# TRANSCRIPTOME ANALYSIS OF MARINE MAMMAL STRESS RESPONSES AND TISSUE ENERGY USAGE (Charge # TG-IBN150010)

# PROGRESS REPORT

**PI: Dr. Jane I. Khudyakov**
University of the Pacific
3601 Pacific Avenue
Stockton, CA 95211
Phone: (626) 826-8040
Email: jkhudyakov@gmail.com

## ABSTRACT

Physiological stress, such as disturbance caused by anthropogenic activity, is a threat to health, reproduction, and survival of wild animals, especially taxa of concern such as marine mammals. However, there is little consensus on how to identify stressed individuals or discriminate between acute and chronic stress states in wildlife. Non-targeted omics technologies such as transcriptomics (RNAseq) offer a powerful approach for rapidly increasing our understanding of downstream consequences of stress in study systems with few other molecular resources (e.g. sequenced genomes, microarrays). I have developed an experimental stress manipulation and RNAseq data analysis pipeline in a tractable marine mammal system, the northern elephant seal. Using amazon elastic cloud computing and XSEDE startup resources, I previously identified tissue-specific molecular markers of acute stress in elephant seal muscle and blubber (Khudyakov et al. 2015, Khudyakov et al. in preparation). I am currently using RNAseq to compare transcriptome profiles of elephant seals exposed to acute and repeated stress manipulations to identify markers of chronic stress, a project that is currently funded by the Office of Naval Research (Award No. N00014-15-1-2773). I received an XSEDE startup allocation of 50,000 service units on Stampede and 500 gigabytes of Ranch storage in October 2015 to complete preliminary transcriptome analysis. However, I was unable to fully utilize the allocated XSEDE resources due to the seasonal nature of my research (biological samples and sequencing data will not be obtained until fall 2016 and spring 2017, respectively) and transfer to another institution during the allocation period. My new institution, University of the Pacific, lacks HPC resources that can support my work. I am requesting a 12-month renewal of my startup allocation to complete chronic stress transcriptome analysis and train graduate students in my new laboratory in bioinformatics. I am also requesting a supplement on PSC-Bridges (1000 SUs) for transcriptome assembly and ECSS support for code optimization and data visualization.

## SCIENTIFIC DISCOVERIES

The XSEDE startup allocation on Stampede enabled annotation of a transcriptome assembly generated from blubber collected from elephant seals during an acute stress challenge. Using the Trinotate[1] pipeline and ultrafast alignment algorithm DIAMOND[2], I identified 140,672 vertebrate homologs of elephant seal transcripts and 147,172 homologs of elephant seal peptides predicted from assembled transcripts. This is an

improvement over my previous annotation using BLAST search of the mouse genome, which identified 105,178 elephant seal homologs of mouse genes. Functional annotation of the elephant seal blubber transcriptome identified a large number of genes involved in oxidative stress signaling and antioxidant defenses, which were discussed in a recent publication (Crocker, Khudyakov and Champagne, 2016).

I used Stampede to identify genes differentially expressed in response to an acute stress challenge in elephant seal blubber with kallisto[3], a transcript abundance estimation algorithm and DESeq2[4], a differential expression statistical package. I found 426 genes that were differentially expressed during the acute response to stress and 106 genes differentially expressed during recovery from stress. These included key regulators of lipid and energy homeostasis, some of which have not been studied extensively in mammalian adipose tissue, and which are being described in a manuscript that will be submitted to Scientific Reports (Khudyakov et al. in preparation). Using this information, I developed seal-specific qPCR markers for profiling metabolic adjustments used by elephant seals to regulate energy usage during prolonged fasting periods.

A significant challenge in the field of transcriptomics is the ability to reliably assess quality of de novo transcriptome assemblies. I used the most up-to-date tools and recommendations to evaluate the quality of the acute stress blubber transcriptome. I found that my assembly accurately represented the sequenced reads (87% of reads mapped correctly) and contained orthologs of 80% of conserved vertebrate genes (Benchmarking Universal Single-Copy Orthologs, or BUSCOs[5]), a measure of completeness. I also used TransRate[6], a new package that calculates quality scores based on detailed contig-by-contig metrics. I obtained a TransRate assembly quality score of 0.42, which is higher than >50% of the transcriptomes deposited in the NCBI Transcriptome Shotgun Assembly database as reported by authors of the software[6].

The XSEDE startup allocation provided me with the ability to evaluate newly developed bioinformatics tools in an HPC environment and to update my RNAseq data analysis pipeline[7] in preparation for a project examining elephant seal tissue transcriptome responses to chronic stress starting in fall 2016. I will use my XSEDE experience to train a University of the Pacific MS student, Jared Deyarmin, in bioinformatics analyses for the chronic stress project. Based on the work I conducted on Stampede, I discussed recommendations for RNAseq data analysis approaches and software during an invited presentation at a marine mammal genomics workshop (Society of Marine Mammalogy Annual Meeting, San Francisco, Dec. 2015), which were included in a manuscript submitted to Journal of Heredity (Cammen et al., in review).

## COMPUTATIONAL ACCOMPLISHMENTS

The startup allocation enabled me to utilize recently developed software and computational pipelines to improve transcriptome annotation and differential expression analysis. Instead of the traditional, computationally intensive sequence alignment algorithm BLAST, I utilized DIAMOND[2], an ultrafast sequence aligner, to search protein and nucleotide similarity between the elephant seal transcriptome and public sequence databases. I used the Trinotate[1] pipeline to identify elephant seal transcripts with homology to conserved vertebrate protein domains, transmembrane regions, and signaling peptides to produce a more complete transcriptome annotation than was generated in previous projects[7]. Assembly quality was evaluated using BUSCO[5] and

TransRate[6], generating relevant metrics that I had not used previously[7]. I used Stampede to evaluate performance of several novel k-mer-based transcript abundance estimation approaches (eXpress[8], kallisto[3]) and differential expression analysis tools (edgeR[9], DESeq[10], DESeq2[4]). Kallisto, a new ultrafast quasi-alignment transcript abundance estimation algorithm[3], was a vast improvement over the computationally intensive read mapping-based RSEM[11], which I had used for another transcriptome project[7]. Differential expression tools tested identified overlapping sets of genes; however, DESeq2 reported the least number of low-expression transcripts and is well supported by literature[4].

Due to the lightweight nature of RNAseq tools used (DIAMOND, kallisto), and the fact that I have not yet conducted de novo transcriptome assembly on Stampede, I have only utilized 3283.0 of the 50,000 SUs requested to date. This work enabled optimization of bioinformatics tools that will be applied to another project (chronic stress transcriptomes) starting in fall of 2016, which requires renewal of my startup allocation.

## PUBLICATIONS

The following were generated using XSEDE resources:

Khudyakov JI, Champagne CD, Meneghetti L, Crocker DE. Blubber transcriptome response to acute corticosteroid elevation in a fasting-adapted phocid, the northern elephant seal. (in preparation)

Cammen KM, Andrews KR, Carroll EL, Foote AD, Humble E, Khudyakov JI, Louis M, McGowen MR, Olsen MT, Van Cise AM. Genomic methods take the plunge: recent advances in next generation sequencing of marine mammals. Journal of Heredity (in review).

Crocker DE, Khudyakov JI, Champagne CD. 2016. Oxidative stress in northern elephant seals: integration of omics approaches with ecological and experimental studies. Comparative Biochemistry and Physiology - Part A: Molecular and Integrative Physiology S1095-6433 (16)30037-X.

## REFERENCES


1. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8(8):1494-512.
2. Buchfink B, Xie C, Huson DH (2015). Fast and sensitive protein alignment using DIAMOND. Nature Methods 12:59-60.
3. Bray NL, Pimentel H, Melsted P, Pachter L (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology 34:525-527.
4. Love MI, Huber W, Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12):550.
5. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 10.1093/bioinformatics/btv351.
6. Smith-Unna RD, Boursnell C, Patro R, Hibberd JM, Kelly S (2015). TransRate: reference free quality assessment of *de-novo* transcriptome assemblies. bioRxiv 021626.
7. Khudyakov JI, Preeyanon L, Champagne CD, Ortiz RM, Crocker DE (2015). Transcriptome analysis of northern elephant seal (Mirounga angustirostris) muscle

tissue provides a novel molecular resource and physiological insights. BMC Genomics 16:64.
8. Roberts A, Pachter L (2012). Streaming fragment assignment for real-time analysis of sequencing experiments. Nature Methods 10:71-73.
9. Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26(1):139-140.
10. Anders S, Huber W. (2010). Differential expression analysis for sequence count data. Genome Biology 11:R106.
11. Li B, Dewey CN. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12: 323.