Sandia National Laboratories
U.S. DEPARTMENT OF ENERGY
SAND2014-0357P

LDRD Project Number: 171001
Project Title: Pathogen Capture
Project Team members: Zachary Bent (PI), Rachelle Hamblin
Related LDRD Investment Area: Biosciences

Abstract:
Yersinia entercolitica is a zoonotic extracellular pathogen that typically causes a self-limiting gastroenteritis in humans. Like all successful bacterial pathogens, *Y. enterocolitica* is able to rapidly respond to changing conditions as it encounters different microenvironments within the host and infects different host cell types. This ability to appropriately respond to the challenges of infection requires rapid and global shifts in gene expression patterns. In this study we analyze these shifts using a new pathogen transcript enrichment strategy followed by whole transcriptome sequencing (RNA-Seq) of *Y. enterocolitica* infecting murine macrophages. By focusing on early time points during infection, we were able to observe the transcriptional shifts that occur as the bacteria move from log-phase growth in a nutrient rich medium at 26°C to an active infection of mammalian cells. We used growth in filter-sterilized spent tissue culture media as a control to focus our analysis on genes that are expressed upon contact with host cells. We also compared the transcriptomes of bacteria located on the surface of host cells to that of cells that have been internalized by the macrophage cells revealing numerous genes involved in intracellular survival. Finally, we analyzed the expression of biovar 1B-specific genes such as those of the plasticity zone to determine how they contribute to its exceptional virulence. This study provides insight into the spatially and temporally dynamic expression patterns of known virulence factors as well as revealing several genes encoding unknown products that likely play roles during infection.

Introduction:
RNA sequencing (RNA-Seq) is an emerging field that has great potential to aid in the understanding of host-pathogen interactions. Prior to the advent of next-generation-sequencing (NGS) researchers relied on cDNA microarrays to study the transcriptomic profiles of both host and pathogen. Many of the host response studies were highly successful, leading to much that is currently understood about the innate immune response to pathogen-associated molecular patterns (PAMPs). Several bacterial gene expression studies elucidated mechanisms of pathogenesis in species such as *Salmonella enterica*, *Yersinia pestis*, and *Listeria monocytogenes*. However, while potentially powerful, microarray technology has several disadvantages when compared to RNA-Seq including the significant cost associated with the creation of both host and pathogen arrays. Using RNA-Seq to answer these questions not only decreases the price, but also significantly increases sensitivity, dynamic range, and the ability to detect un-annotated and non-protein-coding transcripts. Combined with the increasing availability of NGS sequencing facilities and improvements to bioinformatics tools, RNA-Seq is rapidly becoming a feasible method for many labs to study pathogen transcriptomics.

While some studies have attempted to perform RNA-Seq of both the host and the pathogen from the same sample, these studies used a brute force approach by sequencing samples at great depth. Unfortunately, even at great sequencing depth the pathogen transcripts are vastly outnumbered by host transcripts by as much as 200 fold. Therefore, to begin to think about transcriptomic analyses of both the host and pathogen from the same sample, it is useful to first enrich for the pathogen transcripts and bring them to a level more comparable to the host. Raising the number of pathogen reads as little as ten fold would significantly decrease the cost of sequencing as more samples could be multiplexed together while still maintaining the same amount of pathogen information. Several commercially available kits

enrich for pathogen reads indirectly by decreasing the number of host rRNA reads, but these kits either bias the pathogen transcripts or only enrich for non-rRNA host transcripts.

In this study we describe a non-biased, hybridization-based method and device that enriches for pathogen transcripts from infected samples. This technique, which we refer to as pathogen capture, is amenable to the study of any viral or bacterial pathogen and leads to upwards of 100 fold enrichment of the pathogen transcripts. Similar in design to currently available exome or custom capture techniques, our method is able to selectively separate pathogen cDNA transcripts from a host-dominant background into a pathogen-enriched pool for RNA-Seq analysis. Unlike exome or custom capture kits however; our technique is able to capture all possible pathogen transcripts at a dynamic range that allows for accurate determination of expression levels of many pathogen genes and at a fraction of the cost.

Detailed Description of Experiment/Method:

RNA-Seq library preparation and sequencing

200ng of total RNA was fragmented using NEBNext RNA fragmentation buffer (New England BioLabs) with a 3-minute incubation. The fragmented RNA was cleaned using the RNA Clean and Concentrator kit (Zymo Research). Double stranded and tagged cDNA was created as previously described. After determining the optimal cycle number for the indexing PCR, non-captured control samples were barcoded using custom, in-house index primers. Ready to sequence samples were quantitated by Qubit and combined into libraries with approximately equal amounts of each sample. The Vincent J. Coates Genomics Sequencing Laboratory (University of California, Berkeley) performed 100-base, single-end sequencing using an Illumina HiSeq 2000. The RVFV infection samples were also sequenced in-house using an Illumina MiSeq with a 50-cycle kit.

Probe generation and hybridization

Each of the three segments of the RVFV was PCR amplified. Total genomic DNA was extracted from overnight cultures of *Y. enterocolitica* using the DNeasy Blood and Tissue kit (Qiagen) according to the manufactures instructions. 100ng of PCR product or genomic DNA was used as starting material for the BioPrime DNA Labeling System (Life Technologies) also according to the manufacture's protocol. Each reaction was cleaned up using Qiagen's QiaQuick PCR purification kit. 2µg of probe was then mixed with 20ng of double stranded tagged cDNA from the infected cultures and dried using a Vacufuge (Eppendorf). Samples were re-suspended in 10µl of NimbleGen hybridization buffer and placed in a thermocycler at 95°C for 5 minutes followed by incubation at 60°C for at least 16 hours.

Fluidic device for capture and pathogen capture protocol

A simple fluidic system was assembled from commercially available components to speed and facilitate the chromatography steps. Syringe pumps (Cole-Parmer) were used to control fluids and to introduce the sample to the cartridge column. A 7-port, 6-way selection valve (Scivex) allowed switching between the sample and the various buffers flowing into the packed cartridge. Fraction collection was facilitated using a UV detector (Linear UVIS 200) and monitoring A_{260} just prior to the end of the outlet tubing. The entire system was placed in a small incubator/oven to minimize non-specific secondary structure formation (50-60°C).

Polyether ketone (PEEK) microcolumn cartridges, 30 µL volume, were fabricated in-house as previously described or obtained commercially (LabSmith). The cartridges were capped at each end with a PEEK nut threaded to connect to CapTite® fittings to facilitate fluid connections. The column cartridges were slurry packed with monomeric avidin agarose (Pierce/Thermo) in PBS under house vacuum. The agarose gel was retained in the column body with 35µm pore size PEEK mesh (Spectrum Labs). Prior to use the

packed cartridges were washed with 10 column volumes of PBS at 300µL/hr. The sample in 10µL volume in hybridization buffer, was introduced via syringe at the rate of 100µL/hr and flushed through the column with ~3 column volumes of PBS at the rate of 300µL/hr. The column was flushed with 10 column volumes of high salt PBS (250mM sodium phosphate pH 7.4, 750mM NaCl) at 300µL/hr. Biotinylated probe and hybridized pathogen transcripts were eluted with 50µL of elution buffer (2mM D-biotin in PBS) at 200µL/hr.

RNA-Seq analysis

Raw reads were processed using a previously described quality filter designed to remove low quality reads or sections of reads as well as any sequences derived from the sequencing adaptors or primers. The quality filtered FASTQ files were mapped to the *Y. enterocolitica* genome with Bowtie 2 in local alignment mode. The alignments were converted and sorted with the SAMtools package. For the differential expression analysis, read counts were generated for each CDS in the NCBI RefSeq annotation of the *Y. enterocolitica* genome with the BEDTools multicov tool. Differentially expressed genes were identified at each time point with the R package DESeq, by comparing the read counts of each CDS at four and eight hours to those in the culture-grown control. This package tests for differential expression through the application of the negative binomial distribution and a shrinkage estimator for the distribution's variance. Normalized expression levels among the various samples were obtained by estimating the total sequencing depths for each sample as the median of the ratios of the sample's counts to geometric mean across all samples. Further details of the statistical analyses can be found in the DESeq vignette (http://www.bioconductor.org/packages/2.12/bioc/vignettes/DESeq/inst/doc/DESeq.pdf). Genes were identified as differentially expressed when the DESeq calculated adjusted p-value was less than 0.05 and the change in expression was at least two-fold up or down. FPKM values for each annotated CDS were calculated from the alignments by providing Cufflinks with a reference annotation. Each gene's functional category was determined by the J. Craig Venter Institute's Comprehensive Microbial Resource (cmr.jcvi.org).

Results:

Wild type *Y. enterocolitica* strains were grown to mid-log phase and inoculated into 6-well plates of either fresh salt-free LB medium (control), filter-sterilized spent RPMI (RPMI), or centrifuged onto P388 cells (Infection) and grown at the temperatures indicated above. After RNA-Seq analysis comparing the different conditions at each time point, a significant number of differentially expressed genes were discovered. Interestingly, there was a near equal number of up and down regulated genes in each condition analyzed. Note that ~25% of all *Y. enterocolitica* genes are differentially expressed when the control transcriptome is compared to the infection.

To determine the genes that are most affected by contact with host cells, the bacteria grown in the filter-sterilized spent RPMI were compared to the infection at the 1hr time-point. The 10 genes with the largest up and down-regulation are shown in the table. Strikingly, several of the most up-regulated genes were involved in sugar transport and inositol metabolism. Inositol is commonly found within eukaryotic cellular membranes and has been implicated in the activation of the Ysc T3SS secreted effector YopJ. The graph depicts the clustering of all genes with ≥8-fold change in expression across the *Y. enterocolitica* chromosome. Although the Ysa T3SS has previously been shown to be expressed only at 26°C, the clustering analysis reveals that several genes in the system are significantly up-regulated in the filter-sterilized spent RPMI at 37°C

To further analyze the expression of the Ysa T3SS during the infection of murine macrophage cells, FPKM (fragments per kilobase of transcript per million mapped reads) values were determined for several

representative genes within the *ysa* locus. While most of the genes making up the apparatus do not change much in RPMI, they tend to go up dramatically during infection indicating that the Ysa T3SS system is contact dependent. Genes encoding proteins that make up the translocon are significantly up-regulated in RPMI alone with additional increases in expression during the infection.

All differentially expressed genes comparing the RPMI to Infection at 1hr post-infection are grouped according to their cellular function. As would be expected, there is significant down regulation of genes involved in protein synthesis and energy metabolism during the infection. This is likely due to decreased growth rate caused by the innate immune response of the murine macrophage cells. Interestingly, there are over 120 genes with unknown function that are up-regulated in this comparison. However, it should be noted that many known virulence factors are categorized as unknown in this analysis

Discussion:

Y.enterocolitica has the ability to infect multiple cell types. The requirements for survival and proliferation during infection have been studied both *in vitro* and *in vivo*, primarily through mutational analysis to identify the genes that are critical for virulence. While this approach has been highly successful in discovering genes that are required for full virulence of the bacteria in a given model system of infection, it is not without its disadvantages. One significant drawback to these types of studies is that they often fail to determine the stage(s) of infection for which the genes are required. For example, a gene that is required for the initial entry into a host cell will be identified as critical for virulence, however it is typically not possible to determine whether this gene is also involved in replication within the host cell cytosol, as a mutant for that gene will not proceed to those later stages of infection. To understand when and where genes are expressed throughout the course of an infection, transcriptional analyses are required. Global analysis of the transcriptome can be performed using either microarrays designed specifically for the pathogen of interest or, more recently, by sequencing total RNA (RNA-Seq) from an infected sample. While RNA-Seq is a relatively new technology, its sensitivity, dynamic range, low cost, and ability to detect non-protein-coding transcripts is unmatched by microarray-based approaches.

A major consideration for either transcriptomics approach is that the RNA recovered from virtually any infection is primarily host-derived, with the pathogen RNA outnumbered by well over 100-fold. Using RNA-Seq to analyze the pathogen transcriptome under these circumstances becomes expensive, as deeper sequencing is required to get enough reads for a gene-level analysis throughout the course of an infection. Prior work in our lab has demonstrated the effectiveness of a capture-based approach to enrich for pathogen transcripts from infected cells. This technique relies on the use of biotinylated probe sequences randomly generated from the entire bacterial genome, ensuring that all possible transcripts can be captured. Double-stranded and tagged cDNA generated from the infected sample are mixed with a large excess of capture probe. The mixture is denatured, and stringent hybridization conditions are established to allow the pathogen-derived cDNA to anneal to complementary capture probe sequences. The hybridization mixture is then adsorbed to a monomeric avidin column, washed repeatedly to remove cDNA non-specifically bound to the probe, and the remaining cDNA released from the column to generate a pool enriched for pathogen-derived sequences. The short tags at the ends of the cDNA allow PCR-mediated addition of full-length sequencing adaptors, whereas the probes lack these tags preventing inadvertent sequencing of the probe. Using a *F. tularensis* LVS infection model, we previously demonstrated unbiased enrichment of bacterial transcripts by upwards of 50-fold. This enrichment of pathogen transcripts allows for much more efficient sequencing of the bacterial transcriptome at any stage of infection, as compared to brute-force RNA-Seq without enrichment.

Given the proven ability of our pathogen capture approach to enrich for *F. tularensis* transcripts present in infected samples, we employed the technique to perform a differential gene expression analysis comparing the transcriptomes of the bacteria Y. enterocolitica before and after infection. By observing the transcriptional profiles of the bacteria at early time points after infection, we hypothesized that it would be possible to determine sets of genes that were important in two critical stages of the infection: extracellular and intracellular growth. At each time point we analyzed the global transcriptional shifts with respect to changes in expression of functional categories of genes as well as sets of known and putative transcriptional regulators and virulence factors.

The challenge faced by bacteria as they shift from culture to infection of host cells can be summed up as a change from replication in a protected environment to survival in a threatening environment. Consistent with this idea, we found that transition from culture to infection was generally associated with up-regulation of genes involved in virulence and stress response, and down-regulation of genes involved in replication. These results were consistent with expectations, and indicated that our techniques are effective at detecting the previously-identified transcriptional shifts that occur during infection. However, the switch from growth in culture to infection of host cells is both spatially and temporally dynamic, with different cell types and intracellular compartments presenting different challenges to bacterial survival. This is why, when the two stages of infection were analyzed in greater detail, we observed analogous but distinctive gene expression patterns associated with extracellular and intracellular growth.

Impact:

This study has led to several promising outcomes:

1) We were able to publish a paper in PLoS ONE describing the results of a related study analyzing the *Francisella tularensis* transcriptome and 2 key stages during the infection of murine macrophage cells. (ZW Bent et. al, PLoS ONE 8(10): e77834)
2) The intern that this project paid for was able to assist in the development of a new system to perform the pathogen capture technique that doesn't rely on any specialized equipment. This system is better than the previous setup in every possible way and is more effective at capturing pathogen transcripts.
3) The results from this study are currently being put together into a new manuscript for publication in the high impact journal PLoS Pathogens.
4) I presented the preliminary data from this study at the Cold Spring Harbor Microbial Pathogenesis and Host Response Meeting in a poster that was very well received.
5) The abstract for this project was selected for an oral presentation at the 2014 ASM Biodefense Conference.
6) Based on the interest generated by the poster I set up a collaboration with a group out of the University of Colorado, Boulder to perform this type of work on the livers and spleens of mice infected with *Salmonella* Typhimurium.
7) The results of this work have led to a continuation of our work with a collaborator at UC Davis who specializes in Yersinia pathogenesis.
8) Given the initial success of this study we were awarded a three-year LDRD grant to perform similar studies and to develop a device capable of multiplexing the process in a fast, flexible, and cost-effective manner.

Conclusion:

In this study we demonstrate a unique and unbiased approach to enrich for pathogen sequences from infected samples leading to significantly more informative RNA-Seq results. This method is based on the

selective hybridization of pathogen cDNA to complementary biotinylated probe sequences. The cDNA that is bound to the probe can then be removed from the mixed sample by filtering through a monomeric avidin column and then released to form a pathogen-enriched cDNA pool. This technique has several advantages over commercially available custom enrichment kits that operate by similar principles. Custom kits use chemically synthesized probe sequences to known genes, which both significantly increase the cost as well as limits the type of transcripts that can be captured to only those known genes. In addition, due to the cost of probe synthesis, these kits do not provide enough probes to capture the complete dynamic range of the bacterial transcriptome through the course of an infection. By creating probes to the entire viral or bacterial genome using an inexpensive and unbiased process we are able to produce a sufficient amount of probes to not only capture a dynamic range of transcripts, but also to capture any RNA species that may be expressed by the pathogen. In addition, the probe generation process is simple and flexible, allowing for the rapid production of probes to any pathogen of interest.

To demonstrate our pathogen capture technique we performed time course tissue culture infections using *Yersinia enterocolitica*. In all cases, captured samples were significantly enriched for bacterial sequences. This enrichment was unbiased and led to a large increase in information about the pathogen through the course of the infection. This decreases both the cost and time required for sequencing and makes bacterial transcriptome sequencing from infected samples feasible for labs without access to specialized microfluidic devices. Importantly, this study has led to a large increase in our understanding of *Y. enterocolitica* pathogenesis and has especially increased our knowledge of the expression of biovar 1B specific virulence factors. These previously under-studied virulence factors are critical to the enhanced virulence this strain exhibits when compared to more common and less pathogenic strains of the bacteria. Future work will be required to verify these results in the context of the whole organisms and likely we will find that there are tissue-specific gene expression patterns that the bacteria exhibit during the infection of mice.

Funding Statement:
This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.