RESEARCH ARTICLE

A Rapid Spin Column-Based Method to Enrich Pathogen Transcripts from Eukaryotic Host Cells Prior to Sequencing

Zachary W. Bent[1¤a]*, **Kunal Poorey**[1], **Annette E. LaBauve**[2], **Rachelle Hamblin**[1], **Kelly P. Williams**[1], **Robert J. Meagher**[2]*

1 Systems Biology Department, Sandia National Laboratories, Livermore, California, United States of America, **2** Biotechnology and Bioengineering Department, Sandia National Laboratories, Livermore, California, United States of America

¤a Current address: 10X Genomics, Pleasanton, CA, United States of America
* zachwbent@gmail.com (ZWB); rmeaghe@sandia.gov (RJM)

OPEN ACCESS

Citation: Bent ZW, Poorey K, LaBauve AE, Hamblin R, Williams KP, Meagher RJ (2016) A Rapid Spin Column-Based Method to Enrich Pathogen Transcripts from Eukaryotic Host Cells Prior to Sequencing. PLoS ONE 11(12): e0168788. doi:10.1371/journal.pone.0168788

Editor: Yung-Fu Chang, Cornell University, UNITED STATES

Received: June 17, 2016

Accepted: December 6, 2016

Published: December 21, 2016

Data Availability Statement: All raw sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) with the accession number PRJNA317373.

Funding: This work was funded by Sandia National Laboratories Laboratory Directed Research and Development (LDRD) Project 173021 (RJM). Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S.

Abstract

When analyzing pathogen transcriptomes during the infection of host cells, the signal-to-background (pathogen-to-host) ratio of nucleic acids (NA) in infected samples is very small. Despite the advancements in next-generation sequencing, the minute amount of pathogen NA makes standard RNA-seq library preps inadequate for effective gene-level analysis of the pathogen in cases with low bacterial loads. In order to provide a more complete picture of the pathogen transcriptome during an infection, we developed a novel pathogen enrichment technique, which can enrich for transcripts from any cultivable bacteria or virus, using common, readily available laboratory equipment and reagents. To evenly enrich for pathogen transcripts, we generate biotinylated pathogen-targeted capture probes in an enzymatic process using the entire genome of the pathogen as a template. The capture probes are hybridized to a strand-specific cDNA library generated from an RNA sample. The biotinylated probes are captured on a monomeric avidin resin in a miniature spin column, and enriched pathogen-specific cDNA is eluted following a series of washes. To test this method, we performed an *in vitro* time-course infection using *Klebsiella pneumonia*e to infect murine macrophage cells. *K*. *pneumonia*e transcript enrichment efficiency was evaluated using RNA-seq. Bacterial transcripts were enriched up to ~400-fold, and allowed the recovery of transcripts from ~2000–3600 genes not observed in untreated control samples. These additional transcripts revealed interesting aspects of *K*. *pneumoniae* biology including the expression of putative virulence factors and the expression of several genes responsible for antibiotic resistance even in the absence of drugs.

Introduction

Next-generation sequencing of RNA (RNA-Seq) has emerged as a powerful new technology with wide application space in diverse fields such as cancer research and plant science [1, 2]. One area in which RNA-Seq has great, but currently unmet, potential is in the study of

Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist. The authors are current or former employees of Sandia National Laboratories. The specific roles of these authors are articulated in the 'author contributions' section. Sandia National Laboratories did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no competing interests related to their employment with Sandia National Laboratories. This does not alter our adherence to PLOS ONE policies on sharing data and materials. Author ZWB is presently an employee of 10X Genomics. The authors attest that 10X Genomics had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors declare no competing interests related to ZWB's present employment with 10X Genomics.

host-pathogen interactions. Understanding the correlated gene expression of both the host and pathogen together in different tissues and time points during an infection would significantly improve our understanding of the dynamic host-pathogen relationship. Current RNA-Seq methodology is well-suited to measuring the genes expressed by the host in response to infection with a pathogen [3–5]. However, the genes expressed by the pathogen while invading host tissue and evading the host immune response are much harder to discern. This is because it is extremely difficult to obtain sufficient pathogen transcripts from an infected host sample, particularly early in the infection when the pathogen is least abundant, but during which time the pathogen may be actively adapting to the host environment, or evading host immune response. Typically host transcripts outnumber pathogen transcripts by well over 100 fold [6–8], meaning that using a standard RNA-seq library prep to sequence the pathogen transcripts in a mixed sample can be an expensive and computationally wasteful proposition [9, 10].

The large differential between host and pathogen transcript numbers was a known problem facing attempts to profile pathogen expression in microarray experiments [11, 12]. This problem also affects RNA-seq, even though the number of reads and depth of coverage is very high. Enrichment of low-abundance pathogen transcripts is beneficial to achieve sufficient depth of coverage to discern the dynamics of the pathogen transcriptome. We do note that in certain models (*e.g.* experiments performed at high MOI [13], or with a physical separation of pathogen-infected cells [14]) enrichment may not be necessary, and in that event enrichment is undesirable to avoid introduction of any unnecessary bias.

If enrichment is needed, pathogen transcripts can be separated from host transcripts at three stages prior to sequencing, each of which has demonstrated some level of success. Recent work has demonstrated the utility of an up-front separation of bacteria from the host tissue; however, this method still benefits from additional separation at the RNA stage for the depletion of host transcripts [15]. Another recent study demonstrates fluorescence activated cell sorting (FACS) of infected cells prior to sequencing to enrich for an internalized, GFP-expressing bacterial pathogen [14]. Generalizing this approach requires a discernible fluorescent marker of infection, access to a FACS instrument, as well as developing sample prep and sorting techniques for each new infection model. Several methods (some commercially available as kits) are depletive, removing RNA or cDNA classes based on their abundance (host and ribosomal RNAs) or sequence (ribosomal RNAs). These approaches, *e.g.* Ribo-Zero epidemiology [16], microbeEnrich [17], depletion of abundant sequences by hybridization (DASH) [18] and hydroxyapatite chromatography [19], act negatively and do not specifically select for pathogen RNAs. These approaches are overall beneficial by reducing the most abundant transcripts (most notably host ribosomal RNA) but may still be insufficient on their own when the bacterial load is very low. Despite the inclusion of RNAse inhibitors with these kits, any extended processing or handling of RNA at the bench top can lead to degradation of RNA, biasing the sample against unstable transcripts [20]. These kits typically require a large quantity of starting material (> 1 μg) and may also enrich for bacterial and viral transcripts other than the specific pathogen that is being studied. The final stage prior to sequencing in which bacterial transcripts can be enriched is after the RNA has been converted to cDNA. Working with samples at this stage has several advantages including the increased stability of cDNA and the need for less starting material.

We have previously reported an enrichment strategy for pathogen transcripts based on hybridization with biotinylated capture probes that are generated enzymatically from genomic DNA of the pathogen [7]. Processing was performed in a custom-built microfluidic system that handled one sample at a time, and furthermore required significant engineering expertise to construct and operate, which put the technique outside the capabilities of many microbiology laboratories. We report here a method that significantly improves upon our previous enrichment strategy by increasing the throughput to 12 samples simultaneously and decreasing the time required. Furthermore, the new protocol utilizes only commercial, off-the-shelf equipment that is commonly available in molecular biology laboratories. This hybridization-based method specifically enriches for the pathogen of interest at the cDNA level and enables sequencing of both the host and pathogen transcriptome from the same sample with minimal sample input requirements. We demonstrate our approach here with an *in vitro* infection study performed with a multi-drug resistant strain of *Klebsiella pneumoniae*, which is emerging as a significant cause of nosocomial infections.

Methods and Materials

Bacterial strains, cells, and growth conditions

Klebsiella pneumoniae strain ATCC BAA-2146 [21, 22] was obtained from ATCC and routinely grown on LB agar or in LB broth (Gibco) at 37˚C with shaking. P388D1 murine macrophage cells were also obtained from ATCC (ATCC® CCL-46™) and cultured in RPMI supplemented with 10% fetal bovine serum. Cell culture and infections were performed at 37˚C in an atmosphere of 5% CO_2.

Infection of murine macrophage cells

P388D1 murine macrophages were grown in 6-well plates for two days to form a confluent monolayer. Cultures of *K. pneumoniae* were grown overnight and then sub-cultured into fresh medium and grown to mid-log phase. Concentrations of bacteria and macrophages were determined and the P388D1 cells were infected within the 6-well plates in triplicate at MOI (multiplicity of infection) of 0.1, 1, 10, or 100. The plates were centrifuged at 500 × g for 5 minutes to enhance adherence and then placed back in the incubator for 1 hr in the first experiment. In the second experiment the same procedure was followed using an MOI of 10 and incubation periods of 2, 4, 8, or 24 hr. After incubation, the RPMI medium was removed and the cells were washed twice with 37˚C PBS to remove non-adherent bacteria, resulting in primarily macrophages with externally adhered and internalized [23] bacteria (Fig 1A). Cells were lysed and RNA was preserved by adding 1ml of RNAzol (Molecular Research Center, Inc.) to each well. The contents of each well was then transferred to 2 mL cryo-tubes and frozen at -80˚C until the RNA extraction was performed.

RNA extraction and cDNA Synthesis

Samples were thawed on ice. 400 μL of molecular biology grade water was added to each sample, and samples were mixed. After a 15-minute incubation at room temperature, samples were centrifuged at 4˚C for 15 minutes at 16,000 × g. The bulk (800 μL) of the aqueous phase was transferred to a new tube and mixed with an equal volume of 100% ethanol. RNA was extracted with the Direct-zol kit (Zymo Research) according to the manufacturer's instructions. RNA concentrations were determined by Qubit (Life Technologies), purity (A_{260}/A_{230} and A_{260}/A_{280}) was determined by Nano Drop (Thermo), and the RNA integrity number for the total RNA was obtained by BioAnalyzer (Agilent). The RNA was next fragmented using the NEBNext Magnesium RNA Fragmentation Module [24] with a 3 minute incubation at 94˚C followed by a cleanup step using the RNA Clean & Concentrator-5 (Zymo Research). Double-stranded, tagged cDNA was generated from 50 ng of fragmented RNA by the Peregrine method as previously described [7, 25].

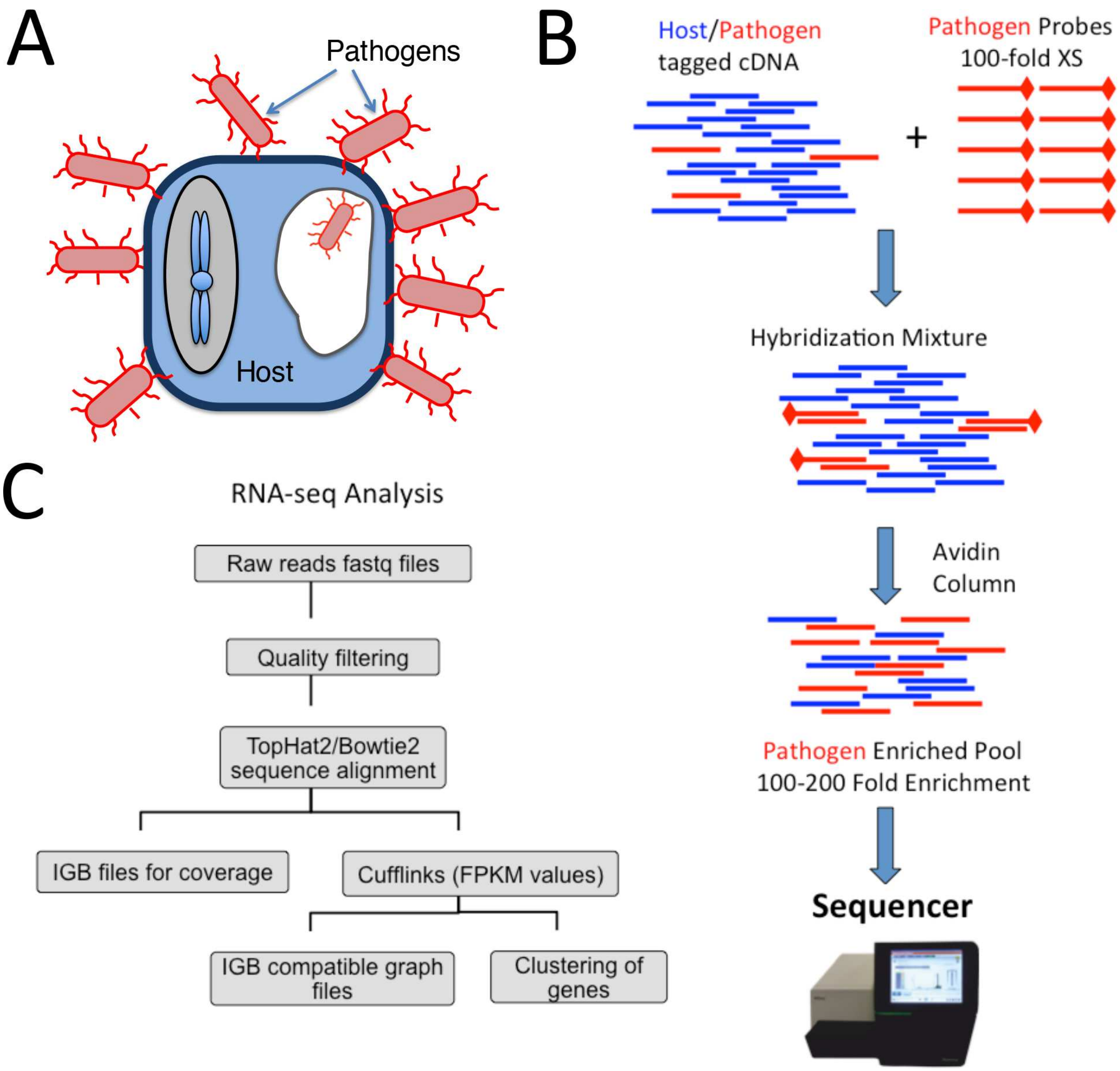

Fig 1. Schematic workflow of capture protocol and analysis. (A) Infection of pathogen (*K*. *pneumoniae*, red) to host P388D1 murine macrophage cells (blue). Non-cell associated bacteria are washed away prior to RNA extraction, such that the bacteria are primarily associated with host cells (both externally and internalized). (B) Capture protocol workflow where host pathogen cDNA with Illumina-compatible adapters is mixed with pathogen-specific capture probes for overnight hybridization. The hybridized mixture is then passed through a biotinylated avidin column to enrich for pathogen transcripts, prior to Illumina library construction. (C) RNA-seq workflow of YAnTra which was used to analyze RNA seq raw data where: Raw reads were quality filtered, aligned to reference genome, and processed with cufflinks to obtain gene expression.

doi:10.1371/journal.pone.0168788.g001

Enrichment of *K. pneumoniae* transcripts

K. *pneumoniae* was grown overnight in LB broth at 37°C with shaking. 1.5 mL of cultured bacteria was then collected by centrifugation and total genomic and plasmid DNA extracted using the DNeasy Blood and Tissue kit (Qiagen). Biotinylated *K*. *pneumoniae* probes were generated from the genomic and plasmid DNA using the BioPrime DNA Labeling System (Life Technologies) according to manufacturer's instructions. The hybridization reaction was set up with 20 ng of double-stranded tagged cDNA mixed with 2 μg of probes and then dehydrated using a vacuum centrifuge and re-suspended in 10 μL of hybridization buffer (NimbleGen). The hybridization mixture was then denatured on a thermocycler at 95˚C for 5 min followed by a 16–18 hr (overnight) incubation at 60˚C. Monomeric avidin agarose (60 μL) (Pierce) was packed into Micro-Spin columns (Pierce) by 30 s of centrifugation at 600 × g. All subsequent steps were carried out in a 60˚C incubator, with all buffers and equipment pre-warmed in the incubator for at least 2 hr. The packed column was washed twice with 200 μL of 0.1 M PBS and then the hybridized sample was applied directly to the packed resin. After 5 min incubation the sample was washed 15 times with 200 μL 2X Stringent Wash buffer (NimbleGen). The probe-bound sequences were then eluted off the column twice with 25 μL of elution buffer (Pierce). The eluted fraction was cleaned and concentrated to a final volume of 13 μL using 1.8X volumes of AMPure XP beads (Beckman Coulter). The enrichment protocol is illustrated schematically in Fig 1B.

Final Library Preparation and Sequencing

A qPCR assay was performed on the captured cDNA samples to determine the correct number of cycles required to generate the final sequencing library as well as to assess the efficacy of normalization and enrichment [25]. Typically, higher C(t) values indicate greater depletion of host transcripts. The final library was created by PCR amplifying the samples with the full length sequencing adapters and custom 9-mer barcodes [25]. Following PCR, the samples were cleaned and size selected in a two-step cleanup using 0.75 volumes of AMPure XP beads followed by 0.15 volumes to achieve a final library with an average size of ~330 bp. Libraries were combined with 12 samples each, in equal amounts and concentrated using the DNA Clean & Concentrator-5 (Zymo Research). Prior to sequencing the final libraries were quantified using qPCR. Non-captured samples were prepared by diluting the double-stranded cDNA 100 fold and then adding sequencing adaptors and barcodes with 12 cycles of PCR. Combined libraries were sequenced on an Illumina MiSeq using 150-cycle kits and custom 9 base index read. Samples were loaded at 18 pM and 151 base single-end reads were obtained. All raw sequencing data has been deposited in the NCBI Sequence Read Archive (SRA) with the accession number PRJNA317373.

RNA-Seq Analysis

The sequencing data was analyzed by our in-house YAnTra software pipeline (Yet Another Transcriptomics pipeline). The fastq files obtained from the MiSeq were analyzed by a previously described Perl script to filter out low-complexity or low-quality sequences, and parts of primer sequences [19, 26]. Briefly: sequence quality filtering is performed by a custom Perl script which removes internal barcodes and trims low quality fragments from the reads, fragments of primers used for the library construction and regions of low quality through Dustmasker. After applying these filters, reads shorter than 30 bp or with an overall quality score of less than 30 are filtered out. The quality filtered reads are then aligned to the mouse or *K. pneumoniae* reference genome using TopHat2 [27] and Bowtie2 [28] producing sam/bam files for alignment maps. The reference genome sequence and gene annotations for *K. pneumoniae* strain BAA-2146 were obtained from our previous study [22] [Genbank accession numbers CP006659.1, CP006660.1, CP006661.1, CP006662.1, and CP006663.1 corresponding to the chromosome and four plasmids]. Cufflinks [29] was used to analyze sam/bam files to get normalized expression values for genes as FPKM (Fragment Per Kilobase of exon per Million fragments mapped) values. A schematic of the YAnTra pipeline is shown in Fig 1C.

Results and Discussion

The recently sequenced *Klebsiella pneumoniae* strain ATCC BAA-2146 (Kpn2146) was the first isolate in the United States found to encode the NDM-1 metallo-β-lactamase, making it resistant to carbapenem antibiotics, a class of broad-spectrum antimicrobials "of last resort" that are typically effective against other multi-drug resistant bacteria [30, 31]. Carbapenem-resistant *Enterobacteriaceae* (CRE) such as Kpn2146 represent a potential public health crisis [32, 33], yet *K. pneumoniae* has historically been labeled an opportunistic pathogen. Compared to more highly pathogenic bacteria, relatively little work has been done to elucidate the mechanisms it uses in host interaction. RNA-Seq of bacteria during an infection represents a novel way to rapidly increase knowledge about the genes bacteria express during infection and could be useful in determining key virulence factors [34].

To our knowledge, the complete transcriptional profile of opportunistic pathogens such as *K. pneumoniae* has not been studied using RNA-seq within the context of infection models, as opposed to pure culture [35]. We thus chose to demonstrate the efficacy of our bacterial transcript enrichment method in a preliminary study in which we infected a murine macrophage cell line with Kpn2146 at infectious doses ranging over four orders of magnitude as well as observing a 24-hour time course infection at a more standard infectious dose (MOI = 10). Because free-swimming bacteria are washed away, the signal we examine comes from bacteria that are primarily associated with the host cells or possibly attached to the plate. (Note that gentamicin treatment, a standard protocol to remove extracellular bacteria in tissue culture infections, is not feasible because this strain of *K. pneumoniae* has high-level resistance to gentamicin.) Our enrichment method is based on the hybridization of biotinylated probes against the bacteria of interest, selecting bacterial cDNA from the mixed host/pathogen cDNAs [6]. Probes were generated without bias from both strands of bacterial genomic DNA so that every bacterial transcript has a corresponding probe. As illustrated in Fig 1B, probes are hybridized with the mixed host and pathogen cDNA at 100-fold excess ensuring that the complete dynamic range of bacterial transcripts can by captured. In a scenario where host transcripts outnumber pathogen transcripts 100-to-1, the capture probes are present in approximately 10^4-fold excess relative to bacterial transcripts, ensuring high capture efficiency. The hybridization mixture is passed through monomeric avidin-packed spin columns to bind all biotinylated probes, including those probes with bound cDNA transcripts. Through a series of washes the non-bound host transcripts are depleted and then the pathogen-enriched pool can be eluted off the column using a concentrated biotin buffer. The final sequencing library is created by PCR amplifying this pool to add full-length Illumina sequencing adaptors and indexes.

Characterization of capture enrichment

In the first experiment a short 1-hour infection was performed at 4 different multiplicities of infection (MOI) ranging from very low (0.1) to very high (100). In the second experiment, a time course was undertaken examining enrichment of Kpn2146 transcripts at two, four, eight and 24 hours post-infection. In each case we show consistently high levels of enrichment, especially evident in both the shorter time and lower MOI infections where the bacteria is less abundant. As shown in Fig 2A, in the one-hour infection, Kpn2146 transcripts were enriched by approximately 150- to 400-fold compared to the non-captured control. We also observed a significant enrichment of reads aligning to the bacterial genome by applying the capture protocol to RNA-seq libraries at all time points from the infection (2, 4, 8, and 24 hours), as shown in Fig 2B. No hybridization-based protocol is 100% selective for its target, because off-target hybridization can occur. In the case of an infection model the host RNA is such an overwhelmingly large fraction of the total RNA that even a highly selective capture protocol with stringent hybridization conditions and washes will result in carryover of host-derived RNA. The mapping of reads to the mouse genome was uniform in the uncaptured samples reflecting no bias. After capture the percentage of host reads decreased with increasing time of infection ranging between 38.7–19.9%.

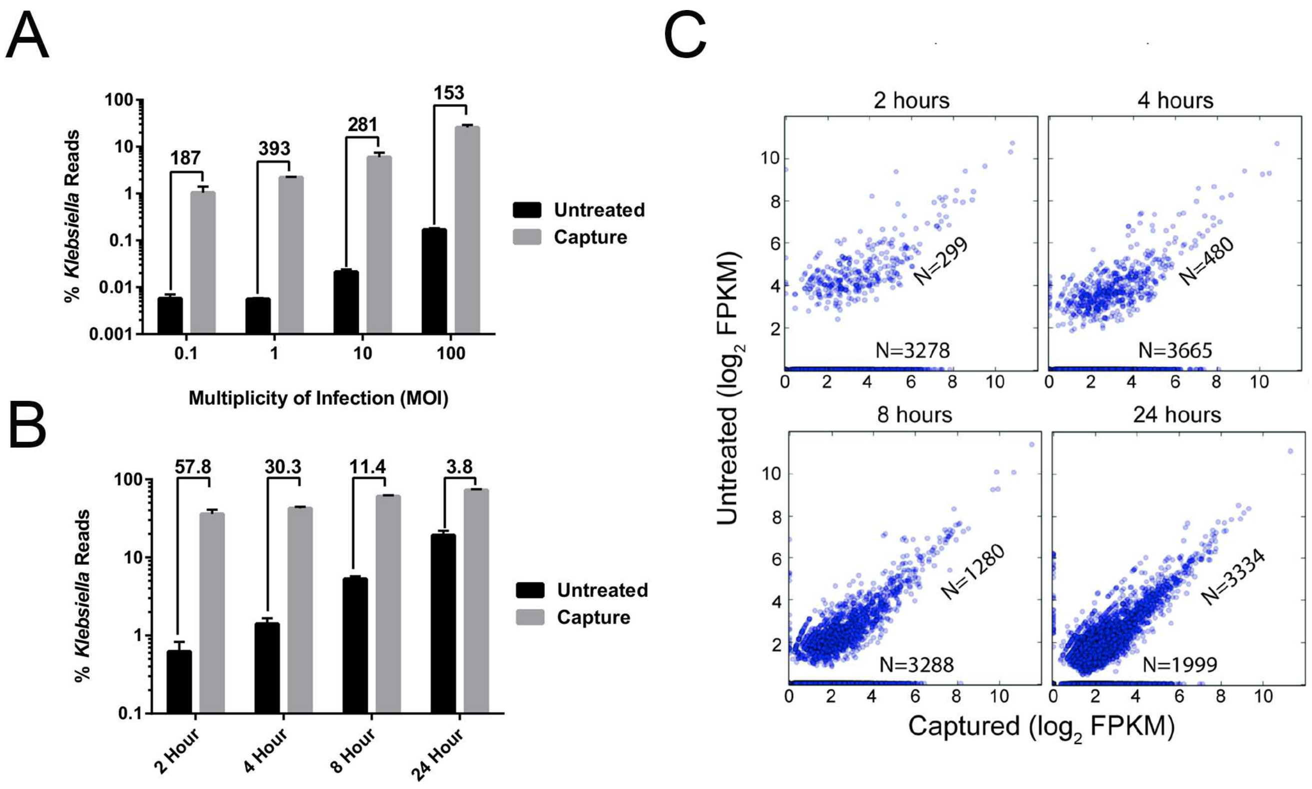

Fig 2. Characterizing performance of capture enrichment. Comparison of transcriptomics coverage over the captured and uncaptured samples for the infection time course. (A) Percentage of reads mapping to *K. pneumoniae* in the captured *vs* untreated controls in 1-hour infections at different MOI. (B) Percentage of reads mapping to *K. pneumoniae* in the captured *vs* untreated controls at different time points of an infection with MOI of 10. In both (A) and (B) the numbers above each group of bars refers to the fold-enrichment achieved in the captured *vs* untreated controls. Error bars refer to the standard deviation of percentage of reads from the triplicate infections. (C) Scatter plots of $\log_2$(FPKMs) for all genes comparing captured and uncaptured samples. The figures are marked by two numbers which show number of genes measured in both capture and uncaptured and genes, which were only measured in captured samples (along the x-axis). *K. pneumoniae* BAA-2146 has 5653 protein-coding genes and 190 RNA genes. The triplicate runs have been combined in each panel.

doi:10.1371/journal.pone.0168788.g002

To verify that the capture is unbiased, we first determined that Kpn2146 reads had a similar distribution over the transcriptome with or without the capture protocol. We compared the normalized expression values measured using cufflinks across the samples. A plot of FPKM values for each sample (S1 Fig) shows that the captured samples have a more complete representation of the pathogen transcriptome than non-captured samples, both in terms of the numbers of reads and the range of gene expression. This is also evident in Fig 2C, which compares normalized gene expression for all observed genes at each time point of infection in the captured *versus* uncaptured libraries. Points lying along the horizontal axis are genes that are observed in the captured library, but not observed in the uncaptured library. The number of genes observed in the captured library is more than 10-fold higher compared to the uncaptured library for the earliest infection time point (2 hours), which indicates that the capture technique is revealing genes that could not be observed by a standard RNA-seq protocol. For genes that are observed in both the captured and uncaptured libraries, the expression pattern of the genes are well correlated ($R^2 \sim 0.83–0.98$) for captured *versus* uncaptured libraries, as indicated by the clustering of genes along the $y = x$ lines in Fig 2B. This demonstrates that the capture protocol increases the sensitivity of RNA-seq without significantly distorting the gene expression levels, consistent with the observation made in our previous study with the lower-throughput version of our capture technique [7]. Uncaptured libraries are biased toward highly expressed genes at early time points in the infection; this bias decreases as the infection progresses. One simple explanation for this phenomenon is that growth of the bacteria over the time course of the infection leading to more bacterial RNA in the sample (up to 68% of total reads by the final time point). With the capture protocol the total reads mapping to bacterial transcripts increases by over two orders of magnitude, which makes it possible to measure a wider range of gene expression in the samples. Both captured and uncaptured samples show an increase in the number of activated genes during the course of the infection. The efficiency and sensitivity of RNA-seq for pathogens is significantly increased by the capture method.

To further assess reproducibility of the technique, and to address the possibilities of bias or probe saturation for highly expressed transcripts, we performed capture enrichment in triplicate using RNA isolated from three pure cultures of Kpn2146. As expected we found similar representation of genes in these libraries with and without the capture approach. A composite scatter plot showing FPKM for RNA-seq with and without the capture protocol with all three replicates combined is presented in S2 Fig. As with the infection models, the points clustered along the $y = x$ line, with no deviation from this line for highly expressed genes (probe saturation would appear as a deviation from this line, with reduced representation of highly expressed genes following capture).

Additional visualizations of the sample-to-sample reproducibility and bias associated with capture are presented in S3–S5 Figs. These figures allow direct comparison of individual replicates from the pure culture experiment and a subset of the infection experiments, in the form of an array of scatter plots of FPKM allowing comparison of each individual replicate against each other replicate (S3 Fig), and "all *versus* all" heat maps showing distance (the inverse of similarity or correlation) between replicates of selected data sets (S4 and S5 Figs). To summarize these supplemental figures: the captured and uncaptured data sets from the pure culture experiments are highly similar to one another. The individual replicates of each particular condition tend to cluster together; *i.e.* are most similar to each other. The infection experiments form a separate cluster that is highly dissimilar to the pure culture experiments. Within each condition, the captured and uncaptured data sets appear dissimilar to each other due to the significantly different representation of the bacterial transcriptome in the captured data sets.

Survey of gene expression in *K. pneumoniae* infection of murine macrophage cells

Kpn2146 is notable for its large arsenal of antimicrobial resistance genes as well as mutations in drug target sites and efflux pump regulatory elements [22], giving rise to a "pandrug-resistant" phenotype (*i.e.* resistant to all 34 drugs and drug/inhibitor combinations tested in a standard automated susceptibility testing panel for Gram-negative bacteria; the results are available at https://www.atcc.org/~/media/BA6C8F7C7C4C4649B2AEF501E51D76B8.ashx, collected 9/6/2016). Expression of genes associated with antimicrobial resistance over the time course of the infection are presented in the heat map in Fig 3A. Genes encoding drug resistance enzymes have been grouped together, with β-lactamase encoding enzymes as a subgroup, and multidrug efflux pumps grouped separately. Due to the low number of reads mapping to the bacteria in the uncaptured libraries, the gene expression data is sparse, whereas in capture samples the expression profile is visible for the whole time course. Expression of β-lactamase-encoding genes, including *bla*$_{\text{NDM-1}}$, was high at all time points for all the samples, even though antibiotics were absent, consistent with previous studies [36]. Follow-up studies are planned to determine which, if any, of the resistance determinants (including efflux pumps) show inducible expression in the presence of antimicrobial drugs. We note further that 19 of the drug resistance genes are encoded on the three large plasmids of Kpn2146, along with

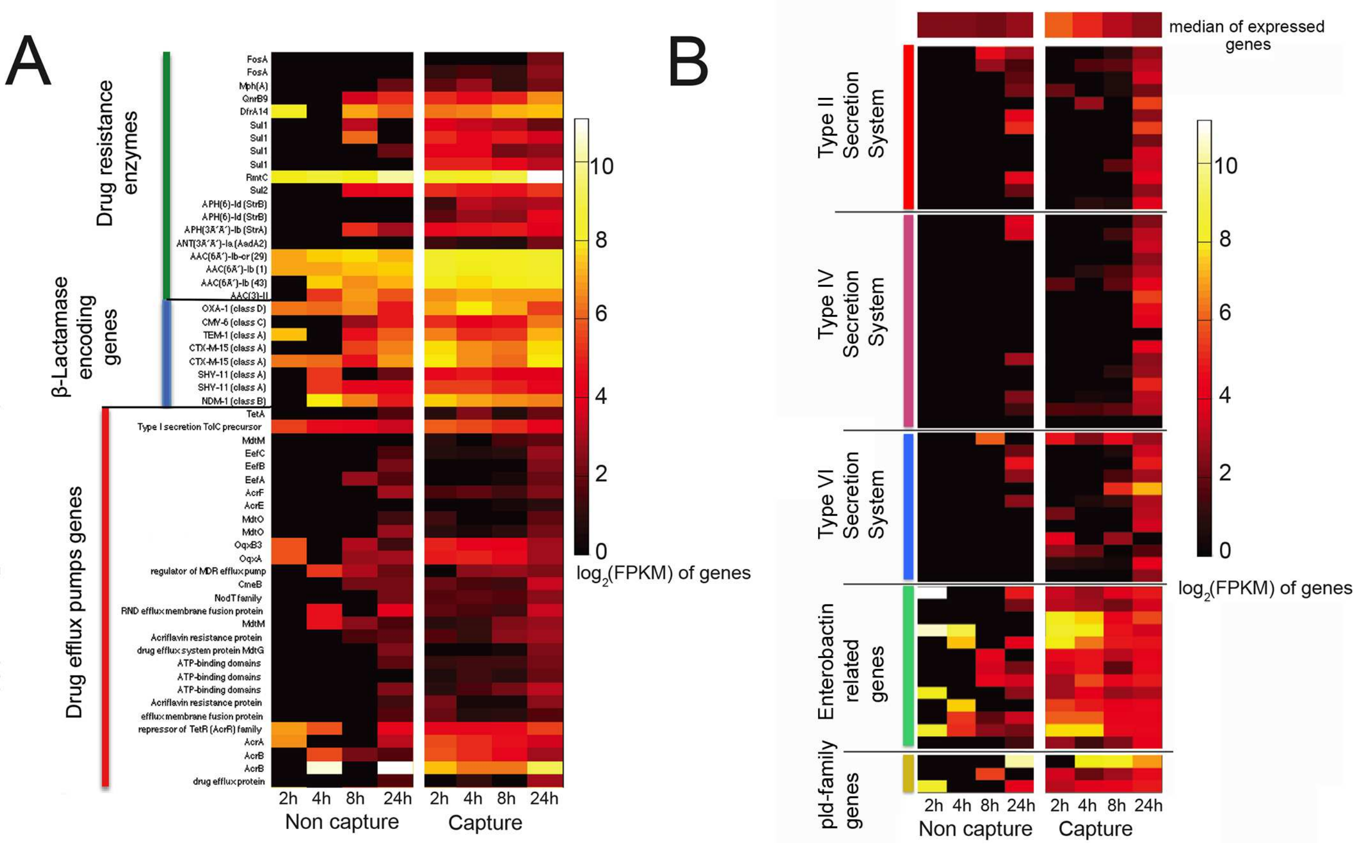

Fig 3. Dynamics of gene expression during infection. Heatmap representation of gene expression (log$_2$(FPKMs)) for A) genes involved in antibiotic resistance over the time course of infection. This includes drug resistance enzymes, β-Lactamase encoding genes and drug efflux pump encoding genes. Genes are clustered according to the annotation class shown on the left. B) gene expression (log$_2$(FPKMs)) pattern for genes putatively associated with virulence in *K. pneumoniae*. Genes are clustered according the annotation class shown on the left. For panel B of the heat map, a list of gene names in the order in which they appear is provided in S1 Table.

doi:10.1371/journal.pone.0168788.g003

numerous other genes associated with stress and survival. Our approach, generating probes directly from the total DNA of the isolate, allows us to enrich for transcripts regardless of plasmid or chromosomal origin. S6 Fig illustrates mapping of reads to the plasmid pNDM-US at the first and last time points, illustrating that the plasmid shows the same trend as the chromosome, of enriching for low-expressed genes.

Several putative virulence factors have previously been described in *Klebsiella pneumoniae* [37], of which capsule synthesis (*cps* genes) is the best characterized [38]. Kpn2146 does not display a hypermucoviscous phenotype, and lacks the *rmpA* regulator of capsule expression frequently observed in virulent *Klebsiella pneumoniae* infections [39, 40]. Heat maps showing gene expression for several other putative virulence factors are shown in Fig 3B [37, 41, 42]. Of these systems, genes associated with the siderophore enterobactin are highly expressed at early time points. Iron acquisition systems are well-known factors for *in vivo* growth of pathogens [43, 44], and enterobactin is the only siderophore we identified in the Kpn2146 genome. Improved coverage across the entire cluster of enterobactin-related genes is observed in the captured data set, compared to incomplete coverage in the non-captured data set. The Type II, Type IV (conjugation) and Type VI secretion systems all have relatively low expression early in the infection, although particularly for the Type VI system, the capture data set reveals low-level expression among several components of the systems that are not visible in the non-captured data set. Among the phospholipase genes, we observe the highest level of expression of *pld2* (the top row of the pld-family genes in the heat map), although *pld1* and *pldA* are expressed as well. The bacterial phospholipases are generally believed to play a role in membrane disruption and cell invasion by Gram-negative bacteria [45, 46]. The *pld1* gene was previously identified as a virulence factor in *K. pneumoniae* [37] although we are not aware of specific roles elucidated for *pldA* and *pld2* in virulence of *K. pneumoniae*.

Kpn2146 has 11 identified genomic islands [22]. Genes encoded within genomic islands are frequently associated with adaptation and pathogenicity [47, 48]. The gene expression profile within genomic islands over the time course of the infection and for both captured and uncaptured libraries is shown in S7 Fig. The capture method reveals that certain genomic islands (Kpn23SapB, Kpn40GuaA, Kpn16Fis) have higher gene expression than others at the 24 hour time point of the infection (S7 and S8 Figs).

For the sake of researchers who wish to analyze our data from the infection experiments for additional features, we provide as supplementary information tables of the read counts (S2 Table) and FPKM (S3 Table), as well as the corresponding annotation file (S4 Table). These tables provide data from each replicate of the time course and MOI experiments described above. The raw sequencing reads are provided in the NCBI Sequence Read Archive (accession number PRJNA317373).

We note that the host transcriptome can be profiled in an unbiased fashion from the original uncaptured library, allowing parallel analysis of the host and pathogen. We do not belabor the host results here, as this preliminary study was performed in a cell line which is not an accurate model of host behavior, *versus* an animal model or even primary cell model. The capture methodology is expected to enable such parallel host- and pathogen RNA-seq (or "Dual RNA-seq") experiments in more complex models, with unprecedented resolution of the pathogen transcriptome. While this study was performed with a bacterial pathogen, based on our previous work [7] we believe viral transcriptomes could be captured using this technique as well.

Pathogen transcript enrichment protocol

The entire capture protocol takes approximately 24 hours, much of which is the overnight incubation for hybridization of capture probes. The protocol requires approximately 1.5 hours of hands-on time to process 12 samples in parallel, most of which is the series of washes performed after capturing the biotinylated probes on the avidin resin in the spin columns. We have greatly simplified the protocol from our previous work, which was performed in a custom-built microfluidic apparatus capable of processing one sample at a time, to the version described here, which uses disposable plastic spin columns in a small centrifuge with a 12-position rotor. This enables the protocol to be carried out in any well-equipped molecular biology laboratory, without any custom-built components. Temperature control is crucial; the most expensive single piece of equipment required is an incubator capable of maintaining at least 60˚C, and large enough to house a small centrifuge and all of the necessary buffers and pipettes. We use an incubator with arm holes to allow a researcher to place his or her arms inside the incubator without opening a door, which helps maintain the temperature within the incubator. A photograph of the setup is illustrated in S9 Fig. Our previous method required approximately three days for the capture step alone, whereas the current approach based on spin columns allows preparation of the sequencing-ready library in two days. Further automation of the protocol is under development.

A notable advantage of our protocol for bacterial transcriptomics is the use of biotinylated capture probes that are synthesized directly from the genomic DNA of the bacterium of interest. Most other targeted capture techniques reported in literature rely upon capture probes or "baits" that comprise large libraries (ranging from hundreds to tens of thousands) of discrete synthetic DNA or RNA oligonucleotides, designed specifically to target sequences of interest. This approach is attractive, for example, for targeted capture of human exome sequences, where the same library of capture probes can be applied to many different studies. Synthetic capture probes may be an attractive option for organisms with very small genomes (*e.g.* RNA viruses), where the entire genome can be covered by a relatively small number of probes. Bacteria such as *K. pneumoniae*, however, have genomes that are both too large, and too highly variable (or mosaic) to be easily covered by a single set of synthetic capture probes: Kpn2146, has a genome of approximately 6 Mbp (including over 300 kbp on four plasmids); this would require a collection of 60,000 oligos 100 bases in length to completely cover the genome at 1X depth, and several times this number would be required to create a tiled capture array for more efficient capture. Given the large genomic variability between strains, and particularly since interesting genes are often located within mobile elements such as plasmids and genomic islands, a probe library synthesized for one strain may not be optimal for another. A study, for example, to compare the global gene expression of two or more strains becomes impractical if relying upon synthetic capture probe libraries. Our enzymatic synthesis technique, by contrast, allows creation of a random library covering the entire genome, in less than a day, for less than $20 per sample, from genomic DNA that can be produced quickly in large quantities from easily cultured organisms such as *K. pneumoniae*. Gene-targeted approaches have missed subsequently-discovered genes [49]; the unbiased genomic coverage achieved here can address unusual RNA genes or small protein genes that may only become delineated subsequent to the experiment. This method provides great savings over a library of chemically synthesized probes which requires careful design and several weeks to a few months of turnaround time, at a much higher cost per sample. Chemical synthesis of a custom probe library would only become cost-effective for large-scale studies (96 samples or more) with only a single, previously-sequenced bacterial strain of interest, where the same library can be used for many samples.

We note that our protocol bears some resemblance to the selective capture of transcribed sequences (SCOTS) technique that has been applied to microarray studies bacterial transcriptomics in infection models [50, 51]. The SCOTS technique relies upon a similar approach to preparing capture probes by biotinylating genomic DNA from the organism of interest. However, to achieve sufficient enrichment for microarray analysis, typically three rounds of hybridization and capture were performed, with an intermediate PCR amplification using a conserved primer. Besides being more time consuming, the intermediate PCR steps (up to 90 cycles across several rounds) may lead to additional bias. By contrast, we demonstrate here that a single round of capture-based enrichment is sufficient to dramatically improve coverage of pathogen sequences using RNA-seq.

Conclusion

We have developed an improved method for targeted capture of bacterial transcripts from a mixed sample, in this case an infection model characterized by an overwhelming abundance of host RNA. The new protocol can be carried out in most well-equipped laboratories, and significantly improves the representation of bacterial genomes in an infection model. We established the capabilities of the technique using a multi-drug resistant *K. pneumoniae* isolate in a murine macrophage cell line. We demonstrated >100-fold enrichment of bacterial transcripts, allowing analysis of many genes that could not observed without enrichment, without dramatically biasing the relative expression between the captured and uncaptured libraries. Although this study was designed primarily to characterize the method, the results indicate expression of numerous drug resistance genes during the infection, with several drug-modifying enzymes expressed at high levels relative to the median of expressed genes, even in the absence of drugs. The results also suggest a role for the siderophore enterobactin early in the infection. Follow-up studies will further explore the biological significance of the expressed genes, including the effect of antimicrobial drugs on gene expression, as well as studying the coupled dynamics of host and pathogen gene expression in a more realistic infection model.

Supporting Information

S1 Fig. Capture increases representation of low-expressed genes. Box scatter plot of FPKM values for captured and uncaptured libraries, for timecourse infections with MOI of 10. The Capture samples reveal many more transcripts with low-level expression, that are not visible in the Uncaptured libraries.
(PNG)

S2 Fig. Capture enrichment does not introduce bias in pure culture samples. (A) $\log_2$ gene FPKM counts were used to make scatter plots for the combined replicates of samples of culture treated with and without capture treatment. The R^2 value of 0.94 was was calculated for the linear model fit of the scatter plot between the mentioned datasets, indicating high correlation between the captured and uncaptured datasets, with a high degree of linearity for highly expressed genes (indicating a lack of probe saturation effects). B-C) similarly the datasets used in A are used to compare with a dataset for infection sample which had a different gene expression profile. The R^2 value computed for B and C are much lower than the culture captured and uncaptured sample in figure A.
(TIF)

S3 Fig. Comparison of biological replicate experiments. This figure presents an array of scatter plots allowing comparison of individual pure culture experiments, including replicates of individual conditions, as well as comparisons between captured and uncaptured. For comparison, a set of infection samples (24 hour time point, with capture) are also included, showing that the infection experiments are dissimilar from the pure culture experiments, but more similar to each other. The plots along the main diagonal are histograms of FPKM for each condition.
(PDF)

S4 Fig. Reproducibility of pure culture capture experiments. Each replicate of the samples tested were analyzed using the YAnTra pipeline to measure the FPKM counts of the gene feature. Sample to sample pairwise Pearson correlation coefficient was calculated for the overall gene expression profile of the samples. The distance matrix generated as 1-p is clustered and represented as a heatmap, with black indicating zero distance between samples, and bright yellow indicating maximum distance. The replicate samples analyzed form two prominent clusters: The samples sequenced coming from the culture samples for both captured and uncaptured, and the samples from the captured infection samples (24 hour time point) which were used to root the dendogram for clustering analysis.
(PNG)

S5 Fig. Reproducibilty of capture enrichment experiments. As in S4 Fig, this figure presents a heatmap of sample to sample Pearson correlation coefficient of selected data sets after the hierarchical clustering, this time including the infection experiments (all time points, all replicates, with and without capture). The pure culture experiments (without capture) are included

as an outgroup.
(PNG)

S6 Fig. Capture improves coverage at early time points in infection. Mapping of reads to the plasmid pNDM-US, with and without capture, at early (2 hour) and late (24 hour) time points in infection.
(PNG)

S7 Fig. Expression within genomic islands. Heatmap representation of gene expression (log2 (FPKMs)) for genes located within genomic islands of Kpn2146.
(PNG)

S8 Fig. Expression within genomic islands compared to median gene expression. Boxplots for gene expression distribution at 24 hr infection time point for the genomic islands. The black line shows the median gene expression of genes in non-island genes.
(PNG)

S9 Fig. Experimental apparatus. Photographs of the equipment used for the spin column-based capture protocol carried out in an incubator with arm holes. The inset at the lower right shows a larger view of the spin column containing the monomeric avidin resin.
(JPG)

S1 Table. Gene names for Fig 3B. Due to size constraints the names of individual genes are omitted from the rows of Fig 3B. The table provides a list of the locus tags with corresponding names from the annotation file, in the order in which they appear in the heat map.
(CSV)

S2 Table. Read counts for infection experiments. The read counts for each sequencing experiment are provided, with the gene locus (corresponding to the annotation file in S4 Table) followed by the number of reads. The table includes three biological replicates for each experiment (time course and MOI experiments), with both captured and uncaptured data sets.
(CSV)

S3 Table. FPKM for infection experiments. The cFPKM for each sequencing experiment are provided, with the gene locus (corresponding to the annotation file in S4 Table) followed by the calculated FPKM. The table includes three biological replicates for each experiment (time course and MOI experiments), with both captured and uncaptured data sets.
(CSV)

S4 Table. Annotation file for Kpn2146. The table provides data on location, sense, and annotation for **each** gene referred to in S1, S2, and S3 Tables.
(CSV)

Author Contributions

Conceptualization: ZWB RJM.

Data curation: KP KPW.

Formal analysis: ZWB KP AEL KPW RJM.

Funding acquisition: RJM.

Investigation: ZWB AEL RH.

Methodology: ZWB KP AEL RH KPW RJM.

Project administration: RJM.

Software: KP KPW.

Supervision: ZWB RJM.

Validation: ZWB AEL.

Visualization: KP AEL.

Writing – original draft: ZWB KP AEL RJM.

Writing – review & editing: ZWB KP AEL RJM.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014; 30(9):418–26. doi: 10.1016/j.tig.2014.07.001 PMID: 25108476
2. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10(1):57–63. PubMed Central PMCID: PMC2949280. doi: 10.1038/nrg2484 PMID: 19015660
3. Ordas A, Hegedus Z, Henkel CV, Stockhammer OW, Butler D, Jansen HJ, et al. Deep sequencing of the innate immune transcriptomic response of zebrafish embryos to Salmonella infection. Fish Shellfish Immunol. 2011; 31(5):716–24. doi: 10.1016/j.fsi.2010.08.022 PMID: 20816807
4. Wilk E, Pandey AK, Leist SR, Hatesuer B, Preusse M, Pommerenke C, et al. RNAseq expression analysis of resistant and susceptible mice after influenza A virus infection identifies novel genes associated with virus replication and important for host resistance to infection. BMC Genomics. 2015; 16:655. PubMed Central PMCID: PMC4557482. doi: 10.1186/s12864-015-1867-8 PMID: 26329040
5. Brady RA, Bruno VM, Burns DL. RNA-Seq Analysis of the Host Response to Staphylococcus aureus Skin and Soft Tissue Infection in a Mouse Model. PLoS One. 2015; 10(4):e0124877. PubMed Central PMCID: PMC4406450. doi: 10.1371/journal.pone.0124877 PMID: 25901897
6. Bent ZW, Brazel DM, Tran-Gyamfi MB, Hamblin RY, VanderNoot VA, Branda SS. Use of a capture-based pathogen transcript enrichment strategy for RNA-Seq analysis of the Francisella Tularensis LVS transcriptome during infection of murine macrophages. PLoS One. 2013; 8(10):e77834. Epub 2013/10/25. PubMed Central PMCID: PMC3796476. doi: 10.1371/journal.pone.0077834 PMID: 24155975
7. Bent ZW, Tran-Gyamfi MB, Langevin SA, Brazel DM, Hamblin RY, Branda SS, et al. Enriching pathogen transcripts from infected samples: A capture-based approach to enhanced host-pathogen RNA sequencing. Anal Biochem. 2013; 438(1):90–6. Epub 2013/03/29. doi: 10.1016/j.ab.2013.03.008 PMID: 23535274
8. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol. 2012; 10 (9):618–30. Epub 2012/08/15. doi: 10.1038/nrmicro2852 PMID: 22890146
9. Tierney L, Linde J, Muller S, Brunke S, Molina JC, Hube B, et al. An Interspecies Regulatory Network Inferred from Simultaneous RNA-seq of Candida albicans Invading Innate Immune Cells. Front Microbiol. 2012; 3:85. PubMed Central PMCID: PMC3299011. doi: 10.3389/fmicb.2012.00085 PMID: 22416242
10. Kawahara Y, Oono Y, Kanamori H, Matsumoto T, Itoh T, Minami E. Simultaneous RNA-seq analysis of a mixed transcriptome of rice and blast fungus interaction. PLoS One. 2012; 7(11):e49423. PubMed Central PMCID: PMC3490861. doi: 10.1371/journal.pone.0049423 PMID: 23139845
11. Eriksson S, Lucchini S, Thompson A, Rhen M, Hinton JCD. Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. Molecular microbiology. 2003; 47(1):103–18. PMID: 12492857
12. Waddell SJ, Butcher PD, Stoker NG. RNA profiling in host-pathogen interactions. Current Opinion in Microbiology. 2007; 10(3):297–302. doi: 10.1016/j.mib.2007.05.013 PMID: 17574903
13. Srikumar S, Kröger C, Hébrard M, Colgan A, Owen SV, Sivasankaran SK, et al. RNA-seq Brings New Insights to the Intra-Macrophage Transcriptome of *Salmonella* Typhimurium. PLoS Pathog. 2015; 11 (11):e1005262. doi: 10.1371/journal.ppat.1005262 PMID: 26561851
14. Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, et al. Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions. Nature. 2016; 529(7587):496–501. doi: 10.1038/nature16547 PMID: 26789254

15. Date SV, Modrusan Z, Lawrence M, Morisaki JH, Toy K, Shah IM, et al. Global gene expression of methicillin-resistant Staphylococcus aureus USA300 during human and mouse infection. J Infect Dis. 2014; 209(10):1542–50. doi: 10.1093/infdis/jit668 PMID: 24286981

16. Baddal B, Muzzi A, Censini S, Calogero RA, Torricelli G, Guidotti S, et al. Dual RNA-seq of Nontypeable Haemophilus influenzae and Host Cell Transcriptomes Reveals Novel Insights into Host-Pathogen Cross Talk. mBio. 2015; 6(6):e01765–15. Epub 2015/11/19. PubMed Central PMCID: PMCPMC4659474. doi: 10.1128/mBio.01765-15 PMID: 26578681

17. Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, Waldor MK. RNA-seq-based monitoring of infection-linked changes in Vibrio cholerae gene expression. Cell host & microbe. 2011; 10(2):165–74.

18. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. Genome biology. 2016; 17:41. Epub 2016/03/06. PubMed Central PMCID: PMCPMC4778327. doi: 10.1186/s13059-016-0904-5 PMID: 26944702

19. Vandernoot VA, Langevin SA, Solberg OD, Lane PD, Curtis DJ, Bent ZW, et al. cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications. Biotechniques. 2012; 53(6):373–80. Epub 2012/12/12. doi: 10.2144/000113937 PMID: 23227988

20. He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, et al. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. Nat Methods. 2010; 7(10):807–12. doi: 10.1038/nmeth.1507 PMID: 20852648

21. Centers for Disease Control and Prevention. Detection of Enterobacteriaceae isolates carrying metallo-beta-lactamase—United States, 2010. MMWR Morb Mortal Wkly Rep. 2010; 59(24):750. Epub 2010/06/26. PMID: 20577157

22. Hudson CM, Bent ZW, Meagher RJ, Williams KP. Resistance determinants and mobile genetic elements of an NDM-1-encoding Klebsiella pneumoniae strain. PLoS One. 2014; 9(6):e99209. PubMed Central PMCID: PMC4048246. doi: 10.1371/journal.pone.0099209 PMID: 24905728

23. Cano V, March C, Insua JL, Aguiló N, Llobet E, Moranta D, et al. Klebsiella pneumoniae survives within macrophages by avoiding delivery to lysosomes. Cellular Microbiology. 2015; 17(11):1537–60. doi: 10.1111/cmi.12466 PMID: 26045209

24. Minato Y, Ghosh A, Faulkner WJ, Lind EJ, Schesser Bartra S, Plano GV, et al. Na+/H+ antiport is essential for Yersinia pestis virulence. Infect Immun. 2013; 81(9):3163–72. Epub 2013/06/19. PubMed Central PMCID: PMC3754208. doi: 10.1128/IAI.00071-13 PMID: 23774602

25. Langevin SA, Bent ZW, Solberg OD, Curtis DJ, Lane PD, Williams KP, et al. Peregrine: A rapid and unbiased method to produce strand-specific RNA-Seq libraries from small quantities of starting material. RNA Biol. 2013; 10(4). Epub 2013/04/06.

26. Bent ZW, Poorey K, Brazel DM, LaBauve AE, Sinha A, Curtis DJ, et al. Transcriptomic Analysis of Yersinia enterocolitica Biovar 1B Infecting Murine Macrophages Reveals New Mechanisms of Extracellular and Intracellular Survival. Infect Immun. 2015; 83(7):2672–85. PubMed Central PMCID: PMC4468540. doi: 10.1128/IAI.02922-14 PMID: 25895974

27. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology. 2013; 14(4): R36. Epub 2013/04/27. doi: 10.1186/gb-2013-14-4-r36 PMID: 23618408

28. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. PubMed Central PMCID: PMC3322381. doi: 10.1038/nmeth.1923 PMID: 22388286

29. Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome biology. 2011; 12(3):R22. PubMed Central PMCID: PMC3129672. doi: 10.1186/gb-2011-12-3-r22 PMID: 21410973

30. Gupta N, Limbago BM, Patel JB, Kallen AJ. Carbapenem-resistant Enterobacteriaceae: epidemiology and prevention. Clin Infect Dis. 2011; 53(1):60–7. Epub 2011/06/10. doi: 10.1093/cid/cir202 PMID: 21653305

31. Jacob JT, Klein E, Laxminarayan R, Beldvas Z, Lynfield R, Kallen AJ, et al. Vital signs: carbapenem-resistant Enterobacteriaceae. MMWR Morb Mortal Wkly Rep. 2013; 62(9):165–70. Epub 2013/03/08. PMID: 23466435

32. Guh AY, Limbago BM, Kallen AJ. Epidemiology and prevention of carbapenem-resistant Enterobacteriaceae in the United States. Expert review of anti-infective therapy. 2014; 12(5):565–80. Epub 2014/03/29. doi: 10.1586/14787210.2014.902306 PMID: 24666262

33. Tangden T, Giske CG. Global dissemination of extensively drug-resistant carbapenemase-producing Enterobacteriaceae: clinical perspectives on detection, treatment and infection control. J Intern Med. 2015; 277(5):501–12. doi: 10.1111/joim.12342 PMID: 25556628

34. Avican K, Fahlgren A, Huss M, Heroven AK, Beckstette M, Dersch P, et al. Reprogramming of Yersinia from virulent to persistent mode revealed by complex in vivo RNA-seq analysis. PLoS Pathog. 2015; 11 (1):e1004600. PubMed Central PMCID: PMC4295882. doi: 10.1371/journal.ppat.1004600 PMID: 25590628

35. Li J, Liu F, Wang Q, Ge P, Woo PCY, Yan J, et al. Genomic and transcriptomic analysis of NDM-1 Klebsiella pneumoniae in spaceflight reveal mechanisms underlying environmental adaptability. Scientific Reports. 2014; 4:6216. doi: 10.1038/srep06216 PMID: 25163721

36. Canton R, Gonzalez-Alba JM, Galan JC. CTX-M Enzymes: Origin and Diffusion. Front Microbiol. 2012; 3:110. PubMed Central PMCID: PMC3316993. doi: 10.3389/fmicb.2012.00110 PMID: 22485109

37. Lery LM, Frangeul L, Tomas A, Passet V, Almeida AS, Bialek-Davenet S, et al. Comparative analysis of Klebsiella pneumoniae genomes identifies a phospholipase D family protein as a novel virulence factor. BMC Biology. 2014; 12(1):1–15.

38. Cortes G, Borrell N, de Astorza B, Gomez C, Sauleda J, Alberti S. Molecular analysis of the contribution of the capsular polysaccharide and the lipopolysaccharide O side chain to the virulence of Klebsiella pneumoniae in a murine model of pneumonia. Infect Immun. 2002; 70(5):2583–90. Epub 2002/04/16. PubMed Central PMCID: PMCPMC127904. doi: 10.1128/IAI.70.5.2583-2590.2002 PMID: 11953399

39. Nassif X, Honore N, Vasselon T, Cole ST, Sansonetti PJ. Positive control of colanic acid synthesis in Escherichia coli by rmpA and rmpB, two virulence-plasmid genes of Klebsiella pneumoniae. Molecular microbiology. 1989; 3(10):1349–59. Epub 1989/10/01. PMID: 2693894

40. Shon AS, Bajwa RPS, Russo TA. Hypervirulent (hypermucoviscous) Klebsiella pneumoniae: A new and dangerous breed. Virulence. 2013; 4(2):107–18. doi: 10.4161/viru.22718 PMID: 23302790

41. Sandkvist M. Type II Secretion and Pathogenesis. Infection and Immunity. 2001; 69(6):3523–35. doi: 10.1128/IAI.69.6.3523-3535.2001 PMID: 11349009

42. Tomas A, Lery L, Regueiro V, Perez-Gutierrez C, Martinez V, Moranta D, et al. Functional Genomic Screen Identifies Klebsiella pneumoniae Factors Implicated in Blocking Nuclear Factor kappaB (NF-kappaB) Signaling. The Journal of biological chemistry. 2015; 290(27):16678–97. Epub 2015/05/15. PubMed Central PMCID: PMCPMC4505419. doi: 10.1074/jbc.M114.621292 PMID: 25971969

43. Lawlor MS, O'Connor C, Miller VL. Yersiniabactin is a virulence factor for Klebsiella pneumoniae during pulmonary infection. Infect Immun. 2007; 75(3):1463–72. Epub 2007/01/16. PubMed Central PMCID: PMCPMC1828572. doi: 10.1128/IAI.00372-06 PMID: 17220312

44. Yancey RJ, Breeding SA, Lankford CE. Enterochelin (enterobactin): virulence factor for Salmonella typhimurium. Infect Immun. 1979; 24(1):174–80. Epub 1979/04/01. PubMed Central PMCID: PMCPMC414280. PMID: 156695

45. Istivan TS, Coloe PJ. Phospholipase A in Gram-negative bacteria and its role in pathogenesis. Microbiology. 2006; 152(5):1263–74.

46. Stahl J, Bergmann H, Göttig S, Ebersberger I, Averhoff B. Acinetobacter baumannii Virulence Is Mediated by the Concerted Action of Three Phospholipases D. PLoS ONE. 2015; 10(9):e0138360. doi: 10.1371/journal.pone.0138360 PMID: 26379240

47. Gyles CL, Palchaudhuri S, Maas WK. Naturally occurring plasmid carrying genes for enterotoxin production and drug resistance. Science. 1977; 198(4313):198–9. PMID: 333581

48. Frost LS, Leplae R, Summers AO, Toussaint A. Mobile genetic elements: the agents of open source evolution. Nat Rev Microbiol. 2005; 3(9):722–32. doi: 10.1038/nrmicro1235 PMID: 16138100

49. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. Genetics. 2001; 158(1):41–64. PubMed Central PMCID: PMC1461638. PMID: 11333217

50. Faucher SP, Porwollik S, Dozois CM, McClelland M, Daigle F. Transcriptome of Salmonella enterica serovar Typhi within macrophages revealed through the selective capture of transcribed sequences. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103(6):1906–11. doi: 10.1073/pnas.0509183103 PMID: 16443683

51. Graham JE, Clark-Curtiss JE. Identification of Mycobacterium tuberculosis RNAs synthesized in response to phagocytosis by human macrophages by selective capture of transcribed sequences (SCOTS). Proceedings of the National Academy of Sciences. 1999; 96(20):11554–9.