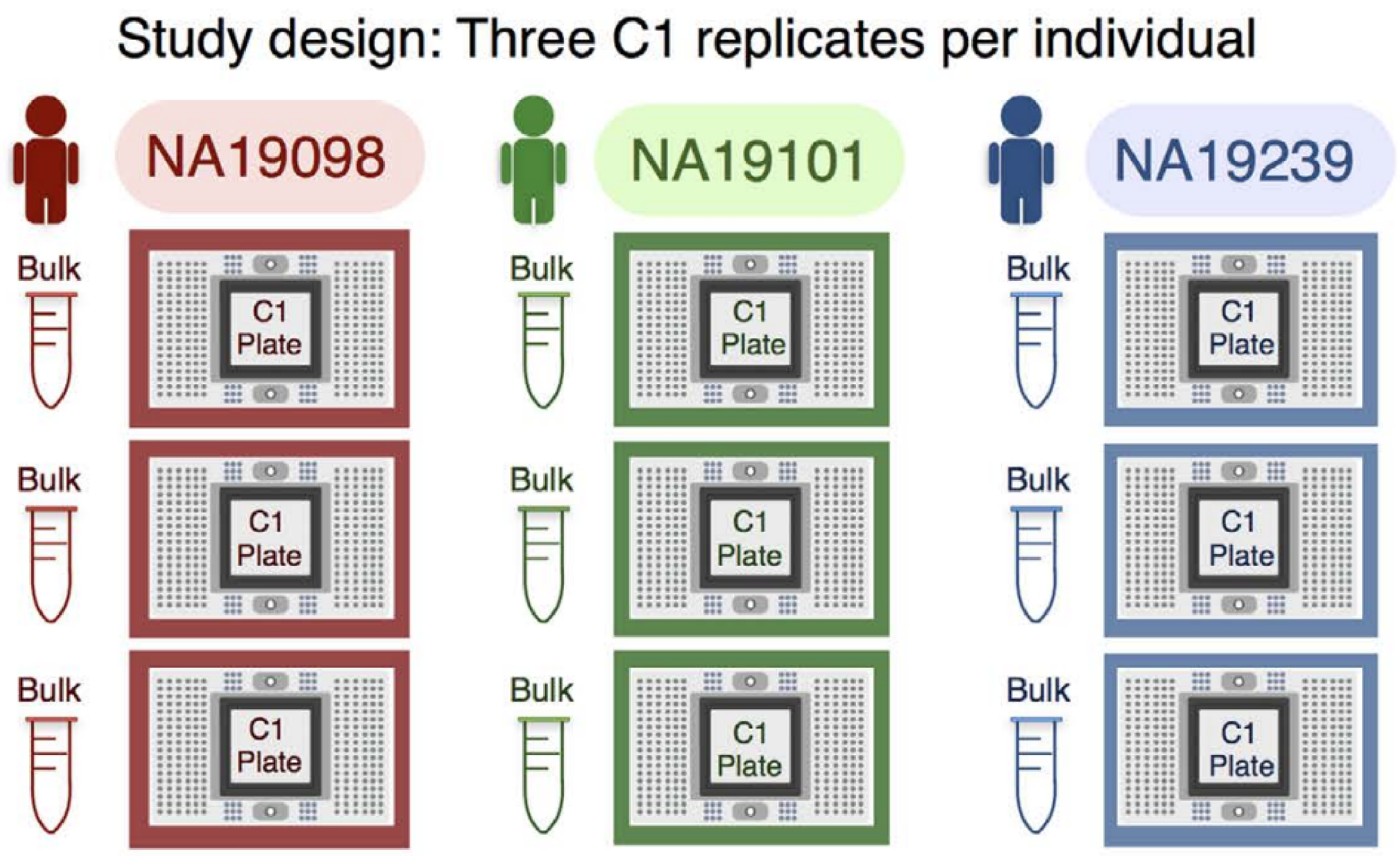
Quality Control

SINGLE-CELL RNA-SEQ WORKFLOWS IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

Tung dataset

6 RNA-sequencing datasets per individual: 3 bulk & 3 single-cell (on C1 Plates).



[1] Batch effects and the effective design of single [2] cell gene expression studies. Tung et al. Figure 1a.

Tung dataset

```
sce
```

```
class: SingleCellExperiment
dim: 18726 864
metadata(0):
assays(1): counts
rownames(18726): ENSG00000237683
  ENSG00000187634 ... ERCC-00170 ERCC-00171
rowData names(0):
colnames(864): NA19098.r1.A01 NA19098.r1.A02
  ... NA19239.r3.H11 NA19239.r3.H12
colData names(5): individual replicate well
  batch sample_id
reducedDimNames(0):
spikeNames(1): ERCC
```

Calculate quality control measures

```r
# load the scater library
library(scater)

# calculate quality control measures
sce <- calculateQCMetrics(
    sce,
    feature_controls = list(ERCC = isSpike(sce, "ERCC"))
```

- ERCC spike-in genes are used to filter out low-quality cells
- High ratio of synthetic spike-in RNAs vs endogenous RNAs means cell is likely dead or stressed

[1] Quality control with scater (Single [2] Cell Analysis Toolkit for Gene Expression Data in R): https://bioconductor.org/packages/3.9/bioc/vignettes/scater/inst/doc/vignette [3] qc.html

Functions used in exercises

- Calculate quality measures: `calculateQCMetrics()`
- Get the count matrix: `counts()`
- Find sum for each row of a matrix: `rowSums()`
- Find elements that follow a pattern: `grepl()`
- Identify spike-in genes: `isSpike()`
- Plot the distribution of `x` : `plot(density(x))`
- Add a line to a plot: `abline()`

Let's practice!

SINGLE-CELL RNA-SEQ WORKFLOWS IN R

Quality Control (continued)

SINGLE-CELL RNA-SEQ WORKFLOWS IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

Calculate quality control measures

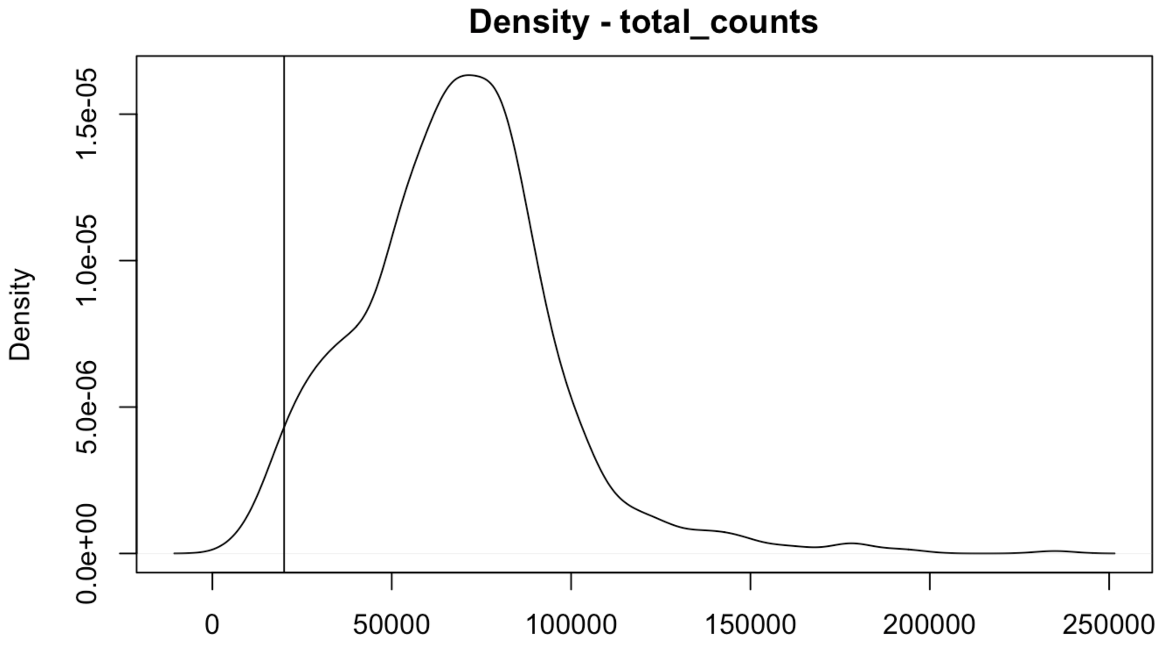
```r
library(scater)
sce <- calculateQCMetrics(
    sce,
    feature_controls = list(ERCC = isSpike(sce, "ERCC")
)
```

Cell filtering - Library size

- Total number of reads for each cell
- In `scatter` : `total_counts`
- Goal: remove cells with few reads

Cell filtering - Library size

```
# plot the density of library size and add a vertical line
plot(density(sce$total_counts), main = "Density - total_counts")
# set the threshold for minimal library size
threshold <- 20000
# plot a vertical line
abline(v = threshold)
```



Cell filtering - Library size

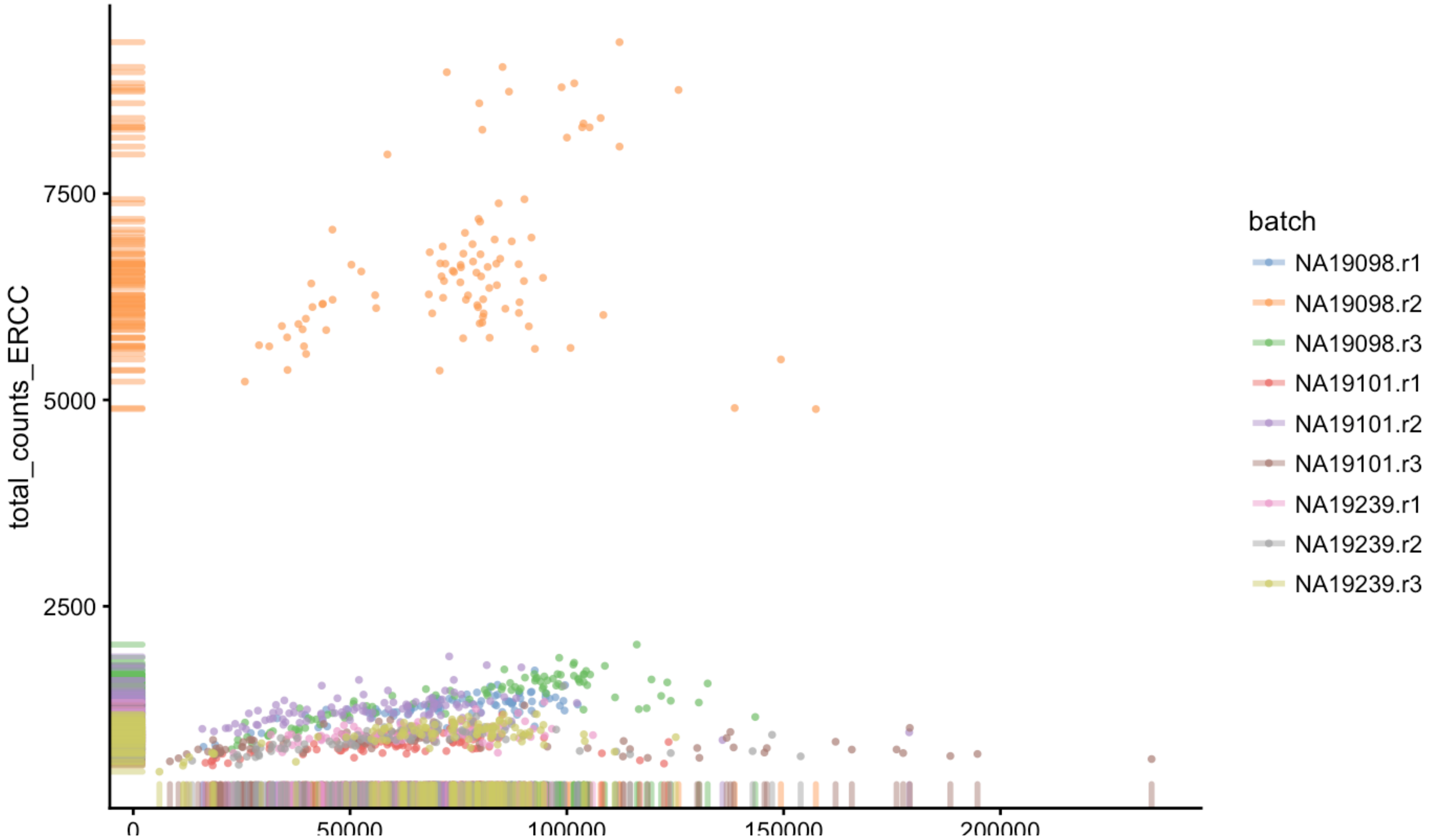
```r
# find entries in the total_counts matrix greater than threshold
keep <- (sce$total_counts > threshold)

# tabulate the keep matrix
table(keep)
```

```
keep
FALSE  TRUE
   27   837
```

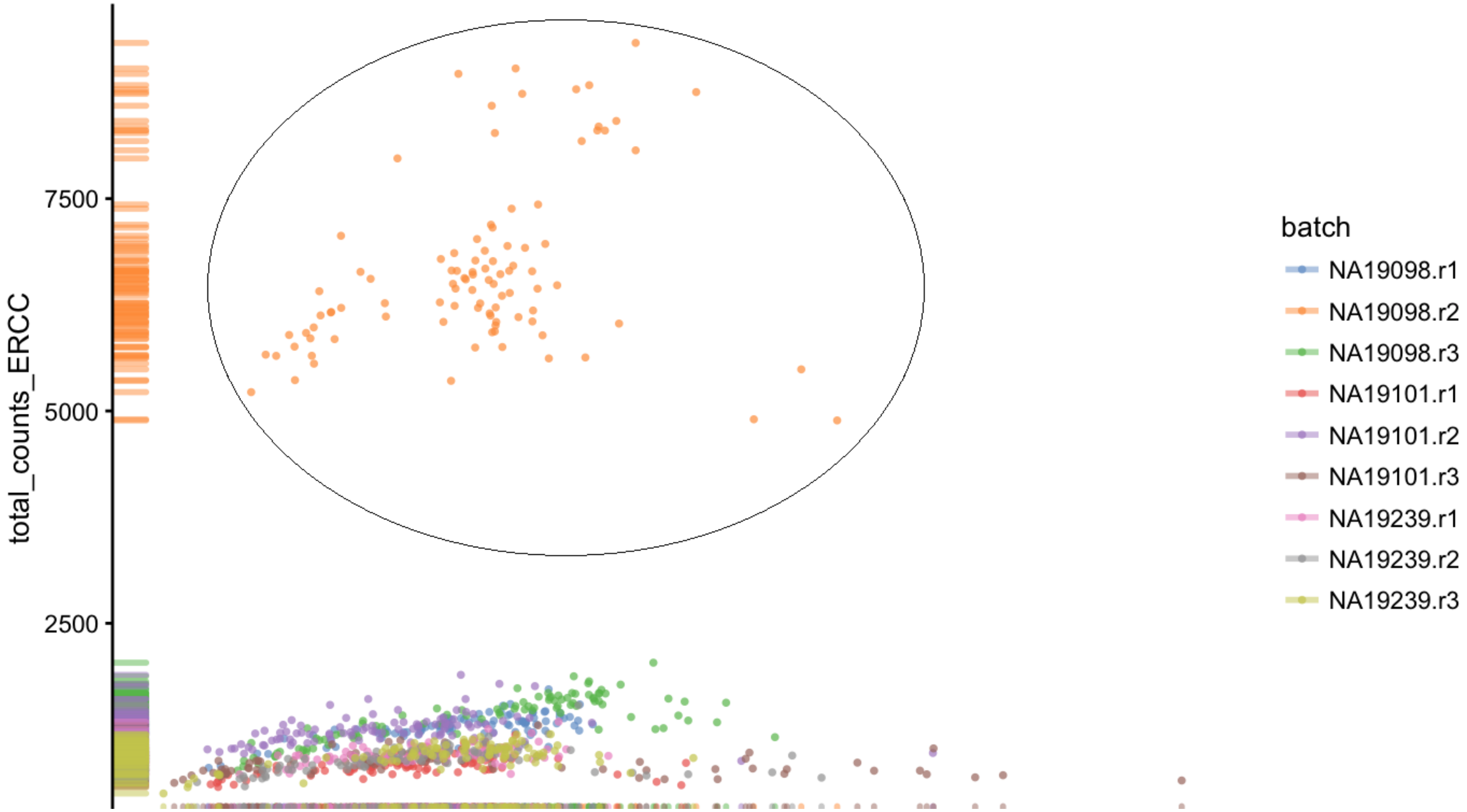
Cell filtering - Batch

```
plotPhenoData(
    sce,
    aes_string(x = "total_counts", y = "total_counts_ERCC", colour = "batch"))
```



Cell filtering - Batch

```
plotPhenoData(
    sce,
    aes_string(x = "total_counts", y = "total_counts_ERCC", colour = "batch"))
```



Cell filtering - Batch

```r
# find batches that are NOT equal to NA19098.r2
keep <- (sce$batch != "NA19098.r2")

# tabulate the keep matrix
table(keep)
```

```
keep
FALSE  TRUE
   96   768
```

Gene filtering

- remove genes mainly not expressed

```
# keep genes with counts of at least 2 in at least 2 cells
filter_genes <- apply(counts(sce), 1, function(x) length(x[x >= 2] >= 2)

# tabulate filter_genes
table(filter_genes)
```

```
filter_genes
FALSE  TRUE
 4512 14214
```

  - performed after cell filtering

Let's practice!

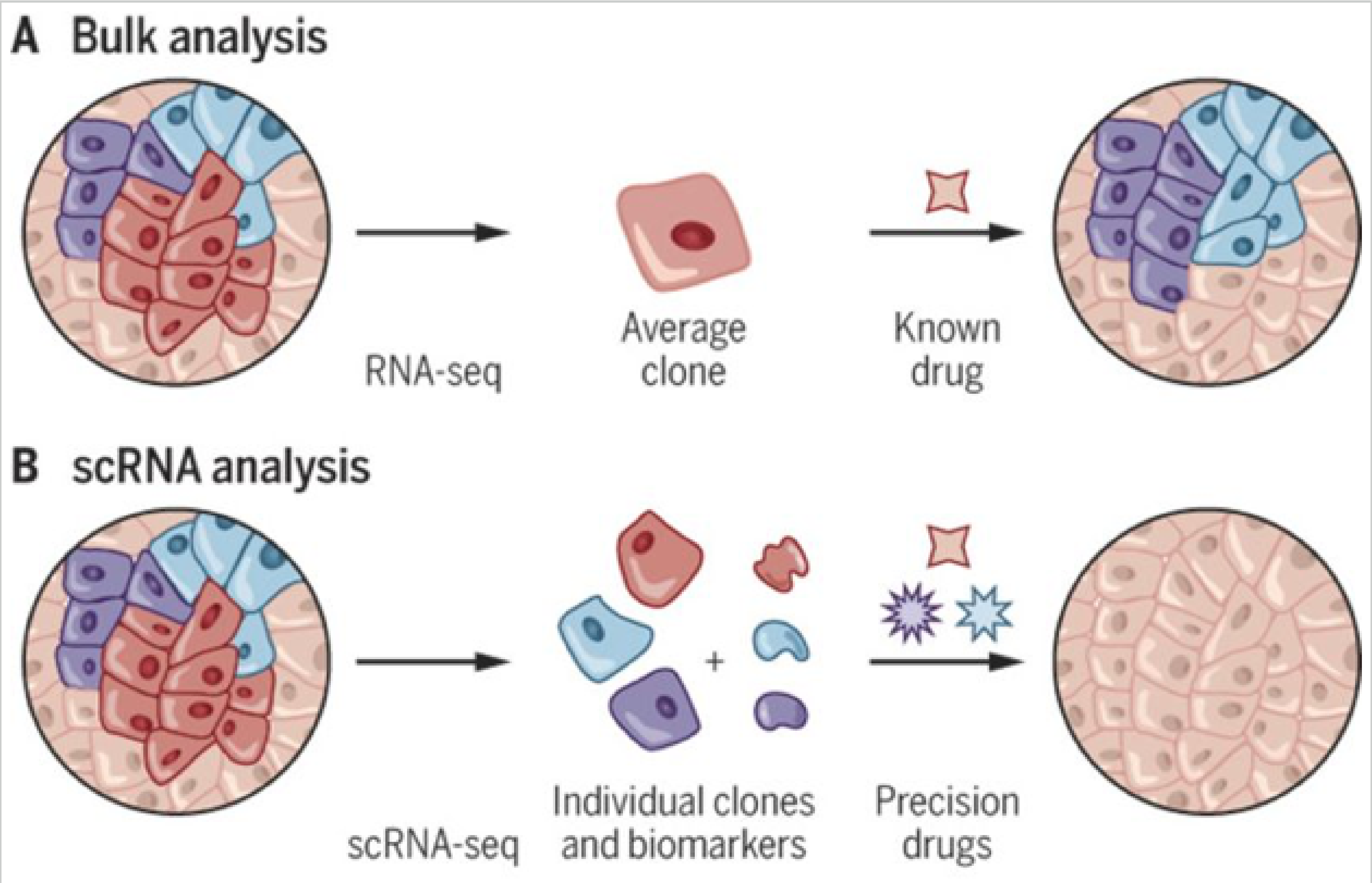
SINGLE-CELL RNA-SEQ WORKFLOWS IN R

Normalization

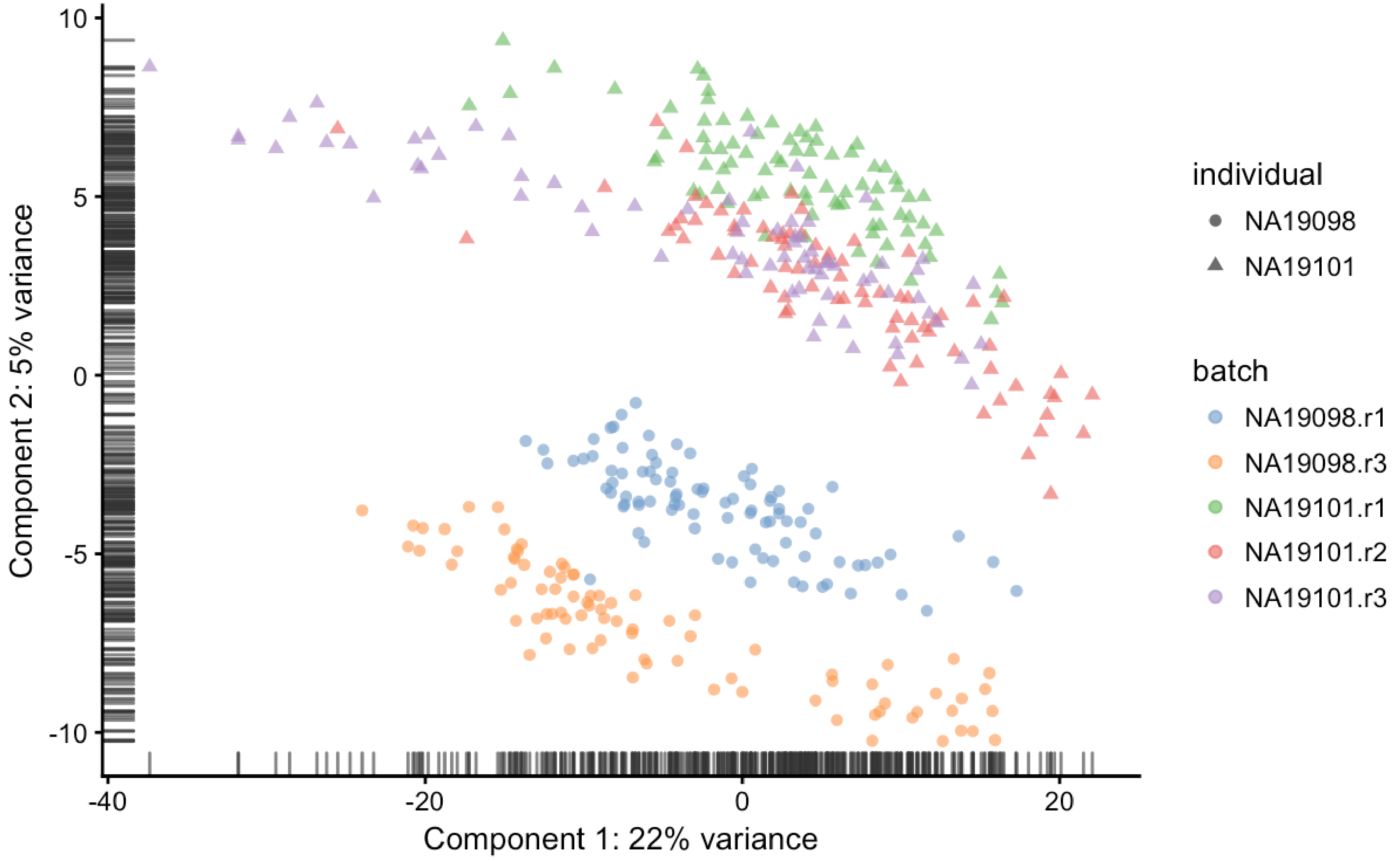
SINGLE-CELL RNA-SEQ WORKFLOWS IN R

Fanny Perraudeau
Senior Data Scientist, Whole Biome

Biological and technical variation



Batch effect



Clustering by batch - undesired technical artifact

Goal of normalization

- remove technical variation (e.g. batch effect)
- ...while preserving biological variation

Normalization methods

- Normalizing by dividing by normalization factor
  - Library size
  - Counts per million (CPM)
- Other common scaling factors
  - Weighted trimmed mean of M-values (TMM) in edgeR
  - DESeq scaling factors
  - Scaling factors accounting for zero inflation in scran

[1] "Normalizing single [2] cell RNA sequencing data Challenges and opportunities" (Vallejos et al 2017)

Functions used in exercises

- Plot principal components: `plotPCA()`
- Get first two principal components: `reducedDim(sce, “PCA”)[, 1:2]`
- Calculate and get the size factors: `computeSumFactors()` , `sizeFactors()`
- Names of the matrices stored in an SCE: `assays()`
- Normalize counts: `normalize()`
- Plot the relative log expression: `plotRLE()`

Let's practice!

SINGLE-CELL RNA-SEQ WORKFLOWS IN R