Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis

Simon A. Hardwick1,2, Wendy Y. Chen1,2, Ted Wong1, Bindu S. Kanakamedala1, Ira W. Deveson1,2, Sarah E. Ongley3,4, Nadia S. Santini5,6, Esteban Marcellin7, Martin A. Smith1,2, Lars K. Nielsen7, Catherine E. Lovelock8, Brett A. Neilan3,4 & Tim R. Mercer1,2,9

The complexity of microbial communities, combined with technical biases in next-generation sequencing, pose a challenge to metagenomic analysis. Here, we develop a set of internal DNA standards, termed "sequins" (sequencing spike-ins), that together constitute a synthetic community of artificial microbial genomes. Sequins are added to environmental DNA samples prior to library preparation, and undergo concurrent sequencing with the accompanying sample. We validate the performance of sequins by comparison to mock microbial communities, and demonstrate their use in the analysis of real metagenome samples. We show how sequins can be used to measure fold change differences in the size and structure of accompanying microbial communities, and perform quantitative normalization between samples. We further illustrate how sequins can be used to benchmark and optimize new methods, including nanopore long-read sequencing technology. We provide metagenome sequins, along with associated data sets, protocols, and an accompanying software toolkit, as reference standards to aid in metagenomic studies.
The sequencing of DNA recovered directly from environmental samples can reveal the presence of microbial communities without requiring prior laboratory cultivation. This approach, termed metagenomics, has exposed previously hidden microbial diversity in a range of different environments, from the open ocean, to complex soil samples, to the human microbiome. Accordingly, metagenomics is often used to determine the profile of the microbes that inhabit a given environment, to diagnose the presence of a microbial pathogen and to identify novel microbial lineages.

Comparisons between microbial communities that inhabit different environmental sites can also distinguish differences in the identity and abundance of microbes. These approaches can identify microbes that confer specific environmental characteristics, or measure the impact of environmental variables on microbial communities, and have been used to discover host–microbe interactions, identify novel microbes with biotechnological value, and measure environmental health.

Despite the promise of this approach, the analysis of metagenomics data remains challenging. The sheer size and complexity of microbial genomes within a sample, many of which may be novel, confounds reliable identification and makes quantification of microbes difficult. Additional technical biases that accrue during next-generation sequencing (NGS) further bias metagenomic analysis. Differences in microbial population structures can also invalidate the assumptions that underlie normalization approaches, and thereby preclude accurate detection of genuine biological differences between samples.

Reference standards can offset these analytical challenges. Reference standards enable the limits of sampling and analysis to be understood, and can measure technical variables that bias analysis with NGS. Reference standards can also evaluate quantitative accuracy, and act as scaling factors by which to normalize between samples. Accordingly, there is a pressing need to develop metagenome reference standards that can benchmark analytical methods and enable comparisons between multiple metagenome samples.

The National Institute of Standards & Technology (NIST) recently released a set of four bacterial reference genomes that can be used to benchmark and validate sequencing-based diagnostic assays. Mock microbial communities—in which multiple microbes are individually cultured and combined at known concentrations—have been used to discover host–microbe interactions, identify novel microbes with biotechnological value, and measure environmental health.

We designed a set of 86 artificial DNA sequences (sequins) that represent the range, features, and complexity of a natural microbial community, despite having an entirely artificial sequence (see Supplementary Data 1 for details of sequences). We sampled a diverse selection of finished microbial genomes (RefSeq) that encompassed a wide representation of taxa (including Eukaryota, Bacteria, and Archaea), size (0.5–10 Mb for prokaryotic genomes), GC content (20–71%), rRNA operon count (1–11), and isolation from a diverse range of environments (human body, aquatic, terrestrial, and extreme physical or chemical conditions) (Supplementary Fig. 1). A representative subsequence of each genome was selected and inverted to remove homology, while maintaining nucleotide composition, GC content, and the distribution of repetitive and unique sequences (Fig. 1a). Inverted sequences were queried against the BLAST non-redundant nucleotide collection (nt) database in order to ensure they had no significant homology (E value < 0.01) with any known natural sequences. To reduce the sequencing depth required to profile the community, we selected DNA subsequences from the microbial genomes that range in size from ~1 to 10 kb, resulting in a synthetic microbial community of total size ~227 kb.

We first simulated read libraries generated from metagenome sequins to validate synthetic sequences without the confounding impact of variables introduced through library preparation and sequencing. For comparison, we also simulated read libraries for a previously published mock community of 23 bacterial and 3 archaean species (MBARC-26)15,16. Simulated libraries were aligned using Bowtie2 to the artificial sequences and MBARC-26 genomes.

The two libraries demonstrated comparable mappability (fraction of reads concordantly mapped) and no cross alignment of reads between sequins and MBARC-26 genomes was observed (Supplementary Fig. 2a). Simulating a range of matched sequencing depths (from 0.1 to 100× coverage) also indicated comparable breadth of alignment coverage for both sequins and MBARC-26 genomes across the full range (Fig. 1b).

We also performed de novo genome assembly of the simulated libraries (using Ray Meta), showing that sequins and MBARC-26 genomes reached near-complete assembly (~295% of reference genomes assembled into contigs) at ~20x mean coverage (Fig. 1c). Below this level, the assembly of sequins and MBARC-26 genomes were comparably affected by decreasing sequence coverage. We also found no evidence of chimeric assemblies between sequins and MBARC-26 genomes. Finally, a search of the simulated sequins library against BLAST’s nt database using Centrifuge returned no significant hits. Collectively, these simulated analyses provided sequence-level validation that sequins align and assemble equivalently to a wide range of microbial genomes, without risk of cross-alignments or chimeric assembly events contaminating downstream analysis.

Synthesis and experimental validation of sequins. Sequins were synthesized, manufactured, and formulated into a staggered mixture (Mix A) that spans a ~3.2 × 10^4-fold concentration range (Supplementary Fig. 3a). We initially sequenced a neat mixture of sequins (i.e., with no natural DNA added) in triplicate to assess quantitative accuracy and technical variation between replicates. By plotting mean fold-coverage against known mixture...
concentration, we observed high quantitative accuracy ($R^2 = 0.979$; slope $= 1.04 \pm 0.02$) and minimal variation between replicates (Fig. 2a), with reproducible sequencing coverage profiles for each sequin (Fig. 2b).

We next aligned the neat sequins library to a combined index comprising sequins and MBARC-26 genomes, finding that only a negligible fraction of the reads (0.26%) aligned to an MBARC-26 genome (Supplementary Fig. 2b). The vast majority (>99%) of these cross-aligning reads aligned to the *E. coli* K-12 genome (NC_000913.3), and likely result from contamination in laboratory reagents and processes. Conversely, no experimental reads derived from a neat preparation of MBARC-26 gDNA (SRR3656745) aligned to any sequin genome (Supplementary Fig. 2b). The two libraries demonstrated comparable mappability, with concordant pair alignment rates of 89.7% and 86.9% for sequins and MBARC-26 genomes, respectively (Supplementary Fig. 2b).

To enable assessment of fold change differences between samples, we also prepared an alternative mixture of sequins (Mix B) containing the same set of 86 DNA standards, but with a subset that undergo known fold changes between mixtures ($n = 50$) and a subset that remain at equimolar concentrations ($n = 36$) (Supplementary Fig. 3b). This alternative mixture design allows the accuracy of fold changes across different samples to be assessed, while also providing negative controls for inter-sample normalization.

Validation of sequins using MBARC-26 mock community. To validate the use of metagenome sequins, we next spiked the staggered mixture into genome DNA from the MBARC-26 mock community at a low fractional abundance (1%). The combined sample then underwent concurrent library preparation and sequencing (Fig. 1a; see Methods). Of the resulting library, 1.49% of reads aligned to sequins, while 98.1% aligned to MBARC-26 genomes. We plotted the breadth of alignment coverage for each sequin and MBARC-26 genome against its measured fold-coverage, showing that they performed comparably across the full range (Fig. 1d). The rate of errors in these sequenced reads was comparable for sequins and MBARC-26 genomes, confirming the use of sequins for estimating run-specific sequencing error rates (Supplementary Fig. 2c).

Since sequins span the full range of GC contents observed in natural microbial genomes (~24–72%), we used sequins to assess the extent of any GC bias in the sequencing data. We plotted the...
measured sequencing error rate for each individual sequin against its respective GC content, finding that the mismatch error rate increased steadily with increasing GC content (Supplementary Fig. 4a, left). This effect was also observed for indel errors, albeit significantly weaker. We observed similar results for the MBARC-26 genomes (Supplementary Fig. 4a, right). We also plotted fold-coverage (observed divided by expected) for each sequin against its respective GC content, observing only very minor GC bias, with slightly elevated coverage for sequins in the middle of the GC content range (Supplementary Fig. 4b, left). Again, similar results were seen for MBARC-26 genomes (Supplementary Fig. 4b, right). While previous studies have found much stronger under-representation of GC-rich and AT-rich fragments, all libraries in the present study were prepared using a PCR-free protocol (see Methods) and thus the effect of any GC bias would be expected to be less obvious. To investigate any length bias, we plotted observed/expected coverage of each sequin against its length, finding slightly reduced coverage at both the small and large ends of the spectrum (Supplementary Fig. 4c).

We next assessed the de novo assembly of the sequins by comparison to the MBARC-26 genomes. We found that, despite their reduced size, sequins assembled similarly to MBARC-26 genomes at matched coverage levels (Fig. 1e). Both sequins and MBARC-26 genomes achieved complete assembly at ~30× coverage, a finding consistent with previous studies of bacterial genomes. Again, no chimeric contigs between sequins and MBARC-26 genomes were observed in the assembly. The comparable performance of sequins and MBARC-26 genomes demonstrated the commutability of sequins as an internal reference for addition to environmental DNA samples during metagenome sequencing and analysis.

**Benchmarking of long-read sequencing using sequins.** Long-read sequencing can resolve repetitive regions within a microbial genome, and is useful for genome assembly. Therefore, we next used the sequins to benchmark and optimize the performance of Oxford Nanopore Technology’s MinION long-read sequencing technology. We first sequenced a neat mixture of metagenome sequins, generating a total of 160,709 base-called reads, of which 74.0% successfully aligned to reference sequences using bwa-mem (Supplementary Table 1). We detected 75 (out of 86) sequins with at least one read, corresponding to a limit of detection (LoD) of 60 attomoles/µL and a dynamic range of 1.6 × 10⁴. Most sequins (61 out of 86) were fully covered by at least one MinION read, thereby obviating the need for de novo assembly (Supplementary Fig. 4b).
Fig. 5a). To assess quantitative accuracy, we plotted the measured fold-coverage of each sequin against its input concentration, noting that the quantitative accuracy of MinION sequencing ($R^2 = 0.914$) was slightly less than matched Illumina short-read data ($R^2 = 0.981$) (Supplementary Fig. 5b).

We next measured sequencing error rates, finding that MinION had a mismatch error rate of 7.12% (compared to 0.127% for Illumina) and an indel rate of 8.71% (compared to 0.0077% for Illumina). Notably, MinION sequencing suffered significantly higher indel rates in homopolymeric sequences (mean indel rate = 16.7%) compared to other regions (mean = 7.69%; unpaired t-test with Welch’s correction, p value <0.0001), with characteristic sequence coverage drops at the upstream end of homopolymer tracts (Supplementary Fig. 5c). This phenomenon has also been reported by others.

We then spiked MBARC-26 mock community DNA with metagenome sequins (5% fractional abundance) and sequenced the combined library to generate 299,050 base-called reads, with a mean length of 2590 bp (Supplementary Fig. 6a). While all MBARC-26 genomes were detected with at least one read, the average fold-coverage varied from as low as 0.0439× for Nocardiopsis dassonvillei up to 24.4× for Fervidobacterium pennisolvens (Supplementary Fig. 6b). We also detected 66 out of 86 sequins, corresponding to an LoD of 120 attomoles/µL and a dynamic range of 8.2 × 10^5. The error rates for sequins and MBARC-26 genomes were comparable, with mismatch rates of 8.30% and 7.67%, respectively, and indel rates of 6.64% and 6.49%, respectively (Supplementary Fig. 6c).

We then used the sequins to benchmark the performance of a range of different read-mapping tools designed for long, error-prone reads (bwa-mem, graphMap, marginAlign, and minimap2; Supplementary Table 2). Across the four mappers, the fraction of reads aligning to sequins ranged from 7.21 to 8.71%, with a corresponding range in base-level sensitivity (i.e., the fraction of reference bases covered by aligned reads) from 68.9 to 73.2% (Supplementary Fig. 6d). We also used sequins to assess quantitative accuracy, finding that average fold-coverage for sequins ranged from 97.3 to 111.7×, and quantitative accuracy ($R^2$) ranged from 0.889 to 0.915. This illustrates how sequins can provide a useful internal metric by which to benchmark and optimize bioinformatic analysis.

Using sequins in a real metagenome experimental context. Having validated the veracity of sequins, we next sought to demonstrate their use in a real experimental context. We spiked sequins into DNA extracted from saltmarsh samples collected from Haslam’s Creek, in Sydney Olympic Park, Australia (see Methods). We sought to compare the microbial composition of (i) non-rooted soil, (ii) rhizosphere, and (iii) root samples taken from saltmarsh sites undisurbed by human development (termed “natural”), with saltmarsh sites that were regenerated as part of the Sydney 2000 Olympics (regenerated) (Supplementary Fig. 7a). Therefore, we spiked Mixes A and B alternately into DNA extracted from three replicate natural and regenerated sites, respectively, at 5% fractional abundance prior to library preparation and sequencing.

To assess the quantitative accuracy of fold change measurements between samples, we plotted observed against expected log fold change (LFC) for each sequin, indicating a strong linear relationship ($R^2 = 0.971$; slope = 1.05 ± 0.02) (Fig. 3a). We then performed differential abundance testing on the sequins (Mix A vs. B) using DESeq2. Of the sequins that change abundance across mixtures, 48 out of 50 returned an adjusted p value <0.05 (sensitivity = 96.0%). Conversely, of the negative control sequins, 32 out of 36 returned a p value (adj) >0.05 (specificity = 88.9%). As expected, diagnostic power increased with expected LFC, as illustrated by receiver operator characteristic (ROC) analysis (Supplementary Fig. 7b).

De novo assembly of metagenome samples can be confounded by both inherent sample complexity, as well as technical factors introduced during library preparation, sequencing, or analysis. Sequins can distinguish between these outcomes and thereby assist in quality control and troubleshooting. To demonstrate this, we performed de novo assembly on all samples using Ray Meta, and evaluated the quality of each assembly using MetaQUAST (supplying the sequin reference sequences; Supplementary Table 3). The N50 values for contigs aligned to sequins (2272–2879 nt) indicated that library preparation, sequencing, and bioinformatic assembly were performed successfully (Supplementary Fig. 7c). Conversely, the N50 values for unaligned contigs were poorer and less consistent (697–3659 nt), indicating that inherent sample complexity limited assembly, rather than technical factors.

We next examined the taxonomic composition of reads using MG-RAST, by searching against the RefSeq database (phylum level). Principal component analysis (PCA) revealed that samples clustered loosely by type (non-rooted soil, rhizosphere, and root) rather than environment (natural and regenerated) (Fig. 3c, left). However, box plots of relative log expression (RLE) showed a clear need for normalization, with large distributional differences between samples that indicated unwanted variation (Fig. 3c, right). While RLE plots were first developed for gene expression data, they can additionally be used to uncover unwanted variation in many other types of high-dimensional data, and are helpful for determining whether a normalization procedure has worked as intended.

To normalize between samples, we employed the RUVg method, which performs factor analysis on suitable sets of control genes (e.g., spike-in controls) to adjust for unwanted technical variation. To perform RUVg normalization, we nominated the subset of sequins that remain at equimolar concentrations across mixtures as negative controls (n = 36). RUVg normalization improved the data, with samples still clustering loosely by type (Fig. 3d, left), and sample RLE plots now centered around zero with most of the excessive variation removed (Fig. 3d, right). While samples still did not cluster perfectly by type, it must be kept in mind that the replicates were collected from three separate natural and regenerated sites. Nonetheless, the fact that samples clustered loosely by type but not at all by environment is an interesting biological observation, and implies that regeneration of these saltmarsh sites after initial development did not have a discernible effect on the microbial ecology of the samples. In order to compare variation among replicates, we plotted the mean coefficient of variation for each of the six treatment groups both before and after RUVg normalization, finding that variation decreased substantially after RUVg normalization (Fig. 3b). This provided further evidence that RUVg normalization was successful. Notably, the use of sequins with the RUVg approach outperformed other normalization methods, including upper-quartile (UQ) normalization, where some samples still displayed excessive variability (see Supplementary Figs. 8 and 9).

Using sequins for normalization between samples. While metagenomics can profile the relative proportions of species (or other taxonomic units) within a sample, it is generally not possible to detect differences in total microbial load between samples. This limitation can be addressed by the addition of spike-in controls in fixed (rather than fractional) amounts to provide a reference by which to rescale and normalize samples.
To demonstrate this principle, we assembled two distinct mock microbial mixtures (A and B) comprising gDNA extracted from four species of Cyanobacteria (Fig. 4a), with the total amount of genome DNA doubling each time (A1, A2, and A3; and B1, B2, and B3). A fixed (as opposed to fractional) amount of sequins was added to each sample before library preparation and sequencing. Comparison of these mixtures provided examples of both relative and absolute fold change comparisons for each microbe.

We initially performed conventional normalization (based on genome size and sequencing depth). While this enabled the relative abundance of a microbe fraction to be measured within a single mixture, this approach was unable to distinguish absolute changes in abundance. Accordingly, the microbial composition of Mixes A1, A2, and A3 were indistinguishable from one another; and likewise for Mixes B1, B2, and B3 (Fig. 4b). By contrast, after normalizing each sample relative to the sequin input, this allowed the increasing absolute abundance of microbes to be detected within the mixtures (Fig. 4c).

The advantage of normalizing by absolute comparison to sequins was further demonstrated by assessing microbe fold change differences between mixtures. For example, if we compare the abundance of the *Synechocystis* sp. PCC 6803 genome between the six mixtures, we observe a total of 15-fold change differences (Fig. 4d). We are unable to accurately resolve either relative or absolute fold change differences when normalizing our sample by genome size and library depth. By contrast, the fold changes in *Synechocystis* abundance could be accurately resolved when normalized relative to absolute sequins.

To demonstrate the breadth of this advantage, we plotted the observed relative to expected fold change for each species across the six samples (total of 60 comparisons), observing a much stronger correlation following normalization with sequins ($R^2 = 0.997$), compared to normalizing only by genome size and library depth ($R^2 = 0.537$) (Fig. 4e). This analysis illustrates how sequins can be used to reveal major global shifts in microbe community abundance that are otherwise imperceptible using conventional normalization procedures.
**Discussion**

Metagenomics can profile the community of microbes within an environmental sample. This analysis of microbes does not require prior laboratory cultivation, can discover new microbial lineages, and diagnose the presence of pathogens within patient samples. Despite this promise, the analysis and comparison of metagenome samples can be affected by a range of unwanted technical variables, such as different library preparation methods or the presence of enzymatic inhibitors. A major advantage of sequins is their ability to measure and mitigate this technical variation that influences sequencing. As internal standards, sequins can be used to normalize different methods, experiments, and batches, and distinguish genuine biological differences from confounding technical variables.

In addition to enabling quality control and inter-sample normalization, reference standards are essential for the development and optimization of new sequencing technologies. In this study, we used sequins to benchmark the performance of both short-read and long-read sequencing technology, as well as a range of different software tools. Emerging long-read technologies offer several advantages for metagenomics, including more accurate resolution of repeat elements, and the ability to carry out real-time, portable genomic surveillance in the field during disease outbreaks. As new technologies continue to develop, sequins provide a constant reference by which to benchmark different tools.

Global differences in the size and complexity of microbial populations can violate the assumptions that underlie technical variation that influences sequencing. As internal standards, sequins can be used to normalize different methods, experiments, and batches, and distinguish genuine biological differences from confounding technical variables.

In addition to enabling quality control and inter-sample normalization, reference standards are essential for the development and optimization of new sequencing technologies. In this study, we used sequins to benchmark the performance of both short-read and long-read sequencing technology, as well as a range of different software tools. Emerging long-read technologies offer several advantages for metagenomics, including more accurate resolution of repeat elements, and the ability to carry out real-time, portable genomic surveillance in the field during disease outbreaks. As new technologies continue to develop, sequins provide a constant reference by which to benchmark different tools.
normalization between samples. For example, the comparison of the taxonomic subpopulations within a microbial community is limited to the relative abundance of taxa, and metagenomics cannot detect shifts in absolute abundance between samples. Here, we demonstrate how sequins can be used to calibrate and rescale different samples, thereby enabling the accurate measurement of both relative and absolute fold changes.

Metagenome studies often sample widely disparate environmental sites, and often at different times and methods, resulting in substantial batch variation between experimental samples. We propose that the routine use of reference standards, such as sequins, provides a reliable reference against which to standardize sample collection across such different sites, and can support more rigorous meta-analysis between different experiments and studies.

Within the laboratory, sequins can be used for the routine surveillance of individual samples, and to measure and improve operational performance. This is particularly important in clinical microbiology, where equipment and processes must be regularly validated, and the diagnosis of pathogens within patient samples can have serious consequences. This is particularly important in clinical surveillance of individual samples, and to measure and improve studies.

Here, we demonstrate how sequins can be used to calibrate and validate procedures. Given these benefits to metagenomic methods, we have made sequins, along with associated data sets, protocols, and an accompanying software toolkit, freely available to researchers on our website at www.sequin.xyz.

Methods

Design of artificial microbial community

To design artificial sequences, we first retrieved all high-quality, finished microbial genome sequences from ReSeq. We then ranked and systematically selected genomes according to a range of features, including taxa (including representatives from Eukaryota, Bacteria, and Archaea), size (0.5–10 Mb for prokaryotes), GC content (20–71%), rRNA operon count (1–11), and isolation from a diverse range of environments (human body, aquatic, terrestrial, and extreme physical or chemical conditions) (Supplementary Fig. 1).

This ensured that the selected genomes provided a proportional representation of these different features. To illustrate the evolutionary relationships between these microbial genomes, we generated a phylogenetic tree using iTOL (v3) with default parameters.

We then aimed to select a subspecies representative of each genome as follows. We first selected a large number (>100) of random –10 kb subsequences of each genome. Subsequences were then inverted to remove primary sequence homology, while still retaining different features of the original microbial genome, e.g., secondary structure, GC content, and the distribution of repetitive and unique sequences. Inverted sequences were then queried against the BLAST nt database for homology to known sequences and any inverted sequence with a significant match (E value < 0.01) was omitted from further analysis. From the remaining subsequences, we then selected a subspecies that had a GC content closest to the original source genome. In addition, some sequences were manually curated to generate sequins with extremely high or low GC content. Upon completion, we generated a set of 86 synthetic DNA standards (sequins), which ranged in length from ~1 to 10 kb, with a combined size of ~227 kb (see Supplementary Data 1 for full details of sequences).

Simulated read libraries

To validate that sequins perform equivalently to real microbial genomes, we compared them to a mock microbial community consisting of 26 bacterial and archaeal species (MBARC-26). We generated simulated read libraries for sequins and MBARC-26 genomes using the ART simulator (v5) (Mount-Rainer 2016–06–05) in Illumina HiSeq mode with 2×125 bp reads. The FASTA sequences inputted to ART were derived from the reported NCBI Reference Sequence Database accession numbers for MBARC-26 genomes. We simulated libraries corresponding to a range of sequencing depths, from 0.1 to 100× coverage.

Synthesis and manufacture of sequins

DNA standards were synthesized by GeneArt (Life Technologies) and cloned into pMA vectors. A list of all sequences, as well as their concentration in each mixture, is provided in Supplementary Data 1. The synthesized plasmid for each sequin was transformed in E. coli (α-select Silver Efficiency, Bioline, Australia), grown up in a 50 mL culture, purified and used for DNA sequence verification by Sanger sequencing. Sequins were excised from the plasmid backbone by restriction digest and subsequent confirmation with gel electrophoresis. Purified sequins were quantified using the BR dQubit Assay on a Qubit 2.0 Fluorometer (Life Technologies) and verified on the Agilent 2100 Bioanalyzer with an Agilent High Sensitivity DNA Kit (Agilent Technologies).

Preparation of sequin mixtures

Staggered mixtures (A and B) each consist of a pool of 86 sequins that are combined at twofold serial dilutions to encompass ~3.2 × 10^4-fold concentration range (Supplementary Fig. 3a). Each mixture has 16 staggered concentration points, with at least 5 sequins (representing a range of GC contents and lengths) per point. Individual sequins in the two alternative mixtures are present at defined molar concentration ratios, allowing for comparison of fold changes in abundances between samples (Supplementary Fig. 3b).

The 86 sequins were pooled using an epMotion 5070 ePBlue” software program to make the final mixtures.

Microbe and metagenome samples

Genomic DNA from the MBARC-26 mock community was obtained from the laboratory of T. Woyke (Joint Genome Institute, CA, USA).

Samples were collected from natural and regenerated sites along Haslam’s Creek within Sydney Olympic Park, Australia (33°50′24.04″S, 151°3′48.57″E and 33°50′24.62″S, 151°3′31.69″E, respectively) in February 2016. We collected triplicates of non-rooted soil, rhizosphere, and root material from the first 10 cm of topsoil at three natural sites (natural 1, natural 2, and natural 3) and three rootless regenerations (regenerated 1, regenerated 2, and regenerated 3). Materials were transported from the field to the laboratory within vials in a dry shipper and stored at −86°C until further analyses were performed. In brief, we term non-rooted soil as root-free sediment, rhizosphere soil samples were obtained by washing up roots with a standard phosphate-buffered saline solution (PBS) and PBS washings were also collected in PBS, root material was collected in PBS. All DNA extractions were carried out using an Mo Bio PowerSoil® DNA Isolation Kit according to the manufacturer’s instructions.

For the cyanobacteria mixtures, we used gDNA extracted from monocultures of four different species for which complete, finished genome references are available: Nostoc punctiforme ATCC 29133 (NC_010628.1), Synechococcus elongatus PCC 7942 (NC_009911.1), Synechococcus elongatus PCC 7942 (NC_007694.1), and Leptolyngbya sp. PCC 7376 (NC_019683.1). We prepared two alternative mixtures of gDNA (A and B), with each species undergoing a known fold change between mixtures (Fig. 4a). Mixture A was sequenced three times, with the amount of DNA reads per species (A1, A2, and A3) determined for Mix A (B1, B2, and B3). A fixed amount of sequins (0.666 ng) was added to each sample prior to library preparation and sequencing. To normalize samples using sequins, we calculated the fractional abundance of each species (first normalized for genome size and library depth) and then divided this by the fraction of reads, which aligned to sequins in each sample.

Preparation of Illumina DNA libraries and sequencing

Initially, metagenomic sequins (Mixes A and B) were sequenced neat (i.e., without any natural DNA admixtures). A was sequenced using a paired-end, in order to allow to calculation of the MBARC-26 for the cyanobacteria sequence, sequins were spiked into microbial gDNA at 1% fraction. For the MBARC-26 validation experiment, sequins were spiked into microbial gDNA at 1% fractional abundance. For the metagenome samples, sequins were spiked into total metagenome DNA at 5% fractional abundance. The Nextera XT Sample Preparation Kit (Illumina®) was used to prepare DNA libraries according to the manufacturer’s instructions. Prepared libraries were quantified on a Qubit Fluorometer (Life Technologies) and verified on an Agilent 2100 Bioanalyzer with an Agilent High Sensitivity DNA Kit (Agilent Technologies). Libraries were sequenced on a HiSeq 2500 instrument (Illumina®) with 2 × 125 bp reads at the Kinghorn Centre for Clinical Genomics, Sydney, Australia.

Read alignment and de novo assembly

FastQC (v0.11.5) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to confirm sequence quality and read lengths for the Illumina reads. Reads were filtered using Cutadapt (v1.8.1) with the Trim Galore wrapper (v0.4.1) in —paired mode. Simulated and experimental reads (FASTQ files) were mapped with Bowtie2 (v2.3.2) to a combined genome comprising sequin DNA sequences and microbial genomes. Average fold-covering of each genome was calculated from mapped BAM files using the BEDTools “coverage” feature (v2.25.0). To assess GC content bias, we calculated observed/expected coverage as the mean coverage of each sequin divided by its expected concentration in the mixture. This controls for the fact that sequin abundances vary over several orders of magnitude within the staggered mixtures. Breadth of alignment coverage was calculated using the BEDTools “coverage” feature. De novo genome assembly was carried out using Ray Meta (v2.3.1) with default parameters. Assemblies were evaluated using MetaQUAST (v4.5) by supplying reference genome sequences. In order to test whether reads derived from real microbial genomes align to any sequins genome, we used a previously published data set, in which MBARC-26 gDNA was sequenced neat (~174 million paired-end 150 bp reads sequenced on Illumina HiSeq 2500; NCBI Accession: SRR3656745). Illumina sequencing error rates (mismatches and indels) were calculated using the “AlignmentSummaryMetrics” tool from the Picard suite.
NATURE COMMUNICATIONS | DOI: 10.1038/s41467-018-05555-0 | www.nature.com/naturecommunications

**Taxonomic binning and differential abundance analysis.** The MG-RAST pipeline was used to examine the taxonomic composition of each saltmarsh sample, by inputting raw sequence reads directly. Default options for quality control, including removal of artificial duplicate reads, trimming, and screening for human DNA, were used. Reads were compared against the RefSeq database (domain and phylum levels) using BLAT parameters of 60% similarity, 15 bp and E value of 10⁻⁵. Receiver operator characteristic (ROC) curves were constructed by modifying the code from the ercc-dashboard R package. Samples were normalized using the RUVg procedure, using the subset of 36 sequences that remain at equilibrium concentrations between mixtures as negative controls, with k = 1 factors of unwanted variation. We used DESeq2 to test for differential abundance between samples at both the domain and phylum levels, including fold changes of sequences between Mixes A and B.

**Software package.** We have developed a software package (termed "ANAQUIN") to assist users in the analysis of data containing sequins (or other spike-in controls) The software is available for download as a Bioconductor package at https://bioconductor.org/packages/release/bioc/html/Anaquin.html.

**Data availability.** DNA-sequencing libraries are available from SRA under BioProject ID PRJNA22917. Saltmarsh metagenome samples have been uploaded to MG-RAST (project number MGPR1258). In addition, selected data sets are available online at our website, www.sequin.xyz.

Received: 21 January 2018 Accepted: 20 June 2018 Published online: 06 August 2018

**References**

1. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. Nat. Biotechnol. 35, 833–841 (2017).

2. Thomas, T., Gilbert, J. & Meyer, F. Metagenomes—a guide from sampling to data analysis. Microb. Inform. Exp. 2, 1–12 (2012).

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

4. Howe, A. C. et al. Tackling soil diversity with the assembly of large, complex metagenomes. PLoS. ONE 9, e8904–e8909 (2014).

5. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).

6. Gargis, A. S., Kalman, L. & Lubin, I. M. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. J. Clin. Microbiol. 54, 2857–2863 (2016).

7. Hug, A. J. et al. A new view of the tree of life. Nat. Microbiol. 2, 16048 (2018).

8. Nayfach, S. & Pollard, K. C. Toward accurate and quantitative comparative metagenomics. Cell 166, 1103–1116 (2016).

9. Gibbons, S. M. & Gilbert, J. A. Microbial diversity—exploration of natural ecosystems and microbiomes. Curr. Opin. Genet. Dev. 35, 66–72 (2015).

10. Stämmler, F. et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4, 28 (2016).

11. Mason, C. E., Afshinnekoo, E., Tighe, S., Wu, S. & Levy, S. International standards for genomes, transcriptomes, and metagenomes. J. Biomol. Tech. 28, 8–18 (2017).

12. Hardwick, S. A., Deveson, I. W. & Mercer, T. R. Reference standards for next-generation sequencing. Nat. Rev. Genet. 18, 473–484 (2017).

13. Goldberg, B., Sichtig, H., Geyer, C., Ledeboer, N. & Weinstein, G. M. Making the leap from research laboratory to clinic: challenges and opportunities for next-generation sequencing in infectious disease diagnostics. mBio 6, e01888–15 (2015).

14. Jumpstart Consortium Human Microbiome Project Data Generation Working. G. Evaluation of 16S rDNA-based community profiling for human microbiome. (2015).

15. Singer, E. et al. High-resolution phylogenetic microbial community profiling. ISME J. 10, 2020–2032 (2016).

16. Singer, E. et al. Next generation sequencing data of a defined microbial mock community. Sci. Data 3, 160081 (2016).

17. Römer, C. et al. Validation of picogram- and fntogram-input DNA libraries for microscale metagenomics. PeerJ 4, e2486 (2016).

18. Bowers, R. M. et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics 16, 856 (2015).

19. Deveson, I. W. et al. Representing genetic variation with synthetic DNA standards. Nat. Methods 13, 784–791 (2016).

20. Sims, D. J. et al. Plasmid-based materials as multiplex quality controls and calibrators for clinical next-generation sequencing assays. J. Mol. Diagn. 18, 336–349 (2016).

21. Hardwick, S. A. et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. Nat. Methods 13, 792–796 (2016).

22. Jiang, L. et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011).

23. Tatusova, T. et al. Update on ReSeq microbial genomes resources. Nucleic Acids Res. 43, D599–D605 (2015).

24. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

25. Bouvier, S., Raymond, F., Godziradis, L., Lavoilette, F. & Corbeil, J. Ray meta: scalable de novo metagenome assembly and profiling. Genome Biol. 13, 1–13 (2012).

26. Kim, D., Song, L., Breitbart, F. P. & Salberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. Genome Res. 26, 1721–1729 (2016).

27. Glassing, A., Dowd, S. E., Galadnikud, S., Davis, B. & Chiodini, R. I. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 8, 24 (2016).

28. Saller, S. J. et al. Reagent and laboratory contamination can critically impact metagenome-based microbiome analyses. BMC Biol. 12, 87 (2014).

29. Munro, S. A. et al. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. Nat. Commun. 5, 5125 (2014).

30. Chen, Y.-C., Liu, T., Yu, C.-H., Chiang, T.-Y. & Hwang, C.-C. Effects of GC content on the metagenomic reconstruction of a mock microbial community. ISME J. 18 (2017).

31. Gandolfo, L. C. & Speed, T. P. RLE plots: visualizing unwanted variation in sequencing datasets. ISME J. 6, 898–901 (2012).

32. Luo, C., Tsementzi, D., Kyrpides, N. C. & Konstantinidis, K. T. Individual genome assembly from complex community short-read metagenomic datasets. ISME J. 6, 286–291 (2012).

33. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 40, e72–e72 (2012).

34. Jain, M. et al. Improved data analysis for the MicroINQ metagenomics. Nat. Methods 12, 351–356 (2015).

35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

36. Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics 17, 78 (2016).

37. Meyer, F. et al. The metagenomics RAST server—a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 1–8 (2008).

38. Gondolfo, L. C. & Speed, T. P. RLE plots: visualizing unwanted variation in high dimensional data. PLoS ONE 13, e0191629 (2018).
Acknowledgements

The authors thank Dr T. Woyke (Joint Genome Institute, CA, USA) for providing MBARC-26 mock community DNA. The authors also thank the following funding sources: Australian National Health and Medical Research Council (NHMRC) Australia Fellowship 1062470 to T.R.M., APP1108254 to B.S.K., and APP1114016 to J.B. I.W.D. is supported by a Cancer Institute NSW Early Career Fellowship (2018/ECF013). T.R.M. and T.W. are supported by the Paramor Family Fellowship. S.A.H. is supported by an Australian Postgraduate Award scholarship. The authors would like to thank Dr A. Verges (UNSW Sydney, Australia) for her supervision and support for N.S.S. The contents of the published material are solely the responsibility of the administering institution, a participating institution, or individual authors, and do not reflect the views of NHMRC.

Author contributions

S.A.H. and T.R.M. designed sequins, mixtures, and experiments. W.Y.C. and B.S.K. prepared sequins mixtures and libraries for Illumina sequencing. M.A.S. performed MinION library preparation, sequencing, and base calling. S.A.H., T.W., and I.W.D. carried out bioinformatics analysis. S.E.O. and E.M. contributed materials. B.A.N., C. E.L., and I.K.N contributed samples and provided supervision. N.S.S. collected salt-marsh samples and performed DNA extractions. S.A.H. and T.R.M. wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05555-0.

Competing interests: Author T.R.M. declares the following competing interests: The Garvan Institute of Medical Research has filed a patent (PCT/AU2015/050797) covering aspects of sequencing controls. The remaining authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.