**Notes to dRNAseq project:**

**RNA-seq data to reveal novel response mechanism to bacterial within host wound tissues**

Shicheng Guo, Thao Le, Jennifer L Anderson, Sanjay Shukla, 2019

Introduction to the research background and I proposed 3 analysis strategies and aims (12/09/2019)

* transcriptome different between wound tissues and normal tissues (skin and immnune cells)
* bacterial transcript nework and host transcript network correlation
* deconvolute mix host transcript patterns to sigle-cell contribution.
* make strain and function distribution analysis based on rRNA and check consistency with mRNA analysis
* check the genes for the domaint spices for each patients ? whether they are toxic genes?

cd ~/hpc/project/RnaseqBacterial/extdata/rnaseq

**Timeline:**

* 02/24: Updated revised manuscript to Sanjay: version 4 with more details about the Tables.
* 02/19: Updated revised manuscript to Sanjay: version 3 with modified figure legends
* 02/18: Batut, B., K. Gravouil, C. Defois, S. Hiltemann, J.-F. Brugère et al., 2018 ASaiM: a Galaxy-based framework to analyze microbiota data. [GigaScience 7: giy057](https://academic.oup.com/gigascience/article/7/6/giy057/5001424).
* 02/18: DIR="/gpfs/home/guosa/hpc/tools/HUMAnN2"
* 02/18: humann2\_databases --download chocophlan full $DIR
* 02/18: humann2\_databases --download uniref uniref90\_diamond $DIR
* 02/18: humann2\_databases --download utility\_mapping full $DIR
* 02/18: 13 million reads(~7 GB fastq), 1 core, 6G memeory with UniRef50 EC filtered database =4 hours
* 02/18: 13 million reads(~7 GB fastq), 1 core, 11G memeory with UniRef50 full database =25 hours
* 02/18: uniref90\_annotated.1.1.dmnd 11G file size to be downloaed with 20 mins.
* 02/18: pip install humann2 [humann2](http://huttenhower.sph.harvard.edu/humann2) is designed for both meta-genomic and trancriptome data analysis.
* Franzosa EA\*, McIver LJ\*, Rahnavard G, Thompson LR, Schirmer M, Weingart G, Schwarzberg Lipson K, Knight R, Caporaso JG, Segata N, Huttenhower C. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 15: 962-968 (2018).
* conda install -c bioconda metaphlan2 metaphlan2 is designed for metagenomic, not metatranscriptome data.
* 02/18: [MetaPhlAn2](https://bitbucket.org/biobakery/biobakery/wiki/metaphlan2): profiling the composition of microbial communities from metagenomic sequencing data.
* 02/14: One issue with including the 16S rRNA is that it is only able to provide a limited resolution for metagenome organism identification. With this pipeline, the goal is to identify down to genus and/or species level resolution, which is where 16S rRNA struggles. Another challenge is that, if an individual is doing metatranscriptome research, they're likely going to run a depletion step in their wet-lab protocol. rRNA depletions don't equally target all 16S sequences, and so the 16S results will be skewed by that step. For this reason, the SortMeRNA step removes rRNA sequences by default. You are always free to customize the pipeline to leave it out and include rRNA results, if you prefer, but it is unlikely to provide additional information when it comes to the metatranscriptome activity profiles.
* 02/13: This should be feasible, although keep in mind that the results would only really be useful for organism identification (since there's not going to be any functional information derived from the 16S sequences). It also may require a different reference, since it would be better to annotate against GreenGenes or SILVA instead of just RefSeq
* 02/12: Staph aureus: GFF: <https://www.ncbi.nlm.nih.gov/genome/proteins/154?genome_assembly_id=299272>
* 02/11: NC\_007793.gff file should be updated since lots of genes don't have name, we can give name for them
* 02/11: For example, SAUSA300\_0020:DNA-binding response regulator don't have name and we can name it as dbrR
* 02/11: For exampleSAUSA300\_0044:metallo-beta-lactamase family protein don't have name as mblfP
* 02/10: rRNAdeletionFunctionFreqHeatmap.pdf is functional annotation for the reads for each sample with heatmap
* 02/10: rRNAdeletionStrainHeatmap.pdf is strain annotation for the reads for each sample with heatmap style
* 02/10: Metatranscriptomes\_Function\_Analysis.pdf is bar-plot to show functional annotation for each sample
* 02/10: Metatranscriptomes\_Strain\_Analysis.pdf is bar-plot to show strain annotation for each sample
* 02/10: strain.tab.matrix.counts.txt and strain.tab.matrix.counts.csv are strain counts information for each sample
* 02/10: strain.tab.matrix.freq.txt and strain.tab.matrix.freq.csv are strain frequency information for each sample
* 02/10: function.tab.matrix.counts.txt and function.tab.matrix.freq.txt are functional annotation to reads
* 02/09: python ~/hpc/tools/samsa2/python\_scripts/DIAMOND\_analysis\_counter.py -I 42\_S3\_L001 -D ~/hpc/db/nr -O
* 02/09: important to install SortMeRNA: wget https://github.com/biocore/sortmerna/archive/2.1.tar.gz
* 02/09: important: pear-0.9.11-linux-x86\_64.tar.gz and pear-src-0.9.11.tar.gz can be download [here](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/bin)
* 02/09: 67G/home/guosa/hpc/db/nr.gz and 126G/home/guosa/hpc/db/nr.dmnd could be found [here](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/home/guosa/hpc/db)
* 02/09: A [repository](https://github.com/Shicheng-Guo/microbiome_helper) of bioinformatic scripts, SOPs, and tutorials for analyzing microbiome data.
* 02/09: [SAMSA2](https://transcript.github.io/samsa2/): a standalone metatranscriptome analysis pipeline
* 02/09: DIAMOND is a sequence aligner for protein and translated [DNA searches](https://github.com/Shicheng-Guo/diamond/blob/master/README.md)
* 02/08: 70G nr.faa + nr.gz, 2972.8s for indexing: wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz
* 02/08: RNA-seq workflow: gene-level exploratory analysis and differential [expression](https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html)
* 02/08: Workflow for Microbiome Data Analysis: from raw reads to community [analyses](https://bioconductor.org/help/course-materials/2017/BioC2017/Day1/Workshops/Microbiome/MicrobiomeWorkflowII.html).
* 02/08: mapping all reads to USA300 genes and found major reads are rRNA, see result [here](https://raw.githubusercontent.com/Shicheng-Guo/RnaseqBacterial/master/extdata/USA300/Rockhopper.result.summary.txt)
* 01/17: make strain and function distribution analysis based on rRNA and check consistency with mRNA analysis
* 01/17: check the genes for the domaint spices for each patients? Whether they are toxic genes?
* 01/17: Now the [strain](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/manuscript/result/Metatranscriptomes_Strain_Analysis.pdf) an [function](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/manuscript/result/Metatranscriptomes_Function_Analysis.pdf) distribution of the RNA-seq reads come out after remove rRNA and adaptors
* 01/17: Okay all the further analysis will be based on deep triming (-e 0.3) to the RNA-seq data to avoid adaptor aff-.
* 01/17: Sanjay: The kit did not specifically remove plasmid DNA. As far rRNA removal is concerned ...
* 01/17: Jennifer confirmed she performed a total DNA prep of the samples Thao provided, no enrichment.
* 01/17: Jennifer shared [adaptor, index and primcers](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/manuscript/Samples_with_indexes_and_adaptors.xlsx) information. Plasmid DNA was sequenced along with the total sample and the plan was to separate it bioinformatically.
* 01/17: check [notrim](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/rnaseqdata/notrimqc) fastq and [slight](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/rnaseqdata/slightrimdata) trim fastq, you can find why 90% reads are adaptors and rRNA.
* 01/16: [USA300 fasta](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master) and saved to /FastQ\_Screen\_Genomes/MW2 and index with bowtie2
* 01/16: [MW2 fasta](https://www.ncbi.nlm.nih.gov/nuccore/BA000033.2?report=fasta) save to /FastQ\_Screen\_Genomes/USA300 and index with bowtie2
* 01/16: Okay. now everything is clear. 90% reads are adaptor,check [figures here](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master), not human, not bacterials....
* 01/16: Fast filtering, mapping and OTU picking, rRNA removing: <https://bioinfo.lifl.fr/RNA/sortmerna/>
* 01/16: result:\\mcrfnas2\bigdata\Genetic\Projects\shg047\project\RnaseqBacterial\extdata\rnaseq\fastqscreen
* 01/16: Ryan told me the he cannot successully do what I want and so I only run it in BIRC7 which is quite slow
* 01/16: A fast and robust protocol for metataxonomic analysis using RNAseq data: Microbiome. 2017; 5: 7.
* 01/16: current result validated my expectation again, rRNA, tRNA, vector, adaptor counts for 80% of the reads.
* 01/16: I notice [Nature Metod](https://www.nature.com/articles/nmeth.3176) also mentioned another method called [DIAMOND](https://www.nature.com/articles/nmeth.3176) in 2015, I found it is powerful.
* 01/16: StringTie is better than Cufflinks, IsoLasso, Scripture and Traph as [Nature Biotechnology](https://www.nature.com/articles/nbt.3122) mentioned
* 01/16: mergeCountFiles: <https://rdrr.io/cran/rnaseqWrapper/man/mergeCountFiles.html>
* 01/16: check again is there any replications were set in this study? so that pca or mds can be done accurately
* 01/16: nothing was mentioned by tRNA and rRNA deplete in the protocol from Thao and Jeniffer (thinking why?)
* 01/15: at least another way to measure the ratio of bacteria to host genome size to compare with reads distribution.
* 01/15: think about the time points of the sampling, early immnue response? or late immune response?
* 01/11/2020, Just find CLARK have a new version and updated to [CLARKSCV1.2.6.1](http://clark.cs.ucr.edu/Tool/), in MCRI CHG1 server
* 01/10/2020, USA300 protein annotation were collected and the most recent one is from [Uniprot:2,607 protein](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/uniprot-proteome_UP000001939.gff)
* 01/09/2020, Let's check what's the difference between USA300(CP000255.1), MW2(BA000033.2) and CP019590.1
* 01/07/2020, Sanjay mentioned the genomic annoation to USA300 is not well prepared. Maybe I need to updated it.
* 01/06/2020, US300 <http://bacteria.ensembl.org/Staphylococcus_aureus_subsp_aureus_usa300_fpr3757/Info/Index>
* 01/06/2020: Sanjay said [US300](http://bacteria.ensembl.org/Staphylococcus_aureus_subsp_aureus_usa300_fpr3757/Info/Index) should be correct reference genome rather than my previous used version
* 12/28:I find [90% reads](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/mapping.md) are not located in human or Staphylococcus. I want to know where these reads come from?
* [Staphylococcus aureus](https://www.ncbi.nlm.nih.gov/genome/154), 10908 genomes, the genus Staphylococcus are pathogens of humans and other mammals
* 12/27/2019: EBV port to receive EBV alignment: <https://ebv.wistar.upenn.edu/tools.html>
* 12/27/2019: try to extract first 1000 reads and submit blast query from the website port NCBI and check microbe.
* 12/27/2019: No response from Ryan, try to install by myself. failed, install require administrative root, give up
* 12/26/2019: installed however the script cannot run since we don't have perl module Archive::Tarcontact Ryan
* 12/25/2019: install local blast in BIRC10 /home/local/MFLDCLIN/guosa/hpc/tools/ncbi-blast-2.10.0/bin/blastn
* 12/24/2019: alignment RNA-seq data with blastn to check what happened? and why the mapping ratio is so low.
* 12/22/2019: [1st-pass STAR](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/STARmanual.pdf) completed and started [2nd-pass STAR](https://groups.google.com/forum/#!msg/rna-star/4dhcEGFMiK0/XoMh6rB7CwAJ) and received sorted bam files ([about 2-pass](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/2pass-star.md))
* 12/18/2019: Deconvolution analysis to RNA-seq data to CD4+, CD8+ and B-cells and other white cells with [deconvSeq](https://github.com/Shicheng-Guo/deconvSeq)
* 12/18/2019: sam and bam files were saved in ~/hpc/project/RnaseqBacterial/extdata/rnaseq/bam
* 12/18/2019: RNA-seq alignment completed and data were saved in ~/hpc/project/RnaseqBacterial/extdata/rnaseq
* STAR2-pass pipeline was also prepared to do the RNA-seq analysis since majority TCGA data are from STAR2
* Gene-level quantifications: read counts and TPM values will be generated [withRNA-SeQC v1.1.9](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3356847/) and [HTSeq - FPKM](https://docs.gdc.cancer.gov/Encyclopedia/pages/HTSeq-FPKM/)
* 12/12/2019: download GENCODE 26 (<https://www.gencodegenes.org/human/release_26.html>) for the alignment
* confirmed with Sanjay to remove 8 from current analysis since it is not Staph Aureus, but Staph capitus
* ask Sanjay to check '8' sample which we cannot find in the clinical informaiton files (8: Staph capitus)
* Jennifer share [some laboratory novel information, N=54](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/102816%20cDNA%20Sample%20list-Sanjay-2019.xlsx), 54 samples -> 40 cDNA -> 24 fastqs.
* 12/12/2019: I ask Tao to prepare filename and smapleid connection file and protocol for RNA-seq
* 12/12/2019: Sanjay sent me the [clinical information n=30](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/Subject%20Info%20Abscess%20Study%20no%20MRN.xlsx) without connection file between filename and sampleid
* 12/10/2019: [24 sample id](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/S24_id.txt) is extracted and wait for clinical information (urgent).
* check the raw data: /mnt/bigdata/Genetic/Projects/Shukla\_cDNA/data/all\_samples/\*.fastq
* 12/09/2019: [IRB approved](https://github.com/Shicheng-Guo/RnaseqBacterial/blob/master/extdata/IRB/Outcome_Letter_SUL10110_IRB00000673.pdf): Role of Putative Toxin Genes in Infections due to Community Associated MRSA
* 11/27/2019: /mnt/bigdata/Genetic/Projects/Shukla\_cDNA/data/all\_samples/\*.fastq
* 11/25/2019: /mnt/bigdata/Genetic/Projects/Shukla\_cDNA/report
* 11/25/2019: Terrie works to add me to IRB of Sanjay's SUL10110
* 11/23/2019: Further Discuss methylation in MS project with Sanjay
* 11/23/2019: Discuss methylation in MS project with Sanjay
* 11/22/2019: Discuss RNA-seq project with Sanjay

**Installation:**

1. DIAMOND, version 0.8.3: <https://github.com/bbuchfink/diamond>
2. Trimmomatic, a flexible read cleaner: <http://www.usadellab.org/cms/?page=trimmomatic>
3. wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-Src-0.39.zip
4. cd TrimGalore-0.4.5
5. PEAR, if using paired-end data (recommended): <https://sco.h-its.org/exelixis/web/software/pear/>
6. SortMeRNA: <http://bioinfo.lifl.fr/RNA/sortmerna/>
7. wget https://github.com/biocore/sortmerna/archive/2.1.tar.gz
8. tar xzvf 2.1.tar.gz
9. cd sortmerna-2.1
10. ./configure
11. make

**Supplementary:**

Data analysis flow-chart was as the following. Its okay even microbe RNA-seq data is not the same pipeline.