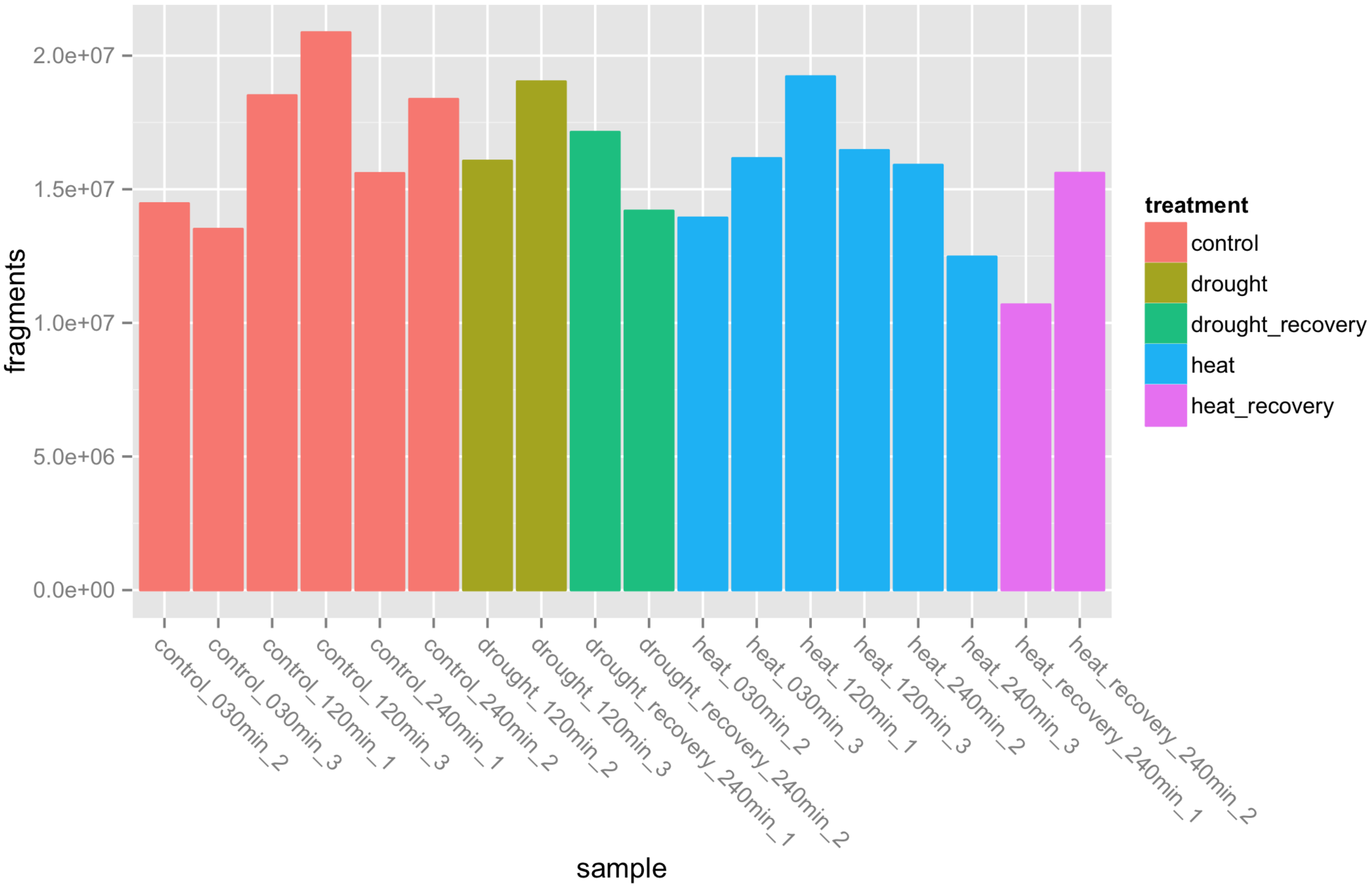
# ATAC-seq Analysis

- Process the reads
- Align the reads
- Call peaks
- Investigate the location of peaks

# ATAC-seq Analysis

- ~~Process the reads~~
- Align the reads
- Call peaks
- Investigate the location of peaks

# Ca. 10M to 20M read pairs per library



# Align reads using bowtie2, process using samtools

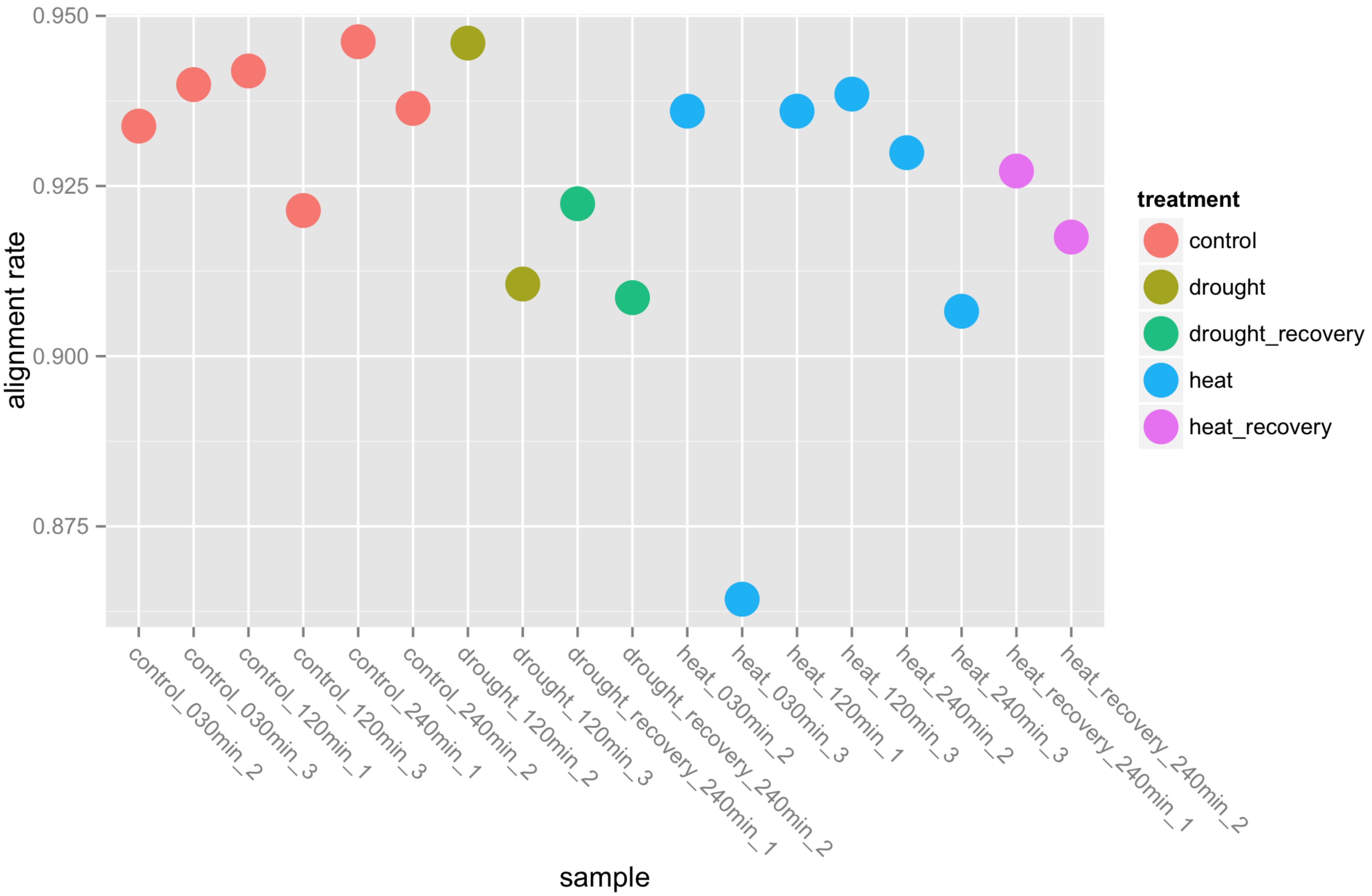
```
bowtie2 --threads 12 \
        --very-sensitive \
        --maxins 2000 \
        --no-discordant \
        -x $genomedir \
        -1 "$left" \
        -2 "$right" | samtools view -bS - -o $bam;
```

Not in effect unless
--no-mixed is also set

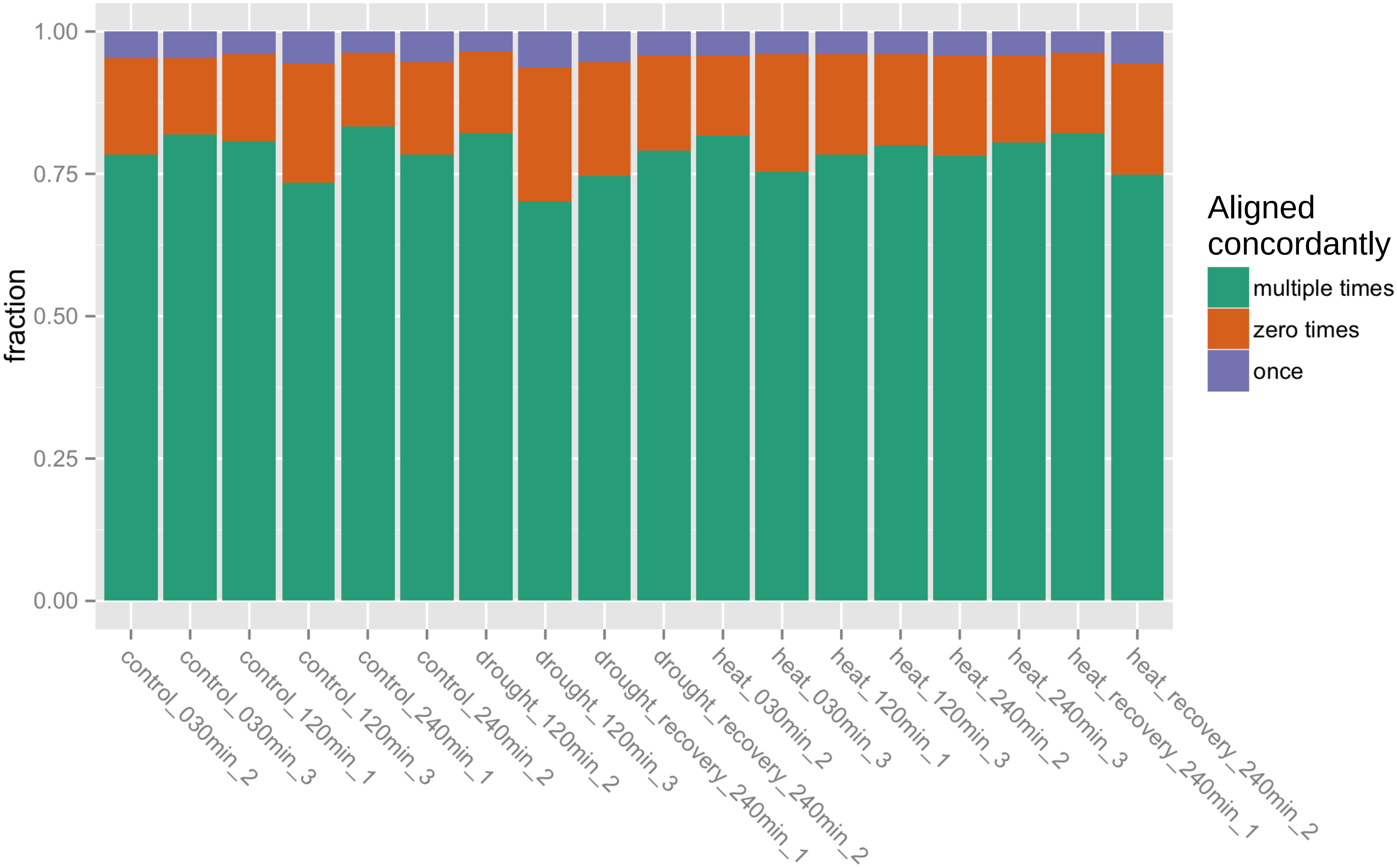
samtools:

- sort by name
- fixmate (Fill in mate coordinates, ISIZE and mate related flags)
- rmdup (Remove potential PCR duplicates)
- sort by position
- index

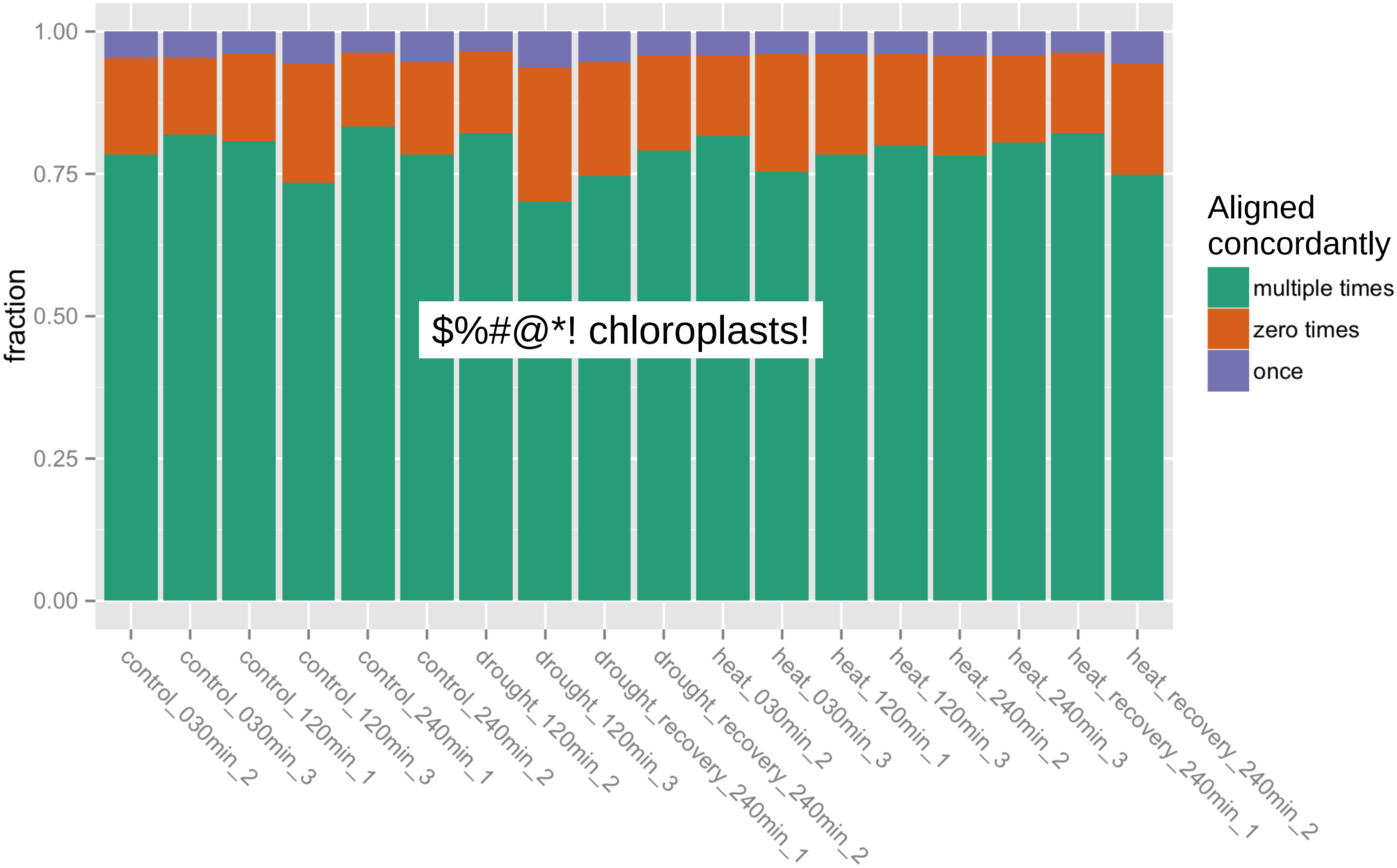
# High alignment rate of around 92%



# But, many low-quality alignments

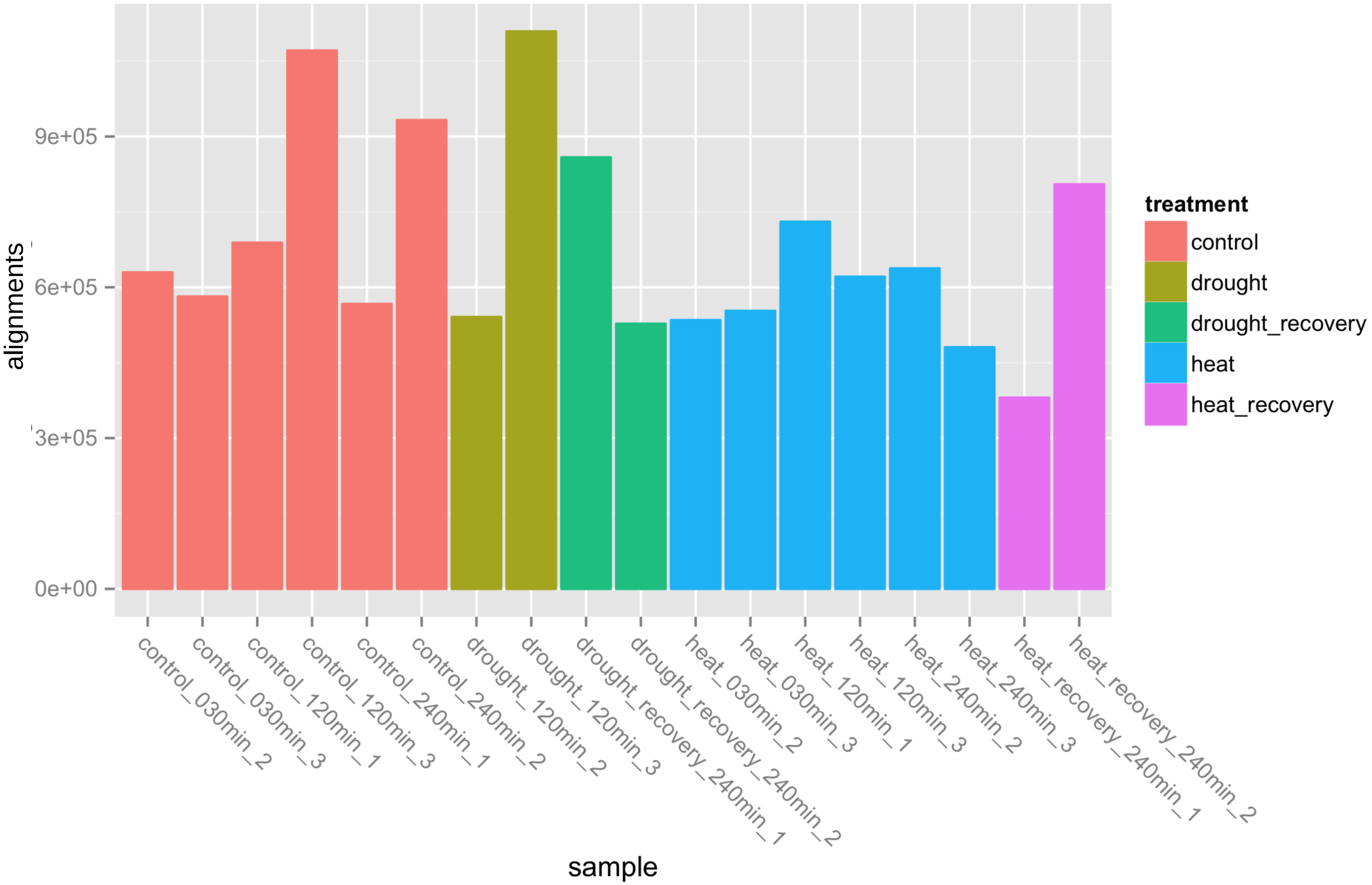


# But, many low-quality alignments

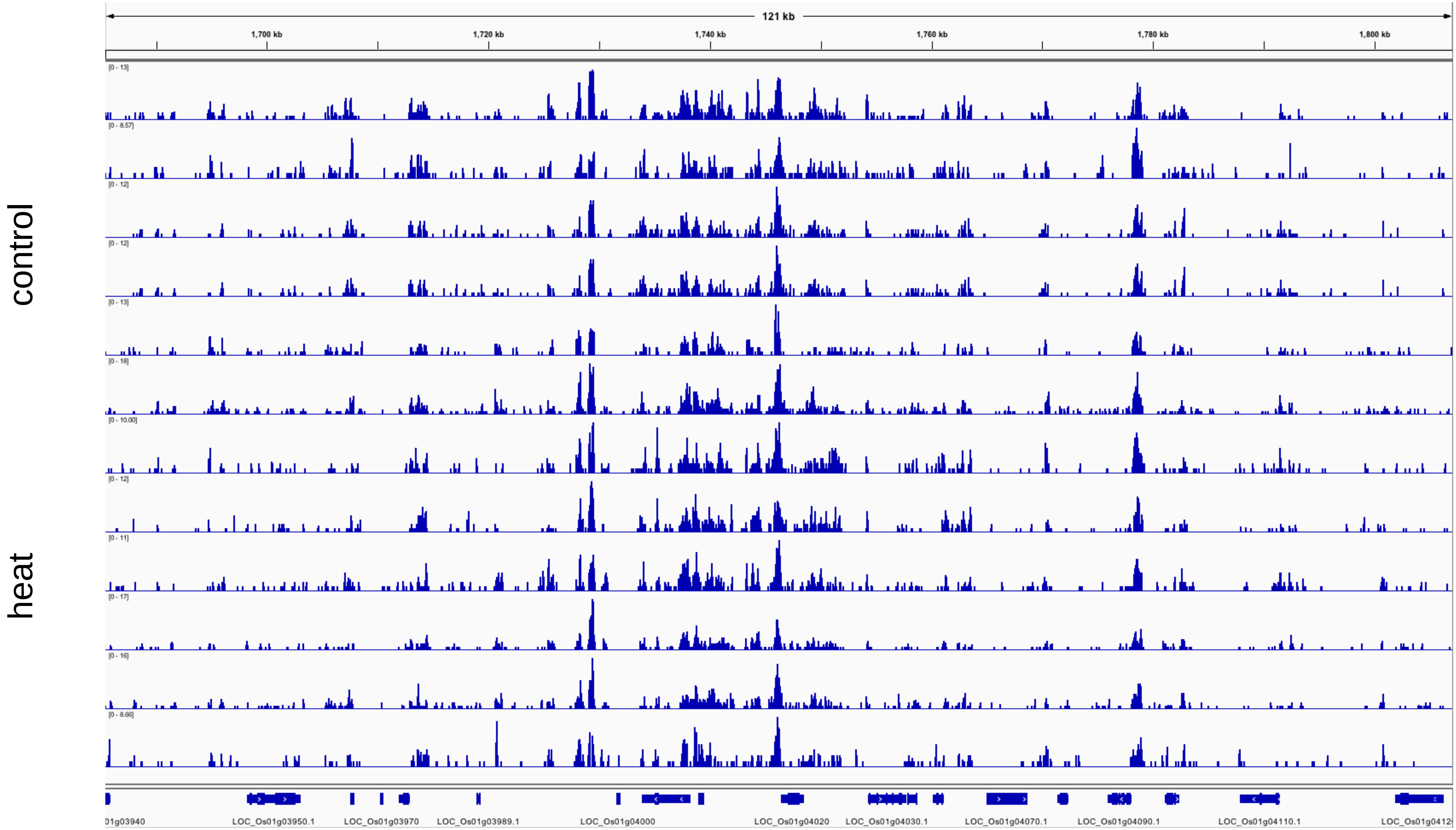


$%#@*! chloroplasts!

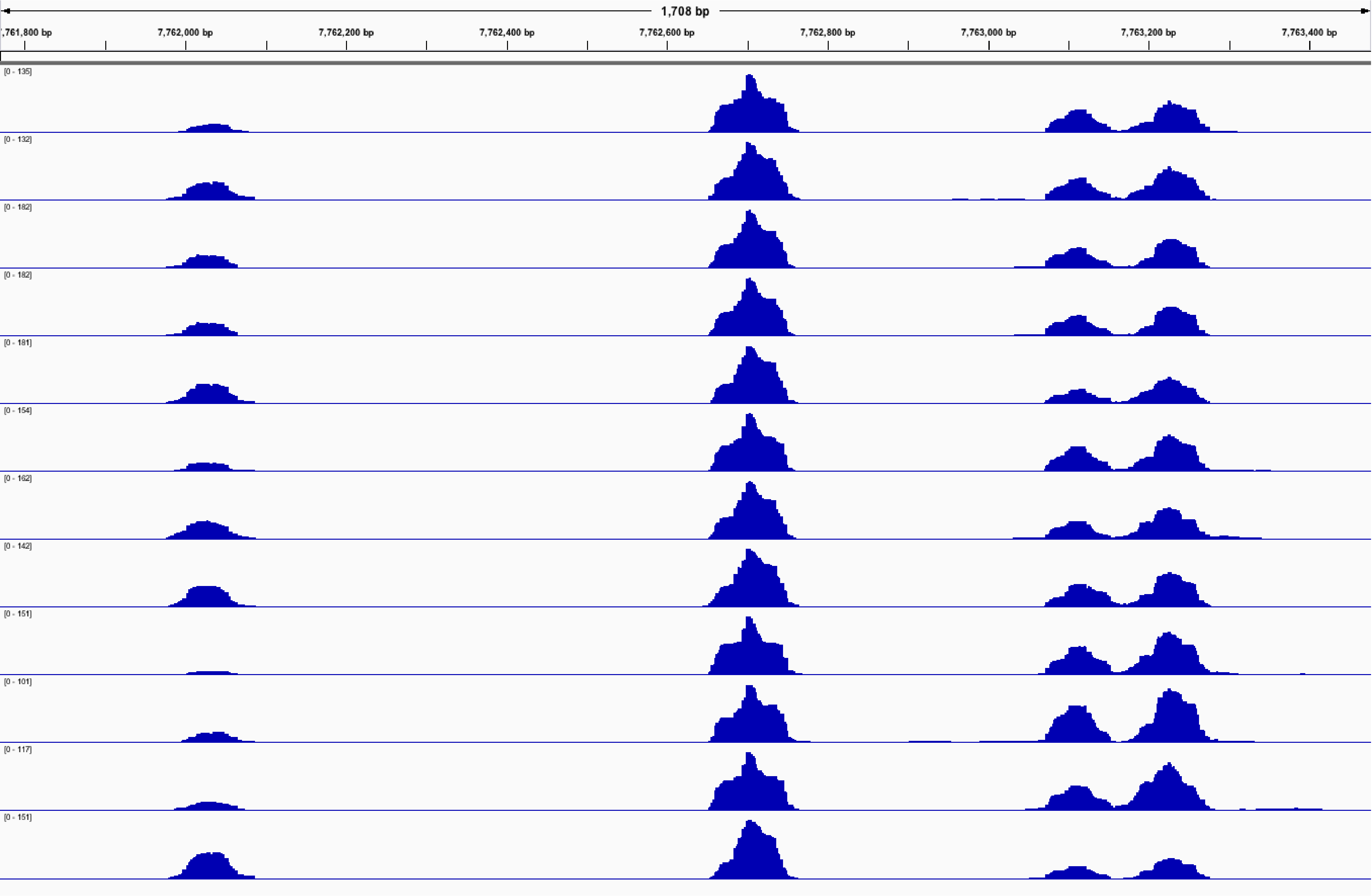
# Ca. 4.6% of alignments have q > 30



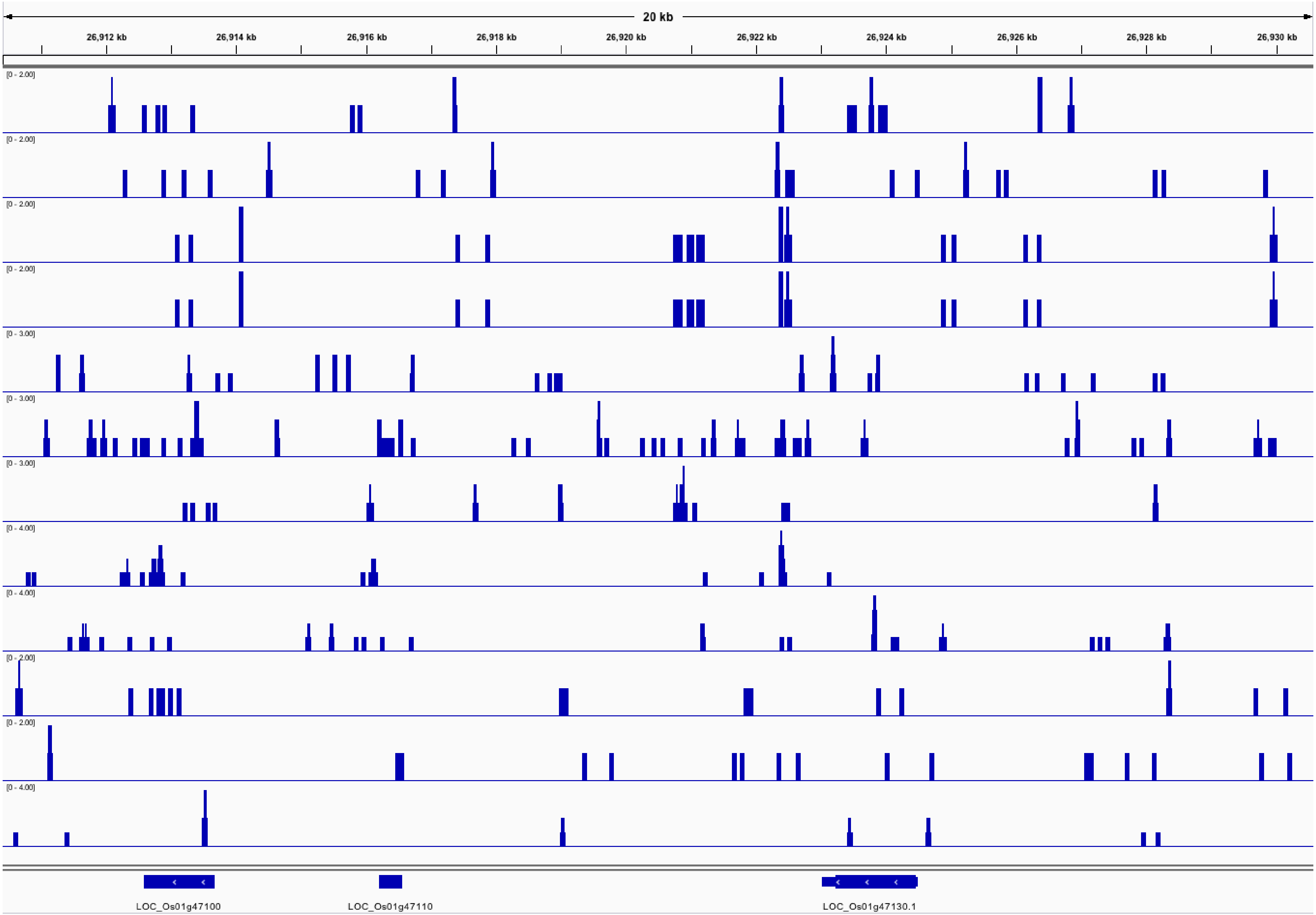
# Consistent read coverage across libraries



# Consistent read coverage across libraries



# Noise in low-coverage regions



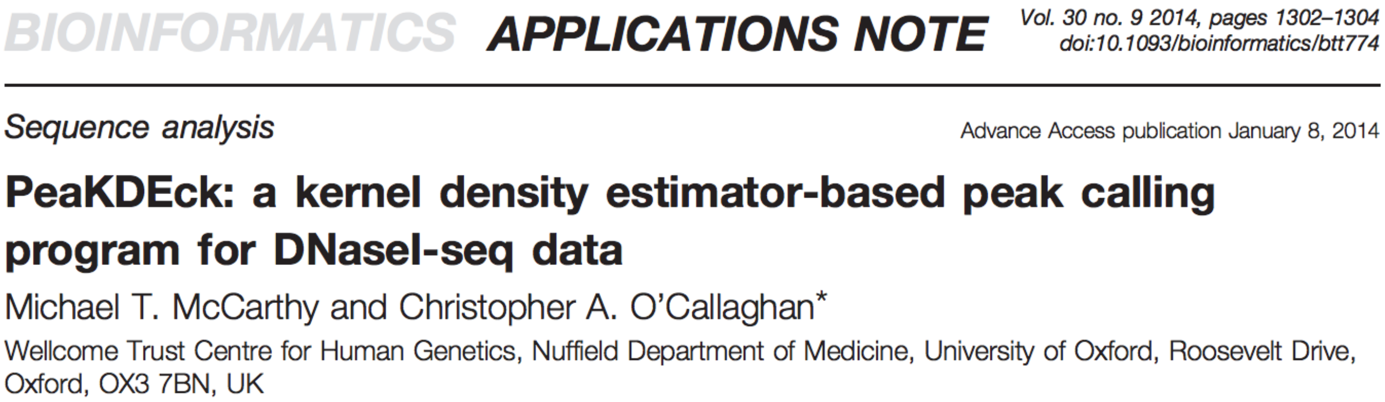
# ATAC-seq Analysis

- ~~Process the reads~~
- ~~Align the reads~~
- ➡ Call peaks
- Investigate the location of peaks

# Many options for peak calling

- PeaKDEck
- Hotspot
- F-Seq
- MACS
- ZINBA
- HOMER

# Many options for peak calling: PeaKDEck



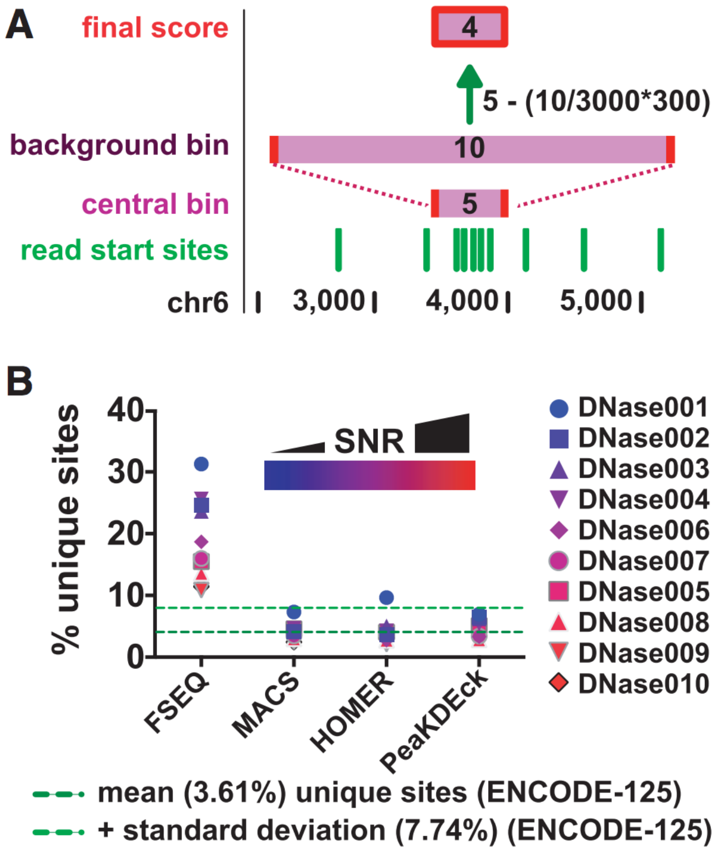
BIOINFORMATICS APPLICATIONS NOTE
Vol. 30 no. 9 2014, pages 1302–1304
doi:10.1093/bioinformatics/btt774

*Sequence analysis*

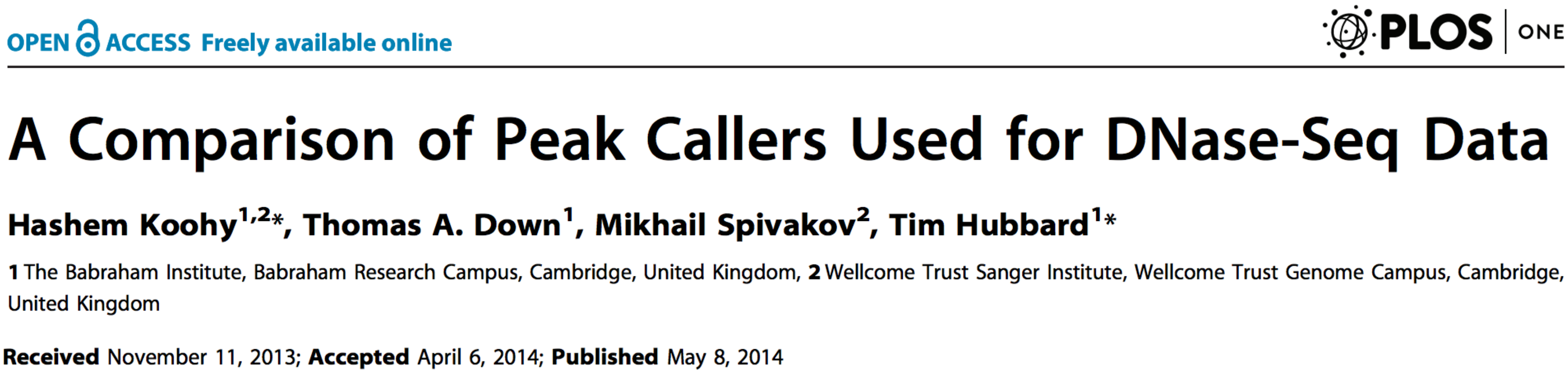
Advance Access publication January 8, 2014

## PeaKDEck: a kernel density estimator-based peak calling program for DNaseI-seq data

Michael T. McCarthy and Christopher A. O'Callaghan*
Wellcome Trust Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Roosevelt Drive,
Oxford, OX3 7BN, UK



# Many options for peak calling: Comparison paper



OPEN ACCESS Freely available online

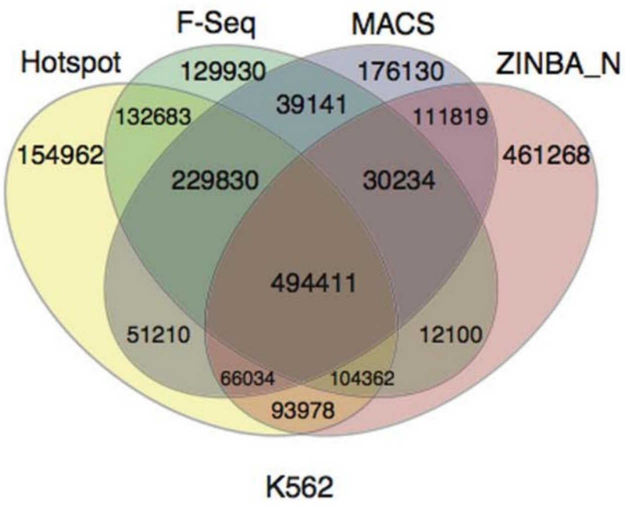
PLOS ONE

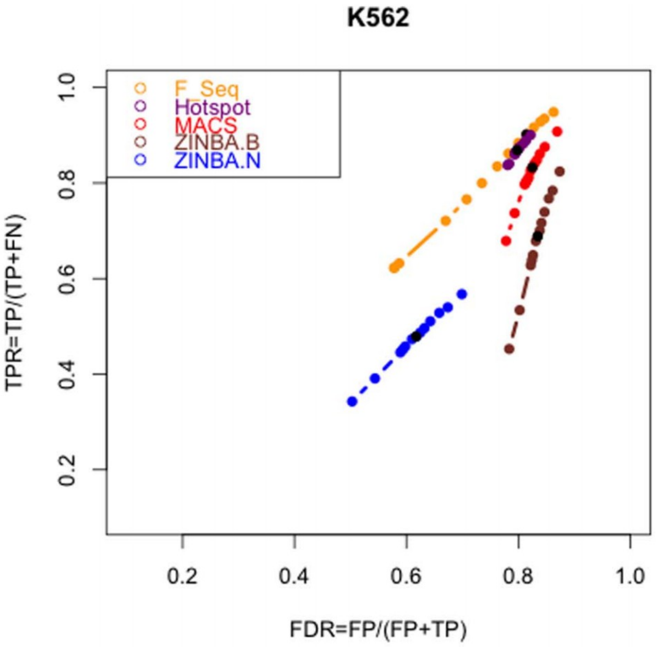
## A Comparison of Peak Callers Used for DNase-Seq Data

Hashem Koohy[1,2]*, Thomas A. Down[1], Mikhail Spivakov[2], Tim Hubbard[1]*

[1] The Babraham Institute, Babraham Research Campus, Cambridge, United Kingdom, [2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, United Kingdom

**Received** November 11, 2013; **Accepted** April 6, 2014; **Published** May 8, 2014

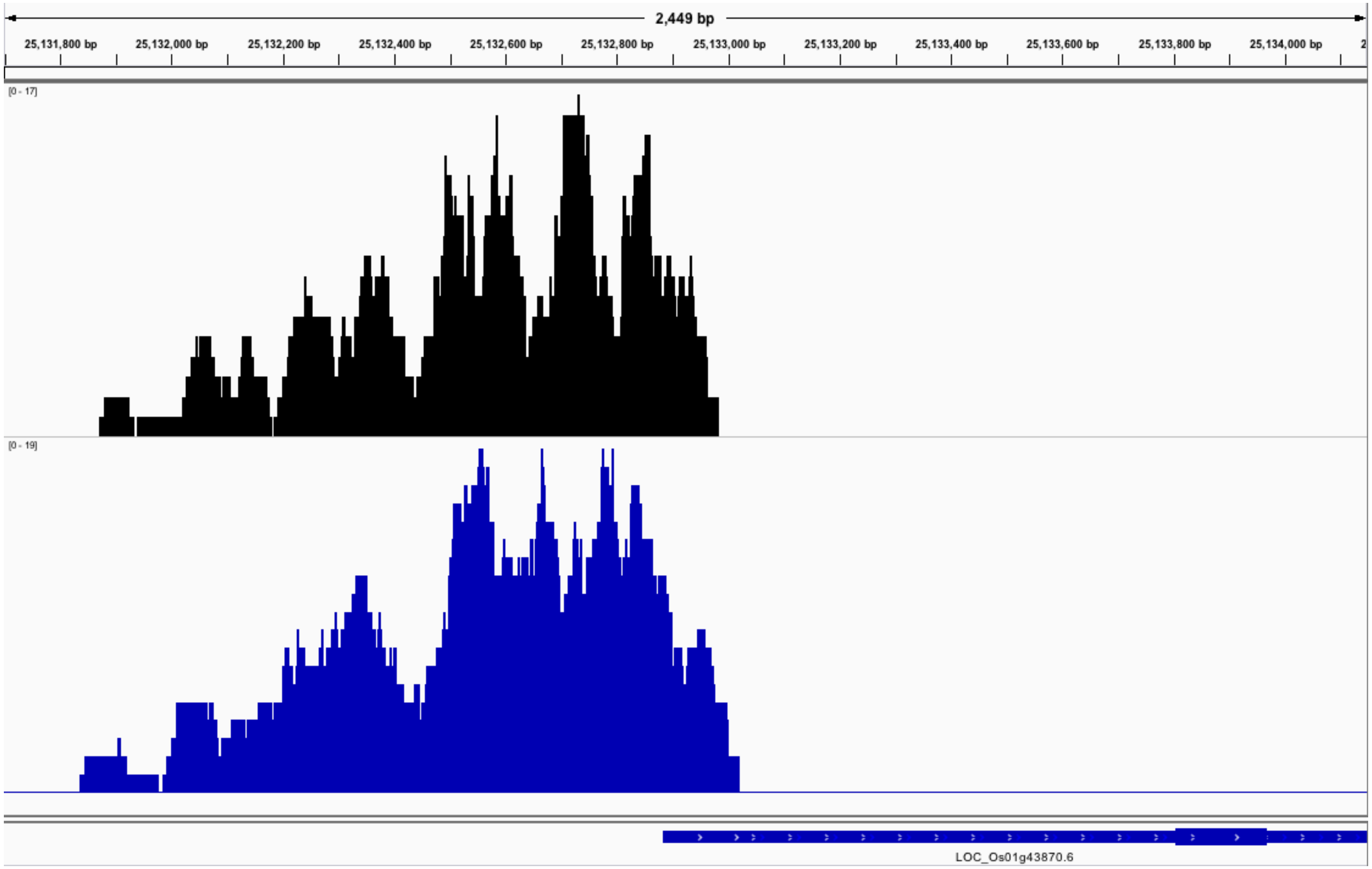




# Transform aligned reads to cut sites

Define cut site as first base of aligned fragment and first base after aligned fragment

Count number of cuts in every (overlapping) 72bp window



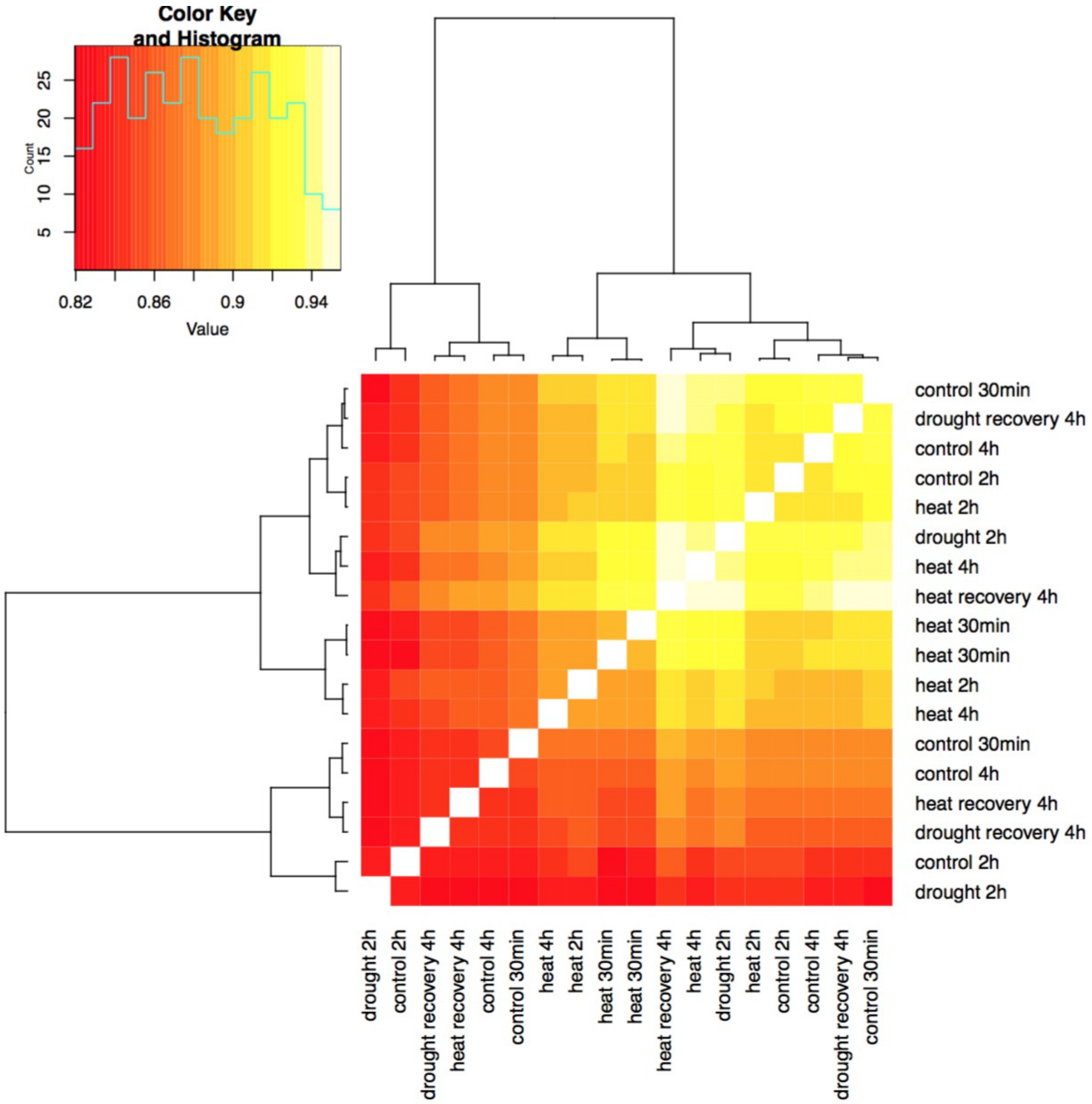
Read coverage

Cut sites

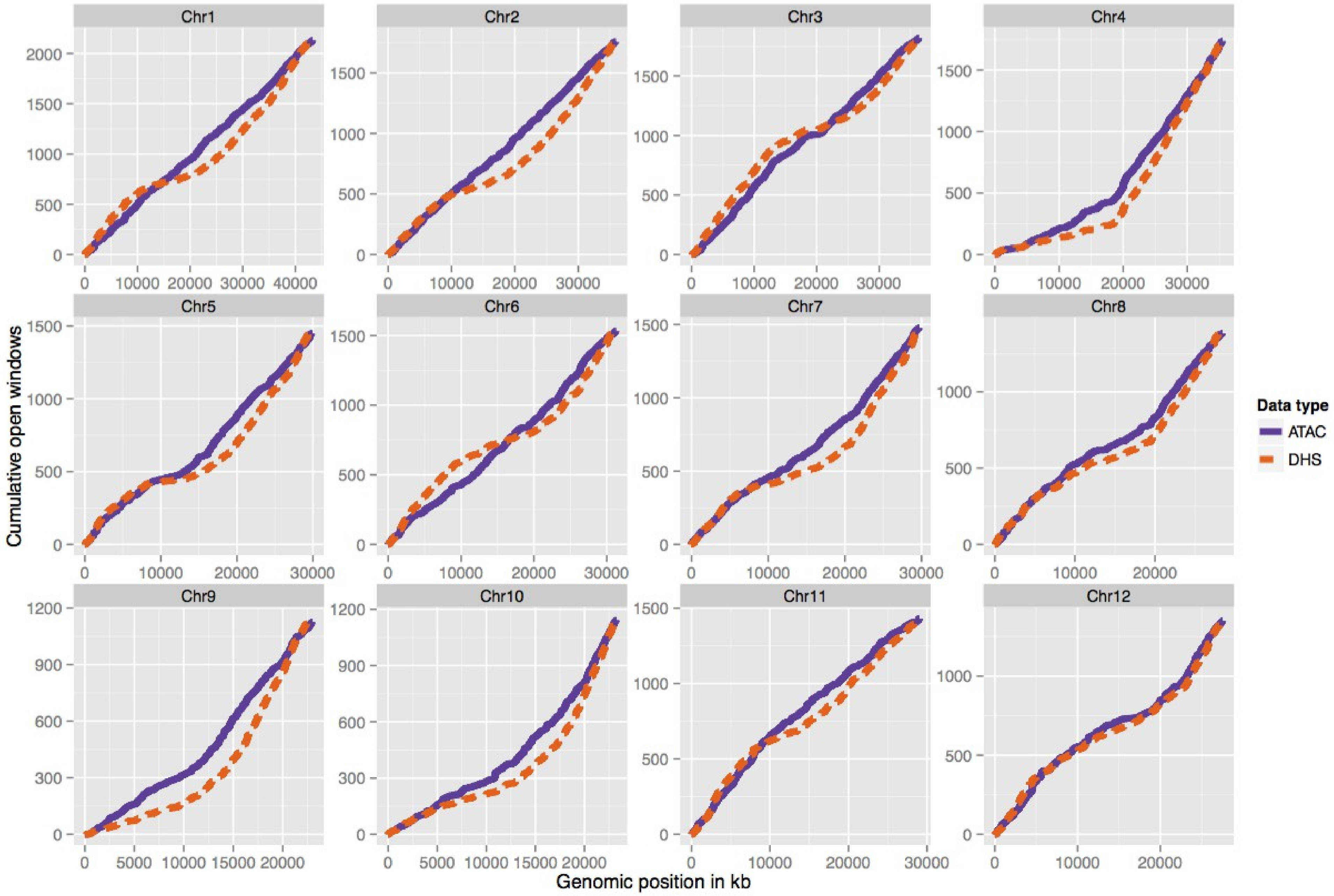
# Replicates not more similar than non-replicates

Pairwise comparison of libraries

Pearson correlation of number of cuts (log10) in all windows that had more than 1 cut in both libraries



# Global distribution of ATAC cut sites similar to DHS



# Custom peak caller based on all libraries

Ignore treatment label, detect open regions based on all 18 libraries

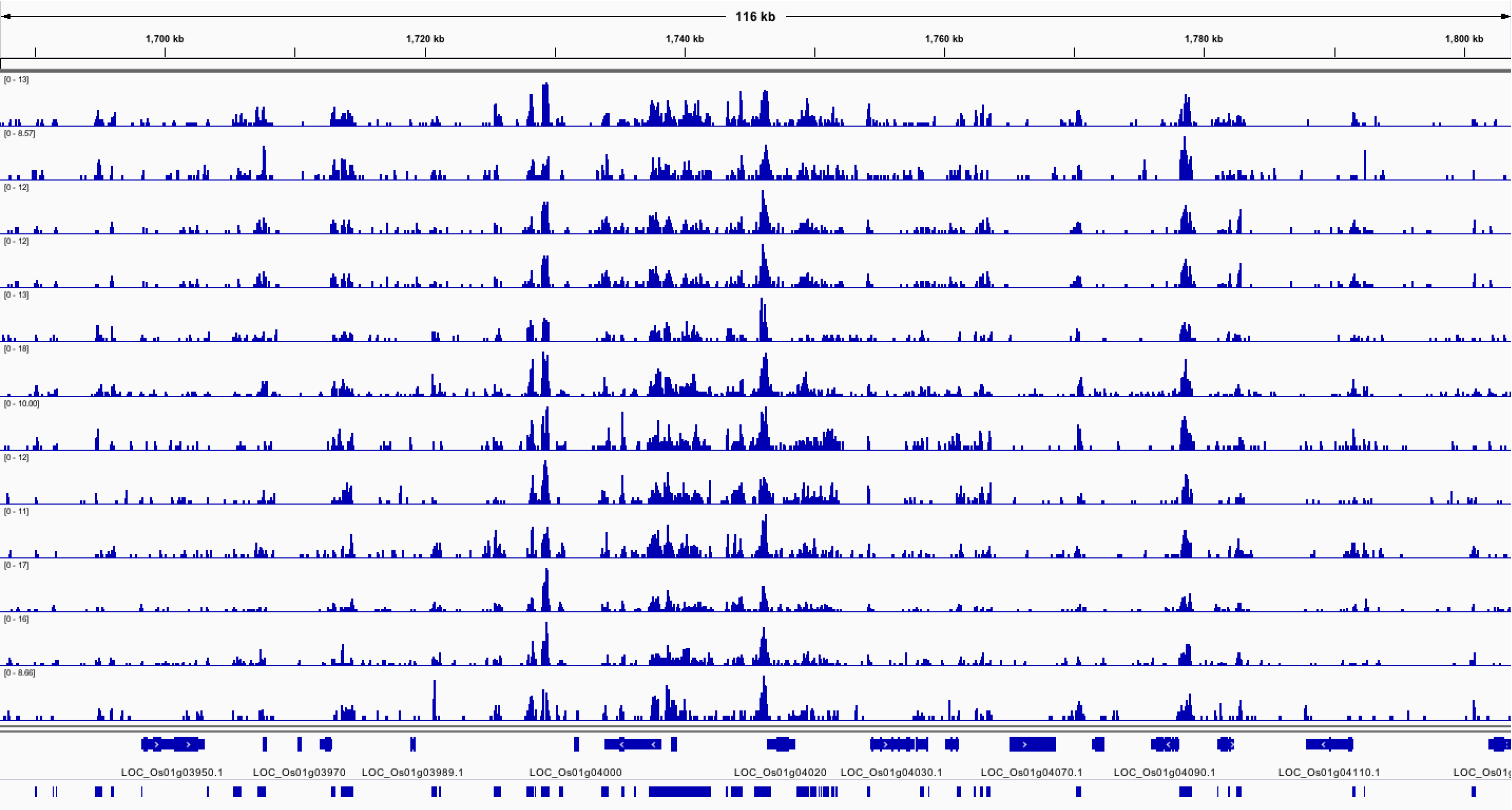
Count number of cut sites that fall into 72 bp window centered on each base

Consider base open if its window contained at least one cut site in more than half of the libraries

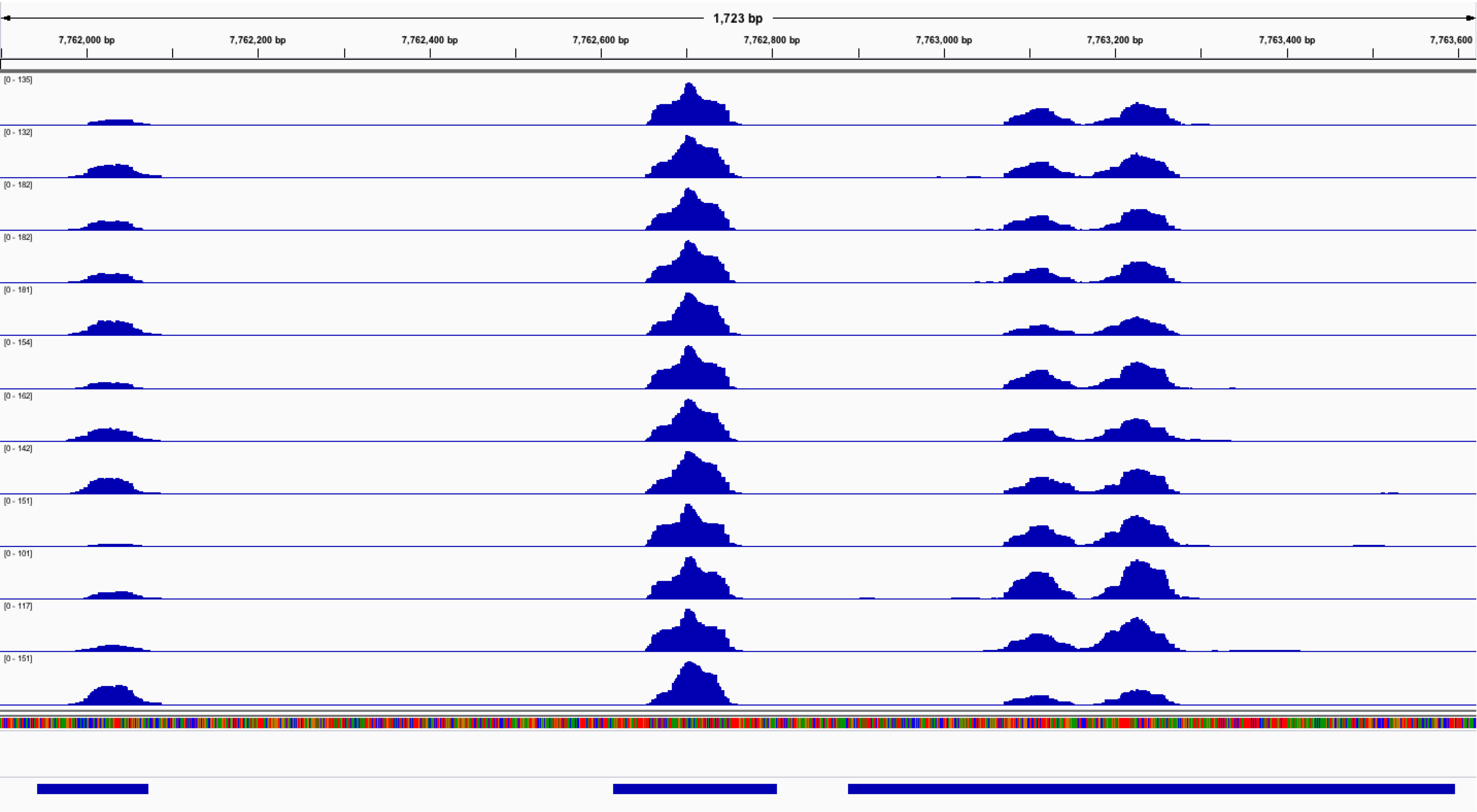
If two open bases are less than 72 bp apart, call all intermediate bases open

Tools used: R & Rsamtools library

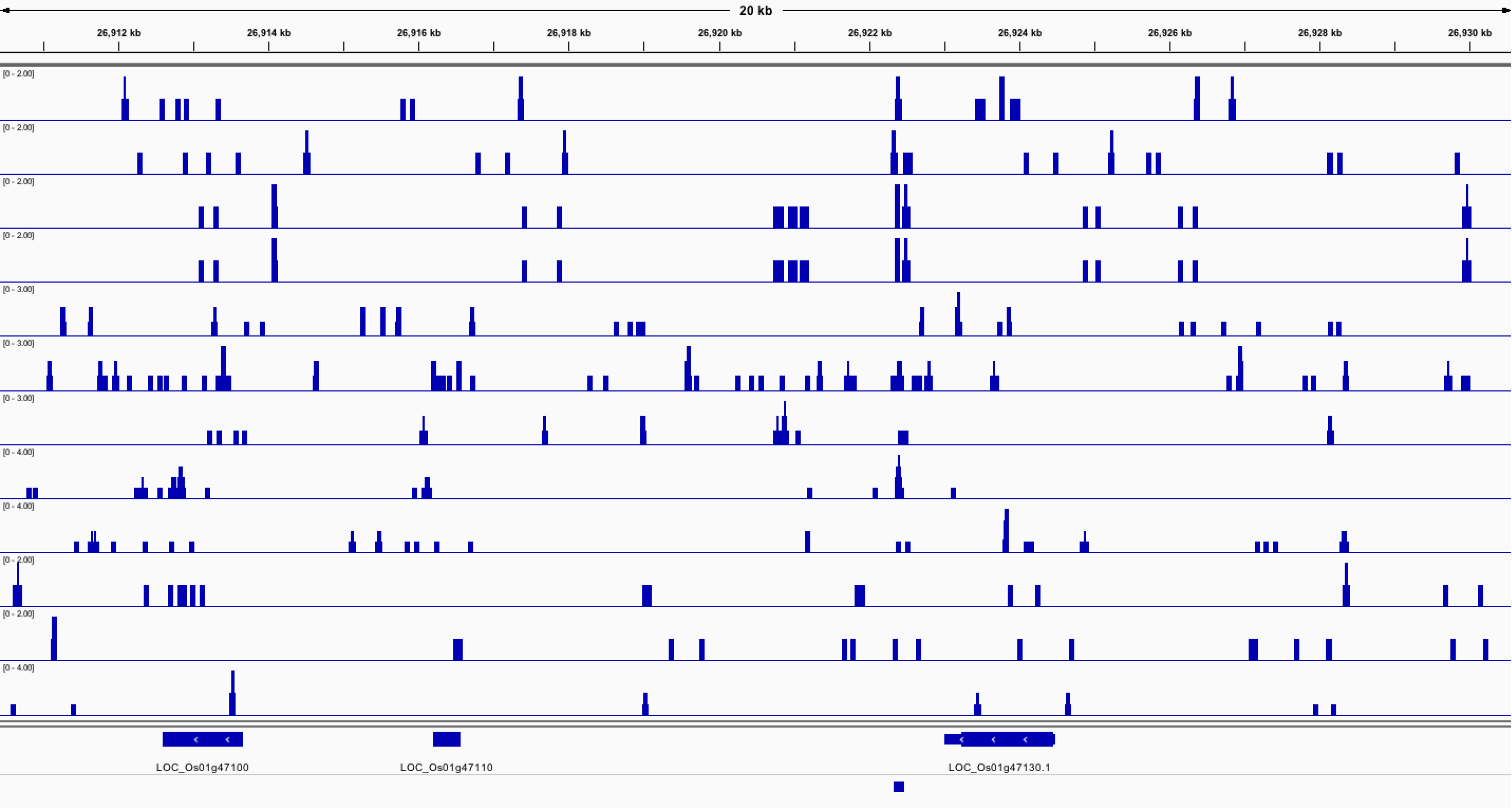
# Peaks (open regions) seem to make sense



# Peaks (open regions) seem to make sense



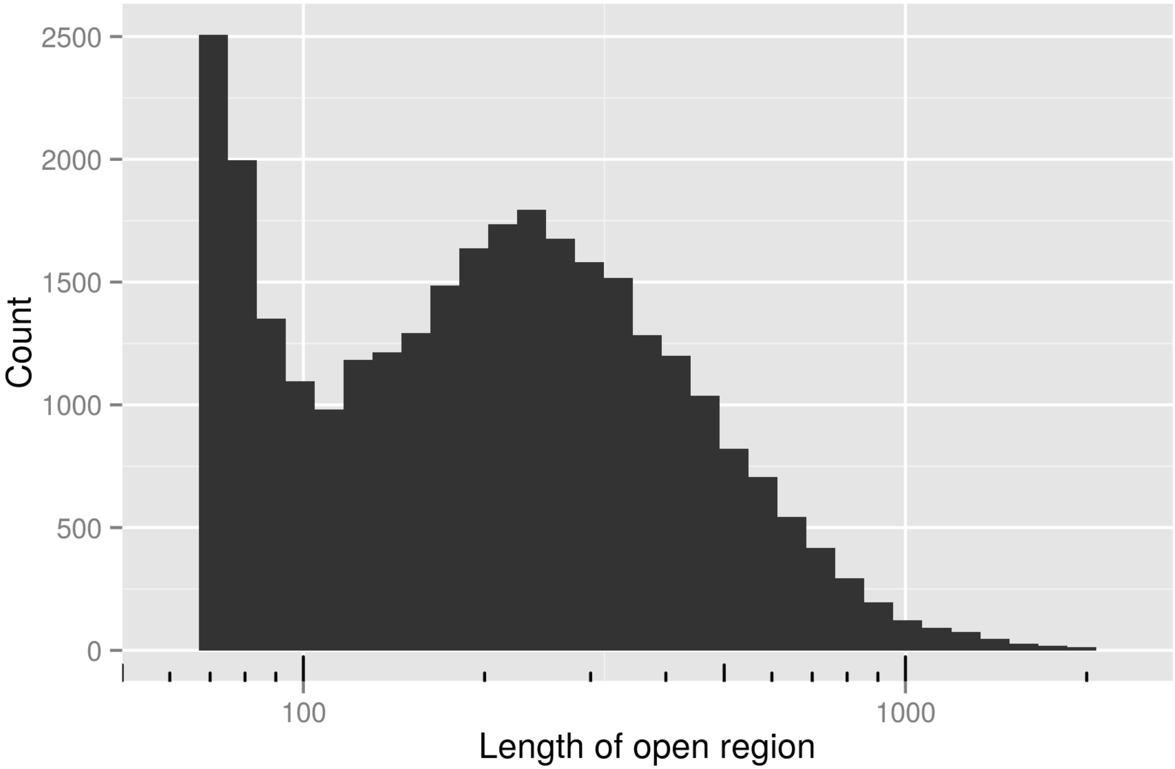
# Peaks (open regions) seem to make sense



# ATAC-seq Analysis

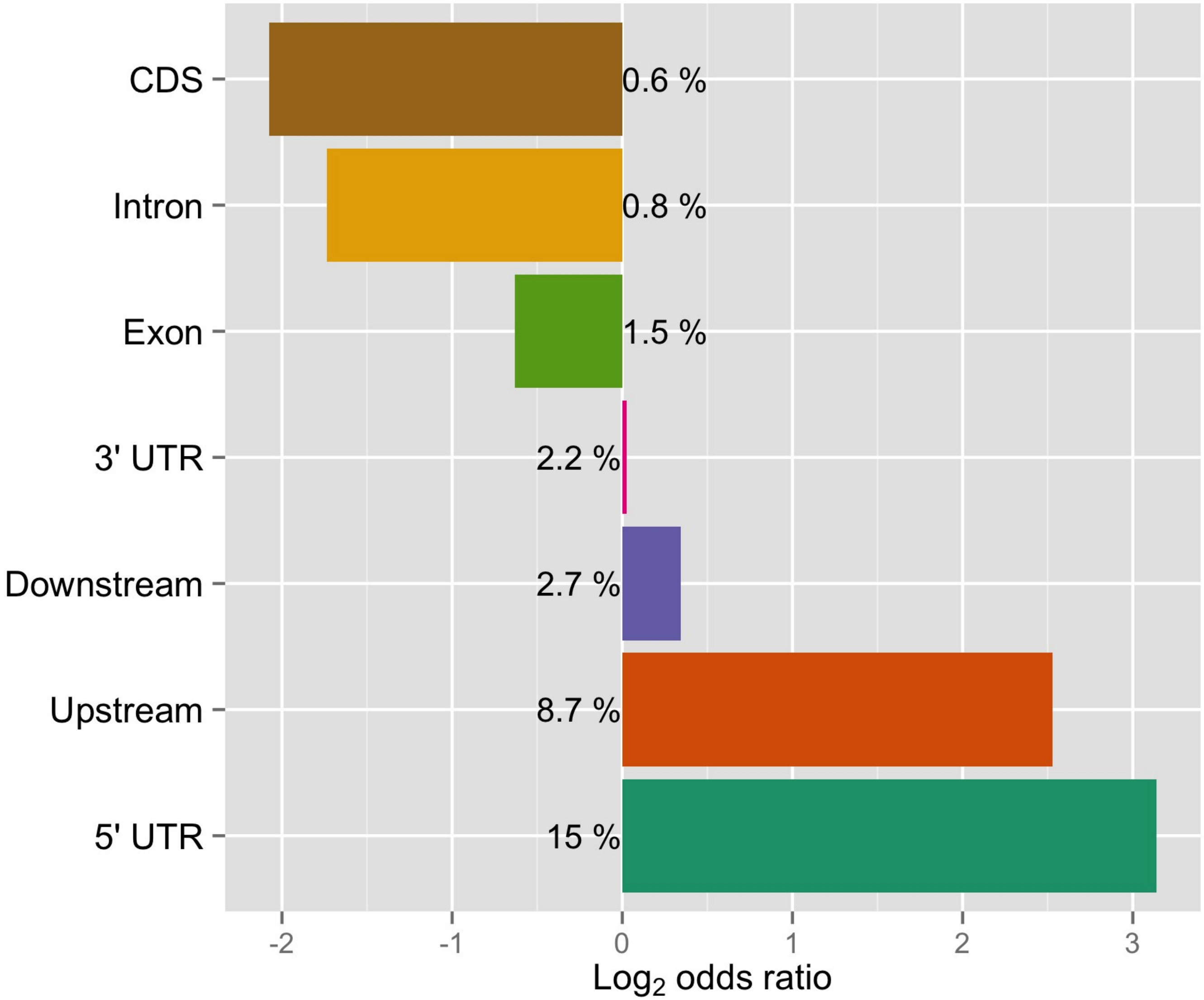
- ~~Process the reads~~
- ~~Align the reads~~
- ~~Call peaks~~
- ➡ Investigate the location of peaks

# We call 2% of the genome open

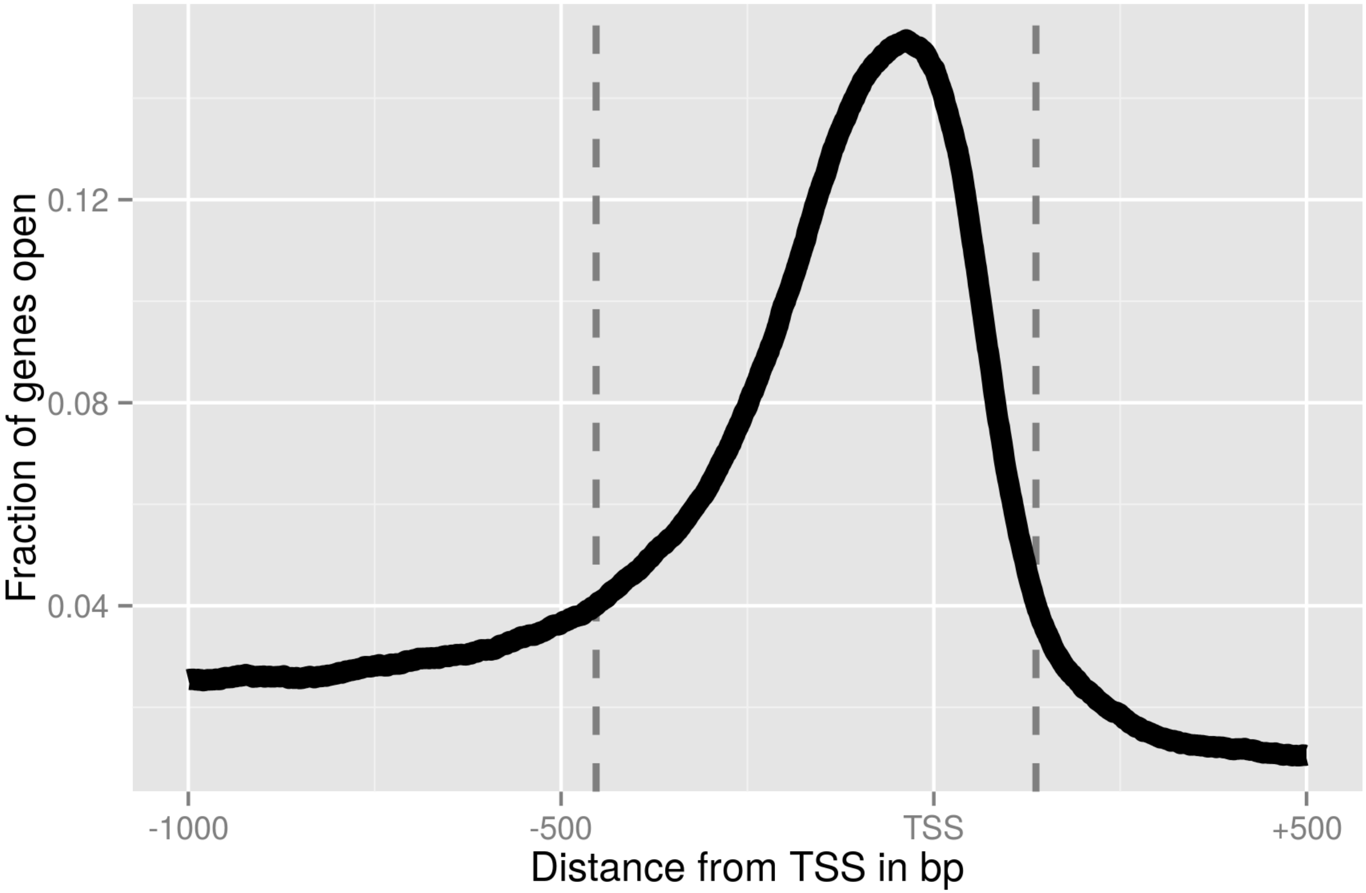


- Ca. 30k open regions
- Covering 8M base pairs
- Average length: 268 bp
- Median length: 206 bp

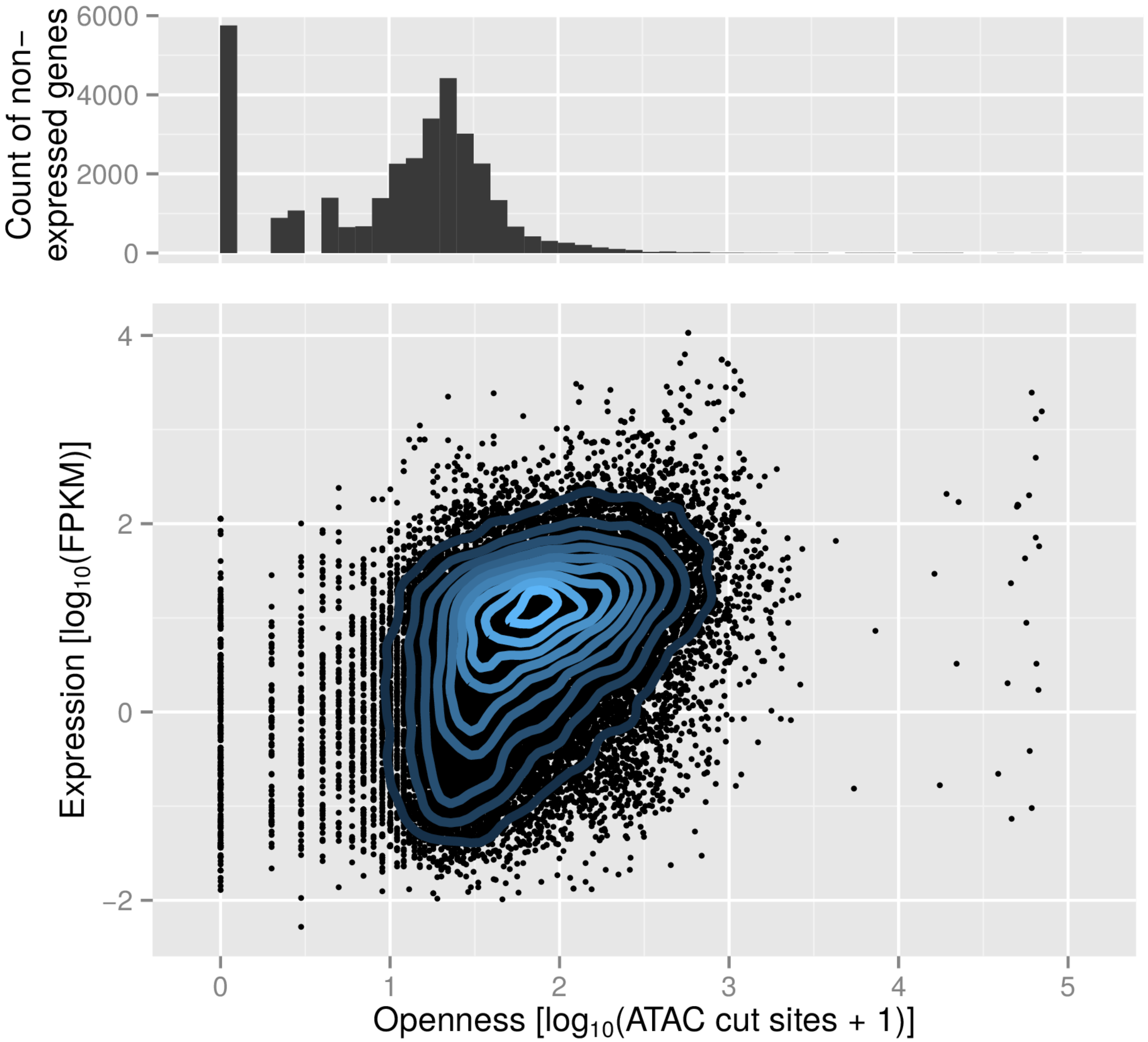
# Open regions highly enriched upstream and in 5' UTR



# Open regions density highest just before TSS



# Number of ATAC cut sites in promoter is correlated with expression


Count of non–
expressed genes
Expression [$\log_{10}$(FPKM)]
Openness [$\log_{10}$(ATAC cut sites + 1)]