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) describing the discovery of extensive horizontal gene transfer in a tardigrade genome 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 used this genomic sequence and RNA-Seq data to identify likely *bona fide* tardigrade contigs. Two other reports have contributed data and analysis: Delmont and Eren (4) used a 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 assembly.

Before discussing the robustness of our finding of extensive horizontal gene transfer in tardigrades, we first note that we mistakenly uploaded an outdated version of our assembly to public databases.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). The 212.3Mb assembly used in our report is available at the link provided below. The originally uploaded 252Mb assembly includes ~40Mb of sequence that we had identified as microbial contamination (see methods detailed in link below). This mistake may have contributed to discrepancies and issues raised by others. For example, Bemm *et al.* (2) identified 39Mb of "untrusted" assembly, similar to the ~40Mb we filtered. Similarly, Delmont and Eren (4) identified whole bacterial genomes in our 252Mb assembly, which were already largely filtered out in our 212.3Mb assembly. We apologize to other groups for this mistake.

A central finding of our original report (1) was a high rate of "foreign" genes in the *H. dujardini* genome, which has been argued by others (2-5) to be an artifact of incomplete filtering of contaminating microbial sequences. The availability of these other groups' independent assemblies and multiple filtering approaches (via k-mer selection (2) or GC% and coverage (5)) allows us to test the robustness of our conclusions. Since any metric used will have its own particular strengths and weaknesses, we used 3 different approaches: i.) the HGT index (8), ii.) identifying genes that align to prokaryotic but not eukaryotic sequences, and iii.) identification of a base level of HGT, i.e. "class C" genes as defined by Crisp and coauthors (9).

Using the HGT index we find that 3.8-7.1% of genes in these independent *H. dujardini* assemblies appear foreign (Fig. 1A) (10, 8, 11). Examination of genes with prokaryotic but not eukaryotic alignments resulted in the identification of 164-384 genes (Fig. 1B). Class C genes in various *H. dujardini* assemblies range from 2.5 to 4.6% (Fig. 1C). For

each of these metrics, the proportion and/or number of foreign genes found in *H. dujardini* is substantially elevated compared to 'typical' animals (Fig. 1).

In favor of these foreign genes not being contamination is the fact that, in these independent assemblies, most genes identified as foreign reside on scaffolds together with tardigrade genes. Furthermore, foreign vs. tardigrade genes from these assemblies are represented at similar proportions in the other assemblies. On the whole, these independent assemblies show robust coverage of scaffolds from genomic and RNAseq reads of 4 independent sequencing efforts (including cleaned, antibiotic-treated individuals (3)) (Fig. 2). If these genes of foreign origin are considered to be contaminants and their scaffolds are removed, the completeness and size of each assembly is severely affected.

Together these data suggest that *H. dujardini* is likely to have an elevated level of foreign genes, most likely acquired by horizontal gene transfer rather than being artifacts of remaining contaminants. The level of HGT that we infer remains higher than in most animals that have been tested (Fig. 1), although not as high a level as we had originally concluded (1).

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). However, 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* species, 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. We speculated that HGT might be important for tardigrade biology, but we restricted our analysis to HGT rather than functional HGT (1). The RNAseq datasets (3, 5) will aid in assessing potential functions of these genes.

We acknowledge that different groups have concluded different proportions of HGT in the *H. dujardini* genome and that different metrics will give different estimates for the proportion of foreign genes in the 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 iterations 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 new sequencing technologies and bioinformatic and phylogenetic approaches.

Additional data, methods, and results of our new analyses, and more detailed responses to letters (2,3) 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 (2016) 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.