Specific responses to Bemm *et al.,* 2016

We agree with the authors of Bemm *et al.* (1) that pre- and/or postfiltering possible contaminants from next-generation sequencing read datasets and assemblies is an important step in any *de novo* sequencing project. While Bemm *et al.* took a prefiltering approach, we took a postfiltering approach (see methods). We thank the authors of Bemm *et al.* for contributing their time and expertise to address the issue of contamination. Below we offer some comments on their approaches with the goal of helping the science move forward.

Bemm *et al.'s* prefiltering approach assumes that short-read datasets were sequenced close to saturation. Otherwise, selecting only k-mers that are represented in all datasets is likely to result in the exclusion of real k-mers/reads.

We found that short-read datasets were far from saturated (Fig. 8). Consistent with this, Bemm *et al.'s* filtering method identified only 9.6% of k-mers as being represented in all short-read datasets. This resulted in a 'Trusted' assembly missing nearly 20% of core eukaryotic genes and lacking >20% of publicly available *H. dujardini* ESTs (Fig. 5). Additionally, by the HGT index (2) 8501 genes (27.7%) are annotated as tardigrade genes in Bemm *et al.'s* Untrusted assembly, which also contained 24.6% of core conserved eukaryotic genes. These findings suggest to us that this method of prefiltering reads from unsaturated datasets may be too stringent, resulting in the removal of true sequences in addition to contaminants.

Our postfiltered assembly, on which our HGT analysis was originally conducted, excluded many of our longest scaffolds, which like Bemm *et al.'s* longest Untrusted scaffolds look to come from Chitinophagaceae-like bacteria (Fig. 9). Similarly, both our postfiltering and Bemm *et al's* prefiltering methods removed ~40Mb of contaminating sequences. This suggests that many of the most obvious contaminants were removed by both methods.

Bemm *et al.* used different annotation methods for their Trusted and Untrusted assemblies. They annotated their Trusted assembly using GeneMark-ES, a program designed to identify and annotate eukaryotic genes, whereas they annotated their Untrusted assembly using GeneMark-S, a program designed to identify and annotate bacterial genes. This difference in annotation methods could contribute to differences in gene attributes reported by Bemm *et al.*, such as gene spacing, and skew classification of prokaryotic and eukaryotic-like genes between their Trusted and Untrusted assemblies.

Finally, Bemm *et al.* performed reciprocal best BLAST analysis to identify genes we annotated as foreign within their Untrusted assembly. However, their Trusted assembly was not included as part of this analysis, meaning that foreign genes with better hits to genes in their Trusted assembly could be mistakenly assigned to their Untrusted assembly.

We have examined relative levels of foreign genes found in independent *H. dujardini* assemblies, including Bemm *et al.'s* Trusted assembly. Each assembly was prepared by groups working independently using different methods to stringently filter, assemble, and annotate their assemblies. Using 3 different metrics we find elevated levels of foreign genes (the majority of which reside on scaffolds with tardigrade genes and are proportionally represented in other assemblies) compared with ‘typical’ invertebrates (Fig. 1).

We are encouraged by, and grateful for the discussions, approaches, and new datasets from the scientific community, which are affording us opportunities and new ways to retest and refine our original findings and hypotheses.

1. Bemm FM, Weiß CL, Schultz J, Förster F (2016) The genome of a tardigrade - Horizontal gene transfer or bacterial contamination? *Proc Natl Acad Sci*.

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