QoRTs Example Dataset

Stephen Hartley
National Human Genome Research Institute
National Institutes of Health

8 August 2014
Revised 9 October 2014
v0.0.20

Contents

1 Overview

The QoRTs software package is a fast, efficient, and portable multifunction toolkit designed to assist in the analysis, quality control, and data management of RNA-Seq datasets. Its primary function is to aid in the detection and identification of errors, biases, and artifacts produced by paired-end high-throughput RNA-Seq technology. In addition, it can produce count data designed for use with differential expression [1] and differential exon usage tools [2], as well as individual-sample and/or group-summary genome track files suitable for use with the UCSC genome browser (or any compatible browser).

The `QoRTsExampleData` package contains QoRTs QC summary output from an example dataset, designed for use with QoRTs and used in the QoRTs vignette. Due to size constraints this package does not include the raw bam-files themselves. The actual bamfiles, along with a step-by-step example walkthrough that covers the entire analysis pipeline, are linked to from the QoRTs github website (https://github.com/hartleys/QoRTs).

[1] Such as *DESeq*, *DESeq2*[1] or *edgeR*[2]
[2] Such as *DEXSeq*[3]

For more information on the QoRTs package, see the QoRTs vignette and reference manual, which can be found at https://github.com/hartleys/QoRTs.

1.1 Example data

The separate R package *QoRTsExampleData* contains an example dataset with an example decoder:

```
directory <- system.file("extdata/", package="QoRTsExampleData",
                         mustWork=TRUE);
decoder.file <- system.file("extdata/decoder.txt",
                            package="QoRTsExampleData",
                            mustWork=TRUE);
decoder.data <- read.table(decoder.file,
                           header=T,
                           stringsAsFactors=F);
print(decoder.data);
```

```
##    sample.ID lane.ID unique.ID  qc.data.dir group.ID input.read.pair.count
## 1      SAMP1      L1 SAMP1_RG1 ex/SAMP1_RG1     CASE                465298
## 2      SAMP1      L2 SAMP1_RG2 ex/SAMP1_RG2     CASE                472241
## 3      SAMP1      L3 SAMP1_RG3 ex/SAMP1_RG3     CASE                500691
## 4      SAMP2      L1 SAMP2_RG1 ex/SAMP2_RG1     CASE                461405
## 5      SAMP2      L2 SAMP2_RG2 ex/SAMP2_RG2     CASE                467713
## 6      SAMP2      L3 SAMP2_RG3 ex/SAMP2_RG3     CASE                492322
## 7      SAMP3      L1 SAMP3_RG1 ex/SAMP3_RG1     CASE                485397
## 8      SAMP3      L2 SAMP3_RG2 ex/SAMP3_RG2     CASE                489859
## 9      SAMP3      L3 SAMP3_RG3 ex/SAMP3_RG3     CASE                516906
## 10     SAMP4      L1 SAMP4_RG1 ex/SAMP4_RG1     CTRL                460968
## 11     SAMP4      L2 SAMP4_RG2 ex/SAMP4_RG2     CTRL                468391
## 12     SAMP4      L3 SAMP4_RG3 ex/SAMP4_RG3     CTRL                484530
## 13     SAMP5      L1 SAMP5_RG1 ex/SAMP5_RG1     CTRL                469884
## 14     SAMP5      L2 SAMP5_RG2 ex/SAMP5_RG2     CTRL                475001
## 15     SAMP5      L3 SAMP5_RG3 ex/SAMP5_RG3     CTRL                494213
## 16     SAMP6      L1 SAMP6_RG1 ex/SAMP6_RG1     CTRL                452429
## 17     SAMP6      L2 SAMP6_RG2 ex/SAMP6_RG2     CTRL                458810
## 18     SAMP6      L3 SAMP6_RG3 ex/SAMP6_RG3     CTRL                477751
```

Due to size constraints the example dataset contained in this R package includes only the QC output data, not the raw bam-files themselves. The actual bamfiles, along with a step-by-step example walkthrough that covers the entire analysis pipeline, are linked to from the QoRTs github website (https://github.com/hartleys/QoRTs).

The example dataset is derived from a set of rat pineal gland samples, which were multiplexed and sequenced across six sequencer lanes. For the sake of simplicity, the example dataset was limited to only six samples and three lanes. However, the bam files alone would still occupy 18 gigabytes of disk space, which would make it unsuitable for distribution as an example dataset. To further reduce

the example bamfile sizes, only reads that mapped to chromosomes chr14, chr15, chrX, and chrM were included. Additionally, all the selected chromosomes EXCEPT for chromosome 14 were randomly downsampled to 30 percent of their original read counts.

THIS DATASET IS INTENDED FOR DEMONSTRATION AND TESTING PURPOSES ONLY. Due to the various alterations that have been made to reduce file sizes and improve portability, it is really not suitable for any actual analyses.

For more information on how to use this dataset, see the QoRTs vignette and reference manual, which can be found at https://github.com/hartleys/QoRTs.

2 References

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. URL: http://genomebiology.com/2010/11/10/R106.

[2] Mark D. Robinson and Gordon K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23:2881, 2007. URL: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/21/2881, arXiv:http://bioinformatics.oxfordjournals.org/cgi/reprint/23/21/2881.pdf, doi:10.1093/bioinformatics/btm453.

[3] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22:2008, 2012. doi:10.1101/gr.133744.111.

3 Session Information

The session information records the versions of all the packages used in the generation of the present document.

```
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS release 6.8 (Final)
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
```

```
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] QoRTs_1.2.26 Cairo_1.5-9  knitr_1.15.1
##
## loaded via a namespace (and not attached):
## [1] BiocStyle_2.2.1 magrittr_1.5    tools_3.3.0     stringi_1.1.2
## [5] highr_0.6       stringr_1.1.0   evaluate_0.10
```

4 Legal

NOTE: The Scala package includes (internally) the sam-JDK library (sam-1.113.jar), from picard tools. The MIT license and copyright information can be accessed using the command:

```
java -jar /path/to/jarfile/QoRTs.jar ? samjdkinfo
```