No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*

Georgios Koutsovoulos[a], Sujai Kumar[a], Dominik R. Laetsch[a,b], Lewis Stevens[a], Jennifer Daub[a], Claire Conlon[a], Habib Maroon[a], Fran Thomas[a], A. Aziz Aboobaker[c] and Mark Blaxter[a*]

a Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3FL, UK

b The James Hutton Institute, Invergowrie, Dundee DD2 5DA, UK

c Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

Supplemental File 1 Contents

S1.1 Supplemental Experimental procedures

Culture of *H. dujardini*

H. dujardini starter cultures were obtained from Sciento, Manchester, and cloned by isolation of single females in vinyl microtitre plates. Cultures were bulked from an individual female. Tardigrades were maintained on *Chlamydomonas reinhardtii* algae, which was grown in 1x Bold's medium, pelleted and resuspended in fresh spring water to be fed to the tardigrades. Cultures were maintained at 19°C and aerated continuously. DNA for sequencing was prepared from tardigrades of mixed ages from bulk cultures maintained in glass baking dishes. These were isolated from *C. reinhardtii* by two rounds of Baermann filtration through two layers of sterile milk filter paper and left without food until remaining green algae and darker digestion products were no longer visible in the gut (3–4 days). Tardigrades were then washed repeatedly in lab artificial freshwater by gentle centrifugation. Pelleted tardigrades were snap frozen while still alive in a minimal volume and stored at -80°C.

Genome size measurement

We estimated the size of the *H. dujardini* genome by propidium idodide staining and flow cytometry, using *C. elegans* (genome size 100 Mb), and *Gallus gallus* red blood cells (1200 Mb) as size controls, following published protocols (1).

RNA and DNA extraction

RNA was isolated from cleaned, pelleted tardigrades using Trizol reagent, after percussive disruption of cleaned tardigrades under liquid nitrogen. Genomic DNA was isolated by a manual phenol-chloroform method, after percussive disruption of cleaned tardigrades under liquid nitrogen.

Expressed sequence tag (EST) sequencing

A directional cDNA library was constructed in pSPORT1 using the SMART cDNA synthesis protocol and transformed into BL21 *E. coli*. Individual recombinant clones were picked into microtitre plates and inserts amplified using universal PCR primers (M13L and M13R). The amplified inserts were sequenced in one direction (using primer T7) after enzymatic clean-up with Exo1 and SAP, using BigDye reagents on an AB 3730 sequencer. All successful sequences were trimmed of low quality sequence and vector using trace2dbest (2) (see Table S3 for software used, version numbers and additional commands) and submitted to dbEST (see Table S2 and SI File 8). Data were publicly released on submission in 2003-2004.

Genome survey sequencing

A 2 kb-insert *H. dujardini* genomic library was constructed in the pCR4Blunt-TOPO vector. Individual recombinant clones were picked to microtitre plates and inserts amplified using M13R and pBACe3.6_T7 primers and sequenced with the T3 primer. Sequences were processed with trace2seq (2) and submitted to dbGSS (see Table S2 and SI File 8). Data were publicly released on submission in 2005.

Genome sequencing with Illumina technology

Purified *H. dujardini* genomic DNA was supplied to Edinburgh Genomics (http://genomics.ed.ac.uk) for Illumina sequencing. We obtained sequence from two libraries: a small insert library (~300 bp

insert size, prepared with Illumina TruSeq reagents by Edinburgh Genomics) and a 4 kb virtual insert mate-pair library (constructed by CGR, Liverpool). These were sequenced on HiSeq2000 at Edinburgh Genomics to generate datasets of 101 base paired end reads. The raw data are summarised in Table ST2 and are available in the ENA under study accession PRJEB11910 (runs ERR1147177 and ERR1147178).

We estimated library insert sizes by mapping read pairs to the initial single-end CLC assembly. This works well for short insert libraries, but less well for mate-pairs, where many read pairs do not map to the same contig. This failure to map to the same contig means that the mate pair reads are under-counted.

The insert size distribution of the “300 base” standard library on the preliminary SE assembly had a median of 292 bases (Standard Deviation (SD) 96 bases) (Figure S4A). The insert size distribution of the 4000 base mate pair library has a median of 1133 bases (1460 bases SD) (Figure S4B). There were many mate-pair fragments with very small virtual inserts in this library.

We were granted early access to Illumina GAIIX RNA-Seq data from Itai Yanai (accession GSE70185) in advance of publication. Lars Hering granted access to assemblies of the RNA-Seq data generated for their analyses of *H. dujardini* opsin genes (3).

Data validation and filtering for genome assembly

Details of programmes and command options used are in Table ST2 below. We performed initial quality control on our raw Illumina data using fastqc (S. Andrews, unpublished), and used trimmomatic (4) to remove low quality and adapter sequence. We screened the quality- and adapter-trimmed data for contaminants using taxon annotated GC-coverage plots (TAGC or blobplots) using an updated version of the blobtools package (https://drl.github.io/blobtools; Dominik Laetsch, in prep.). The paired-end reads were normalised with one-pass khmer (5) and were assembled with Velvet (6, 7) using a k-mer size of 55, and non-normalised reads mapped back to this assembly using the CLC mapper (CLCBio, Copenhagen) or bwa mem (8). For each scaffold, GC% was counted (ignoring N base calls) and read coverage calculated. Each scaffold was compared to the NCBI nucleotide (nt) database using BLAST megablast (9, 10) and to UniRef90 using diamond blastx (11), and the results were filtered by the blobtools script to annotate each scaffold with the taxonomy of the highest scoring matches in these databases. Blobtools estimates taxonomic similarity of a scaffold or contig either by simply recording the taxonomy of the highest match to any segment of the sequence, or assigning taxonomy based on the sum of best match scores across of the scaffold or contig. The taxonomic assignment data are summarised in a text file. The scaffolds were then plotted in a two dimensional scatter plot (X-axis : GC proportion, Y-axis : log coverage), coloured by putative taxon of origin based on the BLAST or diamond results. Using the blobplot, our cleaning process (i) identifies contigs or scaffolds with discordant coverage and taxonomic affiliation, (ii) identifies reads that were used to build those scaffolds and contigs (iii) identifies the paired reads for this set of putative contaminant reads (iv) returns to the “keep” pile any read pairs where the pair has a mapping that is not also flagged for removal. Thus the removal of reads is conservative - reads are only excluded if they and their pair have contaminant (or “no mapping”) status. This strategy is explained in the TAGC plot papers (12, 13).

Subsequent assemblies from filtered and cleaned data were also screened using blobplots. The initial Velvet assembly was used to estimate library insert sizes so that accurate parameters could be passed to subsequent assembly steps. The mate pair library insert distribution was not normally distributed, and the library contained many pairs that appeared to derive from non-mate fragments.

The blobtools cleaning process was repeated two more times, as newly assembled contaminants could be identified. Gaps were filled in the final assembly using GapFiller (14, 15). The mate pair library was used to scaffold the gap-filled assembly with SSPACE (14), accepting only the information from mate pair reads mapping 2 kb from the ends of the scaffolds. The final assembly spans 135 megabases (Mb) with median coverage of 86 fold**.** The completeness of the genome assembly was assessed using CEGMA (16), and by mapping EST, GSS and RNA-Seq data. The assembly was reviewed in a blobplot, generated as above, and the taxonomic assignment file for nHd.2.3 is available as SI File 9.

Genome annotation

We annotated the assembled *H. dujardini* genome nHd.2.3 using a two-pass approach. We used MAKER (17) to generate a first-pass set of gene models, using the ESTs and available transcriptome data as evidence, and then used these to inform a second pass of annotation with Augustus (16). Protein sequences were annotated using BLAST searches against UniRef90 and the NCBI nonredundant protein database. Protein domains and motifs were predicted with InterProScan (18). The genome sequence and annotations were loaded into an instance of BADGER (19) and made publicly available in April, 2014. The genome assembly, predicted transcriptome, predicted proteome and GFF file of annotations are available for download at http://www.tardigrades.org and at http://dx.doi.org/10.5281/zenodo.45436. The genome assemblies and annotations produced in Edinburgh have not been deposited in ENA, as we are still filtering the assembly for contamination, and have no wish to contaminate the public databases with foreign genes mistakenly labelled as “tardigrade”.

Comparison of *H. dujardini* genome assemblies

We compared the UNC *H. dujardinii* assembly (20), downloaded from http://weatherby.genetics.utah.edu/seq_transf/, 27 November 2015) to our raw Illumina data (quality and adapter trimmed but otherwise unfiltered) and the nHd.2.3 genome assembly. We mapped both our read data and the UNC TG-300, TG-500 and TG-800 library raw read data (from http://weatherby.genetics.utah.edu/seq_transf/, 01, 02, and 03 December 2015) to UNC and nHd.2.3 genome assemblies using bwa mem (8). The resulting read mapping files, together with the results of a diamond (11) search against UniRef90 and megablast (9, 10) search against NCBI nt, were used to compute blobplots of both assemblies. Summary taxonomic assignment statistics are available in SI File 3 and 9. We also accessed the UNC PacBio data from http://weatherby.genetics.utah.edu/seq_transf/ (03 December 2015). To explore transcription of putative HGT loci, we assessed gene expression using kallisto (21) and 351 million RNA-Seq reads, estimating expression as transcripts per million (tpm). Normalised, average RNAseq base-coverage for each scaffold in both assemblies was calculated by mapping RNAseq data using GSNAP (22, 23). We mapped two transcriptome assemblies provided by Hering and Mayer (3). These assemblies were based on the same raw data, assembled with CLCBio or IDBA assemblers. We screened both genome assemblies against the SILVA ribosomal RNA database (24) using BLAST.

Horizontal gene transfer into nHd.2.3

We assessed horizontal gene transfer into *H. dujardini* initially by calculating a summed best diamond blastp score for every protein predicted from nHd.2.3 compared to the UniRef90 database. From the summed scores we assessed whether the nHd.2.3.1 protein could be assigned to non-eukaryote, non-metazoan eukaryote, metazoan or unassigned origins, with assignment requiring that the taxonomic origins of $\geq$90% of all the hits returned by diamond were congruent. The label

"NotSure" was attached to proteins that had no diamond hits, or that had conflicting hits (i.e. <90% of hits were to one taxonomic group). We also calculated assignment to metazoan versus non-metazoan within the eukaryote group using the same rule. We then assessed, for each of the 6,863 scaffolds from which we predicted proteins, the presence of proteins with different taxonomic assignments. Descriptions of each of the bacterial and non-metazoan HGT candidates in nHd.2.3 are available in Supplemental Files 6 and 8. We also classified the UNC protein predictions using this pipeline. We used the mapping file provided by UNC between the UNC PacBio assembly and the UNC genome assembly (downloaded from http://weatherby.genetics.utah.edu/seq_transf/pacbio/, 3 December 2015) and scored each potential HGT–metazoan genome junction for confirmation with these long-read data.

Simple BLAST best hit analyses, even if conditioned by filters that assess a range of additional parameters, can be a crude tool. We examined the phylogenetic affinities of each candidate fHGT locus in nHd.2.3 to assess whether the assignment of their apparent taxon-of-origin was supported by a fully parameterised molecular evolutionary analysis (see SI File 6). We first clustered all the proteins in nHd.2.3.1 using OrthoFinder (25), and selected clusters containing candidate fHGT loci. As some fHGT loci were related, the number of clusters was less than the total number of fHGT candidates. For each cluster, we collected up to 20 sequences found in a BLASTp search of UniRef90. We also collected, where present, the top metazoan matches. For each cluster and its top BLAST matches we generated an alignment using Clustal Omega (26). This alignment was subjected to trimming for alignment quality using trimal (27), using two settings (soft, which removed non-conserved blocks, and hard, which additionally removes low identity alignments). The best evolutionary model for each trimmed alignment was assessed using ProtTest (28). A RAxML (29) search using the best model was then conducted using 100 bootstraps. Trees were reviewed by hand after visualisation in FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Clusters were removed from the analytic pipeline at several stages. For many clusters, too few sequences were aligned (i.e. <4) for any meaningful phylogenetic assessment. The trimming procedure resulted in removal of all residues in some alignments, which were then not analysed further. The number of clusters removed from hard trimmed alignments was, as would be expected, greater than that removed from soft trimmed alignments. For some clusters in each fHGT class (from bacteria, and from non-metazoan eukaryotes) we were unable to find credible metazoan homologues, and thus the phylogenetic analyses only included bacterial or non-metazoan sequences, respectively. For those alignments containing metazoan representatives, we sought evidence that a clan on the unrooted trees included a tardigrade sequence and bacterial or non-metazoan sequences, to the exclusion of other metazoan sequences. The proposal for fHGT was assessed as supported if such a clan was present. The alignments and RAxML trees for each cluster assessed are available at http://www.tardigrades.org and http://10.5281/zenodo.45436.

Candidate fHGTs for which phylogenetic support was found ("hard"-trimmed: 55 and "soft"-trimmed: 161) were subjected to GO-term enrichment analysis using Blast2GO (30) (see SI File 10). Given a false-discovery-rate of 0.05, the set of hard-trimmed fHGTs showed enrichment for the functional categories "mandelamide amidase activity" (p-value = 2.79 E-6; 2 in the test set vs. 0 in the reference set) and "beta-glucosidase activity" (p-value = 2.39 E-7; 3 in the test set vs. 5 in the reference set). The set of soft-trimmed fHGTs showed, in addition, enrichment of the functional "category peroxidase activity" (p-value = 2.91571E-14; 11 in the test set vs 70 in the reference set) including non-heme haloperoxidases and non-heme chloroperoxidases. However, we advise caution when drawing conclusions from these results concerning the biology of *H. dujardini* due to the draft nature of the assembly.

Availability of Supporting Data

The raw Illumina sequence read data have been deposited in SRA, and the GSS and EST data in dbGSS and dbEST respectively (see Table S4). The genome assemblies produced in Edinburgh have not been deposited in ENA, as we are still filtering the assembly for contamination, and have no wish to contaminate the public databases with foreign genes mistakenly labelled as “tardigrade”. The assemblies (including GFF files, and transcript and protein predictions), our analysis intermediate files, blobDB for each TAGC plot and high-resolution versions of Figures 1, 2, 3 and S4, the Supplemental files and a range of supplemental data are available at http://www.tardigrades.org and http://dx.doi.org/10.5281/zenodo.45436. Additional code used in the analyses is available from https://github.com/drl/tardigrade and https://github.com/sujaikumar/tardigrade.

Acknowledgements

The Edinburgh tardigrade project was funded by the BBSRC UK (grant reference 15/COD17089). GK was funded by a BBSRC PhD studentship. DRL is funded by a James Hutton Institute/School of Biological Sciences University of Edinburgh studentship. SK was funded by an International studentship and is currently funded by BBSRC award BB/K020161/1. LS is funded by a Baillie Gifford Studentship, University of Edinburgh. We thank Bob McNuff of Sciento for his inspired culturing of *H. dujardini*, Reinhardt Kristensen for guidance with microscopy and identification of *H. dujardini*, Sinclair Stammers of MicroMacro for videomicroscopy assistance, the University of Edinburgh COIL facility for fluorescence microscopy, Itai Yanai for pre-publication access to RNA-Seq data, Lars Hering for access to RNA-Seq assemblies, The University of Liverpool Centre for Genomic Research for mate-pair library construction, and staff of Edinburgh Genomics (and its precursor the GenePool) for sequencing. We thank a wide community of colleagues on twitter, blogs, and email for discussion of the results presented here which were posted on bioRxiv for discussion (http://dx.doi.org/10.1101/033464) in the weeks since the publication of the UNC genome. We thank Gytis Dudas for assistance.

S1.2 Detailed examination of the claims for HGT in *Hypsibius dujardini* by Boothby *et al.*

The claim for extensive HGT by Boothby *et al.* (20) was based on a series of findings concerning features of their assembly (the UNC assembly) and expectations of eukaryotic versus bacterial gene structure.

1 The presence in the assembly of sequences that display features of bacterial and other origins by application of a BLAST bit-score based HGT index (31).

2 The development of phylogenies for possible fHGT genes to show their affinity with non-metazoan taxa.

3 The presence of spliceosomal introns in putatively fHGT loci derived from bacteria.

4 The codon usage bias of the fHGT loci matching eukaryote biases.

5 The coverage of fHGT loci matches the *bona fide* tardigrade genome.

6 The GC proportion of fHGT loci matching the *bona fide* tardigrade genome.

7 PCR-based affirmation of some junctions between candidate fHGT genes and other genes.

8 Long-read PacBio single molecule affirmation of the assembly, and junctions.

We examined these claims in detail, and summarise our findings in the main text and figures. We use additional evidence not available to Boothby *et al.*, but note that most of the claims are poorly supported by the Boothby *et al* data itself.

1 The presence in the assembly of sequences that display features of bacterial and other origins by application of a BLAST bit-score based HGT index (31)*.*

The HGT index (24) compares BLAST bit scores of a candidate HGT sequence derived from searches of recipient and donor taxon databases. It was designed to be applied where there is prior evidence of integration of the tested sequence in a eukaryotic genome. Screening of transcriptomes generated from poly(A)-selected mRNA, which will exclude bacterial and archaeal sequences *a priori*, is thus a credible use (31). Applying the HGT index to loci predicted from poorly-assembled genomic sequence is uninformative, as a bacterial contaminant will have a high HGT index simply because it is a bacterial gene. While the finding that a large number of loci in the UNC identified by Boothby *et al*. have high HGT indices is correct, this does not in itself evidence HGT.

2 The development of phylogenies for possible fHGT genes to show their affinity with non-metazoan taxa.

Phylogenetic analyses are sensitive tests of the likely affinities of sequences, and are *de rigeur* in supporting claims of fHGT. However this test does not in itself prove fHGT in the absence of other data. Phylogenetic analysis of a gene derived from a contaminant will simply affirm the phylogenetic position of the gene within the clade of its source species. Phylogenetic analysis is strongly confirming of fHGT when fHGT is supported by other evidence of integration into a host genome.

3 The presence of spliceosomal introns in putatively fHGT loci derived from bacteria.

Boothby *et al.* noted that many of their fHGT candidates contained spliceosomal introns. As spliceosomal introns are only found in eukaryotes, this feature is strongly supportive of HGT. In the 6,663 loci identified as fHGT, 57.1% were found to have predicted introns. Intronless genes are rare and usually have specialised functions in metazoans, as participation in the intron splicing pathway is required to promote nuclear-cytoplasmic export of mRNAs for translation. The finding of introns in otherwise "bacterial" genes is likely to be an analytic artefact. Eukaryotic gene finders are designed to generate spliced gene models, and will "invent" introns in prokaryotic DNA. We have provided an example of this - where we used CEGMA and genemark to identify genes in the *E. coli* genome, at https://gist.github.com/GDKO/bc507bc9b620e6006a44. *E. coli* scores 13% in CEGMA, and 797 of 2034 genes predicted have introns. These introns conform precisely to the eukaryotic splice donor and acceptor signals. Another example from our work is the identification of apparently spliced gene models in the nuclear insertions of *Wolbachia* DNA in nematode genomes (32). This DNA is non-functional in the nematode, and the gene finders attempt to use positive signal of gene-like features in the insertions (sequence similarity matches to known proteins, biased GC% and codon usages, and open reading frames) to generate eukaryotic, spliced gene predictions. These gene predictions have correct eukaryotic splice junctions, as this is a strong prior used by the algorithms.

Thus the finding of eukaryote-like splice donor and acceptor sequences in prokaryotic DNA can simply be a process artefact resulting from the prediction algorithm. The excess of gene predictions (39,532; one of the highest of any metazoan) and the apparent uncritical acceptance of all models proposed by the variety of gene finders employed by the MAKER pipeline suggests that many of the intron-containing, bacterial-like gene predictions are wrong. We provide additional evidence of bacterial contamination for almost all of the predicted UNC fHGT candidates in SI File 4.

4 The codon usage bias of the fHGT loci matching eukaryote biases.

Codon usage bias is the result by a number of evolutionary pressures (33). The overall GC% of the genome is a strong predictor of bias, and genomes of similar GC% will tend to have similar codon biases. For example, we have demonstrated that the nematode *Caenorhabditis elegans* which has 43% GC in its coding regions, was better modelled by codon usage tables from the plant *Arabidopsis thaliana* (also 43% GC) than it was by codon usage tables from *Brugia malayi* and other filarial nematodes (which have 37% GC) (34). Organisms with the same GC% can differ markedly in synonymous codon usage. This pattern is driven by selection, likely through tRNA abundance profiles, and is most strongly evident in organisms with large effective population sizes. For example *Eschericiha coli* has very strong synonymous codon bias, but *Homo sapiens* does not. Boothby *et al.* present an analysis of codon usage, claiming that codon usage in their fHGT candidates is more similar to metazoan-like tardigrade loci than to homologues from bacterial species. The support for these assertions is not strong, as it is based either single-gene comparisons or on a selection of only 50 loci, is based on a poorly specified "difference in codon usage" metric and the statistical tests used are poorly justified (they analyse codon usage as if the codons were independent of GC% and synonymy). We have not further analysed this dataset.

5 The coverage of fHGT loci matches the bona fide *tardigrade genome.*

and

6 The GC proportion of fHGT loci matching the bona fide *tardigrade genome.*

Coverage and GC% are very informative measures for detection of different replicons in a genome assembly. If the 6,663 genes suggested by Boothby *et al.* (20) are true HGT, the scaffolds on which they reside will have read coverage and GC% similar to that of *bona fide* tardigrade scaffolds, where species origin is unproblematic. The analyses presented by Boothby *et al.* claim to have examined this issue, and report finding unbiased variation in coverage. They assume that contaminants will have lower coverage than the *bona fide* genome (which is not necessarily the case). Only standard deviations and not means were given for coverage distributions, and a surprising number of scaffolds appeared to have very low coverage in both the putative HGT and Met (metazoan) groups in their Figure S2 C and D (20).

We mapped the UNC raw data (which we trimmed and adapter cleaned), and our independent, trimmed and adapter cleaned, but otherwise unfiltered, raw data to the UNC assembly and visualised the UNC assembly in blobplots (Figure 2). These blobplots allowed us to identify both bacterial and eukaryotic sources of contamination in the UNC assembly. Summary taxonomic assignment statistics are available in Supplemental File 3. The *bona fide* tardigrade scaffolds, identified as matching *H. dujardini* GSS and ESTs, had high coverage (~100 fold base coverage in the summed UNC raw data), and formed a blob with GC% centred on ~45%. The majority of the best protein sequence similarity matches of scaffolds in this blob were to Arthropoda and other Metazoa, and a few matched Bacteria. The very high coverage scaffolds that matched existing *H. dujardini* sequences corresponded to the mitochondrion (35) and ribosomal RNAs.

Nearly one third of the UNC assembly (7,334 scaffolds, spanning 68.9 Mb) had zero or very low (<10) coverage of reads mapped from either the UNC raw data (Fig. S4 A [all data] and B-D [split by UNC sequencing library]) or Edinburgh raw data (Figure 2 B), and had sequence similarity to bacteria.

Bacterial genomes in low-complexity metagenomic datasets often assemble with greater contiguity than does the target metazoan genome, even when sequenced at low coverage, because bacterial DNA usually has higher per-base complexity (i.e. a greater proportion is coding) (12, 13), and the longest scaffolds in the UNC assembly were bacterial (Figure 3 A). The largest span of UNC contaminants matched Bacteriodetes, and had uniformly low coverage. A second group from Proteobacteria had a wide dispersion of coverage, from ~10 fold higher than the *H. dujardini* nuclear mean to zero. Most proteobacterial scaffolds had distinct higher GC% compared to *bona fide H. dujardini* scaffolds. It was striking that many of these putatively bacterial scaffolds had close to zero coverage in both UNC and Edinburgh data (Figure 3 B). In Figure 3 B, scaffolds assigned to Eukaryota spanned 173 Mb and had an N50 of 17.3 kb, Bacteria scaffolds spanned 70.7 Mb and had an N50 of 13.3 kb, Archaea scaffolds spanned 0.1 Mb and had an N50 of 12.7 kb, Virus scaffolds spanned 0.01 Mb and had an N50 of 6.3 kb, and scaffolds with no significant similarity matches spanned 8.5 Mb and had an N50 of 7.2 kb.

A similar analysis of nHd.2.3 identified a small number of remaining contaminants (Figure 3 C). In Figure 3 C scaffolds assigned to Eukaryota spanned 118 Mb and had an N50 of 58.3 kb, Bacteria scaffolds spanned 5.0 Mb and had an N50 of 37.7 kb, Archaea scaffolds spanned 0.02 Mb and had an N50 of 1.5 kb, Virus scaffolds spanned 0.03 Mb and had an N50 of 22.3 kb, and scaffolds with no significant similarity matches spanned 11.2 Mb and had an N50 of 2.0 kb.

The wide spread of coverage in the Edinburgh data of UNC proteobacterial scaffolds may reflect the presence of related but not identical contaminants in the UNC and Edinburgh cultures, but the variation in coverage between datasets make it unlikely that these are common symbionts.

7 PCR-based affirmation of some junctions between candidate fHGT genes and other genes.

and

8 Long-read PacBio single molecule affirmation of the assembly, and junctions.

Boothby *et al.* (20) assessed integration by identifying fHGT candidates in their assembly, and affirming correct assembly by direct PCR amplification between the focal gene and neighbouring gene, and by mapping long-read single-molecule PacBio data. We approached this component of identification of fHGT in two ways: by screening the complete UNC assembly, and by reviewing the subset of fHGT candidates tested by Boothby *et al.*.

We classified each UNC protein prediction as viral, bacterial, archaeal, non-metazoan eukaryotic, or metazoan if their matches in the NCBI non-redundant (nr) databases had a dominant taxonomic source, or unclassified in the case of conflicting signal. To generate a primary list of potential fHGT candidates, we screened the neighbourhood of each non-metazoan locus in the UNC assembly and identified 713 non-metazoan to metazoan and 294 non-eukaryote to eukaryote junctions. Long-read PacBio data are ideal for direct confirmation of linkage between fHGT and neighbouring genes for which long-term vertical phylogenetic inheritance is likely. The UNC PacBio data were of relatively low quality (a mean length of 1.8 kb and a N50 length of 2.0 kb), and the PacBio assembly provided by the authors spanned only 120 Mb (with an N50 of 3.3 kb). PacBio data affirmed only 26 non-metazoan to metazoan linkages of the 713 present in the assembly, and 10 non-eukaryote to eukaryote junctions out of the 294 present.

Boothby *et al.* (20) also assessed genomic integration of 107 candidate loci directly, using PCR amplification of predicted junction fragments. Their 107 candidates were reported to include 38 bacterial–bacterial, 8 archaeal–bacterial, and 61 non-metazoan to metazoan or non-eukaryotic to eukaryotic gene pairs (summarised in Supplemental Table S3; details are present in SI File 5). Our assessment of the taxonomic origin of the loci in these pairs suggested some of classifications were in error, and we identified 49 bacterial to bacterial junctions. Confirmation was achieved by Boothby *et al.* for most junctions, but PCR products were only analysed electrophoretically (several had faint or multiple products), and no sequencing to confirm the expected amplicon sequence was reported. Many (49/107) of these analyses did not explicitly assess HGT, as they probed relationships between pairs of bacterial genes. Rather obviously, bacterial genes will have bacterial neighbours in bacterial genomes. We found no expression of the 49 bacterial to bacterial junction loci, further supporting their assignment as contaminants rather than examples of fHGT.

We found the remaining 58 to contain 24 prokaryotic–eukaryotic, 27 non-metazoan eukaryotic–metazoan and 7 viral–eukaryotic junction pairs (Table S3). Of the 24 prokaryotic–eukaryotic junctions, two of the putative eukaryotic neighbours have marginal assignment to Eukaryota. The 27 non-metazoan eukaryotic–eukaryotic junctions include 6 where the assignment of the focal locus (non-metazoan eukaryotic) is unclear. The 7 viral–eukaryotic junctions include one where the assignment of the neighbouring gene is uncertain, and we note that 6 of the 7 viral candidates involved homologues of the same protein (carrying domain of unknown function DUF2828), all from Mimiviridae. Mimiviruses are well known for their acquisition of foreign genes, and thus these scaffolds may derive from mimivirus infection of one of the several species in the multi-xenic culture

rather than tardigrade genome insertions. All 58 loci had read coverage in the Edinburgh raw data, and we observed the same genomic environment in nHd.2.3 in 51 gene pairs. We found evidence of expression from 49 of these loci (Table S3).

We were able to confirm only 32 of the 107 putative fHGT linkages in the UNC PacBio data (Table S3). We summarise our assessment of each of the 107 loci tested by Boothby et al in SI File 5, and our assessment of each of the 6,663 fHGT candidates proposed by Boothby et al. in SI File 4.

S1.3 Abbreviations

BLAST Basic Local Alignment Search Tool

cDNA copy DNA

CEGMA Core Eukaryotic Genes Mapping Approach

ENA European Nucleotide Archive node of the International Nucleotide Sequence Database Consortium

EST expressed sequence tag, stored in the database of ESTs (dbEST)

fHGT functional horizontal gene transfer

GC guanine plus cytosine

GFF genome feature format

GSS genome survey sequence, stored in the database of GSS (dbGSS)

LSU large subunit ribosomal RNA

N50 length length-weighted median contig length

PacBio Pacific Biosciences SMRT single-molecule real time sequencing

PCR polymerase chain reaction

RNA-Seq Transcriptomic RNA sequencing (as performed on the Illumina platform)

SSU small subunit ribosomal RNA

TAGC plot taxon-annotated GC-coverage plot

tpm (length-normalised) transcripts per million as estimated in kallisto

UNC University of North Carolina

UniRef90 UniProt Reference Clusters collapsed at 90% pairwise identity, a product of the UniProt database.

S1.4 Supplemental Figure Legends

Supplemental Figure 1: The tardigrade *Hypsibius dujardini*

A A whole, dorsal view of a living *H. dujardini* adult, taken under differential interference contrast microscopy. The head, with two eyes, is to the top right. The green colouration in the centre is algal food in the gut. Within the body, numerous coleomocytes and strands corresponding to the unicellular muscles (see **D**) can be seen. Six of the eight legs are under the body

B, C Identification of the species in the Sciento culture was confirmed as *H. dujardini* by comparing the morphology of the doubleclaws on the legs (**B**) and of the pharyngeal armature (the stylets and placoids) to the standard systematic key (36) (**C**).

D *H. dujardini* has readily accessible internal anatomy. In this fluorescence micrograph, the animal has been stained with rhodamine-phalloidin to label the actin bundles, especially in the muscles. The arrangement of these muscles can be followed through the three dimensional animal, mapping central and distal attachment points. The bright component to the left is the triradial myoepithelial pharynx. (This image is one of a stacked confocal series).

E, F DIC and matching fluorescence confocal image of a *H. dujardini* stained with bis-benzimide (Hoechst 3342) and biodipy ceramide. The bis-benzimide (blue) labels nuclei, while the biodipy ceramide labels lipid membranes and particularly membranes of neural cells. This ventral view shows the paired ventral nerve cords that link the four segmental ganglia. Each leg has a focus of nuclei associated with gland cells. (This image is one of a stacked confocal series).

The scale bar in F is 40 micrometres.

Supplemental Figure 2: Insert size estimation for Illumina libraries

A Insert size distribution for the short-insert library. Left panel, left graph: read pairs mapping in the expected F-R orientation. Left panel, right graph: read pairs mapping in the unexpected R-F (mate) orientation. The right panel shows the same data with an expanded X-axis.

B Insert size distribution for the mate-pair library. Left panel, left graph: read pairs mapping in the unexpected F-R orientation. Left panel, right graph: read pairs mapping in the expected R-F (mate) orientation. The right panel shows the same data with an expanded X-axis.

Supplemental Figure 3: The BADGER genome exploration environment for *H. dujardini*

The *Hypsibius dujardini* genome has been loaded into a dedicated BADGER genome exploration environment at http://www.tardigrades.org. The explorer will be updated as new analyses are performed.

Supplemental Figure 4: Mapping of UNC raw read data to the UNC assembly of *H. dujardini*

A Blobplot (https://drl.github.io/blobtools) of the UNC *Hypsibius dujardini* assembly contigs with coverage derived from pooled UNC raw genomic sequence data (data files TG-300, TG-500 and TG-800), as in Figure 2 A. All short insert raw reads were mapped back to this assembly, and each

scaffold is plotted based on its GC content (X-axis) and coverage (Y-axis), with a diameter proportional to its length and coloured by its assignment to phylum. The colours used for each phylum, and the number of contigs or scaffolds assigned to that phylum, the span of these contigs or scaffolds, and the N50 length of the contigs or scaffolds is given in the top right quadrant. The histograms above and to the right of the main plot sum contig spans for GC

B Blobplot showing the UNC assembly contigs with coverage derived from the UNC TG-300 raw genomic sequence data. Scaffold points are plotted as in Figure 1 A.

C Blobplot showing the UNC assembly contigs with coverage derived from the UNC TG-500 raw genomic sequence data. Scaffold points are plotted as in Figure 1 A.

D Blobplot showing the UNC assembly contigs with coverage derived from the UNC TG-800 raw genomic sequence data. Scaffold points are plotted as in Figure 1 A.

A high resolution version of this Figure is available as SI File 7.

Table S1. *Hypsibius dujardini* assembly comparison

Genome	***H. dujardini* Edinburgh preliminary assembly**	***H. dujardini* Edinburgh final assembly**	***H. dujardini* UNC**
Filename	*nHd.1.0*	*nHd.2.3.abv500.fna*	*tg.genome.fsa*
Longest scaffold (bp)	676,517	594,143	1,534,183
Scaffold metrics			
Number of scaffolds	88,105	13,202	22,497
Span (bp)	185,756,071	134,961,902	252,538,263 *
Minimum length (bp)	200	500	2,000
Mean length (bp)	2,108	10,222	11,225
N50 length (bp)	9,778	50,531	15,907
Number of scaffolds in N50	4,326	701	4,078
GC proportion	0.479	0.452	0.469
Span of uncalled bases (N) (bp)	191,665	3,548,224	35,835
Metrics for contigs longer than 100 bp (scaffolds split at >= 10 Ns)			
Longest contig	194,006	116,477	1,534,183
Number of contigs	271,185	25,005	22,972
Span (bp)	178,005,367	131,393,004	252,502,428
Minimum length (bp)	100	100	2,000
Mean length (bp)	656	5,254	10,991
N50 length (bp)	970	11,636	15,542
Number of contigs in N50	42,909	3,245	4,197
CEGMA quality assessment (16)			
Complete	n/a	88.7%	89.5%
Average number of copies (complete)	n/a	1.35	3.26
Complete and partial	n/a	97.2	94.8
Average number of copies (complete and partial)	n/a	1.55	3.52
Genome content			
ESTs mapping to assembly †	96.1%	96.4%	92.3%
GSSs mapping to assembly ‡	98.6%	98.9%	93.1%
Proportion of transcriptome (3) mapping to assembly §	92.1% / 93.8%	91.8% / 93.6%	85.2% / 88.2%
Proportion of RNA-Seq reads mapping to assembly	92.6%	92.8%	89.5%
Number of protein-coding genes	n/a	23,021	39,532 ¶
Potential contaminant span in assembly #	n/a	1.5 Mb	68.9 Mb
Potential contaminant proportion	n/a	1.1%	27.3%
Initial number of putative HGT genes	n/a	216 + 385 ‖	6,663
Genes derived from probable bacterial contamination **	n/a	355	9,872 ‡‡
Remaining HGT candidates showing expression >0.1 tpm	n/a	196 + 369 ‖	n/a

* In the published manuscript the assembled genome is described as being 212 Mb in span, but this is an error.

† Proportion of 5235 EST sequences; megablast search with E-value cutoff 1e-65.

‡ Proportion of 1063 GSS sequences; megablast search with E-value cutoff 1e-65.

§ Hering and Mayer (3) generated two assemblies, one with CLCBio (33,530 transcript fragments, left scores) and one with IDBA (29,288 transcripts, right scores).

¶ In their manuscript, Boothby *et al.* (20) state that they predicted 38,145 genes. However in the GFF annotation file there are 39,532 protein coding gene predictions.

Assessed from blobplot analyses (Supplemental Files 3 and 4).

‖ Bacterial + non-metazoan eukaryote, respectively. The “non-metazoan eukaryote” loci may be tardigrade.

** Genes present on scaffolds identified as likely to derive from contaminants.

‡‡ There were 9,872 loci predicted on the 68.9 Mb of contaminant scaffolds. Not all were flagged as fHGT by Boothby *et al.* (20)

n/a: not assessed.

Table S2. Scaffolds containing bacterial and eukaryotic ribosomal RNA sequences in the UNC *H. dujardini* assembly.

rDNA	UNC scaffold name	ribosomal RNA sequence match	percentage identity	alignment length	E-value	Kingdom	Phylum	diagnosis
SSU	scaffold3_size1208507	AF418954.1.1472	98.78	1473	0	Bacteria	Armatimonadetes	bacterial contaminant
SSU	scaffold8370_size10204	EU403982.1.853	98.48	853	0	Bacteria	Armatimonadetes	bacterial contaminant
SSU	scaffold1508_size26732	HM262842.1.1359	99.12	1359	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold20720_size3563	KC424744.1.1516	99.19	1486	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold798_size35300	FM200995.1.867	99.65	867	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold8_size763136	FJ719709.1.1479	99.66	1479	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold9_size589582	EU431693.1.1487	98.99	1484	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold9893_size9116	HQ111170.1.1485	99.39	1484	0	Bacteria	Bacteroidetes	bacterial contaminant
SSU	scaffold117_size78986	JF731636.1.586	100	586	0	Bacteria	Chloroflexi	bacterial contaminant
SSU	scaffold4784_size14796	JF235642.1.1304	99.39	1304	0	Bacteria	Chloroflexi	bacterial contaminant
SSU	scaffold15864_size5630	HM921100.1.1296	99.38	967	0	Bacteria	Planctomycetes	bacterial contaminant
LSU	scaffold20255_size3726	JMIT01000004.442220.445136	99.42	1373	0	Bacteria	Proteobacteria	bacterial contaminant

SSU	scaffold20255_size3726	JN982334.1.1740	98.73	1736	0	Bacteria	Proteobacteria	bacterial contaminant
SSU	scaffold24845_size2328	AY328759.1.1532	98.24	1196	0	Bacteria	Proteobacteria	bacterial contaminant
LSU	scaffold10356_size8852	AF245379.1.2684	99.53	1290	0	Bacteria	Verrucomicrobia	bacterial contaminant
SSU	scaffold10356_size8852	AJ966883.1.1522	99.74	1522	0	Bacteria	Verrucomicrobia	bacterial contaminant
LSU	scaffold5217_size14016	AYZY01065239.1.1194	98.5	1198	0	Bacteria	Verrucomicrobia	bacterial contaminant
SSU	scaffold5217_size14016	JN820219.1.1522	99.34	1521	0	Bacteria	Verrucomicrobia	bacterial contaminant
LSU	scaffold2445_size21317	GQ398061.6667.10164	98.89	3500	0	Eukaryota	Rotifera	rotifer contaminant
SSU	scaffold2445_size21317	GQ398061.4166.5977	99.5	1812	0	Eukaryota	Rotifera	rotifer contaminant
LSU	scaffold2691_size20337	GQ398061.6667.10164	98.89	3500	0	Eukaryota	Rotifera	rotifer contaminant
SSU	scaffold2691_size20337	GQ398061.4166.5977	99.5	1812	0	Eukaryota	Rotifera	rotifer contaminant
LSU	scaffold13679_size6865	GBZR01000520.241.2385	100	2145	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
SSU	scaffold13679_size6865	GBZR01012413.16.1820	99.94	1805	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
SSU	scaffold14700_size6246	GBZR01012413.16.1820	99.94	1805	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
LSU	scaffold4498_size15348	GBZR01009173.1125.4217	99.97	3093	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
SSU	scaffold4498_size15348	GBZR01012413.16.1820	99.94	1805	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
LSU	scaffold4704_size14961	GBZR01000520.241.2385	100	2145	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
LSU	scaffold6057_size12691	GBZR01009173.1125.4217	99.91	1166	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
SSU	scaffold6057_size12691	GBZR01012413.16.1820	99.5	672	0	Eukaryota	Tardigrada	*Hypsibius dujardini*

			5					
LSU	scaffold7913_size10578	GBZR01009173.1125.4217	99.97	3093	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
SSU	scaffold7913_size10578	GBZR01012413.16.1820	99.94	1805	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
LSU	scaffold864_size34133	GBZR01000520.241.2385	100	2145	0	Eukaryota	Tardigrada	*Hypsibius dujardini*

Table S3. Putative HGT loci in the UNC *H. dujardini* assembly

type [*]	revised type [*]	no.	in PacBio [†]	in nHd read data	linkage in nHd.2.3	expressed [‡]	blobplot [§]	informative
P:A-B	B-B	8	0 (0)	n/a	n/a	0	contam.	no
P:B-B	B-B	36	3 (2)	n/a	n/a	0	contam.	no
P:B-?	B-?	3	0 (0)	n/a	n/a	0	contam.	no
P:A-nME	B-B	1	0 (0)	n/a	n/a	0	contam.	no
P:B-M	B-B	1	0 (0)	n/a	n/a	0	contam.	no
P:B-M	?-E	2	1 (0)	2	2	0	tardigrade	no
P:A-M	B-E	1	1 (0)	1	1	1	tardigrade	
P:B-nME	B-nME	1	1 (0)	1	1	1	tardigrade	
P:B-M	B-M	20	7 (2)	20	16	19	tardigrade	
E:F-M	?-M, F-?, ?-?	4	1 (0)	4	4	2	tardigrade	no
E:S-M	S-?	2	0 (0)	2	2	1	tardigrade	no
E:F-M	F-M	17	12 (1)	17	15	17	tardigrade	
E:S-M	F-M	1	0 (0)	1	1	1	tardigrade	
E:S-M	S-M	3	0 (0)	3	2	2	tardigrade	
V:V-M	V-?	1	0 (0)	1	1	0	tardigrade	no
V:V-M	V-M	6	4 (0)	6	6	5	tardigrade	
Total		107	30 (5)	58	51	49		
Total HGT-informative		51	25 (3)	49	42	46		

[*] From Supplemental Information file Dataset_S02 in (20). Junctions are classified by the inferred source of the loci. Thus P:B-B means “Prokaryote source, bacterial–bacterial junction”. A archaeal, B bacterial, E eukaryote, F fungal, M: metazoan, nME non-metazoan eukaryote, P prokaryote, S streptophyte, V viral, ?: the taxonomic placement of locus was poorly supported. [†] Number of junctions affirmed by UNC PacBio data (20) from the end of gene 1 to the start of gene 2. The number affirmed across the full span (the start of gene 1 start to the end of gene 2) is given in brackets. [‡] Genes were counted as being expressed if they had more than 0.1 tpm (0.0001% of all reads) from (37). [§] tardigrade: similar high read coverage and GC% to *bona fide* tardigrade scaffolds; contam: low coverage (<10) and/or GC% divergent from *bona fide* tardigrade scaffolds. n/a: not assessed. An annotated version of Dataset_S02 (20) is available as Supplemental File 4.

Table S4. Raw data for *H. dujardini*

Data type	***Platform***	***Read length***	***Insert size***	***Number of reads (raw)***	***Number of reads (trimmed)***	***Number of bases (trimmed)***	***Accessions***
EST	AB3730 (Sanger)	>100 b	100 bp - 5 kb	n/a	5235	2,916,184	CD449043 to CD449952, CF075629 to CF076100, CF544107 to CF544792, CK325778 to CK326974, CO501844 to CO508720, and CO741093 to CO742088; see Supplemental File 8 for all accession numbers
GSS	AB3730 (Sanger)	>100 b	2 kb	n/a	1063	626,204	CZ257545 to CZ258607; see Supplemental File 8 for all accession numbers
short insert	Illumina HiSeq2000	101 b paired end	300 bp	74,374,353 pairs	67,405,223 pairs	12,839,412,868	see Figure S4 for insert size distribution. Accession ERR1147177
mate pair	Illumina HiSeq2000	101 b paired end	4 kb*	58,825,639 pairs	44,484,447 pairs	4,934,137,897	see Figure S4 for insert size distribution. Accession ERR1147178
RNA-Seq	Illumina GAIIX	101 b paired end	140 b	175,600,991 pairs	144,545,842 pairs	28,053,857,067	Accession GSE70185. These reads are from Levin *et al*. “The phyletic-transition and the origin of animal body plans” (37).

n/a: not applicable

* The mate pair library had a wide mate pair insert size distribution (see Figure S4), such that the median insert size was 1.1 kb (SD 1.4 kb) rather than 4 kb. This deviation from the desired insert size was due to the library containing many fragments that appear to be standard, non-mate-pair derived segments of the genome, as can be common in such libraries.

Table S5. Software used

Software	***version***	***additional parameters***	***source***	***reference***
trace2dbest	3.0.1	default	http://www.nematodes.org/bioinformatics/trace2dbEST/	(2)
fastqc	0.11.4	default	http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc	
trimmomatic	0.35	default	http://www.usadellab.org/cms/?page=trimmomatic	(4)
khmer		one pass default	https://github.com/dib-lab/khmer	(5)
BLAST	2.2.31+	contingent on search	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download	(9, 10)
diamond	0.79	contingent on search	https://github.com/bbuchfink/diamond/	(11)
blobtools	0.9.4	NCBI Taxonomy retrieved 19 October 2015	https://drl.github.io/blobtools	(see (13))
bwa	0.7.12-r1044	bwamem	http://bio-bwa.sourceforge.net/	(8)
MAKER	2.28	default	http://www.yandell-lab.org/software/maker.html	(17)
CEGMA	2.5	default	http://korflab.ucdavis.edu/datasets/cegma/	(16)
Augustus	2.5.5	default	http://bioinf.uni-greifswald.de/augustus/downloads/	(38)
CLC assembler	3.2.2	default	http://www.clcbio.com	
CLC mapper	3.2.2	-l 0.9 -s 0.9	http://www.clcbio.com	
BADGER	1.0	default	https://github.com/elswob/Badger	(19)
InterProScan	5	default	https://www.ebi.ac.uk/interpro/download.html	(18)
Velvet	1.2.06	kmer size 55, -exp_cov auto -cov_cutoff auto	https://www.ebi.ac.uk/~zerbino/velvet/	(6, 7)
GapFiller	1.10	default	http://www.baseclear.com/genomics/bioinformatics/basetools/gapfiller	(14, 15)
SSPACE	3	accepting only information from reads mapping 2 kb from the ends of initial	http://www.baseclear.com/genomics/bioinformatics/basetools/SSPACE	(14)

		scaffolds		
kallisto	0.42.4	-b 10	https://pachterlab.github.io/kallisto/	(21)
GMAP/GSNAP	2015-11-20	--nofails --novelsplicing=1	http://research-pub.gene.com/gmap/	(22, 23)
OrthoFinder	0.3	default	http://www.stevekellylab.com/software/orthofinder	(25)
Clustal omega	1.2.0	default	http://www.clustal.org/omega/	(26)
trimal	1.4	Soft Trim: -gt 0.8 -st 0.001 Hard Trim: -gt 0.8 -st 0.001 -resoverlap 0.75 -seqoverlap 80	http://trimal.cgenomics.org/	(27)
ProtTest	3.4	-JTT -LG -WAG -Dayhoff -G -AICC	https://github.com/ddarriba/prottest3	(28)
RAxML	8.1.20	default	http://sco.h-its.org/exelixis/software.html	(29)
FigTree	1.4.2	default	http://tree.bio.ed.ac.uk/software/figtree/	
blast2GO	3.2	UniProtKB/SwissProt retrieved 26 November 2015	https://www.blast2go.com	(30)

Table S6: Description of Additional Supplemental Files

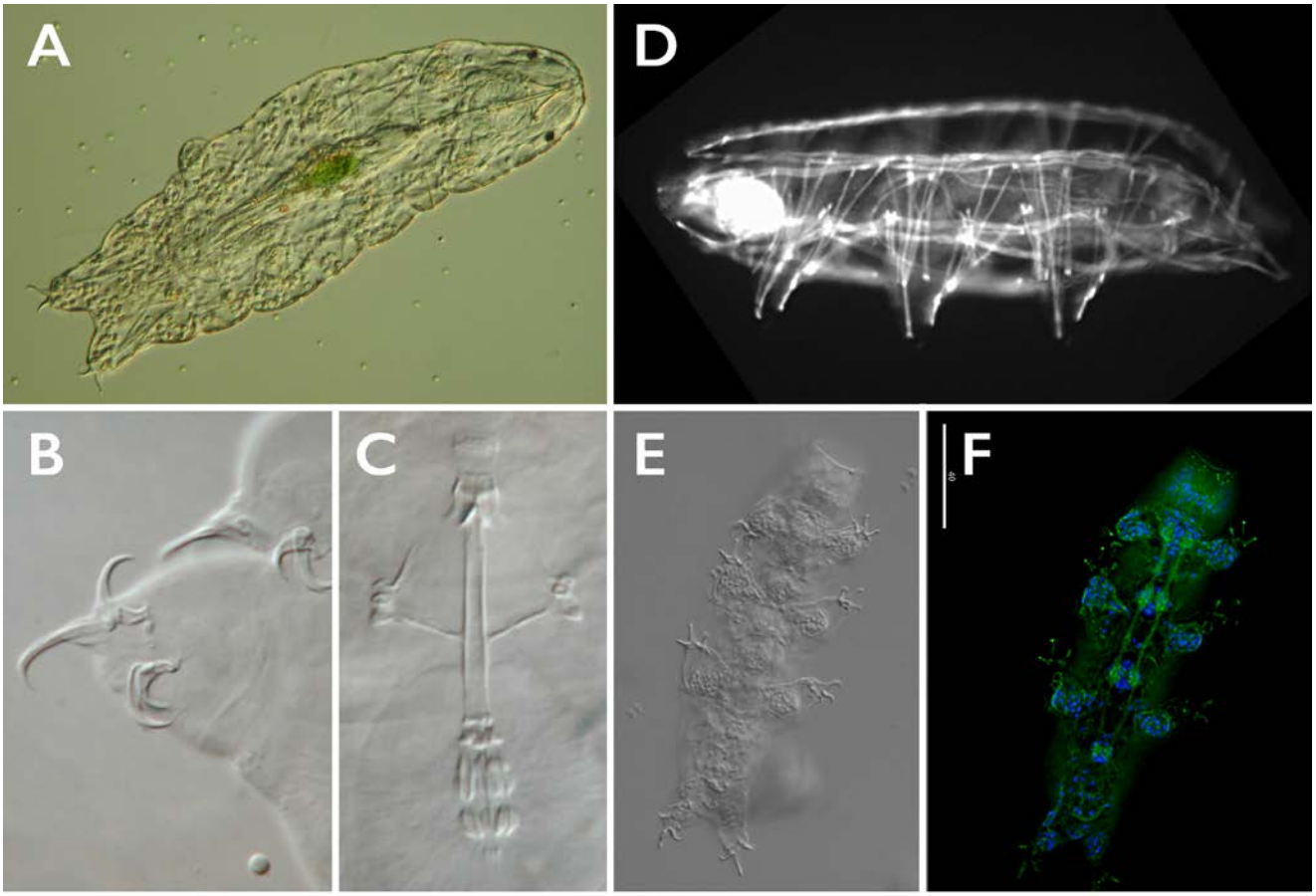
File	Content
Supplemental_File_1	This file. PDF file.
Supplemental_File_2_nHd23_likely_contaminant_scaffolds.xls	File listing scaffolds in nHd.2.3 that were identified as potentially derived from contaminating organisms rather than *H. dujardini*. xls file.
Supplemental_File_3_Summary_stats_from_blobplot_of_UNC_tggenome.xls	Summary statistics, generated in blobtools, for the TAGC plots of the UNC assembly, as presented in Figure 2 A-D. xls file.
Supplemental_File_4_Assignments_of_the_6663_putative_UNC_HGT_loci.xls	Review of properties of 6,663 putative HGT candidates from Boothby et al. xls file.
Supplemental_File_5_HGT_candidates_in_the_UNC_assembly_tested_by_PCR.xls	File reproducing the table presented by Boothby et al. on pages 184-185 of their supplemental file 02 PDF, giving, first all the columns of data presented by Boothby et al. (columns A to AB) and then (in columns AC to AN) our analyses, including data on linkage of candidates in nHd.2.3 and expression levels. xls file.
Supplemental_File_6_Analysis_of_Edinburgh_fHGT_candidates.xls	File giving phylogenetic analysis of Bacterial fHGT candidates, phylogenetic analysis of Non-Metazoan fHGT candidates, All Bacterial fHGT candidates, All Non-metazoan fHGT candidates.xls file.
Supplemental_File_7_High_resolution_figures.pdf	High-resolution, vector file version of Figures 1, 2, 3, S4. PDF.
Supplemental_File_8_EST_and_GSS_accessions.xls	Accession numbers for all EST and GSS sequences. xls file.
Supplemental_File_9_Summary_stats_from_blobplot_of_nHd23.xls	Summary statistics, generated in blobtools, for the TAGC plots of the nHd.2.3 assembly, as presented in Figure 1 B. xls file.
Supplemental_File_10_Blast2GO_results.xls	Blast2GO results for phylogenetically verified fHGT candidates from nHd.2.3. xls file.

References

1. Hare EE, Johnston JS (2011) Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol* 772:3-12.
2. Parkinson J*, et al.* (2004) PartiGene - constructing partial genomes. *Bioinformatics* 20(9):1398-1404.
3. Hering L, Mayer G (2014) Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in panarthropoda. *Genome Biol Evol* 6(9):2380-2391.
4. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
5. Crusoe MR*, et al.* (2015) The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res* 4:900.
6. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18(5):821-829.
7. Zerbino DR, McEwen GK, Margulies EH, Birney E (2009) Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One* 4(12):e8407.
8. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754-1760.
9. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
10. Boratyn GM*, et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Research* 41(Web Server issue):W29-33.
11. Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12(1):59-60.
12. Kumar S, Blaxter ML (2011) Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis* 55(3):119-126.
13. 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. *Frontiers in Genetics* 4:237.
14. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27(4):578-579.
15. Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. *Genome Biol* 13(6):R56.
16. Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23(9):1061-1067.
17. Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491.
18. Jones P*, et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30(9):1236-1240.
19. Elsworth B, Jones M, Blaxter M (2013) Badger--an accessible genome exploration environment. *Bioinformatics* 29:2788-2789.
20. Boothby TC*, et al.* (2015) Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci U S A* 112:15976-15981.
21. Bray N, Pimentel H, Melsted P, Pachter L (2015) Near-optimal RNA-Seq quantification. *arXiv*: arXiv:1505.02710.
22. Wu TD, Nacu S (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26(7):873-881.
23. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859-1875.

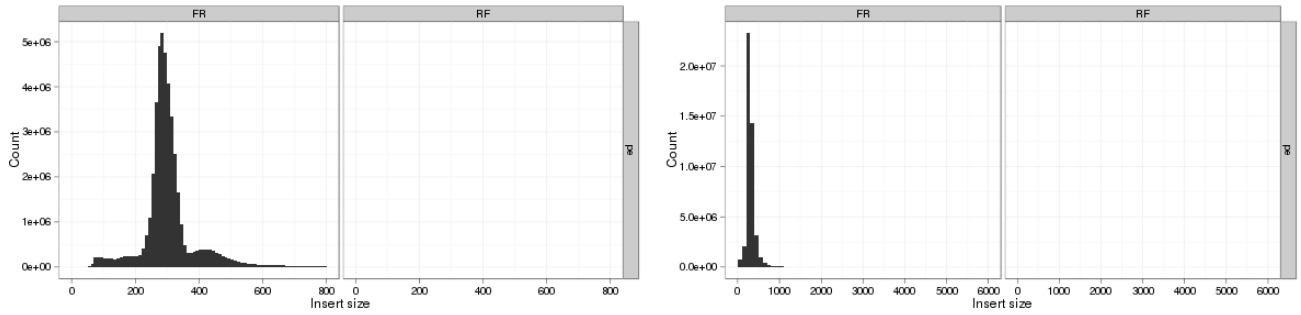
24. Pruesse E*, et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35(21):7188-7196.
25. Emms DM, Kelly S (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16:157.
26. Sievers F, Higgins DG (2014) Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* 1079:105-116.
27. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972-1973.
28. Darriba D, Taboada GL, Doallo R, Posada D (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164-1165.
29. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22(21):2688-2690.
30. Conesa A*, et al.* (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21(18):3674-3676.
31. Boschetti C*, et al.* (2012) Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genet* 8(11):e1003035.
32. Koutsovoulos G, Makepeace B, Tanya VN, Blaxter M (2014) Palaeosymbiosis revealed by genomic fossils of Wolbachia in a strongyloidean nematode. *PLoS Genet* 10(6):e1004397.
33. Lynch M (2007) *The origins of genome architecture* (Sinauer Associates, Sunderland, Mass.) pp xvi, 494 p.
34. Wasmuth JD, Blaxter ML (2004) prot4EST: Translating Expressed Sequence Tags from neglected genomes. *BMC Bioinformatics* 5(1):187.
35. Rota-Stabelli O*, et al.* (2010) Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol Evol* 2:425-440.
36. Morgan CI, King PE (1976) *British tardigrades (Tardigrada): Keys and notes for identification of the species.* (Academic Press, London) p 133.
37. Levin M*, et al.* (in press) The phyletic-transition and the origin of animal body plans.
38. Stanke M*, et al.* (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* 34(Web Server issue):W435-439.

Supplemental Figure 1: The tardigrade *Hypsibius dujardini*

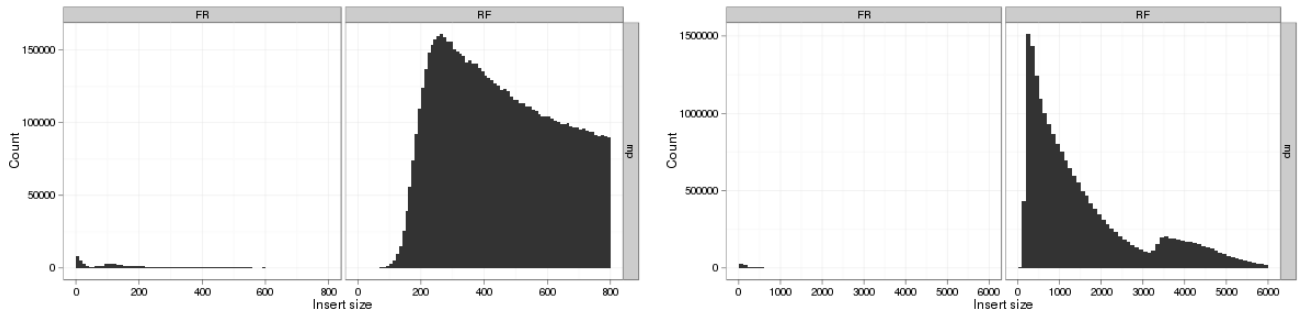


Supplemental Figure 2: Insert size estimation for Illumina libraries

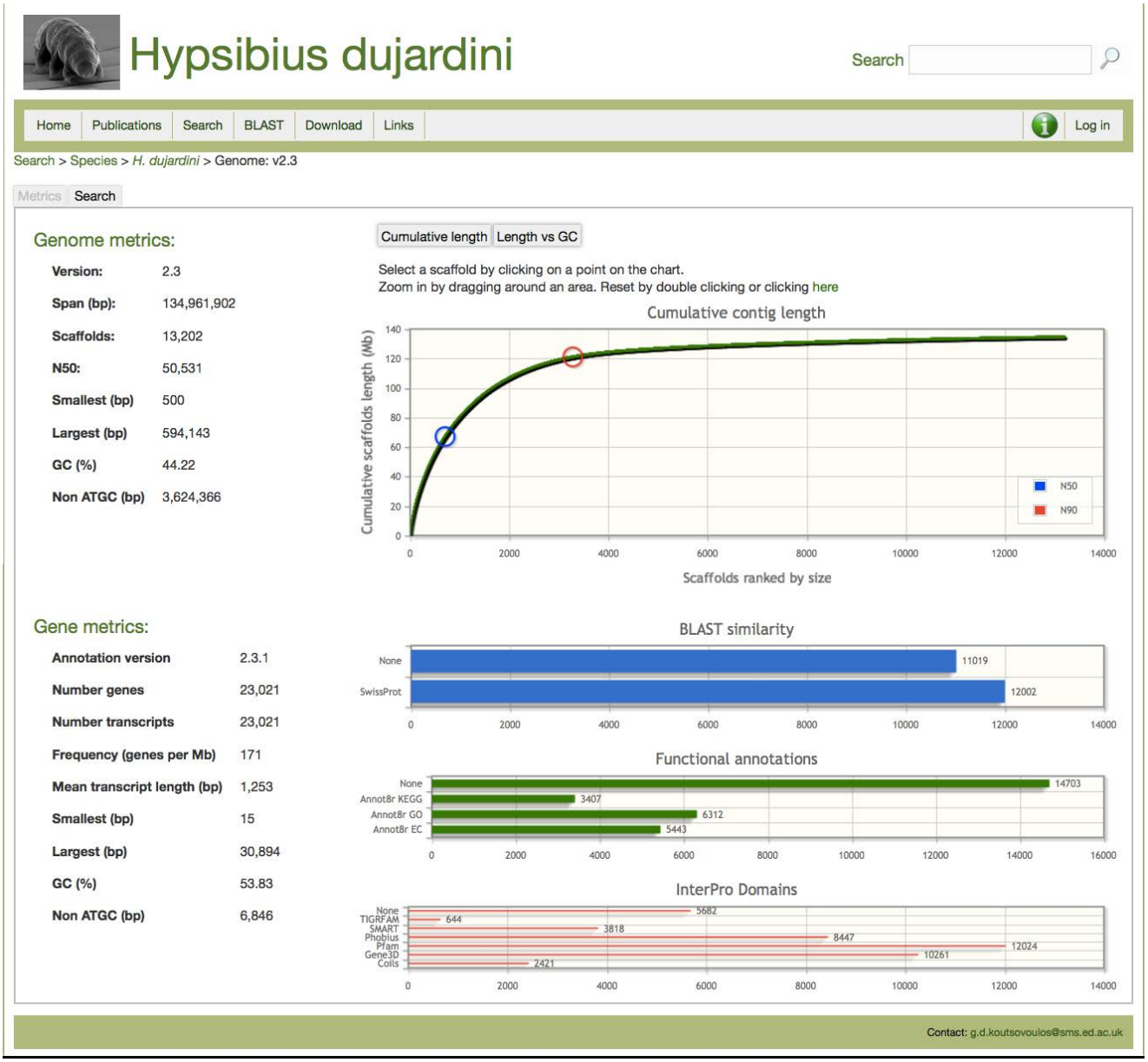
A



B



Supplemental Figure 3: The BADGER genome exploration environment for *H. dujardini*



Supplemental Figure 4: Mapping of UNC raw read data to the UNC assembly of *H. dujardini*