METHOD ARTICLE

ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services [version 1; referees: 1 approved, 2 approved with reservations]

Espen Mikal Robertsen [1], Hubert Denise [2], Alex Mitchell [2], Robert D. Finn [2], Lars Ailo Bongo [1], Nils Peder Willassen[1]

[1]Center for Bioinformatics (SfB), UiT The Arctic University of Norway Bioinformatics, Tromsø, Norway
[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, UK

v1 **First published:** 23 Jan 2017, **6**(ELIXIR):70 (doi: 10.12688/f1000research.10443.1)
Latest published: 23 Jan 2017, **6**(ELIXIR):70 (doi: 10.12688/f1000research.10443.1)

Abstract

Metagenomics, the study of genetic material recovered directly from environmental samples, has the potential to provide insight into the structure and function of heterogeneous microbial communities. There has been an increased use of metagenomics to discover and understand the diverse biosynthetic capacities of marine microbes, thereby allowing them to be exploited for industrial, food, and health care products. This ELIXIR pilot action was motivated by the need to establish dedicated data resources and harmonized metagenomics pipelines for the marine domain, in order to enhance the exploration and exploitation of marine genetic resources. In this paper, we summarize some of the results from the ELIXIR pilot action “Marine metagenomics – towards user centric services”.

This article is included in the EMBL-EBI gateway.

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

This article is included in the ELIXIR gateway.

Open Peer Review

Referee Status: ? ✓ ?

	Invited Referees 1	2	3
version 1 published 23 Jan 2017	? report	✓ report	? report

1 **Marla I. Trindade** , University of the Western Cape South Africa, **Leonardo van Zyl** , , University of the Western Cape South Africa

2 **Takashi Gojobori** , King Abdullah University of Science and Technology (KAUST) Saudi Arabia

3 **Anders Blomberg** , University of Gothenburg Sweden

Discuss this article

Comments (0)

Corresponding author: Espen Mikal Robertsen (espen.m.robertsen@uit.no)

Competing interests: No competing interests were disclosed.

How to cite this article: Robertsen EM, Denise H, Mitchell A *et al.* **ELIXIR pilot action: Marine metagenomics – towards a domain specific set of sustainable services [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2017, **6**(ELIXIR):70 (doi: 10.12688/f1000research.10443.1)

Grant information: Funding was provided from ELIXIR, EMBL-EBI and UiT The Arctic University of Norway.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 23 Jan 2017, **6**(ELIXIR):70 (doi: 10.12688/f1000research.10443.1)

Introduction

Marine microbial genomics and metagenomics are arguably still in their infancies, but each discipline is rapidly expanding in terms of research activity and are converging against each other. At present, the lack of specialized databases for marine metagenomics[1], as well as dedicated data management e-infrastructures and harmonized pipelines, makes implementation of large-scale studies challenging, and replication of analysis close to impossible. In addition, data production from metagenomics projects is growing exponentially due to reducing sequencing costs[2], which demands optimized and flexible solutions for analysis of metagenomic data.

To address these challenges, UiT The Arctic University of Norway (part of ELIXIR Norway), together with the European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), initiated an ELIXIR pilot action with the following aims: (i) Identify the overlap between two existing metagenomics pipelines, EBI Metagenomics Portal (EMG)[3] and META-pipe[4], thereby opening the potential for interoperability; (ii) implement new or improve existing components in each pipeline to enrich the output; and (iii) perform a gap analysis to identify deficient areas of marine metagenomics analysis. The overall outcome of this pilot action was aimed at shaping the foundations from which the marine (meta)genomics community could establish long-term, sustainable service platforms. A description of this pilot action and a webinar video can be found at https://www.elixir-europe.org/about/implementation-studies/marine-metagenomics and https://www.elixir-europe.org/documents/update-elixir-pilot-actions-launched-2014-marine-metagenomics-towards-user-centric, respectively.

The EMBL-EBI has developed EMG, a generic platform, which aims to provide insights into the phylogenetic diversity and functional potential of all environmental samples, while UiT has specifically developed META-pipe towards the marine domain, with a focus on bioprospecting. In this article, we describe a comparison of the two pipelines, using the outputs of equivalent input sequence data to illustrate the similarities and differences.

Overview of pipeline workflows

Figure 1 shows a schematic of the pipeline workflows from META-pipe and EMG. A simple visual comparison of these workflows reveals that while there are some commonalities between the pipelines, there are a series of key differences in the tools and approaches to the analysis.

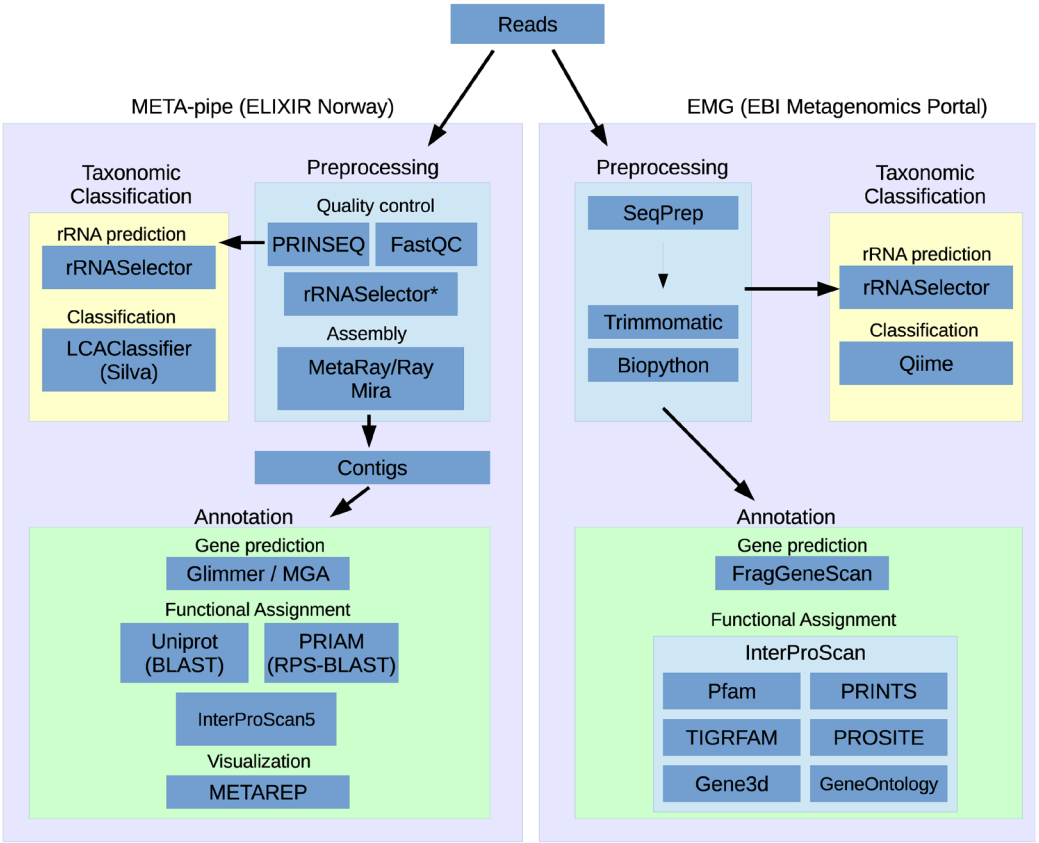

Figure 1. Tools and steps in EMG and META-pipe.

Briefly, the main differences between the pipelines are in the preprocessing and taxonomic classification steps. More specifically, while both preprocessing steps preform filtering of low quality reads and length filtering they diverge thereafter. META-pipe performs assembly of small-subunit (SSU) ribosomal RNAs (rRNAs) filtered reads, whereas EMG merges overlapping pair-end reads into longer single reads (where appropriate) and performs taxonomic classification and functional analysis on unassembled sequences. Despite the presence or absence of sequence assembly, both pipelines use rRNASelector[5] for the identification of 5S, 16S and 23S rRNA sequences, before passing extracted rRNAs to taxonomic classification tools. EMG uses QIIME[6] with Greengenes[7] and a closed-reference OUT picking strategy for taxonomic classification and annotation of 16S rRNA, while META-pipe uses LCAClassifier[8] coupled with a manually curated custom database coined SilvaMod, derived from SILVA[9] and especially created for LCAClassifier. EMG masks all rRNA regions before passing them to the functional analysis section of the pipeline, whereas META-pipe removes all rRNA prior to assembly. There are also some minor differences between the pipelines in functional annotation. EMG uses FragGeneScan[10] for gene prediction and a subset of the InterPro database together with the InterProScan5[11] for functional assignment of predicted coding sequences (CDSs). META-pipe use MetaGeneAnnotator (MGA)[12] for gene prediction and InterProScan5 with the full InterPro database, and BLAST against PRIAM[13] and UniProt[14] for additional functional assignment.

To understand the impact of the outlined differences in functional and taxonomic identifications, and as a prelude to future harmonisation, we undertook a comparison of the results from four different environmental datasets using EMG v2.0 and META-pipe. In addition, as a part of the ELIXIR pilot action project, we performed a gap analysis and concluded on recommendations for developing sustainable ELIXIR services for marine metagenomics domain.

Methods

Metagenomics datasets used in the study

For comparison of the two pipelines, four previously unpublished environmental datasets were selected and run against both pipelines. These include two environmental samples from sediments at two different locations in the Barents Sea. Two samples, “Muddy” from the southeast of Edgøya (N77 08 40, E26 31 16) and “Sandy” from the intertidal zone at Nordenskioeldøya located in the Hinlopen Strait (N79 12 49, E19 18 58), were collected during two research cruises in 2010 and 2012, respectively. Sequencing libraries were constructed using the Nextera XT DNA Library Preparation Kit and the Nextera XT Index Kit (Illumina Inc). The samples from the Barents Sea was sequenced with the Illumina MiSeq platform using the MiSeq Reagent Kit v2 (500 cycle) with 2 × 250 bp paired-end read length configuration. The other two samples were from moose and sea urchin; frozen moose faeces found in Grunnfjorden (N69 59 42, E19 36 16) and faeces from a seawater tank containing sea urchin at Nofima (Tromsø), respectively. These two latter samples were sequenced using MiSeq Reagent Kit v3 (600 cycle) in a 2 × 300 bp paired-end read length configuration (Supplementary Table 1). The metagenomic sequence reads have been deposited at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the sample accession numbers ERS624612 (muddy), ERS624613 (sandy), ERS624611 (moose) and ERS738393 (sea urchin).

Ribosomal RNA filtering and assembly

Filtering of homologous sequences may reduce the assembly complexity and the number of misassembled contigs. Since META-pipe, contrary to EMG, uses assembled reads (contigs) for functional analysis, we wanted to investigate the effect of removing small-subunit (SSU) ribosomal RNAs (rRNA) sequences before assembly.

To filter prokaryotic rRNA reads, including 5S, 16S and 23S rRNA, hidden Markov models (HMMs) from the rRNASelector were implemented as a part of the functional annotation pipeline. HMMs identify metagenomic fragments coding for rRNA genes if they meet the following two conditions: (i) a sequence read shows an overlap (>60 bp) with an rRNA HMM profile and (ii) the E-value is below 10^{-5}. Fragments satisfying these conditions were selected. The unselected fragments are stored for subsequent assembly and functional analyses.

All datasets, with or without rRNA filtering, were assembled using MIRA (version 4.0.2) in *de novo* mode, with kmer 31 and forced non-IUPAC bases[15].

Before filtering and assembly, the datasets were quality checked with FastQC (version 0.11.3; available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and filtered with PRINSEQ (version 0.20.4)[16] (parameters: –trim_left 10 –trim_right 10 –min_len 50 –ns_max_p 10) for all datasets. Additionally, datasets with particularly low quality at the 3’ end (under Q20) were trimmed using parameters –trim_qual_right 20.

Evaluation of metagenome assemblies

For evaluation of the rRNA filtering step, the assemblies with or without filtering, MetaQUAST v3.2 was used[17]. For the two sediment samples, MetaQUAST was run in the reference-based evaluation mode with an in-house generated marine reference database, MarRef, as reference genomes. MarRef consists of 337 manually curated complete prokaryotic genomes (unpublished, curated by Terje Klemetsen), with a total length of 1135 Mb (Supplementary Table 2). For the two faecal metagenomes, moose and sea urchin, MetaQUAST was run in *de novo* evaluation mode. In this case, instead of using a reference database, MetaQUAST downloads reference sequences automatically based on rRNA sequence alignments. To do so, MetaQUAST searches the SILVA rRNA database using BLASTN with contigs as queries, thereby identifying species present in the dataset. The genomes of these species are then downloaded from NCBI and used as a reference database for assembly evaluation. For these latter samples, MetaQUAST identified 64 reference genome sequences with a total length of 262.8 Mb (Supplementary Table 3).

Taxonomic classification

The four datasets selected for comparison were run on both pipelines with default parameters. The “Muddy” and the “Moose” dataset were analysed in depth, as we wanted to examine any particular differences using the two pipelines with respect to both marine and gut biomes. EMG and META-pipe both use rRNASelector for selecting rRNA sequences from metagenomics shotgun reads. META-pipe uses LCAClassifier with default parameters (LCA relative range: 2%; minimum bit score: 155) for rRNA

annotation, which uses the manually curated Silva*Mod* database – a database based on the taxonomical annotation used in SILVA[9] SSURef NR release 106. The Silva*Mod* also includes annotations to the NCBI taxonomy database to increase resolution of eukaryotic classifications based on mitochondrial and plastid 16S rRNA sequences. It offers resolution down to genus rank and has been shown to perform especially well on environmental datasets[8]. EMG uses QIIME for taxonomic classification, with GreenGenes[7] version 13.8 database as a reference for the classification (default closed-reference OTU picking protocol with reverse strand matching enabled). Unique taxa identified were counted for each analyses and results was visualized using Krona charts[18].

Annotation: Gene prediction and functional assignment

META-pipe uses MetaGeneAnnotator (MGA) for prediction of protein-coding (CDS) regions in contigs longer than 500 bp after assembly with MIRA. The MGA uses a self-training model from input sequences for predictions, in addition to statistical models of bacterial, archaeal and prophage genes. The MGA not only sensitively detects typical genes, but also detects atypical genes, such as horizontally transferred genes and prophage genes in prokaryotic sequences. EMG uses FragGeneScan, which combines sequencing error models and codon usages in a hidden Markov model for the prediction of protein-coding region, regardless of species.

For functional assignment of predicted CDSs, EMG uses a sub set of InterPro release 50.0 (Pfam[19,20], TIGRFAM[21], PRINTS[22,23], PROSITE patterns[24], CATH-Gene3d[25]), while META-pipe uses the full InterPro release 5.10-50.0, in addition to BLAST against PRIAM version 2.0 and UniProtKB release 2014_09 databases. Gene ontology (GO) terms for all predicted CDSs in the "Muddy" dataset obtained from InterProScan5 were converted to GO-slim terms using OBO-files maintained by the Gene Ontology Consortium[26,27], and used for functional comparison between META-pipe and EMG. This dataset was selected for in-depth functional comparison to emphasize the marine topic of this pilot project.

Results and discussion

Assembly quality assessment of rRNA filtering

Assembly of metagenomics reads is a complex and challenging task, due to both the computational overheads and biological complexity. Near-identical sequences, such as mobile genetic elements, homologous genes and conserved regions, combined with high diversity, low coverage and short reads, often results in errors and chimeric assembly. To analyse the effect of filtering rRNA in the assembly process in META-pipe, we assembled four datasets with and without rRNA reads.

Using MetaQUAST to evaluate the effect of rRNA filtering (Table 1), we observed a marginal reduction in total number of contigs and total length on the rRNA filtered datasets compared to unfiltered. Both marine sediment rRNA unfiltered datasets contain more possible misassembled contigs compared to the corresponding rRNA filtered dataset, 4 and 8 for Muddy and Sandy datasets, respectively. While the Muddy sample contains no misassemblies, the Sandy dataset contains four misassembles, where flanking sequences may align to different reference genomes, overlap or aligns over 1kb away from each other. Similarly, the unfiltered rRNA datasets have more mismatches and indels compared to the two filtered rRNA assemblies. For the Muddy sample, a reduction of mismatches and indels by a factor of 4 was observed for the filtered dataset, while the Sandy gave a reduction of a factor of 3. We believe these mismatches and indels stem from the inherent conservation of rRNA sequences, which causes spurious contigs in assembly.

Table 1. Effect on assembly with rRNA filtering.

	Muddy[1]	Muddy ΔrRNA[1]	Sandy[1]	Sandy ΔrRNA[1]	Moose[2]	Moose ΔrRNA[2]	Sea Urchin[2]	Sea Urchin ΔrRNA[2]
# Contigs > 0 bp	267 433	266 814	148 228	147 928	973 097	972 462	1 010 649	1 010 610
# Contigs > 500 bp	25 581	25 302	25 294	25 138	211 333	210 348	114 307	114 189
# Contigs > 1000 bp	5 659	5 572	6 213	6 118	57 779	57 147	32 593	32 433
Total length (bp)	25 155 475	24 822 906	25 301 011	25 038 101	248 491 504	246 880 376	143 551 468	143 216 805
Aligned to reference (bp)	291 296 (0.003%)	181 694 (0.001%)	444 398 (0.009%)	286 608 (0.004%)	4 266 146 (0.099%)	3 504 213 (0.083%)	11 923 718 (0.284%)	11 615 439 (0.262%)
N50 (contigs>500)	931	930	976	972	1214	1211	1287	1285
# Misassemblies	0	0	4	1	10	4	38	20
# Possible misassembled contigs	4	2	8	1	23	5	89	68
# Mismatches	918	214	2 836	1 093	5 609	4 202	15 674	14 132
# Indels	83	19	277	96	278	167	1003	814

[1]MarRef database length: 1 135 Mb, [2]MetaQUAST downloaded reference database length: 262 Mb.

A very low percentage of the assembled contigs from the marine sediment samples mapped to the reference genomes (0.001% – 0.009%). However, as MarRef is still relatively small compared to the huge diversity estimated in marine sediments, the low percentage of mapped contigs is not surprising. Consequently, it is difficult to achieve a thorough estimate of misassemblies, mismatches and indels simply because of poor reference coverage. We believe that the number of misassemblies will increase as the marine reference database increases. The marine sediment datasets were also tested using MetaQUAST in *de novo* evaluation mode, where references are identified and downloaded automatically. MetaQUAST generated a reference database of 40 genomes, but assembled contigs only mapped to one of these identified references (in comparison to 159 out of 337 using the in-house marine reference database).

For the faecal datasets, contigs are longer, and more contigs mapped to the reference database (0.083% – 0.286%), which probably is a consequence of higher coverage. However, the number of misassemblies, possible misassemblies, mismatches and indels increased significantly, although the MetaQUAST generated reference database for these samples were considerably smaller than MarRef.

Removal of rRNA before assembly clearly reduces misassemblies, possible misassembled contigs, mismatches and indels, but the lack of specific marine databases hampers the comparison and benchmarking of the different approaches using MetaQUAST.

Taxonomic classification

In general, META-pipe with the LCAClassifier/Silva*Mod* configuration identifies more unique taxa for the marine sediment datasets, while EMG, using Qiime/GreenGenes, identifies more taxa for the faecal datasets, as shown in Table 2. As LCAClassifier generally offers resolution up to genus rank, we also observe that META-pipe is more reluctant to classify at species level, compared to EMG.

Our results are in agreement with Lanzen *et al.*[8], who showed that classification using Silva*Mod* performed better than with GreenGenes, particularly when applied to environmental sequences. META-pipe also offers eukaryotic classifications based on mitochondrial and plastid 16S rRNA sequences. However, in general, SSU rRNA gives limited resolution as a taxonomic marker for eukaryotic sequences compared to internal transcribed spacers (ITS) or large subunit (LSU) rRNA[28].

To obtain a more detailed overview of the difference between the pipelines, we explored the marine "Muddy" dataset and the gut/intestine "Moose" dataset in more depth. While META-pipe was able to predict 6584 16S rRNA sequences, EMG predicted 4339 in the "Muddy" dataset (Figure 2). For the "Moose" dataset, META-pipe predicted 43949 and EMG predicted 25018 (Figure 3). As this step is in practice identical for both pipelines, the dissimilarities in rRNA prediction stems from the preprocessing step in EMG, where overlapping reads are merged and the total read count reduced from 18 to 12 million. Reduction of input sequence reads by one third also reduces predicted rRNA sequences by the same fraction. Although there were dissimilarities in the number of predicted 16S rRNA, the most apparent difference observed between the pipelines was the fraction of unassigned sequences.

In the "Muddy" dataset, EMG classified 2500 sequences (58%), while META-pipe was able to classify 6119 (93%). However, if we ignore unassigned and eukaryotic sequences from the data from META-pipe, most high-level nodes in the taxonomy hierarchy have comparable relative fractions, e.g. like *Planctomycetes*, *Bacteroidetes*, *Acidobacteria*, *Choloflexi*, *Nitospirae* and *Actinobacteria*. The largest inconsistencies were in the *Archaea* and *Protobacteria* were EMG assigned 3.5% (89) and 57.9% (1448), respectively, to these nodes, while META-pipe assigned 5.3% (320) and 49.3% (2913). For the "Moose" dataset EMG classified 15630 sequences (62%), while META-pipe classified 41130 (94%). As in the "Muddy" dataset, trends are similar when ignoring unassigned and eukaryotic fractions, with most identified taxa showing only marginal differences between the two pipelines. The discrepancy observed between the EMG and META-pipe in prediction of rRNA sequences and taxonomic classification relies heavily on the methods, parameters and settings, and the underlying databases used, meaning a more thorough benchmarking of the different methods and databases are needed to determine the sensitivity, specificity and accuracy.

Although we observe comparable results from the taxonomic classification, there is a need for benchmarking of tools for rRNA prediction and classification in addition to dedicated rRNA databases for the marine domain.

Table 2. Number of identified unique taxa. Numbers in parenthesis includes eukaryotic hits classified by META-pipe.

Dataset	Muddy		Sandy		Moose		Sea urchin	
Pipeline	**META-pipe**	**EMG**	**META-pipe**	**EMG**	**META-pipe**	**EMG**	**META-pipe**	**EMG**
Phylum	41 (57)	40	22 (38)	16	16 (26)	17	18 (54)	13
Class	67 (88)	83	38 (62)	36	22 (39)	33	38 (97)	30
Order	126 (150)	111	69 (98)	61	32 (51)	42	72 (135)	68
Family	113 (138)	97	93 (115)	84	44 (62)	61	88 (157)	107
Genus	111 (129)	79	138 (155)	120	79 (92)	85	160 (227)	170
Species	6 (11)	19	31 (40)	38	5 (14)	30	61 (102)	73

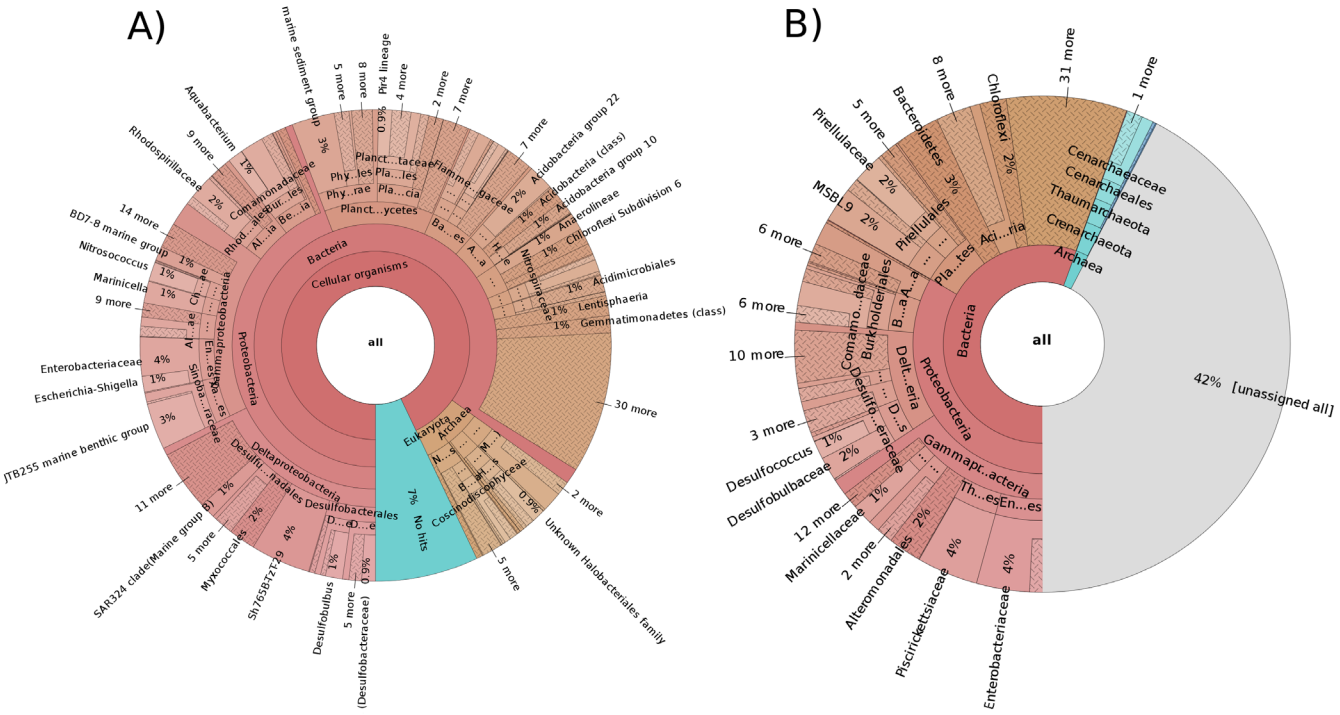

Figure 2. Krona chart representation of taxonomic classification of the "Muddy" dataset from META-pipe (**A**) and EBI Metagenomics Portal (**B**) pipelines.ppl.

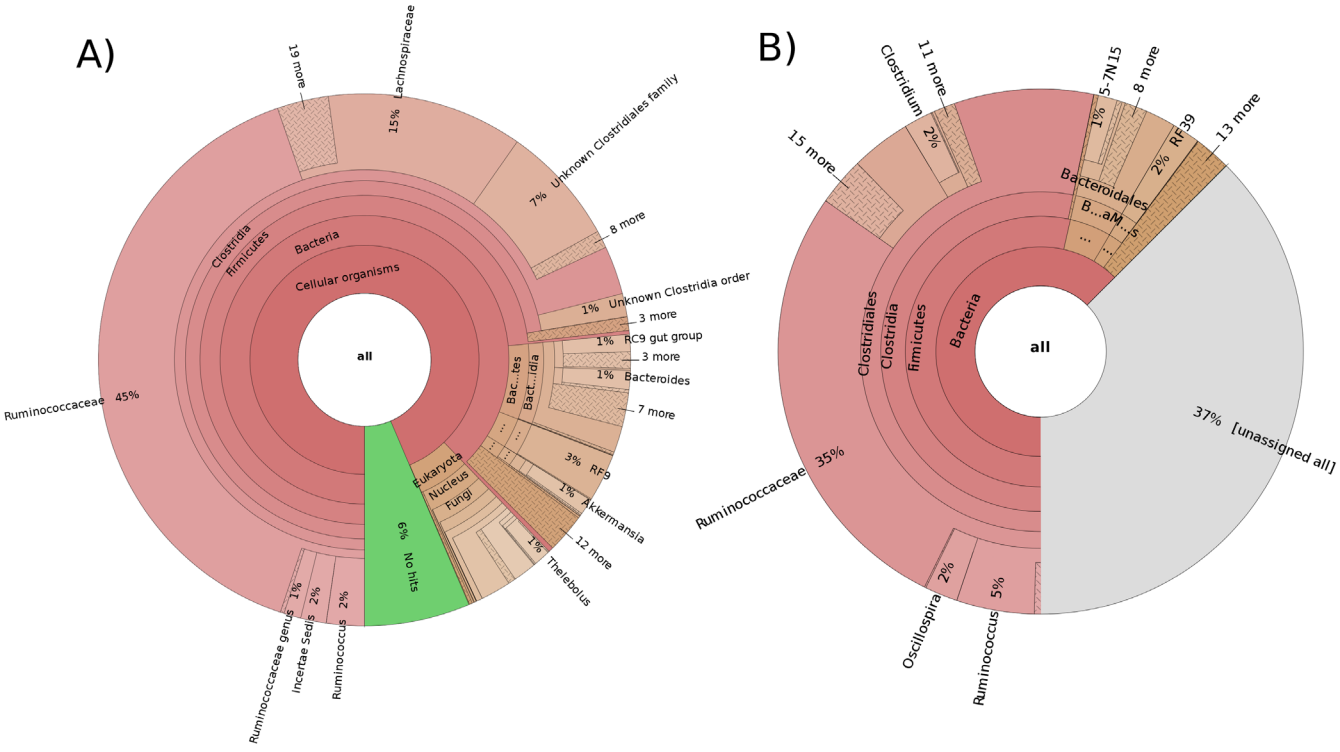

Figure 3. Krona chart representation of taxonomic classification of the "Moose" dataset from META-pipe (**A**) and EBI Metagenomics Portal (**B**) pipelines.

Functional analysis

To gain more insight into the effect of assembling compared to merging paired-end sequencing reads before CDSs prediction and functional assignment, we compared the output results from the functional analysis of the "Muddy" sample (ERS624612) from META-pipe and EMG.

In short, we expect the difference will manifest in three different ways when comparing outputs from the two pipelines. Firstly, longer or full length predicted CDSs would give rise to better functional assignment than shorter CDSs. Secondly, since an assembly will reduce the relative coverage to a consensus sequence (contig), the results will not be quantifiable in the same way as an analysis performed on single or merged reads. Thirdly, assembly will reduce the number of candidate CDSs to a subset containing CDSs from the most abundant organisms in the dataset, depending on the complexity of the dataset, sequencing technology and assembly quality.

EMG predict 11 572 617 CDSs (from 12 103 194 merged reads), while META-pipe predicts a total of 47 434 CDSs (from 25 581 assembled contigs > 500 bp), which accounts for 0.4% compared to EMG. We explored the distribution of predicted gene lengths from both pipelines (Figure 4). On average, META-pipe predicts genes of 155 amino acids in length and the longest gene is 1996 amino acids, while EMG predicts genes of 73 amino acids in length and the longest gene is 162 amino acids.

EMG predicts approximately 1.0 CDS per merged read, while META-pipe predicts 1.9 CDSs on average per contig. Not surprisingly, the longer the contigs the more CDSs predicted, but what effect does this have on the functional assignment of each CDSs and the microbial community as a whole? To answer this question we compared the accumulated number of GO-slim annotations for each analysis using the "Muddy" dataset. In general, the more GO-slim annotations for each CDS, the better description of the molecular function, biological process, and cellular component of gene products will be obtained.

EMG provided a total of 28 942 422 accumulated GO-slim annotations for the predicted CDSs, while META-pipe only provided 565 125 accumulated annotations, which accounts for 0.2% compared to EMG. However, META-pipe provided on average 11.9 GO-slim annotations per CDS, while the number for EMG is 2.5 GO-slim annotations per CDS, which indicated that the longer CDSs predicted a better functional description. Additionally, META-pipe utilizes all available databases shipped with Interproscan5 contrary to the reduced set utilized by EMG, which naturally provides more potential GO annotations per annotated gene. How does this effect the functional assignment of the community as a whole?

As shown in Figure 5, the effects are relatively small on top-level terms when GO-slim annotations are sorted. Most of the top-level terms (e.g. molecular function, biological process and cellular component) in the GO hierarchy ranks similarly due to accumulative counting (accumulated from counts in lower connected nodes in the diacyclic GO-slim graph). Less common terms are ranked somewhat differently e.g. protein binding, cell communication and carbohydrate metabolism. These differences arise from the observed differences in GO-slim annotation for each predicted CDS in the two pipelines. As META-pipe performs assembly, DNA sequences

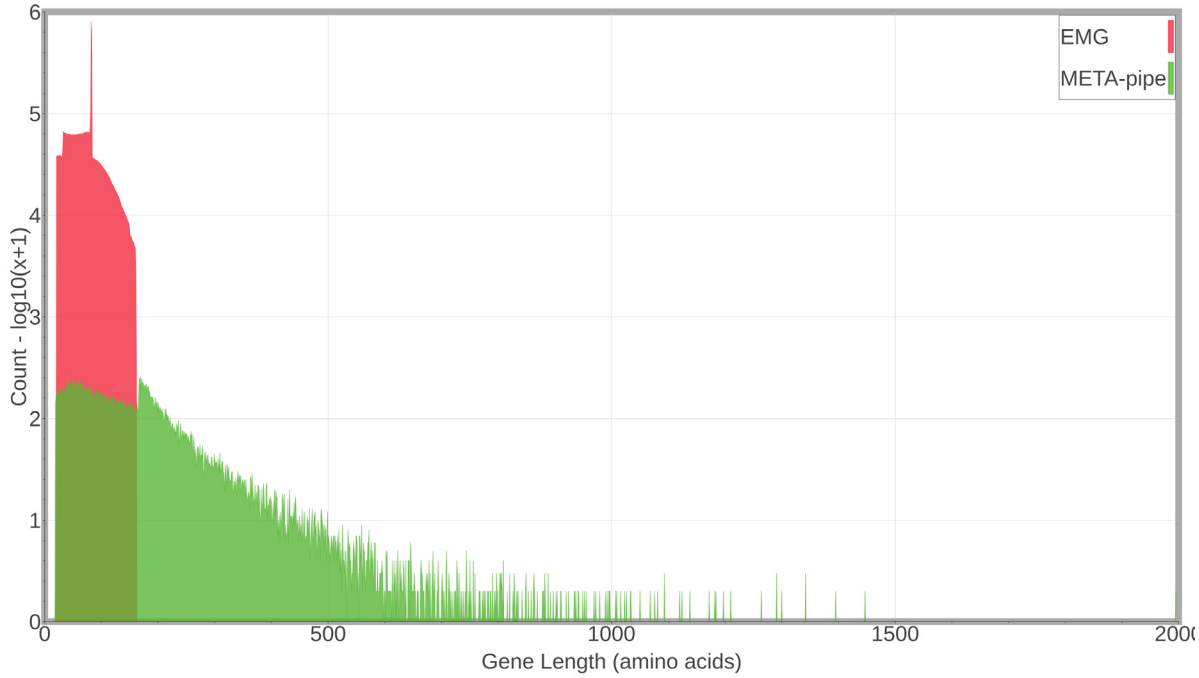

Figure 4. Predicted gene length distribution from META-pipe and EMG pipelines. EMG, EBI Metagenomics Portal.

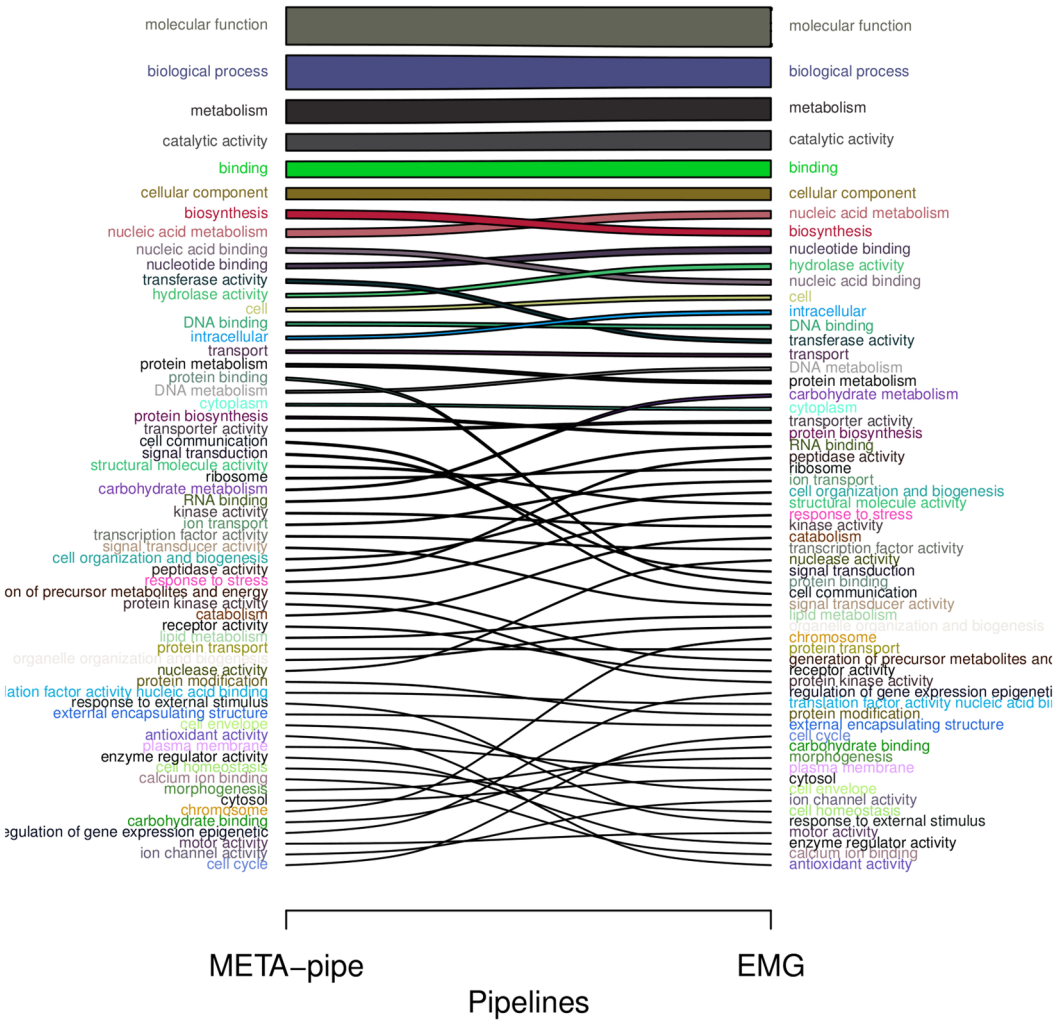

Figure 5. Comparison of counted GO-slim annotations from META-pipe and the EMG pipeline. Thickness of bars corresponds to fraction size of accumulated GO-slim annotations for each pipeline. GO, Gene Ontology; EMG, EBI Metagenomics Portal.

from low abundance organisms will effectively get excluded from the functional analysis due to insufficient coverage, which in turn changes to GO-profile compared to the EMG-analysis.

The functional assignment of the "Muddy" sample is comparable between EMG and META-pipe. However, a more thorough analysis has to be performed understand the differences observed in low-level GO-terms.

Harmonization and interoperability

Throughout the project, several changes and improvements were implemented to harmonize, shorten the process time and enrich the output of the two pipelines. In short, masking of homologous sequences before assembly to reduce misassemblies, new databases to enhance functional annotation, optimization and modifications of databases to reduce wall-time. We identified several key steps and file formats within the respective workflows of each pipeline where intermediate data could be interchanged, allowing for potential interoperability between pipelines. Both pipelines have seen improvements since the start of this project. The EMG pipeline is now in version 3.0[29], while a new version/redesign of Meta-pipe is currently in development to improve computational constrains and functionality.

Gap analysis

In order to develop sustainable ELIXIR services for marine metagenomics, we performed a gap analysis and concluded on four areas where actions are urgently needed. These include the need to: i) standardise metagenomics data generation; ii) establish marine metagenomics resources; iii) develop gold standard pipelines for metagenomics analysis; and iv) explore HPC and storage technologies. A short description on the four recommendations follows.

Metagenomics data standards. The context in which marine metagenomics projects are conducted often gets lost, since these

data are rarely submitted along with the sequence data. If these contextual data are missing, key opportunities for comparison and analysis across studies and environments are hampered or even impossible to conduct. A metagenomics study should report on each processing step, from contextual data of sampling, through experimental variables of sequencing and metadata of sequence analysis to parameters associated with archiving of the analysed data. Over the past five years standards for describing how a sample was captured and sequenced e.g. for sampling and environment packages, these standards need to be extended to include the whole metagenomics experimental workflow from sample gathering to computational results. As illustrated above, analysis pipelines produce different results on the same input, and comparing the results and understanding whether the differences are real, i.e. coming from the biology of the system under investigation, or whether they are artefacts of the analysis methods, is non-trivial to disentangle.

Marine metagenomics data resources. Marine metagenomics research and innovation is limited by the lack of dedicated reference data resources. As indicated by the use of GreenGenes in EMG, existing reference databases are generalized or biased and the contextual data for the records is often incomplete or lacking. Due to the lack of coverage of marine organisms in existing databases, only about one quarter of sequences can be annotated from typical marine samples. To improve the characterization of marine environmental samples, establishment of dedicated data resources for the marine microbial domain is highly needed.

Gold standards pipelines. As with most emerging bioinformatics fields, a myriad of tools that perform different types of metagenomics analysis are constantly being published or updated. Pipelines that aggregate such tools are therefore under constant flux. Knowing which tool is the most appropriate to use for specific tasks can be difficult to assess, particular for new researchers entering the field. There is a need to evaluate several types of analysis tool (e.g. preprocessing of reads, prediction of CDSs and taxonomic assignment), and defining gold standard tools and databases.

High-performance computer and storage technologies. Marine genomic datasets vary in size, ranging from tens of gigabytes for the typical datasets, to terabytes for projects such as Tara Ocean[30], OSD[31] and Malaspina (Information available at http://scientific.expedicionmalaspina.es/). Although some pipelines, such as EMG and META-pipe, have been designed for parallel execution on high-performance computer (HPC) clusters, there is a need for exploring more elastic storage and computation resource allocation, e.g. on academic or commercial clouds.

Conclusion

While there are differences in the respective approaches, EMG and META-pipe provide comparable results. They have their own strengths and weaknesses, and it is clear that the optimal solution for the community would be harmonization and interoperability between the analysis platforms. There is still a need for improvements, e.g. harmonization of the preprocessing step, and improvement of eukaryote taxonomic classification by implementing reference databases for internal transcribed spacers (ITS) and/or large subunit (LSU) rRNAs.

The outcome of the gap analysis has been disseminated to the ELIXIR-EXCELERATE Marine metagenomic infrastructure use case (https://www.elixir-europe.org/excelerate/marine), which will help to define the requirements and specifications for the establishment of a sustainable ELIXIR marine metagenomics infrastructure.

Data and software availability

EBI Metagenomics Portal (EMG): https://www.ebi.ac.uk/metagenomics/

META-pipe: https://galaxy-uit.bioinfo.no (Needs academic user affiliation (FEIDE) or NeLS user login)

The metagenomic sequence reads are available from the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the sample accession numbers ERS624612 (muddy), ERS624613 (sandy), ERS624611 (moose) and ERS738393 (sea urchin).

Author contributions

EMR, RDF and NPW drafted the manuscript. EMR, HD and AM conducted all the experiments. All authors read, revised and approved the manuscript.

Competing interests

No competing interests were disclosed.

Grant information

Funding was provided from ELIXIR, EMBL-EBI and UiT The Arctic University of Norway.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We would like to thank Concetta de Santi for isolating DNA from the environmental samples and Seila Pandur for running the MiSeq sequencer.

Supplementary materials

Supplementary Table 1: Datasets used in the present analysis of the two pipelines.

Click here to access the data.

Supplementary Table 2: Genomes and Genbank accession numbers included in the in-house marine reference database MarRef.

Click here to access the data.

Supplementary Table 3: References identified using rRNA in MetaQUAST.

Click here to access the data.

References

1. Mineta K, Gojobori T: **Databases of the marine metagenomics.** *Gene.* 2016; **576**(2 Pt 1): 724–728.
PubMed Abstract | Publisher Full Text
2. Baker M: **Next-generation sequencing: adjusting to data overload.** *Nat Methods.* 2010; **7**(7): 495–499.
Publisher Full Text
3. Hunter S, Corbett M, Denise H, *et al.*: **EBI metagenomics--a new resource for the analysis and archiving of metagenomic data.** *Nucleic Acids Res.* 2013; **42**(Database issue): D600–6.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Robertsen EM, Kahlke T, Raknes IA, *et al.*: **META-pipe - Pipeline Annotation, Analysis and Visualization of Marine Metagenomic Sequence Data.** *ArXiv 160404103 Cs.* 2016.
Reference Source
5. Lee JH, Yi H, Chun J: **rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries.** *J Microbiol.* 2011; **49**(4): 689–691.
PubMed Abstract | Publisher Full Text
6. Caporaso JG, Kuczynski J, Stombaugh J, *et al.*: **QIIME allows analysis of high-throughput community sequencing data.** *Nat Methods.* 2010; **7**(5): 335–336.
PubMed Abstract | Publisher Full Text | Free Full Text
7. DeSantis TZ, Hugenholtz P, Larsen N, *et al.*: **Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.** *Appl Environ Microbiol.* 2006; **72**(7): 5069–5072.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Lanzén A, Jørgensen SL, Huson DH, *et al.*: **CREST--Classification Resources for Environmental Sequence Tags.** *PLoS One.* 2012; **7**(11): e49334.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Quast C, Pruesse E, Yilmaz P, *et al.*: **The SILVA ribosomal RNA gene database project: improved data processing and web-based tools.** *Nucleic Acids Res.* 2013; **41**(Database issue): D590–D596.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Rho M, Tang H, Ye Y: **FragGeneScan: predicting genes in short and error-prone reads.** *Nucleic Acids Res.* 2010; **38**(20): e191.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Jones P, Binns D, Chang HY, *et al.*: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics.* 2014; **30**(9): 1236–1240.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Noguchi H, Taniguchi T, Itoh T: **MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes.** *DNA Res.* 2008; **15**(6): 387–396.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Claudel-Renard C, Chevalet C, Faraut T, *et al.*: **Enzyme-specific profiles for genome annotation: PRIAM.** *Nucleic Acids Res.* 2003; **31**(22): 6633–6639.
PubMed Abstract | Publisher Full Text | Free Full Text
14. UniProt Consortium: **UniProt: a hub for protein information.** *Nucleic Acids Res.* 2015; **43**(Database issue): D204–D212.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Chevreux B, Pfisterer T, Drescher B, *et al.*: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.** *Genome Res.* 2004; **14**(6): 1147–1159.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets.** *Bioinformatics.* 2011; **27**(6): 863–4.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Mikheenko A, Saveliev V, Gurevich A: **MetaQUAST: evaluation of metagenome assemblies.** *Bioinformatics.* 2016; **32**(7): 1088–90.
PubMed Abstract | Publisher Full Text
18. Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser.** *BMC Bioinformatics.* 2011; **12**: 385.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Bateman A, Birney E, Durbin R, *et al.*: **The Pfam Protein Families Database.** *Nucleic Acids Res.* 2000; **28**(1): 263–266.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Finn RD, Coggill P, Eberhardt RY, *et al.*: **The Pfam protein families database: towards a more sustainable future.** *Nucleic Acids Res.* 2016; **44**(D1): D279–D285.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res.* 2003; **31**(1): 371–373.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Attwood TK: **The PRINTS database: A resource for identification of protein families.** *Brief Bioinform.* 2002; **3**(3): 252–263.
PubMed Abstract | Publisher Full Text
23. Attwood TK, Coletta A, Muirhead G, *et al.*: **The PRINTS database: a fine-grained protein sequence annotation and analysis resource—its status in 2012.** *Database (Oxford).* 2012; **2012**: bas019.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Sigrist CJ, de Castro E, Cerutti L, *et al.*: **New and continuing developments at PROSITE.** *Nucleic Acids Res.* 2013; **41**(Database issue): D344–347.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Buchan DW, Shepherd AJ, Lee D, *et al.*: **Gene3D: Structural Assignment for Whole Genes and Genomes Using the CATH Domain Structure Database.** *Genome Res.* 2002; **12**(3): 503–514.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Ashburner M, Ball CA, Blake JA, *et al.*: **Gene Ontology: tool for the unification of biology.** *Nat Genet.* 2000; **25**(1): 25–29.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Gene Ontology Consortium: **Gene Ontology Consortium: going forward.** *Nucleic Acids Res.* 2015; **43**(Database issue): D1049– D1056.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Santamaria M, Fosso B, Consiglio A, *et al.*: **Reference databases for taxonomic assignment in metagenomics.** *Brief Bioinform.* 2012; **13**(6): 682–695.
PubMed Abstract | Publisher Full Text
29. Mitchell A, Bucchini F, Cochrane G, *et al.*: **EBI metagenomics in 2016 -- an expanding and evolving resource for the analysis and archiving of metagenomic data.** *Nucleic Acids Res.* 2015; **44**(D1): D595–603.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Karsenti E, Acinas SG, Bork P, *et al.*: **A Holistic Approach to Marine Eco-Systems Biology.** *PLoS Biol.* 2011; **9**(10): e1001177.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Kopf A, Bicak M, Kottmann R, *et al.*: **The ocean sampling day consortium.** *GigaScience.* 2015; **4**: 27.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ? ✔ ?

Version 1

Referee Report 31 May 2017

doi:10.5256/f1000research.11253.r19596

? **Anders Blomberg**
Department of Marine Sciences, Lundberg Laboratory, University of Gothenburg , Gothenburg, Sweden

Metagenomics has a great potential to influence our understanding of the complex ecology of biotopes, including marine waters. Despite the impressive speed of generating sequence data, the analyses pipelines are not as well developed and standardized. This article describes a comparison between two analysis pipelines and how they perform on different types of sequence data. The main methodological difference between the two pipe-lines tested is if the anaysis is done on the read-level (EMG) or at the contig-level (META-pipe). This will of course have a major influence on the results obtained, which is in essence what this study aims to outline. The manuscript also has a link to a nice webinar that explains parts of the background, technical details, challenges and some of the results.

Major critique/comments:

1. Why a specific marine metagenomics pipeline? Why could not this service be generic - independent on where the organisms live (marine, soil, stomach, flowers, etc....).

   This issue is addressed in the webinar, but not in the paper, e.g. marine samples/sequences are taxonomically complex and with really high genetic/sequence diversity. There might be more reasons. These reasons for a specific marine pipeline should be outlined in 1-2 sentences in the paper.

2. Why picking unpublished data for the test? Anything specifically general with this data? Or could there be very specific biases and technical problems with this data? This should be outlined and described. They should also consider using some already published data for their comparison.

   In addition, the data analysed is based on comparably long-read Illumina reads - 250nt and 300nt. Plenty of metagenomics data has been, and will be, collected using more standard length reads (≈ 125nt). Please discuss, e.g. in the conclusion part, to what extent this selection of example data could have had an impact on the obtained results.

3. In the Conclusion section they state: " While there are differences in the respective approaches, EMG and META-pipe provide comparable results. "

   But do they really show similar results? There appear to exist huge differences between the two programs that are also highlighted earlier in the text:

   p.6, rc, " While META-pipe was able to predict 6584 16S rRNA sequences, EMG predicted 4339 in

the “Muddy” dataset (Figure 2)."

p.6, rc, " In the “Muddy” dataset, EMG classified 2500 sequences (58%), while META-pipe was able to classify 6119 (93%)."

p.8, lc, " EMG predict 11 572 617 CDSs (from 12 103 194 merged reads), while META-pipe predicts a total of 47 434 CDSs (from 25 581 assembled contigs > 500 bp), which accounts for 0.4% compared to EMG.

p.8, rc, " EMG provided a total of 28 942 422 accumulated GO-slim annotations for the predicted CDSs, while META-pipe only provided 565 125 accumulated annotations, which accounts for 0.2% compared to EMG."

I think there statement about "comparable results" should be modified and differences also highlighted in the Conclusion.

4. Finally, should one recommend that both pipelines are used in analyses before publication, and that results are being reported? And if so, what about other pipelines? How do they see that this challenge (which is a great problem in the comparison of results between studies using different analysis pipelines) should be handled in the future?

Minor comments:

Abstract
In the last sentence of the abstract it says:
"In this paper, we summarize some of the results from the ELIXIR pilot action “Marine metagenomics – towards user centric services”. Shouldn't this be the same as in the title?

Page 3, left column (lc), line 7
I am not sure I see why replication would be hard given the information in publications - databases are not per see a guarantee for higher transparancy in information handling. Even if the access to the data might be easier. The statement should be modified. Or do they mean "results" and not "analyses" are hard to replicate?

p.3, lc, l.15
Please provide a short overview of the types of metagenomics pipeline that are available at this stage. Please explain to the reader why EMG and META-pipe were selected for comparison? Anything that make this comparison particularly valid?

p.4, lc, l.3
preform - > perform

p.4, rc, l.18
Give arguments for why Kmer = 31 was selected. Are there reasons to believe the results would have been different if another Kmer had been used?

p.4, right column (rc), l.4 from bottom
Please explain "biomes".

p.6, lc, l.33
They state that META-pipe is reluctant to classify on species level. Can one explain to the reader why that is?

p.6, lc, l.35
Be more specific - how was "better" defined?

Table 2
Just to be sure - do they mean prokaryotic + eukaryotic?

Table 2
Why can't EMG do eukaryotes?

Figure 4
What do we really learn from this figure?

Figure 5
How have the annotations been sorted?

p.9, lc, l.3
to - > the

p.10, lc, l.10 from bottom
They talk about gold standard tools. But these might differ dependent on the data - technical problems; community complexity; lengths of reads; .... Can they be a bit more specific for how they see that this "golden standard" can be reached?

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
No

Competing Interests: No competing interests were disclosed.

Referee Expertise: Functional genomics, genomics, databases, yeast genetics, yeast phenomics, marine biology, osmoregulation

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 18 April 2017

doi:10.5256/f1000research.11253.r21907

Takashi Gojobori
Computational Bioscience Research Center (CBRC), Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Title and Abstract
The title can be more precise in reflecting the contents. For example, “ELIXIR pilot action: Comparisons of two representative pipelines of metagenomics between EMG and META-pipe”.

In Abstract, the outcome of the two pipeline comparisons may be mentioned.

Article Content
The design, methods and analysis of the results from the study been explained well and they are appropriate for the topic being studied. However, I have the following comments:

1. In the studies of metagenomics, there are essentially two ways of picking up genomic fragments from the DNA samples: (a) Amplicon-oriented approach and (b) Random shotgun approach. For approach (a), rRNAs can be targets for sequencing whereas for approach (b) any genomic fragments can be targets. Naturally, approach (a) can be used for phylogenetic identification only, while approach (b) can be used for not only phylogenetic identification but also functional analysis.

   If the authors explain about these two approaches clearly in the text, the whole context can be easier to understand to the readers who do not have the expertise of metagenomics.

2. Marine metagenomics is one of metagenomic studies. The two pipelines explained in the present paper are basically for metagenomics in general, but not specialized for “marine” metagenomics. Then, the authors may be requested to clarify which points in the pipelines or in the methodologies are keen differences between metagenomics in general and “marine” metagenomics.

3. Although the main part of the present paper is on differences between the two pipelines. In particular, whether the genomic fragment assembly is conducted or not appears to lead to huge differences of the outcome between the two pipelines. This is very important notion. According to our experiences of marine metagenomics, we have already recognized that the fragments assembly produces more identifiable OTUs and function in the annotation process. Therefore, we usually conduct the fragment assembly.

   However, the authors did not mention anything about it in Conclusion (Also see below).

Conclusion
In Conclusion, the authors have stated that while there are differences in the respective approaches, EMG and META-pipe provide comparable results. I do not think so. As mentioned above, the author should state the main different point between two pipelines, because they are not really “comparable”.

Data
I think that enough information has been provided to be able to replicate the experiment.I also think that the data are in a usable format/structure and all the data have been provided.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 13 February 2017

doi:10.5256/f1000research.11253.r19670

? **Marla I. Trindade**[1] , **Leonardo van Zyl** [2]
[1] Institute for Microbial Biotechnology and Metagenomics (IMBM), Department of Biotechnology, University of the Western Cape, Bellville, South Africa
[2] Department of Biotechnology, University of the Western Cape, Bellville, South Africa

The authors, using 4 different metagenome datasets, compare the assembly and annotation results of 2 pipelines (META-pipe and EMG). They additionally compare the assemblies after also filtering out rRNA reads.

As presented by the authors, the discrepancy observed between the 2 pipelines, particularly in predicting the taxonomic classification, was significant, and alarmingly different. While such a limitation is acknowledged by many published studies, and accepted in practice, I 100% agree with the authors that it is crucial for rigorous benchmarking to be conducted, and they rightly conclude that action is urgently needed - it goes against all research principles to present conclusions and hypotheses when results are generated using tools and methods which are accepted to completely bias the outcome. What actual scientific value do such studies offer given the discrepancies? Thus, for this reason, the study by Robertsen et al should be indexed so as to raise further awareness to the danger of not paying attention to the proper curation of next gen sequence data generation and analysis.

I do however have some major and minor questions / corrections which need to be addressed before this study can be approved:

Major:

1. The authors provide absolutely no information of how the samples were collected and how the DNA was prepared. Given that the authors in their gap analysis, themselves recognise that this kind of detail should be reported in every metagenomic study, it is ironic that they do not do so even if the purpose of the manuscript was not to describe the content of these metagenomes. Furthermore, the Gap analysis should also address the biases that the mDNA extraction and sequencing technology introduce, which are also well documented.

2. Please confirm whether the figures presented for the "aligned to reference" in Table 1 represents the combined results when using the MarRef and the de novo generated database? It is confusing because in the methods section it is stated that the sediment samples were analysed using the MarRef whereas the others were done using the MetaQUAST-generated database; however, in the results section the marine samples are reported to have been analysed using both databases. Also, what were the cut-off values used for the reference alignments, and were parameters modified to try and increase the %assemblies, or do these represent the best possible outcome?

3. Irrespective of the answer in the above, the % that could be referenced was very low (the highest was 0.262%), representing a minute proportion of the sequence data generated. If I understood correctly, the #misassemblies refers to only the % of sequences that could be referenced (ie. a minute proportion of the sequence data), and thus I do not think this is a very informative factor with which to compare the different assembly procedures. i.e. if misassembly is only judged on between 0.001% and 0.284% of the dataset this might not be an accurate reflection of assembly issues, as suggested by the authors.

4. The choice of assembler and assembly parameters could have a huge impact on the outcome on the META-pipe pipeline. Have the authors, and pipeline administrators, satisfied themselves that MIRA is the best assembler (SPADES, IDBA-UD, collaboration with CLC Genomics?). The fact that when you had longer contigs (faecal dataset), the number of missasemblies increased significantly, points to assembly issues. See a very recent study be Hesse et al 2017 (with relevant references within) which specifically addresses these issues.

5. The authors use a non-validated dataset to determine the best pipeline (or at least differences). It would’ve been preferable to use a curated dataset of known composition, to know exactly what the ideal outcome should’ve been. A good example of this is the taxonomic assignments of the two pipelines. The authors finish paragraph 7 on page 6 with “more thorough benchmarking of the different methods and databases are needed to determine the sensitivity, specificity and accuracy”. Had they used a curated dataset, they would know which pipeline gives a more accurate picture of the taxonomic composition of the uploaded dataset. The study suffers the same issue when looking at functional classification.

6. I agree that dedicated effort and resources needs to be established to ensure increased population of the sequence databases with marine derived genome data. Please can the authors clarify whether they are proposing there to be dedicated marine databases - if this is the case I do not support this notion. From an ecological perspective it would be interesting to make connections with terrestrial systems, if they exist. Only a comprehensive database could help you establish these links. If purely for bioprospecting, perhaps this is less of an issue. Or did the authors simply mean that increased research effort / focus is needed?

7. In their gap analysis the authors propose the need for evaluating analysis tools and defining gold standard tools and databases. Standardization of metagenomic data generation is a great idea, similar to the MIQE guidelines for real-time PCR, but very difficult to implement in practice. The pipelines should always allow some flexibility to accommodate datasets outside the “ideal”. Can they propose who should take on this responsibility, or how to coordinate such a task?

8. No comparison was made to MG-RAST (278,783 metagenomes), one of the most widely used metagenomic analysis pipelines. If a “gold standard” pipeline is to be created, surely it should also be done in consultation with all large groups involved with such analyses.

9. If I understand Figure 5 correctly, the more parallel lines we have the more similar the predictions of the two pipelines? Can the authors provide some quantitative value to summarize the information given in the Figure? Otherwise the display of the results of this analysis seems a bit arbitrary as my only other two references are a figure with all parallel lines (1), and one where there are no parallel lines (0). I acknowledge this may not be easily doable.

Minor:

1. Pg4, line 13: “OUT” change to “OTU”

2. Pg5, end of the first paragraph: "and results was visualized" should be changed to "and results **were** visualized".

3. Pg5, last paragraph: Figure 4 does not really add value. What it displays is fully expected and will only change for the META-pipe pipeline depending on the environment being sequenced as well as “depending on the complexity of the dataset, sequencing technology and assembly quality”.

4. Pg8, 3rd paragraph: "the longest gene is 1996 amino acids". Although it is understood what the authors mean, it is incorrect terminology and this needs to be corrected in this paragraph and also in the Figure 5 legend. A gene is denoted in base pairs, and protein in amino acids.

5. Pg9, paragraph 2: "analysis has to be performed understand" needs to be changed to "analysis has to be performed **to** understand".

6. Pg10, 3rd paragraph: "difficult to assess, particular for new researchers" should be changed to "difficult to assess, particular**ly** for new researchers".

Competing Interests: No competing interests were disclosed.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.