Towards unified quality verification of synthetic count data with *countsimQC*

Charlotte Soneson
Statistical Bioinformatics (Robinson group)
University of Zurich, SIB Swiss Institute of Bioinformatics

EuroBioc, Cambridge, December 2017

University of Zurich[UZH]

@CSoneson

Swiss Institute of Bioinformatics

Polyester was created to fulfill the need for a tool to simulate RNA-seq reads for an experiment with replicates and well-defined differential expression. Users can easily simulate small experiments from a few genes or a single chromosome. This can reduce computational time in simulation studies when

Next, we designed simulations based on each of the real data sets (Methods). Briefly, the total spike-in count for each well was rescaled by a randomly sampled factor with variance equal to our experimental estimate of spike-in variance.

ata from human and the malaria parasite *Plasmodium falciparum* at three complexity levels (T1, T2 and T3) for each of the two organisms. Each data type was simulated three times, giving a total of 18 data sets, that are used throughout (Online Methods). *P. falciparum* was

applied to transformed gene-expression data. A comprehensive simulation study is conducted to measure the effect of several parameters on model performances, such as overdispersion, sample size, number of genes, number of classes, DE rate and the transformation method.

In addition to real scRNA-seq datasets (Islam et al., 2011; Grün et al., 2014), we used simulated datasets for our assessment. Using simulated data gives some advantages over the use of real data. Namely: (i) it provides a complete knowledge of positive, i.e., truly differentially expressed, and negative, i.e., truly not differentially expressed, genes; (ii) it gives the possibility to run replicated experiments, thus statistically testing the difference of the assessment scores; (iii) it allows testing different data scenarios. In this work, we

Here, we present the Splatter Bioconductor package for simple, reproducible, and v documented simulation of scRNA-seq data. Splatter provides an interface to multiple simulation methods including Splat, our own simulation, based on a gamma-Poisson distribution. Splat can simulate single populations of cells, populations with multiple cell types, or differentiation paths.

Based on this rationale, we simulated read count data to test how the SNR scores are distributed for each replicate model. The technical and biological replicate data were generated using Poisson and negative binomial distributions, respectively; 30% of the genes were chosen and their test group decreased to generate DE genes (see Methods). Then, the SNR

The simulations studies presented below assess Cuffdiff's performance under various hypothetical experimental scenarios. They are all intended to be "realistic", in the sense that users may encounter these experimental settings during routine RNA-Seq analysis. We have performed extensive sequencing using our in-house RNA-Seq read simulator, TuxSim.

2.1 Simulation of RNA-seq Data

We apply the RNASeqReadSimulator (update 2012.04.30) to generate synthetic RNA-seq data because it allows users full flexibility in controlling the read generation process. It also provides information about the transcript of origin for each read. RNASeqReadSimulator uses a number of Python

R package on GitHub

csoneson / countsimQC

Unwatch 2 | Star 15 | Fork 0

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

countsimQC - Compare characteristic features of count data sets

Edit

Add topics

55 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

csoneson Updated README — Latest commit e2521bc 2 days ago

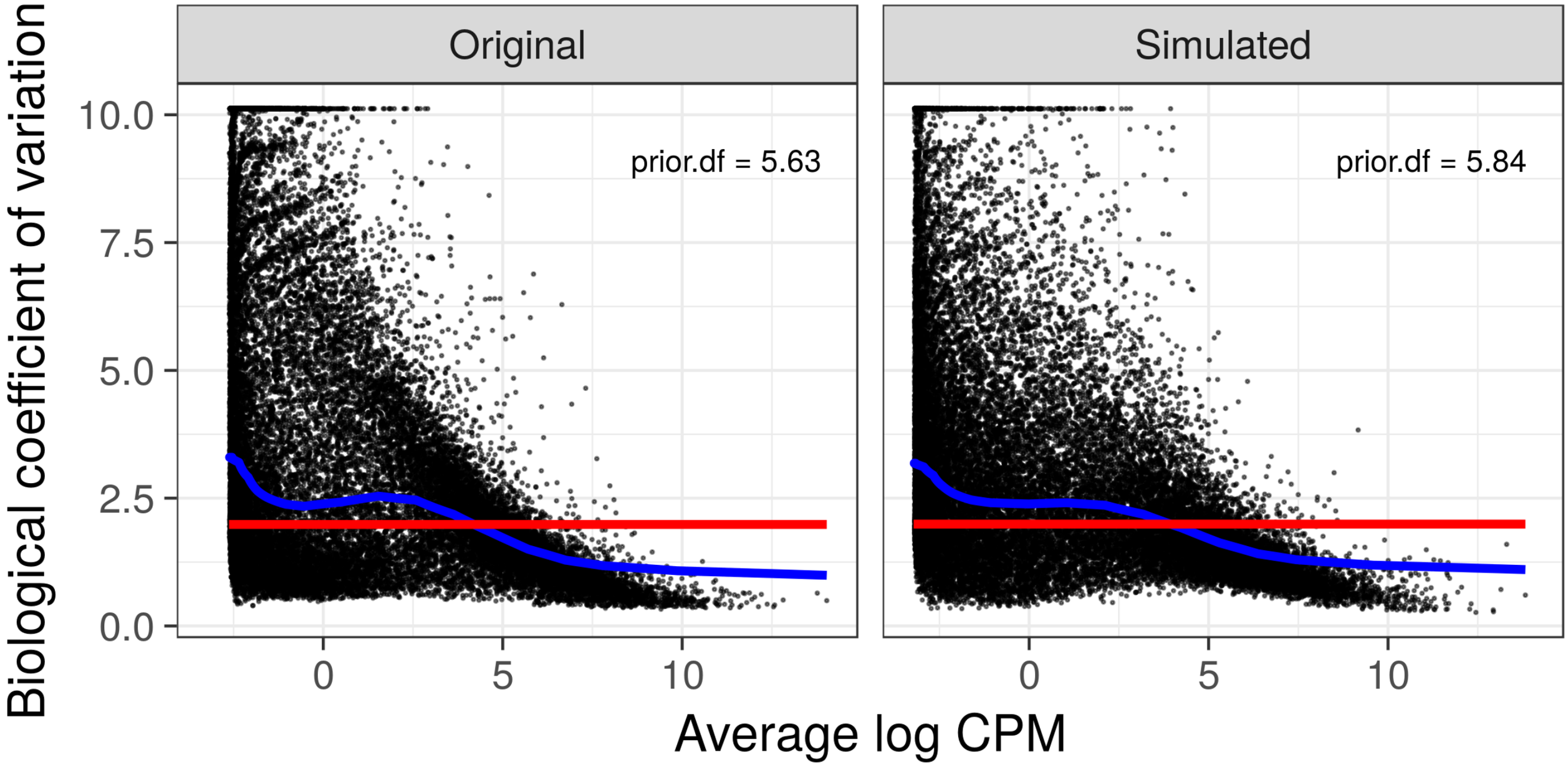
R	Fixed small bug in subsetting	3 months ago
data	v0.4.5. Allow matrices or data frames as input. Calculate area betwee...	4 months ago
inst/extdata	v0.5.2. Fixed dispersion visualizations.	3 months ago
man	Documentation	3 months ago
tests	v0.5.0. Added a number of quantitative evaluation criteria. Improved ...	3 months ago
vignettes	v0.5.1. Add example reports. Add options to silence progress. Small f...	3 months ago
.travis.yml	travis	3 months ago
DESCRIPTION	v0.5.2. Fixed dispersion visualizations.	3 months ago
NAMESPACE	v0.5.1. Add example reports. Add options to silence progress. Small f...	3 months ago
NEWS	Updated NEWS. Added example report.	3 months ago

Simple report generation

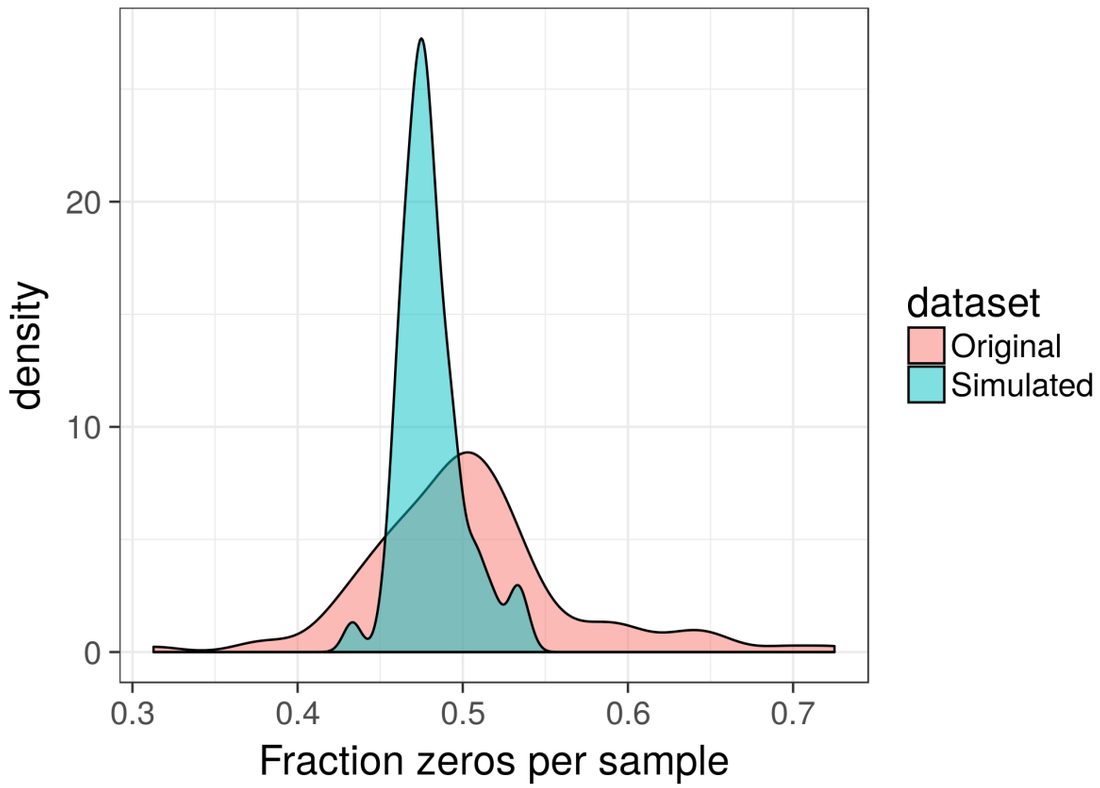
Named list of DESeqDataSet objects, each containing counts, sample info and design formula

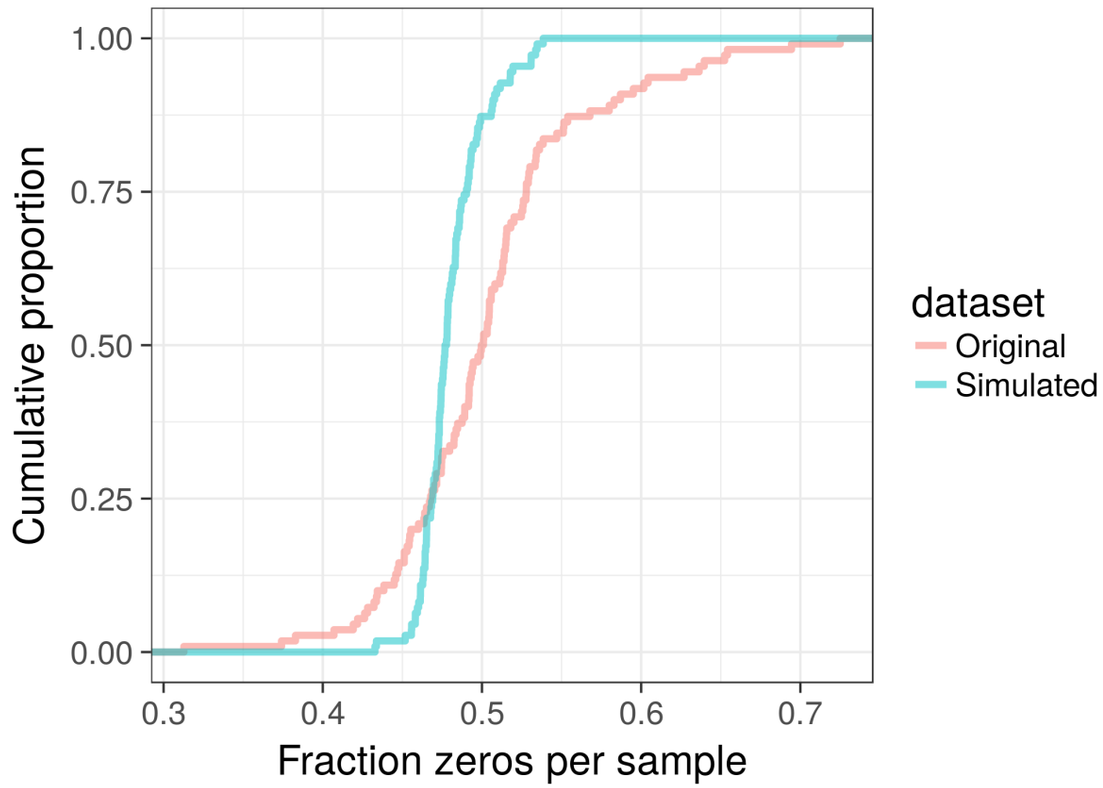
```
library(countsimQC)
data(countsimExample)
countsimQCReport(ddsList = countsimExample,
                 outputFile = "countsimReport.html",
                 outputDir = "./",
                 description = "This is a comparison of three count data sets.")
```

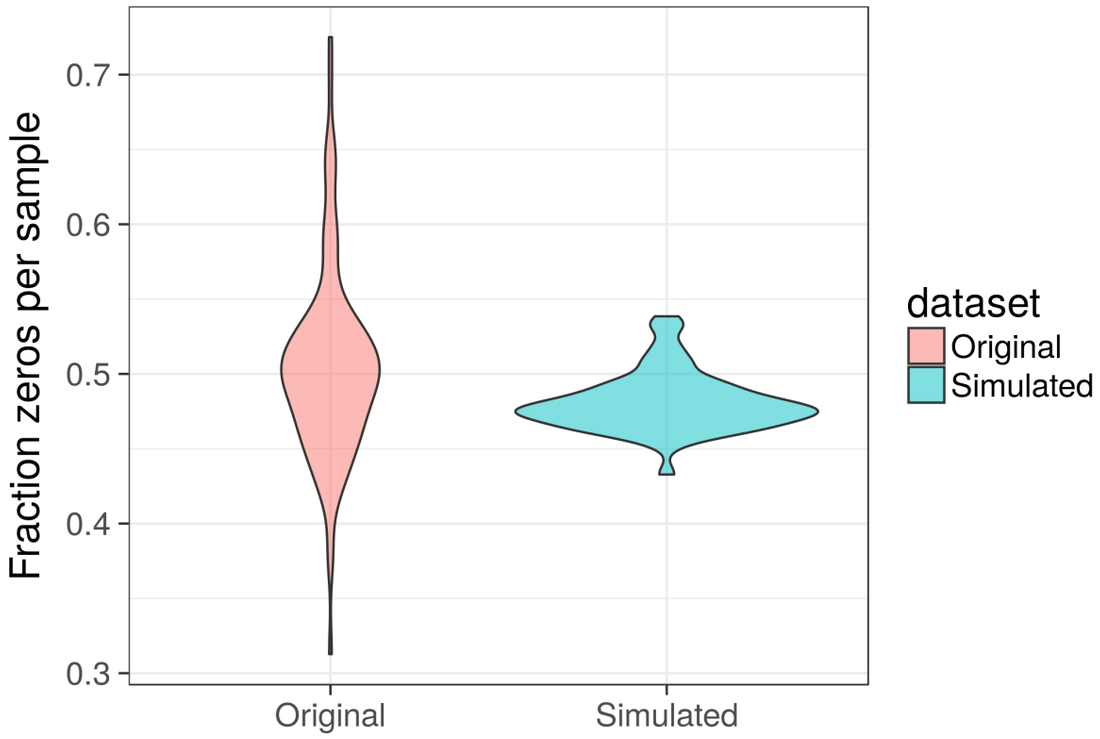
Example plot 1: mean vs dispersion

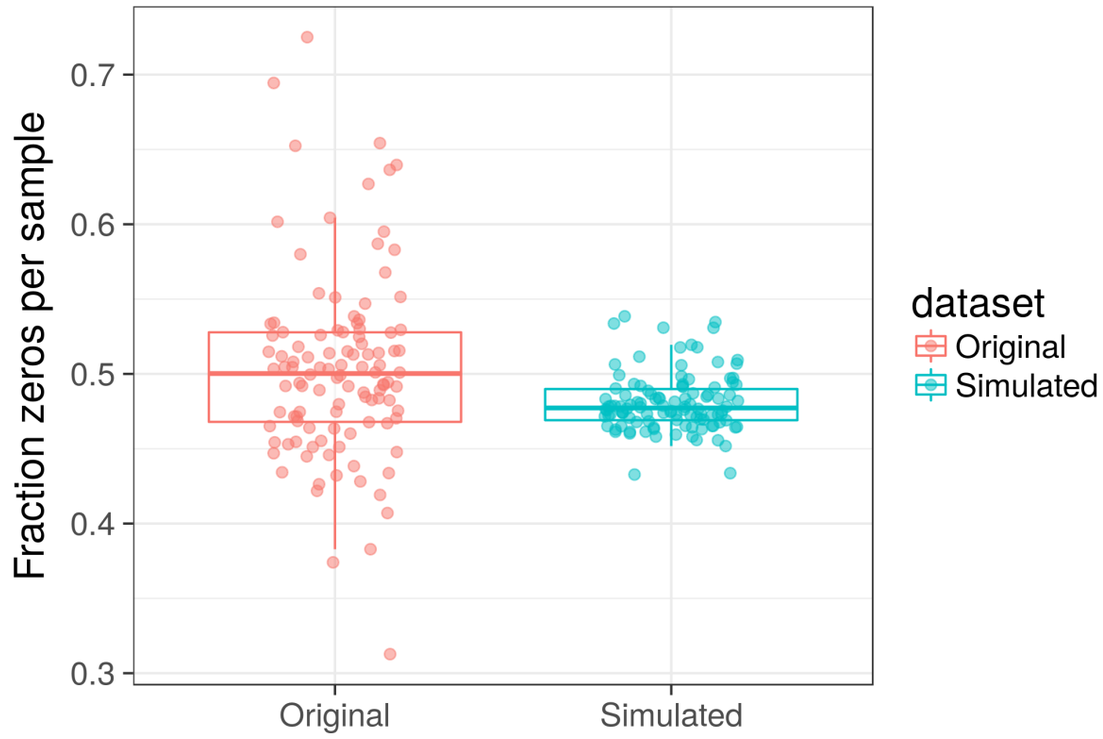


Example plots 2: fraction zeros per sample

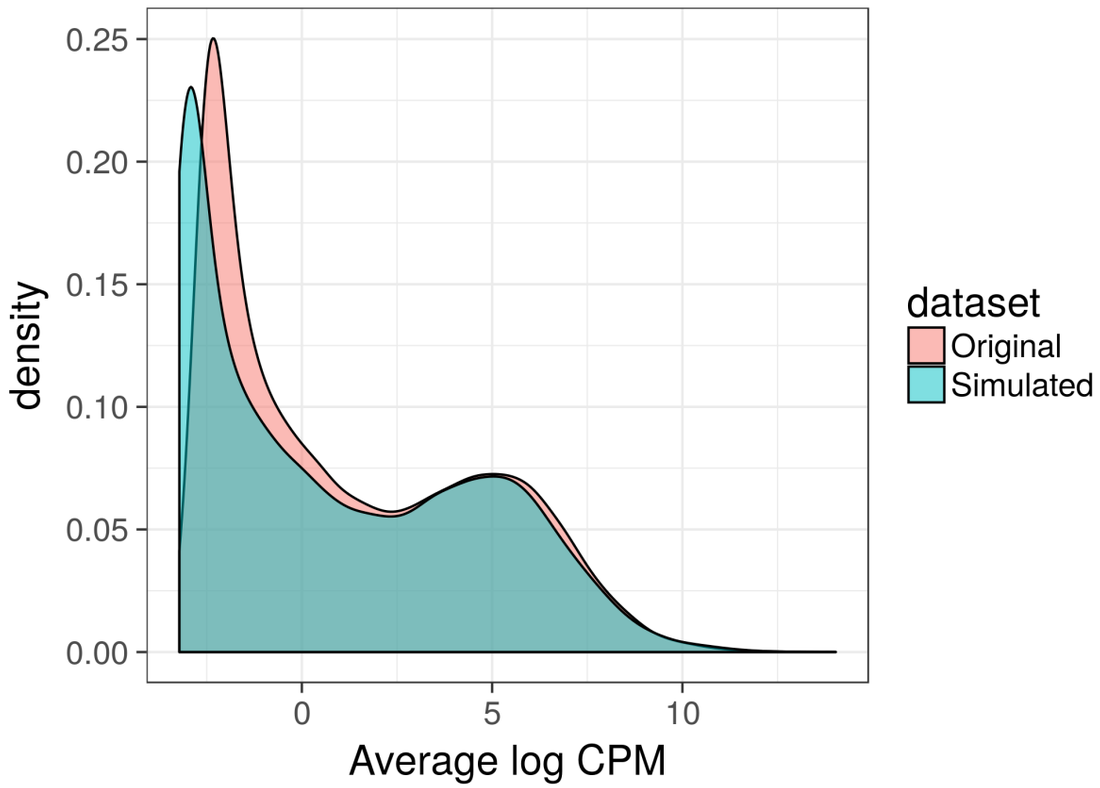


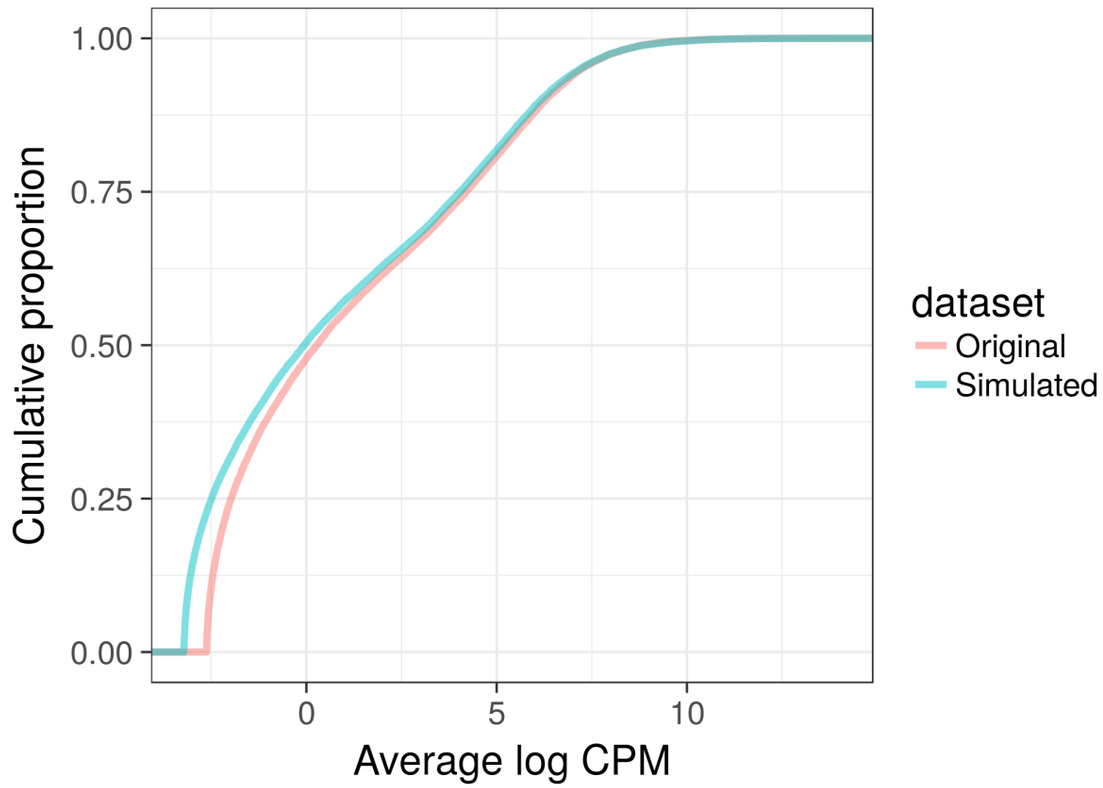


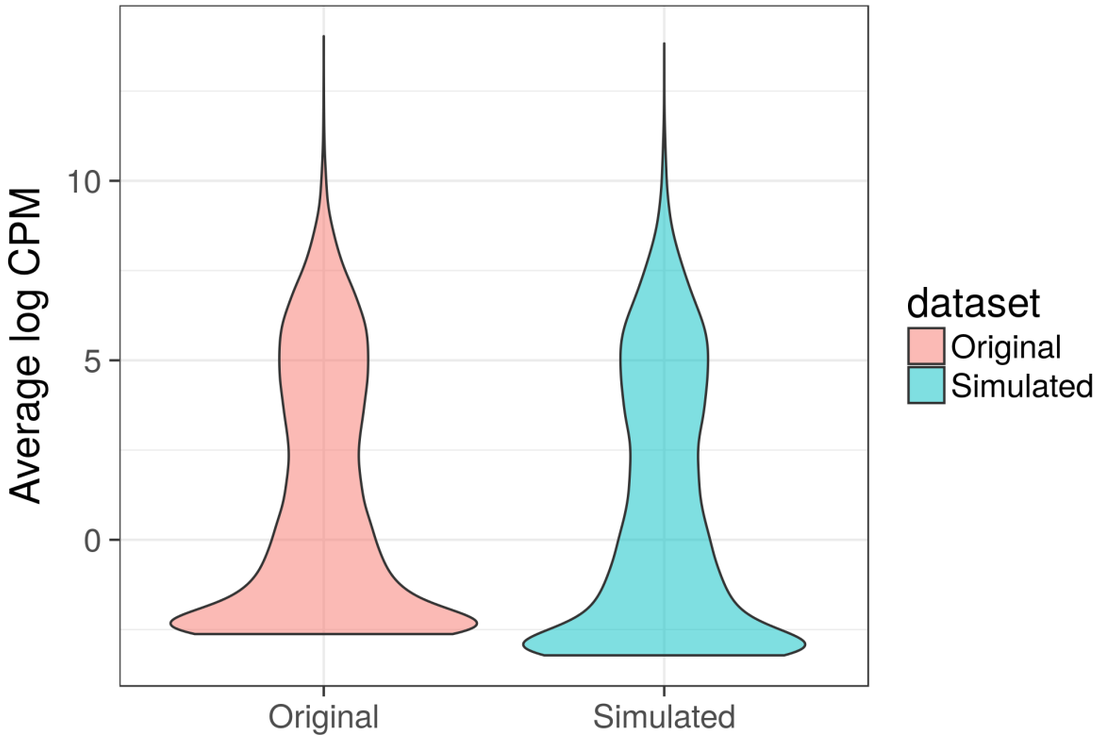


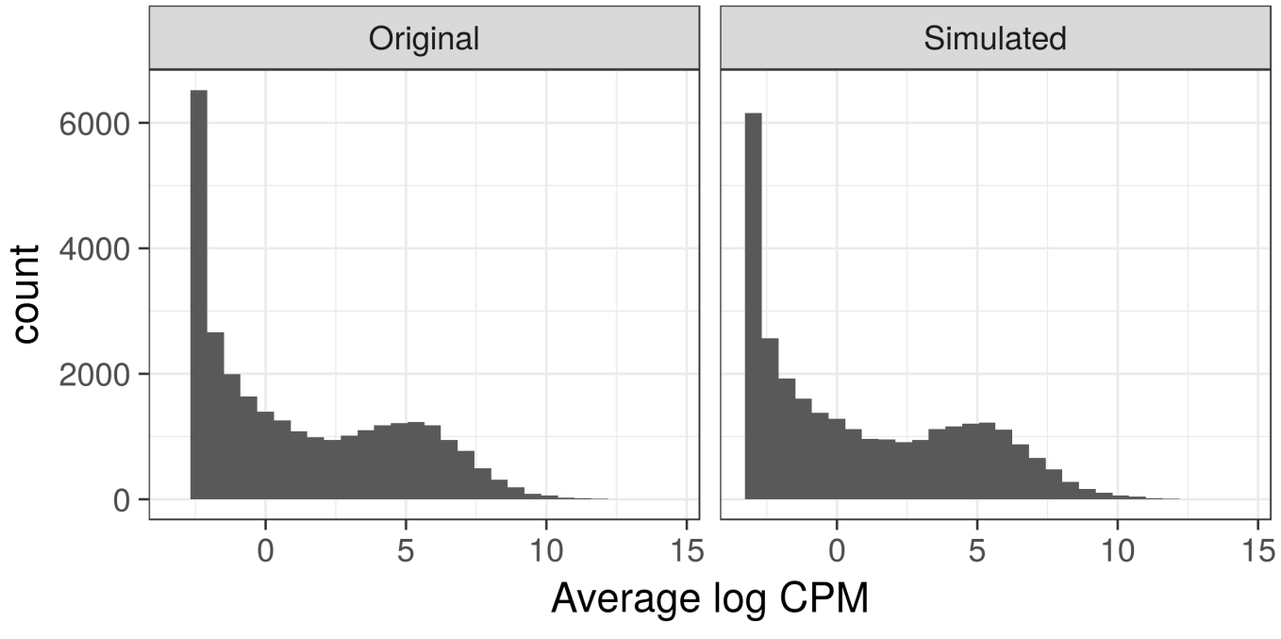


Example plots 3: logCPM distributions

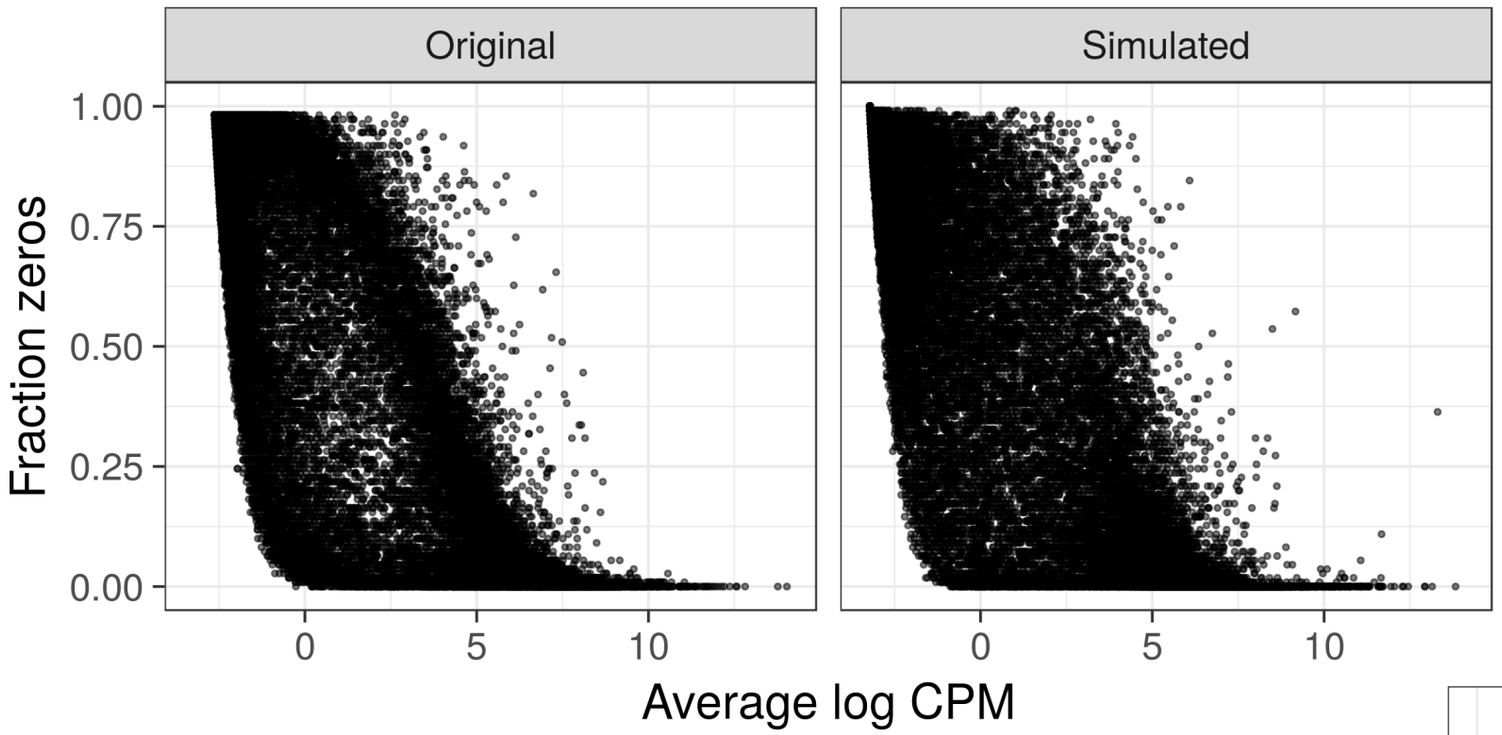


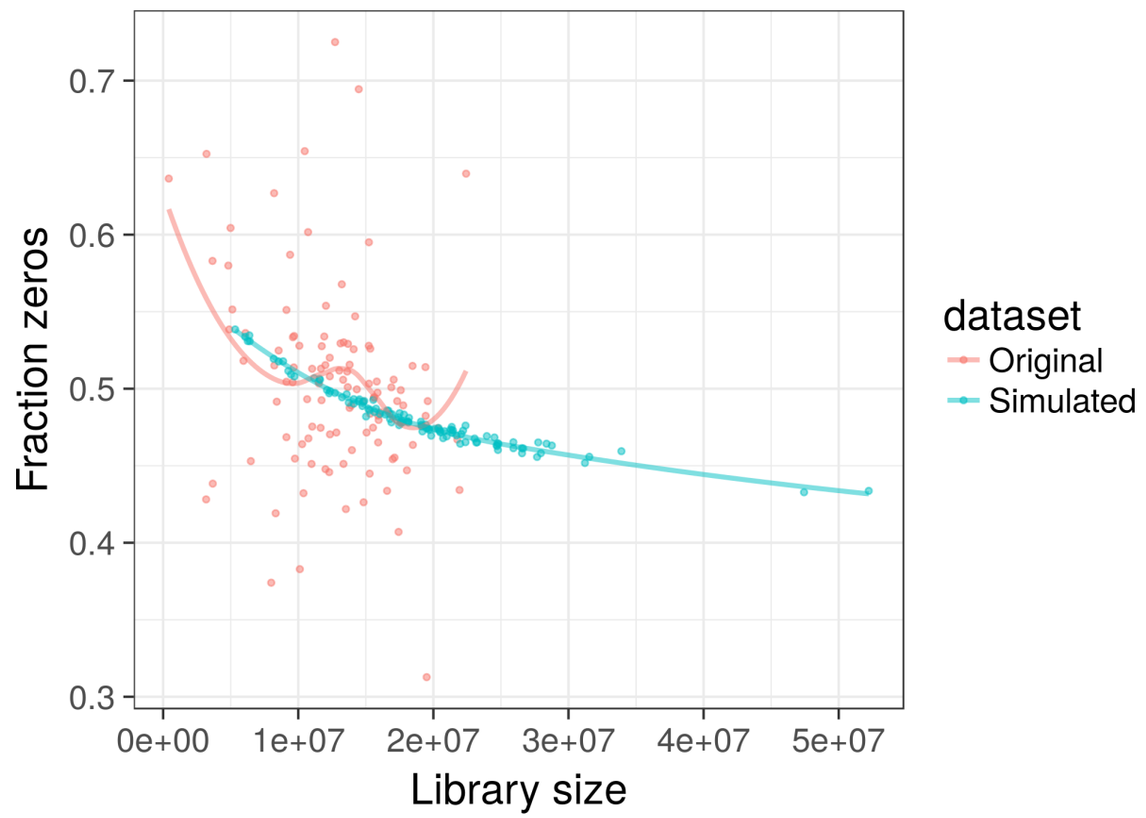






Example plots 4: total count/average expression vs fraction zeros





Full list of characteristics

General

- number of features
- number of samples

Feature-wise

- dispersion vs mean
- variance vs mean
- fraction zeros vs mean
- log CPM
- fraction zeros
- pairwise correlations

Sample-wise

- fraction zeros vs total count
- fraction zeros
- total count
- TMM factors
- effective library sizes
- pairwise correlations

In addition: comparisons of all characteristics between data set pairs

dataset1	dataset2	K-S statistic	K-S p-value	Scaled area between eCDFs	Runs statistic	Runs p-value	NN rejection fraction	Average silhouette width
SRP035988	SRP041538	0.217	0.000381	0.0669	-2.14	0.0162	0.332	0.0734
SRP035988	SRP051848	0.164	0.0146	0.0485	-2.56	0.00526	0.233	0.0406
SRP041538	SRP051848	0.139	0.0538	0.0293	0.207	0.582	0.115	0.0229

Acknowledgements

Statistical Bioinformatics Group @University of Zurich and SIB Swiss Institute of Bioinformatics

Mark D Robinson
Helen Lindsay
Simone Tiberi
Lukas Weber
Katharina Hembach
Stephan Schmeing
Vladimir Barbosa
Ruizhu Huang
Stephany Orjuela

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

University of Zurich[UZH]

https://github.com/csoneson/countsimQC

Bioinformatics

CORRECTED PROOF

Towards unified quality verification of synthetic count data with *countsimQC*

Charlotte Soneson ✉, Mark D Robinson ✉ Author Notes

Bioinformatics, btx631, https://doi.org/10.1093/bioinformatics/btx631
Published: 04 October 2017 **Article history** ▾

View Metrics