RNA sequencing quality control metrics definition

Sebastian Ullrich[1], Sitanshu Gakkhar[3], Martin Hirst[2,3], Roderic Guigo[1]

1.Computational RNA Biology Group, Center for Genomic Regulation (CRG), Carrer del Dr. Aiguader, 88, 08003 Barcelona, Spain.
2.Department of Microbiology and Immunology, Michael Smith Laboratories, Centre for High-Throughput Biology, University of British Columbia, 2125 East Mall, Vancouver BC V6T1Z4, Canada.
3. Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, 675 W. 10th Avenue, Vancouver, BC V5Z 1L3, Canada.

This document aims to define the RNA Sequencing Quality Control Core Metrics and how to compute them. In general we highly recommend using only chastity passed reads.

1. Input RNA quality

The quality of the input sample RNA material is crucial for obtaining meaningful estimates for gene expression. As a wieldy used measure the **RNA integrity number** (RIN) indicates the degradation status of a given RNA sample. GTEx (www.**gtex**portal.org/) suggests a threshold for primary samples of 6 (range: 0 – 10) For samples coming from cell culture higher values should be expected (above 8).

2. Genomic contamination

Besides degradation of the input material, RNA samples can have remaining DNA molecules due to problems in the purification procedure. Those contaminations can be evaluated by the amount of reads from intergenic regions. The suggested measure "**proportion of intergenic reads**" is calculated as the following:

$$\frac{\text{number of mapped reads not overlaping with any transcript coordinates (including introns)} +/- L}{\text{number of all mapped reads per sample}}$$

L : Overhang extending transcript region by +/-500bp

Limitations mostly depend on the used annotation and are therefore prone to bias by novel transcripts. As a common standard we recommend to use the provided BED file derived from gencode annotation version 22 http://www.gencodegenes.org/releases/22.html

3. Library enrichment (riboZero, polyA+)

Ribosomal RNAs are highly abundant in most cells and would take a high fraction of the sequenced reads if they were not removed, resulting in a low representation of less abundant transcripts. No matter if cells are enriched for polyadenylated tails or specifically depleted for ribosomal RNAs, levels of ribosomal RNAs should be very low in the resulting libraries. In order to access successful depletion we recommend to use the "**fraction of reads mapping to ribosomal genes**" as measure.

$$\frac{\text{number of reads mapped to ribosomal genes}}{\text{number of mapped reads to the entire genome}}$$

We recommend using the Fasta files (humRibosomal.fa, hum5SrDNA.fa) provided within Illumina iGenomes for the corresponding genome assembly.

http://support.illumina.com/sequencing/sequencing_software/igenome.html

4. Library amplification/diversity

Most common library preparation procedures include PCR amplification steps of the fragments to increase the amount of material for the sequencing process. Over-amplification can occur when the number of cycles is high and the input material had a low RNA concentration. Resulting libraries have a low diversity of fragments whereas duplicates with identical start and end positions are abundant. We suggest using the “**fraction of reads flagged as duplicates**” as measure for library diversity, which is computed as the following:

$$\frac{\text{number of mapped reads}^{*}\text{ with the same start and end position as any other read from the same sample}}{\text{number of all mapped reads per sample}}$$

* read pairs in the case of paired end data

5. Mappability

The amount of reads of a RNA sequencing sample that can be mapped to the genome of the given species can be biased by a variety of factors from contamination by RNA or DNA of other samples coming from different organisms to errors during base calling. As a measure of those biases we suggest to use the “**proportion of mapped reads**“ which is computed as the following:

$$\frac{\text{number of all mapped reads per sample}}{\text{number of all reads in the provided sample fastq library file}}$$

The amount of added spike-ins can account for significant amount of reads not being mapped if not added to the genome towards which mapping is done.