Bioinformatics and Biology Insights

ORIGINAL RESEARCH

A Microbial Metagenome (*Leucobacter* sp.) in *Caenorhabditis* Whole Genome Sequences

Riccardo Percudani

Department of Biosciences, Laboratory of Biochemistry and Molecular Biology, University of Parma, Parma, Italy.
Corresponding author email: riccardo.percudani@unipr.it

Abstract: DNA of apparently recent bacterial origin is found in the genomic sequences of *Caenorhabditis angaria* and *Caenorhabditis remanei*. Here we present evidence that the DNA belongs to a single species of the genus *Leucobacter* (high-GC Gram+ Actinobacteria). Metagenomic tools enabled the assembly of the contaminating sequences in a draft genome of 3.2 Mb harboring 2,826 genes. This information provides insight into a microbial organism intimately associated with *Caenorhabditis* as well as a solid basis for the reassignment of 3,373 metazoan entries of the public database to a novel bacterial species (*Leucobacter* sp. AEAR). The application of metagenomic techniques can thus prevent annotation errors and reveal unexpected genetic information in data obtained by conventional genomics.

Keywords: host-microbe interactions, organism identification in DNA sequences, contamination in sequence database, next-generation sequencing, purine degradation

Bioinformatics and Biology Insights 2013:7 55–72

doi: 10.4137/BBI.S11064

This article is available from http://www.la-press.com.

Introduction

During a routine search in the Genbank database for proteins involved in allantoin (a purine derivative) catabolism, a significant similarity of a typical bacterial enzyme to a protein sequence from *Caenorhabditis remanei* (XP_003087095; hypothetical protein CRE_29417) was detected. In the corresponding genomic scaffold (NW_003316202), another open reading frame (ORF) was present (XP_003087094; hypothetical protein CRE_29416) coding for another typical bacterial enzyme of allantoin catabolism. The genomic scaffold had as a best hit a whole genome shotgun (wgs) sequence from *Caenorhabditis angaria* (AEHI01035490). Other genes involved in purine catabolism and similar to bacterial genes were found in *C. angaria* and *C. remanei* wgs sequences, suggesting the presence in these two species of a whole metabolic pathway absent in other *Caenorhabditis*.

Our first hypothesis was the horizontal transfer of bacterial genes to a common ancestor of *C. angaria* and *C. remanei*. The acquisition of genes involved in the recycling of nitrogen from purines[1,2] made sense in light of the ecological niches of some *Caenorhabditis*: ureidic purine derivatives allantoin and allantoic acid can be the primary, if not the only nitrogen source[3] of certain decomposed vegetables on which these species normally feed.[4] This situation was reminiscent of the recently reported adaptive acquisition in the coffee beetle of a bacterial gene allowing the utilization of the coffee berry galactomannan.[5]

However, this initial hypothesis was called into question by further evidence suggesting the presence of a bacterial contamination in the genomic sequences. Firstly, *C. angaria* and *C. remanei* are not sister species in the *Caenorhabditis* phylogeny—*C. remanei* belongs to the *elegans* super-group and *C. angaria* to the *drosophilae* super-group[4]—implying independent events of horizontal transfer or the loss of the transferred gene in many other *Caenorhabditis* species. In the latter scenario, however, the genes identified in *Caenorhabditis. angaria* and *C. remanei* appeared too similar to each other (>95% identity) in relation to the phylogenetic distance of the two species. Secondly, no significant similarity was found in any part of the genomic scaffolds containing the purported horizontally transferred genes with eukaryotic genomes other than *C. angaria* and *C. remanei*.

Here we describe the analyses that allowed us to establish the presence of exogenous DNA in the *Caenorhabditis* wgs sequences, and to identify the contaminant DNA in both *Caenorhabditis* species as belonging to a novel species of the genus *Leucobacter*. By taking advantage of metagenomic techniques, the contaminant sequences have been segregated from the bulk assembly and brought together in a draft genome of 3.2 Mb. This information allowed the identification of the sequences erroneously attributed to *C. remanei* and *C. angaria* in the public database, while the functionally annotated gene catalog deriving from the draft genome made it possible to predict biological properties of the novel microbial species and the nature of its association with *Caenorhabditis.* Based on the evidence presented here, we propose that certain metagenomic analyses could be routinely applied to conventional genomics in order to improve accuracy and quality of the genomic information.

Methods

Identification of exogenous DNA

A reference set of *E. coli* ribosomal protein was obtained from GenBank and searched using tblastn against the wgs contigs of *Caenorhabditis* species. Sequences with significant similarity ($E < 10^{-2}$) were searched using blastx to identify the best hits in the Refseq database. The MGTAXA Galaxy server (http://mgtaxa.jcvi.org/) was used for the taxonomic prediction[8] of *Caenorhabditis* wgs contigs, and to obtain charts of taxonomic assignments for wgs contigs of different *Caenorhabditis.* species. The GC content of *Caenorhabditis* wgs contigs was determined with the bp_gccalc.pl script of the Bioperl package. For GC content comparison, artificial 1 Kb contigs were extracted from the complete genomes of *C. elegans* and *L. xyli* using the splitter program of the EMBOSS package. The analysis of sequence similarity between wgs contigs of different *Caenorhabditis* species was conducted with a local version of blast (Ver. 2.2.26) using blastn with r (reward for a match) and q (penalty for a mismatch) parameters set to 5 and –4, respectively. Species determination was based on phylogenetic reconstruction of 16S rRNA genes. Sequences were aligned with ClustalW,[33] and the alignment manually edited using GeneDoc (http://www.nrbsc.org/gfx/genedoc/). Maximum-likelihood phylogeny[34] was obtained with the dnaml program of the PHYLIP package. Bootstrap analysis (n = 100) was conducted with Mega 5.1 (http://www.megasoftware.net). The tree was rooted by midpoint and visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Genome assembly

Sanger reads, quality values, and ancillary information of *C. remanei* were downloaded from the NCBI Trace Archive FTP site (ftp://ftp.ncbi.nih.gov/pub/TraceDB). Reads were trimmed with the Figaro_trim_seqs utility of the Figaro package[35] using the vector clipping information provided by the Trace Archive. Illumina reads of *C. angaria* were downloaded from the NCBI Sequence Read Archive FTP site (ftp://ftptrace.ncbi.nlm.nih.gov/sra) using the Axel download accelerator (http://freecode.com/projects/axel) and extracted in FASTA format using the fastqdump utility of the SRA Toolkit. *C. angaria* Illumina reads of individual sequencing runs were assembled with Velvet[9] (ver. 1.2.03) using a k-mer value of 41. The contigs assembled with automatic cutoff values (-exp_cov = auto) were blasted against the high-CG fraction (GC $>$ 0.55) of *C. remanei* contigs to reveal the presence of *Leucobacter* reads. The run revealing the presence of *Leucobacter* reads (SRR065714) was used for further analyses. Optimal cutoffs for the bacterial assembly were established through metagenomic analysis of the contig coverage distribution. The R package ggplot2 (http://had.co.nz/ggplot2/) was used to graph the length-weighted frequency of contig coverage based on Velvet statistics, and the density distribution of the per sample taxonomic predictions of the MGTAXA server. The contigs obtained with optimized cutoffs were selected based on MGTAXA predictions and blastn similarity with draft genome of *Leucobacter chromiiresistens*. To provide mate pair information *C. remanei* Sanger reads with GC $>$ 0.55 were mapped to the selected contigs using the minimus2 program of the Amos package (ver. 3.1). Other contig links were established through gene syntheny with *L. chromiiresistens* inferred by comparing the encoded proteins using the promer program of the MUMer package.[36] Contigs were scaffolded with the Bambus program (Ver. 2.3)[35] giving higher priority to the 167 mate pair links and lower priority to the 442 syntheny links. All the assembly and scaffolding procedures were carried out with a cluster of six-core AMD Opteron 8425 HE processors (24 CPUs and 132 GB RAM) running Scientific Linux.

Genome annotation

The draft genome of *L.* sp. AEAR was annotated with the Rapid Annotation using Subsystem Technology (RAST) server.[12] tRNA sequences were identified in short contigs not included in the draft genome using a local version of tRNAscan-SE,[37] and assigned to *L.* sp. AEAR or to *C. angaria* through homology searches. Functional distinction of tRNA genes with CAT anticodon were determined with a standalone version of the TFAM 1.3 program.[38] Codon usage was tabulated from the complete set of protein coding sequences using the codcopy program of the EMBOSS package. To map genes into the purine degradation pathway, candidate proteins were first selected by blasting proteins representative of the known families of the pathway against the complete collection of *L.* sp. AEAR proteins. Candidate proteins were then blasted against a set of proteins involved in the pathway ("in-pathway" set), and a set of homologous proteins not involved in the pathway ("out-pathway" set). Sequences were assigned a particular reaction of the pathway if the best in-pathway score was significant ($E < 10^{-6}$) and higher than the best out-pathway score.

Results

Identification of exogenous DNA in *Caenorhabditis* genomic sequences

To test the hypothesis of a bacterial contamination, we searched the *Caenorhabditis* wgs sequences with a collection of 23 *Escherichia coli* ribosomal proteins, a class of proteins in which only exceptional cases of horizontal transmission have been documented.[6,7] The *E. coli* ribosomal proteins found matches in *C. angaria* (and sometimes in *C. remanei*) with a much higher significance than in other *Caenorhabditis* species (Table S1). When searched against the Refseq database, those anomalous hits found the best matches with ribosomal genes from bacteria (particularly Actinobacteria), whereas the other hits found the best matches with ribosomal genes from metazoa.

To gain further insight on the presence of exogenous DNA, wgs contigs of the six *Caenorhabditis* species present in Genbank were submitted to MGTAXA, a software for the taxonomic assignment of metagenomic sequences.[8] The fraction of wgs sequences from

C. brenneri, *C. japonica*, *C. sp.* 11, and also *C. remanei* assigned to the bacterial domain was equal or less than 1%, whereas a much higher proportion of putative bacterial sequences (12%) was detected in *C. angaria* (Fig. S1); the relative majority (38%) of the *C. angaria* contigs of putative bacterial origin were assigned to Actinobacteria, a class of high GC Gram+ bacteria.

Homology results obtained with the ribosomal proteins and the genome-wide taxonomic prediction suggested a phylogenetic relationship with GC-rich Actinobacteria for some *Caenorhabditis* wgs sequences. As the *Caenorhabditis* genomes have typically a low GC content (<40% GC), a distinct nucleotide composition can be anticipated for sequences deriving from both sources. Indeed, the analysis of nucleotide composition revealed the presence of an atypical GC content in some *Caenorhabditis* wgs sequences (Fig. 1). The plots of the GC content of individual contigs showed in *C. angaria* and *C. remanei* (and to a lesser extent in *C.* sp. 11) a wider variation than other *Caenorhabditis*, with a range that would be expected from a combination of sequences from nematodes and high CG bacteria (Fig. 1A). The frequency of distribution of the GC content in the contigs of *C. angaria* and *C. remanei* clearly showed the presence of two distinct populations: one with a content (around 38% GC) typical of nematodes, and another with a content (around 70% GC) typical of GC-rich Actinobacteria (Fig. 1B).

We used the 16S ribosomal RNA (16S rRNA) of a representative of these bacteria (*Leifsonia xyli*) to search the *Caenorhabditis* wgs sequences, and found highly similar sequences in *C. angaria*, *C. remanei*, and *C.* sp. 11. These sequences where only distantly related to authentic nematode rRNAs (~50% identity), but had high similarity (>95% identity) in the reference RNA sequence database with the 16S rRNAs of various species belonging to the genus *Leucobacter* (class: Actinobacteria, family: Microbacteriaceae). The phylogenetic reconstruction of the 16S rRNA allowed the identification of the bacterial DNA at the species level (Fig. 2). The sequence found in *C.* sp. 11 was

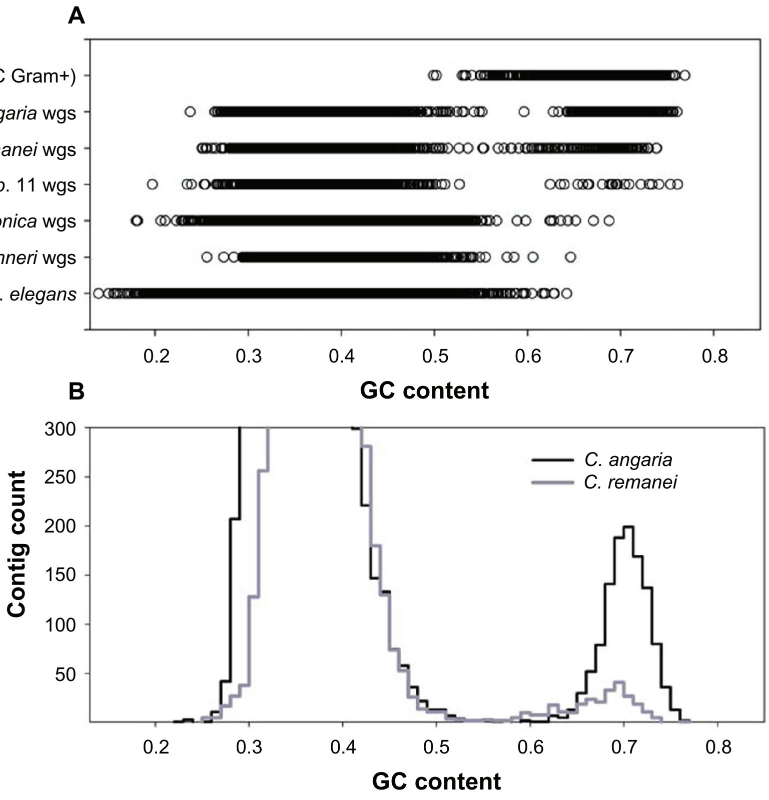

Figure 1. GC content of *Caenorhabditis* wgs sequences. (**A**) Horizontal plot comparing the GC content of individual wgs contigs ≥1 Kbp. from *C. brenneri* (11,925), *C. japonica* (34,475), *C.* sp.11 (6,678), *C. remanei* (11,919), and *C. angaria* (19,618) with the GC content of artificial non-overlapping 1 Kbp. contigs extracted from the complete genomes of *C. elegans* (100,264) and the high GC Gram+ bacterium *Leifsonia xyli* (2,584). (**B**) Frequency distribution of *C. remanei* and *C. angaria* wgs contigs showing the presence of two separated populations with different GC content.

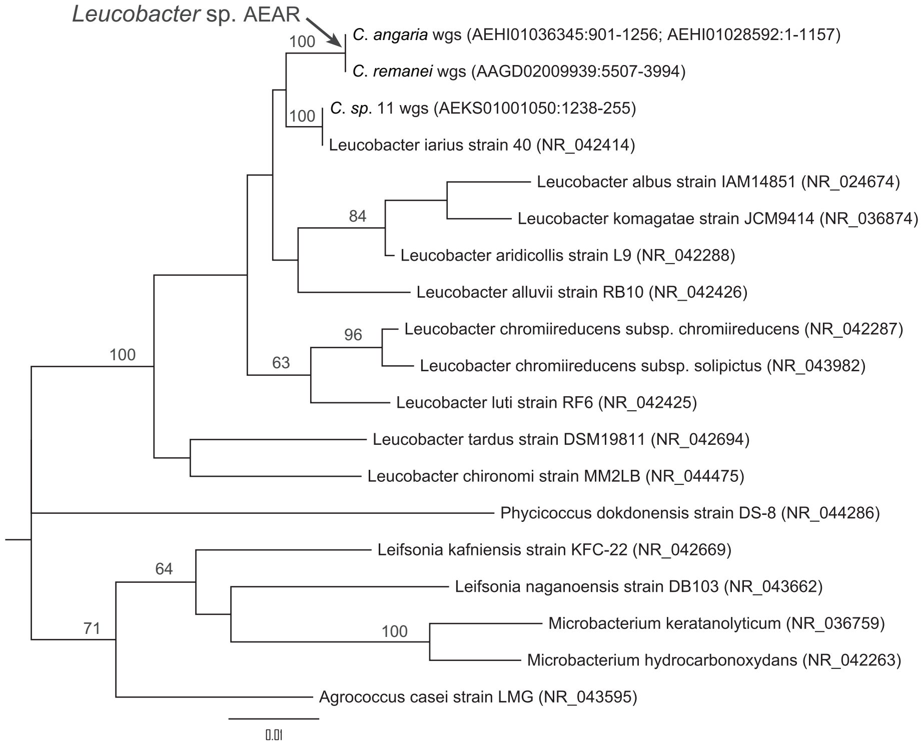
Leucobacter sp. AEAR
100
C. angaria wgs (AEHI01036345:901-1256; AEHI01028592:1-1157)
C. remanei wgs (AAGD02009939:5507-3994)
100
C. sp. 11 wgs (AEKS01001050:1238-255)
Leucobacter iarius strain 40 (NR_042414)
Leucobacter albus strain IAM14851 (NR_024674)
84
Leucobacter komagatae strain JCM9414 (NR_036874)
Leucobacter aridicollis strain L9 (NR_042288)
Leucobacter alluvii strain RB10 (NR_042426)
96
Leucobacter chromiireducens subsp. chromiireducens (NR_042287)
100
63
Leucobacter chromiireducens subsp. solipictus (NR_043982)
Leucobacter luti strain RF6 (NR_042425)
Leucobacter tardus strain DSM19811 (NR_042694)
Leucobacter chironomi strain MM2LB (NR_044475)
Phycicoccus dokdonensis strain DS-8 (NR_044286)
Leifsonia kafniensis strain KFC-22 (NR_042669)
64
Leifsonia naganoensis strain DB103 (NR_043662)
71
100
Microbacterium keratanolyticum (NR_036759)
Microbacterium hydrocarbonoxydans (NR_042263)
Agrococcus casei strain LMG (NR_043595)
0.01

Figure 2. Species identification of the actinobacterial DNA found in *Caenorhabditis* wgs sequences.
Notes: Shown is the midpoint-rooted maximum likelihood tree of 16S rRNA sequences of selected Actinobacteria together with the 16S rRNA sequences identified in *C. angaria*, *C. remanei*, and *C.* sp. 11. The scale bar represents a 1% difference in nucleotide sequence. Tree leaves are denoted by species names followed by the accession numbers of the Genbank entries (in parenthesis) used for the phylogenetic reconstruction. Bootstrap support higher than 50% is indicated at respective nodes. The leaf corresponding to the newly identified *Leucobacter* sp. AEAR is indicated by an arrow. The complete 16S rRNA sequence from *C. angaria* wgs sequences was obtained by joining two overlapping contigs.

found to be identical to *L. iarius*, while the sequences of *C. angaria and C. remanei*, though identical to each other, were not attributable to any of the known 16S rRNAs, also when partial sequences of uncultured *L.* sp. were included in the comparison (Fig. S2). The perfect identity of the 16S rRNAs found in the wgs sequences of the two *Caenorhabditis* species (1514/1514 bp) rules out the possibility that the differences observed with other 16S rRNAs derive from sequencing errors or the chimeric assembly of DNA from different sources, and supports the assignment of the bacterial DNA to a novel *Leucobacter* species closely related to *L. iarius* (1492/1508 bp). We tentatively name this species *Leucobacter* sp. AEAR. As evident from an all versus all comparison of wgs sequences (Fig. S3), the similarity between the bacterial DNA fraction found in *C. angaria* and *C. remanei* extends outside the 16S rRNA. However, the two DNA are not perfectly identical in other genomic regions (statistical mode: 99% similarity), suggesting that they derive from different strains of the same species.

Assembly of the *Leucobacter* sp. AEAR genome

To gain insights into the association between the identified microbe and *Caenorhabditis* species, and to segregate contigs of bacterial and eukaryotic origin, we decided to obtain a draft genome of the *Leucobacter* sp. AEAR using available data. Although most of the information on the bacterial genome was present in the *C. angaria* sequences, as could be judged from the analysis of the ribosomal proteins and the size of the contig fraction with high GC content (Table 1), this information was spread in many short contigs (N50 = 1344). We thus recurred to the original reads with the aim to improve the assembly of the bacterial fraction. We performed separated Velvet assemblies[9] of the different sequencing runs, and found the presence of the *Leucobacter* DNA in a single experiment (SRR065714), consisting of 8,887,206 single reads of 76 bp. (over a total of 157,491,568 single/paired reads).

The presence of the exogenous DNA in this sequencing run was evident from the near sequence identity

Table 1. Statistics of the high-GC contigs in different assemblies.

Assembly (GC > 0.55)[a]	**Size**	**Num. contigs**	**N50**
Wgs	3,435,010	4157	1344
SRR065714 + automatic cutoffs[b]	3,634,456	3651	2283
SRR065714 + optimized cutoffs[c]	3,419,437	2052	2852
SRR065714 + optimized cutoffs[c] + long reads[d]	3,376,200	1960	3354

Notes: [a]Contigs with CG content > 0.55 obtained by Velvet with k-mer = 41; [b]exp_cov = 3.2, cov_cutoff = 2; [c]exp_cov=18, cov_cutoff = 9; [d]Sanger reads of *C. remanei.*

of a number of contigs with the high-GC fraction of *C. remanei* and also from the analysis of k-mer coverage distribution (Fig. 3). This analysis revealed a bimodal distribution with frequency peaks around 3× and 18× coverage (Fig. 3A). Because the genome size of the bacterium is expected to be much smaller than that of *Caenorhabditis*, the DNA fraction with greater coverage was suspected to be of bacterial origin. Consistent with this supposition, the plotting of the cumulative GC content over the coverage distribution evidenced an unusually high GC content for the contig fraction with greater coverage (Fig. 3B), and the contigs assigned by MGTAXA to Actinobacteria showed a density peak in correspondence with the fraction with high coverage (Fig. 3C). These figures served for the choice of the optimal parameters for the assembly of the bacterial fraction. Using the same k-mer length (41) of the *C. angaria* wgs assembly[10] and the N50 of high-GC contigs as a proxy measure of the quality of the bacterial assembly, an improved assembly was obtained (N50 = 2852) with coverage cutoff values corresponding to the peak of the bacterial density distribution. A further improvement (N50 = 3354) was obtained by using the Sanger sequences of *C. remanei* as long reads to help the solving of repetitive regions (see Table 1).

Not surprisingly, the vast majority (91%) of the contigs obtained with the optimized assembly were assigned to bacteria, with 95% of the bacterial fraction assigned to Actinobacteria (Fig. S4). We selected 1368 contigs assigned to Actinobacteria, plus an additional 53 contigs having highly significant similarity at the nucleotide level (E < 1e-20) with the draft genome of the congeneric *L. chromiiresistens,*[11] obtaining a total of 3,227,180 bp. These contigs were scaffolded using mate pair information provided by the *C. remanei* reads and synteny with *L. chromiiresistens*, thus linking and orienting 619 contigs in 36 scaffolds with the largest scaffold (1.7 Mb) containing >50% of the base pairs. In the final assembly (Supplementary information), we additionally included unlinked contigs with >1000 bp. for a total of 3.2 Mbp, a genome size near to the size (3.3 Mbp) of the *L. chromiiresistens* genome.

We used the draft genome as a reference for the identification of nucleotide sequences of the public database that were misannotated as metazoan sequences. Through homology searches, we identified 75 reference sequences from *C. remanei*, and 229 and 3069 wgs

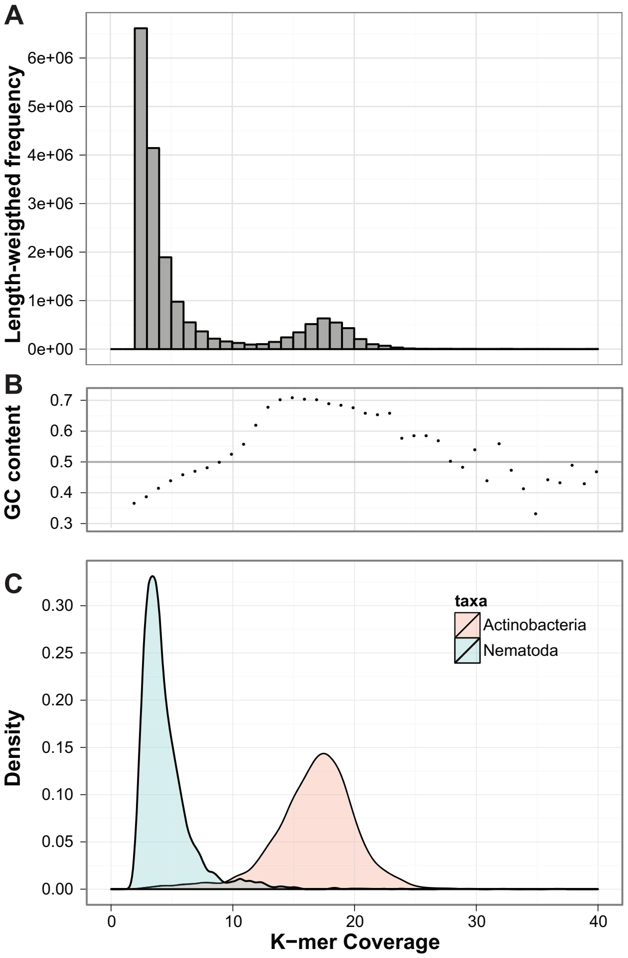

Figure 3. Features of the k-mer coverage distribution in *C. angaria* genomic reads (SRR065714). (**A**) Length-weighted count of the k-mer (k = 41) coverage of contigs assembled by Velvet with automatic cut-off values. (**B**) Cumulative GC content of the contigs grouped by their k-mer coverage using a bin width of 1. (**C**) Density distribution over the k-mer coverage of the contigs assigned by MGTAXA to Nematoda (n = 2900) or Actinobacteria (n = 1810).

contigs from *C. remanei* and *C. angaria*, respectively, having >95% identity over an alignment length > 100 bp. with the bacterial assembly (Supplementary information). No such hits were found in other eukaryotic genomes besides 6 sequences in *Anopheles gambiae* and 19 in *Caenorhabditis* sp. 11, most likely attributable to the closely related *L. iarius* (see Fig. 2). We found that the 99.8% of the '*Caenorhabditis'* contigs with local similarity to the *L.* sp. AEAR genome had an overall GC content > 0.55 suggesting that the presence of chimeric contigs in the wgs (and our) assemblies is negligible.

Features of the *Leucobacter sp*. AEAR genome

The automatic annotation[12] of the *Leucobacter* sp. AEAR draft genome identified 2826 genes, of which 2778 encode a protein product. The GC content of the protein coding genes (70.2%) is similar to the overall genomic content (70.1%) and much higher than the GC content (58.1%) of the genes producing functional RNAs. As expected, the usage of codons in protein coding gens is very biased, with a strong prevalence of GCending (90.2%) and GC-starting (71.2%) triplets (Fig. S5A). Consequently, the use of codons containing neither C nor G is extremely low, with the leucine UUA codon (which occurs 49 times in 40 genes) being the rarest one, similar to other actinobacterial species.[13]

Genome annotation provides evidence that the draft assembly contains a nearly complete genetic information of the microorganism. Of the 123 protein coding genes included in the "minimal" gene set of Actinobacteria,[14] eight were not found in the *L.* sp. AEAR genome. Noticeably, three of those missing genes encode proteins involved in the de novo purine nucleotide biosynthesis, a pathway that appears to be absent in this organism (see below). Among the genes producing functional RNAs, the annotation revealed a complete ribosomal RNA (rRNA) operon, with rRNAs 5S (144 bp.), 23S (3033 bp.), and 16S (1514 bp.), and a set of 44 tRNAs (41 in *L. chromiiresistens*) accounting for the entire decoding capacity after the addition of two tRNA genes found in short contigs and not included in the assembly (Fig. S5B). Given that RNA genes are often present in repeated units which are not resolved in short-read assemblies, we used read coverage values to estimate the copy number of those genes.[15] Using this criterion, the rRNA operon (54× coverage) is expected to be present in three copies, whereas most tRNA genes are estimated to be present in a single copy (see Fig S5B). A balanced tRNA gene set with limited redundancy further supports the notion that codon bias in this organism is shaped by directional mutational pressure rather than translational selection.[16,17]

A general function could not be determined by bioinformatic analysis for 738 "hypothetical proteins" (37%) encoded in the *L.* sp. AEAR genome. Of the remaining 2040 proteins, 933 were assigned to specific biochemical pathways or functional roles (ie, "subsystems"). No genes involved in motility and chemotaxis are found in the *L.* sp. AEAR genome. Likewise, the genome does not appear to encode functions related to spore formation. It contains several genes involved in oxygen respiration, comprising terminal cytochrome *d* and cytochrome *c* oxidases (EC 1.10.3.- and 1.9.3.1, respectively) and genes involved in oxidative stress such as catalase (EC 1.11.1.6) and superoxide dismutase (EC 1.15.1.1). Conversely, there are no genes such as nitrate reductase, dimethyl sulfoxide reductase, fumarate reductase, which enable anaerobic respiration in facultative aerobe Actinobacteria (eg, *Propionibacterium acnes*).[18] The genome contains genes involved in amino acid metabolism suggesting the presence of biosynthetic pathways for all the standard proteinogenic amino acids. Conversely, no genes are present in the subsystems of de novo purine and pyrimidine biosyntheses, suggesting that, at variance with most Actinobacteria (including the congeneric *L. chromiiresistens*), *L.* sp. AEAR lacks pathways for the biosynthesis of nucleobases.

The comparison of the gene distribution in the different subsystems highlights similarities and peculiarities of *L.* sp. AEAR with respect to other Actinobacteria (Fig. 4 and Table S2). The *L.* sp. AEAR genome is significantly enriched in genes involved in amino acid metabolism and in genes involved in membrane transport. The first feature, reflecting the abundance of biosynthetic and degradative pathways for amino acids, is probably characteristic of the genus because it is shared with *L. chromiiresistens*, whereas the second feature, reflecting an extreme abundance of ATP-binding cassette (ABC) transporters for di- and oligo-peptides, appears to be distinctive of the AEAR species (see Table S2). On the other hand, the *L.* sp.

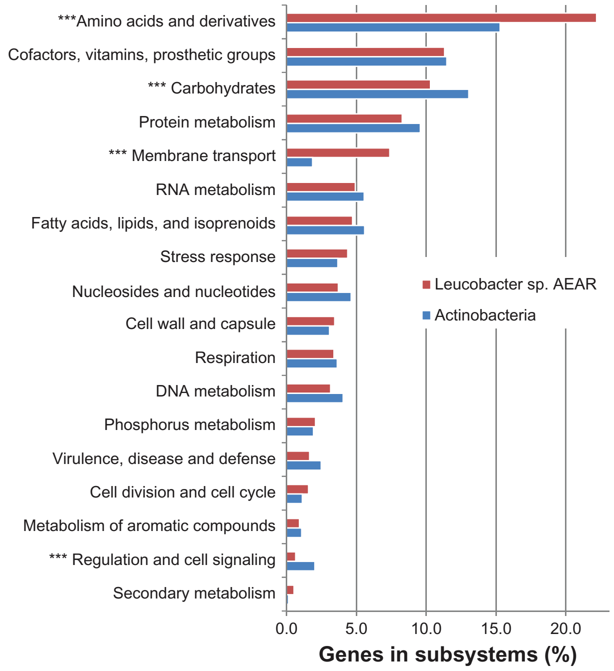

Figure 4. Functional category distribution of *Leucobacter* sp. AEAR genes compared to Actinobacteria. Genes are assigned to functional subsystems according to the RAST annotation.[12]

Notes: The percentage of genes assigned to each category in *L.* sp. AEAR is compared to the average of ten species representing nine different families of Actinobacteria (as listed in Table S2). Categories significantly different ($P < 0.001$) from the mean according to the one-sample Student's *t*-test are indicated by a triple asterisk.

AEAR genomes is significantly depleted of genes involved in carbohydrate metabolism and cell signaling and regulation. The percentage of genes in the nucleotide metabolism subsystem is not significantly reduced because the absence of gens for purine and pyrimidine biosyntheses is compensated by the abundance of genes for nucleobase uptake and salvage.

Finally, following the observation that initiated this study, the genes involved in purine degradation were identified using a custom search procedure (see Methods). *L.* sp. AEAR does not possess genes for purine oxidation and is thus not able to produce the allantoin intermediate. It contains, however, a metabolic pathway (absent in *L. chromiiresistens*) for the utilization of both *R* and *S* enantiomers of allantoin (Fig. S6), comprising an allantoin racemase,[19] a metal-dependent allantoinase,[20] and a novel type of allantoicase (R. Percudani, unpublished). The urea produced by allantoin degradation is not directly converted to ammonia (urease is missing in the genome), but carboxylated to allophanate for subsequent hydrolysis, similar to the pathway that enable *Saccharomyces cerevisiae* to use allantoin as a nitrogen source.[21,22]

Discussion

This study was motivated by the accidental discovery of DNA of apparently recent bacterial origin in *Caenorhabditis* species. The evidence was puzzling at first because the best similarity for the sequences identified in a *Caenorhabditis* species (*C. remanei*) was found in another *Caenorhabditis* species (*C. angaria*), as one would expect for affirmed *Caenorhabditis* sequences. Further analyses established, however, that the rationale for the observation was not the horizontal transfer of bacterial genes to the genome of a nematode ancestor, but the presence of exogenous DNA from the same bacterial species in two different *Caenorhabditis* species.

The exogenous DNA has been assigned to a bacterium tentatively named *Leucobacter* sp. AEAR. With a similarity of $<$99% in the 16S rRNA to the most closely related species (*L. iarius*), the microorganism first described here at the molecular level can be considered as a novel species according to current taxonomic standards.[23] Interestingly, the related species *L. iarius* has been originally isolated from the bacterial flora associated with a soil nematode (*Steinernema thermophilum*).[24] This evidence, together with the finding of this species in two *Caenorhabditis* genomes of different origin and sequenced with different techniques, strongly suggest that *L.* sp. AEAR is an organism naturally associated to *Caenorhabditis* rather than an accidental contaminant.

L. sp. AEAR belongs to a large and variegated bacterial phylum, Actinobacteria, which has been subjected to several genomic studies (reviewed in[14]). However, the genus *Leucobacter* is poorly known at the cellular and molecular level, with a single draft genome (*L. chromiiresistens*) published only recently.[11] Although the species described here is unknown at the cellular level, the analysis of its draft genome makes it possible to predict biological properties and ecological traits of the organism. *L.* sp. AEAR is anticipated to be a non motile, non sporulating aerobic bacterium. It is, however, probably adapted to microaerobic conditions as suggested by the presence of a cytochrome *d* terminal oxidase, a respiratory protein which is known to predominate when *E. coli* cells are grown at low aeration.[25] It has an overdeveloped amino acid metabolism, with a high proportion of genes involved in biosynthetic and degradative pathways and an exceedingly high number of peptide transporters, a restricted carbohydrate metabolism, and a possible dependence on the external supply of nucleobases.

The gene complement (2826 genes) is very similar to related free-living species (eg, 2802 genes in *L. chromiiresistens*). This argues against the possibility that *L.* sp. AEAR is an intracellular mutualist (ie, endosymbiont) or parasite, because in Actinobacteria and other bacteria these lifestyles are typically associated with a great reduction of the number of genes.[26,27] An intracellular localization is also inconsistent with the finding of the DNA of this bacterium in only one of several sequencing runs of *C. angaria*. The available data instead supports the hypothesis that *L.* sp. AEAR is a free-living bacterium able to establish a non obligatory, extracellular association with *Caenorhabditis* species. Based on the existing evidence, it is difficult to determine whether the interaction is beneficial or detrimental to the nematode host. The latter possibility is suggested by the fact that another species of the same genus, *L. chromiireducens* subsp. solipictus, is known to cause uterine infections in *C. elegans,*[28] but it is questioned by the relatively low proportion of genes assigned to the virulence and disease category (see Fig. 4 and Table S2). On the other hand, some of the metabolic capabilities identified in the bacterial genome could improve the host fitness in its natural environment. Among these capabilities are the production of amino acids that are essential for the nematode (ten amino acids are nutritionally essential for *C. elegans*), and the production of usable nitrogen from a compound (allantoin) which is not accessible to the nematode although typically present in its natural environment. Experimental protocols established for the closely related *L. iarius*[24] and the molecular information provided by this work may enable the future isolation of the bacterium from its natural hosts and the elucidation of its interaction with *Caenorhabditis.*

From a methodological standpoint, the results presented here demonstrate the convenience of the application of techniques developed for metagenomic analysis to conventional genomics. When the sequencing involves obtaining collective data from an ecological niche, or a metagenome, specific techniques are required for the assembly and organism classification of DNA sequences of different origin. In contrast, organism source is generally considered prior knowledge in genomic sequencing. A known issue, however, is the frequent contamination of human DNA in genomic sequences of other organisms.[29,30] In other cases, the presence of DNA of heterogeneous origin can be anticipated because the sequencing involves an organism known to establish an obligatory association with another organism.[31] In the case of the *C. angaria* sequencing, the presence of *E. coli* DNA was expected because the bacterium is used to feed the nematode; all the *E. coli* sequences were then identified by homology and eliminated from the assembly.[10] (Intriguingly, we found *E. coli* DNA in all *C. angaria* sequencing runs except that containing *L.* sp. AEAR).

However, the presence of organism associations in genomic sequencing is not always recognized a priori, but needs to be determined a posteriori with unbiased methods of analysis. The results presented in this work can suggest some effective analyses for the determination of unexpected DNA in genomic sequences. The GC content analysis provided clear indications of the presence of DNA populations with distinct nucleotide compositions. The applicability of this analysis, however, is limited to cases in which the contaminant DNA has a very different GC content. A genome-wide taxonomic prediction method used in metagenomic analysis,[8] readily revealed the presence of substantial contaminations, but was less effective in the case of moderate contaminations (compare *C. angaria*, *C. remanei*, and other *Caenorhabditis* species in Figure S1). The analysis of 16S rRNA allowed the identification of contaminant DNA and also a species-level classification of the organism. However, 16S rRNA genes, which are often present as repeated units (see additional file in[26] for a comprehensive compilation), typically have high read coverage and can be found in the assembly also in cases of very modest contaminations (eg, *C.* sp. 11 in Figs. 2 and S1). Finally, homology searches with a reference set of vertically transmitted genes revealed moderate contamination and also provided an estimate of the contamination level (compare *C. angaria*, *C. remanei*, and other *Caenorhabditis* species in Table S1).

Once the presence of heterogeneous DNA in genomic reads has been assessed, taxonomic predictions applied to the read coverage distribution (see Fig. 3) can be used to target the assembly towards a particular DNA fraction. The method that we used, inspired by a metagenomic assembly procedure,[32] allowed us to improve the quality of the bacterial DNA assembly while excluding the DNA fraction of the eukaryotic host. In other cases a similar strategy could be used to exclude the contaminant bacterial DNA from the host genome assembly. As illustrated by the example shown here, however, the presence of DNA of unexpected origin in genomic sequences should not be regarded in principle as unwelcome information, but as additional data that can potentially improve the understanding of an organism's biology.

Acknowledgements

I thank the Centro di Calcolo di Ateneo of the University of Parma for computer time and Anastasia Liuzzi for help with genome analysis.

Funding

This work has been supported by the University of Parma.

Author Contributions

Conceived and designed the experiments: RP. Analyzed the data: RP. Wrote the first draft of the manuscript: RP. Contributed to the writing of the manuscript: RP. Agree with manuscript results and conclusions: RP. Jointly developed the structure and arguments for the paper: RP. Made critical revisions and approved final version: RP.

Competing Interests

The author discloses no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Ramazzina I, Costa R, Cendron L, et al. An aminotransferase branch point connects purine catabolism to amino acid recycling. *Nat Chem Biol.* Nov 2010;6(11):801–6.
2. Werner AK, Witte CP. The biochemistry of nitrogen mobilization: purine ring catabolism. *Trends Plant Sci.* Jul 2011;16(7):381–7.
3. Reinbothe H, Mothes K. Urea, ureides, and guanidines in plants. *Annu Rev Plant Phys*. 1962;13(1):129–49.
4. Kiontke KC, Felix MA, Ailion M, et al. A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evol Biol*. 2011;11:339.
5. Acuna R, Padilla BE, Florez-Ramos CP, et al. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc Natl Acad Sci U S A*. Mar 13, 2012;109(11):4197–202.
6. Brochier C, Philippe H, Moreira D. The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet*. Dec 2000;16(12):529–33.
7. Garcia-Vallve S, Simo FX, Montero MA, Arola L, Romeu A. Simultaneous horizontal gene transfer of a gene coding for ribosomal protein l27 and operational genes in Arthrobacter sp. *J Mol Evol*. Dec 2002;55(6):632–7.
8. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. Sep 2009;6(9):673–6.
9. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. May 2008;18(5):821–9.
10. Mortazavi A, Schwarz EM, Williams B, et al. Scaffolding a Caenorhabditis nematode genome with RNA-seq. *Genome Res*. Dec 2010;20(12):1740–7.
11. Sturm G, Buchta K, Kurz T, Rensing SA, Gescher J. Draft genome sequence of Leucobacter chromiiresistens, an extremely chromium-tolerant strain. *J Bacteriol*. Jan 2012;194(2):540–1.
12. Aziz RK, Bartels D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*. 2008;9:75.
13. Chater KF, Chandra G. The use of the rare UUA codon to define "expression space" for genes involved in secondary metabolism, development and environmental adaptation in streptomyces. *J Microbiol.* Feb 2008;46(1):1–11.
14. Ventura M, Canchaya C, Tauch A, et al. Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev*. Sep 2007;71(3):495–548.
15. Iben JR, Epstein JA, Bayfield MA, et al. Comparative whole genome sequencing reveals phenotypic tRNA gene duplication in spontaneous Schizosaccharomyces pombe La mutants. *Nucleic Acids Res*. Jun 2011;39(11):4728–42.
16. Shields DC, Sharp PM. Synonymous codon usage in Bacillus subtilis reflects both translational selection and mutational biases. *Nucleic Acids Res*. Oct 12, 1987;15(19):8023–40.
17. Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *J Mol Biol*. May 2, 1997;268(2):322–30.
18. Bruggemann H, Henne A, Hoster F, et al. The complete genome sequence of Propionibacterium acnes, a commensal of human skin. *Science*. Jul 30, 2004;305(5684):671–3.
19. French JB, Neau DB, Ealick SE. Characterization of the structure and function of Klebsiella pneumoniae allantoin racemase. *J Mol Biol*. Jul 15, 2011;410(3):447–60.
20. Kim K, Kim MI, Chung J, Ahn JH, Rhee S. Crystal structure of metal-dependent allantoinase from Escherichia coli. *J Mol Biol*. Apr 17, 2009;387(5):1067–74.
21. Cooper TG, Lam C, Turoscy V. Structural analysis of the dur loci in S. cerevisiae: two domains of a single multifunctional gene. *Genetics*. Mar 1980;94(3):555–80.

22. Wong S, Wolfe KH. Birth of a metabolic gene cluster in yeast by adaptive gene relocation. *Nat Genet.* Jul 2005;37(7):777–82.
23. Stackebrandt E, Ebers J. Taxonomic parameters revisited: tarnished gold standards. *Microbiol Today.* 2006;33:152–5.
24. Somvanshi VS, Lang E, Schumann P, et al. Leucobacter iarius sp. nov., in the family Microbacteriaceae. *Int J Syst Evol Microbiol*. Apr 2007;57(Pt 4): 682–6.
25. Cotter PA, Melville SB, Albrecht JA, Gunsalus RP. Aerobic regulation of cytochrome d oxidase (cydAB) operon *expression in Escherichia coli: roles of Fnr and ArcA in repression and activation. Mol Microbiol*. Aug 1997; 25(3):605–15.
26. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D. Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. *Biol Direct*. 2009;4:13.
27. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* Jan 2012;10(1):13–26.
28. Muir RE, Tan MW. Virulence of Leucobacter chromiireducens subsp. solipictus to Caenorhabditis elegans: characterization of a novel host-pathogen interaction. *Appl Environ Microbiol*. Jul 2008;74(13):4185–98.
29. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS One*. 2011;6(3): e17288.
30. Longo MS, O'Neill MJ, O'Neill RJ. Abundant human DNA contamination identified in non-primate genome databases. *PloS One*. 2011;6(2):e16410.
31. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)*. Mar 15, 2007;23(6):673–9.
32. Toshiaki N, Tsuyoshi H, Hideaki T, Yasubumi S. *MetaVelvet: An Extension of Velvet Assembler to De Novo Metagenome Assembly from Short Sequence Reads*. Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine; 2011; Chicago, Illinois.
33. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res*. Nov 11, 1994;22(22):4673–80.
34. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol.* 1981;17(6):368–76.
35. White JR, Roberts M, Yorke JA, Pop M. Figaro: a novel statistical method for vector sequence removal. *Bioinformatics (Oxford, England)*. Feb 15, 2008;24(4):462–7.
36. Kurtz S, Phillippy A, Delcher AL, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12.
37. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. Mar 1, 1997; 25(5):955–64.
38. Ardell DH, Andersson SG. TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Res*. 2006;34(3):893–904.

Supplementary Data

Lsp41_contigs.fa: Velvet contigs—optimized bacterial assembly.

Lsp41_scaffolds.fa: Bambus scaffolds + unlinked contigs (>1000 bp.).

Lsp41_reads.fa.gz: *C. angaria* (Illumina) and *C. remanei* (Sanger) reads in the bacterial assembly.

Lsp41_ncRNAs.fa: 16S rRNA, 23S rRNA, 5S rRNA, 44 tRNAs.

Lsp41_annotated.fa: 2778 annotated protein sequences.

C_remanei_refseq.blasttable: 75 *C. remanei* Refseq entries with >95% id. over >100 bp.

C_remanei_wgs.blasttable: 229 *C. remanei* wgs entries with >95% id. over >100 bp.

C_angaria_wgs.blasttable: 3069 *C. angaria* wgs entries with >95% id. over >100 bp.

Table S1. Blast hits of *Escherichia coli* ribosomal proteins in *Caenorhabditis* wgs sequences.

***E. coli* protein**	***C. angaria*[a]**	***C. remanei*[a]**	**Other *Caenorhabditis*[b]**
L1	1E-27 (AEHI01034406.1)	>0.01	>0.01
L11	2E-37 (AEHI01033769.1)	2E-12	4E-15
L13	9E-44 (AEHI01033983.1)	>0.01	7E-03
L14	3E-37 (AEHI01033293.1)	2E-07	2E-08
L16	3E-37 (AEHI01033293.1)	>0.01	>0.01
L18	7E-05 (AEHI01033934.1)	>0.01	>0.01
L22	4E-21 (AEHI01033058.1)	>0.01	>0.01
L3	5E-61 (AEHI01033448.1)	8E-64 (AAGD02011036.1)	1E-06
L5	2E-57 (AEHI01033293.1)	>0.01	>0.01
L6	3E-24 (AEHI01033293.1)	>0.01	>0.01
S11	2E-34 (AEHI01035048.1)	2E-10	1E-13
S12	8E-54 (AEHI01035015.1)	3E-06	8E-11
S13	4E-18 (AEHI01034574.1)	>0.01	>0.01
S15	6E-22 (AEHI01033268.1)	>0.01	>0.01
S17	9E-20 (AEHI01033293.1)	>0.01	>0.01
S2	7E-67 (AEHI01034302.1)	2E-67 (AAGD02009919.1)	4E-07
S3	9E-51 (AEHI01033293.1)	>0.01	>0.01
S4	3E-20 (AEHI01033940.1)	>0.01	>0.01
S5	2E-48 (AEHI01033571.1)	1E-08	5E-09
S7	7E-38 (AEHI01035015.1)	8E-13	2E-14
S8	3E-38 (AEHI01033293.1)	>0.01	>0.01
S9	1E-26 (AEHI01033983.1)	4E-27 (AAGD02006963.1)	>0.01

Notes: [a]Accession numbers of *Caenorhabditis* wgs entries having best reciprocal hits in bacteria are indicated in parenthesis, next to the E-values obtained by the best tblastn match of *E. coli* ribosomal proteins; [b]E-values of the best tblastn matches of *E. coli* ribosomal proteins in wgs contigs of *C. brenneri*, *C. japonica*, *C.* sp.11.

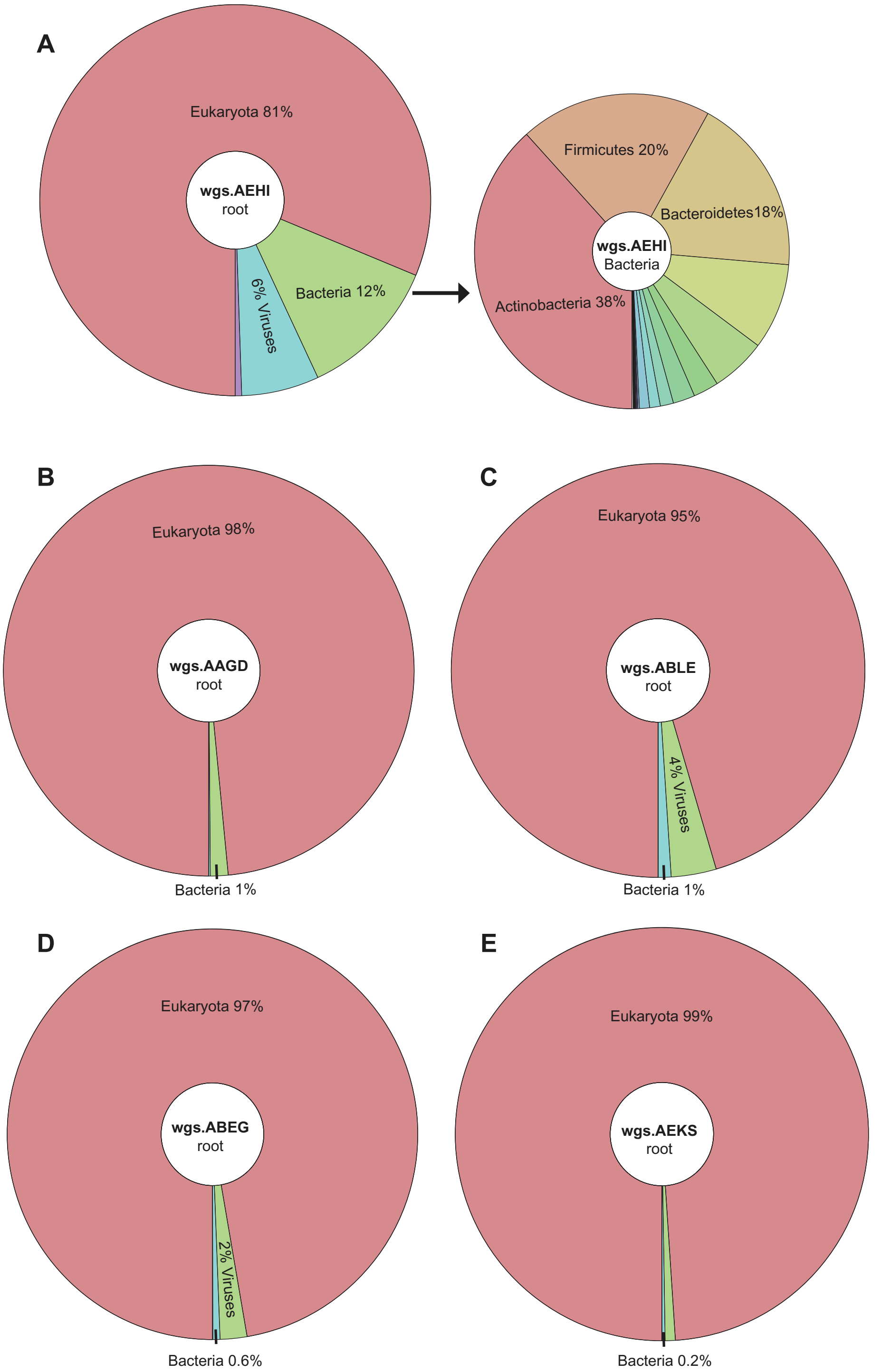

Figure S1. Krona charts of aggregated taxonomic assignments for *Caenorhabditis* wgs contigs. (**A**) *C. angaria* wgs contigs assigned by MGTAXA at the domain level; subdivision of the bacterial sequences in main phyla is shown in the smaller circle. (**B**) *C. remanei* wgs contigs. (**C**) *C. brenneri* wgs contigs. (**D**) *C. japonica* wgs contigs. (**E**) *C.* sp. 11 MAF-2010 wgs contigs.

Table S2. Gene distribution (%) in functional categories for *Leucobacter* sp. AEAR and other Actinobacteria.

Lifestyle[a]	***Leucobacter* sp. AEAR**	***Leucobacter chromiiresistens***	***Leifsonia xyli***	***Bifidobact. longum***
	N.D.	**FL—Aer**	**FHA—Aer**	**FHA—An**
Family	**Micro-bacteriaceae**	**Micro-bacteriaceae**	**Micro-bacteriaceae**	**Bifido-bacteriaceae**
Amino acids and derivatives	**22.22*****	21.52	12.91	15.59
Cofactors, vitamins, prost. groups	11.27	7.87	11.27	7.63
Carbohydrates	***10.23******	13.07	13.37	12.60
Protein metabolism	8.23	8.32	10.48	13.35
Membrane transport	**7.35*****	3.90	1.57	1.16
RNA metabolism	4.80	4.03	6.23	7.96
Fatty acids, lipids, and isoprenoids	4.64	5.85	7.08	1.82
Stress response	4.32	3.77	2.95	2.99
Nucleosides and nucleotides	3.68	4.94	4.91	6.30
Cell wall and capsule	3.44	3.06	4.52	4.06
Respiration	3.36	3.64	2.88	1.00
DNA metabolism	3.12	4.29	5.37	5.14
Miscellaneous	3.04	5.07	6.49	8.71
Phosphorus metabolism	2.08	1.69	1.83	2.99
Virulence, disease and defense	1.60	2.86	1.31	3.15
Cell division and cell cycle	1.52	1.37	1.44	1.49
Metabolism of aromatic compounds	0.88	1.04	0.07	0.08
Sulfur metabolism	0.88	0.78	0.39	0.66
Iron acquisition and metabolism	0.72	0.46	0.33	0.00
Potassium metabolism	0.64	0.39	0.66	1.33
Regulation and cell signaling	***0.64******	0.59	2.10	1.49
Secondary metabolism	0.48	0.39	0.00	0.08
Nitrogen metabolism	0.48	0.39	0.46	0.00
Dormancy and sporulation	0.24	0.13	0.13	0.17
Phages, prophages, transposable elements	0.16	0.26	0.20	0.25
Photosynthesis	0.00	0.00	0.00	0.00
Motility and chemotaxis	0.00	0.33	1.05	0.00

Continues in the next page

Continued from previous page

Propionibact. acnes	***Mycobact. tuberculosis***	***Nocardia farcinica***	***Corynebact. glutamicum***	***Streptomyces avermitilis***	***Frankia* sp. EAN1pec**	***Thermobifida fusca* YX**
FL—An/micr	**FHA—Aer**	**FL—Aer**	**FL—Aer/An**	**FL—Aer**	**FHA—Aer**	**FL—Aer**
Propioni-bacteriaceae	**Myco-bacteriaceae**	**Nocardiaceae**	**Coryne-bacteriaceae**	**Strepto-mycetaceae**	**Frankiaceae**	**Nocardio-psaceae**
13.04	14.29	15.86	13.26	15.59	14.41	13.26
15.97	13.46	12.36	12.30	10.69	12.70	14.62
14.16	10.22	12.14	11.45	16.77	15.38	14.77
11.04	8.96	7.11	9.76	7.54	8.19	9.53
1.75	1.62	0.93	2.65	1.27	0.93	1.92
6.55	4.37	4.32	6.47	4.27	4.32	5.14
3.43	7.34	7.98	4.40	6.78	7.80	5.70
2.31	3.80	5.19	3.66	4.18	2.87	3.53
5.43	3.63	3.27	4.51	3.84	3.00	3.98
2.62	2.58	1.70	3.61	2.45	3.13	2.87
4.55	4.89	4.71	3.29	3.66	4.93	4.64
3.12	3.89	3.07	4.19	3.18	3.80	3.83
6.30	6.82	6.85	7.90	7.51	6.22	6.55
1.75	2.32	1.41	1.80	1.33	0.48	1.71
1.68	3.76	3.04	1.75	2.00	1.52	1.56
1.12	0.92	0.64	1.17	0.79	0.97	1.16
0.06	0.66	2.88	2.33	1.30	1.55	0.05
0.69	0.96	2.08	1.59	1.85	2.16	1.51
0.25	0.17	0.32	0.64	0.70	0.71	0.30
1.19	0.66	0.83	0.11	0.70	0.00	0.71
2.00	3.71	1.89	2.07	2.00	2.32	2.02
0.12	0.00	0.13	0.21	0.12	1.23	0.00
0.75	0.74	0.74	0.64	1.03	0.55	0.35
0.12	0.09	0.16	0.16	0.36	0.23	0.20
0.00	0.17	0.35	0.05	0.03	0.03	0.05
0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.03	0.05	0.06	0.58	0.05

Notes: ***Significantly higher (Bold) or lower (Bold italics) than the average value ($P < 0.001$; one-sample Student's *t*-test).
Abbreviations: [a]FL, free-living; FHA, free-living/host associated; Aer, aerobic; An, anaerobic; Micr, microaerobic; N.D., not determined.

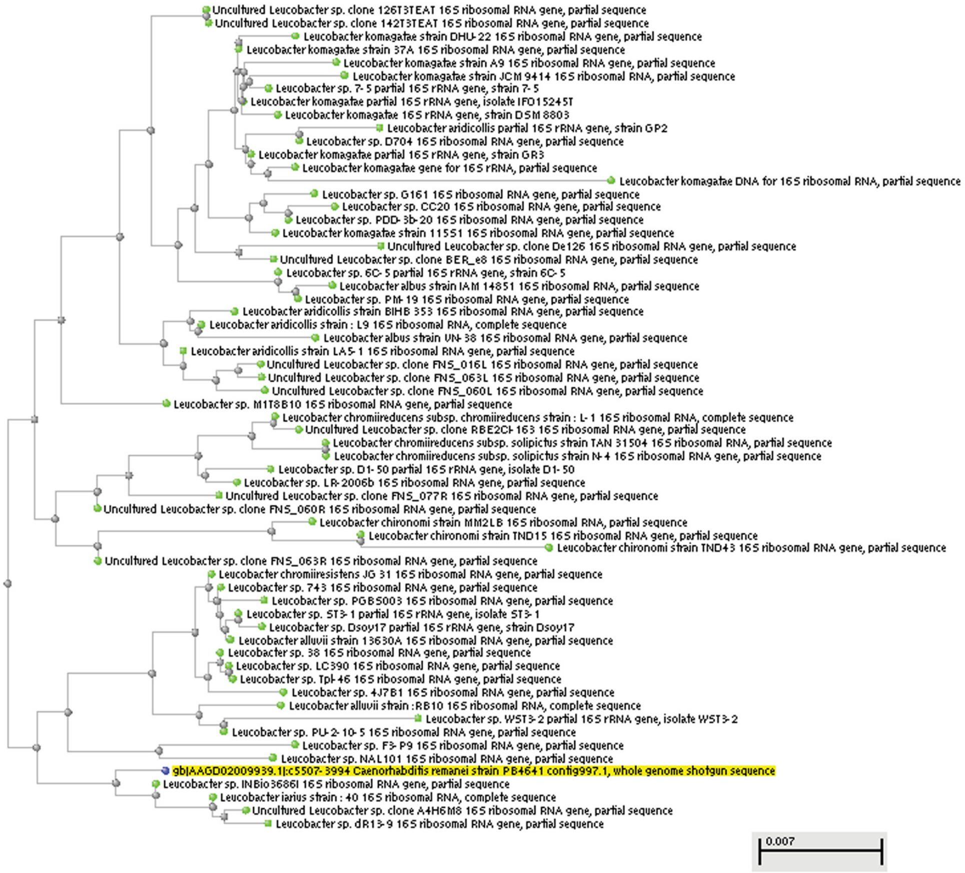
Uncultured Leucobacter sp. clone 126T3TEAT 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone 142T3TEAT 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae strain DHU-22 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae strain 37A 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae strain A9 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae strain JCM 9414 16S ribosomal RNA, partial sequence
Leucobacter sp. 7-5 partial 16S rRNA gene, strain 7-5
Leucobacter komagatae partial 16S rRNA gene, isolate IFO15245T
Leucobacter komagatae 16S rRNA gene, strain DSM 8803
Leucobacter aridicollis partial 16S rRNA gene, strain GP2
Leucobacter sp. D704 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae partial 16S rRNA gene, strain GR3
Leucobacter komagatae gene for 16S rRNA, partial sequence
Leucobacter komagatae DNA for 16S ribosomal RNA, partial sequence
Leucobacter sp. G161 16S ribosomal RNA gene, partial sequence
Leucobacter sp. CC20 16S ribosomal RNA gene, partial sequence
Leucobacter sp. PDD-3b-20 16S ribosomal RNA gene, partial sequence
Leucobacter komagatae strain 115S1 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone De126 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone BER_e8 16S ribosomal RNA gene, partial sequence
Leucobacter sp. 6C-5 partial 16S rRNA gene, strain 6C-5
Leucobacter albus strain IAM 14851 16S ribosomal RNA, partial sequence
Leucobacter sp. PM-19 16S ribosomal RNA gene, partial sequence
Leucobacter aridicollis strain BIHB 353 16S ribosomal RNA gene, partial sequence
Leucobacter aridicollis strain : L9 16S ribosomal RNA, complete sequence
Leucobacter albus strain UN-38 16S ribosomal RNA gene, partial sequence
Leucobacter aridicollis strain LA5-1 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_016L 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_063L 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_060L 16S ribosomal RNA gene, partial sequence
Leucobacter sp. M1T8B10 16S ribosomal RNA gene, partial sequence
Leucobacter chromiireducens subsp. chromiireducens strain : L-1 16S ribosomal RNA, complete sequence
Uncultured Leucobacter sp. clone RBE2CI-163 16S ribosomal RNA gene, partial sequence
Leucobacter chromiireducens subsp. solipictus strain TAN 31504 16S ribosomal RNA, partial sequence
Leucobacter chromiireducens subsp. solipictus strain N-4 16S ribosomal RNA gene, partial sequence
Leucobacter sp. D1-50 partial 16S rRNA gene, isolate D1-50
Leucobacter sp. LR-2006b 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_077R 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_060R 16S ribosomal RNA gene, partial sequence
Leucobacter chironomi strain MM2LB 16S ribosomal RNA, partial sequence
Leucobacter chironomi strain TND15 16S ribosomal RNA gene, partial sequence
Leucobacter chironomi strain TND43 16S ribosomal RNA gene, partial sequence
Uncultured Leucobacter sp. clone FNS_063R 16S ribosomal RNA gene, partial sequence
Leucobacter chromiiresistens JG 31 16S ribosomal RNA gene, partial sequence
Leucobacter sp. 743 16S ribosomal RNA gene, partial sequence
Leucobacter sp. PGBS003 16S ribosomal RNA gene, partial sequence
Leucobacter sp. ST3-1 partial 16S rRNA gene, isolate ST3-1
Leucobacter sp. Dsoy17 partial 16S rRNA gene, strain Dsoy17
Leucobacter alluvii strain 13630A 16S ribosomal RNA gene, partial sequence
Leucobacter sp. 38 16S ribosomal RNA gene, partial sequence
Leucobacter sp. LC390 16S ribosomal RNA gene, partial sequence
Leucobacter sp. TPI-46 16S ribosomal RNA gene, partial sequence
Leucobacter sp. 4J7B1 16S ribosomal RNA gene, partial sequence
Leucobacter alluvii strain :RB10 16S ribosomal RNA, complete sequence
Leucobacter sp. WST3-2 partial 16S rRNA gene, isolate WST3-2
Leucobacter sp. PU-2-10-5 16S ribosomal RNA gene, partial sequence
Leucobacter sp. F3-P9 16S ribosomal RNA gene, partial sequence
Leucobacter sp. NAL101 16S ribosomal RNA gene, partial sequence
gb|AAGD02009939.1|:c5507-3994 Caenorhabditis remanei strain PB4641 contig997.1, whole genome shotgun sequence
Leucobacter sp. INBio3686I 16S ribosomal RNA gene, partial sequence
Leucobacter iarius strain : 40 16S ribosomal RNA, complete sequence
Uncultured Leucobacter sp. clone A4H6M8 16S ribosomal RNA gene, partial sequence
Leucobacter sp. dR13-9 16S ribosomal RNA gene, partial sequence
0.007

Figure S2. Distance tree of results of the Blast search with the *C. remanei* 16S rRNA in the nr database.
Notes: The tree is based on Blast pairwise alignments between the query sequence (highlighted in yellow) and the hits obtained through a homology search against all the *Leucobacter* sequences in the nr database. A partial sequence of a *Leucobacter* species (sp. INBio2553H, sequence accession HM771025) associated with the guts of beetle larvae was found identical to both the *C. remanei* and *L. iarius* sequences.

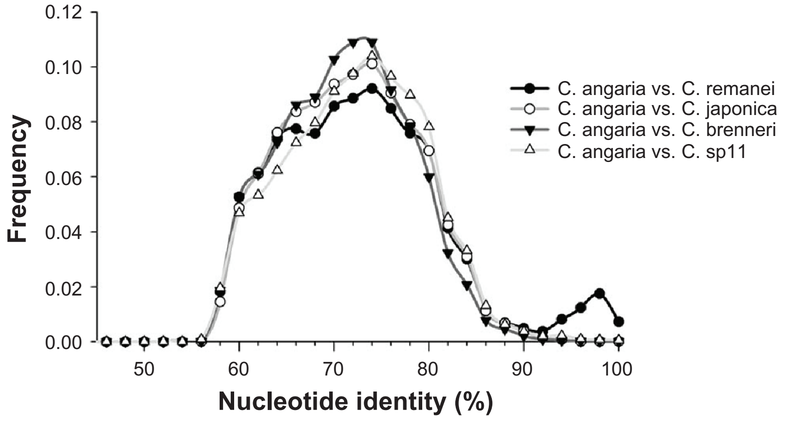

Figure S3. Frequency distribution of pairwise nucleotide identity in local alignments of *C. angaria* contigs with other *Caenorhabditis* wgs contigs of *C. angaria* were split in segments of 120 bp. and compared using blastn (r = 5; q = –4) with wgs contigs of other *C.* species.
Note: The sequence identity distribution of significant best hits ($E < 10^{-3}$) with an alignment length ≥ 100 is shown with a 2% bin width.

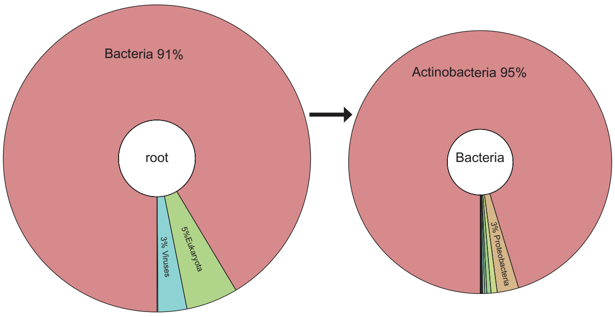

Figure S4. Krona charts of aggregated taxonomic assignments for the bacterialoptimized SRR065714 assembly.
Notes: Velvet contigs obtained with optimized coverage cutoffs assigned by MGTAXA at the domain level; subdivision of the bacterial sequences in main phyla is shown in the smaller circle.

A

```
UUU  0.4(   272)  UCU  0.8(   642)  UAU  1.3(  1023)  UGU  0.2(   153)
UUC 31.6( 24379)  UCC 13.2( 10150)  UAC 16.3( 12547)  UGC  5.1(  3953)
UUA  0.1(    49)  UCA  1.1(   837)  UAA  0.3(   264)  UGA  2.4(  1816)
UUG  1.5(  1184)  UCG 24.2( 18705)  UAG  0.9(   720)  UGG 13.3( 10284)

CUU  2.0(  1572)  CCU  1.6(  1202)  CAU  2.5(  1900)  CGU  6.5(  5007)
CUC 58.8( 45365)  CCC 22.7( 17503)  CAC 17.4( 13427)  CGC 44.3( 34149)
CUA  0.5(   420)  CCA  1.4(  1089)  CAA  1.2(   895)  CGA  6.3(  4841)
CUG 39.9( 30771)  CCG 28.2( 21779)  CAG 27.0( 20868)  CGG 19.4( 14956)

AUU  0.9(   693)  ACU  1.5(  1120)  AAU  1.3(   990)  AGU  1.1(   838)
AUC 45.2( 34913)  ACC 33.0( 25501)  AAC 17.1( 13227)  AGC 15.3( 11808)
AUA  0.2(   186)  ACA  1.1(   836)  AAA  1.2(   910)  AGA  0.6(   489)
AUG 16.5( 12697)  ACG 19.2( 14851)  AAG 19.8( 15290)  AGG  1.6(  1233)

GUU  2.3(  1790)  GCU  6.1(  4694)  GAU 12.9(  9969)  GGU 10.0(  7701)
GUC 45.1( 34774)  GCC 54.8( 42289)  GAC 44.4( 34293)  GGC 56.3( 43469)
GUA  2.0(  1516)  GCA  8.0(  6151)  GAA 10.4(  8055)  GGA  9.8(  7571)
GUG 36.9( 28510)  GCG 64.9( 50076)  GAG 50.4( 38928)  GGG 17.6( 13589)
```

B

```
Phe  AAA    - | Ser  AGA    - | Tyr  ATA    - | Cys  ACA    - |
Phe  GAA    1 | Ser  GGA    1 | Tyr  GTA    1 | Cys  GCA    1 |
Leu  TAA    1 | Ser  TGA    1 | stop TTA    - | stop TCA    - |
Leu  CAA    1 | Ser  CGA    - | stop CTA    - | Trp  CCA    1 |

Leu  AAG    - | Pro  AGG    - | His  ATG    - | Arg  ACG    1 |
Leu  GAG    1 | Pro  GGG    1 | His  GTG    1 | Arg  GCG    - |
Leu  TAG    1 | Pro  TGG    1 | Gln  TTG    1 | Arg  TCG    - |
Leu  CAG    - | Pro  CGG    1 | Gln  CTG    2 | Arg  CCG    1 |

Ile  AAT    - | Thr  AGT    - | Asn  ATT    - | Ser  ACT    - |
Ile  GAT    1 | Thr  GGT    1 | Asn  GTT    1 | Ser  GCT    1 |
Ile  TAT    - | Thr  TGT    1 | Lys  TTT    1 | Arg  TCT    1 |
Met  CAT*   4 | Thr  CGT    1 | Lys  CTT    2 | Arg  CCT    1 |

Val  AAC    - | Ala  AGC    - | Asp  ATC    - | Gly  ACC    - |
Val  GAC    1 | Ala  GGC    1 | Asp  GTC    1 | Gly  GCC    1 |
Val  TAC    1 | Ala  TGC    1 | Glu  TTC    1 | Gly  TCC    1 |
Val  CAC    1 | Ala  CGC    - | Glu  CTC    1 | Gly  CCC    2 |
```

Figure S5. Codon and anticodon usage in *Leucobacter* sp. AEAR. (**A**) Codon frequencies inferred from the protein coding sequences identified in the draft genome of *L.* sp AEAR. The codon frequency per thousand is reported next to each triplet followed by the absolute frequency of the codon (in parenthesis); (**B**) anticodon frequencies inferred from the tRNA genes identified by tRNAscan.
Notes: The tRNAs with CAT* anticodon comprise two tRNAf^{Met} initiators, a tRNA^{Met} elongator, and a tRNA^{Ile} with 2-lysyl cytidine at the wobble position (reading the TAT codon), as determined by the tFAM program.

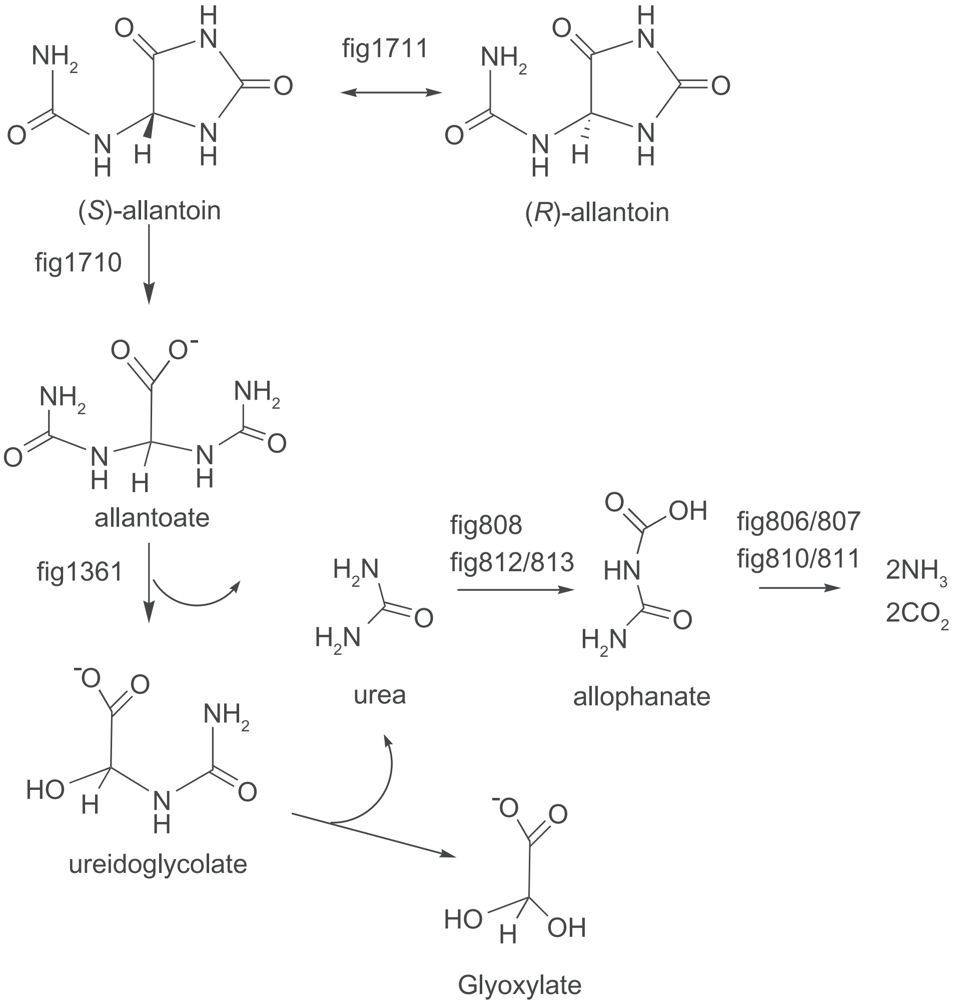

Figure S6. Putative allantoin utilization pathway in *L.* sp. AEAR as reconstructed from the draft genome.
Notes: Genes assigned to enzymatic steps are indicated with RAST numbers; the corresponding sequences are reported in the Supplementary files. No genes where assigned to the conversion of ureidoglycolate to glyoxylate; this reaction, however, is known to occur also in the absence of a protein catalyst.