Perform Quality Control on Raw Reads

Question 1: How many sequences are in the sample HIV\_12hr\_rep1? What is their average length?

Question 2: Which metrics show one or more failed samples?

Question 3: Were any reads completely removed from the samples? Note:The MultiQC "General Statistics" tables shows a rounded value, so use the "Sequence Counts" graph.

Question 4: Is the adapter problem solved? What about the GC content? Note: HIV replication is ramping up rapidly in these cells in the first 24 hours.

**Read Alignment**

5. In RNAseq, the percentages of uniquely aligned reads are typically lower than for DNAseq, due to the presence of unremoved ribosomal RNA. These are are present in multiple copies throughout the genome and cause reads not to be mapped confidently. RNAseq is expected to be above 75% for an uncontaminated human sample. Is the "% Aligned" above 75% for these samples? You can optionally check to see which percentage of the reads align to the HIV genome by re-running STAR using the HIV genome with built-in gene model hiv\_nc001802.

6. Which samples appear to show higher expression of MYC, the Mock or HIV?

7. How many exons does this gene have?

**Gene Quantification**

8. Locate the "featureCounts:Assignment" plot, which shows whether reads were assigned to genes (features) or whether they failed to be assigned. What is the main reason for reads not being "Assigned"?

**Testing for Differential Expression using DESeq2**

9. What are the top two most significant genes? Does the direction of change for gene MYC agree with our observation in Question 6?

10. What observations can you make from the PCA plot? Do samples cluster as expected?

11. What observation can you make about the pvalue distribution, does it look like there are many true significant results? Note that the published dataset has been downsampled for instructional purposes.