Specific responses to Bemm et al., 2016

We agree with the authors of Bemm *et al* (1) thatpre- 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 constructive feedback, comments, and suggestions that may improve their approaches further and help the science move forward more quickly.

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. 9). Consistent with this, Bemm *et al’s* filtering method identified only 9.6% of kmers as being represented in all short-read dataset. This resulted in a ‘Trusted’ assembly missing nearly 20% of core eukaryotic genes and lacking >20% of publically available *H. dujardini* ESTs. ###Additionally, by the HGT index (2) 7160 genes are annotated as non-foreign in Bemm *et al’s* Untrusted assembly, which also contained 24.6% of core conserved eukaryotic genes. These finding suggest to us that this method of prefiltering reads from unsaturated datasets maybe too stringent, resulting in the removal of true sequences in addition to possible contamination.

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. 10). This suggests to us that many of the most obvious contaminants were removed by both methods.

Bemm *et al.* use 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. The difference in annotation methods could contribute to differences in gene attributes reported by Bemm *et al.*, such as gene spacing, and could impact the skewed representation 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 HGT within their Untrusted assembly. However, their Trusted assembly was not included as part of this analysis, meaning that HGT 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 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. None the less, using 3 different metrics to identify foreign genes we find levels of foreign genes, the majority of which reside on scaffolds with non-foreign genes and are proportionally represented in other assemblies, 2-27X proportions found in ‘typical’ invertebrates.

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

1. Bemm FM, Weiß CL, Schultz J, Förster F 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.