environmental microbiology

Environmental Microbiology (2022) **24**(4), 2136–2156 doi:10.1111/1462-2920.15947

Disentangle genus microdiversity within a complex microbial community by using a multi-distance long-read binning method: example of *Candidatus* Accumulibacter

Aline Adler,[1] **Simon Poirier**,[1,3] **Marco Pagni**,[2] **Julien Maillard**[1] **and Christof Holliger**[1*]

[1]*Laboratory for Environmental Biotechnology, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.*

[2]*Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland.*

[3]*IFP Energie nouvelles, 1 et 4 avenue de Bois-Préau, 92852, Rueil-Malmaison Cedex, France.*

Summary

Complete genomes can be recovered from metagenomes by assembling and binning DNA sequences into metagenome assembled genomes (MAGs). Yet, the presence of microdiversity can hamper the assembly and binning processes, possibly yielding chimeric, highly fragmented and incomplete genomes. Here, the metagenomes of four samples of aerobic granular sludge bioreactors containing *Candidatus* (*Ca*.) Accumulibacter, a phosphate-accumulating organism of interest for wastewater treatment, were sequenced with both PacBio and Illumina. Different strategies of genome assembly and binning were investigated, including published protocols and a binning procedure adapted to the binning of long contigs (MuLoBiSC). Multiple criteria were considered to select the best strategy for *Ca*. Accumulibacter, whose multiple strains in every sample represent a challenging microdiversity. In this case, the best strategy relies on long-read only assembly and a custom binning procedure including MuLoBiSC in metaWRAP. Several high-quality *Ca*. Accumulibacter MAGs, including a novel species, were obtained independently from different samples. Comparative genomic analysis showed that MAGs retrieved in different samples harbour genomic rearrangements in addition to accumulation of point mutations. The microdiversity of *Ca*. Accumulibacter, likely driven by mobile genetic elements, causes major difficulties in recovering MAGs, but it is also a hallmark of the panmictic lifestyle of these bacteria.

Received 16 August, 2021; accepted 19 February, 2022. *For correspondence.E-mail christof.holliger@epfl.ch; Tel. (+41) 21 693 47 24; Fax (+41) 21 693 47 22.

Introduction

Microbes represent the functional backbone of all ecosystems. They have crucial roles in complex and highly dynamic communities as diverse as soils, oceans, human bodies and bioprocesses. To maintain and manage ecosystems services, microbial ecology research is facing a key emerging challenge: predicting how functions and structures of microbial communities are ruled within constantly fluctuating environments (Antwis *et al*., 2017). The study of microbiomes through whole-metagenome shotgun sequencing (WGS) enables the analysis of unexplored and uncultivated microbial populations by targeting the reconstruction of genomes from environmental shotgun DNA sequence data. Thus, WGS allows new perspectives for the characterization of their taxonomic composition and the prediction of their metabolic potential (Sangwan *et al*., 2016). Moreover, the knowledge acquired on bacterial genomes serves as a basis to investigate their *in situ* metabolic activity with various techniques: microautoradiography-fluorescent *in situ* hybridization (MAR-FISH, Okabe *et al*., 2004; Xia *et al*., 2008), Raman-FISH (Fernando *et al*., 2019), qPCR or metatranscriptomic and metaproteomic studies (Oyserman *et al*., 2016; Brotto *et al*., 2018).

Metagenomic studies can be conducted with two different sequencing approaches. Short-read WGS, mainly carried out with Illumina technology, produces reads that are inexpensive, accurate (error rate ~0.1%), but short (<500 bp). However, due to their short-read length, assemblies often report very fragmented contigs caused by repeated elements that exceed the length of sequenced reads. Long-read WGS overcomes this issue, resulting in *de novo* genome assemblies with greater contiguity. This approach is carried out with two

technologies from Pacific Biosciences (PacBio) (Eid *et al*., 2009) and Oxford Nanopore (Jain *et al*., 2015) that produce read lengths >20 kb. However, the high error rates of both technologies (up to ∼13%–15% per base) complicate the assembly, which require adapted strategies (Kono and Arakawa, 2019; Wickramarachchi *et al*., 2020). Long-reads are still rarely used alone for the assembly of microbial communities, but it proves to be a valuable method to extract MAGs from low-complexity metagenomes (Ahlgren *et al*., 2017; Driscoll *et al*., 2017; Moss *et al*., 2020; Somerville *et al*., 2019). New approaches for long-read correction (Rang *et al*., 2018) and synthetic (mock) community standards are emerging (Dilthey *et al*., 2019), allowing to detect biases associated with DNA extraction, sequencing and downstream analysis. Hybrid assembly combining short-read and long-read sequencing has become a common strategy improving contiguity of long-read assembly (Chen *et al*., 2020; Sanders *et al*., 2019), but miss-assemblies can occur between closely related genomes (Sczyrba *et al*., 2017; Sevim *et al*., 2019).

To allow recovering complete genomes of high quality from microbial community samples, ingenious bioinformatics methods are required (Schmid *et al*., 2018). Advance computational metagenomics have delivered many tools assembling billions of reads into contigs that are subsequently grouped into draft genomes by metagenome binning (Breitwieser *et al*., 2019; Chen *et al*., 2020). However, as evidenced by the inconsistencies of genomes that can be observed between WGS data analysis approaches, assemblers and binning tools development is still an actively improving field. Binning is a critical step to establish an accurate genome from a metagenome assembly. Binning tools cluster contigs based on compositional properties (e.g. GC content, tetranucleotide frequencies) or sequence coverage across multiple samples (Albertsen *et al*., 2013; Imelfort *et al*., 2014). Some of them additionally use marker genes from a reference database and scoring strategies (Parks *et al*., 2015; Simao *et al*., 2015). Yet, recent publications showed that no single binning approach such as ABAWACA (https://github.com/CK7/abawaca), CONCOCT (Alneberg *et al*., 2014), MaxBin (Wu *et al*., 2016) and MetaBAT (Kang *et al*., 2019) is superior for all samples/environment types or even for all populations within one sample. Therefore, bin consolidation tools such as Binning_refiner (Song and Thomas, 2017) and DAS Tool (Sieber *et al*., 2018) attempt to combine the strengths and minimize the weaknesses of different binning tools. A similar strategy has been used in a modular pipeline software called metaWRAP (Uritskiy *et al*., 2018) combining Binning_refiner and a subsequent refinement module. This module uses the completion and contamination metrics estimated with CheckM, which is based on collocated sets of nearly ubiquitous and single-copy genes (Parks *et al*., 2015).

Despite the success in extracting reconstructed genomes from a variety of environmental metagenomes, significant challenges still exist. Notably, microdiversity inherent to many bacterial species leads to major difficulties in the accurate resolution of metagenome-assembled genomes (MAGs), as analysis techniques are insufficient for distinguishing many closely related genomes (Sharon *et al*., 2013; Kashtan *et al*., 2014; Sczyrba *et al*., 2017; Sevim *et al*., 2019; Ayling *et al*., 2020). Microdiversity refers to the diversity of organisms that are closely related phylogenetically yet potentially displaying different metabolic activities and therefore occupying distinct niches. Genomic studies comparing multiple strains of the same species have revealed that while much of the genome sequence is highly conserved, functional variations can arise from introduction of genes by horizontal gene transfer, or changes in gene regulation due to mutations or genome rearrangements (Nelson *et al*., 2016; Sangwan *et al*., 2016).

Microdiversity is especially challenging within *Candidatus (Ca.)* Accumulibacter, one of the key functional group in the enhanced biological phosphorus removal (EBPR) process. This predominant polyphosphate accumulating organism (PAO) is as-yet-uncultivated. Moreover, clusters in *Ca*. Accumulibacter cannot be clearly defined by 16S rRNA genes since they share high identities (over 97%). Based on the phylogenetic distance of the polyphosphate kinase gene *pkk1*, two *Ca*. Accumulibacter clades I and II, with several subclades (IA-ID and IIA-II-I) were proposed (Peterson *et al*., 2008; Mao *et al*., 2015). Many studies underlined that they differed in certain abilities to accumulate phosphorus and reduce nitrate (Flowers *et al*., 2013; Gao *et al*., 2019; Rubio-Rincon *et al*., 2019), but it appears that different metabolisms can co-exist in a single clade (Rubio-Rincon *et al*., 2019). Despite numerous studies conducted on *Ca*. Accumulibacter, only one complete genome was published so far, *Ca*. Accumulibacter phosphatis str. UW-1 (Martin *et al*., 2014), hereafter referred to as *Ca*. Accumulibacter UW1. It has been assembled from a highly enriched sludge and is composed of a chromosome and three plasmids. In addition, several fragmented MAGs are currently available in RefSeq (O'Leary *et al*., 2016). They were constructed by using abundance-based binning methods on sequences obtained from the biomass of multiple reactors (Skennerton *et al*., 2015; Singleton *et al*., 2021) or multiple time points (Albertsen *et al*., 2016). Eight of them contain at least a 16S rRNA gene sequence; *Ca*. Accumulibacter sp. BA-93 (Skennerton *et al*., 2015), hereafter referred to as *Ca*. Accumulibacter BA93, *Ca*. Accumulibacter aalborgensis (Albertsen *et al*., 2016), *Ca*.

Accumulibacter sp. SSA1 (Arumugam *et al*., 2021) and five recently assembled MAGs (Singleton *et al*., 2021).

Combining the coverage information from multiple samples has proven to be efficient to recover MAGs and it has become a common practice (Albertsen *et al*., 2013; Skennerton *et al*., 2015; Sangwan *et al*., 2016). This approach is based on the hypothesis that bacterial populations are clonal. Yet, clonality is an uneven trait within the bacterial kingdom (Shapiro, 2016). Some bacteria have a rather panmictic lifestyle, mediated by intra-population genomic recombinations via transfer and integration of mobile genetic elements, such as plasmids (Wiedenbeck and Cohan, 2011; Rosen *et al*., 2015). This property was the motivation to assemble in the present study *Ca*. Accumulibacter MAGs from individual samples (Supporting Information Fig. S1). The assembly of long-reads and the use of an in-house binning strategy adapted to long contigs enabled to recover several high-quality *Ca*. Accumulibacter MAGs with a high contiguity and complete 16S rRNA gene sequences. The comparison of MAGs extracted in several copies from different samples revealed putative genomic recombination events and provided a first glimpse into the structural plasticity of *Ca*. Accumulibacter.

Results

Sampling and sequencing of metagenomic DNA

Total DNA was extracted from four biomass samples collected from the AGS sequencing batch reactor on days 71 (d71), 322 (d322), 427 (d427) and 740 (d740) and operated with synthetic influents with different composition of carbon substrates (Adler and Holliger, 2020). The monitoring of the bacterial community of the AGS by amplicon sequencing and the performance of the sludge were described by Adler and Holliger (2020). The composition of the bacterial communities corresponding to days 71, 322, 427 and 740 is presented in the Supporting Information Fig. S2. Changes in the microbial community composition were observed throughout the experiment, also within the *Ca*. Accumulibacter population, which already indicated a certain microdiversity of this genus (Adler, 2019; Adler and Holliger, 2020). In order to introduce additional abundance differences among the extracted metagenomic DNAs, two different extraction methods were used, extraction A, a gentle extraction preserving long DNA fragments, and extraction B, a standard method. The DNA was sequenced with a short-read (Illumina HiSeq; extractions A and B) and a long-read (PacBio Sequel; extraction A) sequencing technology. This produced 39.6 Gbp of paired-end short-read data and 26.5 Gbp of long-read data.

Assembly into contigs

Three different strategies were evaluated to produce contigs: short-read, hybrid, and long-read assembly. Statistics on contig size, average gene length and presence of complete 16S rRNA gene sequences were compared in order to evaluate the potential of each assembly type for the study of *Ca*. Accumulibacter (Supporting Information - Table S5). The number of long-read contigs per sample ranged from 548 (d322) to 1997 (d427), with an N50 ranging from 55 509 bp to 211 407 bp and a maximal contig length of 3.9 Mbp (Supporting Information - Table S5). The short-read assemblies produced much smaller contigs with a N50 ranging from 9056 bp (d71) to 19 743 bp (d740) and a maximal contig length of 2.17 Mbp. Compared with the short-read assemblies, the hybrid co-assemblies produced longer contigs with a N50 ranging from 15 754 bp (d427) to 21 228 bp (d71), which is 3.5–10 times shorter than the N50 of the long-read contigs. Several long-read contigs affiliated to *Saccharibacteria* (phylum), *Dechloromonas* (genus), *Dokdonella* (genus) or Sbr-gs28 (phylotype) were likely forming complete or nearly complete MAGs. The symmetric cumulative GC skew can be an indicator of the sequence accuracy of a MAG. The GC skew plots of these MAGs were compared with those of reference genomes (Supporting Information Figs. S3–S6). Not only the cumulative GC skew of the MAGs composed of a unique contig were close to zero, but the evolution of the GC skew was very similar to the one of the reference genomes. Average gene lengths were computed to check the quality of the assemblies, since assembly errors can introduce artificial stop codons, thus shortening gene lengths. The average gene lengths are in the same order of magnitude for all the assemblies of all four samples. It is the highest with the short-read assemblies (859–897 bp), followed by the long-read assemblies (798–863 bp) and the hybrid co-assemblies (691–786 bp).

Complete 16S rRNA gene sequences are often missing from metagenome assemblies. In the long-read assemblies (Supporting Information Table S5), the proportion of complete 16S rRNA genes is equal or close to 100%, whereas it is lower than 50% with the short-read and the hybrid assemblies. Moreover, the number of complete 16S rRNA genes per 10 Mbp of assembly is the highest in the long-read assemblies and the lowest in the short-read assemblies.

Eight distinct (with more than 2 bp differences) 16S rRNA genes that affiliated with *Ca*. Accumulibacter, were extracted from the long-read assemblies (Fig. 1). They were named as ACC001, ACC003, ACC004, ACC005, ACC007, ACC009, ACC010 and ACC012. Some of them were found in multiple samples and/or in multiple contigs

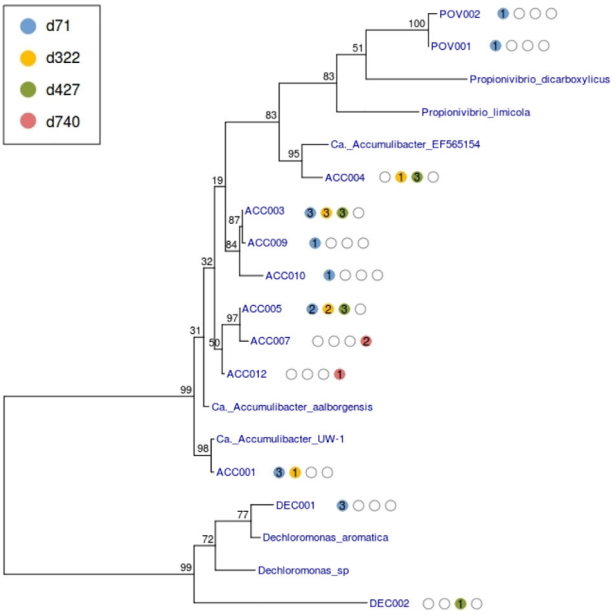


Fig. 1. Phylogenetic tree of 16S rRNA gene sequences related to *Ca*. Accumulibacter, *Propionivibrio* and *Dechloromonas* in chosen reference genomes and PacBio long-read contigs of aerobic granular sludge samples collected on day 71 (d71), 322 (d322), 427 (d427) and 740 (d740) in a lab-scale sequencing batch reactor fed with volatile fatty acids (VFA) for d71, VFA, glucose and amino acids for d322 and d427, VFA glucose, amino acids, starch and peptone for d740. The number of sequences per sample is indicated in the coloured dots (sequences with <3 bp of differences were considered as the same). The 16S rRNA gene sequence ACC005 and POV001 are identical to the 16S rRNA gene sequences in *Ca*. Accumulibacter sp. BA-93 and *Ca*. Propionivibrio aalborgenis, respectively.

of the same sample. ACC001 was found in sample d71 and d322. It is phylogenetically close to the 16S rRNA genes of *Ca*. Accumulibacter UW1 (Martin *et al*., 2014). ACC003 and ACC005 were found in the assemblies of the three first samples. ACC005 is identical to the 16S rRNA gene of the MAG *Ca*. Accumulibacter BA93 (Skennerton *et al*., 2015). ACC007 and ACC012 were found only in sample d740. ACC004, found in the assemblies of samples d322 and d427, was easier to assemble than the other *Ca*. Accumulibacter sequences included in this study. The best match for its 16S rRNA gene was a *Ca*. Accumulibacter with 99.0% of identity in MiDAS database vS123_2.1.3 (McIlroy *et al*., 2017). This affiliation was confirmed with online 16S rRNA databases MiDAS v4.8 (Dueholm *et al*., 2021) and EzBioCloud (Yoon *et al*., 2017), but GTDB-tk v.1.4.1 by (Chaumeil *et al*., 2020) places ACC004 within the genus *Propionivibrio*, with an ANI value of 78.94%, an AF value of 0.39 and a RED value of 0.94. Only three different rRNA gene sequences were extracted from the hybrid co-assemblies, ACC002, ACC006 and ACC008.

Since the apparent microdiversity within the genus *Ca*. Accumulibacter was only detected with the long-read assemblies, which also provides a higher contiguity, they were chosen for the extraction of the MAGs of this genus. Although the median coverage of long-reads mapping on the contigs related to *Ca*. Accumulibacter was generally above 30, and even up to around 280, these contigs were short compared with those related to other taxa (Supporting Information Figs. S7–S11). For example, in sample d71 (Supporting Information Fig. S7), a single contig affiliated to *Dechloromonas* is 3.5 Mbp long with a median coverage of 83. In the same sample, several contigs affiliated with *Ca*. Accumulibacter have a coverage above 200, and yet they are all shorter than 1 Mbp.

Multi-distance long-read binning with silhouette coefficient optimization

The long-read assembly was achieved in winter 2019, before the release of MetaBAT2 (Adler, 2019; Kang *et al*., 2019). The popular binning tools tested at the time were not successful at separating the different *Ca*. Accumulibacter MAGs. These tools, designed for short-read assemblies, have shown good results when combining the coverage information of numerous datasets (Alneberg *et al*., 2014). They were therefore not adapted to our experimental design consisting in grouping MAGs from long-read assemblies without combining the coverage information from other samples. With long-read assemblies, the scaffolding was not possible since all the non-ambiguous information provided by the long reads is used to build the contigs. Our in-house binning tool MuLoBiSC was therefore developed in order to group the contigs into bins by taking advantage of the length of long-read contigs. In addition to tetranucleotide frequency and mapping information of short and long reads from the same sample, it includes a metric based on the complementarity of benchmarking single copy orthologous genes (BUSCO; Seppey *et al*., 2019). The clustering of the contigs was optimized by maximizing the sum of silhouette coefficients weighted by the length of the contigs (WSC). It enabled to reduce MAG contamination, discard unreliable contigs, and recover more robust bins. This multi-distance binning strategy allowed to recover 19 high-quality MAGs (>90% completeness, <5% contamination) from the four samples, including taxa lacking a complete reference genome from the same genus in public databases (e.g. Sbr-gs28, *CPC_C22&F32*, *Cythophagaceae*) and five *Ca*. Accumulibacter MAGs, two of them present in three different samples.

Comparison of bins obtained with individual binning tools

The characteristics of all *Ca*. Accumulibacter bins resulting from the four binning tools applied in this study (MetaBAT2, MaxBin2, CONCOCT and MuLoBiSC) are detailed in the Supporting Information Table S6. Based on CheckM metrics, quality thresholds were applied to evidence only *Ca*. Accumulibacter bins with a completeness >80% and contamination <10% (Supporting Information Table S7). This filtering step specifically removed highly contaminated bins obtained with MaxBin2 and CONCOCT. Despite this filtering, a few abnormal bins, considered as barely contaminated by CheckM, were still present in this table. At day 71, MetaBAT2 gathered in the same bin contigs assigned to ACC003 and ACC010. In the same sample, MaxBin2 provided a chimeric bin with contigs assigned to *Ca*. Accumulibacter (ACC009), *Ca*. Competibacter (COM001) and *CPB S18* (CPS001). This binning tool also provided a bin (ACC003) with an aberrant total length of 6.8 Mbp in sample d427. Similarly, within sample d322, CONCOCT provided two contaminated *Ca*. Accumulibacter bins (ACC003 and ACC005) with abnormal genome size of 6.8 Mbp and 8.1 Mbp, respectively. Overall, MuLoBiSC and MetaBAT2 provided more reliable *Ca*. Accumulibacter bins (eight and six, respectively) compared to CONCOCT (two) and MaxBin2 (one).

Comparison of bins obtained with metaWRAP

In order to improve *Ca*. Accumulibacter bin quality, the metaWRAP pipeline was used to integrate results from the different binning tools. The combination of the three binning tools MetaBAT2, MaxBin2 and CONCOCT, that we call here metaWRAP_BXC and that has originally been implemented in the metaWRAP framework, was taken as reference and compared with the combination consisting of MetaBAT2, MaxBin2 and MuLoBiSC called metaWRAP_BXM. The bins characteristics are presented in Table 1. Other combinations were also tested and the results are presented in the Supporting Information - Table S9. MetaWRAP_BXM provided eight good quality *Ca*. Accumulibacter bins, whereas metaWRAP_BXC only provided five good quality bins and two contaminated ones. Within sample d71, metaWRAP_BXC kept the MetaBAT2 bin gathering ACC003 and ACC010, and within sample d427 it kept the MaxBin2 bin with an abnormal genome size of 6.7 Mbp. For the combination of MetaBAT2 and MuLoBiSC, called metaWRAP_BM, the results were identical to metaWRAP_BXM for seven out of eight bins. Only the bin ACC012 from day 740 comprised two extra contigs. From tests with the integrative binner DAS Tool, we concluded that it was not adapted to investigate microdiversity of *Ca*. Accumulibacter from our initial contigs dataset. Overall, the results showed that the MAGs provided by metaWRAP and the combination metaWRAP_BXM had a higher quality than those

Table 1. Characteristics of *Ca*. Accumulibacter related MAGs obtained with metaWRAP and the default combination of binning tools, MetaBAT2, MaxBin2 and CONCOCT (BXC) and the combination MetaBAT2, MaxBin2, MuLoBiSC (BXM).

Sample[a]	Binning tool	Contigs containing at least one *Ca*. Accumulibacter 16S rRNA gene[b]			Number of contigs	Bin length (Mbp)	Completeness (%)	Contamination[c] (%)	WSC[d]
d71	BXC	ACC003a	ACC003b	ACC010	57	4.8	87.7	1.8	2.5
	BXM	ACC003a	ACC003b		51	4.6	86.4	1.3	2.6
		ACC005a	ACC005b		14	4.3	84.1	1.4	3.1
d322	BXC	ACC003a	ACC003b		13	5.4	98.6	0.4	4.0
	BXM	ACC003b			13	5.4	98.6	0.4	2.8
		ACC003a	ACC003b	ACC003c	45	6.7	98.2	2.7	2.0
	BXC	ACC004a			4	5.2	98.0	1.4	3.8
d427		ACC005a	ACC005b	ACC005c	35	5.2	98.5	3.9	2.6
		ACC003a	ACC003b	ACC003c	17	5.4	98.1	2.7	4.4
	BXM	ACC004a			5	5.2	98.6	1.5	3.6
		ACC005a	ACC005b	ACC005c	19	4.7	97.9	2.2	4.2
	BXC	ACC007a			15	4.9	95.7	5.4	2.7
d740		ACC007b	ACC012		27	5.6	98.3	5.6	2.2
	BXM	ACC007a			13	4.8	94.8[e]	4.4[e]	3.0
		ACC007b	ACC012		27	5.6	98.3[f]	5.6[f]	2.2

Only the MAGs with a completeness >80% and a contamination <10% are shown here. More details, including the lower quality bins are in the Supporting Information Table S8.

[a]Metagenomic samples taken at four different days of reactor operation (d71, d322, d427, d740).

[b]The *Ca*. Accumulibacter 16S rRNA genes with different numbers have a sequence difference of at least three nucleotides. The letters indicate different contigs containing the same *Ca*. Accumulibacter 16S rRNA gene.

[c]Completeness and contamination percentages were determined with CheckM.

[d]WSC = weighted silhouette coefficient (expressed in millions).

[e]The completeness and contamination after the correction of the chimeral contig are 98.1% and 4.4%, respectively.

[f]The completeness and contamination after the correction of the chimeral contig are 95.9% and 4.6%, respectively.

obtained with the combination metaWRAP_BXC or with individual binning tools. Hence, these high-quality bins were chosen for the subsequent analyses, unless specified otherwise.

pkk1 *sequences and the consistency of MAGs in the different samples*

Polyphosphate kinase 1 *(pkk1*) gene sequences are traditionally used to classifiy *Ca*. Accumulibacter into clades. In order to assess the consistency between the 16S rRNA gene and *pkk1* classification in our MAGs, the *pkk1* sequences were extracted from the *Ca*. Accumulibacter, *Propionivibrio* and *Dechloromonas* MAGs and corresponding reference genomes. A phylogenetic tree of these sequences together with *pkk1* sequences of the existing *Ca*. Accumulibacter clades was built with sequences from *Dechloromonas* MAGs and reference genomes to root the tree (Fig. 2). This tree is very consistent with the 16S rRNA gene tree. The three ACC005 MAGs containing 16S rRNA gene sequences identical to *Ca*. Accumulibacter BA93 (clade IA) have a *pkk1* sequence similar to *Ca*. Accumulibacter BA93,

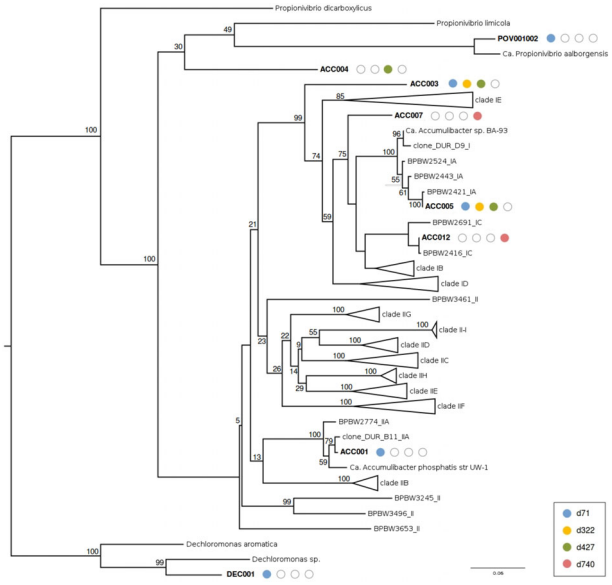


Fig. 2. Phylogenetic tree built from a selection of polyphosphate kinase 1 (*pkk1*) from different *Ca*. Accumulibacter clades (He *et al*., 2007; Peterson *et al*., 2008; Mao *et al*., 2015) listed in the Supporting Information Table S3 and *ppk1* gene sequences related to *Ca*. Accumulibacter, *Propionivibrio* and *Dechloromonas* in chosen references and PacBio long-read contigs of aerobic granular sludge samples collected on day 71 (d71), 322 (d322), 427 (d427) and 740 (d740) in a lab-scale sequencing batch reactor fed with volatile fatty acids (VFA) for d71, VFA, glucose and amino acids for d322 and d427, VFA glucose, amino acids, starch and peptone for d740. The presence of sequences in the samples is indicated in the coloured dots. The *ppk1* sequence indicated as ACC001 is located in a contig belonging to ACC001 bin in MetaBAT2 and in bin.4 in metaWRAP_BXM.

placing the ACC005 MAG in clade IA. These three MAGs have different values of completeness and contamination, yet, they are highly similar to one another (Supporting Information Fig. S12), with an ANI from 98.6% to 99.6%. The percentage of aligned bases between two ACC005 MAGs goes from 68.8% between sample d71 and d322 to 89.8% between sample d322 and d427, respectively (Supporting Information Table S10).

The three ACC003 MAGs also share a unique *pkk1* sequence. This sequence places ACC003 in type I but no specific clade could be assigned since no other sequence with an identity percentage higher than 86% was found in the NCBI database for this *pkk1*. Again, these three MAGs have different completeness and contamination values, but they have an ANI above 98.5% (Supporting Information Table S11 and Fig. S13).

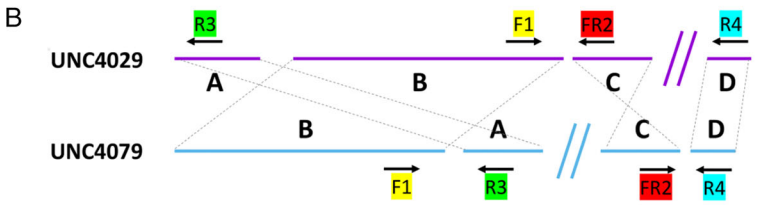
The *pkk1* sequence of MAG ACC004 is found on the same contig as the two ACC004 16S rRNA gene sequences; therefore, it provides information about the consistence of the assembly but not about the quality of the binning. As in the 16S rRNA gene tree, the ACC004 *pkk1* is placed apart from the other *Ca*. Accumulibacter *pkk1* genes from other MAGs and references. For this *pkk1*, the best match on NCBI database is with *Ca*. Accumulibacter UW1 but with an identity of 80.6% on 95% of the sequence. The ANI between ACC004 and the other *Ca*. Accumulibacter MAGs and chosen references, assessed by the percentage of aligned bases, is between 77.9% and 78.6% (Supporting Information Table S12). The similarity of this MAG with the (*Ca*.) Propionivibrio MAGs and references is slightly higher, with ANIs ranging from 78.8% to 79.2%.

Genomic rearrangements in Ca*. Accumulibacter MAG*

MUMmer plots were used to compare similar *Ca*. Accumulibacter MAGs assembled from multiple samples. Potential genomic rearrangements are visible on most of these plots (Supporting Information Figs. S14 and S15) even between very similar MAGs (Supporting Information Figs. S12 and S13). Predicted recombinase (e.g. *recR*, *xerCD*) and transposase genes (e.g. *tnpA*) *were* found in many of the rearrangement ends. To investigate the impact of recombinases and transposases in *Ca*. Accumulibacter compared to other genera for which MAGs were easier to assemble, the frequency of these genes was computed in the MAGs of the four samples (Supporting Information Table S13). Recombinase and transposase genes are present in relatively high frequency in the MAGs affiliated to *Ca*. Accumulibacter. Indeed, the five MAGs with the highest frequency of transposase/recombinase were all affiliated to *Ca*. Accumulibacter, and all *Ca*. Accumulibacter MAGs were within the top 30% when ordered by transposase/

A

Fragment	UNC4029 coordinates	UNC4079 coordinates	Primer
A	123495 → 125718	130925 → 133148	R3
B	130822 → 138281	121486 → 128936	F1
C	139037 → 142976	286449 ← 282506	FR2
D	294725 → 294853	287534 →287663	R4



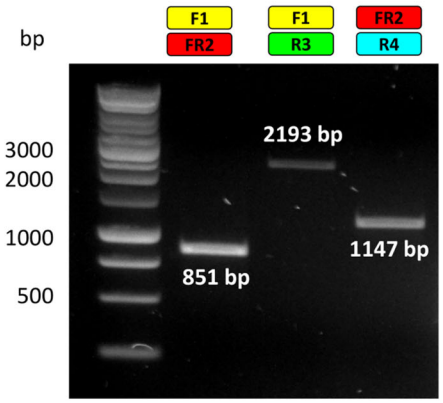


Fig. 3. Validation of a selected genomic rearrangement between contigs UNC4029 and UNC4079. A PCR amplification strategy was designed from the alignment of both contigs across a large genomic rearrangement (Supporting Information Fig. S15).
A. Matching fragments of the two contigs indicating their coordinates and the designed primers.
B. PCR strategy and primer positions on the conserved fragments from both contigs across the selected putative rearrangement. Please note that primer arrows are not depicted on the same scale as the contig fragments.
C. Gel electrophoresis after PCR amplification of targeted fragments across the predicted genomic rearrangement between contigs UNC4029 and UNC4079. Unique PCR products amplified with the above-mentioned primer combinations were obtained with the expected sizes, thus confirming the predicted genomic rearrangement.

recombinase frequency. When comparing similar MAGs, the ones with a higher completeness are generally above the ones with a lower completeness. For example, ACC003 of sample d427 (98% complete) has six recombinases and 42 transposases per 1000 CDS, whereas it has only two recombinases and 33 transposases per 1000 CDS in its less complete version of sample d71 (86% complete). Similar observation can be made for the ACC005 MAGs. This suggests that the recombinase/transposase-rich regions of these genomes were the most difficult to bin.

As an example to validate the results of genome assembly, primer design and PCR reactions were performed to confirm a possible major genomic rearrangement observed between contigs UNC4029 and UNC4079 (Supporting Information Fig. S15, Fig. 3A and B). All three targeted DNA fragments could be amplified from sample d740 and displayed the expected size (Fig. 3C). The three PCR products were then cloned and sequenced, and the obtained sequences matched exactly the corresponding regions in contigs UNC4029 and UNC4079 (data not shown).

Ca*. Accumulibacter MAGs in metagenomes of individual granules*

By analysing the sequences of metagenomes of 142 individual granules collected in the same AGS reactor few months before the start of our sampling, Leventhal *et al*. (2018) revealed that the AGS biomass was partitioned among individual granules into different community types determined by a dominant genotype of *Ca*. Accumulibacter. The reference genomes available at the time of that study had been used for this analysis. The short-reads of the 142 granule metagenomes were remapped on the *Ca*. Accumulibacter MAGs obtained here (Supporting Information Fig. S16). The *Ca*. Accumulibacter populations that were related to *Ca*. Accumulibacter UW1 in the previous study, had the highest number of matches with the MAG ACC001a provided by MetaBAT2. The patterns of mapping on ACC001a and ACC001bc are similar but not totally correlated. This indicates that these two bins contain the sequences of at least two sub-populations. The MAG ACC005 was the dominant *Ca*. Accumulibacter in another set of granules. Four granules were dominated by ACC007-related populations and four others by *Ca*. Accumulibacter sp. BA-91. However, these two MAGs were only detected in abundance in the granules collected during the first sampling.

Identification of plasmids

The most complete published genome of *Ca*. Accumulibacter contains a chromosome of 5.1 Mbp and three plasmids of 167.6, 42.3 and 37.7 kbp. Therefore, the presence of plasmids in the assembly, and more particularly within *Ca*. Accumulibacter MAGs, was investigated. A total of 49 complete or partial plasmids, ranging from 16 kbp to 147 kbp, identified by the presence of a plasmid replication initiator gene, *trfA*, and a BUSCO count lower than four, were assembled (Supporting Information Table S14). All these contigs are classified in the phylum of Proteobacteria, 39 of them to *Ca*.

Accumulibacter. One quarter of these contigs is indicated as circular, the others are either incomplete or having regions of high similarity with another contig in the assembly, preventing the circularization. Some contigs were found in multiple assemblies [UNC0157 (d71), UNC1481 (d322), UNC2046 (d427)] or [UNC1469 (d322), UNC2058 (d427)]. Since only four of these contigs were binned, an attempt was made to assign the others to a particular MAG by inspecting the coverage and the tetranucleotide frequency. Unfortunately, the coverage information were not corresponding to any of the *Ca*. Accumulibacter MAGs. The examination of the tetranucleotide signature was not very conclusive either since the contigs with the most similar tetranucleotide frequencies (Supporting Information Additional files 2, 3, 4 and 5) were identified as other putative plasmids (presence of *trfA* or *traM* genes, contiguity with plasmid in the assembly graph, low BUSCO content).

New Ca*. Accumulibacter MAGs*

None of the binning strategy succeeded to form a good quality bin containing the 16S rRNA sequence ACC001. MetaBAT2 produced two bins containing one and two ACC001 rRNA genes (Supporting Information Table S6) and having 96.9% and 48.3% of completeness and 20.5% and 36.2% of contamination, respectively. MetaWRAP_BXM produced a bin (bin.4) with contigs similar to *Ca*. Accumulibacter UW1 with a completeness of 66.9% and a contamination of 0.0% but containing no 16S rRNA gene sequence. The analysis of the assembly graphs (Supporting Information Figs. S17–S19) revealed that some of these contigs with a coverage higher than 200 are linked to multiple other contigs with a lower coverage containing similar gene contents. This suggested that multiple closely related strains of *Ca*. Accumulibacter clade IIA were co-existing in the same sample and that their genomes share parts that were similar enough to be assembled (single contig with higher coverage) and others that were not (multiple contigs 'in parallel' with lower coverage). Such contigs are very difficult to bin automatically.

To assess the phylogeny of the contigs, the blast search option was used in Bandage (Wick *et al*., 2015) with the reference genomes of *Ca*. Accumulibacter UW1 and BA93. The contigs with an identity above 90% with *Ca*. Accumulibacter UW1 or BA93, on more than half of their length were considered as *Ca*. Accumulibacter type I or type II, respectively. This allowed to identify several confusions between the MAGs ACC001 (type II) and ACC005 (type I). In sample d71, nine contigs of type I were placed in a type II bin. With metaWRAP_BXM, only one of these potentially misplaced contigs remained in the bin. In sample d322, a total of 72, 22, and 21 contigs of type II were placed in a type I bin by MetabAT2, MuLoBiSC, and metaWRAP_BXM, respectively (Supporting Information Fig. S12A,C). This might be the reason for the lower similarity between ACC005 from sample d322 and ACC005 from the two other samples (Supporting Information Table S10). In the sample d322, the coverage of the type I (ACC005) and type II contigs was very similar (Supporting Information Fig. S7). Therefore, the distinction between the MAGs based on the coverage was not working. This phenomenon was not observed in sample d427 where the type II contigs had a much lower coverage than type I contigs.

The visualization of the blast results of *Ca*. Accumulibacter BA93 on the contigs of sample d740 allowed detecting a potentially chimeric contig (Supporting Information Fig. S20). The contigs belonging to the ACC007 MAG have a high similarity (>90%) with *Ca*. Accumulibacter BA93 on most of their length, whereas the contigs of the ACC012 MAG have this similarity only on small parts. The contig containing the second ACC007 16S rRNA gene of the assembly belongs to the ACC012 MAG. This contig has a high similarity with *Ca*. Accumulibacter BA93 on its first third (ending with the rRNA genes operon), but not on the other two thirds. Moreover, a change in the long-read coverage on this contig can be observed on the first third of the contig (Supporting Information Fig. S21). The most likely hypothesis to explain this observation is that our assembly strategy produced a chimera between two *Ca*. Accumulibacter populations and that the first third of the contig belongs to the ACC007 MAG and the other two thirds belong to the ACC012 MAG.

The manually corrected MAGs ACC003 from sample d322 (ERS6225847), ACC004 (ERS6225849) and ACC005 (ERS6225848) from sample d427, ACC007 (ERS6225850) and ACC012 (ERS6225851) from sample d740 and DEC001 (ERS6225852) from sample d71 are deposited in the ENA under the project accession number PRJEB38840 with the PacBio and Illumina runs. All the submitted MAGs meet the standards of high-quality draft MAGs (Bowers *et al*., 2017).

Discussion

Ca. Accumulibacter is a very diversified bacterial genus for which few good quality reference genomes are available (Martin *et al*., 2014; Skennerton *et al*., 2015; Albertsen *et al*., 2016; Bowers *et al*., 2017; Arumugam *et al*., 2021; Sayers *et al*., 2021; Singleton *et al*., 2021). In the present study, an original approach using long-read assembly and an adapted in-house binning protocol was used to recover five *Ca*. Accumulibacter-related MAGs from four AGS samples collected in a lab-scale reactor operated with different influents over a period of

several months. They meet the minimum information about MAG (MIMAG) requirements for high-quality draft metagenomes, according to Bowers *et al*. (2017).

The advantage of long-read only assembly

Long-read only assemblies of metagenomes are still rare (Ahlgren *et al*., 2017; Somerville *et al*., 2019; Moss *et al*., 2020; Singleton *et al*., 2021). Long-read sequences are commonly combined with short-read data in hybrid co-assemblies (Daims *et al*., 2015; Slaby *et al*., 2017; Bertrand *et al*., 2019) increasing the contiguity of the short-read contigs but also introducing misassemblies between similar sequences (Sczyrba *et al*., 2017; Sevim *et al*., 2019). Here, the PacBio long-read assembly confirmed to display interesting advantages compared to the corresponding hybrid or short-read assemblies such as long size and a high recovery rate of complete 16S rRNA genes (Driscoll *et al*., 2017; Moss *et al*., 2020). The average gene length of the long-read assemblies is slightly lower than in the short-read assemblies but slightly higher than in the hybrid assemblies. This suggests that the long-read assemblies contain slightly more sequencing errors than the short-reads assemblies, but also less sequencing errors than the hybrid assemblies. The higher error rates in the hybrid assemblies very likely comes from the way the metaspades assembler (Nurk *et al*., 2013) works in hybrid mode. It uses the long-reads to scaffold short-read contigs and sequencing error will likely be frequent on these junctions because they are not polished. In the long-read assemblies, we deduced from our systematic examination of the contigs quality after each round of polishing with Arrow© (2011–2018, Pacific Bioscience of California), that the remaining sequencing errors where mainly located in low coverage contigs (below 10), which is not the case for the contigs in *Ca*. Accumulibacter bins. The rRNA genes, which contain highly conserved sequences between related taxa, are known to be problematic for short-read assemblies of metagenomes (Driscoll *et al*., 2017; Hugoson *et al*., 2020). These genes are often missing in the *Ca*. Accumulibacter MAGs assembled from short-read sequences (Sayers *et al*., 2021). All the high-quality *Ca*. Accumulibacter MAGs assembled in this study contain at least one 16S rRNA gene.

Binning strategy

Binning tools were optimized for short-read or hybrid assemblies. Our in-house binning tool MuLoBiSC was developed in order to take advantage of the big length of long-read contigs. It was specifically designed to our data consisting for each sample of one dataset of long-reads and two of short-reads and would need to be further adapted in order to be used with other data configurations. The particularity of MuLoBiSC is that it initiates the binning with the longest contigs, giving more weight to the information they contain and use single-copy genes (BUSCO) complementarity to reconstruct the metagenomic puzzle. To objectively assess the classification of each contig within each bin, we used Silhouette coefficient (SC). SC has been successfully used after clustering as a cluster validity measure for gene expression data (Gat-Viks *et al*., 2003) or SNP genotypes (Lovmar *et al*., 2005). Here, it helped to refine the binning of the contigs by increasing the consistency of the global clustering. It also provides a metric on how 'well' a contig is classified within a bin.

If the performance of MuLoBiSC with our dataset was similar to the one of MetaBAT2, the combination of the two binning tools and MaxBin2 with metaWRAP (metaWRAP_BXM) resulted in MAGs of higher quality (good estimated completeness and contamination, plausible genome size) than the default combination proposed in metaWRAP (MetaBAT2, MaxBin2 and CONCOCT). metaWRAP_BXM provided a good resolution of the different *Ca*. Accumulibacter MAGs, and the 16S rRNA genes and *pkk1* information was consistent. Even so, a careful examination of the MAGs obtained with metaWRAP_BXM detected remaining contamination. It confirms that the evaluation of contamination and completeness with single-copy orthologs, although essential and useful, has its limitations (Nelson *et al*., 2020), and that a human supervised curation is often necessary to complete the evaluation of MAGs.

Taxonomic classification of ACC004

The 16S rRNA sequence of ACC004 was affiliated with a *Ca*. Accumulibacter with 99.0% of identity in MiDAS and EzBiocloud (Yoon *et al*., 2017; Dueholm *et al*., 2021). The identity threshold for 16S rRNA gene identity between the members of a genus proposed by Yarza *et al*. (2014) is here fully respected. The 16S rRNA gene sequence generally provides valuable information for taxonomic classification, in particular for genus and higher taxonomic levels (Zhi *et al*., 2012; Thompson *et al*., 2013; Konstantinidis *et al*., 2017). Nevertheless, the placement of ACC004 in the 16S rRNA gene tree or the *pkk1* tree along with the ANI value being higher with *Propionivibrio* than with *Ca*. Accumulibacter genomes, motivated further analysis.

Commonly accepted consensus are now used to define the species of bacteria based on its genome (Richter and Rossello-Mora, 2009; Thompson *et al*., 2013; Qin *et al*., 2014; Whitman, 2015; Konstantinidis *et al*., 2017; Barco *et al*., 2020; Parks *et al*., 2020), but although some methods have been

proposed for higher classification levels, such as assessing the percentage of conserved proteins (Qin *et al*., 2014), or detecting the inflexion points in ANI vs alignment fraction (AF) graphs (Barco *et al*., 2020), none is actually accepted as a standard. The question of the existence of a natural clustering of genomes at the different taxonomic level is not easily resolved (Qin *et al*., 2014; Barco *et al*., 2020; Parks *et al*., 2020), in particular in taxonomic group containing few representatives. If the classification of new MAGs is straightforward when a closely related genome with a defined taxonomy is available, it becomes difficult otherwise. This is illustrated here by the case of ACC004 which appears to be distantly related to *Ca*. Accumulibacter and *Propionivibrio* based on ANI, *pkk1* and 16S rRNA gene sequence similarities.

GTDB-tk places ACC004 within the genus *Propionivibrio*, with an ANI value of 78.94%. The genus *Propionivibrio* gather bacteria with dissimilar metabolisms and the minimum ANI between its members is below 78%, which was proposed as an 'observed' lower bound for members of a genus (Parks *et al*., 2020). It is therefore possible that the discovery of new MAGs inside the existing *Propionivibrio* genus along with the study of their metabolism will provide a new landmark for the classification of ACC004 (Barco *et al*., 2020; Parks *et al*., 2020).

The elusive microdiversity of Ca*. Accumulibacter*

From the four *Ca*. Accumulibacter reconstructed from our metagenomes, ACC005 is very similar to *Ca*. Accumulibacter BA93, with an identical 16S rRNA gene sequence, similar *pkk1* sequences (97.9%) and a global ANI value of 98.9%. ACC007, found in the sample collected when the reactor was treating complex polymeric wastewater, is classified in clade IC. Based on the comparison of *pkk1* and 16S rRNA gene, ACC012 can be classified in type I. The *pkk1* sequence of ACC003 also places this MAG outside of the actual clade classification. Comparing ACC003 *ppk1* sequence to the NCBI nucleotide database, the highest identity (93.4%) is obtained for *Ca*. Accumulibacter phosphatis clone PPK8H07, with a query cover of 47%. This ACC003 represents therefore a valuable new reference in the genus *Ca*. Accumulibacter. Its robustness is confirmed through the repeated independent reconstruction from three different samples. The 16S rRNA gene and *pkk1* sequences are exactly the same in the three ACC003 MAGs, and the ANI value between ACC003 from sample d322 and d427 is 99.9%. Yet, the comparison of these two MAGs suggests that genomic rearrangements may have occurred in ACC003 between the two time points. A second possible explanation for these rearrangements is that the two conformations were present in the *Ca*. Accumulibacter population at both time points, but one of them was dominant on day 322 and the other on day 427.

Compared to other genera, *Ca*. Accumulibacter was very difficult to assemble and bin, despite a sufficient sequence coverage. These difficulties are well known for closely related strains for all types of assembly (Sczyrba *et al*., 2017; Ayling *et al*., 2020). In sample d71, numerous contigs related to *Ca*. Accumulibacter UW1 were assembled, among them, a *pkk1* sequence and three 16S rRNA genes (ACC001), but no good quality MAG of type II was recovered. An important microdiversity that is not homogeneously distributed across the genome was identified as the main cause of this failure. This *Ca*. Accumulibacter type II population was already abundant in the AGS reactor nine months before the start of the experiment (Leventhal *et al*., 2018). Moreover, aerobic granular sludge samples increase the microdiversity of the overall biomass because each granule has its own microbial community with its particular dominant populations (Leventhal *et al*., 2018). Lowering the diversity of the microbial community is a way to obtain high-quality MAGs from long-read sequences of metagenomes (Yu *et al*., 2017a). In case of aerobic granular sludge, long-read sequencing of individual granules could be a solution. However, the mapping of individual sequences from Leventhal *et al*. (2018) did not bring any strong conclusion on the intra-granule diversity of *Ca*. Accumulibacter type II since both the MAGs ACC001a and ACC001bc used for the analysis likely contain a mix of different strains. Our hypothesis is that intra-population genetic material transfers and recombinations created a diversity that is more complicated for the assembler to deal with than the accumulation of point mutations. Indeed, a comparison of the different MAGs obtained in this study suggests frequent recombination events in the *Ca*. Accumulibacter genomes that would not have been captured with short read or hybrid assemblies. In one of them, the distance between the matching sequences was small enough to perform a PCR on the DNA of the original sample (d740) showing that the rearrangement was not an artefact of the assembly.

The role of mobile genetic elements and plasmids in the diversity of Ca*. Accumulibacter*

The high density of transposases and recombinases among the coding sequences of *Ca*. Accumulibacter can be part of the reasons for the frequent structural variations observed in these genomes (Thomson *et al*., 2004; Konstantinidis and Tiedje, 2005). Transposases and recombinases can be part of mobile genetic elements, which may contribute to genomic rearrangements in bacterial cells. The comparison of *Bordetella pertussis* genomes has revealed multiple structural variations,

facilitated by mobile genetic elements, between genomes with low ANI (Weigand *et al*., 2017). However, since mobile genetic elements generally do not contain conserved single-copy genes, their absence remains unnoticed by the traditional completeness estimation software (Parks *et al*., 2015; Simao *et al*., 2015; Nelson *et al*., 2020) and their frequency can be largely underestimated in short-read assembled MAGs (Driscoll *et al*., 2017; Maguire *et al*., 2020; Nelson *et al*., 2020). Even though, they can carry genes of importance such as antimicrobial resistance genes (Konstantinidis and Tiedje, 2005; Zowawi *et al*., 2015). For example, plasmids can play an important role in strain-level adaptation to a specific niche or virulence factor (Arredondo-Alonso *et al*., 2020; Koch *et al*., 2020; Liao *et al*., 2020), but they are often not included in MAGs. It has been suggested that plasmids contain a common pool of genes that contribute to the adaptation of the whole bacterial community to specific selective conditions (Sentchilo *et al*., 2013). Here, we identified numerous type F plasmids from the long-read assembly, most of them affiliated to *Ca*. Accumulibacter by CAT (von Meijenfeldt *et al*., 2019). Only four of them were included in bins. The difficulty of assigning plasmid-containing contigs to the corresponding genomes has also been observed on hybrid assemblies of metagenomic mock communities (Nelson *et al*., 2020). Here, the attempt to assign the unbinned plasmids to a MAG based on the tetranucleotide signature or coverage was unfruitful. Indeed, the closest tetranucleotide signatures of a plasmid were mostly other plasmid-related contigs and no correlation could be found between plasmids and MAGs coverages. This is consistent with the existence of clusters of similar plasmids able to colonize a range of related taxa (Redondo-Salvo *et al*., 2020). Our hypothesis is that the *Ca*. Accumulibacter related plasmids are also highly micro-diversified, due to recombination events between them (Sen *et al*., 2013; Redondo-Salvo *et al*., 2020). Moreover, these plasmids can also play a role in the nonclonality of *Ca*. Accumulibacter through genomic recombination between plasmids and chromosomes (Frost *et al*., 2005; Sentchilo *et al*., 2013; Redondo-Salvo *et al*., 2020).

Implications and perspectives

The decision to analyse each sample independently is validated by the observation of multiple genomic rearrangements in MAGs originating from different samples. It allowed to compare and assess the robustness of MAGs obtained independently from four different samples and get a first insight into genomic rearrangements in *Ca*. Accumulibacter. High rates of recombination in bacterial populations have been observed in diverse taxa such as *Cyanobacteria* (Rosen *et al*., 2015) or *Enterobacterales* (Redondo-Salvo *et al*., 2020). This 'population-level genotypic heterogeneity' (Nelson *et al*., 2020) in *Ca*. Accumulibacter populations could explain why several studies on the denitrification capabilities of the various clades of *Ca*. Accumulibacter came up with different conclusions (Rubio-Rincon *et al*., 2019). Yet, the mechanisms of these rearrangements and their dynamic evolution warrant further investigations. More generally, the synthetic mock communities are probably too simple to mimic the true diversity of environmental metagenomes (Nicholls *et al*., 2019; Sevim *et al*., 2019). Adding some intra-strain genomic transferts and rearrangements would probably increase the relevance of the mock datasets and help increasing the fitting of MAGs reconstruction pipelines to real metagenomic data. Since sequencing technologies are evolving fast (Frank *et al*., 2016), we can expect that the length and accuracy of sequences will continue to increase in the future. The use of innovative techniques such as chromosome conformation capture may help solving the challenges brought by microdiversity and mobile genetic elements (Marbouty and Koszul, 2015). However, the pipeline used for MAGs reconstruction needs adaptations in order to address the real complexity of metagenomes.

Conclusions

The PacBio long-reads from samples of aerobic granular sludge collected at four different time points, with different feeding conditions, were assembled independently by using the coverage information from the same sample. The introduction of MuLobiSC in the default combination of binning tools from metaWRAP allowed us to disentangle four high-quality *Ca*. Accumulibacter MAGs, two of them present in multiple samples. Through an enhanced contiguity, the comparison of these *Ca*. Accumulibacter highlights multiple genomic recombination events within the *Ca*. Accumulibacter genus. These events are proposed to be largely responsible for the challenges in assembling genomes of *Ca*. Accumulibacter by creating a microdiversity that is uneven across the genomes. They also pose difficulties when using short-reads or hybrid assemblies and binning, but the small size of the contigs does not allow this to be detected. Plasmids related to *Ca*. Accumulibacter have been assembled. They probably participate in the plasticity of their genome and in its adaptation to different ecological niches. The intra-population transfers and recombination of genetic material in *Ca*. Accumulibacter may explain why their phenotype is sometimes decoupled from the clades as defined by a single gene, *pkk1*. The magnitude, mechanisms, and kinetics of genomic rearrangements in *Ca*. Accumulibacter and other bacteria with panmictic

evolution deserves further study. The progress of sequencing technology combined with adapted assembly and binning strategies should make it possible to assemble MAGs of better quality and contiguity along with the pool of associated mobile genetic elements, which play an important role in the adaptation of bacteria to external changing conditions.

Experimental procedures

Biomass sampling

Biomass samples of an aerobic granular sludge (AGS) reactor were collected every week to monitor the microbial communities through 16S rRNA gene amplicon sequencing, as previously described (Adler and Holliger, 2020). Since the microbial community composition was previously shown to be variable from one granule to another (Leventhal *et al*., 2018), a minimum of 20 granules was estimated sufficient to have a representative sample of the bacterial community (Adler, 2019). Therefore, at least 1 ml of wet biomass per sample was collected and homogenized following the protocol described in the study by Adler and Holliger (2020). In addition, four samples were chosen for metagenome DNA sequencing and assembly. The first sample was collected on day 71 (d71) when the reactor was treating simple wastewater, the second and third on day 322 and 427 (d322 and d427) when the reactor was treating complex monomeric wastewater, and the fourth on day 740 (d740) when the reactor was treating complex polymeric wastewater. All biomass samples were stored at −80°C until further use.

DNA extractions

Two different protocols were used for the extraction of metagenomic DNA. The bacterial genomic DNA isolation CTAB protocol (hereafter referred to as extraction A) was applied as follows. To 740 µl of cell suspension, 20 µl of 100 mg ml^{-1} lysozyme solution was added and the sample was incubated for 5 min at room temperature. After addition of 40 µl of 10% SDS and 4 µl of Proteinase K solution (23 mg ml^{-1}, Sigma-Aldrich), the sample was incubated for 1 h at 37°C. Then, 100 µl of 5 M NaCl and 100 µl of a 65°C pre-warmed CTAB solution (4% NaCl, 10% hexadecyl trimethyl ammonium bromide) were added and the reaction was incubated for 10 min at 65°C. Extraction of DNA was obtained by adding 0.5 ml of a chloroform:isoamyl alcohol (24:1) solution and by centrifugation for 10 min at 16 000*g*. The upper aqueous phase was transferred to a new tube and mixed with 0.6 volume of cold isopropanol. After 30 min incubation at room temperature, the DNA was collected by 15 min centrifugation at 16 000*g*. The pellet was washed with 70% ethanol and centrifuged again. The pellet was air-dried for 5 min and resuspended in 97.5 µl of ddH_2O. A volume of 2.5 µl of RNase (4 mg ml^{-1}, Promega) was added and the sample was incubated for 20 min at 37°C. The DNA was then precipitated by adding 20 µl of 3 M potassium acetate solution (pH 5.2) and 400 µl of 100% ethanol for 15 min at −80°C. The DNA was recovered by centrifugation at 4°C and full speed in a table top centrifuge (16 000*g*) for 15 min. The pellet was washed three times in 250 µl of 70% ethanol and centrifuged again. Finally, the pellet was resuspended in 100 µl of 10 mM Tris–HCl buffer (pH 8.0). Maxwell® 16 Tissue DNA Purification Kit (hereafter referred to as extraction B) was applied according to the manufacturer's instructions with the following modifications. Elution of the DNA from the magnetic beads was performed in TE buffer. Eluate was treated with RNase as follows. To 150 µl of eluate, 50 µl of ddH_2O and 5 µl of RNase A solution (4 mg ml^{-1}) were added and the reaction was incubated for 20 min at 37°C. The DNA was then precipitated by adding 20 µl of 3 M potassium acetate solution (pH 5.2) and 400 µl of 100% ethanol for 15 min at −80°C. The DNA was recovered centrifugation at 4C and full speed in a table top centrifuge (16 000*g*) for 15 min. The pellet was washed three times in 250 µl of 70% ethanol and centrifuged again. The DNA pellet was air-dried for 5 min and resuspended in 50 µl of 10 mM Tris–HCl buffer (pH 8.0). All DNA samples were stored at −80°C until further use.

DNA sequencing

The DNA samples obtained with extractions A and B were transmitted to the Lausanne Genomic Technologies Facility (University of Lausanne, Switzerland) for PacBio long-reads (extraction A) and Illumina short-reads (extraction A and B) sequencing. For PacBio sequencing, 4.5 µg of the DNA from sample d71 and 4 µg of the DNA from samples d322, d427 and d740 were used to prepare a SMRTbell library with the PacBio SMRTbell Template Prep Kit 1 (Pacific Biosciences, Menlo Park, CA, USA) according to the manufacturer's recommendations. DNA fragments were selected by size on a Blue Pippin system (Sage Science, Beverly, MA, USA) for molecules larger than 7 kb for sample d71 and d427, 8 kb for sample d322 and 10 kb for sample d740. The resulting libraries were sequenced with v2/v2.1 chemistry and diffusion loading on a PacBio Sequel instrument (Pacific Biosciences) at 600 min movie length using SMRT cells v2. For Illumina sequencing, DNA extractions A and B from samples d71, d322, d427 and d740 were sequenced in multiplex on an Illumina HiSeq platform in paired-end mode (2 × 100 bp).

Long-read assembly

The long-read assembly of each sample was performed by combining minimap2 v2.12 (Li, 2018) and miniasm v0.2.r159 (Li, 2016). Bamtools v2.4.1 (Barnett *et al*., 2011) was used to convert the long-reads from 'bam' to 'fasta' format and keep only the sequences longer than 500 bp. For each data-set, a mapping of the long-reads against themselves was performed with minimap2 and the option -x ava-pb. *De novo* assemblies of the long-reads and assembly graphs were created by using miniasm on the minimap2 output. Several combinations of options were tested in order to maximize the contigs length while keeping the probability of creating misassemblies as low as possible. The quality of the assemblies was assessed by mapping the trimmed long-reads on the assemblies with minimap2 and the option -ax map-pb and samtools 1.8 (Li *et al*., 2009) to convert, sort and index the mapping. The visual inspection of the mapping was performed on the genome viewer IGV v2.4.16 (Robinson *et al*., 2011). After comparison of the assembly statistics and the quality of the mapping, the combination of options -h 700 -s 3000 -g 500 -r 0.75 -n 5 in miniasm was chosen. The used software did not require the sequencing errors to be fixed beforehand. The sequencing errors in the assembled contigs were removed by 10 rounds of polishing: the trimmed long-reads were mapped on the current assemblies with minimap2 and the wrapper especially designed for PacBio data: pbmm2. Arrow© (2011–2018, Pacific Bioscience of California) was used to establish consensus sequences based on this indexed mapping and the current assembly. The sequences of the assembly graphs were replaced by the polished sequences, yet the edges of the graph are the relations between the contigs before the polishing. Some of them may therefore be 'chimeric', but it still provides precious information about the assembly process and the similarity between the contigs before correction. Bandage v0.8.1 (Wick *et al*., 2015) was used in combination with blast+ v2.9.0 (Camacho *et al*., 2009) to visualize the similarities with chosen reference genomes (Supporting Information Table S1).

Short-read and hybrid assemblies

The trimming and quality filtering of Illumina short-reads were performed with trimmomatic v0.36 (Bolger *et al*., 2014) with a sliding window of 10 bp, a minimal quality score threshold (phred33) of 15 and a minimal sequence length of 50 bp. A *de novo* assembly of each dataset was performed with the trimmed sequences using Spades v3.12.0 (Nurk *et al*., 2013) and the script metaspades.py with increasing kmer sizes of 21, 33, 55, 77. For each day, an hybrid assembly the long-reads (extraction A) and short-reads (extraction A) of the corresponding sample was performed using Spades and the command metaspades.py with kmer lengths 21, 33, 55 and 77 and the option --pacbio.

Taxonomic affiliation of contigs

A first taxonomic classification of long-read contigs was performed by CAT v5.2 (von Meijenfeldt *et al*., 2019). CAT uses DIAMOND (Buchfink *et al*., 2015) to align predicted open reading frames to the NCBI non-redundant protein database (Sayers *et al*., 2021) (released 23 November 2020) and computes the most likely probable ancestor. The 16S rRNA genes on the contigs were used to confirm or refine the taxonomic classification. A new tool for taxonomic classification based on ANI, relative evolutionary divergence (RED) and AF, named GTDB-tk v1.4.1, proposed by (Chaumeil *et al*., 2020) was used to propose an alternative classification for ACC004.

Multi-distance long-read binning with silhouette coefficient optimization

The long-read contigs were grouped into MAGs with an in-house script inspired from binning techniques used with Illumina data (Albertsen *et al*., 2013; Alneberg *et al*., 2014), but taking advantage of the large size of the contigs. Four different distance measures were averaged with weighting to obtain a global distance between contigs: a distance based on the tetranucleotide signatures of the DNA sequences, a distance based on the coverage of the short-reads from the same sample, a distance based on the coverage of long-reads and a distance based on the overlapping of BUSCO v4 (Betaproteobacteria BUSCO set). Finally, the binning was optimized using silhouette coefficients. In the following, we use the acronym MuLoBiSC for multi-distance long-read binning with silhouette coefficient optimization, to refer to this in-house binning method.

The pairwise distance based on tetranucleotide characteristic frequencies vector between two DNA sequences was computed as follows: a tetranucleotide median frequency vector was computed on sliding windows of 2000 bp and a resolution of 200 bp over the whole DNA sequences. The euclidean distance between the median frequency vectors of two DNA sequences was computed.

The pairwise distance based on BUSCO was computed as the number of common BUSCO (complete or duplicated) over their total in the two DNA sequences. The presence of BUSCO in the sequences was assessed by BUSCO v4 (Seppey *et al*., 2019) with the bacteria BUSCO set 'betaproteobacteria_odb10'

containing consensus hidden Markov model (HMM) profiles of 569 BUSCO and the option -m geno.

The pairwise distance based on the short-reads coverage between two sequences was computed as follows: the two short-reads datasets corresponding to two different extraction methods (A and B) were aligned separately on the PacBio contigs from the same sample with Bowtie v2.3.4.1 (Langmead and Salzberg, 2012) and the options -k 1 -p 8 -R 3 -D 20 -N 1 -X 1000 -q --phred33. The coverage on each nucleotide was extracted from the mapping by using BEDtools v2.26.0 (Quinlan and Hall, 2010) and the options genomecov -d. The short-reads coverage distance between two sequences was computed as the euclidean distance of their two-dimension median coverage vectors (coverage with extraction A, coverage with extraction B).

The pairwise distance based on the long-reads coverage between two sequences was computed as follows: the trimmed long-reads were mapped on the sequences with pbalign\blasr (Chaisson and Tesler, 2012) and a minimum similarity threshold of 85%, with the command pbalign --minAccuracy 85. The long-reads coverage distance was computed as the euclidean distance between their median coverages (extracted from the mapping output file as for the short-reads coverage distance).

A global distance between two sequences was obtained by averaging these four distances with empirically chosen weightings: 1.146597 for the tetranucleotide distance, 0.00061 for the short-read coverage distance, 0.00067 for the long-read coverage distance and 0.4 for the BUSCO distance. The binning was performed with the following greedy algorithm. The longest contig was used to create the first bin. For each contig ordered with decreasing length, the global distance between the contig and the concatenated sequence of each bin was computed, the contig was attributed to 'closest' bin, if the distance was lower than 0.035. Otherwise, it was used to create a new bin. When a contig was added to an existing bin, new attributes were extracted from the concatenated contigs in the bin for BUSCO and tetranucleotide attributes or by averaging the coverages of the contigs in the bin weighted by their lengths.

To objectively assess the classification of each contig within each bin, we propose to use Silhouette Coefficient (SC). SC was originally introduced as a general graphical aid for interpretation and validation of cluster analysis (Rousseeuw, 1987). SC values were calculated for each contig using the silhouette function in the R package 'cluster' v2.1.0 (Maechler *et al*., 2019). This coefficient, ranging for −1.0 to 1.0, provides a measure of how well a contig is classified within a cluster/bin according to the tightness of the clusters and the separation between them (Van Craenendonck and Blockeel, 2015). For each contig c, SC is computed as:

$$\frac{b(c)-a(c)}{\max(a(c),b(c))} \quad (1)$$

where $a(c)$ is the average dissimilarity between the contig c and all the other contigs in the same bin, $b(c)$ is the smallest mean dissimilarity between c and the contig of another bin. Each SC was weighted by its corresponding contig length before summing them. This weighted sum of silhouette coefficients (WSC) was used as an objective function for optimizing the binning. In each round of optimization, every contig with an SC inferior to 0.0 is reassigned to the new bin that maximize the global WSC increase. The optimization is performed over several rounds until WSC reaches a plateau.

Comparison with recently published binning tools

Automatic binning of the long-read contigs of each sampling day was performed using independently MaxBin2 v2.2.4, MetaBAT2 v2.15, and CONCOCT v1.1.0 through the metaWRAP v1.2 pipeline (Uritskiy *et al*., 2018) with the two short-read datasets from the corresponding day. MetaWRAP's integrative approach enables the refining of bins generated with different combinations of the three published binning tools and the in-house binning tool MuLoBiSC. All tested combinations are presented in the Supporting Information, but in the Results section, we will focus on the comparison of the original combination of MetaBAT2, MaxBin2 and CONCOCT (referred to as metaWRAP_BXC), and the one of MetaBAT2, MaxBin2 and MuLoBiSC (metaWRAP_BXM).

Metrics related to completeness and contamination of bins were compared by using CheckM v1.1.2 (Parks *et al*., 2015) with the options lineage_wf --reduced_tree. WSC associated with the bins provided by each binning tool were calculated. When contigs were not binned, they were automatically gathered into a 'trash' bin to produce WSC consistent with MuLoBiSC.

In silico *16S rRNA gene extraction, sequence and MAG naming*

Potential 16S rRNA gene sequences were extracted from the assemblies by using infernal v1.1.2 (Nawrocki and Eddy, 2013). The contigs were compared to a consensus RNA profile of bacterial 16S rRNA genes (RF00177.cm), provided by Rfam (Bateman *et al*., 2017) with the infernal command cmsearch and the option -A for the multiple alignment output. The DNA sequences matching the 16S rRNA gene consensus were collected using samtools with the option faidx and compared by using blastn v2.5.0+ (Camacho *et al*., 2009) against the MiDAS 16S rRNA gene sequence database vS123_2.1.3 (McIlroy *et al*., 2017).

A name with a three letters acronym followed by a three digits number was attributed to each rRNA sequence with a length higher than 1400 bp. The similarity of the sequences between them was assessed with blastn, the rRNA gene sequences with less than <3 bp differences were given the same name. The MAGs containing at least one 16S rRNA gene sequence were named based on the accronym of this sequences. A list of the acronyms and their corresponding taxa is provided in the Supporting Information Table S2.

Annotation and statistics on predicted genes

The genomic contigs were annotated by using prokka v1.14.6 (Seemann, 2014), and the dependency prodigal (Hyatt *et al*., 2010), with the option --metagenome. The reference genomes were re-annotated in the same way. The average gene lengths per assembly and sample were computed in R from the lengths of the coding DNA sequences (CDS) in the prokka output files of the different assemblies. The numbers of recombinase and transposase gene sequences per MAG were assessed from the output of the bash commands grep -c 'recombinases' and grep -c 'transposase', respectively, on the prokka '.tsv' output files of the bins (metaWRAP_BXM). The table obtained was sorted by ordering the MAGs by the normalized sum of the recombinase and the transposase genes frequencies (i.e. the number of recobinases divided by the average number of recombinases plus the number of transposases divided by the average number of transposases).

Extraction of pkk1

Sequences homologous to the polyphosphate kinase *pkk1* were searched by using hmmer v3.3.1 (http://hmmer.org, 2018), comparing hidden Markov models (HMM) TIGR03705, from the databases TIGRFAMs v15 (Haft *et al*., 2001), with the coding DNA sequences annotated by prokka in the long-read contigs and reference genomes. The *pkk1* sequences corresponding to *Ca*. Accumulibacter, *Propionivibrio* and *Dechloromonas* were kept and aligned by using MAFFT v7.271 (Katoh and Standley, 2013) with a selection of *pkk1* from different *Ca*. Accumulibacter clades (He *et al*., 2007; Peterson *et al*., 2008; Mao *et al*., 2015) listed in the Supporting Information Table S3. The alignment was cut in Jalview v2.7 (Waterhouse *et al*., 2009) to fit the size of the extra *pkk1* sequences.

Phylogenetic trees

The phylogenetic trees were constructed from the sequences alignments with Raxml v8.2.10 (Stamatakis, 2014) and the general time reversible model of nucleotide substitution under the Gamma model of rate heterogeneity (*−m GRTGAMMA*) and the *Dechloromonas* sequences as root (option -o). The corresponding bootstrap values were calculated based on 100 resampling. The 16S rRNA genes phylogenetic tree was edited in R with the packages ggtree (Yu *et al*., 2017b) and treeio (Yu, 2018) and the *pkk1* tree was edited in FigTree v1.4.2 (Rambaut and Drummond, 2015).

Mummer plot and MAGs similarity statistics

NUCmer (NUCleotide MUMmer) v3.1 (Kurtz *et al*., 2004) was used to compare MAGs. Mummer plot comparing two MAGs were drawn with the NUCmer command mummerplot, generally with the options - layout -f with the delta file filtered to plot only alignments >4 kb. The proportion of aligned bases between two MAGs is a weighted mean of the proportion of aligned bases. It is computed as:

$$\text{Aligned\_bases} = \frac{\text{AlignedBases\_ref} + \text{AlignedBases\_qry}}{\text{TotalBases\_ref} + \text{TotalBases\_qry}} \quad (2)$$

The average nucleotide identity (ANI) was computed by Fastani (Jain *et al*., 2018).

GC skew plots

Chen *et al*. (2020) showed that the cumulative GC skew pattern is rather symmetric in most bacterial genomes (~85%). Asymmetric patterns can help identify potential misassemblies. In combination with others, the cumulative GC skew is a valuable indicator to assess the sequence accuracy of MAGs. GC skew plots were created with the script gc_skew.py (Brown *et al*., 2016) on single contig MAG DNA sequences obtained with the long-reads assemblies and chosen reference genomes (Supporting Information Table S1).

Identification of plasmids

Contigs belonging to type F plasmids were identified in the assemblies by looking for the gene plasmid replication initiator (*trfA*) in the prokka annotation. *TrfA* is a plasmid replication initiator protein widespread in Gram-bacteria. It was shown to be essential for the replication of the RK2 plasmid in *Escherichia coli* and other gram-bacteria (Thomas *et al*., 1980). BUSCO are rather essential genes, they are not expected in abundance in plasmids. Contigs with more than two BUSCO from the Betaproteobacteria set, were excluded. Contigs indicated as circular unitig by miniasm where considered as

circularized. In order to detect identical plasmids in the different assemblies, the identified F plasmids were compared two by two by using NUCmer.

PCR-based validation of a recombination

From the Nucmer alignment report of the contigs UNC4029 (corresponding to ACC012) and UNC4079 (ACC007), four primers were designed on conserved regions located up- and downstream of the selected genetic rearrangement (Supporting Information Table S4).

PCR reactions were performed with MyTaq™ DNA polymerase (Bioline, LABGENE Scientific, Chatel- St-Denis, Switzerland) according to the manufacturer's instructions. Reactions were conducted in 100 µl with 10 ng of DNA from sample d740 as template. Aliquots of 10 µl of PCR products were analysed by agarose gel electrophoresis with a gel containing 0.8% agarose in Tris-acetate-EDTA buffer using standard procedure. The remaining PCR products were purified using the QIAquick PCR Purification Kit (QIAGEN AG, Hombrechtikon, Switzerland) and eluted in 30 µl of elution buffer. DNA concentration was measured with the NanoDrop 1000 spectrophotometer (Fisher Scientific AG, Reinach, Switzerland). Cloning of PCR products was done with the pGEM-T easy vector (Promega AG, Dubendorf, Switzerland) following the manufacturer's instructions. An insert:vector ratio of 5 was applied and ligation was incubated for 4 h at 16°C. Aliquots for 5 µl of ligation products were transformed in 50 µl of JM109 competent cells (Promega), and cells were plated on LB-agar plates containing 50 $pgml^{-1}$ ampicillin. Clones were selected by colony PCR performed with MyTaq™ DNA polymerase (Bioline) and 1 µl from 10 µl colony suspension that were previously boiled for 5 min. Two clones with the expected insert size were selected and cultivated in 5 ml of LB-ampicillin medium overnight at 37°C and 160 rpm. Plasmids were purified using the QIAprep Spin Miniprep Kit (QIAGEN) following the manufacturer's instructions. Aliquots of 200 ng of plasmid DNA were sent to Fasteris (Fasteris SA, Plan-les-Ouates, Switzerland) for sequencing using T7 and SP6 standard primers. DNA sequences were analysed using the SnapGene Viewer software (from Insightful Science; available at snapgene.com).

Identification of Ca*. Accumulibacter MAGs in individual granules*

In 2018, Leventhal *et al*. (2018) identified that the dominant genotypes of *Ca*. Accumulibacter differed from one granule to another. The short-read sequences of the 142 investigated metagenomes obtained from samples taken 9 and 5 months before the start of our experiment in the same reactor were mapped on references following the procedure described by Leventhal *et al*. (2018). The set of reference genomes was replaced with Ca. Accumulibacter MAGs provided by MetaBAT2 (ACC001a and ACC001bc) for MAGs related to *Ca*. Accumulibacter UW1 and with metaWRAP_BXM for the other MAGs. *Ca*. Accumulibacter str. BA-91 was kept in the reference since it is not similar to any of our MAGs. A heatmap of the proportions of mapping reads per granule was plotted with the function heatmap.2 from the package gplots v3.1.1 (Warnes *et al*., 2020) in R.

Availability of data and materials

The data generated and analysed during this study are included in this published article and its supporting information files. The long-read and short-read runs and the high-quality drafts ACC003 from sample d322, ACC004 and ACC005 from sample d427, ACC007 and ACC012 from sample d740 and DEC001 from sample d71 are available in the ENA repository under the project accession number PRJEB38840. The script of the in-house binning MuLoBiSC, which was specially adapted for the experimental design of this study, can be consulted at https://github.com/Aline-Git/MuLoBiSC.

Author's contributions

A.A., M.P. and C.H. planned the experiments, J.M. designed and conducted the molecular work, A.A. ran the SBR reactors, A.A. performed the assembly, A.A., S.P. and M.P. performed the binning, A.A. and S.P. drafted the paper, A.A., S.P., M.P., J.M. and C.H. reviewed it and provided valuable edits. All authors contributed to the article and approved the submitted version.

Acknowledgements

The authors thank Vincent Jeannet (LBE, EPFL), for the tests on DNA extraction. The work of Emmy Oppliger, Idriss Hendaoui, Marie Horisberger and Valerie Berclaz (LBE, EPFL), on the maintenance of the reactors is kindly acknowledged. We thank Emmanuelle Rohrbach and Stephane Marquis (LBE, EPFL) for the training of the apprentice and their help in the laboratory and Marc Deront (LBE, EPFL) for the installation of the computing machine. The authors thank Emanuel Schmid-Siegert for his precious guidance on the assembly of long reads. They thank Gabriel E. Leventhal and Otto X. Cordero (DCEE, MIT) for their work on individual granules. This research was financed by the Swiss National Science Foundation (SNSF), grant number 200021–152963. Open Access Funding provided by Ecole Polytechnique Federale de Lausanne. [Correction added on 6 June 2022, after first online publication: CSAL funding statement has been added.]

References

Adler, A., and Holliger, C. (2020) Multistability and reversibility of aerobic granular sludge microbial communities upon changes from simple to complex synthetic wastewater and Back. *Front Microbiol* **11**: 574361. https://doi.org/10.3389/fmicb.2020.574361.

Adler, A.S. (2019) *The Effect of Different Organic Substrates on the Microbial Communities of Aerobic Granular Wastewater Treatment Sludge*, Vol. **261**. Lausanne: EPFL. https://doi.org/10.5075/epfl-thesis-9678.

Ahlgren, N.A., Chen, Y., Needham, D.M., Parada, A.E., Sachdeva, R., Trinh, V., *et al*. (2017) Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. *Environ Microbiol* **19**: 2434–2452. https://doi.org/10.1111/1462-2920.13768

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G.W., and Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31**: 533–538. https://doi.org/10.1038/nbt.2579.

Albertsen, M., McIlroy, S.J., Stokholm-Bjerregaard, M., Karst, S.M., and Nielsen, P.H. (2016) "Candidatus Propionivibrio aalborgensis": a novel glycogen accumulating organism abundant in full-scale enhanced biological phosphorus removal plants. *Front Microbiol* **7**: 1033. https://doi.org/10.3389/fmicb.2016.01033.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., *et al*. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144–1146. https://doi.org/10.1038/NMETH.3103.

Antwis, R.E., Griffiths, S.M., Harrison, X.A., Aranega-Bou, P., Arce, A., Bettridge, A.S., *et al*. (2017) Fifty important research questions in microbial ecology. *FEMS Microbiol Ecol* **93**: 5. https://doi.org/10.1093/femsec/fix044.

Arredondo-Alonso, S., Top, J., McNally, A., Puranen, S., Pesonen, M., Pensar, J., *et al*. (2020) Plasmids shaped the recent emergence of the major nosocomial pathogen enterococcus faecium. *MBio* **11**: e03284-19. https://doi.org/10.1128/mBio.03284-19.

Arumugam, K., Bessarab, I., Haryono, M.A.S., Liu, X., Zuniga-Montanez, R.E., Roy, S., *et al*. (2021) Recovery of complete genomes and non-chromosomal replicons from activated sludge enrichment microbial communities with long read metagenome sequencing. *npj Bioflim Microb* **7**: 23. https://doi.org/10.1038/s41522-021-00196-6

Ayling, M., Clark, M.D., and Leggett, R.M. (2020) New approaches for metagenome assembly with short reads. *Brief Bioinform* **21**: 584–594. https://doi.org/10.1093/bib/bbz020.

Barco, R.A., Garrity, G.M., Scott, J.J., Amend, J.P., Nealson, K.H., and Emerson, D. (2020) A genus definition for bacteria and Archaea based on a standard genome relatedness index. *MBio* **11**: e02475-19. https://doi.org/10.1128/mBio.02475-19.

Barnett, D.W., Garrison, E.K., Quinlan, A.R., Stroemberg, M. P., and Marth, G.T. (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691–1692. https://doi.org/10.1093/bioinformatics/btr174.

Bateman, A., Kalvari, I., Argasinska, J., Finn, R.D., Petrov, A.I., Quinones-Olvera, N., *et al*. (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**: 335–342. https://doi.org/10.1093/nar/gkx1038.

Bertrand, D., Shaw, J., Kalathiyappan, M., Ng, A.H.Q., Kumar, M.S., Li, C., *et al*. (2019) Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* **37**: 937–944. https://doi.org/10.1038/s41587-019-0191-2

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., *et al*. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**: 725–731. https://doi.org/10.1038/nbt.3893.

Breitwieser, F.P., Lu, J., and Salzberg, S.L. (2019) A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* **20**: 1125–1139. https://doi.org/10.1093/bib/bbx120.

Brotto, A.C., Annavajhala, M.K., and Chandran, K. (2018) Metatranscriptomic investigation of adaptation in NO and N2O production from a lab-scale nitrification process upon repeated exposure to anoxic-aerobic cycling. *Front Microbiol* **9**: 3012. https://doi.org/10.3389/fmicb.2018.03012.

Brown, C.T., Olm, M.R., Thomas, B.C., and Banfield, J.F. (2016) Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol* **34**: 1256–1263. https://doi.org/10.1038/nbt.3704.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60. https://doi.org/10.1038/nmeth.3176.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST plus: architecture and applications. *BMC Bioinformatics* **10**: 421. https://doi.org/10.1186/1471-2105-10-421.

Chaisson, M.J., and Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**: 23. https://doi.org/10.1186/1471-2105-13-238.

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., and Parks, D. H. (2020) GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**: 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A.M., and Banfield, J.F. (2020) Accurate and complete genomes from metagenomes. *Genome Res* **30**: 315–333. https://doi.org/10.1101/gr.258640.119.

Daims, H., Lebedeva, E.V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., *et al*. (2015) Complete nitrification by Nitrospira bacteria. *Nature* **528**: 504. https://doi.org/10.1038/nature16461.

Dilthey, A.T., Jain, C., Koren, S., and Phillippy, A.M. (2019) Strain-level metagenomic assignment and compositional

estimation for long reads with MetaMaps. *Nat Commun* **10**: 3066. https://doi.org/10.1038/s41467-019-10934-2

Driscoll, C.B., Otten, T.G., Brown, N.M., and Dreher, T.W. (2017) Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand Genomic Sci* **12**: 9. https://doi.org/10.1186/s40793-017-0224-8

Dueholm, M.S., Nierychlo, M., Andersen, K.S., Rudkjøbing, S.V., Knutsson S., Albertsen, M., *et al*. (2021) Midas 4: a global catalogue of full-length 16s rrna gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *bioRxiv*. https://doi.org/10.1101/2021.07.06.451231.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., *et al*. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138. https://doi.org/10.1126/science.1162986.

Fernando, E.Y., McIlroy, S.J., Nierychlo, M., Herbst, F.-A., Petriglieri, F., Schmid, M.C., *et al*. (2019) P.H resolving the individual contribution of key microbial populations to enhanced biological phosphorus removal with Raman-FISH. *ISME J* **13**: 1933–1946. https://doi.org/10.1038/s41396-019-0399-7.

Flowers, J.J., He, S., Malfatti, S., del Rio, T.G., Tringe, S.G., Hugenholtz, P., and McMahon, K.D. (2013) Comparative genomics of two '*Candidatus* Accumulibacter' clades performing biological phosphorus removal. *ISME J* **7**: 2301–2314. https://doi.org/10.1038/ismej.2013.117.

Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G.H., McHardy, A.C., Nederbragt, A.J., and Pope, P.B. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci Rep* **6**: 25373. https://doi.org/10.1038/srep25373.

Frost, L., Leplae, R., Summers, A., and Toussaint, A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* **3**: 722–732. https://doi.org/10.1038/nrmicro1235.

Gao, H., Mao, Y., Zhao, X., Liu, W.-T., Zhang, T., and Wells, G. (2019) Genome-centric metagenomics resolves microbial diversity and prevalent truncated denitrification pathways in a denitrifying PAO-enriched bioprocess. *Water Res* **155**: 275–287. https://doi.org/10.1016/j.watres.2019.02.020.

Gat-Viks, I., Sharan, R., and Shamir, R. (2003) Scoring clustering solutions by their biological relevance. *Bioinformatics* **19**: 2381–2389. https://doi.org/10.1093/bioinformatics/btg330.

Haft, D., Loftus, B., Richardson, D., Yang, F., Eisen, J., Paulsen, I., and White, O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* **29**: 41–43. https://doi.org/10.1093/nar/29.1.41.

He, S., Gall, D.L., and McMahon, K.D. (2007) "*Candidatus* Accumulibacter" population structure in enhanced biological phosphorus removal sludges as revealed by polyphosphate kinase genes. *Appl Environ Microbiol* **73**: 5865–5874. https://doi.org/10.1128/AEM.01207-07.

hmmer.org, H.H.M : HMMER 3.2.1 (June 2018); hmmscan: search sequence(s) against a profile database Copyright (C) Howard Hughes Medical Insitute. Freely distributed under the BSD open source license (2018).

Hugoson, E., Lam, W.T., and Guy, L. (2020) miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* **36**: 936–937. https://doi.org/10.1093/bioinformatics/btz664.

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bionformatics* **11**: 119. https://doi.org/10.1186/1471-2105-11-119.

Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., and Tyson, G.W. (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**: e603. https://doi.org/10.7717/peerj.603

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**: 5114. https://doi.org/10.1038/s41467-018-07641-9

Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., and Akeson, M. (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* **12**: 115–351. https://doi.org/10.1038/NMETH.3290.

Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**: e7359. https://doi.org/10.7717/peerj.7359.

Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J. W., Biller, S.J., Coe, A., *et al*. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* **344**: 416–420. https://doi.org/10.1126/science.1248575.

Katoh, K., and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780. https://doi.org/10.1093/molbev/mst010.

Koch, H., Germscheid, N., Freese, H.M., Noriega-Ortega, B., Luecking, D., Berger, M., *et al*. (2020) Genomic, metabolic and phenotypic variability shapes ecological differentiation and intraspecies interactions of *Alteromonas macleodii*. *Sci Rep* **10**: 809. https://doi.org/10.1038/s41598-020-57526-5

Kono, N., and Arakawa, K. (2019) Nanopore sequencing: review of potential applications in functional genomics. *Dev Growth Differ* **61**: 316–326. https://doi.org/10.1111/dgd.12608.

Konstantinidis, K., and Tiedje, J. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**: 2567–2572. https://doi.org/10.1073/pnas.0409727102.

Konstantinidis, K.T., Rossello-Mora, R., and Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. *ISME J* **11**: 2399–2406. https://doi.org/10.1038/ismej.2017.113.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12. https://doi.org/10.1186/gb-2004-5-2-r12.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nat Methods* **9**: 354–357. https://doi.org/10.1038/NMETH.1923.

Leventhal, G.E., Boix, C., Kuechler, U., Enke, T.N., Sliwerska, E., Holliger, C., and Cordero, O.X. (2018)

Strain-level diversity drives alternative community types in millimetre-scale granular biofilms. *Nat Microbiol* **3**: 1295–1303. https://doi.org/10.1038/s41564-018-0242-3

Li, H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110. https://doi.org/10.1093/bioinformatics/btw152.

Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Liao, J., Orsi, R.H., Carroll, L.M., and Wiedmann, M. (2020) Comparative genomics reveals different population structures associated with host and geographic origin in antimicrobial-resistant salmonella enterica. *Environ Microbiol* **22**: 2811–2828. https://doi.org/10.1111/1462-2920.15014.

Lovmar, L., Ahlford, A., Jonsson, M., and Syvanen, A. (2005) Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics* **6**: 35. https://doi.org/10.1186/1471-2164-6-35.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. (2019). Cluster: cluster analysis basics and extensions (R package version 2.1.0)

Maguire, F., Jia, B., Gray, K.L., Lau, W.Y.V., Beiko, R.G., and Brinkman, F.S.L. (2020) Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microbial Genomics* **6**: 000436. https://doi.org/10.1099/mgen.0.000436.

Mao, Y., Graham, D.W., Tamaki, H., and Zhang, T. (2015) Dominant and novel clades of *Candidatus* Accumulibacter phosphatis in 18 globally distributed full-scale wastewater treatment plants. *Sci Rep* **5**: 11857. https://doi.org/10.1038/srep11857.

Marbouty, M., and Koszul, R. (2015) Metagenome analysis exploiting high-throughput chromosome conformation capture (3C) data. *Trends Genet* **31**: 673–682. https://doi.org/10.1016/j.tig.2015.10.003.

Martin, H.G., Ivanova, N., Kunin, V., Warnecke, F., Barry, S. K. & Salamov, A. et al. (2014) Complete sequence of chromosome of Candidatus Accumulibacter phosphatis clade IIA str. UW-1, accession CP001715.

McIlroy, S.J., Kirkegaard, R.H., McIlroy, B., Nierychlo, M., Kristensen, J.M., Karst, S.M., *et al*. (2017) MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. *Database*. bax016. https://doi.org/10.1093/database/bax016

Moss, E.L., Maghini, D.G., and Bhatt, A.S. (2020) Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* **38**: 701–707. https://doi.org/10.1038/s41587-020-0422-6

Nawrocki, E.P., and Eddy, S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**: 2933–2935. https://doi.org/10.1093/bioinformatics/btt509.

Nelson, W.C., Maezato, Y., Wu, Y.-W., Romine, M.F., and Lindemann, S.R. (2016) Identification and resolution of microdiversity through metagenomic sequencing of parallel consortia. *Appl Environ Microbiol* **82**: 255–267. https://doi.org/10.1128/AEM.02274-15.

Nelson, W.C., Tully, B.J., and Mobberley, J.M. (2020) Biases in genome reconstruction from metagenomic data. *PeerJ* **8**: e10119. https://doi.org/10.7717/peerj.10119.

Nicholls, S.M., Quick, J.C., Tang, S., and Loman, N.J. (2019) Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **8**: giz043. https://doi.org/10.1093/gigascience/giz043.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A.A., Korobeynikov, A., Lapidus, A., *et al*. (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology*: 714–737. https://doi.org/10.1089/cmb.2013.0084

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., McVeigh, D.H.R., Rajput, B., *et al*. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: 733–745. https://doi.org/10.1093/nar/gkv1189.

Okabe, S., Kindaichi, T., and Ito, T. (2004) Mar-fish—an ecophysiological approach to link phylogenetic affiliation and *in situ* metabolic activity of microorganisms at a single-cell resolution. *Microb Environ* **19**: 83–98. https://doi.org/10.1264/jsme2.19.83.

Oyserman, B.O., Noguera, D.R., del Rio, T.G., Tringe, S.G., and McMahon, K.D. (2016) Metatranscriptomic insights on gene expression and regulatory controls in *Candidatus* Accumulibacter phosphatis. *ISME J* **10**: 810–822. https://doi.org/10.1038/ismej.2015.155.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., and Hugenholtz, P. (2020) A complete domain-to-species taxonomy for bacteria and Archaea. *Nat Biotechnol* **38**: 1079–1086. https://doi.org/10.1038/s41587-020-0501-8

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055. https://doi.org/10.1101/gr.186072.114.

Peterson, S.B., Warnecke, F., Madejska, J., McMahon, K.D., and Hugenholtz, P. (2008) Environmental distribution and population biology of *Candidatus* Accumulibacter, a primary agent of biological phosphorus removal. *Environ Microbiol* **10**: 2692–2703. https://doi.org/10.1111/j.1462-2920.2008.01690.x.

Qin, Q.-L., Xie, B.-B., Zhang, X.-Y., Chen, X.-L., Zhou, B.-C., Zhou, J., *et al*. (2014) A proposed genus boundary for the prokaryotes based on genomic insights. *J Bacteriol* **196**: 2210–2215. https://doi.org/10.1128/JB.01688-14.

Quinlan, A.R., and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rambaut, A. & Drummond, A. (2015) FigTree, ver. 1.4.2. http://tree.bio.ed.ac.uk/software/figtree/

Rang, F.J., Kloosterman, W.P., and de Ridder, J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.

Genome Biol **19**: 90. https://doi.org/10.1186/s13059-018-1462-9

Redondo-Salvo, S., Fernandez-Lopez, R., Ruiz, R., Vielva, L., de Toro, M., Rocha, E.P.C., *et al*. (2020) Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat Commun* **11**: 3602. https://doi.org/10.1038/s41467-020-17278-2

Richter, M., and Rossello-Mora, R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**: 19126–19131. https://doi.org/10.1073/pnas.0906412106.

Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26. https://doi.org/10.1038/nbt.1754.

Rosen, M.J., Davison, M., Bhaya, D., and Fisher, D.S. (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science* **348**: 1019–1023. https://doi.org/10.1126/science.aaa4456.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**: 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Rubio-Rincon, F.J., Weissbrodt, D.G., Lopez-Vazquez, C.M., Welles, L., Abbas, B., Albertsen, M., *et al*. (2019) "Candidatus Accumulibacter delftensis": a clade IC novel polyphosphate-accumulating organism without denitrifying activity on nitrate. *Water Res* **161**: 136–151. https://doi.org/10.1016/j.watres.2019.03.053.

Sanders, J.G., Nurk, S., Salido, R.A., Minich, J., Xu, Z.Z., Zhu, Q., *et al*. (2019) Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biol* **20**: 226. https://doi.org/10.1186/s13059-019-1834-9

Sangwan, N., Xia, F., and Gilbert, J.A. (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**: 8. https://doi.org/10.1186/s40168-016-0154-5

Sayers, E.W., Beck, J., Bolton, E.E., Bourexis, D., Brister, J.R., Canese, K., *et al*. (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **49**: 10–17. https://doi.org/10.1093/nar/gkaa892.

Schmid, M., Frei, D., Patrignani, A., Schlapbach, R., Frey, J.E., Remus-Emsermann, M.N.P., and Ahrens, C.H. (2018) Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res* **46**: 8953–8965. https://doi.org/10.1093/nar/gky726.

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droege, J., *et al*. (2017) Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods* **14**: 1063–1071. https://doi.org/10.1038/NMETH.4458.

Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**: 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Sen, D., Brown, C.J., Top, E.M., and Sullivan, J. (2013) Inferring the evolutionary history of IncP-1 plasmids despite incongruence among backbone gene trees. *Mol Biol Evol* **30**: 154–166. https://doi.org/10.1093/molbev/mss210.

Sentchilo, V., Mayer, A.P., Guy, L., Miyazaki, R., Tringe, S.G., Barry, K., *et al*. (2013) Community-wide plasmid gene mobilization and selection. *ISME J* **7**: 1173–1186. https://doi.org/10.1038/ismej.2013.13.

Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing genome assembly and annotation completeness. In Kollmar, M. (Ed). *Gene Prediction*, Methods in Molecular Biology. New York, NY, 1962, 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Sevim, V., Lee, J., Egan, R., Clum, A., Hundley, H., Lee, J., *et al*. (2019) Shotgun metagenome data of a defined mock community using Oxford Nanopore PacBio and Illumina technologies. *Sci Data* **6**: 285. https://doi.org/10.1038/s41597-019-0287-z

Shapiro, B.J. (2016) How clonal are bacteria over time? *Curr Opin Microbiol* **31**: 116–123. https://doi.org/10.1016/j.mib.2016.03.013.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120. https://doi.org/10.1101/gr.142315.112.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., and Banfield, J.F. (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**: 836–843. https://doi.org/10.1038/s41564-018-0171-1

Simao, F.A., Waterhouse, R.M., Loannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

Singleton, C.M., Petriglieri, F., Kristensen, J.M., Kirkegaard, R.H., Michaelsen, T.Y., Andersen, M.H., *et al*. (2021) Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing. *Nat Commun* **12**: 2009. https://doi.org/10.1038/s41467-021-22203-2.

Skennerton, C.T., Barr, J.J., Slater, F.R., Bond, P.L., and Tyson, G.W. (2015) Expanding our view of genomic diversity in *Candidatus* Accumulibacter clades. *Environ Microbiol* **17**: 1574–1585. https://doi.org/10.1111/1462-2920.12582.

Slaby, B.M., Hackl, T., Horn, H., Bayer, K., and Hentschel, U. (2017) Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *ISME J* **11**: 2465–2478. https://doi.org/10.1038/ismej.2017.101.

Somerville, V., Lutz, S., Schmid, M., Frei, D., Moser, A., Irmler, S., *et al*. (2019) Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol* **19**: 143. 10.1186/s12866-019-1500-0

Song, W.-Z., and Thomas, T. (2017) Binning refiner: improving genome bins through the combination of different binning programs. *Bioinformatics* **33**: 1873–1875. https://doi.org/10.1093/bioinformatics/btx086.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Thomas, C., Meyer, R., and Helinski, D. (1980) Regions of broad-host-range plasmid RK2 which are essential for replication and maintenance. *J Bacteriol* **141**: 213–222. https://doi.org/10.1128/JB.141.1.213-222.1980.

Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E., and Thompson, F.L. (2013) Global analyses of Ceratocystis cacaofunesta mitochondria: from genome to proteome. *BMC Genomics* **14**: 91. https://doi.org/10.1186/1471-2164-14-91.

Thomson, N., Baker, S., Pickard, D., Fookes, M., Anjum, M., Hamlin, N., *et al*. (2004) The role of prophage-like elements in the diversity of salmonella enterica serovars. *J Mol Biol* **339**: 279–300. https://doi.org/10.1016/j.jmb.2004.03.058.

Uritskiy, G.V., DiRuggiero, J., and Taylor, J. (2018) MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**: 158. https://doi.org/10.1186/s40168-018-0541-1

Van Craenendonck, T. & Blockeel, H. Using internal validity measures to compare clustering algorithms, Benelearn 2015 Poster presentations (online), pp. 1–8, 2015.

von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2019) Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* **20**: 217. https://doi.org/10.1186/s13059-019-1817-x

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., *et al*. (2020). Gplots: various R Programming Tools for Plotting Data R package version 3.1.1. https://CRAN.R-project.org/package=gplots

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009) Jalview version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

Weigand, M.R., Peng, Y., Loparev, V., Batra, D., Bowden, K. E., Burroughs, M., *et al*. (2017) The history of *Bordetella pertussis* genome evolution includes structural rearrangement. *J Bacteriol* **199**: e00806-16. https://doi.org/10.1128/JB.00806-16.

Whitman, W.B. (2015) Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst Appl Microbiol* **38**: 217–222. https://doi.org/10.1016/j.syapm.2015.02.003.

Wick, R.R., Schultz, M.B., Zobel, J., and Holt, K.E. (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**: 3350–3352. https://doi.org/10.1093/bioinformatics/btv383.

Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., and Lin, Y. (2020) MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics* **36**: 3–11. https://doi.org/10.1093/bioinformatics/btaa441.

Wiedenbeck, J., and Cohan, F.M. (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol Rev* **35**: 957–976. https://doi.org/10.1111/j.1574-6976.2011.00292.x.

Wu, Y.-W., Simmons, B.A., and Singer, S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**: 605–607. https://doi.org/10.1093/bioinformatics/btv638.

Xia, Y., Kong, Y., and Nielsen, P.H. (2008) *In situ* detection of starch-hydrolyzing microorganisms in activated sludge. *FEMS Microbiol Ecol* **66**: 462–471. https://doi.org/10.1111/j.1574-6941.2008.00559.x.

Yarza, P., Yilmaz, P., Pruesse, E., Gloeckner, F.O., Ludwig, W., Schleifer, K.-H., *et al*. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**: 635–645. https://doi.org/10.1038/nrmicro3330.

Yoon, S.-H., Ha, S.-M., Kwon, S., Lim, J., Kim, Y., Seo, H., and Chun, J. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol* **67**: 1613–1617. https://doi.org/10.1099/ijsem.0.001755.

Yu, F.B., Blainey, P.C., Schulz, F., Woyke, T., Horowitz, M. A., and Quake, S.R. (2017a) Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *Elife* **6**: e26580. https://doi.org/10.7554/eLife.26580.

Yu, G. (2018). Treeio: base classes and functions for phylogenetic tree input and output R package version 1.4.1. https://guangchuangyu.github.io/software/treeio

Yu, G., Smith, D., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017b) ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**: 28–36. https://doi.org/10.1111/2041-210X.12628.

Zhi, X.-Y., Zhao, W., Li, W.-J., and Zhao, G.-P. (2012) Prokaryotic systematics in the genomics era. *Antonie Van Leeuwenhoek* **101**: 21–34. https://doi.org/10.1007/s10482-011-9667-x

Zowawi, H.M., Forde, B.M., Alfaresi, M., Alzarouni, A., Farahat, Y., Chong, T.-M., *et al*. (2015) Stepwise evolution of pandrug-resistance in *Klebsiella pneumoniae*. *Sci Rep* **5**: 15082. https://doi.org/10.1038/srep15082.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Supporting Information file 1 — Supplementary material and Supplementary tables and figures.

Supporting Information file 2 — Tetranucleotide frequency distances between long-read contigs from sample d71

Supporting Information file 3 — Tetranucleotide frequency distances between long-read contigs from sample d322

Supporting Information file 4 — Tetranucleotide frequency distances between long-read contigs from sample d427

Supporting Information file 5 — Tetranucleotide frequency distances between long-read contigs from sample d740

Supporting Information file 6 — Collection of *Rhodocyclaceae* MAGs analysed in this study.