The genome of the tardigrade *Hypsibius dujardini*

Georgios Koutsovoulos[1], Sujai Kumar[1], Dominik R. Laetsch[1,2], Lewis Stevens[1], Jennifer Daub[1], Claire Conlon[1], Habib Maroon[1], Fran Thomas[1], Aziz Aboobaker[3] and Mark Blaxter[1]*

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

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

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

* corresponding author: mark.blaxter@ed.ac.uk

Keywords

tardigrade genome blobplots contamination Ecdysozoa

Abstract

Tardigrades are meiofaunal ecdysozoans and are key to understanding the origins of Arthropoda. We present the genome of the tardigrade *Hypsibius dujardini*, assembled from Illumina paired and mate-pair data. While the raw data indicated extensive contamination with bacteria, presumably from the gut or surface of the animals, careful cleaning generated a clean tardigrade dataset for assembly. We also generated an expressed sequence tag dataset, a Sanger genome survey dataset and used these and Illumina RNA-Seq data for assembly validation and gene prediction. The genome assembly is ~135 Mb in span, has an N50 length of over 50 kb, and an N90 length of 6 kb. We predict 23,021 protein-coding genes in the genome, which is available in a dedicated genome browser at http://www.tardigrades.org. We compare our assembly to a recently published one for the same species and do not find support for massive horizontal gene transfer. Additional analyses of the genome are ongoing.

Introduction

Tardigrades are a rather neglected phylum of endearing, microscopic animals [1]. They are members of the superphylum Ecdysozoa [2], and moult during both pre-adult and adult growth. They are part of the Panarthropoda, and current thinking places them as a sister phylum to Onychophora (velvet worms) and Arthropoda [3,4]. They, like onychophorans, have lobopod limbs, but all species have four pairs. There are about 800 described species of tardigrade [1]. All are small (tardigrades are usually classified in the meiofauna) and are found in sediments and on vegetation from the Antarctic to the Arctic, and from mountain ranges to the deep sea. Their wide dispersal in terrestrial habitats may be associated with the ability of many (but not all) species to enter environmentally resistant stasis, where the tardigrade can lose almost all body water, and thus resist extremes of temperature, pressure and desiccation. Research interest in tardigrades ranges from their utility as environmental and biogeographic marker taxa, the insight their cryptobiotic mechanisms may yield for biotechnology, to exploration of their development compared to other Ecdysozoa, especially the well-studied Nematoda and Arthropoda.

Hypsibius dujardini (Doyère, 1840) is a limnetic tardigrade that is an emerging model for evolutionary developmental biology [4-14]. It is easily cultured in the laboratory, is largely see-through (aiding analyses of development and anatomy; Figure 1), and has a rapid life cycle. *H. dujardini* is a parthenogen, and so is intractable for traditional genetic analysis, though reverse-genetic approaches are being developed [10]. We and others have been using *H. dujardini* as a genomic study system, revealing the pattern of ecdysozoan phylogeny [3,4] and the evolution of small RNA pathways [15]. *H. dujardini* is not known to be cryptobiotic, but serves as a useful comparator for tardigrades that have evolved this fascinating physiology.

Here we describe the genome of *H. dujardini* and present our assessment of the completeness and credibility of our assembly compared to another recently published version [6]. Additional analyses are ongoing, and will be added to this bioarxiv manuscript as they are completed.

Materials and Methods

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 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 genome size controls, following published protocols [16].

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 (see Table 1 for software used, version numbers and additional commands), and 5235 sequences over 100 bases in length submitted to dbEST. 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 trace2dbest as above, and 1063 submitted to dbGSS. 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 size 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 available in the ENA under study accession PRJEB11910 (runs ERR1147177 and ERR1147178). We were granted early access to Illumina GAIIX RNA-Seq data from Itai Yanai in advance of publication. Lars Hering granted access to assemblies of the RNA-Seq data generated for their analyses of *H. dujardini* opsin genes [5].

Data validation and filtering for genome assembly

We performed initial quality control on our raw Illumina data using fastqc, and addressing any issues flagged. Low quality and adapter sequence was removed using trimmomatic. 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 (available from https://github.com/DRL/blobtools/blob/master/README.md) (Dominik Laetsch, unpublished). [On a historical note, the first ever blobplot was drawn to explore contamination in our initial GAIIX data (Supplemental File 4).] The paired-end reads were normalised with one-pass khmer and were assembled with Velvet using a k-mer size of 55, and non-normalised reads mapped back to this assembly using the CLC mapper. For each scaffold, the GC% was counted (ignoring N base calls) and the read coverage calculated. Each scaffold was compared to the NCBI nucleotide database and to UNIREF90 using BLAST, and the results were filtered by the blobtools script to annotate each scaffold with the taxonomy of the highest scoring match in these databases. 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 results. Using the blobplot we identified likely contaminant reads, and removed these (and their pairs) from the quality- and adapter-trimmed data. The assembly from the filtered and cleaned data was also screened using blobplots. We also used blobplots to explore the coverage in our data of the recently published *H. dujardini* UNC assembly [6]. The new blobtools program is being prepared for publication. The initial Velvet assembly was also 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. The mate pair library was used to scaffold the gap-filled assembly with SSPACE, 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, and by mapping EST, GSS and RNA-Seq data.

Genome annotation

We annotated the assembled *H. dujardini* genome nHd.2.3 using a two-pass approach. We used MAKER 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. Protein sequences were annotated using BLAST searches against UNIREF90 and the NCBI nonredundant protein database. Protein domains and motifs were predicted with InterProScan. The genome sequence and annotations were loaded into an instance of BADGER [17] and made publicly available in late 2014. The genome assembly, predicted transcriptome, predicted proteome and GFF file of annotations are available for download on the http://www.tardigrades.org website.

Comparison of *H. dujardini* genome assemblies

We compared the University of North Carolina *H. dujardinii* assembly (UNC) [6], 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 mappeds both our read data, the Yanai RNA-Seq data and the UNC TG-300 library raw read data (from http://weatherby.genetics.utah.edu/seq_transf/, 01 December 2015) to the UNC assembly. The resulting read mapping files, together with the results of a Diamond search against UniRef90 and megablast search against NCBI nt were used to compute blobplots of the UNC assembly.

The genome of *H. dujardini*

We sought to sequence the genome of *H. dujardini* to assist the growing community of researchers using this tardigrade model, and to better place Tardigrada in metazoan phylogeny. We estimated the genome of *H. dujardini* to be ~110 Mb by propidium iodide flow cytometry. This is of the same order of magnitude as a measure published previously [13]. Other tardigrade genomes have been estimated at 40 Mb to 800 Mb (http://www.genomesize.com/; [18]).

Genomic sequencing of small target organisms that cannot be grown axenically differs from projects focused on larger species (where careful dissection can yield contamination-free, single-species samples), or from species where fully axenic samples (such as cell cultures) are available. In these small organisms it is necessary to pool many individuals, and thus also pool their associated microbiota. This associated microbiota will include gut as well as adherent and infectious organisms. Adult *H. dujardini* have only ~10^3 cells, and thus only a very small mass of bacteria would be required to yield equivalent representation of their genomes. Despite careful cleaning, genomic DNA samples prepared for sequencing of *H. dujardini* were contaminated with other taxa: bacteria and algal food.

Contaminants can negatively affect assembly in a number of ways. They generate contigs or scaffolds that do not derive from the target genome, and can compromise downstream analyses. Because the contaminants are unlikely to be at the same stoichiometry as the target genome, assemblers that try to optimise assembly by tracing paths through the De Bruijn graph based on expected coverage may be misled. These contaminants can also result in chimaeric contigs in the assembly that contain contaminant and target genome in apparent physical linkage. Our experience is that cleaned datasets also result in better assemblies (as judged by numerical scores such as N50 length), but that care must be taken not to accidentally eliminate real target genome data (for example that resulting from horizontal gene transfer (HGT)). We used taxon-annotated CG-coverage plots (TAGC or blobplots) to pre-screen raw and assembled data for contaminants [19,20]. On the blobplots of the raw (trimmed and adapter cleaned) data we identified at least five distinct sets ("blobs") of likely contaminant data, deriving from a variety of Bacteria (Figure 2). These sets of contigs were identified, and reads mapping to them identified. These reads and their pairs (if they did not have a conflicting taxonomic assignment) were removed from the dataset. We identified minimal contamination with *C. reinhardtii*, the food source, and this was removed using the *C. reinhardtii* reference genome. Further rounds of assembly and blobplot analysis revealed (as is usual) a small number of new likely contaminant contigs (which had assembled from the previously unassembled reads and unannotated contigs). These were also removed. We further removed very small contigs and scaffolds (those below 500 bp). The resultant assembly, nHd.2.3.abv500, may still contain contaminant data (see below) but is largely coherent with respect to coverage, GC% and taxonomic identity of best BLAST matches.

Our assembly spans 135 Mb (Table 3). We validated the assembly using biological and numerical criteria. The assembly had a good N50 length (>50 kb). Issues with the mate pair library insert distribution may have resulted in mis-scaffolding in some areas, but our cautious use of these data we hope has reduced this possibility. The assembly had good representation of core conserved eukaryotic genes (CEGMA) and of our EST and GSS data. The vast majority of the RNA-Seq data mapped credibly to the genome.

We annotated protein-coding genes on the *H. dujardini* nHd.2.3 assembly, and produced a higher-confidence Augustus-predicted set of 23,021 proteins. We note that this number may be inflated because of the fragmentation of our assembly. For example, only 20,370 of the proteins were predicted to have a start methionine. Many of the 2,651 proteins lacking methionine were on short scaffolds, are themselves short, and may be either fragments or mispredictions. We are working to improve the assembly and gene predictions.

The assembly of the *H. dujardini* genome was not a simple task, and the nHd.2.3 assembly is likely to still contain bacterial contaminants despite our data filtering efforts. We identified 657 scaffolds spanning 9.3 Mb where the sum of best BLAST or Diamond matches across all the genes on the scaffold suggested attribution to bacterial or fungal source genomes (Supplementary file 5). We identified ribosomal RNA sequences in the nHd.2.3 assembly using the SILVA database of curated bacterial 12S and 16S and eukaryote 18S (small subunit, SSU) and 28S (large subunit, LSU) sequences. We identified three scaffolds in the nHd.2.3 assembly that contained two instances each of the *H. dujardini* SSU and LSU (one scaffold contained both). We also identified an 11 kb scaffold that had best matches to ribosomal RNA sequences from bodonid kinetoplastid protozoa. We additionally screened the predicted protein set for sequences with higher similarity to bodonid proteins than to proteins from Ecdysozoa (using *C. elegans* and *D. melanogaster* as representatives) and identified two additional small scaffolds, both encoding putatively kinetoplastid-derived multicopy genes. In comparisons to the SILVA bacterial 12S and 16S databases, no scaffolds with matches were identified in nHd.2.3. Thus even with our stringent cleaning it is possible that contaminating scaffolds are still present. As we are currently reassembling *H. dujardini* with improved assembly algorithms, we have currently simply flagged these likely contaminant contigs in the BADGER genome explorer. Putatively contaminant scaffolds will be investigated fully as part of the next cycle of assembly optimisation

The genome is available to browse and download (including raw data, filtered subsets thereof, and intermediate analysis files) on a dedicated BADGER genome exploration environment [17] server at http://www.tardigrades.org (Figure 3). Because of our delay in preparing a formal publication we released the data publicly on this server in April 2014.

Comparing genome assemblies

Boothby *et al* [6] recently published a genome assembly for *H. dujardini* (here called the UNC assembly) based on a subculture of the same Sciento stock as our tardigrades. Surprisingly, the genome is very different in span and gene content (Table 3). We thus explored the differences between our nHd.2.3 assembly and the Boothby *et al.* one to identify likely reasons for the discrepancies. The UNC *H. dujardini* genome was reported to contain a surprisingly high proportion (17%) of putatively horizontally transferred protein-coding genes, and so we also compared the putative HGT gene set predictions with ours.

We compared GC-coverage plots for the UNC raw data and our raw data (including contaminant-derived reads) to the UNC assembly (made available in advance of NCBI GenBank release by the Goldstein laboratory, from http://weatherby.genetics.utah.edu/seq_transf/) and generated standard blobplots with blobtools (Figure 4A, B). Our raw data were derived from an independent subculture of the Sciento stock of *H. dujardini* also used by the UNC team. In both blobplots (Edinburgh, Figure 4A and UNC, Figure 4B), a large blob at ~90 fold coverage and GC% of ~45% likely corresponded to the *H. dujardini* genome: the best BLAST matches were to existing *H. dujardini* sequences (ESTs and GSS from our laboratory), arthropods and other Metazoa. The preponderance of Chordata hits (spanning 9 Mb) are likely to be derive from transposon and retrotransposon entries from species in this phylum as well as marginal similarities that, possibly interestingly, tend to match chordate, mollusc and annelid proteins better than those from arthropods. The high-coverage scaffolds (200-

300 fold, and 2000-3000 fold) that match existing *H. dujardini* sequences, corresponded to the mitochondrion [4] and ribosomal RNAs, as would be expected.

A significant proportion of the UNC scaffolds had zero or very low coverage of reads mapped in both Edinburgh and UNC raw data. These scaffolds contain matches to Bacteriodetes, spanning 27.7 Mb, that included most of the largest scaffolds in the UNC assembly. Bacterial genomes in low-complexity metagenomic datasets often assemble with greater contiguity than does the target metazoan genome, because bacterial DNA usually has higher per-base complexity (i.e. does not contain so many repeats). A second group of bacterial contigs, that appeared to derive from Proteobacteria, had a wide dispersion of coverage, from ~10 fold higher than the *H. dujardini* nuclear mean to zero. Again these are likely to derive from one or more genomes (they span 21.5 Mb). Most of the Proteobacteria scaffolds all had distinct GC%, and grouped separately from the true *H. dujardini* scaffolds. Comparing coverage between read sets, it is striking that many of the putatively bacterial scaffolds had zero coverage in both UNC and Edinburgh data. We presume that these scaffolds were assembled from UNC data from other libraries (we have not yet screened their 500 and 800 base libraries) containing additional contaminants. Scaffolds with some coverage in the UNC 300 data often had zero coverage in Edinburgh data. The wide spread of proteobacterial scaffolds suggests some sharing of contaminants between the UNC and Edinburgh cultures, but it is likely that these are different taxa, as the coverages vary widely between datasets. It is thus unlikely that these are common symbionts.

We identified seven scaffolds with matches to *H. dujardini* SSU and LSU (with four containing both subunits; Table 4) in the UNC assembly. We also identified two very similar ~20 kb scaffolds (scaffold2445_size21317 and scaffold2691_size20337) that both contained two tandemly repeated copies of the ribosomal cistron of a bdelloid rotifer closely related to *Adineta vaga*, for which a genome sequence is available. All the top 50 matches to these scaffolds in a megablast search of NCBI nr were to bdelloid rotifer SSU and LSU sequences. We screened the UNC genome for additional matches to the *A. vaga* genome, and found many, but given that *A. vaga* is robustly reported to contain a large proportion of bacterially-derived HGT genes we treated these matches with caution. However a total of 0.5 Mb of scaffolds had best sum matches to Rotifera rather than to any bacterial source (Supplementary file 5). Rotifers are commonly observed in the initial multi-xenic cultures supplied by Sciento. In the UNC assembly, we identified 15 scaffolds with robust matches to 12S and 16S genes from bacteria (Table 4). The scaffolds matched Armatimonadetes, Bacteroidetes, Chloroflexi, Planctomycetes, Proteobacteria and Verrucomicrobia, as was expected from the blobplot analyses.

We also mapped embryo RNA-Seq read data (Yanai *et al.*, in press) to the assemblies. As these RNA-Seq data were derived *via* poly(A) selection, transcriptional evidence would be strong evidence of eukaryotic transcription. Very few of the UNC scaffolds that had low or no read coverage in our raw genome data had any RNA-Seq reads mapped (Figure 4B). Those that did, had very low levels of mapping. Comparison to the same plot for the nHd.2.3 assembly showed that the pattern of high, low and no RNA-Seq expression scaffolds observed in the area we attributed to the tardigrade genome in the UNC assembly blobplots was reflected in the nHd2.3 blobplots. The RNA-Seq data thus give no support to gene expression from the low coverage, bacterial-genome like contigs in the UNC assembly.

These analyses thus positively identified much of the UNC assembly as likely derived from contaminant bacteria not represented in the sample sequenced in Edinburgh (Supplementary file 5). The total of the UNC assembly that may be attributable to contamination is ~78 Mb, or ~30% of the assembly span. However some of this span may be truly resident in the *H. dujardini* and derive from HGT.

No evidence of massive horizontal gene transfer in *H. dujardini*

Long known to be very important in prokaryotic taxa, horizontal gene transfer (HGT) has emerged as an exciting new component of the evolutionary trajectories of the genomes of animals. The UNC genome draft for *H. dujardini* was reported to contain a surprising ~17% of putative HGT genes [6]. The "foreign" origin of a subset of these loci was confirmed by phylogenetic analyses, and they were reported to have similar GC% and codon usage as did loci with clear metazoan affinities. A subset was analysed by PCR to confirm linkage to "resident" genes.

HGT can potentially bring to a recipient genome an array of new biochemical capacities, and contrasts with gradualist evolution of endogenous genes to new function. Surveys of published genomes have revealed many cases of HGT [21], including several where the new genes confer important new functions on the host. For example, tylenchomorph plant parasitic nematodes carry and deploy a suite of plant cell wall degrading enzymes and other effectors that they have acquired from bacterial and fungal sources [22-24]. These effectors are intimately involved in the parasitic biology of these species. However, claims of functional HGT must be carefully backed up by several lines of evidence [25]. Animal genomes can accrete horizontally transferred DNA from a range of sources, especially symbionts that travel with the germline [26], but the majority of these transfers are non-functional, with the DNA fragments "dead on arrival", and subsequently evolving neutrally. The common nuclear insertions of mitochondrial genes are one example of this kind of HGT, but other examples from a range of bacteria, especially *Wolbachia*, are well established [26,27]. While it is theoretically possible that noncoding HGT fragments may still affect host genome regulation, examples of "functional noncoding HGT" are largely lacking (and one high-profile published example [26] is a laboratory artefact).

Biological incorporation of a bacterial gene in an animal genome requires a series of adaptations to the new transcriptional environment [25]. The gene must acquire transcriptional regulatory signals, or be inserted in-frame in an exon of a host gene. If inserted independent of a host gene, the HGT gene must acquire spliceosomal introns, as intron-free genes are rare and specialised in animal genomes, as introns in pre-mRNAs direct these for processing and nuclear export. Adding to the potential for confusion, given a prokaryotic sequence to process (and especially dead on arrival ones that have acquired disabling mutations) eukaryotic gene finding algorithms will frequently predict apparently spliced gene models just because the algorithm has a strong prior to expect introns [27]. Once resident, the HGT gene will acclimatise to both the codon usage of the new host transcriptome, and the overall GC-bias of the host genome. Thus candidate HGT fragments in an assembly that have distinct GC% and codon usage and lack genes with spliceosomal introns should be regarded with suspicion. In the end, it is necessary to acquire additional evidence of the "foreign" origin of putative HGT genes. This evidence can include linkage to other, known host-genome-resident genes, ecological or phylogenetic perdurance (presence in all, or many individuals of a species, and presence in related taxa), phylogenetic proof of foreignness, and evidence of active transcription (for example in mRNA sequencing data). It is unusual for contaminants to be as the same genome stoichiometry as the target genome, and thus genome read coverage can also identify potential contaminant sequences [19,20].

As it is increasingly easy to generate and assemble genome sequences from target species, and it may not be possible to perform rigorous pre-extraction and post-sequencing cleaning of samples and data, it is possible to coassemble a target genome and an array of contaminants. The use of short-read data and De Bruijn graph assemblers then can result in assembly artefacts that include target and contaminant sequence. Claims of HGT thus need to be tempered by careful cleaning - treating new genome assembly projects as, initially, low complexity metagenome projects - and also careful post assembly validation [19,20].

The scaffolds identified as likely bacterial contaminants in the UNC assembly likely encompass most of the "HGT" candidates that had been identified in the UNC assembly [6]. It was not possible to confirm this as the genome annotation files for the UNC assembly had not been released at the time of writing. We note that these analyses constitute sensitive tests of two of the expected characteristics of true HGT sequences: perdurance and physical linkage to sequences clearly identifiable as host. Their absence from our raw data at the least shows that the hypothetical UNC "HGT" scaffolds are not found in all animals subcultured from the Sciento stock, and that if they are "HGT" the transfer event happened since the separation of the two cultures (less than fifteen years). The dominance of the bacterial matches over those from other (metazoan) taxa and the overall low coverage also suggests that these contigs are largely not coassemblies including bacterial-like and metazoan-like components. The RNA-Seq mapping data also fail to support expression of genes on UNC bacterial-genome-like scaffolds as eukaryotic poly(A) mRNAs (Figure 4C).

We note that these analyses were performed by comparing the Edinburgh raw genomic data to the UNC assembly, as the UNC raw reads were not yet available. We were also unable to confirm directly expression of the UNC genome HGT candidates because we did not have genome coordinates for the gene predictions. The analyses will be repeated when the UNC raw data and genome annotation files do become available and will clarify the HGT *versus* contaminant status of the scaffolds that match bacterial sequences.

Low levels of horizontal gene transfer in *H. dujardini*

As part of our analysis of the nHd.2.3 assembly we screened the 23,021 protein coding genes for signatures of HGT. We identified 496 proteins that had best matches only to bacterial, and not eukaryote, proteins. Of these, 313 were predicted to contain a spliceosomal intron, and 140 were on scaffolds that also had at least one other protein prediction that was unequivocally of eukaryote origin. We mapped the Yanai RNA-Seq data to the nHdu.2.3 genome, and extracted read counts for all predicted genes. Many of the 496 bacterial-like genes had very few or no reads mapping (less than 10 reads out of 65 million reads mapped). Only 36 bacterial-like genes had more than 1000 reads mapping (ie ~0.0015% of total reads mapped) and these are the best candidates for functional HGT (Table 5).

We compared our list of putative HGT candidates with the scaffold regions identified by Boothby *et al.* in their PCR confirmation of HGT events [6]. Many of the PCR confirmations merely showed the linkage of two genes likely to be of bacterial origin. The data presented above suggests that these "bacterial scaffolds" are likely to be derived from contaminant organisms. Several putative "archaeal – bacterial" gene pairs that were tested by PCR were identified in our analyses as also being bacterial – bacterial pairs, and placed in the same class. Thus of the 107 PCR-tested HGT events, we considered 57 to be of real interest. We confirmed the presence in our assembly of 54 of this subset. Ten of these UNC candidates map to seven unique nHd.2.3 HGT candidates that had high levels of RNA-Seq read mapping (Table 5)

Discussion

We have generated a good quality, first-draft genome for the model tardigrade *H. dujardini*. The assembly has good numerical and biological credibility scores. We have identified areas for improvement of our assembly, particularly with respect to removal of contaminant-derived sequences. Our analyses of gene content and the phylogenetic position of *H. dujardini* and by inference Tardigrada, are at an early stage, but are already yielding useful insights. *H. dujardini*, as a limnetic species, does not readily enter cryptobiosis, and thus analysis of its genome for functions associated with this striking phenomenon, often erroneously associated with all tardigrades, await sequencing of species that do. By comparing data from cryptobiotic and non-cryptobiotic tardigrades, we will be better able to identify the loci expressed and functioning in freeze-tolerance, anhydrobiosis and other resistance phenotypes.

The *H. dujardini* EST data have already been used by others in deep phylogeny analyses that place Tardigrada in Panarthropoda [3]. These analyses, and our own based on the genome sequence, do however find that *H. dujardini* sequences have an "affinity" for Nematoda, and a *H. dujardini*-plus-Nematoda clade is readily found. Whether this is because of long-branch attraction, or reflects poor resolution in the base of Ecdysozoa because of real complexity in this part of the tree or short time periods between divergence events is not clear. The *H. dujardini* mitochondrial genome was isolated and fully sequenced based on the EST data and phylogenetic analysis along with onychophoran and diverse arthropod mitochondrial genomes gave support for Panarthropoda [4]. A P2X receptor identified in the ESTs was shown to have an intriguing, unique mix of electrophysiological properties [9]. The presence of this ancient class of ligand gated ion channels in a tardigrade implies that it has been lost independently in nematodes and arthropods. The ESTs were briefly summarised elsewhere [13]. A study of the evolution of opsin loci in *H. dujardini* compared sequences derived from RNA-Seq transcriptomics to the nHd.2.3 assembly, and identified all the target genes (albeit with misprediction of alternate splicing from our genome) [5].

Our assembly, and inferences from it, conflict with a recently published draft genome (UNC) [6] for what is essentially the same strain of *H. dujardini*. Our assembly, despite having superior assembly statistics, is ~120 Mb shorter than the UNC assembly. Our genome size estimate from sequence assembly is congruent with the values we obtained by direct measurement. We find 15,000 fewer protein-coding genes, and a hugely reduced impact of predicted HGT on gene content in *H. dujardini*. These HGT candidates await detailed validation. While resolution of the conflict between these assemblies awaits detailed examination based on close scrutiny of the raw UNC data, our analyses suggest that the UNC assembly is compromised by sequences that derive from bacterial contaminants, and that the expanded genome span, additional genes, and HGT candidates are likely to be artefactual.

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. LS is funded by a Baillie Gifford Studentship, University of Edinburgh. We thank Bob Duff 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, members of the Aboobaker lab for culture support, The Centre for Genomic Research for mate-pair library construction, and staff of Edinburgh Genomics (and its precursor the GenePool) for sequencing.

Author contributions

FT, JD, AA, CC and HM developed the *H. dujardini* culture system and prepared nucleic acids. AA supplied additional DNA samples for Illumina sequencing. The cDNA library was made by JD, and EST sequencing performed by JD and FT. The GSS library was made by CC, and GSS sequencing performed by CC and FT. JD, CC, FT and HM analysed the EST and GSS data. The initial genome assemblies were made by MB and SK, and the final assembly and annotation by GK. Analyses of the genome were performed by MB, GK, SK, DRL and AA. LS constructed the BADGER genome browser instance and managed data. The manuscript was written by MB with input from all authors.

Table 1 Software used

Software	***version***	***additional parameters***	***source***	***reference***
trace2dbest	3.0.1	the program was supplied with species and library name, PCR and sequencing primer names and length cutoffs	http://www.nematodes.org/bioinformatics/trace2dbEST/	[28]
fastqc	0.11.4	default	http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc	
khmer		one pass default	https://github.com/dib-lab/khmer	[29]
BLAST	2.2.31+	contingent on search	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download	[30,31]
Diamond	0.79	contingent on search	https://github.com/bbuchfink/diamond/	[32]
blobtools	0.9.4	using NCBI Taxonomy retrieved 19 October 2015	http://github.com/DRL/blobtools	(see also [20])
MAKER	2.28	default	http://www.yandell-lab.org/software/maker.html	[33]
CEGMA	2.5	default	http://korflab.ucdavis.edu/datasets/cegma/	[34]
Augustus	2.5.5	default	http://bioinf.uni-greifswald.de/augustus/downloads/	[35]
CLC assembler	3.2.2	default	http://www.clcbio.com/products/clc-assembly-cell/	
CLC mapper	3.2.2	-l 0.9 -s 0.9	http://www.clcbio.com/products/clc-assembly-cell/	
BADGER	1.0	default	https://github.com/elswob/Badger	[17]
InterProScan	5	default	https://www.ebi.ac.uk/interpro/download.html	[36]
Velvet	1.2.06	kmer size 55, -exp_cov auto -cov_cutoff auto	https://www.ebi.ac.uk/~zerbino/velvet/	[37,38]
GapFiller	1.10	default	http://www.baseclear.com/genomics/bioinformatics/basetools/gapfiller	[39,40]
SSPACE	3	accepting only information from reads mapping 2 kb from the ends of the scaffolds	http://www.baseclear.com/genomics/bioinformatics/basetools/SSPACE	[39]

Table 2 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 1 for all accession numbers
GSS	AB3730 (Sanger)	>100 b	2 kb	n/a	1063	626,204	CZ257545 to CZ258607; see Supplemental File 2 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 Supplemental File 3 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 Supplemental File 3 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 Itai Yanai, and will be released when the manuscript they are a part of is published.

n/a: not applicable

* The mate pair library had a wide mate pair insert size distribution (see Supplemental File 3), such that the median insert size is 1.1 kb (SD 1.4 kb)rather than 4 kb. This deviation from the desired insert size is due to the library containing many fragments that appear to be standard, non-mate-pair derived segments of the genome, as is common in such libraries.

Table 3 *Hypsibius dujardini* assembly comparison

Genome	***H. dujardini* Edinburgh**	***H. dujardini* UNC**
Filename	*nHd.2.3.abv500.fna*	*tg.genome.fsa*
Longest scaffold (bp)	594143	1534183
For scaffolds longer than 500 bp		
Number of scaffolds	13202	22497
Span (bp)	134961902	252538263
Minimum length (bp)	500	2000
Mean length (bp)	10222	11225
N50 length (bp)	50531	15907
Number of scaffolds in N50	701	4078
GC proportion	0.452	0.469
For scaffolds longer than 1000 bp		
Number of scaffolds	7999	22497
Span (bp)	131304830	252538263
Minimum length (bp)	1000	2000
Mean length (bp)	16415	11225
N50 length (bp)	52541	15907
Number of scaffolds in N50	666	4078
For contigs longer than 100 bp (scaffolds split at >= 10 Ns)		
Longest contig	116477	1534183
Number of contigs	25005	22972
Span (bp)	131393004	252502428
Minimum length (bp)	100	2000
Mean length (bp)	5254	10991
N50 length (bp)	11636	15542
Number of contigs in N50	3245	4197
For runs of undetermined base calls (>= 10 N)		
Number of spans	12197	475
Span (bp)	3548224	35835
N50 length (bp)	788	106
CEGMA completeness		
Complete	88.7%	89.5%
Average number of copies	1.35	3.26
Partial	97.2	94.8
Average number of copies	1.55	3.52
Genome content		
ESTs mapping to assembly *	95.9%	91.8%
GSSs mapping to assembly **	96.6%	90.9%
Number of protein-coding genes	23,021	38,145
Potential contaminant span in assembly ¶	9.3 Mb	77.5 Mb
Potential contaminant proportion	6.8%	30.4%

* of 5235 sequences; megablast search with E-value cutoff 1e-65.

** of 1063 sequences; megablast search with E-value cutoff 1e-65.

¶ assessed from blobplot analyses (Supplemental Files 5 and 6).

Table 4: 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	99.55	672	0	Eukaryota	Tardigrada	*Hypsibius dujardini*
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 5: Expressed putative HGT genes in the genome of *Hypsibius dujardini*

nHd gene name	UNC scaffold	best match in SwissProt	ID line of best match	EC number of enzymatic activity	EC functional description	GO annotation	GO functional description	KEGG annotation	KEGG functional description	InterProScan annotation
nHd.2.3.1.t10700-RA		BGLB_CLOTH	Thermostable beta-glucosidase B	3.2.1.21	Beta-glucosidase.					Coil
nHd.2.3.1.t15477-RA	scaffold634_size55776	TTC32_HUMAN	Tetratricopeptide repeat protein 32	3.1.3.16	Phosphoproteinphosphatase.					Coil
nHd.2.3.1.t15990-RA		RBP2A_PLAF7	Reticulocyte-binding protein 2 homolog a			GO:0016020	membrane			Coil
nHd.2.3.1.t08832-RA	scaffold653_size38539, scaffold1407_size27578									Coil
nHd.2.3.1.t01320-RA		Y705_DEIRA	Uncharacterized protein DR_0705					K01884	Aminoacyl-tRNA biosynthesis	CYTOPLASMIC_DOMAIN
nHd.2.3.1.t16051-RA		YJDF_ECOLI	Inner membrane protein yjdF							CYTOPLASMIC_DOMAIN
nHd.2.3.1.t09364-RA				3.1.3.2	Acidphosphatase.					CYTOPLASMIC_DOMAIN
nHd.2.3.1.t15665-RA		M2DH_ASPTN	Mannitol 2-dehydrogenase	1.1.1.67	Mannitol2-dehydrogenase.	GO:0019594	mannitol metabolic process	K00045	Fructose and mannose metabolism	G3DSA:1.10.1040.10
nHd.2.3.1.t07813-RA		SNDH_ACELI	L-sorbosone dehydrogenase							G3DSA:2.120.10.30
nHd.2.3.1.t09693-RA										G3DSA:2.120.10.30
nHd.2.3.1.t16830-RA										G3DSA:2.60.120.10
nHd.2.3.1.t15207-RA		BOLA3_BOVIN	BolA-like protein 3							G3DSA:2.90.10.10
nHd.2.3.1.t13881-RA										G3DSA:3.10.450.50
nHd.2.3.1.t15428-RA	scaffold5875_size12964	BGLB_CLOTH	Thermostable beta-glucosidase B	3.2.1.21	Beta-glucosidase.					G3DSA:3.20.20.300
nHd.2.3.1.t18868-RA		INO1_MYCTU	Inositol-3-phosphate synthase	5.5.1.4	Inositol-3-phosphatesynthase.					G3DSA:3.30.360.10
nHd.2.3.1.t18578-RA		CBS_HUMAN	Cystathionine beta-synthase	4.2.1.22	Cystathioninebeta-synthase.					G3DSA:3.40.50.1100
nHd.2.3.1.t01864-RA										G3DSA:3.40.50.150

nHd.2.3.1.t18825-RA										G3DSA:3.40.50.150
nHd.2.3.1.t10233-RA		HYIN_AGRRH	Indoleacetamide hydrolase	3.5.1.86	Mandelamideamidase.					G3DSA:3.90.1300.10
nHd.2.3.1.t10953-RA		UBIG_PSYA2	3-demethylubiquinone-9 3-methyltransferase	2.1.1.64	3-demethylubiquinol 3-O-methyltransferase.			K00568	Ubiquinone and other terpenoid-quinone biosynthesis	NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t13129-RA		Y705_DEIRA	Uncharacterized protein DR_0705					K01884	Aminoacyl-tRNA biosynthesis	NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t03012-RA		E13B_BACCI	Glucan endo-1,3-beta-glucosidase A1	3.2.1.39	Glucanendo-1,3-beta-D-glucosidase.					NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t02992-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t03161-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t08069-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t10701-RA	scaffold187_size82271									NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t11031-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t14106-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t14687-RA										NON_CYTOPLASMIC_DOMAIN
nHd.2.3.1.t05037-RA	scaffold301_size53378, scaffold2358_size21708	YL728_MIMIV	Uncharacterized protein L728							PF11443 (DUF2828)
nHd.2.3.1.t00994-RA	scaffold741_size41061, scaffold3315_size18320	CYNS_BURCM	Cyanate hydratase	4.2.1.104	Cyanase.			K01725	Nitrogen metabolism	TIGR00673 cyanate hydratase
nHd.2.3.1.t04854-RA		FBN2_HUMAN	Fibrillin-2							
nHd.2.3.1.t04323-RA										
nHd.2.3.1.t04485-RA										
nHd.2.3.1.t04660-RA	scaffold1372_size27931									
nHd.2.3.1.t04897-RA										

Figure 1: The tardigrade *Hypsibius dujardini*

A A whole, dorsal view of a living *H. dujardini* adult, talken 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) (**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.

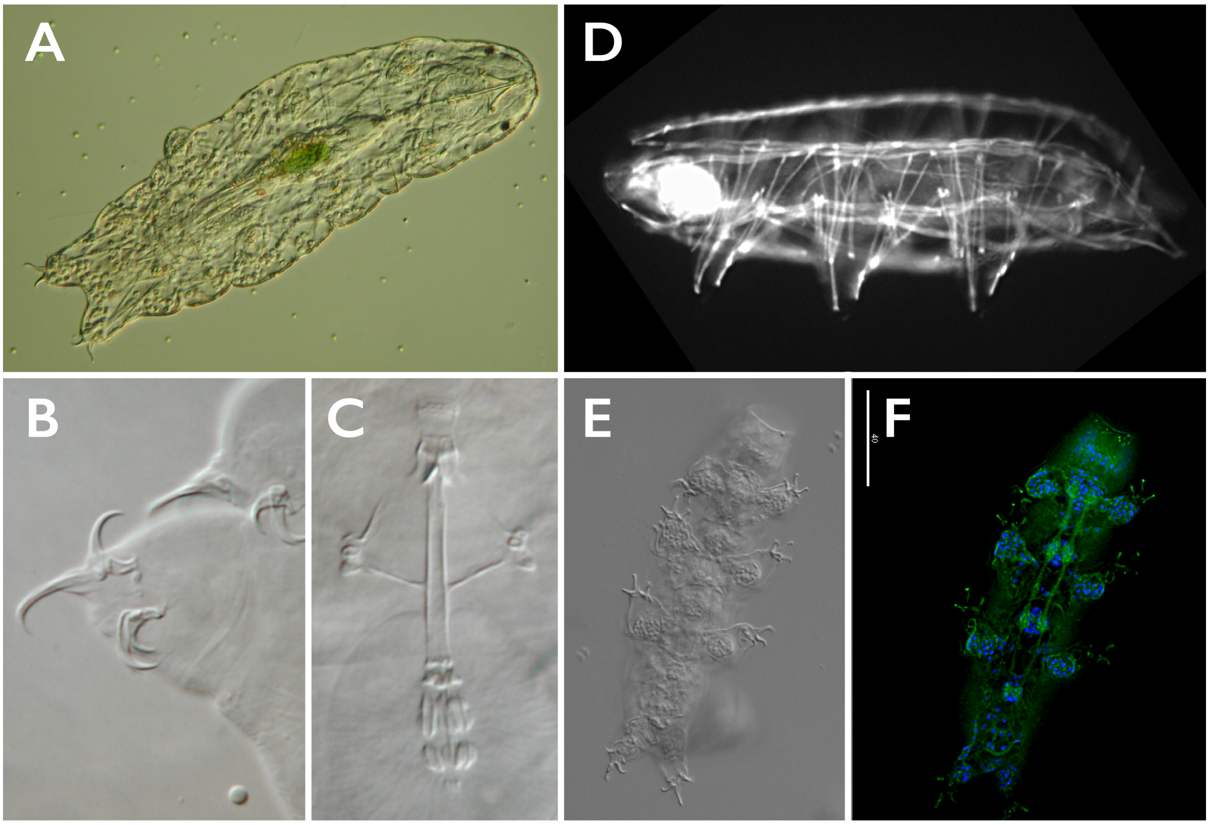
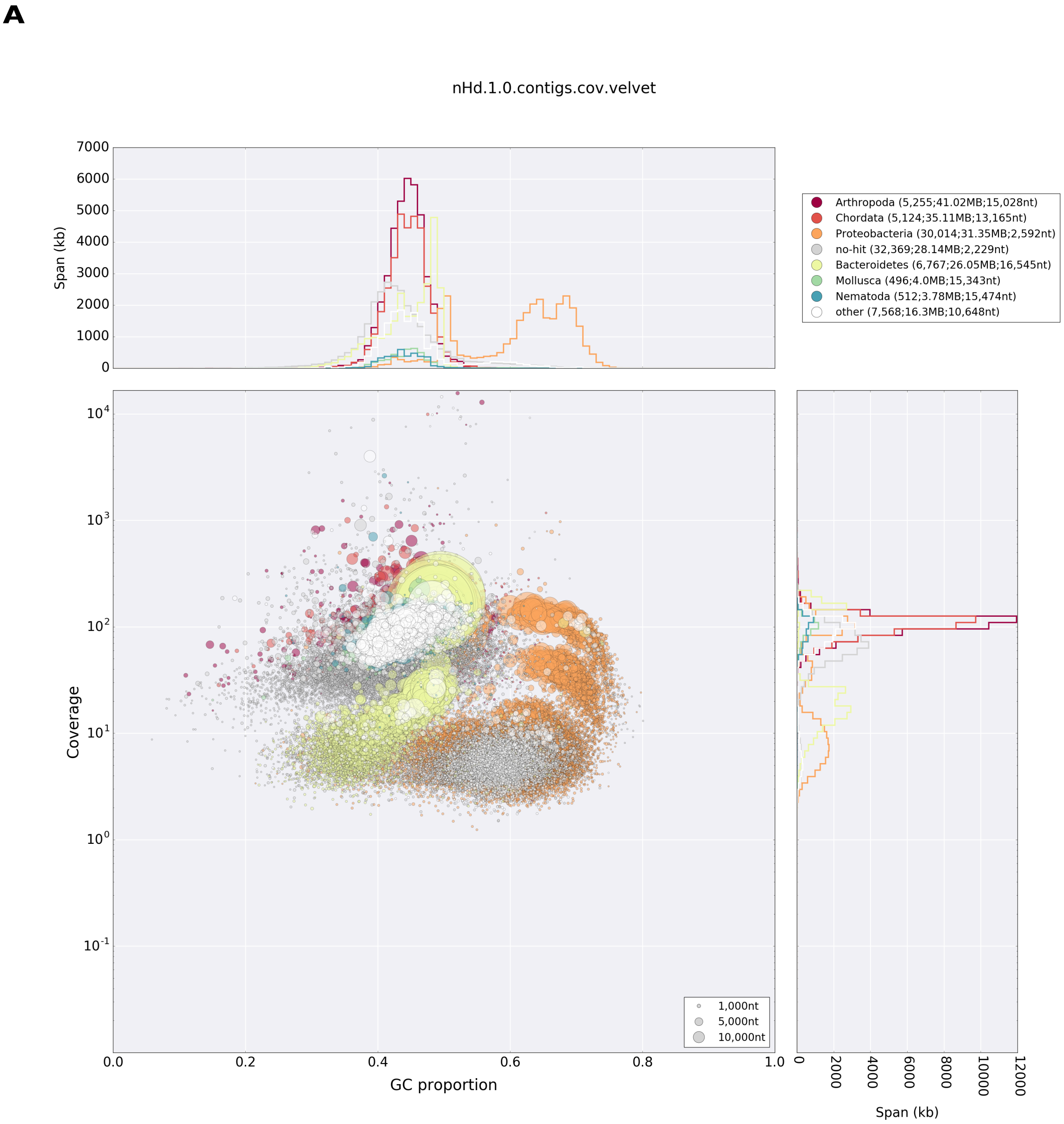

A
D
B
C
E
F
40

Figure 2: Contaminant removal and assembly validation via blobplots.

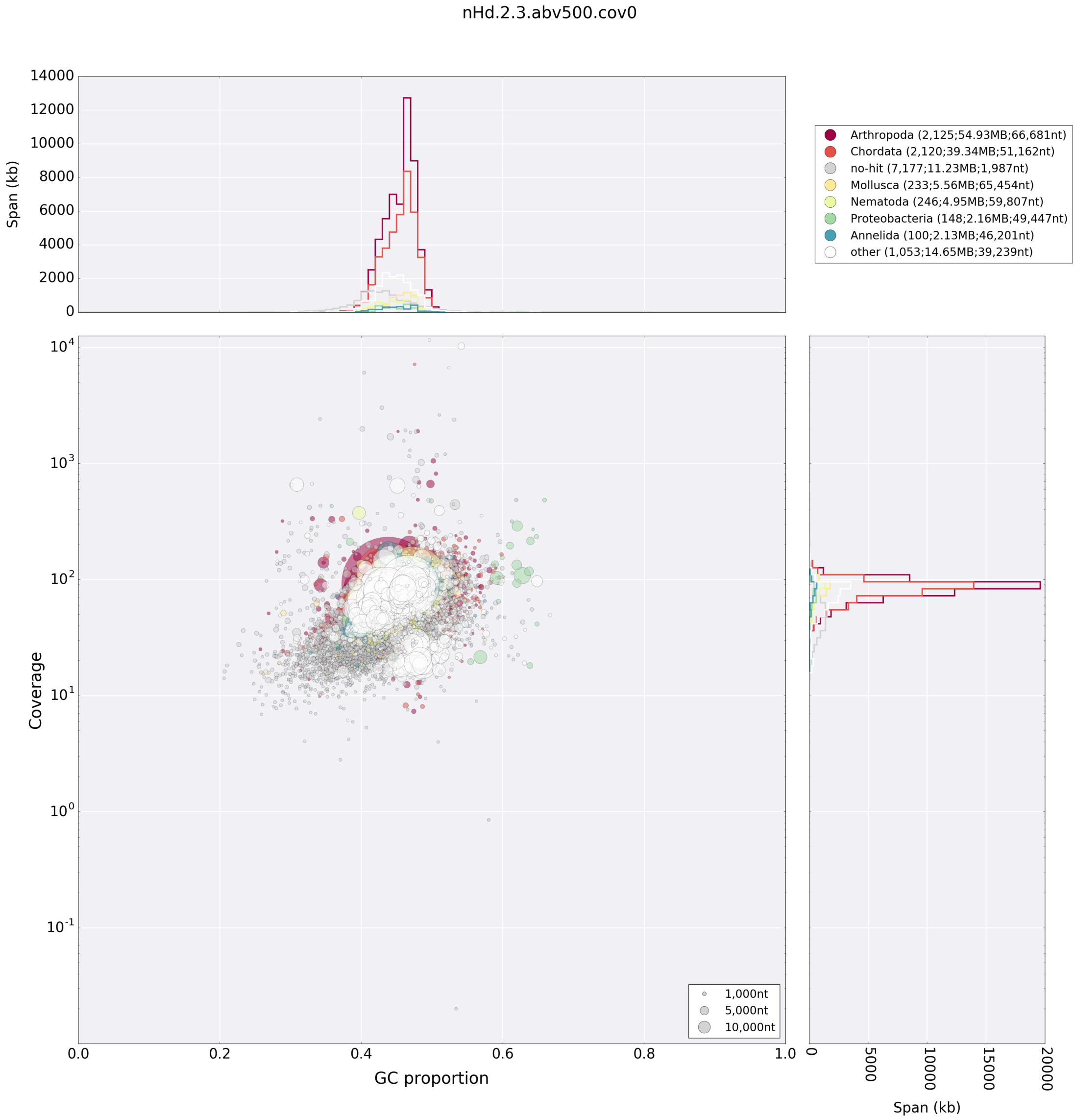
A Blobplot of the initial assembly of the trimmed raw read data, identifying significant contamination with a variety of bacterial genomes.

B Blobplot of the nHd.2.3 assembly derived from the third cycle of read cleaning. All raw reads were mapped back to this assembly. Remaining scaffolds that have sequence similarities to Bacteria rather than Metazoa have very similar coverage to the *H. dujardini* scaffolds.

C Blobplot of the nHd.2.3 assembly, with scaffold points plotted as in **B** but coloured by read coverage from mapping of RNA-Seq data. (High RNAseq cov : ≥ 10 reads/kb; Low RNAseq cov : < 10 reads/kb, No RNAseq cov : 0 reads/kb).


A
nHd.1.0.contigs.cov.velvet
Span (kb)
7000
6000
5000
4000
3000
2000
1000
0
Arthropoda (5,255;41.02MB;15,028nt)
Chordata (5,124;35.11MB;13,165nt)
Proteobacteria (30,014;31.35MB;2,592nt)
no-hit (32,369;28.14MB;2,229nt)
Bacteroidetes (6,767;26.05MB;16,545nt)
Mollusca (496;4.0MB;15,343nt)
Nematoda (512;3.78MB;15,474nt)
other (7,568;16.3MB;10,648nt)
Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
2000
4000
6000
8000
10000
12000
Span (kb)

B


nHd.2.3.abv500.cov0
Span (kb)
14000
12000
10000
8000
6000
4000
2000
0
● Arthropoda (2,125;54.93MB;66,681nt)
● Chordata (2,120;39.34MB;51,162nt)
● no-hit (7,177;11.23MB;1,987nt)
● Mollusca (233;5.56MB;65,454nt)
● Nematoda (246;4.95MB;59,807nt)
● Proteobacteria (148;2.16MB;49,447nt)
● Annelida (100;2.13MB;46,201nt)
○ other (1,053;14.65MB;39,239nt)
Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
5000
10000
15000
20000
Span (kb)

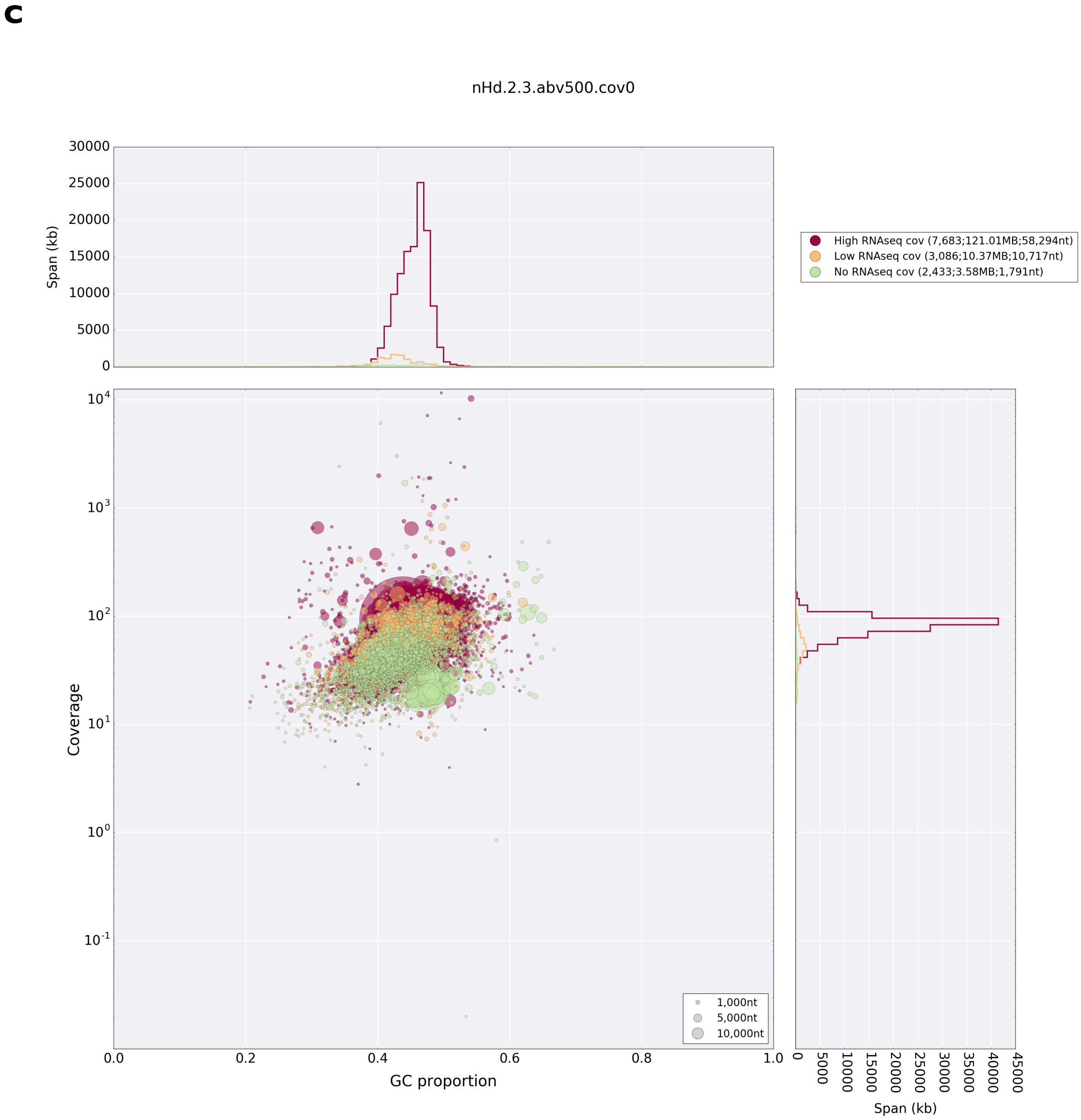
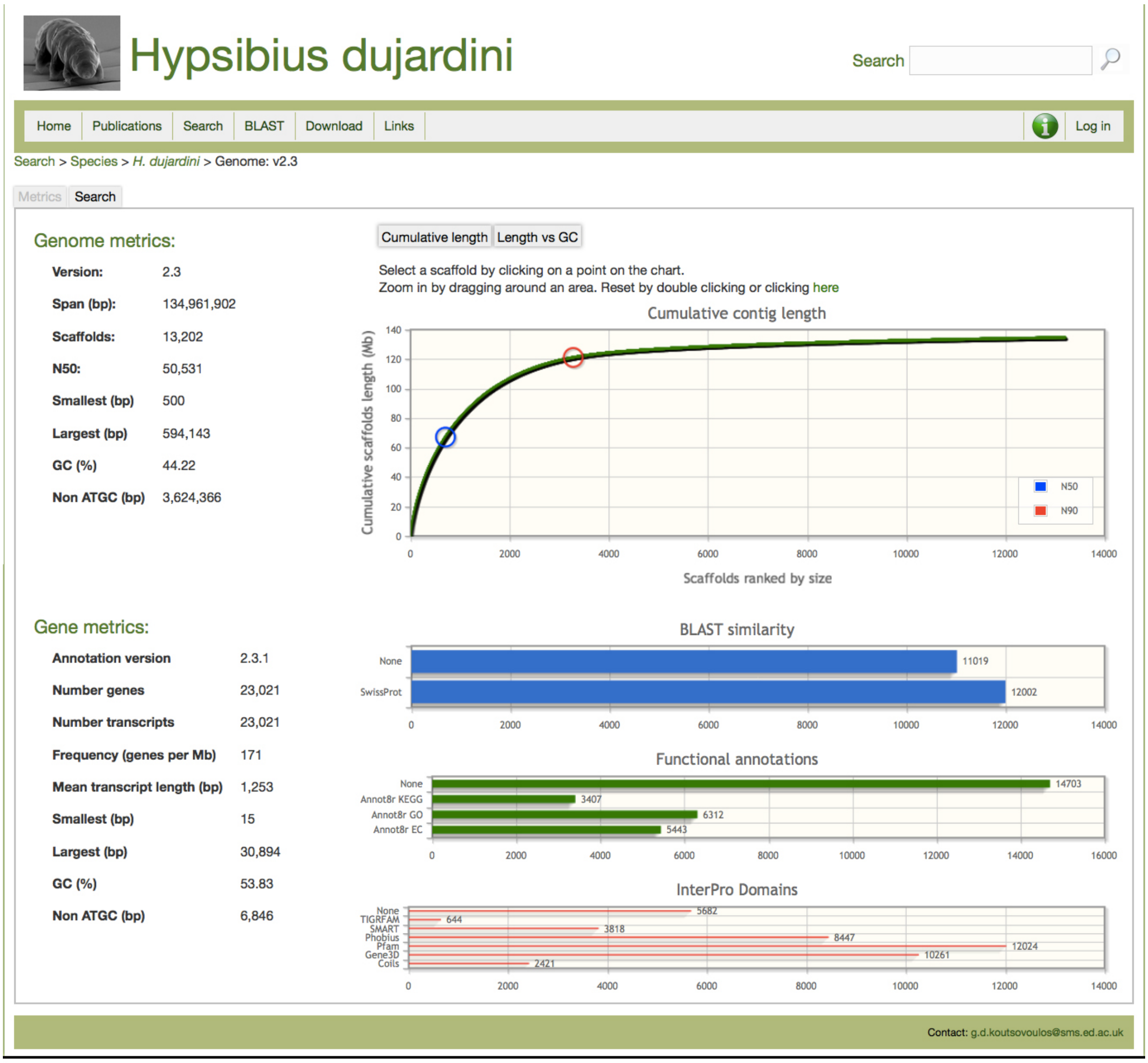

c
nHd.2.3.abv500.cov0
Span (kb)
30000
25000
20000
15000
10000
5000
0
High RNAseq cov (7,683;121.01MB;58,294nt)
Low RNAseq cov (3,086;10.37MB;10,717nt)
No RNAseq cov (2,433;3.58MB;1,791nt)
Coverage
10^4
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
Span (kb)

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.



Hypsibius dujardini

Search

Search > Species > *H. dujardini* > Genome: v2.3

Genome metrics:

Version:	2.3
Span (bp):	134,961,902
Scaffolds:	13,202
N50:	50,531
Smallest (bp)	500
Largest (bp)	594,143
GC (%)	44.22
Non ATGC (bp)	3,624,366

Select a scaffold by clicking on a point on the chart.
Zoom in by dragging around an area. Reset by double clicking or clicking here

Gene metrics:

Annotation version	2.3.1
Number genes	23,021
Number transcripts	23,021
Frequency (genes per Mb)	171
Mean transcript length (bp)	1,253
Smallest (bp)	15
Largest (bp)	30,894
GC (%)	53.83
Non ATGC (bp)	6,846

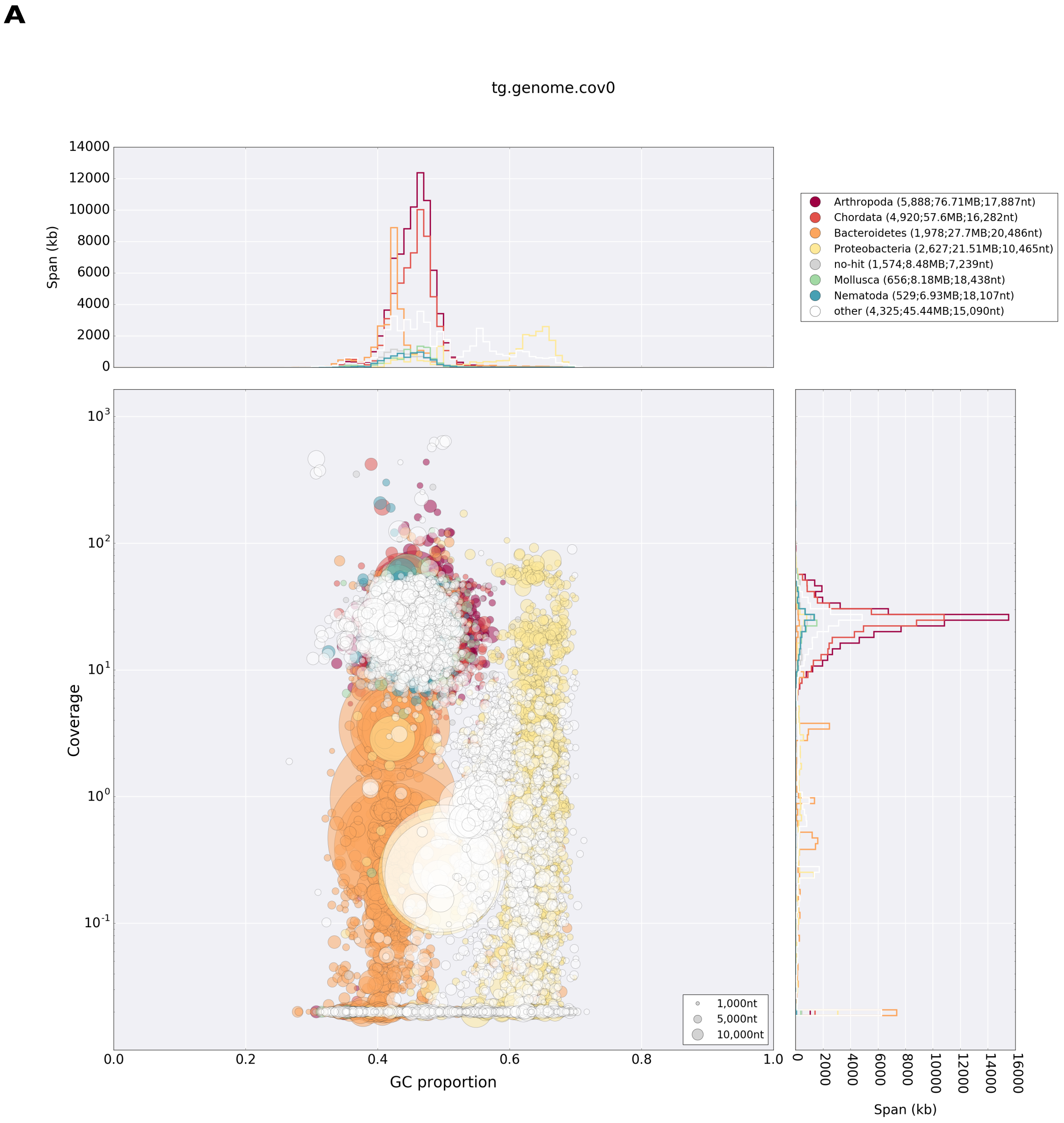
Contact: g.d.koutsovoulos@sms.ed.ac.uk

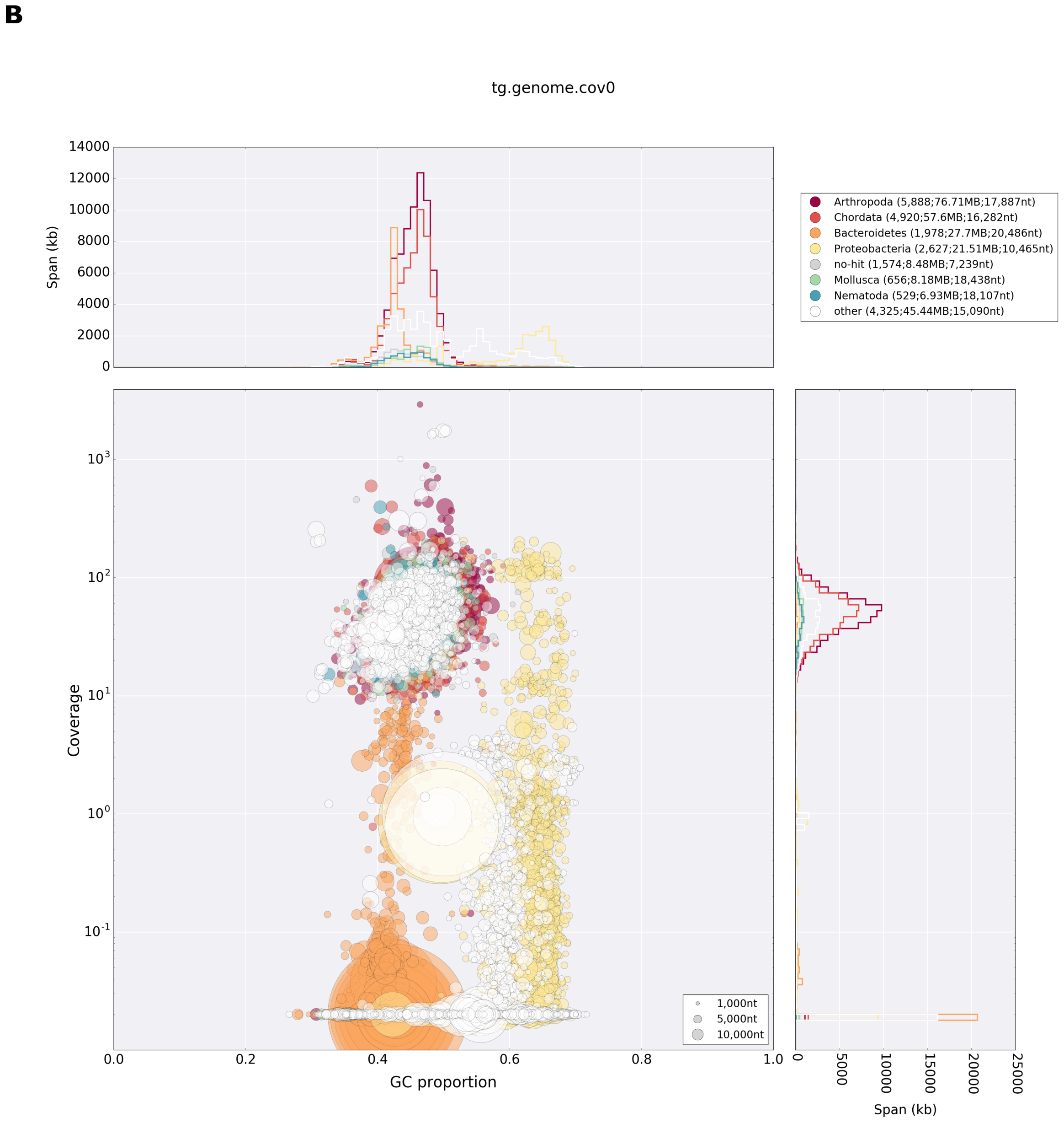
Figure 4: Mapping of read data to UNC assembly identifies non-shared contaminants and no expression from bacterial scaffolds

A Blobplot showing the UNC assembly contigs distributed by GC proportion and coverage derived from the UNC raw genomic sequence data (data file TG-300). Scaffold points are scaled by length, and coloured based on taxonomic assignment of the sum of the best BLAST and Diamond matches for all the genes on the scaffold. Taxonomic assignments are summed by phylum.

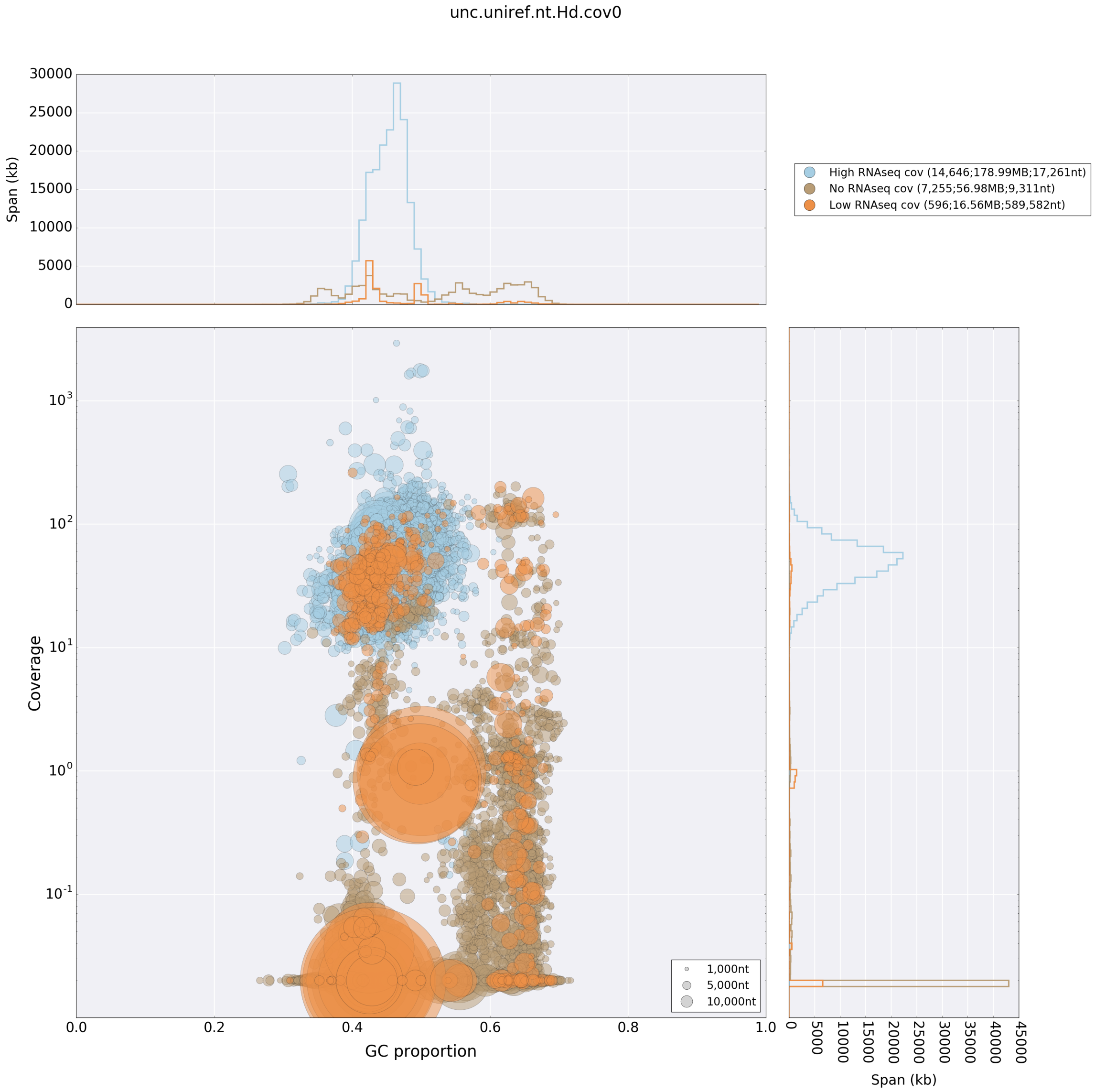
B Blobplot showing the UNC assembly contigs distributed by GC proportion and coverage derived from the Edinburgh raw genomic sequence data. Scaffold points are scaled by length, and coloured based on taxonomic assignment of the sum of the best BLAST and Diamond matches for all the genes on the scaffold. Taxonomic assignments are summed by phylum.

C Blobplot (as in **A**) with the scaffold points coloured by RNA-Seq read coverage. (High RNAseq cov : ≥ 10 reads/kb; Low RNAseq cov : < 10 reads/kb, No RNAseq cov : 0 reads/kb).


A
tg.genome.cov0
Span (kb)
14000
12000
10000
8000
6000
4000
2000
0
Arthropoda (5,888;76.71MB;17,887nt)
Chordata (4,920;57.6MB;16,282nt)
Bacteroidetes (1,978;27.7MB;20,486nt)
Proteobacteria (2,627;21.51MB;10,465nt)
no-hit (1,574;8.48MB;7,239nt)
Mollusca (656;8.18MB;18,438nt)
Nematoda (529;6.93MB;18,107nt)
other (4,325;45.44MB;15,090nt)
Coverage
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
2000
4000
6000
8000
10000
12000
14000
16000
Span (kb)


B
tg.genome.cov0
Span (kb)
14000
12000
10000
8000
6000
4000
2000
0
● Arthropoda (5,888;76.71MB;17,887nt)
● Chordata (4,920;57.6MB;16,282nt)
● Bacteroidetes (1,978;27.7MB;20,486nt)
● Proteobacteria (2,627;21.51MB;10,465nt)
● no-hit (1,574;8.48MB;7,239nt)
● Mollusca (656;8.18MB;18,438nt)
● Nematoda (529;6.93MB;18,107nt)
○ other (4,325;45.44MB;15,090nt)
Coverage
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
5000
10000
15000
20000
25000
Span (kb)

c


unc.uniref.nt.Hd.cov0
Span (kb)
30000
25000
20000
15000
10000
5000
0
High RNAseq cov (14,646;178.99MB;17,261nt)
No RNAseq cov (7,255;56.98MB;9,311nt)
Low RNAseq cov (596;16.56MB;589,582nt)
Coverage
10^3
10^2
10^1
10^0
10^{-1}
1,000nt
5,000nt
10,000nt
0.0
0.2
0.4
0.6
0.8
1.0
GC proportion
0
5000
10000
15000
20000
25000
30000
35000
40000
45000
Span (kb)

References

1 Kinchin, I. M. *The Biology of Tardigrades*. (Portland Press, 1994).

2 Garey, J. R., Krotec, M., Nelson, D. R. & Brooks, J. Molecular analysis supports a tardigrade-arthropod association. *Invertebrate Biology* **115**, 79-88 (1996).

3 Dunn, C. W. *et al.* Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745-749, doi:10.1038/nature06614 (2008).

4 Rota-Stabelli, O. *et al.* Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol Evol* **2**, 425-440, doi:evq030 [pii]

10.1093/gbe/evq030 (2010).

5 Hering, L. & Mayer, G. Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in panarthropoda. *Genome Biol Evol* **6**, 2380-2391, doi:10.1093/gbe/evu193 (2014).

6 Boothby, T. C. *et al.* Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.1510461112 (2015).

7 Gross, V. & Mayer, G. Neural development in the tardigrade Hypsibius dujardini based on anti-acetylated alpha-tubulin immunolabeling. *Evodevo* **6**, 12, doi:10.1186/s13227-015-0008-4 (2015).

8 Mayer, G., Kauschke, S., Rudiger, J. & Stevenson, P. A. Neural markers reveal a one-segmented head in tardigrades (water bears). *PLoS One* **8**, e59090, doi:10.1371/journal.pone.0059090 (2013).

9 Bavan, S., Straub, V. A., Blaxter, M. L. & Ennion, S. J. A P2X receptor from the tardigrade species Hypsibius dujardini with fast kinetics and sensitivity to zinc and copper. *BMC Evol Biol* **9**, 17, doi:1471-2148-9-17 [pii]

10.1186/1471-2148-9-17 (2009).

10 Tenlen, J. R., McCaskill, S. & Goldstein, B. RNA interference can be used to disrupt gene function in tardigrades. *Dev Genes Evol* **223**, 171-181, doi:10.1007/s00427-012-0432-6 (2013).

11 Gabriel, W. N. & Goldstein, B. Segmental expression of Pax3/7 and engrailed homologs in tardigrade development. *Dev Genes Evol* **217**, 421-433, doi:10.1007/s00427-007-0152-5 (2007).

12 Mayer, G. *et al.* Selective neuronal staining in tardigrades and onychophorans provides insights into the evolution of segmental ganglia in panarthropods. *BMC Evol Biol* **13**, 230, doi:10.1186/1471-2148-13-230 (2013).

13 Gabriel, W. N. *et al.* The tardigrade Hypsibius dujardini, a new model for studying the evolution of development. *Dev Biol* **312**, 545-559, doi:10.1016/j.ydbio.2007.09.055 (2007).

14 Goldstein, B. & Blaxter, M. Tardigrades. *Curr Biol* **12**, R475 (2002).

15 Sarkies, P. *et al.* Ancient and Novel Small RNA Pathways Compensate for the Loss of piRNAs in Multiple Independent Nematode Lineages. *PLoS Biol* **13**, e1002061, doi:10.1371/journal.pbio.1002061 (2015).

16 Hare, E. E. & Johnston, J. S. Genome size determination using flow cytometry of propidium iodide-stained nuclei. *Methods Mol Biol* **772**, 3-12, doi:10.1007/978-1-61779-228-1_1 (2011).

17 Elsworth, B., Jones, M. & Blaxter, M. Badger--an accessible genome exploration environment. *Bioinformatics*, doi:10.1093/bioinformatics/btt466 (2013).

18 Gregory, T. R. *et al.* Eukaryotic genome size databases. *Nucleic Acids Res* **35**, D332-338, doi:10.1093/nar/gkl828 (2007).

19 Kumar, S. & Blaxter, M. L. Simultaneous genome sequencing of symbionts and their hosts. *Symbiosis* **55**, 119-126, doi:10.1007/s13199-012-0154-6 (2011).

20 Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. & Blaxter, M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in Genetics* **4**, 237, doi:10.3389/fgene.2013.00237 (2013).

21 Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. & Micklem, G. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol* **16**, 50, doi:10.1186/s13059-015-0607-3 (2015).

22 Danchin, E. G. & Rosso, M. N. Lateral gene transfers have polished animal genomes: lessons from nematodes. *Frontiers in cellular and infection microbiology* **2**, 27, doi:10.3389/fcimb.2012.00027 (2012).

23 Danchin, E. G. What Nematode genomes tell us about the importance of horizontal gene transfers in the evolutionary history of animals. *Mob Genet Elements* **1**, 269-273, doi:10.4161/mge.18776 (2011).

24 Bird, D. M., Jones, J. T., Opperman, C. H., Kikuchi, T. & Danchin, E. G. Signatures of adaptation to plant parasitism in nematode genomes. *Parasitology* **142 Suppl 1**, S71-84, doi:10.1017/S0031182013002163 (2015).

25 Blaxter, M. Symbiont Genes in Host Genomes: Fragments with a Future? *Cell Host and Microbe* **2**, 211-213 (2007).

26 Dunning-Hotopp, J. C. *et al.* Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* **317**, 1753-1756, doi:10.1126/science.1142490 (2007).

27 Koutsovoulos, G., Makepeace, B., Tanya, V. N. & Blaxter, M. Palaeosymbiosis revealed by genomic fossils of Wolbachia in a strongyloidean nematode. *PLoS Genet* **10**, e1004397, doi:10.1371/journal.pgen.1004397 (2014).

28 Parkinson, J. *et al.* PartiGene - constructing partial genomes. *Bioinformatics* **20**, 1398-1404 (2004).

29 Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res* **4**, 900, doi:10.12688/f1000research.6924.1 (2015).

30 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

31 Boratyn, G. M. *et al.* BLAST: a more efficient report with usability improvements. *Nucleic Acids Research* **41**, W29-33, doi:10.1093/nar/gkt282 (2013).

32 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

33 Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, doi:10.1186/1471-2105-12-491 (2011).

34 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-1067, doi:10.1093/bioinformatics/btm071 (2007).

35 Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435-439, doi:10.1093/nar/gkl200 (2006).

36 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

37 Zerbino, D. R., McEwen, G. K., Margulies, E. H. & Birney, E. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One* **4**, e8407, doi:10.1371/journal.pone.0008407 (2009).

38 Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**, 821-829, doi:10.1101/gr.074492.107 (2008).

39 Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579, doi:10.1093/bioinformatics/btq683 (2011).

40 Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol* **13**, R56, doi:10.1186/gb-2012-13-6-r56 (2012).