Systematic and Applied Microbiology 42 (2019) 30–40

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Recovering microbial genomes from metagenomes in hypersaline environments: The Good, the Bad and the Ugly

María Dolores Ramos-Barbero[a,1], Ana-B. Martin-Cuadrado[a,1], Tomeu Viver[b], Fernando Santos[a], Manuel Martinez-Garcia[a], Josefa Antón[a,c,*]

[a] *Department of Physiology, Genetics and Microbiology, University of Alicante, Alicante, Spain*
[b] *Department of Animal and Microbial Biodiversity, Marine Microbiology Group, Mediterranean Institute for Advanced Studies (IMEDEA, CSIC-UIB), Esporles, Spain*
[c] *Multidisciplinary Institute of Environmental Studies Ramon Margalef, University of Alicante, Alicante, Spain*

ARTICLE INFO

Article history:
Received 4 May 2018
Received in revised form 6 November 2018
Accepted 7 November 2018

Keywords:
Microdiversity
Binning
Hypersaline
Haloquadratum
Salinibacter
Halorubrum

ABSTRACT

Current metagenomic tools allow the recovery of microbial genomes directly from the environment. This can be accomplished by binning metagenomic contigs according to their coverage and tetranucleotide frequency, followed by an estimation of the bin quality. The public availability of bioinformatics tools, together with the decreasing cost of next generation sequencing, are democratizing this powerful approach that is spreading from specialized research groups to the general public. Using metagenomes from hypersaline environments, as well as mock metagenomes composed of Archaea and Bacteria frequently found in these systems, we have analyzed the advantages and difficulties of the binning process in these extreme environments to tackle microbial population diversity. These extreme systems harbor relatively low species diversity but high intraspecific diversity, which can compromise metagenome assembly and therefore the whole binning process. The main goal is to compare the output of the binning process with what is previously known from the analyzed samples, based on years of study using different approaches. Several scenarios have been analyzed in detail: (i) a good quality bin from a species highly abundant in the environment; (ii) an intermediate quality bin with incongruences that can be solved by further analyses and manual curation, and (iii) a low-quality bin to investigate the failure to recover a very abundant microbial genome as well as some possible solutions. The latter can be considered the "great metagenomics anomaly" and is mainly due to assembly problems derived from the microdiversity of naturally co-existing populations in nature.

Introduction

Metagenomics is the genomic analysis of microorganisms by direct extraction and sequencing of DNA from a natural sample [18]. Since its beginning, 20 years ago, it has spurred a revolution in microbial ecology similar to that propelled by the rRNA approach [3], which unveiled the widespread distribution and abundance of so-called microbial dark matter [47]. This fraction of the biosphere encompasses all the microbes that have not been brought into pure culture and represents the overwhelming majority of the microbes in nature [45,55]. Metagenomics is unveiling not only the phylogenetic identity of this uncultured majority but also its metabolic capabilities and roles in biogeochemical cycles. Furthermore, the "judicious use of bioinformatics tools" [52] has allowed the retrieval of (almost) complete microbial genomes from metagenomes (known as MAGs or metagenomic assembled genomes), which in turn allows linking specific microbes to their metabolic capabilities and paves the way to evolutionary and biogeography studies. The first reconstructed MAGs were *Leptospirillum* GII and *Ferroplasma* t.II from an acidophilic biofilm from acid mine drainage [61]. In that case, not much sequence was necessary as the biofilm was dominated by a small number of populations and the genetic diversity was shown to be quite restricted. Since then, the number of MAGs in databases has been growing exponentially [8,41] and is now over 8445 (Genomes Online Database, August 2018 [46]).

Although there are no strict guidelines on how to retrieve MAGs from metagenomes, the Genomic Standards Consortium,

[*] Corresponding author at: Department of Physiology, Genetics and Microbiology, University of Alicante, Apartado 99, 03080 Alicante, Spain.
E-mail address: anton@ua.es (J. Antón).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.syapm.2018.11.001
0723-2020/© 2018 Elsevier GmbH. All rights reserved.

which includes the world's leading research scientists in the field, has issued standards for reporting bacterial and archaeal genome sequences assembled from metagenomes [8]. These MIMAG (minimum information about a MAG) include parameters related to the quality of the assembly of the metagenome reads, and the completion and contamination of the retrieved genomes, determined by comparison with a set of conserved marker genes (see below) that should be present in single copy in the MAG. High quality MAGs meeting the most strict quality criteria have been proposed to be used as the basis of a "new taxonomy" [25,64] (or, at least, of new ways of naming these prokaryotes), that would include the uncultured majority, since indeed this is a highly relevant part of the microbial biosphere.

The mandatory standards for considering a MAG as a high-quality draft genome are completion over 90%, contamination under 5%, and convincing proofs of the assembly quality (such as the presence of rRNA genes and at least 18 tRNAs). Conversely, low quality draft MAGs are composed of contigs than constitute less than 50% of the genome and have less than 10% contamination, with little or no review of the assembly, other than standard statistics. Assemblages of sequences with contamination above 10% are not considered as MAGs in [8]. Retrieval of MAGs is thus based on two pillars: good assembly and good selection of the marker genes used to define completion and contamination. In fact, in the MIMAG proposal [8], these two aspects are mentioned as fields for improvement in the future, as new methods of metagenomics analysis appear, and new phylogenetic clades are discovered.

A powerful approach to recover (nearly) complete genomes from metagenomes is the so-called binning approach in which metagenomic contigs are binned into groups which likely correspond to the same population in the original samples. Frequently, this binning is based on coverage and tetranucleotide frequency of the contigs, followed by an estimation of the quality of the bin. Among the different tools available for binning and quality assessment, MaxBin [65] and CheckM [40], respectively, stand out due to their high performance [52,53]. Normally, binning requires further manual inspection and curation as well as the downstream analysis of recovered MAGs [52].

Initial works on recovery of MAGs from metagenomes focused on the description of a few genomes per environment, as in the acidic biofilm described above [61], or as in the first study using a binning approach similar to the one described here, which allowed the recovery of 31 MAGs from activated sludge samples [1]. However, MAG recovery is currently used on a wider scale and publications describing hundreds or even thousands of MAGs from a given environment are becoming more common [4,9,57,60]. In parallel, the public availability of the bioinformatic tools needed for MAG recovery, together with the lowering of the costs of next generation sequencing (NGS), are democratizing this approach. Thus, the approach is moving from the global tool developers' hands [53] to that of end-users, who can be more concerned about the questions than about the methods used to answer them. Therefore, it seems likely that this approach will become widely used, much like the use of high throughput sequencing to analyze 16S rRNA genes (or other phylogenetic marker genes) PCR amplified from community DNA [24]. This "metabarcoding" is a good example of how technologies which were once only accessible to a few laboratories have become almost routinely used in many labs worldwide.

Here, we present some examples of the use of the binning approach to study hypersaline environments. The aim is to illustrate some issues we have encountered in retrieving ecologically meaningful MAGs from systems that we know well. This is a sort of benchmarking of MAG recovery in samples characterized by a high cell density (around 10^7 cells/ml), a low diversity and a high strain heterogeneity (or microdiversity) [12,39,42]. This high microdiversity is especially relevant since it may hamper good quality assembly of the metagenomic reads [1,9,53,59,60] which, as mentioned above, is one of the key steps in the recovery of high quality MAGs. The final goal of this work is to explore the binning approach as end-users of the available tools and make suggestions that we hope will help to make automatic pipelines more accessible for microbial ecologists.

We have analyzed a range of situations that go from a perfect recovery of specific MAGs with a straightforward identification to the failure to recover highly abundant species. This last case offers, however, many interesting points to address that we think will be especially helpful to discuss (and maybe push forward) the limits of the techniques and to understand the structure of microbial communities in nature.

Materials and methods

Sampling and sequencing

Different water samples were collected from brines "Campos Saltern" (CS; Mallorca, Spain), "Valle Salado de Añana" (AV; Basque Country, Spain) and "Bras del Port" (CR30; Alicante, Spain). One liter from the brines in CS and CR30 and 4 l from AV were first centrifuged at 30,000 g during 40 min and later filtered through 0.22 microns hydrophilic Durapore membranes (Millipore®). DNA was extracted using the Ultraclean soil DNA isolation kit (Mobio®) following the manufacturer's instructions and purity was checked by agarose gels. DNA was sequenced using the Illumina platform, Miseq 250 bp (pair-end) Nextera XT. Table 1 shows details for each of the three metagenome collections constructed (CS, CR30 and AV).

Cell counts and FISH

Cells were recovered from Santa Pola saltern CR30 crystallizer as follows: one milliliter of the sample was fixed overnight at 4 °C with 7% formaldehyde (final concentration) and diluted to 10 ml with PBS buffer (1×). 500 μl were then filtered through 0.2 μm GTTP filters (Millipore) to collect the cells. Fluorescence *in situ* hybridizations were performed at 46 °C for 2 h in the hybridization buffer (0.9 M NaCl, 20 mM Tris–HCl, 0.01% SDS and 35% formamide) with 5 ng/μl of *Archaea* specific probe ARCH915 [3]. After hybridization and DAPI staining [5], samples were examined with an Axioplan microscope (Leica DMLA).

Assembly and binning

The pipeline followed in this work is summarized in Fig. 1. The sequenced reads from each collection were quality trimmed by PRINSEQ and assembled using two different assemblers, metaSPADES [6] and IDBA 1.1.1 [43]. Different K-mers were assayed for both assemblers. The IDBA "hybrid-option" was also tested. Finally, the sequences obtained with the assembler IDBA using the pre_correction option were chosen as it offered the largest number of big-contigs (>1 kb) and the highest N50 parameter. Details of each of the IDBA assemblies are in Table 1. Unless otherwise indicated, only contigs larger than 1 kb were considered for the binning analysis. For each of the metagenomes, automatic binning were performed using MaxBin program (v.2.2.1) [65], based on the oligonucleotide frequencies, GC content, and the differential coverage. Possible bin contamination and strain heterogeneity was estimated with CheckM v.0.9.7 [40] and MiGA platform [50] ("*NCBI prok*" mode). In parallel, ORFs contained in the contigs were detected using Prodigal [22] and their taxonomy and functional annotation was performed comparing the predicted protein sequences against the NCBI-nr database using diamond [10]. Megan (v.6.6.7) [21] was used to interactively analyze the taxonomy of each of the bins. The CR30 contigs (>1 kb) were also

Table 1
Metagenomes used in this study.

Metagenome ID	Saltern	% Salinity	No. contigs	Metagenome size (bp/reads)	%GC	No. contigs >1000 bp	N50	Max. contig length (bp)
CS_4h	Campos, Mallorca (Spain)	32.4	223,802	195,589,618/1,053,686	56.4	43,284	1098	111,625
CS_10h	Campos, Mallorca (Spain)	32.4	6681	11,528,608/793,286	58.5	2721	2523	126,394
AV	Valle Salado de Añana, Araba (Spain)	23	24,950	34,657,252 / 12,384,730	60.0	9629	1812	64,386
CR30	Bras del Port, Alicante (Spain)	34.4	198,186	156,993,243 / 14,993,022	55.9	20,414	466	42,846

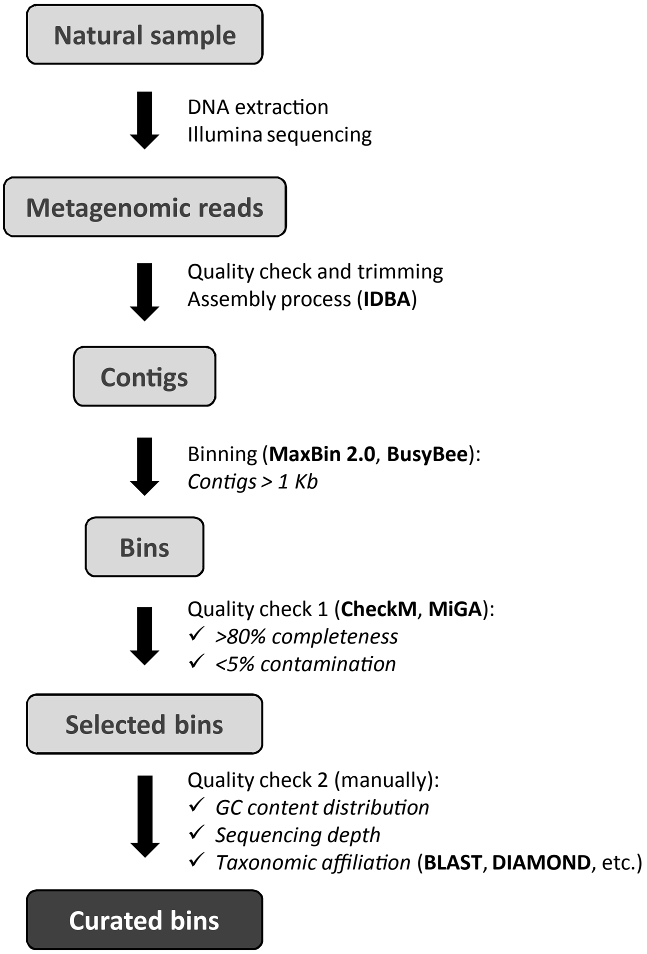


Fig. 1. Pipeline used in this work.

uploaded to Busybee program [26] to compare the results with those obtained with MaxBin (parameters used: minimum sequence length 500, minimum sequence length for border points: 1000, minimum sequence length for cluster points: 2000, K-mer length: 5, probability: 0, minimum points in neighborhood: 30, transformation: standard, compression: 0 and seed: 0). The bins obtained with MaxBin2 and Busybee were submitted to DasTool [54] to try to obtain better quality MAGs. Although DasTool is not itself an *ab-initio* binner, it has been reported to enhance the quality of the bins starting from previously bins obtained with other programs. Sub-binnings of the CR30 metagenome were performed using *FastA.subsampling.pl* from the Enveomics package [49] with the −r 10 option and using fractions of the library to be sampled (as percentage, 0.5, 1, 2.5, 5, 10, 20, 25, 30, 40, 50, 60 and 80%). Completeness of 2140 genomes (complete genomes downloaded from the Genbank, March 2018) was estimated by the presence of 35, 112 or 53 essential/core genes [1,35,44] using HMMER. An evalue<10^{-5} and an alignment coverage >65% were used as *cut-off* to define homolog genes. In parallel, evaluation with CheckM [40] was also performed.

Other bioinformatic analyses

Metagenome diversity index of the raw sequences for each collection was estimated by *nonpareil curves* [51]. 16S rRNA genes were extracted from each collection using RNAscan software [27] and OTUs were calculated with *pick_de_novo_otus.py* from Qiime [11]. Shannon diversity index was calculated with the *alpha_diversity.py* (-m shannon) tool from Qiime. The presence of tRNAs in each of the bin sequences were searched using t-RNAscan program [29]. GC content of the contigs was calculated with GEECEE tool from the EMBOSS package [47]. Abundance and sequencing depth for each of the contigs was calculated with BLASTN comparisons [2] (*cut-off*: >70% coverage, −evalue 0.1 and besthit). The *BlastTab.seqdepth.pl* script from the Enveomics package [49] was used to calculate the sequencing depth of the contigs. Recruitment plots were drawn with R (enveomics.R library).

"In-silico" *binning*

ART program was used to construct the "*in-silico*" short-reads metagenomes (150 pb, pair-end) [19]. The art-illumina simulation tool generated synthetic sequencing reads based on a reference genome (parameters used: -ss HS25 -sam -p -l 150 -f 10 -m 200 -s X, where X is the coverage). For each reference genome chosen, a metagenome was constructed to simulate a coverage of 10X (in Supplementary Figs. S1 and S2 in the online version at DOI: 10.1016/j.syapm.2018.11.001) or simulate the proportion found in CR30 (in Supplementary Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.11.001). For this purpose, the short-reads of the CR30 metagenome were previously plotted against each of the reference genomes through BLASTn comparisons. The number of "reads per kilobase of genome per gigabase of the collection" (rpkg) were calculated and used as the coverage in ART. Short-read metagenomes for each reference genome were then joined as desired and introduced in the MaxBin2 program as one single pool of sequences. To simulate the assembled contigs to be binned, each of the reference genomes were fragmented in non-overlapping pieces of 1, 5 or 10 kb. For the experiment shown in Supplementary Fig. S1 in the online version at DOI: 10.1016/j.syapm.2018.11.001, seven genomes of *Salinibacter* plus seven of *Halorubrum* were mixed in the same proportion ("equally mixed"). Also, equal amounts of sequences were mixed from 1, 2, 3, 4 and 5 genomes of *Haloquadratum* for the experiment shown in Supplementary Fig. S2 in the online version at DOI: 10.1016/j.syapm.2018.11.001. In the experiment shown in Supplementary Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.11.001, a total of 51 genomes were used mimicking the species proportion found in CR30. The reference genomes used in this analysis were: *Haloquadratum walsbyi* DSM 16790, *Hqr. walsbyi* C23, *Hqr. walsbyi* J07HQX50, *Hqr. walsbyi* J07HQW1, *Hqr. walsbyi* J07HQW2, *Salinibacter altiplanensis* AN4, *S. altiplanensis* AN15, *S. altiplanensis* LL19, *Salinibacter ruber* DSM13855, *S. ruber* SP273, *S. ruber* UBA968, *Salinibacter* sp. 10B, *Halorubrum* *ezzemoulense* Ec15, *Hrr. ezzemoulense* Fb21, *Hrr. ezzemoulense* G37, *Hrr. ezzemoulense* Ga2p, *Hrr. ezzemoulense* Ga36, *Hrr. ezzemoulense* LD3, *Hrr. ezzemoulense* LG1, *Hrr. ezzemoulense* DSM17463, *Halorubrum halophilum*, *Halorubrum hochstenium*, *Halorubrum kocurii* JCM14978, *Halorubrum lacusprofundi* DL18, *Hrr. lacusprofundi* HLS1, *Halorubrum lipolyticum* DSM21995, *Halorubrum litoreum* JCM13561, *Halorubrum saccharovorum* H3, *Hrr. saccharovorum* DSM1137, *Halorubrum aethiopicum* SAH-A6, *Halorubrum aidingense* JCM13560, *Halorubrum arcis* JCM13916, *Halorubrum californiensis* DSM19288, *Halorubrum coriense* DSM10284, *Halorubrum distributum* JCM9100 and JCM10118, *Halorubrum tebenquichense* DSM14210, *Halorubrum terrestre* JCM10247, *Halorubrum tropicale*, *Halorubrum* sp. AJ67, *Halorubrum* sp. BV1_EP28, *Halorubrum* sp. Ea8, *Halorubrum* sp. Eb13, *Halorubrum* sp. Hd13, *Halorubrum* sp. Ib24, *Halorubrum* sp. J07HR59, *Halorubrum* sp. Br10_E2g5, *Halorubrum* sp. SD612, *Halorubrum* sp. SD626R, *Halorubrum* sp.SD683 and *Halorubrum* sp. T3. In order to analyze the hypervariable regions of *Salinibacter*, 14 different genomes (Supplementary Table S2) were fragmented in 5 Kb pieces and were submitted to Maxbin2 using the CS_4h metagenome collection. ART program was used to construct the *"in-silico"* short-reads metagenomes (150 pb, pair-end) [19]. The art-illumina simulation tool generated synthetic sequencing reads based on a reference genome (parameters used: -ss HS25 -sam -p -l 150 -f 10 -m 200 -s X, where X is the coverage). For each reference genome chosen, a metagenome was constructed to simulate a coverage of 10X (in Supplementary Figs. S1 and S2) or simulate the proportion found in CR30 (in Supplementary Fig. S3). For this purpose, the short-reads of the CR30 metagenome were previously plotted against each of the reference genomes through BLASTn comparisons. The number of "reads per kilobase of genome per gigabase of the collection" (rpkg) were calculated and used as the coverage in ART. Short-read metagenomes for each reference genome were then joined as desired and introduced in the MaxBin2 program as one single pool of sequences. To simulate the assembled contigs to be binned, each of the reference genomes were fragmented in non-overlapping pieces of 1, 5 or 10 kb. For the experiment shown in Supplementary Fig. S1, seven genomes of *Salinibacter* plus seven of *Halorubrum* were mixed in the same proportion ("equally mixed"). Also, equal amounts of sequences were mixed from 1, 2, 3, 4 and 5 genomes of *Haloquadratum* for the experiment shown in Supplementary Fig. S2. In the experiment shown in Supplementary Fig. S3, a total of 51 genomes were used mimicking the species proportion found in CR30. The reference genomes used in this analysis were: *Haloquadratum walsbyi* DSM 16790, *Hqr. walsbyi* C23, *Hqr. walsbyi* J07HQX50, *Hqr. walsbyi* J07HQW1, *Hqr. walsbyi* J07HQW2, *Salinibacter altiplanensis* AN4, *S. altiplanensis* AN15, *S. altiplanensis* LL19, *Salinibacter ruber* DSM13855, *S. ruber* SP273, *S. ruber* UBA968, *Salinibacter* sp. 10B, *Halorubrum ezzemoulense* Ec15, *Hrr. ezzemoulense* Fb21, *Hrr. ezzemoulense* G37, *Hrr. ezzemoulense* Ga2p, *Hrr. ezzemoulense* Ga36, *Hrr. ezzemoulense* LD3, *Hrr. ezzemoulense* LG1, *Hrr. ezzemoulense* DSM17463, *Halorubrum halophilum*, *Halorubrum hochstenium*, *Halorubrum kocurii* JCM14978, *Halorubrum lacusprofundi* DL18, *Hrr. lacusprofundi* HLS1, *Halorubrum lipolyticum* DSM21995, *Halorubrum litoreum* JCM13561, *Halorubrum saccharovorum* H3, *Hrr. saccharovorum* DSM1137, *Halorubrum aethiopicum* SAH-A6, *Halorubrum aidingense* JCM13560, *Halorubrum arcis* JCM13916, *Halorubrum californiensis* DSM19288, *Halorubrum coriense* DSM10284, *Halorubrum distributum* JCM9100 and JCM10118, *Halorubrum tebenquichense* DSM14210, *Halorubrum terrestre* JCM10247, *Halorubrum tropicale*, *Halorubrum* sp. AJ67, *Halorubrum* sp. BV1_EP28, *Halorubrum* sp. Ea8, *Halorubrum* sp. Eb13, *Halorubrum* sp. Hd13, *Halorubrum* sp. Ib24, *Halorubrum* sp. J07HR59, *Halorubrum* sp. Br10_E2g5, *Halorubrum* sp. SD612, *Halorubrum* sp. SD626R, *Halorubrum* sp.SD683 and *Halorubrum* sp. T3. In order to analyze the hypervariable regions of *Salinibacter*, 14 different genomes (Supplementary Table S2) were fragmented in 5 Kb pieces and were submitted to Maxbin2 using the CS_4h metagenome collection.

Nucleotide sequence accession number

The raw sequences of the metagenomics datasets have been deposited in the repository data of Discovery Environment in the Cyverse structure platform and are publicly available through the following link:

https://de.cyverse.org/dl/d/36DABF23-D12E-4CA9-B0A3-F2529C204E3C/Ramos-Barbero.zip

Results and discussion

We have analyzed the metagenomes of three hypersaline environments whose characteristics are shown in Table 1. As a general rule, this kind of system is dominated by *Euryarchaeota* followed by different groups of *Bacteria*, with *Bacteroidetes* as the most frequently retrieved phylum [17]. The recently described Nanohaloarchaeota [15,35] can be also numerically relevant members of the community [34]; however, their ubiquity and abundance is very variable among different saline environments.

Bins (Table 2) have been chosen to cover different scenarios: (i) a good quality bin from a species highly abundant in the environment ("the good", CS_4h BIN1); (ii) an intermediate quality bin with incongruences ("the ugly", CS_10h BIN1) that can be easily solved, and (iii) a low quality bin ("the bad", CR30 BIN4) to illustrate the failure to recover a very abundant microbial genome. This last phenomenon, which has been widely discussed [15,36,48], can be considered as the "great metagenomics anomaly", by comparison to the "great plate count anomaly" or the failure of culturing to retrieve the most abundant microbes in nature [56]. In addition, several mock metagenomes were also analyzed with the binning pipeline to address the issues derived from the analysis of the natural samples.

The Good and the Ugly: *Salinibacter ruber* added to a pond

The first system under analysis was a crystallizer pond in Campos salterns (Mallorca, Spain) in August 2014. As part of a set of experiments to be described elsewhere, a pure culture of the extremely halophilic bacteroidetes *Salinibacter ruber* strain M8 (with a genomic GC content of 66.12%) was added to the pond to a final concentration of around 15% of the total cells. Before the addition, *S. ruber* was present in the pond but the specific strain M8 was below the detection limit, as indicated by the failure to recover the strain specific hypervariable regions [42] from the metagenome. The community was monitored by temporal sampling and metagenomic analysis before and after the addition of the strain. From these analyses, we have selected two MAGs, obtained immediately after the addition of M8 (CS_4h) and 6 h later (CS_10h).

Right after the addition of *S. ruber* M8 (CS_4h), an estimate based on the abundance of M8 specific islands in the pond indicated that the strain accounted for 13.7% of the metagenomic reads. By using the bioinformatic tools described in the material and methods Section, *S. ruber* M8 reads were recovered by binning from the metagenome CS_4h (Fig. 2a). The MAG meets all the requirements, with high completeness (95.34%) and low contamination (4.35%), but it still includes a contig containing a 16S rRNA gene of the archaeon *Halorubrum*. Besides, MyTaxa scan tool (implemented in MiGA) suggested that there were some other contigs, 10%, that were not *S. ruber*. However, BLASTp of the ORFs in the bin indicated that the contamination with microbes other than *Salinibacter* was low (3.4%). In this case, the "contamination" with the external 16S rRNA gene containing contig was easy to detect since (i) MiGA retrieved

Table 2
General characteristics of bins obtained from hypersaline samples.

Bin ID	No. contigs	Sequence size (bp)	N50 (bp)	%GC	No. essential genes	Predicted proteins	% of metagenome recruited	Closest relative in database (MiGA NCBI PROK)	%AAI
CS_4h BIN001	285	3,903,183	30,751	65.8	106/111	3490	55.97	*Salinibacter ruber* M8 NC 014032	96.34
CS_10h BIN001	655	4,257,022	15,325	65.8	106/111	4141	44.17	*Salinibacter ruber* M8 NC 014032	99.77
AV BIN021	1016	2,926,965	3761	55.7	87/111	3605	1.00	*Thiohalobacter thiocyanaticus* NZ AP018052	48.68
CR30 BIN004	4385	8,718,229	2083	62.2	92/111	11,424	3.86	*Salinibacter ruber* DSM 13855 NC 007677	64.84

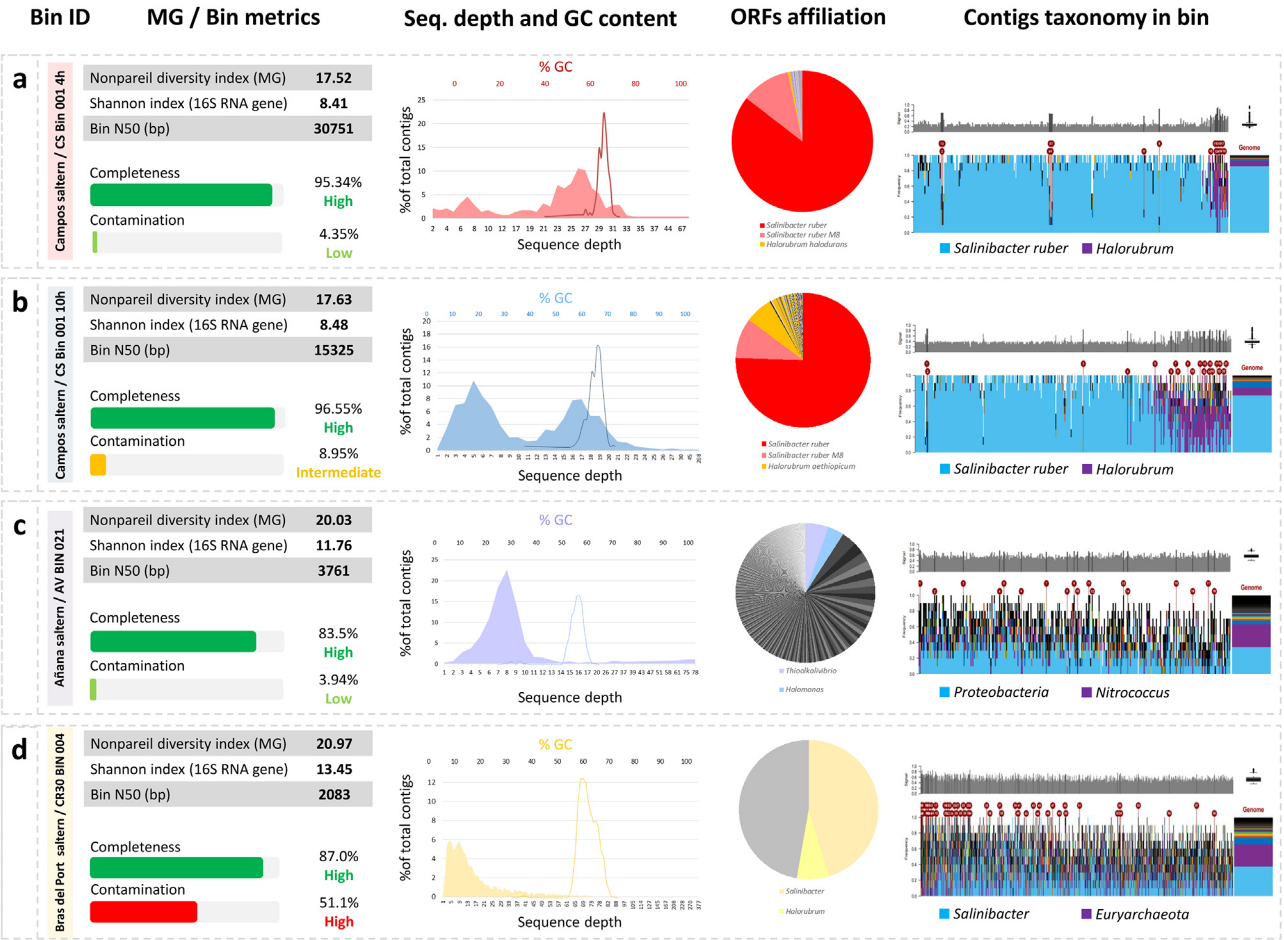


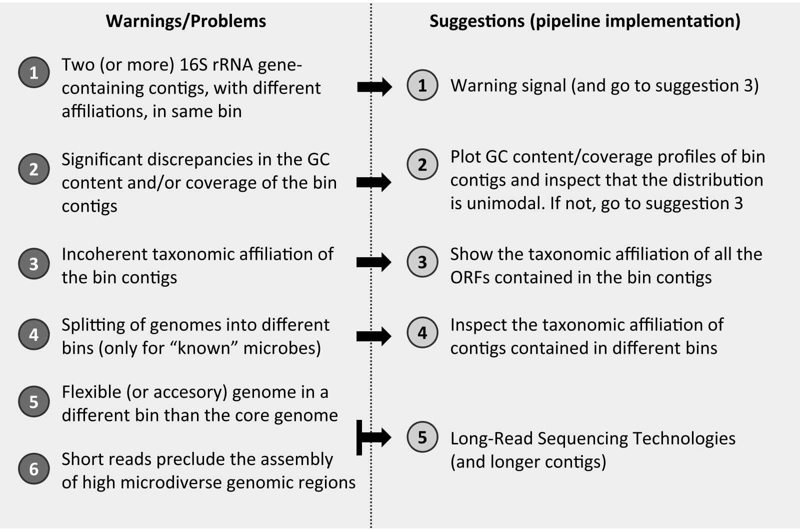
Fig. 2. Bins from extremely halophilic microorganisms obtained from different salterns in Spain: Campos salterns in panels (a) (CS_4h) and (b) (CS_10h), (c) "Valle Salado de Añana" and (d) Bras del Port salterns (CR30 crystallizer). For each panel, from left to right, the following information is shown: (i) bin ID, (ii) metagenome (MG) metrics (nonpareil diversity indexes and Shannon indexes inferred from 16S rRNA gene sequences), and bin metrics (N50 for the obtained bins, bin completeness and degree of contamination), (iii) sequencing depth (filled area) and GC content (solid line) distribution of contigs within the bins, and (iv) taxonomic affiliation of detected ORFs and contigs. N50 values and taxonomic affiliation of contigs were obtained from MiGA platform. Percentages of completeness and contamination were extracted from CheckM.

two 16S rRNA genes from different domains, and (ii) the system under study was dominated by an organism whose genome was previously known. However, this is not the case with most ecological studies where the detection of contaminating contigs is likely more challenging. Therefore, it would be helpful (see Box 1) to get some warning signs when 16S rRNA genes incongruent with the taxonomic environment of the binned contigs are detected.

The binning process was repeated with the metagenome retrieved 6 h after the addition of the strain *S. ruber* M8 (CS_10h), when it accounted for 7.8% of the metagenomic reads. A MAG with a fair quality corresponding to *S. ruber* was again retrieved, with a sharp GC content profile (Fig. 2b) and including again an *Halorubrum* 16S rRNA gene, in addition to that of *S. ruber*. In this case, however, the coverage plot differs markedly from the *S. ruber* bin recovered from the CS_4h sample, taken only 6 h before from the same pond (Fig. 2a). BLASTp analyses indicated that 96% of the genes of the high coverage fraction had hits (identity above 99%) to *S. ruber* M8 genome; however, only 11% of the genes in the low coverage peak had their best match with this genome (with an average identity of 93.8%). A more detailed inspection indicated that the low coverage peak was heavily contaminated with the archaeon *Halorubrum* spp. sequences, which have a high GC content (67.4%) similar to that of *Salinibacter* (Table 2). *Halorubrum* spp. are extremely halophilic euryarchaea frequently present in hypersaline environments, which generally harbor a high intraspecific diversity due to frequent recombination [39].

Finally, the manual curation of the bin CS_10h, by removing the low coverage contigs, allowed the recovery of *S. ruber* M8 genome (99.98% ANI with the reference genome) with a completeness of 95.06% and a contamination of 1.13%, thus improving considerably the quality of the initial CS_10h bin shown in Fig. 2b. Thus, manual inspection and, if needed, curation of different coverage contigs

Box 1: Detected warnings after binning process and proposal to be implemented in automatic binning pipelines.



Warnings/Problems

1. Two (or more) 16S rRNA gene-containing contigs, with different affiliations, in same bin
2. Significant discrepancies in the GC content and/or coverage of the bin contigs
3. Incoherent taxonomic affiliation of the bin contigs
4. Splitting of genomes into different bins (only for “known” microbes)
5. Flexible (or accesory) genome in a different bin than the core genome
6. Short reads preclude the assembly of high microdiverse genomic regions

Suggestions (pipeline implementation)

1. Warning signal (and go to suggestion 3)
2. Plot GC content/coverage profiles of bin contigs and inspect that the distribution is unimodal. If not, go to suggestion 3
3. Show the taxonomic affiliation of all the ORFs contained in the bin contigs
4. Inspect the taxonomic affiliation of contigs contained in different bins
5. Long-Read Sequencing Technologies (and longer contigs)

within a bin can help to recover ecologically meaningful genomes. For this purpose, the automatic implementation in the pipeline of the display of a coverage plot would certainly help researchers to evaluate the quality of their results (Box 1).

Here, it is worth mentioning that, in spite of the high AAI of the manually cured bins in Fig. 2a-b with the genome of the strain M8, strictly speaking, we cannot state that the strain was recovered since the hypervariable regions (belonging to the species accessory genome [32]), were only partially assembled. For instance, among the CS_4h contigs, the hypervariable region 2 (HVR2) was not assembled while in the CS_10 h collection, a partial region with only 83% identity to HVR2 was identified (Supplementary Fig. S4d in the online version at DOI: 10.1016/j.syapm.2018.11.001). Obviously, this is expected since binning is based on tetranucleotide frequencies and coverages, which are also ways of distinguishing accessory from core genome [42]. As stated in Ref. [25], bins from metagenomes represent consensus population genomes, which may lack the strain-specific accessory genome. In this context it is interesting to recall that completeness refers mainly to the “core-genome”, not taking into account that the flexible (or “accessory”) genome may represent about 20% of a complete genome [12,13,58]. These areas of the genomes may confer slightly different phenotypes among microbial lineages within the same microbial species sharing the same habitat. Normally, the low recruitment (lower coverage) of these genome regions impede their assembly from metagenomes and therefore MAGs lack these interesting microbial regions (see for example, Ref. [23]). In any case, their different GC content or oligonucleotide frequency is likely to place them in a different bin than the rest of the genome, as shown below.

To explore how strain-specific regions would be binned in the case of *Salinibacter*, we performed an *in silico* experiment where all the available *Salinibacter* genomes were fragmented in 5 kb pieces and submitted to binning using the CS_4h metagenome to mimic the real coverage of *Salinibacter* (Supplementary Fig. S4e in the online version at DOI: 10.1016/j.syapm.2018.11.001). Only strains *S. ruber* 10B and M8 showed more than 90% of the genome classified in bins (Supplementary Table S2 in the online version at DOI: 10.1016/j.syapm.2018.11.001). Plotting of *S. ruber* M8 bins along the reference genome showed that the two HRVs clustered in a bin (BIN1) different from those harboring the rest of the genome (BIN2 and 3). In the same way, when the *Hqr. walsbyi* HSBQ001 genome was submitted to a similar “artificial” binning experiment (see below), the genomic islands were found in a different bin than the rest of the genome (Supplementary Fig. S2a in the online version at DOI: 10.1016/j.syapm.2018.11.001). This is intriguing since the binning together of the different hypervariable regions indicates that they share some genomic traits which in turn could be pointing to some kind of co-evolution. With the available data, this remains an open question.

The binning of the so-called *S. ruber* conserved region, CR, is also worth mentioning. This is a 376 kb region with a high degree of sequence conservation (99.5% between strains M8 and M31, with no nonsynonymous mutations) which contains the “hypersalinity island” harboring a cluster of 19 genes coding for transporters of crucial importance to life in hypersaline systems [33]. Unexpectedly, fragments of this conserved region were found dispersed among 12 bins or were not classified. This heterogeneous distribution among so many different bins may be explained by the mosaicism found in this area of the core genome. Many of the genes coded in this region could have been acquired by lateral gene transfer from other members of the microbial community [33] and thus could have different coverages in the metagenome. However, these transfers, essential for life in high salt systems, are likely to be old enough to have allowed the homogenization of codon adaptation indexes and GC with the rest of the core genome.

In order to further explore the performance of the binning tools with microbes with very similar GC contents, we analyzed several mock metagenomes composed of mixtures of 14 genomes from three different species of *Halorubrum* and *Salinibacter* with an equal coverage of 10X. Contigs of 1, 5 and 10 kb were obtained from those genomes and binned in independent analysis (see methods) (Supplementary Fig. S1 in the online version at DOI: 10.1016/j.syapm.2018.11.001). Although, as expected, the number of retrieved bins decreased for longer contigs, in no case was a bin containing a single species nor a bin above the quality threshold recovered. However, for contigs above 5 kb, only species from the same genus were binned together. These results highlight two important issues: the relevance of assembly quality and thus of contig length (widely discussed previously [1,9,53,60,61]) and the challenges posed by the analysis of communities composed of members of similar GC contents and abundances, which is a common scenario in hypersaline environments [15].

Another aspect that has been pointed out as a way of improving MAG recovery is to co-assemble metagenomes from different samples containing identical target community members [1]. As shown above, this has not been necessary in the above-mentioned examples (in fact, sub-sampling can be a way of improving MAG detection with our samples, as discussed below) although this is not needed in all the hypersaline environments we have analyzed, as illustrated by the MAG recovered from Añana salterns, shown in Fig. 2c.

Añana salterns are located in inland Northern Spain and are used to obtain salt from underground brine that emerges at different points and is stored in wells before feeding the crystallizer ponds. Overall, although microbial communities in these salterns harbor a remarkable diversity, they are not dominated by a few species as in most coastal salterns studies, and harbor lower concentrations of cells and viruses (Ramos-Barbero *et al.*; unpublished results). We sampled the spring and three wells and analyzed their corresponding metagenomes. In this case, the co-assembly of the four metagenomes was necessary to retrieve MAGs meeting the quality criteria. One example is shown in Fig. 2c. This MAG contains an unclassified 16S rRNA gene sequence based on RDP and SILVA platforms, and therefore its taxonomic affiliation poses many more challenges than in the previous examples, given that its AAI to the closest relative (Table 2) is below the threshold for assigning MAGs to genera, which is 45–65% [25].

The BAD: "The great metagenomics anomaly"

The hypersaline system used in this case study is the crystallizer pond CR30 from the Alicante salterns "Bras del Port", which has been characterized for years using a whole suit of techniques from culture to metagenomics [63]. FISH (Fig. 3a) indicated that in the analyzed sample, taken in November 2014, *Archaea* constituted 56% of the total cell counts (1.35×10^7 cells/ml). *Archaea* (see Fig. 3a) frequently displayed the square morphology characteristic of *Hqr. walsbyi*. This euryarchaeon (with a 47.9% GC in its genome [7]) dominates the microbial communities in many close to saturation systems over the world, albeit with relevant exceptions [38,62]. Metagenomic analysis of this sample showed that indeed *Hqr. walsbyi* was the most abundant species in CR30, as indicated by recruitment of the *Hqr. walsbyi* available genomes (see below). In terms of abundance of 16S rRNA gene reads in the metagenome, *Hqr. walsbyi* accounted for 48.5% of the reads followed by *Salinibacter* spp. (7.2%), and the high GC euryarchaeal *Halobellus*, *Halorubrum* and *Halonotius* (between 1.7 and 0.5%). These data are in good agreement with previous reports on the community of CR30 [15].

However, when the metagenome was launched to the binning pipeline, 11 bins (Fig. 3b) were obtained, none of which met the quality criteria for completeness and contamination. The "best" of them, which still has a high contamination (51.1%), is shown in Fig. 2d. Furthermore, *Hqr. walsbyi* genome was split into 4 different bins. The binning was repeated with the program BusyBee [26] and the recently published DasTool approach [54]. Both of them failed to recover *Haloquadratum* bins.

This was unexpected given (i) the distinctive GC content of *Hqr. walsbyi* and (ii) its high abundance and sequencing depth in the CR30 metagenome (328.6X and 223.35X, for strains C23 and DSM 16790, isolated from an Australian saltern and the same pond analyzed here, respectively). Again, the assembly was likely the cause, as previously shown for the SS37 metagenome obtained from the same pond, CR30, by Ghai *et al.* [15]. These authors analyzed a 454 metagenome of 309 Mb that was highly enriched in *Hqr. walsbyi* (64% of the reads, as determined by comparison with genomes from cultured strains) but the largest fragments assembled of this organism were shorter than 11 kb. This illustrates how metagenomics assemblies can fail to recover the most abundant microbes, which has been widely documented not only for bacteria and archaea [15,36,48] but also for viruses [31]. This phenomenon could be called the "great metagenomics anomaly" (much like the "great plate count anomaly"). To be more precise, it is an anomaly of short-read metagenomics, not of metagenomics itself, since longer sequencing reads allow longer contig recovery [14]. Besides technical issues, microdiversity is at the heart of the assembly problem since it hampers the assembly of heterogeneous genetic sequences and repeats [1,9,53,60].

In order to improve the recovery of *Hqr. walsbyi* MAGs from our CR30 sample, the binning process was repeated after different sub-sampling (0.5%, 1%, 2.5%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, and 80% of total reads in metagenome), as previously described [20]. The best bin corresponding to *Hqr. walsbyi* genome was recovered in subsamples containing only 1% of the metagenome, with a median of 91.3% completeness and no contamination (n = 10; Supplementary Fig. S5d in the online version at DOI: 10.1016/j.syapm.2018.11.001). The best of these sub-binnings (with a 100% completeness) contained 892 contigs. A total of 797 (89.34%) had a match with some of the *Haloquadratum* genomes published with a coverage > 70%. The metagenomics reads from this bin were recovered and recruited against *Hqr. walsbyi* DSM 16790 genome (Supplementary Fig. S5 in the online version at DOI: 10.1016/j.syapm.2018.11.001). As expected, the recovered bin did not include the hypervariable regions of the species [12,28], but also failed to recover core genome regions, as shown in Supplementary Fig. S5 in the online version at DOI: 10.1016/j.syapm.2018.11.001. Besides, the calculated size of the bin (2.7 Mb) was still far from the 3.1–3.2 Mb of the reference genomes. Obviously, here the amount of "unrecovered" genome can be ascertained but would go unnoticed for previously unknown genomes.

Along these lines, it is worth mentioning that some published complete prokaryotic genomes may present completeness values <80% (calculated by four different approximations, see methods) when, due to the high divergence, they do not provide reliable matches with some of the universal protein-coding genes (Supplementary Fig. S6 in the online version at DOI: 10.1016/j.syapm.2018.11.001). As shown in Supplementary Table S1 in the online version at DOI: 10.1016/j.syapm.2018.11.001, most of them are small genomes from symbionts (for example, genomes of *Hodgkinia, Portiera, Sulcia* or *Carsonella* spp.) or not very well described taxa such as *Nanoarchaeota* (*Nanoarchaeum equitans* and *Nanopusillus acidilobi*) or *Candidatus* Wolfebacteria. The lack of complete reference genomes might most directly influence completion estimates.

In parallel to the analysis of *Hqr. walsbyii* in the CR30 metagenome, the Mallorca metagenome used for recovering the MAG shown in Fig. 2a ("the good") was searched for *Haloquadratum* related bins. In this case, a bin of a fair quality was readily recovered. The reason for this discrepancy can be found in the different levels of intraspecific diversity of *Hqr. walsbyi* in the metagenomes of CR30 and CS, as shown in Fig. 4a-b. This microdiversity is inversely correlated to the average of the identity found (96 and 99%) in the distribution identity percentages graphs (Fig. 4d). In both systems, the abundance of *Hqr. walsbyi* was high (35.37% and 21.2%, as percentage of total nucleotides recruited by the reference genome of the cultured strain DSM 16790, Fig. 4c), although the population in Campos salterns was very homogeneous in contrast to that in CR30 (Fig. 4d). In addition, the sequencing depth of the genome was very different in both systems, with only 4.7X in Campos salterns, as opposed to 223X in CR30. However, a sequencing depth of 50X has been recommended as a threshold to successfully recover MAGs [1], which is in apparent contradiction to our results.

To further explore the effect of the intraspecific diversity of *Hqr. walsbyi* (that can be estimated on the performance of the binning tools used here), we constructed mock metagenomes containing one to five strains of this archaeon, with ANI values ranging from 98.6% to 70.70%. Results are shown in Supplementary Fig. S2 in the online version at DOI: 10.1016/j.syapm.2018.11.001. When the binning was performed with one single genome, two bins were retrieved that were enriched, respectively, in the core genome and

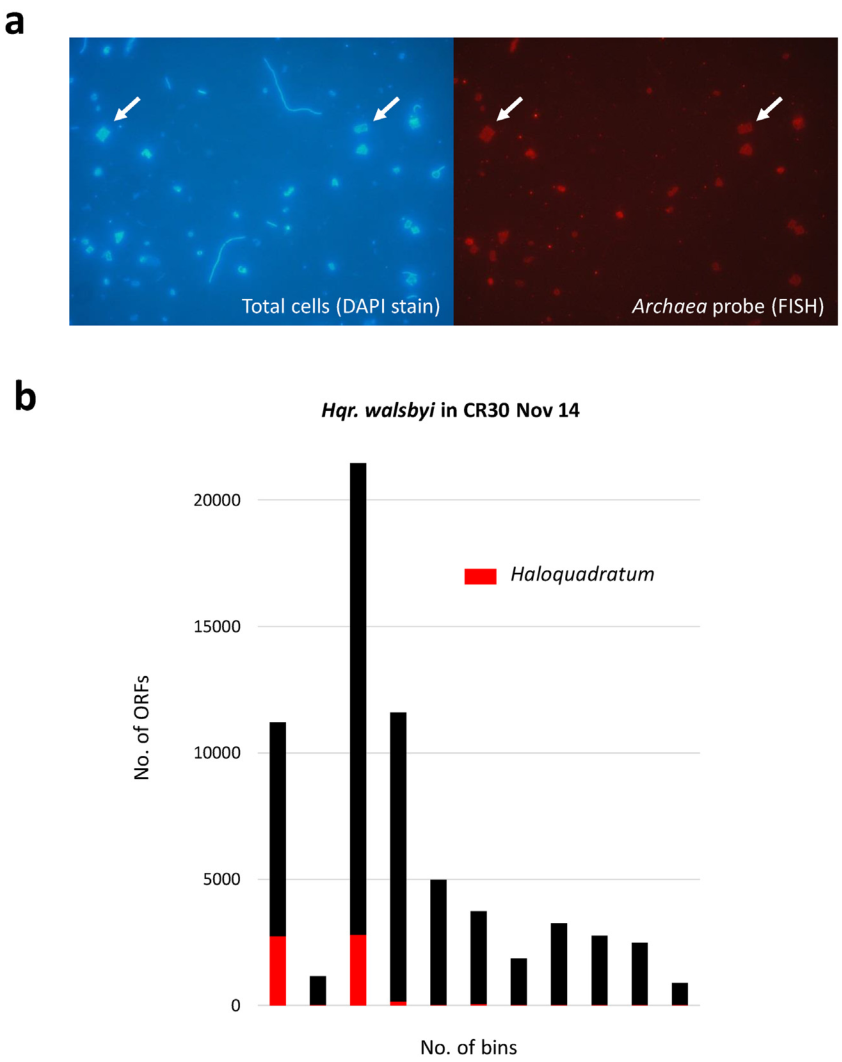


Fig. 3. Presence of *Haloquadratum walsbyi* in a hypersaline water sample from CR30 crystallizer (Bras del Port salterns). (a) Total cells by DAPI staining (left) and fluorescence *in situ* hybridization (FISH) of *Archaea* with probe ARCH915 (right). Arrows point to *Hqr. walsbyi* cells. (b) Binning of CR30 metagenome. *Hqr. walsbyi* genome fragments were split and distributed in four mixed bins.

the islands (i.e. accessory genome) of *Hqr. walsbyi,* as shown in Supplementary Fig. S2a in the online version at DOI: 10.1016/j.syapm.2018.11.001. This is to be expected since core and accessory genomes differ in their tetranucleotide frequency [30], one of the traits considered for binning.

The addition of more strains to the mock metagenome yielded an increasing number of bins. The high proportion of sequences not classified into bins (Supplementary Fig. S2b-c in the online version at DOI: 10.1016/j.syapm.2018.11.001) is especially noteworthy. This phenomenon has been previously described in a benchmarking study of metagenomics tools [53]. Finally, a complex mock metagenome mimicking the composition of a sample from CR30 [15] was analyzed. In this mock metagenome, reads corresponding to species of *Haloquadratum*, *Halorubrum* and *Salinibacter*, accounted for 71.52, 8.29 and 20.18% of the total reads, respectively. Again, a high proportion of sequences (64%) could not be classified and the genome of *Hqr. walsbyi* could not be recovered from this dataset (Supplementary Fig. S3 in the online version at DOI: 10.1016/j.syapm.2018.11.001), nor could the genome of the other components of the community. It thus seems that the particularities of each environment is a key factor determining the feasibility of recovering a given MAG from different metagenomes.

Conclusions

Our examples underline the relevance of population microdiversity in the retrieval of high quality genomes from metagenomic sets, even for very abundant microbes such as *Haloquadratum* and *Salinibacter* in hypersaline environments (as depicted in Fig. 5). Microdiversity negatively affects the quality of metagenome assembly, which is a critical step in the overall binning process. This phenomenon is likely not restricted to extreme environments but could be affecting other natural systems like SAR11 in the marine environment and freshwater low GC *Actinobacteria* [16,37]. In hypersaline systems other issues arise derived from the similar GC content and abundance of otherwise very distant microbes, like high GC *Euryarchaea* and *Bacteroidetes*, which hinder the binning of their genomes (Fig. 5).

As mentioned in the introduction, the binning approach will soon become a standard tool. In this study, we highlight critical issues in the process of binning that we have encountered when targeting well-known components of the community. This study-obviously not feasible when dealing with previously undescribed systems- has allowed us to evaluate the limits of the approach in our hands. Based on our results, we have detected some points in the process of binning that should be made more obvious or

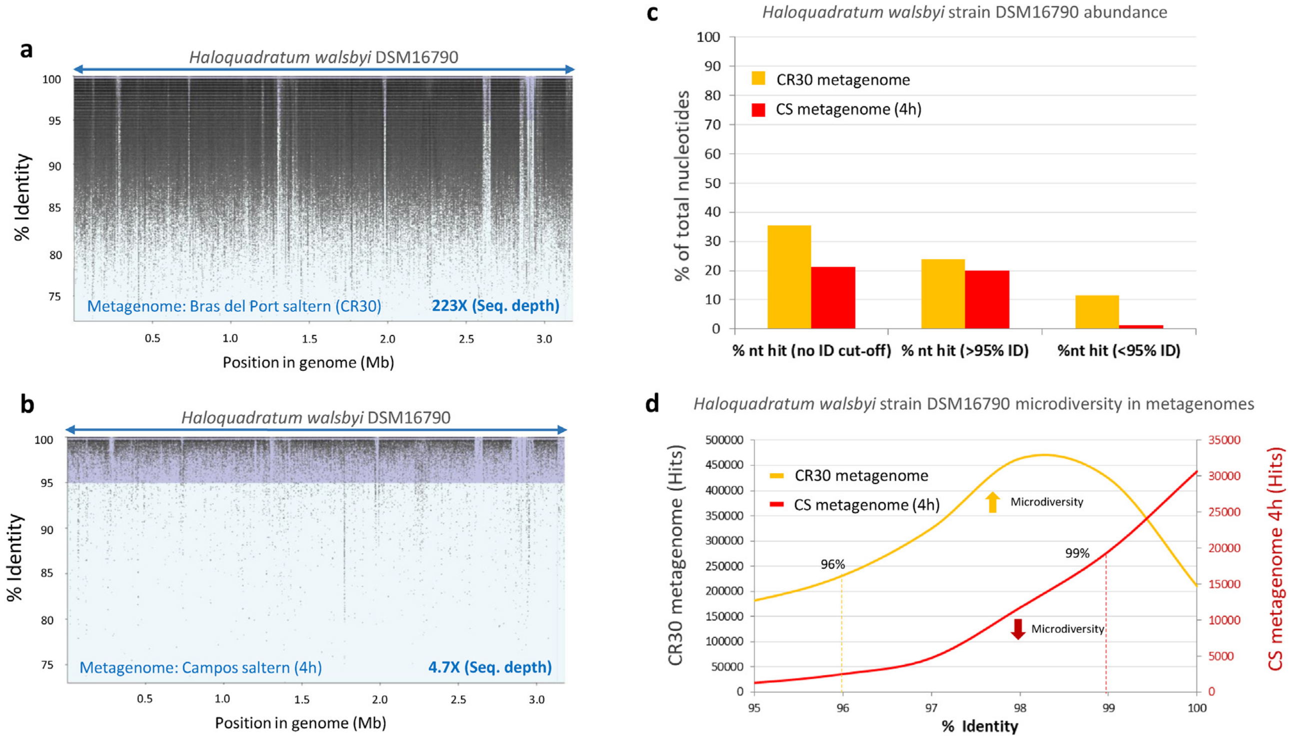


Fig. 4. Abundances and microdiversity patterns of *Haloquadratum walsbyi* (strain DSM 16790) in the metagenomes of Campos (CS_4h) and Bras del Port (CR30) salterns. Recruitment plots indicating the presence of *Hqr. walsbyi* in both systems are shown in (a) and (b) respectively, while the percentage of total recruited nucleotides, according to different identity cut-offs, is shown in (c). The degree of microdiversity within the *Hqr. walsbyi* assemblage was also analyzed in the two studied metagenomes (d). Dashed lines indicate the reads average identity mapping in the reference genomes.

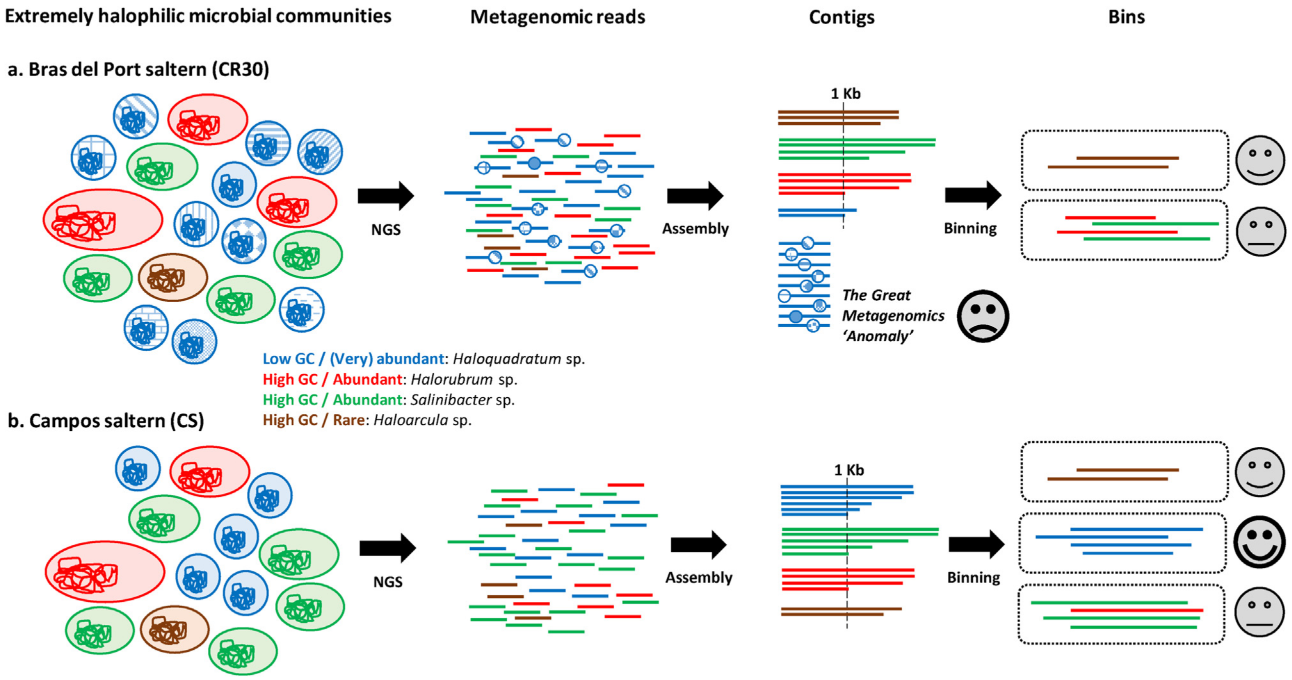

Extremely halophilic microbial communities
Metagenomic reads
Contigs
Bins
a. Bras del Port saltern (CR30)
NGS
Assembly
1 Kb
Binning
The Great Metagenomics 'Anomaly'
Low GC / (Very) abundant: Haloquadratum sp.
High GC / Abundant: Halorubrum sp.
High GC / Abundant: Salinibacter sp.
High GC / Rare: Haloarcula sp.
b. Campos saltern (CS)
NGS
Assembly
1 Kb
Binning

Fig. 5. Microbial populations may be differentiated by their abundances and GC contents. In (a), a prokaryotic community inhabiting the CR30 crystallizer from Bras del Port salterns is dominated by heterogeneous low GC-strains of *Haloquadratum*. Even though this microbe is highly abundant and provides high numbers of metagenomics reads (blue reads), its microdiversity or variation at the genomic level (represented by different symbols in blue reads) impedes the reconstruction of appropriate genomic fragments which hinders subsequent analyses. Since contigs above 1 kb in size are necessary for binning, the most abundant microbe in the system is ruled out from subsequent analyses. It is the "great metagenomics anomaly" (sad emoji). Moreover, chimeric bins related to microbes with similar GC contents and sequencing depths, although phylogenetically divergent, can also be produced, as in the case of the *Salinibacter-Halorubrum* mixed bin represented by the red-and-green bin (confused emoji). In (b), *Haloquadratum* populations from Campos salterns are neither as abundant nor as diverse (blue reads) and *Salinibacter* (in green) outnumbers *Halorubrum* populations (in red). Under these circumstances a bin containing *Haloquadratum* genomic fragments is obtained (very happy emoji) while the *Salinibacter-Halorubrum* mixed bin is more enriched in bacterial contigs (confused emoji).

implemented in the pipelines (such as the plotting of the coverage profiles in the bins, the presence of phylogenetically incongruent genes, or the splitting of known genomes into different bins, among others, see Box 1). These minor modifications could be of great help for microbial ecologists. Standardization of the binning process, although could be regarded in principle as a good solution, could also derive in systematic biases. Furthermore, different microbial assemblages may require different binning approaches, as shown above. Therefore, in order to understand their limitations and for reproducibility purposes, papers describing MAG recovery should provide a careful description of the methods and the rationale of the criteria included in the "manual curation" process. This will certainly contribute to the unveiling of a meaningful classification of the uncultured majority.

Acknowledgements

The help of the owners of the Salines d'es Trenc and Bras del Port and the Fundación Valle Salado de Añana/Añanako Gatz Harana is greatly appreciated. This research was supported by the Spanish Ministry of Economy project CLG2015_66686-C3-3 (to JA), which also included European Regional Development Fund (FEDER) funds. We thank Karen Neller for her professional English editing of the manuscript and Luis M. Rodriguez-R and two additional anonymous reviewers for they thorough revision of the manuscript.

References

[1] Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., Nielsen, P.H. (2013) Genome sequences of rare: uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat. Biotechnol. 31 (6), 533–538.

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) *Gapped BLAST* and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

[3] Amann, R.I., Ludwig, W., Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59 (1), 143–169.

[4] Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., Thomas, B.C., Singh, A., Wilkins, M.J., Karaoz, U., Brodie, E.L., Williams, K.H., Hubbard, S.S., Banfield, J.F. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat. Commum. 7, p. 13219.

[5] Anton, J., Llobet-Brossa, E., Rodriguez-Valera, F., Amann, R. (1999) Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. Environ. Microbiol. 1 (6), 517–523.

[6] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19 (5), 455–477.

[7] Bolhuis, H.H., Palm, P.P., Wende, A.A., Falb, M.M., Rampp, M.M., Rodriguez-Valera, F.F., Pfeiffer, F.F., Oesterhelt, D.D. (2006) The genome of the square archaeon "Haloquadratum walsbyi": life at the limits of water activity. BMC Genomics 7 (1), p. 169.

[8] Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., Tringe, S.G., Ivanova, N.N., Copeland, A., Clum, A., Becraft, E.D., Malmstrom, R.R., Birren, B., Podar, M., Bork, P., Weinstock, G.M., et al. (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35 (8), 725–731.

[9] Brown, C.T. (2015) Strain recovery from metagenomes. Nat. Biotechnol. 33 (10), 1041–1043.

[10] Buchfink, B., Xie, C., Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12 (1), 59–60.

[11] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Koenig, D., Ley, R.E., Lozupone, C.A., Lozupone, D., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 (5), 335–336.

[12] Cuadros-Orellana, S., Martin-Cuadrado, A.B., Legault, B., D'Auria, G., Zhaxybayeva, O., Papke, R.T., Rodriguez-Valera, F. (2007) Genomic plasticity in prokaryotes: the case of the square haloarchaeon. Isme J. 1 (3), 235–245.

[13] Fernandez-Gomez, B., Fernandez-Guerra, A., Casamayor, E.O., Gonzalez, J.M., Pedros-Alio, C., Acinas, S.G. (2012) Patterns and architecture of genomic islands in marine bacteria. BMC Genomics 13, p. 347.

[14] Frank, J.A., Pan, Y., Tooming-Klunderud, A., Eijsink, V.G., McHardy, A.C., Nederbragt, A.J., Pope, P.B. (2016) Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. Sci. Rep. 6, p. 25373.

[15] Ghai, R., Pašić, L., Fernández, A.B., Martin-Cuadrado, A.-B., Mizuno, C.M., McMahon, K.D., Papke, R.T., Stepanauskas, R., Rodriguez-Brito, B., Rohwer, F., Sánchez-Porro, C., Ventosa, A., Rodríguez-Valera, F. (2011) New abundant microbial groups in aquatic hypersaline environments. Sci. Rep., 1.

[16] Ghylin, T.W., Garcia, S.L., Moya, F., Oyserman, B.O., Schwientek, P., Forest, K.T., Mutschler, J., Dwulit-Smith, J., Chan, L.K., Martinez-Garcia, M., Sczyrba, A., Stepanauskas, R., Grossart, H.P., Woyke, T., Warnecke, F., Malmstrom, R., Bertilsson, S., McMahon, K.D. (2014) Comparative single-cell genomics reveals potential ecological niches for the freshwater acI Actinobacteria lineage. ISME J. 8 (12), 2503–2516.

[17] Gomariz, M., Martinez-Garcia, M., Santos, F., Rodriguez, F., Capella-Gutierrez, S., Gabaldon, T., Rossello-Mora, R., Meseguer, I., Anton, J. (2015) From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic Bacteroidetes generalists. ISME J. 9 (1), 16–31.

[18] Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol. Mol. Biol. Rev. 68 (4), 669–685.

[19] Huang, W., Li, L., Myers, J.R., Marth, G.T. (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28 (4), 593–594.

[20] Hug, L.A., Thomas, B.C., Sharon, I., Brown, C.T., Sharma, R., Hettich, R.L., Wilkins, M.J., Williams, K.H., Singh, A., Banfield, J.F. (2016) Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ. Microbiol. 18 (1), 159–173.

[21] Huson, D.H., Mitra, S. (2012) Introduction to the analysis of environmental sequences: metagenomics with MEGAN. Methods Mol. Biol. 856, 415–429.

[22] Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, p. 119.

[23] Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., Armbrust, E.V. (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335 (6068), 587–590.

[24] Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., Glockner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. 41 (1), p. e1.

[25] Konstantinidis, K.T., Rossello-Mora, R., Amann, R. (2017) Uncultivated microbes in need of their own taxonomy. ISME J. 11 (11), 2399–2406.

[26] Laczny, C.C., Kiefer, C., Galata, V., Fehlmann, T., Backes, C., Keller, A. (2017) BusyBee Web: metagenomic data analysis by bootstrapped supervised binning and annotation. Nucleic Acids Res. 45 (W1), W171–W179.

[27] Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35 (9), 3100–3108.

[28] Legault, B.A., Lopez-Lopez, A., Alba-Casado, J.C., Doolittle, W.F., Bolhuis, H., Rodriguez-Valera, F., Papke, T.R. (2006) Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. BMC Genomics 7 (1), p. 171.

[29] Lowe, T.M., Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25 (5), 955–964.

[30] Martin-Cuadrado, A.B., Pasic, L., Rodriguez-Valera, F. (2014) Diversity of the cell-wall associated genomic island of the archaeon "Haloquadratum walsbyi". BMC Genomics 16, p. 603.

[31] Martinez-Hernandez, F., Fornas, O., Lluesma Gomez, M., Bolduc, B., de la Cruz Pena, M.J., Martinez, J.M., Anton, J., Gasol, J.M., Rosselli, R., Rodriguez-Valera, F., Sullivan, M.B., Acinas, S.G., Martinez-Garcia, M. (2017) Single-virus genomics reveals hidden cosmopolitan and abundant viruses. Nat. Commun. 8, p. 15892.

[32] Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R. (2005) The microbial pan-genome. Curr. Opin. Genet. Dev. 15 (6), 589–594.

[33] Mongodin, E.F., Nelson, K.E., Daugherty, S., Deboy, R.T., Wister, J., Khouri, H., Weidman, J., Walsh, D.A., Papke, R.T., Sanchez Perez, G., Sharma, A.K., Nesbo, C.L., MacLeod, D., Bapteste, E., Doolittle, W.F., Charlebois, R.L., Legault, B., Rodriguez-Valera, F. (2005) The genome of "Salinibacter ruber": convergence and gene exchange among hyperhalophilic bacteria and archaea. Proc. Natl. Acad. Sci. U S A 102 (50), 18147–18152.

[34] Mora-Ruiz, M.D.R., Cifuentes, A., Font-Verdera, F., Perez-Fernandez, C., Farias, M.E., Gonzalez, B., Orfila, A., Rossello-Mora, R. (2018) Biogeographical patterns of bacterial and archaeal communities from distant hypersaline environments. Syst. Appl. Microbiol. 41 (2), 139–150.

[35] Narasingarao, P., Podell, S., Ugalde, J.A., Brochier-Armanet, C., Emerson, J.B., Brocks, J.J., Heidelberg, K.B., Banfield, J.F., Allen, E.E. (2012) De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. Isme J. 6, 81–93.

[36] Newton, R.J., Griffin, L.E., Bowles, K.M., Meile, C., Gifford, S., Givens, C.E., Howard, E.C., King, E., Oakley, C.A., Reisch, C.R., Rinta-Kanto, J.M., Sharma, S., Sun, S., Varaljay, V., Vila-Costa, M., Westrich, J.R., Moran, M.A. (2010) Genome characteristics of a generalist marine bacterial lineage. ISME J. 4 (6), 784–798.

[37] Newton, R.J., Jones, S.E., Eiler, A., McMahon, K.D., Bertilsson, S. (2011) A guide to the natural history of freshwater lake bacteria. Microbiol. Mol. Biol. Rev. 75 (1), 14–49.

[38] Oren, A. (2015) Halophilic microbial communities and their environments. Curr. Opin. Biotechnol. 33, 119–124.

[39] Papke, R.T., Koenig, J.E., Rodriguez-Valera, F., Doolittle, W.F. (2004) Frequent recombination in a saltern population of Halorubrum. Science 306 (5703), 1928–1929.

[40] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2014) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25 (7), 1043–1055.

[41] Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., Tyson, G.W. (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat. Microbiol. 2 (11), 1533–1542.

[42] Pena, A., Teeling, H., Huerta-Cepas, J., Santos, F., Yarza, P., Brito-Echeverria, J., Lucio, M., Schmitt-Kopplin, P., Meseguer, I., Schenowitz, C., Dossat, C., Barbe, V., Dopazo, J., Rossello-Mora, R., Schuler, M., Glockner, F.O., Amann, R., Gabaldon, T., Anton, J. (2010) Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting "Salinibacter ruber" strains. Isme J. 4 (7), 882–895.

[43] Peng, Y., Leung, H.C., Yiu, S.M., Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28 (11), 1420–1428.

[44] Raes, J., Korbel, J.O., Lercher, M.J., von Mering, C., Bork, P. (2007) Prediction of effective genome size in metagenomic samples. Genome Biol. 8 (1), R10.

[45] Rappe, M.S., Giovannoni, S.J. (2003) The uncultured microbial majority. Annu. Rev. Microbiol. 57, 369–394.

[46] Reddy, T.B., Thomas, A.D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E.A., Kyrpides, N.C. (2015) The Genomes OnLine Database (GOLD) v. 5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res. 43, D1099–D1106 (Database issue).

[47] Rice, P., Longden, I., Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16 (6), 276–277.

[48] Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. Nature 499 (7459), 431–437.

[49] Rodriguez-R, L.M., Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints 4, e1900v1.

[50] Rodriguez, R.L., Gunturu, S., Harvey, W.T., Rossello-Mora, R., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T. (2018) The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res. 46 (W1), W282–W288.

[51] Rodriguez, R.L., Konstantinidis, K.T. (2014) Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. Bioinformatics 30 (5), 629–635.

[52] Sangwan, N., Xia, F., Gilbert, J.A. (2016) Recovering complete and draft population genomes from metagenome datasets. Microbiome 4, p. 8.

[53] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Droge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jorgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M.Z., et al. (2017) Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat. Methods 14 (11), 1063–1071.

[54] Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F. (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat. Microbiol. 3 (7), 836–843.

[55] Solden, L., Lloyd, K., Wrighton, K. (2016) The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr. Opin. Microbiol. 31, 217–226.

[56] Staley, J.T., Konopka, A. (1985) Measurement of in situ activities of non-photosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39, 321–346.

[57] Stewart, R.D., Auffret, M.D., Warr, A., Wiser, A.H., Press, M.O., Langford, K.W., Liachko, I., Snelling, T.J., Dewhurst, R.J., Walker, A.W., Roehe, R., Watson, M. (2018) Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. Nat. Commun. 9 (1), p. 870.

[58] Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., et al. (2005) Genome analysis of multiple pathogenic isolates of "Streptococcus agalactiae": implications for the microbial pan-genome. Proc. Natl. Acad. Sci. U S A 102 (39), 13950–13955.

[59] Tully, B.J., Emerson, J.B., Andrade, K., Brocks, J.J., Allen, E.E., Banfield, J.F., Heidelberg, K.B. (2015) De novo sequences of "Haloquadratum walsbyi" from Lake Tyrrell, Australia, reveal a variable genomic landscape. Archaea 2015, p. 875784.

[60] Tully, B.J., Sachdeva, R., Graham, E.D., Heidelberg, J.F. (2017) 290 metagenome-assembled genomes from the Mediterranean sea: a resource for marine microbiology. PeerJ 5, p. e3558.

[61] Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428 (6978), 37–43.

[62] Ventosa, A., de la Haba, R.R., Sanchez-Porro, C., HabaPapke, R.T. (2015) Microbial diversity of hypersaline environments: a metagenomic approach. Curr. Opin. Microbiol. 25, 80–87.

[63] Ventosa, A., Fernandez, A.B., Leon, M.J., Sanchez-Porro, C., Rodriguez-Valera, F. (2014) The Santa Pola saltern as a model for studying the microbiota of hypersaline environments. Extremophiles 18 (5), 811–824.

[64] Whitman, W.B. (2016) Modest proposals to expand the type material for naming of prokaryotes. Int. J. Syst. Evol. Microbiol. 66 (5), 2108–2112.

[65] Wu, Y.W., Simmons, B.A., Singer, S.W. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32 (4), 605–607.