

QoRTs Package User Manual

Stephen Hartley
National Human Genome Research Institute
National Institutes of Health

August 8 2014
February 20, 2015
v0.1.13

Contents

1	Overview	3
2	Requirements	3
2.1	Recommendations	4
3	Preparations	4
3.1	Alignment	4
3.2	Sorting	5
4	Quick Start	5
5	Dataset Organization	6
5.1	Example data	7
6	Processing of aligned RNA-Seq data	8
6.1	Memory Usage	9
7	Visualization	10
7.1	Generating all default plots	10
7.2	Plotting by sample, lane, or group	11
7.2.1	Summary Plots	12
7.2.2	Colored by Sample	15
7.2.3	Colored by Lane/Batch	16
7.2.4	Colored by Group/Phenotype	17
7.2.5	Basic Sample Highlight	18
7.2.6	Sample Highlight, Colored by Lane	19
7.3	Description of Individual Plots	21

7.3.1	Phred Quality Score	22
7.3.2	GC Content	23
7.3.3	Clipping Profile	24
7.3.4	Cigar Op Profile	25
7.3.5	Insert Size	26
7.3.6	Gene-Body Coverage	28
7.3.7	N-Rate	29
7.3.8	Cigar Length Distribution	30
7.3.9	Cumulative Gene Diversity	31
7.3.10	Nucleotide Rates, by Cycle	33
7.3.11	Aligned Nucleotide Rates, by Cycle	34
7.3.12	Leading Clipped Nucleotide Rates	35
7.3.13	Trailing Clipped Nucleotide Rates	36
7.3.14	Mapping location rates	37
7.3.15	Splice Junction Loci	39
7.3.16	Number of Splice Junction Events	40
7.3.17	Splice Junction Event Rates per Read-Pair	42
7.3.18	Breakdown of Splice Junction Events	43
7.3.19	Breakdown of Splice Junction Events, by locus type	44
7.3.20	Strandedness test	45
7.3.21	Mapping stats	46
7.3.22	Chromosome counts	47
7.3.23	Normalization Factors	48
7.3.24	Normalization Factors	49
7.3.25	Read drop rate	50
8	Secondary Utilities	50
8.1	Generating a flattened annotation file	50
8.2	Merging Count Data	51
8.3	Generating genome browser tracks	51
8.3.1	Generating wiggle tracks	53
8.3.2	Merging wiggle tracks	53
8.3.3	Generating splice-junction tracks	54
8.3.4	Merging splice-junction tracks	54
8.3.5	Generating splice-junction tracks	54
8.3.6	Merging splice-junction tracks	55
8.4	Importing data into other tools	55
9	References	56
10	Session Information	56
11	Legal	57

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).

In its primary role as a QC tool it can produce a wide variety of graphs, plots, and tables that allow the data to be visualized in various ways. Data can be compiled and contrasted in multiple ways to allow systematic errors or artifacts to reveal themselves more easily. While it will not directly assign pass/fail status, it is a powerful tool for bioinformaticians to detect and identify features in the data. In (hopefully) most cases, these plots and graphs will not reveal anything other than mixed statistical noise. Next-Gen sequencing technologies have matured to the point where gross systematic errors and batch-specific biases are relatively modest and rare. However: mistakes can still occur, and basing conclusions on flawed data can be disastrous.

The QoRTs package is composed of two parts: a java jar-file (for data processing) and a companion R package (for generating tables, figures, and plots). The java utility is written in the Scala programming language (v2.11.1), however, it has been compiled to java byte-code and does not require an installation of Scala (or any other external libraries) in order to function. The entire QoRTs toolkit can be used in almost any operating system that supports java and R.

This vignette primarily covers the quality control functionality of QoRTs, and briefly covers the other functions and capabilities.

The most recent release of QoRTs is available on the QoRTs github page (<http://github.com/hartleys/QoRTs>). Additional help and documentation is available online (<https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/index.html>). A comprehensive walkthrough covering the entire process from post-alignment all the way to differential expression analysis, along with a full example dataset and example output can be found online (<https://dl.dropboxusercontent.com/u/103621176/qorts/exData/QoRTsFullExampleData.zip>).

2 Requirements

Hardware: The java utility performs the bulk of the data processing, and will generally require at least 4gb of RAM. In general at least 8gb is recommended if available. The R package is only responsible for some light data processing and for plotting/visualization, and thus has much lower resource requirements. It should run adequately on any reasonably-powerful workstation.

Software: The QoRTs software package requires R version 3.0.2 or higher, as well as java 6 or higher.

¹Such as [DESeq](#), [DESeq2](#) [1] or [edgeR](#) [2]

²Such as [DEXSeq](#) [3]

Annotation: QoRTs requires transcript annotations in the form of a gtf file. If you are using a annotation guided aligner (which is STRONGLY recommended) it is likely you already have a transcript gtf file for your reference genome. We recommend you use the same annotation gtf for alignment, QC, and downstream analysis. We have found the Ensembl "Gene Sets" gtf³ suitable for these purposes. However, any format that adheres to the gtf file specification⁴ will work.

Dataset: QoRTs requires aligned and sorted RNA-Seq data. Data can be paired-end or single-end, unstranded or stranded (using either strandedness rule, see Section 6).

2.1 Recommendations

Clipping: For the purposes of Quality Control, it is generally best if reads are NOT hard-clipped prior to alignment. This is because hard clipping, especially variable hard-clipping from both the start and end of reads, makes it impossible to determine sequencer cycle from the aligned bam files, which in turn can obfuscate cycle specific artifacts, biases, errors, and effects. If undesired sequence must be removed, it is generally preferred to replace such nucleotides with N's, as this preserves cycle information. Note that many advanced RNA-Seq aligners will "soft clip" nonmatching sequence that occurs on the read ends, so hard-clipping low quality sequence is generally unnecessary and may reduce mapping rate and accuracy.

Replicates: Using barcoding, it is possible to build a combined library of multiple distinct samples which can be run together on the sequencing machine and then demultiplexed afterward. In general, it is recommended that samples for a particular study be multiplexed and merged into "balanced" combined libraries, each containing equal numbers of each biological condition. If necessary, these combined libraries can be run across multiple sequencer lanes or runs to achieve the desired read depth on each sample.

3 Preparations

There are a number of processing steps that must occur prior to the creation of usable bam files. We will briefly go over the required steps here:

3.1 Alignment

QoRTs is designed to run on paired-end next-gen RNA-Seq data. The data must be aligned (or "mapped") to a reference genome before QoRTs can be run. RNA-Star [4], GSNAP [5], and TopHat2 [6] are all popular and effective aligners for use with RNA-Seq data. The use of short-read or unspliced aligners such as BowTie, ELAND, BWA, or Novoalign is NOT recommended.

³Which can be acquired from the Ensembl website at <http://www.ensembl.org>

⁴See the gtf file specification at <http://genome.ucsc.edu/FAQ/FAQformat.html>

3.2 Sorting

For paired-end data, QoRTs requires that the bam files be sorted by either read name or by position. Sorting by name will result in a smaller memory footprint, but sorting by coordinate is more common in most production pipelines. Sorting can be accomplished via the samtools or novosort tools (which are NOT included with QoRTs). Sorting is unnecessary for single-end data.

To sort by coordinate:

```
samtools sort unsorted.bam sorted
OR
novosort unsorted.bam > sorted.bam
```

Or, to sort by read name:

```
samtools sort -n unsorted.bam sortedByName
OR
novosort -n unsorted.bam > sortedByName.bam
```

Technically, unsorted bam files can be run through QoRTs using the `--coordSorted` option. However, this is NOT RECOMMENDED, as the memory footprint may grow very large. As QoRTs iterates through the bam file it generates a (hash-table based) ref-map of unpaired reads. Thus, if too many read pairs are too far apart in the bam file then out-of-memory errors may occur. Similarly, certain unusual datasets containing large quantities of chimeric read-pairs (in which many read pairs are mapped to distant loci) may use unacceptable quantities of memory. QoRTs will print several warnings if the number of unpaired reads grows too high. In this case, it is recommended that the bam file be sorted by name.

4 Quick Start

In order to produce quality control metrics, plots, and pdf reports on a single replicate, simply use the command:

```
java -Xmx4G -jar /path/to/jarfile/QoRTs.jar QC
    --generatePlots
    mybamfile.bam
    transcriptAnnotationFile.gtf.gz
    /qc/data/dir/path/
```

Note: This command must be executed as a single line. Additional options may be required depending on the nature of the dataset. For single-ended data, the `--singleEnded` option should be included. For strand-specific data, the `--stranded` option should be included, and possibly also the `--fr_secondStrand` option depending on the stranded library type. For position-sorted data, the `--coordSorted` option should be used. See 6, or the QC utility documentation⁵ for more information on the available options.

⁵Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/QC.html>

5 Dataset Organization

Several QoRTs functions will require "decoder" information in some form, which describes each sample and all of its technical replicates (if any). The simplest method is to provide a decoder file. All of the columns are optional except for `unique.ID`, however if `group`, `lane`, and/or technical replicate information is not supplied then QoRTs obviously will not be able to produce plots that organized and/or grouped by these factors.

Fields:

- *unique.ID*: A unique identifier for the row. QoRTs will also accept the synonym "lanebam.ID". **THIS IS THE ONLY MANDATORY FIELD.**
- *lane.ID*: The ID of the lane or batch. By default this will be set to "UNKNOWN".
- *group.ID*: The ID of the "group". For example: "Case" or "Control". By default this will be set to "UNKNOWN".
- *sample.ID*: The ID of the biological sample from which the data originated. Each sample can have multiple rows, representing technical replicates (in which the same sample is sequenced on multiple lanes or runs). By default QoRTs will assume that every row comes from a separate sample, and will thus set the `sample.ID` to equal the `unique.ID`.
- *qc.data.dir*: The directory in which the java utility saved all the output data. If this column does not exist, by default it will be set to be `unique.ID`.
- *input.read.pair.count*: (Optional) The number of reads in the original fastq file, prior to alignment. If this field is left out, then QoRTs will skip comparisons and plotting of mapping rates. There are a number of other ways to input this value. See Section 7.3.21.
- *multi.mapped.read.pair.count*: (Optional) The number of reads that were multi-mapped by the aligner. This field should only be used if filtering for multi-mapped reads is performed prior to analysis with QoRTs (which is not recommended). Even in this case, this field can simply be left out and QoRTs will skip plotting and comparisons of multi-mapping rates. See Section 7.3.21.

In addition, the decoder can contain any other additional columns as desired, as long as all of the column names are distinct.

While QoRTs is primarily designed to allow comparisons between biological groups, lanes, sequencing-runs, etc, it can also be used on simpler datasets, or even individual samples. Thus, only the `unique.ID` variable is actually required. For testing purposes, you can produce a completed decoder (with all default values filled in) using the `completeAndCheckDecoder` function.

The simplest example would just be a character vector of `unique.ID`'s:

```
completeAndCheckDecoder(c("SAMPLE1", "SAMPLE2", "SAMPLE3"));

##   unique.ID sample.ID lane.ID group.ID qc.data.dir
## 1  SAMPLE1  SAMPLE1 UNKNOWN  UNKNOWN  SAMPLE1
## 2  SAMPLE2  SAMPLE2 UNKNOWN  UNKNOWN  SAMPLE2
## 3  SAMPLE3  SAMPLE3 UNKNOWN  UNKNOWN  SAMPLE3
```

Alternatively, any of the optional fields can be included or left out, as desired:

```
incompleteDecoder <- data.frame(unique.ID = c("SAMPLE1", "SAMPLE2"),
                                group.ID = c("CASE", "CONTROL"));
completeAndCheckDecoder(incompleteDecoder);

##   unique.ID group.ID sample.ID lane.ID qc.data.dir
## 1  SAMPLE1     CASE  SAMPLE1 UNKNOWN   SAMPLE1
## 2  SAMPLE2    CONTROL  SAMPLE2 UNKNOWN   SAMPLE2
```

5.1 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.

6 Processing of aligned RNA-Seq data

The first step is to process the aligned RNA-Seq data. The bulk of the data-processing is performed by the QoRTs.jar java utility. This tool produces an array of output files, analyzing and tabulating the data in various ways. This utility requires about 10-20gb of RAM for most genomes, and takes roughly 4-7 minutes to process 1 million read-pairs.

```
java -jar /path/to/jarfile/QoRTs.jar QC
      mybamfile.bam
      transcriptAnnotationFile.gtf.gz
      /qc/data/dir/path/
```

In the above command (which must be entered as a single line), you must replace /path/to/jarfile/ with the file-path to the directory in which the jar file is kept. The path /qc/data/dir/path/ should be replaced with the path you want the QC data to be written. This should match the path located in the decoder in the qc.data.dir column for this sample-run.

The bam processing tool includes numerous options. A full description of these options can be found in the online documentation of the jar utility⁶, or by entering the command:

```
java -jar /path/to/jarfile/QoRTs.jar QC --man
```

There are a number of crucial points that require attention when using the QoRTs.jar QC command.

- *Stranded Data*: By default, QoRTs assumes that the data is *NOT* strand-specific. For strand-specific data, the `--stranded` option must be used.
- *Stranded Library Type*: The `--fr_secondStrand` option may be required depending on the stranded library type. QoRTs does not attempt to automatically detect the platform and protocol used for stranded data. There are two types of strand-specific protocols, which are described by the TopHat/CuffLinks documentation at <http://cufflinks.cbc.b.umd.edu/manual.html#>

⁶Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

library as `fr-firststrand` and `fr-secondstrand`. In HTSeq, these same library type options are defined as `-s reverse` and `-s yes` respectively. According to the CuffLinks manual, `fr-firststrand` (the default used by QoRTs for stranded data) applies to dUTP, NSR, and NNSR protocols, whereas `fr-secondstrand` applies to "Directional Illumina (ligation)" and "Standard SOLiD" protocols. If you are unsure which library type applies to your dataset, don't worry: one of the tests will report stranded library type. If you use this test to determine library type, be aware that you may have to re-run QoRTs with the correct library type set.

- *Read Groups*: Depending on the production pipeline, each biological sample may be run across multiple sequencer lanes. These separate files can be merged together either before or after analysis with QoRTs (and maybe even before alignment). However, if the merger occurs before analysis with QoRTs, then each bam file will consist of multiple separate lanes or runs. In this case, it is **STRONGLY** recommended that separate QC runs be performed on each "read group", using the `--readGroup` option. This will prevent run- or lane-specific biases, artifacts, or errors from being obfuscated.
- *Read Sorting*: For paired-end data reads must be sorted. By default, QoRTs assumes that the bam file is sorted by *NAME*. If the bam file is instead sorted by position, then the `--coordSorted` option must be used. Sorting is unnecessary for single-end data.
- *Single-end vs paired-end*: By default, QoRTs assumes the input bam file consists of paired-end data. For single-end data, the `--isSingleEnd` option must be used.

For example, to read the first read group bam-file for SAMP1 from the example dataset (which is stranded, coordinate-sorted, and uses the `fr_firstStrand` stranded library type), one would use the following command:

```
java -jar /path/to/jarfile/QoRTs.jar QC
      --stranded
      --coordSorted
      inputData/bamFiles/SAMP1_RG1.bam
      inputData/annoFiles/anno.gtf.gz
      outputData/qortsData/SAMP1_RG1/
```

This command must be run on each bam file (and possibly more than once on each, if each bam file consists of multiple separate read-groups).

6.1 Memory Usage

Memory usage: The QoRTs QC utility requires at least 4gb or RAM for most genomes / datasets. Larger genomes, genomes with more annotated genes/transcripts, or larger bam files may require more RAM. You can set the maximum amount of RAM allocated to the JVM using the options `-Xmx4000M`. This should be included *before* the `-jar` in the command line. For example:

```
#Set the maximum to the minimum recommended 4 gigabytes:
java -Xmx4000M -jar /path/to/jarfile/QoRTs.jar QC
      --stranded
      --coordSorted
      inputData/bamFiles/SAMP1_RG1.bam
```

```
        inputData/annoFiles/anno.gtf.gz
        outputData/qortsData/SAMP1_RG1/

#Or Set the maximum to 16 gigabytes:
java -Xmx16G -jar /path/to/jarfile/QoRTs.jar QC
        --stranded
        --coordSorted
        inputData/bamFiles/SAMP1_RG1.bam
        inputData/annoFiles/anno.gtf.gz
        outputData/qortsData/SAMP1_RG1/
```

This option can be used with any and all of the QoRTs java utilities.

7 Visualization

Next, you must read in all the QC output from the java utility, using the command below. This command requires 2 arguments: a root directory and a decoder (which can be either a data frame or a file). We will be using the example data found in package *QoRTexampleData*, which is described in Section 5.1

```
res <- read.qc.results.data(directory, decoder = decoder.data,
                           calc.DESeq2 = TRUE, calc.edgeR = TRUE);
```

Note that the `calc.DESeq2` and `calc.edgeR` options are optional, and tell QoRTs to attempt to load the *DESeq2* and *edgeR* packages (respectively) and use the packages to calculate additional normalization size factors. This is not strictly needed for most purposes, but allows QoRTs to plot the normalization factors against one another. See section 7.3.23 for more information.

7.1 Generating all default plots

To generate all the default compiled plots all at once, use the command:

```
makeMultiPlot.all(res, outfile.dir = ".");
```

This will usually take some time to run, but will produce all the compiled summary plots described in the rest of this section, including separate highlight plots for every sample in the dataset. By default all images will be saved to file as pngs. There are a number of alternatives, which can be selected using the `plot.device.name` parameter. For example:

```
#Generate multi-page pdf reports:
makeMultiPlot.all(res, outfile.dir = ".", plot.device.name = "pdf");
#Generate svg vector drawings:
makeMultiPlot.all(res, outfile.dir = ".", plot.device.name = "svg");
```

Note: The R PDF device primarily uses vector drawings, however, some of the plots are too large to be efficiently stored as vectors. If pdf reports are desired, we recommend installing the *png* package. If this package is installed, then QoRTs will automatically rasterize the plotting areas of certain large plots (in particular: the gene diversity plots and the various NVC plots). Setting the `rasterize.large.plots` parameter to `FALSE` will deactivate this behavior. The `raster.height` and `raster.width` parameters can be used to increase the pixel resolution of the rasterized plotting regions, if desired.

The *png* package can be installed with the R command:

```
install.packages("png");
```

7.2 Plotting by sample, lane, or group

QoRTs includes automated methods for organizing and plotting the results in numerous different ways. The intent of these tools is to make any patterns and biases more visible to the user.

All plotting functions in QoRTs require a `QoRTs_Plotter` object. A `QoRTs_Plotter` is a `RefClass` object that contains all the QC data along with a set of parameters that determine how to color and draw each replicate's data. A full accounting of all possible options available in the is beyond the scope of this manual, but can be found in the help docs for the `QoRTs_Plotter` class.

7.2.1 Summary Plots

The most basic `QoRTs_Plotter` can be created using the command:

```
basic.plotter <- build.plotter.basic(res);
```

This `QoRTs_Plotter` object can be used to plot all replicates on top of one another in semi-transparent gray. For example:

```
makePlot.insert.size(basic.plotter);
```

Which produces Figure 1:

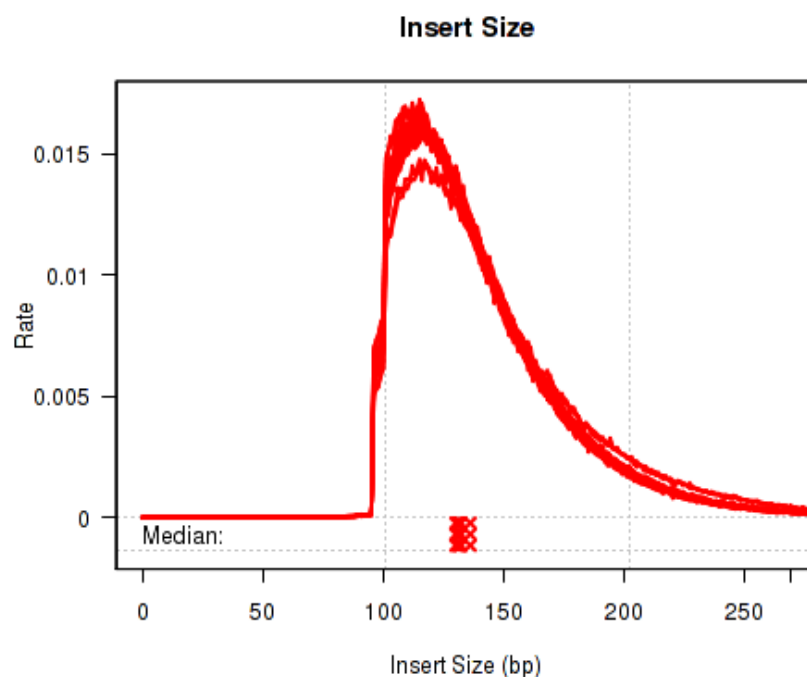


Figure 1: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.basic(res);
```

Which produces Figure 2:

This plot includes many sub-plots, all in a single frame. The sub-plots are:

- (a): Minimum phred quality score, by read position. Described in section 7.3.1
- (b): Lower-quartile phred quality score, by read position. Described in section 7.3.1
- (c): Median phred quality score, by read position. Described in section 7.3.1

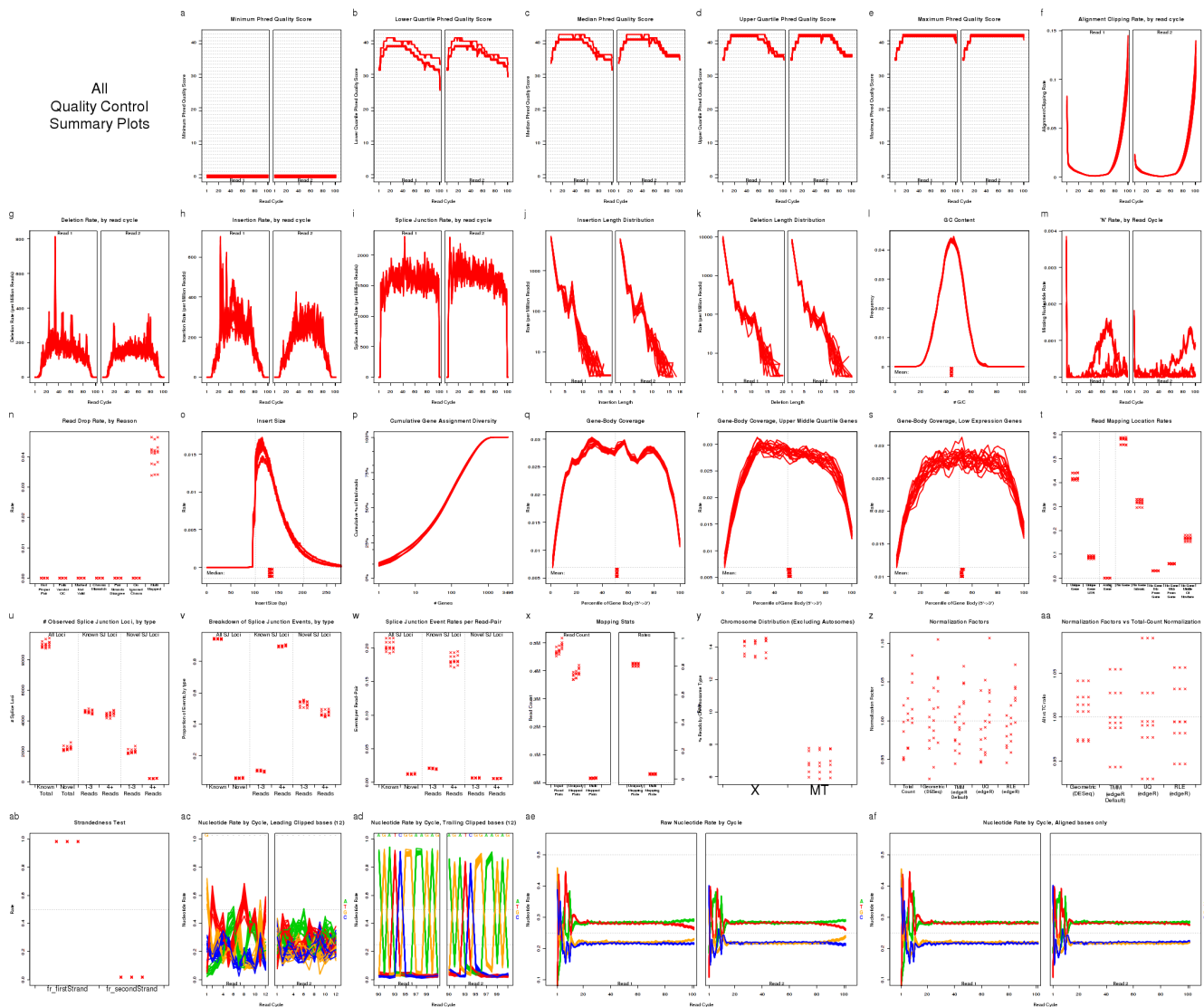


Figure 2: Compiled summary multi-plot

- (d): Upper-quartile phred quality score, by read position. Described in section 7.3.1
- (e): Maximum phred quality score, by read position. Described in section 7.3.1
- (f): Clipping profile. Described in section 7.3.3
- (g): Deletion profile. Described in section 7.3.4
- (h): Insertion profile. Described in section 7.3.4
- (i): Splicing profile. Described in section 7.3.4
- (j): Insertion length distribution. Described in section 7.3.8
- (k): Deletion length distribution. Described in section 7.3.8
- (l): GC content distribution. Described in section 7.3.2
- (m): N-rate, by read position. Described in section 7.3.7
- (n): Read drop rate. Described in section 7.3.25
- (o): Insert size distribution. Described in section 7.3.5
- (p): Cumulative gene assignment diversity. Described in section 7.3.9

- *(q)*: Gene body coverage, overall. Described in section [7.3.6](#)
- *(r)*: Gene body coverage, upper-middle quartile genes. Described in section [7.3.6](#)
- *(s)*: Gene body coverage, low expression genes. Described in section [7.3.6](#)
- *(t)*: Read mapping location rates. Described in section [7.3.14](#)
- *(u)*: Observed splice junction loci counts. Described in section [7.3.15](#)
- *(v)*: Splice junction event distribution. Described in section [7.3.17](#)
- *(w)*: Splice junction events per read-pair. Described in section [7.3.19](#)
- *(x)*: Read-mapping statistics. Described in section [7.3.21](#)
- *(y)*: Chromosome counts. Described in section [7.3.22](#)
- *(z)*: Comparison of normalization factors. Described in section [7.3.23](#)
- *(aa)*: Comparison of normalization factors relative to TC normalization. Described in section [7.3.24](#)
- *(ab)*: Strandedness test. Described in section [7.3.20](#)
- *(ac)*: Leading-clipped nucleotide rates. Described in section [7.3.12](#)
- *(ad)*: Trailing-clipped nucleotide rates. Described in section [7.3.13](#)
- *(ae)*: Raw nucleotide rate by read position. Described in section [7.3.10](#)
- *(af)*: Aligned nucleotide rate by read position. Described in section [7.3.11](#)

A printable pdf version of this multi-plot, with 6 plots on each page, can be generated with using the options:

```
makeMultiPlot.basic(res, plot.device.name = "pdf");
```

7.2.2 Colored by Sample

For small datasets, it can be useful to simply color each sample a distinct color, so that outliers can be easily identified. For this, you first generate a `QoRTs_Plotter` using the command:

```
bySample.plotter <- build.plotter.colorBySample(res);
```

This `QoRTs_Plotter` can be used to draw all the replicate on top of one another, but color them based on their sample.ID. The plotter can then be used to create various QC plots, for example:

```
makePlot.insert.size(bySample.plotter);  
makePlot.legend.over("topright",bySample.plotter);
```

Which produces Figure 3:

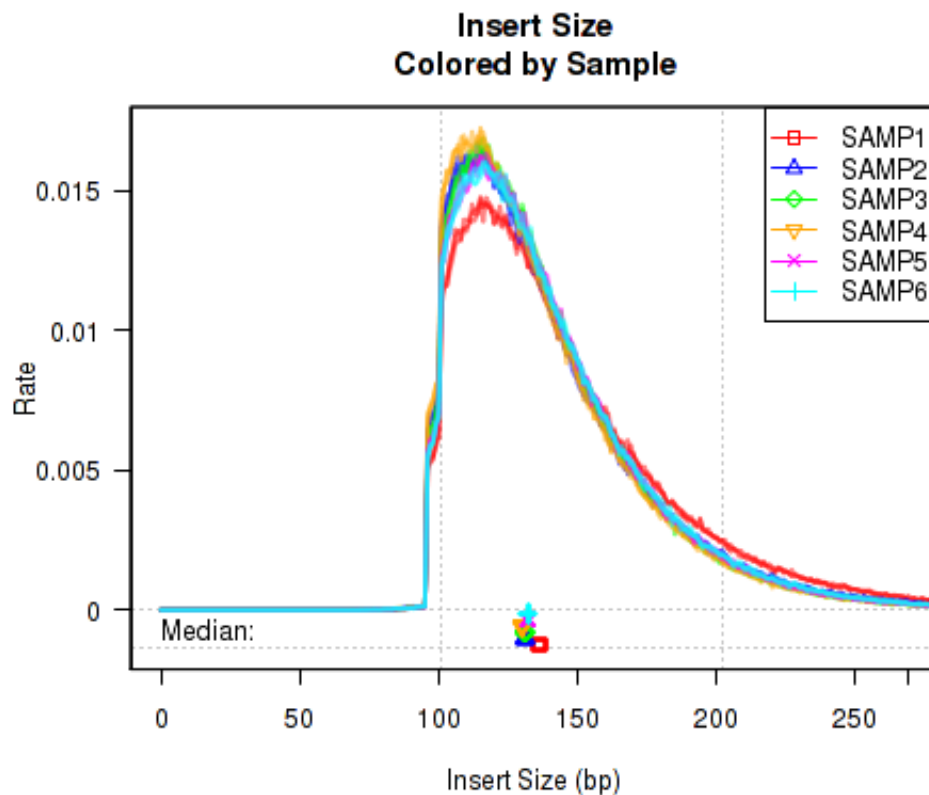


Figure 3: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.colorBySample(res);
```

7.2.3 Colored by Lane/Batch

In order to more easily detect batch effects, it is possible to color each replicate by lane/batch. For this, you can generate a `QoRTs_Plotter` with the command:

```
byLane.plotter <- build.plotter.colorByLane(res);
```

This `QoRTs_Plotter` can be used to color replicates based on `lane.ID`. The `QoRTs_Plotter` can then be used to create various QC plots, for example:

```
makePlot.insert.size(byLane.plotter);  
makePlot.legend.over("topright",byLane.plotter);
```

Which produces Figure 4:

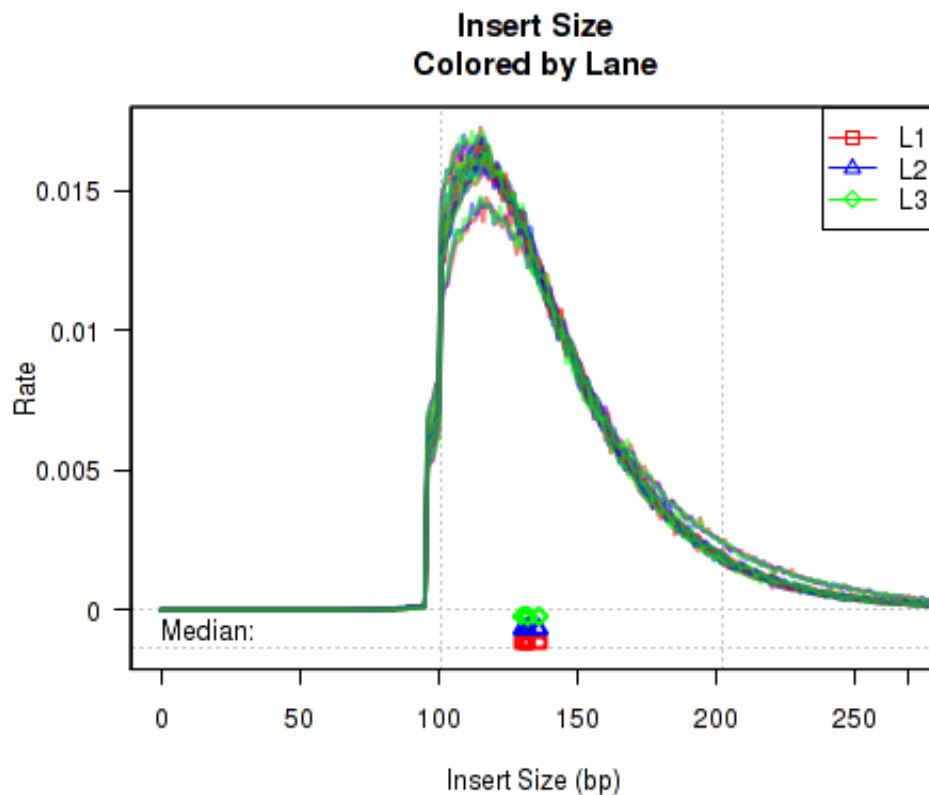


Figure 4: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.colorByLane(res);
```


7.2.4 Colored by Group/Phenotype

To detect variations caused by biological conditions (or artifacts and errors that occur disproportionately in certain biological conditions), it is sometimes useful to color samples by group.ID.

```
byGroup.plotter <- build.plotter.colorByGroup(res);
```

This QoRTs.Plotter can then be used to create various QC plots, for example:

```
makePlot.insert.size(byGroup.plotter);  
makePlot.legend.over("topright",byGroup.plotter);
```

Which produces Figure 5:

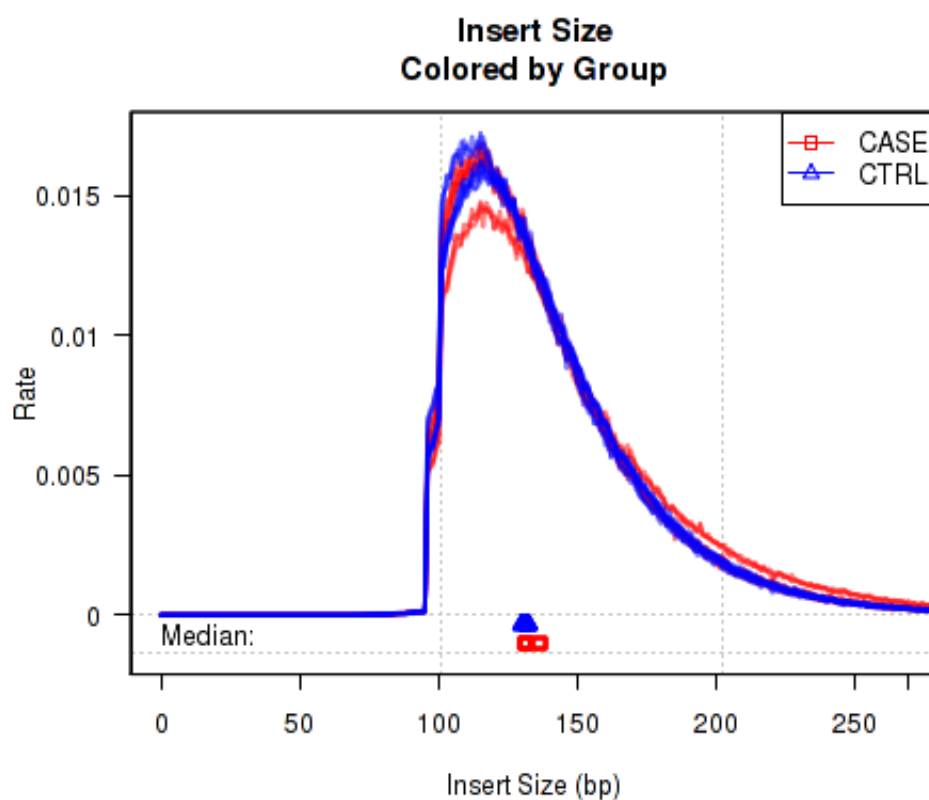


Figure 5: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.colorByGroup(res);
```

7.2.5 Basic Sample Highlight

Sometimes it is useful to "highlight" an individual sample.

```
sample.SAMP1.plotter <- build.plotter.highlightSample("SAMP1",res);
```

This QoRTs_Plotter can then be used to create various QC plots, for example:

```
makePlot.insert.size(sample.SAMP1.plotter);  
makePlot.legend.over("topright",sample.SAMP1.plotter);
```

Which produces Figure 6:

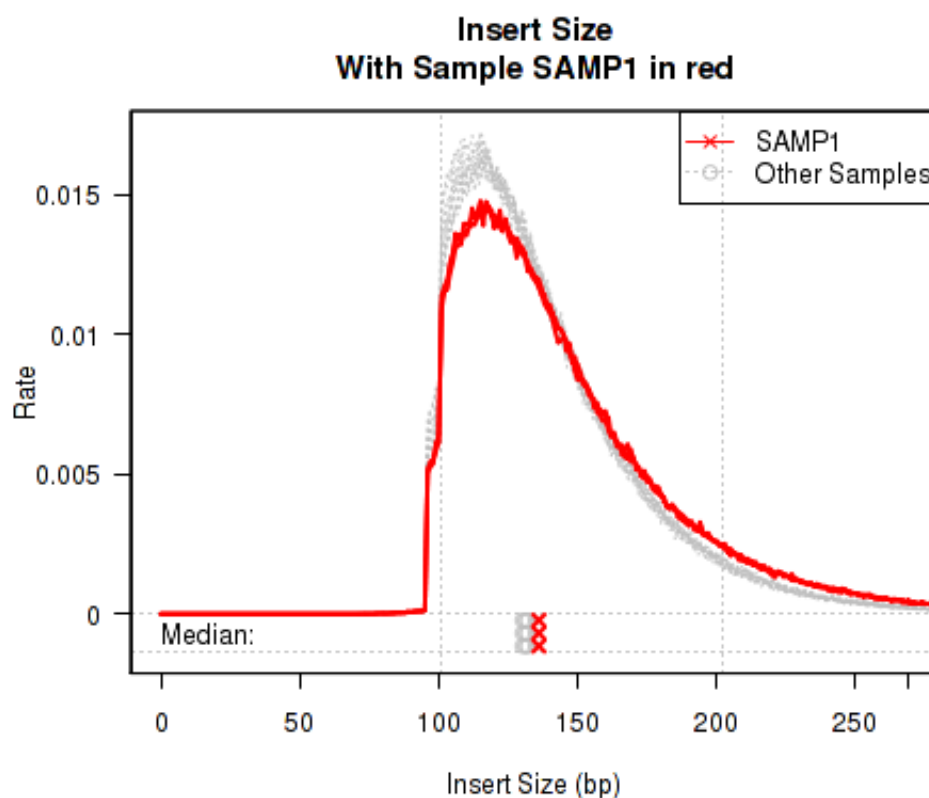


Figure 6: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.highlightSample(res,  
    curr.sample = "SAMP1");
```

7.2.6 Sample Highlight, Colored by Lane

Sometimes it can be useful to highlight an individual sample. However, if that sample has multiple "technical replicates" (derived from multiple separate lanes/runs on the same library), it can be useful to color the different runs with different distinct colors. With this plotter, only the "highlighted" sample is colored, all other samples are colored Gray.

```
sample.SAMP1.colorByLane.plotter <-  
  build.plotter.highlightSample.colorByLane("SAMP1",res);
```

This QoRTs.Plotter can then be used to create various QC plots, for example:

```
makePlot.insert.size(sample.SAMP1.colorByLane.plotter);  
makePlot.legend.over("topright",sample.SAMP1.colorByLane.plotter);
```

Which produces Figure 7:

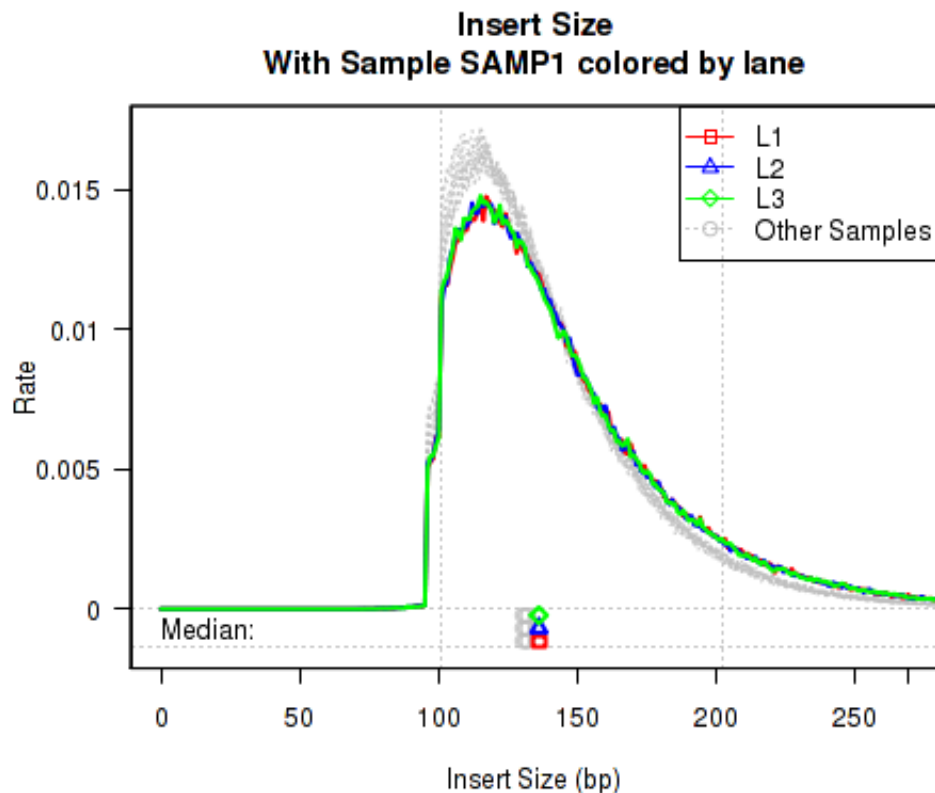


Figure 7: Phred Quality Score Plots

The above example plot displays the "Insert Size" of each replicate, as described in Section 7.3.5.

In addition, a compiled multi-plot in this style, containing all the standard QC plots, can be generated with the command:

```
makeMultiPlot.highlightSample.colorByLane(res,  
  curr.sample = "SAMP1");
```

7.3 Description of Individual Plots

QoRTs is capable of producing a wide variety of different plots and graphs. While most of these plots will not be particularly interesting or informative in the majority of cases, they may reveal artifacts or errors if and when they occur.

The example plots in the following section all use the `byLane.plotter` `QoRTs_Plotter` (from Section [7.2.3](#)), which colors each replicate by its lane ID.

7.3.1 Phred Quality Score

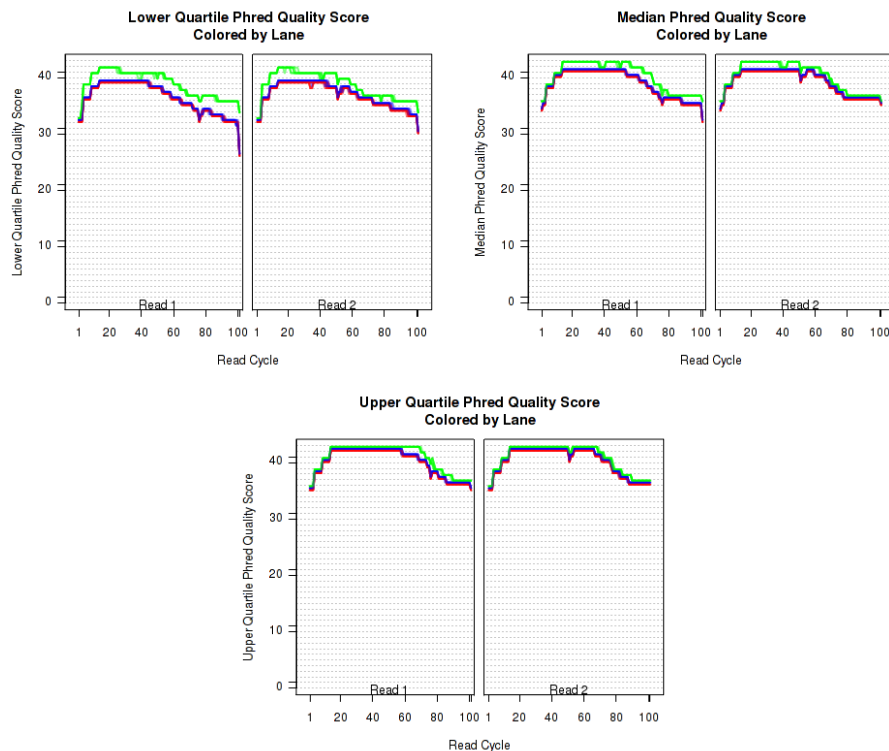


Figure 8: Phred Quality Score Plots

The plots shown in Figure 8 displays information about the phred quality score (y-axis) as a function of the position in the read (x-axis). Five statistics can be plotted: minimum, maximum, upper and lower quartiles, and median. These statistics are calculated individually for each replicate and each read position (ie, each plotted line corresponds to a replicate).

Note that the Phred score is always an integer, and as such these plots would normally be very difficult to read because lines would be plotted directly on top of one another. To reduce this problem, the plots are vertically offset from one another.

These plots can be generated individually with the commands:

```
makePlot.qual.pair(byLane.plotter, "lowerQuartile");
makePlot.qual.pair(byLane.plotter, "median");
makePlot.qual.pair(byLane.plotter, "upperQuartile");
```

Additional options (Not shown):

```
makePlot.qual.pair(byLane.plotter, "min");
makePlot.qual.pair(byLane.plotter, "max");
```

7.3.2 GC Content

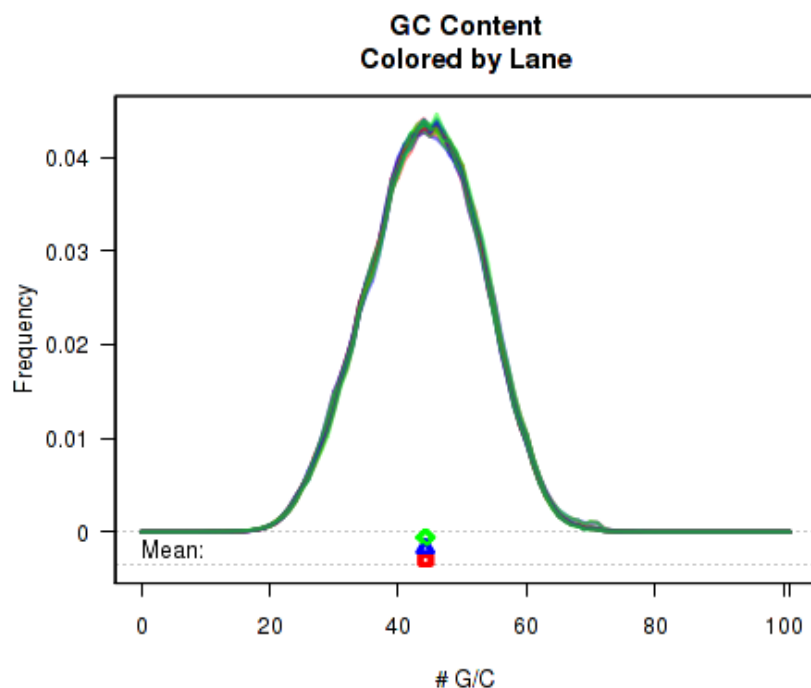


Figure 9: GC Bias

For each replicate, Figure 9 displays a histogram showing the frequency that different proportions of G and C (versus A, T, and N) appear in the replicate's reads. Each plotted line corresponds to a replicate. At the bottom of the plot the mean average G/C content is also plotted. Once again, the means are offset from one another by lane, to allow for easy detection of batch effects.

This plot can be generated individually with the command:

```
makePlot.gc(byLane.plotter);
```

The `byPair` option can be used to calculate the GC-distribution for read-pairs rather than for all reads individually. This is disabled by default because it often results in a jagged distribution when a appreciable proportion of the reads have an insert size equal to or smaller than the read length. When this occurs, the read-pair will almost always have an even number of G/C nucleotides.

7.3.3 Clipping Profile

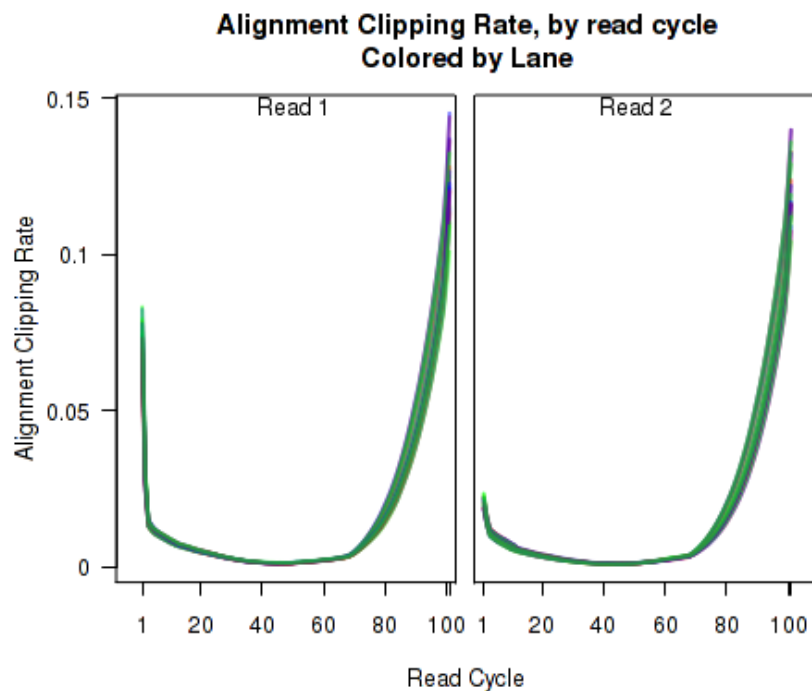


Figure 10: Clipping Profile

For each replicate, Figure 10 displays the rate (y-axis) at which the aligner soft-clips the reads as a function of read position (x-axis). Note that this will only be informative when using aligners that are capable of soft-clipped alignment (such as RNA-Star or GSNAP, but *not* TopHat).

This plot can be generated individually with the command:

```
makePlot.clipping(byLane.plotter);
```


7.3.4 Cigar Op Profile

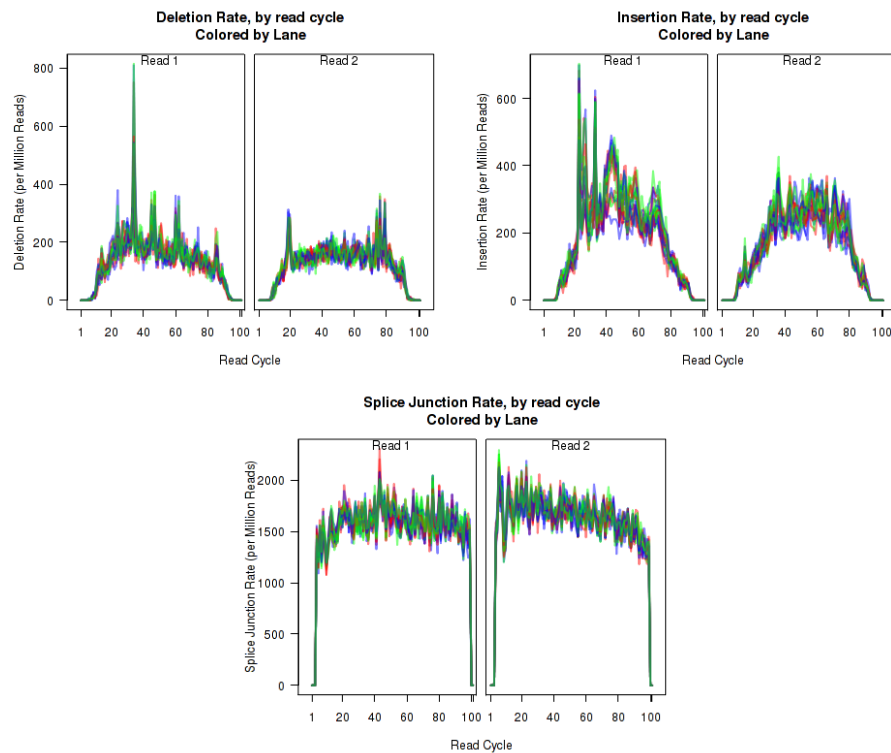


Figure 11: Cigar Operation Profiles

For each replicate, Figure 11 displays the rate (y-axis) of various cigar operations as a function of read position (x-axis). All 9 legal cigar operations can be plotted, but for most purposes only Deletions, Insertions, and Splice junctions will be informative.

This plot can be generated with the command:

```
makePlot.cigarOp.byCycle(byLane.plotter, "Del");
makePlot.cigarOp.byCycle(byLane.plotter, "Ins");
makePlot.cigarOp.byCycle(byLane.plotter, "Splice");
```

7.3.5 Insert Size

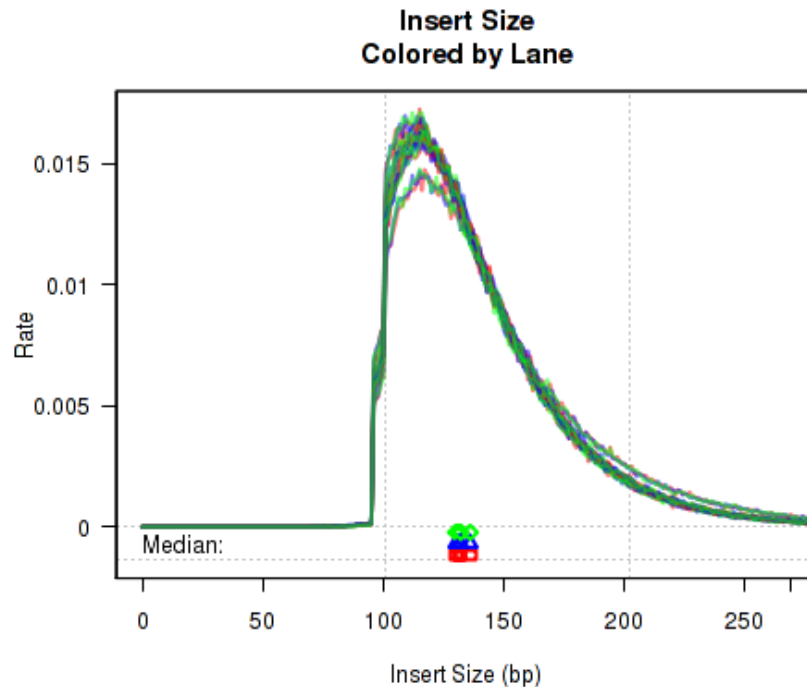


Figure 12: Insert Size

For each replicate, Figure 12 displays a histogram of the "insert size". Each line corresponds to one replicate, and displays the rate (y-axis) at which that replicate's reads possess a given insert size (x-axis).

Definition: "Insert Size": The "insert size" is the length (in base-pairs) between the two sequencing adapters for a pair of paired-end reads. In other words, it is the size of the original RNA fragment.

Insert Size Estimation: The Insert size is calculated using the alignment of the paired reads. When the two paired reads are aligned such that they overlap with one another the insert size can be calculated exactly. In such cases, the calculation of the insert size does *not* depend on the transcript annotation. However, when there is no overlap the exact insert size can be uncertain. Multiple splice junctions may lie in the region between the endpoints of the two paired reads, and there is no real way to determine which junctions the fragment used, if any. QoRTs uses the set of all splice junctions found between the endpoints of the two reads, and uses the shortest possible path from endpoint to endpoint. In some cases this may *under-estimate* the insert size, as the actual path may not be the shortest possible path. In other cases this may also *over-estimate* the insert size, if the RNA fragment includes novel splice junctions not found in the transcript annotation. However, in most cases this method appears to produce a reasonably good approximation of the insert size.

Note that the median average insert sizes for each replicate are plotted below the main plot. Each point corresponds to one replicate.

This plot can be generated individually with the command:

```
makePlot.insert.size(byLane.plotter);
```

If the dataset is single-ended, this will generate a placeholder plot.

7.3.6 Gene-Body Coverage

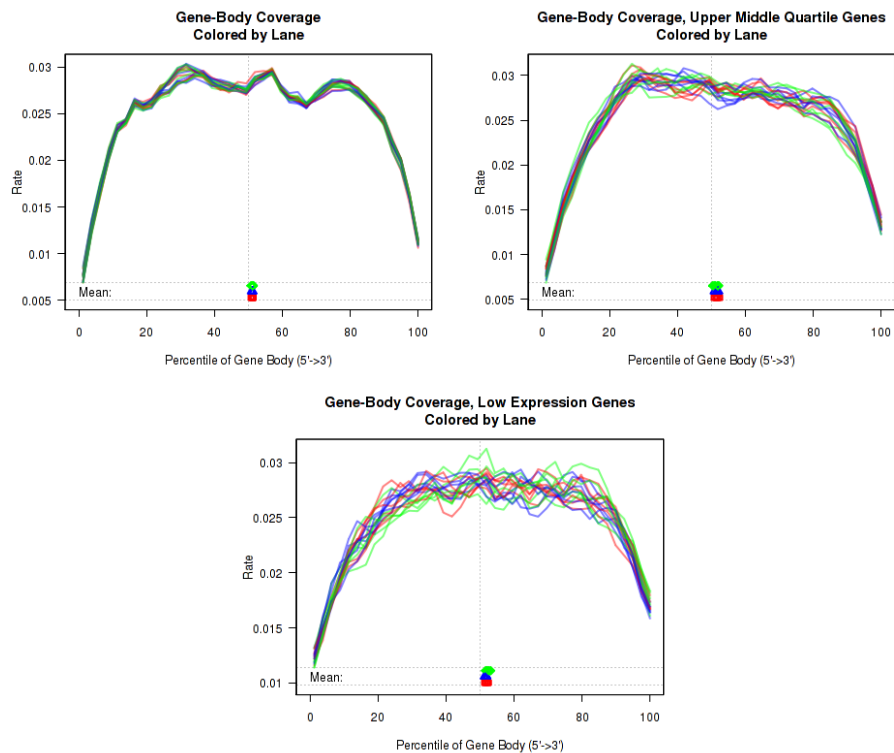


Figure 13: Gene-Body Coverage

For each replicate, the leftmost plot of Figure 13 displays the coverage profile across quantiles of all genes' lengths, from 5' to 3'. The middle plot displays the coverage profile for only the genes that are in the upper-middle quartile by read-count. The leftmost plot displays the coverage profile for the genes that are in the two lower quartiles.

Minor notes: To calculate the coverage profile, all the transcripts for each gene are merged together into a single "flat" pseudo-transcript which contains all exonic regions belonging to the gene. For each gene, the pseudo-transcript is broken up into 40 equal-length counting bins, so that each bin contains 2.5% of the total gene length. Each read-pair is counted once for every counting bin with which it overlaps. Any genes that overlap with other genes are automatically excluded. Any reads that overlap with more than one gene are automatically excluded. Any genes that have zero reads on a given replicate are automatically excluded for the purposes of finding the gene quantiles.

This plot can be generated individually with the command:

```
makePlot.genebody.coverage(byLane.plotter);
makePlot.genebody.coverage.UMQuartile(byLane.plotter);
makePlot.genebody.coverage.lowExpress(byLane.plotter);
```

7.3.7 N-Rate

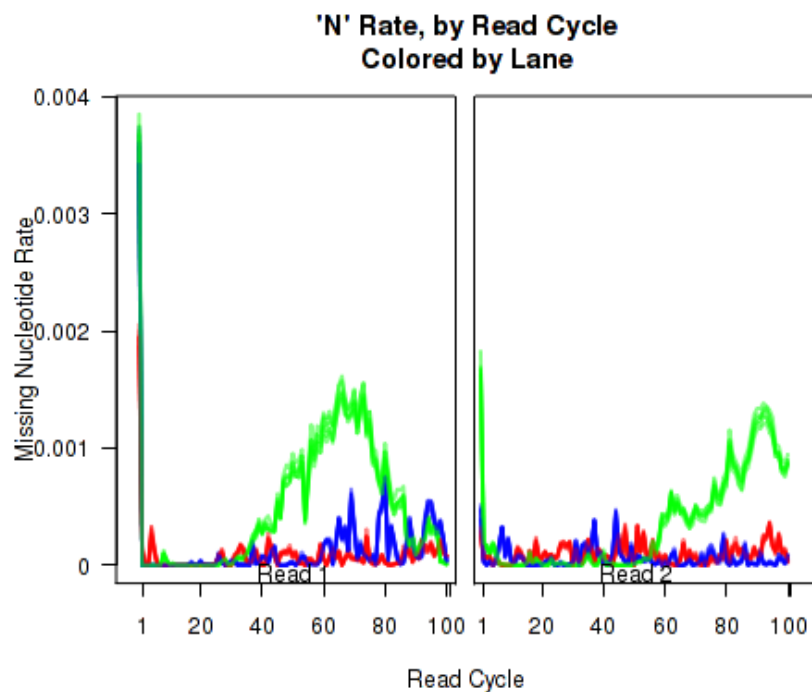


Figure 14: PLOT NAME

Figure 14 displays the rate (y-axis) at which the read sequence is "N" (or "missing"), as a function of the read position (x-axis). Each line corresponds to one replicate.

This plot can be generated individually with the command:

```
makePlot.missingness.rate(byLane.plotter);
```

7.3.8 Cigar Length Distribution

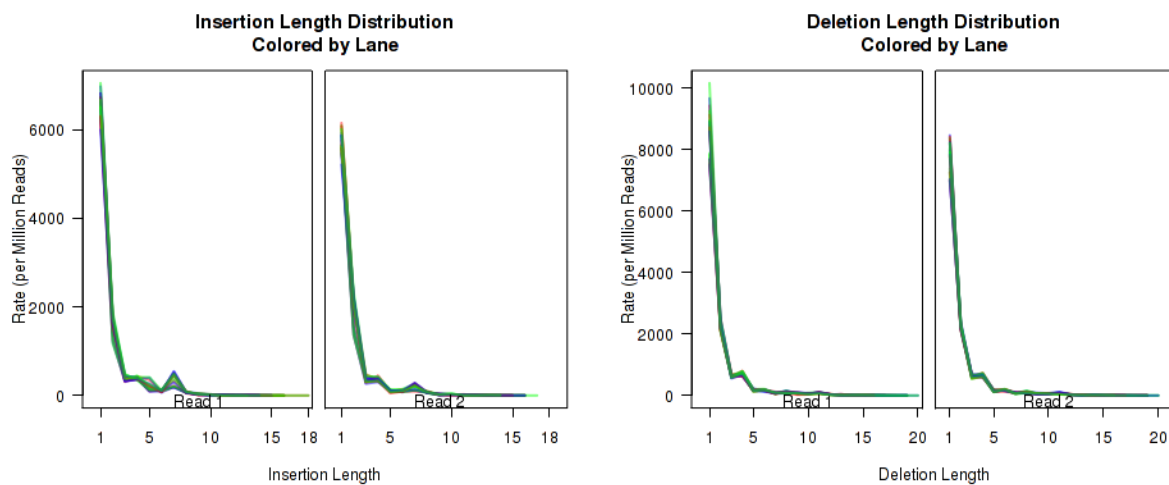


Figure 15: Cigar Length Distribution

The plots in Figure 15 display histograms of cigar operation length.

These plots can be generated individually with the commands:

```
makePlot.cigarLength.distribution(byLane.plotter, "Ins");  
makePlot.cigarLength.distribution(byLane.plotter, "Del")
```

7.3.9 Cumulative Gene Diversity

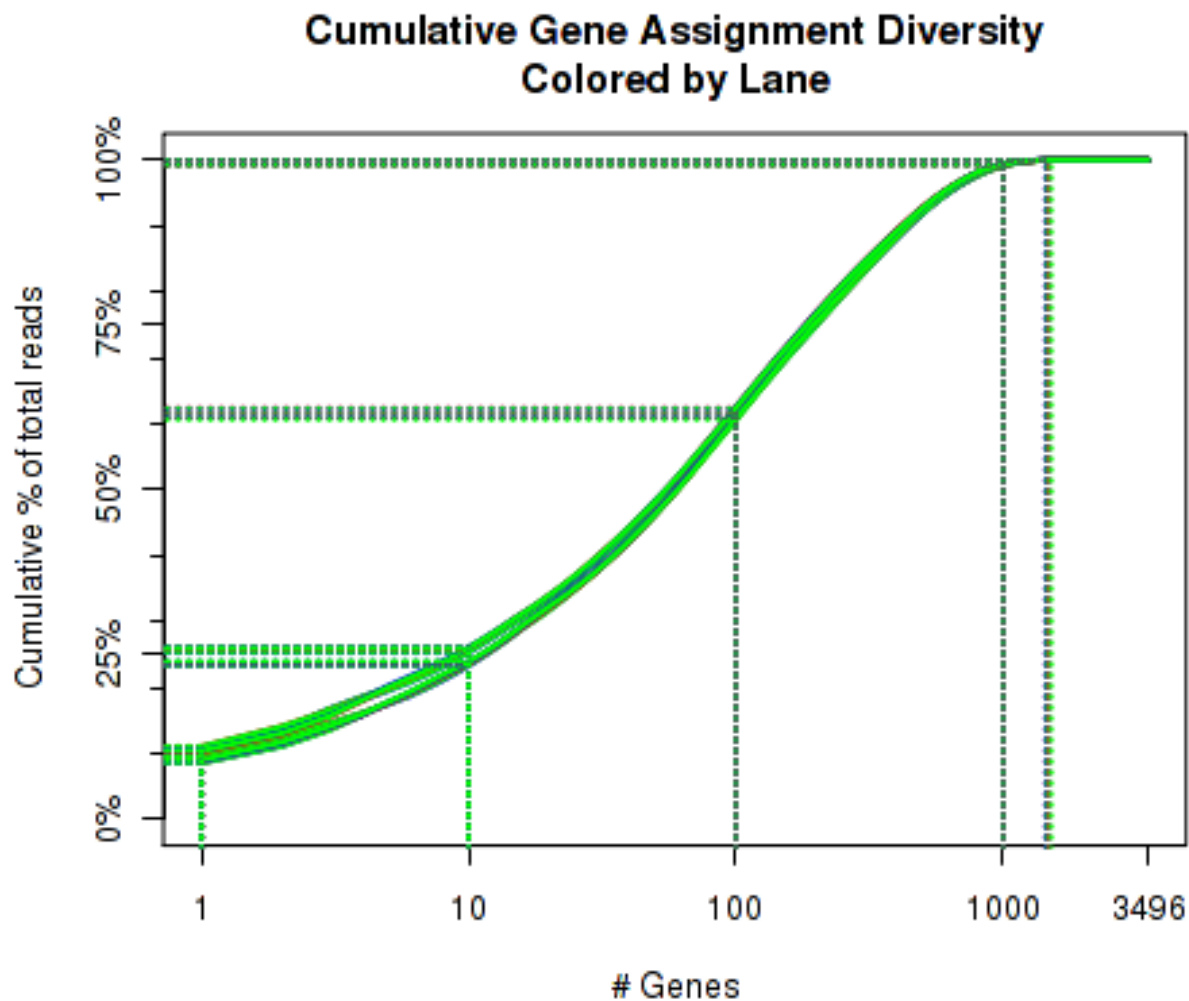


Figure 16: Cumulative Gene Diversity

For each replicate, Figure 16 displays the cumulative gene diversity. For each replicate, the genes are sorted by read-count. Then, a cumulative function is calculated for the percent of the total proportion of reads as a function of the number of genes. Intercepts are plotted as well, showing the cumulative percent for 1 gene, 10 genes, 100 genes, 1000 genes, and 10000 genes.

So, for example, across all the replicate, around 50 to 55 percent of the read-pairs were found to map to the top 1000 genes. Around 20 percent of the reads were found in the top 100 genes. And so on.

This can be used as an indicator of whether a large proportion of the reads stem from of a small number of genes.

Note that this is restricted to only the reads that map to a single unique gene. Reads that map to more than one gene, or that map to intronic or intergenic areas are ignored.

This plot can be generated individually with the command:

```
makePlot.gene.cdf(byLane.plotter);
```


7.3.10 Nucleotide Rates, by Cycle

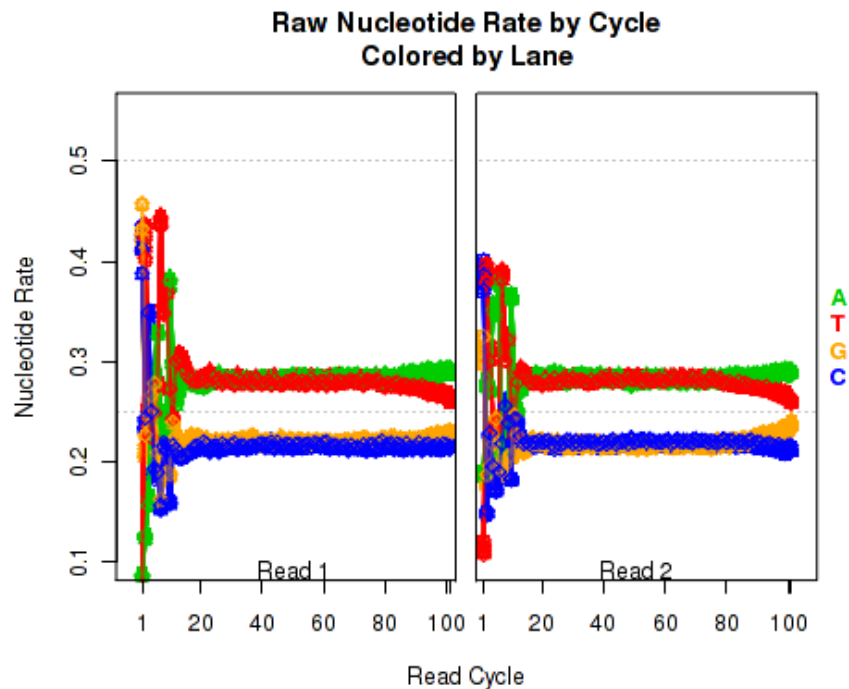


Figure 17: Nucleotide rates, by cycle

For each replicate, Figure 17 displays the rate at which each nucleotide appears (y-axis), as a function of the position in the read (x-axis). The color scheme for NVC plots is different from the other plots. Rather than being used for emphasis or to allow cross-comparisons by sample, biological-condition, or lane, the colors are used to indicate the four nucleotides: A (green), T (red), G (orange), or C (blue). Depending on the type of plotter being used, sample-runs will be marked and differentiated by marking the lines with shapes (R points). In many cases the points will be unreadable due to overplotting, but clear outliers that stray from the general trends can be readily identified.

When used with a "sample.highlight" type plotter (see 7.2.5), "highlighted" samples will be drawn with a deeper shade of the given color.

This plot displays the "raw" nucleotide rates, including bases that are soft-clipped by the aligner.

This plot can be generated individually with the command:

```
makePlot.raw.NVC(byLane.plotter);
```

7.3.11 Aligned Nucleotide Rates, by Cycle

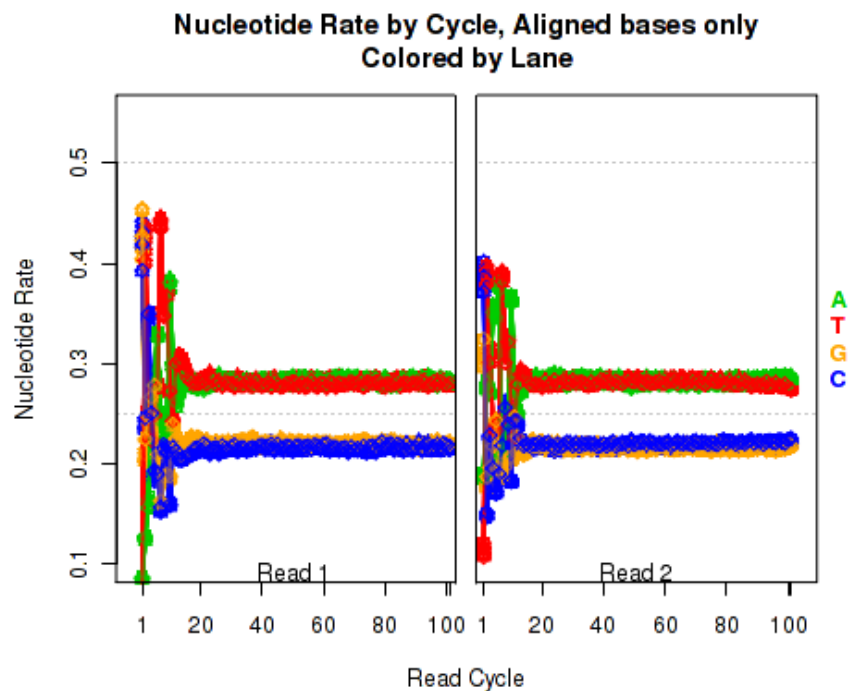


Figure 18: Aligned nucleotide rates, by cycle

Figure 18 is identical to Figure 17 (described in section 7.3.10), except that it only counts bases that are *not* soft clipped off by the aligner.

This plot can be generated individually with the command:

```
makePlot.minus.clipping.NVC(byLane.plotter);
```

7.3.12 Leading Clipped Nucleotide Rates

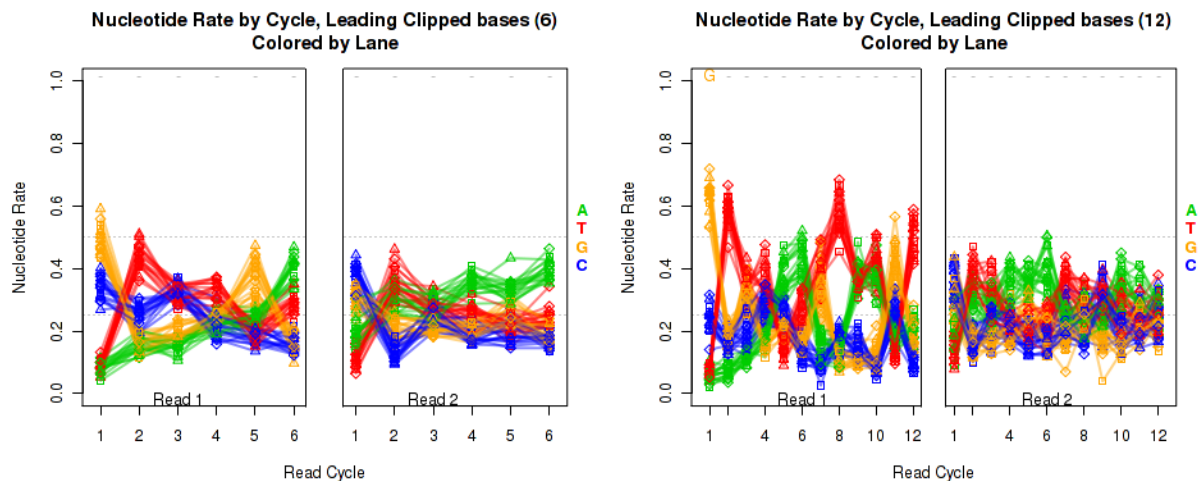


Figure 19: Leading-clipped nucleotide rates

The left plot in Figure 19 displays the nucleotide rate (y-axis) as a function of read position (x-axis), for the first 6 bases of reads in which *exactly* 6 bases were clipped off the 5' end. The right plot displays the nucleotide rate (y-axis) as a function of read position (x-axis), for the first 12 bases of reads in which *exactly* 12 bases were clipped off the 5' end.

This plot can be generated individually with the command:

```
makePlot.NVC.lead.clip(byLane.plotter, clip.amt = 6);
makePlot.NVC.lead.clip(byLane.plotter, clip.amt = 12);
```

Any integer can be used as the clip.amt value.

7.3.13 Trailing Clipped Nucleotide Rates

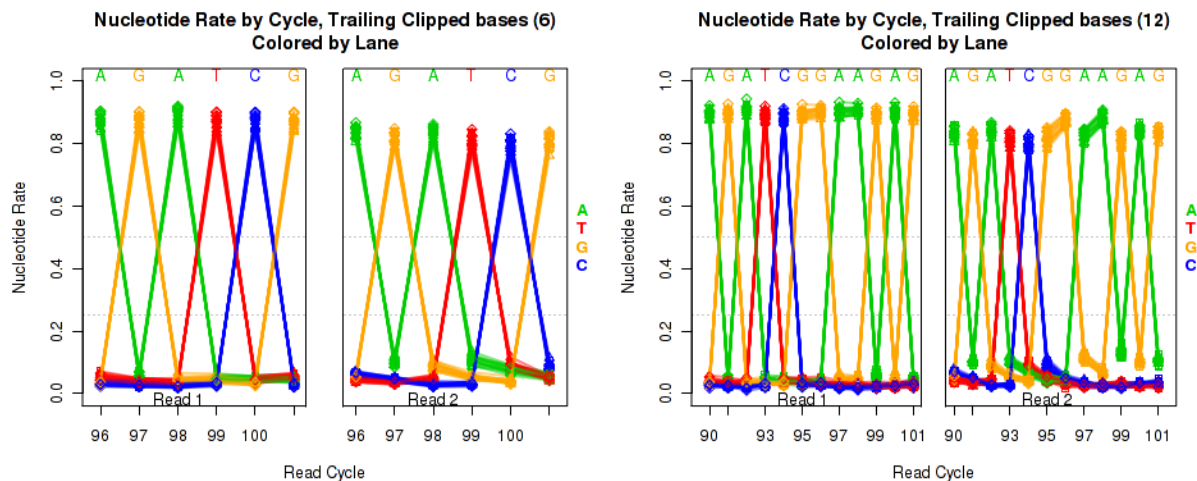


Figure 20: Trailing-clipped nucleotide rates

The left plot in Figure 20 displays the nucleotide rate (y-axis) as a function of read position (x-axis), for the last 6 bases of reads in which *exactly* 6 bases were clipped off the 3' end. The right plot displays the nucleotide rate (y-axis) as a function of read position (x-axis), for the last 12 bases of reads in which *exactly* 12 bases were clipped off the 3' end.

Note concerning the example data: In the example dataset an extremely strong trend is easily visible. The specific sequence observed matches that of the sequencing adapter used. The pattern appears in reads coming from fragments that are smaller than the read length. In these cases, the 3' end of each read will continue into the adapter sequence after sequencing the entire template fragment. Thus: for the left and right plots the sequence comes from reads with an insert size of exactly 95 and 89, respectively (ie 101 base pairs minus 6 or 12).

These plots can be generated individually with the command:

```
makePlot.NVC.tail.clip(byLane.plotter, clip.amt = 6);
makePlot.NVC.tail.clip(byLane.plotter, clip.amt = 12);
```

Any integer can be used as the clip.amt value.

7.3.14 Mapping location rates

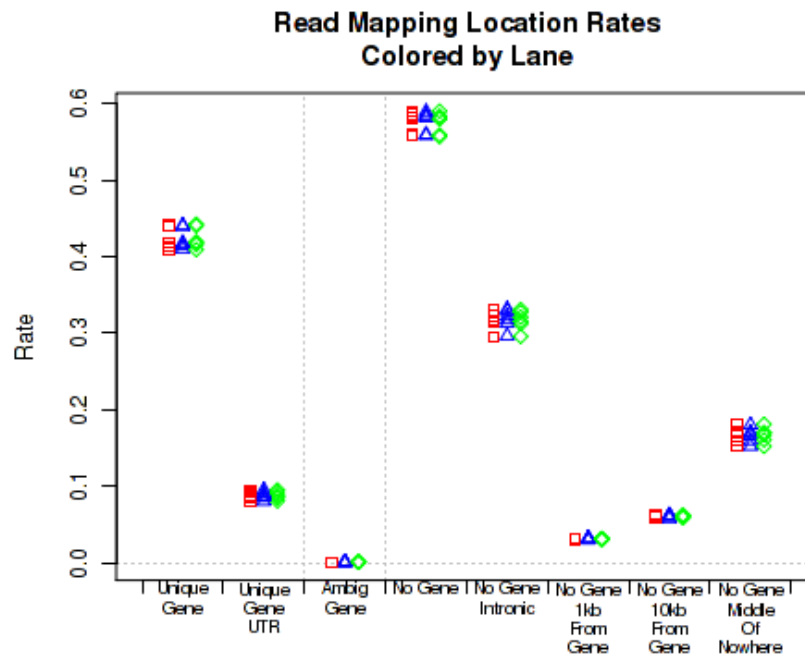


Figure 21: Gene assignment rates

For each replicate, Figure 21 displays the rate (y-axis) for which the replicate's read-pairs are assigned to the given categories.

The categories are:

- *Unique Gene*: The read-pair overlaps with the exonic segments of one and only one gene. For many downstream analyses tools, such as DESeq, DESeq2 [1] and *EdgeR* [2], only read-pairs in this category are used.
- *Ambig Gene*: The read-pair overlaps with the exons of more than one gene.
- *No Gene*: The read-pair does not overlap with the exons of any annotated gene.
- *No Gene, Intronic*: The read-pair does not overlap with the exons of any annotated gene, but appears in a region that is bridged by an annotated splice junction.
- *No Gene, 1kb from gene*: The read-pair does not overlap with the exons of any annotated gene, but is within 1 kilobase from the nearest annotated gene.
- *No Gene, 10kb from gene*: The read-pair does not overlap with the exons of any annotated gene, but is within 10 kilobases from the nearest annotated gene.
- *No Gene, middle of nowhere*: The read-pair does not overlap with the exons of any annotated gene, and is *more* than 10 kilobases from the nearest annotated gene.

This plot can be generated individually with the command:

```
makePlot.gene.assignment.rates(byLane.plotter);
```

7.3.15 Splice Junction Loci

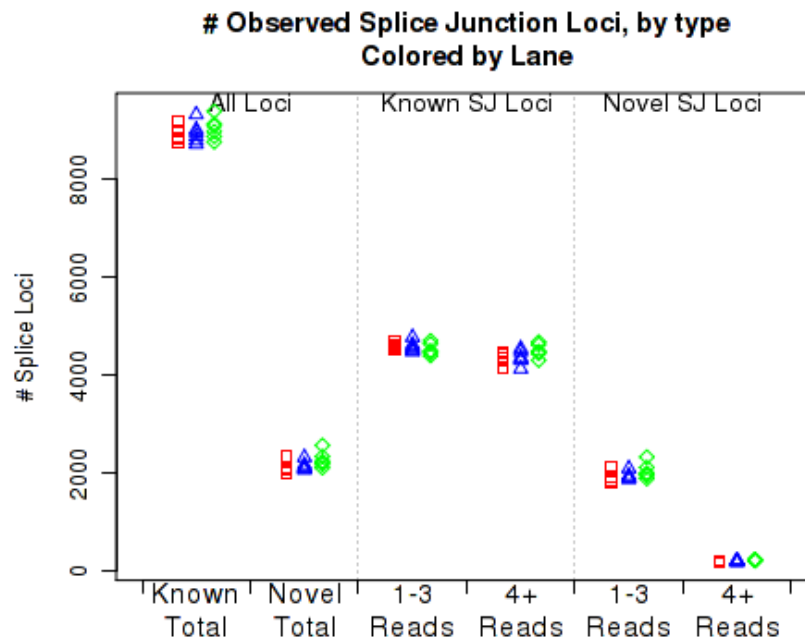


Figure 22: Splice junction loci

For each replicate, Figure 22 displays the number (y-axis) of splice junction *loci* of each type that appear in the replicate's reads. Splice junctions are split into 4 groups, first by whether the splice junction appears in the transcript annotation gtf ("known" vs "novel"), and then by whether the splice junction has 4 or more reads covering it, or 1-3 reads ("Hi" vs "Lo").

This plot can be used to detect a number of anomalies. For example: whether mapping or sequencing artifacts caused a disproportionate discovery of novel splice junctions in one sample or batch.

This plot can be generated individually with the command:

```
makePlot.splice.junction.loci.counts(byLane.plotter);
```

7.3.16 Number of Splice Junction Events

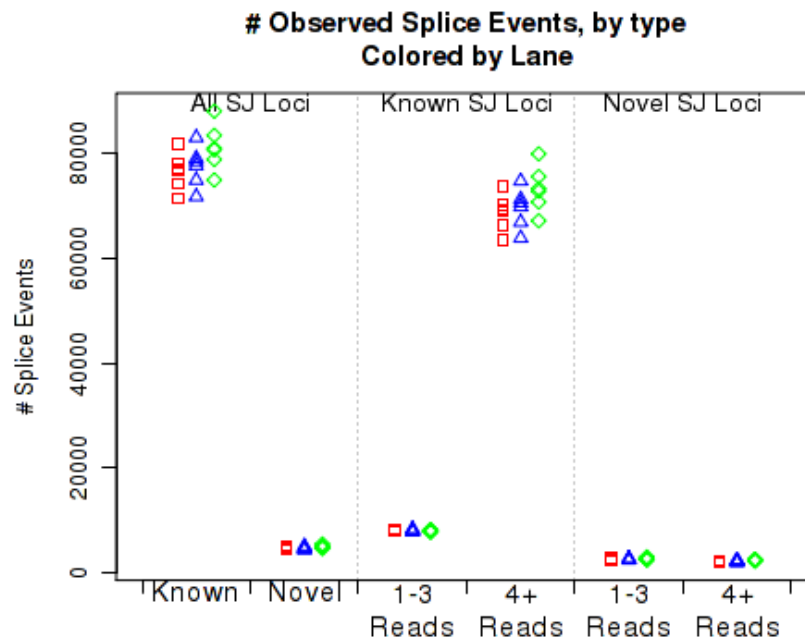


Figure 23: Number of splice junction events

For each replicate, Figure 23 displays the number (y-axis) of all splice junction events falling into each of six categories. A splice junction "event" is one instance of a read-pair bridging a splice junction. Some reads may contain multiple splice junction events, some may contain none. If a splice junction appears on both reads of a read-pair, this is still only counted as a single "event".

The six categories are:

- Known: The splice junction locus is found in the supplied transcript annotation gtf file.
- Novel: The splice junction locus is NOT found in the supplied transcript annotation gtf file.
- Known, 1-3 reads: The locus is known, and is only covered by 1-3 read-pairs.
- Known, 4+ reads: The locus is known, and is covered by 4 or more read-pairs.
- Novel, 1-3 reads: The locus is novel, and is only covered by 1-3 read-pairs.
- Novel, 4+ reads: The locus is novel, and is covered by 4 or more read-pairs.

Note that because different samples/runs may have different total read counts and/or library sizes, this function is generally not the best for comparing between samples. In general, the event rates per read-pair should be used, see the next section, 7.3.17.

This plot is used to detect whether sample-specific or batch effects have a substantial or biased effect on splice junction appearance, either due to differences in the original RNA, or due to artifacts that alter the rate at which the aligner maps across splice junctions.

This plot can be generated individually with the command:


```
makePlot.splice.junction.event.counts(byLane.plotter);
```

7.3.17 Splice Junction Event Rates per Read-Pair

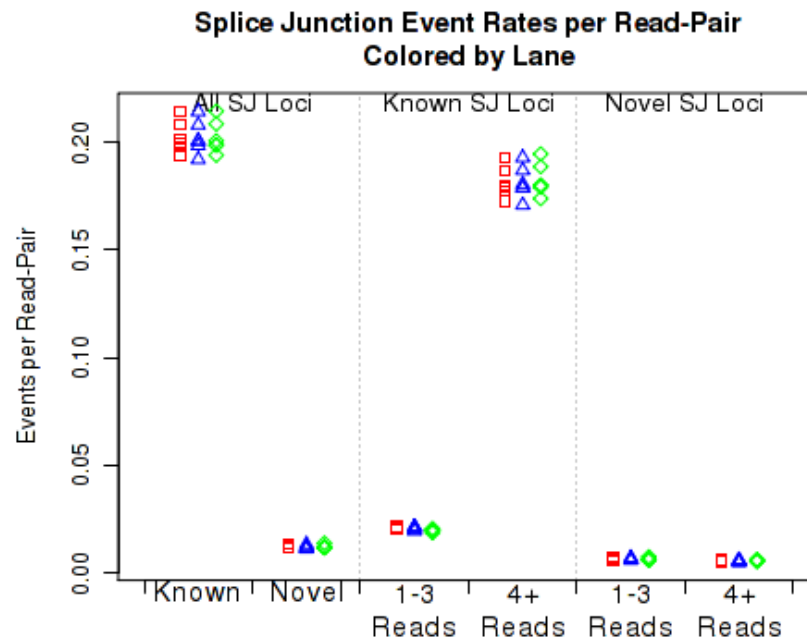


Figure 24: Splice junction events

For each replicate, Figure 24 displays the rate (y-axis) at which each type of splice junction events appear, per read-pair.

This plot is used to detect whether sample-specific or batch effects have a substantial or biased effect on splice junction appearance, either due to differences in the original RNA, or due to artifacts that alter the rate at which the aligner maps across splice junctions.

This plot can be generated individually with the command:

```
makePlot.splice.junction.event.ratesPerRead(byLane.plotter);
```

7.3.18 Breakdown of Splice Junction Events

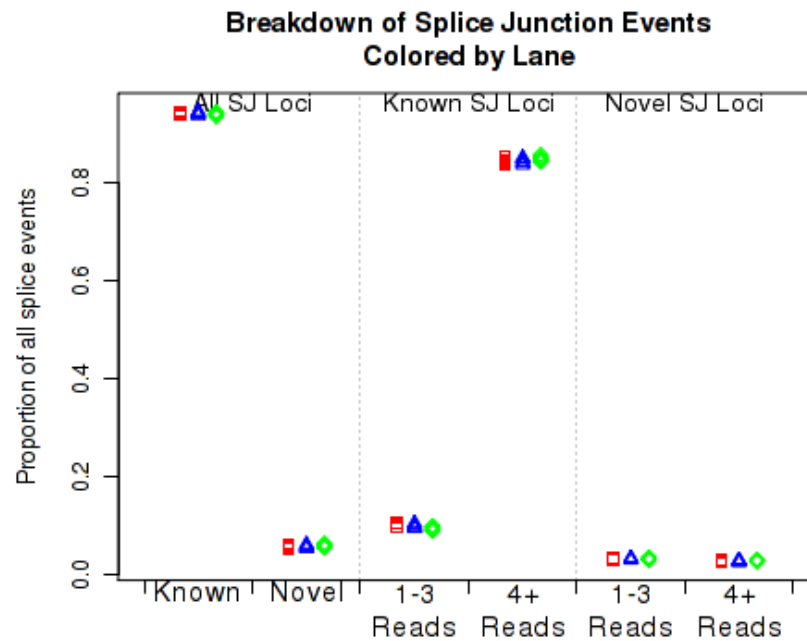


Figure 25: Proportions of splice junction events

For each replicate, Figure 25 displays the proportion (y-axis) of all splice junction events falling into the six categories.

This plot is used to detect whether sample-specific or batch effects have a substantial or biased effect on splice junction appearance, either due to differences in the original RNA, or due to artifacts that alter the rate at which the aligner maps across splice junctions.

This plot can be generated individually with the command:

```
makePlot.splice.junction.event.proportions(byLane.plotter);
```

7.3.19 Breakdown of Splice Junction Events, by locus type

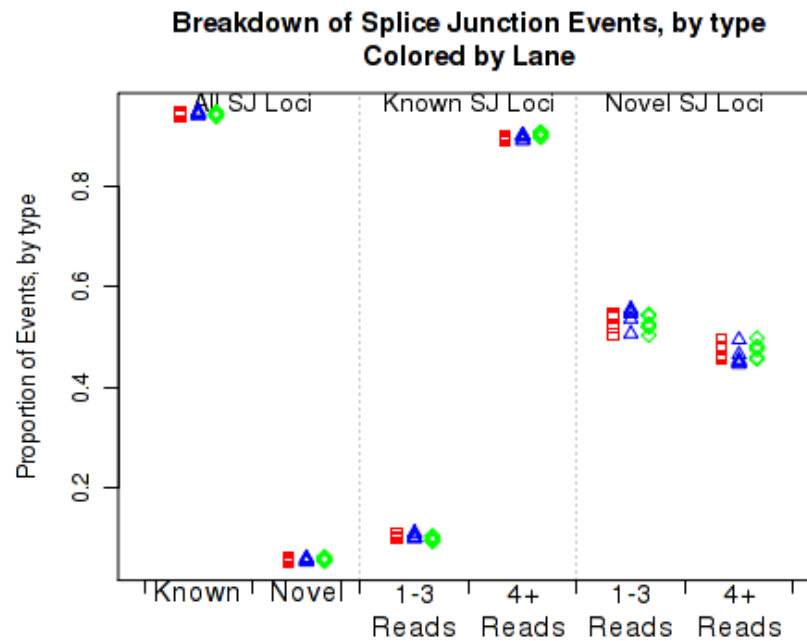


Figure 26: Splice junction events

For each replicate, Figure 26 displays the proportion (y-axis) at which splice junction events appear on known vs novel splice junction loci, the proportion of known splice junction events that occur on low-coverage junctions (1-3 read-pairs) vs high-coverage junctions (4 or more read-pairs), and the proportion of novel splice junction events that occur on low vs high coverage junctions.

This plot is used to detect whether sample-specific or batch effects have a substantial or biased effect on splice junction appearance, either due to differences in the original RNA, or due to artifacts that alter the rate at which the aligner maps across splice junctions.

This plot can be generated individually with the command:

```
makePlot.splice.junction.event.proportionsByType(byLane.plotter);
```

7.3.20 Strandedness test

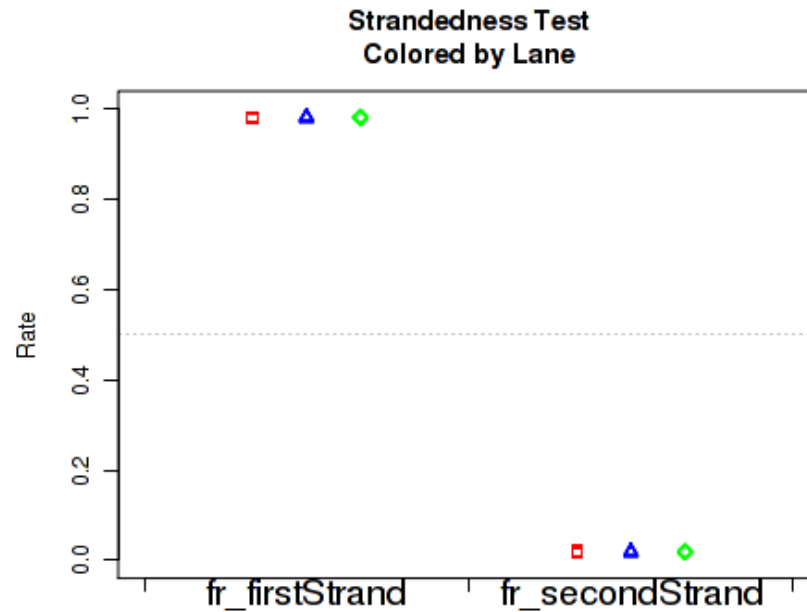


Figure 27: Strandedness

Figure 27 displays the rate at which reads appear to follow the two possible library-type strandedness rules. (See section 6 for more information on stranded library types).

This plot is used to detect whether your data is indeed stranded, and whether you are using the correct stranded data library type option. For unstranded libraries, one would expect all points to fall very close to the 50-50 center line. For stranded libraries, all points should fall closer to 99. If (and only if) all the replicates were run using the same strandedness and library type options, then green target boxes will be drawn around the areas where the points should appear. If points appear substantially outside these boxes, then you may be running QoRTs using the wrong library type options.

This plot can be generated individually with the command:

```
makePlot.strandedness.test(byLane.plotter);
```

7.3.21 Mapping stats

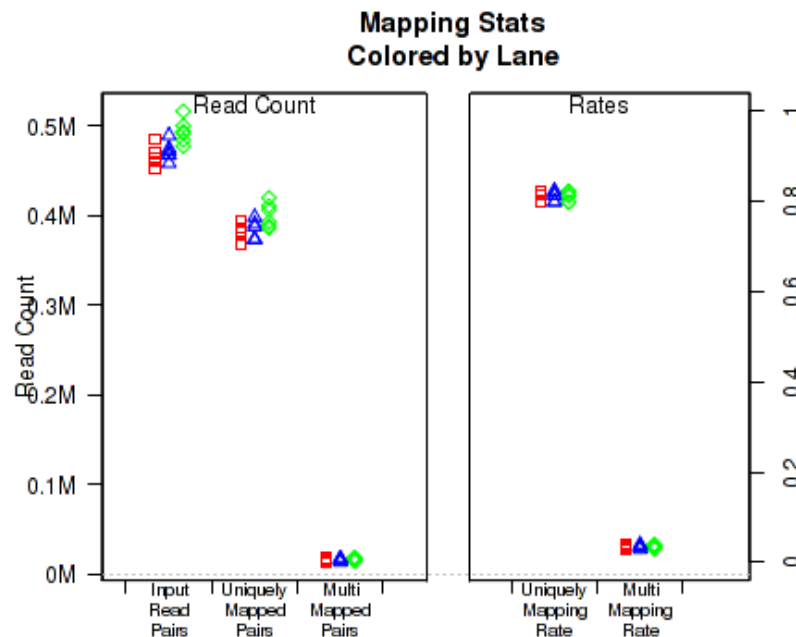


Figure 28: Mapping stats

For each replicate, Figure 28 displays the mapping rates and counts. In order to plot this data, QoRTs must be provided with the pre-alignment read-count for each replicate. There are a number of ways to provide this information to QoRTs. The easiest method is to list it specifically in the replicate decoder (see Section ??). Alternatively, this information can be provided at the initial processing stage (see Section 6), either by setting the input read count explicitly using the `--seqReadCt` parameter, or by providing one of the unaligned fastq files via the `--rawfastq` parameter, in which case the input read count is calculated simply by dividing the number of lines in the fastq file by 4. For paired-end data, only one of the two fastq files needs to be provided, as both will have the same number of reads.

If the dataset contains multi-mapped reads, then numbers and rates of multi-mapping will be included in this plot. If multi-mapped reads were filtered out of the dataset prior to analysis with QoRTs, the multi-mapping rates can still be specified explicitly using the decoder (see Section ??).

This plot can be generated individually with the command:

```
makePlot.mapping.rates(byLane.plotter);
```

If the input read count data is not found, this will generate a placeholder plot.

7.3.22 Chromosome counts

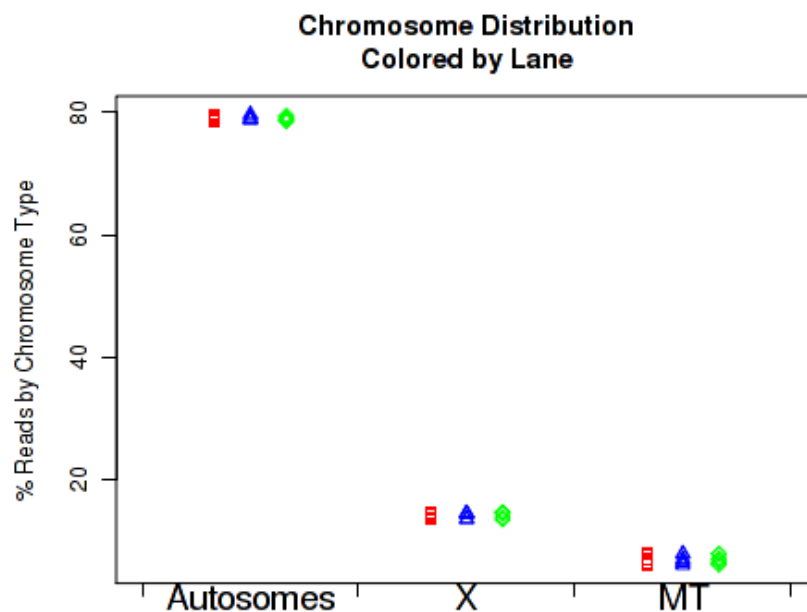


Figure 29: Mapping stats

For each replicate, Figure 29 displays the number of read-pairs mapping to each category of chromosome.

The `chromosome.name.style` must be set to match the style of your chromosome names. By default it assumes the chromosomes are named `chr1`, `chr2`, `chr3`, etc.

For more information, see the help document using the command `help(makePlot.chrom.type.rates)`. This plot can be generated individually with the command:

```
makePlot.chrom.type.rates(byLane.plotter);
```

7.3.23 Normalization Factors

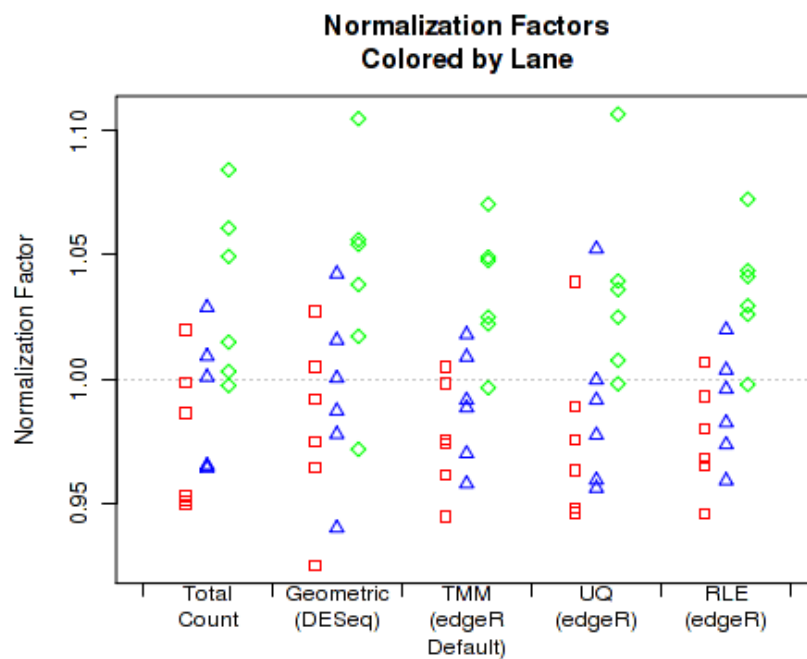


Figure 30: Normalization Factors

For each replicate, Figure 30 displays the normalization factors.

By default, QoRTs will automatically detect whether *DESeq2* and *edgeR* are installed and will use these tools to calculate their respective normalization size factors. If neither package is found, then it will only plot the total count normalization.

This plot can be generated individually with the command:

```
makePlot.norm.factors(byLane.plotter);
```


7.3.24 Normalization Factors

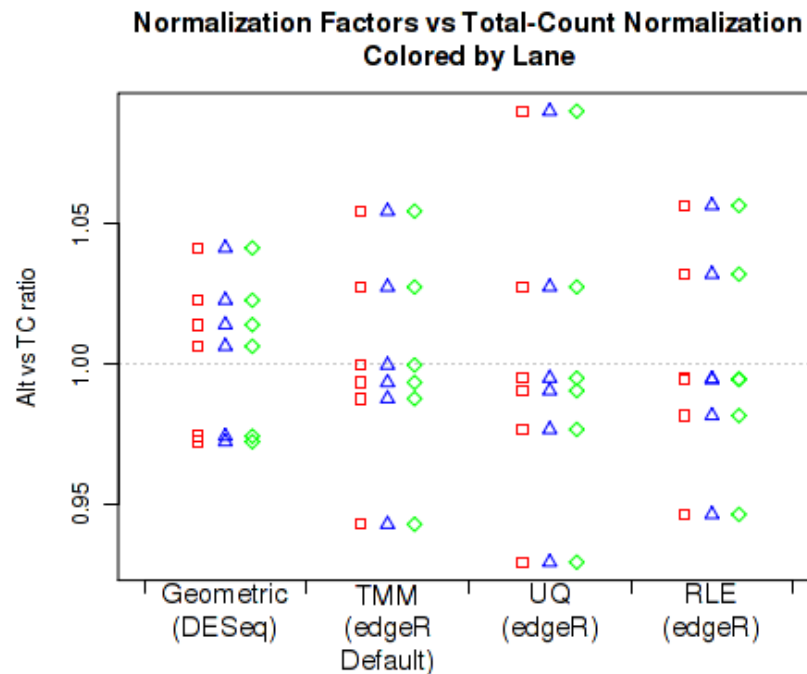


Figure 31: Normalization Factors vs TC

For each replicate, Figure 31 displays the ratio of the alternate normalization factors to the Total Count normalization factors.

By default, QoRTs will automatically detect whether *DESeq2* and *edgeR* are installed and will use these tools to calculate their respective normalization size factors. If neither package is found, then it will only plot the total count normalization.

This plot can be generated individually with the command:

```
makePlot.norm.factors.vs.TC(byLane.plotter);
```

7.3.25 Read drop rate

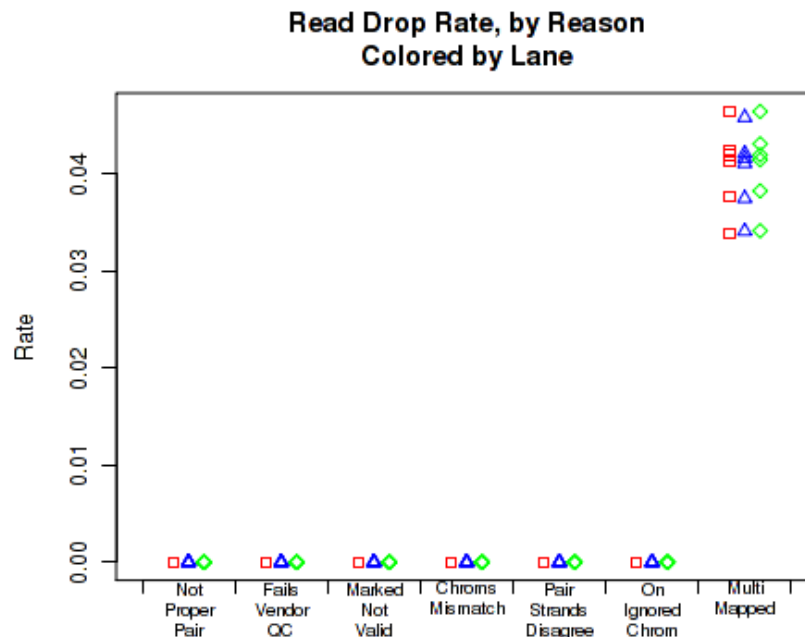


Figure 32: Drop rates

For each replicate, Figure 32 displays the rates and reasons for reads being dropped from QC analysis. Note that in the example dataset reads were never dropped. This is a consequence of the pre-processing steps in the example pipeline.

This plot can be generated individually with the command:

```
makePlot.dropped.rates(byLane.plotter);
```

8 Secondary Utilities

In addition to the standard quality-control tools described in the previous sections, QoRTs also includes a number of other minor utilities intended to assist in data visualization, cleaning, and preparation for downstream analyses.

8.1 Generating a flattened annotation file

Before counting exons and splice junctions, QoRTs generates a set of non-overlapping exonic fragments out of all the exons in the genome annotation gtf file. It then assigns each exonic fragment a unique identifier. Similarly, it assigns every splice junction its own unique identifier. A gtf file listing all these genomic features and their unique identifiers can be created using the following command:

```
java -jar /path/to/jarfile/QoRTs.jar makeFlatGtf
                                input.gtf
                                flattened.gff
```

Both the input and output annotation files can be either .zip or .gz compressed. Compression is autodetected from the file extension.

strandedness: You must use the `--stranded` option to create the flattened gff for use with stranded datasets. DO NOT mix stranded flattened gff with unstranded data, or vice versa.

DEXSeq: DEXSeq also requires a flattened annotation file, which is formatted similarly. In order to produce a flattened gff file that DEXSeq can read, include the `--DEXSeqFmt` option.

This gtf file conforms to the UCSC gff file definition, (found here: <http://genome.ucsc.edu/FAQ/FAQformat.html>). It will contain 4 different feature types (column 3): "aggregate_gene", "exonic_part", "splice_site", and "novel_splice_site".

8.2 Merging Count Data

For the purposes of quality control it is generally preferable to run QoRTs on each sample-run individually, so that potential technical artifacts related to sequencing run or lane can be identified. However, for most downstream purposes these "technical replicates" will be combined and treated as a single sample. Differential expression tools like *DESeq*, *DESeq2* [1], *DEXSeq* [3], and *EdgeR* [7] assume that each set of gene counts (or exon counts, for *DEXSeq*) is derived from a different biological sample.

Thus, the java utility includes a function for quickly and easily calculating merged sample-wise counts:

```
java -jar /path/to/jarfile/QoRTs.jar mergeAllCounts
                                decoder.txt
                                /path/to/qc/results/dir/
                                ./merged/
```

This decoder MUST contain the unique.ID and sample.ID columns.

Alternatively, the merger can be performed for a single sample directly, via the command:

```
java -jar /path/to/jarfile/QoRTs.jar mergeCounts
    ./SAMP1_RG1/,./SAMP1_RG2/,./SAMP1_RG3/
    ./merged/SAMP1/
```

The list of QC data directories must be separated by commas and contain no whitespace.

More information and a full accounting of all parameters and options can be found in the online documentation⁷, or by using the commands:

```
java -jar /path/to/jarfile/QoRTs.jar mergeAllCounts--man
and
java -jar /path/to/jarfile/QoRTs.jar mergeCounts --man
```

8.3 Generating genome browser tracks

In addition to the standard QC plots, which examine the data as a whole, it is sometimes desirable to be able to query and examine coverage information at specific genetic loci. In particular, when

⁷Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

identifying candidate genes via genome-wide analyses, it is often vital to examine the locus for artifacts before carrying out costly and time-consuming validation experiments.

Figure 33 is just one example of the tracks that can be produced with QoRTs. A full description of how this particular splice junction track can be generated can be found in the example dataset walkthrough, which is linked to on the QoRTs github site.

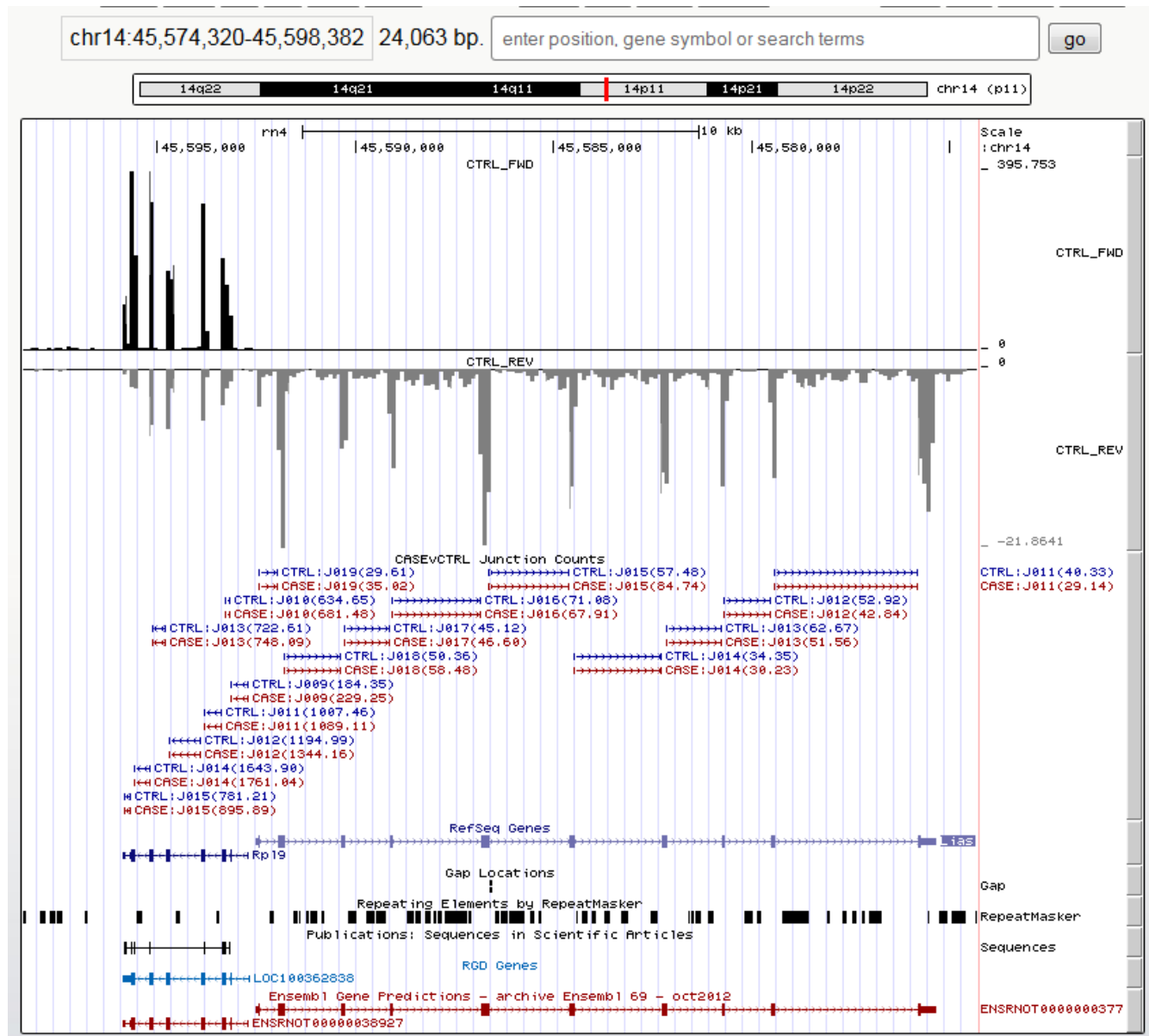


Figure 33: An example of some of the tracks that can be generated via QoRTs. The first track ("CTRL_FWD") shows the mean normalized coverage depth for 100-base-pair windows for the forward (genomic) strand. The second track reveals the coverage depth on the negative strand (plotted as negative values). The third track ("CASEvCTRL Junction Counts") displays the locus ID for each splice junction, along with the mean normalized coverage bridging the splice junction loci in cases and controls. Note that cases and controls are colored distinctly, for ease of use.

8.3.1 Generating wiggle tracks

QoRTs includes a utility to generate ".wig", or "wiggle plot" files. These wiggle plot files include counts for the mean coverage for each equal-sized window across the whole genome. These files are designed to be used with the UCSC browser or similar interfaces, and allow easy and intuitive visualization of your data.

```
java -jar /path/to/jarfile/QoRTs.jar bamToWiggle
                                     infile.bam
                                     trackName
                                     chromLengthFile
                                     outfilePrefix
```

The chromLengthFile is a simple tab-delimited text file that includes each chromosome in the first column and the chromosome's length (in base-pairs) in the second column. If the wiggle file is intended for use with a standard genome on the UCSC genome browser, then the UCSC utility fetchChromSizes should be used to generate this file. (see <http://genome.ucsc.edu/goldenPath/help/bigWig.html> for more information on fetchChromSizes, as well as information on how to compress your wig files into smaller and more efficient bigWig files). It also takes

Common options and flags for this function include:

- --sizefactor 1.0: A float value. All the coverage values will be divided by this factor. Useful for comparing two samples that may have different normalization factors.
- --stranded: Flag to indicate that data should be treated as stranded.
- --fr_secondStrand: Flag to indicate that the data is of the fr_secondstrand stranded library type. (See section 6 for more information on the two stranded library types).
- --negativeReverseStrand: If this flag is set, then the negative strand will be counted in negative numbers. This can be useful for plotting both strands in a single multiwig track, via a trackhub. (see <http://genome.ucsc.edu/goldenPath/help/trackDb/trackDbDoc.html>)
- --simpleCountByRead: If this flag is raised, then each read of each read-pair will be counted separately. Thus the wiggle plot will count simple read coverage depth rather than read-pair coverage depth. This means that when read-pairs overlap they will be counted twice over the overlapping region.

Many of the other parameters are identical to those used by the QC tool: --singleEnded, --coordSorted, etc.

More information and a full accounting of all parameters and options can be found in the online documentation⁸, or by using the command:

```
java -jar /path/to/jarfile/QoRTs.jar bamToWiggle --man
```

8.3.2 Merging wiggle tracks

QoRTs includes a utility for summing or averaging multiple wiggle files, either with or without normalization factors. For example, to calculate the normalized mean coverage for each 100-bp window across all CASE samples in the example dataset:

```
java -jar /path/to/jarfile/QoRTs.jar mergeWig
```

⁸Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

```
--calcMean \
--filenames outputData/countTables/SAMP1/QC.wiggle.fwd.wig.gz,outputDa
--sizeFactors 1.057995,0.999932,1.015372 \
path/to/output/CASE.fwd.wig.gz
```

There are a number of other allowed parameterizations. The `--sampleList` parameter (which can be either a comma-delimited list or a list .txt file) can be used along with the `--infilePrefix` and `--infileSuffix` to specify the file names if all of the wiggle files are in the same parent directory. If the `--sizeFactors` is omitted then the non-normalized means will be calculated.

Common options and flags for this function include:

- `--calcMean`: If this flag is raised, the utility will calculate the average rather than the total coverage for each window.
- `--makeNegative`: If this flag is raised, the output will be multiplied by -1.
- `--sizeFactorFile`: A file containing at least two columns, the sample.ID and the size.factor. This file must include all samples in the sample list, but can include other samples that are not included in the sample list.

More information and a full accounting of all parameters and options can be found in the online documentation⁹, or by using the command:

```
java -jar /path/to/jarfile/QoRTs.jar mergeWig --man
```

8.3.3 Generating splice-junction tracks

To visualize splice junction data, QoRTs can produce bed files that show splice junction counts:

```
java -jar /path/to/jarfile/QoRTs.jar makeJunctionTrack
--stranded \
--filenames outputData/countTables/SAMP1/QC.spliceJunctionAndExonCount
flattened.gff \
path/to/output/CASE.bed.gz
```

Common options and flags for this function include:

- `-rgb`: The color to use for each bed entry.

8.3.4 Merging splice-junction tracks

More information and a full accounting of all parameters and options can be found in the online documentation¹⁰, or by using the command:

```
java -jar /path/to/jarfile/QoRTs.jar makeSpliceBed --man
```

8.3.5 Generating splice-junction tracks

Merged splice junction tracks can be created using the same utility used to create single-sample splice junction tracks. This uses syntax similar to the syntax used for merging wiggle files:

```
java -jar /path/to/jarfile/QoRTs.jar makeJunctionTrack
--calcMean \
--stranded \
```

⁹Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

¹⁰Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

```
--filenames outputData/countTables/SAMP1/QC.spliceJunctionAndExonCount
--sizeFactors 1.057995,0.999932,1.015372 \
flattened.gff \
path/to/output/CASE.bed.gz
```

Common options and flags for this function include:

- *rgb*: The color to use for each bed entry.

8.3.6 Merging splice-junction tracks

More information and a full accounting of all parameters and options can be found in the online documentation¹¹, or by using the command:

```
java -jar /path/to/jarfile/QoRTs.jar makeSpliceBed --man
```

8.4 Importing data into other tools

In addition to providing quality control information, QoRTs also provides the requisite input files needed for the DESeq/DESeq2 [1], *DEXSeq* [3], and *EdgeR* [2,7,8] analysis tools. These files will be identical to those that would be generated by HTSeq (using the default "union rule" option). All the data files can be found in the qc.data.dir directory. The files for use with *DESeq*, *DESeq2*, and *EdgeR* will be named QC.geneCounts.formatted.for.DESeq.txt.gz and the files for use with *DEXSeq* will be named QC.exonCounts.formatted.for.DEXSeq.txt.gz

A note on the DEXSeq counts: The DEXSeq counts may not be perfectly identical to those produced by the DEXSeq dexseq_prepare_annotation.py and dexseq_count.py scripts. There are two reasons for these differences, both relating to the treatment of aggregate genes. The first reason is minor: Aggregate genes will be named slightly differently. When multiple genes overlap with one another, DEXSeq produces "aggregate genes" which include all transcripts for all these overlapping genes. It names the aggregate gene using the set of genes in the aggregate, delimited with + characters. Unfortunately, the genes are not listed in any particular order. Thus, it is not possible for QoRTs to produce the exact same aggregate gene names. QoRTs lists the contained genes in lexicographic order.

Secondly, for UNSTRANDED data, the QoRTs and DEXSeq annotation "flattening" step will behave slightly differently under default conditions. DEXSeq's dexseq_prepare_annotation.py script always operates in "stranded" mode, and explicitly distinguishes between genes on opposing strands. QoRTs, on the other hand, prepares the annotation in stranded or unstranded modes. If the DEXSeq-style behavior is desired, a stranded flat gff file can be produced by the makeFlatGff utility, then passed explicitly to the QoRTs QC utility running in non-stranded mode using the --flatgff parameter. This will override the default (and recommended) behavior in which the flattened gff will use the same stranded rule as the counting utility.

In general, testing aggregate genes in DEXSeq is not recommended, as the two genes are likely to be independently regulated and will likely produce false positives. For most purposes, it is preferable to drop such genes from the count tables prior to DEXSeq analysis.

¹¹Found here: <https://dl.dropboxusercontent.com/u/103621176/QoRTs/helpDocs/jarHtml/index.html>

9 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>, <http://arxiv.org/abs/http://bioinformatics.oxfordjournals.org/cgi/reprint/23/21/2881.pdf> 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.
- [4] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013. URL: <http://bioinformatics.oxfordjournals.org/content/29/1/15.abstract>, <http://arxiv.org/abs/http://bioinformatics.oxfordjournals.org/content/29/1/15.full.pdf+html> arXiv:<http://bioinformatics.oxfordjournals.org/content/29/1/15.full.pdf+html>, doi:10.1093/bioinformatics/bts635.
- [5] Thomas D. Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 2010. URL: <http://bioinformatics.oxfordjournals.org/content/26/7/873.abstract>, <http://arxiv.org/abs/http://bioinformatics.oxfordjournals.org/content/26/7/873.full.pdf+html> arXiv:<http://bioinformatics.oxfordjournals.org/content/26/7/873.full.pdf+html>, doi:10.1093/bioinformatics/btq057.
- [6] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven Salzberg. Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, 14(4):R36, 2013. URL: <http://genomebiology.com/2013/14/4/R36>, <http://dx.doi.org/10.1186/gb-2013-14-4-r36> doi:10.1186/gb-2013-14-4-r36.
- [7] McCarthy DJ Robinson MD and Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139–140, 2010.
- [8] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40:4288–4297, 2012.

10 Session Information

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

```
sessionInfo()  
## R version 3.1.1 (2014-07-10)  
## Platform: x86_64-unknown-linux-gnu (64-bit)
```



```
##
## locale:
## [1] C
##
## attached base packages:
## [1] parallel stats4 stats graphics grDevices utils datasets
## [8] methods base
##
## other attached packages:
## [1] edgeR_3.8.2 limma_3.22.1 DESeq2_1.6.1
## [4] RcppArmadillo_0.4.500.0 Rcpp_0.11.3 GenomicRanges_1.18.1
## [7] GenomeInfoDb_1.2.2 IRanges_2.0.0 S4Vectors_0.4.0
## [10] BiocGenerics_0.12.0 QoRTsExampleData_0.1.10 QoRTs_0.1.13
## [13] Cairo_1.5-6 knitr_1.7
##
## loaded via a namespace (and not attached):
## [1] AnnotationDbi_1.28.1 BBmisc_1.8 BatchJobs_1.5
## [4] Biobase_2.26.0 BiocParallel_1.0.0 BiocStyle_1.4.1
## [7] DBI_0.3.1 Formula_1.1-2 Hmisc_3.14-5
## [10] MASS_7.3-35 RColorBrewer_1.0-5 RSQLite_1.0.0
## [13] XML_3.98-1.1 XVector_0.6.0 acepack_1.3-3.3
## [16] annotate_1.44.0 base64enc_0.1-2 brew_1.0-6
## [19] checkmate_1.5.0 cluster_1.15.3 codetools_0.2-9
## [22] colorspace_1.2-4 digest_0.6.4 evaluate_0.5.5
## [25] fail_1.2 foreach_1.4.2 foreign_0.8-61
## [28] formatR_1.0 genefilter_1.48.1 geneplotter_1.44.0
## [31] ggplot2_1.0.0 grid_3.1.1 gtable_0.1.2
## [34] highr_0.4 iterators_1.0.7 lattice_0.20-29
## [37] latticeExtra_0.6-26 locfit_1.5-9.1 munsell_0.4.2
## [40] nnet_7.3-8 plyr_1.8.1 proto_0.3-10
## [43] reshape2_1.4 rpart_4.1-8 scales_0.2.4
## [46] sendmailR_1.2-1 splines_3.1.1 stringr_0.6.2
## [49] survival_2.37-7 tools_3.1.1 xtable_1.7-4
```

11 Legal

This software is "United States Government Work" under the terms of the United States Copyright Act. It was written as part of the authors' official duties for the United States Government and thus cannot be copyrighted. This software is freely available to the public for use without a copyright notice. Restrictions cannot be placed on its present or future use.

Although all reasonable efforts have been taken to ensure the accuracy and reliability of the software and data, the National Human Genome Research Institute (NHGRI) and the U.S. Government does not and cannot warrant the performance or results that may be obtained by using

this software or data. NHGRI and the U.S. Government disclaims all warranties as to performance, merchantability or fitness for any particular purpose.

In any work or product derived from this material, proper attribution of the authors as the source of the software or data should be made, using "NHGRI Genome Technology Branch" as the citation.

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
```