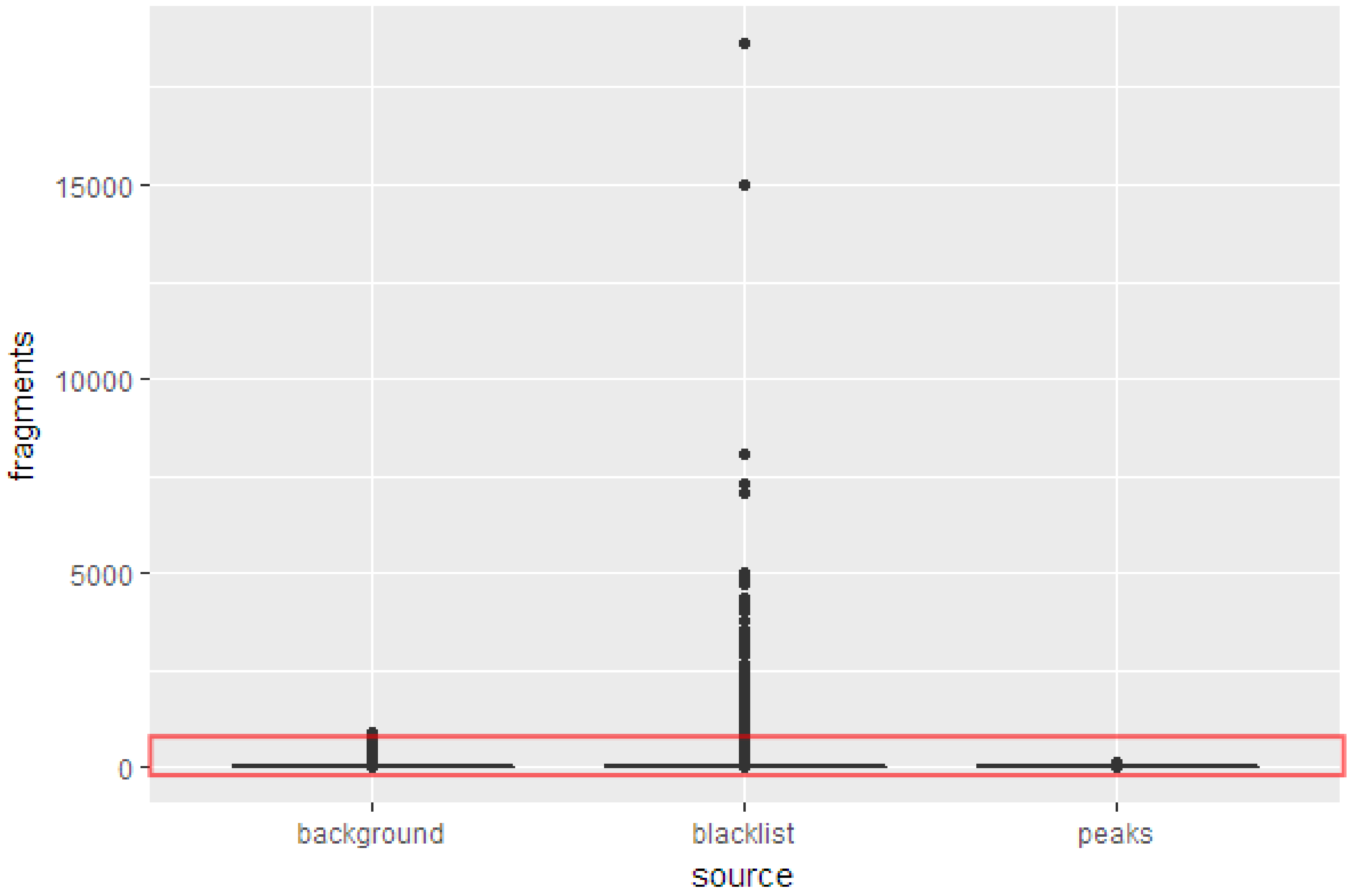
```
bkg_bins <- subset(bins, !bins %in% peak_bins & !bins %in% bl_bins)
```

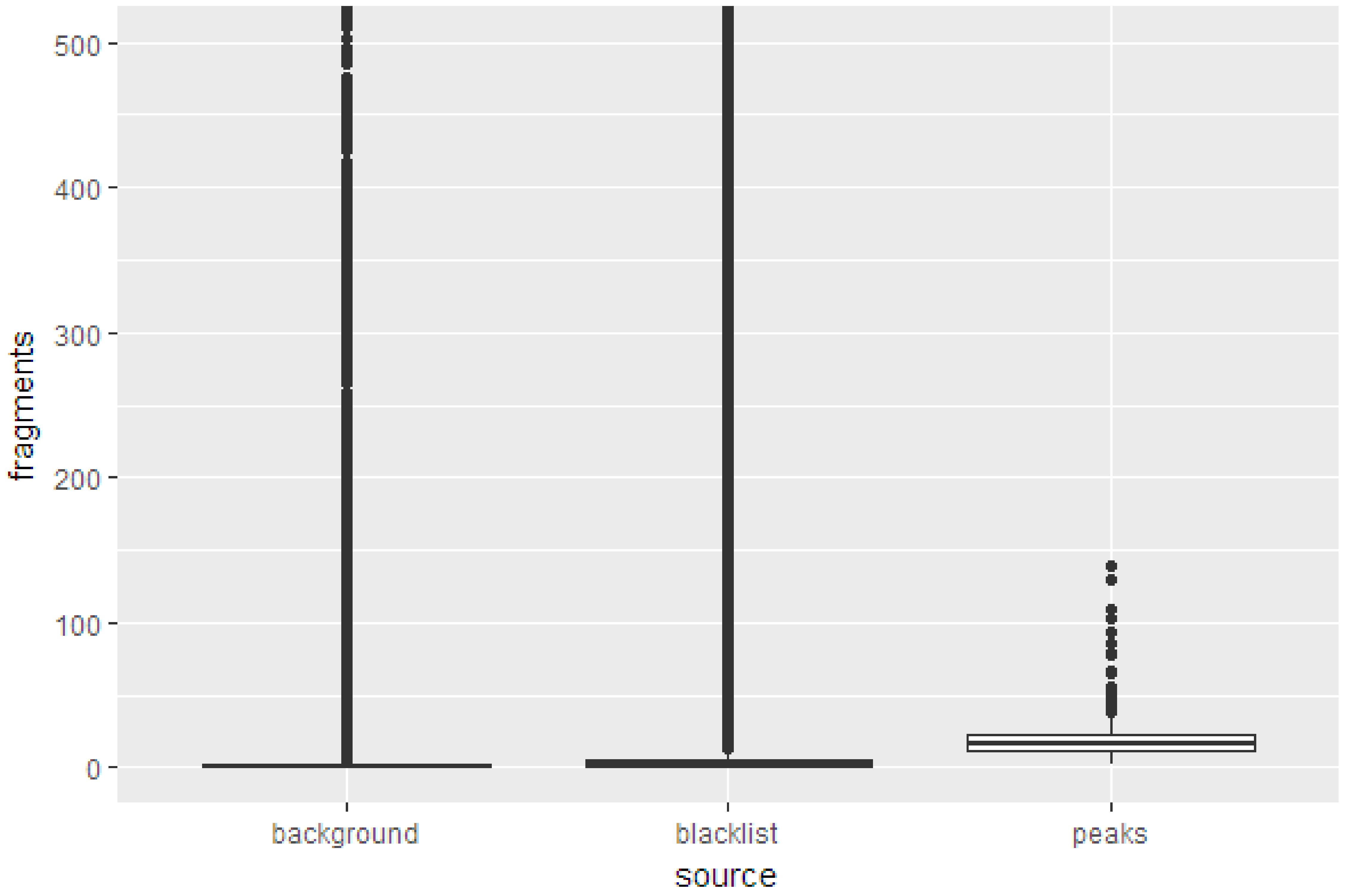
Count number of reads overlapping with each remaining bin.

```
bkg_bins$score <- countOverlaps(bkg_bins, reads_ext)
```



fragments
15000
10000
5000
0
background
blacklist
peaks
source


15000
10000
5000
0
fragments
background
blacklist
peaks
source


fragments
500
400
300
200
100
0
background
blacklist
peaks
source

Let's practice!

CHIP-SEQ WITH BIOCONDUCTOR IN R