Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture

Connor B. Driscoll, Timothy G. Otten, Nathan M. Brown and Theo W. Dreher

Abstract

Here we report three complete bacterial genome assemblies from a PacBio shotgun metagenome of a co-culture from Upper Klamath Lake, OR. Genome annotations and culture conditions indicate these bacteria are dependent on carbon and nitrogen fixation from the cyanobacterium *Aphanizomenon flos-aquae*, whose genome was assembled to draft-quality. Due to their taxonomic novelty relative to previously sequenced bacteria, we have temporarily designated these bacteria as *incertae sedis Hyphomonadaceae* strain UKL13-1 (3,501,508 bp and 56.12% GC), *incertae sedis Betaproteobacterium* strain UKL13-2 (3,387,087 bp and 54.98% GC), and *incertae sedis Bacteroidetes* strain UKL13-3 (3,236,529 bp and 37.33% GC). Each genome consists of a single circular chromosome with no identified plasmids. When compared with binned Illumina assemblies of the same three genomes, there was ~7% discrepancy in total genome length. Gaps where Illumina assemblies broke were often due to repetitive elements. Within these missing sequences were essential genes and genes associated with a variety of functional categories. Annotated gene content reveals that both *Proteobacteria* are aerobic anoxygenic phototrophs, with Betaproteobacterium UKL13-2 potentially capable of phototrophic oxidation of sulfur compounds. Both proteobacterial genomes contain transporters suggesting they are scavenging fixed nitrogen from *A. flos-aquae* in the form of ammonium. *Bacteroidetes* UKL13-3 has few completely annotated biosynthetic pathways, and has a comparatively higher proportion of unannotated genes. The genomes were detected in only a few other freshwater metagenomes, suggesting that these bacteria are not ubiquitous in freshwater systems. Our results indicate that long-read sequencing is a viable method for sequencing dominant members from low-diversity microbial communities, and should be considered for environmental metagenomics when conditions meet these requirements.

Keywords: Aerobic anoxygenic phototroph, *Aphanizomenon flos-aquae*, *Hyphomonadaceae*, *Betaproteobacterium*, *Bacteroidetes*, PacBio SMRT sequencing
Introduction
Metagenomic sequencing is the process of sampling DNA sequences from multiple genomes in a community of organisms, and has been applied to many environmental samples to assess both functional diversity and species richness of microbial communities [1, 2]. Recently, there has been a progression in metagenomic approaches associated with advances in sequencing technologies. Next-generation sequencing methods [3] such as 454 and Illumina HiSeq/MiSeq have greatly reduced sequencing costs per base relative to Sanger sequencing due to increased throughput, which has facilitated shotgun metagenomics (randomly sequencing all DNA in a sample). This has provided several advantages over amplicon sequencing. For example, focus has shifted from assigning taxa using single genes to using multiple genes and sequence composition instead [2, 4]. It has also permitted functional characterization of individual representatives or whole microbial communities [5, 6].

However, there are technical hurdles associated with short-read sequencing. Specifically, assembling short reads (50–300 bp) into contiguous sequences (contigs) rarely leads to complete genome assemblies due to repetitive genomic elements such as 16S rRNA genes and IS elements [7] that are 1 kb or greater in length. There are two consequences as a result. First, closing draft genomes by primer walking requires considerable manual effort and time. Second, if closure is not possible, contigs must be clustered and binned using methods like differential coverage [8], co-abundance [5, 9, 10], or gene/nucleotide composition [11]. While useful, these methods are often not comprehensive and become even more difficult to implement when used in a metagenomic context, where multiple genomes (sometimes from closely related organisms) must be delineated [12].

Single-molecule real time sequencing technologies, such as PacBio and Oxford Nanopore, are part of the third-generation sequencing wave [7]. These sequencers produce average read lengths in the 5–50 kb range, with ~50% of reads longer than 14 kb [13], which exceeds the size of repetitive elements in the average bacterial genome. Although more error-prone, these longer reads have proven advantageous for assembling closed genomes if sequencing depth is high enough to allow error correction [14]. To date, long-read sequencing has rarely been used for metagenomics for several reasons: 1) the amount of sequence data returned is a fraction of an Illumina run (up to 750 Gb/flow cell of Illumina HiSeq 3000 vs. up to 1 Gb/SMRT cell of PacBio Sequel based on company specifications), 2) the sequencing cost per base pair is higher, and 3) PacBio does not rely upon DNA amplification, so high concentrations of raw DNA are required. Due to these limitations, long-read metagenomics has so far been limited to whole-16S amplicon sequencing [15] and to improving binning from fragmented (short-read) assemblies [16].

Here, we have generated a PacBio shotgun metagenome from a non-axenic cyanobacterium culture established in summer 2013 originating from Upper Klamath Lake, OR. In this freshwater lake, the $N_2$-fixing filamentous cyanobacterium *Aphanizomenon flos-aquae* blooms annually. These blooms are harvested and sold as nutritional supplements. Little is known about the co-occurring microbial community in this lake, whose composition could be influenced by the presence of *A. flos-aquae* as the dominant primary producer [17, 18]. By applying a selective growth medium lacking nitrogen, our goal was to sequence and assemble complete genomes from a relatively simple community, in turn assessing the possibility for using PacBio shotgun sequencing for environmental metagenomics. We closed three novel bacterial genomes, which provide insight into putative metabolic dependencies of these bacteria on *A. flos-aquae* in the co-culture. However, we were unable to close the *A. flos-aquae* genome, which is in draft quality and will be discussed elsewhere.

Organism information
Classification and features
The taxonomic placement of each genome was assessed three ways (Additional file 1: Table S1). We used the SILVA SSU Ref NR database (accessed on March 9, 2016) to search for significant 16S rDNA matches [19]. Also, we generated 16S phylogenetic trees for each genome, using the SINA aligner [20] and FastTree [21], with all classified *Alphaproteobacteria, Betaproteobacteria*, and *Bacteroidetes* representatives in SILVA, shown with their nearest groups in Fig. 1a–c. For the second taxonomic placement method, we used PhyloPythiaS+ [4], which searches for genomes with similar k-mer composition. The third method, Phylosift [22], is a pipeline that aligns 40 marker genes to generate a weighted probability score for specific taxonomic assignments. Due to lack of similarity with previously-classified bacterial representatives, these approaches were unable to assign these genomes to genus or species levels. Phylogenetic analysis of 16S rRNA genes placed each of the novel bacteria between established clades within or between families, with 0.11–0.13 average substitutions per site to the nearest neighbor’s 16S gene (Fig. 1a–c). Phylogenomic analyses (PhylopythiaS+ and Phylosift) were also unable to find close relatives, resulting in the genomes being placed at higher taxonomic levels than genus (Additional file 1: Table S1). We have therefore used this information to designate these organisms as *incertae sedis* (of uncertain placement). While *Candidatus* designations are often assigned to unplaced taxa, the International Code of Nomenclature for Prokaryotes requires the *Candidatus*
Fig. 1 (See legend on next page.)

Driscoll et al. Standards in Genomic Sciences (2017) 12:9
Page 3 of 16
usage to be accompanied by phenotypic information [23], which we did not have available. We assigned the temporary strain names Hypomonadaceae UKL13-1, Betaproteobacterium UKL13-2, and Bacteroidetes UKL13-3 until further representative sequences become available to guide the naming of new genera as appropriate. Minimum Information about the Genome Sequences is summarized in Table 1.

Although we initiated and maintained this mixed-community culture for 1 year, the culture was lost and we did not obtain physiological information regarding these organisms. Sustaining long-term A. flos-aquae cultures is often difficult, and it is common for cultures to die. In the absence of phenotypic information, we discuss insights from the genome annotations for the three novel bacteria.

Genome sequencing information

Genome project history

Cultures were initiated from UKL, where annual A. flos-aquae blooms constitute a serious ecological disturbance but are also harvested and sold as nutritional supplements. The genome sequences were deposited to DDBJ/EMBL/GenBank under the accessions CP012156, CP012157, and CP012155 for the Hypomonadaceae UKL13-1, Betaproteobacterium UKL13-2, and Bacteroidetes UKL13-3 genomes, respectively. Project information is summarized in Table 2.

Growth conditions and genomic DNA preparation

One Aphanizomenon flos-aquae colony from a depth-integrated water sample from the UKL Link Dam site collected on 6 August 2013 was transferred to Bold 3 N0 medium [24] without NaN3 (or any other form of N). This medium consisted of 0.17 mM CaCl2, 0.3 mM MgSO4, 0.43 mM KH2PO4, 1.29 mM KH2PO4, 0.43 mM NaCl, P IV trace metals, and 0.1 μM vitamin B12 at pH 8.0. The culture was maintained under cool white fluorescent light (20 μE m-2 s-1) with a light/dark cycle of 16 h/8 h at 24 °C. Three separate DNA extractions were performed from this culture (Table 3). A sample taken in November 2013 was collected on a 1.2 μm GF/C filter (Whatman), and DNA was extracted for Illumina sequencing using a DNA extraction kit (GeneRite DNA-EZ RWOC1). A similarly collected sample (Nov 2013) was extracted using phenol-chloroform [25] and pooled with phenol-chloroform extracted DNA from an unfiltered sample of the culture collected during March 2015 (to balance the proportion of sequencing capacity associated with cyanobacteria and heterotrophic bacteria). This pooled sample was quantified with the Q32850 Quant-iT dsDNA BR Assay Kit. Approximately eight micrograms of DNA were submitted for PacBio sequencing.

Genome sequencing and assembly

The November 2013 sample was processed using a Nextera XT kit and sequenced using the Illumina HiSeq 2000 at the Oregon State University Center for Genome Research and Biocomputing to generate 17,617,259 paired-end reads (101 bp). The pooled (11/2013 & 3/2015) sample was processed for PacBio sequencing by the Molecular Biology and Genomics Core at Washington State University. Eight SMRT cells of PacBio RS sequencing generated 348,623 reads with an average length of 7,737 bp. PacBio sequences were assembled using HGAP [26] with three different parameter sets to optimize for assembly of different genomes (Additional file 2: Table S2). Initially, only the Bacteroidetes genome assembled from two SMRT cells (167,289 PacBio reads), at a seed read length cutoff of 12.8 kb. The less abundant Hypomonadaceae UKL13-1 and Betaproteobacterium UKL13-2 genomes required all eight SMRT cells to close (348,623 reads). While the Betaproteobacterium genome closed with a seed read-length cutoff of 13.6 kb, the Hypomonadaceae genome only assembled completely when this cutoff was lowered to 6 kb, likely since it had the lowest coverage of the three genomes. A lower cutoff directs more reads towards use in assembling, thereby improving chances of completing low-coverage assemblies [27]. However, this also reduces the number of aligned reads used for error correction, which can in turn affect assembly quality. These tradeoffs should be considered before performing assemblies, but it is notable that we would not have completed the Hypomonadaceae UKL13-1 genome without lowering this cutoff. The Hypomonadaceae, Betaproteobacterium, and Bacteroidetes genomes were of finished
quality (Table 4), with each having average Phred scores (ASCII base 33) of 75.9, 76.0, and 81.9, respectively. Although HGAP assembly involves chimera detection, we additionally evaluated the possibility of chimeric assembly of each genome by mapping the Illumina reads to the completed genomes using REAPR [28], which breaks incorrect assemblies by assessing the paired-read coverage distribution at each base; no chimeras were identified. We were unable to complete other genomes in the culture, including the draft-quality A. flos-aquae genome assembly (Table 4).

The Illumina-sequenced culture was assembled using the IDBA-Hybrid [29] software. We binned Illumina-assembled contigs from the three completed genomes by differential coverage of reads from both PacBio and Illumina samples. That is, Illumina and PacBio reads were separately mapped to each assembly using BWA-MEM [30] and BLASR [31], respectively.
Contigs were then binned using the mmgenome R package [8] (Table 5).

**Genome annotation**

All genomes were annotated with NCBI’s PGAP [32] and PROKKA [33]. Counts of features (Genes, CDS, pseudogenes, rRNAs, tRNAs, ncRNAs, and CRISPR arrays) come from PGAP annotations. Amino acid sequences were assigned to COG categories by searching against the COG protein database [34] using RAP-SEARCH [35], taking only the top hits above an E-value of 1E-30. Amino acid sequences from each genome were also annotated using the KEGG database [36] with the “genus_prokaryotes” database on September 3, 2015.

**Genome properties**

Each genome assembled into one closed contig, whose properties and statistics are shown in Table 6. The *Hyphomonadaceae* UKL13-1 genome consists of a single circular chromosome 3,501,508 bp long and a GC content of 56.12%. The genome contains a total of 3255 predicted genes, including 2934 predicted protein-coding sequences, 277 pseudogenes, and 44 RNA genes (40 tRNAs, one copy of the 16S-23S-5S rRNA operon, and 1 ncRNA) (Fig. 2). The Betaproteobacterium UKL13-2 genome consists of a single circular chromosome 3,387,087 bp long and a GC content of 54.98%. The genome contains a total of 3087 predicted genes, including 2772 predicted protein-coding sequences, 265 pseudogenes, and 50 RNA genes (43 tRNAs, two copies of the 16S-23S-5S rRNA operon, and 1 ncRNA) (Fig. 3). The *Bacteroidetes* UKL13-3 genome consists of a single circular chromosome 3,236,529 bp long and a GC content of 37.33%. The genome contains a total of 2850 predicted genes, including 2598 protein-coding sequences, 211 pseudogenes, and 41 RNA genes (35 tRNAs and two copies of the 16S-23S-5S rRNA operon) (Fig. 4). The distribution of genes into COG functional categories is summarized in Table 7.

**Insights from the genome sequence**

**PacBio metagenome and comparison to Illumina metagenome**

The bacterial community associated with the *Aphanizomenon flos-aquae* culture was subjected to metagenomic analysis with sequencing on eight PacBio SMRT cells, resulting in three completed novel bacterial genomes: *Hyphomonadaceae* UKL13-1, Betaproteobacterium UKL13-2, and *Bacteroidetes* UKL13-3 (Table 4). There were insufficient reads to close the genome of *A. flos-aquae,* although 67 contigs could be clustered to represent an estimated 97% of the genome (Table 4). The *A. flos-aquae* genome was sequenced with lower coverage than the three completed genomes, and additional sequencing would be needed for genome completion. Contigs from partial genomes of two additional bacteria were also clustered: a novel *Flavobacterium* (63% estimated genome

### Table 2 Project information

| MIGS ID | Property | Hyphomonadaceae UKL13-1 | Betaproteobacterium UKL13-2 | Bacteroidetes UKL13-3 |
|---------|----------|-------------------------|-----------------------------|-----------------------|
| MIGS-31 | Finishing quality | Complete | Complete | Complete |
| MIGS-28 | Libraries used | SMRT library prep | SMRT library prep | SMRT library prep |
| MIGS-29 | Sequencing platform | PacBio | PacBio | PacBio |
| MIGS-31.2 | Fold coverage | 94x | 143x | 112x |
| MIGS-30 | Assemblers | HGAP | HGAP | HGAP |
| MIGS-32 | Gene calling method | GeneMarkS+ | GeneMarkS+ | GeneMarkS+ |
|          | Locus tag | AEM38 | AEM42 | AEM51 |
|          | GenBank ID | CP012156 | CP012157 | CP012155 |
|          | GenBank date of release | March 30, 2016 | March 30, 2016 | March 30, 2016 |
|          | GOLD ID | Gp0126808 | Gp0126809 | Gp0126810 |
|          | BIOPROJECT | PRJNA290648 | PRJNA290650 | PRJNA290651 |
| MIGS-13  | Source material identifier | UKL13 | UKL13 | UKL13 |
|          | Project relevance | Environmental | Environmental | Environmental |

### Table 3 DNA extraction procedures and respective sequencing technologies

| Extraction | Handling | Sample date(s) | Sequencing |
|------------|----------|----------------|------------|
| 1          | 1.2 μm GF/C filtration | GeneRiteKit (silica beads) | Illumina 100 bp paired-end HiSeq 2000 |
| 2          | 1.2 μm GF/C filtration and whole sample | Phenol-chloroform | PacBio RS |

**Table 2 Project information**

**Table 3 DNA extraction procedures and respective sequencing technologies**
and Bacteroidetes bin contained more contigs than the Hyphomonadaceae (~100 or more contigs. The Betaproteobacterium genome predominant genomes separated into bins containing blies. When assembled with Illumina reads, the three comparison of PacBio-only and Illumina-only assem-
cate the presence of at least six separate bacterial taxa in this identified in the Brevundimonas contigs. Our results indi-
present the absence of at least six separate bacterial taxa in this non-axenic culture.

A parallel Illumina HiSeq 2000 metagenome allowed comparison of PacBio-only and illumina-only assem-
blies. When assembled with illumina reads, the three predominant genomes separated into bins containing ~100 or more contigs. The Betaproteobacterium genome bin contained more contigs than the *Hyphomonadaceae* and *Bacteroidetes* genomes, although it was sequenced at the highest Illumina depth of the three (63x coverage vs. 23x and 58x coverage, respectively) (Table 5). There was a ~200 kb discrepancy between Illumina bin length and completed genome length for each of the three genomes. The total binned contig lengths for the *Bacteroidetes* and Betaproteobacterium were each shorter than the completed genomes, while the *Hyphomonadaceae* total binned contig length was longer (Table 5). The additional sequences in the *Hyphomonadaceae* bin were primarily contigs shorter than 10 kb that were not part of the PacBio-assembled *Hyphomonadaceae* genome. The bin quality control program CheckM [38] overesti-
mated genome completeness or underestimated contamination when compared with the finished genome size. For example, CheckM estimated that the *Hyphomonadaceae* UKL13-1 bin contained ~2% contamination, while comparing the bin length with the completed genome length suggests ~6% contamination (Table 5). These discrepan-
cies indicate that genome binning has a tendency to exclude important sequences or include extraneous se-
quencies, and reveals the difficulty of assessing binned genome completeness and contamination without a reference. Incomplete binning is common for draft genomes, particularly from metagenomic assemblies [12].

We also assessed the extent to which genome repeats affected Illumina assemblies. Repeats in each genome were identified by using BLASTN to align each genome with itself, with a minimum E-value cutoff of 1E-30. Both intragenome BLASTN hits and missing Illumina coverage were then visualized with a circular genome plot (Additional file 3: Figure S1a-c). Breaks in Illumina assemblies commonly co-localized with intragenomic repeats in each genome. In particular, the Betaproteobacteri-
um UKL13-2 genome is enriched for repeat sequences relative to the other two genomes and contains larger regions unassembled by Illumina reads, factors that possibly contributed to the greater genome fragmenta-
tion (Table 5).

We then analyzed gene functions in sequences missing from illumina bins to assess the extent to which critical gene content was missing (Fig. 5). Most annotated genes in these regions were assigned to the mobilome category (X, esp. transposases), although genes from most other COG categories were also represented. Annotations within these regions included essential genes such as tRNAs, rRNA operons, translation-associated genes, and nucleotide metabolism genes, in addition to a variety of enzymes and transporters (Table 8; Additional file 4: Tables S3, Additional file 5: Table S4 and Additional file 6: Table S5). The presence of duplicated essential genes (DNA ligase, EF-Tu) resulted in both copies being absent from the Betaproteobacterium genome (Table 8); the pres-
ence of multiple rDNA sequences commonly produces

### Table 4 Genomes identified from PacBio assemblies. PacBio read coverage calculated by mapping with BLASR

| Genome                | Assembly Length (bp) | No. contigs | PB read coverage | Completeness estimate | Contamination estimate |
|-----------------------|----------------------|-------------|------------------|-----------------------|------------------------|
| *Hyphomonadaceae* UKL13-1 | 3,501,508            | 1           | 94x              | -                     | -                      |
| Betaproteobacterium UKL13-2 | 3,387,087            | 1           | 143x             | -                     | -                      |
| *Bacteroidetes* bacterium UKL13-3 | 3,236,529            | 1           | 112x             | -                     | -                      |
| Aphanizomenon flos-aquae  | 4,250,721            | 67          | 40x              | 96.67%                | 0.22%                  |
| Unknown *Flavobacterium* | 2,347,065            | 96          | 22x              | 62.67%                | 0.25%                  |
| Unknown *Caulobacterales* bacterium | 487,875              | 53          | 6x                | 17.15%                | 0.00%                  |

Completeness and contamination estimates for incomplete genomes are from CheckM.

### Table 5 Illumina assembly statistics for each genome. Contig number and assembly length are from extracted bins. Illumina coverage calculated by mapping with BWA-MEM. Bin coverage parameters used to bin illumina assemblies with mmgenome

| Genome                | Illumina coverage | # Illumina contigs | Bin coverage parameters | Bin assembly length (bp) | Bin assembly (% of genome) | Bin estimated completeness | Bin estimated contamination |
|-----------------------|-------------------|--------------------|-------------------------|--------------------------|-----------------------------|---------------------------|-----------------------------|
| *Hyphomonadaceae* UKL13-1 | 23x               | 122                | Illumina: 15-40x PacBio: >49x | 3,716,244               | 106.13%                     | 98.48%                    | 2.19%                       |
| Betaproteobacterium UKL13-2 | 63x               | 162                | Illumina: 37-87x PacBio: 71-211x | 3,131,899               | 92.47%                     | 96.15%                    | 1.42%                       |
| *Bacteroidetes* bacterium UKL13-3 | 58x               | 96                 | Illumina: 44-103x PacBio: >228x | 3,009,740               | 92.99%                     | 97.81%                    | 0.55%                       |

Assembly as % of genome is comparison of contig bin length with actual genome length. Completeness and contamination estimated with CheckM.
breaks in short-read assemblies [39]. In such cases, rDNA sequences confined to small contigs lose their linkage to other genes. This makes assigning 16S sequences to draft genomes difficult when multiple organisms are present in the same sample, and can make it difficult to link 16S amplicon information to binned genomes from shotgun metagenomes. Also, the functional variety of non-mobilome-associated missing genes within these assembly breaks suggests they hold informative sequences regarding physiology or lifestyle.

**Novel completed genomes**

To functionally characterize the three novel genomes, we searched all protein-coding sequences against the COG database using Rapsearch 2.16 and a 1E-30 E-value cutoff. We then repeated this for all bacterial genomes in GenBank (collected on November 3, 2015) and compared these to our novel genomes to assess enrichment of protein-coding sequences associated with each COG functional category. These are shown as a percentage of all protein-coding sequences from each respective genome (Additional file 7: Figure S2). Our results indicate that the *Hyphomonadaceae* UKL13-1 genome contains more lipid metabolism genes than most bacteria (at 5.01% of predicted coding sequences vs. a mean of 2.96%), while the *Bacteroidetes* UKL13-3 genome contains more cell wall/envelope/membrane biogenesis genes (7.39%, vs. a mean of 4.61%).

We then searched the KEGG database to identify complete and partial pathways in each genome. Identification of additional genes was aided by using Mauve whole- or partial-genome alignments [40] to reference genomes (*Cytophaga hutchinsonii, Roseobacter denitrificans, Rubrivivax gelatinosus, and Rhodobacter capsulatus*) and between *Hyphomonadaceae* UKL13-1 and Betaproteobacterium UKL13-2. The *Hyphomonadaceae* UKL13-1 and Betaproteobacterium UKL13-2 genomes contain anoxygenic photosynthesis and reaction center genes, as well as genes for bacteriochlorophyll and carotenoid synthesis (Additional file 8: Table S6). Phylogenies of their 16S genes reveal they do not cluster near groups containing phototrophic bacteria (Fig. 1a, b). Neither genome contains RuBisCO genes, consistent with these bacteria being aerobic anoxygenic phototrophs. These are a class of heterotrophs that use phototrophy to drive ATP and NAD(P)H production, but are unable to fix net carbon through photosynthesis [41, 42]. For Betaproteobacterium UKL13-2, the presence of genes for thiosulfate or sulfite oxidation (*soxABCDXYZ*) suggests that these sulfur compounds can serve as electron donors for ATP synthesis [42], perhaps in addition to organic compounds (Fig. 6). Both *A. flos-aquae* and Betaproteobacterium appear to be capable of assimilatory sulfate reduction of MgSO₄ (provided as the only S source in the growth medium) (Additional file 8: Table S6), which is often used as the pathway for amino acid synthesis. Photolithotrophic oxidation of reduced S compounds obtained from *A. flos-aquae* by Betaproteobacterium could be energetically advantageous. Since neither genes for oxidation of reduced sulfur or nitrogen compounds are evident in the *Hyphomonadaceae* genome, organic compounds likely serve as electron donors in this bacterium [41, 42] (Fig. 6). In contrast with the

| Attribute                        | *Hyphomonadaceae UKL13-1* | *Betaproteobacterium UKL13-2* | *Bacteroidetes UKL13-3* |
|----------------------------------|---------------------------|-------------------------------|------------------------|
| Genome size (bp)                 | 3,501,508                 | 3,387,087                     | 3,236,529              |
| DNA coding (bp)                  | 3,166,294                 | 3,017,556                     | 2,922,707              |
| DNA G+C (bp)                     | 1,964,937                 | 1,862,116                     | 1,208,228              |
| DNA scaffolds                    | 1                         | 1                             | 1                      |
| Total genes                      | 3255                      | 3087                          | 2850                   |
| Protein coding genes             | 2934                      | 2772                          | 2598                   |
| RNA genes                        | 44                        | 50                            | 41                     |
| Pseudo genes                     | 277                       | 265                           | 211                    |
| Genes in internal clusters       | -                         | -                             | -                      |
| Genes with function prediction   | 2459                      | 2300                          | 1872                   |
| Genes assigned to COGs           | 2156                      | 2078                          | 1696                   |
| Genes with Pfam domains          | 2697                      | 2489                          | 2066                   |
| Genes with signal peptides       | 382                       | 235                           | 301                    |
| Genes with transmembrane helices | 310                       | 271                           | 255                    |
| CRISPR repeats                   | 0                         | 2                             | 1                      |

Table 6 Properties and statistics for each genome
proteobacterial genomes, *Bacteroidetes* UKL 13-3 contains no autotrophy genes, consistent with the typical lifestyle of bacteria from this phylum as feeding on cellular detritus [43]. However, fewer genes were annotated from *Bacteroidetes* UKL13-3, and fewer completed KEGG pathway modules were identified than for the *Hyphomonadaceae* or Betaproteobacterium genomes (38 vs. 72 and 80, respectively). This could be due to protein-coding sequences carrying distant homology to those currently deposited in KEGG, limiting the ability to identify metabolic genes and pathways.

The *A. flos-aquae* genome was the only identified source of nitrogen fixing genes in the culture. Since the growth medium was nitrogen-deplete, all other bacteria in the community likely depend on reduced N provided by the cyanobacterium. It has been shown that *A. flos-aquae* from the Baltic Sea fixes N$_2$ and releases it as NH$_4^+$, which is then taken up by surrounding heterotrophic or phototrophic bacteria [44, 45]. Both proteobacterial genomes contain the ammonia transporter gene *amtB*, which would allow uptake of NH$_4^+$ released by *A. flos-aquae* (Fig. 6). No ammonia channel transport genes were annotated in the *Bacteroidetes* UKL13-3 genome. The proteobacterial genomes both contain chemotaxis and flagellar genes, and the Betaproteobacterium genome also contains type IV pilus genes for twitching motility (Additional file 8: Table S6; Fig. 6). Motility may be necessary for these organisms to stay associated with and obtain benefits from *A. flos-aquae*, similar to other host-associated bacteria [46].

We searched the novel genomes for the presence of other transporters to inform of the needs for survival and growth. Both proteobacterial genomes contain transporters for alkanesulfonate, iron(III), phosphate, and phosphonate. The *Hyphomonadaceae* genome also contains a transporter for putrescine, while the
Betaproteobacterium genome contains complete transporter modules for tungstate, molybdate, glutamate/aspartate, and branched-chain amino acids. Few, and only broadly functional, transporter modules were identified in the Bacteroidetes genome. All three genomes appear to carry complete genetic pathways for nucleotide biosynthesis, as well as genes for synthesis of all 20 amino acids, indicating these organisms are self-sufficient in this regard. Because the Flavobacterium and Brevundimonas genomes were so incomplete (Table 4), their gene content is not reported here.

We were unable to identify any plasmids in the assemblies. The distribution of all plasmids in GenBank shows that the majority are found in Proteobacteria (~47%), although most of these are associated with Gammaproteobacteria (~63%), rather than Alphaproteobacteria (~22%) or Betaproteobacteria (~8.7%) [47]. Plasmids from Bacteroidetes were much rarer at ~1.6%. It may then be unsurprising that these bacteria lack plasmids.

![Fig. 3 Circular map of the chromosome of Betaproteobacterium UKL13-2. See Fig. 2 legend for explanation](image)

**Betaproteobacterium UKL13-2**

**Freshwater bacteria associated with cyanobacterial blooms**

Bacteria from the Alphaproteobacteria, Betaproteobacteria, and Bacteroidetes are common in freshwater systems [43], are known to be commonly associated with cyanobacterial blooms, and can directly influence the growth of cyanobacteria in culture [48]. We therefore propose that the three newly sequenced genomes represent a bacterial community that is dependent on Aphaniizomenon flos-aquae (Fig. 6). Some Alphaproteobacteria have been identified in such cyanobacterial-associated communities [18]. For example, Alphaproteobacteria 16S sequences have been detected in association with the nitrogen-fixing cyanobacterium Gloeotrichia echinulata [49]. Interestingly, 16S rDNA from Hyphomonadaceae UKL13-1 shared significant identity (Additional file 1: Table S1) with one of these sequences (A0904), suggesting that bacteria related to Hyphomonadaceae UKL13-1 are associated with various bloom-forming cyanobacteria.
However, the extent to which such co-occurrences reflect physiological interdependencies remains to be explored. Betaproteobacteria have been seen physically associated with cyanobacteria [18, 49]. The predicted chemotaxis and flagellar and twitching motility genes (Additional file 8: Table S6) would assist both Hyphomonadaceae UKL13-1 and Betaproteobacterium UKL13-2 to remain associated with *A. flos-aquae* colonies and obtain the benefits of ammonium and organic nutrients released by the cyanobacterium (Fig. 6). We have detected no genes by which these photoheterotrophic bacteria could obviously benefit *A. flos-aquae*.

Bacteria from the *Bacteroidetes* phylum are commonly identified in, and sometimes dominate, freshwater lake systems [50]. They are also frequently found in particle-associated communities and commonly degrade extracellular polysaccharide matrices that are grazed via gliding motility [51]. *Bacteroidetes* UKL13-3 possesses annotated gliding motility genes, which may indicate physical association with the originally isolated *A. flos-aquae* colony. Extracellular mucilage, as well as a range of nutrients (reduced C, N and S compounds) released by *A. flos-aquae*, may support the growth of *Bacteroidetes* UKL13-3, whose genome seems to lack many functionally annotated pathways. *Bacteroidetes* UKL13-3 has the only annotated extracellular peroxidase gene in the three genomes, which could protect against reactive oxygen species generated by photosynthesis in *A. flos-aquae* (Fig. 6), which itself lacks annotated peroxidase genes. This may indicate a mutual benefit for both bacteria, and conform to the Black Queen Hypothesis described for interactions between the unicellular cyanobacterium *Prochlorococcus* with other interacting bacteria [52]. On the other hand, large populations of *Bacteroidetes* bacteria have been observed following cyanobacterial bloom decline [53] due to subsequently favorable
Table 7 Number and proportion of genes associated with COG functional categories

| Code | Hyphomonadaceae UKL13-1 | Betaproteobacterium UKL13-2 | Bacteroidetes UKL13-3 | COG category |
|------|-------------------------|-----------------------------|-----------------------|--------------|
|      | Value | % of total | Value | % of total | Value | % of total |            |
| J    | 184   | 4.91       | 187  | 5.07       | 175  | 5.34       | Translation |
| A    | 1     | 0.03       | 1    | 0.03       | 1    | 0.03       | RNA processing and modification |
| K    | 128   | 3.41       | 100  | 2.71       | 85   | 2.6        | Transcription |
| L    | 109   | 2.91       | 100  | 2.71       | 126  | 3.85       | Replication |
| B    | 2     | 0.05       | 2    | 0.05       | 1    | 0.03       | Chromatin structure and dynamics |
| D    | 25    | 0.67       | 46   | 1.25       | 28   | 0.85       | Cell cycle control |
| Y    | 0     | 0.00       | 0    | 0.00       | 0    | 0.00       | Nuclear structure |
| V    | 69    | 1.84       | 77   | 2.09       | 74   | 2.26       | Defense mechanisms |
| T    | 216   | 5.76       | 168  | 4.56       | 81   | 2.47       | Signal transduction mechanisms |
| M    | 165   | 4.4        | 181  | 4.91       | 242  | 7.39       | Cell wall /membrane/biogenesis |
| N    | 66    | 1.76       | 80   | 2.17       | 18   | 0.55       | Cell motility |
| Z    | 0     | 0.00       | 18   | 0.49       | 1    | 0.03       | Cytoskeleton |
| W    | 11    | 0.29       | 30   | 0.81       | 2    | 0.06       | Extracellular structures |
| U    | 49    | 1.31       | 58   | 1.57       | 34   | 1.04       | Intracellular trafficking |
| O    | 131   | 3.49       | 121  | 3.28       | 124  | 3.79       | Posttranslational modification |
| C    | 135   | 3.6        | 187  | 5.07       | 114  | 3.48       | Energy production and conversion |
| G    | 133   | 3.55       | 95   | 2.58       | 79   | 2.41       | Carbohydrate transport and metabolism |
| E    | 197   | 5.25       | 224  | 6.08       | 127  | 3.88       | Amino acid transport and metabolism |
| F    | 66    | 1.76       | 68   | 1.85       | 74   | 2.26       | Nucleotide transport and metabolism |
| H    | 137   | 3.65       | 134  | 3.64       | 94   | 2.87       | Coenzyme transport and metabolism |
| I    | 188   | 5.01       | 120  | 3.26       | 96   | 2.93       | Lipid transport and metabolism |
| P    | 153   | 4.08       | 143  | 3.88       | 85   | 2.6        | Inorganic ion transport and metabolism |
| Q    | 101   | 2.69       | 66   | 1.79       | 38   | 1.16       | Secondary metabolites biosynthesis |
| R    | 223   | 5.95       | 213  | 5.78       | 211  | 6.44       | General function prediction only |
| S    | 125   | 3.33       | 98   | 2.66       | 95   | 2.9        | Function unknown |
| NA   | 1104  | 29.44      | 1083 | 29.39      | 1201 | 36.67      | Not in COGs |

*The total is based on the total number of predicted protein coding genes in the annotated genomes

Conditions for copiotrophs [54] and cell turnover of *A. flos-aquae* and other cells may provide organic material for *Bacteroidetes* UKL13-3 growth in co-culture, with similar benefits from lysed cells for the two *Proteobacteria*.

Search for the novel genomes in freshwater metagenomes

We searched for the occurrence of the three novel bacteria in 62 freshwater lake metagenomes from 8 sampling sites across the United States, including Oregon, Washington state, California, Texas, and Kansas (BioProject accessions: PRJNA312985, PRJNA282166, PRJNA312830, PRJNA312986, and PRJNA294203, respectively). To do so, we mapped reads from these metagenomes to the references with BWA-MEM with default parameters (~0.067% mismatch rate) and calculated average genome coverage. Matches were found in two samples. A metagenome from Copco Reservoir, CA, on the Klamath River downstream of *UKL* on September 19, 2007 contained ~86x average read coverage depth of the *Hyphomonadaceae* UKL13-1 genome and ~151x coverage depth of the *Bacteroidetes* UKL13-3 genome from 398,356,734 Illumina read pairs. Additionally, a metagenome from Cranberry Lake, WA on August 11, 2014 contained the Betaproteobacterium UKL13-2 genome at ~99x average coverage depth from 13,955,857 Illumina read pairs. We also searched in 50 additional freshwater lake metagenomes in the IMG, MG-RAST, and SRA databases. The only detection found was the Betaproteobacterium UKL13-2 genome at ~19x average coverage depth in a metagenome consisting of 319,415,720 Illumina read pairs labeled “vibrio metagenome HEM-04” from a freshwater lake (BioProject accession: PRJNA64039). This initial analysis shows that the three novel bacteria are found elsewhere in freshwater habitats, although they do not appear to be ubiquitous or widely abundant.
Taxonomic placement and naming of genomes from uncharacterized bacteria

Currently, there is a lack of guidance and standardization for assigning taxonomic nomenclature to genomic sequences lacking phenotypic information. Until now, most of these sequences have been amplicons or draft assemblies from shotgun metagenomes, in which the genomes are usually incomplete and there are likely to be some contaminating contigs. We observed both of these defects in the draft genomes assembled from the short-read Illumina sequencing that paralleled the assembly of complete genomes from the PacBio long-read sequencing (Tables 4 and 5). Critical genes—including ribosomal RNA operons—were missing from the draft genomes (Table 8), making it perhaps premature to assign new taxonomic designations to such bacteria. Our work demonstrates that as long-read sequencing depth increases, so will the likelihood of assembling complete microbial genomes from uncultured samples of high relevance to extant microbial communities. Especially for circular chromosomes, the assembly of such complete genomes would seem to carry a low risk of artifactual assembly of the genome from a mixed-DNA sample in a metagenomic study, particularly when no PCR amplifications are used in the sequencing protocol, as in the present case. It would seem beneficial to the advancement of microbial ecology to develop new guidelines by which currently unclassifiable complete genomes derived from metagenomic data can be taxonomically placed with new genus and species names as appropriate.

Conclusions

Here, we have shown that completing multiple genome assemblies is possible from a simple microbial community using PacBio sequencing, a feat that is nearly
impossible with short-read shotgun sequencing alone. There are several advantages to this approach. Completing genome assemblies from a shotgun metagenome avoids genome gaps and excludes contaminant sequences, which are significant issues with binned draft genomes. Absent sequences can contain functionally relevant information, such as gene clusters encoding secondary metabolites [55] or antibiotic resistance genes near mobile elements [56]. Here we observed that key essential genes (Table 8) were missing from each short-read assembly. Also, short-read assemblers can compress small repeats, potentially removing important functional information [57]. In addition to providing more complete genomic information, long-read sequencing of communities such as mixed cultures or environmental samples creates possibilities for new experimental designs. For example, complete genomes from novel organisms sequenced from the environment can be used as new references for culture-free resequencing efforts, such as to explore gene linkage patterns among alleles in a population. Further, long-read sequencers often detect DNA modifications, such as methylation, allowing capture of epigenetic information from environmental sequencing runs.

Although PacBio sequencing is low-throughput compared with short-read sequencers, our results suggest that the current state of this technology allows genome sequencing from communities with relatively low diversity, such as those in extreme environments [58] or when dominated by one or a few organisms [59]. Platform improvement, such as the recently released PacBio Sequel instrument, is expected to make long-read sequencing increasingly desirable for shotgun metagenomics in the future.

Here, we have sequenced three novel genomes that may be associated with _Aphanizomenon flos-aquae_ as part of the cyanobacterial phycosphere (Fig. 6). Based on gene annotations and growth medium, both _Proteobacteria_ are motile aerobic anoxygenic phototrophs that utilize fixed nitrogen and carbon provided by _A. flos-aquae_. _Betaproteobacterium_ is a heterotroph that likely has similar nutritional requirements, and may exist in a mutual relationship with _A. flos-aquae_ through provision of an extracellular peroxidase. In future work, it will be
interesting to explore the possible existence and nature of dependencies between these novel bacteria and *A. flos-aquae* colonies in blooms in Upper Klamath Lake and elsewhere.

**Additional files**

**Additional file 1: Table S1.** Taxonomic placement of each novel genome by 16S similarity, composition (PhyloPythiaS+), and multiple marker gene similarities (Phylosift). (CSV 312 bytes)

**Additional file 2: Table S2.** Assembly parameters for genome assemblies from PacBio reads. Minimum read length cutoff is lowest read-length used for assembly, with remaining reads used for error correction. (CSV 195 bytes)

**Additional file 3: Figure S1a.** *Hyphomonadaceae* UKL13-1 genome repeats and Illumina Breaks. Blue lines signify intragenomic repeats (based on BLASTN with a minimum E-value cutoff of 1E-30), and red bars mark sequences missing from Illumina assemblies. b. *Betaproteobacterium* UKL13-2 genome repeats and Illumina Breaks. c. *Bacteroides* bacterium UKL13-3 genome repeats and Illumina Breaks. (ZIP 4776 kb)

**Additional file 4: Table S3.** Annotated genes in *Hyphomonadaceae* UKL13-1 Illumina breaks. Genes called and annotated with PROKKA. (XLXS 13 kb)

**Additional file 5: Table S4.** Annotated genes in *Betaproteobacterium* UKL13-2 Illumina breaks. Genes called and annotated with PROKKA. (XLXS 19 kb)

**Additional file 6: Table S5.** Annotated genes in *Bacteroides* bacterium UKL13-3 Illumina breaks. Genes called and annotated with PROKKA. (XLXS 16 kb)

**Additional file 7: Figure S2.** Percentage of protein-coding sequences from all bacterial genomes assigned to COG categories. Novel genomes are highlighted. (PDF 73 kb)

**Additional file 8: Table S6.** Percentage of protein-coding sequences from UKL13-2 genome repeats and Illumina breaks. Genes called and annotated with PROKKA. (XLSX 16 kb)

**Additional file 9: Table S7.** Percentage of protein-coding sequences from UKL13-3 genome repeats and Illumina breaks. Blue lines signify intragenomic repeats and Illumina breaks. (ZIP 4776 kb)

**Additional file 10: Table S8.** Percentage of protein-coding sequences from UKL13-1 genome repeats and Illumina breaks. Genes called and annotated with PROKKA. (CSV 312 bytes)

**Abbreviations**

AAP: Aerobic anoxygenic phototroph; ATP: Adenosine triphosphate; COG: Clusters of orthologous genes; DOM: Dissolved organic matter; IS: Insertion sequence; KEGG: Kyoto encyclopedia of genes and genomes; NGS: Next-generation sequencing; PROKKA: Prokaryotic genome annotation pipeline; Rubisco: Ribulose-1,5-bisphosphate Carboxylase; SMRT: Single molecule real-time; UKL: Upper Klamath Lake

**Acknowledgments**

We thank Sara Eldridge at the USGS for providing samples from UKL used for culture inoculum, Kim Haley for discussions and Mark Wildung and Derek Pouchnik from the WSU Molecular Biology and Genomics Core for DNA preparation and PacBio sequencing in addition to providing HGAPE assemblies. We also thank the NL Tartar Research Fellowship, the Oregon State University Agricultural Experiment Station, and the Mabel E. Pernot Trust for providing funding.

**Authors’ contributions**

Conceived the project: TWD, TGO, NMB, CBD. Provided strains: TGO. Annotated the genomes: CBD. Analyzed and interpreted results: CBD, TWD, NMB. Wrote the manuscript: CBD, TWD, TGO. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

Received: 28 April 2016 Accepted: 9 December 2016 Published online: 19 January 2017

**References**

1. Gilbert JA, Dupont CL. Microbial Metagenomics: Beyond the Genome. Annu Rev Mar Sci. 2011;3:347–71.
2. Escobar-Zepeda A, de Leon AV-P, Sanchez-Flores A. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Front Genet. 2015;6:348.
3. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402.
4. Gregor ID J, Schirmer M, Quince C, McHardy AC. PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. arXiv. 2014;1406:7123.
5. Sharon I, Morowitz MJ, Thomas, BC, Costello EK, Relman DA, Banfield JF. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res. 2013;23(1):110–20.
6. Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Colding SD, Tyson GW. Methane metabolism in the archaeal phylum Bathyrarchaeota revealed by genome-centric metagenomics. Science. 2015;350(6259):494–8.
7. Koren S, Phillips AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2013;26:110–20.
8. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31(6):533–8.
9. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. Gropht: an automated tool for the recovery of population genomes from related metagenomes. Peer J. 2014;2:e2603.
10. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. Peer J. 2015;3:e1165.
11. Cleary B, Brito IL, Huang K, Gevers D, Shea T, Young S, Alm EJ. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol. 2013;33:1023–30.
12. Hess M, Szyba A, Egan R, Kim T-W, Chokhawala H, Schroth G, Luo S, Clark DS, Chen F, Zhang T. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science. 2011;331(6016):463–7.
13. Lee H, Gurtowski J, Yoo S, Mascarin S, McCombie WR, Sztach M. Error correction and assembly complexity of single molecule sequencing reads. BioRxiv. 2014:doi: 10.1101/006995.
14. Koren S, Dohlen MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 2012;30(7):693–700.
15. Fichot EB, Norman RS. Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome. 2013;1(1).
16. Frank JA, Pan Y, Tooming-Klunderud A, Eijssen WG, McHardy AC, Nederbakt AJ, Pope PB. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci Rep. 2016;6:25373.
17. Bagatini IL, Eiler A, Bertilsson S, Kvenerv K, Vieira AAH. Host-specificity and dynamics in bacterial communities associated with bloom-forming freshwater phytoplankton. PLoS ONE. 2014;9:e85950.
18. Louati I, Pascoua N, Doebros D, Bernard C, Humbert JF, Leloup J. Structural diversity of bacterial communities associated with bloom-forming freshwater cyanobacteria differs according to the cyanobacterial genus. PLoS ONE. 2015;10(11):e0140614.
19. Quast C, Pruesse E, Yilmaz P, Gerken T, Schweer T, Yaraa P, Pellels J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(11590–6.
20. Pruesse E, Pellels J, Glockner FO. SNA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28(14):1823–9.
21. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):e940.
22. Darling AE, Jospin G, Lowe E, Matsen IV FA, Bk HM, Eisen JA. Phylosift: phylogenetic analysis of genomes and metagenomes. Peer J. 2014;2:e243.
23. Parker CT, Tindall BJ, Garrity GM. International code of nomenclature of microbes and virology. Int J Syst Evol Microbiol. 2015;doi: 10.1099/ijsem.0.0000778.
24. Culture Collection of Algae at the University of Texas at Austin. B3N media recipe. [https://utex.org/products/bold-in-medium]. Accessed 12 Jan 2017.
25. Sambrook J, Russell DW. Purification of nucleic acids by extraction with phenol: chloroform. Cold Spring Harbor Protocols 2006, 2006(1):doi: 10.1101/pdb.prot4455.
