LETTER

REPLY TO BEMM ET AL. AND ARAKAWA: Identifying foreign genes in independent *Hypsibius dujardini* genome assemblies

Thomas C. Boothby[a,1] and Bob Goldstein[a]

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 *Hypsibius 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, prefiltered genome assembly. Arakawa (3) has sequenced genomes of starved, washed, individual animals that had been treated with antibiotics for 48 h, 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of our finding of extensive horizontal gene transfer in tardigrades, we first note that we mistakenly uploaded an outdated version of our assembly to public databases.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 252 Mb, larger than the 212.3-Mb assembly that we had used in our report (1), as pointed out correctly by others (5). The 212.3-Mb assembly used in our report is available at the link provided below. The originally uploaded 252-Mb assembly includes ∼40 Mb of sequence that we had identified as microbial contamination (see methods detailed at https://github.com/Hd-tg-genome/PNAS_response). This mistake may have contributed to discrepancies and issues raised by others. For example, Bemm et al. (2) identified 39 Mb of untrusted assembly, similar to the ∼40 Mb we filtered. Similarly, Delmont and Eren (4) identified whole bacterial genomes in our 252-Mb assembly, which were already largely filtered out in our 212.3-Mb assembly. We apologize to other groups for this mistake.

A central finding of our original report (1) is 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. Because any metric used will have its own particular strengths and weaknesses, we used three different approaches: (*i*) employing the horizontal gene transfer (HGT) index (8), (*ii*) identifying genes that align to prokaryotic but not eukaryotic sequences, and (*iii*) identifying a base level of HGT, i.e., "class C" genes as defined by Crisp et al. (9).

Using the HGT index, we find that 3.8–7.1% of genes in these independent *H. dujardini* assemblies appear foreign (Fig. 1*A*). Examination of genes with prokaryotic but not eukaryotic alignments resulted in the identification of 164–384 genes (Fig. 1*B*). Class C genes in various *H. dujardini* assemblies range from 2.5% to 4.6% (Fig. 1*C*). For each of these metrics, the proportion and/or number of foreign genes found in *H. dujardini* is substantially elevated compared with typical animals (Fig. 1) (8–11).

In favor of these foreign genes not being contaminants 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 four 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.

[a]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599

Author contributions: T.C.B. and B.G. designed research and wrote the paper.

The authors declare no conflict of interest.

Data deposition: 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]To whom correspondence should be addressed. Email: tboothby@gmail.com.

Together, these data suggest that *H. dujardini* is likely to have an elevated level of foreign genes, most likely acquired by HGT 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% (28/1,206) 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, although 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 https://github.com/Hd-tg-genome/PNAS_response), 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.

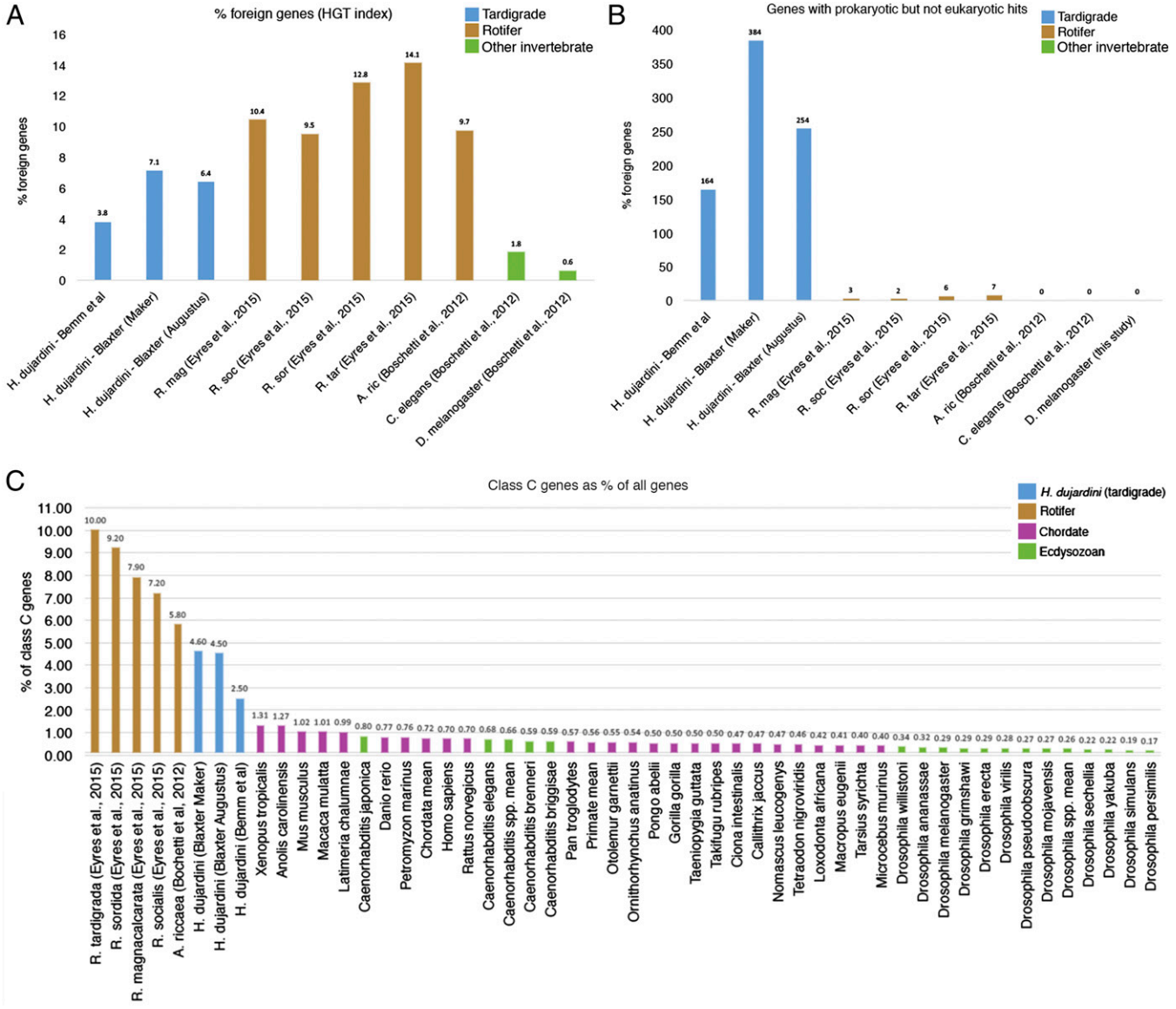

Fig. 1. Identification of foreign sequences in independent *H. dujardini* assemblies by different methods. (*A*) Foreign genes in independent *H. dujardini* genome assemblies were identified using the HGT index (8). The percent of all foreign genes in each dataset is shown. Tardigrade data are labeled in blue. Rotifer (animals with high levels of HGT) data are shown in brown. Data for other invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*) are labeled in green. Rotifer, *C. elegans,* and *D. melanogaster* data were obtained from ref. 8. (*B*) Foreign genes were identified by selecting only those genes with BLAST hits to prokaryotes, but not eukaryotes (Evalue cutoff 1e-5). Note that this is a significantly more stringent approach for identifying foreign genes than the HGT index, and it excludes identification of prokaryotic genes with metazoan homologs, nonmetazoan eukaryotic genes, or metazoan genes that have been horizontally transferred. The raw number of all foreign genes in each dataset is shown. Tardigrade data are labeled in blue. Rotifer (animals with high levels of HGT) data are shown in brown. Data for other invertebrates (*C. elegans* and *D. melanogaster*) are labeled in green. Rotifer and *C. elegans* data were derived from ref. 8. *D. melanogaster* data were obtained by performing HGT index analysis (see methods at https://github.com/Hd-tg-genome/PNAS_response) and the applying selection criteria detailed above. (*C*) Class C genes were identified in various datasets using the parameters detailed in ref. 9. Data for chordate and ecdysozoan animals were obtained from ref. 9, and the source of other datasets is noted parenthetically on the *x* axis. Plotted on the *y* axis is the percent of total genes that are classified as Class C foreign genes according to the methods used in ref. 9. Data for rotifer species are colored brown. Data from various tardigrade assemblies are colored blue. Data for other ecdysozoan animals (the group of animals to which tardigrades belong) are colored in green. Data for chordate animals are colored purple. The numbers above bars denote the percent of all genes classified as Class C foreign genes.

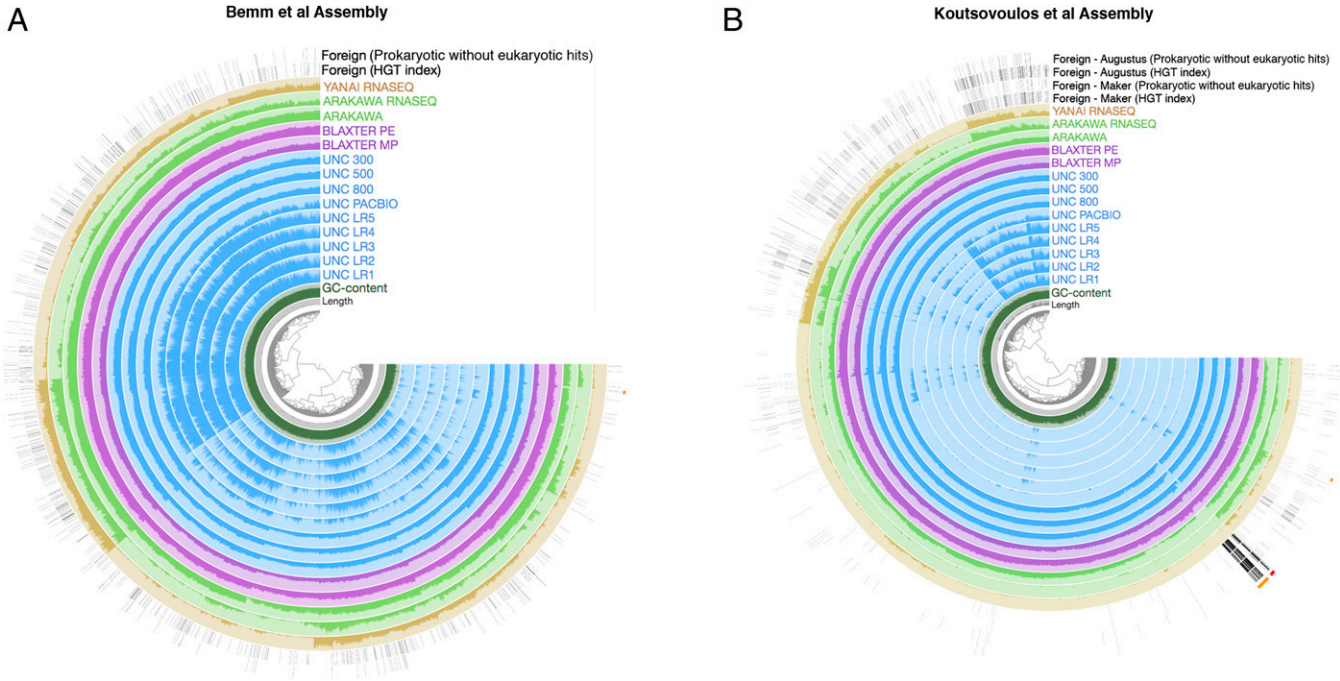

Fig. 2. Foreign genes are contained in scaffolds confirmed by multiple datasets. 14 next-generation sequencing read datasets, originating from 4 independent sequencing projects, were mapped against (*A*) Bemm et al.'s (2) and (*B*) Koutsovoulos et al.'s (5) assembly and visualized using Anvi'o (6). Tracks showing coverage of Moleculo long-read (LR), Pacbio, and short-insert library reads [University of North Carolina (UNC) 300, 500, and 800] from our original study (1) are colored blue. Tracks showing coverage by genomic reads from ref. 5 are colored purple. Tracks showing coverage by genomic and pooled RNAseq reads generated by Dr. Kazuharu Arakawa are colored green. The track showing coverage by pooled RNAseq reads generated by Dr. Itai Yanai is colored yellow. Black tick marks in the outer rings indicate scaffolds that contain foreign sequences identified using the method indicated in parentheses. Highlighted in red are scaffolds that are covered by genomic reads originating from only one group's sequencing effort. Highlighted in orange are scaffolds covered by two of three groups' genomic reads.

1 Boothby TC, et al. (2015) Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA* 112(52):15976–15981.
2 Bemm F, Weiß CL, Schultz J, Förster F (2016) Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? *Proc Natl Acad Sci USA* 113:E3054–E3056.
3 Arakawa K (2016) No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci USA* 113:E3057.
4 Delmont TO, Eren AM (2016) Identifying contamination with advanced visualization and analysis practices: Metagenomic approaches for eukaryotic genome assemblies. *PeerJ* 4:e1839.
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 USA* 113(18):5053–5058.
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* 4:237.
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):90.
12 Dunning Hotopp JC, et al. (2007) Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317(5845):1753–1756.
13 Klasson L, et al. (2014) Extensive duplication of the Wolbachia DNA in chromosome four of *Drosophila ananassae*. *BMC Genomics* 15(1):1097.
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.