# 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* 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.


**Key words**: genomics, assembly, curation, visualization, contamination, anvi'o, HGT

## 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; Brown et al. 2015; Schleper et al. 2005). 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 et al. 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 (Albertsen et al. 2013; 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 et al. 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 et al. 2008). To optimize the assembly, researchers often employ multiple library preparations for sequencing (Ekblom & Wolf 2014; Gnerre et al. 2010), which may increase the potential sources of post-DNA extraction contamination. Contaminants in assembly results can eventually contaminate public databases (Merchant et al. 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 et al. 2005, 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 assembly results (which were also contaminated with bacterial sequences), Koutsovoulos et al. 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) using anvi'o, an analysis and visualization platform originally designed for the identification and assessment of bacterial genomes in metagenomic assemblies (Eren et al. 2015). In our analysis, we relied on bacterial single-copy genes to assess the occurrence of bacterial genomes in assembly results, used k-mer frequencies to organize contigs, combined all sequencing data for each library preparation method from both groups into a single display, and overlaid RNA-Seq data (courtesy of Itai Yanai) over contigs to confirm the origin of contigs.

## 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 (Koutsovoulos et al. 2015). These authors subsequently curated a 135 Mbp draft genome by removing potential bacterial contamination (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. Itai Yanai (Technion - Israel Institute of Technology, http://yanailab.technion.ac.il/) graciously provided RNA-seq data generated from a *H. dujardini* culture, which will be available under the accession ID accession GSE70185 upon their publication.

**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 et al. 2011). Bowtie2 v2.2.4 (Langmead & Salzberg 2012) with default parameters mapped all reads to assemblies. We used samtools v1.2 (Li et al. 2009) to generate BAM files from mapping results.

**Processing of contigs, visualization and genome binning**. We processed BAM files and raw genome assemblies using anvi'o v1.2.2 (available from http://github.com/meren/anvio), generated anvi'o contig databases, profiled BAM files, and merged resulting profiles using default parameters and following the metagenomic workflow outlined in Eren et al (2015). In addition, we mapped and profiled the RNA-seq data to identify scaffolds with transcriptomic activity, and exported the table for proportion of each scaffold covered by transcripts using anvi'o 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. We included the RNA-seq results and scaffolds with HGTs into our visualization as an additional data file. The URL http://merenlab.org/data/ reports anvi'o files to regenerate Figure 1 and Figure 2, our curation of the tardigrade genome from Boothby et al.'s assembly (which is also available in NCBI via the bioproject ID PRJNA309530), and the FASTA files for bacterial genomes we identified in the Boothby et al. and Koutsovoulos et al. assemblies. To finalize the anvi'o generated SVG files for publication, we used Inkscape v0.91 (available from https://inkscape.org/).

**Predicting number of bacterial genomes**. To estimate the number of bacterial genomes in a given collection of scaffolds in a raw assembly or in a curated genome bin, and to visualize the distribution of HMM hits for each bacterial single-copy gene, we used the anvi'o script 'gen-stats-for-single-copy-genes', which reports the most frequent number in the list of number of hits per single-copy gene as the estimated number of bacterial genomes in a collection of scaffolds. The script uses HMMer v3.1b2 (Eddy 2011) to search for Hidden Markov Profiles (HMMs) of 139 bacterial single-copy genes identified by Campbell et al (2013), and the R library 'ggplot' v1.0.0 (R Development Core Team 2011; Ginestet 2011) to plot results.

**Taxonomical and functional annotation of bacterial genomes.** After binning, we uploaded bacterial draft genomes recovered from the assembly 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.

## 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 without any curation, 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 larger 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 is 47 Mbp larger than the draft genome released by Koutsovoulos et al. (2015). The portion of scaffolds covered by RNA-Seq data suggests that the additional 47 Mbp still originate from the tardigrade genome. Thus, our selection is likely to be a more complete draft genome for *H. dujardini* than that of Koutsovoulos et al., most probably due to Boothby et al.'s inclusion of longer reads.

## 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, a recent study shows that the majority of long reads from Moleculo libraries originated from low-abundance organisms in samples (Sharon et al. 2015), while another study suggests relatively more sequencing bias in Moleculo library preparation results (Kuleshov et al. 2015). Therefore another interpretation of the mapping results can be that the bacterial contaminants were present in the sample in low abundances pre-DNA extraction, and individual Moleculo library preparations resulted in 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. Although this falls outside of the scope of our study,

this oddity may indicate a potential flaw in metrics commonly used to quantify foreign DNA in eukaryotic genomes. We also processed and visualized the raw assembly 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 tryptophan, an essential amino acid for animals (Crawford 1989; Zelante et al. 2013), proteorhodopsin, host invasion and intracellular resistance, dormancy and sporulation, and oxidative stress. 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. The need for *de novo* reconstruction of microbial genomes from environmental samples through shotgun metagenomics data has given raise to advanced techniques and software platforms that can make sense of complex assemblies (Wu et al. 2014; Dick et al. 2009; Alneberg et al. 2014; Kang et al. 2015; Eren et al. 2015). Our study demonstrates that these approaches can be effectively used in eukaryotic assembly projects for curation purposes.

## 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. This work was supported by the Frank R. Lillie Research Innovation Award, and startup funds from the University of Chicago.

## Figure and table legends

Figure 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).

Figure 2. Figure 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.

Figure S1. Visualization and curation of the raw tardigrade genome assembly from Koutsovoulos et al. (2015). In the left panel (curation step I), 24,841 scaffolds that were longer than 1 kbp from the raw assembly were clustered based on sequence composition and coverage values in data from the two Illumina sequencing libraries (the inner dendrogram). Scaffolds longer than 40 kbp were split into sections of 20 kbp for visualization purposes. The second layer shows the GC-content for each scaffold. The next two view layers represent the log-normalized mean coverage values for scaffolds in the two sequencing datasets. Finally, our scaffold selections (tardigrade draft 01 and six bacterial draft genomes) are displayed in the outer layer. In the right panel (curation step II), the 15,839 scaffolds from the tardigrade selection from step I were clustered based on sequence composition only for more precise curation. Additional scaffold selections (tardigrade draft 02 and two bacterial draft genomes) are displayed in the outer layer.

Table S1. Summary of *H. dujardini* and bacterial genomes identified from the raw assembly results of Boothby et al. (2015) and Koutsovoulos et al. (2015). * Inferred from Boothby et al. (2015) and Koutsovoulos et al. (2015) publications. ** Scores were calculated using bacterial single copy genes from Campbell et al. (2013) and are only used to assess bacterial contamination levels in the eukaryotic assembly results.

Table S2. Summary of functions identified by RAST in the bacterial draft genome #2 (selection #3 in Fig. 1).

Table S3. Summary of HMM hits for each bacterial single-copy gene (collection of 139 from Campbell et al. (2013)) identified in 1) the raw assembly by Boothby et al. (2015), 2) the raw assembly by Koutsovoulos et al. (2015), 3) the curated draft genome of Hypsibius dujardini from Boothby et al. assembly in this study, and 4) the curated draft genome of *H. dujardini* from Koutsovoulos et al. (2015).

## 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. Nat. Biotechnol. 31:533–8.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. (2014). Binning metagenomic contigs by coverage and composition. Nat. 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. Environ. Microbiol. 17:2203–2208.

Artamonova II, Mushegian AR. (2013). Genome sequence analysis indicates that the model eukaryote Nematostella vectensis harbors bacterial consorts. Appl. Environ. Microbiol. 79:6868–73.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. (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, et al. (2015). Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc. Natl. Acad. Sci. 112:201510461.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. (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, et al. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proc. Natl. Acad. Sci. U. S. A. 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, et al. (2010). The dynamic genome of Hydra. Nature 464:592–6.

Crawford IP. (1989). Evolution of a biosynthetic pathway: the tryptophan paradigm. Annu. Rev. Microbiol. 43:567–600.

Delmont TO, Eren AM, Maccario L, Prestat E, Esen ÖC, Pelletier E, et al. (2015). Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Front. Microbiol. 6:358.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, et al. (2009). Community-wide analysis of microbial genome sequence signatures. Genome Biol. 10:R85.

Eddy SR. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. 7:e1002195.

Ekblom R, Wolf JBW. (2014). A field guide to whole-genome sequencing, assembly and annotation. Evol. Appl. 7:n/a–n/a.

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

Eren AM, Vineis JH, Morrison HG, Sogin ML. (2013). A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology Jordan, IK, ed. PLoS One

8:e66643.

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

Ginestet C. (2011). ggplot2: Elegant Graphics for Data Analysis. J. R. Stat. Soc. Ser. A (Statistics Soc. 174:245–246.

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. (2010). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. 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. Integr. Comp. Biol. 55:241–52.

Horikawa DD, Cumbers J, Sakakibara I, Rogoff D, Leuko S, Harnoto R, et al. (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, Harms-Ringdahl M, Torudd J. (2005). Radiation tolerance in the eutardigrade Richtersius coronifer. Int. J. Radiat. Biol. 81:649–56.

Jönsson KI, Rabbow E, Schill RO, Harms-Ringdahl M, Rettberg P. (2008). Tardigrades survive exposure to space in low Earth orbit. Curr. Biol. 18:R729–R731.

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, et al. (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. Nat. Biotechnol. 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. Front. Genet. 4:237.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. Nat. 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.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Loman NJ, Pallen MJ. (2015). Twenty years of bacterial genome sequencing. Nat. Rev. Microbiol. 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 Biol. 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 Biol. 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 Res. 25:1043–55.

Percudani R. (2013). A Microbial Metagenome (Leucobacter sp.) in Caenorhabditis Whole Genome Sequences. Bioinform. Biol. Insights 7:55–72.

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

R Development Core Team R. (2011). R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. 1:409.

Ramløv H, Westh P. (2001). Cryptobiosis in the Eutardigrade Adorybiotus (Richtersius) coronifer: Tolerance to Alcohols, Temperature and de novo Protein Synthesis. Zool. Anzeiger - A J. Comp. Zool. 240:517–523.

Richard G-F, Kerrest A, Dujon B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. 72:686–727.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol. 5:e77.

Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12:87.

Schleper C, Jurgens G, Jonuscheit M. (2005). Genomic studies of uncultivated archaea. Nat. Rev. Microbiol. 3:479–88.

Schloss PD, Handelsman J. (2003). Biotechnological prospects from metagenomics. Curr. Opin. Biotechnol. 14:303–10.

Sharon I, Kertesz M, Hug LA, Pushkarev D, Blauwkamp TA, Castelle CJ, et al. (2015). Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. gr.183012.114.

Strong MJ, Xu G, Morici L, Splinter Bon-Durant S, Baddoo M, Lin Z, et al. (2014). Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. PLoS Pathog. 10:e1004437.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. 6:938–47.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. (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, et al. (2001). The sequence of the human genome. Science 291:1304–51.

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

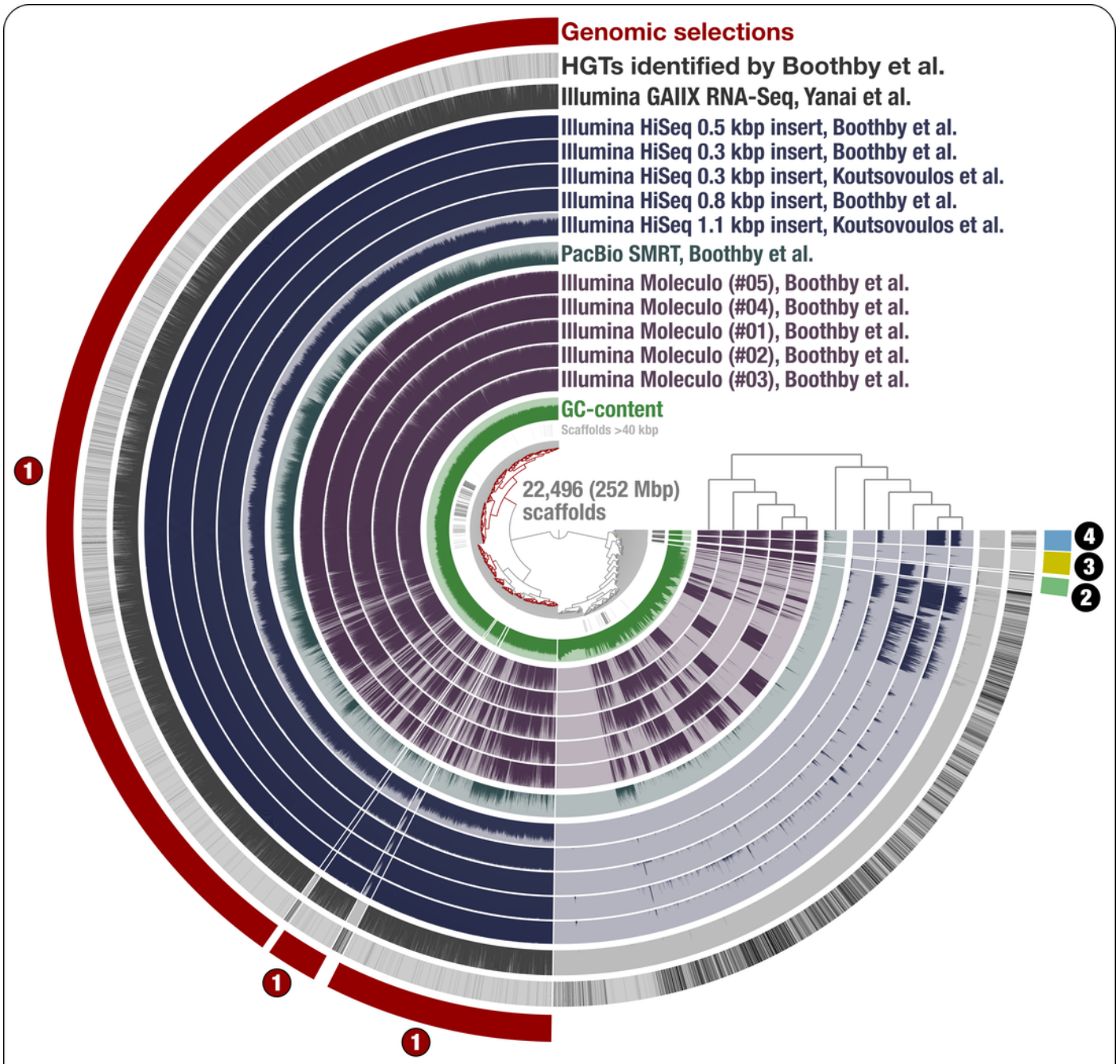
Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:26.

Zelante T, Iannitti RG, Cunha C, De Luca A, Giovannini G, Pieraccini G, et al. (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.