**Guide to interpreting MOC RNA-Seq data**

Files in Results directory:

1. **<ProjectID>\_<Accession>.tsv**

A table of the number of total fragments aligning to all features in all samples. Features representing transcribed regions include all annotated protein coding sequences (CDS:), ncRNAs (ncRNAs:), miscellaneous RNAs (misc\_RNAs), rRNAs (rRNA:), and tRNAs (tRNA:). By default the boundaries of CDSs are extended 20 and 30 bp 5’ and 3’, respectively, to ensure that fragments aligning to the UTRs of mRNAs are assigned to their associated gene. Intergenic features (IGR:) represent the intervening regions between annotated genes. Each transcribed feature also has a cognate antisense feature (AS\_\*:) that corresponds to the opposite strand from which the feature is transcribed. By default IGR features are assigned to the plus strand and AS\_IGRS to the minus strand. For fragments spanning 2 or more features, the fragment count for each feature is incremented by the fraction of the fragment length covering that feature.

1. **<ProjectID>\_<Accession>\_fpkm.tsv**

These are the NORMALIZED fragment counts aligning to all features in all samples.   
Fragment counts are normalized by dividing the total fragment counts by the length of the feature in kb then diving by the total millions of reads aligning to the sense strand of all CDSs in that dataset:

Fragments per kilobase per million reads (FPKM) = (# of fragments aligning to feature)/(length of feature in kb \* millions of fragments aligning to CDS)

1. **<ProjectID>\_<Accession>\_corr.tsv**

Pearson correlation (R2) of CDS fpkm from all pairwise combinations of samples.

1. **<ProjectID>\_AlignmentSummaryMetrics.txt**

Output of picard metrics for all samples. For details see:

[https://broadinstitute.github.io/picard/picard-metric-definitions.html - AlignmentSummaryMetrics](https://broadinstitute.github.io/picard/picard-metric-definitions.html#AlignmentSummaryMetrics)

1. **<ProjectID>\_metrics.txt**

**Sample:** Sample ID

**Total\_reads:** The total number of reads (not read pairs) for this sample

**pcnt\_aligned\_to\_any\_replicon:** % of reads aligned to any replicons (or contig) included in fasta/gff files included in analysis

**pcnt\_properly\_mapped\_pairs:** % of reads properly mapped in pairs (based on picard metrics)

**For\_replicon...:** The replicon to which the metrics following this column pertain. For samples aligned to a single replicon/contig, there will be 1 row of metrics. For those aligned to multiple replicons, there will be a line for each replicon plus a line providing the metrics for all replicons. All metrics below pertain to the indicated replicon(s)/contig(s)

**pcnt\_aligned:** % of reads aligned. For a reference strain grown in monoculture this number should be >90%. Lower values may indicate technical problems with library construction and/or sequencing or that the wrong reference sequence/annotations were used in analysis. For non-reference strains, bacterial communities, or samples that include both host and bacterial cells, this number can vary considerably even when libraries and sequencing data are of high quality.

**Total\_frags\_counted:** For paired end reads, reads properly aligned in pairs are counted as a single fragment spanning the lengths of both reads as well as the region of the genome between them. Each read aligned but not properly paired is counted as an individual fragment covering the entire length of the read.

**pcnt\_sense:** % of fragments aligned to the sense strand of all transcribed features. For most reference strains in monoculture this should exceed 80% and in most cases be above 90%.

**CDS\_total\_counts\_for\_replicon:** The total # of fragments aligned to both CDS and AS\_CDS features features in replicon. For standard applications our target for this number is > 1M.

**CDS\_pcnt\_of\_counted:** This number is usually 40-60% for reference strains in log phase cultures but can be significantly lower at stationary phase and in certain growth conditions such as during starvation or following exposure to transcriptional inhibitors. Alternatively, lower CDS % may be due to bad extractions, particularly when accompanied by a higher than expected % of ncRNAs or IGRs. It may also be an indication of alignment to the wrong reference strain if the % of total reads aligned is low and the % of reads aligned to rRNA is high, since as a whole rRNAs are better conserved than protein coding genes.

**CDS\_pcnt\_sense:** % of reads aligning to both CDS and AS\_CDS features that are aligned to CDS features. This value should exceed 85%.

**rRNA\_pcnt\_of\_counted:** This number should be below 30% for all strains, < 10% for most strains, and < 2% for some strains. High rRNA percentage can reflect poor depletion during library construction. Alternatively, rRNA can be higher under certain conditions such as following exposure to transcriptional inhibitors or can reflect poor RNA extractions.

**rRNA\_pcnt\_sense:** % of reads aligning to both rRNA and AS\_rRNA features that are aligned to rRNA features. This value should exceed 95%.

**misc\_RNA\_pcnt\_of\_counted:** Annotations for many strains do not include misc\_RNAs so this number is often 0.

**misc\_RNA\_pcnt\_sense:** % of reads aligning to both misc\_RNA and AS\_misc\_RNA features that are aligned to misc\_RNA features. This value should exceed 95%.

**ncRNA\_pcnt\_of\_counted:** Annotations for many strains do not include misc\_RNAs so this number is often 0. If ncRNAs are annotated in a reference strain they often include highly conserved, highly expressed structural RNAs such as tmRNA and rnpB. Collectively these ncRNAs can add up to a significant % of the data (5-15%). If the numbers are significantly higher, it can indicate poor extraction or poor size selection during library construction. Alternatively, stable ncRNAs can be enriched under certain conditions such as starvation or following exposure to transcriptional terminators.

**ncRNA\_pcnt\_sense:** % of reads aligning to both misc\_RNA and AS\_misc\_RNA features that are aligned to misc\_RNA features. This value should exceed 95%.

**tRNA\_pcnt\_of\_counted:** Size selection during extraction or during library construction removes most tRNAs. tRNAs should make up <5% of the total reads. If this is significantly higher it usually reflects problems during extraction or library construction.

**tRNA\_pcnt\_sense:** % of reads aligning to both tRNA and AS\_tRNA features that are aligned to tRNA features. This value should exceed 95%.

**IGR\_pcnt\_of\_counted:** The number of reads aligning to both IGR and AS\_IGR. This number is usually 10-20%. When annotations do not include misc\_RNAs and ncRNAs, this number will be towards the higher side of this range.

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

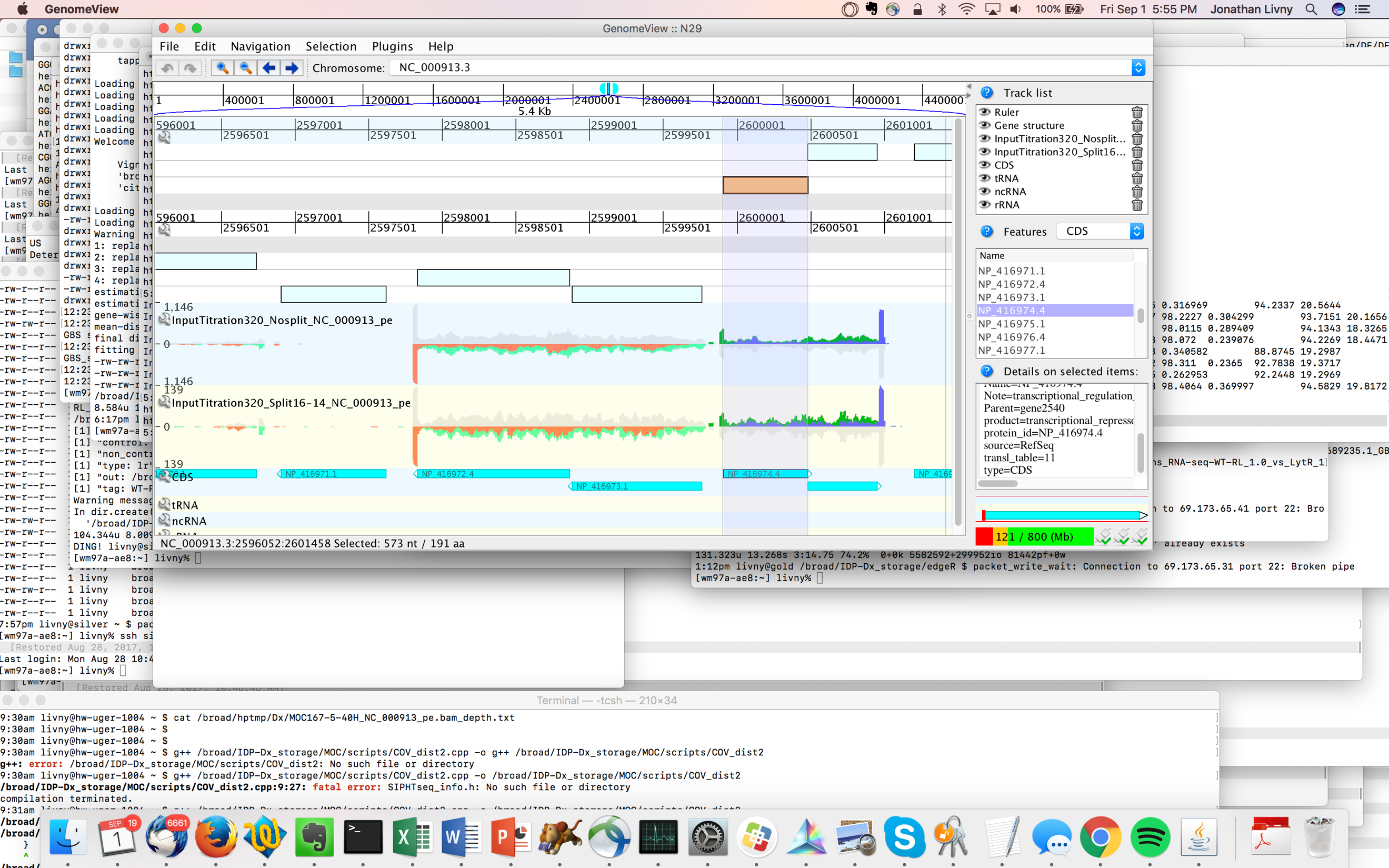
Files in DE directory:

This directory contains files generated by DESeq2 and edgeR for the sets of samples denoted in the CG\_Pairs column in the key file. For DESeq2, the CG\_IDs are included in the file name in the format A\_vs\_B with upregulated/downregulated features in the results file corresponding to those features upregulated/downregulated in A compared to B. Files ending in AllFeatures.txt and CDS.txt contain results for all features and only CDS features, respectively. The \*stats.txt files in the DESeq2 directory is a summary of the # of features whose abundance is significantly increased or decreased (fold > 2, Padj< 0.05).

Files in GV directory:

These files facilitate visualization of coverage plots in the context of genome sequences and and annotations using the GenomeView browser. Directions to utilize these files follow the following steps:

1. Go to <http://genomeview.org/> and press the “Launch” button or try to do it by pressing the button here: [aunch](http://genomeview.org/start/launch.jnlp)
2. This will open a the GV browser through a java app. Dismiss the initial dialogue box.
3. Enter command-O, select “Local files”, navigate to the GV directory, and select the .fna file.
4. Repeat for .gff file and all .tdf files of interest.
5. Here is an example of what the coverage plots should look like:



Blue and dark green represent coverage of read 1 and read 2 on the plus strand, respectively. Orange and light green represent coverage of read 1 and read 2 on the minus strand, respectively. Gray represent the composite of read 1 and 2 coverage on the opposite strand. All gene-encoding features are included in thec.gff track. Use control-F/control-G to find features/coordinates of interest.

**Method text for RNAtag-Seq**

**Generation of RNA-Seq data**

Illumina cDNA libraries were generated using a modified version of the RNAtag-seq protocol [1]. Briefly, 500ng-1 μg of total RNA was fragmented, depleted of genomic DNA, dephosphorylated, and ligated to DNA adapters carrying 5’-AN8-3’ barcodes of known sequence with a 5’ phosphate and a 3’ blocking group. Barcoded RNAs were pooled and depleted of rRNA using the RiboZero rRNA depletion kit (Epicentre). Pools of barcoded RNAs were converted to Illumina cDNA libraries in 2 main steps: (i) reverse transcription of the RNA using a primer designed to the constant region of the barcoded adaptor with addition of an adapter to the 3’ end of the cDNA by template switching using SMARTScribe (Clontech) as described [2]; (ii) PCR amplification using primers whose 5’ ends target the constant regions of the 3’ or 5’ adaptors and whose 3’ ends contain the full Illumina P5 or P7 sequences. cDNA libraries were sequenced on the Illumina [NextSeq 500] platform to generate paired end reads.

**Analysis of RNA-Seq data** Sequencing reads from each sample in a pool were demultiplexed based on their associated barcode sequence using custom scripts.  Up to 1 mismatch in the barcode was allowed provided it did not make assignment of the read to a different barcode possible. Barcode sequences were removed from the first read as were terminal G’s from the second read that may have been added by SMARTScribe during template switching.

**FOR MICROBIAL SINGLE ORGANISM DATA**

Reads were aligned to [reference sequence] using BWA [3] and read counts were assigned to genes and other genomic features using custom scripts. Differential expression analysis was conducted with DESeq2 [4] and/or edgeR [5].  Visualization of raw sequencing data and coverage plots in the context of genome sequences and gene annotations was conducted using GenomeView [6].

*To add to acknowledgments…*

RNA-Seq libraries were constructed and sequenced at the Broad Institute of MIT and Harvard by the Microbial ‘Omics Core and Genomics Platform, respectively.  The Microbial ‘Omics Core also provided guidance on experimental design and conducted preliminary analysis for all RNA-Seq data.

1. https://www.ncbi.nlm.nih.gov/pubmed/25730492

2. <https://www.ncbi.nlm.nih.gov/pubmed/11314272>

3. <https://www.ncbi.nlm.nih.gov/pubmed/19451168>

4. <https://www.ncbi.nlm.nih.gov/pubmed/25516281>

5. https://www.ncbi.nlm.nih.gov/pubmed/19910308

6. https://www.ncbi.nlm.nih.gov/pubmed/22102585