Review for PeerJ

"Brain transcriptomes of harbor seals demonstrate gene expression patterns of animals undergoing a metabolic disease and a viral infection" by Rosales and Thurber presents a valuable gene sequence resource for a commonly studied phocid species from a tissue that is not easily available to marine mammal researchers. The results show interesting changes in gene expression consistent with tissue responses to viral infection in PhV-1 infected seals and changes in metabolism in seals that died of unknown causes. I thought that overall this was a very interesting study that used robust methods for transcriptome analysis, although I have some concerns about the assembly quality (see below).

Major comments:

1. Data sharing

The raw sequenced reads should also be submitted to NCBI's SRA. The transcriptome assembly should be uploaded to a public site such as TSA or made available through a permanent host (figshare, for instance). The complete lists of DEGs (list for PhV-1-upregulated, list for UCD-upregulated, for instance) with annotation (not just GO categories) should be made available on figshare or added as supplementary files.

2. Assembly quality

My biggest concern is the very low rate of alignment of sequenced reads to both the transcriptome assembly and the Weddell seal reference genome. This suggests several possibilities: 1) there is a glitch somewhere in the alignment protocol, producing artificially low alignment rates, 2) the quality of the transcriptome assembly is poor, or 3) the quality of the sequenced reads is poor. Actually, regarding #3 - I wonder if the trimming may have been too aggressive, leading to loss of information (see http://journal.frontiersin.org/article/10.3389/fgene.2014.00013/full). However, the fact that 87.5% of the assembled transcripts have hits to known proteins in UniProt is somewhat reassuring. I think the authors need to explore this further and double-check their alignment parameters (perhaps the parameters are too strict) as well as conduct an additional QC test on the assembly. If the alignment rates are still <50%, I would recommend using an additional method to re-align the unmapped reads and add them to the assembly so that valuable sequence information is not lost. For example, see http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-16-S5-S8

For current recommendations on evaluating assemblies, see http://biorxiv.org/content/biorxiv/early/2016/02/18/035642.full.pdf and https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Another point for future reference: the current draft of the Weddell seal genome assembly is still pretty messy, so I would have recommended doing de novo assembly only and skipping the alignment to the genome step - Trinity's de novo assemblies are generally fairly high in quality (for transcriptome referenced in citation 10, the alignment rates were >85%: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1253-6).

Minor comments:
I have some minor comments that mainly aim to improve clarity and grammar and include additional information about bioinformatics methods. In addition, I found that the enrichment in expression of genes involved in fatty acid metabolism in the UCD animals was very interesting, especially since many stranded harbor seal pups also experience chronic malnutrition at time of death. I think the authors can expand this idea some more in the discussion (Line 286), whereas discussion of the *Burkholderia* correlation could be condensed. The latter seemed somewhat speculative to me, as the correlation was weak and less biologically interesting, especially since the authors did detect high expression of genes involved in response to bacterial infection in the samples from UCD animals. In addition, I think correlation (r) is more applicable that regression analysis (R) in this case (see below).

Line 52: Change "illnesses" to "illness".

Line 58: Change "evolvement in marine mammal transcriptomics" to something like "increase in application of transcriptomics to marine mammal studies".

Lines 61-62: I think the use of the terms "biotic" and "abiotic" to describe gene expression changes in mammals is inaccurate (I have seen these most commonly used in plant literature). I may be mistaken, but I thought "abiotic" refers to non-biological factors such temperature, season, length of day, etc. I would refer to the study referenced in citation 10 as "used transcriptomics to measure gene expression during a physiological challenge" and the end of the sentence to something like "…there have been no studies that examined gene expression during tissue responses to infection by pathogens".

Lines 110-111: How soon after death were the tissues sampled? If sampling did not occur immediately, for how long and at what temperature were the carcasses frozen? Were they frozen immediately after death? How were the brain tissue samples stored (at -80C, in RNAlater, etc)? Please include these details.

Lines 137-140: I would include a sentence to mention that transcriptome assembly was conducted using a combination of genome-guided and de novo methods. Change "…reads in both libraries were combined and aligned against the hypothetical transcript…" to "reads from both libraries were combined and aligned to the Weddell seal reference genome assembly…"

Lines 143-145: Information on the software used for transcript abundance estimation (RSEM, eXpress, kallisto, or Salmon) need to be included. I assume the authors used RSEM or eXpress since they aligned the reads to the reference first (kallisto and Salmon do not require alignment). Software versions that were used need to be included for bowtie2 and RSEM/eXpress. Were the transcript counts normalized, and if so - how (by FPKM or TPM)?

Line 151: Which versions of R and Bioconductor were used for DESeq2?

Line 155-156: Minor grammatical point: The last sentence is an incomplete fragment. Combine with the previous sentence. I suggest not starting sentences with "while". This also applies to lines 336 and 373.

Line 158: What do you mean by "composite transcriptome"? I suggest simply calling it the "transcriptome assembly".

Line 160: The UniProt database version that was used needs to be included as public repositories are constantly updated. Usually the version is the date the database was downloaded.

Line 164: Include version of ErmineJ.

Line 180-185: 1) I would suggest re-running the statistical analysis using Pearson's or Spearman correlation (depending on whether data meets the assumptions of the test) rather than linear regression. See: http://www.graphpad.com/support/faqid/1141/ and http://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression.
2) Change "fatty acid normalized transcripts" to "normalized counts of fatty acid metabolism genes" (it is the *abundance* that was normalized, not the transcripts). How were they normalized? How were the *Burkholderia* transcript abundance values normalized? (by TPM or FPKM?)

Line 188: I suggest changing "…expressed during a PhV-1 infection and compare them to harbor seals…" to "expressed in the brain during PhV-1 infection from those expressed in brains of harbor seals..."

Line 196-199: I suggest changing the first 2 sentences to something like "The sequenced libraries were used to build an assembly using a combination of genome-guided and de novo methods. The *Leptonychotes weddellii* genome assembly (accession #) was used as the reference…" Not sure what is meant by "remove any potential DNA sequences in our data". Do the authors mean non-coding DNA? Intronic? How were these sequences removed? I think this requires more explanation and additional detail in methods.

Lines 200-201: Is the 27.43% the percent of reads aligning to the Weddell genome? Are the 17.1-28.05% the percentages of reads aligning to the harbor seal transcriptome assembly? This was not clear. Also, see major comment regarding alignment rates.

Line 206: I suggest changing "…transcripts had similarities to the UniProt database…" to "transcripts had hits to proteins in the UniProt database".

Line 210: I suggest changing the heading to "Functional annotation of differentially expressed genes distinguishes UCD from PhV-1-infected harbor seals". Double-check the rest of the manuscript for consistency in referring to the PhV sample group as "PhV-1-infected" rather than "comparative".

Lines 212-214: 1) What do the authors mean by "after filtering" in line 213? Filtering out duplicated GO terms? Or filtering out terms that were not enriched at $p<0.05$?
2) For the enrichment analysis, what were the genes enriched *relative to*? Was it enrichment of GO categories in DEG relative to the entire transcriptome assembly? Or enrichment of GO terms in the harbor seal transcriptome relative to another genome, like human or mouse? I would also change "enriched in ErmineJ" to "enriched in the differentially expressed gene set" or "enriched in the harbor seal transcriptome relative to the … genome", depending on what the enrichment is relative to. The rest of this section reads great and the results are quite compelling.

Lines 230-240: Again, make sure to specify what the enrichment of KO terms was relative to. Were the five KEGG pathways most abundant in all of the DEG or all of the transcriptome assembly? Were the KO terms were upregulated in PhV-1 samples relative to those in UCD samples or the entire transcriptome?

Line 249: I suggest changing "correlation of these metabolism genes with *Burkholderia* RNA" to "correlation of these metabolism genes with abundance of *Burkholderia*-specific transcripts".

Lines 253-255: I think it is slightly inaccurate to say that transcriptomics can be used to identify "culprits" or "mechanisms" of disease. Transcriptomics may be able to detect infectious agents or characterize cellular/tissue gene expression profiles during disease or in response to infection, but I think to truly understand the *mechanism* of the disease you would need time-course sampling of multiple tissues. I would also change "between two disease states" in line 255 to "known and unknown disease states".

Line 259: Change "UCDs" to "UCD seals".

Lines 267-278: See comment in first paragraph of "Minor comments". The correlation is statistically significant, but quite weak - I would mention this here.

Lines 279-285: This paragraph did not make sense to me.

Line 362: I suggest changing "controversy to" to "controversy about".

Line 389: I suggest changing "more pronounced" to "more numerous".

Line 391: Change "said to be up-regulated" to "identified as upregulated"

Line 393: Change "the disease infecting agent" to "infectious agent".

Line 397: What do the authors mean by "mitigate these caveats"? I would think that additional RNA-seq studies examining gene expression over the time-course of disease and in additional tissues would be more informative.

Line 406: I recommend changing “interactions and an unknown disease” to “interactions and brain tissue response to an unknown disease”. I think the authors need to make sure to specify that the response to disease they are observing in this study is specific to brain tissue, and not representative of the entire organismal response. Double-check that this is specified throughout the manuscript.

Line 408: I think the statement “We now have a better understanding of PhV-1” is too great of a logical leap. One can say that we have a better understanding of gene expression in brain tissue of animals undergoing a viral infection, but more work (including functional labwork, not just more sequencing) is needed to understand the *mechanism* and *progression* of the disease, or even virus-host *interaction*. The genes and GO categories enriched in PhV-1 infected samples are not novel, nor do they seem unique to this particular virus (unless I am mistaken, but the GO categories seem fairly generalized to host response to viral infections). What I mean is that HTS is not the be-all and end-all that is going to solve all of our biological problems (although many papers claim this in broad overreaching conclusions). I would caution against making such broad concluding statements.

Comments on figure legends (figures themselves look great):
Figure 1: What does “the most variable 2500 transcripts” mean? Are these the DEGs? I would rephrase this, being more specific about which transcripts are referred to here. Is the PCA analysis done on normalized gene counts?

Figure 2: Which GO terms are these? The ones enriched in the DEGs, or in the whole transcriptome? Again, be more specific - what do you mean by “significant”?

Figure 3: Significantly enriched relative to what? What exactly is the heat map showing - what does the z-score and colors represent? Is it normalized gene counts, log2 fold-change in expression, significance? Again, this needs to be more specific.

Figure 4: You are referring to GO *categories*, not transcripts. Again, what are they enriched relative to? And what do the colors and z-score represent?

Figure 5: I would rephrase this legend to something like “KEGG pathways involved in (human?) herpes simplex infection. Highlighted boxes (grey) represent terms that were significantly enriched in DEGs in PhV-1-infected harbor seals…”

Table 1:
I would change the legend for this table to something like “Stranding information for samples used in the study”. Also, it may be interesting to add mass at time of death (and percent adiposity, if available) since it is likely that chronic malnutrition in UCD animals contributed to the expression of fatty acid metabolism genes seen in the brain.