Rapid, inexpensive measurement of synthetic bacterial community composition by Sanger sequencing of amplicon mixtures

Nathan Cermak, Manoshi Sen Datta, Arolyn Conwill

PII: S2589-0042(20)30099-7
DOI: https://doi.org/10.1016/j.isci.2020.100915
Reference: ISCI 100915

To appear in: ISCIENCE

Received Date: 16 October 2019
Revised Date: 22 January 2020
Accepted Date: 10 February 2020

Please cite this article as: Cermak, N., Datta, M.S., Conwill, A., Rapid, inexpensive measurement of synthetic bacterial community composition by Sanger sequencing of amplicon mixtures, ISCIENCE (2020), doi: https://doi.org/10.1016/j.isci.2020.100915.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 The Author(s).
Rapid, inexpensive measurement of synthetic bacterial community composition by Sanger sequencing of amplicon mixtures

Nathan Cermak, Manoshi Sen Datta, Arolyn Conwill

These authors contributed equally to this work. Author order determined by coin toss.

No institutional affiliation.

Department of Biology, Technion - Israel Institute of Technology, Haifa, Israel.

Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Lead contact and corresponding author: mdatta8788@gmail.com.

Summary

Synthetic bacterial communities are powerful tools for studying microbial ecology and evolution, as they enable rapid iteration between controlled laboratory experiments and theoretical modeling. However, their utility is hampered by the lack of fast, inexpensive, and accurate methods for quantifying bacterial community composition. While next-generation amplicon sequencing can be very accurate, high costs (> $30 per sample) and turnaround times (> 1 month) limit the nature and pace of experiments. Here, we quantify amplicon composition in synthetic bacterial communities through Sanger sequencing. We PCR amplify a universal marker gene, then we sequence this amplicon mixture in a single Sanger sequencing reaction. We then fit the “mixed” electropherogram with contributions from each community member as a linear combination of time-warped single-strain electropherograms, allowing us to estimate the fractional amplicon abundance of each strain within the community. This approach can provide results within one day and costs ~$5 USD per sample.

Introduction

Model microbial communities – comprised of a small number of pre-defined, culturable taxa – are emerging as powerful tools in microbial ecology and biotechnology. Unlike wild microbial communities, whose underlying design principles are often obscured by complex environmental conditions and thousands of microbial “parts”, simple synthetic consortia can be studied precisely under controlled laboratory conditions. Through this approach, numerous studies have uncovered principles of microbial community interactions, assembly, organization, and evolution (Celiker and Gore, 2014; Friedman et al., 2017; Goldford et al., 2018; Harcombe et al., 2014; Momeni et al., 2011, 2017; Ratzke and Gore, 2018; Wolfe et al., 2014). Furthermore, simple synthetic consortia hold great promise for biotechnology (Brenner et al., 2008), including synthesis of natural products that would be difficult to achieve with a single species (Zhou et al., 2015).
Despite the importance of model microbial communities, characterizing their composition (the proportional abundances of their constituent strains) quickly and cheaply remains challenging, since most standard methods have significant drawbacks (Table 1). On one hand, counting individual cells through colony formation on agar plates or with fluorescent labeling and flow cytometry is both cost- and time-effective and provides a direct measurement of population size. However, these methods can only be applied when strains are morphologically distinct or genetically tractable. On the other hand, next-generation sequencing can provide precise abundance estimates for arbitrary microbial communities, regardless of their composition, but typically has large up-front costs and can take weeks to months to receive results. Notably, all DNA-based methods provide estimates of gene or amplicon abundances, which are distinct from cell abundances because strains differ in their gene (Větrovský and Baldrian, 2013) and genome copy number (Akerlund et al., 1995; Schæchter et al., 1958), as well as extraction (Abusleme et al., 2014; Yuan et al., 2012) and amplification efficiency (Polz and Cavanaugh, 1998).

Sanger sequencing has long been a cheap and effective method to characterize the taxonomy of bacterial strains in isolation, often by sequencing the 16S rRNA gene. This process typically begins by PCR-amplifying the 16S rRNA gene(s) from a pure bacterial culture containing a single strain. The result is a homogeneous pool of 16S rRNA amplicons (unless the strain has multiple copies of the 16S rRNA gene). Subsequently, the amplicon pool is subjected to a linear amplification process that yields DNA segments of different lengths (Sanger et al., 1977), where all segments of a given length have a fluorescent color label corresponding to the final (3′) base (Smith et al., 1986). Then, DNA segments are sorted by length via capillary electrophoresis (Swerdlow and Gesteland, 1990), and the nucleotide sequence is determined from the corresponding sequence of fluorescent colors. Data are produced in the form of an electropherogram, in which fluorescent signal is plotted as a function of electrophoretic time (roughly corresponding to sequence position). Once characterized, the 16S rRNA gene sequence is often used as a taxonomic marker for a bacterial isolate.

In multi-strain bacterial communities where each member has a distinct 16S rRNA sequence, Sanger sequencing can be extended to characterize the presence and/or fractional abundance of each community member. The full complement of 16S rRNA genes present within a multi-strain community can also be PCR-amplified (typically with degenerate universal primers) and analyzed via Sanger sequencing. This process results in a “mixed” electropherogram. Like the single-strain electropherogram, a mixed electropherogram records the fluorescent signal as a function of electrophoretic time, but it now includes contributions from each of the strains present. Two approaches to characterize multi-strain community composition from mixed electropherograms have been developed previously (described below). However, unlike the new method we propose here, both prior approaches sought to characterize community composition without any prior knowledge of which strains were present. In the first method (Kommedal et al., 2008), a novel base-calling method was developed to preserve ambiguity at positions where multiple nucleotides were present, thereby allowing the authors to enumerate every possible constituent sequence. They then compared possible sequences to a database of known 16S rRNA gene sequences. Using this method, they reliably identified the bacteria present in numerous two- and three-species mixtures, including clinical
samples (Kommedal et al., 2009, 2011; Wolff et al., 2013). However, this approach has not been used for quantification of strain abundance, and it is unclear how accurately the members of more complex communities (>3 strains) can be resolved.

In the second method (Amir and Zuk, 2011), the authors developed an algorithm to find a sparse set of strains whose combined DNA would be expected to generate the observed mixed electropherogram. To do this, they first created a database of predicted electropherograms (based on a statistical model of how gene sequences determine electropherograms) for 16S rRNA sequences of nearly 20,000 bacterial strains. They then used computationally solve for a small set of strains that could best reproduce the observed electropherogram. Applying this method to a mixture of five equally abundant strains, they detected at least eight strains, of which seven were closely related to strains in the actual mixture. However, their fractional abundance estimates were noisy, varying from 5%-15% when the actual abundances were 20% each.

Here we develop and evaluate a new and distinct method for analyzing Sanger sequencing traces from amplicon mixtures as a fast (1 day) and inexpensive (~$5/sample) method for quantifying the fractional abundance of individual strains within simple model communities. It differs from previous approaches in two main ways. First, it assumes that one knows the full set of strains that might be in the mixture and experimentally measures their individual Sanger electropherograms. For model systems consisting of cultured isolates, this requirement is easily fulfilled. Second, our method accounts for a common mode of run-to-run variability not previously accounted for, which we show is necessary for accurate compositional estimates. We benchmark this method with multiple 2-, 4-, and 7-member communities of marine bacterial isolates, achieving a root-mean-square error of roughly 1% and yielding results similar to Illumina sequencing. We also demonstrate the utility of this method by quantifying time dynamics of five model communities over two weeks. Overall, given its accuracy and broad applicability, we believe that this method will enable experiments with a wide range of simple synthetic microbial communities that were previously time- or cost-prohibitive.

We have also implemented our method in a free and open-source package for the open-source language R (R Core Team, 2017), under the name “CASEU” for Community/Compositional Analysis via Sanger Electropherogram Unmixing. We provide functions for fitting and evaluating fit quality, both via the R language/terminal and through a graphical user interface.

### Approach

Our approach is to fit mixed-strain electropherograms as linear combinations of time-warped single-strain electropherograms. For a model bacterial community in which all component strains are known, it is possible to measure its mixed electropherogram, as well as each single-strain electropherogram. We thus sought to find a function relating the two that would allow us to extract the relative proportions of individual strains in the mixed electropherogram.

A simple linear model is insufficient due to retention-time variability. Naively, it is reasonable to fit a mixed Sanger electropherogram as an abundance-weighted linear combination of single-strain electropherograms. However, this approach yields poor fits due to between-
sample and within-sample variability in the run speed – that is, the rate at which molecules migrate during electrophoresis. This phenomenon, referred to as “retention-time variability”, is a well-known confounding factor in electrophoretic methods (Eilers, 2004; Nielsen et al., 1998) including Sanger sequencing. Indeed, we observed substantial retention-time variability in our measurements: technical replicates of the same sample sequenced on different days were often temporally offset from each other (by roughly ±1 base) and were sometimes stretched or contracted relative to one another by ±0.3% (see Figs. S1 and S2).

Instead, our fitting procedure conceptually involves two components: time-warping, which accounts for retention-time variability, and fitting a linear model. First, we warp (locally shift and stretch or contract) the time axis of single-strain electropherograms (Fig. 1). Second, we estimate strain abundances by fitting the mixed electropherogram as a linear combination of time-warped single-strain electropherograms. In practice, we do these steps simultaneously, by identifying warping parameters and abundance fractions that minimize the sum-of-squares difference between the observed and model-predicted mixed electropherogram, as follows:

$$\arg\min f_1, \ldots, f_n, g_1(t), \ldots, g_n(t), x_0 \sum_{t=t_0}^{t_{\text{end}}} \sum_{c=1}^{c_{\text{end}}} \left( Y[t, c] - \left( x_0 + \sum_{i=1}^{n} f_i g_i(t), c \right) \right)^2 + \lambda \sum_{i=1}^{n} R(g_i)$$ (1)

where

- $f_i$ is the abundance of strain $i$, where $i$ ranges from 1 to $n$;
- $t$ is an index of time, ranging from $t_0$ to $t_{\text{end}}$;
- $Y[t, c]$ is a matrix of the mixed electropherogram, with one row per timepoint and one column for each of the four fluorescence channels, $c$;
- $X_i[t, c]$ is a matrix of strain $i$’s electropherogram, with one row per timepoint and one column for each of the four fluorescence channels, $c$;
- $x_0$ is a scalar accounting for constant background fluorescence;
- $g_i(t)$ is a warping function for strain $i$; and
- $R(g_i)$ is a quadratic penalty function for shifting and stretching individual electropherograms.

Because it is not possible to have negative abundances, we constrain strain abundances to be non-negative ($f_i > 0$) by using non-negative least-squares fitting (Lawson and Hanson, 1995).

A piecewise-linear time-warping function can account for retention-time variability. Within a given electropherogram, the relative run “speed” may vary substantially, such that certain sections are stretched, and others are contracted, compared to the average speed. To account for within-electropherogram variability, we use a continuous piecewise linear warping function $g_i(t)$ (see Fig. 1B), which divides the electropherogram into several segments, each of which can be locally stretched or contracted (Nielsen et al., 1998). To prevent unreasonably large stretching or shifting, our software package includes the option of a quadratic penalty for moving the end of each segment from its original location ($R(g) = \sum_j (\tau_j - b_j)^2$) where $\tau_j$ and $b_j$ are as shown in Fig. 1B). However, for our data analysis in this work, we did not use this regularization term ($\lambda = 0$). To determine the optimal number of segments, we systematically varied the number of segments and aligned technical replicates to each other. We found that using five segments enabled us to align all samples precisely to either of their two technical replicates over
a region of ~630 bases (Fig. S2). Using only a single segment yielded poor alignments between technical replicates (Fig. S2) and produced mediocre estimates of known mixture fractions (Fig. S3). Using more than five segments did not improve the alignments between technical replicates (Fig. S2), but increased computation time.
Results

To benchmark CASEU’s performance, we analyzed a series of mock bacterial communities of 2, 4, or 7 bacterial strains with known fractional abundances. We prepared these communities by PCR-amplifying the 16S rRNA gene from each single strain (here called “A” through “H”), and mixing together amplicons from different strains in known fractions. By analyzing mixtures of amplicons, rather than mixtures of cells, we could measure sequencing and algorithmic performance independent of biases due to DNA extraction efficiency, PCR efficiency, or 16S rRNA gene copy number. Using mock communities, we assessed the following metrics of algorithmic performance:

- accuracy of fractional abundance estimates, by systematically varying the abundance of community members between 1.3%-95%;
- reproducibility, by sequencing each sample three times on separate days;
- ability to differentiate closely related strains, by varying phylogenetic distance between strains; and
- ability to correctly reject the presence of “decoy strains”, which are included as potential community members in the fits but were absent in reality.

We first analyzed two-strain mixtures for which the proportion of a single strain varied from 5% to 95% (Fig. 2). Across all two-strain communities (except the mixtures of strains A and B, see below), fractional abundance estimates were accurate with an average absolute deviation between the expected fraction and the observed fraction of 0.9 percentage points (range 0.05% to 3%). Furthermore, abundance estimates were consistent across independently sequenced technical replicates; the average standard deviation of triplicate measurements was 0.59 percentage points (range 0.06% to 1.17%).

In larger communities (4- and 7-strain mixtures), fractional abundance estimates were similarly accurate, even for low-abundance community members. To test the effect of strain evenness on fractional abundance estimates, we prepared 4- or 7-strain communities whose strain abundances were distributed according to a power law ($f_i \propto i^{-\alpha}$), where we varied the value of the exponent $\alpha$. This allowed us to assemble communities of varying evenness (Fig. 3), ranging from those in which all strains were at equal abundance ($\alpha = 0$) to those in which the dominant strain was 50-fold more abundant than the least abundant strain ($\alpha = 2$). Across these communities, abundance estimates were similarly accurate compared to the two-strain communities, with root-mean-square (RMS) errors of 0.75 and 1.14 percentage points (maximum errors of 2.2 and 3.4 percentage points), respectively for the 4- and 7-strain communities. Furthermore, the magnitude of error in a strain’s abundance was nearly independent of that strain’s abundance in the community (Fig. S5A). The standard deviations we observe between triplicate results were comparable to what would be attained by counting based methods (e.g. next-generation sequencing or plate counts) with ~5000 counts (reads or colonies) per sample (Fig. S5B).

It is important to not only estimate the abundance of a strain known to be present, but also to correctly determine when a strain is absent. To test whether CASEU is susceptible to erroneously finding strains that were not present, we re-fit all our two- and four- species communities, this
time including all strains (except for B) as possible “decoy” community members. In nearly all
cases, CASEU correctly rejected the presence of strains that were not included in the community
(Fig. 4). Notably, CASEU erroneously found non-zero amounts of strain D in some samples
where only strains A and C were present. We attribute this to the similarity between
electropherograms of strains C and D (discussed below).

To differentiate strains, CASEU requires that their electropherograms are dissimilar. We
quantified similarity as the correlation between two electropherograms after aligning one to the
other. Our mock communities contained mixtures of strains with varying degrees of
electropherogram similarity, ranging from 0.98 (for strains A and B) to 0.09 (for strains A and G)
over a 630-basepair region of the 16S rRNA gene (Fig. S4). While CASEU failed to differentiate
strains A and B which have an average post-alignment correlation of 0.98 (Fig. 2), it accurately
estimated fractional abundances for all other communities (Figs. 2 and 3) containing between-
strain correlations of up to 0.82 (strains C and D; Fig. 3). However, strain D was sometimes
mistakenly found in the mixtures of strain A and C, suggesting it may sometimes be mistaken for
strain C. Therefore, we suggest that strains with correlations of ~0.8 or greater may not be
clearly resolvable with CASEU and should be analyzed with caution. In our dataset, this
corresponds to roughly within-genus distances or closer, but the relationship between CASEU
resolvability and phylogeny may depend on the specific strains of interest.

We also note that we expect electropherogram correlations to be strongly affected by indels,
because our alignment approach has insufficient flexibility to accommodate large gaps. We note
that while strains A and B differ by only SNPs, strain C possesses a 12-base deletion near the
beginning of the gene (Table S1). This likely explains the low correlation and subsequent ability
to differentiate A and C (which otherwise have only 29 SNPs in their 930 bases of high-quality
sequence), but not A and B (which have no indels and only a dozen SNPs in one gene region).

We next investigated whether we could improve our results for the mixture of strains A and B by
focusing on the region in which these two strains’ electropherograms differ. Our logic was that if
most of the electropherogram is uninformative and subject to some amplitude noise, then
removing the uninformative regions should improve the signal-to-noise ratio. We thus fit only a
small region of the electropherogram (roughly 59 bp) that included the differing bases between
strains A and B (Fig. S4). This enabled us to obtain far more accurate fractional abundance
estimates for mixtures of strains A and B (Fig. S4D). We thus suggest that CASEU users seeking
to differentiate highly similar strains restrict their analysis to the region in which their
electropherograms vary.
Evaluation on synthetic model communities

We envision CASEU as a rapid, inexpensive alternative to Illumina sequencing for characterizing the structure of simple synthetic microbial communities. To demonstrate this use case, we performed experiments with seven four-strain model communities of unknown fractional composition. We extracted DNA from each community, then amplified and sequenced each sample twice, once via Sanger sequencing (16S rRNA V1-V9 hypervariable regions) and once via next-generation Illumina sequencing (16S rRNA V4-V5 hypervariable regions) (Fig. 5). Importantly, this analysis does not compare the accuracy of the two methods, since the true fractional abundances are unknown, but rather whether their fractional abundance estimates are consistent.

We found that CASEU provided community composition estimates consistent with next-generation Illumina sequencing of 16S rRNA amplicons (Fig. 5), despite differences in library preparation procedure and sequencing technology. Across all communities, fractional abundance estimates between the two methods were highly correlated (Pearson correlation 0.88, Fig. 5H). Furthermore, in five out of seven communities, we observed strong quantitative agreement between Illumina estimates and CASEU estimates, with an RMS difference of 7.0 percentage points (Fig. 5A,B,C,E,G).

In the two model communities where CASEU and Illumina sequencing disagreed (Fig. 5D,F; RMS differences of 15 and 28 percentage points, respectively), the differences can be attributed to a single group of closely related strains. These three strains (I3M06, I2M14, and I2M19) are very closely related (electropherograms cannot be distinguished by CASEU) and in both model communities, these strains were estimated by CASEU to be at substantially lower fractions than was estimated by Illumina sequencing. While we remain uncertain as to why these strains are detected less with CASEU than Illumina, it may be a result of CASEU and Illumina relying on different primers and amplification protocols.

In our CASEU analyses performed with these four-strain model communities, we additionally observed two cases in which CASEU produced poor fits as quantified by the correlation between the observed and predicted traces (Fig. S7). In the first case, the predicted electropherogram had a correlation of only 0.63 to the observed electropherogram, compared to >0.9 for all other samples. This poor fit alerted us to a low-quality Sanger sequencing electropherogram for one strain in the community, which contained a large anomalous fluorescence spike (Fig. S7A). In the second case, CASEU yielded a correlation between predicted and observed traces of 0.45, compared to >0.95 for other samples from the same model community. This poor fit was caused by the presence of a contaminating strain, which was not included in the fit (Fig. S7B). Including the contaminating strain increased the fit correlation to 0.95. Thus, while we only rarely observed poor fits, CASEU includes a simple metric that enables users to identify and exclude problematic samples.
**Discussion**

In microbial ecology, model communities have emerged as a useful intermediate between single-species microbiology and complex natural communities. Here, we demonstrate that Sanger sequencing can be used for **rapid, inexpensive, and accurate** quantification of model community composition.

- **CASEU can provide rapid results.** Sanger sequencing requires a simple sample preparation protocol with a single PCR step, followed by outsourced Sanger sequencing. Therefore, the time to acquire results is largely limited by sequencing time, which is often less than one day. In contrast, next-generation sequencing requires a more time-consuming library preparation protocol, often with multiple PCR steps for adaptor ligation and barcoding. Furthermore, runtime for an Illumina MiSeq routinely exceeds one day (e.g., 40 hours for paired-end 150x150 sequencing), but can require weeks to months if outsourced.

- **CASEU can be inexpensive.** Sanger sequencing has a fixed cost per sample (here, $4 USD for sequencing and roughly $1 for PCR and cleanup), whereas Illumina sequencing has large upfront cost (typically more than $1,000 USD per sequencing lane), plus per sample costs for library preparation.

- **CASEU is accurate for simple model communities.** Sanger sequencing provides an accurate and reproducible means to quantify amplicon composition for model communities, achieving similar results as Illumina sequencing for model communities, and errors of less than one percentage point for mock communities.

More broadly, we believe that our Sanger sequencing demixing approach can be extended beyond the 16S gene. For example, CASEU might be used with model communities containing closely-related strains by using other marker genes (e.g. Vibrio communities that are poorly resolved by 16S but easily differentiated by hsp60 sequences (Hunt et al., 2008)), or even communities containing both fungal and bacterial members (for example, cheese rind model communities (Wolfe et al., 2014)) by amplifying both 16S and 18S or ITS sequences simultaneously through multiplexed PCR. Beyond microbes, CASEU might extended to quantify aneuploidy using marker sequences with conserved primer sites present on all chromosomes (Kinde et al., 2012). Overall, we believe CASEU provides a versatile tool to assess sequence-variant composition in multiple contexts.

**Limitations of study**

CASEU has important limitations. To determine if CASEU is appropriate for your application, we recommend considering the following factors as they pertain to your model community.

1. **Number of strains.** Here, we demonstrate that CASEU can provide accurate fractional abundance estimates for communities of 2, 4, and 7 strains. However, CASEU may be suitable for larger model communities, as we did not identify an upper bound on the number of resolvable members.

2. **Resolvability of strains.** We found that strain resolvability depends on the correlation of their electropherograms, which is distinct from their aligned sequence similarity.
Therefore, for your particular community, we recommend Sanger sequencing each individual strain and verifying that their electropherograms cannot be aligned to be highly correlated, which can be done with our R package.

3. **Low-abundance strains.** Given typical errors of 1-2%, CASEU cannot resolve community members at fractional abundances below 1%. If this dynamic range is needed, alternatives like qPCR or next-generation sequencing may be more suitable.

4. **Sources of bias.** CASEU shares the same limitations of all DNA-based approaches for quantifying community composition, including bias in DNA extraction and amplification efficiency. Importantly, next-generation sequencing, qPCR, and CASEU do not yield cell counts, but instead yield sequence abundance. While sequence abundance is expected to be roughly proportional to cell count for any given strain, this relationship may vary between strains depending on gene copy number (Větrovský and Baldrian, 2013), growth phase (Akerlund et al., 1995; Hildenbrand et al., 2011; Schaechter et al., 1958), DNA extraction efficiency (e.g., Abusleme et al., 2014; Yuan et al., 2012), and amplification efficiency (e.g., Polz and Cavanaugh, 1998).

**Data and code availability.**

We have implemented our method in a free and open-source R package called CASEU (“Community/Compositional Analysis by Sanger Electropherogram Unmixing”), available at https://bitbucket.org/DattaManoshi/caseu.

**Acknowledgments**

Thanks to Andrea Velenich and Jeff Gore for helpful discussions, and to Otto Cordero for generously sharing strains. AC acknowledges support from the Simons Collaboration on Principles of Microbial Ecosystems (PriME).

**Contributions**

MSD and NC designed the model and fitting algorithm, performed the experiments with amplicon mixtures, and analyzed all Sanger sequencing data. AC performed experiments with synthetic model bacterial communities. AC and MSD analyzed Illumina sequencing data. MSD and NC wrote the paper with input from AC.

**Declaration of interests**

The authors declare no competing interests.
Legends:

Fig. 1. CASEU quantifies the fraction of individual strains in mixed communities by fitting mixed Sanger electropherograms as linear combinations of time-warped single-strain electropherograms. (A) Schematic of CASEU approach. Electropherograms shown are simulated for illustration purposes. For clarity, only a single fluorescence channel is illustrated. (B) Example of the continuous piecewise warping function used for alignment. The warping function is parameterized by six numbers, \( b_0-b_5 \) (the values of the function at \( \tau_0, \tau_1, \ldots, \tau_5 \)). The figure shows an exaggerated warping with simulated electropherograms for illustration purposes.

Fig 2. CASEU accurately resolves composition of two-strain mock communities. (A) An example alignment and fit over approximately 630 bases, showing a single fluorescence channel. Top and middle traces show reference electropherograms of individual strains (after warping). Bottom trace shows the electropherogram of a 1:1 mixture (black) and best-fit weighted sum of aligned references (red). (B) Zoom-in of segment of (A), showing alignments and fit over approximately 120 bases. (C) Phylogenetic tree of genes chosen for analysis, made using nearly full-length 16S sequences from ref (Datta et al., 2016). We aligned these sequences using the SINA Alignment Server (Pruesse et al., 2012) and made an approximate maximum-likelihood tree using FastTree 2.1.10 with the default options (Price et al., 2010). (D) Similarity matrix between all strains used in Figs. 2-4. Similarity was calculated as the Pearson correlation between electropherograms after aligning one to the other. Because each strain was measured in triplicate, each strain pair consists of a 3x3 submatrix of similarity values. The average of these replicate pairs is written in the figure. (E) Estimated mixture fractions plotted against the true ratio at which the sequences were mixed (circles). We also note the similarity (Pearson correlation) between strains. These mixture fractions have been corrected for errors in stock concentration (uncorrected fraction data shown in Fig. S3, bottom row).

Fig 3. CASEU provides reliable estimates of community composition in mixtures of 16S amplicons from four (left) or seven (right) strains. Solid bars show measurements after accounting for stock concentration error (uncorrected data is given in Fig. S6), red lines show true mixture proportions based on power law distributions. In power law distributions, the abundance of the \( i^{th} \) most abundant strain is proportional to \( \frac{1}{\alpha^i} \) where \( \alpha \) is the power law exponent.

Fig. 4. CASEU correctly infers the absence of strains that were not present in the mixture. (A) Community composition estimates of two-strain mixtures (as in Fig. 2D), in which five extra “decoy” strains were included as potential community members to test CASEU’s ability to infer strain absence. Bars indicate CASEU estimates (average of three replicates); open circles indicate each of the three replicate estimates, and red lines indicate the true values. Estimates are corrected for errors in stock concentrations. (B) Community composition estimates of four-strain mixtures (as in Fig. 3; \( \alpha \) is the power law exponent), in which three extra “decoy” strains were included as potential members. Bars and points are as in (A).

Fig. 5. CASEU yields estimates of community composition that are typically consistent with Illumina 16S sequencing. (A-G) Bar plots indicate results of CASEU analyses of mixtures of saturated cultures of four bacterial strains. Solid blue lines show estimates obtained from Illumina sequencing of the 16S rRNA V4-V5 hypervariable region. (H) Fractional abundance estimates for CASEU vs Illumina sequencing. Solid line shows equality.

Table 1. Comparison of methods for determining strain composition in simple model microbial communities. Sanger sequencing prices and turnaround times were obtained from Genewiz.
(https://www.genewiz.com/Public/Services/Sanger-Sequencing/Purified-Templates, accessed 2018 Apr 10).

* CGEB - Integrated Microbiome Resource (http://cgeb-imr.ca/pricing.html, accessed 2018 Feb 17).

** BioMicroCenter:Pricing - OpenWetWare. (https://openwetware.org/wiki/BioMicroCenter:Pricing, accessed 2018 Feb 17).

*** CGEB - Integrated Microbiome Resource (IMR). (http://cgeb-imr.ca/queue.html, accessed 2018 Feb 17).
| Method                                | Cost considerations                               | Speed                  | Measurement uncertainty          | Biological limitations                                                                 |
|--------------------------------------|--------------------------------------------------|------------------------|----------------------------------|----------------------------------------------------------------------------------------|
| Illumina sequencing of marker gene amplicons | Entirely outsourced >$20/sample for library prep and sequencing*  
In-house >$1500/lane at university facility, plus library prep costs ($37.5/sample)** | Typically weeks, sometimes months to get results*** | Ideally limited by Poisson (counting) error. Given 50,000 reads, can detect members with abundance <0.01% | Requires marker gene that has unique sequence but conserved primer sites for all strains (e.g., 16S rRNA gene) |
| qPCR of marker genes                  | <$1/sample for PCR                                | Same day               | Large dynamic range but low accuracy | Similar to Illumina sequencing. Requires designing specific primers or probes for each strain. |
| Plate counts of CFUs                 | Low (requires only agar plates)                   | Typically 2-3 days, depending on growth rates | Ideally limited by Poisson error | Strains must produce morphologically distinct colonies. Communities must be dissociable to single cells. |
| Fluorescent labeling of cells         | Flow cytometer or microscope use                  | Same day               | Ideally limited by Poisson error | Requires genetically tractable strains and spectrally distinct labels for each strain, potentially limiting communities to a few strains. |
| CASEU (this work)                    | $4-6/sample for sequencing, plus <$1/sample for PCR | As fast as next day     | Fractional abundance error typically one percentage point | Similar to Illumina sequencing for smaller consortia. |

*Note: Prices are approximate and subject to change.**Note: Additional costs may apply.***Note: Results may vary depending on specific conditions.
References

Abusleme, L., Hong, B.-Y., Dupuy, A.K., Strausbaugh, L.D., and Diaz, P.I. (2014). Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing. J. Oral Microbiol. 6, 23990.

Akerlund, T., Nordström, K., and Bernander, R. (1995). Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of Escherichia coli. J. Bacteriol. 177, 6791–6797.

Amir, A., and Zuk, O. (2011). Bacterial Community Reconstruction Using Compressed Sensing. J. Comput. Biol. 18, 1723–1741.

Brenner, K., You, L., and Arnold, F.H. (2008). Engineering microbial consortia: a new frontier in synthetic biology. Trends Biotechnol. 26, 483–489.

Celiker, H., and Gore, J. (2014). Clustering in community structure across replicate ecosystems following a long-term bacterial evolution experiment. Nat. Commun. 5, 4643.

Datta, M.S., Sliwerska, E., Gore, J., Polz, M.F., and Cordero, O.X. (2016). Microbial interactions lead to rapid micro-scale successions on model marine particles. Nat. Commun. 7, 11965.

Eilers, P.H.C. (2004). Parametric Time Warping. Anal. Chem. 76, 404–411.

Enke, T.N., Datta, M.S., Schwartzman, J., Cermak, N., Schnitz, D., Barrere, J., Pascual-García, A., and Cordero, O.X. (2019). Modular Assembly of Polysaccharide-Degrading Marine Microbial Communities. Curr. Biol. 29, 1528-1535.e6.

Friedman, J., Higgins, L.M., and Gore, J. (2017). Community structure follows simple assembly rules in microbial microcosms. Nat. Ecol. Evol. 1, 109.

Goldford, J.E., Lu, N., Bajić, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., Segrè, D., Mehta, P., and Sanchez, A. (2018). Emergent simplicity in microbial community assembly. Science 361, 469–474.

Harcombe, W.R., Riehl, W.J., Dukovski, I., Granger, B.R., Betts, A., Lang, A.H., Bonilla, G., Kar, A., Leiby, N., Mehta, P., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. Cell Rep. 7, 1104–1115.

Hildenbrand, C., Stock, T., Lange, C., Rother, M., and Soppa, J. (2011). Genome Copy Numbers and Gene Conversion in Methanogenic Archaea. J. Bacteriol. 193, 734–743.

Hill, J.T., Demarest, B.L., Bisgrove, B.W., Su, Y.-C., Smith, M., and Yost, H.J. (2014). Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. Dev. Dyn. Off. Publ. Am. Assoc. Anat. 243, 1632–1636.

Hunt, D.E., David, L.A., Gevers, D., Preheim, S.P., Alm, E.J., and Polz, M.F. (2008). Resource Partitioning and Sympatric Differentiation Among Closely Related Bacterioplankton. Science 320, 1081–1085.

Kinde, I., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2012). FAST-SeqS: A Simple and Efficient Method for the Detection of Aneuploidy by Massively Parallel Sequencing. PLOS ONE 7, e41162.

Kommedal, Ø., Karlsen, B., and Sæbø, Ø. (2008). Analysis of Mixed Sequencing Chromatograms and Its Application in Direct 16S rRNA Gene Sequencing of Polymicrobial Samples. J. Clin. Microbiol. 46, 3766–3771.

Kosmol, Ø., Kvello, K., Skjåstad, R., Langeland, N., and Wiker, H.G. (2009). Direct 16S rRNA Gene Sequencing from Clinical Specimens, with Special Focus on Polybacterial Samples and Interpretation of Mixed DNA Chromatograms. J. Clin. Microbiol. 47, 3562–3568.
Kommedal, Ø., Lekang, K., Langeland, N., and Wiker, H.G. (2011). Characterization of polybacterial clinical samples using a set of group-specific broad-range primers targeting the 16S rRNA gene followed by DNA sequencing and RipSeq analysis. J. Med. Microbiol. 60, 927–936.

Lawson, C.L., and Hanson, R.J. (1995). Solving Least Squares Problems (SIAM).

Momeni, B., Chen, C.-C., Hillesland, K.L., Waite, A., and Shou, W. (2011). Using artificial systems to explore the ecology and evolution of symbioses. Cell. Mol. Life Sci. CMLS 68, 1353–1368.

Momeni, B., Xie, L., and Shou, W. (2017). Lotka-Volterra pairwise modeling fails to capture diverse pairwise microbial interactions. ELife 6, e25051.

Nielsen, N.-P.V., Carstensen, J.M., and Smidsgaard, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. J. Chromatogr. A 805, 17–35.

Polz, M.F., and Cavanaugh, C.M. (1998). Bias in Template-to-Product Ratios in Multitemplate PCR. Appl. Environ. Microbiol. 64, 3724–3730.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PloS One 5, e9490.

Pruesse, E., Peplies, J., and Glöckner, F.O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinforma. Oxf. Engl. 28, 1823–1829.

R Core Team (2017). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).

Ratzke, C., and Gore, J. (2018). Modifying and reacting to the environmental pH can drive bacterial interactions. PLOS Biol. 16, e2004248.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74, 5463–5467.

Schaechter, M., Maaløe, O., and Kjeldgaard, N.O. (1958). Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. J. Gen. Microbiol. 19, 592–606.

Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B., and Hood, L.E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679.

Swerdlow, H., and Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. Nucleic Acids Res. 18, 1415–1419.

Větrovský, T., and Baldrian, P. (2013). The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. PLOS ONE 8, e57923.

Wolfe, B.E., Button, J.E., Santarelli, M., and Dutton, R.J. (2014). Cheese rind communities provide tractable systems for in situ and in vitro studies of microbial diversity. Cell 158, 422–433.

Wolff, T.Y., Eickhardt, S., Björnsdottir, M.K., Moser, C., Bjarnsholt, T., Høiby, N., and Thomsen, T.R. (2013). Direct sequencing and RipSeq interpretation as a tool for identification of polymicrobial infections. J. Clin. Microbiol. JCM.00190-13.

Yuan, S., Cohen, D.B., Ravel, J., Abdo, Z., and Forney, L.J. (2012). Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome. PLOS ONE 7, e33865.
Zhou, K., Qiao, K., Edgar, S., and Stephanopoulos, G. (2015). Distributing a metabolic pathway among a microbial consortium enhances production of natural products. Nat. Biotechnol. 33, 377–383.
A

1) Measure Sanger electropherograms of individual strains

Strain 1

Strain 2

2) Measure Sanger electropherogram of mixture

Mixed electropherogram

3) Warp electropherograms to align them to mixed electropherogram

Strain 1 (aligned)

Strain 2 (aligned)

4) Fit mixed electropherogram to determine composition

Mixed electropherogram

Best fit linear combination (50% strain 1, 50% strain 2)

B

$g(t)$

$b_0, b_1, b_2, b_3, b_4, b_5$

Before warping

After warping

Time, t
Our method “CASEU” estimates amplicon composition from mixed Sanger electropherograms

CASEU uses a marker gene and is fast (1 day turnaround) and inexpensive ($5/sample)

CASEU is accurate to ±1% abundance in mock microbial communities with up to 7 strains

CASEU is available as a free, open-source R package