Delgado and Andersson *Microbiome (2022)10:72*
https://doi.org/10.1186/s40168-022-01259-2

Microbiome

RESEARCH

Open Access

Evaluating metagenomic assembly approaches for biome-specific gene catalogues

Luis Fernando Delgado and Anders F. Andersson*

Abstract

Background: For many environments, biome-specific microbial gene catalogues are being recovered using shotgun metagenomics followed by assembly and gene calling on the assembled contigs. The assembly is typically conducted either by individually assembling each sample or by co-assembling reads from all the samples. The co-assembly approach can potentially recover genes that display too low abundance to be assembled from individual samples. On the other hand, combining samples increases the risk of mixing data from closely related strains, which can hamper the assembly process. In this respect, assembly on individual samples followed by clustering of (near) identical genes is preferable. Thus, both approaches have potential pros and cons, but it remains to be evaluated which assembly strategy is most effective. Here, we have evaluated three assembly strategies for generating gene catalogues from metagenomes using a dataset of 124 samples from the Baltic Sea: (1) assembly on individual samples followed by clustering of the resulting genes, (2) co-assembly on all samples, and (3) mix assembly, combining individual and co-assembly.

Results: The mix-assembly approach resulted in a more extensive nonredundant gene set than the other approaches and with more genes predicted to be complete and that could be functionally annotated. The mix assembly consists of 67 million genes (Baltic Sea gene set, BAGS) that have been functionally and taxonomically annotated. The majority of the BAGS genes are dissimilar (< 95% amino acid identity) to the Tara Oceans gene dataset, and hence, BAGS represents a valuable resource for brackish water research.

Conclusion: The mix-assembly approach represents a feasible approach to increase the information obtained from metagenomic samples.

Keywords: Gene catalogue, Brackish water, Metagenomics, Assembly approach, Mix assembly, Baltic Sea

Background

High-throughput sequencing has led to the establishment of the field of metagenomics, allowing the direct analysis of genetic material contained within an environmental sample [1]. This approach offers a detailed characterization of complex microbial communities without the need for cultivation. It can be used to address questions like *which* microorganisms are present, *what* are they capable of doing, and *how* do they interact. Metagenomics has been used for studying several ecosystem types, such as soils, human guts, and oceans [2–4].

For many environments, biome-specific gene catalogues have been recovered using shotgun metagenomics, followed by assembly and gene calling on the assembled contigs. Examples are the integrated reference catalogue of the human microbiome [4] and the Tara Oceans gene catalogue [2]. Gene catalogues facilitate the discovery of novel gene functions and gene variants. Annotated gene catalogues can also serve as genomic backbones onto which sequencing reads from metagenomes and metatranscriptomes, as well as

*Correspondence: anders.andersson@scilifelab.se

Department of Gene Technology, Science for Life Laboratory, School of Engineering Sciences in Chemistry, Biotechnology and Health, KTH Royal Institute of Technology, Stockholm, Sweden

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

mass-spectrometry spectra from metaproteomics, can be mapped, which enables fast and accurate taxonomic and functional profiling with such datasets.

The assembly can be carried out either by co-assembling reads from all the samples (or groups of samples) or individually assembling reads from each sample. The co-assembly approach has the advantage that some genes displaying too low abundance to be assembled from individual samples may reach enough coverage to be recovered. However, combining data from many samples often means mixing data from a diversity of closely related strains (from the same species). This fine-scale genomic variation can compromise the assembly process because the de Bruijn graph will include many alternative paths. Consequently, the assembler may decide to break the graph into smaller pieces, which can result in fragmentation of genes.

An alternative approach is to perform assembly on each sample individually. The individually assembled samples approach will minimize the mixing of data from different strains and therefore potentially result in more completely assembled genes, at least for fairly abundant genomes. However, another problem arises, which is that (more or less) identical genes from multiple samples will be reconstructed. To serve as a reference dataset, it is desirable to have a nonredundant set of genes. Sequence redundancy removal can be achieved by clustering the gene sequences (or their protein translations [5]) resulting from the different assemblies based on sequence similarity, using some cutoff criteria. For each gene cluster, a representative sequence is then chosen based on, e.g. gene completeness, centrality in the cluster, or abundance in the dataset.

Recently, a Baltic Sea specific gene catalogue with 6.8 million genes was constructed based on the metagenomic data from 81 water samples spanning the spatiotemporal gradients of the Baltic Sea [6]. For the construction of the Baltic Sea specific gene catalogue, all the 2.6 billion (i.e. 10^9) reads were co-assembled, and genes called on all contigs > 1000 bp. While this gene catalogue has established itself as a useful resource for analysing metagenome and metatranscriptome datasets from brackish environments [7–11], only ca 10% of the shotgun reads from a typical Baltic Sea metagenome sample are mapping to genes with a functional annotation [6]. A reason for the seemingly low coverage could be that the co-assembly approach has resulted in a fragmented assembly. A more comprehensive reference gene catalogue would hence be desirable for this environment. In this study, we conduct an extensive comparison of three assembly approaches on an expanded set of metagenome samples from the Baltic Sea and present an updated gene catalogue for the Baltic Sea microbiome.

Methods

Metagenome samples

Five previously published sample sets [6, 7, 12] were used in this study. The sampling locations are shown in Additional file 1, and a brief description of sample retrieval and sequencing is given in Additional file 2; for further details, we refer to the original publications. Sequencing of all sample sets was conducted using Illumina HiSeq 2500.

Preprocessing of reads

Removal of low-quality bases was performed earlier [7] using Cutadapt [13] (parameters-q 15, 15) followed by adapter removal (parameters-n 3 — minimum length 31). The resulting read files were thereafter screened for PCR duplicates using FastUniq [14] with default parameters.

Assembly

Individual assemblies on the 124 samples were performed earlier [7], using MEGAHIT [15] v.1.1.2 with the "--presets meta-sensitive" option. For the co-assembly conducted here, all preprocessed reads were first combined and normalized using BBnorm of BBmap v.38.08 (https://sourceforge.net/projects/bbmap/) with the following parameters: target = 70, mindepth = 2, and prefilter = t. Also, the normalized read set was too extensive to allow co-assembly with the tag "presets –meta-sensitive" with MEGAHIT. Therefore, they were assembled with "--presets meta-large" (using MEGAHIT v.1.1.2), as recommended for complex metagenomes in the MEGAHIT documentation.

Gene prediction

Genes were predicted on contigs (from the co-assembly and from the individual assemblies) using Prodigal [16] v.2.6.3 with the -p meta option. Gene completeness is based on Prodigal gene prediction. Complete genes refer to predicted genes having a predicted start and a stop codon (Prodigal indicator "00"); partial genes are predicted genes with either no start or stop codon (Prodigal indicator "01" or "10"), typically due to that the gene runs off the edge of a contig; and incomplete genes are predicted genes without a start and a stop codon (Prodigal indicator "11").

Protein clustering

Clustering of the proteins stemming from the different samples for the individual assembly, and from the co-assembly for the mix-assembly strategy, was performed using MMseqs2 [17] v9.d36de using the cascaded clustering mode (MMseqs2 cluster, https://mmseqs.com/latest/userguide.pdf). Clustering was first performed on the proteins from the individual assemblies, and the cluster-representative proteins were subsequently clustered with the co-assembly proteins. The following parameters were used in the two MMseqs2 runs: -c 0.95, --min-seq-id 0.95, --cov-mod 1, and --clust-mod 2. This means proteins displaying ≥ 95% amino acid identity were clustered. Strains belonging to the same prokaryotic species generally display > 95% average amino acid identity [18]. As recommended in the MMseq2 user guide, -cov-mod 1 was used, since it allows clustering of fragmented proteins (as often occurs in metagenomic datasets). With --cov-mode 1 only, sequences are clustered that have a sequence length overlap greater than the percentage specified by -c (i.e. 95% with -c 0.95) of the target sequence. In MMseqs2, the query is seen as the representative sequence, and the target is a member sequence. To lower the risk for fragmented proteins becoming cluster-representative sequences, -cluster-mode 2 was used, again following the recommendations of the MMseq2 user guide. It sorts sequences by length and in each clustering step forms a cluster containing the longest sequence and the sequences that it matches.

Read mapping and counting

To reduce the computational burden of the read mapping, random subsets of 10,000 non-normalized forward reads per sample were created using seqtk v.1.2-r101-dirty (https://github.com/lh3/seqtk), with seed 100 (-s 100). These reads (12.4 million in total) were mapped to the representative gene sequences from the individual, co-, and mix assembly, respectively, using Bowtie2 v.2.3.4.3 [19], with the parameter "--local." The resulting SAM files were converted to BAM with SAMtools v.1.9 [20]. The htseq-count script from HTSeq [21] v.0.11.2 was used to obtain raw counts per gene, with the parameters "-f bam -r pos -t CDS -i ID -s no -a 0". For the counting, GFF input files were used, created using the script create_gff.py available at https://github.com/EnvGen/toolbox/tree/master/scripts. In order to estimate read depth coverage of the genes in the total metagenome, we multiplied the counts per gene by the average read-pair length divided by the length of the gene and multiplied this number with the total number of read pairs in the whole dataset divided by the total number of randomly sampled forward reads. This is a rough estimation of the coverage of each gene in the total metagenome; however, after normalisation with BBnorm, high coverage genes will get a lower coverage.

Functional annotations

Functional annotation of proteins was conducted using EggNOG [22], Pfam [23], and dbCAN [24]. Annotations against Pfam v.31.0 and dbCAN v.5.0 were conducted with hmmsearch and hmmscan [25], respectively, in HMMER v.3.2.1, selecting hits with *E*-value < 0.001. Annotations against EggNOG v.4.5.1 were performed using eggNOG-mapper v.1.0.3 [26], using accelerated profile HMM searches [27], following the recommendation for setting up large annotation jobs.

Taxonomic affiliation

MMseqs2 (v13.45111) taxonomy [28], with parameters "--orf-filter 0 --tax-lineage 1", was used to assign taxonomic labels to contigs from which representative genes were predicted. MMseqs2 taxonomy uses an approximate 2bLCA (lowest common ancestor, LCA) approach. GTDB [29, 30] v.202 was used as a reference database for bacteria and archaea and Uniprot90 [31] (downloaded on June 4, 2021) for eukaryotes and viruses. An interactive chart for the gene set's taxonomic information was generated using Krona (Ondov et al. 2011) (see Additional file 3).

RNA gene screening

Barrnap v.0.9 [32], using default parameters, was used to identify potential rRNA genes, and identification of rRNA and other potential RNA genes in the mix-assembly gene set was conducted using the Rfam v.14.6 [33] database, with hmmsearch [25], in HMMER v.3.3.2, with flag "--cut_ga". The union of genes identified as rRNA by Barnap and Rfam/hmmsearch was removed from the final gene set.

Results

We used a set of 124 metagenome samples from the Baltic Sea ([6, 7, 12]; see Additional file 1) to evaluate three assembly approaches for generating a nonredundant gene catalogue: co-assembly on all samples ("co-assembly"), assembly on individual samples ("individual assembly"), and a combination of the previous two ("mix assembly"). For the co-assembly, due to the complexity of the dataset, direct co-assembly of all reads was not possible, even on a server with 1 TB of memory. Therefore, the reads were first normalized such that reads stemming from highly abundant genomes (with high-frequency *k*-mers) were downsampled (to a depth of 70× coverage), and those presumably derived from errors (with a depth below 2×) were removed. This reduced the total number of read pairs from 5.4 to 2.9 billion.

Since the contigs of the co-assembly are derived from reads from all samples, it will result in a nonredundant set of genes. In contrast, genes from the individually assembled samples may overlap between samples. To reduce this redundancy, clustering was conducted on the encoded proteins [17]. We used a cutoff of 95% amino acid identity, conforming to that strains belonging to the same species typically display more than 95%

Table 1 Summary statistics for the different assembly approaches

Assembly approach	Total bps	Number of genes	Num. of genes ≥ 100 bp	Num. of complete genes	Num. of partial genes	No. of incompletegenes
Individual	18,770,879,205	50,045,582	45,859,319	6,258,868	27,073,554	16,713,160
Co	20,347,887,912	45,455,222	42,278,556	11,443,584	23,815,733	10,195,905
Mix	27,043,772,505	67,583,055	61,576,531	12,690,647	37,345,617	17,546,791

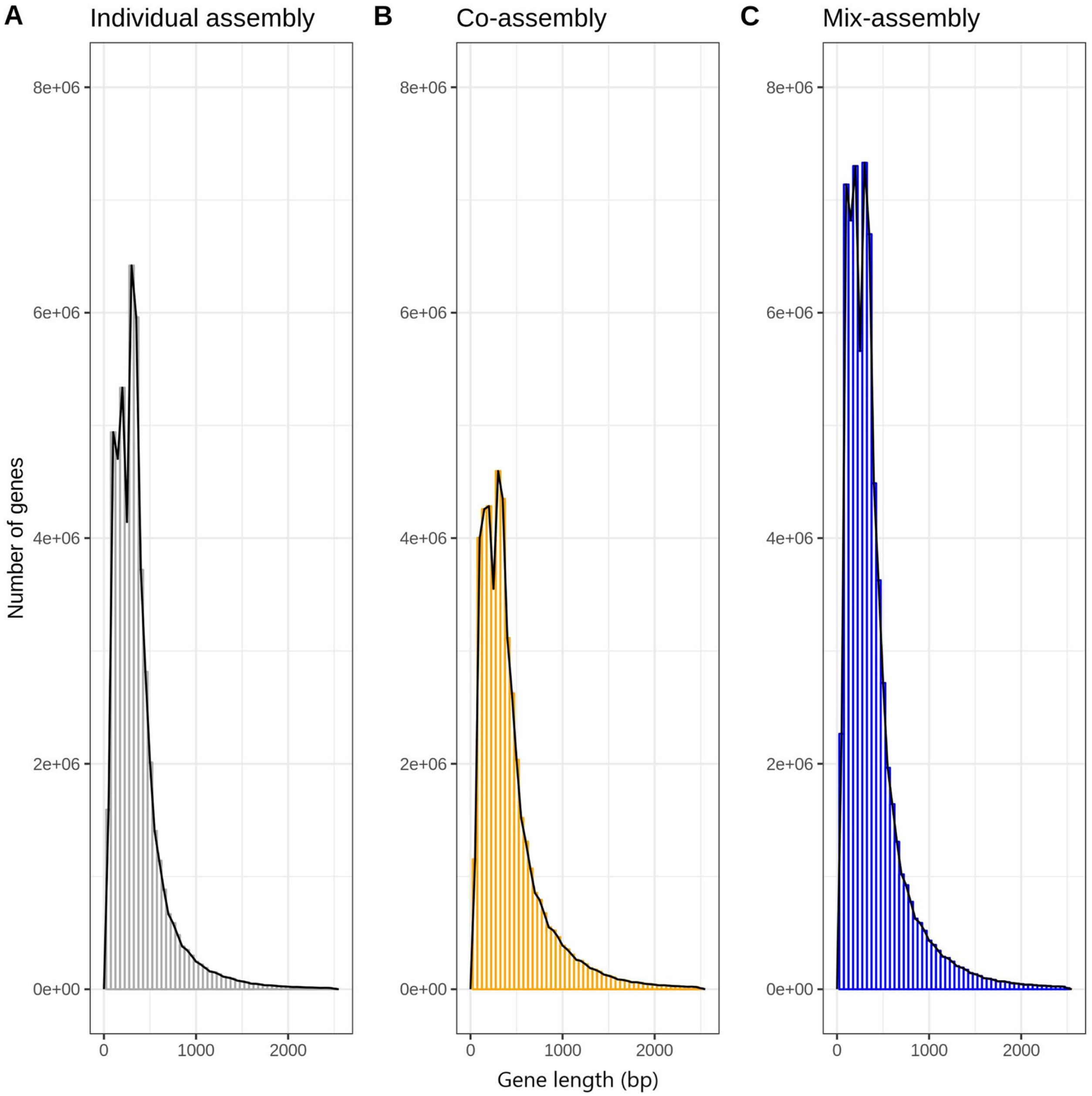

Fig. 1 Gene length distributions of the three assembly approaches. **a** Co-assembly. **b** Individual assembly. **c** Mix assembly. Only genes ≤ 2500 bp are included in the histograms

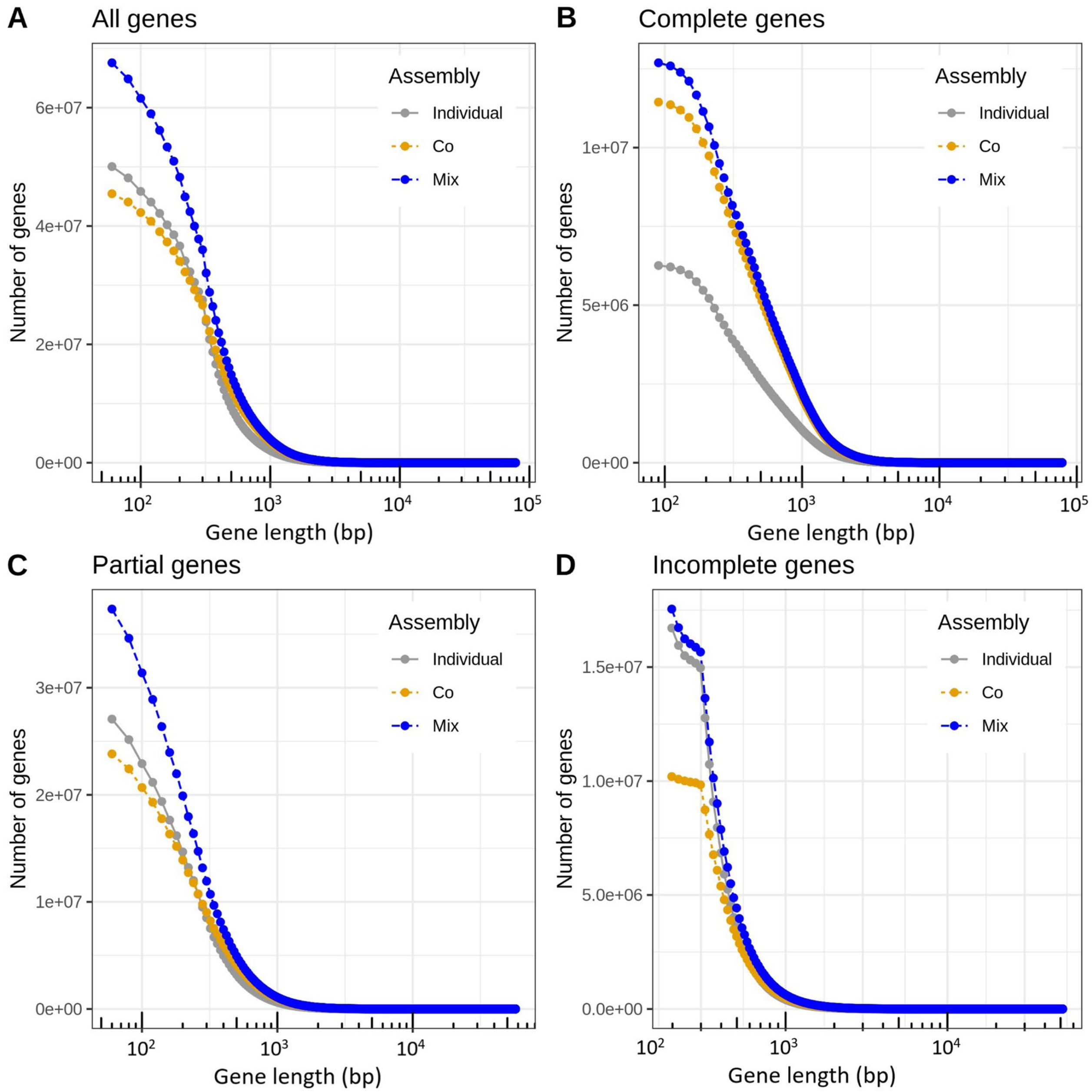

Fig. 2 Cumulative distribution of gene lengths for the three assembly approaches. **a** All genes. **b** Complete genes. **c** Partial genes. **d** Incomplete genes. Complete genes refers to genes predicted to be complete (having a predicted start codon and a stop codon), partial genes to genes that lack either a start or a stop, and incomplete genes to genes that lack both start and stop. Gene length is given in logarithmic scale

Table 2 Pfam annotation statistics for the different assembly approaches

Assembly approach	Total number of annotated genes	Number of annotated complete genes	Number of annotated partial genes	Number of annotated incomplete genes
Individual	11,930,617	2,422,526	4,751,188	4,756,903
Co	13,343,858	4,514,607	5,128,252	3,700,999
Mix	15,566,195	4,584,290	5,751,705	5,230,200

average amino acid identity [18]. This reduced the number of individual-assembly genes from 134 to 50 million. Likewise, clustering was conducted on the co-assembly proteins together with the nonredundant set of individual-assembly proteins, to generate the mix-assembly gene set.

The mix-assembly approach resulted in the largest number of nonredundant genes (67 M), followed by individual assembly (50 M) and co-assembly (45 M; Table 1). Mix assembly also had the largest number of genes predicted to be complete (12 M) followed closely by co-assembly (11 M) but twice as many as individual assembly (6 M; Table 1).

The gene length distributions were fairly similar for the three approaches (Fig. 1), with peaks in the distributions between 300 and 350 bp. Co-assembly had the largest median gene length (336 bp), although mix assembly had the largest number of genes along the full range of gene lengths (Fig. 2).

Annotating the proteins against Pfam [23] gave the largest number of annotated genes for mix assembly (15 M) followed by co-assembly (13 M) and individual assembly (12 M), despite that co-assembly had a higher proportion of genes with annotation (29.4%) compared to the other two (23.0% for mix assembly, 23.8% for individual assembly; Table 2).

Since biome-specific gene catalogues are often used as reference sequences for mapping of shotgun reads from metagenomes or transcriptomes, we further evaluated the gene sets by mapping reads from the metagenome samples to them. The average mapping rates for the 124 samples were 83.9, 84.7, and 87.7% for individual-, co-, and mix assembly, respectively, with numbers ranging from 47.5, 49.2, and 53.2% to 96.2, 96.1, and 97.3% for individual-, co-, and mix assembly. The mix-assembly read-mapping rate was significantly higher than the individual- (Wilcoxon signed-rank test, $P < 10^{-21}$) and co-assembly ($P < 10^{-21}$) rates (Fig. 3a). Figure 4 presents

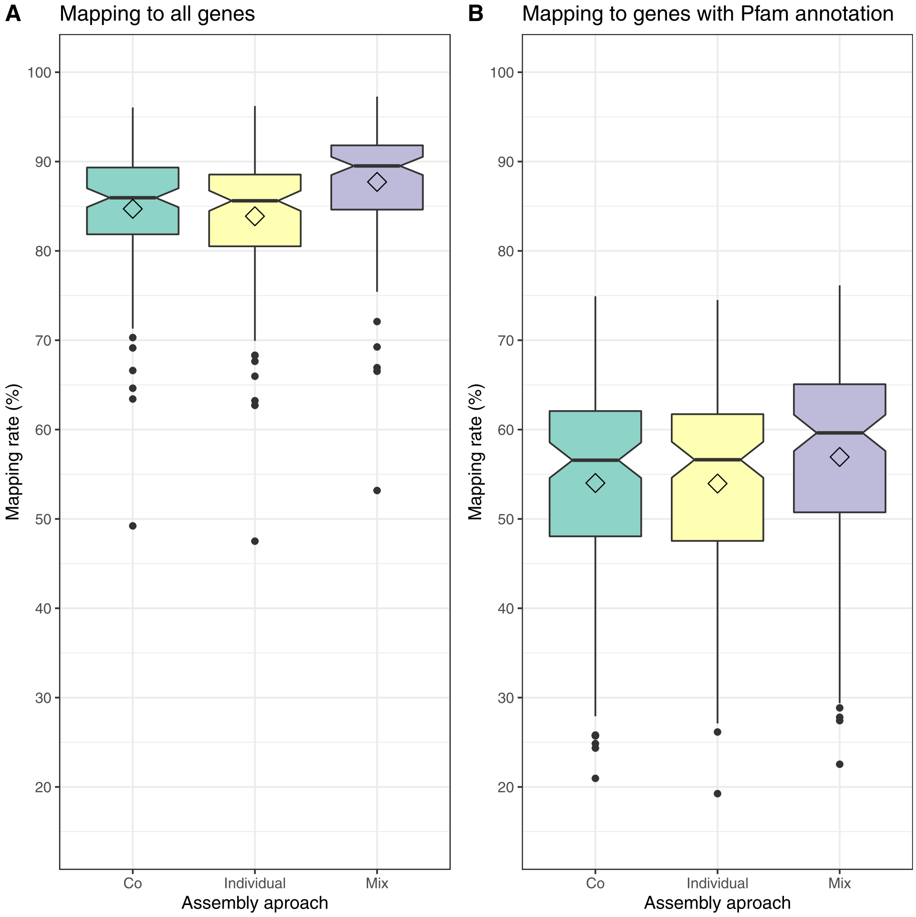

Fig. 3 Read mapping rates to genes from the three assembly approaches. The boxplots show the distribution of mapping rate (% of reads) for the 124 samples, based on a random subset of 10,000 forward reads per sample. **a** When mapping to all genes. **b** When mapping to genes with Pfam annotation

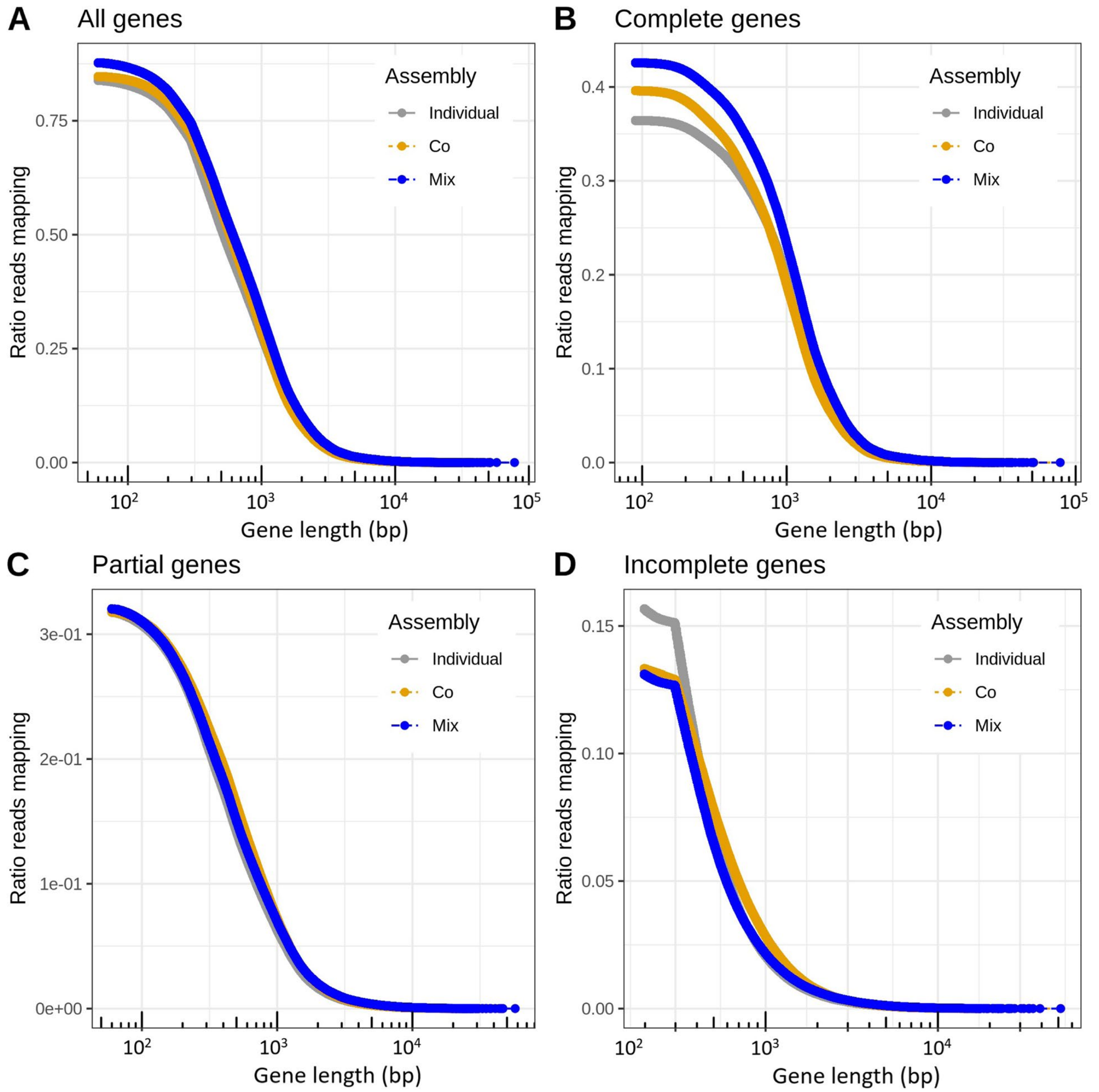

Fig. 4 Read mapping rate as a function of gene length cutoff. The plots show the proportion of reads mapping at different cutoffs on minimum gene length. **a** All genes. **b** Complete genes. **c** Partial genes. **d** Incomplete genes. Complete genes refer to genes predicted to be complete (having a predicted start codon and a stop codon), partial genes to genes that lack either a start or a stop, and incomplete genes to genes that lack both start and stop. Gene lengths are given in logarithmic scale

the cumulative mapping rate by gene length, showing the proportion of reads mapping at different gene length cutoffs. For all three assembly strategies, the highest fraction of reads mapping corresponds to complete genes, followed by partial genes. Of the three, mix assembly had the highest fraction of mapping reads mapping to complete genes (42.6%) and the lowest to partial (32.0%) and incomplete (13.1%) genes (see 'Methods' for definitions of partial and incomplete genes). Mix assembly also had the highest proportion of reads mapping to genes with a Pfam annotation (56.9%, $P < 10^{-21}$), followed by co-assembly (54.0%) and individual assembly (54.0%) (Fig. 3b).

The contribution of genes from the individual- and co-assembly to the mix-assembly set of genes is shown in Fig. 5. A majority (52%) of the mix-assembly genes

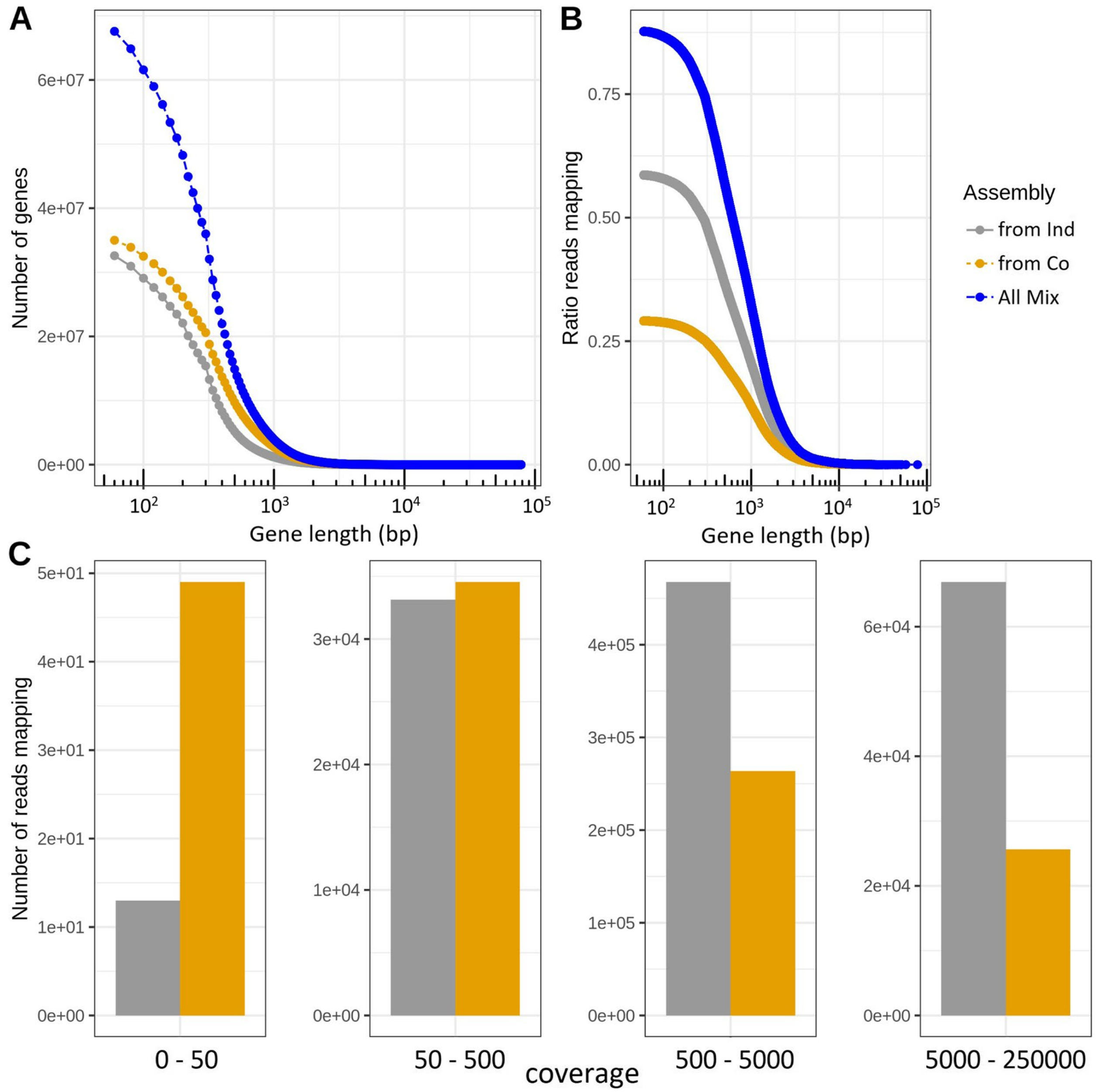

Fig. 5 Contribution of genes from individual assembly and co-assembly to the mix-assembly gene set. **a** Cumulative distribution of gene lengths for the mix-assembly genes: for all ("All mix") and for those derived from individual-assembly ("from Ind") and co-assembly ("from Co"). Gene length is given in logarithmic scale. **b** Read mapping rate as a function of gene length cutoff. **c** Total number of reads mapping to mix-assembly genes derived from either individual assembly or co-assembly, for four bins of genes binned by their estimated coverage in the total metagenome (see "Methods"): low (0–50 ×), median (50–500 ×), high (500–5000 ×), and very high (5000–250,000 ×) read depth coverage

originates from co-assembly genes (Fig. 5a), representing 67% of the complete and 50% and 45% of the partial and incomplete genes, respectively (data not shown). However, among the reads that map to the mix-assembly genes, a larger fraction of reads map to genes derived from the individual assembly than to genes derived from the co-assembly (Fig. 5b). These seemingly conflicting results may reflect that mix-assembly genes derived from the individual assembly tend to be of higher abundance in the microbial communities than those from the co-assembly. This was confirmed by grouping the mix-assembly genes in low, median, and high coverage genes, where the majority of mapping reads mapped to genes derived from co-assembly for low coverage genes, but to

Table 3 Number of mix-assembly representative genes annotated using different databases

Gene completeness	dbCAN	EggNOG	Pfam
Complete	420,422	5,354,169	4,582,506
Partial	562,445	8,374,034	5,751,622
Incomplete	603,580	7,865,395	5,230,173
Total	1,586,447	21,593,598	15,564,301

genes derived from individual assembly for high coverage genes (Fig. 5c).

The mix-assembly gene set is significantly more extensive than the previously published Baltic Sea gene catalogue (BARM [6];) and may serve as a valuable resource for brackish water research. We compared the mix-assembly protein set with the Tara Ocean Microbial Reference Gene Catalog (OM-RGC.v2 [34]). Of the 67.5 M mix-assembly proteins, only 1.4 M were > 95% identical to Tara proteins, and, vice versa, of the 46.7 M Tara proteins, 1.3 M were > 95% identical to the mix-assembly proteins. Hence, the vast majority of the mix-assembly gene sequences are distinct from Tara genes. To increase the usefulness of the mix-assembly gene set, we removed genes potentially encoding ribosomal RNA and thus falsely predicted as protein coding (n = 16,804) and conducted taxonomic and functional annotation on the remaining genes. A subset of the genes (n = 70,223) was predicted to include encodings of other structural RNAs (in Rfam [33]), but we decided to keep these since they may also encode important protein-coding regions. The resulting gene set, which we call BAltic Gene Set (BAGS.v1), encompasses 67,566,251 genes, of which 31.0 M have a taxonomic affiliation (see Additional file 3) and 23.4 M have at least one type of functional annotation: 15.5 M with Pfam, 21.5 M with EggNOG [22], and 1.5 M with dbCAN [24] annotation (Table 3).

Twenty-seven percent of the BAGS.v1 genes were predicted to be of eukaryotic origin. It should however be noted that the gene predictions were conducted with a gene caller for prokaryotic genes (Prodigal), and that a fraction of the eukaryotic genes has likely been imperfectly predicted.

Discussion

Metagenome assembly is commonly carried out either by individually assembling reads from each sample [35] or by co-assembling reads from all the samples of a dataset [2, 6]. Here, the performance of these assembly approaches was compared. Although the number of genes was lower for the co-assembly, the total length (in number of base pairs) was higher than for the individual assembly. The two gene sets reported a similar mapping rate, although the co-assembly set had a higher number of genes predicted to be complete and a lower number of partial and incomplete genes than the individual-assembly set. In this study, we also proposed a new approach for assembly, aiming to combine the advantages of the individual- and co-assembly approaches, referred to as mix assembly. The mix-assembly strategy resulted in significantly more genes than the other approaches and also in the largest number of complete genes. It further gave the highest mapping rates and the greatest number of genes with a Pfam annotation. The reason why not only the number of genes but also the number of complete genes increased compared to the other approaches is likely because in the protein clustering process, the longest proteins were selected to form cluster seeds. Thus, if for example, an incomplete or partial protein from the co-assembly set forms a cluster with a complete protein from the individual assembly, the complete protein will likely represent this cluster in the mix assembly, since it is longer. Thereby, the clustering step that combines the two gene sets enriches for complete proteins. However, it may also to some extent enrich for artificially long proteins that may stem from sequencing, assembly or gene calling errors.

Analysing the contribution of individual- and co-assembly genes in the set of mix-assembly genes showed that genes with relatively low coverage (low number of mapping reads) in the samples were mainly stemming from the co-assembly. This likely reflects that co-assembly sometimes is able to recover genes that display too low coverage to be assembled from individual samples. On the other hand, genes with relatively high coverage were mostly originating from the individual assembly, which may be caused by the co-assembly sometimes breaking in such genes due to strain variation. If strain variation for such a gene is less pronounced in at least one of the individual samples, a longer fraction of the gene could be recovered in the individual assembly.

The 67 million genes of the mix assembly are based on 124 metagenome samples that span the salinity and oxygen gradients of the Baltic Sea and also capture seasonal dynamics at two locations [7]. This dataset (BAGS.v1) is a tenfold expansion compared to our previous gene set [6] and has the potential to serve as an important resource for exploring gene functions and serve as a backbone for mapping of meta-omics data from brackish environments. Consistent with our earlier study showing that the prokaryotes of the Baltic Sea are closely related to but genetically distinct from freshwater and marine relatives [35], only a small fraction of the mix-assembly genes displayed > 95% amino acid similarity to genes of the Tara

Ocean gene catalogue. This implies that the Tara Ocean catalogue is not suitable for mapping of meta-omics data from the Baltic Sea and emphasizes the need for a brackish water microbiome reference gene catalogue. The gene catalogue BAGS.v1, including gene and protein sequences, and taxonomic and functional annotations, is publicly available at the SciLifeLab Data Repository, https://doi.org/10.17044/scilifelab.16677252.

Conclusion

In this study, we have evaluated three metagenome assembly approaches for biome-specific gene catalogues. The mix-assembly approach, which combines assembly on individual samples with co-assembly on all samples, outperformed the other two approaches in terms of number of nonredundant genes, number of complete genes, mapping rates, and number of genes with a Pfam annotation. Hence, the mix-assembly approach represents a feasible approach to increase the information gained from metagenomic samples.

Abbreviations

BAGS: Baltic Sea gene set; LCA: Lowest common ancestor.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-022-01259-2.

Additional file 1. Map with sampling locations. The marker colour shows the salinity of the water sample and its size, the sampling depth. The contour lines indicate depth with 50 m intervals.

Additional file 2. Table with brief description of sampling and sequencing. For detailed descriptions, see references in the table.

Additional file 3. Interactive chart of the BAGS gene set taxonomic affiliations.

Acknowledgements

Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Authors' contributions

AFA designed the study. LFD conducted the bioinformatics analyses. Both authors wrote the manuscript. The authors read and approved the final manuscript.

Funding

Open access funding provided by Royal Institute of Technology. This work is part of the Swedish Biodiversity Data Infrastructure (SBDI), funded by its partner organizations, and the Swedish Research Council VR through grant no. 2019-00242. Funding was also provided by the Swedish Research Council VR grant no. 2021-05563 and through the 2019–2020 BiodivERsA joint call for research proposals under the BiodivClim ERA-Net COFUND programme together with the funding organisation FORMAS, grant number 2020-02366.

Availability of data and materials

The shotgun reads and individual sample assemblies have been published earlier [6, 7, 12]. The co-assembly contigs and the mix-assembly gene set (BAGS) together with annotations are available at the SciLifeLab Data Repository powered by Figshare, https://doi.org/10.17044/scilifelab.16677252. The contigs for the individual assemblies were published earlier [7] and are available at ENA hosted by EMBL-EBI under the study accession number PRJEB34883. When using the BAGS gene set in your work, please cite Alneberg et al. (2020) [7] in addition to this study.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 18 October 2021 Accepted: 17 March 2022
Published online: 06 May 2022

References

1. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform Biol Insights. 2015;9:75–88 Available from: https://doi.org/10.4137/BBI.S12462.
2. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348. American Association for the Advancement of Science. [cited 2021 Aug 11]. Available from: https://doi.org/10.1126/science.1261359.
3. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al. Strategies to improve reference databases for soil microbiomes. ISME J. 2017;11:829–34 Available from: https://doi.org/10.1038/ismej.2016.168.
4. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41 Available from: https://doi.org/10.1038/nbt.2942.
5. Steinegger M. Ultrafast and sensitive sequence search and clustering methods in the era of next generation sequencing [Internet]. Technische Universität München; 2018. Available from: http://mediatum.ub.tum.de/doc/1435187/678546.pdf.
6. Alneberg J, Sundh J, Bennke C, Beier S, Lundin D, Hugerth LW, et al. BARM and BalticMicrobeDB, a reference metagenome and interface to meta-omic data for the Baltic Sea. Sci Data. 2018;5:180146 Available from: https://doi.org/10.1038/sdata.2018.146.
7. Alneberg J, Bennke C, Beier S, Bunse C, Quince C, Ininbergs K, et al. Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes [Internet]. Commun Biol. 2020; Available from: https://doi.org/10.1038/s42003-020-0856-x.
8. Bunse C, Israelsson S, Baltar F, Bertos-Fortis M, Fridolfsson E, Legrand C, et al. High frequency multi-year variability in Baltic Sea microbial plankton stocks and activities [Internet]. Front Microbiol. 2019; Available from: https://doi.org/10.3389/fmicb.2018.03296.
9. Markussen T, Happel EM, Teikari JE, Huchaiah V, Alneberg J, Andersson AF, et al. Coupling biogeochemical process rates and metagenomic blueprints of coastal bacterial assemblages in the context of environmental change. Environ Microbiol. 2018;20:3083–99 Available from: https://doi.org/10.1111/1462-2920.14371.
10. Capo E, Bravo AG, Soerensen AL, Bertilsson S, Pinhassi J, Feng C, et al. Deltaproteobacteria and spirochaetes-like bacteria are abundant putative mercury methylators in oxygen-deficient water and marine particles in the Baltic Sea. Front Microbiol. 2020;11:574080 Available from: https://doi.org/10.3389/fmicb.2020.574080.
11. Grossart H-P, Massana R, McMahon KD, Walsh DA. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnol Oceanogr. Wiley; 2020. p. 65. Available from: https://doi.org/10.1002/lno.11382.
12. Larsson J, Celepli N, Ininbergs K, Dupont CL, Yooseph S, Bergman B, et al. Picocyanobacteria containing a novel pigment gene cluster dominate

the brackish water Baltic Sea. ISME J. 2014;8:1892–903 Available from: https://doi.org/10.1038/ismej.2014.35.
13. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10–2. [cited 2021 Aug 11]. Available from: https://doi.org/10.14806/ej.17.1.200.
14. Xu H, Luo X, Qian J, Pang X, Song J, Qian G, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. PLoS One. 2012;7:e52249 Available from: https://doi.org/10.1371/journal.pone.0052249.
15. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6 Available from: https://doi.org/10.1093/bioinformatics/btv033.
16. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119 Available from: https://doi.org/10.1186/1471-2105-11-119.
17. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time [Internet]. Nat Commun. 2018; Available from: https://doi.org/10.1038/s41467-018-04964-5.
18. Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes [Internet]. J Bacteriol. 2005:6258–64 Available from: https://doi.org/10.1128/jb.187.18.6258-6264.2005.
19. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9 Available from: https://doi.org/10.1038/nmeth.1923.
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9 Available from: https://doi.org/10.1093/bioinformatics/btp352.
21. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2014;31:166–9. Oxford Academic. [cited 2021 Aug 11]. Available from: https://doi.org/10.1093/bioinformatics/btu638.
22. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93 Available from: https://doi.org/10.1093/nar/gkv1248.
23. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9 Available from: https://doi.org/10.1093/nar/gkaa913.
24. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012;40:W445–51 Available from: https://doi.org/10.1093/nar/gks479.
25. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching [Internet]. Nucleic Acids Res. 2011:W29–37 Available from: https://doi.org/10.1093/nar/gkr367.
26. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C. et al, Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper [Internet]. Mol Biol Evol. 2017:2115–22 Available from: https://doi.org/10.1093/molbev/msx148.
27. Eddy SR. Accelerated profile HMM searches [Internet]. PLoS Comput Biol. 2011:e1002195 Available from: https://doi.org/10.1371/journal.pcbi.1002195.
28. Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy KE. Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics. 2021; Available from: https://doi.org/10.1093/bioinformatics/btab184.
29. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for bacteria and archaea [Internet]. Nat Biotechnol. 2020:1079–86 Available from: https://doi.org/10.1038/s41587-020-0501-8.
30. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36:996–1004 Available from: https://doi.org/10.1038/nbt.4229.
31. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9 Available from: https://doi.org/10.1093/nar/gkaa1100.
32. Seemann T. barrnap 0.9 : rapid ribosomal RNA prediction [Internet]. 2018. Available from: https://github.com/tseemann/barrnap.
33. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2021;49:D192–200 Available from: https://doi.org/10.1093/nar/gkaa1047.
34. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh H-J, Cuenca M, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell. 2019;179:1068–83.e21 Available from: https://doi.org/10.1016/j.cell.2019.10.014.
35. Hugerth LW, Larsson J, Alneberg J, Lindh MV, Legrand C, Pinhassi J, et al. Metagenome-assembled genomes uncover a global brackish microbiome. Genome Biol. 2015;16:279 Available from: https://doi.org/10.1186/s13059-015-0834-7.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC