Title:

Optimizing microbial community metagenomic hybrid assembly correction and polishing approaches informed by reference-independent analyses

Authors:

Garrett J. Smith[1,#], Theo van Alen[1], Maartje van Kessel[1], Sebastian Lücker[1]

Affiliations:

[1] Department of Microbiology, Radboud Institute for Biological and Environmental Sciences, Radboud University, Nijmegen, The Netherlands

[#] Corresponding author

Corresponding author and contact information:

gsmith@science.ru.nl

Abstract

Hybrid metagenomic assembly, leveraging both long- and short-read sequencing technologies, of microbial communities is becoming an increasingly accessible approach yet offers challenges that existing benchmarks and pipelines do not comprehensively handle well. One specific methodological knowledge gap is selecting the ideal number of iterations of long read correction and/or short read polishing, and in particular which the features of an assembly to examine to empirically determine them. To account for biological and technological variation, in this study the microbial communities of two laboratory-scale bioreactors were sequenced with both short and long read platforms and assembled with two different, commonly used software packages. Following assembly, long read error correction was iterated ten times, with subsequent ten rounds of short read polishing after common (0 and 2) and extensive (5 and 10) correction iterations. Several assembly characteristics were tracked for each assembly and iteration of correction and polishing: gene fragmentation, short read recruitment, automated binning yields, as well as observed beta-diversity. While long read correction was necessary, the greatest improvements occurred within the first few iterations of short read polishing, and extensive correction and polishing did not tangibly improve assembly quality. Furthermore, essentially all tracked characteristics followed the same patterns; simpler statistics can serve as decent proxies for more complex analyses to save computational resources assessing assembly quality. Hybrid sequencing approaches will likely remain relevant due to the low costs of short read sequencing, therefore it is imperative users are equipped to estimate assembly quality prior to downstream gene- and genome-centric analyses.

Background/Introduction

Though the increasing variety of high-throughput short- and long-read (meta)genomic sequencing technologies are only within their first decades of existence, both the sequencing technologies and software development have flourished and been implemented to study microbial ecosystems[1–3]. Hybrid sequencing integrating both short- and long-read platforms for single microorganisms or microbial communities is becoming increasingly popular[4]

because they compensate for the other's weaknesses – shorter reads achieve higher accuracy while longer reads offer better contiguity. Despite the technical costs decreasing and technologies improving, hybrid strategies currently are cost-effective for recovering high-quality Metagenome-Assembled Genomes (MAGs) from microbial communities[5]. However, many challenges remain that may deter users from seeking relatively complicated approaches like hybrid (meta)genomic sequencing of microorganisms and complex communities.

Hybrid (meta)genomic assembly typically requires iterative steps to resolve various issues derived from sequencing errors, but in the absence of "gold standard" reference genomes for less well-studied microbial consortia, communities, or ecosystems, the determinants of a high-quality assembly are unclear. Applying published methods may not be suitable for disparate ecosystems, and common community guidance of "two iterations of two tools" likely oversimplifies the needs of uncertain users. Therefore, reliable metrics to determine the ideal approach to hybrid metagenomic assembly, correction, and polishing is imperative. Many genome and assembly quality algorithms rely on comparison to high-quality references, *e.g.*, mis-assemblies or specific marker gene presences[6,7], which can lead to poor interpretation of divergent genomes and high resource consumption costs. Other common statistics, for example contig counts, total base pairs assembled, and L/N50, are reference-independent but provide only a glimpse of contiguity, which should not substantially change during iterative processes that fix relatively small-scale errors[8]. A broadly applicable and simple to compute assembly trait to help estimate the quality of hybrid assemblies during long read correction and/or short read polishing, for example predicted average gene length[9,10], would aid both novice and experienced users with complex sequencing projects.

Here, we performed multiple variations of correction and polishing strategies to examine the effects of these processes on microbial community metagenomic datasets. Two long-term laboratory-scale nitrifying bioreactors were sequenced using both Illumina MiSeq (short-read, SR) and Oxford Nanopore Technology (long-read, LR) platforms and assembled with multiple programs to allow for initial biological and computational variation. LR assemblies were then corrected with LRs and polished with SRs using free, common tools in the field. We primarily monitored the following four features

throughout the iterative correction and polishing processes to empirically determine the optimum draft community assembly quality: gene fragmentation, SR recruitment, medium-quality or better automated bin yields, and beta-diversity of single copy marker genes. All metrics followed the same trends that supported common conventions in literature and online guidance – the greatest improvements across nearly all possible quality metrics occurred within the first two iterations of LR correction followed by two iterations of SR polishing. However, LR correction generally was noisy, and neither LR correction nor SR polishing beyond 5 iterations showed substantial improvements in any assembly quality metric. Fortunately, the patterns for simpler, reference-independent features (*i.e.*, gene fragmentation and SR recruitment) were essentially the same as more complex, reference-dependent features (*i.e.*, automated bin quality, beta diversity), which could serve as less resource-exhaustive proxies and easier to implement in automated processes. Finally, we offer practical guidelines to aid users to empirically determine their own optimal high-quality data yields via hybrid metagenomic assembly.

Methods

Long and short read sequencing

Biomass was collected in November 2020 and March 2021 from long-term, autotrophic nitrifying enrichment cultures maintained in either oxygen- (OLR) or nitrogen-limited (NLR) conditions in a tandem laboratory-scale bioreactor system inoculated in 2015 with wastewater treatment plant activated sludge from the Bavaria Brewery in Lieshout, The Netherlands (51.518666 N, 5.613009 E). Comprehensive analysis of these bioreactor communities is an on-going project and will be published separately. Genomic DNA was extracted from the biomass using a conventional N-cetyl-N,N,N,-trimethyl ammonium bromide (CTAB) protocol in 2020, and the Powersoil DNA Isolation kit (Qiagen) in 2021 with minimal modifications to the manufacturer's directions to reduce DNA shearing such as inverting rather than vortexing or pipetting to mix.

In total 1 ng of DNA for both the OLR and NLR reactors was used to prepare a library using the Nextera XT kit (Illumina, San Diego, California U.S.A.) according to manufacturer's instructions. After quality and quantity check of the libraries, they were paired-end sequenced (2x 300bp) using the

Illumina MiSeq sequencing machine and the MiSeq Reagent Kit v3 (San Diego, California USA) according to manufacturer's protocols. Oxford Nanopore Technologies (ONT) sequencing was done with 840 and 1670 ng DNA for the OLR and NLR reactors, respectively, after library preparation using the Ligation Sequencing Kit 1D (SQK-LSK108) and the Native Barcoding Expansion Kit (EXP-NBD104), according to the manufacturer's protocols (Oxford Nanopore Technologies, Oxford, UK). The libraries were loaded on a Flow Cell (R9.4.1) and sequenced on a MinION device (Oxford Nanopore Technologies, Oxford, UK), according to the manufacturer's instructions. Guppy (version 4.0.11)[11] was used to basecall fast5 files using the dna_r9.4.1_450bps_hac.cfg model, both provided by Oxford Nanopore Technologies.

Raw sequencing reads for both bioreactors and technologies are available at NCBI via BioProject PRJNA1005948 as the following BioSamples: SAMN37004618, raw MiSeq reads for the OLR; SAMN37004620, raw Oxford Nanopore Technologies reads for the OLR; SAMN37004619, raw MiSeq reads for the NLR; SAMN37004621, raw Oxford Nanopore Technologies reads for the NLR.

Long-read, short-read, and hybrid microbial community metagenomic assembly

An overview of the experimental design is shown in Fig. 1. All computational programs were employed with default settings unless explicitly stated. Generic example code is available in File S1.

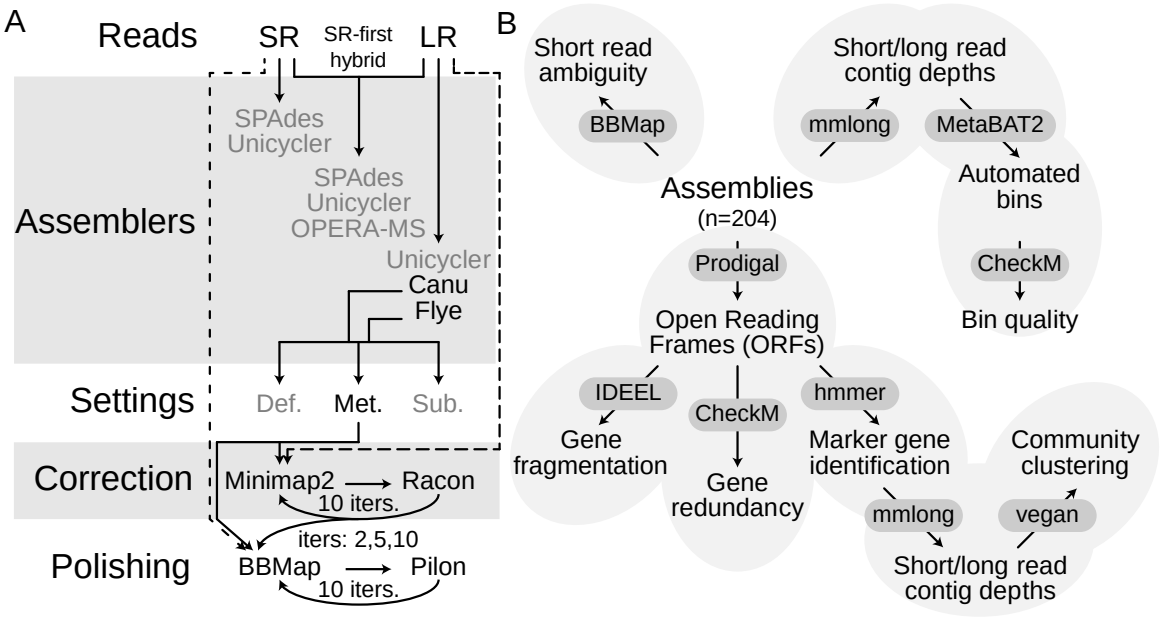

Figure 1. Schematic overview of assembly, correction, and polishing. (A) Assembly, correction, and polishing experimental design. Broken lines represent read recruitment steps for either correction or polishing. Acronyms: SR, short read; LR, long read; Def., default settings; Met., metagenome-optimized settings; Sub., sub-sampled LRs with metagenome-optimized settings. (B) Simplified diagram of analyses performed and the tools used to perform them for each assembly and after each iteration of correction or polishing.

Sequencing yielded 3.1-4.6 Gbp per library (Table S1) following read trimming and length and quality control using bbduk (BBtools version 37.76)[12] with a minimum phred score 18 and length of 200 bps for the MiSeq reads, and porechop (version 0.2.3_seqan2.1.1)[13] with minimum split length of 3000 bps for the ONT reads.

Three common SPAdes-dependent programs were used to generate SR-only, LR-only, and/or SR-first hybrid, *i.e.*, short read assembly followed by connecting contigs using long read alignments, assemblies: metaSPAdes (version 3.15.4)[14] known for high-quality metagenomic assemblies, Unicycler (version 0.4.9b)[15] with utilities for optimizing SPAdes to enable recovery of circular and high-quality single genomes, and OPERA-MS (version 0.9.0)[3] capable of automated refinement of high-quality individual genomes within a

multi-species metagenome (hybrid only). The k-mer list generated by automated selection during Unicycler hybrid assembly was used for metaSPAdes assemblies. While the focus here was optimizing a hybrid approach that leveraged the continuity of the long-read assemblies, these SPAdes-dependent assemblies served as a baseline for implementing SR-focused assembly approaches.

Two programs were used for assembly of LR data: Canu (version 1.8)[16] that performs read correction prior to assembly and is generally thought to result in more accurate assemblies, and Flye (version 2.9-b1768)[17] that performs correction of the assembly using the input reads and is generally thought to sacrifice some accuracy for better resource efficiency. Both assemblers were used in three different ways: (1) default settings ("def") developed for single-genomes, (2) metagenomic settings ("meta") to improve assembly of sequences with uneven depths[2,17], and (3) metagenomic settings for approximately even bp sub-samples ("sub") of the reads to artificially reduce the sequencing depth and possibly uncover biological variation. Long reads were sub-sampled into 12 read pools with approximately the same quantity of bps using the "subsample" utility of Trycycler (version 0.4.1)[18]. All assemblies were filtered to a 4 kbp minimum contig length prior to any downstream analyses.

Long read correction and short read polishing

Consensus correction of individual LR "meta" assemblies was performed using Racon (version 1.3.1)[19] up to 10 iterations, the greatest number identified among literature and published automated tools[15]. For each iteration of Racon correction, minimap2 (version 2.17-r941)[20] was used with default settings to recruit LRs to the original assembly or previous iteration's corrected assembly to generate the overlap information. Contigs that were not corrected were written with the optional flag "-u", and though it was not done here the optional flag "-f" may allow multi-mapping of the LRs. SR pile-up polishing of LR "meta" assemblies was performed using Pilon (version 1.23)[21] to 10 iterations after 0 (original assembly), 2, 5, and 10 rounds of LR correction. These stages were chosen to include common (0 and 2) LR correction endpoints found in literature and online guidance, and relatively extensive (5 and 10) LR correction

endpoints that are observed in existing pipelines like Unicycler. We did not test SR polishing after 1 iteration of LR correction because it was essentially never observed in published literature or online guides. For each iteration of SR polishing, bbmap (BBtools version 37.76) was used to recruit SRs to the original assembly, Racon-corrected assemblies, or previous iteration's polished assembly to generate alignment fragments, using a 97.5% identity filter and retaining all ambiguous alignments. Because we found no examples of LR correction after SR polishing and it is expected to re-introduce errors, this was not performed here.

Assembly quality estimation

Programs for estimating quality were used with default settings except when explicitly stated otherwise. Assemblies were compared using metaQuast (version 5.0.2)[22] without references to examine the distribution of contig counts and lengths. Recruitment of SRs to the assemblies to quantify aligned reads and ambiguity was performed using BBmap (BBtools version 37.76) with retaining only perfect mappings of paired reads and ambiguous mappings randomly assigned. Open reading frames (ORFs) were predicted using Prodigal (version 2.6.3)[23] using the "meta" procedure for quantification of gene counts and lengths, and as input for downstream analyses of assembled genes. The "meta" procedure of Prodigal was implemented prior to downstream analytical pipelines because it led to increased marker gene recovery estimates, though following similar trends (data not shown), compared to the outputs of the analytical pipelines that assume inputs are single genomes and therefore run prodigal in "single" mode by default. Phylogenetic marker gene recovery, fragmentation, and redundancy were estimated by Benchmarking Universal Single-Copy Orthologs (BUSCO, version 5.1.2)[24] using "genome" mode and only the "bacteria_odb10" lineage, as well as CheckM (version 1.1.3)[25] using the taxonomy workflow for the domain "Bacteria". Gene fragmentation of entire assemblies was estimated by comparing to a Diamond (version 0.9.31)[26] database of the uniref50 (downloaded 2023-03-01)[27] dataset using IDEEL (downloaded 2023-05-13)[10]. To more concretely associate potential complete genome contigs with the domain bacteria, circular and long (>1 Mbp) contigs were also analyzed individually using the Genome Taxonomy

DataBase toolkit (GTDB-tk, version 1.6.0 with reference database version r202)[28]. Based on the GTDB-tk output for circular contigs alone, there were members of at least 6 and 5 distinct Bacterial phyla present in the OLR and NLR, respectively (data not shown). Microbiomes with abundant Archaea or Eukaryotes may need to adjust these taxonomically constrained pipelines to better suit their ecosystem.

Automated binning and beta diversity

Programs for automated binning and estimation of genome quality, as well as identification and coverage calculation of marker genes, were used with default settings except where noted otherwise.

Both LRs and SRs from both bioreactor datasets were used for read depth calculation of contigs using the mmlong (version 0.1.2)[29] utility *readcoverage2*. Automated binning using composition (*i.e.*, tetranucleotide frequency) and coverage (*i.e.*, read depths) was then performed using Metabat2 (version 2.12.1)[30]. The quality of automated bins was estimated using CheckM (version 1.1.3)[25], with cutoffs of >50% completeness and <10% contamination scores as “medium quality" (MQ). These oversimplified thresholds are no longer *en vogue*, but represent computationally simple, rapid, and bulk estimates compared to contemporary, thorough requirements that include other information like rRNA and tRNA counts[31].

The RNA polymerase subunit B (*rpoB*) is a protein-coding gene typically found only once in a genome and is universally conserved among Bacteria, Archaea, and Eukarya, and therefore is tractable to serve as a marker gene for individual species to complement the often poorly assembling 16S rRNA gene for gene-centric phylogenetic analyses of metagenomes. *RpoB* genes were identified in assembled contigs using hmmer (version 3.1b2)[32] with the available model and thresholds for the Protein FAMily (pfam) identifier PF04563.15 (downloaded 2020-06-09). To observe beta-diversity, rpoB-containing contigs throughout the iterative correction and polishing processes were compiled for the each bioreactor and assembler separately, and then the LRs and SRs from both bioreactor datasets were used for coverage estimation for *rpoB*-containing contigs using the mmlong (version 0.1.2)[29] utility *readcoverage2*. The Bray-Curtis abundance-rank dissimilarity between each

read set was calculated for subsequent two-axis Non-Metric multi-Dimensional Scaling (NMDS) performed using vegan (version 2.5-7)[33] in R (version 4.1.2)[34], and the values were then manually mean-centered and scaled for better comparability. These NMDS analyses were unable to converge after 50 tries due to low stress, but here the species scores were used to view the clustering and trajectories of the *rpoB*-containing contigs' "species scores" for each assembly rather than the four reads' "site scores".

Further data analysis and visualization

Logs and outputs were mined for data using various bash commands. Data were ultimately imported into R (versions 4.1.2)[34] for analysis and visualization relying primarily on tidyverse (version 1.3.1)[35]. Most calculations were additionally normalized to contig length in Mbps to make bioreactors and assemblers more comparable. No specific code was developed for this project due to the focus on applying end-user tools.

Results

Short read, long read, and hybrid assembly baselines

SR sequencing yielded 7.56 million paired-end reads totaling 4.06 Gbp, and 8.73 million paired-end reads totaling 4.60 Gbp of data after low quality read removal and base trimming for the OLR and NLR, respectively. LR sequencing yielded 318 thousand reads totaling 3.13 Gbp, and 416 thousand reads totaling 4.01 Gbp of data after low quality read removal and base trimming data for the OLR and NLR, respectively (Table S1). Reads were either assembled alone (SR-alone or LR-alone) or used as inputs for automated SR-first hybrid assembly. There were substantial differences between assemblies of the bioreactors of approximately 63 total assembled Mbp on average (Fig. S1), which we infer was due to the greater sequencing depths for the NLR compared to the OLR rather than any biological differences.

In addition to biological and/or sequencing depth effects on the assemblies, the assembly approach implemented by the different programs also had a substantial effect on the assembly outcome. This was apparent across many aspects of these assemblies: LR assemblies were larger in terms of total bps (on average 1.35-fold), they were less fractured such that a greater

fraction of bps were on larger contigs (1.69-fold), and they yielded more circularized contigs (2.34-fold for count and 5.5-fold for total bps) (Text S1, Figs. S1, S2A, S3). However, probable complete microbial genomes, *i.e.*, a single contig greater than 0.5 Mbp and ideally circular, reconstructed with SR and SR-first hybrid assemblies tended to have higher completeness scores on average (1.73-fold for long contigs and 1.38-fold for circular, Text S1, Fig. S2B, S3). Thus, baseline assemblies recapitulate most benchmarking studies that LR assemblies lead to better contiguity, but SR assemblies are more accurate. Still, the contiguity of the LR assemblies offers advantages for microbial community reconstruction with sufficient LR correction and/or SR polishing to resolve mis-assemblies.

From here, we explored these metagenome-optimized LR assemblies by monitoring the assembly quality while performing LR correction with Racon and SR polishing with Pilon in order to determine the optimal approach for these communities. We sought metrics that should noticeably change, in contrast to for example the size and contiguity of the assembly, for assessing microbial community-scale quality, and that ideally would also be relatively independent of reference and database composition. Given known errors in LR assemblies, we focused on the following assembly-level information criteria to help determine possible optimal correction and polishing iterations: gene fragmentation because indels in LR lead to frameshifts that fracture genes, and SR recruitment because these reads are more accurate. Additionally, we followed automated bin recovery and single phylogenetic marker gene beta diversity to understand how these LR correction and SR polishing might influence typical genome- and gene-centric analyses performed on metagenomic datasets.

Gene fragmentation

Small scale errors (*i.e.,* insertions, deletions, and substitutions) in LRs are well-documented issues for using these platforms that cause gene fragmentation in the resulting assemblies[10]. Studies have already demonstrated that average gene length can be a practical parameter for tracking assembly quality, for example with viral populations[9] or MAG collections[10]. As a precursor to many downstream analyses, coding gene identification with programs like Prodigal

allow direct estimates of coding gene contents, while other programs like IDEEL and BUSCO can more directly estimate gene fragmentation by comparison to a reference database.

Consistent with overall assembly statistics (see above), the NLR contained both more genes (1.34-fold) and more bps in genes (1.39-fold) than the OLR, likely due to sequencing depth. The Flye assemblies contained more bps in genes (1.23-fold) but only slightly more total genes (1.10-fold) compared to Canu, indicating better initial assembly quality (Text S1). On average, gene fragmentation did not markedly improve, either in reducing the number of coding genes or increasing the quantity of bps in coding genes, due to LR correction but did substantially improve after just two iterations of SR polishing (0.81-fold for gene count, 1.12-fold for bps in genes; Fig. S4). The same patterns were noticeable not only for the entire assembly, but also for several relevant assembly fractions principally ready for downstream analyses – circular and long contigs that may represent essentially complete microbial sequences, and medium quality or better automated bins (Text S1, Figure S4, Fig. S5).

Comparison of these coding genes to databases showed similar patterns. After just two iterations of SR polishing, assessing the entire assemblies with BUSCO indicated that bacterial marker genes were no longer found only fragmented (Text S1, Fig. S5), and IDEEL indicated that the number of coding genes with length within 5% of their most similar reference sequence considerably increased (2.7-fold, Text S1, Fig. S6). All gene length analyses followed similar patterns and linear regression demonstrated that the number of coding genes within 5% of their most similar reference gene length was strongly correlated to the total number of base pairs in coding genes ($R^2 \geqq 0.94$, p-adj $<<0.05$, Fig. 2). This means that less resource-exhaustive analyses can act as an equally good assembly quality indicator. We acknowledge that the resulting data are unevenly distributed across both dependent and independent variables in this analysis, but the correlation holds when looking at only the first three SR polishing iterations as well, albeit weaker (adjusted R^2 0.6-0.75, p-value $<<0.05$, data not shown). Furthermore, several similar calculations such as total number of coding genes, and median number of coding genes and bps in coding genes normalized to contig length (data not shown) were also

correlated, and therefore these simpler, less resource-exhaustive assembly quality indicators are robust proxies to reduce computational resource requirements and database-dependency for gene fragmentation estimates from BUSCO or IDEEL.

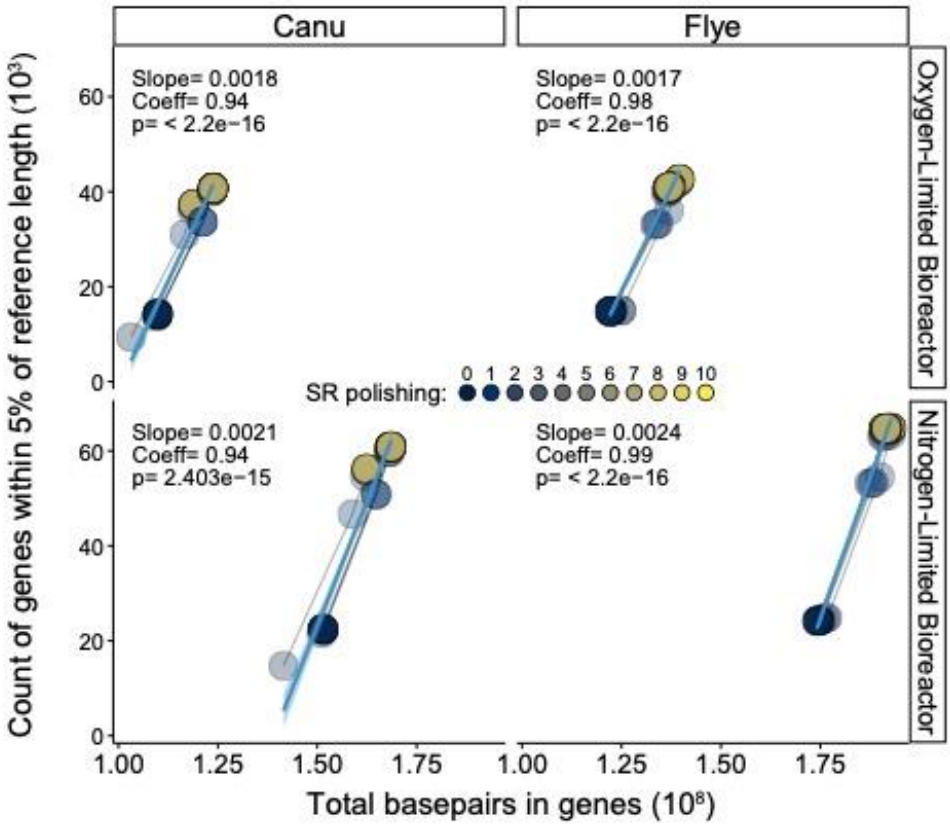

Figure 2. Linear correlation between calculated bps in genes and the number of these genes estimated to be similar length to reference database entries for each bioreactor and Long Read (LR) assembly throughout the iterative LR correction and Short Read (SR) polishing processes.

The two bioreactors are separated over vertical panels, and the two LR assemblers are separated over the horizontal panels. Each point is colored by the SR polishing iteration, with lines connecting the points from the same LR correction iteration, all of which are partially transparent. The light blue solid lines and shaded regions are the linear regressions for the displayed data and their 95% confidence intervals. Slopes, correlation coefficients (R^2-adjusted) and p-values are displayed in upper-left corner of each panel.

Short read recruitment

Multiple iterations of correction and polishing are performed in published literature and suggested by community guidance because it is expected that newly fixed errors allow different read pools to align. Then theoretically, read pools would stabilize when the assembly achieves high accuracy. Additionally, in a community where multiple strains co-exist and/or species encode duplicated genes, the quantity of reads that are ambiguously recruited (*i.e.*, aligning to two or more positions) should decrease as the accuracy of the sequences improves to resolve minor sequence differences. Here, information-rich outputs from BBmap were analyzed for the total number of aligned SR bps, assembled contigs to which no SRs aligned, and ambiguously aligned SRs to observe the magnitudes in which these features change during LR correction and SR polishing.

Consistent with gene fragmentation changes, LR correction offered minimal improvement and SR polishing led to drastic improvement in SR recruitment profiles. Unexpectedly, sequencing depth had little effect on SR recruitment results, in contrast to the used assembler program. Here, Flye assemblies initially recruited more SR bps (1.64-fold) and had smaller percentage of contigs that did not recruit any SRs (2.59-fold, Text S1). These data along with the larger proportion of SRs that align ambiguously (2.02-fold, Text S1) suggests that the Flye assemblies may be of higher quality and possibly capable of reconstructing similar yet distinct strains. On average, LR correction led to approximately 25-35% increases in total SR bps recruited, SR bps recruited normalized to contig length, and SR alignment ambiguity, but only about a 10% reduction in contigs lacking SR recruitment (Text S1). In contrast, SR polishing greatly improved SR recruitment across all measured metrics, including increased total SR bps recruited (3-fold) and median SR bps recruited normalized to contig length (10-fold), fewer contigs lacking any SR recruitment (30% reduction), and a smaller fraction of ambiguously aligned SRs (26% reduction, Fig. 3, Text S1).

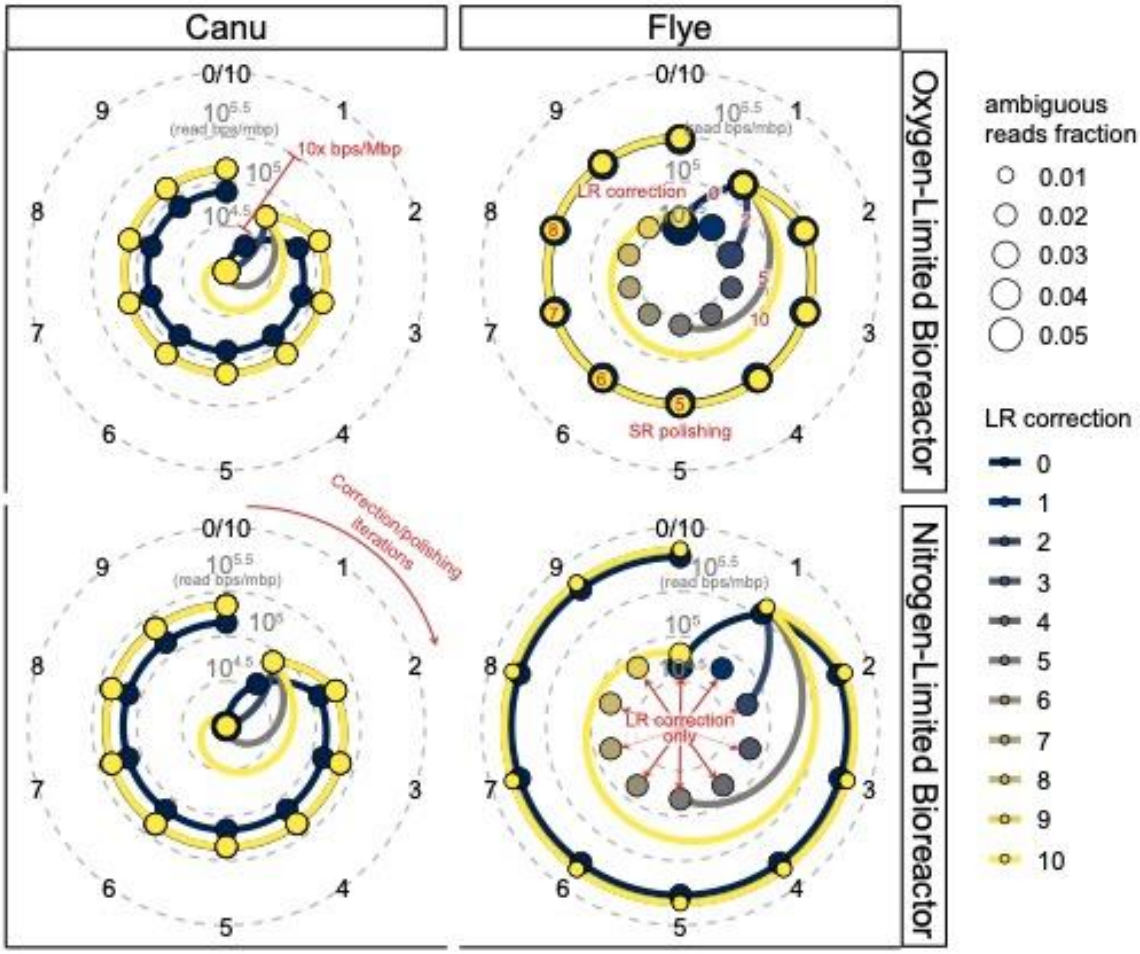

Figure 3. Polar coordinate plot of Short Read (SR) recruitment for each bioreactor and Long Read (LR) assembly throughout the iterative LR correction and SR polishing processes.

The two bioreactors are separated over vertical panels, and the two long read (LR) assemblers are separated over the horizontal panels. Each panel shows a polar coordinate plot of the LR correction or SR polishing iteration in a circle with the bps in SR recruited per contig per Mbp sequence length as the distance from the center. The innermost ten points (due to lower values) indicate the amount of bps in SR recruited after up to ten LR correction iterations without SR polishing. The colored lines connecting points of the same color link assemblies with the same number of LR correction steps preceding SR polishing. The size of the points indicates the fraction of SR that were ambiguously recruited, *i.e.*, SR aligning to 2 or more positions.

Automated bin recovery

A primary goal for many microbial community metagenomic sequencing approaches is to reconstruct high-quality, ideally essentially complete genomes of as many of the members as possible. With these data, it is possible, and is becoming more tractable for a wider range of scientists, to link genetic information with (non)biological phenomena from gene or protein expression to biochemistry or kinetics and even incorporating it into larger-scale modeling efforts. Therefore, attaining optimal microbial genomic data is the desired outcome for a hybrid metagenomic assembly and error-fixing strategy. Assemblies at each stage of LR correction and SR polishing were automatically binned, and both the assemblies and bins were assessed with CheckM to estimate the quality of microbial genomes and estimate the redundancy by counting the 104 bacterial marker genes and reporting the copy number of each gene and the estimated completion and contamination percentages.

Automated bin recovery followed similar patterns as gene fragmentation and SR recruitment, displaying major improvements due to SR polishing and minor due to LR correction. LR assemblies themselves yielded very few medium-quality or better (MQ) automated bins, only 9.5 on average. LR correction increased the number of MQ bins by three to 12.3 on average, but SR polishing alone almost doubled the number of MQ bins, and in combination with LR correction more than doubled the initial MQ bin count on average (Fig. 4, Text S1, Fig. S7). Additionally, the mean completeness scores for MQ bins remained in the low 70% range after only LR correction but increased to over 80% as a result of SR polishing (Fig. 4, Text S1, Fig. S7), indicating an increase in the quality of recovered genomic contents. Furthermore, the substantial increases in MQ bin count and quality due to SR polishing were matched by large increases in the quantity of both number of contigs (7.17-fold) and bps (1.92-fold) contained in MQ bins (Fig. 4, Text S1, Fig. S7), suggesting that SR polishing improved the ability for contigs to be binned. Akin to previous analyses, MQ bin recovery in terms of count, average quality, and data included tended to peak or stabilize as a result of both LR correction and SR polishing, but the exact yields were noisy for clear saturation signal like most other analyses, indicating that automated binning at any of these points should still undergo manual curation prior to downstream analyses.

Bacterial marker gene redundancy estimated by CheckM for the full assemblies followed similar patterns as automated bin recovery and most other analyses. The estimated copy number of marker genes, as well as redundancy, *i.e.*, the sum of CheckM completion and contamination scores reached their maximum values after two iterations of SR polishing, with greatest redundancy after both LR correction and SR polishing (Fig S7, Text S1). Intriguingly, this estimated redundancy of the assemblies was strongly correlated ($R^2 \geqq 0.85$, $p \ll 0.05$) to the number of MQ bins recovered (Fig S8, Text S1), and could be used as a preliminary indicator of genome yield prior to automated binning.

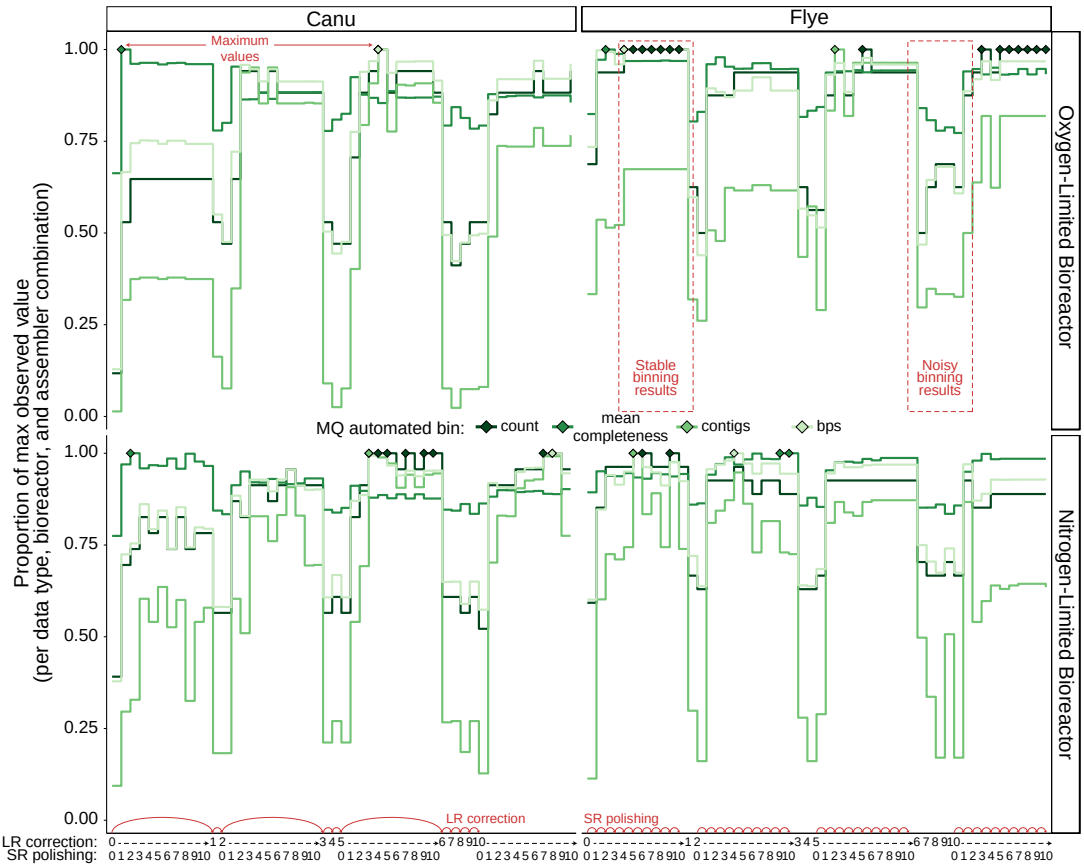

Figure 4. Stair plots of automated medium-quality or better (MQ) bin recovery for each bioreactor and Long Read (LR) assembly throughout the iterative LR correction and Short Read (SR) polishing processes.

The two bioreactors are separated over vertical panels, the two LR assemblers are separated over the horizontal panels. The long read (LR) correction and SR polishing iterations are spread across the x-axis so that the ten SR polishing steps are immediately to the right of the respective preceding LR correction step. Each colored line represents the proportion of the maximum observed

value per combination of bioreactor and assembler of one of the following characteristics MQ automated bins: the number of MQ bins recovered ("count"), the mean completeness score of the bins estimated by CheckM ("mean completeness"), the number of contigs in MQ bins ("contigs"), and the quantity of bps in MQ bins ("bps"). Matching color diamonds indicate the LR correction and/or SR polishing stage with the maximum observed value for each characteristic.

Microbial community beta diversity

Complementing genome-centric approaches, many studies perform gene-centric analyses that better handle poorly assembled community members. The RNA Polymerase subunit B (*rpoB*) gene is a phylogenetic marker for all domains of life and was used here for beta-diversity clustering analysis by using their SR recruitment profiles via Non-Metric multi-Dimensional Scaling (NMDS). This would then also combine the various types of data already discussed, particularly improvements for gene fragmentation should yield better microbial community representation and changes in SR recruitment pools should affect the abundance of these genes.

Inferred microbial community beta diversity was strongly influenced by both LR correction and SR polishing, but only SR polishing led to stabilized clusters. Initial LR assemblies and assemblies with LR correction alone contained fewer (2 total genes) and more fractured (1/3 the mean length) *rpoB* genes compared to assemblies after SR polishing (Text S1, Figs. S9). Unexpectedly, read depths on average appeared to decrease due to SR polishing (Text S1, Figs. S10), presumably due to the identification of additional *rpoB* genes on low abundance contigs after SR polishing made these genes intact. Both improved *rpoB* recovery and depth profiles had clear effects on the trajectories of their clustering: assemblies with only LR correction were noisy compared to the convergence observable from SR polishing (Fig. 5). In most cases, assemblies with SR polishing alone were distant from assemblies with both LR correction and SR polishing, highlighting the beneficial impact of LR correction only after SR polishing.

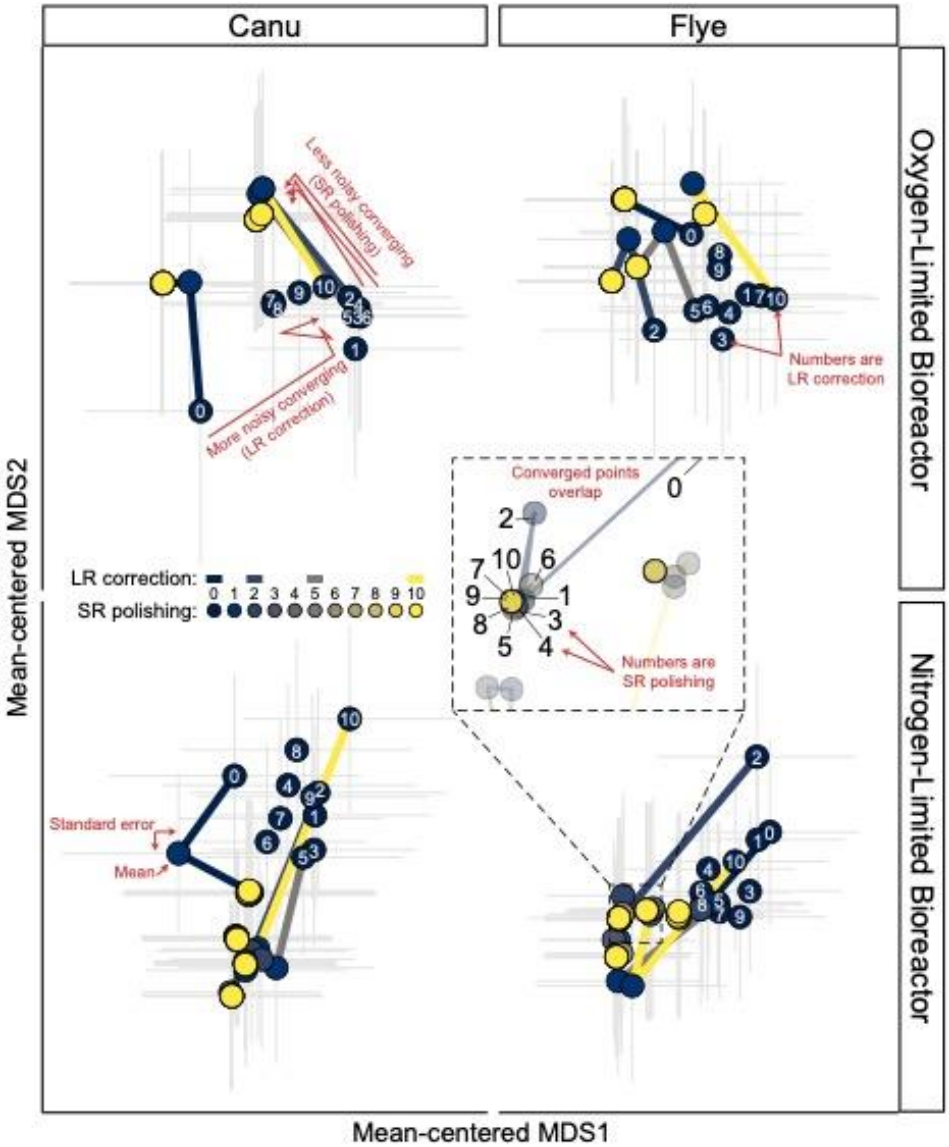

Figure 5. Beta diversity estimated from RNA Polymerase subunit B (*rpoB*) gene profiles for each bioreactor and Long Read (LR) assembly throughout the iterative LR correction and Short Read (SR) polishing processes.

The two bioreactors are separated over vertical panels, the two LR assemblers are separated over the horizontal panels. Points are the mean-centered species scores for all *rpoB* genes per assembly, with gray lines showing the standard errors of the mean. Each point is colored by the SR polishing iteration, with colored lines connecting points to the preceding LR correction and subsequent SR polishing iteration. White text over points indicates the LR correction iteration prior to any SR polishing. Inset shows the overlapping points of the

second LR correction stage as the communities converge due to SR polishing for the NLR bioreactor's Flye assembly.

Discussion

Microbial community hybrid assembly optimization leveraging long reads for correction and short reads for polishing

Assembly of LRs offers a major advantage over SR data alone, due to its greater contiguity. However, it comes at a significant cost due to lower accuracy. Integrated, or hybrid, assembly approaches leverage complementary beneficial attributes of both LR and SR platforms to offset their weaknesses. Programs already exist to perform this integration, often by SR assembly followed by LR bridging the assembly graphs[15], and some can even handle mixed microbial communities[3]. Unfortunately, community complexity appears to be an obstacle leading to fractured SR assemblies that these programs cannot overcome (Fig. S1), likely, at least in part, because these programs were not developed for these purposes. Therefore, we focused on leveraging the contiguity of LR assemblies and with further LR correction and SR polishing to yield high quality microbial community metagenomic data.

To gauge the quality of an assembly, often mis-assemblies are quantified by comparison to reference databases, which however is not always ideal for poorly characterized microorganisms or ecosystems. We thus monitored relatively reference-independent assembly characteristics. Specifically changes in gene fragmentation and SR recruitment were used as indicators of assembly accuracy, and automated bin recovery and beta diversity showed how these processes would affect microbial community reconstruction for downstream analyses. Fortunately, essentially all four of these characteristics followed the same overall patterns of improvements in sequence accuracy and community reconstruction: (1) LR correction alone was insufficient; (2) the greatest improvements occurred within the first three iterations of SR polishing; (3) although the effects were not always immediately apparent, the assemblies benefited from two to five iterations LR correction and any changes after could be attributed to noise; (4) beyond five iterations of SR polishing after LR correction provided severely diminished returns. Furthermore, simple and reference-independent analyses were robustly

correlated with more complicated and/or reference-dependent analyses, for example, total bps in genes and count of genes within 5% of the most similar reference length. Therefore, computational costs can be reduced by using relatively simple calculations predictive of improvements to genome- and gene-centric analyses. Rather than inform other users of the best approach for their own hybrid metagenomic microbial community projects, this study emphasizes that some, but not exhaustive LR correction and SR polishing are necessary to obtain high-quality metagenomic assemblies and equips users with simple analyses to assess the optimum assembly without relying on reference datasets.

Unexplored factors and issues in this study

First, recent technological advances for both LR sequencing and application may ease the approach for integrating hybrid metagenomic assembly of microbial communities, but this study was performed using older sequencing chemistry and basecalling algorithms. Both major LR platforms, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have achieved large improvements in chemistry and basecalling that have increased both depth and accuracy of LR sequencing[36,37]. Therefore, not only will LR-alone assembly be more accurate as a result of these advances, it may even become sufficient for retrieving HQ genomes without need for any SR polsihing[5,38]. However, resources may restrict users to older and more accessible technologies, or projects may already have older datasets that still need analysis. Additionally, for the foreseeable future, SR accompaniments to LR data will likely continue to be implemented as a far cheaper and more accessible means to increase the sample count for differential coverage binning. Furthermore, SRs may still aid assembly quality if applied properly by implementing contemporary hybrid assembly and error-fixing approaches[39]. Therefore, combining both sequencing technologies may be more cost-effective than either platform alone[5] to yield the highest quality data.

Second, the myriad of programs and pipelines to process LRs from read correction, to assembly, to assembly correction, SR recruitment and polishing, and more continuously improve and expand[4]. To reasonably perform this study and analyze the results, only two assembly, one LR correction, and one SR

polishing programs were tested. These programs were chosen based on their observed prevalence in literature and online resources, but are beginning to be outcompeted by others that yield better results[40]. Additionally, not only are there more and likely better program options, but these programs also come with a suite of settings to fine-tune performance that were not explored. One setting of particular interest that was not investigated here was read recruitment stringency, which could greatly impact SR polishing of communities with multiple similar strains, or genomes with multicopy genes or repeat regions that will recruit reads ambiguously. However, it is not realistic for most labs to explore all options, and the intent of this study was to explore assembly statistics that were both relatively reference- and tool-agnostic.

Third, though relatively reference-agnostic, many analyses here explored only bacterial phylogenetic markers. It is possible that communities with, or a focus upon, eukaryotic or archaeal members may require modified approaches or analyses given sometimes large differences in genomic architectures. Preliminary analyses of the bioreactors sequenced in our study suggested an absence of archaea, but cultivation work repeatedly overrun by recurrent fungi indicates the possibility of Eukaryote presence. Cursory analyses of the *rpoB* gene affiliations from these and related datasets include at least three distinct probable amoeba endosymbionts[41–43] (70-100% amino acid identity) and at least three distinct ciliates (48-59% amino acid identity), but still no archaea. Additionally, viruses and plasmids, which often assemble poorly but can significantly impact microbial community ecology and physiology, were largely ignored in this study. Given the plentiful small circular contigs, or larger sometimes circular contigs with essentially no bacterial markers, these non-bacterial community members may be prevalent in these bioreactor communities, and may require different approaches to attain optimal data quality[6,9]. Additionally, many cutoffs for genomic quality also excluded high quality but unusual bacterial genomes. For example, members of Candidate Phyla Radiation lineages tend to be small with low completion estimates despite being complete and circularized[44]. Cursory phylogenetic investigation of circular contigs from the initial assemblies included members of the Gracilibacteria (formerly BD1-5)[45]. For most of these cases, other algorithms or

settings can be implemented to better identify Eukaryotes, Archaea, viruses, plasmids, and unusual bacteria.

Fourth, contig pools were not separated for some analyses in order to simplify the workflow. Specifically, in this workflow circular contigs that likely constituted nearly complete bacterial genomes were not removed from the assemblies prior to binning, which could result in chimeric automated bins. We note here that across all LR assemblies, the mean and median automated bin (completion >0%) contamination were 0.6% and 0%, respectively, though 51 of 14,965 bins did have contamination greater than 20%. The high degree of strain heterogeneity (35-43%) in automated bins with contamination >20% indicates that a substantial fraction of chimeric automated bins were composed of similar strains. Similar to separating circular and large contigs prior to binning, performing LR correction and SR polishing on contigs that are essentially complete genomes separately from the rest of the assembly may lead to greater improvements by potentially reducing read recruitment ambiguity.

Fifth and lastly, we acknowledge that the biomass collected for DNA isolation and sequencing with the different platforms were separated by several months. Though largely stable after enrichment for over five years, the communities still shift slowly, thought largely due to strains competing for the same niches, which will be examined in follow-up publications. These time differences could impact ideal integration the LR and SR data if strains substantially shifted in abundance, leading to improper error-fixing events and could produce inaccurate consensus genomes rather than a realistic strain representation.

Conclusions

Integrating LR and SR sequencing data for a microbial community is a complex task. Not only can the biology be difficult to analyze for a variety of reasons, but complementing the strengths and weaknesses of sequencing platforms or computational workflows is challenging. Despite the improvements to LR sequencing data that appear to preclude the need for SRs, we expect that SRs will still be beneficial as low cost additional data for differential coverage, and users may still desire to explore SR polishing to maximize their assembly quality. Here, we've tested multiple assemblers and variations on iterative

correction and polishing, and performed analyses that aid empirically driven decisions for integrating these datasets. Fortunately, commonly applied workflows and word-of-mouth conventions were validated, briefly that use of two tools (typically one tool using LRs and one SRs) with two iterations each is sufficient, because the majority of improvement occurred within the first two iterations of LR correction and SR polishing. It was apparent from all analyses that the combination of LR correction and SR polishing yielded higher quality assemblies than either alone, and also that neither process needed to be iterated more than five times. These patterns were similar for gene fragmentation and SR recruitment, as well as automated bin recovery and community clustering trajectories, which enables simple calculations estimate the results more complex analyses to save computational resources. Finally, we offer guidelines to enable other users to integrate LR and SR sequencing platforms for microbial community metagenomes with fewer obstacles and guided by empirical data from their own projects.

Practical recommendations for users

- Use the same pool or replicate extractions of DNA for both SR and LR, which ideally minimizes community differences that might lead to mis-assembly.
- Calculate read depth using bps rather than counts when mixing LR and SR profiles, otherwise LR depth will be vastly and unevenly underestimated.
- If performing assembly-level marker gene inventory analyses using programs like BUSCO or CheckM, use a separate gene-caller to identify ORFS as an input to save resources and improve accuracy.
- If using these tools and parameters, limit LR correction to 2-5 iterations and SR polishing to 5 iterations to preserve resources.
    - Choose any LR correction iteration to follow with SR polishing; there were few trends or notable differences between them.
    - Choose at least 2 and up to 5 SR polishing iterations; 1 iteration led to great improvements but 2 or 3 led to saturation of most observed features, and the quantity of changes and bps changed

during SR polishing largely plateaued on a log-scale by iteration 4 (Fig. S11). Increasing the read recruitment stringency, which we suggest, likely requires more iterations due to theoretically fewer mis-assemblies being corrected each time.

    - A mixture of different tools may improve the overall quality[40], possibly due to differential abilities to fix certain errors, but LR correction should only be performed prior to SR polishing to avoid the (re-)introduction of errors.

- While the ultimate goal might be genome recovery, we suggest monitoring gene fragmentation or data collected during LR correction or SR polishing for empirically guiding the endpoints rather than genome recovery.
    - Essentially all gene- and genome-centric analyses followed the same trajectories, so simpler analyses should serve as reasonable proxies for more complex options.
    - Gene sizes are relatively simple to compute after running a gene-caller like Prodigal, which can also be precursor to downstream analyses.
    - Logged data from programs like Pilon are information-rich and contain, for example, the number of changes and the amount of bps changed, which could guide users.
    - SR recruitment is an additional valuable tool to guide users' hybrid approach, because many downstream analyses depend on read depths, but costs more computational resources than other options.
- If possible, use multiple assembly programs prior to LR correction and SR polishing, and then consolidate the genome outputs (circular, MQ automated bins) because different algorithms may have a substantial impact on which community members assemble well[18].

References

(1) Overholt, W. A.; Hölzer, M.; Geesink, P.; Diezel, C.; Marz, M.; Küsel, K. Inclusion of Oxford Nanopore Long Reads Improves All Microbial and Viral Metagenome-Assembled Genomes from a Complex Aquifer System. *Environmental Microbiology* **2020**, *22* (9), 4000–4013. https://doi.org/10.1111/1462-2920.15186.

(2) Singleton, C. M.; Petriglieri, F.; Kristensen, J. M.; Kirkegaard, R. H.; Michaelsen, T. Y.; Andersen, M. H.; Kondrotaite, Z.; Karst, S. M.; Dueholm, M. S.; Nielsen, P. H.; Albertsen, M. Connecting Structure to Function with the Recovery of over 1000 High-Quality Metagenome-Assembled Genomes from Activated Sludge Using Long-Read Sequencing. *Nat Commun* **2021**, *12* (1), 2009. https://doi.org/10.1038/s41467-021-22203-2.

(3) Bertrand, D.; Shaw, J.; Kalathiyappan, M.; Ng, A. H. Q.; Kumar, M. S.; Li, C.; Dvornicic, M.; Soldo, J. P.; Koh, J. Y.; Tong, C.; Ng, O. T.; Barkham, T.; Young, B.; Marimuthu, K.; Chng, K. R.; Sikic, M.; Nagarajan, N. Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes. *Nat Biotechnol* **2019**, *37* (8), 937–944. https://doi.org/10.1038/s41587-019-0191-2.

(4) Amarasinghe, S. L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M. E.; Gouil, Q. Opportunities and Challenges in Long-Read Sequencing Data Analysis. *Genome Biology* **2020**, *21* (1), 30. https://doi.org/10.1186/s13059-020-1935-5.

(5) Sereika, M.; Kirkegaard, R. H.; Karst, S. M.; Michaelsen, T. Y.; Sørensen, E. A.; Wollenberg, R. D.; Albertsen, M. Oxford Nanopore R10.4 Long-Read Sequencing Enables the Generation of near-Finished Bacterial Genomes from Pure Cultures and Metagenomes without Short-Read or Reference Polishing. *Nature Methods* **2022**, *19* (7), 823–826. https://doi.org/10.1038/s41592-022-01539-7.

(6) Wick, R. R.; Holt, K. E. Benchmarking of Long-Read Assemblers for Prokaryote Whole Genome Sequencing. F1000Research February 1, 2021. https://doi.org/10.12688/f1000research.21782.4.

(7) Chen, Z.; Erickson, D. L.; Meng, J. Benchmarking Hybrid Assembly Approaches for Genomic Analyses of Bacterial Pathogens Using Illumina and Oxford Nanopore Sequencing. *BMC Genomics* **2020**, *21* (1), 631. https://doi.org/10.1186/s12864-020-07041-8.

(8) Chen, Z.; Erickson, D. L.; Meng, J. Polishing the Oxford Nanopore Long-Read Assemblies of Bacterial Pathogens with Illumina Short Reads to Improve Genomic Analyses. *Genomics* **2021**, *113* (3), 1366–1377. https://doi.org/10.1016/j.ygeno.2021.03.018.

(9) Zablocki, O.; Michelsen, M.; Burris, M.; Solonenko, N.; Warwick-Dugdale, J.; Ghosh, R.; Pett-Ridge, J.; Sullivan, M. B.; Temperton, B. VirION2: A Short- and Long-Read Sequencing and Informatics Workflow to Study the Genomic Diversity of Viruses in Nature. *PeerJ* **2021**, *9*, e11088. https://doi.org/10.7717/peerj.11088.

(10) Stewart, R. D.; Auffret, M. D.; Warr, A.; Walker, A. W.; Roehe, R.; Watson, M. Compendium of 4,941 Rumen Metagenome-Assembled Genomes for Rumen Microbiome Biology and Enzyme Discovery. *Nat*

Biotechnol **2019**, *37* (8), 953–961. https://doi.org/10.1038/s41587-019-0202-3.

(11) *Nanopore Community*. Oxford Nanopore Technologies. https://nanoporetech.com/community (accessed 2023-04-05).

(12) *BBMap*. SourceForge. https://sourceforge.net/projects/bbmap/ (accessed 2023-02-14).

(13) Wick, R. Porechop, 2023. https://github.com/rrwick/Porechop (accessed 2023-02-14).

(14) *metaSPAdes – Center for Algorithmic Biotechnology*. https://cab.spbu.ru/software/meta-spades/ (accessed 2023-02-14).

(15) Wick, R. R.; Judd, L. M.; Gorrie, C. L.; Holt, K. E. Unicycler: Resolving Bacterial Genome Assemblies from Short and Long Sequencing Reads. *PLOS Computational Biology* **2017**, *13* (6), e1005595. https://doi.org/10.1371/journal.pcbi.1005595.

(16) *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. https://genome.cshlp.org/content/27/5/722 (accessed 2023-02-14).

(17) Kolmogorov, M.; Bickhart, D. M.; Behsaz, B.; Gurevich, A.; Rayko, M.; Shin, S. B.; Kuhn, K.; Yuan, J.; Polevikov, E.; Smith, T. P. L.; Pevzner, P. A. MetaFlye: Scalable Long-Read Metagenome Assembly Using Repeat Graphs. *Nat Methods* **2020**, *17* (11), 1103–1110. https://doi.org/10.1038/s41592-020-00971-x.

(18) Wick, R. R.; Judd, L. M.; Cerdeira, L. T.; Hawkey, J.; Méric, G.; Vezina, B.; Wyres, K. L.; Holt, K. E. Trycycler: Consensus Long-Read Assemblies for Bacterial Genomes. *Genome Biology* **2021**, *22* (1), 266. https://doi.org/10.1186/s13059-021-02483-z.

(19) Vaser, R.; Sovic, I.; Nagarajan, N.; Sikic, M. Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads. *Genome Res.* **2017**, gr.214270.116. https://doi.org/10.1101/gr.214270.116.

(20) Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **2018**, *34* (18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

(21) Walker, B. J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C. A.; Zeng, Q.; Wortman, J.; Young, S. K.; Earl, A. M. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **2014**, *9* (11), e112963. https://doi.org/10.1371/journal.pone.0112963.

(22) Mikheenko, A.; Saveliev, V.; Gurevich, A. MetaQUAST: Evaluation of Metagenome Assemblies. *Bioinformatics* **2016**, *32* (7), 1088–1090. https://doi.org/10.1093/bioinformatics/btv697.

(23) Hyatt, D.; Chen, G.-L.; LoCascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **2010**, *11* (1), 119. https://doi.org/10.1186/1471-2105-11-119.

(24) Manni, M.; Berkeley, M. R.; Seppey, M.; Simão, F. A.; Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **2021**, *38* (10), 4647–4654. https://doi.org/10.1093/molbev/msab199.

(25) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res.* **2015**, *25* (7), 1043–1055. https://doi.org/10.1101/gr.186072.114.
(26) *Fast and sensitive protein alignment using DIAMOND | Nature Methods*. https://www.nature.com/articles/nmeth.3176 (accessed 2023-06-01).
(27) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; the UniProt Consortium. UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches. *Bioinformatics* **2015**, *31* (6), 926–932. https://doi.org/10.1093/bioinformatics/btu739.
(28) Chaumeil, P.-A.; Mussig, A. J.; Hugenholtz, P.; Parks, D. H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. *Bioinformatics* **2020**, *36* (6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.
(29) SorenKarst. Mmlong, 2022. https://github.com/SorenKarst/mmlong (accessed 2023-02-14).
(30) Kang, D. D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*, e7359. https://doi.org/10.7717/peerj.7359.
(31) Shaffer, M.; Borton, M. A.; McGivern, B. B.; Zayed, A. A.; La Rosa, S. L.; Solden, L. M.; Liu, P.; Narrowe, A. B.; Rodríguez-Ramos, J.; Bolduc, B.; Gazitúa, M. C.; Daly, R. A.; Smith, G. J.; Vik, D. R.; Pope, P. B.; Sullivan, M. B.; Roux, S.; Wrighton, K. C. DRAM for Distilling Microbial Metabolism to Automate the Curation of Microbiome Function. *Nucleic Acids Research* **2020**, *48* (16), 8883–8900. https://doi.org/10.1093/nar/gkaa621.
(32) Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **2011**, *7* (10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
(33) Oksanen, J.; Simpson, G. L.; Blanchet, F. G.; Kindt, R.; Legendre, P.; Minchin, P. R.; O'Hara, R. B.; Solymos, P.; Stevens, M. H. H.; Szoecs, E.; Wagner, H.; Barbour, M.; Bedward, M.; Bolker, B.; Borcard, D.; Carvalho, G.; Chirico, M.; Caceres, M. D.; Durand, S.; Evangelista, H. B. A.; FitzJohn, R.; Friendly, M.; Furneaux, B.; Hannigan, G.; Hill, M. O.; Lahti, L.; McGlinn, D.; Ouellette, M.-H.; Cunha, E. R.; Smith, T.; Stier, A.; Braak, C. J. F. T.; Weedon, J. *Vegan: Community Ecology Package*; 2022.
(34) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2021.
(35) Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; Yutani, H. Welcome to the Tidyverse. *Journal of Open Source Software* **2019**, *4* (43), 1686. https://doi.org/10.21105/joss.01686.
(36) *PacBio Revio | Long-read sequencing at scale*. PacBio. https://www.pacb.com/revio/ (accessed 2023-04-04).

(37) *Continuous development and improvement*. Oxford Nanopore Technologies. https://nanoporetech.com/about-us/continuous-development-and-improvement (accessed 2023-04-04).
(38) Zhao, W.; Zeng, W.; Pang, B.; Luo, M.; Peng, Y.; Xu, J.; Kan, B.; Li, Z.; Lu, X. Oxford Nanopore Long-Read Sequencing Enables the Generation of Complete Bacterial and Plasmid Genomes without Short-Read Sequencing. *Frontiers in Microbiology* **2023**, *14*.
(39) Zhang, H.; Jain, C.; Aluru, S. A Comprehensive Evaluation of Long Read Error Correction Methods. *BMC Genomics* **2020**, *21* (6), 889. https://doi.org/10.1186/s12864-020-07227-0.
(40) Wick, R. R.; Holt, K. E. Polypolish: Short-Read Polishing of Long-Read Bacterial Genome Assemblies. *PLOS Computational Biology* **2022**, *18* (1), e1009802. https://doi.org/10.1371/journal.pcbi.1009802.
(41) Schulz, F.; Lagkouvardos, I.; Wascher, F.; Aistleitner, K.; Kostanjšek, R.; Horn, M. Life in an Unusual Intracellular Niche: A Bacterial Symbiont Infecting the Nucleus of Amoebae. *ISME J* **2014**, *8* (8), 1634–1644. https://doi.org/10.1038/ismej.2014.5.
(42) Birtles, R. J.; Rowbotham, T. J.; Michel, R.; Pitcher, D. G.; Lascola, B.; Alexiou-Daniel, S.; Raoult, D. “Candidatus Odyssella Thessalonicensis” Gen. Nov., Sp. Nov., an Obligate Intracellular Parasite of Acanthamoeba Species. *International Journal of Systematic and Evolutionary Microbiology* **2000**, *50* (1), 63–72. https://doi.org/10.1099/00207713-50-1-63.
(43) Horn, M.; Wagner, M.; Müller, K.-D.; Schmid, E. N.; Fritsche, T. R.; Schleifer, K.-H.; Michel, R. Neochlamydia Hartmannellae Gen. Nov., Sp. Nov. (Parachlamydiaceae), an Endoparasite of the Amoeba Hartmannella Vermiformis. *Microbiology (Reading)* **2000**, *146 (Pt 5)*, 1231–1239. https://doi.org/10.1099/00221287-146-5-1231.
(44) Chen, L.-X.; Anantharaman, K.; Shaiber, A.; Eren, A. M.; Banfield, J. F. Accurate and Complete Genomes from Metagenomes. *Genome Res.* **2020**, *30* (3), 315–333. https://doi.org/10.1101/gr.258640.119.
(45) Sieber, C. M. K.; Paul, B. G.; Castelle, C. J.; Hu, P.; Tringe, S. G.; Valentine, D. L.; Andersen, G. L.; Banfield, J. F. Unusual Metabolism and Hypervariation in the Genome of a Gracilibacterium (BD1-5) from an Oil-Degrading Community. *mBio* **2019**, *10* (6), e02128-19. https://doi.org/10.1128/mBio.02128-19.