Reply to Bemm et al. and Arakawa: Identifying foreign genes in independent *H. dujardini* genome assemblies.

Thomas C. Boothby and Bob Goldstein

Our report (1) has raised questions from other groups who were sequencing the *H. dujardini* genome in parallel or who have done new experiments and analyses since our report (2–5). Bemm et al. (2) now report filtering our data for likely contaminants, resulting in a new, pre-filtered genome assembly. Arakawa (3) has sequenced genomes of starved, washed, individual animals that had been treated with antibiotics for 48h, and he has removed contigs not covered by new RNA-Seq data that he provided from animals in multiple stages and conditions. Two other reports have contributed data and analysis: Delmont and Eren (4) used a valuable, newly published analysis and visualization platform, Anvi'o (6), to identify likely contaminants in our genome assembly, and Koutsovoulos et al. (5) applied useful Taxon-annotated GC-coverage plots (Blobplots) (7) to our data and reported an independent genome sequence and assembly.

We regret that an outdated version of our assembly was mistakenly uploaded to public databases. This assembly was 252Mb, larger than the 212.3Mb assembly that we had used in our report (1) as pointed out correctly by others (5) and that retained likely contaminants that we had identified and removed by methods detailed in the link below—contaminants that would have greatly elevated the proportion of foreign genes in our 212.3Mb assembly to well above the 17.5% that we reported. This mistake may have contributed in large part to discrepancies and issues raised by others. For example, the Untrusted assembly from (2) is 39Mb similar to the ~40Mb postfiltered from out 252Mb assembly. Similarly, (4) identified whole bacterial genomes in our prefiltered assembly, which were largely missing from our postfiltered assembly. We apologize to other groups for this mistake, and we appreciate that the concerns they have raised by and large remain as valid concerns for our 212.3Mb assembly, which is available at the link provided below.

Each group (2–5) pointed out some appropriate limitations with our analysis (1), and we agree that the new assemblies (2, 4, 5) are more extensively filtered than was our 212.3Mb assembly. Extensive filtering could lead to the exclusion of foreign genes from assemblies. For example, if HGT is occurring between species with substantially different GC content, then the recipient genome will have regions of GC content that differ from the rest of the genome. To *a priori* exclude reads or scaffolds on the basis of GC content is to bias against detecting HGT. In this light, these third-party assemblies provide an excellent opportunity to examine the result of other groups' filtering via k-mer selection (2) or via GC% and coverage (5) on the level of foreign genes found in different *H. dujardini* genome assemblies. Therefore, we assessed foreign genes in these assemblies. Since any metric used to assess foreignness of a gene will have its own particular strengths and weaknesses we used 3 different approaches: the HGT index (8), identifying genes that align to prokaryotic but not eukaryotic sequences, and identification of C class genes (9).

Using the HGT index we find elevated proportions of foreign genes in the new assemblies, ranging from 2-4X the proportion of HGT in *C. elegans* and ~6-12X the proportion in *D. melanogaster* (Fig. 1A) (10, 8, 11). Examination of genes with prokaryotic but not eukaryotic alignments also shows elevated levels of foreign genes in these assemblies (Fig. 1B). The number of foreign C class genes in various *H. dujardini* assemblies range from ~3-27X those in other Ecdysozoan animals (Fig. 2).

In favor of these foreign genes not being contamination is the fact that, in these independent assemblies, the majority of genes identified as coming from foreign sources resides on scaffolds together with non-foreign genes (Fig. 3). Furthermore, foreign vs. non-foreign genes from these assemblies are represented at similar proportions in the other assemblies (Fig. 4). On the whole, these independent assemblies show robust coverage of scaffolds—with genomic and RNAseq reads of 4 independent sequencing efforts (including Arakawa’s datasets prepared from extensively cleaned and antibiotic treated specimens) suggesting that widespread contamination is not an issue (Fig. 5). If these genes of foreign origin are considered to be contaminants and their scaffolds are post-filtered from their respective assembly, the completeness and size of the overall assembly is severally affected (Fig. 6 - *will refer to Blaxter v. of BioRxiv with Maker too*).

The HGT index has been demonstrated to identify genes that form monophyletic groups with non-metazoan but not metazoan sequences, indicating that the foreign genes identified above are likely actually from foreign sources (1, 8). The majority of the foreign genes identified above are assembled in scaffolds with non-foreign genes that are covered by reads from multiple independent sequencing efforts, suggesting that *H. dujardini* is likely to have an elevated level of foreign genes, most likely acquired by horizontal gene transfer rather than remaining contaminants, similar to that of rotifers.

It has been suggested that many foreign genes in the *H. dujardini* genome might not be functional or might be contaminants because RNAseq reads do not map to some of these genes (3, 5). It is well documented in the literature that many HGT genes are expressed at low levels or in some cases are not expressed at all. For example, it is known that essentially the entire *Wolbachia* genome has been transferred into the genome of *Drosophila ananassae,* yet only ~2% (28/1206) of these horizontally acquired genes are transcribed at detectable levels (12). Follow-up investigations confirmed extensive HGT into the nuclear genome of this *Drosophila* line, but failed to detect biologically relevant expression of any foreign genes (13, 14). Thus, while identification of an expressed foreign gene could be viewed as evidence in favor of HGT, lack of expression alone is not a criterion for disproving HGT. While we speculated that HGT might be important for tardigrade biology, we restricted our analysis to HGT rather than functional HGT (1). RNAseq datasets (3,5) will aid in assessing potential functions of these genes.

We acknowledge that different groups using a variety of metrics have concluded different proportions of HGT in the *H. dujardini* genome. We have our own concerns about some of the methods used by other groups to exclude genes from our assembly (see link below), but we appreciate that the work of multiple groups is moving the science forward rapidly. All genomes are interations and we fully expect that new data and new technologies will refine this genome much as they have for the human genome and others. The true proportion of HGT may well lie between the various current estimates and may best be resolved with newer or future sequencing technologies, bioinformatic, and phylogenetic approaches.

Data, methods and results of our new analyses, and more detailed responses to specific issues raised by Bemm et al. (2) and Arakawa (3) are available at https://github.com/Hd-tg-genome/PNAS\_response.

1. Boothby TC, et al. (2015) Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci* 112(52):15976–15981.

2. Bemm FM, Weiß CL, Schultz J, Förster F The genome of a tardigrade - Horizontal gene transfer or bacterial contamination? *Proc Natl Acad Sci*.

3. Arakawa K (2016) No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci*.

4. Delmont TO, Eren AM (2016) *Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies* (PeerJ PrePrints) Available at: https://peerj.com/preprints/1695.pdf [Accessed March 25, 2016].

5. Koutsovoulos G, et al. (2016) No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc Natl Acad Sci*:201600338.

6. Eren AM, et al. (2015) Anvi’o: an advanced analysis and visualization platform for ‘omics data. *PeerJ* 3:e1319.

7. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M (2013) Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* 4. doi:10.3389/fgene.2013.00237.

8. Boschetti C, et al. (2012) Biochemical Diversification through Foreign Gene Expression in Bdelloid Rotifers. *PLoS Genet* 8(11):e1003035.

9. Crisp A, Boschetti C, Perry M, Tunnacliffe A, Micklem G (2015) Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol* 16(1):50.

10. Gladyshev EA, Meselson M, Arkhipova IR (2008) Massive Horizontal Gene Transfer in Bdelloid Rotifers. *Science* 320(5880):1210–1213.

11. Eyres I, et al. (2015) Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats. *BMC Biol* 13(1). doi:10.1186/s12915-015-0202-9.

12. Dunning Hotopp JC, et al. (2007) Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes. *Science* 317(5845):1753–1755.

13. Klasson L, et al. (2014) Extensive duplication of the Wolbachia DNA in chromosome four of Drosophila ananassae. *BMC Genomics* 15(1):1.

14. Kumar N, et al. (2012) Efficient subtraction of insect rRNA prior to transcriptome analysis of Wolbachia-Drosophila lateral gene transfer. *BMC Res Notes* 5(1):230.