Final Progress Report: STIR 2014
Metabolism of the Lung Microbiome: A Three-Level Comparison
PI: Laura Tipton

We set out to compare the predicted metagenome, metatranscriptome, and metabalome of the lung microbiome in 10 samples. With combined funding from STIR and another grant, we were able to successfully sequence the metatranscriptomes of 37 samples. To identify over- and under-expressed genes, (as opposed to genes that are over- or under-abundant), we also sequenced the metagenomes of these 37 samples. All sequencing data will be made publicly available via the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI) prior to publication. This completes aim 1A, to perform metatranscriptomic experiments.

To fully characterize the metatranscriptome of the lung microbiome, we analyzed both the metatranscriptomic and metagenomic reads using the HMP Unified Metabolic Analysis Network 2 (HUMAnN2) pipeline (1). An update to the original HUMAnN pipeline, HUMAnN2 determines what genes are being transcribed by the bacteria present in a community by matching the metatranscriptomic reads to the UniRef50 reference database of gene families and normalizing the expression levels by the abundance of the metagenomics reads. The resulting UniRef50 transcript expression tables for each sample were joined and regrouped by KEGG terms. The most expressed KEGG terms include several tRNA synthesases, and the average normalized expression of the top 10 KEGG terms can be found in **Table 1**.

Of the 37 samples that were successfully sequenced for metatranscriptomics, we had previously sequenced the 16S rRNA V4 amplicons and analyzed the metabolome of 25 samples. These 25 samples were used to compare the three levels at which the metabolism of the community of bacteria in the human lungs can be studied: the genes or metagenomics level, the RNA transcripts or metatranscriptomic level, and the metabolic protein or metabolomics level. Each of the three levels were processed to assign KEGG terms. We compared the abundance/expression levels of each KEGG term using correlation measures. This comparison was not included in the original proposal but is a more direct comparison than the original aim 2A to compare enriched pathways in the lung microbiome. We were able to match 3,490 KEGG orthology terms between the metatranscriptome and the predicted metagenome. Correlations between them ranged from -0.3 to 1.0, with 325 (9.3%) KEGG terms being significantly correlated (**Figure 1A**), where significance is defined as a Pearson correlation test with a p-value < 0.05, equivalent to a Pearson correlation of 0.4. Due to the processing of the metabolome, we have only managed to match 30 KEGG pathway terms between it and the predicted metagenome. Correlations between them ranged from -0.3 to 0.64, with 8 (26.7%) KEGG terms being significantly correlated (**Figure 1B**). The poor correlation with the predicted metagenome likely stems from the low similarity between the microbial community in the lung and the reference communities used to predict the metagenome, most of which originate in the human gut. PICRUSt, which was used to predict the metagenome from the 16S amplicon reads (2), estimates this similarity, which can range from 0 to 1, to be 0.03 on average for our samples (SD 0.02). This hypothesis will be further supported if the metatranscriptome and metabolome are more correlated; this analysis is ongoing.

The KEGG terms identified by each level were used to look for differential abundance/expression between HIV infected individuals (N=6) and HIV uninfected individuals (N=19). We then compared the list of KEGG terms identified as differentially abundant or expressed between biologic levels, as well as the direction (over abundant/expressed or under abundant/expressed) of all KEGG terms, completing aim 2B, (and aim 1C, to identify differentially expressed transcripts, in the process). Even when KEGG terms were identified as differentially abundant/expressed by more than one biologic level, the

direction was not always the same. The predicted metagenome and metatranscriptome both called 4 KEGG terms differentially abundant/expressed: K00067 dTDP-4-dehydrorhamnose reductase; K00163 pyruvate dehydrogenase E1 component; K00845 glucokinase; K14205 phosphatidylglycerol lysyltransferase. Over all KEGG terms, they had 54% agreement on direction (**Figure 2A**). The predicted metagenome did not call any KEGG pathway terms differentially abundant but had 59% agreement on direction with the metabolome (**Figure 2B**). The lack of agreement on direction of over- or under-abundance/expression lends credence to the theory that important genes and proteins are coming from the rare members of the community.

While each biologic level of study can be used to answer different questions, they can all be used to look at the metabolism of a microbial community, such as the one found in the lungs. However, the results here show that the metagenomic level is not well suited to look at the metabolism, especially when the metagenome is predicted from 16S rRNA amplicon sequencing. Therefore, we recommend that microbiome studies that intend to look at the metabolism of the community use metatranscriptomics or metabolomics, despite the increased costs associated with these methods. This recommendation, results presented here, and the results of an integrative analyses of these three biologic levels will be presented in the forthcoming manuscript "A Tri-omics Comparison and Integration of the Lung Microbiome in COPD" (Tipton, et. al., *in prep*).

1. **Abubucker S**, **Segata N**, **Goll J**, **Schubert AM**, **Izard J**, **Cantarel BL**, **Rodriguez-Mueller B**, **Zucker J**, **Thiagarajan M**, **Henrissat B**, **White O**, **Kelley ST**, **Methé B**, **Schloss PD**, **Gevers D**, **Mitreva M**, **Huttenhower C**. 2012. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput Biol **8**:e1002358.

2. **Langille MGI**, **Zaneveld J**, **Caporaso JG**, **McDonald D**, **Knights D**, **Reyes JA**, **Clemente JC**, **Burkepile DE**, **Vega Thurber RL**, **Knight R**, **Beiko RG**, **Huttenhower C**. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol **31**:814–821.

Table 1 Average expression of the top 10 KEGG terms, as determined by the HUMAnN2 pipeline. Average expression is determined by normalizing the number of metatranscriptome reads on the number of metagenome reads and is in a variant of Reads Per Kilobase transcript per Million mapped reads (rpkm).

KEGG Orthology ID	Description	Average Expression (SD)
K01872	Alanyl-tRNA synthetase	390.3 (957.2)
K01845	Glutamate-1-semialdehyde 2,1-aminomutase	390.7 (943.4)
K03043	DNA-directed RNA polymerase subunit beta	444.6 (727.7)
K01873	Valyl-tRNA synthetase	540.5 (886.2)
K03046	DNA-directed RNA polymerase subunit beta'	552.2 (930.7)
K00962	Polyribonucleotide nucleotdyltransferase	564.2 (1733.5)
K01883	Cysteinyl-tRNA sytehetase	593.8 (1638.3)
K17570	Hydrocephalus-inducing protein	677.6 (4121.6)
K01890	Phenylalanyl-tRNA synthesase beta chain	707.5 (1575.1)
K01870	Isoleucyl-tRNA synthetase	813.9 (1600.7)