BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 3 2015, pages 311–317
doi:10.1093/bioinformatics/btu546

Genome analysis

Advance Access publication October 11, 2014

Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism

Jonathan D. Magasin[1] and Dietlind L. Gerloff[2,*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064 and [2]Foundation for Applied Molecular Evolution (FfAME), Gainesville, FL 32604, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Despite advances in high-throughput sequencing, marine metagenomic samples remain largely opaque. A typical sample contains billions of microbial organisms from thousands of genomes and quadrillions of DNA base pairs. Its derived metagenomic dataset underrepresents this complexity by orders of magnitude because of the sparseness and shortness of sequencing reads. Read shortness and sequencing errors pose a major challenge to accurate species and functional annotation. This includes distinguishing known from novel species. Often the majority of reads cannot be annotated and thus cannot help our interpretation of the sample.

Results: Here, we demonstrate quantitatively how careful assembly of marine metagenomic reads within, but also across, datasets can alleviate this problem. For 10 simulated datasets, each with species complexity modeled on a real counterpart, chimerism remained within the same species for most contigs (97%). For 42 real pyrosequencing ('454') datasets, assembly increased the proportion of annotated reads, and even more so when datasets were pooled, by on average 1.6% (max 6.6%) for species, 9.0% (max 28.7%) for Pfam protein domains and 9.4% (max 22.9%) for PANTHER gene families. Our results outline exciting prospects for data sharing in the metagenomics community. While chimeric sequences should be avoided in other areas of metagenomics (e.g. biodiversity analyses), conservative pooled assembly is advantageous for annotation specificity and sensitivity. Intriguingly, our experiment also found potential prospects for (low-cost) discovery of new species in 'old' data.

Contact: dgerloff@ffame.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 17, 2014; revised on July 8, 2014; accepted on August 7, 2014

1 INTRODUCTION

High-throughput DNA sequencing of microbial communities (metagenomics) has greatly expanded our knowledge (reviewed broadly in Schloss and Handelsman, 2005). For example, new marine biology, a better understanding of biogeochemical cycles and more are keenly anticipated from this (currently still main) option for characterizing unculturable species at the molecular level (Béjà *et al.*, 2000; Biller *et al.*, 2014; Dinsdale *et al.*, 2008; Iverson *et al.*, 2012; Rusch *et al.*, 2007; Venter *et al.*, 2004). In marine metagenomics (and for samples derived from other habitats), full realization of the potential of next-generation sequencing (NGS) relies on high-quality species and functional annotation of the now hundreds of datasets available to the scientific community (Gilbert and Dupont, 2011). However, metagenome annotation is confounded by (i) the shortness of NGS reads (Wommack *et al.*, 2008; Temperton and Giovannoni, 2012 review this challenge and others), which merely characterize small fragments of genes, and (ii) by the small fraction of known microbes represented in sequence databases, often described as 'the culturable 1%' (Amann *et al.*, 1990; Wooley *et al.* 2010). In traditional genomics, assembling reads into contigs is common practice. By contrast, metagenomes are primarily released to the community as reads, predominantly from Sanger or pyrosequencing ('454') to date, and even some 'gold standard' annotation pipelines are optimized for unassembled data. Concern over chimeric contigs may explain the reluctance to assemble. For example, only 3 of the 42 datasets in this work were assembled according to the publications in which they first appeared. This may seem justified to an extent by pyrosequencing read lengths, which typically range from around 100–600 nt in publically available sets, depending on how recently the data were obtained. While some annotation can be derived from individual reads in this range (Thomas *et al.*, 2012; Wommack *et al.*, 2008), it is plausible that assembly should add value to the data by improving annotation. However, the potential benefits for annotation have not been quantified rigorously.

Previous smaller-scale work with simulated assemblies already suggests that chimeric contigs may not be as common as feared. For example, Mavromatis *et al.* (2007) reported that 85–95% of contigs from assemblies of the most species rich set of their simulated Sanger-sequencing data did not mix genomes. Neither did the majority of contigs assembled from simulated pyrosequencing reads in other work by the Moya group [94% in Pignatelli *et al.* (2011); 68–97% in Vázquez-Castellanos *et al.* (2014)]. Below, we give this question a deeper look, but we also take a step further to ask: Is it advisable for deriving species and function annotation, to cautiously consider not only data from one sample, but also from others? By allowing limited chimerism within and across samples we may strike a balance between information gain and precision (with respect to an individual cell's genome) that is beneficial if used strictly for annotation. For example, a contig identified as *Trichodesmium erythraeum* is informative even if the contig mixes strains of this bacterium, whereas refusal to call the species because a contig fails to match precisely a known strain tells us nothing.

*To whom correspondence should be addressed.

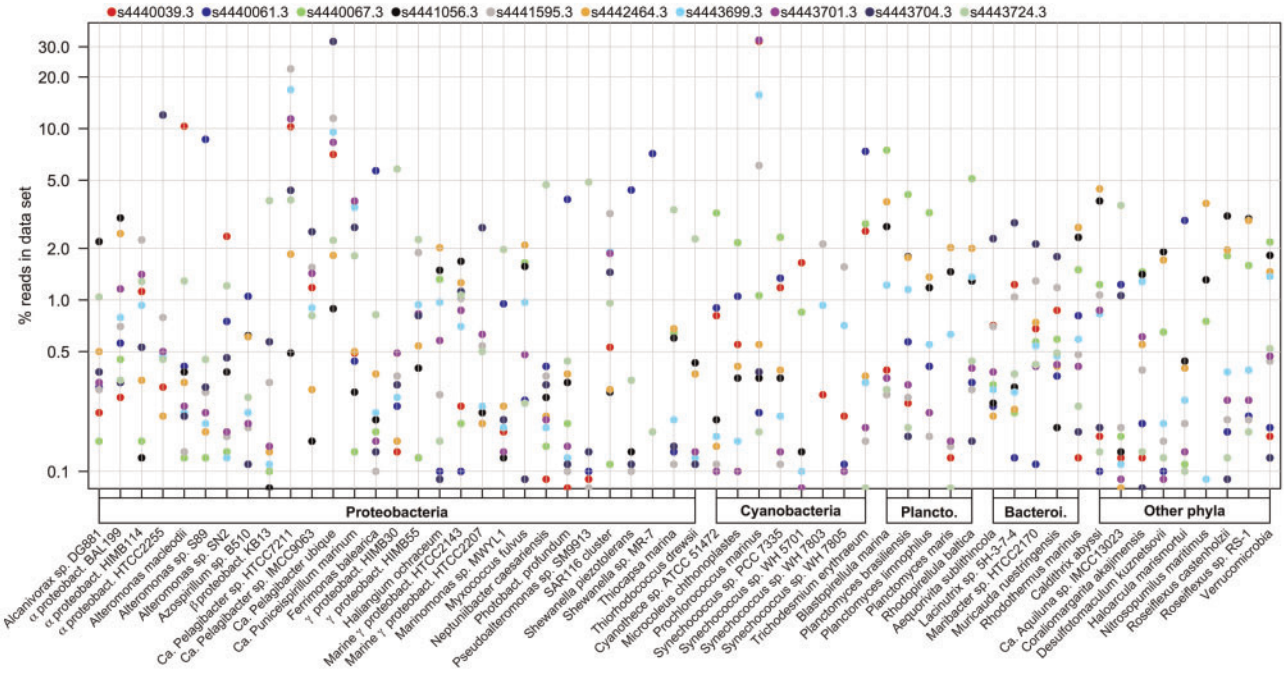

Fig. 1. Species composition of simulated datasets. The 56 species shown contributed ~99.9% of the total reads across all 10 simulated datasets. Sixty-three species (51 of them shown above) contributed ~99.9% in the real data counterparts (data not shown). On average, each simulated set and its real counterpart shared 12 of their top 20 species. The commonly used cutoff (UBLASTN top hit >50 bit score) was used for this illustration, which may yield false-positive matches (see text). Abbreviations: Plancto, Planctomycetes; Bacteroi, Bacteroidetes; bact, bacterium. See also Supplementary Data

2 METHODS

Marine RefSeq database: To assign reads and contigs (simulated and real) to marine taxa, we created a searchable, non-redundant database of all known marine bacterial and archaeal protein sequences in NCBI RefSeq (August 2012). The database included 2 585 466 sequences from 754 taxa. See Supplementary Methods.

Simulated and real data: See Supplementary Data and Fig. 1.

Assembly: Intra-set assemblies were done with Roche's GS *de novo* Assembler ('Newbler') v2.7 and required read overlaps of ≥40 nt at ≥90% identity. Pooled assemblies used the same parameters. However, because of the large number of reads (8 730 323) in the 42 real sets, we pooled via an iterative approach in which contigs aggregated reads over multiple rounds of assembly. Simulated sets were also assembled with Genovo (Laserson *et al.*, 2011) to investigate algorithm impact on chimerism. See Supplementary Methods.

Species annotation: To obtain conservative species calls for this discussion (Section 3.3), we applied the following protocol, to real and simulated data. Marine RefSeq was searched with each read or contig using translated nucleotide UBLAST (Edgar, 2010) with *E*-value cutoff 1E-1. Only hits exceeding bit score 110 and 90% id were used for consensus-based species calls. For reads, perfect species consensus of the hits was required. For contigs, calls were made only if the species entropy of the hits was ≤0.242—a 100-read contig with three minority reads, each from a different species, would have this entropy—and only if the hits covered >33% of the contig. (Coincidentally, Charuvaka and Rangwala (2011) measured chimerism the same way.) For simulated sets, correctness of consensus-based species annotation was evaluated by comparison with the true species of the sequence. For reads, the source genome was always known. For contigs, the true species was the majority species of the constituent reads. See Supplementary Methods.

Estimating the odds of correct species calls: To compare the odds of correct species calls for unassembled versus intraset-assembled versus pooled-assembled reads, we counted correct and incorrect calls (but not non-calls) for reads and contigs from the 10 simulated datasets. We also modeled the impact of species unknown to Marine RefSeq (to estimate the risk of incorrect species calls because of homologous hits in Marine RefSeq) by creating 10 decoy sets in which non-marine NCBI RefSeq genomes served as proxies for unknown marine species. Application of Bayes' Theorem produced the Figure 3 odds curves, with probabilities of a correct or incorrect species call estimated from the observed counts. See Supplementary Data and Methods.

Protein domain annotation: Open reading frames (ORFs) were called on reads and contigs with FragGeneScan v1.16 (model 454_10 was used allowing a 1% error rate; Rho *et al.*, 2010) and used to search Pfam A release 26.0 (Punta *et al.*, 2012) with HMMER v3.0 hmmscan [default parameters with trusted cutoff bit scores defined for each hidden Markov model (HMM); Eddy *et al.*, 2009]. The per-domain 'independent' *E*-value served as proxy for annotation confidence. Hits to Pfam domains of unknown function were ignored in our evaluation.

Gene family annotation: ORFs were also searched against PANTHER 8.1 (Mi *et al.*, 2013) gene family and subfamily HMMs using the PANTHER script `pantherScore.pl`. Reads were annotated only if they or the contig they belonged to aligned to an HMM with *E*-value <1E-23, recommended by PANTHER.

Pooled contig follow-up analyses: See Supplementary Methods.

3 RESULTS AND DISCUSSION

3.1 Simulated sets approximated species diversity of their real counterparts

To evaluate chimerism and annotation, we created simulated metagenomic datasets from reference sequences (complete genome sequences from known species). Simulated metagenomes in which the provenance of individual reads is known have contributed greatly to method development and discussion of

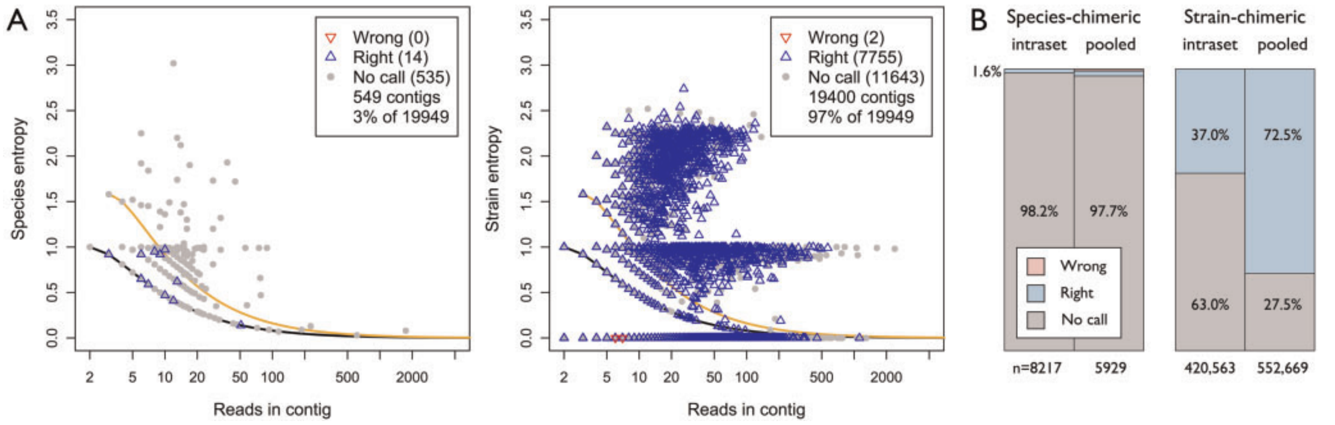

Fig. 2. Contig chimerism and its impact on species call correctness in simulated data. (**A**) Size and chimerism are shown for contigs from all 10 simulated datasets, separately assembled. There were 549 species-chimeric contigs (A, left) and 19 400 species-homogenous contigs (A, right). For reference, the black curves show the Shannon entropy if exactly one read differs in taxon from all others in the contig. The orange curves show entropy if a contig has two minority reads each from a different taxon. Of the 549 (3% of 19 949) species-chimeric contigs, 513 have entropy ≤1 bit. There are 237 contigs within 0.01 bits of the black curve and 157 on the curve. Of the 19 400 (97%) contigs that are chimeric below the level of species, 14 019 (70%) have entropy 0 bits (not chimeric) and another 3160 have entropy ≤1 bit. Corresponding plots after pooled assembly are in the Supplementary Results. (**B**) Despite chimerism species calls for reads, based on the calls for contigs into which they assembled, were usually correct. Pooled assembly of the 10 sets increased the number of assembled reads (*n*) and nearly doubled the proportion of correct calls for reads in strain-chimeric contigs

experimental data in this field (Charuvaka and Rangwala, 2011; Mavromatis *et al.,* 2007; Mende *et al.,* 2012; Pignatelli *et al.,* 2011; Wang *et al.,* 2014). To ensure that any impact of chimerism in simulation is likely predictive for real data, we created 10 simulated sets ('s' prepended to names) using the program MetaSim (Richter *et al.,* 2008), based on the taxonomic profiles of 10 real datasets in which 559 species (727 taxa) were identified. Thus, we aimed to approximate specific real sets with respect to species composition and total base pairs. In effect the simulated sets included data from 317 species in total (~394 taxa; Fig. 1). This deficit in complexity, compared with the real sets, was caused by MetaSim reaching the target dataset size before sampling all genomes requested in each input profile. Nonetheless, to our knowledge, our simulated sets are superior in complexity and 'realism' to previously published studies, and approximated their real counterparts well, in aggregate and individually. Both in the real and the simulated sets, a small number of identifiable species contributed ~99.9% of the reads [63 species (real), 56 (simulated), with 51 species in common]. Individually, each real set and its simulated version overlapped by 12 species on average in their respective 20 most abundant species (data not shown).

Several patterns are evident in Figure 1. First, many species are found in several sets, as is evident from the nearly 10 datasets shown in each species column. Indeed, 73% (230/317) of species appeared in more than eight of the simulated sets. This was also true of the real sets: 76% (425/559) of species appeared in more than eight sets. Nine of the sets derived from subtropical or tropical waters, four of them from the mid-Pacific, but the high proportion of shared species seems noteworthy nonetheless. This may also reflect that marine environmental samples harbor many rare species in the tail of their taxa distributions (Pedrós-Alió *et al.,* 2006) and that homology-based species annotation at standard stringency risks to (mis)associate their reads with a related species. Based on this observation, we use only a highly conservative species calling protocol in our analyses below (see Section 2).

Second, there is similarity across some sets for species within the same genus. For example, sets s4441595.3 (gray), s4443699.3 (cyan), s4443701.3 (purple), s4440039.3 (red), s4443724.3 (light green) and s4442464.3 (orange) show the same order by read abundance of three *Candidatus Pelagibacter* species: HTCC7211, *ubique*, IMCC9063. These proteobacterial species are members of the SAR11 cluster, ubiquitous and abundant in marine microbial communities (Morris *et al.*, 2002), but only 13 complete or near-complete *Candidatus Pelagibacter* genomes have been released (seven of them of *ubique* strains). One daring explanation for the pattern could be that just one uncharacterized *Candidatus Pelagibacter* species is actually present in the real samples, which has three close relatives among the species represented in our database (which is deemed 'closest' may vary depending on which parts of the uncharacterized genome are represented by the reads). This might be a more plausible explanation than three species appearing in the same rank order in six samples, which is also possible, of course. However, two of the six datasets (4443724.3 and 4442464.3) are from distant sample sites from the other four (mid-Pacific). We cannot resolve the cause conclusively but note that we found reminiscent rank order patterns in the corresponding real datasets as well, while for example Cyanobacteria seemed devoid of this phenomenon. Our conservatism in species calling in the analyses described below serves to ensure the validity of their results in either case, although it necessarily leads to higher proportions of 'non-calls' than metagenomicists are used to seeing in the literature [50–80% (Thomas *et al.*, 2012)].

3.2 Assembly of simulated sets yielded mostly strain- and few species-chimeric contigs

3.2.1 Chimerism and impact on species calls Intuitively, we categorize chimerism at three levels: whether a contig combined reads from different species, or different strains (from the same species), or different individuals (cells). While it is extremely relevant for biodiversity questions and even systems biology, the third level is of less interest here (neither positively nor negatively) than the first two levels. Admittedly species and strain classification criteria can neither be expected to be straightforward, especially for microorganisms, nor consistent. However, this distinction respects the taxonomical classification of species, and pursues our intuitive premise that strain chimerism will unlikely impact on annotation negatively (when annotation protocols are largely homology-based). When Newbler was used to assemble the 10 simulated sets individually ('intra-set'), with a minimum required read overlap of 40 nt at 90% identity, i.e. standard parameters, this yielded mostly small contigs (Fig. 2A; average N50 contig size of 1734 ± 897 nt), but of high quality, given the hundreds of species present in each sample. Of the 19 949 intra-set contigs 97% (19 400) were species-homogenous and 70% (14 019) were strain-homogenous (Fig. 2A). Only 3% (549) were chimeric at the level of species or higher. Moreover, the degree of chimerism, measured as Shannon entropy of the species or strains of the reads within a contig in consideration of its length (plot in Fig. 2A), is low. For example, 157 of the 549 species-chimeric contigs had only one read that differed in species from all others in the contig. Accordingly, chimerism did not impact the correctness of species calls made with our conservative protocol (see Section 2). If we defined the true species of a contig to be the majority species of its reads (there was almost always a strong majority, which is evident by the low entropies) few contigs elicited incorrect species calls: 0 species-chimeric contigs and only 2 strain-chimeric contigs (Fig. 2A), all short (<10 reads).

We noted that 43 reads received species calls before assembly (i.e. 'raw' reads) that were challenged by the calls made for the contigs into which they were later assembled. The contig-based call was correct for 27 of these reads. These are marginal occurrences by reference to the data volume analyzed (~1.1 M total reads in the 10 sets) of which >428K we found in contigs (Fig. 2B).

When probing dependency of our findings on the type of assembler and data used (Supplementary Methods and Results; Figs S5–S7), we noted higher species-chimerism rates (10%) for intra-set contigs with Genovo as a non-overlap-layout-consensus (non-OLC) algorithm example. Species-misannotation with non-OLC (8.7% as opposed to 0.0% with Newbler) was containable by eliminating short contigs, but we recommend OLC. Generally, assembly seems worthwhile for annotation if such caution is applied, although the extent of improvement will vary.

Next, we assembled all 10 sets in a 'pooled assembly' (see Section 2) and observed a modest increase in the number of contigs (19 949 intra-set versus 21 408 pooled; Fig. S4). This is not surprising given that intra-set contigs that recruit additional reads when pooled leave the count unchanged and intra-set contigs that merge decrease the count, i.e. only newly assembled reads from different datasets would be expected to increase the contig count in our experiment. More relevant, the read count assembled in contigs increased by 12% through pooling (429K intra-set versus 559K pooled, Fig. 2B) and the proportion of in-contig reads with correct species calls with our conservative species caller nearly doubled (37% intra-set versus 73% pooled). While we acknowledge that metagenomic contigs will mask some diversity (Desai *et al.,* 2012), these are striking improvements due to pooling.

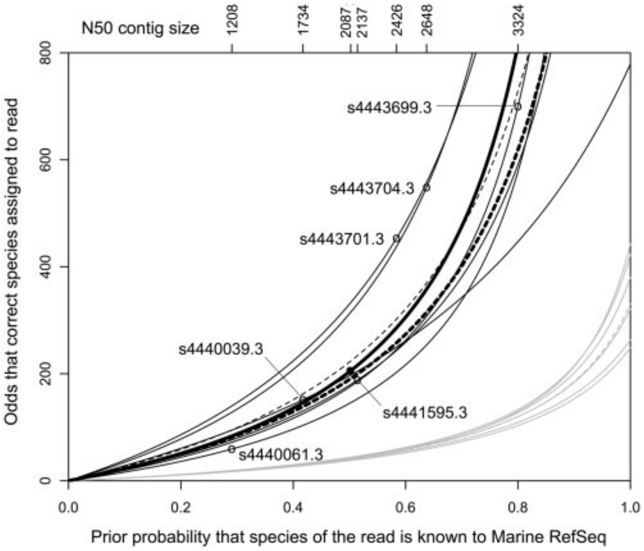

Fig. 3. Odds of correct species calls for reads that are unassembled, intraset-assembled and pooled-assembled. The odds ratio curves account for whether a read is from a species represented in Marine RefSeq or not. Each gray curve is based on measured rates of correct and incorrect species calls for unassembled reads in a simulated set and 10 decoy sets (Supplementary Methods). Similarly black curves represent calls for assembled reads. Only datasets for which >1% of reads assembled are shown. The dashed curves show the mean odds ratios with no assembly (gray), intra-set assembly (black), or 10 pooled assemblies with each set withheld (bold). The solid bold curve shows the odds when all 10 sets are pooled. The discs show the N50 contig size for each assembly and suggest that the odds of a correct species call for a read increase with N50

3.2.2 Species calls with consideration of novel marine species Figure 3 answers the question: Does assembly improve the odds of a correct species call for a read, given that only a fraction of species are represented in Marine RefSeq? The odds ratio curves weigh correct against incorrect species calls, by counting the reads in three categories: unassembled, intraset-assembled and pooled-assembled. (Reads are not reflected if no species calls were made with our conservative protocol (see also Fig. 2B)). The six marine sets for which >1% of the reads assembled show a clear improvement from unassembled (gray) to intra-set (black), in step with the N50 contig size, i.e. the higher the N50, the better the odds. The odds curve for pooled assembly of all 10 sets (bold) is consistent with its N50 contig size. This is consistent with a population of pooled contigs that somewhat resembles the intra-set populations but has contigs with more deeply aligned reads.

3.3 Assembly of 42 real datasets improved species annotation

Encouraged by the low species-level chimerism we observed in simulation, we selected 42 real marine microbial metagenomic sets from the MG-RAST public repository (Meyer *et al.,* 2008) and assessed the impact of assembly on species identification. All data were obtained with pyrosequencing ('454') technology. Figure 4 (left panel) shows the proportion of reads that elicited calls before assembly (gray) and after (colored). For each dataset,

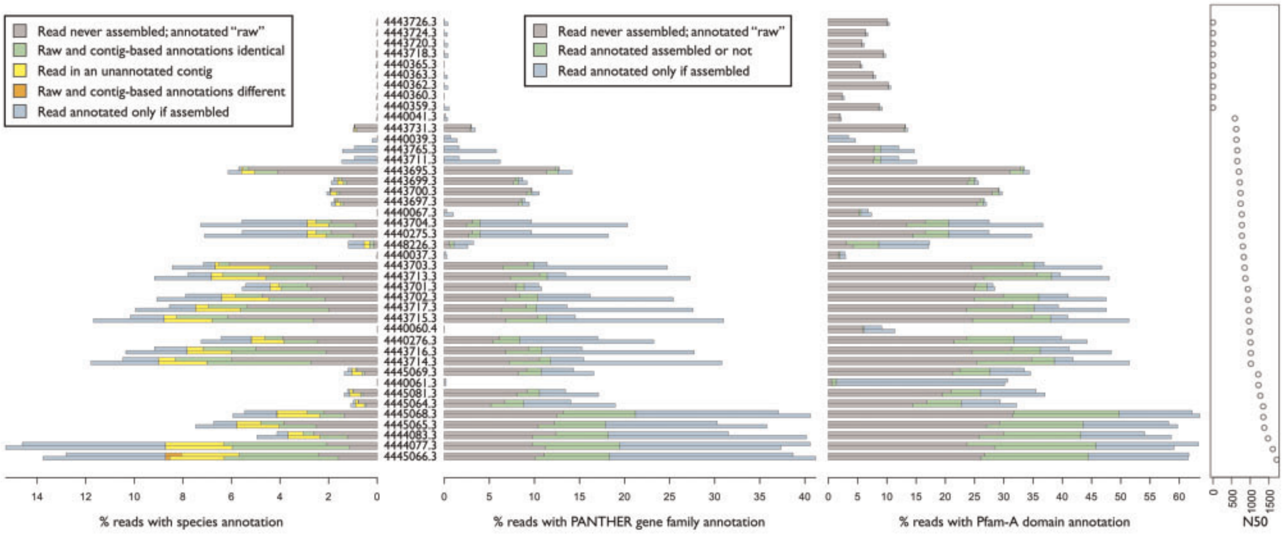

Fig. 4. Species, PANTHER and Pfam-A annotation changes because of intra-set and pooled assembly of real datasets. **Left:** Species call change (quantity) relative to the total read count in each of the 42 real datasets. Intra-set bars are above the corresponding pooled assembly bars in each pair. Datasets are sorted vertically by increasing intra-set N50 contig size (far right): 589 nt for 4440041.3, and 1711 nt for 4445066.3. Gray: reads that did not assemble i.e. elicited annotation only as individual 'raw' reads. Colored: reads that assembled into contigs and elicited annotation (see legend). **Middle:** A comparable graphic of gene family annotation increases for intra-set and pooled contigs, based on alignments to PANTHER HMMs, mirrors the species results. Annotation increases with N50 with some of the same exceptions (e.g. 4440061.3). **Right:** Pfam-A domain annotation also increases with assembly. This was especially pronounced for some sets with poor species and gene family annotation despite N50 > 0 nt (e.g. 4440061.3)

the proportions are illustrated as two bars, for intra-set (upper bar) and pooled (lower) assembly, respectively. Sets are arranged top to bottom by increasing degrees of assembly, measured as intra-set N50 contig size (right panel), and tended to experience greater increases in reads for which species calls were made (colored) the more they assembled. The correlation between intra-set call rate and N50 was strong (Kendall's $\tau = 0.53$), in spite of being impacted by some sets that assembled well but yielded few (or no) species calls with our protocol (see below).

The percentage of new species calls (blue), i.e. calls possible only after assembly, increased with intra-set N50 ($\tau = 0.48$; max 5.9% for 4444077.3). Pooling further increased the new call rate (max 6.6% for 4444077.3) by adding reads to intra-set contigs or creating new ones. In this context, the ratios of new calls to other kinds (green, yellow) depend on the conservativeness of the species caller, of course. A less conservative protocol would more readily assign species to raw reads and thus preempt potential new calls on assembly, though at a cost in misidentifications of homologs for known species as we discussed above (Section 3.1).

Often, individual reads elicited species calls before assembly (gray) that could be compared with calls based on contig membership. Usually, a contig-based call provided corroboration for the same species as was called beforehand (green), and more so for sets with a higher intra-set N50 ($\tau = 0.57$). Pooling increased the proportion of corroborated calls in each set. For example, the intra-set assembly of 4443715.3 produced contigs that supported calls for 2.1% of reads in this set (green); pooled assembly increased support to 4.1% of these reads. Pooling also tended to increase the proportion of reads for which raw species calls could not be corroborated (yellow), i.e. no call was made for the relevant contig. On inspection, such contigs typically recruited too few translated UBLAST hits from our Marine RefSeq protein database to cover 33% of their length (required for a species call, see Section 2). Rarely, a discrepancy arose between the contig-based and the pre-assembly calls for a read (orange). With these real data, we cannot firmly establish which (if either) of the calls is accurate. However, from the trends observed with simulated data (Fig. 3) it seems reasonable to predict that assembly will have helped rectify an incorrect species assignment more often than not.

The species annotation impact of pooling over intra-set assembly for each dataset is also captured by the ratio of pooled to intraset-assembled reads in a set (i.e. the ratio of colored regions in Fig. 4). On average, pooling increased species calls by 15% (minimum 0% for 4448226.3, maximum 57% for 4443711.3), in the 27 sets with >0.5% total species calls. Novel species might be responsible for the lack of improvement in the other 15 sets where <0.5% of reads elicited a species call. These 15 sets also yielded comparatively low annotation rates in MG-RAST (0–26.1%). For example, sets 4440060.4, 4440061.3 and 4440067.3 received no calls by our species caller in spite of assembling well (with N50 values of 991 nt, 1221 nt and 773 nt, respectively). These sets derive from a phage study of marine microbialites in which viral metagenomes from the same locations also had low annotation rates (>97% of reads unidentified; Desnues et al., 2008).

3.4 Assembly of 42 real datasets improved domain and gene family annotation

3.4.1 Pfam-A protein domain annotation We investigated whether assembly would improve annotation of Pfam-A domains (Fig. 4, right panel). As with species calls, domain assignments increased after assembly, in positive correlation with N50 ($\tau = 0.75$). Notably though, the number of post-assembly domain assignments far exceeded those of species. For the 33 intra-set assemblies with acceptable N50 values, the proportion of corroborated or new domain assignments was 0.1–40% (mean = 10.6%), compared with the 0.0–12.5% (mean = 2.1%) for species calls. Considering only new annotations, domain assignments ranged from 0.0–29% (mean = 5.9%) and species calls from 0.0–5.9% (mean = 1.0%). The conservation of domains across taxa is illustrated nicely by this increase and domain/family annotation is where we anticipated (and saw) the potential benefits of assembling chimeric contigs from close relatives best reflected. Set 4440061.3 is an extreme example: Neither raw reads nor contigs matched to the data/species in Marine RefSeq, yet 30% of reads assembled into contigs for which domain assignments resulted (using a standard HMMER3 protocol, see Section 2). Interestingly, for this set ~40% of the domain assignments were to Bacteriophage Replication Gene A (PF05840) and ~1% to Capsid Protein F (PF02305), possibly reflecting an abundance of prophages, which were not in Marine RefSeq.

Some reads that assembled could not be mapped to pooled contigs for technical reasons (This caused the fractions of reads in pooled-assembled contigs shown in Fig. 4 to fall below the corresponding intra-set fractions (e.g. 4445066.3 Pfam domains), even though we define pooling as a superset of intra-set results. We note that the consequence with respect to our evaluation is merely that the annotation gains depicted in Fig. 4 are in fact underestimates.

3.4.2 PANTHER gene family annotation To gauge whether assembly would improve identification of multidomain protein architectures (a step toward functional annotation), we used raw and assembled reads in searches against PANTHER 8.1 gene family HMMs. The PANTHER results strongly mirrored the species call results, as is evident in Fig. 4. Pearson correlations corroborated this symmetry for total annotations ($r_{intraset} = 0.68$, $r_{pooled} = 0.71$) and also for only the new calls enabled by assembly (blue; $r_{intraset} = 0.85$, $r_{pooled} = 0.80$). The surprising symmetry between the plots was not reasonably explained by shared reference species between PANTHER HMMs and our database. Only five of the 147 PANTHER species (UniProt Reference Proteomes) were in our database. Therefore, the much higher annotation rates for gene families versus species (Fig. 4) suggest that the PANTHER models identified contigs representing marine species not in our database. For example, set 4445081.3 has few species calls (even in MG-RAST, only 14.9% of reads have protein or rRNA annotation) but many gene calls, likely because of conserved, homologous genes in uncharacterized species.

3.5 Pooling—a prospect for species discovery?

Can pooled assembly of metagenomic data help discovery of novel species, especially those that are ubiquitous? To answer this, we first checked whether pooled contigs that lacked species calls by reference to our database MarineRefSeq, which is a protein database and also excluded, e.g. marine viruses (see Section 2), would perhaps elicit hits to NCBI RefSeq genome data. We searched with all 3806 unannotated contigs from pooled assembly with lengths >5kb using a CAMERA BLASTN workflow (E-value < 0.1, Sun *et al.,* 2011). The vast majority of the contigs (3688) failed to align at >80% sequence identity over >80% of their lengths to any reference genomes (3779 failed if 90% cutoffs were set). Thus, most of the contigs likely represented novel microbial species.

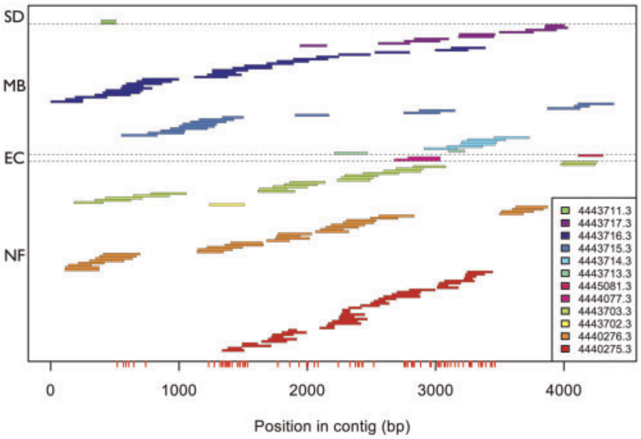

Fig. 5. Pooled contig03272 was composed of reads from 12 samples from four geographic regions: San Diego (SD), Monterey Bay (MB), western English Channel (EC), near a Norwegian fjord (NF). Although reads lacked significant hits to reference databases, the contig was mappable and may represent a novel, geographically diverse type of γ proteobacterium. Ticks on the bottom axis show 76 positions where reads from each geographic region were nearly unanimous but differed across regions (as described in the text). These proportions are not likely to occur by chance and may reflect geographically specific differences

To illustrate the potential of pooled assembly in this context, we investigated a single pooled contig that lacked a species call, contig03272, in detail. This 4388 bp contig was picked at random from a list of contigs that had not elicited a species call, contained reads from multiple samples, and were >4kb in length. These criteria served to enrich for contigs that might represent novel ubiquitous species and that would overlap multiple genes. Our example contig03272 was built from 148 reads from 12 surface water samples from widely separated oceanic regions (Fig. 5). Read depth was relatively uniform (2 reads minimum; mean = 8.9 reads) and all reads mapped to the contig (BLASTN) over nearly their full lengths (96.5% minimum; mean = 99.7%) and with high sequence identity (92.7% minimum; mean = 97.5%). Two computationally predicted protein sequences aligned to over 78% of contig03272 (BLASTX search of NCBI/nr, E-value <10), both of them proteins from SAR86 clade members (B and A): a Dehydroquinate Synthase (71% and 65% id) and a Membrane Carboxypeptidase/Penicillin-binding protein (Mrca; 59% and 62% id). These predicted proteins are encoded by neighboring genes in both SAR86 genomes. At the time of analysis, these observations in combination suggested to us that contig03272 might represent a geographically widespread, novel marine microbe.

Indeed, during the preparation of this manuscript, a draft genome assembly for the γ proteobacterium SCGC AAA076-P13 (NCBI BioProject accession PRJNA195664) was submitted to the NCBI as results of a Joint Genome Institute (JGI) project applying single-cell sequencing technology to 30 ubiquitous, uncultured marine microbes. The timing is fortuitous, as the new data corroborated both the ubiquity of the microbe represented by contig03272 and that our contig had been assembled correctly. BLASTN of contig03272 against the draft SCGC AAA076-P13 genome assembly yielded coverage of 97% of the contig, in three match alignments with 97%, 97% and 95% sequence identity, respectively.

Interestingly, the draft genome assembly was based on a sample collected in the Gulf of Maine, far from the four oceanic regions that contributed to contig03272 (Fig. 5). As the latter were widely separated, population-level differences between the isolates of this new species (SCGC AAA076-P13) should be apparent in our data. A cursory inspection of non-unanimous positions in the contig corroborated this. Of the 106 non-unanimous positions covered by $\geq$8 reads, 76 were near-unanimous (we allowed one read not to match) within oceanic regions. Random reshuffling of the geographic labels in the same contig (same read positions and label proportions), as a null model, failed to attain this degree of segregation coincidentally ($P \approx 0$; 0 instances in 10 000 trials).

4 CONCLUSIONS

While the notion that annotation (quality and quantity) should improve with assembly is intuitive, this is to our knowledge the first deep effort at quantifying the benefits of strain-chimeric contigs from a practical perspective for metagenomics. In simulation and for real datasets, we showed that intra-set and pooled assembly of marine metagenomic data (i) produce chimeric contigs that rarely violate marine species ‘boundaries’; (ii) lead to quality and quantity improvements in species, protein domain and gene family annotation; (iii) may help identify geographically diverse but novel species and population differences within those species. We focused exclusively on data from pyrosequencing (‘454’) efforts as the currently predominant type in marine metagenomic repositories. However, most key concepts and the strategy of our analyses are transferrable to other sequencing technologies. With their read lengths now surpassing those from the first ‘454’ generation, Illumina-sequenced sets may be the next data to investigate.

ACKNOWLEDGEMENTS

The authors thank Jonathan Zehr (UCSC Ocean Sciences) and Jim Tripp (JGI-DOE) for helpful scientific discussions, and the UCSC School of Engineering for financial assistance for J.D.M.

Conflict of interest: none declared.

REFERENCES

Amann,R.I. *et al.* (1990) Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Appl. Environ. Microb.*, **56**, 1919–1925.

Béjà,O. *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**, 1902–1906.

Biller,S.J. *et al.* (2014) Bacterial vesicles in marine ecosystems. *Science*, **343**, 183–186.

Charuvaka,A. and Rangwala,H. (2011) Evaluation of short read metagenomic assembly. *BMC Genomics*, **12**, S8.

Desai,N. *et al.* (2012) From genomics to metagenomics. *Curr. Opin. Biotech.*, **23**, 72–76.

Desnues,C. *et al.* (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, **452**, 340–343.

Dinsdale,E.A. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.

Eddy,S.R. *et al.* (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Gilbert,J.A. and Dupont,C.L. (2011) Microbial metagenomics: beyond the genome. *Ann. Rev. Mar. Sci.*, **3**, 347–371.

Iverson,V. *et al.* (2012) Untangling genomes from metagenomes: revealing an uncultured class of marine *Euryarchaeota*. *Science*, **335**, 587–590.

Laserson,J. *et al.* (2011) Genovo: *de novo* assembly for metagenomes. *J. Comput. Biol.*, **18**, 429–443.

Mavromatis,K. *et al.* (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods*, **4**, 495–500.

Mende,D.R. *et al.* (2012) Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS One*, **7**, e31386.

Meyer,F. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Mi,H. *et al.* (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.

Morris,R.M. *et al.* (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, **420**, 806–810.

Pedrós-Alió,C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.*, **14**, 257–263.

Pignatelli,M. and Moya,A. (2011) Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS One*, **6**, e19984.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Rho,M. *et al.* (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191–e191.

Richter,D. *et al.* (2008) MetaSim—a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.

Rusch,D.B. *et al.* (2007) The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.

Schloss,P.D. and Handelsman,J. (2005) Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol.*, **6**, 229.

Sun,S. *et al.* (2011) Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res.*, **39**, D546.

Temperton,B. and Giovannoni,S.J. (2012) Metagenomics: microbial diversity through a scratched lens. *Curr. Opin. Microbiol.*, **15**, 605–612.

Thomas,T. *et al.* (2012) Metagenomics – a guide from sampling to data analysis. *Microb. Inform. Exp.*, **2**, 3.

Vázquez-Castellanos,J.F. *et al.* (2014) Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics*, **15**, 37.

Venter,J.C. *et al.* (2004) Environmental shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.

Wang,Y. *et al.* (2014) MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. *BMC Genomics*, **15** (**Suppl. 1**), S12.

Wommack,K.E. *et al.* (2008) Metagenomics: read length matters. *Appl. Environ. Microb.*, **74**, 1453–1463.

Wooley,J.C. *et al.* (2010) A primer on metagenomics. *PLoS Comput. Biol.*, **6**, e1000667.