The effects of variable sample biomass on comparative metagenomics

Meghan Chafee\textsuperscript{1,2}, Loïs Maignien\textsuperscript{1,3}, Sheri L. Simmons\textsuperscript{1*}

*Corresponding author:

Dr. Sheri Simmons
Marine Biological Laboratory
7 MBL St
Woods Hole, MA 02543
Tel. 781-366-1286
sherisim@gmail.com

Running title: Metagenomics of low biomass microbial communities

\textsuperscript{1} Bay Paul Center, Marine Biological Laboratory, Woods Hole MA 02543

\textsuperscript{2} Current address: Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

\textsuperscript{3} Current address: Laboratory of Microbiology of Extreme Environments (LM2E), UMR 6197, UBO IUEM, Plouzané, France
Summary

Longitudinal studies that integrate samples with variable biomass are essential to understand microbial community dynamics across space or time. Shotgun metagenomics is widely used to investigate these communities at the functional level, but little is known about the effects of combining low and high biomass samples on downstream analysis. We investigated the interacting effects of DNA input and library amplification by PCR on comparative metagenomic analysis using dilutions of a single complex template from an Arabidopsis-associated microbial community. We modified the Illumina Nextera kit to generate high-quality large-insert (680 bp) paired-end libraries from 50 pg to 50 ng of input DNA. Using assembly-based metagenomic analysis, we demonstrate that DNA input level has a significant impact on community structure due to overrepresentation of low-GC genomic regions following library amplification. In our system, these differences were largely superseded by variations between biological replicates, but our results advocate verifying the influence of library amplification on a case-by-case basis. Overall, this study provides recommendations for quality-filtering and de-replication prior to analysis, as well as a practical framework to address the issue of low biomass or biomass heterogeneity in longitudinal metagenomic surveys.

Keywords: library preparation / low biomass / metagenomics / phyllosphere

Introduction

Many important environmental and host-associated microbial communities contain low biomass, which presents significant challenges for standard metagenomic library preparation and analysis. Communities with low cell density
include atmospheric microbes (Bowers et al., 2011), permafrost soils (Yergeau et al., 2010), marine viruses (Duhaime et al., 2012), and the deep subsurface (Kallmeyer et al., 2012). Human-associated microbial communities on body sites such as nasal passages (Bogaert et al., 2011) and skin (Grice et al., 2009) are often low-density. Targeted sampling of these communities at small spatial scales using techniques like laser capture microdissection (Wang et al., 2010) and single-cell isolation (Rinke et al., 2013) is often necessary, but generates very low biomass.

Metagenomic techniques are critical to investigate the diversity and dynamics of these important low-biomass communities. In recent years, several commercial and laboratory techniques have become available to generate libraries from lower amounts of input DNA. Recommended inputs for commercial systems range from 1ng to 100ng, which is still greater than typical DNA yields from many low biomass samples. Both the Ovation Ultralow library system (NuGen) and the Nextera XT kit (Illumina) require 1ng input DNA. (Adey et al., 2010) describe custom modifications to a prior version of the Nextera kit (from Epicentre) that allowed library construction from 10pg.

Recent laboratory methods include multiple displacement amplification (MDA), linear amplification for deep sequencing (LADS), and linker amplification (LA). MDA, which uses phi29 DNA polymerase to generate copies of template DNA prior to library construction, is commonly used in single-cell microbial genomics (Rinke et al., 2013) but introduces significant bias in quantitative comparisons of low biomass microbial community structure (Yilmaz et al., 2010). LADS was designed to reduce PCR bias against AT- and GC-rich sequences from 3-40 ng input DNA (Hoeijmakers et al., 2011) but requires extensive sample manipulation and is not readily scalable. Linker amplification (LA), designed for next-gen sequencing of phage DNA, combines
Covaris shearing, linker ligation, size-fractionation, and subsequent PCR amplification to generate shotgun libraries from as little as 1 pg input (Duhaime et al., 2012).

Despite these technical advances, few studies have examined the effects of low input Illumina library construction protocols and variable sample to sample biomass on comparative metagenomic analyses. Recent work on 454 shotgun library generation from environmental viral communities, which often contain very low biomass, demonstrates that libraries from a single template generated using 1 pg to 10 ng input DNA are similar with respect to GC bias, read depth, and the diversity of protein clusters (Duhaime et al., 2012). (Solonenko et al., 2013) also used environmental viral concentrates to examine effects of DNA input, amplification and sequencing platform using the same set of metrics. On the Illumina platform, the largest differences occurred between viral metagenomes prepared with > 1 µg of input DNA and those prepared with 10 or 100 ng. (Parkinson et al., 2012) found that tagmented libraries generated with <1 ng input material were biased towards AT-rich sequences and showed an increasing level of read duplication at input levels of tens of pg DNA. In addition, a recent study using 16s rRNA profiling of a single diluted sample indicated that templates with less than $10^6$ cells diverged from the source community (Biesbroek et al., 2012), suggesting that similar differences may appear in shotgun metagenomes.

These studies provide insight into some issues involved in low-input sequencing, but the consequences of such biases on comparison of complex metagenomes are not well understood. Assembly-based methods, which aim to recover composite microbial “population genomes” from complex metagenomic samples, are increasingly popular in metagenomic studies (Albertsen et al., 2013; Nielsen et al.,
These reference genomes can then be used as a basis for comparing community structures through mapping of reads onto assembled contigs and subsequent analyses of taxonomic and functional diversity. Because large longitudinal or spatial studies analyzed with this method may include hundreds of samples with varying biomass levels, it is important to investigate whether sample biomass hinders the discovery of true biological variation.

A frequent approach for handling low biomass samples is to amplify sequencing libraries with additional PCR cycles. Expected biases from input level and PCR amplification interact in a synergistic fashion and can also generate separate effects, making it important to consider both factors independently. Template complexity, which is a function of both input level and the source genome, interacts with PCR amplification level to determine the frequency of duplicate reads (Kozarewa et al., 2009). Duplicate reads are defined as those sharing the exact same start and end location and are artificially generated by overamplification of a finite pool of starting molecules; natural duplicates could occur from genomic DNA shearing at the same position in separate template molecules. In complex environmental metagenomes natural duplicate frequency is expected to be quite low (Gomez-Alvarez et al., 2009), although the inherent lower complexity associated with low biomass could generate a higher frequency of both natural and artificial duplicate reads. PCR amplification also has significant effects on sequenced libraries, including significant base-composition bias (Aird et al., 2011) and artifactual molecules impacting downstream sequencing performance (NuGen, 2013). Because most studies do not optimize library amplification for each sample, it is likely that samples with different biomass will be amplified with the same number of PCR cycles and compared on a one-to-one basis.
This study aims to better understand the potential biases associated with comparative metagenomic studies from samples with heterogeneous biomass, provide a practical framework to process these samples for metagenomic sequencing, and finally examine how data quality filtering and processing can mitigate the impact of such biases on downstream data analysis. We thus systematically test the effects of template DNA concentration and PCR cycle number on community structures recovered from shotgun metagenomic data. We use a single DNA extract from an Arabidopsis-associated microbial community serially diluted over 3 orders of magnitude, tagmented with the Nextera library preparation kit, and amplified using 3 different PCR cycle numbers at each input level (Table 1). We present protocols for the preparation of size-selected Illumina shotgun libraries from limited biomass using the Illumina Nextera transposase-based method and analyze several types of artifacts generated during library prep and sequencing. Using both read- and assembly-based metagenomic methods, we examine effects of library preparation methods on comparative analyses. We also demonstrate the feasibility of metagenomic assembly, contig binning and re-mapping of reads to assembled contigs as a method for comparative analysis of metagenomic community structure in the phyllosphere.

Results

Library preparation

Careful manipulation of tagmentation enzyme concentration and alteration of PCR cycle number allowed for a 3 order of magnitude range (0.05ng to 50ng) of input DNA to be tagmented, amplified and precisely size-selected (see Supplementary Results and Fig. S1). With increasing PCR amplification, dramatically
different insert size distributions are observed following sequencing and mapping to the Arabidopsis genome (Fig. 1 and Fig. S4). Mapping to bacterial reference genomes showed the same pattern (data not shown). Differences in library size distribution are not immediately obvious on Bioanalyzer traces. However, a small population of high molecular weight (HMW) molecules >900bp is present in high cycle libraries and absent from low and medium cycle libraries (Fig. S5A). The HMW product contains molecules composed of small fragments trapped within heteroduplexes, hindering agarose gel migration during Bioanalyzer visualization and size selection. Once denatured for sequencing, these small molecules preferentially bind to and amplify on the HiSeq flow cell, resulting in a bimodal insert size distribution in high cycle samples with an abundance of small molecules (Fig. 1). An average of 14%, 23% and 67% of mapped reads were below 450 bp for low, medium and high cycle samples, respectively across the 4 input levels (Fig. S4).

**Sequencing**

In total, 197.8 million chastity-filtered reads (passing the Illumina quality filter) were obtained for all 12 samples (5.76 – 35.6 million per library) (Table 2). All raw sequence data is available via the NCBI SRA (BioProject ID PRJNA225633). High cycle samples contributed 41-66% of the total reads for each input level, despite attempts to pool in equimolar amounts, because molar calculations based on Bioanalyzer trace data excluded small molecules hidden within the HMW fraction. An average of 11% of chastity-filtered reads (stdev = 1.3%) were removed by quality filtering from each library using published criteria (Minoche et al., 2011). High cycle samples across all input levels exhibited an elevated percentage of reads passing quality filters (Table 2).
Dereplication of reads revealed a positive correlation between DNA input, cycle number and percent artificial read duplication (Fig. 2). The relationship between log(input level) and percent duplicated reads was highly significant (p=0.0001, $r^2=0.74$). Notably, three pairs of samples with different input levels in our experimental matrix were amplified for the same number of PCR cycles (Table 1). For these sample pairs, ten-fold increases in duplication level closely paralleled ten-fold decreases in input. The lowest input level (50pg) generated nearly 70% artificial duplicate reads in the low cycle treatment while medium and high cycle treatments at that input level generated ~50% duplication, substantially reducing effective sample coverage for a given sequencing effort. The size of sets of identical PCR duplicates also differed between the 50pg input libraries and all other input levels. There were up to 36 identical reads/set in the 50pg libraries, but a maximum of 11, 10, and 8 reads per set for the 0.5ng, 5ng, and 50ng libraries respectively.

Within each input level, there was an inverse correlation between the percentage of artificial duplicate reads and cycle number. This effect was the most pronounced for input levels of 50pg and 0.5ng, with 16% and 14.7% more duplicated reads in the low cycle treatment relative to high cycle, respectively. The effect was still present at 5ng (4% difference) and negligible at 50ng input (0.9%). We also observed a very slight but significant effect of cycle number on the GC content of reads across all input levels. Average read GC content decreased from 59.6% ± 11.6% at 8 cycles to 58.6%±11.6% at 24 cycles. The small size of this effect is consistent with observations by (Duhaime et al., 2012). Overall, only 0.22 ± 0.08% of reads mapped to the *A. thaliana* nuclear and plastid genomes.

*rRNA taxonomic profiling*
We analyzed the taxonomic profile of all metagenomic samples using reads identified as 16s rRNA using usearch5.2 (Edgar, 2010) and classified with the GAST method (Huse et al., 2008). We compared these profiles to results to 16s rRNA v4v6 variable region amplicon tags generated from the same sample (Maignien et al., 2014). Dominant populations in all metagenomic libraries and the amplicon sample were very similar (Fig. 3A). Differences were driven by low abundance taxa (< 1% relative abundance, Fig. 3B). Reads from low cycle libraries were assigned to significantly fewer genera than reads from medium and high cycle libraries (t-test; p-values 0.04 and 0.03 respectively). Bray-Curtis PCoA analysis of 16s reads derived from low input metagenomes indicates a slight effect of DNA input level on standard beta diversity-based comparison of community structures. Low and high cycle libraries derived from 50pg input and the high cycle library derived from 0.5ng input are separate from all others (Fig. S6). However, PERMANOVA analysis (“adonis” function in the vegan R package) using the Bray-Curtis metric on normalized 16s read count data demonstrates that there is no significant effect of input or cycle treatment on taxonomic community structure.

At the genus level, metagenomic samples were far more diverse than the amplicon sample. Across all metagenomic samples, we observed 146 different genera while only 41 of these were present in the amplicon sample. Most but not all of the non-overlapping genera are present at very low abundances in the metagenomic libraries (< 10 reads total per library). However, a subset of five genera (Arsenophonus, Xanthomonas, Acidovorax, Delftia, and Mycoplasma) absent in the amplicon sample have on average >10 reads per metagenomic library, suggesting a possible 16s rRNA universal primer bias, or low-level contamination from molecular kits (Salter et al., 2014).
Metagenomic de novo assembly and read mapping for comparative genomics

We used Metavelvet (Namiki et al., 2012) to generate a meta-assembly from 68 million pooled low and medium cycle reads across all input levels (~68 million reads). Reads from high cycle treatments were excluded from the meta-assembly due to the large fraction of reads falling out of the expected insert size range (Fig. 1). The final assembly was generated with a kmer length of 51, which produced 19,002 contigs with an N50 length of 27kb, a maximum contig length of 980kb, and a total length of 73.6 Mb. Metavelvet identified six distinct kmer coverage peaks (Figs. 4 and S8) used for targeted downstream analysis (Fig. S7). We then mapped reads from all libraries -including high-cycle ones- to quality-filtered contigs, resulting in a contig x library matrix of coverage data. This data matrix was used to assess effects of input and cycle number on apparent community structure at the genomic level. The fraction of perfectly matched reads averaged 68.4%, with high cycle samples showing a slightly lower level of mapping than medium and low cycle samples (63.8% vs 70.8%; Table 2). We did not perform further manual curation on this assembly; detailed analysis of gene content will be reported separately.

Contig classification and tetranucleotide binning

PhymmBL uses a combination of intrinsic sequence information (Phymm) and homology searches (Blast) to phylogenetically classify contigs (Brady and Salzberg, 2009). Out of 19,002 contigs ranging from 101 - 979,654bp (N50 of 27Kbp), 80% were classified to genus level with ≥70% confidence and 32% with ≥90% confidence.

To assess whether the six kmer coverage peaks identified by Metavelvet corresponded to distinct populations, we binned contigs based on their
Chafee et al. (2009) demonstrated that tetranucleotide composition using emergent self-organizing maps (Dick et al., 2009). Tetranucleotide-based binning, which is based solely on sequence composition, largely recapitulated the dominant Rhodococcus, Sphingomonas, Chryseobacter, Acinetobacter, and Pseudomonas populations identified by assembly coverage peaks (Fig. 5) with the exception of Variovorax. Contigs identified by PhymmBL as genus Variovorax had a range of kmer coverages in the assembly but formed a coherent tetranucleotide bin ("Mixed" bin in Fig. 4B; see Supplementary Results for a more detailed analysis). Overall, 85%-99% of contigs within each manually defined tetramer bin shared PhymmBL annotations at the genus level.

Differential abundance across input levels

We used the general linear modeling feature of DESeq (Anders and Huber, 2010) to identify contigs that showed a significant difference in the number of mapped reads due to input level (linear model: count ~ cycle + input compared to count ~ cycle). The number of reads mapped to meta-assembly contigs was negatively correlated with input level (Fig. 5A), with samples from higher input levels showing fewer mapped reads across contigs ranging from 5 kb to 702 kb. In addition, the larger contig coverage changes observed in contigs < 10 kb (Fig. 5A) suggests the presence of mapping artifacts due to short, low coverage and potentially mis-assembled contigs, thus providing a ground truth for the exclusion of contigs < 10 kb in comparative metagenomic mapping-based analyses. In total, we detected 62 contigs larger than 10 kb with statistically significant differential coverage due to DNA input levels. We tested if differential contig coverage across input levels was related to various parameters such as contig length, assembly kmer coverage, average contig
mapped read depth, PhymmBL contig taxonomic assignment, or GC content. There was a marginally significant relationship between the degree of fold change in relative abundance and contig length (p=0.07), but not assembly kmer coverage (p=0.3) or average mapped read depth (p=0.4). The strongest effect driving differential coverage appears to be contig GC content (Fig. 5B). Contigs showing significant differential coverage across input levels have a mean GC content of 42%, compared with a mean GC content of 61% for contigs not showing significant differences (p < 2e-16, two sample t-test with pooled variance). This effect logically correlated with a taxonomic bias toward the low-GC microorganisms, with 16 of these contigs assigned to *Chryseobacterium*, 11 to *Acinetobacter*, and 1 each to the genera *Azospirillum* and *Blautia* (Fig. 5C).

**Principal components and PERMANOVA analysis**

We normalized counts of mapped reads to library size in DESeq (Fig. S9), filtered out contigs <10kb and with kmer coverage <12, and used this normalized matrix in subsequent principal components and PERMANOVA analyses. We calculated sample-to-sample distances using the Bray-Curtis metric and used the distance matrix as input to standard principal components analysis (Fig. 6A). Additionally, we mapped reads from five phyllosphere communities sampled from different plants as part of the same original experiment onto our meta-assembly contigs (Fig. 6B).

When our serially diluted samples are considered in isolation, there are clear differences in Bray-Curtis distance between input levels (Fig. 6A). The two lowest input samples (50pg and 0.5ng) appear more similar to each other than to the higher input samples (5ng and 50ng); the 50ng sample is the most distinct. The first two
components explain a substantial portion of total variance (PC1=57% and PC2=21%). The values of both components are strongly related to input (PC1: p=0.07, PC2: p=0.01) and not to cycle level (low, medium, and high; p>0.2 for both components).

To further test the statistical significance of these differences, we performed PERMANOVA analysis on the Bray-Curtis distance matrix using the adonis function in the R vegan package (Oksanen, 2011). We tested effects of both treatments in isolation (cycle and input), in an additive linear model (cycle + input), and a model with an interaction term (cycle + input + input:cycle). We found that DNA input level has a statistically significant effect on Bray-Curtis distances over and above the effects of cycle number (p = 0.04). Cycle number did not have a significant effect on distances, either in isolation or when added to input level, and there was no significant interaction between cycle number and input level. Additionally, cycle number did not have a significant effect on Bray-Curtis distances between libraries derived from the same input level.

However, differences in community structure due to input level within our single serially diluted sample are small compared with the variance between biological replicate communities (Fig. 6B). All communities shown in Fig. 6B, including the base community for our serial dilution, were previously shown to have a highly similar community structure based on 16S rRNA pyrosequencing (Maignien et al., 2014). Bray-Curtis distances between biological replicates greatly exceed differences due to input levels and cycle number within a single sample of template DNA.

**Discussion**
Comparative studies of microbial community structure often require amplification of low biomass samples by PCR, but little is known about how the library preparation procedures necessary to generate sufficient material for sequencing affect downstream shotgun metagenomic analysis. Here, we generated paired-end Illumina libraries from a single complex community template from a plant-associated microbial community diluted over three orders of magnitude. Each input level (50pg, 0.5ng, 5ng and 50ng) was amplified with three different PCR cycle numbers (Table 1) to robustly measure effects of input and amplification on community structure. We generated a metagenomic assembly from pooled reads, which captured 74.6 Mbp of contigs from the six dominant populations in our sample (Fig. 4). We then mapped 142 million paired end reads onto this internal reference set to assess whether these libraries, derived from the same original source template, differed in the relative abundance of microbial taxa. Our results go beyond previous work evaluating library construction effects on viral metagenomes, which did not separately consider input and amplification (Duhaime et al., 2012; Solonenko et al., 2013). Additionally, to our knowledge this work provides the first evaluation of the effects of input level on comparative assembly-based metagenomic analyses.

We identified a statistically significant effect of input DNA amount on the composition of metagenomic libraries generated from a single diluted template (Fig. 6A), as measured with the Bray-Curtis (BC) distance metric. PCR cycle number within each input level did not significantly affect the distance between sequenced libraries, but had substantial effects on library insert size (Figs. 1, S4) and the level of artificial read duplication (Fig. 2). The effects of input level on the distance between libraries were primarily driven by a small, but significant decrease in read GC content with increased amplification levels. While cycle number effects within an input level
were negligible, the increase in cycle number from 8-12 for the 50 ng library to 20-24 for the 50 pg library was sufficient to introduce a bias towards amplification of AT rich sequences.

Although our assembly captured large contigs across a wide range of GC content (~35% to 75%, Fig. 5C), the set of contigs showing statistically different numbers of mapped reads with input level were significantly biased towards low GC content (Fig. 5B). We noted that fewer reads mapped onto contigs from high input samples than low input samples (Fig. 5A). This effect was small, accounting for only a ~15% increase in normalized mapped reads from the 50 ng library to the 50 pg library, but statistically significant across a wide range of contig lengths. Our conclusion is that the slight AT bias observed in the 50 pg libraries was sufficient to cause a larger number of reads to map at high identity onto AT-rich metagenomic contigs.

Because the low GC content contigs belong primarily to two distinct populations (Acinetobacter and Chryseobacterium, Fig. 5C), the outcome of this bias is a change in the estimation of population abundances based solely on an artifact of the library preparation procedure. The enrichment of AT-rich sequences with increased amplification has been previously noted (Duhaime et al., 2012; Parkinson et al., 2012), but the surprising finding here is that even very small differences in read GC content due to input DNA levels are sufficient to drive statistically significant changes in the number of reads mapping to a reference metagenome.

In the case of the phyllosphere community, this amplification bias was weak enough to be negligible compared to differences between biological replicates (Fig. 6B). Our results suggest, however, that longitudinal studies involving samples with heterogenous DNA concentration should assess the significance of this particular
bias. We therefore suggest to include serial dilutions of the highest input level libraries to verify, on a case-by-case basis, that PCR amplification-driven beta-diversity remains negligible compared to inter sample beta-diversity.

\textit{De novo assembly and mapping as a method for comparative metagenomics}

Using 68 million reads, we were able to assemble long contigs (up to 980kb) from six dominant populations (Fig. 4) yielding the largest set of assembled metagenomic data collected from the phyllosphere to date. The Metavelvet assembler (Namiki et al., 2012) identified these populations using kmer coverage information (Fig. 4). They include the genera \textit{Rhodococcus}, \textit{Sphingomonas}, \textit{Chryseobacter}, \textit{Acinetobacter}, \textit{Variovorax}, and \textit{Pseudomonas}, also identified by 16s rDNA tag sequencing (Fig. 3). With the exception of \textit{Variovorax}, each phylogenetically coherent set of contigs is also contained within a restricted kmer range (Fig. 5A). Binning of assembled contigs based on tetranucleotide composition also recovered the same six populations (Fig. 5B).

Contigs identified as \textit{Variovorax} by PhymmBL formed a coherent tetranucleotide-based grouping despite the range of assembly kmer coverages (pink dots, Fig. 4A). This suggests a partial failure of \textit{de novo} assembly, likely due to the existence of two closely related strains in roughly equal abundances (Supplementary Results). A subassembly of reads mapping to contigs identified as \textit{Variovorax} clearly shows the presence of two populations with distinct kmer coverages (~40x and ~52x), as well as a set of larger contigs with kmer coverage roughly the sum of these (~75-100x), suggesting the co-assembly of some genomic regions conserved between populations (Fig. S11).
The robust identification of dominant phyllosphere populations with \textit{de novo} assembly of reads from multiple libraries allowed us to use these “meta-assembly” contigs as a reference for comparative analysis. We mapped reads from all libraries onto these contigs for use in comparative analysis of input and cycle effects on library structure. Although the idea of “reference-independent” comparative metagenomics is not new (Dutilh et al., 2012), it is infrequently used in microbial metagenomics relative to annotation-dependent analysis of individual reads and ultimately allows for inclusion of a much larger fraction of reads in the analysis. The widely used metagenomic pipeline MG-RAST typically annotates only \(~1/3\) of reads (Delmont et al., 2012) while we were able to map \(~70\%\) of reads from each library onto meta-assembly contigs (Table 2) using stringent mapping parameters.

Additionally, we present a general method for the comparative statistical analysis of mapped metagenomic data beginning with library normalization using the \texttt{R} package \texttt{DESeq}, principally designed for use in differential expression (DE) analyses of RNAseq data. We reasoned that mapped metagenomic count data is similar to RNAseq data and thus also requires normalization to prevent inflated contributions of highly abundant contigs to the identification of differences between samples. \texttt{DESeq} is well designed for this application and has been tested and validated as a robust methods for normalization of high-throughput count data (Dillies et al., 2012). We show that \texttt{DESeq} normalization allows comparison of libraries with very different count distributions (Fig. S9) that can then be used for standard principal components analysis using a variety of distance metrics.

\textit{Effects of library preparation on metagenomic library quality}
Our results suggest several modifications in library preparation procedures to optimize assembly-based metagenomic analysis of samples from a single study with variable biomass. In particular, the strong dependence of insert size distribution (Fig. 1) and read duplication levels (Fig. 2) on input DNA amount and PCR amplification levels can significantly impact downstream analyses. Insert size distributions that deviate from expected values can negatively impact assembler performance, and artificially duplicated reads can bias mapping and count data if not corrected.

High cycle samples uniformly showed a bimodal insert size distribution—a peak at the target size (650bp) and at ~150bp despite narrow size-selection (Fig. S2). High cycle samples also consistently generated higher read counts despite normalization of library concentration prior to sequencing, as well as a small but noticeable increase in the percentage of reads passing quality control (Table 2). These effects likely both arise from artifacts generated as primer becomes limiting in late stage PCR cycles, particularly heteroduplexes of ssDNA and dsDNA that cannot be effectively size selected with agarose gel-based methods (Fig. S4). When denatured prior to sequencing, heteroduplexes dissolve to generate a hidden, highly abundant population of small molecules that preferentially bind to and amplify on the Illumina flow cell. Normalization of multiplexed samples fails because molarity calculations are based on a single insert size distribution across samples. Shifts in insert size due to overamplification are problematic if the goal is to generate large-insert libraries with a narrow size distribution, as is optimal for de novo assembly, although they are still suitable for mapping. Reconditioning PCR (Duhaime et al., 2012) can help reduce these artifacts (Fig. S3).

Read duplication was one of the most dramatic effects of input level, with a smaller contribution from cycle number (Fig. 2). In ultra-low input (50pg) libraries, we
detected a very high level of duplication (50-70%) with only 0.5-2% for high input (50ng) libraries, which is consistent with previous studies (Solonenko et al., 2013). Duplicated reads can arise from two sources: over-amplification of a limited pool of input molecules (“artificial”) or fragmentation of two identical genomic molecules in exactly the same place during library preparation (“sampling-induced”). It is desirable to informatically remove artificial duplicates prior to any count or mapping-based analysis to obtain more accurate estimates of true population sizes or allele frequencies. Artificial and sampling-induced duplicates cannot be informatically separated, although the use of both mate paired reads in the dereplication process (as performed here) can significantly improve the likelihood of removing artificial rather than sampling-induced duplicates (Zhou et al., 2014). Moreover, due to the high genomic complexity of metagenomic samples, the frequency of sampling-induced duplicates is likely to be quite low (Gomez-Alvarez et al., 2009).

Read duplication was also affected by cycle number, with high cycle samples consistently showing the lowest level of duplication within each input level (Fig. 2). Because high cycle reads had uniformly higher sequencing quality (Table 2), this is unlikely due to sequencing error. Taxonomically identified shotgun reads from the 16s rDNA gene (Figs.3, S6) also revealed that high cycle samples have significantly more unique hits at the family level than low-cycle counterparts. Increased unique reads in high cycle samples could stem from enhanced amplification of rare molecules and/or the accumulation of PCR error in more abundant molecules. However, the Kapa polymerase used in library amplification has an error rate of only 1 in 3.54x10⁶ bases, making PCR error an unlikely explanation. More likely, rare molecules that fall below the detection limit in low-cycle samples are amplified to a level above the threshold with additional cycles of PCR.
In summary, we found that the percentage of duplicated reads in a library depended on two factors: total input DNA and cycle number within an input level. We identified a significant inverse log-linear relationship between input DNA and the percentage of duplicated reads, which could be used to estimate the expected fraction of read duplication given a particular input level and cycle number. Additionally, we suggest that libraries generated from limited input material (< 1 ng) be sequenced to a higher depth than libraries generated from larger amounts of input material. This will allow for the increased sampling needed to capture an equivalent number of unique molecules from low-input libraries.

Conclusion

An original motivation for this study was to evaluate strategies for comparing metagenomic libraries derived from samples with several orders of magnitude variations in DNA concentration. We showed that despite identifiable biases, metagenomic libraries generated with different input DNA levels can in fact be statistically comparable provided the appropriate dereplication, quality control, and normalization methods are employed. An improved understanding of input and cycle effects on library structure provides a groundwork for more accurate interpretation of metagenomic data and comparison of samples with very different biomass. An important application of this study will be to minimize bias in exploration of the the earliest colonization events in microbial community assembly.

Experimental Procedures

Phyllosphere microbes were obtained from mature A. thaliana rosettes using a gentle wash method; the experiment from which the samples are drawn is described
separately (Maignien et al., 2014). Briefly, rosettes grown in soil were immersed in a solution of 0.2% Silwet in TE and bath sonicated for 10 min. Washes were filtered with a 5um filter to remove plant debris followed by 0.2um filtration to capture microbes. Genomic DNA was extracted from the filter with the Biostic Bacteremia Kit (MoBio, Carlsbad, CA).

**Optimization of low-input library preparation**

Our goal was to develop a robust scalable protocol for generation of large insert (650bp) libraries for paired-end Illumina sequencing using picogram to nanogram levels of environmental DNA. We modified the Nextera DNA Sample Preparation Kit (Illumina cat # FC-121-1030 and FC-121-1011) to generate paired-end libraries through an extensive optimization process (see Supplementary Methods).

Following optimization trials, shotgun Illumina libraries were generated using four dilutions of a single community DNA template (0.05 to 50ng; Table 1). Each tagmented dilution was amplified with three different PCR cycle numbers. The minimum (‘low’) PCR cycle number required to generate 2nM product within our target range (minimum required for sequencing), was calculated based on Bioanalyzer region analysis of a model sample. “Medium” and “high” cycle treatments added two and four cycles, respectively (see Supplementary Methods).

Read mapping to the *Arabidopsis* reference genome revealed a large population of small insert library molecules due to PCR artifacts generated by primer limitation in late cycles. A 1-2 cycle "reconditioning PCR” involves addition of fresh primer to an amplified library to break down partially hybridized species (Thompson et al., 2002) that inhibit accurate size selection (NuGen, 2013). With replenished primers, primer-adapter binding outcompetes adapter-adapter binding and restores ssDNA/dsDNA
hybrid molecules to dsDNA, which we confirmed using a test sample (see Supplementary Methods and Results).

**Bioinformatic analyses**

*Read processing and quality analysis*

We provide a schematic overview of the bioinformatic pipeline in Fig. S7. Reads were de-multiplexed and filtered using Illumina's CASAVA processing pipeline (version 1.8.2), then quality filtered and trimmed using custom Perl scripts implementing criteria described by (Minoche et al., 2011). Adapter and transposase sequence was trimmed using cutadapt (Martin 2011). We removed PCR-generated artificial duplicate reads using a custom pipeline (see Supplementary Methods). Duplication levels varied greatly between treatments (see Results). Paired reads corresponding to artificial duplicates were removed from quality-filtered fastq files; all downstream analyses included only quality-filtered, de-replicated, paired reads.

To determine the level of host plant DNA in our metagenomes, we mapped reads to A. thaliana nuclear chromosomal, chloroplast and mitochondrial gDNA with Bowtie2 (Langmead et al., 2009) (Supplementary Methods) and calculated insert size distributions with Picard Tools (http://picard.sourceforge.net/).

*Read-based analysis of 16s rRNA*

100-bp 16s rRNA reads were identified using usearch5.2 with 95% minimum query length and 80% minimum percent identity against the full-length SILVA 105 reference database (Quast et al., 2013). rRNA reads were then taxonomically classified using GAST (Huse et al., 2008). To avoid double-counting hits, only read 1 was analyzed. Additionally, we compared these 16s reads from the shotgun
metagenome with v4v6 variable region 16s rRNA tags sequenced on a 454 GS Ti from the same DNA sample (Maignien et al. 2014) classified with the same method (See Supplementary Results). Data were analyzed using the R packages vegan (Oksanen, 2011) and ggplot2 (Wickham, 2009).

**De novo assembly and mapping**

Our approach for assembly-based analysis was to generate a meta-assembly using pooled libraries and then to map reads back onto metagenomic contigs (see Supplementary Methods). Mapped read counts for each contig from individual libraries were used as input for comparative statistical analyses. Due to the small insert size of high PCR cycle libraries (see Results), 68.4 million reads from the 4 low and 4 medium cycle libraries were pooled and assembled using velvet (v. 1.2.08) (Zerbino and Birney, 2008) and Metavelvet (v.1.2.02) (Namiki et al., 2012). Weighted histograms of kmer coverage vs. length-weighted frequency were used to identify populations with discrete kmer coverage in the meta-assembly.

Reads from all 12 individual libraries were mapped to the meta-assembly with Bowtie2. Custom BioPerl scripts were used to identify reads mapping at 100% identity over 100% of read length. We constructed a count matrix based on mapping output (rows = contigs, columns = sample) and filtered out contigs < 5kb and Metavelvet kmer coverage < 12x as criteria for quality control to remove small, low coverage and potentially mis-assembled contigs.

To assess variation among biological replicates, we mapped reads from five phyllosphere metagenomes collected from the same experiment as the serially diluted sample. All samples were prepared with the same modified Nextera protocol using 3-16ng of input and shown to have similar taxonomic structures with v4v6
amplicon pyrosequencing (Maignien et al., 2014). A more detailed analysis of these and 15 additional metagenomes will be reported separately (Maignien et al., in prep).

**Contig classification and tetranucleotide binning**

PhymmBL (Brady and Salzberg, 2011) was used to assign genus-level taxonomic identity to assembled contigs. We used complete bacterial genomes in RefSeq (version 60) plus 15 draft genomes from plant-associated bacteria (Brown et al., 2012) as a training set.

We also binned contigs with kmer coverage >14x based on tetranucleotide frequency, using Emergent Self-Organizing Maps (ESOM) as previously described (Dick et al., 2009). Contigs were split into 5kb fragments prior to calculation of tetranucleotide counts, with a minimum length of 2.5kb. Vectors of tetranucleotide frequencies were input into Databionic ESOM Tools software (Ultsch and Moerchen, 2005) for clustering and visualization. Boundaries between discrete bins were manually selected through visual inspection of the ESOM. Contigs within each ESOM group were used to assess the agreement between tetranucleotide binning, PhymmBL classification and kmer populations defined by Metavelvet.

**Statistical analysis of mapping data**

We used the R package DESeq (Anders and Huber, 2010) to identify individual contigs displaying significant differential abundance (as measured by coverage) between treatments (input level and cycle number). We then used the R package vegan (Oksanen, 2011) to perform beta-diversity and permutational ANOVA analyses (see Supplementary Methods). DESeq is designed to work with count data from high-throughput sequencing experiments. As input, we used a raw count matrix...
of perfectly matched reads in every sample mapped to a subset of quality-filtered contigs from our meta-assembly. DESeq performs two distinct data transformations on raw count data. The first normalizes counts by library size, which is necessary to compare libraries with very different count profiles (Fig. S7). The second normalization step estimates the dispersion-mean relationship based on a model that accounts for counting noise and biological variation across samples, which is necessary to test for differential abundance of contigs across experimental treatments (analogous to RNA-seq differential expression (DE)). Using DESeq, we tested the fit of several general linear models to the count data and compared them to determine whether the factor in which the two differ is significant. A Benjamini-Hochberg multiple testing correction was applied to the resulting p-values.

We used size-factor corrected data from DESeq as input to principal components analysis (PCA) and PERMANOVA-based testing of treatment effects. We calculated distances between samples using the Bray-Curtis dissimilarity coefficient, which is an appropriate metric to use with semi-raw abundance data (Legendre and Legendre, 1998) and is widely used for metagenomic community structure analyses (e.g. (Fierer et al., 2012)). We used the “princomp” function in R for standard PCoA and the “ordiellipse” function in the vegan package to calculate ellipses for each input grouping showing a 95% confidence interval around the standard error of the mean.

Additionally, we used Bray-Curtis distance matrices to test effects of input level and cycle number on mapped read counts using the “adonis” function in the R vegan package (Oksanen, 2011), which uses permutation tests to test for significant effects of experimental treatments on dissimilarity values (analogous to multivariate analysis of variance). We tested for significance using a linear model with and without interaction terms with 1000 permutations.
Acknowledgements

We thank M. Sogin, J. Vineis, and A.M. Eren for critical discussions and reading of the manuscript, Z. Cardon for use of the MBL Greenhouse facility, and H. Morrison and J. Vineis for assistance with sequencing. Funding was provided by the J. Unger Vetleson Foundation to S.S. Correspondence and request for materials should be addressed to S.S. (sherisim@gmail.com).

References


Table 1

<table>
<thead>
<tr>
<th>DNA input (ng)</th>
<th>Final enzyme %&lt;sup&gt;a&lt;/sup&gt;</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
<th>18</th>
<th>20</th>
<th>22</th>
<th>24</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>4.7</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.47</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.049</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>0.006</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Standard Nextera protocol for 50ng of input uses 10% final enzyme concentration; our target size range of 800bp required less enzyme for 50ng.

Experimental design illustrating the matrix of input level and cycle number treatments used to generate metagenomic libraries. A single sample of community DNA was diluted to generate lower input libraries.
### Table 2

<table>
<thead>
<tr>
<th>Sample name</th>
<th>DNA</th>
<th>Cycle #</th>
<th>Chastity-filtered reads</th>
<th>QF, MP reads</th>
<th>% Passed QF</th>
<th>DR, MP reads</th>
<th>% mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>D73.50pg.L</td>
<td>50pg</td>
<td>L*</td>
<td>5,764,490</td>
<td>5,043,186</td>
<td>87.5%</td>
<td>1,596,770</td>
<td>64.9%</td>
</tr>
<tr>
<td>D73.50pg.M</td>
<td>50pg</td>
<td>M*</td>
<td>12,216,138</td>
<td>10,928,888</td>
<td>89.5%</td>
<td>4,979,328</td>
<td>64.5%</td>
</tr>
<tr>
<td>D73.50pg.H</td>
<td>50pg</td>
<td>H</td>
<td>35,644,994</td>
<td>32,465,120</td>
<td>91.1%</td>
<td>16,029,256</td>
<td>59.5%</td>
</tr>
<tr>
<td>D73.0.5ng.L</td>
<td>0.5ng</td>
<td>L*</td>
<td>12,514,310</td>
<td>11,095,082</td>
<td>88.7%</td>
<td>8,524,202</td>
<td>72.2%</td>
</tr>
<tr>
<td>D73.0.5ng.M</td>
<td>0.5ng</td>
<td>M*</td>
<td>17,134,074</td>
<td>15,297,762</td>
<td>89.3%</td>
<td>12,531,020</td>
<td>69.4%</td>
</tr>
<tr>
<td>D73.0.5ng.H</td>
<td>0.5ng</td>
<td>H</td>
<td>22,395,548</td>
<td>20,370,518</td>
<td>91.0%</td>
<td>18,860,686</td>
<td>66.1%</td>
</tr>
<tr>
<td>D73.5ng.L</td>
<td>5ng</td>
<td>L*</td>
<td>13,951,816</td>
<td>12,336,096</td>
<td>88.4%</td>
<td>11,531,638</td>
<td>73.8%</td>
</tr>
<tr>
<td>D73.5ng.M</td>
<td>5ng</td>
<td>M*</td>
<td>14,092,092</td>
<td>12,416,596</td>
<td>88.1%</td>
<td>11,714,346</td>
<td>73.6%</td>
</tr>
<tr>
<td>D73.5ng.H</td>
<td>5ng</td>
<td>H</td>
<td>19,477,838</td>
<td>17,507,578</td>
<td>89.9%</td>
<td>17,135,150</td>
<td>67.0%</td>
</tr>
<tr>
<td>D73.50ng.L</td>
<td>50ng</td>
<td>L*</td>
<td>9,493,114</td>
<td>8,337,300</td>
<td>87.8%</td>
<td>8,193,180</td>
<td>74.2%</td>
</tr>
<tr>
<td>D73.50ng.M</td>
<td>50ng</td>
<td>M*</td>
<td>10,882,398</td>
<td>9,489,402</td>
<td>87.2%</td>
<td>9,316,208</td>
<td>73.6%</td>
</tr>
<tr>
<td>D73.50ng.H</td>
<td>50ng</td>
<td>H</td>
<td>24,229,768</td>
<td>21,868,568</td>
<td>90.3%</td>
<td>21,706,212</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

Summary of sequencing results from all 12 libraries. Only low and medium input libraries (L,M marked with *) were used to build the meta-assembly. Chastity filter=automated Illumina quality filter. QF= custom quality filtering, see text; MP= mate-paired reads; DR= dereplicated reads, after removal of artificial duplicate reads (see text). Only mate-paired, dereplicated reads were mapped to meta-assembly contigs. Reads were filtered to retain only those matching at 100% identity over 100% of the read length.
**Figure Legends**

**Figure 1:** Insert size distributions in sequenced libraries show effects of overamplification in high cycle treatments. Paired-end reads from the 5ng input level libraries were mapped to the *Arabidopsis thaliana* Col-0 nuclear and plastid genomes using Bowtie2. The y-axis shows frequency of each insert size and the x-axis shows insert size in bp. Shaded regions shows approximate size range of insert sizes prior to sequencing, as determined by Bioanalyzer analysis. Other input levels show similar patterns (Fig. S5).

**Figure 2:** Levels of artificial read duplication are strongly dependent on input level. Identical paired-end reads resulting from PCR duplication were identified using a custom pipeline (see text). The percentage of clusters consisting of artificial duplicates is shown for each input level and cycle treatment. Cycles are labeled with 'L','M',and 'H' for low, medium, and high treatments at each input level.

**Figure 3:** 16s rRNA analysis shows consistency of dominant populations and variability of low-abundance populations across input and cycle treatments. We used usearch to identify metagenomic reads (read 1 only) mapping to 16s rDNA genes in the SILVA reference database (see text). Amplicon data is derived from pyrosequencing of the v4v6 region in the same sample from which the metagenomic libraries were drawn. Sequences from both methods were classified with GAST. (A) Stacked bar chart showing relative abundance of all taxa and color coding for the six
most abundant taxa; (B) Stacked bar chart showing the relative abundance of all taxa present at a relative abundance of less than 1%.

Figure 4: **Metavelvet de novo assembly identified six populations with distinct kmer coverage peaks.** (A) Contig length, kmer coverage, and annotation. Each point represents a single contig. The contig length (y-axis) is plotted relative to the kmer coverage of that contig calculated by the Metavelvet assembler (x-axis). Dots are colored by the genus-level PhymmBL identification of the contig. Visually distinct coverage peaks are highlighted by black rectangles. Group 1: *Rhodococcus*, Group 2: *Sphingomonas*, Group 3: *Chryseobacterium*, Group 4: *Acinetobacter*, Group 5: *Variovorax*, Group 6: *Pseudomonas*. (B) Contigs were broken into 5kb pieces and clustered based on tetranucleotide composition (see text). Each dot is colored according to its kmer coverage grouping (1-6). The map represents distance between tetranucleotide frequencies using a topographic representation; ridges shown in brown indicate large distances. Kmer peaks 1, 2, 3, 4, and 6 correspond to distinct ESOM clusters. The “Mixed” ESOM cluster contains contigs from many kmer coverage groups; 85% are *Variovorax*.

Figure 5: **Contig GC content is strongly correlated with input level effects on the number of reads mapped to contigs.** (A) Contigs identified by DESeq as showing significant differences in read mapping between input levels. The y-axis indicates the fold change between the 0.5ng, 5ng, and 50ng input levels relative to the 50pg input level as a baseline. Fold change is shown in log base 2 units. Green, 0.5 ng/50 pg; Blue, 5 ng/50 pg; Red, 50 ng/ 50 pg. (B) Boxplot showing the distribution of GC content in contigs identified by DESeq as significantly different between input levels.
(blue) and not significantly different (red). (C) Plot of GC content versus kmer coverage for all contigs > 10 kb in the metagenomic assembly. Dots are colored by the PhymmBL genus identification and scaled by the contig length. Circles are drawn to indicate the six major populations observed in the set of assembled contigs. Note that *Acinetobacter* and *Chryseobacterium* have low GC content and are disproportionately represented in the set of significant contigs shown in blue in (B).

**Figure 6:** Input level affects apparent community structure but differences are small relative to the distance between related samples. (A) Reads from all input/cycle treatments were mapped to a subset of meta-assembly contigs at 100% identity over 100% length. Resulting counts were normalized using DESeq (see text) and used to calculate Bray-Curtis library-to-library distances. Principal components analysis shows clustering of libraries by input level. Ellipses show a 95% confidence interval around the standard error of the mean. (B) Reads from five libraries constructed from distinct *Arabidopsis* phyllosphere communities (pink dots) were mapped to the same subset of meta-assembly contigs. PCoA of Bray-Curtis distances between these samples and the input/cycle libraries from a single sample shows that sample-to-sample variation greatly exceeds input/cycle effects. Labels indicate the sampling day and experimental tray (T1/T2); all input/cycle libraries (blue dots) are derived from a plant sampled on Day 73, Tray 1 (Maignien et al., 2014).