White Paper Application

Project Title: Using next-generation sequencing technology to define the *Vibrio cholerae* transcriptome

Authors: Matthew Waldor, Anjali Mandlik, Jonathan Livny

Primary Investigator Contact:

Name	Matthew Waldor
Position	Professor of Medicine
Institution	Brigham and Women's Hospital/Harvard Medical School
Address	181 Longwood Avenue, Boston
State	MA
ZIP Code	02115
Telephone	617-525-4646
Fax	617-525-4660
E-Mail	mwaldor@rics.bwh.harvard.edu

1. Executive Summary

Vibrio cholerae is an intestinal pathogen that causes cholera, a severe and sometime lethal diarrheal disease. *V. cholerae* is a Gram-negative facultative pathogen that can inhabit two distinct niches; this highly motile bacterium lives in the natural aquatic reservoir and can also survive and multiply in the human intestine, causing devastating illness. A primary mechanism by which the organism adapts to these different environments is through changes in global patterns of transcription. These changes alter the organism's metabolic and protein repertoire (Mekalanos, Rubin et al. 1997; Faruque, Albert et al. 1998). Thus, information about global changes in the *V. cholerae* 'transcriptome' in different *in vitro* as well as *in vivo* conditions will be invaluable to enable analyses of these adaptive processes and provide potential insight for therapeutics.

The current information about the *V. cholerae* transcriptome comes from recombination based *in vivo* technology (RIVET) (Camilli, Beattie et al. 1994) and microarray analyses (Xu, Dziejman et al. 2003). The first technique is extremely laborious and is very low throughput and does not allow for comprehensive global analyses of the transcriptome. Microarrays and other hybridization-based techniques have several important limitations (Morozova, Hirst et al. 2009; van Vliet 2009). In nearly all cases, microarrays do not contain complete representations of both strands of the entire genome. For example, small noncoding RNA species, which are often encoded in 'intergenic' regions, are usually not present in microarrays. Furthermore, most microarrays do not contain DNA that corresponds to 'antisense' transcripts. However, recent studies suggest that antisense transcription may be common (Brantl 2007). In addition, microarrays do not allow for identification of primary versus processed transcripts. Finally, cross-hybridizations can limit the specificity of microarray studies.

The development of 'next-generation sequencing technology'(Mardis 2008) and 'RNA sequencing technology' (Wang, Gerstein et al. 2009; Wilhelm and Landry 2009) is enabling a new level of depth and accuracy in high-throughput transcriptome analyses, without the above mentioned drawbacks. We propose to utilize Illumina next-generation RNA sequencing technology to define the complete *V. cholerae* transcriptome under a

variety of conditions to further our understanding of how *V. cholerae* adapts to the varied niches it can inhabit. Here, we propose to:

1. Define the *in vitro* (lab culture) and *in vivo* (infant rabbit model) *V. cholerae* transcriptome. This will allow us to compare and analyze the sets of genes expressed in these different conditions, thus giving insights into virulence regulation and adaptation.
2. Define the primary transcriptome of *V. cholerae* by combining rapid amplification of cDNA ends (RACE) with RNA-seq to allow global determination of transcription start sites. This will enable differentiation between *de novo* and modified transcripts providing information about post-transcriptional regulatory processes.
3. Identify novel RNA regulatory species including *trans*-acting transcripts (small RNA), *cis*-acting RNAs and antisense RNAs.

Our work will be the first comprehensive study of the *V. cholerae* transcriptome. This sequencing project will provide valuable insights into virulence regulation, post-transcriptional regulation, adaptive mechanisms and host-pathogen interactions. Furthermore, the tools we develop here will be broadly applicable.

2. Justification

1. *V. cholerae* is an extracellular pathogen of the small intestine and causes cholera which is a diarrheal disease. Infected individuals can become rapidly dehydrated from the severe watery diarrhea and without treatment and rehydration, can die within 24 h. Since 1817, there have been seven cholera pandemics. The seventh and present pandemic began in 1961 and the burden of cholera is estimated to reach several million cases a year in Asia and Africa with fewer cases in Latin America (Faruque, Albert et al. 1998).

2. Whole genome sequencing projects have been completed for several strains of *V. cholerae.* 15 different clinical and environmental isolates have been sequenced through the NIAID GSC projects at JCVI and TIGR[1]. Further genome sequencing is being carried out on 11 more strains at The Broad Institute[2]. Apart from this, several whole genome sequencing projects are ongoing at The National Microbial Pathogen Data Resource (NMPDR) team at Los Alamos National Laboratory and TEDA School of Biological Sciences and Biotechnology at Nankai University[3]. Several comparative genomics, evolutionary genetic analysis, metabolic pathway prediction, microarray analyses have been carried out on these generated sequences (Dziejman, Balon et al. 2002; Xu, Dziejman et al. 2003; Beyhan, Tischler et al. 2006; Shi, Romero et al. 2006). However, up to this point, there have been no reports of comprehensive transcriptome data or of global analyses of precise transcription start sites.

[1] http://www3.niaid.nih.gov/LabsAndResources/resources/mscs/completed.htm.
[2] http://www3.niaid.nih.gov/LabsAndResources/resources/mscs/ongoing.htm
[3] http://www.ncbi.nlm.nih.gov/sites/entrez?db=genomeprj&cmd=Retrieve&dopt=Overview&list_uids=15707

3. The proposed study aims to generate global transcriptome data for *V. cholerae* under both lab conditions as well as in an animal model. The study is also designed to obtain information regarding the primary transcriptome and thus will yield a comprehensive assessment of *V. cholerae's* transcription start sites. This transcriptome data will provide valuable insights into virulence mechanisms as well as new knowledge of global transcription patterns and novel RNA regulatory mechanisms.

3. Rationale for Strain Selection

In this study, we propose to obtain the transcriptome data for C6706, a clinical isolate from the ongoing seventh cholera pandemic. C6706 is a *Vibrio cholerae O1* biotype El Tor strain which was isolated from Peru in 1991. This strain is closely related to the sequenced strain N16961, the first *V. cholerae* strain sequenced and annotated (Heidelberg, Eisen et al. 2000). The rationale for selecting C6706 is the presence of a functional quorum sensing system in this strain(Zhu, Miller et al. 2002) and the availability of a defined transposon mutant library (Cameron, Urbach et al. 2008). The availability of a defined transposon mutant library will aid in downstream validation of the transcriptomics data. C6706 strain is present in the Waldor Lab strain collection. It is present in ATCC (#55331) and will be deposited into the BEI repository.

We aim to define the complete transcriptome of this strain in different conditions. For the animal studies, we have developed an infant rabbit model which very closely mimics the cholera pathology observed in humans. In this model, the rabbits suffer from profuse watery diarrhea following oro-gastric infection with *V. cholerae*. There is pronounced fluid accumulation in the cecum of the rabbit where the *V. cholerae* titres have been observed to be around $1*10^8$cfu/ml (Ritchie J and Waldor M, unpublished data). This model was developed using N16961 and C6706 strains and the disease condition caused by these strains in rabbits is under extensive study. Also our group has developed several genetic tools that will enable further analysis and validation of the transcriptome data.

4a. Approach to Data Production: **Data Generation**

1. Library preparation
   i) For each of the growth conditions we will study, we plan to define *V. cholerae's* global transcriptome:
      a) For the *in vitro* conditions, we will extract RNA from *V. cholerae* that has been grown in LB (Luria-Bertani) medium or (M9) minimal medium, which more closely resembles conditions found in the intestinal lumen. The extracted RNA will then be enriched for mRNA using a MICROBExpress (Ambion) kit.
      b) For the *in vivo* conditions, we will extract RNA from the cecal fluid of an infected rabbit. The RNA will be enriched for prokaryotic mRNA using a combination of MICROBEnrich (Ambion) and the MICROBExpress kits.
      c) The bacterial mRNA will then be processed to yield first strand cDNA using random hexamers to cover the entire transcriptome. To maintain strand information, we will produce strand specific libraries in accordance to the protocol published in Parkhomchuk et al. (Parkhomchuk, Borodina et al. 2009). Briefly, incorporation of deoxy-UTP during the second strand synthesis and

subsequent destruction of the uridine-containing strand in the sequencing library will allow us to identify orientation of the transcripts. Illumina specific adapters will be added to the cDNA and this cDNA will then be amplified to yield the "strand specific random hexamer cDNA library".

ii) To define the transcription start sites, total RNA obtained from *in vitro* conditions and enriched for mRNA using MICROBExpress kit will be our starting material for library generation. To distinguish between primary and modified transcripts we will use Terminator 5'-Phosphate-dependent Exonuclease (TEX, Epicenter Biotechnologies), a processive 5'-3' exonuclease that digests RNA having a 5'-monophosphate. Prokaryotic primary transcripts have a 5'-triphosphate and thus will not be affected by the exonuclease, whereas the degraded and modified transcripts will be digested. This will be followed by treatment with Tobacco acid pyrophosphatase (TAP, Epicenter Biotechnologies), which will convert the 5'-triphosphates to 5'-monophosphates. Illumina specific adapters will then be ligated to both ends of the mRNA, which will then be reverse transcribed to cDNA and finally amplified using Illumina specific primers to generate the "5'-specific cDNA library".

2. Data generation

   Each amplified cDNA library will be sequenced using Illumina GA sequencer at the Broad Institute. We estimate that for our analyses we will need around 5-10 million reads/library. A single lane of the Illumina GA for each library should be sufficient to generate such read yields.

   i. We propose to have a technical and biological replicate for each library and anticipate using 16 lanes in total for the phase I study. For the transcription start site identification, we will have 3 +TEX/+TAP libraries and 3 –TEX/+TAP libraries. For the global transcriptome analyses, there will be 3 libraries from cells grown in LB, 3 from cells grown in minimal media-M9 (this media more closely resembles conditions found *in vivo*) and 3 from *in vivo* conditions.
   ii. If the above analyses are successful, we will then proceed to phase II and define the transcriptome under different environmental conditions and with different defined mutants. The phase II studies will at least in part be predicated on the phase I results and analyses. However, we anticipate the phase II studies will include the following mutants and environmental conditions:
      a. *In vitro* virulence inducing conditions (AKI media) with WT and *toxT* mutant. ToxT is the major activator of virulence gene expression in *V. cholerae*.
      b. Comparison of the transcriptome of mid-log versus early stationary phase. This will be done in either LB or minimal media depending on the results of the phase I analyses.
      c. The transcriptome of a *luxO* mutant will be determined under *in vitro* and *in vivo* growth. LuxO governs quorum sensing in *V. cholerae*.
      d. The *in vivo* transcriptomes of a *cheY* and a *flaA* mutant will be determined. CheY is a chemotaxis response regulator and FlaA is a flagellin. These studies will allow us to see how the *in vivo* transcriptome of *V. cholerae* is governed by the organism's ability to sense and swim to host derived stimuli.

Each of the proposed experiments will be done in duplicates. Thus, in total the phase II studies will require 16 additional lanes.

4b. Approach to Data Production: **Data Analysis**

The Illumina sequencing described above will produce numerous large and complex datasets. Analysis of these data to define transcription start sites, generate whole transcriptome profiles, and compare these profiles across samples will be undertaken using existing scripts developed by Jonathan Livny in our laboratory (Liu, Livny et al. 2009) as well as several pipelines being developed by his group at the Broad. This analysis will be conducted in close coordination with The Broad GSCID so that its computational infrastructures and expertise can also be leveraged in extracting biologically relevant information from these datasets.

Illumina sequencing reads will be aligned to the N16961 reference genome sequence using Maq. Scripts developed by Jonathan Livny along with the IGV and Artemis browsers will be used to visualize and annotate the location of these alignments relative to predicted and confirmed ORFs, regulatory RNA-encoding genes, and transcription signals such as promoters and transcription terminators. This will enable us to visually screen our datasets for transcripts of interest as well as mine them for particular types of transcripts such as those encoded in intergenic regions of the genome or antisense to protein-encoding sequences. Other scripts will be used to systematically compare datasets derived from different cDNA libraries. This will facilitate more reliable mapping of transcription start sites by allowing us to identify transcripts that have consistent 5' ends across independent samples and enable us to efficiently search our datasets for previously annotated and novel genes that are differentially expressed in culture vs. the rabbit model.

The transcript abundance profiles derived from each of the libraries will lead to new insights into *V. cholerae* biology and will provide many opportunities for future research. These insights/opportunities include:

1. Global definitions of transcription start sites.
2. Insights into global RNA processing.
3. Validations of existing gene annotations.
4. Insights into operon structures.
5. A variety of opportunities for gene discovery including noncoding RNAs, small ORFs, antisense transcripts and study of regulatory RNAs.
6. New insights into global changes in transcription induced during infection.
7. A variety of opportunities for development of infrastructures for analyses of deep sequencing transcriptome datasets.

5. Community Support and Collaborator Roles:

In the U.S. alone there are at least 50 NIH funded labs that study *V. cholerae* biology. The data and associated analyses will serve as a valuable resource for the whole community. Furthermore, to garner community input, we contacted and received enthusiastic support and endorsement from the following researchers that study *V. cholerae biology:*

1. John Mekalanos (Harvard Medical School)
2. Paula Watnick (Children's Hospital, Boston)
3. Victor DiRita (University of Michigan)
4. Deb Hung (Broad and MGH).

The protocols and tools which we will develop for this project will be widely applicable. All the data that is generated from this project will be made available to the community as per the NIAID data release policies. The raw primary sequence data as well as the associated quality scores will be made available in accordance to the Broad protocol.

We anticipate that the data produced in this study will contain a wealth of biological information beyond the scope of our analyses that will be of great interest and value to the many researchers studying the biology and pathogenesis of *V. cholerae*. However, analyses of these datasets require bioninformatic capabilities that exceed those possessed by most research groups. Therefore, to maximize the accessibility of the cDNA HTS data generated in this project to the broad scientific community, we will make these data available in the following formats:

a. Standard BAM and SAM files of the Illumina reads aligned to the N16961 reference genome. These files will include the sequences and quality score of each read.

b. Sortable tab-delimited Excel files of all putative transcriptional start sites, including their 5' coordinate, strand, normalized abundance, and distance from and name of the 3' gene.

c. Sortable tab-delimited Excel files of all putative transcriptional units, including their 5' and 3' coordinates, strand information, normalized abundance, and annotation of position relative to annotated genes.

d. Sortable tab-delimited Excel files of all annotated genes and their corresponding RPKM (Reads per Kb per Million reads) values. This will facilitate identification of genes differentially expressed under *in vivo* and *in vitro* conditions.

e. Text files that can be directly loaded into the IGV (Broad) and Artemis (Sanger) genome viewers (both freely available and requiring no specialized bioinformatic expertise) that will enable a histogram of read counts at every position on either strands of the genome to be superimposed on the sequence and gene annotations of any reference genome.

Project collaborators:

a. Anjali Mandlik (HHMI/Brigham and Women's Hospital), will generate the libraries for the project as well as participate in analyses and validation of the data.
b. Jonathan Livny (The Broad Institute), will help with the computational analyses of the large datasets generated in this project.
c. Jorg Vogel (Max Planck Institute for Infection Biology, Berlin, Germany), a pioneer in bacterial transcriptomics will provide advice for the project.
d. The Broad Institute GSCID, will carry out the sequencing on the Illumina Genome Analyzer and assist in the computational analyses as required.

Funding sources:

The Waldor lab is funded by NIAID R37-42347 and HHMI. Jonathan Livny is funded by NIAID R00-076608.

6. Availability & Information of Strains:

Vibrio cholerae C6706 is part of the Waldor lab strain collection. We will prepare customized libraries designed for RNA-seq by the Illumina sequencer. The RNA will be extracted from cultures grown in lab media or bacteria isolated from the rabbit cecal fluid after infection. We intend to submit the libraries to the GSC for subsequent sequencing.
Information of the strain:

Name	*Vibrio cholerae O1 biovar El Tor str. C6706*
Identifier	Not applicable
Material type	Not applicable
Genus	Vibrio
Species	*Vibrio cholerae*
Biotype	*El Tor*
Isolated from	Peru 1991
Select agent status	Level 2
International permit requirement	Not applicable
BEIR accession number	Will be deposited.
ATCC accession number	55331

We plan to determine the *V. cholerae* transcriptome under variety of growth conditions. These conditions will be linked to the datasets that will be made publicly available.

7. Compliance Requirements:

7a. Review NIAID's Reagent, Data & Software Release Policy:

NIAID supports rapid data and reagent release to the scientific community for all sequencing and genotyping projects funded by NIAID GSC. It is expected that projects will adhere to the data and reagent release policy described in the following web sites.

http://www3.niaid.nih.gov/research/resources/mscs/data.htm

http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html

<Each Center to include their website that describes/points to the guidelines>

Once a white paper project is approved, NIAID GSC will develop with the collaborators a detailed data and reagent release plan to be reviewed and approved by NIAID.

Accept ☐√ Decline ☐

7b. Public Access to Reagents, Data, Software and Other Materials:

The strain of interest will be deposited in the BEI repository.

All primary sequences and associated quality scores will be made available in compliance with the NIAID data release policy in the formats as described above.

7c. Research Compliance Requirements

Upon project approval, NIAID review of relevant IRB/IACUC documentation is required prior to commencement of work. Please contact the GSC Principal Investigator(s) to ensure necessary documentation are filed for / made available for timely start of the project.

Investigator Signature:

Investigator Name: Dr. Matthew Waldor **Date: November 16, 2009**

References:

Beyhan, S., A. D. Tischler, et al. (2006). "Differences in gene expression between the classical and El Tor biotypes of Vibrio cholerae O1." Infect Immun **74**(6): 3633-42.

Brantl, S. (2007). "Regulatory mechanisms employed by cis-encoded antisense RNAs." Curr Opin Microbiol **10**(2): 102-9.

Cameron, D. E., J. M. Urbach, et al. (2008). "A defined transposon mutant library and its use in identifying motility genes in Vibrio cholerae." Proc Natl Acad Sci U S A **105**(25): 8736-41.

Camilli, A., D. T. Beattie, et al. (1994). "Use of genetic recombination as a reporter of gene expression." Proc Natl Acad Sci U S A **91**(7): 2634-8.

Dziejman, M., E. Balon, et al. (2002). "Comparative genomic analysis of Vibrio cholerae: genes that correlate with cholera endemic and pandemic disease." Proc Natl Acad Sci U S A **99**(3): 1556-61.

Faruque, S. M., M. J. Albert, et al. (1998). "Epidemiology, genetics, and ecology of toxigenic Vibrio cholerae." Microbiol Mol Biol Rev **62**(4): 1301-14.

Heidelberg, J. F., J. A. Eisen, et al. (2000). "DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae." Nature **406**(6795): 477-83.

Liu, J. M., J. Livny, et al. (2009). "Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing." Nucleic Acids Res **37**(6): e46.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet **9**: 387-402.

Mekalanos, J. J., E. J. Rubin, et al. (1997). "Cholera: molecular basis for emergence and pathogenesis." FEMS Immunol Med Microbiol **18**(4): 241-8.

Morozova, O., M. Hirst, et al. (2009). "Applications of new sequencing technologies for transcriptome analysis." Annu Rev Genomics Hum Genet **10**: 135-51.

Parkhomchuk, D., T. Borodina, et al. (2009). "Transcriptome analysis by strand-specific sequencing of complementary DNA." Nucleic Acids Res **37**(18): e123.

Shi, J., P. R. Romero, et al. (2006). "Evidence supporting predicted metabolic pathways for Vibrio cholerae: gene expression data and clinical tests." Nucleic Acids Res **34**(8): 2438-44.

van Vliet, A. H. (2009). "Next generation sequencing of microbial transcriptomes: challenges and opportunities." FEMS Microbiol Lett.

Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Wilhelm, B. T. and J. R. Landry (2009). "RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing." Methods **48**(3): 249-57.

Xu, Q., M. Dziejman, et al. (2003). "Determination of the transcriptome of Vibrio cholerae during intraintestinal growth and midexponential phase in vitro." Proc Natl Acad Sci U S A **100**(3): 1286-91.

Zhu, J., M. B. Miller, et al. (2002). "Quorum-sensing regulators control virulence gene expression in Vibrio cholerae." Proc Natl Acad Sci U S A **99**(5): 3129-34.