

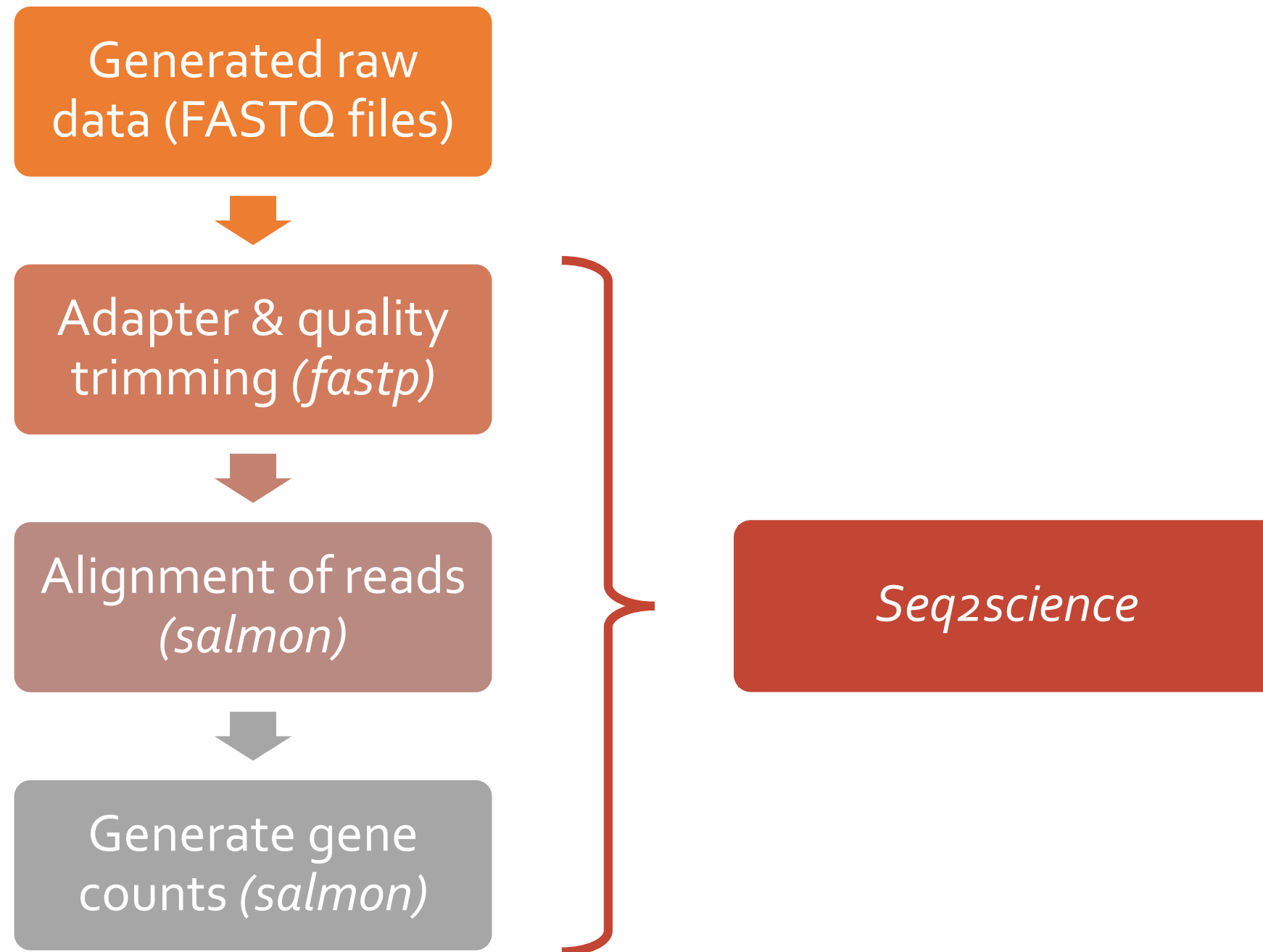
MultiQC: RNA-seq data quality reports

03-10-2022

Dr. Janou Roubroeks

RNA-seq data analysis

Steps taken so far:



seq2science

install with [bioconda](#) [Anaconda.org](#) 0.9.5 downloads 21k stars 107 [Test Status](#) [docs](#) [passing](#)

DOI [10.5281/zenodo.7040780](#)

Seq2science is the attempt of the *van heeringen lab* to generate a collection of generic pipelines/workflows which can be used by complete beginners to bioinformatics and experienced bioinformaticians alike. Please take a look at our [docs](#) for help with installation, how to run it, and best practices.



Our supported workflows:

- Downloading of fastq
- Alignment
- ATAC-seq
- RNA-seq
- ChIP-seq
- scATAC-seq
- scRNA-seq

Performs all of these steps (and many more!)
in a reproducible manner

Seq2science

Samples.tsv

sample	assembly	technical_replicates	descriptive_name
Sample1	GRCh38	Sample1	Heart
Sample1_2	GRCh38	Sample1	Heart
Sample2	GRCh38	Sample2	Lung
Sample2_2	GRCh38	Sample2	Lung
Sample3	GRCh38	Sample3	Brain
Sample3_2	GRCh38	Sample3	Brain

&

config.yaml

```
# tab-separated file of the samples
samples: samples.tsv

# pipeline file locations
result_dir: ./results # where to store results
genome_dir: ../genome # where to look for or download the genomes
fastq_dir: ../fastqs_students # where to look for or download the fastqs

# contact info for multiqc report and trackhub
email:

# produce a UCSC trackhub?
create_trackhub: true
```



Seq2science



Results

Great! What could possibly go wrong?

Potential problems:

Already taken care of with fastp:

1. Removed adapter reads

Normal read:



Adapter read:



2. Removed reads with many low confidence bases (Ns)
3. Removed reads with low quality scores

```
@NS500173:815:HKN2YBGXL:1:11101:25366:1020 1:N:0:CAACCACA  
TTCCTNCTCACTCCTGGCTTCCCACCCCTTTGTGGGAANNNN  
+  
AAAAA#EEEEEE<EEEEAEAEAEAEAEAE/EEEEEEEE####
```

The sequence "AANNNN" in the second line is circled in red, with a red arrow pointing to the number 2. The quality string "####" in the third line has a red arrow pointing to the number 3.

Other issues to look out for:

Duplicates: reads that maps to the same gene as other reads

- Natural: the same RNA fragment occurs and is sequenced twice
- Artificial: during sequencing a copy of the same read is created and sequenced

Some duplication is expected, but very high rates of duplication may be related to:

- Low starting material in PCR amplification
- Large variance of fragment size → overrepresentation of smaller fragments

→ Duplicates are usually not removed, especially not when there is sufficient library complexity

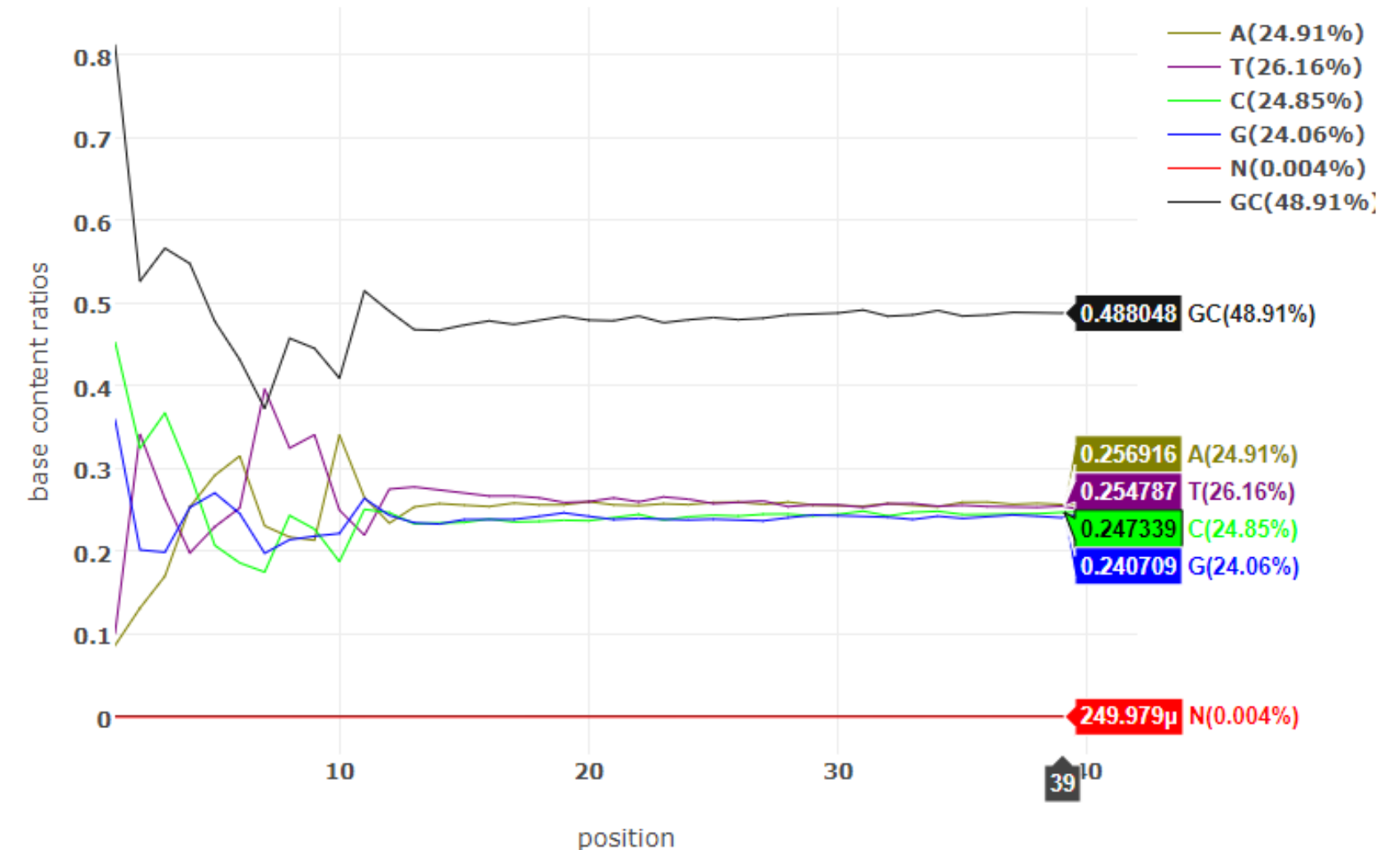
Other issues to look out for:

Sequence specific bias: random primers used during library prep do not have equal binding efficiencies to each read, regardless of their nucleotide sequence

- For each base across all reads, the percentages of A, C, G, and T should be roughly equal
 - Due to the primers, some bias is expected in the first 10-12 bases

GC content: check whether the GC content matches that expected in your organism

- Generally between 41-50% for humans



Data Quality Control (QC)

For each processing step that seq2science runs, it generates reports on the quality of your data.

- **MultiQC (also run with seq2science) aggregates results from all these reports and creates a clear overview of all QC metrics**
 - + Interactive
 - + Highly customisable

Example report

How the report was generated, when, and based on what data

MultiQC
v1.11

General Stats

Workflow explanation

fastp

Filtered Reads

Duplication Rates

Insert Sizes

Sequence Quality

GC Content

N content

Picard

MultiQC

These samples were run by [seq2science](#) v0.5.4, a tool for easy preprocessing of NGS data.

Take a look at our [docs](#) for info about how to use this report to the fullest.

Workflow	rna-seq
Date	September 29, 2021
Project	ghe_2021
Contact E-mail	

Report generated on 2021-09-29, 09:30 based on data in:

Sections of your report, matching the modules run using seq2science which produce QC metrics. The 'Workflow explanation' section tells you which modules were used.

General statistics

Shows key statistics from all modules in a single table

General Statistics

Copy table | Configure Columns | Plot

Showing 4/4 rows and 3/5 columns

Sample Name	% Dups	% GC	M Seqs
ERR458493	32.3%	43%	1.1
ERR458493.qc	32.3%	43%	1.1
ERR458500	32.5%	42%	1.9
ERR458500.qc	32.5%	42%	1.9

FastQC: Average % GC Content

Tells you which module generated this QC metric

Toolbox

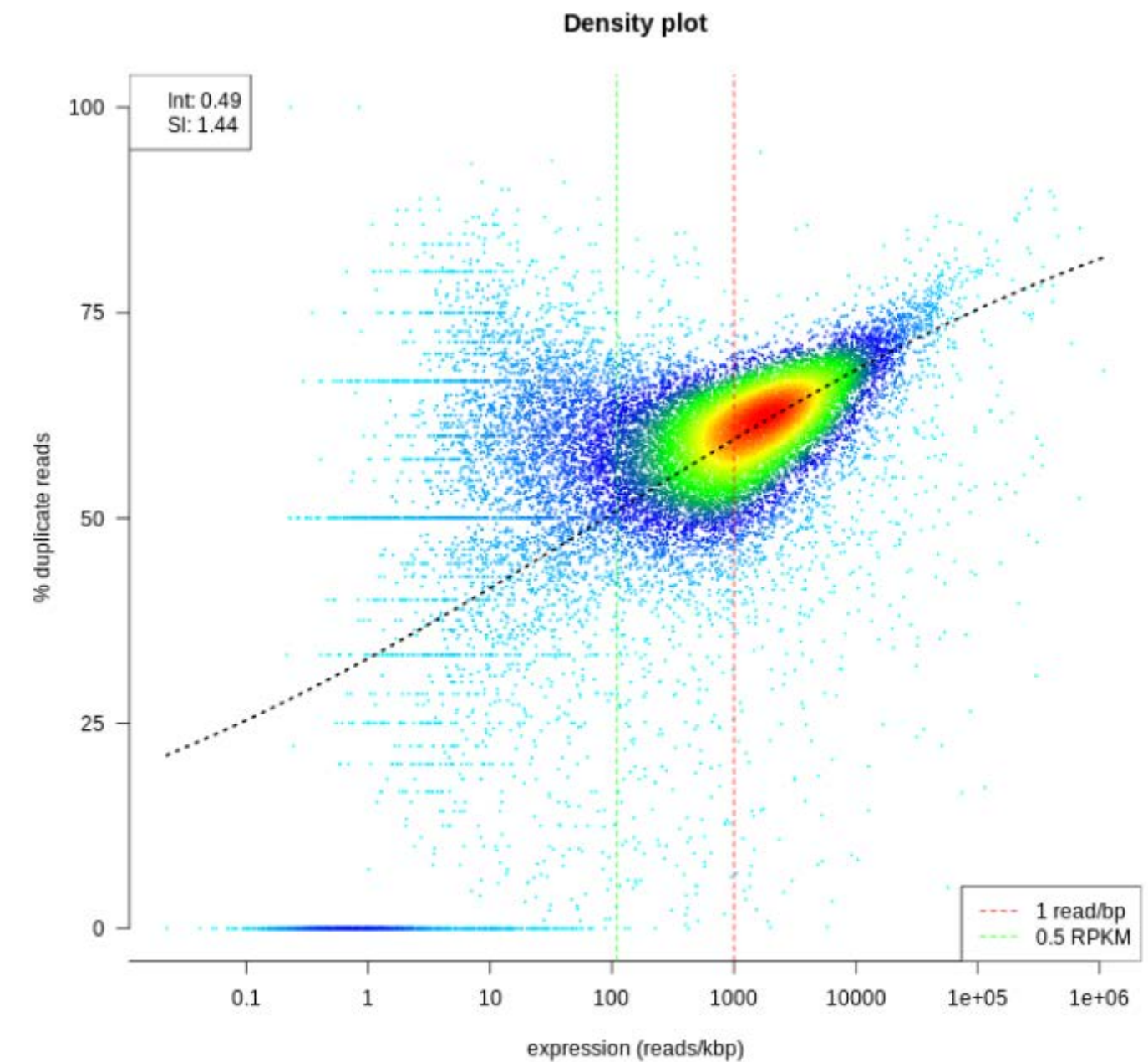
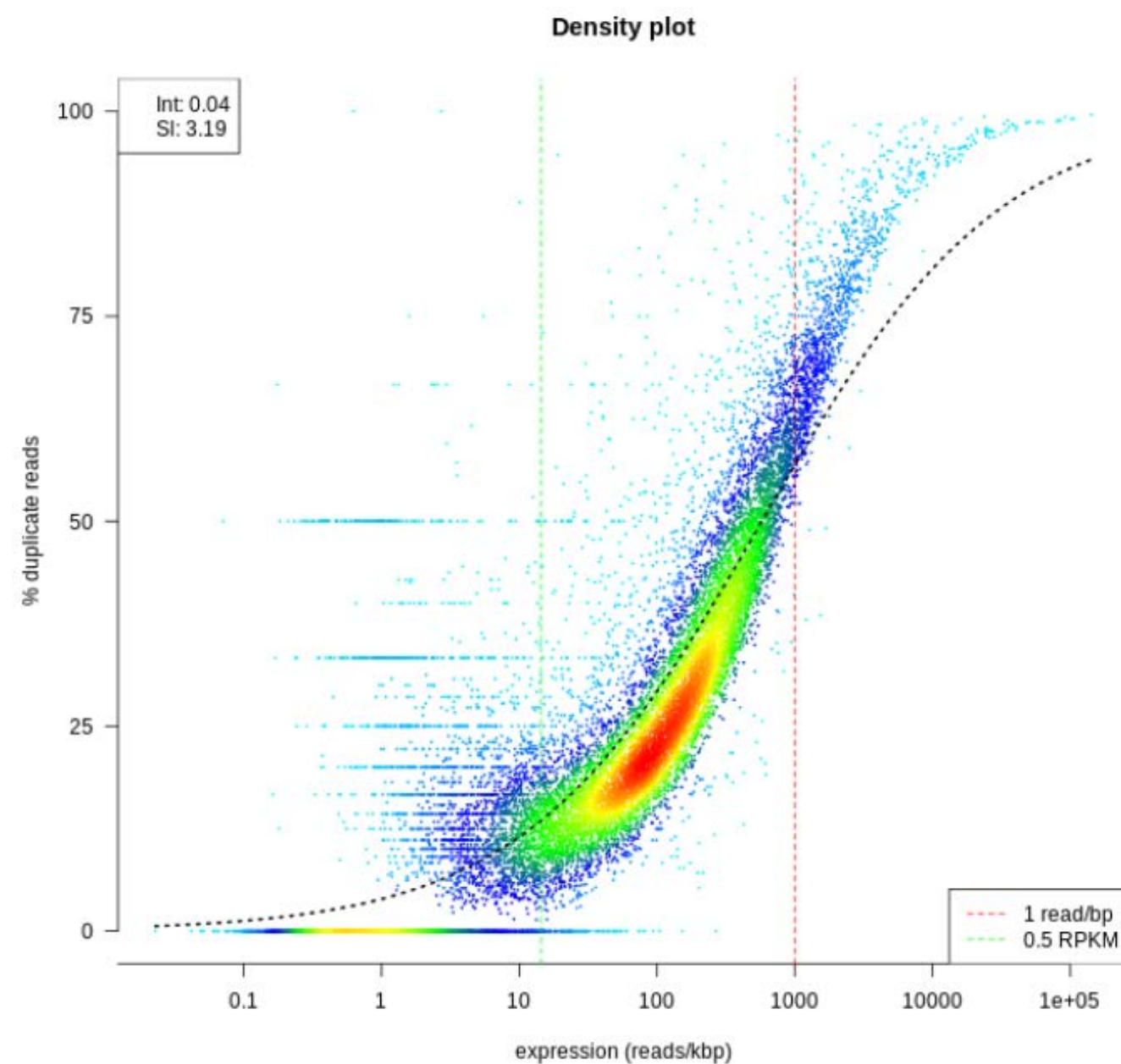


Interactive: select which columns to view/compare, and select columns to plot against each other

Lets you highlight, hide, or rename samples throughout the report

Other QC metrics: dupRadar

- Plots expression of genes versus the number of duplicate reads



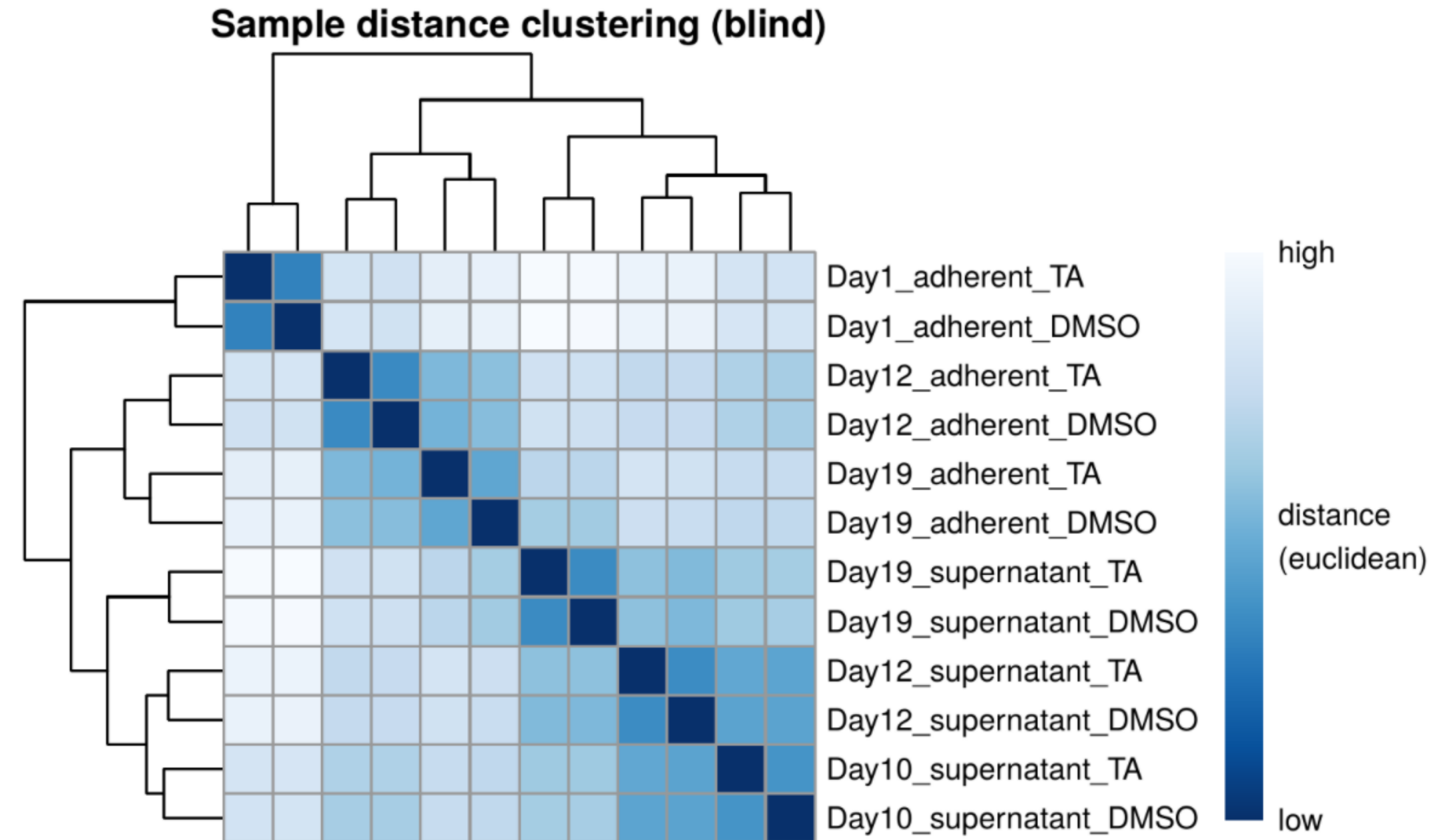
Other QC metrics: Heatmap of counts

Sample clustering


- Another form of quality control
- Shows how different each sample is
 - Samples you expect to be similar should cluster together
- Can reveal
 - Switched samples
 - Contaminations
 - Technical variations

DESeq2 - Sample distance cluster heatmap of counts

Euclidean distance between samples, based on variance stabilizing transformed counts (RNA: expressed genes, ChIP: bound regions, ATAC: accessible regions). Gives us an overview of similarities and dissimilarities between samples.



MultiQC reports

- Most reports will show many more QC metrics which you can inspect!
- Some figures and tables have a  button in the top right corner, which tells you more about what the metric shows and what is expected.

Useful link: <https://multiqc.info/docs/>

Contains further information on:

- Using and editing multiQC reports (e.g. selecting only samples of interest)
- Modules that you may see metrics from

Your own multiQC report:

https://mbdata.science.ru.nl/ghe_2022/day1/

See whether you can assess the quality of your own data. Are all samples OK for further processing?