Universal Transcriptomic Analysis of Host-Microbiome Interactions in Psoriasis

Teal Furnholm¹, Marcus Foo¹, Laura Reingold¹, Kerby Shedden² and Andrew Johnston¹.

¹Department of Dermatology, ²Consulting for Statistics Computing and Analytical Research, University of Michigan, Ann Arbor, MI 48109, USA. Andjoh@umich.edu

This research was made possible with support from:

Background

- Perturbations in host-microbiome interactions underlie many diseases.
- Previous studies revealed changes to the microbial community in psoriatic skin lesions, and the change in dynamics between host and microbiome are increasingly identified as major factors in chronic inflammatory diseases, including many psoriasis co-morbidities^{2,3}.
- In addition, Streptococcal pharyngitis is the best-characterized environmental trigger for psoriasis, occurring ten times more frequently in psoriasis patients than controls, and psoriasis patients have double the rate of streptococcal carriage regardless of pharyngitis¹.
- Knowledge of host-microbiome interactions in the skin and tonsils of psoriasis patients may identify factors that drive this inflammatory skin disease.
- Using RNA-seq, we set out to take a snapshot of this dialogue in both the skin and tonsils of psoriasis patients with and without recurrent streptococcal pharyngitis.

Approach

- Advances in *16SrRNA* sequencing for phylogenetic analysis have been instrumental in describing the composition of the microbiome at various epithelial barriers. However, *16SrRNA* fails to identify bacteria at the species/strain level and reveals little of the functional changes in the microbiome in disease.
- Metatranscriptomic analysis using RNA-seq has emerged as an avenue for studying the dialogue between a host and its microbiome, however, current alignment tools are inadequate because:
 - Alignment of 100+ million RNAseq reads/sample to the 100+ million genes in target organisms is computationally non-trivial (months/sample).
 - The gene indexing required by some software (e.g. Bowtie2) fail to complete successfully with large numbers of genes (100+ million).
 - A large proportion of reads (~80%) map to more than one gene and/or organism due to conserved genes and orthologous genes, and existing software choose a 'most-likely' gene source, disregarding all others, a critical drawback for microbiome analyses.
- To address these issues, we developed new RNA-seq alignment software (Muscato) specifically for multi-genome sequence alignment.
- By coupling Muscato to a hand-curated functionally-annotated phylogenetic database we can perform species/strain level identification of organisms present, assess gene expression, and functional/metabolic pathway usage in any RNA-seq sample.

Database Curation

- In order to align RNA-seq reads to a set of target genes we first curated a database of genomes of all organisms likely to be within our samples.
- Gene sequences were extracted from NCBI GenBank files, GENCODE rel 27, and from Joint Genome Institute's Genome Portal Globus archive using custom perl scripts.
- Missing database identifiers were added as required: proteins were annotated with KEGG functions using GhostKOALA, and MetaCyc IDs were obtained from the BioCyc Pathway Tools source data. Additional protein information (COG, PubMed, RefSeq, EMBL, Gene Ontology, and BioCyc) were downloaded from UniProt. During curation, multiple mis-annotations (including a variety of updated genera for Clostridium) and missing or irregular taxonomic assignments were corrected.
- The database contains the 7 levels of taxonomy for each genome, summarized in Fig. 1.

	1. Bacteria	2. Eukaryota	3. Archaea	4. INA
Phyla	143	27	15	2
Species	38022	1271	870	4614
Gene Count	163593544	16752885	2080828	220050
Unique Gene Count	80533144	15130051	1677079	171470

Figure 1. Phylogeny of Species with Genes in the Muscato Reference Database.

[left panel] A phylogenetic tree of all species with genes in the Muscato reference database. There are four kingdoms: 1. Bacteria, 2. Archaea, 3. Eukaryota, and 4. Invasive Nucleic Acid sequences (INA; plasmids, viruses, transposons). The phyla with the largest proportion of sequenced species include: a. Actinobacteria, b. Firmicutes, c. Proteobacteria, and d. Bacteroidetes, with highly over-represented genera (# species/strains) including: e. Mycobacterium (2158), f. Staphylococcus (4485), g. Streptococcus (1223), h. Salmonella (1204), i. Escherichia (2088), j. Pseudomonas (1048), and Acinetobacter (885). Phylogenetic levels are colored according to the average percentage G+C nucleotide composition of the genes. Nodes are sized according to total number of genes per level. The table shows that there are approximately 10X more bacterial genes than eukaryotic genes, though approximately half of the bacterial genes are identical orthologs (non-unique) between different species.

[right panels] Zoom in on the Archaeal kingdom (K), showing the organization of the phylum (P), class (C), order (O), family (F), genus (G), and species (S) phylogenetic levels along the tree. [top right] Distribution of species in the Archaeal kingdom, showing how gene features such as GC composition can be significantly different as seen for species in the Halobacteria class (black arrow), or even between species of the same genus as seen for Thermococcus [bottom right].

RNA isolation and Alignment

- Swabs were taken from lesional skin (LS, n=11), uninvolved skin (US, n=10), and tonsils (PT, n=11) of psoriasis patients and normal controls (NS, NT n=7).
- Total RNA was extracted and stranded RNA libraries were sequenced.
- Trimmed high quality reads were then aligned to the universal database (work flow pictured below).
- All reads were allowed to multi-map and the last common ancestor (LCA) for each read was identified using our phylogenetic database. In-house perl and R scripts were used to generate differential gene and organism output.
- Muscato allowed us to align >400 million RNA-seq reads against our functionally-annotated phylogenetic database of ~182 million genes.

RNA-seq alignment workflow. On average 116 million, 100nt RNA-seq reads/sample were filtered and aligned using Muscato, yielding ~2 million genes which are fed into a Voom-Limma pipeline to ascertain differential gene expression between samples.

Results : Microbiome Community Analysis

Microbiome phylogenetic trees were constructed using the taxonomic lineage of genes in the metatranscriptome (Fig 2). *Streptococcus* and *Prevotella* species account for a significant proportion of psoriasis community differentially abundant organisms (DAOs). Psoriasis tonsils yielded 333 other DAOs, including *Tannerella forsythia*, *Bulleidia extructa* W1219, *Dialister invisus*, *Gemella sanguinis*, and *Staphylococcus sciuri*, all previously implicated in inflammatory diseases. These data were validated using *16SrRNA* phylotyping using V4 primers on the Illumina MiSeq platform (not shown). The same analysis was carried out for skin swabs which showed reduced microbial diversity in lesional skin (Shannon Index: LS=7.2, US=8.6, NS=8.3), with decreased *Propionibacterium* and >4-fold increased *S. pyogenes* colonization.

Figure 2. Differentially abundant organisms in PRT (n=8) vs RT (n=10) microbiome communities. Differentially abundant organisms (DAOs) are enumerated from fold-difference of normalized reads mapping to each of the phylogenetic levels (K kingdom, P phylum, C class, O order, F family, G genus, S species). *Streptococcus* (1) and *Prevotella* species (5) accounted for many of the differentially abundant organisms found, with 17 out of the 36 detected streptococcal species more than two-fold more abundant than in controls, with *S. anginosus* and *S. parva* predominant. Fold-change is indicated by box color and significance (inverse p-value) by box size.

Figure 3. Functional mapping of DEGs in lesional vs. uninvolved psoriasis skin. Using our annotated gene database we can map DEGs onto metabolic/immunologic pathways to understand which systems are up- or down-regulated in disease. In lesional skin, 799 DEGs (>2-fold) mapped to KEGG functions including inflammation (IL1B, IL8, NFKB, TLR4), proliferation (EGF, STAT3, CCND1, MYC), vascularization (VEGF, TEK, ANGPT1), and T cell response (TGFB1, ITGB2, ICAM1, SGK1). Fold-change is indicated by fill color and source kingdom by edge color.

Figure 4. Functional mapping of DEGs in the tonsils of psoriasis patients who exhibit skin disease worsening with pharyngitis (PRT vs RT). Multiple streptococcal species had differentially expressed (>5-fold) virulence factors, including sialidase (NEU1), hemolysin (HlyC), exfoliative toxin (EtxA), antimicrobial peptide resistance (DltA-D), sec- and TypeIV protein secretion systems, and secreted proteases. Mapping the host and microbe DEGs to KEGG pathways revealed possible urea/CagA/T4SS-mediated induction of host NF-kB activation and cytokine secretion. Fold-change is indicated by fill color and kingdom by edge color.

References

1. Thorleifsdottir RH, et. al. 2016. Acta dermato-venereologica, 96(6), pp.788-792.
2. Alekseyenko AV, et. al. 2013. Microbiome, 1(1), p.31.
3. Diamanti AP, et. al. 2016. Journal of Translational Medicine, 14(1), p.233.

4. Kanehisa M, et. al. 2002. Nucleic acids research, 30(1), pp.42-46.
5. Elder JT, et. al. 2010. JID 130(5), pp.1213-1226.
6. Ferran M, et. al. 2013. JID, 133(4), pp.999-1007.
7. Rohde M, Cleary PP. 2016. In: Streptococcus pyogenes: basic biology to clinical manifestations. Ferretti JJ, Stevens DL, Fischetti VA, Eds. Oklahoma City (OK).

Results : Functional Analysis of Skin DEGs

- Differentially expressed genes (DEGs) were determined from kingdom-normalized read counts using a Voom-Limma pipeline in R.
- Using KEGG⁴ gene function annotations from our database, we detected the interactions of both host and microbes in metabolic and regulatory pathways.
- Since multiple DEGs (from different bacterial species) may have the same KEGG function, the inverse variance weighted means (IVWM) were used to determine collective changes in community functions (Fig 3).
- Many up-regulated gene functions detected in skin lesion samples have been previously identified as DEGs in other psoriasis studies⁵, validating our new method to detect psoriasis skin DEGs.

Results : Functional Analysis of Tonsil DEGs

- Functional analysis of psoriasis tonsil DEGs highlighted many up-regulated streptococcal proteins, including several previously shown to activate skin-homing Th/Tc17 cells⁶, and participate in the systemic invasion and dissemination of *Streptococcus*⁷.
- By comparing psoriasis tonsils to recurrent tonsillitis, viridans-type streptococci (rather than Group A strep) and *Prevotella* spp. were responsible for the top-expressed DEGs including additional toxins and adhesins not identified through the KEGG analysis (Fig 4).

Conclusions

- This new bioinformatics pipeline allows simultaneous host and microbiome phylogenetic, transcriptomic, and functional analysis.
- Streptococci produce some of the top differentially expressed genes in both skin and tonsils, which may result in greater tissue colonization and immune activation.
- Although Group A streptococci have long been thought to be central to triggering psoriasis, our data suggest that viridans streptococci, *Prevotella* spp., and several other systemically invasive bacterial species may be the drivers in the pathogenesis of psoriasis and its co-morbidities.