What it means and what to look for:

In general, when examining these plots, users should scan for a number of potential anomalies:

- **"Spikes"**: In which one of the metrics jumps up or down abruptly, then returns to baseline.
- **"Shelves"**: In which one of the metrics jumps up or down abruptly, then continues at an increased or decreased level.
- **Outliers**: In which one or more particular samples or replicates are very different from all or most of the others. The presence of outliers may be an indicator for sample-collection, library-prep, or sequencing errors or artifacts.
- **Systematic Biases**: In which consistent differences appear between subsets of the data (eg. "lane.ID" or "group.ID"). Many of the biases measured by QoRTs are well-characterized, and many downstream analysis tools are robust against them when they are consistent and uniform. However: biases that vary disproportionately between sample groups may still drive false associations downstream.
- **Inconsistent samples**: In which the technical replicates of a specific biological sample shows substantial variation. In most studies, technical variation is very small relative to biological variation. In the example dataset (for example) the technical replicates are plotted almost on top of one another across many of the plots. If technical replicates do not cluster tightly, or if they cluster with the wrong replicates, then this may be an indicator of a sample swap.

Some "anomalous" metrics may be fundamental to the dataset and may not be indicative of any quality issues. For example, when profiling two different cell types one would expect the two groups to have very different profiles across a number of metrics. However: a single sample that is wildly different from the others within the same group may be cause for concern. In many cases variations may be observed across multiple metrics, all driven by the same underlying phenomenon. The breadth and depth of the metrics provided are intended to provide the tools necessary to identify the most likely underlying source(s) of the aberration(s).

Some aberrations may not be relevant to the study analysis even when they are representative of a real data quality issue. A moderate increase in the deletion rate may not have a noticable impact on expression quantification in a simple differential expression study. The same issue, however, might be catastrophic in a study focused on quantifying the rate of RNA transcription errors or RNA editing events. The number of combinations of study design, sample set structure, sequencing technology, and observed quality control issues are myriad, and all of these factors would inform decisionmaking.

Ultimatiely, bioinformaticians must use their own judgement in deciding how to proceed when unexplained abnormalities are discovered in their dataset.

What it means and what to look for: These plots can be used to detect sequencer problems, bad lanes, or similar hardware-level artifacts and errors. Look for spikes or shelves, and ensure that the quality score is relatively consistent across samples and lanes, and that any differences that do exist are not disproportionate with respect to the study condition (or "group.ID").

What it means and what to look for: GC bias has been indicated as a potential driver of false discoveries. Under certain circumstances the GC bias may vary by batch or by sample. If this is apparent in your dataset, particularly if it is associated with study conditions, one may need to apply a GC bias correction method (such as CQN).

What it means and what to look for: Abnormalities ("spikes", "shelves", or cases where sample groups are visibly different) in these plots can be caused by adaptor sequencing, gene fusions, mutations, population stratification, or differences in insert size.

What it means and what to look for: These plots are most often used in conjuction with the plots in Section 7.4.5. Among other things these plots can reveal sequencer errors: a sequencer cycle-skip might result in a spike in the deletion rate at a particular cycle, whereas an incomplete wash may result in a spike in the insertion rate. These plots may also reveal biological differences like population stratification (eg: a sub-population disproportionately mismatches the reference genome), or broad genome rearrangements/editing (eg: in cancer cells).

In the example dataset, apparent spikes are plainly visible, stemming from deletions in one short and highly-expressed mitochondrial gene. The plots are also fairly noisy due to the small number of reads used in the example dataset and the extremely low frequency of deletions/insertions.

What it means and what to look for: These plots are most often used in conjuction with the plots in Section 7.4.4. They can elucidate the nature of any oddities observed in the previous plots. For example: a large spike at one particular length may suggest that an apparent spike may be due simply to an unannotated variant in one particular high-expression gene, although further investigation is likely merited, to confirm that this is the case.

What it means and what to look for: Spikes in the insert size are common, generally the result of short, highly-expressed (often mitochondrial) transcripts. Because the size-selection of many RNA-Seq protocols is somewhat random, it is important to ensure that the resultant size selection is (relatively) consistent, and that variations are not associated with study condition status. If one study group has disproportionately high or low insert size, this could cause fragment bias that could drive the false discovery of differential effects.

What it means and what to look for: A number of potential sequencer issues can cause an abrupt "spike" or "shelf" in this plot. In one real sample assessed by QoRTs by the software author, it was determined that the sequencer camera was slightly offset on one specific cycle of one specific run. All the reads at the bottom or right edges were lost from that cycle forward, causing the rate of "N" calls to increase more than a hundred-fold. Once the problem was recognized the affected reads were identified and removed.

What it means and what to look for: When run on degraded RNA and/or when using poly-A selection, RNA-Seq often tends to have "3' bias", in which read coverage is higher on the 3' end of transcripts. The degree of this 3' bias tends to be dependent on the degree of degradation. Many analysis tools are robust against this issue when it occurs uniformly across the dataset; However, if some samples are substantially more degraded than others then this may cause problems downstream, particularly if RNA degradation is associated to the experimental condition(s). When studying this plot, check to make sure the gene-body coverage is consistant and/or matches your expectations.

Note that the overall gene-body coverage may be strongly influenced by extreme high-coverage genes (and potentially sequence-specific biases on those specific transcripts). Therefore the upper-middle-quartile plot is generally the preferred general metric for assessing overall gene-body coverage.

What it means and what to look for: This plot can reveal a number of phenomena. First of all: if the top few genes dominate a sample (representing a large percentage of the total reads), oddities may appear in many of the other plots produced by QoRTs, as traits specific to these particular genes are dominant over the variation found across the genome.

It can also reveal biological variations: Different cell types or cells that are healthy, dying, or under stress often have very different diversity profiles from one another. If one sample of within a group is an extreme outlier it may suggest that something is wrong with that sample.

Or technical issues: These plots will often reveal inefficiency of hemoglobin or ribosome depletion protocols, and can also clearly reveal low library complexity (indicated by a very small number of genes being represented).

What it means and what to look for: This can reveal sequence-specific biases such as hexamer or primer bias. Additionally, it can reveal adaptor sequencing. Such issues are generally not a problem as long as they are consistent across samples and groups.

What it means and what to look for: This can reveal sequence-specific biases such as hexamer or primer bias. Such issues are generally not a problem as long as they are consistent across samples and groups. Unlike the raw NVC plot, adaptor sequence will generally be absent from this plot as it usually will not align to the reference.

What it means and what to look for: If a large proportion of the reads are shorter than the read length then this can reveal the adaptor sequence.

What it means and what to look for: Outliers in these plots can indicate biological variations or the presence of large mapping problems. They may also suggest the presence of large, highly-expressed, unannotated transcripts or genes.

What it means and what to look for: 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. It can also be used as an indicator of the comprehensiveness the genome annotation. Replicates that are obvious outliers may have sequencing/technical issues causing false detection of splice junctions.

Abnormalities in the splice junction rates are generally a symptom of larger issues which will generally be picked up by other metrics. Numerous factors can reduce the efficacy by which aligners map across splice junctions, and as such these plots become very important if the intended downstream analyses include transcript assembly, transcript deconvolution, differential splicing, or any other form of analysis that in some way involves the splice junctions themselves. These plots can be used to assess whether other minor abnormalities observed in the other plots are of sufficient severity to impact splice junction mapping and thus potentially compromise such analyses.

What it means and what to look for: This plot is useful for identifying mapping and/or annotation issues, and can indicate the comprehensiveness the genome annotation. Replicates that are obvious outliers may have sequencing/technical issues causing false detection of splice junctions.

Abnormalities in the splice junction rates are generally a symptom of larger issues which will generally be picked up by other metrics. Numerous factors can reduce the efficacy by which aligners map across splice junctions, and as such these plots become very important if the intended downstream analyses include transcript assembly, transcript deconvolution, differential splicing, or any other form of analysis that in some way involves the splice junctions themselves. These plots can be used to assess whether other minor abnormalities observed in the other plots are of sufficient severity to impact splice junction mapping and thus potentially compromise such analyses.

What it means and what to look for: 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. It can assist in identifying mapping and/or annotation issues, and can indicate the comprehensiveness the genome annotation. Replicates that are obvious outliers may have sequencing/technical issues causing false detection of splice junctions.

Abnormalities in the splice junction rates are generally a symptom of larger issues which will generally be picked up by other metrics. Numerous factors can reduce the efficacy by which aligners map across splice junctions, and as such these plots become very important if the intended downstream analyses include transcript assembly, transcript deconvolution, differential splicing, or any other form of analysis that in some way involves the splice junctions themselves. These plots can be used to assess whether other minor abnormalities observed in the other plots are of sufficient severity to impact splice junction mapping and thus potentially compromise such analyses.

What it means and what to look for: This plot can indicate the efficiency of the strand-specific selection protocol, and reveal variations in such efficiency. They can also be used to determine the ”strandedness rule”, which is required by many downstream analysis tools.

What it means and what to look for: Presence of outliers in the mapping rate statistics may be an indicator of large sample-prep, library-prep, or sequencer errors.

What it means and what to look for: Presence of outliers in these plots may point to variable inefficiency in a ribosomal/mitochondrial depletion protocol. In most datasets the Y chromosome counts can be used to determine sample sex (but not in this case, since the Y chromosome is not included in the example dataset's genome assembly). The raw metrics generated by QoRTs can also be used to generate counts for the ERCC spike-ins or similar.

What it means and what to look for: These normalization factors can be used for a number of downstream analyses, including the generation of summary browser tracks.

What it means and what to look for: Large variations in these ratios can indicate large-scale differences between the samples.

What it means and what to look for: This can be used to assess the occurrance rates of a number of failure modes.

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