Optimizing *de novo* genome assembly from PCR-amplified metagenomes

Simon Roux [Corresp., 1], **Gareth Trubl** [2], **Danielle Goudeau** [1], **Nandita Nath** [1], **Estelle Couradeau** [3], **Nathan A Ahlgren** [4], **Yuanchao Zhan** [5], **David Marsan** [5], **Feng Chen** [5], **Jed A Fuhrman** [6], **Trent R Northen** [1], **Matthew B Sullivan** [2, 7], **Virginia I Rich** [2], **Rex R Malmstrom** [1], **Emiley A Eloe-Fadrosh** [Corresp. 1]

[1] Department of Energy Joint Genome Institute, Walnut Creek, California, United States
[2] Department of Microbiology, Ohio State University, Columbus, Ohio, United States
[3] Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, California, United States
[4] Department of Biology, Clark University, Worcester, Massachusetts, United States
[5] Institution of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Cambridge, Maryland, United States
[6] Department of Biological Sciences, University of Southern California, Los Angeles, California, United States
[7] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, Ohio, United States

Corresponding Authors: Simon Roux, Emiley A Eloe-Fadrosh
Email address: sroux@lbl.gov, eaeloefadrosh@lbl.gov

Background.

Metagenomics has transformed our understanding of microbial diversity across ecosystems, with recent advances enabling de novo assembly of genomes from metagenomes. These metagenome-assembled genomes are critical to provide ecological, evolutionary, and metabolic context for all the microbes and viruses yet to be cultivated. Metagenomes can now be generated from nanogram to subnanogram amounts of DNA. However, these libraries require several rounds of PCR amplification before sequencing, and recent data suggest these typically yield smaller and more fragmented assemblies than regular metagenomes.

Methods.

Here we evaluate de novo assembly methods of 169 PCR-amplified metagenomes, including 25 for which an unamplified counterpart is available, to optimize specific assembly approaches for PCR-amplified libraries. We first evaluated coverage bias by mapping reads from PCR-amplified metagenomes onto reference contigs obtained from unamplified metagenomes of the same samples. Then, we compared different assembly pipelines in terms of assembly size (number of bp in contigs ≥ 10kb) and error rates to evaluate which are the best suited for PCR-amplified metagenomes.

Results.

Read mapping analyses revealed that the depth of coverage within individual genomes is

significantly more uneven in PCR-amplified datasets versus unamplified metagenomes, with regions of high depth of coverage enriched in short inserts. This enrichment scales with the number of PCR cycles performed, and is presumably due to preferential amplification of short inserts. Standard assembly pipelines are confounded by this type of coverage unevenness, so we evaluated other assembly options to mitigate these issues. We found that a pipeline combining read deduplication and an assembly algorithm originally designed to recover genomes from libraries generated after whole genome amplification (single-cell SPAdes) frequently improved assembly of contigs ≥ 10kb by 10 to 100-fold for low input metagenomes.

Conclusions.
PCR-amplified metagenomes have enabled scientists to explore communities traditionally challenging to describe, including some with extremely low biomass or from which DNA is particularly difficult to extract. Here we show that a modified assembly pipeline can lead to an improved de novo genome assembly from PCR-amplified datasets, and enables a better genome recovery from low input metagenomes.

Optimizing *de novo* genome assembly from PCR-amplified metagenomes

Simon Roux[1*], Gareth Trubl[2], Danielle Goudeau[1], Nandita Nath[1], Estelle Couradeau[3], Nathan A Ahlgren[4], Yuanchao Zhan[5], David Marsan[5], Feng Chen[5], Jed A Fuhrman[6], Trent R. Northen[1,3], Matthew B. Sullivan[2,7], Virginia I Rich[2], Rex R. Malmstrom[1], Emiley A. Eloe-Fadrosh[1*]

[1] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA, USA
[2] Department of Microbiology, The Ohio State University, Columbus, OH, USA
[3] Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4] Department of Biology, Clark University, Worcester, MA, USA
[5] Institution of Marine and Environmental Technology, University of Maryland Center for Environmental Science, Cambridge, MD, USA
[6] Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA
[7] Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA

* Corresponding Authors:
Emiley A. Eloe-Fadrosh, Simon Roux
Lawrence Berkeley National Laboratory
1 Cyclotron Road Mailstop 100PGF100
Berkeley, CA, 94720, USA
Email address: EAE-F eaeloefadrosh@lbl.gov, SR sroux@lbl.gov

Abstract

Background.

Metagenomics has transformed our understanding of microbial diversity across ecosystems, with recent advances enabling *de novo* assembly of genomes from metagenomes. These metagenome-assembled genomes are critical to provide ecological, evolutionary, and metabolic context for all the microbes and viruses yet to be cultivated. Metagenomes can now be generated from nanogram to subnanogram amounts of DNA. However, these libraries require several rounds of PCR amplification before sequencing, and recent data suggest these typically yield smaller and more fragmented assemblies than regular metagenomes.

Methods.

Here we evaluate *de novo* assembly methods of 169 PCR-amplified metagenomes, including 25 for which an unamplified counterpart is available, to optimize specific assembly approaches for PCR-amplified libraries. We first evaluated coverage bias by mapping reads from PCR-amplified metagenomes onto reference contigs obtained from unamplified metagenomes of the same samples. Then, we compared different assembly pipelines in terms of assembly size (number of bp in contigs ≥ 10kb) and error rates to evaluate which are the best suited for PCR-amplified metagenomes.

Results.

Read mapping analyses revealed that the depth of coverage within individual genomes is significantly more uneven in PCR-amplified datasets versus unamplified metagenomes, with regions of high depth of coverage enriched in short inserts. This enrichment scales with the number of PCR cycles performed, and is presumably due to preferential amplification of short inserts. Standard assembly pipelines are confounded by this type of coverage unevenness, so we evaluated other assembly options to mitigate these issues. We found that a strategy combining read deduplication and an assembly algorithm originally designed to recover genomes from libraries generated after whole genome amplification (single-cell SPAdes) frequently improved assembly of contigs ≥ 10kb by 10 to 100-fold for low input metagenomes.

Conclusions.

PCR-amplified metagenomes have enabled scientists to explore communities traditionally challenging to describe, including some with extremely low biomass or from which DNA is particularly difficult to extract. Here we show that a modified assembly strategy can lead to an improved *de novo* genome assembly from PCR-amplified datasets, and enables a better genome recovery from low input metagenomes.

Introduction

Microbes and their associated viruses dominate all ecosystems on Earth and drive major biogeochemical cycles [1,2]. The vast majority of this microbial and viral diversity has not yet been cultivated [3,4], hence metagenomics, i.e. the sequencing of genomes directly from environmental samples, has emerged as a key method to explore these communities [5,6]. Briefly, DNA is extracted from an environmental sample, sometimes after selecting a subset of the community (e.g. the viruses), and sequenced, typically as short sequencing "reads". These reads are assembled into larger contigs, interpreted as genome fragments, which provides the foundation to investigate functional, ecological, and evolutionary patterns of the largely uncultivated microbial and viral diversity [7–16].

Problematically, as metagenomics is applied to a broader set of samples, some yield very little DNA (e.g. a few nanograms), which poses a challenge for library construction [17]. Examples include low-biomass environments like ice cores or clean rooms [18,19], tough-to-sample locations like hydrothermal vents [11], and sampling procedures that target subsets of the community, e.g. virus particles or labeled metabolically active microbes [20,21]. Sequencing libraries from these types of samples require a DNA amplification step either before or after adapter ligation. In the former, extracted DNA is subjected to whole genome amplification (WGA), typically as Multiple Displacement Amplification (MDA)[22] or Sequence-Independent, Single-Primer Amplification (SISPA)[23]. The resultant amplified product is then sufficient for a standard library preparation and sequencing. However, strong amplification biases make these approaches unsuitable for quantitative estimations of taxa or genes relative abundance [24,25]. Alternatively, tagmentation or adaptase protocols allow sub-nanogram DNA inputs for adapter ligation, and then use PCR (typically $\geq$ 9 cycles) to amplify the ligated DNA [17,26]. In contrast to whole genome amplification, these protocols yield metagenomes (hereafter "PCR-amplified metagenomes") for which read mapping enables a quantification of taxa and/or genes, and are thus the methods of choice for low-input metagenomes. [17,25].

While the impact of PCR amplification, sequencing library choice, and sequencing platforms on metagenome reads composition has been extensively studied (e.g. [17,25,27,28]), and specific assemblers have been developed for unamplified and MDA-amplified metagenomes (e.g. [29,30]), evaluation of *de novo* genome assembly from PCR-amplified metagenomes is needed. Here we compared different approaches for *de novo* assembly of PCR-amplified metagenomes generated with two library preparation kits commonly used on low input samples (Nextera XT and Accel-NGS 1S Plus). We show that preferential amplification of short inserts can lead to uneven genome coverage and sub-optimal assembly. We then highlight alternative sequence processing approaches that maximize *de novo* genome assembly for PCR-amplified libraries, which will enable scientists to extract as much information as possible from these datasets.

Materials & Methods

Origin of samples

Samples and libraries generated as part of 6 different projects were used in this study (Table S1). Most of these samples yielded a low amount of DNA, mainly because they targeted a specific community subset such as viruses, cyanobacteria, or metabolically active cells.

The data analyzed here included:

(i) A set of 20 samples from virus fractions along a natural permafrost thaw gradient ("Permafrost-associated viruses" in Table S1). These were generated using a protocol optimized for recovery of soil viruses [33] with minor amendments. Briefly, viruses were resuspended from triplicate soil samples using a combination of chemical and physical dispersion, filtered through a 0.2 µm polyethersulfone membrane filter, and viral DNA was extracted using DNeasy PowerSoil DNA extraction kit (Qiagen, Hilden, Germany, product 12888).

(ii) A set of 14 samples from the viral fraction of Delaware Bay Estuary surface water ("Delaware Bay viruses"). These surface water viral metagenomes were collected during different seasons from the Delaware estuary and Chesapeake estuary using a Niskin bottle on board of the RV Hugh R Sharp. Details of environmental conditions can be found at http://dmoserv3.bco-dmo.org/jg/serv/BCO-DMO/Coast_Bact_Growth/newACT_cruises_rs.html0%7Bdir=dmoserv3.whoi.edu/jg/dir/BCO-DMO/Coast_Bact_Growth/,info=dmoserv3.bco-dmo.org/jg/info/BCO-DMO/Coast_Bact_Growth/new_ACT_cruises%7D. Viral communities were concentrated from 0.2 µm filtrates following the FeCl3 flocculation method [34]. Briefly, 10 L of seawater was prefiltered through a 142 mm-diameter glass fiber filter GA-55 (~0.6µm-pore size, Cole-Parmer) and a 0.22 µm-pore-size Millipore polycarbonate membrane filter (142mm, Millipore) to remove larger organisms and bacteria. One mL of 10g/L FeCl3 stock solution was added to the 10 L filtrate. After incubating with FeCl3 for 1 hr, the concentrated viral fraction was collected using a 0.8 µm-pore-size Millipore polycarbonate membrane filter (Millipore). The concentrated viruses were resuspended using a resuspension buffer, dialyzed to remove the resuspension buffer, and treated with DNase to remove free DNA. The viral DNA was extracted using the phenol-chloroform-isoamyl alcohol method.

(iii) A set of 11 samples from the viral fraction of surface water at the San Pedro Ocean-time Series site (33°33′N, 118°24′W), off the coast of Los Angeles ("SPOT viruses"). Surface water was collected using a Niskin bottle rosette (5 m) or by bucket (0 m). Viral fraction (<0.22 µm) material was obtained using a peristaltic pump to prefilter seawater through a 0.22 µm Sterivex filter cartridge (EMD Millipore) then collection of 0.5 to 1 L of filtrate on a 25 mm 0.02 µm Whatman Anotop filter cartridge (GE Life Sciences). DNA from the Anotop cartridge was extracted using the protocol "Extracting nucleic acids from viruses on a filter" in ref. [35].

(iv) A set of 18 samples from North-American freshwater lakes (Lake Erie, Lake Michigan, and Lake Superior) from which cyanobacteria were selectively sorted using fluorescence activated single-cell sorting flow cytometry ("Freshwater cyanobacteria" in Table S1). For each sample, approximately 100,000 cells were sorted, and DNA was extracted using prepGEM (ZyGEM; Hamilton, New Zealand) on the cells pellet after 1h centrifugation at 7,200g and subsequent removal of supernatant.

(v) A set of 34 samples from Lake Mendota surface water, for which mini-metagenomes were generated by sorting individual gates using fluorescence activated single-cell sorting flow cytometry

("Mendota communities"). Briefly, subsets of the total microbial cells were defined based on a combination of fluorescence, forward scatter, and size scatter, to generate mini-metagenomes from 75,000 to 150,000 "similar" cells. DNA from these different cell pools was extracted using prepGEM (ZyGEM; Hamilton, New Zealand) on the cells pellet after 1h centrifugation at 7,200g and subsequent removal of supernatant..

(vi) A set of 20 samples from desert soil microbial communities, from which mini-metagenomes were generated following incubation with a bio-orthogonal non-canonical amino acid (BONCAT, "Soil BONCAT", [21,36]). These samples were then sorted via fluorescence activated single-cell flow cytometry to separate active from inactive microbial cells. DNA was extracted from 100,000 sorted cells using prepGEM (ZyGEM; Hamilton, New Zealand) on the cells pellet after 1h centrifugation at 7200g and subsequent removal of supernatant.

Library construction and sequencing

Three library preparation methods were used here, including TruSeq DNA PCR-Free DNA Sample Preparation Kit (Illumina, San Diego, CA, USA), Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA, USA), and Accel-NGS 1S Plus (Swift BioSciences, Ann Arbor, MI, USA). The only samples which contained enough DNA to create a TruSeq DNA PCR-Free library were some samples from the "Delaware Bay viruses" project, for which both Nextera XT and 1S Plus libraries were also created (Table S1). For the two other virus projects ("Permafrost-associated viruses" and "SPOT viruses"), both Nextera XT and 1S Plus libraries were created. Finally, Nextera XT libraries were created for all other projects ("Freshwater cyanobacteria", "Mendota communities", "Soil BONCAT", Table S1). All libraries were prepared according to manufacturer's instructions, and included as many PCR cycles as necessary to obtain 200 pM of DNA for sequencing, with a maximum of 20 cycles for viral metagenomes and 25 cycles for targeted microbial metagenomes. Finally, viral metagenomes were sequenced on either Illumina HiSeq-2500 or Illumina HiSeq-2000, and targeted microbial metagenomes with Illumina NextSeq HO, all with 2x151 reads (Table S1).

Reads contamination filtering and trimming

For all libraries, BBDuk adapter trimming (bbduk.sh https://sourceforge.net/projects/bbmap/ v35.79, parameters: ktrim=r, minlen=40, minlenfraction=0.6, mink=11, tbo, tpe, k=23, hdist=1, hdist2=1, ftm=5) was used to remove known Illumina adapters. The reads were then processed using BBDuk quality filtering and trimming (parameters: maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=27, hdist=1, trimq=12, qtrim=rl). At this stage reads ends were trimmed where quality values were less than 12, and read pairs containing more than three 'N', or with quality scores (before trimming) averaging less than 3 over the read, or length under 51bp after trimming, as well as reads matching Illumina artifact, spike-ins or phiX were discarded. Remaining reads were mapped to a masked version of human HG19 with BBMap (bbmap.sh v35.79, parameters: fast local minratio=0.84 maxindel=6 tipsearch=4 bw=18 bwr=0.18 usemodulo printunmappedcount idtag minhits=1), discarding all hits over 93% identity. Finally, for all Accel NGS 1S Plus libraries, the first 10 bases of forward and reverse reads were discarded to avoid contamination by the low complexity adaptase tail, per manufacturer's instruction.

Comparison of different assembly pipelines

The different assembly pipelines tested here included combinations of two types of read correction, two types of read selection or no read selection, and two types of assemblies (Table S3). The two methods used for read correction were chosen to represent either a "strict" or "relaxed" read correction. The "strict" correction used bfc (v. r181 [31]) to remove reads with unique kmers (parameters: "-1 -s 10g -k 21"), followed by seqtk (v. 1.2-r101-dirty https://github.com/lh3/seqtk) to remove reads for which paired sequences was removed by bfc (parameters: "dropse"). The "relaxed" read correction aimed at keeping as many reads as possible, and used tadpole.sh (v. 37.76 https://jgi.doe.gov/data-and-tools/bbtools/) to correct sequencing errors by leveraging kmer frequency along each read (parameters "mode=correct ecc=t prefilter=2").

An additional read selection step was tested to check whether removing some of the reads associated with regions of high coverage could help *de novo* genome assembly. The two approaches evaluated here included read normalization with bbnorm.sh (v. 37.76 https://jgi.doe.gov/data-and-tools/bbtools/) in which the kmer-based read depth is leveraged to identify high-depth reads and normalized these to a defined depth (here 100x, parameters: "bits=32 min=2 target=100"), as well as a deduplication approach with clumpify.sh (v37.76, https://jgi.doe.gov/data-and-tools/bbtools/), in which identical reads are identified and only one copy retained (parameters: "dedupe subs=0 passes=2"). These parameters identify reads as duplicated only if they are an exact match (i.e. no substitution allowed). The ratio of duplicated reads was calculated by comparing the number of reads after deduplication to the number of input reads for each library (Table S1).

Finally, two different modes of the SPAdes assembler (v. 3.11 [29,30]) were tested to assess whether this could also influence assembly. Specifically, the two modes tested were metaSPAdes (option "--meta") and single-cellSPAdes (option "--sc"). In both cases, SPAdes was run with the error correction step skipped ("--only-assembler") and a fixed set of kmers ("-k 21,33,55,77,99,127").

Assemblies were evaluated using a standard set of metrics computed with stats.sh from the bbtools suite (https://jgi.doe.gov/data-and-tools/bbtools/) and a custom perl script. These included cumulative size of all contigs, cumulative size of all contigs ≥ 10kb, total number of contigs, minimal contig length among contigs making up to 50% of assembly size (N50), minimal contig length among contigs making up to 90% of assembly size (N90), and size of the largest contig (Table S3). Kolmogorov–Smirnov test (from the R package stats [37]) and Cohen's effect size (as implemented in the R package effsize [38]) were used to compare distributions of cumulative size of all contigs ≥ 10kb between different pipelines.

Assembly errors were estimated for the 25 libraries for which an unamplified library was available (Table S2) using QUAST [32]. All contigs ≥ 1kb were included in this analysis, with contigs assembled from the corresponding unamplified library with a standard metagenome assembly pipeline ("strict" read correction, no read selection, and metaSPAdes assembly) used as a reference genome. QUAST was run with the "--fast" option enabled, all other parameters left to default. QUAST provides counts for three types of misassemblies: "relocation" in which two contiguous sections from a newly assembled contig map to the same reference sequence but non-contiguously, "inversion" in which two contiguous sections from a newly assembled contig map to the same reference sequence with one

fragment being reversed, and "translocation" in which two contiguous sections from a newly assembled contig map to different contigs in the reference assembly. Because the assembly from unamplified libraries are not true reference genomes, i.e. each contig is not an independent chromosome, we ignored the misassemblies identified as "translocation", as these could represent cases where both assemblies are correct and produced distinct but overlapping contigs. Instead, the estimated rate of misassemblies was calculated for each assembly as the sum of the number of "relocations" and "inversions" provided by QUAST, divided by the total length of all contigs ≥ 1kb.

Coverage bias analysis

Quality-checked reads were mapped to reference assemblies to estimate contigs coverage and assess potential coverage biases along these contigs. For libraries for which an unamplified metagenome was available (i.e. the 11 samples from the "Delaware Bay viruses" project, Table S2), contigs from a standard metagenome assembly of the unamplified library were used as reference. For every other PCR-amplified library, contigs obtained through the "best" assembly pipeline, i.e. relaxed read correction with tadpole.sh (https://jgi.doe.gov/data-and-tools/bbtools/), read deduplication with clumpify.sh (https://jgi.doe.gov/data-and-tools/bbtools/), and assembly with SPAdes single-cell (error correction turned off, k-mers of 21, 33, 55, 77, 99, 127 [30]) were used as reference. The mapping was computed using BBMap (bbmap.sh https://jgi.doe.gov/data-and-tools/bbtools/) with random assignment of ambiguously mapped reads (parameters: "mappedonly=t interleaved=t ambiguous=random").

For contig coverage comparison to unamplified libraries (Fig. S1A), individual contig coverage was normalized by the library size (i.e. total number of bp in library). For estimation of coverage bias associated with high and low depth of coverage regions along individual contigs, bam files were parsed using a custom perl script to (i) identify unique mapping events, i.e. combinations of unique mapping start coordinate and insert size, and (ii) calculate for each unique mapping the number of different reads providing this exact mapping, the corresponding GC% of the insert, and the size of the insert. This was performed on all contigs ≥ 10kb if these totaled ≥ 50kb, or on all contigs ≥ 2kb otherwise. For 3 libraries (BYXNC, BYXNG, and COHNO), no contigs ≥ 2kb were generated, and the coverage bias was thus not estimated (Table S1).

To quantify the insert size bias, high and low depth regions were first defined for each contig as follows: inserts with a read depth ≥ 70% the maximum read depth of the contig were considered as high depth, while inserts with a read depth ≤ 30% of the contig maximum read depth were considered as low depth. For each library, the distribution of insert size for each of these two types of inserts was gathered, and these were compared using the non-parametric Kolmogorov–Smirnov test (from the R package stats [37]). Cohen's effect size (as implemented in the R package effsize [38]) was also used to assess the magnitude of the difference between the means of the two distributions.

All graphical representations were generated with R [37] using the following packages: ggplot2 [39], dplyr [40], and RColorBrewer [41].

Data availability

Reads for the different metagenomes are available on https://genome.jgi.doe.gov/portal/, using the links listed in Table S1. Results from the different assembly pipelines are available for each library at http://portal.nersc.gov/dna/microbial/prokpubs/BenchmarksPCRMetagenomes/.

Results & Discussion

Coverage biases and assembly pipelines were evaluated across 169 PCR-amplified metagenomes (Table S1). These included 87 viromes, i.e. virus-particle-enriched metagenomes, and 82 targeted microbial metagenomes, i.e. generated after flow cytometry cell sorting and representing only a small subset of the microbial community. Paired PCR-amplified metagenomes generated with the two common library preparation kits (Nextera XT and 1S Plus) were available and could be directly compared for 42 samples (Table S1). In addition, unamplified (TruSeq) libraries were available for 11 samples and used as a reference "standard metagenome" for these samples (Table S2).

Insert length bias of PCR amplification leads to uneven coverage along genomes

Contrary to protocols including an amplification of the DNA pool prior to library construction (e.g. MDA, SISPA), the read composition of a PCR-amplified metagenome should accurately reflect the original community composition. This has been previously demonstrated [17], and could be verified here by observing the coverage of reference contigs (obtained from unamplified metagenomes) in PCR-amplified metagenomes. Overall, nearly all contigs assembled from unamplified metagenomes were detected in PCR-amplified datasets (>90% of contigs with ≥ 5x average coverage depth, Table S2), and there was a strong correlation between unamplified and PCR-amplified coverage for shared contigs (average Pearson correlation r^2=0.77 for Nextera XT and 1S Plus library methods, Table S2, Fig. S1).

In contrast, PCR-amplified metagenomes displayed a relatively high percentage of duplicated reads compared to unamplified datasets (~ 25-85%, Fig. S1B), which contribute to an uneven depth of coverage along individual contigs (Fig. 1A). This unevenness can be measured through the coefficient of variation of coverage depth (standard deviation divided by average coverage, for each contig) which was relatively low for unamplified metagenomes (34% on average) but higher in all PCR-amplified libraries (58% average, 20-357% range, Table S1). Regions with high depth of coverage were not linked to any systematic GC bias but were enriched for short inserts (Fig. S2). As for the ratio of duplicated reads, the difference in insert size between high and low depth regions tended to increase with the number of PCR cycles performed (Fig. 1B). This suggests that some of the uneven coverage along genomes is due to over-amplification of short inserts, which make up a larger proportion of the read pool with each additional PCR cycle.

De novo genome assembly can be improved using tailored read curation and assembly pipeline

Uneven coverage can hamper assembly because standard metagenome assembly pipelines expect a uniform coverage along each genome, and leverage this signal to solve repeats and ambiguities [29]. We thus looked at three data processing steps that could be customized for PCR-amplified libraries. First, standard metagenome assemblies typically use a strict read correction and remove reads with low depth which are potentially erroneous [31]. Even if these low-depth reads are correct, they represent

low abundance sequences that would likely not assemble well anyway, and removing them reduces the time and resources (CPU and memory) required for the assembly. In the case of PCR-amplified libraries however, these low-depth reads might be important to retain, in order to correctly assemble even high-depth contigs (Fig. 1A). Second, read selection tools have been developed to either remove duplicated reads, or computationally normalize libraries, i.e. cap at a defined maximum depth. These tools have been primarily designed for MDA datasets, the majority of which deriving from single cell amplification, however these could be helpful as well for PCR-amplified metagenomes. Finally, some assemblers offer customized options for metagenomes and for single-cell (MDA) libraries, and we tested whether single-cell options might perform better on these PCR-amplified metagenomes.

Over the 12 combinations tested, a pipeline including "relaxed" read correction, read deduplication, and single-cell assembly parameters provided the largest assemblies, although the level of improvement varied (Fig. 2A, Table S3). While the cumulative length of contigs ≥ 1kb only moderately increased compared to a standard assembly (median: 1.17x, Fig. S3A), the cumulative length of contigs ≥ 10kb showed a much larger improvement (median: 3.6x, range: 0.95–3,806x, ks-test p-value: 1e-07, cohen's effect size: 0.66, Fig. S3B). Since large contigs tend to be more relevant for downstream applications, such as genome binning and annotation, systemically applying this alternative assembly strategy on PCR-amplified metagenomes maximizes the information recovered from these datasets. Overall, when considering contigs ≥ 10kb, the alternative strategy provided the largest assembly for 130 samples, and was within 80% of the largest assembly for another 17 samples (Fig. S3C), suggesting it would be a suitable default choice for any PCR-amplified metagenome.

The level of assembly improvement observed was in part linked to the number of PCR cycles performed for each metagenome (Fig. 2B, Table S3). Specifically, samples that required 9 to 12 PCR cycles typically assembled well with the standard metagenome pipeline, with 8Mb in contigs ≥ 10kb on average, which was improved with the alternative assembly to an average of 26Mb (cohen's effect size: 0.68). Samples that required 14 to 18 PCR cycles were improved further as standard assemblies yielded an average of 2Mb in contigs ≥ 10kb per metagenome as compared to 15Mb from alternative assemblies (cohen's effect size: 0.9). Lastly, the assembly of samples that required 20 to 25 PCR cycles remained limited with either approach, though still slightly improved from 562kb to 2Mb in contigs ≥ 10kb for the standard versus alternative approaches (cohen's effect size: 0.68).

Finally, we analyzed the samples for which both unamplified and PCR-amplified metagenomes were available to evaluate the error rate in assemblies obtained from the alternative strategy (Table S1). Specifically, we used QUAST [32] to identify "relocation", i.e. cases in which contiguous regions of a newly assembled contig are non-contiguous in the reference assembly, and "inversion", i.e. cases in which the orientation of contiguous regions differs between the new assembly and reference contigs. This suggested that the alternative assembly strategy generated more erroneous contigs than a standard assembly pipeline (cohen's effect size: 0.7, Fig. 2C, Fig. S3D). For these metagenomes, the amount of additional errors (median: 2x) remains much lower than the additional number and size of long contigs (median: 24x, Table S3), so the alternative assembly strategy still seems relevant for most applications, yet this higher error rate must be considered when analyzing these datasets.

Conclusions

The ability to prepare and sequence libraries from samples containing nanograms or less of DNA has been a tremendous advance for the fields of metagenomics and microbial ecology, and many biological insights have already been derived from these data. Here we highlight how a PCR amplification bias for shorter inserts can hamper standard *de novo* genome assembly for viral and microbial low-input metagenomes, and propose an alternative assembly strategy able to reduce its impact. This will aid scientists in maximizing genomic context from low input metagenomes, and should help improve understanding of challenging ecosystems and targeted subsets of microbial and viral communities.

Acknowledgments

We thank Barbara J Campbell, Mengqi Sun, Maureen L Coleman, and Katherine D McMahon for providing samples and accompanying data used in this work.

Figure & Supplementary Material legends

Figure 1. Coverage bias within individual contigs for unamplified and PCR-amplified libraries. A. Example of coverage bias along a single contig from sample 1064195 (contig 1064195_contig_573). Reads from libraries ASXXB, BWNCO, and BWWYG (Table S2) were mapped to the same contig, and read depth along sliding windows of 100bp is displayed for each library on the y-axis. Windows on the edges of the contig (within 200bp of the 5' or 3'end) were excluded as read depth is not as reliable in these end regions. B. Illustration of the insert size bias associated with high depth of coverage regions in PCR-amplified libraries. For each library, the number of PCR cycles performed for the library is indicated on the x-axis, while the Kolmogorov–Smirnov distance between the insert size distribution of low- versus high-depth regions is indicated on the y-axis. The magnitude of the difference between the means of the two distributions was also estimated using Cohen's effect size (d) and is indicated by the dot color. For clarity, only libraries for which the mean insert size was lower in high depth regions are included in the plot, and the 22 libraries which showed the opposite trend are not plotted (Table S1). KS: Kolmogorov–Smirnov

Figure 2. Optimized pipeline for assembly of PCR-amplified metagenomes. A. Distribution of the cumulative size of long (≥ 10kb) contigs (y-axis) obtained across all PCR-amplified libraries from different assembly pipelines (x-axis). Assembly pipelines are indicated along the x-axis (see Table S3). B. Cumulative size of long (≥ 10kb) contigs obtained with a standard (green) or optimized (purple) assembly pipeline for different ranges of library PCR amplifications (x-axis). Coloring of the assembly pipelines is identical as in panel A. C. Estimated error rate (y-axis) from different assembly pipelines (x-axis) across all PCR-amplified libraries. These assembly errors were estimated for the 25 libraries for which an unamplified reference assembly was available (Table S2). Coloring of the assembly pipelines is identical as in panels A and B. Dedup.: Deduplication, Meta: metaSPAdes, SC: single-cell SPAdes.

Table S1. Description of samples and libraries analyzed. The first tab lists information about individual samples including the list of all libraries generated for each sample, and the second tab includes information about each library.

Table S2. Samples including both unamplified and PCR-amplified libraries. List of the 25 PCR-amplified for which an unamplified dataset was available, alongside specific metrics that could be calculated using the unamplified dataset as reference, i.e. correlation of average depth of coverage of contigs, and percentage of contigs from the unamplified assembly detected in the PCR-amplified library. A contig was considered as detected if ≥ 1 read(s) from the PCR-amplified library mapped to it.

Table S3. Results from the different assembly pipelines tested. The first tab lists the different steps and tools tested. The second tab includes the results of *de novo* genome assembly with the different pipelines for each PCR-amplified library. For the 25 PCR-amplified libraries for which an unamplified reference was available, this second tab also includes estimates of assembly errors for each assembly pipeline obtained with QUAST.

Figure S1. PCR-amplified metagenomes are quantitative but include a significant amount of duplicated reads. A. Comparison of depth of coverage between unamplified (TruSeq, x-axis) and PCR-amplified (Nextera XT or Accel-NGS 1S Plus, y-axis) libraries. The average depth of coverage was computed for each contig as the average read depth normalized by the total size of the library. The 1:1 equivalence is indicated with a black line, while a linear best fit is shown in blue. For clarity, only 1,000 contigs randomly selected from each sample are plotted. Contigs with no reads mapped in the PCR-amplified library were not included. To be able to directly compare the two plots, only samples for which both a Nextera XT and 1S Plus libraries were available are included (Table S1). The subpanels show the correlation coefficient (Pearson and Spearman) of a sample-by-sample correlation between depth of coverage in unamplified and PCR-amplified libraries, either for all contigs or only for contigs ≥ 10kb with a depth of coverage ≥ 10x. B. Percentage of duplicated reads (y-axis) as a function of the number of PCR cycles performed during library creation (x-axis). Underlying data are availabe in Table S1.

Figure S2. Insert size and GC content distribution for all vs high-depth regions. A & B. Distribution of insert size for all regions (green) or only regions with high depth of coverage (orange) across PCR-amplified libraries. In panel A, all insert sizes were centered around 500bp to enable a more direct comparison between libraries. Panel B shows the same data without this transformation (i.e. raw insert size). C & D. Distribution of GC % for all regions (green) or only regions with high depth of coverage (orange). For panel C, each library GC% was centered around 50%, while panel D shows the same data without this transformation.

Figure S3. Assembly size and estimated error rates for different assembly pipelines. Comparison of the output of different assembly pipelines applied to PCR-amplified libraries. Panels A & B show the cumulative length of all contigs (A) or contigs ≥ 10kb (B) across assembly pipelines (x-axis). Panel C

displays the cumulative length of contigs ≥ 10kb relative to the largest value for each library, i.e. as a percentage of the "best" assembly for this library ("best" being defined as the largest cumulative length of contigs ≥ 10kb). Panel D displays the distribution of estimated error rates across the different assembly pipelines, for the 25 libraries for which error rates could be estimated (Table S2 & S3). Norm.: Normalization, Dedup.: Deduplication, Meta: metaSPAdes, SC: single-cell SPAdes.

References

1. Falkowski PG, Fenchel T, Delong EF. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science. 2008;320:1034–9. doi:10.1126/science.1153213

2. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. 2007;5:801–12. doi: 10.1038/nrmicro1750

3. Schloss PD, Girard RA, Martin T, Edwards J, Thrash JC. Status of the archaeal and bacterial census: An update. MBio. 2016;7:1–10. doi:10.1128/mBio.00201-16

4. Lloyd KG, Steen AD, Ladau J, Yin J, Crosby L. Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. mSystems. 2018;3:e00055-18. doi:10.1128/mSystems.00055-18

5. Raes J, Bork P. Molecular eco-systems biology: towards an understanding of community function. Nat Rev Microbiol. 2008;6:693–9. doi:10.1038/nrmicro1935

6. Edwards RA, Rohwer F. Viral metagenomics. Nat Rev Microbiol. 2005;3:504–10. doi: 10.1038/nrmicro1163.

7. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493:45–50. doi:10.1038/nature11711

8. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science. 2015;348:1261359. doi: 10.1126/science.1261359

9. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004;428:37–43. doi:10.1038/nature02340

10. Parks DH, Rinke C, Chuvochina M, Chaumeil P, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1. doi:10.1038/s41564-017-0012-7

11. Anantharaman K, Breier JA, Dick GJ. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. ISME J. 2016;10:225–39. doi:10.1038/ismej.2015.81

12. Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, et al. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat Commun. 2016;7:10613. doi:10.1038/ncomms10613

13. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048. doi:10.1038/nmicrobiol.2016.48

14. Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5:4498. doi:10.1038/ncomms5498

15. Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, et al. Ecogenomics and potential biogeochemical impacts of uncultivated globally abundant ocean viruses. Nature. 2016;537:689–93. doi:10.1101/053090

16. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015; doi:10.1038/nature14447

17. Rinke C, Low S, Woodcroft BJ, Raina J-B, Skarshewski A, Le XH, et al. Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. PeerJ. 2016;4:e2486. doi:10.7717/peerj.2486

18. Knowlton C, Veerapaneni R, D'Elia T, Rogers S. Microbial Analyses of Ancient Ice Core Sections from Greenland and Antarctica. Biology (Basel). 2013;2:206–32. doi:10.3390/biology2010206

19. Weinmaier T, Probst AJ, La Duc MT, Ciobanu D, Cheng JF, Ivanova N, et al. A viability-linked metagenomic analysis of cleanroom environments: eukarya, prokaryotes, and viruses. Microbiome. 2015;3:62. doi:10.1186/s40168-015-0129-y

20. Duhaime MB, Deng L, Poulos BT, Sullivan MB. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. Environ Microbiol. 2012;14:2526–37. doi:10.1111/j.1462-2920.2012.02791.x

21. Hatzenpichler R, Connon SA, Goudeau D, Malmstrom RR, Woyke T, Orphan VJ. Visualizing in situ translational activity for identifying and sorting slow-growing archaeal−bacterial consortia. Proc Natl Acad Sci. 2016;113:E4069–78. doi:10.1073/pnas.1603757113

22. Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, Matsunaga T. Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using φ29 polymerase. Environ Microbiol. 2006;8:1155–63. doi:10.1111/j.1462-2920.2006.01005.x

487 23. Reyes GR, Kim JP. Sequence-independent, single-primer amplification (SISPA) of complex DNA
488 populations. Mol Cell Probes. 1991;473:473–81.

489 24. Marine R, McCarren C, Vorrasane V, Nasko D, Crowgey E, Polson SW, et al. Caught in the middle
490 with multiple displacement amplification: the myth of pooling for avoiding multiple displacement
491 amplification bias in a metagenome. Microbiome. 2014;2:1–8. doi:10.1186/2049-2618-2-3

492 25. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation
493 protocols and template quantity on the metagenomic reconstruction of a mock microbial community.
494 BMC Genomics. BMC Genomics; 2015;16:1–12. doi:10.1186/s12864-015-2063-6

495 26. Roux S, Solonenko NE, Dang VT, Poulos BT, Schwenck SM, Goldsmith DB, et al. Towards
496 quantitative viromics for both double-stranded and single-stranded DNA viruses. PeerJ. 2016;4:e2777.
497 doi:10.7717/peerj.2777

498 27. Duhaime MB, Sullivan MB. Ocean viruses: Rigorously evaluating the metagenomic sample-to-
499 sequence pipeline. Virology. Elsevier; 2012;434:181–6. doi:10.1016/j.virol.2012.09.036

500 28. Solonenko S A, Ignacio-Espinoza JC, Alberti A, Cruaud C, Hallam S, Konstantinidis K, et al.
501 Sequencing platform and library preparation choices impact viral metagenomes. BMC Genomics.
502 2013;14:320. doi:10.1186/1471-2164-14-320

503 29. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic
504 assembler. Genome Res. 2017;5:824–34. doi:10.1101/gr.213959.116

505 30. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling
506 single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol.
507 2013;20:714–37. doi:10.1089/cmb.2013.0084

508 31. Li H. BFC: Correcting Illumina sequencing errors. Bioinformatics. 2015;31:2885–7.
509 doi:10.1093/bioinformatics/btv290

510 32. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly
511 evaluation with QUAST-LG. Bioinformatics. 2018;34:i142–50. doi:10.1093/bioinformatics/bty266

512 33. Trubl G, Solonenko N, Chittick L, Solonenko SA, Rich VI, Sullivan MB. Optimization of viral
513 resuspension methods for carbon-rich soils along a permafrost thaw gradient. PeerJ. 2016;4:e1999.
514 doi:10.7717/peerj.1999

515 34. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, et al. A simple and efficient
516 method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep.
517 2011;3:195–202. doi:10.1111/j.1758-2229.2010.00208.x

518 35. Steward GF, Culley AI. Extraction and purification of nucleic acids from viruses. in Man Aquat
519 viral Ecol Am Soc Limnol Oceanogr Waco, TX. 2010:154--165. doi: 10.4319/mave.2010.978-0-
520 9845591-0-7.154.

521 36. Couradeau E, Sasse J, Goudeau D, Nath N, Hazen TC, Bowen BP, et al. Study of Oak Ridge soils
522 using BONCAT-FACS-Seq reveals that a large fraction of the soil microbiome is active. bioRxiv. 2018;
523 doi:10.1101/404087

524 37. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R
525 Foundation for Statistical Computing; 2018.

526 38. Torchiano M. effsize: Efficient Effect Size Computation. 2017.

527 39. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer Publishing Company; 2016.

528 40. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2018.

529 41. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2014.

Figure 1

Coverage bias within individual contigs for unamplified and PCR-amplified libraries.

A. Example of coverage bias along a single contig from sample 1064195 (contig 1064195_contig_573). Reads from libraries ASXXB, BWNCO, and BWWYG (Table S2) were mapped to the same contig, and read depth along sliding windows of 100bp is displayed for each library on the y-axis. Windows on the edges of the contig (within 200bp of the 5' or 3'end) were excluded as read depth is not as reliable in these end regions. B. Illustration of the insert size bias associated with high depth of coverage regions in PCR-amplified libraries. For each library, the number of PCR cycles performed for the library is indicated on the x-axis, while the Kolmogorov–Smirnov distance between the insert size distribution of low- versus high-depth regions is indicated on the y-axis. The magnitude of the difference between the means of the two distributions was also estimated using Cohen's effect size (d) and is indicated by the dot color. For clarity, only libraries for which the mean insert size was lower in high depth regions are included in the plot, and the 22 libraries which showed the opposite trend are not plotted (Table S1). KS: Kolmogorov–Smirnov

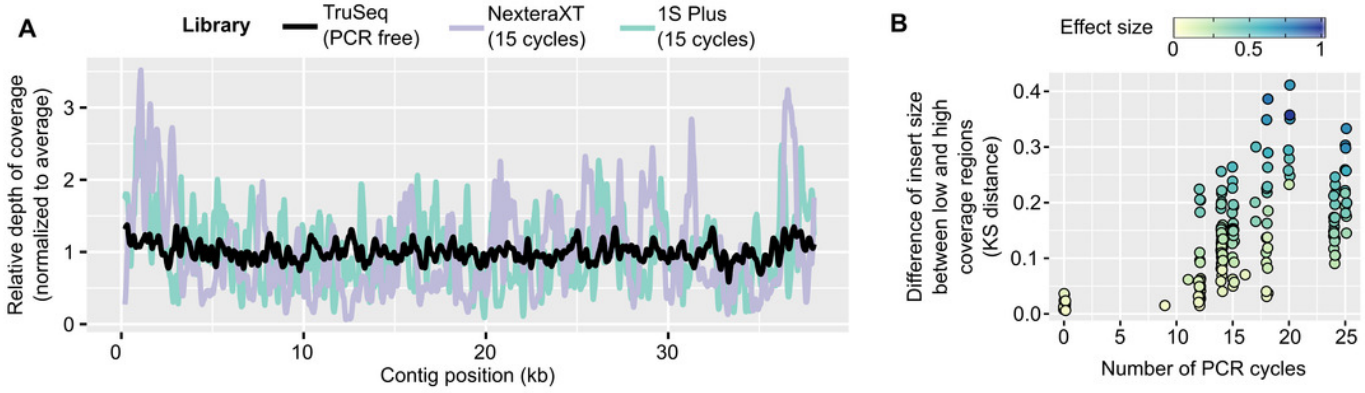

Figure 2

Optimized pipeline for assembly of PCR-amplified metagenomes.

A. Distribution of the cumulative size of long (≥ 10kb) contigs (y-axis) obtained across all PCR-amplified libraries from different assembly pipelines (x-axis). Assembly pipelines are indicated along the x-axis (see Table S3). B. Cumulative size of long (≥ 10kb) contigs obtained with a standard (green) or optimized (purple) assembly pipeline for different ranges of library PCR amplifications (x-axis). Coloring of the assembly pipelines is identical as in panel A. C. Estimated error rate (y-axis) from different assembly pipelines (x-axis) across all PCR-amplified libraries. These assembly errors were estimated for the 25 libraries for which an unamplified reference assembly was available (Table S2). Coloring of the assembly pipelines is identical as in panels A and B. Dedup.: Deduplication, Meta: metaSPAdes, SC: single-cell SPAdes.