Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies

Tom O Delmont, A. Murat Eren

High-throughput sequencing provides a fast and cost effective mean to recover genomes of organisms from all domains of life. However, adequate curation of the assembly results against potential contamination of non-target organisms requires advanced bioinformatics approaches and practices. Here, we re-analyzed the sequencing data generated for the tardigrade *Hypsibius dujardini* using approaches routinely employed by microbial ecologists who reconstruct bacterial and archaeal genomes from metagenomic data. We created a holistic display of the eukaryotic genome assembly using DNA data originating from two groups and eleven sequencing libraries. By using bacterial single-copy genes, k-mer frequencies, and coverage values of scaffolds we could identify and characterize multiple near-complete bacterial genomes, and curate a 182 Mbp draft genome for *H. dujardini* supported by RNA-Seq data. Our results indicate that most contaminant scaffolds were assembled from Moleculo long-read libraries, and most of these contaminants have differed between library preparations. Our re-analysis shows that visualization and curation of eukaryotic genome assemblies can benefit from tools designed to address the needs of today’s microbiologists, who are constantly challenged by the difficulties associated with the identification of distinct microbial genomes in complex environmental metagenomes.

Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies

Tom O. Delmont[1] and A. Murat Eren[1,2]

[1] Department of Medicine, The University of Chicago, Chicago, IL , United States.

[2] Josephine Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA , United States.

Corresponding Author:

A. Murat Eren[1]
Knapp Center for Biomedical Discovery,
900 E. 57th St., MB 9, RM 9118, Chicago, IL 60637 USA

Email address: meren@uchicago.edu

Abstract

High-throughput sequencing provides a fast and cost effective mean to recover genomes of organisms from all domains of life. However, adequate curation of the assembly results against potential contamination of non-target organisms requires advanced bioinformatics approaches and practices. Here, we re-analyzed the sequencing data generated for the tardigrade *Hypsibius dujardini,* and created a holistic display of the eukaryotic genome assembly using DNA data originating from two groups and eleven sequencing libraries. By using bacterial single-copy genes, k-mer frequencies, and coverage values of scaffolds we could identify and characterize multiple near-complete bacterial genomes from the raw assembly, and curate a 182 Mbp draft genome for *H. dujardini* supported by RNA-Seq data. Our results indicate that most contaminant scaffolds were assembled from Moleculo long-read libraries, and most of these contaminants have differed between library preparations. Our re-analysis shows that visualization and curation of eukaryotic genome assemblies can benefit from tools designed to address the needs of today's microbiologists, who are constantly challenged by the difficulties associated with the identification of distinct microbial genomes in complex environmental metagenomes.

Introduction

Advances in high-throughput sequencing technologies are revolutionizing the field of genomics by allowing researchers to generate large amount of data in a short period of time (Loman & Pallen, 2015). These technologies, combined with advances in computational approaches, help us understand the diversity and functioning of life at different scales by facilitating the rapid recovery of bacterial, archaeal, and eukaryotic genomes (Venter et al., 2001; Schleper, Jurgens & Jonuscheit, 2005; Brown et al., 2015). Yet, the recovery of genomes is not straightforward, and reconstructing bacterial and archaeal versus eukaryotic genomes present researchers with distinct pitfalls and challenges that result in different molecular and computational workflows.

For instance, difficulties associated with the cultivation of bacterial and archaeal organisms (Schloss & Handelsman, 2003) have persuaded microbiologists to reconstruct genomes directly from the environment through assembly-based metagenomics workflows and genome binning. This workflow commonly entails (1) whole sequencing of environmental genetic material, (2) assembly of short reads into contiguous DNA segments (contigs), and (3) identification of draft genomes by binning contigs that originate from the same organism. Due to the extensive diversity of bacteria and archaea in most environmental samples (Gans, Wolinsky & Dunbar, 2005; Rusch et al., 2007), the field of metagenomics has rapidly evolved to accurately delineate genomes in assembly results. Today, microbiologists often exploit two essential properties of bacterial and archaeal genomes to improve the "binning" step: (1) k-mer frequencies that are somewhat preserved throughout a single microbial genome (Pride et al., 2003), to identify contigs that likely originate from the same genome (Teeling et al., 2004), and (2) a set of genes that occur in the vast majority of bacterial genomes as a single copy, to estimate the level of completion and contamination of genome bins (Wu & Eisen, 2008; Campbell et al., 2013; Parks et al., 2015). These properties, along with differential coverage of contigs across multiple samples when such data exist, are routinely used to identify coherent microbial draft genomes in metagenomic assemblies (Dick et al., 2009; Albertsen et al., 2013; Wu et al., 2014; Alneberg et al., 2014; Kang et al., 2015; Eren et al., 2015).

On the other hand, researchers who study eukaryotic genomes generally focus on the recovery of a single organism, which, in most cases, simplifies the identification of the target genome in assembly results. However, sequences of bacterial origin can contaminate eukaryotic genome assembly results due to their occurrence in samples (Chapman et al., 2010; Artamonova & Mushegian, 2013), DNA extraction kits (Salter et al., 2014), or laboratory environments (Laurence, Hatzis & Brash, 2014; Strong et al., 2014). One of the major challenges of working with eukaryotic genomes is the extent of repeat regions that complicate the assembly process (Richard, Kerrest & Dujon, 2008). To optimize the assembly, researchers often employ multiple library preparations for sequencing (Gnerre et al., 2010; Ekblom & Wolf, 2014), which may increase the potential sources of post-DNA extraction contamination. Contaminants in assembly results can eventually contaminate public databases (Merchant, Wood & Salzberg, 2014), and impair scientific findings (Artamonova et al., 2015). The detection and removal of contaminants poses a major bioinformatics challenge. To identify undesired contigs in a genomic assembly, scientists can simply compare their assembly results to public sequence databases for positive hits to unexpected taxa (Ekblom & Wolf, 2014), use k-mer coverage plots to identify distinct genomes (Percudani, 2013), or employ scatter plots to partition contigs based on their GC-content and coverage (Kumar et al., 2013). However, advanced solutions developed for accurate identification of microbial genomes in complex metagenomic assemblies can leverage these approaches further, and offer enhanced curation options for eukaryotic assemblies.

The first release of a tardigrade genome by Boothby et al. (2015) demonstrates a striking example of the importance of careful screening for contaminants in eukaryotic genome assemblies. Tardigrades are microscopic animals occurring in a wide range of ecosystems and they exhibit extended capabilities to survive in harsh conditions that would be fatal to most animals (Ramløv & Westh, 2001; Jönsson, Harms-Ringdahl & Torudd, 2005; Jönsson et al., 2008; Horikawa et al., 2013). Boothby and his colleagues generated a composite DNA sequencing dataset from a culture of the tardigrade *Hypsibius dujardini* by exploiting some of the best practices of high-throughput sequencing available today (Boothby et al., 2015). In their assembled tardigrade genome, the authors detected a large number of genes originating from bacteria, making up approximately one-sixth of the gene pool, and suggested that horizontal gene transfers (HGTs) could explain the unique ability of tardigrades to withstand extreme ranges of temperature, pressure, and radiation. However, Koutsovoulos et al.'s subsequent analysis of Boothby et al.'s assembly suggested that it contained extensive bacterial contamination, casting doubt on the extended HGT hypothesis (Koutsovoulos et al., 2015). By applying two-dimensional scatterplots on their own raw assembly results, Koutsovoulos et al. also reported a curated draft genome of *H. dujardini*.

Here we re-analyzed the raw sequencing data generated by Boothby et al. (2015) and Koutsovoulos et al. (2015), in combination with an independent RNA-Seq dataset generated by Levin et al. (2016) for *H. dujardini*. Using anvi'o, an analysis and visualization platform originally designed for the identification of bacterial genomes in metagenomic assemblies (Eren et al., 2015), we employed bacterial single-copy genes to assess the occurrence of bacterial genomes in the raw and curated assembly results, utilized k-mer frequencies and coverage values across multiple sequencing libraries to organize scaffolds, and visualized our findings in in a single display.

Material and methods

Genome assemblies, and raw sequencing data for DNA and RNA. Boothby et al. constructed three paired-end Illumina libraries (insert sizes of 0.3, 0.5 and 0.8 kbp) for 2 x 100 paired-end sequencing on a HiSeq2000 and six single-end long-read libraries (five Illumina Moleculo libraries sequenced by the Illumina "long read" DNA sequencing service, and one PacBio SMRT library sequenced using the P6-C4 chemistry and a 1 X 240 movie), which altogether provided a co-assembly of 252.5 Mbp (Boothby et al., 2015). The tardigrade genome released by Boothby et al. (2015), along with the nine sequencing data used for its assembly, are available at http://weatherby.genetics.utah.edu/seq_transf. Independently, Koutsovoulos et al. generated a 0.3 kbp insert library and a 1.1 kbp insert mate-pair library for 2 x 100 paired end sequencing on a HiSeq2000 that provided a co-assembly of 185.8 Mbp (nHd.1.0) (Koutsovoulos et al. 2015). These authors subsequently curated a 135 Mbp draft genome (nHd.2.3) by removing potential contamination and re-assembling filtered short reads (Koutsovoulos et al., 2015). The tardigrade raw assembly and curated draft genome released by Koutsovoulos et al. (2015) are available at http://badger.bio.ed.ac.uk/H_dujardini, and their two sequencing datasets are available from the ENA, under study accession PRJEB11910.

RNA-seq data. We obtained the RNA-seq data using the NCBI accession id PRJNA272543 (Levin et al., 2016). Briefly, Levin et al. isolated RNA from *H. dujardini* using the Trizol reagent (Invotrogen), constructed paired-end Illumina libraries according to the TruSeq RNA-seq protocol, and sequenced their cDNA libraries with a read length of 100 bp.

Quality filtering and read mapping. We used illumina-utils (Eren et al., 2013) (available from http://github.com/meren/illumina-utils) for quality filtering of short Illumina reads using 'iu-filter-quality-minoche' script with default parameters, which implements the quality filtering described by Minoche et al. (Minoche, Dohm & Himmelbauer, 2011). Bowtie2 v2.2.4 (Langmead & Salzberg, 2012) with default parameters mapped all reads to the scaffolds, and we used samtools v1.2 (Li et al., 2009) to convert reported SAM files to BAM files.

Overview of the anvi'o workflow. Our workflow with anvi'o to identify and remove contamination from a given collection of scaffolds consists of four main steps. The first step is the processing of the FASTA file of scaffolds to create an anvi'o contigs database (CDB). The resulting database holds basic information about each scaffold in the assembly (such as the k-mer frequency, or GC-content). The second step is the profiling of each BAM file with respect to the CDB we generated in the previous step. Each anvi'o profile describes essential statistics for each scaffold in a given BAM file, including their average coverage, and the portion of each scaffold covered by at least one read. The third step is to merge all anvi'o profiles. Merging step combines all statistics from individual profiles, and uses them to compute hierarchical clusterings of scaffolds. The default organization of scaffolds is determined by the average coverage information from individual profiles, and the sequence composition information from the CDB. This organization makes it possible to identify scaffolds that distribute similarly across different library preparations. The final step is to visualize the merged data on the anvi'o interactive interface. The anvi'o interactive interface provides a holistic perspective of the combined data, and allows the identification of draft genome bins, and removal of contaminants.

Processing of scaffolds, and mapping results. We used anvi'o v1.2.2 (available from http://github.com/meren/anvio) to process scaffolds and mapping results, visualize the distribution of scaffolds, and identify draft genomes following the workflow outlined in the previous section, and detailed in Eren et al (2015). We created an anvi'o contigs database CDB for each scaffold collection using the 'anvi-gen-contigs-database' program with default parameters (where k equals 4 for k-mer frequency analysis). We then annotated scaffolds with myRAST (available from http://theseed.org/) and imported these results into the CDB using the program 'anvi-populate-genes-table' to store the information about the locations of open reading frames (ORFs) in scaffolds, and their taxonomical and functional inference. We profiled individual BAM files using the program 'anvi-profile' with a minimum contig length of 1 kbp, and the program 'anvi-merge' combined resulting profiles with default parameters. For the analysis of Boothby et al. (2015) assembly, we also profiled the RNA-Seq data published by Levin et al. (2016) to identify scaffolds with transcriptomic activity, and exported the table for proportion of each scaffold covered by transcripts using the script 'get-db-table-as-matrix'. We used the supplementary material published by Boothby et al. (2015) ("Dataset S1" in the original publication) to identify scaffolds with proposed HGTs. Finally, we used the program 'anvi-interactive' visualize the merged data, and to identify genome bins. We included RNA-Seq results and scaffolds with HGTs into our visualization using the ` --additional-layers` flag. To finalize the anvi'o generated SVG files for publication, we used Inkscape v0.91 (available from https://inkscape.org/).

Predicting the number of bacterial genomes in an assembly. We used the occurrence of bacterial single-copy genes as a proxy to the expected number of bacterial genomes in a raw assembly or in a curated genome bin. First, we ran on each CDB generated in this study the anvi'o program 'anvi-populate-search-tables' to search using HMMer v3.1b2 (Eddy, 2011) for bacterial single-copy genes Campbell et al. (2013) published. Then, we used the anvi'o script 'gen-stats-for-single-copy-genes' to report the number of hits per single-copy gene as an array of integers from each CDB. We finally used mode (i.e., the most frequently occurring number) of this array as the expected number of complete bacterial genomes in a given collection of scaffolds. The script 'gen-stats-for-single-copy-genes' also used the R library 'ggplot' v1.0.0 (R Development Core Team, 2011; Ginestet, 2011) to plot the occurrence of single-copy genes.

Taxonomical and functional annotation of bacterial genomes. We uploaded bacterial draft genomes identified from the raw tardigrade genomic assembly results into the RAST server (Aziz et al., 2008), and used the RAST best taxonomic hits and FigFams to infer the taxonomy of genome bins and functions they harbor.

Data availability: The URL http://merenlab.org/data/ reports (1) anvi'o files to regenerate Figure 1 and Figure 2, (2) our curation of the tardigrade genome from Boothby et al.'s assembly (which is also available through the NCBI under the bioproject ID PRJNA309530), and (3) the FASTA files for bacterial genomes we identified in the raw assemblies from Boothby et al. and Koutsovoulos et al..

Results and Discussion

Boothby et al. generated sequencing data from a tardigrade culture using three short read (Illumina) and six long read (Moleculo and PacBio) libraries, which altogether provided a co-assembly of 252.5 Mbp (Boothby et al., 2015). Using this assembly, the authors suggested that 6,663 genes were entered into the tardigrade genome through HGTs. Independently, Koutsovoulos et al. generated sequencing data from another tardigrade culture using two short read Illumina libraries that provided a co-assembly of 185.8 Mbp, from which they could curate a 135 Mbp tardigrade draft genome by removing potential bacterial contamination using two-dimensional scatterplots of scaffolds with respect to their GC-content and coverage (Koutsovoulos et al., 2015).

A holistic view of the data

The use of multiple library preparations and sequencing strategies is likely to result in more optimal assembly results (Gnerre et al., 2010). Hence, we focused on the scaffolds generated by Boothby et al. (2015) as a foundation to maximize the recovery of the tardigrade genome. To provide a holistic understanding of the composite sequencing data generated by the two teams, we mapped the raw data from the nine DNA sequencing libraries from Boothby et al., and the two Illumina libraries from Koutsovoulos et al. (2015) on this assembly. Anvi'o generated a hierarchical clustering of scaffolds by combining the tetra-nucleotide frequency and coverage of each scaffold across the 11 DNA sequencing libraries (Eren et al., 2015). Besides visualizing the coverage of each scaffold in each sample, we highlighted scaffolds with HGTs identified by Boothby et al. on the resulting organization of scaffolds, and visualized RNA-seq mapping results. Figure 1 displays the anvi'o merged profile that represents all this information in a single display.

A draft genome for *H. dujardini*

Through the anvi'o interactive interface we selected 14,961 scaffolds from the Boothby et al. assembly that recruited large number of short-reads in a consistent manner (Fig. 1). This 182.2 Mbp selection with consistent coverage (#1 in Fig. 1) represents our curation of the tardigrade draft genome from Boothby et al.'s assembly. The remaining 7,535 scaffolds, which total about 70 Mbp of the assembly, harbored 96.1% of HGTs identified by Boothby et al. These scaffolds recruited only 0.05% of the reads from the RNA-Seq data, highlighting the extent of contamination in the original assembly. This finding is in agreement with Koutsovoulos et al.'s findings; however, our curated draft genome from the Boothby et al.'s assembly is 47 Mbp larger than the draft genome released by Koutsovoulos et al. (2015), most probably due to Boothby et al.'s inclusion of longer reads from Moleculo libraries. While the portion of scaffolds covered by RNA-Seq data suggests that this additional 47 Mbp still originate from the tardigrade genome, the biological relevance of this information (or lack thereof) for the characterization of the tardigrade genome falls outside of the scope of our study.

The origin of bacterial contamination

Our mapping results indicate the presence of non-target sequences in the assembly that recruit reads only from long-read libraries. One interpretation could be that most of the contamination in Boothby et al.'s assembly originated from Moleculo libraries, post DNA-extraction (Fig. 1). However, while a recent study shows that the majority of long reads from Moleculo libraries originated from low-abundance organisms in the analyzed samples (Sharon et al., 2015), another study suggests relatively more sequencing bias in Moleculo library preparation results (Kuleshov et al., 2015). Therefore, an alternative interpretation of the mapping results can be that the bacterial contaminants were present in the sample pre-DNA extraction at very low abundances, and each Moleculo library preparation included long reads originating from different parts of this rare community. Regardless, long reads considerably improved Boothby et al.'s assembly, which resulted in a larger tardigrade genome following the removal of non-target sequences. While these results reiterate that the use of long-read libraries is essential to generate more comprehensive assemblies, they also suggest that extra care should be taken to better mitigate the presence of non-target sequences in assembly results when long-read libraries are used for sequencing.

We identified three near-complete bacterial genomes affiliated to *Chitinophaga* and *Thermosinus* in Boothby et al.'s assembly (Fig. 1). Surprisingly, Boothby et al. identified only a small portion of these complete bacterial genomes as sources of HGTs while applying a metric specifically designed to detect foreign DNA in eukaryotic genomes. For instance, none of the 4,459 genes in bacterial draft genome #2 (selection #3 in Fig. 1) were reported in Boothby et al.'s findings as HGTs. We also processed and visualized the raw assembly (nHd.1.0) from Koutsovoulos et al. (2015) using anvi'o (Figure S1), and recovered eight bacterial genomes. However, we found no taxonomical overlap between high-completion bacterial genomes from the two sequencing projects (Table S1).

Interestingly, one bacterial genome (selection #2 in Fig. 1) was detected in DNA libraries from both groups, as well as in the RNA-seq data, suggesting that the related bacterial population was in all samples prior to the DNA/RNA extraction step. This genome is affiliated to *Chitinophaga*, and harbors genes coding for chitin degradation and utilization (Table S2). Chitin occurs naturally in the feeding apparatus of tardigrades (Guidetti et al., 2015), and might be a source of carbon for its microbial inhabitants. The genome also harbors genes coding for the biosynthesis of proteorhodopsin, host invasion and intracellular resistance, dormancy and sporulation, oxidative stress, and tryptophan, which is an essential amino acid for animals (Crawford, 1989; Zelante et al., 2013). Although this genome may belong to a tardigrade symbiont, the generation of the data does not allow us to rule out the possibility that it may be associated with the food source. Nevertheless, this finding suggests that there may be cases where non-target genomes in an assembly can provide clues about the lifestyle of a given host.

Best practices to assess bacterial contamination

Initial assessment of the occurrence of bacterial single-copy genes in eukaryotic assemblies can provide a quick estimation of the number of bacterial genomes that occur in assembly results. The use of bacterial single-copy genes can give much more accurate representation of potential bacterial contamination than screening for 16S rRNA genes alone, as they are less likely to be found in co-assembly results (Miller et al., 2011; Delmont et al., 2015). Although Boothby et al. reported the lack of 16S rRNA genes in their assembly (Boothby et al., 2015), anvi'o estimated that it contained at least 10 complete bacterial genomes (Fig. 2) using a bacterial single-copy gene collection (Campbell et al., 2013). This simple yet powerful step could identify cases of extensive contamination, and alert researchers to be diligent in identifying scaffolds originating from bacterial organisms. Figure 2 also summarizes the HMM hits in scaffolds found in curated tardigrade genomes from our analysis and Koutsovoulos et al.'s study. We observed that the average significance score for the remaining HMM hits for bacterial single-copy genes in curated genomes was 4.2 times lower in average compared to the HMM hits in assembly results (Table S3). The decrease in the significance scores, and the very similar patterns of occurrence of HMM hits between the two curation efforts suggest that some of the HMM profiles may not be specific enough to be identified only in bacteria.

Two-dimensional scatterplots have a long history of identifying distinct genomes in assembly results (Tyson et al., 2004) and continue to be used for delineating microbial genomes in metagenomic assemblies (Albertsen et al., 2013; Cantor et al., 2015), as well as detecting contamination in eukaryotic assembly results (Kumar et al., 2013). Although scatterplots can describe the organization of contigs in assembly results, they suffer from limited number of dimensions they can display, and their inability to depict complex supporting data that can improve the identification of individual genomes. These limitations are particularly problematic in sequencing projects covering multiple sequencing libraries, where displaying mapping results from each library can help detecting sources of contaminants. Despite their successful applications, two dimensional scatter plots limit researchers to the use of simple characteristics of the data that can be represented on an axis (such as GC-content). In contrast, clustering scaffolds, and overlaying multiple layers of independent information produce more comprehensive visualizations that display multiple aspects of the data.

Conclusions

The field of genomics requires advanced computational approaches to take best advantage of constantly evolving ways to generate sequencing data, and to identify and remove contamination from genome assemblies. Our study indicates that some of these advanced approaches may emerge from the field of metagenomics, where the need for *de novo* reconstruction of microbial genomes from environmental samples has given raise to techniques and software platforms that can make sense of complex assemblies. Here we used k-mer frequencies to organize scaffolds, the occurrence of bacterial single-copy genes to estimate the extent of contamination, and advanced visualization strategies to detect and remove contamination in a eukaryotic assembly project while simultaneously characterizing the sources of contamination. Our results also suggest that metagenomic binning strategies can be used to recover near-complete bacterial genomes from raw eukaryotic assemblies, which can provide insights into the potential host-microbe interactions during the curation step.

Acknowledgments

We are grateful to Thomas C. Boothby, Georgios Koutsovoulos, Sujai Kumar, and their colleagues for making their data available and answering our questions. We thank Itai Yanai for providing us with the RNA-Seq data. We also thank Hilary G. Morrison for her invaluable suggestions.

References

Albertsen M., Hugenholtz P., Skarshewski A., Nielsen KL., Tyson GW., Nielsen PH. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology* 31:533–8.

Alneberg J., Bjarnason BS., de Bruijn I., Schirmer M., Quick J., Ijaz UZ., Lahti L., Loman NJ., Andersson AF., Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods* 11:1144–1146.

Artamonova II., Lappi T., Zudina L., Mushegian AR. 2015. Prokaryotic genes in eukaryotic genome sequences: when to infer horizontal gene transfer and when to suspect an actual microbe. *Environmental Microbiology* 17:2203–2208.

Artamonova II., Mushegian AR. 2013. Genome sequence analysis indicates that the model eukaryote Nematostella vectensis harbors bacterial consorts. *Applied and environmental microbiology* 79:6868–73.

Aziz RK., Bartels D., Best AA., DeJongh M., Disz T., Edwards RA., Formsma K., Gerdes S., Glass EM., Kubal M., Meyer F., Olsen GJ., Olson R., Osterman AL., Overbeek RA., McNeil LK., Paarmann D., Paczian T., Parrello B., Pusch GD., Reich C., Stevens R., Vassieva O., Vonstein V., Wilke A., Zagnitko O. 2008. The RAST Server: rapid annotations using subsystems technology. *BMC genomics* 9:75.

Boothby TC., Tenlen JR., Smith FW., Wang JR., Patanella KA., Osborne Nishimura E., Tintori SC., Li Q., Jones CD., Yandell M., Messina DN., Glasscock J., Goldstein B. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences* 112:201510461.

Brown CT., Hug LA., Thomas BC., Sharon I., Castelle CJ., Singh A., Wilkins MJ., Wrighton KC., Williams KH., Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523:208–211.

Campbell JH., O'Donoghue P., Campbell AG., Schwientek P., Sczyrba A., Woyke T., Söll D., Podar M. 2013. UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proceedings of the National Academy of Sciences of the United States of America* 110:5540–5.

Cantor M., Nordberg H., Smirnova T., Hess M., Tringe S., Dubchak I. 2015. Elviz – exploration of metagenome assemblies with an interactive visualization tool. *BMC Bioinformatics* 16:130.

Chapman JA., Kirkness EF., Simakov O., Hampson SE., Mitros T., Weinmaier T., Rattei T., Balasubramanian PG., Borman J., Busam D., Disbennett K., Pfannkoch C., Sumin N., Sutton GG., Viswanathan LD., Walenz B., Goodstein DM., Hellsten U., Kawashima T., Prochnik SE., Putnam NH., Shu S., Blumberg B., Dana CE., Gee L., Kibler DF., Law L., Lindgens D., Martinez DE., Peng J., Wigge PA., Bertulat B., Guder C., Nakamura Y., Ozbek S., Watanabe H., Khalturin K., Hemmrich G., Franke A., Augustin R., Fraune S., Hayakawa E., Hayakawa S., Hirose M., Hwang JS., Ikeo K., Nishimiya-Fujisawa C., Ogura A., Takahashi T., Steinmetz PRH., Zhang X., Aufschnaiter R., Eder M-K., Gorny A-K., Salvenmoser W., Heimberg AM., Wheeler BM., Peterson KJ., Böttger A., Tischler P., Wolf A., Gojobori T., Remington KA., Strausberg RL., Venter JC., Technau U., Hobmayer B., Bosch TCG., Holstein TW., Fujisawa T., Bode HR., David CN., Rokhsar DS., Steele RE. 2010. The dynamic genome of Hydra. *Nature* 464:592–6.

Crawford IP. 1989. Evolution of a biosynthetic pathway: the tryptophan paradigm. *Annual review of microbiology* 43:567–600.

Delmont TO., Eren AM., Maccario L., Prestat E., Esen ÖC., Pelletier E., Le Paslier D., Simonet P., Vogel TM. 2015. Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. *Frontiers in microbiology* 6:358.

Dick GJ., Andersson AF., Baker BJ., Simmons SL., Thomas BC., Yelton AP., Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome biology* 10:R85.

Eddy SR. 2011. Accelerated Profile HMM Searches. *PLoS computational biology* 7:e1002195.

Ekblom R., Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications* 7:n/a–n/a.

Eren AM., Vineis JH., Morrison HG., Sogin ML. 2013. A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS ONE* 8:e66643.

Eren AM., Esen ÖC., Quince C., Vineis JH., Morrison HG., Sogin ML., Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3:e1319.

Gans J., Wolinsky M., Dunbar J. 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science (New York, N.Y.)* 309:1387–90.

Ginestet C. 2011. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174:245–246.

Gnerre S., MacCallum I., Przybylski D., Ribeiro FJ., Burton JN., Walker BJ., Sharpe T., Hall G., Shea TP., Sykes S., Berlin AM., Aird D., Costello M., Daza R., Williams L., Nicol R., Gnirke A., Nusbaum C., Lander ES., Jaffe DB. 2010. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* 108:1513–1518.

Guidetti R., Bonifacio A., Altiero T., Bertolani R., Rebecchi L. 2015. Distribution of Calcium and Chitin in the Tardigrade Feeding Apparatus in Relation to its Function and Morphology. *Integrative and comparative biology* 55:241–52.

Horikawa DD., Cumbers J., Sakakibara I., Rogoff D., Leuko S., Harnoto R., Arakawa K., Katayama T., Kunieda T., Toyoda A., Fujiyama A., Rothschild LJ. 2013. Analysis of DNA repair and protection in the Tardigrade Ramazzottius varieornatus and Hypsibius dujardini after exposure to UVC radiation. *PloS one* 8:e64793.

Jönsson KI., Rabbow E., Schill RO., Harms-Ringdahl M., Rettberg P. 2008. Tardigrades survive exposure to space in low Earth orbit. *Current biology : CB* 18:R729–R731.

Jönsson KI., Harms-Ringdahl M., Torudd J. 2005. Radiation tolerance in the eutardigrade Richtersius coronifer. *International journal of radiation biology* 81:649–56.

Kang DD., Froula J., Egan R., Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165.

Koutsovoulos G., Kumar S., Laetsch DR., Stevens L., Daub J., Conlon C., Maroon H., Thomas F., Aboobaker A., Blaxter M. 2015. *The genome of the tardigrade Hypsibius dujardini*. Cold Spring Harbor Labs Journals.

Kuleshov V., Jiang C., Zhou W., Jahanbani F., Batzoglou S., Snyder M. 2015. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. *Nature Biotechnology* 34:64–69.

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.

Langmead B., Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* 9:357–9.

Laurence M., Hatzis C., Brash DE. 2014. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. *PloS one* 9:e97876.

Levin M., Anavy L., Cole AG., Winter E., Mostov N., Khair S., Senderovich N., Kovalev E., Silver DH., Feder M., Fernandez-Valverde SL., Nakanishi N., Simmons D., Simakov O., Larsson T., Liu S-Y., Jerafi-Vider A., Yaniv K., Ryan JF., Martindale MQ., Rink JC., Arendt D., Degnan SM., Degnan BM., Hashimshony T., Yanai I. 2016. The mid-developmental transition and the evolution of animal body plans. *Nature* advance on.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* 25:2078–9.

Loman NJ., Pallen MJ. 2015. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* 13:787–794.

Merchant S., Wood DE., Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675.

Miller CS., Baker BJ., Thomas BC., Singer SW., Banfield JF. 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome biology* 12:R44.

Minoche AE., Dohm JC., Himmelbauer H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology* 12:R112.

Parks DH., Imelfort M., Skennerton CT., Hugenholtz P., Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* 25:1043–55.

Percudani R. 2013. A Microbial Metagenome (Leucobacter sp.) in Caenorhabditis Whole Genome Sequences. *Bioinformatics and biology insights* 7:55–72.

Pride DT., Meinersmann RJ., Wassenaar TM., Blaser MJ. 2003. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome research* 13:145–58.

R Development Core Team R. 2011. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* 1:409.

Ramløv H., Westh P. 2001. Cryptobiosis in the Eutardigrade Adorybiotus (Richtersius) coronifer: Tolerance to Alcohols, Temperature and de novo Protein Synthesis. *Zoologischer Anzeiger - A Journal of Comparative Zoology* 240:517–523.

Richard G-F., Kerrest A., Dujon B. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews : MMBR* 72:686–727.

Rusch DB., Halpern AL., Sutton G., Heidelberg KB., Williamson S., Yooseph S., Wu D., Eisen JA., Hoffman JM., Remington K., Beeson K., Tran B., Smith H., Baden-Tillson H., Stewart C., Thorpe J., Freeman J., Andrews-Pfannkoch C., Venter JE., Li K., Kravitz S., Heidelberg JF., Utterback T., Rogers Y-H., Falcón LI., Souza V., Bonilla-Rosso G., Eguiarte LE., Karl DM., Sathyendranath S., Platt T., Bermingham E., Gallardo V., Tamayo-Castillo G., Ferrari MR., Strausberg RL., Nealson K., Friedman R., Frazier M., Venter JC. 2007. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology* 5:e77.

Salter SJ., Cox MJ., Turek EM., Calus ST., Cookson WO., Moffatt MF., Turner P., Parkhill J., Loman NJ., Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* 12:87.

Schleper C., Jurgens G., Jonuscheit M. 2005. Genomic studies of uncultivated archaea. *Nature reviews. Microbiology* 3:479–88.

Schloss PD., Handelsman J. 2003. Biotechnological prospects from metagenomics. *Current opinion in biotechnology* 14:303–10.

Sharon I., Kertesz M., Hug LA., Pushkarev D., Blauwkamp TA., Castelle CJ., Amirebrahimi M., Thomas BC., Burstein D., Tringe SG., Williams KH., Banfield J. 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. *Genome Research*:gr.183012.114.

Strong MJ., Xu G., Morici L., Splinter Bon-Durant S., Baddoo M., Lin Z., Fewell C., Taylor CM., Flemington EK. 2014. Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS pathogens* 10:e1004437.

Teeling H., Meyerdierks A., Bauer M., Amann R., Glöckner FO. 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental microbiology* 6:938–47.

Tyson GW., Chapman J., Hugenholtz P., Allen EE., Ram RJ., Richardson PM., Solovyev V V., Rubin EM., Rokhsar DS., Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428:37–43.

Venter JC., Adams MD., Myers EW., Li PW., Mural RJ., Sutton GG., Smith HO., Yandell M., Evans CA., Holt RA., Gocayne JD., Amanatides P., Ballew RM., Huson DH., Wortman JR., Zhang Q., Kodira CD., Zheng XH., Chen L., Skupski M., Subramanian G., Thomas PD., Zhang J., Gabor Miklos GL., Nelson C., Broder S., Clark AG., Nadeau J., McKusick VA., Zinder N., Levine AJ., Roberts RJ., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian AE., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman TJ., Higgins ME., Ji RR., Ke Z., Ketchum KA., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G V., Milshina N., Moore HM., Naik AK., Narayan VA., Neelam B., Nusskern D., Rusch DB., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., Yao A., Ye J., Zhan M., Zhang W., Zhang H., Zhao Q., Zheng L., Zhong F., Zhong W., Zhu S., Zhao S., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng ML., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers YH., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E., Thomas R., Tint NN., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril JF., Guigó R., Campbell MJ., Sjolander K V., Karlak B., Kejariwal A., Mi H., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang YH., Coyne M., Dahlke C., Mays A., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M., Wu D., Wu M., Xia A., Zandieh A., Zhu X. 2001. The sequence of the human genome. *Science (New York, N.Y.)* 291:1304–51.

Wu Y-W., Tang Y-H., Tringe SG., Simmons BA., Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26.

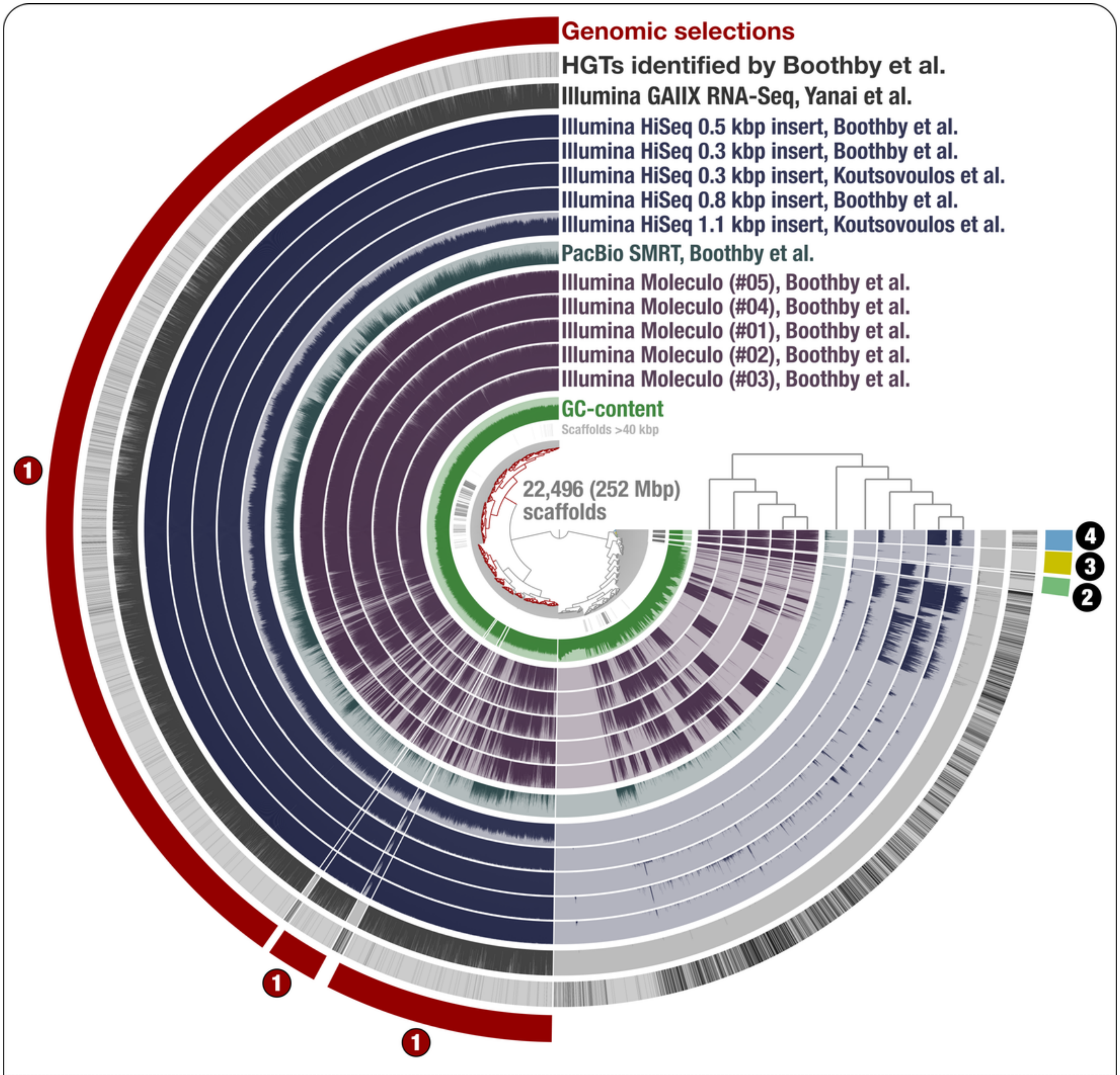
Wu M., Eisen JA. 2008. A simple, fast, and accurate method of phylogenomic inference. *Genome Biology* 9:R151.

Zelante T., Iannitti RG., Cunha C., De Luca A., Giovannini G., Pieraccini G., Zecchi R., D'Angelo C., Massi-Benedetti C., Fallarino F., Carvalho A., Puccetti P., Romani L. 2013. Tryptophan catabolites from microbiota engage aryl hydrocarbon receptor and balance mucosal reactivity via interleukin-22. *Immunity* 39:372–85.

1

Holistic assessment of the tardigrade genome release from Boothby et al. (2015).

Dendrogram in the center organizes scaffolds based on sequence composition and coverage values in data from 11 DNA libraries. Scaffolds larger than 40 kbp were split into sections of 20 kbp for visualization purposes. Splits are displayed in the first inner circle and GC-content (0-71%) in the second circle. In the following 11 layers, each bar represents the portion of scaffolds covered by short reads in a given sample. The next layer shows the same information for RNA-Seq data. Scaffolds harboring genes used by Boothby et al. to support the expended HGT hypothesis is shown in the next layer. Finally, the outermost layer shows our selections of scaffolds as draft genome bins: the curated tardigrade genome (selection #1), as well as three near-complete bacterial genomes originating from various contamination sources (selection #2, #3, and #4).



1 ***Hypsibius dujardini* curated draft genome. 182.2 Mbp** (14,961 scaffolds).

2 **Bacterial draft genome #1. 4.8 Mbp** (4 scaffolds; 100% complete with 5.9% redundancy. RAST Taxonomy: *Chitinophaga pinensis*).

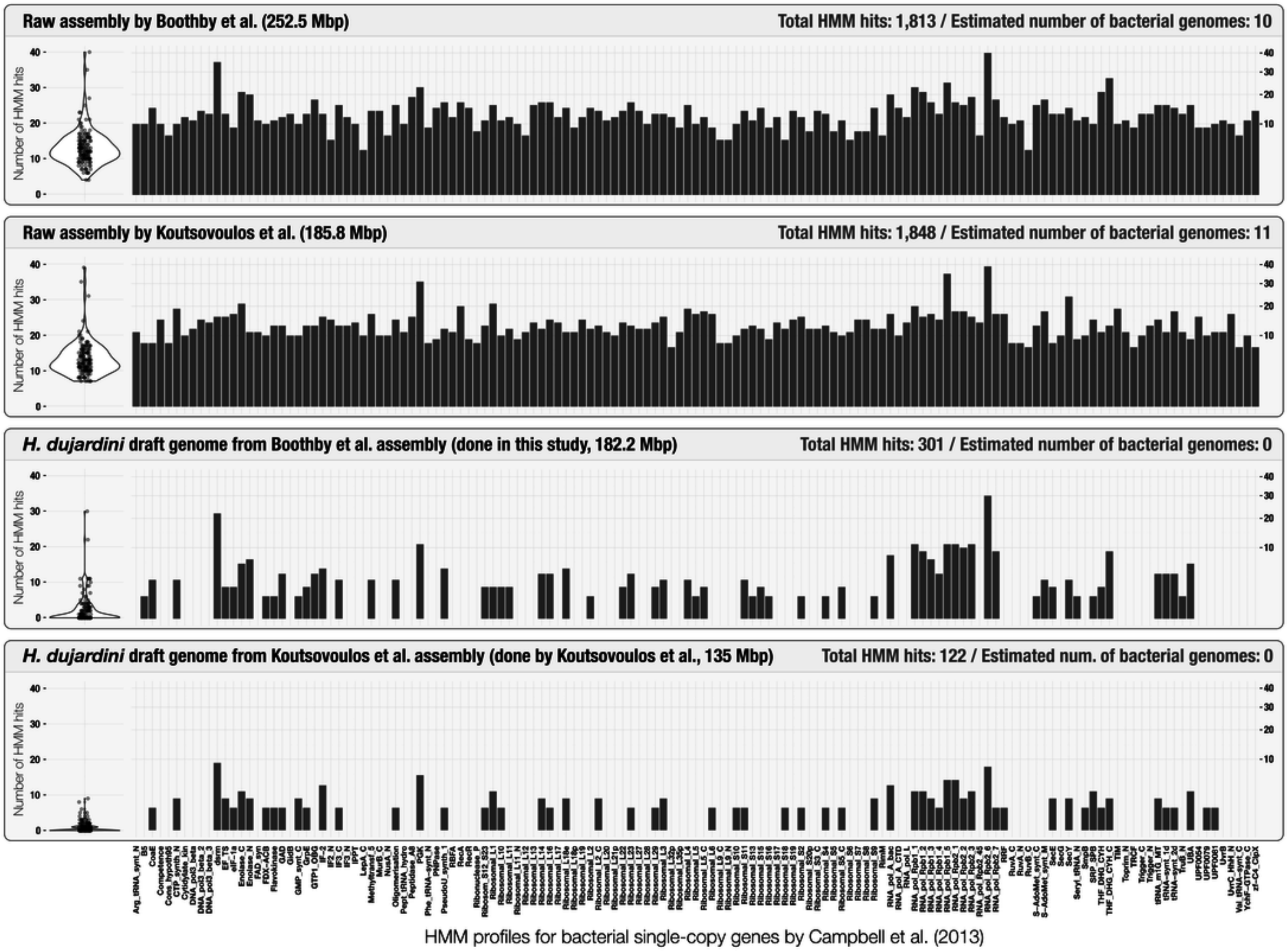
3 **Bacterial draft genome #2. 4.5 Mbp** (29 scaffolds; 97% complete with 0% redundancy. RAST Taxonomy: *Chitinophaga pinensis*).

4 **Bacterial draft genome #3. 3.8 Mbp** (5 scaffolds; 97% complete with 5.9% redundancy. RAST Taxonomy: *Thermosinus carboxydivorans*).

2

Occurrence of the 139 bacterial single-copy genes reported by Campbell et al. (2013) across scaffold collections.

The top two plots display the frequency and distribution of single-copy genes in the raw tardigrade genomic assembly generated by Boothby et al. (2015), and Koutsovoulos et al. (2015), respectively. The bottom two plots display the same information for each of the curated tardigrade genomes. Each bar represents the squared-root normalized number of significant hits per single-copy gene. The same information is visualized as box-plots on the left side of each plot.


Raw assembly by Boothby et al. (252.5 Mbp)
Total HMM hits: 1,813 / Estimated number of bacterial genomes: 10
Raw assembly by Koutsovoulos et al. (185.8 Mbp)
Total HMM hits: 1,848 / Estimated number of bacterial genomes: 11
H. dujardini draft genome from Boothby et al. assembly (done in this study, 182.2 Mbp)
Total HMM hits: 301 / Estimated number of bacterial genomes: 0
H. dujardini draft genome from Koutsovoulos et al. assembly (done by Koutsovoulos et al., 135 Mbp)
Total HMM hits: 122 / Estimated num. of bacterial genomes: 0
Number of HMM hits
Arg_tRNA_synt_N
B5
CoaE
Competence
Cons_hypoth95
CTP_synth_N
Cytidylate_kin
DNA_pol3_beta
DNA_pol3_beta_2
DNA_pol3_beta_3
dsrm
EF_TS
eIF-1a
Enolase_C
Enolase_N
FAD_syn
FDX-ACB
Flavokinase
GAD
GidB
GMP_synt_C
GrpE
GTP1_OBG
IF-2
IF2_N
IF3_C
IF3_N
IPPT
LepA_C
Methyltransf_5
MurB_C
NusA_N
Oligomerisation
Pept_tRNA_hydro
Peptidase_A8
PGK
Phe_tRNA-synt_N
PNPase
PseudoU_synth_1
RBFA
RecA
RecR
Ribonuclease_P
Ribosom_S12_S23
Ribosomal_L1
Ribosomal_L10
Ribosomal_L11
Ribosomal_L11_N
Ribosomal_L12
Ribosomal_L13
Ribosomal_L14
Ribosomal_L16
Ribosomal_L17
Ribosomal_L18e
Ribosomal_L18p
Ribosomal_L19
Ribosomal_L2
Ribosomal_L2_C
Ribosomal_L20
Ribosomal_L21p
Ribosomal_L22
Ribosomal_L23
Ribosomal_L27
Ribosomal_L28
Ribosomal_L29
Ribosomal_L3
Ribosomal_L32p
Ribosomal_L35p
Ribosomal_L4
Ribosomal_L5
Ribosomal_L5_C
Ribosomal_L6
Ribosomal_L9_C
Ribosomal_L9_N
Ribosomal_S10
Ribosomal_S11
Ribosomal_S13
Ribosomal_S15
Ribosomal_S16
Ribosomal_S17
Ribosomal_S18
Ribosomal_S19
Ribosomal_S2
Ribosomal_S20p
Ribosomal_S3_C
Ribosomal_S4
Ribosomal_S5
Ribosomal_S5_C
Ribosomal_S6
Ribosomal_S7
Ribosomal_S8
Ribosomal_S9
RimM
RNA_pol_A_bac
RNA_pol_A_CTD
RNA_pol_L
RNA_pol_Rpb1_1
RNA_pol_Rpb1_2
RNA_pol_Rpb1_3
RNA_pol_Rpb1_4
RNA_pol_Rpb1_5
RNA_pol_Rpb2_1
RNA_pol_Rpb2_2
RNA_pol_Rpb2_3
RNA_pol_Rpb2_45
RNA_pol_Rpb2_6
RNA_pol_Rpb2_7
RRF
RuvA_C
RuvA_N
RuvB_C
S-AdoMet_synt_C
S-AdoMet_synt_M
SecE
SecG
SecY
Seryl_tRNA_N
SmpB
SRP_SPB
THF_DHG_CYH
THF_DHG_CYH_C
TIM
Toprim_N
TRCF
Trigger_C
Trigger_N
tRNA_m1G_MT
tRNA-synt_1d
tRNA-synt_2d
TruB_N
UBA
UPF0054
UPF0079
UPF0081
UvrB
UvrC_HhH_N
Val_tRNA-synt_C
YchF-GTPase_C
zf-C4_ClpX
HMM profiles for bacterial single-copy genes by Campbell et al. (2013)