DASeq: A Differential Abundance Analysis of Bacterial Spiked Blood

Zachery Dickson

April 27, 2019

Abstract

The generation of sequencing data invariably produces sequencing libraries which vary widely in size. This presents a challenge in the analysis of results and comparisons between studies. Due to historical quirks, this challenge has been traditionally 'solved' in metagenomic studies by rarefying sequencing libraries down to some common read depth. This method unnecessarily discards data, as statistical tools for solving this challenge have been developed and matured for the analysis of RNA-Seq data. I have a data-set composed of a community of known bacteria spiked onto a pooled blood background, then subjected to three DNA extraction methods in triplicate. Both shotgun sequencing and targeted enrichment were performed for all treatments. I would like to apply the statistical tools developed: for differential expression analysis to examine the differential abundance of bacteria between these treatments. Unfortunately, the data chosen for this project do not lend themselves well to this analysis, with a lack of replication of blanks and controls being the primary issue. Despite this, The application of the techniques themselves is simple, with potential to aid in future, better designed metagenomic studies.

Introduction

One of the challenges in performing analysis of metagenomic NGS data is the variability in library size. This can confound several effects, and one of the most common ways of dealing with this is to subsample reads from differently sized libraries down to the same size. [1] This process, also called rarefaction, discards data and is statistically inadmissible.

As McMurdie and Holmes discuss [1] the metagenomics field simply needs to look outside of itself for a solution. RNA-Seq has several well developed tools for examining differential expression, which account for several effects including library size in a rigorous, statistical manner.

In this project I aim to apply DESeq2 [2] to a set of samples where known pathogens of known concentration were spiked onto blood. This experiment is both a test of a novel bait design strategy, and extraction methods for sepsis blood samples.

Methods

Data Source

The Sequencing data used in this project is intended as a test of a pathogen bait set, as well as extraction methods for the pathogen DNA from human blood samples. Laura Rossi from the Surette Lab prepared acterial communities as a mixture of 7 strains of bacteria with varying gram staining, and genomic GC content (see Table 1). Two community sizes were prepared: 1000 and 10 CFU respectively of each strain (i.e. 7000 or 70 CFU total).

Strain	Gram Staining	GC Content (%)
Burkholderia multivorans ATCC_17616	negative	67
Escherichia coli BW25113	negative	50
Klebsiella pnemoniae N25C9	negative	57
Staphylococcus aureus IIDR-C0017	positive	33
Streptococcus constellatus C1050	positive	39
Streptococcus intermedius B196	positive	38
Streptococcus pneumoniae R6	positive	40

Table 1: Spike Strains

The bacterial communities were then spiked onto a pool of blood from intensive care unit patients. This was performed in triplicate for each sample condition in Figure 1. Additionally one high and low concentration community was spiked onto water rather than blood. There was also pair of extraction blanks where water was taken through the extraction process, one for the Saponin treatment, and one for the CFD treatment.

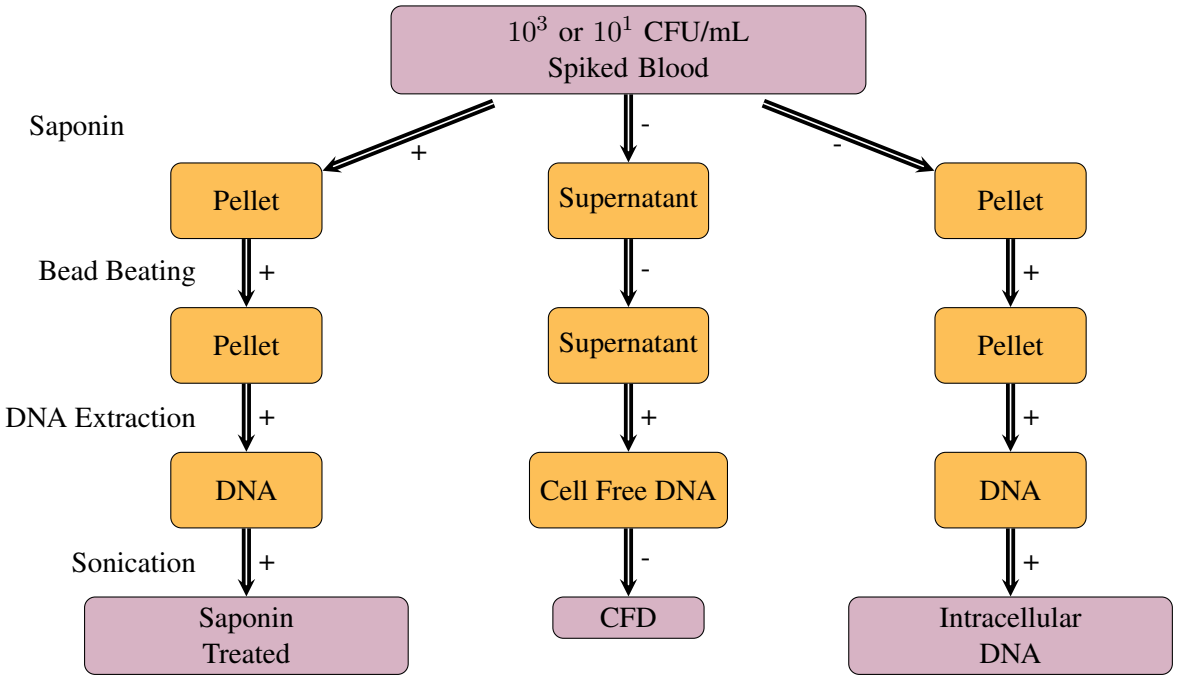


Figure 1: Study Design

The motivation behind the three extractions is as follows:

- **Saponin** - This a pretreatment with a detergent, which will lyse cells lacking a cell wall (i.e. human cells in the blood) This is intended to reduce the human background in the sample by allowing the human DNA to be removed in the supernatant after pelleting. [3]
- **Intracellular** - This is a standard genomic extraction procedure as a baseline

- **CFD** - Rather than extracting the DNA from the pelleted bacteria an human cells, DNA is extracted from the supernatant. There is evidence that pathogen DNA can be found free floating in a patient's blood [4]. This DNA is easier to process and could result in a more rapid diagnosis.

After Laura completed DNA extraction the DNA was purified by Melanie Kuch from the Poinar Lab using the Dabney protocol [5]. After extraction another blank was added; an empty tube to go through library preparation. Library preparation was performed as described by Meyer and Kircher [6, 7].

After library preparation each sample was split in two. One half was enriched with a pathogen baitset according the myBaits Kit Manual. [8] I designed the baits that were used in the enrichment by using an adapted version of BOND [9] to find 75bp oligos in a particular pathogen's genome which was unique among a database of known sepsis pathogens. Here unique means sharing at most 30 consecutive bp of identity, and at most 75% identity across the entire oligo.

The enriched and non-enriched (hereafter referred to as shotgun) samples were sequenced at 2x90bp on an Illumina HiSeq Run. Each sample was run on both lanes.

Preprocessing

Adapter removal, quality trimming, merging of overlapping read pairs, and error correction in overlapping regions was performed with fastp. [10] The TruSeq3-PE2 adapters were used, even though fastp has the ability to automatically detect the adapter sequence. This was because detection failed in some libraries, such as the library blank, likely due to low read depth. Reads with any N's were not accepted. A minimum length of 30 bp was specified, the last residue from every read was trimmed. All other settings were default. Fastp also produces fastqc-style html output, examples of which are listed in the appendix.

The both the merged and un-merged reads which passed through the trimming were de-duplicated based on sequence using PrinSeq. [11] Both Exact duplicates and reverse complement exact duplicates filtered. All other settings were default.

Mapping

All mapping was performed with BWA [12]. Three indices were prepared:

- GRCh38 - An index for the human genome
- Regions - An index for the baits
- WholeGenome - An index for the entire genomes for the complete Sepsis Database.

All bait sequences were padded with up to 100bp of DNA upstream and downstream of the actual bait. Any padded baits which overlapped were collapsed together into a bait region, noting where within the region the actual baits were located.

The de-duplicated reads were first mapped to the human genome, and any reads which mapped were excluded from further analysis. The filtering, as well as several file manipulations was performed with samtools. [13]

Reads from enriched samples were mapped to the bait regions, while the shotgun reads were mapped to the whole genomes.

Counting

Reads are assigned to the organism to which they mapped for both the shotgun and enrichment samples. The baits are assumed to be sufficiently unique that reads which map to the bait are from that organism. As there is leeway in the definition of unique, and it is known that perfect sequence identity is not necessary for DNA hybridization, only reads which mapped with no mismatches were counted. This requirement applied to both shotgun and enrichment. Soft-clipping of reads was allowed only if the soft-clipped region of the read was entirely outside of the reference. Enriched reads also had an extra requirement that at least one base pair of the read overlapped with the actual bait within the padded sequence. This was accomplished with a custom perl script(See appendix) which took advantage of the NM flag, CIGAR String, and other data in the SAM entry for each read. An illustration of situations when a read would and would not be counted is in Table 2.

Alignment	Enrichment	Shotgun	Alignment	Enrichment	Shotgun
$\frac{BBBB}{BBBB}$	✓	✓	$\frac{SSBBBB}{\_BBBB}$	✓	✓
$\frac{BBCB}{BBBB}$	✗	✗	$\frac{SSBBBB}{BBBBBB}$	✗	✗
$\frac{\_AAAB\_\_}{AAAABBBB}$	✓	NA	$\frac{AAAA\_\_\_}{AAAABBBB}$	✗	NA

Table 2: Different alignment possibilities and whether they are counted

Differential Abundance

Analysis of Differential abundance was performed using the R package DESeq2. [2] An in-depth description of of the analysis performed is described in the R markdown file listed in the appendix. The output of that file is also attached at the end of the this document.

Results

QC with fastp

Fastp generates a wide array of summary values I would like to highlight three of these:

- Fragment Length Distribution
- Duplication Rate

- Total reads after filtering

The Fragment length distribution can be viewed in the fastp output, and is summarized with the insert size peak. This is insert size with the highest frequency among overlapping reads, where both the forward and reverse read are at least 30bp long. The proportion of reads for which an insert size could not be determined is also calculated. Apart from examining the distribution for aberrations between samples the FLD can be useful in classifying reads. DNA from different sources may have differing fragment length distributions. For example, contaminants may have a shorter average fragment lengths than DNA of interest. Which does appear to be the case. In Figure 2 the samples which are spiked onto blood have much higher peak sizes than the library blank (all contaminants) water blanks, or samples spiked onto water. This is likely entirely driven by the fragment length distribution of the human DNA.

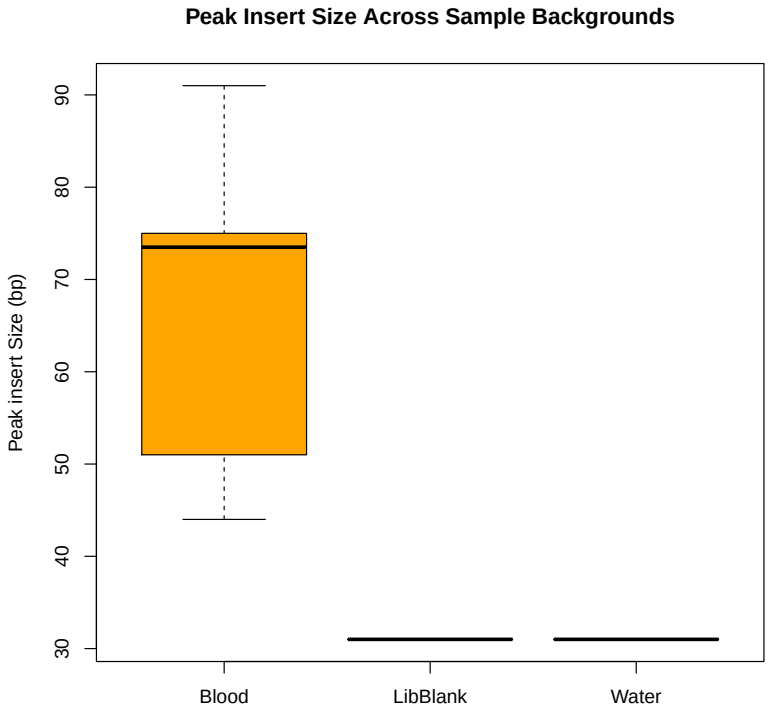


Figure 2: Comparing the insert size peaks of samples with different backgrounds. Note: The minimum acceptable read size was 30bp; An insert size peak at 31bp, likely means the true peak is even shorter.

Duplication rate is expected to differ between the enriched and shotgun sequenced samples. The Enriched samples go through additional rounds of PCR after enrichment, and even without that the variety of molecules is reduced after enrichment, making it more likely to sequence PCR duplicates from the same fragment. This holds up in the data as demonstrated in Figure 3.

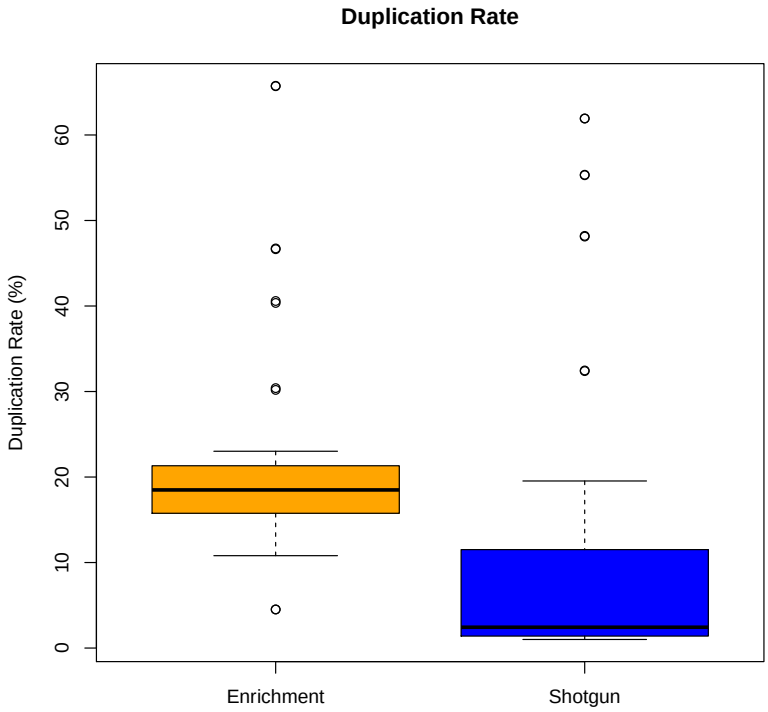


Figure 3: Enriched Samples have higher duplication rates than shotgun samples

The proportion of reads which are discarded by fastp can be a general indicator of the quality of the sample. In Figure 4 that a larger proportion of the reads from water backgrounds are discarded by fastp. This is likely tied to the higher proportion of short fragments as noted above. Interestingly The spiked samples tend to have more discarded reads as well.

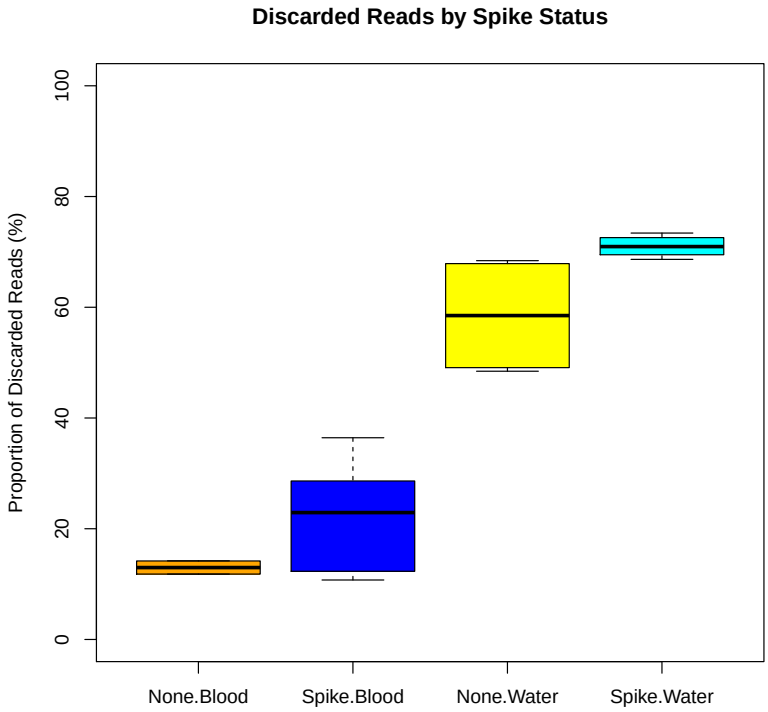


Figure 4: Water background and samples which were spiked have more reads which did not pass the QC filters

Human Filtering

Similar to the results for the fastp QC we can examine how the number of human reads filtered out varies with extraction.

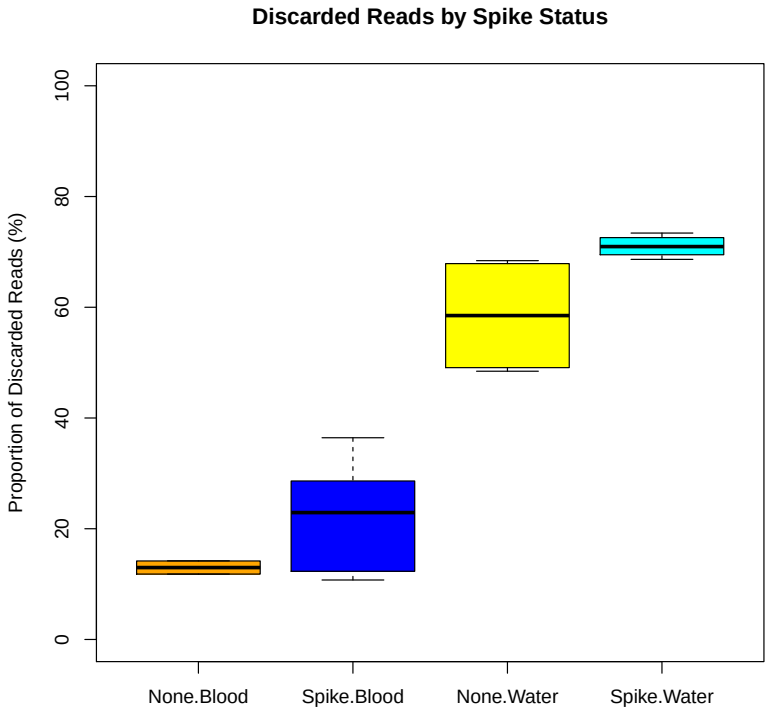


Figure 5: Samples which did not contain blood have lower human read proportions, however the water blanks still have high proportions compared to the Bacterial only. Saponin does not have lower human content than the intracellular samples. Some other work with these samples has noted DNA from the lysed human cells may be in the pellet.

Differential Abundance

The results of the differential abundance analysis are presented in included R markdown file. The output of which is included at the end of this report. One Separate analysis will be discussed here. A lack of replication in the blanks is a major issue with the design of this study. In order to develop a recommendation for the number of blanks required to be able to meaningfully identify differential abundance between the blank and the samples, a simulation was performed with random variations of the blood blank included as pseudo-replicates. A very similar analysis is also done in the R markdown results. This analysis differs in that technical replicates are not collapsed together, and libraries are normalized only bey the number of non-human reads. The results are shown in Figure 6, where by 5 simulated blanks all expected spike species are differentially abundant at least some of the time. Another run was done with fewer trials, but up to 10 simulated blanks, and the non-spike species never displayed significant differential abundance.

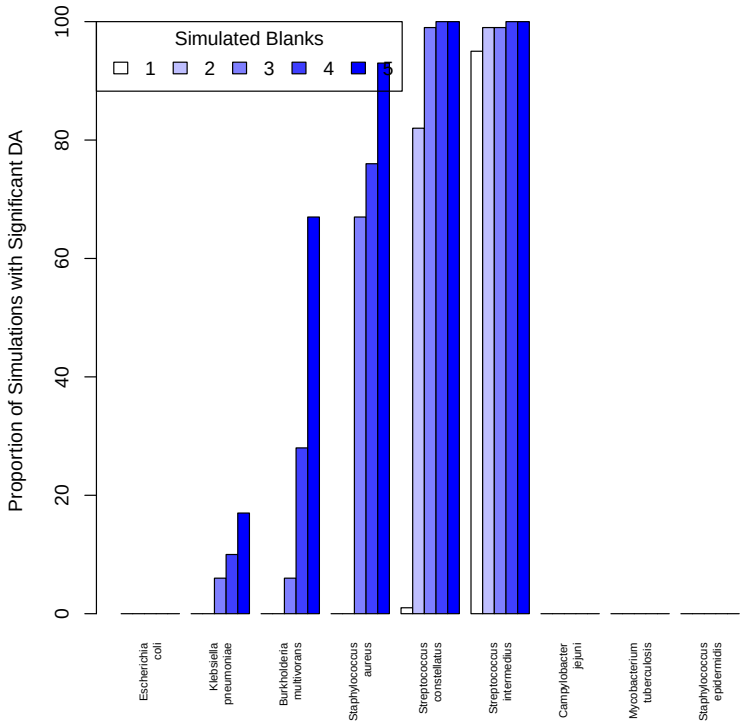


Figure 6: Increasing the number of simulated blanks increases the significance of the spike species in order of their read counts. *E.coli* is not expected to be significant in this setting, as there are no baits at the strain or species level for it.

Discussion and Conclusions

The major conclusion from this analysis is that Blanks should be replicated. The lack of reads mapping to anything is less informative in this type of analysis, especially given the overall low read counts. Blanks should likely be replicated to a higher level than the samples, in a clinical setting this would not be difficult, as period collection and testing of 'clean' blood would be a regular part of quality control for the procedure.

Despite The lack of information from the non-replicated blanks there is evidence that when comparing the low-spike samples to the high-spike samples the difference in abundance of the spiked species could be detected. In this situation replicated samples with non-zero abundances were being compared.

The majority of the pains from applying DESeq2 to this metagenomic data were related more to the experimental design, and idiosyncrasies of the dataset than to the overall goal of the project. Given a properly designed experiment, with adequate replication, and higher read counts, it would not be difficult to apply the techniques developed over this project.

As this experiment is primarily as test of the baitset, and the extraction techniques. On the question of is enrichment with these baits better than shotgon sequencing: The limited differential abundance results seem to indicate higher abundances in the shotgun samples for the spike samples. This may be related to the method in which reads were assigned to taxa, as indicated by

shifts in the abundance of the unmapped reads. As the bait regions are smaller than the whole genomes, there were fewer opportunities for a read to be assigned to a taxa in the enrichment.

On the question of extraction treatments: Overall these data are inconclusive. Some of the spiked species are differentially abundant due due to saponin, but two of these also appeared in the water blanks. The third *S.intermedius* was more abundant in the saponin treatments. The Cell Free DNA treatments are difficult to interpret, as None of the Spike organisms would have been expected in the supernatant, as the spike was intact, pellet-able cells. Lack of a difference of the spike species is may be due to DNA released from the spike during storage, or growth on the culture plates. The differential abundance of *H influenzae* may indicate that this method may be able to identify infections, when whole bacteria are not in the blood sample.

References

[1] P. J. McMurdie and S. Holmes, "Waste not, want not: Why rarefying microbiome data is inadmissible," *PLOS Computational Biology*, vol. 10, pp. 1–12, 04 2014.

[2] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for rna-seq data with deseq2," *Genome Biology*, vol. 15, p. 550, Dec 2014.

[3] M. Faria, J. Conly, and M. Surette, "The development and application of a molecular community profiling strategy to identify polymicrobial bacterial dna in the whole blood of septic patients," *BMC Microbiology*, vol. 15, p. 215, Oct 2015.

[4] A. E. Armstrong, J. Rossoff, D. Hollemon, D. K. Hong, W. J. Muller, and S. Chaudhury, "Cell-free dna next-generation sequencing successfully detects infectious pathogens in pediatric oncology and hematopoietic stem cell transplant patients at risk for invasive fungal disease," *Pediatric Blood & Cancer*, vol. 0, no. 0, p. e27734.

[5] J. Dabney, M. Knapp, I. Glocke, M.-T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. García, S. Pääbo, J.-L. Arsuaga, and M. Meyer, "Complete mitochondrial genome sequence of a middle pleistocene cave bear reconstructed from ultrashort dna fragments," *Proceedings of the National Academy of Sciences*, vol. 110, no. 39, pp. 15758–15763, 2013.

[6] M. M and K. M., "Illumina sequencing library preparation for highly multiplexed target capture and sequencing.," *Cold Spring Harb Protoc.*, no. 6, 2010.

[7] M. Kircher, S. Sawyer, and M. Meyer, "Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform," *Nucleic Acids Research*, vol. 40, pp. e3–e3, 10 2011.

[8] Arbor Biosciences, *myBaits Manual v4 (4.01)*, 2018.

[9] L. Ilie, H. Mohamadi, G. B. Golding, and W. F Smyth, "Bond: Basic oligonucleotide design," *BMC Bioinformatics*, vol. 14, p. 69, Feb 2013.

[10] Y. Zhou, Y. Chen, S. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, pp. i884–i890, 09 2018.

[11] R. Edwards and R. Schmieder, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, vol. 27, pp. 863–864, 01 2011.

[12] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, pp. 1754–1760, 05 2009.

[13] . G. P. D. P. Subgroup, A. Wysoker, B. Handsaker, G. Marth, G. Abecasis, H. Li, J. Ruan, N. Homer, R. Durbin, and T. Fennell, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, pp. 2078–2079, 06 2009.

Appendix

Scripts

Script	Arguments	Description
runfastp.sh	Forward_Reads.fq.gz Num_threads [Adapters.fasta]	Runs fastp on the specified reads, with parameters as described in methods. Reverse_Read file name is generated from Forward Reads by assuming the only difference is _R1_ vs _R2_
ExactFilterSAM.pl	[-hSso{0,1} -m{0,1}] file.sam	Filters a sam file to only those reads matching the specified conditions
OrgCounting.pl	HeaderMap.tab ExactFilter.sam	Counts the number of reads mapping to a particular organism. Note the HeaderMap.tab file must map all possible headings in the sam file.
Process.sh	Num_Threads OutputDir InputDir	Performs the entire preprocessing pipeline, producing a single file with the counts for each organism in each sample at the end. Notes: Requires several files not provided, including the raw sequencing data and BWA sequences.

Table 3: Scripts Used for Preprocessing

Examples

The Examples folder contains three html output files from fastp:

- One lane of the enriched Library blank,
- One lane of enriched LS2, and
- The same lane of shotgun LS2

Other Files

- HeaderMap.tab - Contains a mapping from the sam sequence headers to a trinary organism name. This is used OrgCounting.pl
- SampleDesc.csv - Contains the experimental design for use by DESeq. Additionally contains information about various measures of library size. Used and described in more detail in DASeq.Rmd
- OrgQuant.tab - A file which contains all the read counts for every organisms in every sample. Organisms are in rows, Samples are in columns. Used and Described in more detail by DASeq.Rmd

R files

- DASeq.R - Contains custom functions for applying DESeq2 to the dataset, more fully described in DASeq.Rmd
- DASeq.Rmd - Contains a detailed description of what was done to apply DESeq2 to the data. The output of this file is included below.

DASeq

Zachery Dickson

April 26, 2019

Experimental Design

The data being analyzed for differential abundance is a set of samples and controls of known bacteria spiked onto a blood background. The samples have either a high (1000 CFU) or low (10 CFU) spike, and are treated in three ways.

- A standard extraction procedure where the the sample is pelleted and the DNA extracted from the pellet then sonicated prior to library preparation.
- The same procedure with a saponin pretreatment step to lyse human cells, hopefully reducing background DNA levels.
- A procedure where the supernatant is taken instead, and the extracted DNA was not sonicated.

Each combination of extraction method and spike concentration was split and either shotgun sequenced, or enriched with a baitset targeted towards Sepsis Pathogens. All Samples were performed in triplicate. Additionally there are two positive controls and 4 blank samples:

- One each of High and Low concentration bacteria spiked onto water
- One water extraction blank treated as either a pellet or supernatant
- One blood blank
- One library blank which was a library preparation performed with an empty tube

Each of the blanks and controls were both enriched and shotgun sequenced, and performed in singlet.

All samples and controls were run on two lanes of the sequencer.

Data

The specific input data is the number of reads with map to a particular organism perfectly. That is maps with no mismatches, and with any soft-clipping occuring on portions of the read outside of the reference sequence. Enriched samples were mapped to the sequence of the baits used during enrichment. The baits were padded with up to 100bp upstream and downstream. The mapped reads had to be overlapping the actual bait to be counted. The Shotgun sequenced samples were mapped to the whole genomes of the pathogens used to design the bait set. escriptions of the samples, the experimental design, as well as information about the library size, human background reads, and unmapped reads is all contained in the SampleDesc.csv file.

```
Quant = read.table("OrgQuant.tab",sep="\t",header = T,stringsAsFactors =
FALSE,check.names = F)
SampleDesc = read.csv("SampleDesc.csv")
```

Subsampling to a particular Taxonomical Level

The quantification file has the number of reads mapping to each taxon without mismatches. The 'Organism' in each row is denoted with a trinary name, in the format "Genus|Species|Strain". In the cases where a bait is desigend at the genus or species levels, the lower taxonomic levels have an NA. The entire Quant file can then be subset to a particular taxonomic level by truncating the organism name to the desired level and summing the reads for all rows with the same name. Rows with no information at the desired level may be discarded. This is acheived with the following function:

```
SubSampleToTaxonLvl

## function (CountData, TaxonLvl = c("Strain", "Species", "Genus"),
##     delim = "|", na.rm = TRUE, zero.rm = TRUE)
## {
##     TaxonLvl = match.arg(TaxonLvl)
##     TaxonIndex = switch(TaxonLvl, Strain = 3, Species = 2, Genus = 1)
##     tmp = strsplit(CountData[, 1], delim, fixed = TRUE)
##     tmp = lapply(tmp, "[", 1:TaxonIndex)
##     CountData[1] = sapply(tmp, paste0, collapse = delim)
##     if (na.rm) {
##         TaxonhasNA = sapply(tmp, function(x) {
##             length(grep("^NA$", x)) > 0
##         })
##         CountData = CountData[!TaxonhasNA, ]
##     }
##     CountData = aggregate(CountData[-1], by = list(Organism = CountData[,
##         1]), FUN = sum)
##     if (zero.rm) {
##         RowIsEmpty = apply(CountData[-1], 1, sum, na.rm = T) ==
##             0
##         CountData = CountData[!RowIsEmpty, ]
##     }
##     CountData
## }
```

Normalizing for variable library size

The Quant file only contains data on reads which mapped to a pathogen, but does not include direct information about the size of the library. This information is stored in the Sample file. In order for DESeq to normalize the samples, it will require some information on how to accomplish this. Exactly how to normalize the samples could be done in a few ways:

- Based on the Total Size of the library, i.e. The number of reads with the index for each sample
- Based on the Trimmed reads for the library, i.e. The number of reads which pass QC
- Based on the Nonhuman library size, i.e. The number of trimmed reads which do not map to the human Genome

One could also not normalize at all, but include a row of pseudocounts to prevent any one sample from having no counts.

The following function extracts the appropriate information from the SampleDesc object to add the requested normalization data:

```
AddNormalizationRow

## function (CountData, colData, NormValue = c("Total", "Trimmed",
##     "NonHuman", "Pseudo"), pseudo = 1, nrowName = "Normalization")
## {
##     NormValue = match.arg(NormValue)
##     NormRow = switch(NormValue, Total = colData$LibrarySize,
##         Trimmed = colData$TrimmedSize, NonHuman = colData$TrimmedSize -
##             colData$HumanReads, Pseudo = colData$PathogenReads +
##             pseudo)
##     NormRow = NormRow - colData$PathogenReads
##     CountData[nrow(CountData) + 1, ] = c(0, NormRow)
##     CountData[nrow(CountData), 1] = nrowName
##     CountData
## }
```

Selected Data

The analysis that will be presented will be done at the level of Species, normalizing by the trimmed library size, with Human and Non-Human counts split into two rows. This allows for examining the effect on the human background due to the Saponin pretreatment. There are several common contaminants which are in the same genus as species we have spiked with (eg. Burkholderia Cenocepacia). Operating at the genus level would make distinguishing these contaminants from the spiked organisms impossible. Recent work has demonstrated that there is some strain level non-specificity in the baits which were used which makes strain level counts suspect. Additionally the Library Blank was removed from the analysis, as it confounded the experimental design. Additionally as all samples would be expected to display any contaminants from the library preparation, the library blank should ahve no effect on differential abundance.

```
Quant_Spec = SubSampleToTaxonLvl(Quant,"Species")
Quant_Spec = AddNormalizationRow(Quant_Spec,SampleDesc,"NonHuman",nrowName =
"Unmapped|reads")
Quant_Spec = AddNormalizationRow(Quant_Spec,SampleDesc,"Pseudo",pseudo =
SampleDesc$HumanReads,nrowName = "Homo|sapiens")
Quant_Spec = Quant_Spec[!grepl("ID-LBS",colnames(Quant_Spec))]
SampleDesc = SampleDesc[!grepl("ID-LBS",SampleDesc$Names),]
```

Initial Analysis

Due to the very low number of mapped reads across the samples, some deviations from the standard DESeq2 function were required. These changes are focused on the step of estimating Size Factors. The standard method of calculating geometric means will zero out in almost all cases. Therefore the poscounts method is used: Take the sum of the values where the log is defined, and divided by the total length including zero-counts. Additionaly the DESeq2 documentation recommends using the shorth function rather than the median to calculate the location of the size factors when working with low count data. The adapted workflow, as well as a heatmap of the normalized counts is generated with the following function:

```
DASeq
```

```
## function (CountData, colData, formula = ~Lane + condition, normRows = 1,
##     fitType = c("parametric", "local", "mean"), locType = c("shorth",
##         "median"), collapseBy = NULL)
## {
##     fitType = match.arg(fitType)
##     locType = match.arg(locType)
##     locfunc = switch(locType, shorth = genefilter::shorth, median =
stats::median)
##     if (sum(colnames(CountData[-1]) != colData$Names)) {
##         colData = colData[match(colnames(CountData[-1]), colData$Names),
##             ]
##     }
##     suppressMessages(das <- DESeqDataSetFromMatrix(CountData,
##         colData, tidy = T, design = formula))
##     if (!is.null(collapseBy)) {
##         das = collapseReplicates(das, collapseBy)
##     }
##     trial = try(das <- estimateSizeFactors(das, type = "poscounts",
##         locfunc = locfunc), silent = TRUE)
##     if (inherits(trial, "try-error") & locType == "shorth") {
##         message("Tie occured while caclculating size Factor location with
shorth, switching to median")
##         das = estimateSizeFactors(das, type = "poscounts", locfunc =
median)
##     }
##     das = estimateDispersions(das, fitType = fitType, quiet = T)
##     das = nbinomWaldTest(das)
##     das
## }
```

As each sample has two technical replicates, one for each lane, We can initially check if there is any effect due to lane.

```
Design = ~ Lane + Extraction + Background * Spike + Sequencing
DAObj = DASeq(Quant_Spec,SampleDesc,formula = Design)
resultsNames(DAObj)
```

```
##  [1] "Intercept"                          "Lane_L002_vs_L001"
##  [3] "Extraction_Saponin_vs_Normal"       "Extraction_Supernatant_vs_Normal"
##  [5] "Background_Water_vs_Blood"          "Spike_Low_vs_High"
##  [7] "Spike_None_vs_High"                 "Sequencing_Shotgun_vs_Enrichment"
##  [9] "BackgroundWater.SpikeLow"           "BackgroundWater.SpikeNone"
```

```
res = results(DAObj,name="Lane_L002_vs_L001")
res[sort.list(res$padj),c("log2FoldChange","padj")]
```

```
## log2 fold change (MLE): Lane L002 vs L001
##
## DataFrame with 72 rows and 2 columns
##                                    log2FoldChange              padj
##                                         <numeric>         <numeric>
## Achromobacter|xylosoxidans   -0.438937253196748 0.999427032518165
## Acinetobacter|baumannii      -0.316144899311122 0.999427032518165
## Acinetobacter|nosocomialis   -0.130890158912899 0.999427032518165
## Bacillus|anthracis            0.0543590331588925 0.999427032518165
## Bordetella|bronchiseptica      -0.18296544958837 0.999427032518165
## ...                                          ...               ...
## Vibrio|parahaemolyticus       0.0414576947480478 0.999427032518165
## Vibrio|vulnificus             -0.239168963187484 0.999427032518165
## Yersinia|enterocolitica       -0.205863694364789 0.999427032518165
## Unmapped|reads               -0.00281595058254243 0.999427032518165
## Homo|sapiens                  -0.0462997112726769 0.999427032518165
```

None of the Organisms are Differntially Abundant due to Lane, so the technical replicates can be collapsed together, and the analysis rerun without Lane in the design.

```
Design = ~ Extraction + Background * Spike + Sequencing
techRep =
sapply(strsplit(as.character(DAObj$Names),"_"),function(x){paste0(x[-
length(x)],collapse = "_")})
SampleDesc$Lane=NULL
DAObj = DASeq(Quant_Spec,SampleDesc,formula=Design,collapseBy =
techRep,normRows = 2)
```

At this point we can also examine the normalized counts with a heatmap generated by pheatmap. The noteable features demonstrated by the heatmap:

- Even after normalizing for mappable reads, the shotgun samples have higher read counts than the enrichment
- The Enrichment has lower background signals than the shotgun samples
- Six of the 7 spiked species cluster apart from almost every other species
  - *Strepococcus pneumonia*(Top Cyan) was spiked but was not captured in most enrichment samples
  - *Mycobacterium tuberculosis*(Bottom Pink) was not spiked but appears in most samples in nearly condition invariant manner

  - The sister group to the larger spike cluster is primarily composed of species in the same genus as the spiked organisms
- Low Spike Samples have lower read counts than High Spike Samples
- The Water blanks (last two columns) have more pathogen reads than the Blood Blank (Third last column)

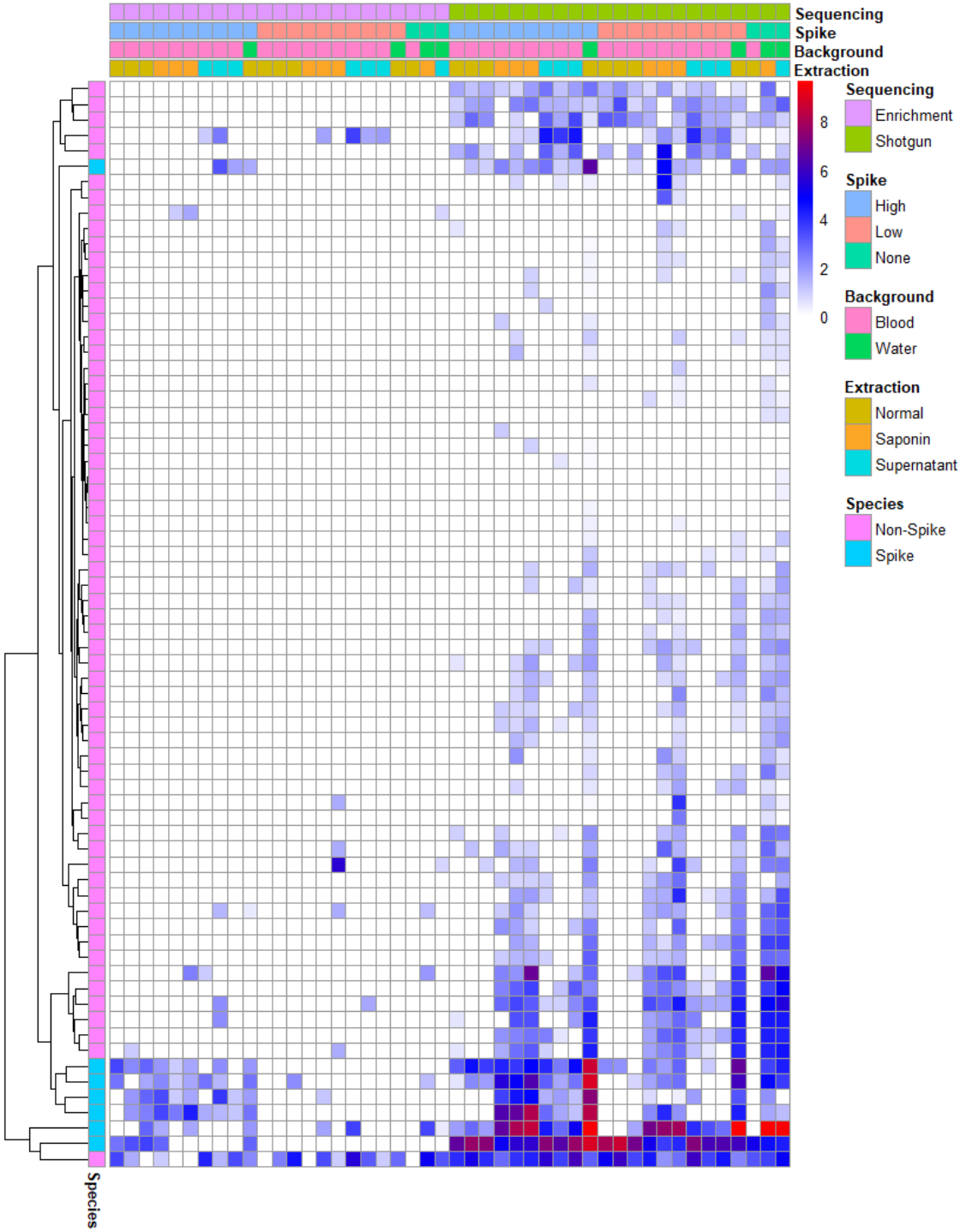


Fig 1 - Heatmap of All Normalized Read Counts

Differential Abundance

We can interate over each of the result Names of the DAObject and extract the differentially abundant organisms. As done with this function:

```
DAOrgList
```

```
## function (DAObj, rNames = resultsNames(DAObj)[-1], threshold = 0.05)
## {
##     extractSig = function(r, DAObj) {
##         res = results(DAObj, name = r)
##         res = res[ifelse(is.na(res$padj), FALSE, res$padj < threshold),
##             ]
##         setNames(res$log2FoldChange, rownames(res))
##     }
##     sapply(rNames, extractSig, DAObj = DAObj)
## }
```

```
resultsNames(DAObj)
```

```
## [1] "Intercept"                          "Extraction_Saponin_vs_Normal"
## [3] "Extraction_Supernatant_vs_Normal" "Background_Water_vs_Blood"
## [5] "Spike_Low_vs_High"                  "Spike_None_vs_High"
## [7] "Sequencing_Shotgun_vs_Enrichment" "BackgroundWater.SpikeLow"
## [9] "BackgroundWater.SpikeNone"
```

Before going over each factor individually, the following is importnat information:

- The seven spiked Strains were
  - *Burkholderia multivorans* ATCC_17616
  - *Escherichia coli* BW25113
  - *Klebsiella pneumoniae* N25C9
  - *Staphylococcus aureus* IIDRC0017
  - *Streptococcus constellatus* C1050
  - *Streptococcus intermedius* C196
  - *Streptococcus pneumoniae* R6
- Most of the spike bacteria are extracellular parasites, but *Burkholderia* and *Streptococcus* species have been known survive intracellularly
- Due to a lack of a closed genome no baits were designed for the spiked *E.coli* at the strain or species level, and it should ideally not appear in the enriched samples, only in the shotgun samples
- The blood spiked onto was a pool of blood from patients in the intensive care unit. Any donor to the ppol may or may not have had a blood infection.
- The numerical values are the log2 fold change of the first level vs the second level

Saponin PreTreatment Vs Normal Extraction

```
DAOrgList(DAObj,resultsNames(DAObj)[2])
```

```
##                             Extraction_Saponin_vs_Normal
## Escherichia|coli                                3.165430
## Klebsiella|pneumoniae                          -2.424375
## Streptococcus|intermedius                       3.330608
```

Only as subset of the Spike Organisms appear to be affected by the Saponin pre-treatment, and the effect does not have a consistent direction. Indicating Saponin may bias the bacterial component of the sample.

Supernatant vs Normal Extraction

```
DAOrgList(DAObj,resultsNames(DAObj)[3])
```

```
## Extraction_Supernatant_vs_Normal.Haemophilus|influenzae
##                                                4.258757
```

The only organism with differetial abundance is *Haemophilus influenzae*, which is more abundant in the supernatant. While not a spike organism, other work with this data hs identified this organism as potentially being a component of the pooled blood. Having been one of the only organisms found in the blood, and found only in the supernatant.

Water vs Blood Background

```
DAOrgList(DAObj,resultsNames(DAObj)[4])
```

```
##                         Background_Water_vs_Blood
## Escherichia|coli                         4.749604
## Staphylococcus|aureus                    4.976353
## Unmapped|reads                          -4.442035
```

A couple of the spike organisms appear to be more abundant in a water background than blood. This likely has less to do with the blanks being contaminated with the spike, than the fact that in the experimental design the water blanks and the positive control have the same background. The unmapped reads may be sourced from the blood background.

Low Spike vs High Spike

```
DAOrgList(DAObj,resultsNames(DAObj)[5])
```

```
##                              Spike_Low_vs_High
## Burkholderia|multivorans             -2.838893
## Staphylococcus|aureus                -2.873053
## Streptococcus|constellatus           -3.164297
## Streptococcus|intermedius            -3.475413
## Unmapped|reads                       -2.027946
```

As expected several of the spiked bacteria are negatively associated with reduced abundance in the low spike compared to the high spike. The reduced abundance of unmapped, non-human reads may be indicitive that some of these reads are simply failed assignments of reads originating from the spiked organisms.

No Spike vs High Spike

No Organsisms were significantly differentially abundant. This is a concerning result, as thes spiked samples should be most different from the unspiked smaples. However, the Water blanks have higher abundance for many organisms than expected.

Shotgun vs Enrichment

```
DAOrgList(DAObj,resultsNames(DAObj)[7])
```

```
##                             Sequencing_Shotgun_vs_Enrichment
## Achromobacter|xylosoxidans                         3.139562
## Burkholderia|cenocepacia                           3.041597
## Burkholderia|multivorans                           3.313516
## Escherichia|coli                                   5.685156
## Klebsiella|pneumoniae                              6.539955
## Micrococcus|luteus                                 3.283858
## Mycobacterium|tuberculosis                         1.826369
## Neisseria|meningitidis                             3.085851
## Pseudomonas|aeruginosa                             2.964886
## Serratia|marcescens                                3.110409
## Staphylococcus|aureus                              2.504544
## Streptococcus|anginosus                            2.718475
## Streptococcus|intermedius                          2.052689
```

This factor has the most differentially abundant species, all of which are more abundant in the shotgun samples. Larger library sizes, no loss of DNA via Enrichment, as well as more DNA to which to map (whole genomes vs padded baits) for the shotgun samples likely account for these differences.

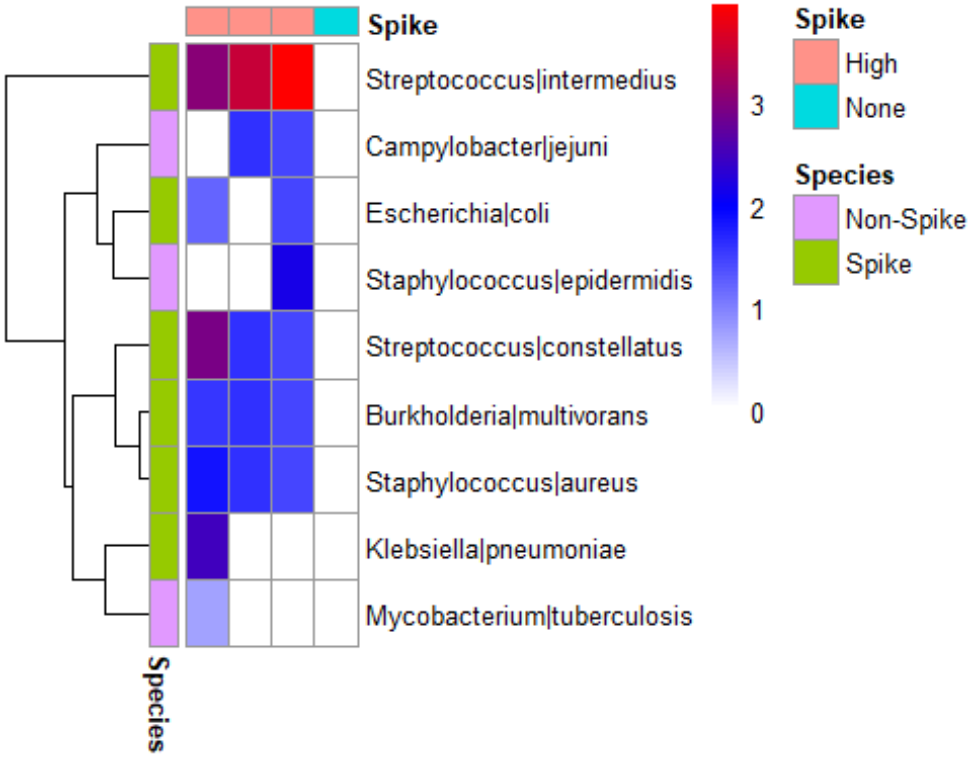
Neither of the interactions had significantly differentially abundant species.

Test Case

In the intended use case for this analysis, differntial abundance between a known blood blank, and the patients samples would be the intended comparison. This will be simulated with our data by restricting to the high concentration spike enrichment on a blood background with a saponin pre-treatement as the Sample and the blood blank. The Sample was selected based on other work done with this dataset indicating Saponin as the best extraction method.

```
InTest = SampleDesc$Names[
     InTest = SampleDesc$Description %in% c("Saponin","BloodBlank") &
SampleDesc$Sequencing == "Enrichment" & SampleDesc$Spike %in%
c("High","None")]
Quant_Test = Quant_Spec[,c(1,match(InTest,colnames(Quant_Spec)))]
Quant_Test = Quant_Test[apply(Quant_Test[-1],1,sum) > 0,]
TestDesc = SampleDesc[match(InTest,SampleDesc$Names),]
techRep =
sapply(strsplit(as.character(TestDesc$Names),"_"),function(x){paste0(x[-
```

```
length(x)],collapse = "_")})
Design = ~ Spike
DAObj = DASeq(Quant_Test,TestDesc,formula = Design,collapseBy = techRep)
SpikeSpecies = data.frame(Species =
as.factor(ifelse(grepl("Escherichia\\|coli|\\|multivorans|\\|pneumoniae|\\|co
nstellatus|\\|intermedius|\\|aureus",
                                Quant_Test$Organism),"Spike","Non-
Spike")),row.names = Quant_Test$Organism)
PlotHeatmap(DAObj,normRows = 2,row_annote = SpikeSpecies,show_rownames = T)
```




```
DAOrgList(DAObj)

## $Spike_None_vs_High
## named numeric(0)
```

None of the organisms which appear in the high saponin sample are differentially abundat from the blood blank given the data at hand. This is almost certainly because there is no replication of the blood blank. To evaluate how many replicates of a near empty sample would be required in practice we can simply copy the blank until some significance is noted.

```
for (copies in 1:5){
     print(DAOrgList(SimDASeq(DAObj,normrows = 2,sim = copies, p=0)))
}

## $Spike_None_vs_High
## named numeric(0)
```

```
##
## Spike_None_vs_High.Streptococcus|intermedius
##                                    -4.871577
## Spike_None_vs_High.Streptococcus|intermedius
##                                    -4.869428
## Spike_None_vs_High.Streptococcus|intermedius
##                                    -4.867769
## Spike_None_vs_High.Streptococcus|intermedius
##                                    -4.867014
```

After Two copies of the current blood blank *Streptococcus Intermedius* becomes significant, but even with increasing copies, nothing else becomes significant. These copied blanks are also not particularly valid. The same test could be done with simulated blanks. Blanks could be simulated with a negative binomial distribution, with a target of 1 success and a probability of failure equal to 1/ (9 + 2). This probability uses the pseudo-observations of 1 read across all 9 species, and 0 reads across all 9 species. Where the simulated reads fall across the 9 species may effect the observed significance. To make conclusions with this in mind, multiple simulations will be performed for each number of simulated blanks, and the number of simulations in which each organism is significant will be counted.

```
DA_UnderSimulatedBlanks(DAObj,normrows = 2,Trials = 100,p=1/11)
```

```
##                            1  2  3  4  5
## Burkholderia|multivorans   0  0  0  0  0
## Campylobacter|jejuni       0  0  0  0  0
## Escherichia|coli           0  0  0  0  0
## Klebsiella|pneumoniae      0  0  0  0  0
## Mycobacterium|tuberculosis 0  0  0  0  0
## Staphylococcus|aureus      0  0  0  0  0
## Staphylococcus|epidermidis 0  0  0  0  0
## Streptococcus|constellatus 0  0  0  0  0
## Streptococcus|intermedius  0 79 77 88 93
```

The result of simulation is essentially the same as duplicating the blanks. *Streptococcus intermedius* is differentially abundant after just two simulated blanks, but there is no indication that any other species will become significant. Of note is that significance is less likely at 3 simulated blanks than either 2 or 4.

In the main paper a similar analysis is done, where technical replicates are not collapsed, and the reads are normalized only by the number of non-human reads. The quite different results are indicative that the low read counts make this analysis very sensitive to cohoices such as normalization.