Supplementary Information – Sharon *et al.*, 2014

Table of contents

Table of figures

Table of tables

Table S1: sequencing statistics for the short- and long-read data. Read length for the short-read data is 150 bp, numbers in the table are for trimmed reads.

	Short reads			**Synthetic long reads**		
Sample	**# reads**	**N50**	**Amount**	**# reads**	**N50**	**Amount**
4 m	263,119,764	135	33.5 gbp	70,342	8,229	490 mbp
5 m	497,853,726	125	49.6 gbp	86,934	7,962	534 mbp
6 m	191,665,440	133	23.8 gbp	67,592	8,222	450 mbp

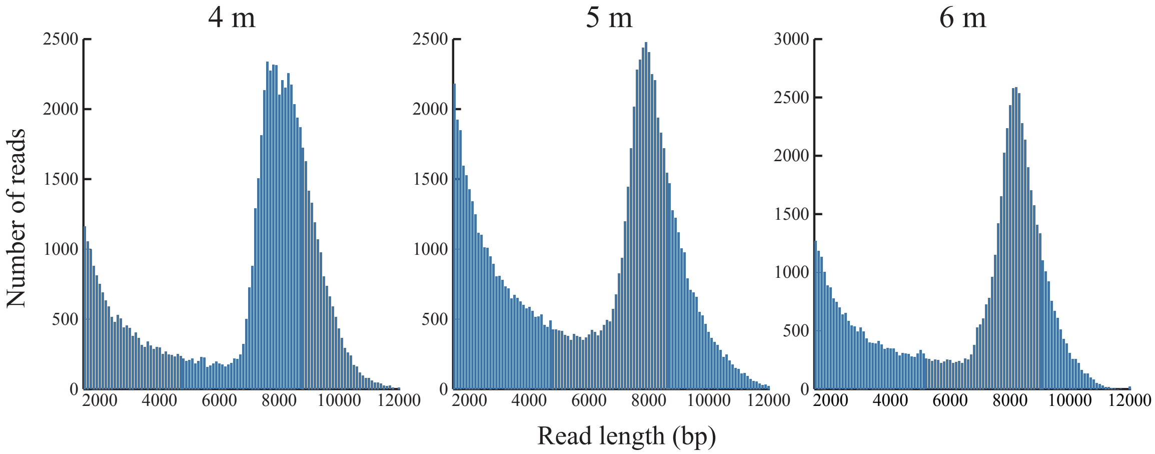

Figure S1: distribution of synthetic long-read lengths for the three samples.

Table S2: statistics for short- and long-read assemblies. Short-read assemblies refer only to scaffolds and contigs longer than 1,500 (400) bp. For the long-read data, overall N50 refers to both assembled and unassembled data.

	Short reads			**Long reads**		
Sample	**% mapped reads**	**Assembly N50**	**Assembly size (mbp)**	**% assembled reads**	**Overall N50**	**Amount**
4 m	27 (35)	4,262 (1,752)	931 (1,694)	11	8,324	470 mbp
5 m	33 (47)	5,128 (2,590)	1,456 (2,207)	14	8,071	505 mbp
6 m	18 (25)	3,747 (1,474)	366 (957)	6	8,264	440 mbp

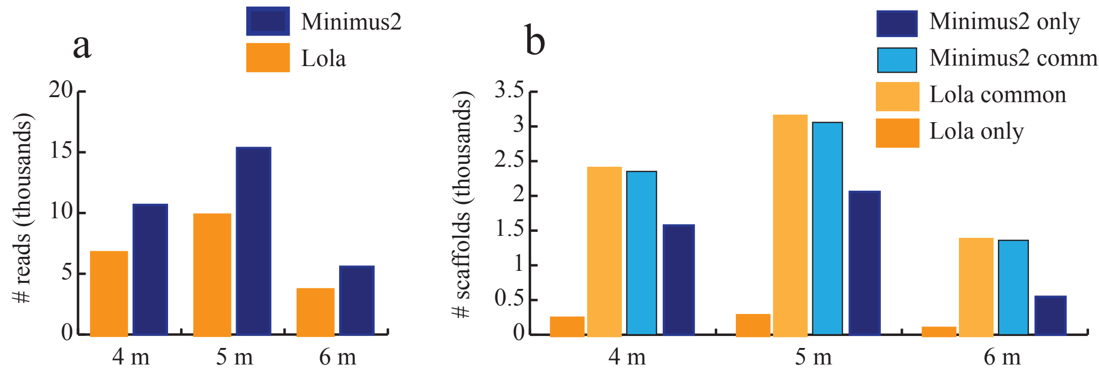

Figure S2: (a) Number of reads assembled by the assembler Lola compared to the number of reads assembled by Minimus 2. (b) Number of assembled contigs that are common to both assemblies and number of reads that were assembled by each of the assemblers alone. Common contigs for each assembler are those that have identical or containing contigs in the other assembler's assembly.

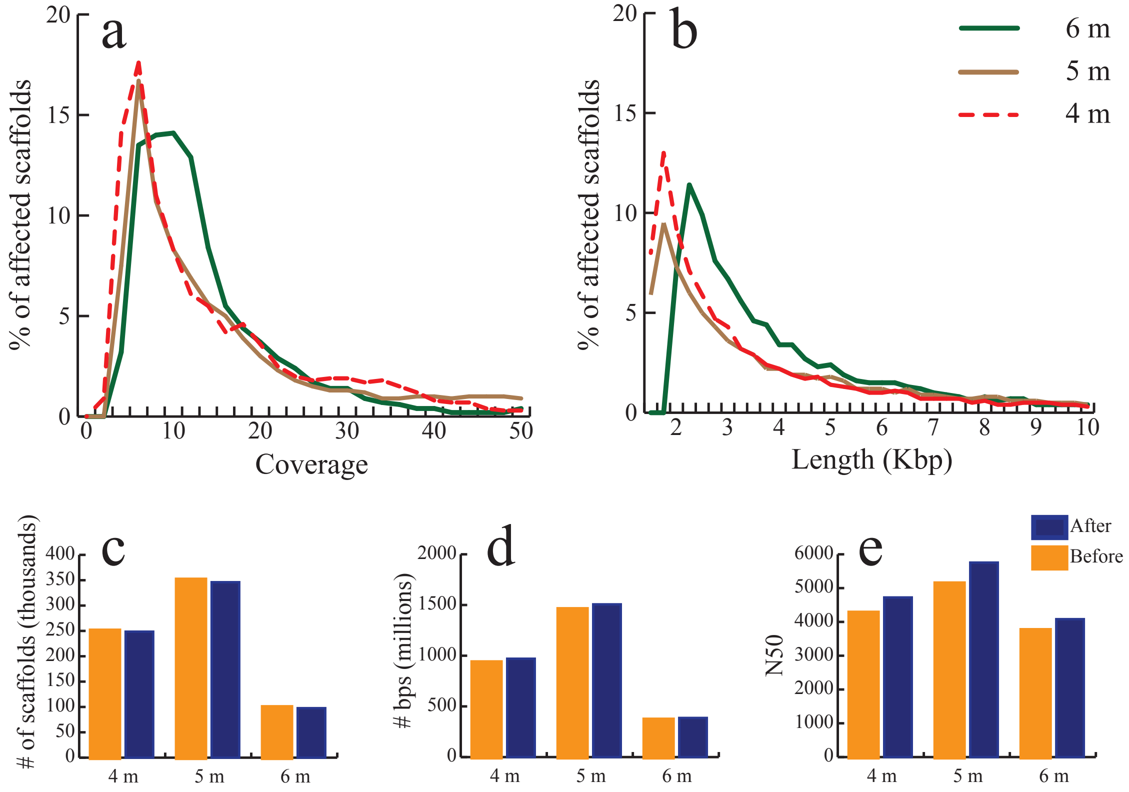

Figure S3: Short-read assemblies before and after scaffolding by Minimus 2 using the synthetic long-reads. synthetic long-reads that did not contribute to the scaffolding or extending of short-read sequences are not included. (a, b) Distribution of coverage values (a) and lengths (b) for assembled scaffolds and contigs affected by long-read data. (c, d, e) change in number of scaffolds (c), total number of bps (d) and N50 (e) for short-read assemblies as a result of scaffolding with long reads.

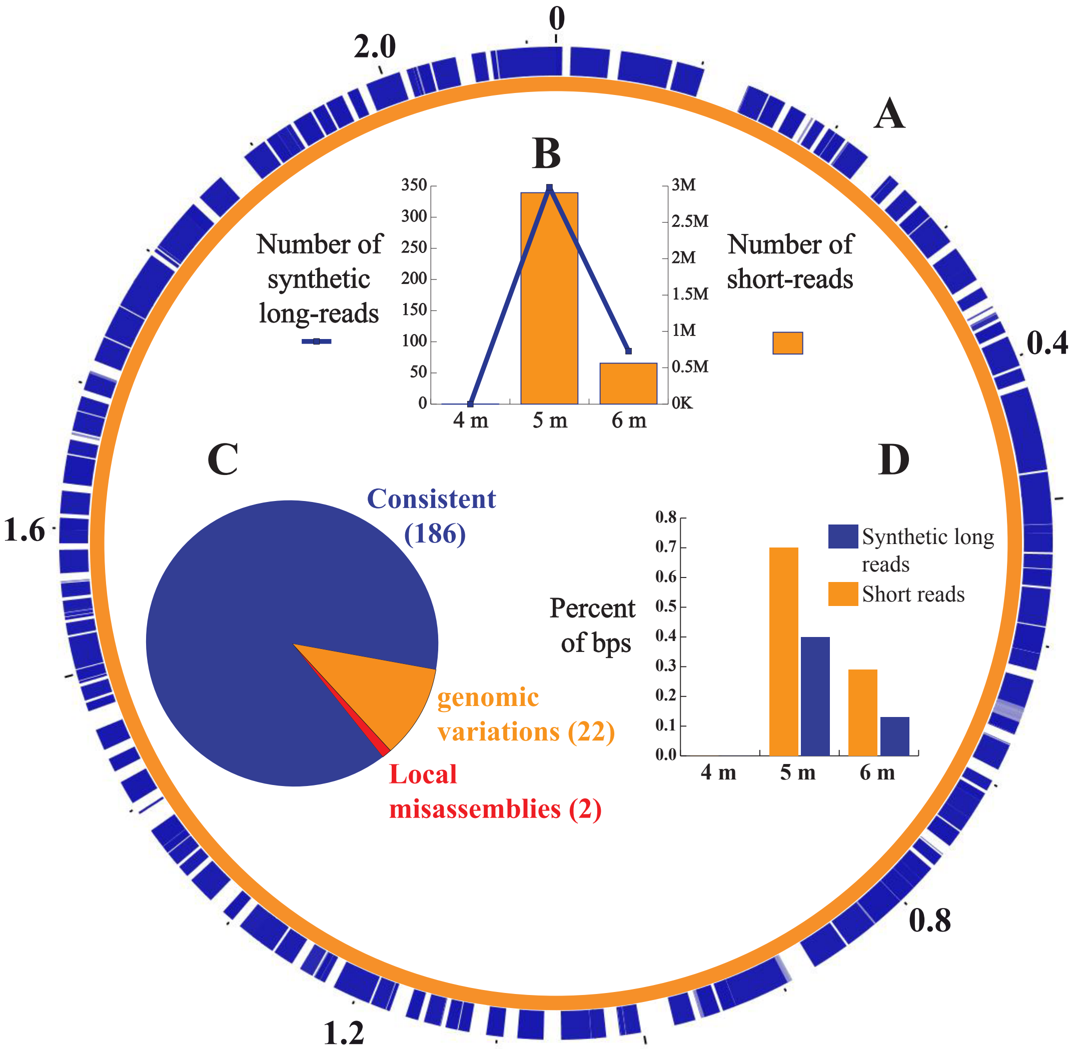

Figure S4: comparison of long- and short-read RBG-1 assemblies. (a) 210 synthetic long-reads and contigs (440 reads) could be aligned to the assembled RBG-1 genome [1] covering 75% of its length. (b) The 5 m sample contributed the majority of long- and short-reads, no reads were found in the 4 m sample. (c) 89% of the aligned long-read sequences were consistent with the assembled short-read genome, 1 % revealed local mis-assemblies in the genome. (d) RBG-1 associated reads account for less than 1 % of both the long- and short-reads in all cases.

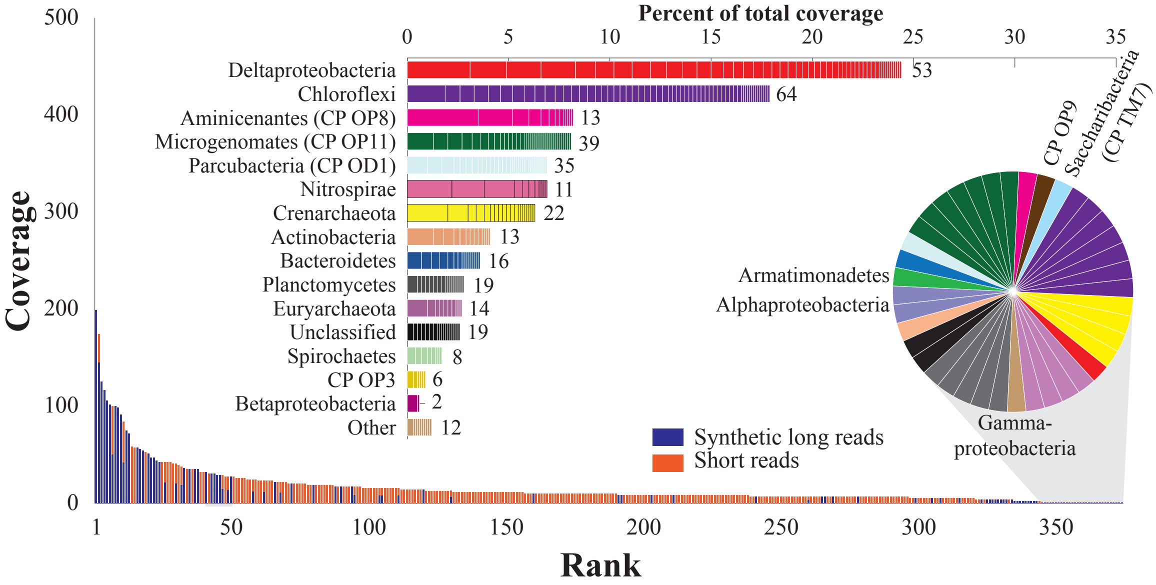

Figure S5: Community composition for the 4 m sample, based on species-level clustering (99% identity) of *rpS3* genes recovered from the long- and short-assemblies. Colors in rank abundance curve (bottom) indicate origin of genes (short-/long-read sequences), coverages are normalized by number of bps in the 5 m sample. Stacked bar graph (top) shows abundance of phyla and Proteobacteria classes in the community, stacked boxes indicate abundance of individual species (number of species indicated). Phylum/class affiliations for *rpS3* genes recovered from long-read sequences with zero coverage by the short-read data are provided in the pie chart (phylum colors are identical to stacked bar graph).

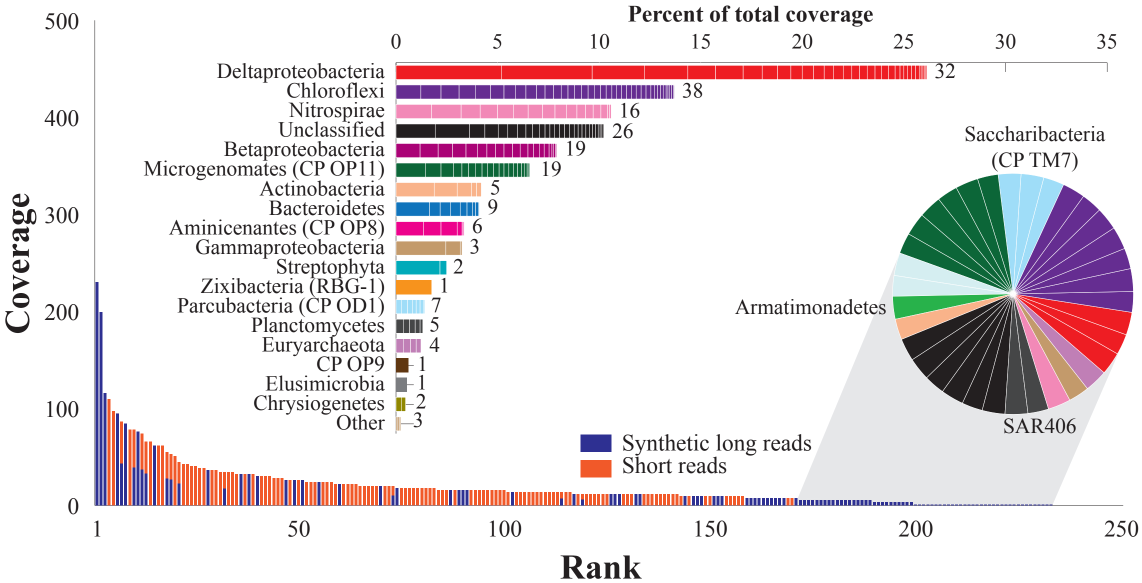

Figure S6: Community composition for the 6 m sample, based on species-level clustering (99% identity) of *rpS3* genes recovered from the long- and short-read assemblies. Colors in rank abundance curve (bottom) indicate origin of genes (short-/ long-read sequences), coverages are normalized by number of bps in the 5 m sample. Stacked bar graph (top) shows abundance of phyla and Proteobacteria classes in the community, stacked boxes indicate abundance of individual species (number of species indicated). Phylum/class affiliations for *rpS3* genes recovered from long-read sequences with zero coverage by the short-read data are provided in the pie chart (phylum colors are identical to stacked bar graph).

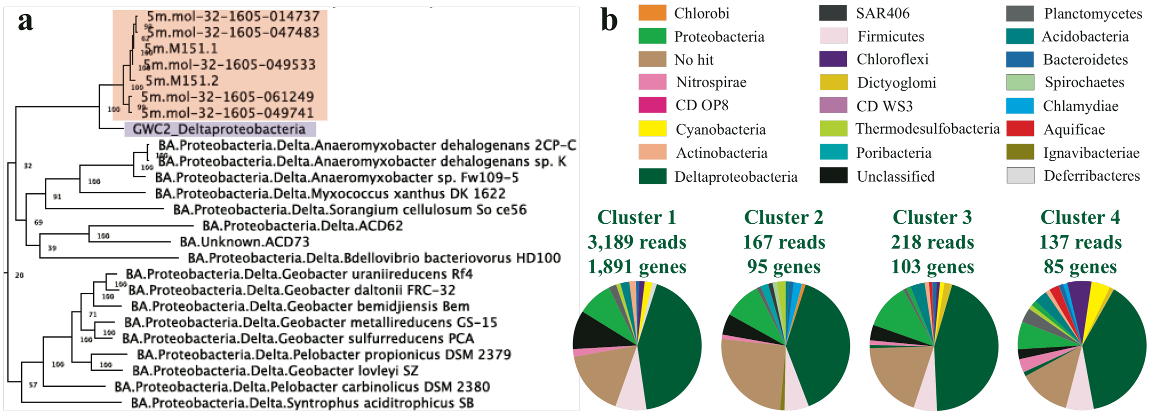

Figure S7: (a) Concatenated ribosomal protein tree containing sequences from cluster 1 that carry ribosomal protein genes places these sequences within the Deltaproteobacteria (refer to concatenated_rp_tree.pdf for complete tree). (b) Taxonomic affiliations of the four long-read clusters with 100 reads or more based on best blast hits of protein coding genes.

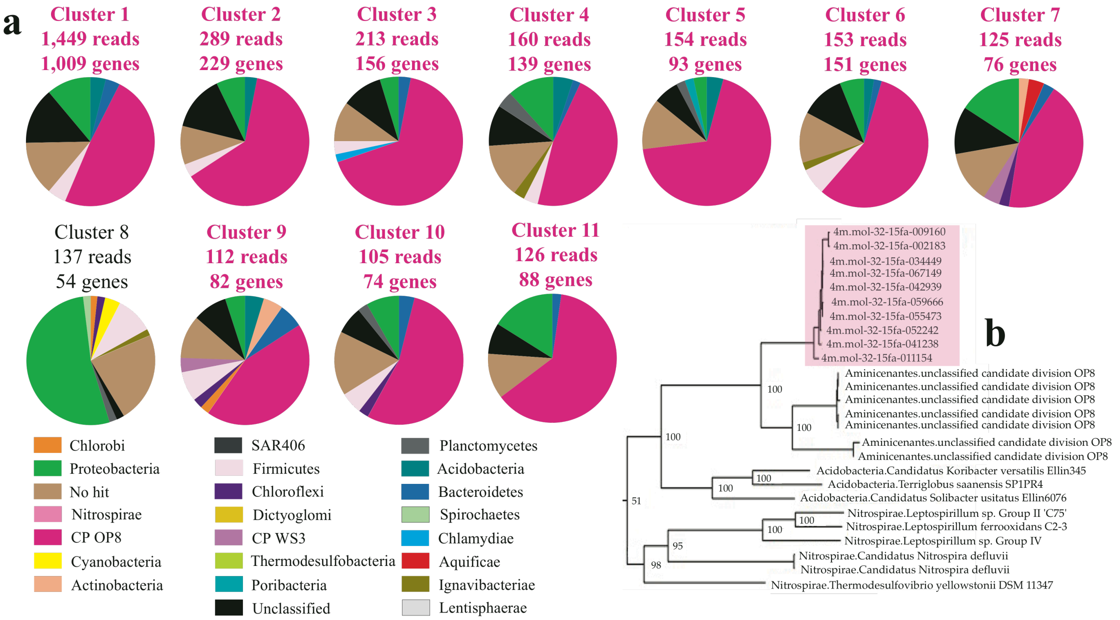

Figure S8: (a) taxonomic profiles for 11 clusters with 100 reads or more from the 4 m sample (based on best blast hits of protein coding genes). Profiles of all clusters but cluster 8 have CP OP8 (Aminicenantes) as the dominant phylum. (b) Concatenated ribosomal protein tree that contain sequences from cluster 6 carrying ribosomal protein genes supports CP OP8 assignment for these organisms (refer to concatenated_rp_tree.pdf for complete tree).

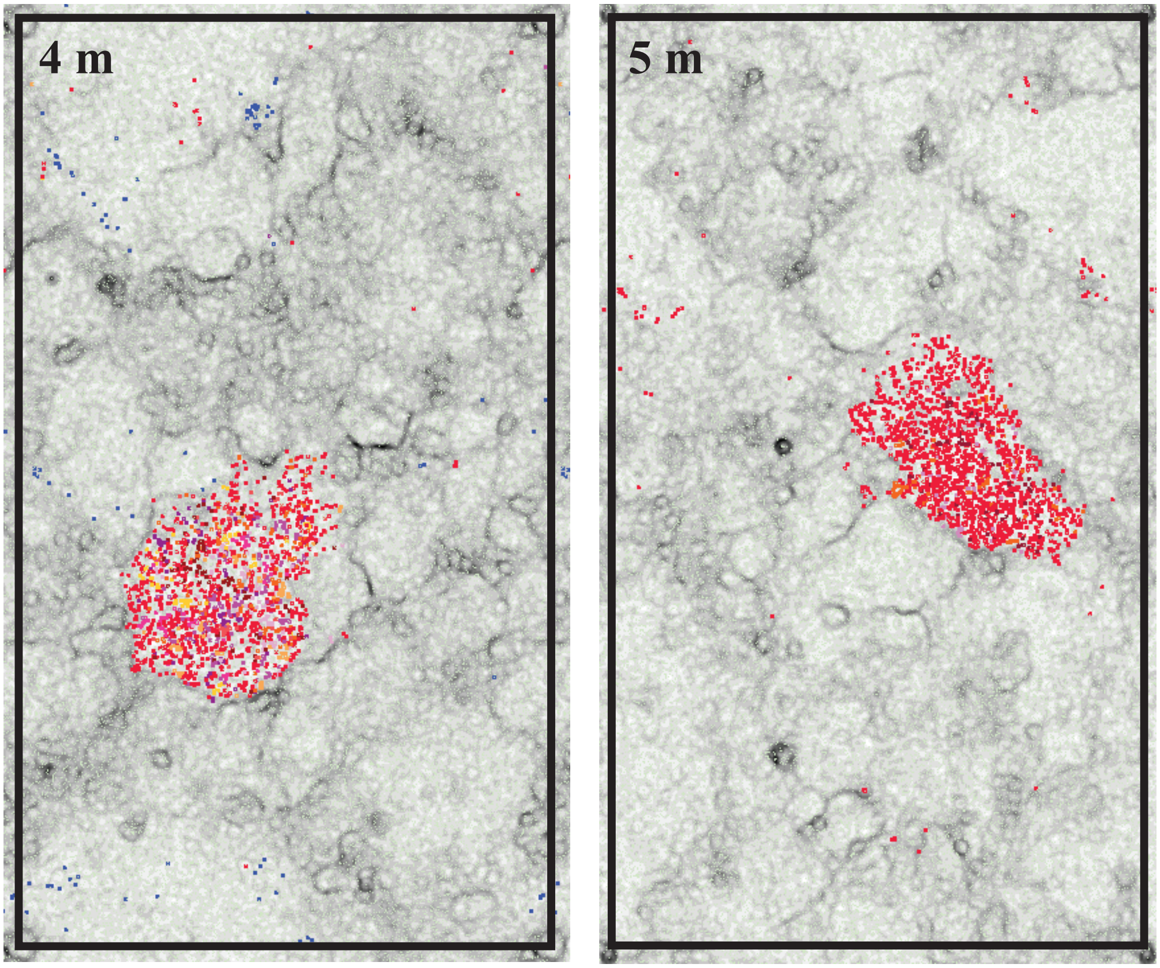

Figure S9: Emergent Self Organizing Map (ESOM) using 3-mer frequencies for synthetic long-reads in the 4 m (left) and 5 m (right) samples. Each datapoint represents a read, colored datapoints show sequences from one of the overlap-based clusters (**Fig. S7** for the 5 m sample and **Fig. S8** for the 4 m). Blue datapoints in the 4 m sample (spread outside the main cluster of colored datapoints) belong to cluster 8 whose taxonomic profile is different from the rest of the clusters.

Assembly of synthetic long-reads with Lola

In order to assemble the data we wrote a program that implements an overlap strategy for assembly. The assembly process consists of 2 steps: (i) identification of overlaps between synthetic long-reads and (ii) assembly of reads based on these overlaps. Here we describe the two steps.

Step I: identifying overlaps. Given a % identity threshold *p* and an overlap size threshold *q* we consider the following five possible overlap types between a pair of sequences A and B that align at *p* or more percent identity over at least *q* bps (**Fig. S10**):

1. **Identity** – both A and B overlap throughout their whole lengths.
2. **Contained** – either A or B but not both are covered throughout their whole length.
3. **End-end** – overlapping regions includes exactly one of A's ends (5' or 3') and exactly one of B's ends such that A's covered end falls inside B and vice versa.
4. **End-mid** – any other overlap that involves exactly one of A or B's ends.
5. **Shared** – any other overlap.

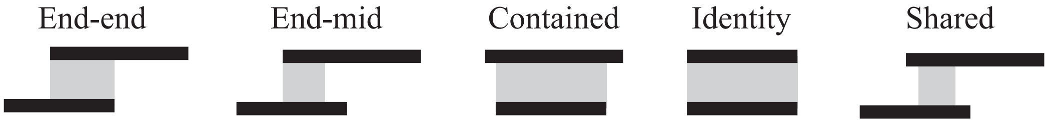

Figure S10: types of overlap (gray regions) between sequences (black lines).

The goal of the first step is to identify all overlaps in the data and decide the type of each one. Step (i) is performed as follows:

1. Remove all reads shorter than *q* bps.
2. Run a self-blast of the read database using parameters –F F –r 1 –q -5 –e 1e-20. These parameters should result with alignments that have high percent identity and are potentially short.
3. Keep all alignments that are at least 0.6**q* bps long (we used *q*=500) with percent identity of at least (*p*-1) % (we used *p*=99). These alignments serve as seeds and will be further checked.
4. For each seed compute the coordinates for the maximal overlap possible between the two reads. Coordinates are computed considering the seed alignment. Based on calculated coordinates decide which overlap types are possible for the pair of reads. Go over the next steps until the conditions for one of the overlap types are met.
5. For potentially identical read pairs align the two sequences using the Needleman-Wunsch global alignment algorithm [7]. Set overlap type to "identity" if sequences align at (*p*-1)% or more.

6. For potential contained overlaps use a variant of the Needleman-Wunsch algorithm that finds the best alignment between a whole sequence A vs. part (or whole) of sequence B. The algorithm uses the following scoring scheme:

   ```
   for(i=0; i<=length(A); i++)
           F(i,0) = d*i;
   for(j=0; j<=length(B); j++)
           F(0,j) = 0;
   for(i=1; i<=length(A); i++)
           for(j=1; j<=length(B); j++) {
                   Match = F(i-1,j-1) + S(Ai, Bj)
                   Delete = F(i-1, j) + d
                   Insert = F(i, j-1) + d
                   F(i,j) = max(Match, Insert, Delete)
           }
   }
   ```

   Search for best alignment starts at the position with the highest score in row F(length(A), *). If alignment at (*p*-1)% or more was determined overlap type was set to “contained”.
7. For potential edge-edge overlaps we used another variant of the Needleman-Wunsch dynamic programming algorithm that finds the best alignment between two sequence ends. The following scoring scheme is applied in order to find the best alignment between the suffix of A and the prefix of B that is at least *q* bps long:

   ```
   for(i=0; i<=length(A); i++)
           F(i,0) = d*i;
   for(j=0; j<=length(B)-1; j++)
           F(0,j) = 0;
   for(i=1; i<=length(A); i++)
           for(j=1; j<=length(B); j++) {
                   Match = F(i-1,j-1) + S(Ai, Bj)
                   Delete = F(i-1, j) + d
                   Insert = F(i, j-1) + d
                   F(i,j) = max(Match, Insert, Delete)
           }
   }
   ```

   Search for best alignment starts at the position with the highest score in column F(*, length(B)). The overlap type is set to edge-edge if the aligned region is at least *q* bps long at % identity of at least (*p*-1).
8. Otherwise the two reads are aligned using the Smith-Waterman local alignment algorithm [8]. Overlap type is set to “shared” if overlapping region is at least *q* bps long at (*p*-1)% identity.

This part of the tool was implemented in C++. Code is available in **Supplementary file overlap-1.02.tar.gz** and is maintained under https://github.com/CK7/overlap.

Step II: assembly. The following process was repeated until no more reads were left that could be used as contig starters.

1. Remove all reads that are contained in another read. Also remove one of every pair of identical reads.
2. Identify an unused read r^{ext} with one or more end-end overlaps at p % identity and no end-mid overlaps on one side $s^{ext} \in \{5, 3\}$ and no end-end connections on the other end. r^{ext}:s^{ext} is the extension end, set r^{ext} as the current assembled contig. Stop assembly if no extension end found.
3. Pick an unused read r^{new} and its end s^{new} that has an end-end overlap with r^{ext}:s^{ext} such that adding r^{new} to the current assembled contig elongates it the most.
4. For every other read that overlaps with r^{ext}:s^{ext} check whether it also has an end-end overlap with r^{new}:s^{new} at (p-1)% identity over its other end. If not – stop assembling the current contig and go to (2).
5. For every other read that overlaps with r^{new}:s^{new} check whether it also has an end-end overlap with r^{ext}:s^{ext} at (p-1)% identity over its other end. If not – stop assembling the current contig and go to (2). Do the same if r^{new}:s^{new} has an end-mid overlap.
6. Add r^{new} to the current assembled contig and mark it as used. Set $r^{ext} = r^{new}$ and $s^{ext} = 8-s^{new}$ (the opposite end of r^{new}).
7. If r^{ext}:s^{ext} has no end-end connections with unassembled reads, or if it has an end-mid connection – stop assembling current contig and go to (2).
8. Otherwise go back to (3) and continue assembling the current contig.

This part of the program was implemented in perl and is available in **Supplementary file Lola-1.02.tar.gz** (program is maintained under https://github.com/CK7/Lola).

Scaffolding of short-read assemblies using synthetic long-reads

Scaffolding of the short-read assembly was performed similarly to the assembly of the synthetic long-reads. Only synthetic long-reads that overlapped with scaffolds from the same depth sample's corresponding short-read assembly were used for this analysis. Once the set of synthetic long-reads that overlaps with the corresponding short-read assemblies were identified, we used the assembly program described above in order to assemble the data.

Reconstruction of syntenic regions

Read clustering. Unassembled synthetic long-reads and assembled contigs from each sample were clustered based on end-end, identity or contained overlaps only, using a 90 % threshold. This threshold roughly represents the expected % identity between similar regions of close species.

Gene prediction. Genes for all reads longer than 5 kbp were predicted using prodigal [10] with parameters –m and –c. Next, all sequences shorter than 5 kbp as well as regions on >5 kbp sequences with no predicted genes were blasted (blastx) against predicted proteins. All regions that aligned at 75 % identity over at least 90 % of the hit length, or with end-end overlap, were marked as new genes. This process was repeated iteratively until no new genes were found. Note that exact start/end coordinates of predicted genes were not important for the purpose of the synteny analysis, only the presence of the genes was utilized.

Clustering of proteins to protein families. All predicted proteins were clustered into families using uclust [11] with a 75 % identity threshold. Singleton families were excluded from further analysis.

Reconstructing gene order in syntenic regions. The following steps were taken for all reads in each cluster (separately):

1. Construct neighborhood graph based on gene location on long-read sequences. Nodes in the graph represent gene families and edges connect two nodes whose genes were found to be neighbors on at least one sequence. Keep weights (number of times each gene/neighborhood relation) were observed.
2. Remove from the graph all edges with weight = 1.
3. Resolve junctions in the graph.
4. Identify bubbles, namely components in the graph with exactly one "in" and one "out" node connecting them to the rest of the graph and two or more paths connecting the in and out nodes within the component.
5. Identify linear components, i.e. components in the graph with one in and one out node that have exactly one path connecting them.
6. Connect bubbles and linear components that share the same in/out nodes such that no other components share the node with them.

Junctions are formed when the same gene family appears in multiple locations on the genome (e.g., due to duplication events). Junctions were resolved based on context of the members of junction protein families: if certain members only appear together with certain genes while other members of the family appear in context with different genes then the family could be split into two sub-families with members being clustered based on their neighbors.

Bubbles were identified based on Depth First Search (DFS) with maximal number of steps = 10. This is performed using each end of each node as root with all nodes visited on each path being kept. Once all paths were visited we try the root with every node visited, from closest to furthest, as the pair of in/out nodes to the component: if all paths that start at the root node go through the other node, then these two nodes bound a bubble component.

Code for the program that implements the above algorithm for gene synteny reconstruction is available in **Supplementary file synteny-1.02.tar.gz** and is maintained under https://github.com/CK7/synteny.

Table S3 Statistics for protein clusters representing marker genes for the Deltaproteobacterium recovered from the 5 m sample. Colors indicate median % identity (dark green: >95, light green: >90, yellow: >85, orange: >80, red: missing).

Marker gene	# of complete/all proteins	Median % identity (Low, High)
ribosomal protein L14	35/37	100.0 (96.7, 100.0)
ribosomal protein L18	13/13	91.5 (86.9, 100.0)
ribosomal protein S12	6/11	92.2 (89.6, 100.0)
gyrA	3/6	99.2 (99.1, 99.9)
ribosomal protein S20	9/9	86.5 (83.5, 100.0)
ribosomal protein L13	2/2	96.0 (96.0, 96.0)
ribosomal protein S11	7/8	96.9 (93.8, 100.0)
ribosomal protein S4	7/9	93.2 (91.7, 100.0)
ribosomal protein S7	11/11	91.7 (87.8, 100.0)
ribosomal protein L5	22/25	93.3 (91.2, 100.0)
ribosomal protein L1	3/3	100.0 (100.0, 100.0)
ribosomal protein L6P-L9E	12/19	91.1 (89.9, 100.0)
ribosomal protein L21	5/5	95.2 (85.6, 100.0)
ribosomal protein L11	3/4	99.3 (99.3, 100.0)
ribosomal protein S5	12/12	96.4 (86.7, 100.0)
ribosomal protein L23	15/15	90.5 (87.2, 100.0)
ribosomal protein S15	2/2	96.6 (96.6, 96.6)
ribosomal protein L10	2/3	100.0 (100.0, 100.0)
alanyl tRNA synthetase	6/17	91.2 (89.1, 99.5)
recA	12/12	97.1 (92.7, 100.0)
ribosomal protein L3	16/16	93.9 (91.5, 100.0)
ribosomal protein S2	10/11	96.7 (90.2, 100.0)
Preprotein translocase subunit SecY	8/11	95.9 (94.5, 100.0)
ribosomal protein L30	10/10	88.5 (87.1, 100.0)
ribosomal protein S17	42/45	97.8 (81.7, 100.0)
leucyl-tRNA synthetase	7/11	96.5 (89.5, 99.9)
arginyl tRNA synthetase	20/24	97.1 (84.7, 100.0)
ribosomal protein S3	43/46	98.2 (92.6, 100.0)
ribosomal protein S19	14/14	96.8 (90.3, 100.0)
ribosomal protein L16-L10E	45/49	98.5 (91.7, 100.0)
ribosomal protein L22	14/34	96.4 (93.6, 100.0)
Valyl-tRNA synthetase	11/15	98.2 (90.2, 100.0)
Histidyl-tRNA synthetase	10/12	87.9 (84.7, 100.0)
ribosomal protein S13	8/9	92.9 (90.6, 100.0)
ribosomal protein S9	2/3	99.2 (99.2, 99.2)
ribosomal protein S6	5/5	84.1 (82.6, 100)
ribosomal protein L27	5/5	100.0 (90.6, 100.0)
ribosomal protein L29	49/49	98.4 (84.6, 100.0)
Phenylalanyl-tRNA synthetase alpha	4/6	98.3 (98.0, 98.8)
ribosomal protein L15	9/10	89.0 (87.0, 100.0)
ribosomal protein L20	4/4	98.3 (96.6, 100.0)
ribosomal protein S10	13/15	98.1 (97.1, 100.0)
aspartyl tRNA synthetase	9/9	92.4 (90.8, 99.8)
ribosomal protein S18	4/4	92.8 (91.1, 100.0)
ribosomal protein L24	29/33	95.5 (86.6, 100.0)
ribosomal protein L4	15/16	91.8 (89.9, 100.0)
ribosomal protein S8	17/20	88.6 (83.3, 100.0)
ribosomal protein S16	3/3	94.7 (93.6, 98.9)
ribosomal protein L17	7/8	82.3 (68.4, 100.0)
ribosomal protein L2	11/16	94.1 (92.3, 100.0)
ribosomal protein L19	7/7	95.4 (88.7, 100.0)

Table S4: Statistics for protein clusters representing marker genes for the CP OP8 phylotype recovered from the 4 m sample. Color scheme is similar to Table S3.

Marker gene	Complete/Total	Median % identity (Low, High)
ribosomal protein L14	10/10	100.0 (98.3, 100.0)
ribosomal protein L18	9/9	98.4 (87.1, 100.0)
ribosomal protein S12	9/9	96.9 (93.8, 100.0)
gyrA	3/7	99.3 (99.2, 99.9)
ribosomal protein S20	9/9	100 (98.8, 100)
ribosomal protein L13	No hits found	
ribosomal protein S11	14/15	96.9 (93.8, 100.0)
ribosomal protein S4	15/15	95.2 (91.3, 100.0)
ribosomal protein S7	7/10	94.9 (93.6, 100.0)
ribosomal protein L5	8/11	99.4 (98.3, 100.0)
ribosomal protein L1	7/7	94.3 (93.9, 100.0)
ribosomal protein L6P-L9E	9/9	98.9 (89.5, 100.0)
ribosomal protein L21	4/6	97.8 (96.1, 99.4)
ribosomal protein L11	7/7	95.0 (94.3, 100.0)
ribosomal protein S5	9/11	97.7 (87.9, 100.0)
ribosomal protein L23	9/9	93.7 (91.8, 100.0)
ribosomal protein S15	7/7	100.0 (96.6, 100.0)
ribosomal protein L10	7/7	87.6 (85.9, 99.4)
alanyl tRNA synthetase	3/7	98.1 (97.7, 99.4)
recA	10/13	96.9 (95.6, 100.0)
ribosomal protein L3	9/9	95.8 (92.5, 100.0)
ribosomal protein S2	4/4	99.2 (99.2, 100.0)
Preprotein translocase subunit SecY	10/15	96.5 (95.2, 100.0)
ribosomal protein L30	12/12	93.0 (88.0, 100.0)
ribosomal protein S17	10/10	100.0 (93.9, 100.0)
leucyl-tRNA synthetase	4/9	97.6 (97.2, 99.6)
arginyl tRNA synthetase	No hits found	
ribosomal protein S3	10/10	99.1 (96.8, 100.0)
ribosomal protein S19	8/8	100.0 (95.8, 100.0)
ribosomal protein L16-L10E	10/10	99.3 (94.3, 100.0)
ribosomal protein L22	9/10	99.2 (92.5, 100.0)
Valyl-tRNA synthetase	3/5	98.3 (98.1, 98.6)
Histidyl-tRNA synthetase	5/6	96.0 (94.8, 99.8)
ribosomal protein S13	14/14	97.6 (95.9, 100.0)
ribosomal protein S9	No hits found	
ribosomal protein S6	6/8	97.8 (96.4, 100.0)
ribosomal protein L27	6/6	98.8 (97.6, 100.0)
ribosomal protein L29	10/10	100.0 (92.2, 100.0)
Phenylalanyl-tRNA synthetase alpha	7/9	93.9 (88.6, 100.0)
ribosomal protein L15	12/12	92.6 (84.5, 100.0)
ribosomal protein L20	7/7	97.5 (96.6, 100.0)
ribosomal protein S10	10/10	98.1 (96.2, 100.0)
aspartyl tRNA synthetase	4/6	95.1 (94.6, 99.5)
ribosomal protein S18	8/8	100.0 (97.6, 100.0)
ribosomal protein L24	11/11	100.0 (90.9, 100.0)
ribosomal protein L4	9/9	98.1 (92.3, 100.0)
ribosomal protein S8	8/12	98.5 (93.2, 100.0)
ribosomal protein S16	8/8	97.6 (92.9, 100.0)
ribosomal protein L17	14/15	92.1 (83.3, 100.0)
ribosomal protein L2	7/8	97.1 (94.9, 100.0)
ribosomal protein L19	8/8	99.1 (98.2, 100.0)

Strain variation for the Deltaproteobacteria clade

Multiple sequence alignment of the 46 *rpS3* genes from the Deltaproteobacteria clade reveals several groups of identical and near identical genes (**Fig. S11**). Clustering of these sequences using a 100 % threshold (DNA level) resulted with 28 clusters, 10 of which consist of multiple members. Mapping of short reads reads to the *rpS3* genes recovered from the long reads resulted with many reads that could not be mapped perfectly to any of the genes, suggesting that even more closely related versions of this gene may be present in the sample. Based on these results we estimate that Deltaproteobacteria clade in the 5 m sample consists of several dozen different strains. While the total coverage of the *rpS3* genes from these strains by the Illumina short read data approaches 1,000x, we did not find any *rpS3* genes in the Illumina assembly for the 5 m sample that align to any of the 46 *rpS3* genes considered here at more than 85% identity. Therefore, we attribute the lack of short read assembly for these genomes to high sequence heterogeneity that cannot be handled by short read assemblers (**Fig. S12** and **Fig. S13**). For comparison, mapping of Illumina reads to the *rpS3* gene of RBG-1 shows very few SNPs, none of them appears in more than a few reads (**Fig. S14**). This suggests that RBG-1 is represented by a single abundant strain, which enables assembly.

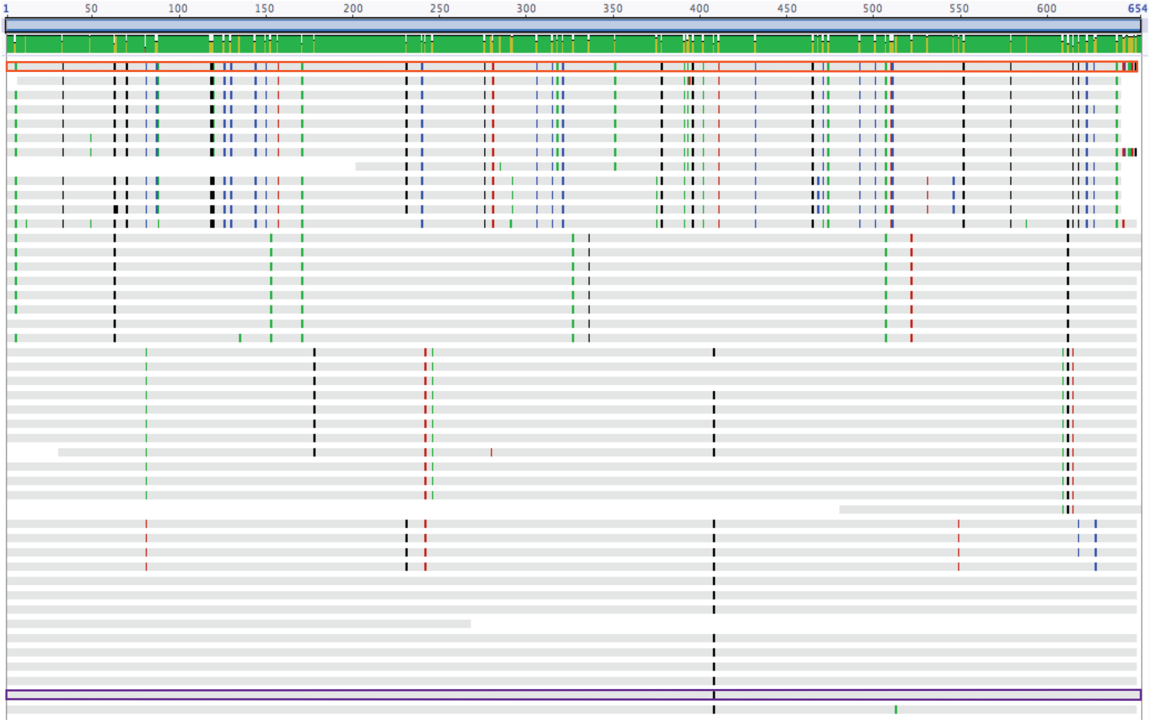

Figure S11: multiple sequence alignment of the 46 genes from the *rpS3* cluster of the Deltaproteobacteria clade (5 m sample). Genes used for mapping of Illumina short reads in **Fig. S12** and **S13** are encircled in purple and orange, respectively.

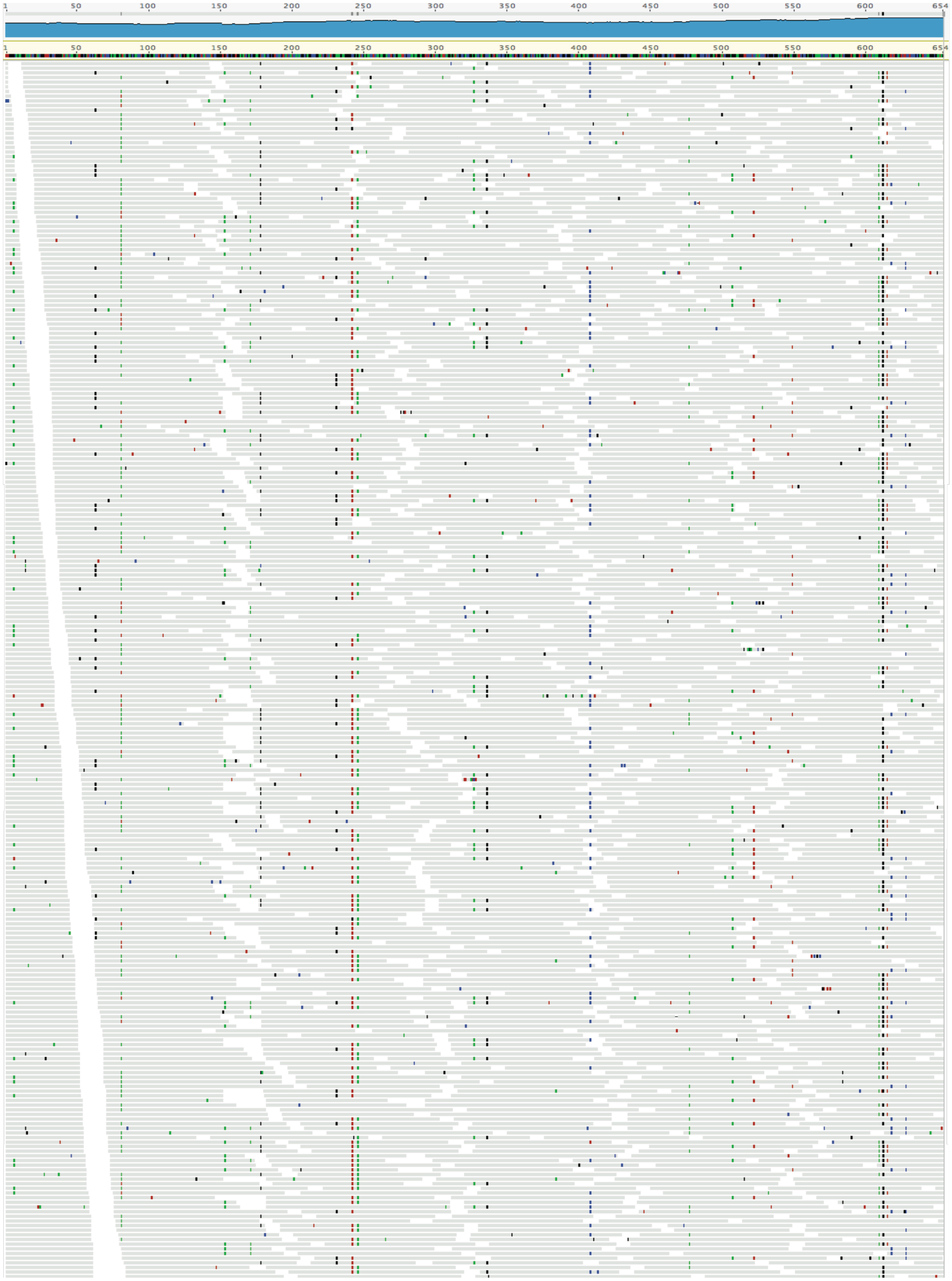

Figure S12: read mapping to one of the *rpS3* genes from the Deltaproteobacteria clade (encircled in purple in **Fig. S11**) reveals high degree of strain variation, indicated by multiple SNPs (colored dots).

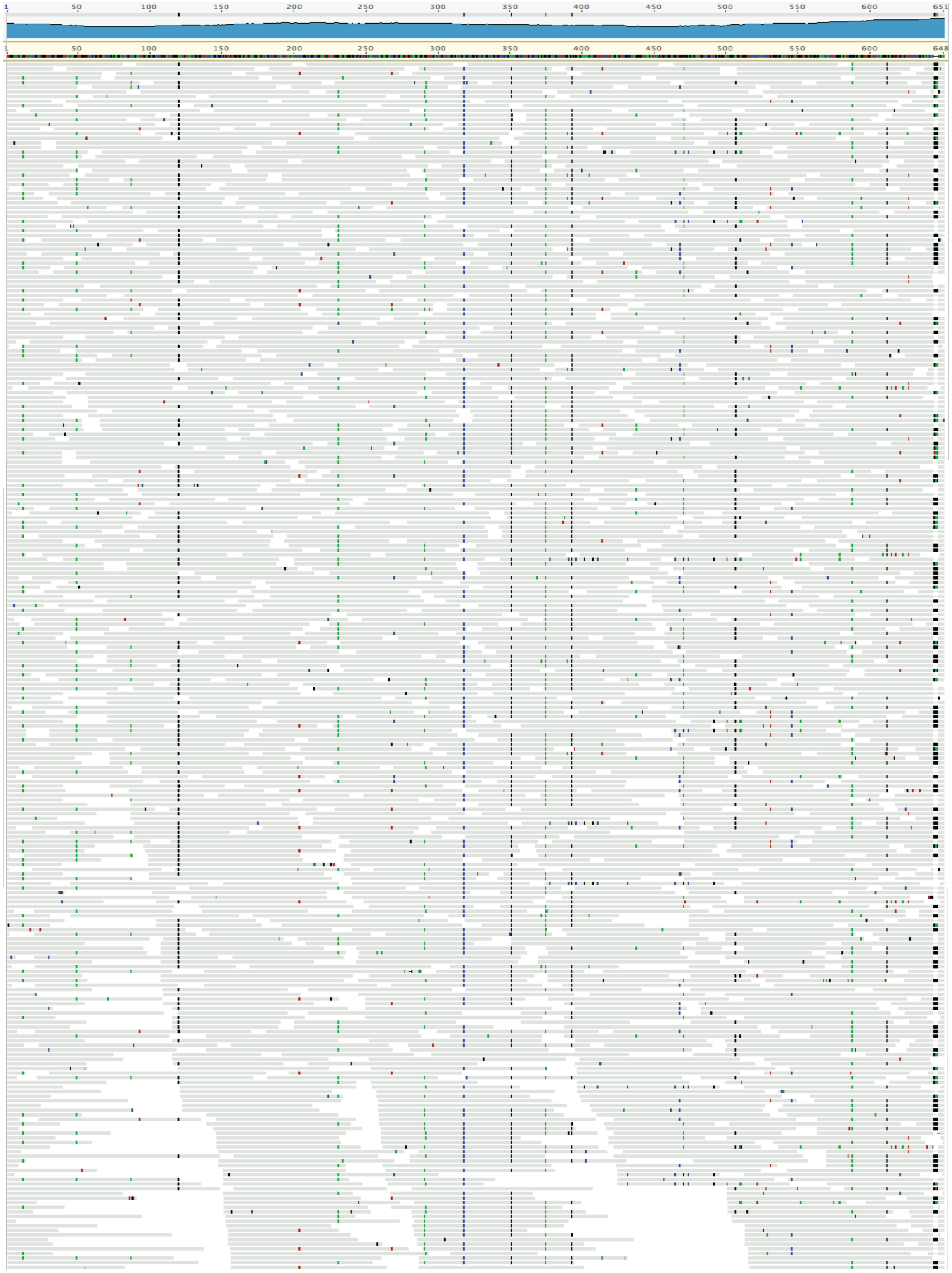

Figure S13: read mapping to a different *rpS3* genes from the Deltaproteobacteria clade (encircled in orange in **Fig. S11**) shows high number of SNPs as well (colored dots).

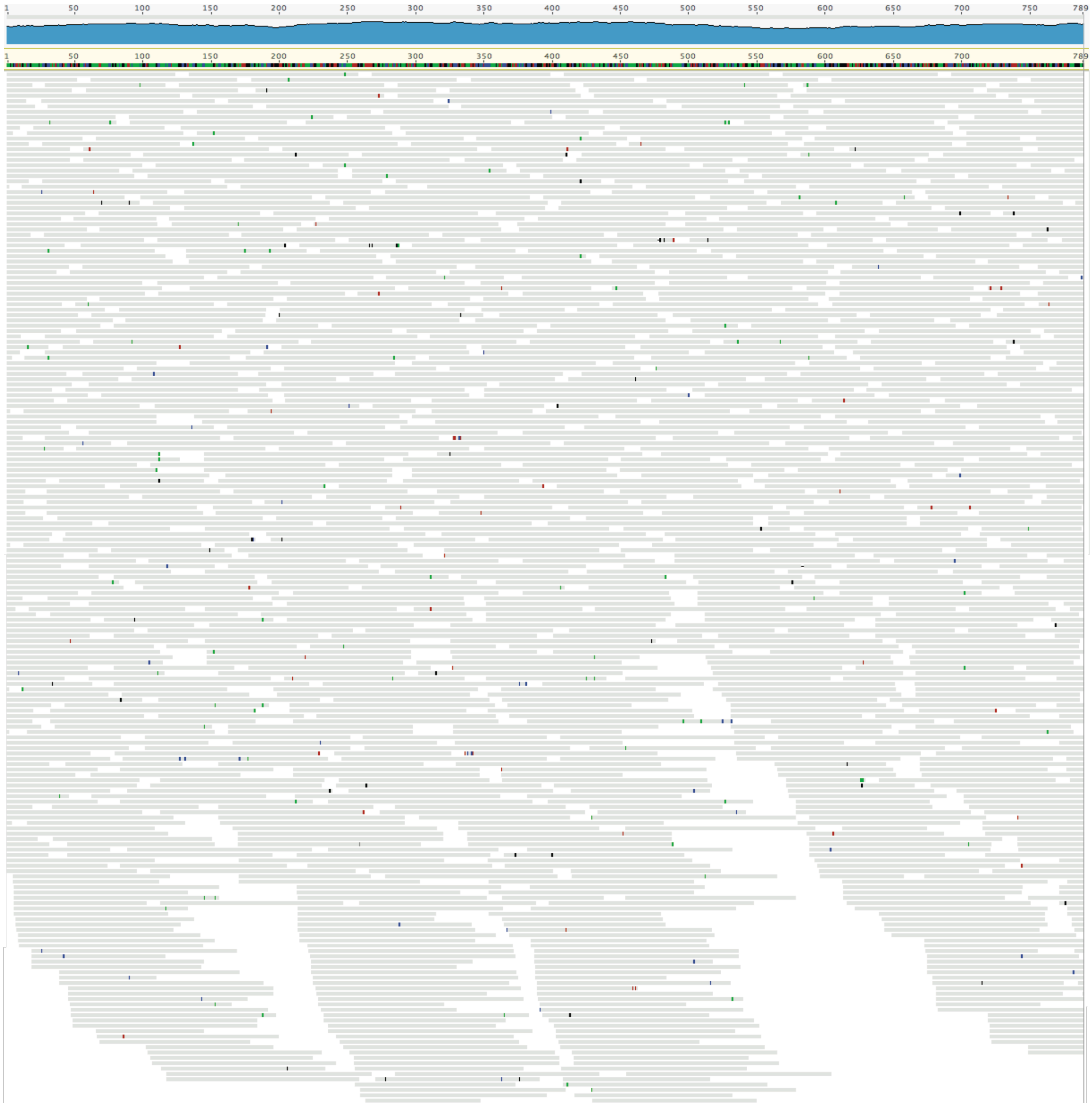

Figure S14: reads from the 5 m sample that are mapped to the RBG-1 *rpS3* gene are consistent with the assembled genome, suggesting the presence of a single abundant strain in the sample.

Recovery of 16S rRNA genes and tree construction

16S rRNA genes were recovered from the short- and long-read assemblies using rnammer [12] with parameters –multi –S bac,arc. All genes longer than 800 bps were clustered using uclust with 99% identity threshold with only one representative from each cluster being used. The three best hits for each representative from the SILVA database ([13], release 115) were used for phylogenetic tree construction. In addition, we included a set of 762 reference 16S sequences from sequenced genomes spanning different phyla. Sequences were aligned using SSU-ALIGN [14] and manually inspected using Geneious. Maximum likelihood trees were reconstructed using RAxML [15] with parameters -f a -m GTRCAT -N 100.

Using the *rpS3* gene for community structure inference

To evaluate community structure we considered both the genes encoding the 16S SSU rRNA and the ribosomal protein S3 (*rpS3*). The 16S SSU rRNA gene is a commonly accepted phylogenetic marker gene [16], however assembly of this gene from metagenomic data often fails due to its high conservation. We were able to identify 355, 495 and 289 16S rRNA genes from the 4, 5 and 6 m respectively with the majority of genes being recovered from the long-read data. Clustering the genes using a 99% identity cutoff (capturing roughly the same genus) revealed that roughly one quarter of the genes recovered from the long-read sequences, in all samples, were clustered with at least one other long-read 16S rRNA gene, compared to none of the genes recovered from short reads (**Fig. S15**). The largest cluster in the 5 m sample, which represented the most abundant genus in both the 5 and 6 m samples, did not have any representatives in the short-read assemblies but contained 14 (5 m) and 9 (6 m) genes from the long-read datasets. In fact, the majority of clusters with at least 3 members in all samples (23/32) and the vast majority of clusters with at least 4 members (8/10) were represented by long-read sequences only. Low clustering levels of the short-read sequences are not due to coverage as many genes recovered from synthetic long-read had robust coverage from mapped short reads. This highlights the fact that the 16S SSU rRNA gene is not a reliable choice for inferring community structure when used directly from assembled metagenomic short read data.

We also considered using the gene cassette of the 16 ribosomal proteins previously proposed [17] and used by Castelle *et al.* [1] for the 5 m sample in this study. This set of genes provides a reliable phylogenetic placement that is required for taxonomic assignment of novel genomes. On the other hand, at least eight of the genes need to be present on the same sequence in order for their sequence to be considered, which may be too restrictive for genomes whose assembly is fragmented. As a result, the number of organisms that can be detected using concatenated ribosomal proteins may be artificially low (see below).

Here we evaluated community structure for the three samples using the ribosomal protein S3 (RpS3), which is the product of a universal single-copy gene. This protein is less reliable for high-resolution phylogeny than both the gene encoding the 16S SSU rRNA and concatenated ribosomal proteins as it is shorter than either of those options, and thus has fewer alignment positions to direct placement. Nevertheless this gene is divergent enough to be recovered through metagenomic assembly (unlike the 16S gene) and can be recovered from fragmented genomes as well (unlike the ribosomal proteins). The number of *rpS3* genes recovered from the short-read assembly was significantly higher than the 16S rRNA genes for all samples (**Fig. S15**). For comparison, the number of *rpS3* genes recovered from synthetic long-reads was smaller than the number of 16S genes recovered from the same data, both because the 16S gene is twice as long as the *rpS3* gene and also because the 16S gene sometimes appears in more than one copy per genome (**Fig. S15**). This result suggests that many more species can be detected using the *rpS3* gene compared to the 16S SSU rRNA gene for short-read assemblies. The number of *rpS3* genes that were recovered from the short-read assembly of the 5 m sample is roughly twice the number of ribosomal protein sets reported by Castelle et al. However no *rpS3* genes were assembled from the short-read data for most of the abundant species in all samples, probably due to a high degree of strain variation.

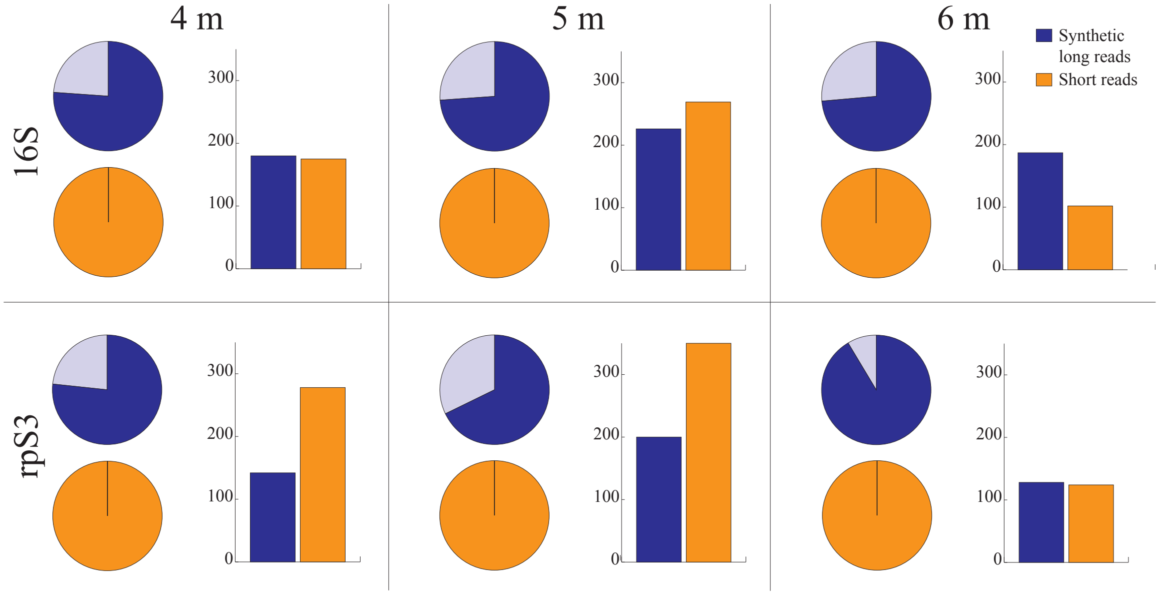

Figure S15: Pie charts: fraction of 16S SSU rRNA (top row) and *rpS3* (bottom row) genes in clusters with no other gene of the same technology (dark) or with at least one other gene from the same technology (light) for the 4, 5 and 6 m samples. Column bars: total number of unclustered 16S SSU rRNA (top) and *rpS3* (bottom) genes found using each technology.

Computing % identity thresholds for *rpS3* for species- and genus-level clustering

In order to generate the rank abundance curves in **Fig. 2**, **S5**, and **S6**, we clustered *rpS3* genes into species and genera groups using thresholds that were calculated according to the following model. Our goal was to compute thresholds that will allow us to determine whether a pair of *rpS3* genes represents two genomes from the same species, same genus or something else, based on the % identity of their global alignment (DNA sequence). In order to do this we developed the following model that enabled us to compute, for each % identity, the fraction of Rps3 protein pairs that are expected to belong to the same species and genus. Once these were calculated we chose thresholds for which the majority of pairs aligned at this % identity or higher indeed belonged to the same species/genus.

Theory: Two *rpS3* genes s1 and s2 that are x% similar can be in exactly one of two states with respect to a taxonomic level t∈{species, genus, family, order, class, phylum, domain, other}:

- State A_t – s1 and s2 belong to the same taxonomic group at level t but not to any taxonomic group at level lower than t (e.g. same genus but not the same species).
- State B_t - s1 and s2 belong to a taxonomic group at a level different than t

Given % identity x for the global alignment of s1 and s2 it is possible to determine the probability that s1 and s2 are in state A_t given the following information:

- $p(x|A_{t'})$ – the probability that s1 and s2 are x% similar for taxonomic level t'∈{species, genus, family, order, class, phylum, domain, other}

- $p(A_{t'})$ – the probability that a random pair s1 and s2 will be in state $A_{t'}$ for taxonomic level t'.

Both these probabilities may change between different communities and taxonomic groups. In fact, it is practically impossible to compute either of them, however we provide here rough estimations for both of them (see below).

In order to compute the probability $p(A_t|x)$ we will estimate the fraction of pairs in state A_t from all pairs of *rpS3* genes that are x% similar. If our sample consists of S cells (and thus S copies of the *rpS3* gene) then there are S*(S-1)/2 different pairs, of which

$$N(A_{t'})= p(A_{t'})*S*(S-1)/2$$

Are in state $A_{t'}$, and

$$N^x(A_{t'})= p(x|A_{t'})*N(A_{t'}) = p(x|A_{t'})*p(A_{t'})*S*(S-1)/2$$

are both in state $A_{t'}$ and the % identity of their global alignment is x. The probability $p(A_t|x)$ can then be estimated through

$$p(A_t|x) = N^x(A_{t'})/\Sigma\ N^x(A_{t'}) = [p(x|A_t)*p(A_t)*S*(S-1)/2]/[\ \Sigma p(x|A_{t'})*p(A_{t'})*S*(S-1)/2]$$
$$= p(x|A_t)*p(A_t)/\Sigma(p(x|A_{t'})*p(A_{t'}))$$

Which is the probability we are looking for.

Computing $p(x|A_{t'})$. In order to compute these probabilities we used a set of *rpS3* genes from 1,978 genomes downloaded from the NCBI website. Global alignments for the genes were calculated using usearch (-allpairs_global program), and distributions of % identity were calculated as follows:

```
For t in {species, genus, order, class, phylum} do
   For taxonomic group o at level t do
      Collect all % identities for all pairs of sequences s1, s2 $\in$ o
      Find the median value and add it to the distribution of t
      Find s with sum of % identities vs all other s'$\in$o is the highest
      Remove all s' $\in$ o except s.
   End for
End for
```

The above procedure was designed to avoid biases that are related to the number of sequenced genomes from each taxonomic group by taking one representative from each group.

Our computations showed that no % identity for any pair from t' ∈ {order, class, phylum} was high enough to impact the computation of thresholds for species and genus. Consequently we discarded these taxonomic levels as well as the domain level from further analysis.

Calculating $p(A_{t'})$. Obtaining these probabilities is practically impossible for various reasons: not only the number of cells from each taxonomic group is unknown but even the assignment to taxonomic groups is not available yet for most of the organisms in our communities. Note also that p($A_{t'}$) may be different for different environments. Here we estimated p($A_{t'}$) using the 16S sequences extracted from the long-read sequences, assuming that these were sampled randomly from the sample. We clustered the 16S sequences based on the following thresholds:

- Species: 99% [18]
- Genus: 94% [19,20]
- Family: 90% [19,20]

These values are in no way exact but are expected to roughly divide our 16S rRNA SSU sequences into phylotypes at these levels. Sequences were aligned using usearch. Next, the different probabilities were estimated based on 100,000 random samplings of sequence pairs. Results are summarized in Table S5.

Table S5: frequency of 16S SSU rRNA gene pairs in the 3 depths. Frequency was computed through simulations.

	4 m: # (%)	**5 m: # (%)**	**6 m: # (%)**
Species	829 (0.83)	824 (0.82)	940 (0.94)
Genus	692 (0.69)	544 (0.54)	1,205 (1.21)
Family	833 (0.83)	552 (0.55)	1121 (1.12)
Other	97,646 (97.65)	98,080 (98.08)	96,734 (96.73)

Calculating thresholds. Using the above we calculated the expected fraction of *rpS3* gene pairs that belong to the same species and genus for the 3 samples (Table S6). Based on these calculations we conclude that thresholds of 99% for species and 90% for genus levels provide a reasonable level of confidence (> ~80%) in the relatedness of pairs of *rpS3* genes.

Table S6: fraction of *rpS3* gene pairs that align at % identity and share the same species (genus) for each sample. For example: 95.1% of the *rpS3* pairs that align at 100 % identity in the 4 m sample share the same species while 100% of these pairs share the same genus. This means that 4.9% of *rps3* genes that align at 100% identity belong to different species under the same genus. From this table we conclude that the majority (~70% or more) of *rps3* genes that align at 99% or more belong to the same species and therefore use this number as a threshold for species-based clustering. Similarly, >70% of pairs that align at 88% identity or more belong to the same genus.

	4 m		**5 m**		**6 m**	
% identity	**Species**	**Genus**	**Species**	**Genus**	**Species**	**Genus**
100.0	95.1	100.0	96.1	100.0	92.6	100.0
99.0	77.5	100.0	81.3	100.0	68.9	100.0
98.0	50.5	100.0	56.3	100.0	39.7	100.0
97.0	61.2	91.3	67.4	93.5	51.8	91.2
96.0	38.4	100.0	44.1	100.0	28.7	100.0
95.0	0.0	100.0	0.0	100.0	0.0	100.0
94.0	22.8	79.7	28.0	83.3	16.9	82.1
93.0	0.0	100.0	0.0	100.0	0.0	100.0
92.0	27.7	100.0	32.6	100.0	19.8	100.0
91.0	11.2	100.0	13.8	100.0	7.6	100.0
90.0	0.0	83.6	0.0	85.7	0.0	86.9
89.0	0.0	70.6	0.0	73.9	0.0	75.7
88.0	0.0	70.6	0.0	73.9	0.0	75.7
87.0	0.0	43.9	0.0	48.1	0.0	50.4
86.0	0.0	61.7	0.0	65.6	0.0	67.7
85.0	0.0	53.8	0.0	57.9	0.0	60.2
84.0	18.6	64.9	23.6	70.1	14.1	68.4

Estimating a lower bound for the number of species in the samples

Our sequencing efforts were far from being exhaustive, as indicated by the relatively low rates of both long and short reads going into assemblies. *rpS3*-based rarefaction curves for the three samples (**Fig. S16**) provided showed that our sequencing efforts are indeed far from being sufficient. In order to get a general sense of how complex our samples are we tried to put a lower bound on the number of most abundant species which we were able to sample. In other words, we were trying to answer the following question: assuming that all species in the sample are sorted by their rank (as in a rank abundance curve), what would be the rank of the least abundant species we were able to sample?

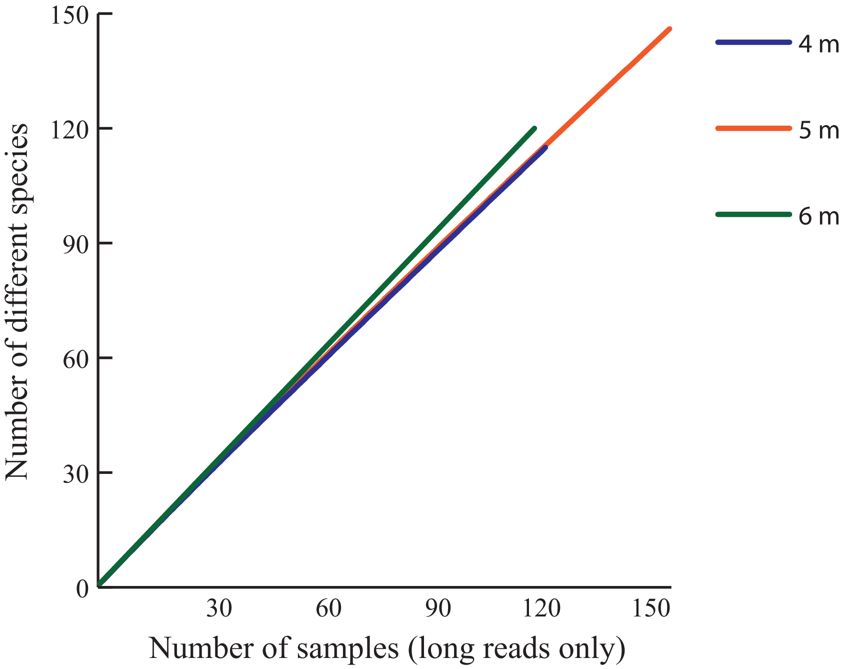

Figure S16: rarefaction curves for the three samples based on *rpS3* genes found on synthetic long reads only. For the 6 m sample each species was detected exactly once. Illumina assembled *rpS3* genes were not used because they represent only the fraction of community that is abundant enough.

To answer this question we "divided" the population into two fractions based on a coverage threshold and estimated a lower bound on the number of species in each. The threshold that was chosen is 2x because this is about the coverage in which short reads start to assemble. We estimated a lower bound on the number of species as follows:

1. Count the number of species with coverage > 2x from both the short-read assemblies and synthetic long reads, based on *rpS3* clustering (we will denote this number $N^{>2}$). We assume that for the > 2x the combination of short- and long-read datasets provide a good approximation for community structure because the short-read data is expected to be assembled for species represented by single strains and the long-read data is expected to uncover multi-strain species (at least the abundant ones). In any case the number computed here is not expected to exceed the true number.

The following steps are taken to determine $N^{\leq 2}$, an upper bound on the number of species with coverage ≤ 2x.

2. Summarize the coverage for species with coverage > 2x based on the coverage of *rpS3* genes from both the short-read assemblies and synthetic long-reads.
3. Use the fraction of short reads that mapped to single copy genes from synthetic long-reads with coverage > 2x as an estimator for the portion of the community occupied by these organisms. (complementary frequencies to the gray bars in **Fig. S18**). Both long and short **reads** are assumed to be sampled randomly with equal probability from the entire community regardless of genome abundance and could therefore be used for this purpose (unlike the **assembled** short-read sequences that represent only genomes with sufficient coverage).
4. Compute the relative abundance of the least abundant species with > 2x coverage by dividing the coverage of the least abundant species with coverage > 2x by the sum of coverages for all species with coverage > 2x (computed in (2)) and normalizing this value by the share of these species in the community (from (3)).
5. The relative abundance of any of the species with coverage ≤ 2x will be smaller than the relative abundance computed in (4). Dividing the relative frequency of species with ≤ 2x by the relative abundance of the least abundant species from the > 2x group will therefore give us a lower bound on the number of species in this fraction. We will denote this number by $N^{\leq 2x}$.
6. The total number of species in the "pool" of species that were sampled is given by $N= N^{>2x} + N^{\leq 2x}$.

The calculated lower bounds on the number of species in the 3 samples were around 900 for the 4 and 6 m samples and around 3,000 for the 5 m samples. We estimate that the true numbers are significantly higher because dozens of species with less that 2x coverage were detected in each of the samples, suggesting that the lowest relative abundance in the >2x fraction is in fact a very strict estimator.

Internal controls on contamination

In order to identify potential contamination in our datasets we used two different approaches. First, we aligned (using BLAST) the long reads against an extended version of the uniref90 database (refer to **Fig. S17** for a summary of the taxonomic profiling for the three samples). The extended version of uniref90 includes also other genomes recovered from Rifle. Using this analysis it is possible to identify DNA that is clearly foreign to the samples. Second, we manually checked the data using the ggKBase platform and other tools in order to search for suspicious DNA. This approach proved to be efficient in the past in identifying contaminations. For the current study we identified a few dozen reads (out of several tens of thousands) that originated from a cloning vector, as well as a few dozens identical long reads that seemed to be an artifact of the sequencing process. The vast majority of the data, however, appears to be reliable

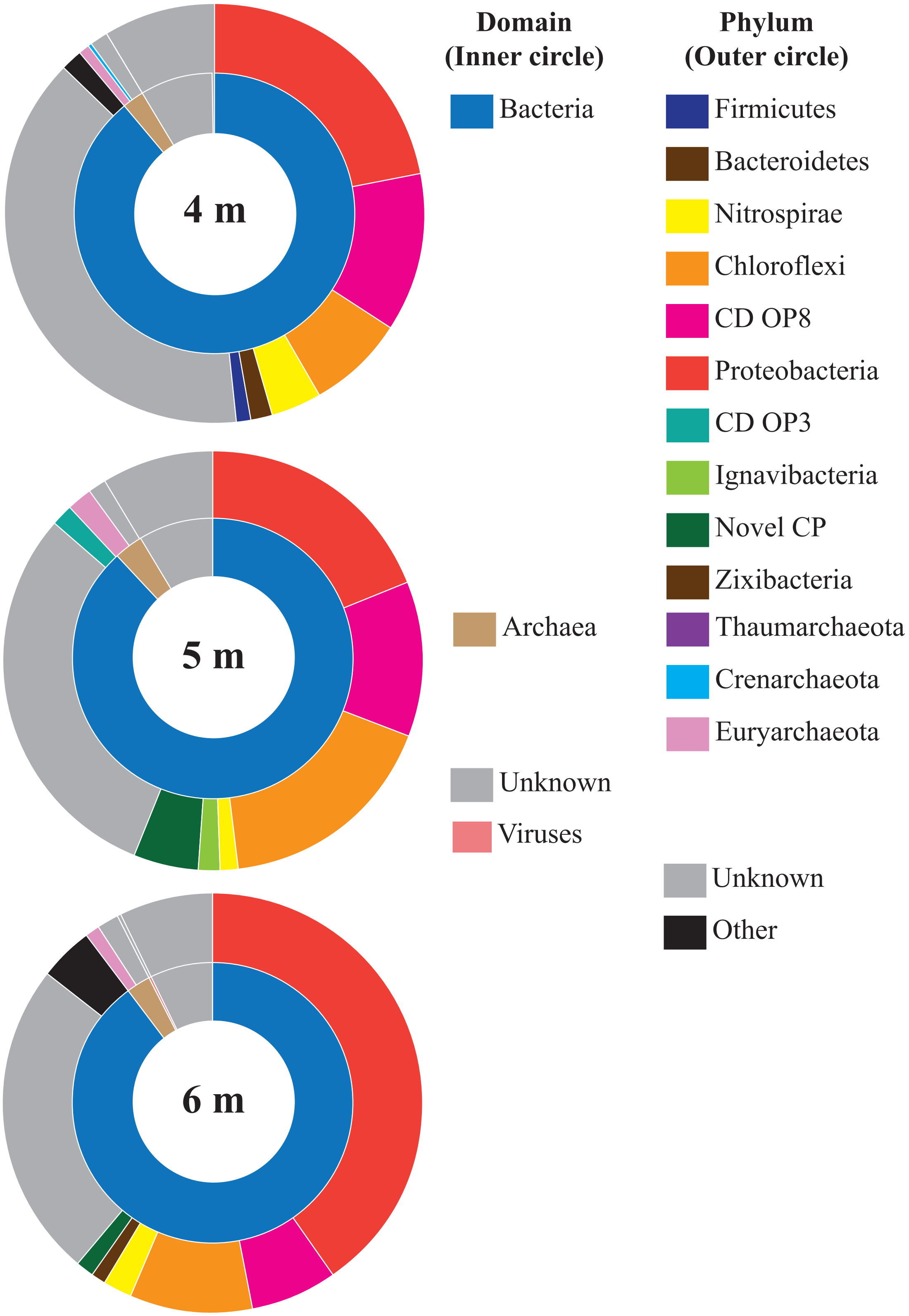

Figure S17: Distribution of domains and phyla of best hits for predicted proteins on assembled and unassembled synthetic long-reads.

Table S7: Statistics of 51 marker gene families in the 4 m sample. Expected, observed columns: expected (based on 37.9 % of bps) and observed number of proteins on ≤ 2x coverage reads.

	Total	Expected	Observed	Observed freq	p-value
Total	5603	1939	1751	0.31	1.0000
Histidyl-tRNA synthetase	136	51.5	47	0.35	0.7615
Phenylalanyl-tRNA synthetase alpha	85	32.2	36	0.42	0.1688
Preprotein translocase subunit SecY	134	50.7	46	0.34	0.7764
Valyl-tRNA synthetase	120	45.4	47	0.39	0.3496
alanyl tRNA synthetase	106	40.1	43	0.41	0.2514
arginyl tRNA synthetase	70	26.5	36	0.51	0.0077
aspartyl tRNA synthetase	100	37.9	37	0.37	0.5295
gyrA	115	43.5	42	0.37	0.5796
leucyl-tRNA synthetase	106	40.1	37	0.35	0.7017
recA	129	48.8	46	0.36	0.6656
ribosomal protein L1	100	37.9	22	0.22	0.9995
ribosomal protein L10	89	33.7	24	0.27	0.9799
ribosomal protein L11	119	45.1	30	0.25	0.9977
ribosomal protein L13	113	42.8	35	0.31	0.9237
ribosomal protein L14	117	44.3	35	0.30	0.9557
ribosomal protein L15	119	45.1	32	0.27	0.9924
ribosomal protein L16-L10E	106	40.1	31	0.29	0.9605
ribosomal protein L17	116	43.9	41	0.35	0.6791
ribosomal protein L18	117	44.3	35	0.30	0.9557
ribosomal protein L19	87	32.9	36	0.41	0.2169
ribosomal protein L2	104	39.4	33	0.32	0.8849
ribosomal protein L20	104	39.4	39	0.38	0.4900
ribosomal protein L21	97	36.7	28	0.29	0.9600
ribosomal protein L22	99	37.5	31	0.31	0.8949
ribosomal protein L23	113	42.8	32	0.28	0.9789
ribosomal protein L24	107	40.5	27	0.25	0.9961
ribosomal protein L27	99	37.5	26	0.26	0.9901
ribosomal protein L29	109	41.3	29	0.27	0.9913
ribosomal protein L3	111	42	34	0.31	0.9322
ribosomal protein L30	91	34.4	23	0.25	0.9924
ribosomal protein L4	107	40.5	31	0.29	0.9661
ribosomal protein L5	105	39.7	31	0.30	0.9541
ribosomal protein L6P-L9E	125	47.3	36	0.29	0.9790
ribosomal protein S10	116	43.9	29	0.25	0.9977
ribosomal protein S11	131	49.6	36	0.27	0.9920
ribosomal protein S12	120	45.4	32	0.27	0.9936
ribosomal protein S13	139	52.6	40	0.29	0.9846
ribosomal protein S15	108	40.9	38	0.35	0.6829
ribosomal protein S16	87	32.9	26	0.30	0.9253
ribosomal protein S17	122	46.2	36	0.30	0.9670
ribosomal protein S18	90	34.1	26	0.29	0.9528
ribosomal protein S19	101	38.2	33	0.33	0.8364
ribosomal protein S2	91	34.4	27	0.30	0.9362
ribosomal protein S20	92	34.8	33	0.36	0.6125
ribosomal protein S3	97	36.7	30	0.31	0.9063
ribosomal protein S4	137	51.9	46	0.34	0.8301
ribosomal protein S5	137	51.9	40	0.29	0.9792
ribosomal protein S6	105	39.7	32	0.30	0.9304
ribosomal protein S7	139	52.6	41	0.29	0.9761
ribosomal protein S8	118	44.7	35	0.30	0.9616
ribosomal protein S9	118	44.7	33	0.28	0.9848

Table S8: Statistics of 51 marker gene families in the 5 m sample. Expected, observed columns: expected (based on 34.6 % of bps) and observed number of proteins on ≤ 2x coverage reads.

	Total	**Expected**	**Observed**	**Observed freq**	**p-value**
Total	5852	2025	1827	0.31	1.0000
Histidyl-tRNA synthetase	134	46.3	42	0.31	0.7570
Phenylalanyl-tRNA synthetase alpha	118	40.8	40	0.34	0.5213
Preprotein translocase subunit SecY	107	37	29	0.27	0.9389
Valyl-tRNA synthetase	151	52.2	49	0.32	0.6782
alanyl tRNA synthetase	103	35.6	37	0.36	0.3466
arginyl tRNA synthetase	92	31.8	30	0.33	0.6108
aspartyl tRNA synthetase	105	36.3	39	0.37	0.2559
gyrA	135	46.7	51	0.38	0.1924
leucyl-tRNA synthetase	103	35.6	35	0.34	0.5071
recA	118	40.8	45	0.38	0.1825
ribosomal protein L1	91	31.4	28	0.31	0.7425
ribosomal protein L10	127	43.9	44	0.35	0.4548
ribosomal protein L11	127	43.9	50	0.39	0.1114
ribosomal protein L13	83	28.7	24	0.29	0.8347
ribosomal protein L14	149	51.5	46	0.31	0.8073
ribosomal protein L15	105	36.3	33	0.31	0.7168
ribosomal protein L16-L10E	139	48	44	0.32	0.7373
ribosomal protein L17	57	19.7	21	0.37	0.3066
ribosomal protein L18	113	39	36	0.32	0.6936
ribosomal protein L19	84	29	37	0.44	0.0281
ribosomal protein L2	142	49.1	40	0.28	0.9378
ribosomal protein L20	115	39.7	34	0.30	0.8504
ribosomal protein L21	95	32.8	37	0.39	0.1589
ribosomal protein L22	111	38.4	28	0.25	0.9780
ribosomal protein L23	160	55.3	42	0.26	0.9851
ribosomal protein L24	139	48	44	0.32	0.7373
ribosomal protein L27	87	30.1	37	0.43	0.0494
ribosomal protein L29	108	37.3	22	0.20	0.9991
ribosomal protein L3	142	49.1	39	0.27	0.9572
ribosomal protein L30	67	23.1	16	0.24	0.9598
ribosomal protein L4	138	47.7	36	0.26	0.9797
ribosomal protein L5	134	46.3	41	0.31	0.8109
ribosomal protein L6P-L9E	136	47	40	0.29	0.8822
ribosomal protein S10	163	56.3	46	0.28	0.9501
ribosomal protein S11	89	30.7	25	0.28	0.8820
ribosomal protein S12	143	49.4	39	0.27	0.9622
ribosomal protein S13	110	38	31	0.28	0.9072
ribosomal protein S15	94	32.5	31	0.33	0.5836
ribosomal protein S16	83	28.7	30	0.36	0.3370
ribosomal protein S17	146	50.5	42	0.29	0.9199
ribosomal protein S18	68	23.5	21	0.31	0.6938
ribosomal protein S19	131	45.3	35	0.27	0.9664
ribosomal protein S2	111	38.4	39	0.35	0.4099
ribosomal protein S20	88	30.4	30	0.34	0.4907
ribosomal protein S3	133	46	41	0.31	0.7940
ribosomal protein S4	94	32.5	25	0.27	0.9383
ribosomal protein S5	133	46	39	0.29	0.8835
ribosomal protein S6	61	21.1	25	0.41	0.1193
ribosomal protein S7	154	53.2	41	0.27	0.9787
ribosomal protein S8	142	49.1	43	0.30	0.8398
ribosomal protein S9	94	32.5	28	0.30	0.8078

Table S9: Statistics of 51 marker gene families in the 6 m sample. Expected, observed columns: expected (based on 59.5 % of bps) and observed number of proteins on ≤2x coverage reads.

	Total	**Expected**	**Observed**	**Observed freq**	**p-value**
Total	5125	3049.375	2486	0.49	1.0000
Histidyl-tRNA synthetase	111	66.2	61	0.55	0.8106
Phenylalanyl-tRNA synthetase alpha	100	59.7	57	0.57	0.6600
Preprotein translocase subunit SecY	103	61.4	52	0.50	0.9602
Valyl-tRNA synthetase	108	64.4	63	0.58	0.5616
alanyl tRNA synthetase	91	54.3	49	0.54	0.8394
arginyl tRNA synthetase	75	44.7	44	0.59	0.5147
aspartyl tRNA synthetase	101	60.2	58	0.57	0.6289
gyrA	123	73.4	73	0.59	0.4793
leucyl-tRNA synthetase	98	58.5	49	0.50	0.9642
recA	125	74.6	65	0.52	0.9463
ribosomal protein L1	87	51.9	38	0.44	0.9979
ribosomal protein L10	109	65	49	0.45	0.9985
ribosomal protein L11	116	69.2	54	0.47	0.9967
ribosomal protein L13	100	59.7	58	0.58	0.5831
ribosomal protein L14	108	64.4	43	0.40	1.0000
ribosomal protein L15	98	58.5	48	0.49	0.9775
ribosomal protein L16-L10E	103	61.4	41	0.40	1.0000
ribosomal protein L17	50	29.8	31	0.62	0.3097
ribosomal protein L18	96	57.3	42	0.44	0.9987
ribosomal protein L19	108	64.4	52	0.48	0.9889
ribosomal protein L2	116	69.2	46	0.40	1.0000
ribosomal protein L20	101	60.2	61	0.60	0.3903
ribosomal protein L21	71	42.3	38	0.54	0.8177
ribosomal protein L22	111	66.2	40	0.36	1.0000
ribosomal protein L23	125	74.6	49	0.39	1.0000
ribosomal protein L24	98	58.5	39	0.40	0.9999
ribosomal protein L27	71	42.3	40	0.56	0.6656
ribosomal protein L29	93	55.5	37	0.40	0.9999
ribosomal protein L3	108	64.4	47	0.44	0.9994
ribosomal protein L30	82	48.9	36	0.44	0.9969
ribosomal protein L4	115	68.6	47	0.41	1.0000
ribosomal protein L5	101	60.2	42	0.42	0.9998
ribosomal protein L6P-L9E	110	65.6	49	0.45	0.9989
ribosomal protein S10	108	64.4	44	0.41	0.9999
ribosomal protein S11	96	57.3	52	0.54	0.8317
ribosomal protein S12	94	56.1	51	0.54	0.8242
ribosomal protein S13	99	59.1	51	0.52	0.9344
ribosomal protein S15	80	47.7	48	0.60	0.4216
ribosomal protein S16	79	47.1	38	0.48	0.9735
ribosomal protein S17	115	68.6	45	0.39	1.0000
ribosomal protein S18	76	45.3	45	0.59	0.4769
ribosomal protein S19	114	68	44	0.39	1.0000
ribosomal protein S2	102	60.8	58	0.57	0.6724
ribosomal protein S20	92	54.9	47	0.51	0.9371
ribosomal protein S3	108	64.4	39	0.36	1.0000
ribosomal protein S4	109	65	64	0.59	0.5300
ribosomal protein S5	114	68	53	0.46	0.9966
ribosomal protein S6	96	57.3	49	0.51	0.9426
ribosomal protein S7	118	70.4	55	0.47	0.9969
ribosomal protein S8	108	64.4	45	0.42	0.9999
ribosomal protein S9	105	62.6	60	0.57	0.6546

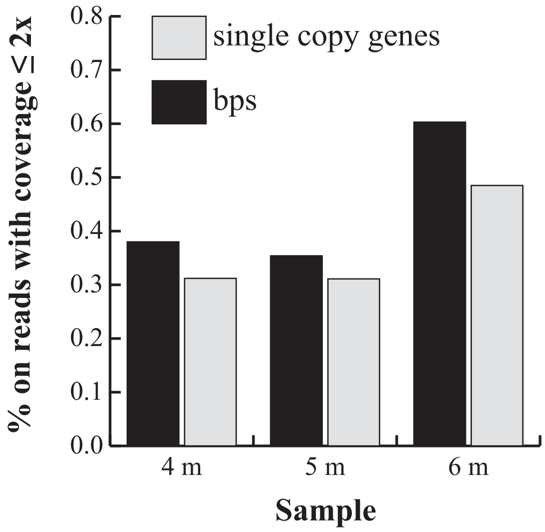

Figure S18: fraction of single copy genes (gray) and base-pairs on synthetic long-reads ≥ 5 kbp with ≤2x coverage. Fraction of single copy genes can be used as a proxy for the fraction of cells represented in the two sets.

Table S10: protein MCL clusters significantly more abundant on synthetic long-reads with ≤2x coverage, 4 m sample. Bonferroni correction was applied to adjust the 0.05 p-value threshold (adjusted p-value is 1.7e-5).

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
1248	Prepilin-type N-terminal cleavage/methylation domain-containing protein	40	15	39	0.98	0
39	Oxidoreductase	245	93	147	0.60	8.6E-12
1982	Hypothetical	26	10	25	0.96	1.1E-11
1463	Hypothetical	34	13	30	0.88	1.3E-10
98	LacI family transcriptional regulator	175	66	105	0.60	1.1E-09
88	Uroporphyrinogen-III decarboxylase	163	62	98	0.60	3.0E-09
18	Aminotransferase DegT/DnrJ/EryC1/StrS aminotransferase	404	153	209	0.52	6.3E-09
745	Tetratricopeptide repeat protein (a structural motif, does not say much about the function)	62	24	43	0.69	1.3E-07
2509	ABC transporter substrate binding protein	23	9	20	0.87	1.5E-07
1101	Glycoside hydrolase family 4	46	17	33	0.72	7.6E-07
43	ABC transporter	240	91	126	0.53	1.6E-06
37	Glycosyl transferase family 1	256	97	133	0.52	1.9E-06
2622	Transposase	20	8	17	0.85	2.0E-06
1427	Sulfatase	36	14	26	0.72	6.7E-06
654	Glycosyl transferase group 1	68	26	43	0.63	6.9E-06
90	Oxidoreductase	170	65	91	0.54	1.3E-05

Table S11: protein MCL clusters significantly more abundant on synthetic long-reads with ≤ 2x coverage, 5 m sample. Bonferroni correction was applied to adjust the 0.05 p-value (adjusted p-value is 1.8e-5).

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
2105	TonB-dependent receptor	24	9	21	0.88	8.9E-09
625	Hypothetical	67	24	44	0.66	5.7E-08
672	Mandelate racemase	64	23	41	0.64	4.2E-07
2555	Von Willebrand factor type A	20	7	17	0.85	4.3E-07
375	Oxidoredctase	88	31	52	0.59	8.3E-07
1861	Hypothetical	29	10	22	0.76	1.1E-06
1563	N-acetyltransferase GCN5	30	11	22	0.73	3.1E-06
1175	Hypothetical/membrane protein	40	14	27	0.68	5.4E-06
776	Coenzyme F390 synthetase/ Capsular polysaccharide biosynthesis protein	59	20	36	0.61	9.9E-06
399	Oxidoredctase	87	30	49	0.56	1.1E-05
1249	Uncharacterized	39	14	26	0.67	1.1E-05

Table S12: protein MCL clusters significantly more abundant on synthetic long-reads with ≤ 2x coverage, 6 m sample. Bonferroni correction was applied to adjust the 0.05 p-value (adjusted p-value is 2.1e-5).

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
866	Permease	49	29	49	1.00	0
2157	ABC-type Fe3+ transport system	21	13	21	1.00	0
2252	Uncharacterized	20	12	20	1.00	0
1724	Outer membrane receptor	26	15	26	1.00	0
1796	RNA polymerase, sigma subunit, ECF family	25	15	25	1.00	0
101	Oxidoreductase	128	76	114	0.89	2.5E-14
911	Sulfatase	47	28	45	0.96	8.3E-10
130	Oxidoreductase	117	70	98	0.84	4.0E-09
383	Mandelate racemase	81	48	71	0.88	5.4E-09
1268	Sulfatase	36	21	35	0.97	7.6E-09
204	MmgE/PrpD family protein	98	58	83	0.85	1.6E-08
1386	Prepilin-type N-terminal Cleavage/methylation domain	34	20	33	0.97	2.2E-08
1121	Lyase	40	24.1	37	0.93	6.0E-07
1026	Oxidoreductase	42	25.3	38	0.90	2.2E-06
601	Phage integrase	61	36.8	52	0.85	5.3E-06
54	Glycosyl transferase family 1	178	107.3	134	0.75	8.9E-06
1159	PadR family transcriptional regulator	39	23.5	35	0.90	8.0E-06
21	Transporter related binding/receprot protein	293	175	210	0.72	5.5E-06
2	Reductase	1096	653	722	0.66	6.1E-06
1943	Pyrrolo-quinoline quinone	23	13.9	22	0.96	8.9E-06
1668	Transporter	27	16	25	0.93	1.6E-05
137	Asparagine synthetase	115	69	89	0.77	1.6E-05

Table S13: protein families with more than 1,000 members, 4 m sample.

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
1	ABC transporter ATP-binding protein	1811	697.2	503	0.27	1
2	Dehydrogenase	1327	510.9	481	0.36	0.95
3	ABC transporter ATP-binding protein	1073	413.1	380	0.35	0.98

Table S14: protein families with more than 1,000 members, 5 m sample.

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
1	ABC transporter ATP-binding protein	1580	548	427	0.27	1.0000
2	Dehydrogenase	1186	411	456	0.38	0.0026
3	ABC transporter ATP-binding protein	1098	381	321	0.29	0.9999

Table S15: protein families with more than 1,000 members, 6 m sample.

Family	Annotation	Total	Expected	Observed	Observed freq	p-value
1	ABC transporter ATP-binding protein	1167	695	601	0.51	1.0000
2	Dehydrogenase	1096	653	722	0.66	6.1E-06
3	ABC transporter ATP-binding protein	1049	625	581	0.55	0.9962

Table S16: Enriched KEGG-terms in synthetic long-reads with ≤ 2x coverage, 4 m sample. Bonferroni correction was applied to adjust the 0.05 p-value threshold (adjusted p-value is 3.8e-5).

KEGG-term	Annotation	Total	Expected	Observed	Observed freq	p-value
K07406	alpha-galactosidase [EC:3.2.1.22]	49	18.5	36	0.73	9.6E-08
K01599	uroporphyrinogen decarboxylase [EC:4.1.1.37]	345	130.7	177	0.51	1.6E-07
K02025	multiple sugar transport system permease protein	251	95.1	133	0.52	4.6E-07
K02392	flagellar basal-body rod protein FlgG	29	10.9	22	0.75	6.5E-06
K02026	multiple sugar transport system permease protein	259	98.1	131	0.5	1.3E-05

Table S17: enriched KEGG-terms in synthetic long-reads with ≤ 2x coverage, 5 m sample. Bonferroni correction was applied to adjust the 0.05 p-value threshold (adjusted p-value is 3.7e-5).

KEGG-term	Annotation	Total	Expected	Observed	Observed freq	p-value
K07406	alpha-galactosidase [EC:3.2.1.22]	49	18.5	36	0.73	9.6E-08
K01599	uroporphyrinogen decarboxylase [EC:4.1.1.37]	345	130.7	177	0.51	1.6E-07
K02025	multiple sugar transport system permease protein	251	95.1	133	0.52	4.6E-07
K02392	flagellar basal-body rod protein FlgG	29	10.9	22	0.75	6.5E-06
K02026	multiple sugar transport system permease protein	259	98.1	131	0.5	1.3E-05

Table S18: enriched KEGG-terms in synthetic long-reads with ≤ 2x coverage, 6 m sample. Bonferroni correction was applied to adjust the 0.05 p-value threshold (adjusted p-value is 3.8e-5).

KEGG-term	Annotation	Total	Expected	Observed	Observed freq	p-value
K07812	trimethylamine-N-oxide reductase (cytochrome c) 2 [EC:1.7.2.3]	47	27.9	43	0.91	1.4E-07
K01953	asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]	166	98.7	126	0.75	2.9E-06
K07031	D-glycero-alpha-D-manno-heptose-7-phosphate kinase [EC:2.7.1.168]	48	28.5	41	0.85	2.3E-05
K02026	multiple sugar transport system permease protein	264	157	188	0.71	2.9E-05

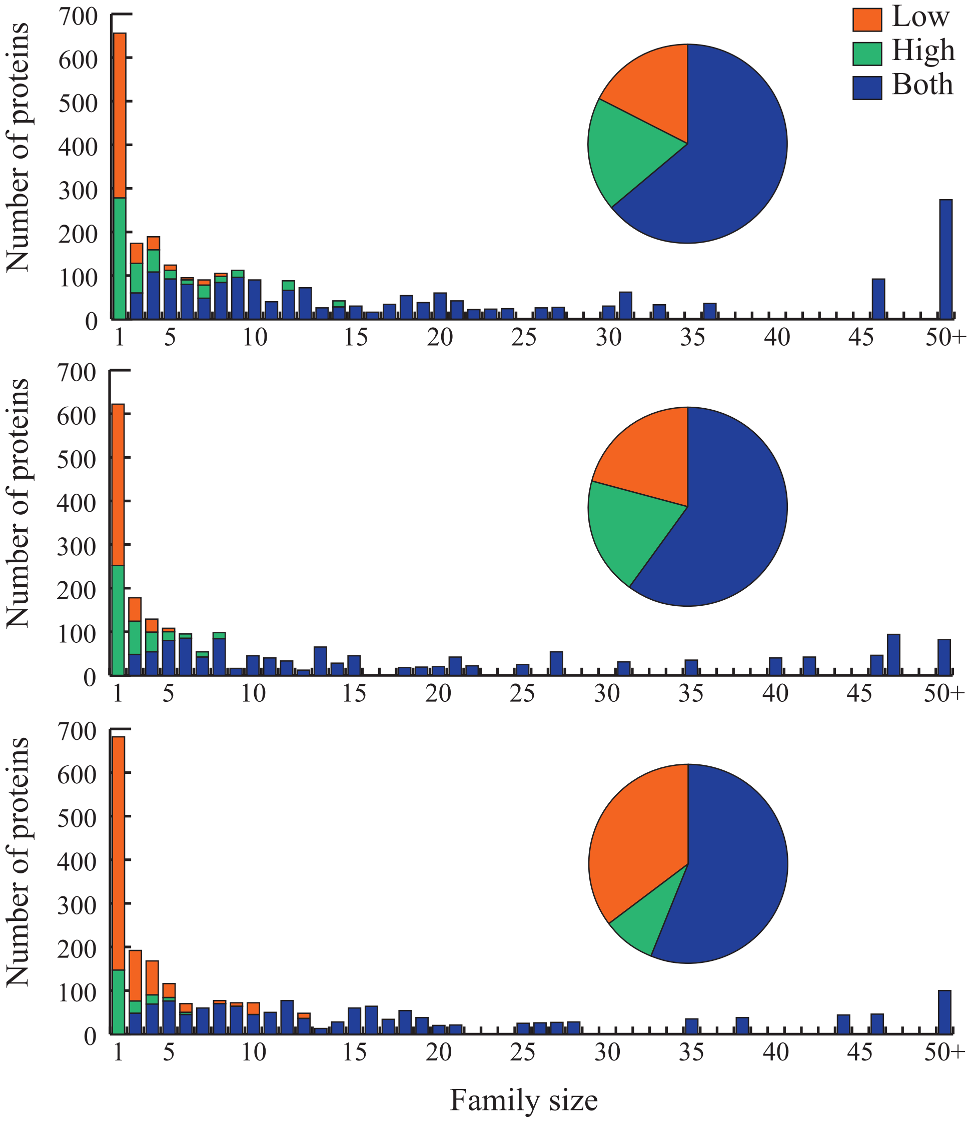

Figure S19: Total number of proteins in glycosyl hydrolase families of different sizes. Families with representatives in low (≤ 2x coverage), high (> 2x coverage) and both scaffolds are reported. Number of families that are unique to low coverage scaffolds (418, 409 and 637 for the 4, 5 and 6 m samples, respectively) was higher than the number of families unique to high coverage scaffolds (348, 316, 141) and the number of families common to both fractions (198, 140, 153); however most of the families that are unique to one of the coverage fractions are singletons.

Table S19: list of terms used for identifying glycosyl hydrolases in the ggKBase platform (refer to http://ggkbase.berkeley.edu/custom_lists/5572-Glycosyl_hydrolase for the implementation of the list in ggKBase).

cellobiosidase	Dextranase	acetylmuramidase
glycosidase	Polygalacturonase	isomaltosidase
glycosyl	lysozyme	isomaltotriosidase
hydrolase	sialidase	maltohexaosidase
endoglucanase	fructofuranosidase	mannobiosidase
glycoside hydrolase	trehalase	lactase
cellulase	hyaluronoglucosaminidase	endogalactosaminidase
chitinase	arabinosidase	maltotriohydrolase
2.4.1.18	pullulanase	EC:3.2.1
glycogen debranching enzyme	glucosylceramidase	polymannuronate hydrolase
galacturonase	galactosylceramidase	octulosonidase
mannosidase	acetylgalactosaminidase	glucuronosidase
arabinase	acetylglucosaminidase	chitosanase
glucuronidase	acetylhexosaminidase	maltohydrolase
xyloglucanase	cyclomaltodextrinase	difructose-anhydride synthase
xyloglycosyltransferase	maltotetraohydrolase	biosidase
mannanase	mycodextranase	cellobiohydrolase
xylanase	glycosylceramidase	alpha-neoagaro-oligosaccharide hydrolase
xylosidase	levanbiohydrolase	glucosaminidase
arabinofuranosidase	levanase	GlcNAcase
galactanase	quercitrinase	mannosylglycerate hydrolase
galactosidase	galacturonidase	rhamnogalacturonan hydrolase
glucoronidase	licheninase	rhamnogalacturonyl hydrolase
rhamnosidase	isoamylase	galacturonohydrolase
fucosidase	iduronidase	rhamnohydrolase
amylase	fructosidase	xylohydrolase
glucosidase	agarase	porphyranase
glucanase	galacturonosidase	glucuronyl hydrolase
Inulinase	carrageenase	chondroitin disaccharide hydrolase

References

1. Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, et al. (2013) Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. Nat Commun 4: 2120.
2. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, et al. (2013) Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome 1: 22.
3. Voskoboynik A, Neff NF, Sahoo D, Newman AM, Pushkarev D, et al. (2013) The genome sequence of the colonial chordate, Botryllus schlosseri. Elife 2: e00569.
4. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, et al. (2014) Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol.
5. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, et al. (2014) Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly repetitive transposable elements. BioRxivorg.
6. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28: 1420-1428.
7. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453.
8. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147: 195-197.
9. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, et al. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res 23: 111-120.
10. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, et al. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11: 119.
11. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.
12. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T, et al. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100-3108.
13. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41: D590-596.
14. Nawrocki PE (2009) Structural RNA Homology Search and Alignment using Covariance Models. PhD thesis, Washington University in Saint Louis, School of Medicine.
15. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics.
16. Hugenholtz P, Pitulle C, Hershberger KL, Pace NR (1998) Novel division level bacterial diversity in a Yellowstone hot spring. J Bacteriol 180: 366-376.
17. Wu M, Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. Genome Biol 9: R151.

18. Janda JM, Abbott SL (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45: 2761-2764.
19. Ribeca P, Valiente G (2011) Computational challenges of sequence classification in microbiomic data. Brief Bioinform 12: 614-625.
20. Vandamme P, Pot B, Gillis M, de Vos P, Kersters K, et al. (1996) Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev 60: 407-438.