**Proposal for a transcriptomic analysis of bacteriome-specific expression profiles by sex in the tsetse fly *Glossina morsitans***

Copeland HC, Lynch BG, Pollio AR, Spencer NJ

***1. Background***

Tsetse flies (Diptera:glossinidae) have been the subject of decades of research, owing in part to several distinguishing aspects of their physiology including obligate blood-feeding in both males and females as well as a viviparous reproductive strategy in which females harbor the developing larva *in utero* (*1*, *2*). In particular, the blood-feeding habit of tsetse flies is of public health and agronomic concern for its role in the transmission and life cycle of the protozoan parasite *Trypanosoma bruceii*, the causative agent of African sleeping sickness (*1*).

Another biological player in these interactions is a bacterium called *Wigglesworthia glossinidia* (hereafter *Wigglesworthia*)*,* which provides essential supplementation for its host’s limited diet (*3*). The symbiotic functions of this organism have been demonstrated to affect trypanosome transmission by the host fly (*4*). One consequence of the ancient association between tsetse and *Wigglesworthia* is the strictly vertical mode of symbiont transmission in which parental *Wigglesworthia* are passed on to developing larvae during host reproduction (*5*). This involves the translocation of these bacteria from specialized gut cells to the secretory cells of the milk glands (*5*, *6*). The exact means by which this translocation occurs and the mechanisms by which the host facilitates this remain poorly understood.

This process demonstrates a unique way that female tsetse flies interact with their *Wigglesworthia* populations, but different roles in symbiosis between males and females have been largely unexplored at other life stages. Here, we propose an investigation of sex-based differences in symbiosis maintenance in the tsetse fly *Glossina morsitans* by identifying differential expression in bacteriomes (specialized, *Wigglesworthia-*housing midgut tissue) from male and female *G. morsitans* at the teneral stage (between emergence from pupa and the first blood meal).

Unpublished RNA-seq data from these bacteriomes will be used for exon-aware mapping to a *G. morsitans* reference genome (*7*), transcript assembly and quantification, and differential expression analysis with HISAT2, StringTie, and Ballgown, respectively. While this pipeline is well-established for such analyses, our application of these tools to this unpublished dataset represents an innovation in the application of molecular biology tools to this organism’s physiology. Given the similarity between the life habits of male and female *G. morsitans* adults prior to mating, differentially expressed exons in teneral bacteriomes could provide insights into either novel sets of interactions between the symbiont and sex-specific physiological processes or early-life cycle signatures of host facilitation of symbiont transmission.

***2. Methods***

*2.1 Data Retrieval and Quality Control*

The RNA-seq data for this experiment represent bacteriomes from only 6 individuals. While a larger number of individuals were originally samples, not all RNA libraries produced usable sequences. We acknowledge that because our analysis is based on host expression in bacteriome extractions and not the microbial populations this sample size may impact the statistical power of our data in identifying definitive patterns of expression. However, it should be noted that there are two technical replicates for each sample, and gut extractions and extractions from different tsetse species as part of the same set of experiments can help corroborate these initial findings through further analyses.

Because the data is private and unpublished, it will be accessed from the Illumina sequence data distribution platform BaseSpace™ Sequence Hub (BSSH) using BaseMount, which will allow us to work with the data through a Unix-based command line interface. This is important because our analyses will make use of the computing power of WVU’s Spruce Knob High Performance Cluster.

Once the data — with each sample consisting of roughly a gigabyte of paired-end, 51bp Illumina reads in a FastQ file format — are accessed from BSSH with BaseMount, they will be subject to quality control analysis using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FastQC will output per-base sequence quality scores, indicating any downward trends or other issues in the sequencer’s confidence in correct base calls over the length of the reads. It will also identify any adapter sequences print general summary statistics. It is worth noting that, because this is RNA-seq data, FastQC will likely identify overrepresented sequences because transcripts that appear in higher copy numbers will appear more times in the sequence data. As a result, any overrepresented sequences not actively identified as adapters will be ignored for the time being.

Our chosen pipeline for expression analysis may be sensitive to low quality sequences. However, since the reads are so short, we do not anticipate any significant decrease in quality at read ends. However, if a small number of positions at the ends of reads show a substantial decrease in Phred quality scores relative to the rest of the read, it may be necessary to crop several base pairs from the read ends using Trimmomatic v0.39 if such processing would not come at too great a cost to the amount of usable sequence data (*8*). Should such processing be employed, the Trimmomatic outputs will be checked with FastQC again to ensure that any quality issues have been remedied. If this is not the case, further processing (and subsequent follow-up QC checking) may be required.

*2.2 HISAT2*

HISAT2 is an RNA seq analysis tool that maps reads against a reference genome to identify genomic positions (*9*). HISAT2 uses global, whole genome-index and tens of thousands of small local indices. In results HISAT2 generating spliced alignments two times faster than Bowtie and BWA. With the use of HISAT2, the exon-aware alignments that are present within the data will be detected from the *G. morsitans* gene and include higher accuracy of alignments. The default minimum and maximum intron lengths recognized by HISAT2 (20 and 500000, respectively) should be appropriate given the average intron length of just under 25000 in *G. morsitans* (*7*).

The output alignment map (a BAM file) can be assessed for various mapping metrics using SamTools flagstat (*10*). Given the localization of *Wigglesworthia* to the tissue sampled, it is anticipated that there will be a considerable amount of unmapped sequence data attributable to *Wigglesworthia* RNA. Reads mapped to multiple locations, however, may warrant a return to the quality control steps or a further investigation of *G. morsitans* genome features that could be implicated in the mapping issues given that *G. morsitans* is not known to be polyploid.

*2.3 StringTie*

In order to accurately quantify expression, the exon-aware sequence alignments output by HISAT2 must be assembled into transcripts, each of which will represent one isoform of a *G. morsitans* gene product. During transcript assembly, the *MERGE* function for StringTie will be used to generate unified, non-redundant set of transcripts present across the samples, which will be important to any analysis of differential expression (*11*). This will require the reference annotation from *G. morsitans* as a GFF or GTF file. Failure to execute this function may result in a transcript assembly that reduces the usable data. Once the transcripts have been merged, StringTie will output transcript abundances.  
 After re-estimation reads might have been re-allocated for transcripts and altered providing additional recount data for each transcript that will be needed for downstream analysis with Ballgown.

*2.4 Ballgown*

Ballgown is a multiuse R package designed to simplify RNA-seq post-mapping analysis by utilizing a unique linear model reducing the computational burden (*12*). With the use of StringTie as the previous step, our input file includes transcript estimates allowing the R object to be built directly within Ballgown. Further, once built, the R object can be manipulated downstream to elucidate specific trends seen comparing our two experimental groups (male and female). Within this project we propose the use of fragments per kilobase of transcript per million (FPKM) as our gene level measurement to help guide further analysis.

We plan to assess differences between male and females by utilizing a tandem function within bioconductor GenomicRanges (*13*). This tool will allow us to further organize our R object with additional information, for example tagging overlapping transcripts. In order to present any findings within the transcription data we will build gene level transcript maps. As an additional feature of ballgown, these maps allow for a clear image that can simplify any findings.

We feel confident that the use of this well described pipeline will provide a significant quantity of resources that will help with any troubleshooting necessary. Additionally, our team has extensive training working with R packages which will add additional troubleshooting tools.

***3. Expected Outcomes***

We expect to have individual gene maps that define genes of interest with specific variation. These maps will help elucidate any differential expression observed between the two experimental groups. We expect transcriptome-wide analyses like Principal Components Analysis (PCA) to show segregation between the samples by sex since fly species, age, and other conditions were otherwise constant. While our dataset presents some limitations in broadly characterizing sex-based physiological differences in bacteriome transcription profiles, signatures of differential expression in genes with potential functions in mediating interactions with gut microbiota like *Mesh* and *Duox* or in pathways related to immunity and reproduction provide a starting point to address our questions about the insect’s development and interactions with its endogenous bacteria. Parallel transcriptomic analysis of *Wigglesworthia* expression and verification of key differentially-expressed genes by RT-qPCR at various tsetse life stages can further elucidate the intricacies of these relationships and the related processes.

***4. Investigator Roles***

Noah Spencer, who has access to the dataset of interest, will be responsible for data acquisition, quality control, and any necessary initial processing of the raw sequencing data. Brianna Lynch will be responsible for mapping the reads to the reference genome using HISAT2. Haylee Copeland will be responsible for transcript assembly using StringTie. Adam Pollio will be responsible for the transcriptome analysis in Ballgown.

Because Noah Spencer’s task falls at the beginning of the pipeline and is a straightforward process with which all involved researchers are familiar, he will also provide troubleshooting support to the group throughout the process based on his prior experience with this dataset and some of the analyses involved.

Each section of the proposal’s methods was authored by the researcher assigned to the corresponding task, demonstrating the familiarity of each team member with the necessary protocols for successful completion of the project goals.

While the ability of individual investigators to complete their assigned duties independently will be useful in limiting conflict, the complexity of these analyses will inevitably require troubleshooting and decision-making that will involve all group members. If the fluency of the investigator heading each portion of the pipeline with the corresponding technique is insufficient to establish a clear direction and a conflict regarding next steps arises, we will waste no time in seeking outside assistance from researchers with the appropriate expertise to inform the final decisions.

While it is important that all investigators be transparent about findings and issues at every step, the final report on the project findings will be authored in a similar fashion as this proposal with each investigator contributing the findings and relevant methodological adjustments from the corresponding portion of the pipeline.

***References***

1. A. Geiger, I. Malele, A. M. Abd-Alla, F. Njiokou, Blood feeding tsetse flies as hosts and vectors of mammals-pre-adapted African Trypanosoma: current and expected research directions. *BMC Microbiology*. **18**, 162 (2018).

2. J. B. Benoit, G. M. Attardo, A. A. Baumann, V. Michalkova, S. Aksoy, Adenotrophic Viviparity in Tsetse Flies: Potential for Population Control and as an Insect Model for Lactation. *Annu Rev Entomol*. **60**, 351–371 (2015).

3. X. Bing, G. M. Attardo, A. Vigneron, E. Aksoy, F. Scolari, A. Malacrida, B. L. Weiss, S. Aksoy, Unravelling the relationship between the tsetse fly and its obligate symbiont *Wigglesworthia* : transcriptomic and metabolomic landscapes reveal highly integrated physiological networks. *Proc. R. Soc. B*. **284**, 20170360 (2017).

4. R. V. M. Rio, A. K. S. Jozwick, A. F. Savage, A. Sabet, A. Vigneron, Y. Wu, S. Aksoy, B. L. Weiss, Mutualist-Provisioned Resources Impact Vector Competency. *mBio*. **10**, e00018-19 (2019).

5. R. V. M. Rio, R. E. Symula, J. Wang, C. Lohs, Y. Wu, A. K. Snyder, R. D. Bjornson, K. Oshima, B. S. Biehl, N. T. Perna, M. Hattori, S. Aksoy, Insight into the Transmission Biology and Species-Specific Functional Capabilities of Tsetse (Diptera: Glossinidae) Obligate Symbiont Wigglesworthia. *mBio*. **3**, e00240-11 (2012).

6. S. Balmand, C. Lohs, S. Aksoy, A. Heddi, Tissue distribution and transmission routes for the tsetse fly endosymbionts. *J. Invertebr. Pathol.* **112 Suppl**, S116-122 (2013).

7. International Glossina Genome Initiative, Genome sequence of the tsetse fly (Glossina morsitans): vector of African trypanosomiasis. *Science*. **344**, 380–386 (2014).

8. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114 (2014).

9. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. **37**, 907–915 (2019).

10. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

11. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. **33**, 290–295 (2015).

12. A. C. Frazee, G. Pertea, A. E. Jaffe, B. Langmead, S. L. Salzberg, J. T. Leek, Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat Biotechnol*. **33**, 243–246 (2015).

13. M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, V. J. Carey, Software for computing and annotating genomic ranges. *PLoS Comput Biol*. **9**, e1003118–e1003118 (2013).