iScience

CellPress

Article

# Rapid, Inexpensive Measurement of Synthetic Bacterial Community Composition by Sanger Sequencing of Amplicon Mixtures

Nathan Cermak,[1,5] Manoshi Sen Datta,[2,5,6,*] and Arolyn Conwill[3,4]

## 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. Although 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 per sample.**

## INTRODUCTION

Model microbial communities, comprising 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 the 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 be applied only 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; Schaechter 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

[1] Fremont, CA, USA, 94555

[2] Department of Biology, Technion - Israel Institute of Technology, Haifa, Israel

[3] Physics of Living Systems, Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4] Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[5] These authors contributed equally

[6] Lead Contact

*Correspondence: mdatta8788@gmail.com

https://doi.org/10.1016/j.isci.2020.100915

| Method | Cost Considerations | Speed | Measurement Uncertainty | Biological Limitations |
|---|---|---|---|---|
| Illumina sequencing of marker gene amplicons | Entirely outsourced<br>>$20/sample for library prep and sequencing[a]<br>In house<br>>$1,500/lane at university facility, plus library prep costs ($37.5/sample)[b] | Typically weeks, sometimes months to get results[c] | 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 1 percentage point | Similar to Illumina sequencing for smaller consortia |

**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).

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

[b] BioMicroCenter:Pricing—OpenWetWare. (https://openwetware.org/wiki/BioMicroCenter:Pricing, accessed 2018 Feb 17).

[c] CGEB—Integrated Microbiome Resource (http://cgeb-imr.ca/queue.html, accessed 2018 Feb 17).

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 with 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 computationally solved 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 2 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; Hill et al., 2014) 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 owing 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 Figures 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 (Figure 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

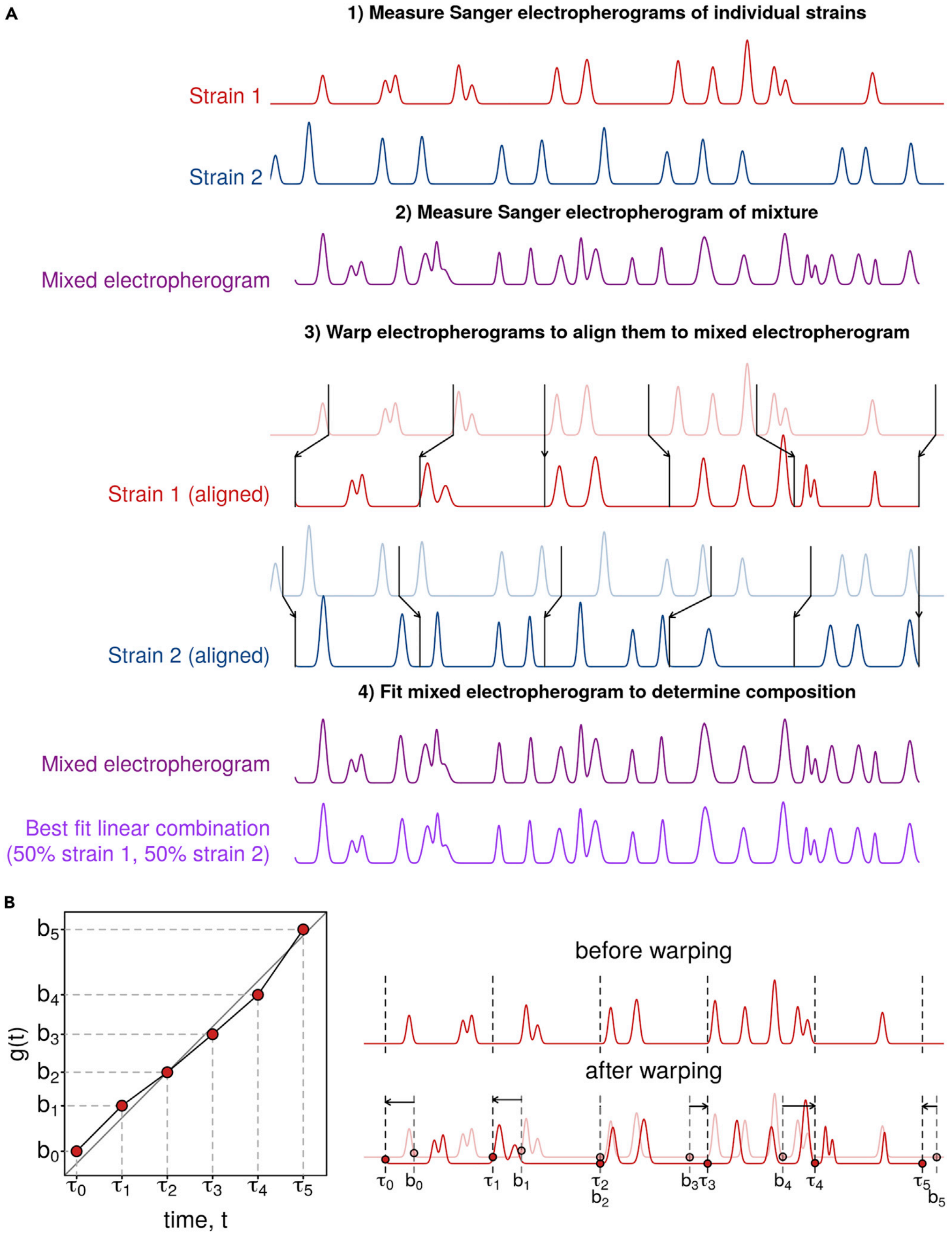
A
1) Measure Sanger electropherograms of individual strains
Strain 1
Strain 2
2) Measure Sanger electropherogram of mixture
Mixed electropherogram
3) Warp electropherograms to align them to mixed electropherogram
Strain 1 (aligned)
Strain 2 (aligned)
4) Fit mixed electropherogram to determine composition
Mixed electropherogram
Best fit linear combination
(50% strain 1, 50% strain 2)
B
g(t)
$b_5$
$b_4$
$b_3$
$b_2$
$b_1$
$b_0$
$\tau_0$ $\tau_1$ $\tau_2$ $\tau_3$ $\tau_4$ $\tau_5$
time, t
before warping
after warping
$\tau_0$ $b_0$ $\tau_1$ $b_1$ $\tau_2$ $b_2$ $b_3$ $\tau_3$ $b_4$ $\tau_4$ $\tau_5$ $b_5$

**Figure 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, *b0-b5* (the values of the function at τ0, τ1, …, τ5). The figure shows an exaggerated warping with simulated electropherograms for illustration purposes.

fractions that minimize the sum-of-squares difference between the observed and model-predicted mixed electropherogram, as follows:

$$\text{argmin}_{f_1,\ldots,f_n,g_1(t),\ldots,g_n(t),\, x_0} \sum_{t=t_0}^{t_{end}} \sum_{c=1}^{4} \left( Y[t,c] - \left( x_0 + \sum_{i=1}^{n} f_i X_i[g_i(t),c] \right) \right)^2 + \lambda \sum_{i=1}^{n} R(g_i) \qquad \text{(Equation 1)}$$

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 $t_{end}$;
- $Y[t,c]$ is a matrix of the mixed electropherogram, with one row per time point 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 time point 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 with the average speed. To account for within-electropherogram variability, we use a continuous piecewise linear warping function $g_i(t)$ (see Figure 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 Figure 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 (Figure S2). Using only a single segment yielded poor alignments between technical replicates (Figure S2) and produced mediocre estimates of known mixture fractions (Figure S3). Using more than five segments did not improve the alignments between technical replicates (Figure 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% and 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.

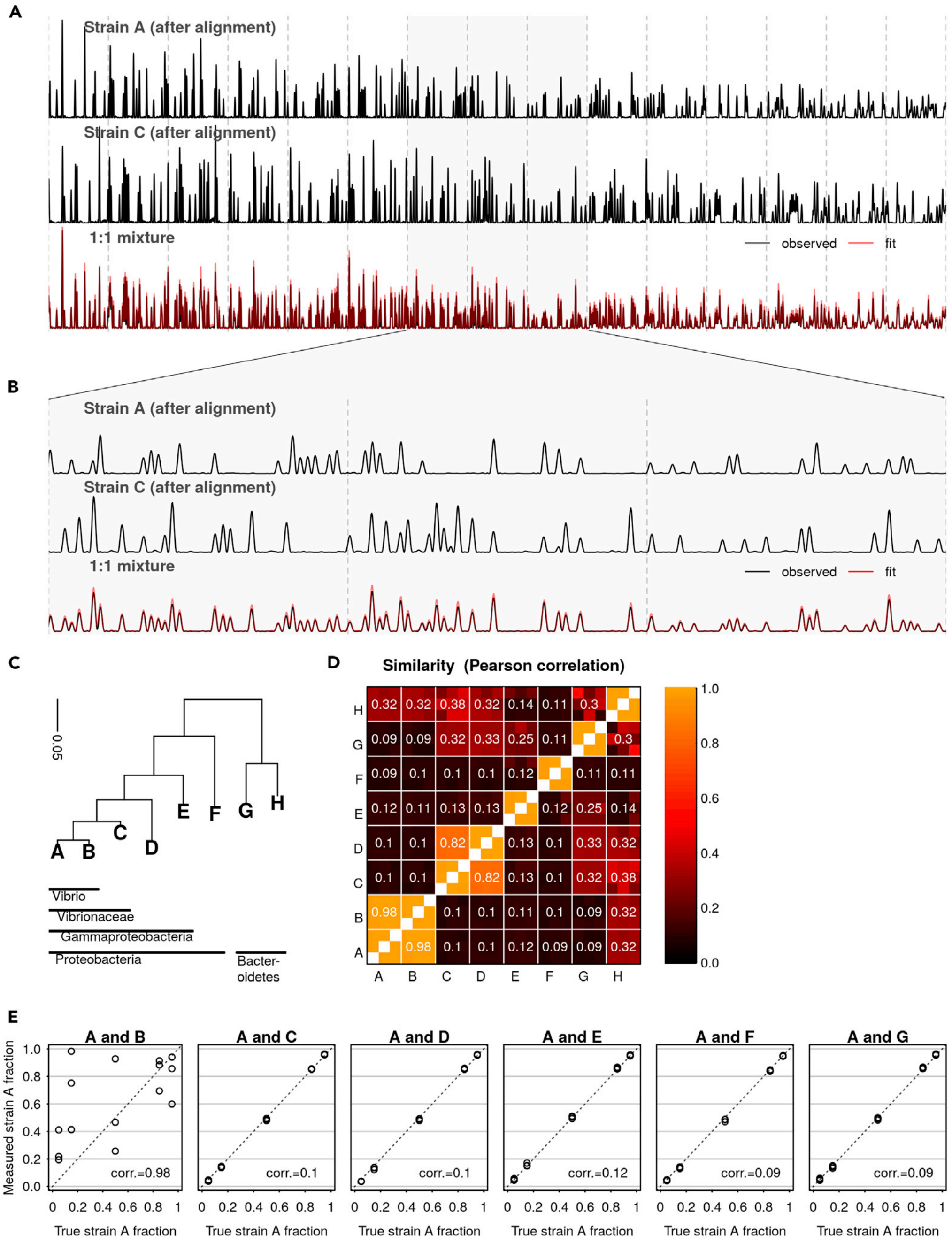

**Figure 2. CASEU Accurately Resolves Composition of Two-Strain Mock Communities**

(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 Datta et al. (2016). We aligned these sequences using the SINA Alignment Server (Pruesse et al., 2012) https://www.arb-silva.de/aligner/ 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 Figures 2, 3, and 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 3 × 3 submatrix of similarity values. The average of these replicate pairs is written in the figure.

**Figure 2. *Continued***

(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 Figure S3, bottom row).

We first analyzed two-strain mixtures for which the proportion of a single strain varied from 5% to 95% (Figure 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%–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%–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 (Figure 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 with 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 (Figure S5A). The standard deviations we observe between triplicate results were comparable with what would be attained by counting-based methods (e.g., next-generation sequencing or plate counts) with ~5,000 counts (reads or colonies) per sample (Figure 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 (Figure 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 (Figure S4). Although CASEU failed to differentiate strains A and B, which have an average post-alignment correlation of 0.98 (Figure 2), it accurately estimated fractional abundances for all other communities (Figures 2 and 3) containing between-strain correlations of up to 0.82 (strains C and D; Figure 3). However, strain D was sometimes mistakenly found in the mixtures of strains 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. Strains A and B differ by only SNPs, whereas 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 (Figure S4). This enabled us to obtain far more accurate fractional

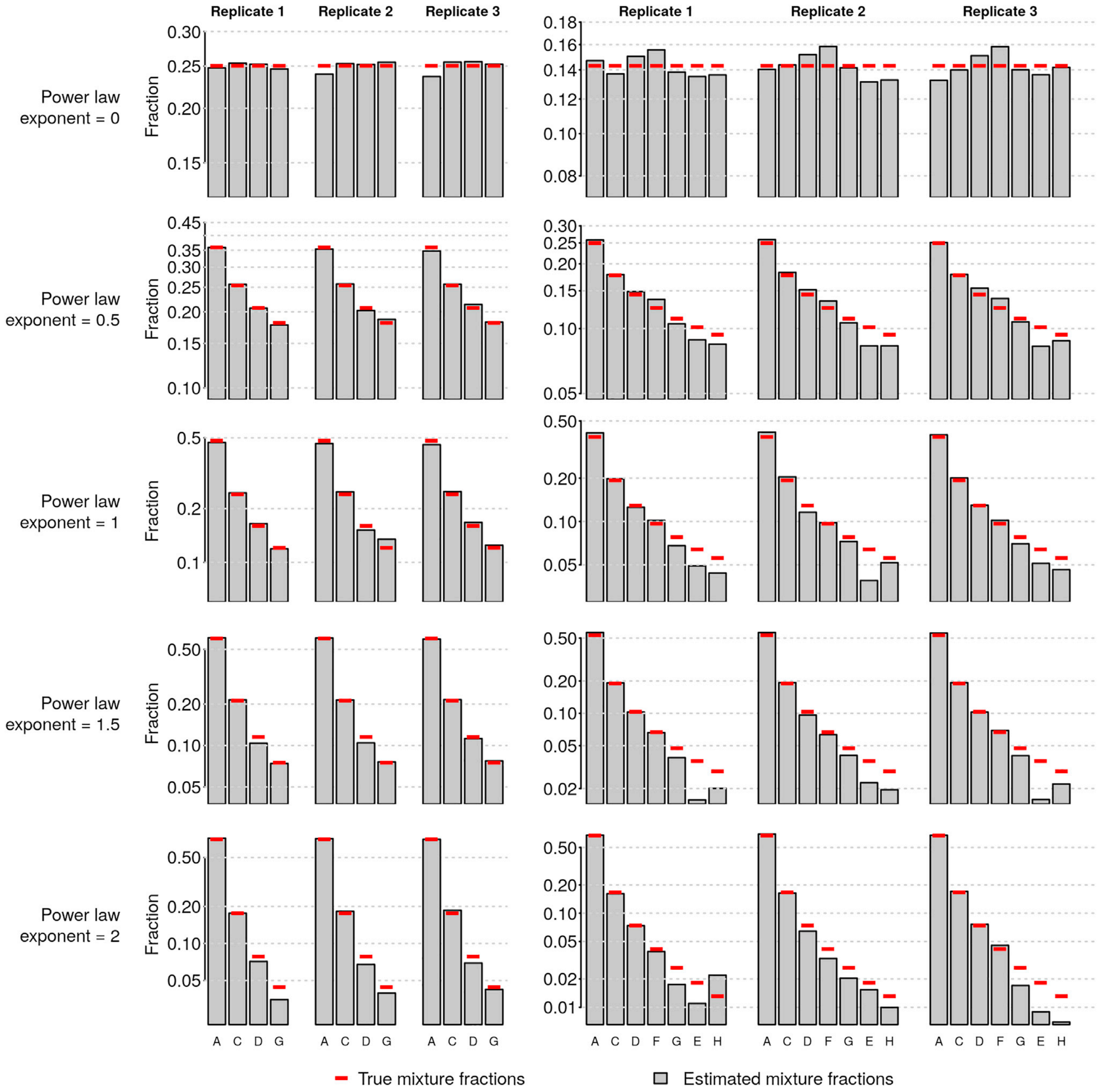

**Figure 3. CASEU Provides Reliable Estimates of Community Composition in Mixtures of 16S Amplicons from Four (Left) or Seven (Right) Strains**

Solid bars show measurements after accounting for stock concentration error (uncorrected data are given in Figure S6); red lines show true mixture proportions based on power law distributions. In power law distributions, the abundance of the ith most abundant strain is proportional to $\frac{1}{i^{\alpha}}$ where $\alpha$ is the power law exponent.

abundance estimates for mixtures of strains A and B (Figure 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

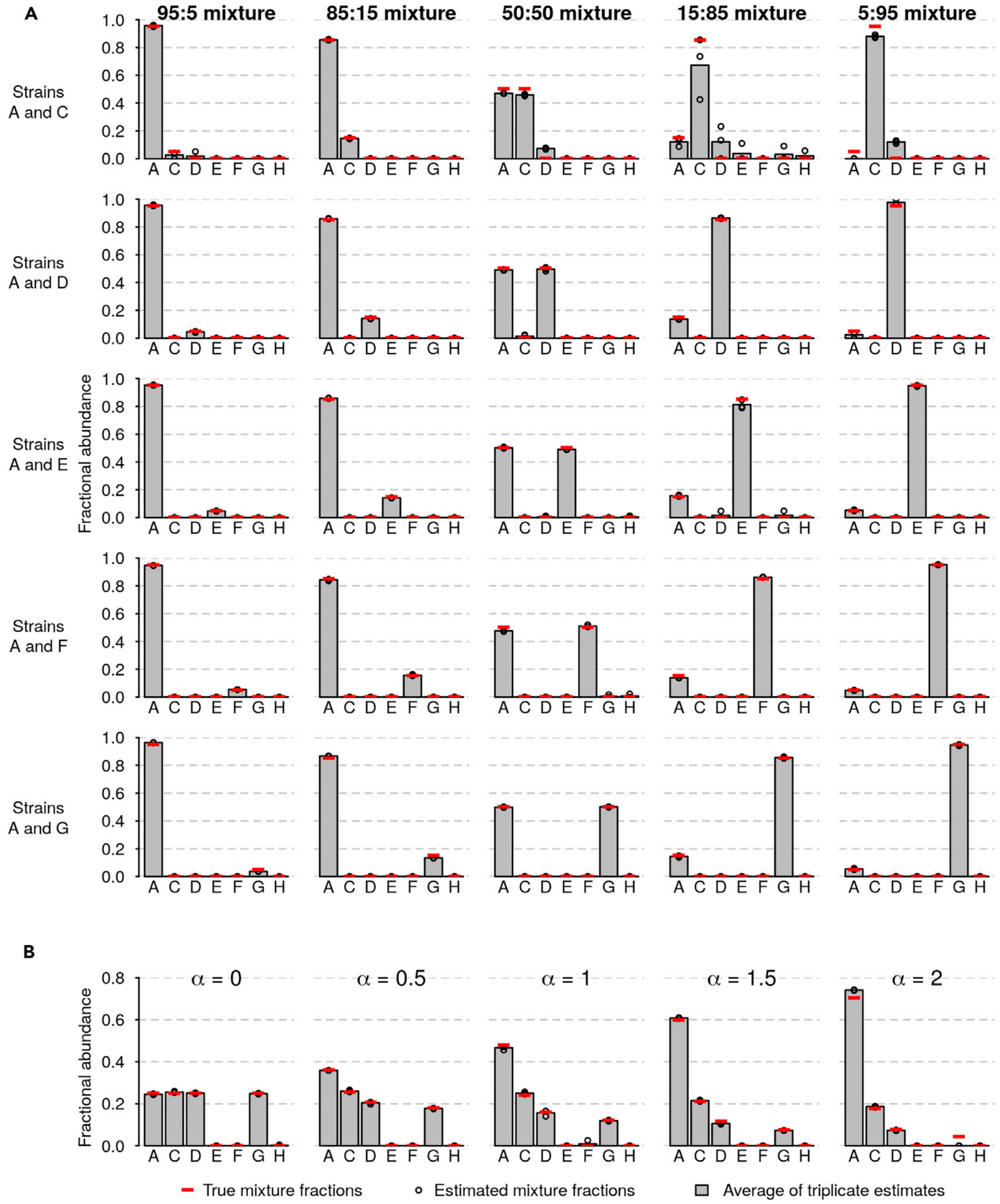

**Figure 4. CASEU Correctly Infers the Absence of Strains that Were Not Present in the Mixture**

(A) Community composition estimates of two-strain mixtures (as in Figure 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 Figure 3; α is the power law exponent), in which three extra "decoy" strains were included as potential members. Bars and points are as in (A).

four-strain model communities of unknown fractional composition, derived from strains isolated in Enke et al. (2019). 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) (Figure 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.

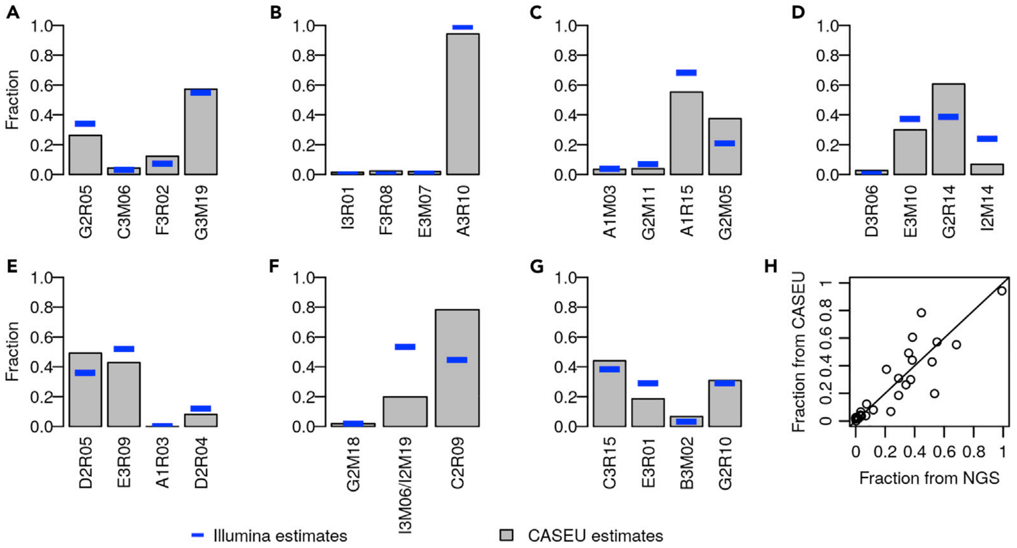

**Figure 5. CASEU Yields Estimates of Community Composition That Are Typically Consistent with Illumina 16S Sequencing**

(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 versus Illumina sequencing. Solid line shows equality.

We found that CASEU provided community composition estimates consistent with next-generation Illumina sequencing of 16S rRNA amplicons (Figure 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, Figure 5H). Furthermore, in five of seven communities, we observed strong quantitative agreement between Illumina estimates and CASEU estimates, with an RMS difference of 7.0 percentage points (Figure 5A–5C, 5E, and 5G).

In the two model communities where CASEU and Illumina sequencing disagreed (Figures 5D and 5F; 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. Although 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 (Figure S7). In the first case, the predicted electropherogram had a correlation of only 0.63 to the observed electropherogram, compared with >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 (Figure S7A). In the second case, CASEU yielded a correlation between predicted and observed traces of 0.45, compared with >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 (Figure 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 1 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 1 day (e.g., 40 h for paired-end 150 × 150 sequencing) but can require weeks to months if outsourced.

## CASEU Can be Inexpensive

Sanger sequencing has a fixed cost per sample (here, $4 for sequencing and roughly $1 for PCR and cleanup), whereas Illumina sequencing has large upfront cost (typically more than $1,000 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 1% 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 be 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 the 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.

### *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.

### *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.

### *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.

### *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. Although 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).

## METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

## 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.

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.100915.

## ACKNOWLEDGMENTS

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

## AUTHOR CONTRIBUTIONS

M.S.D. and N.C. designed the model and fitting algorithm, performed the experiments with amplicon mixtures, and analyzed all Sanger sequencing data. A.C. performed experiments with synthetic model bacterial communities. A.C. and M.S.D. analyzed Illumina sequencing data. M.S.D. and N.C. wrote the paper with input from A.C.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 16, 2019
Revised: January 22, 2020
Accepted: February 10, 2020
Published: March 27, 2020

## 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., Schmitz, 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. *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.

Kommedal, 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). https://www.dropbox.com/s/9ad2ilzw4u09mob/solving-least-squares-problems.pdf?dl=0.

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. *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 Smedsgaard, 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. Bioinformatics *28*, 1823–1829.

R Core Team (2017). R: A Language and Environment for Statistical Computing (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. U S A *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. *51*, 1281–1284.

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.

**iScience, Volume *23***

## Supplemental Information

# Rapid, Inexpensive Measurement of Synthetic Bacterial Community Composition by Sanger Sequencing of Amplicon Mixtures

**Nathan Cermak, Manoshi Sen Datta, and Arolyn Conwill**

# Transparent Methods

**Strains.** Strains A-H are marine isolates collected in (Datta et al., 2016) from Canoe Beach, Nahant, MA, USA. The taxonomic identities of the isolates (classified using SINA (Pruesse et al., 2012)) are as follows: Strain A (1A01), Vibrio; Strain B (4B04), Vibrio; Strain C (6D03), Vibrionaceae; Strain D (6C06), Psychromonas; Strain E (4A09), Oceanospirillaceae; Strain F (4A10), Rhodobacteraceae; Strain G (4C08), Polaribacter; Strain H (6B07). Strains in Figs. 5 and S7 are isolates from (Enke et al., 2019), collected from the same location. Communities are denoted by their subfigure label in Fig. 5.

A: G2R05 Cellulophaga, C3M06 Rhodobacteraceae, F3R02 Neptunomonas, G3M19 Celeribacter
B: I3R01 Vibrio, F3R08 Shewanella, E3M07 Paraglaciecola, A3R10 Tenacibaculum
C: A1M03 Alteromonas, G2M11 Colwellia, A1R15 Pseudoalteromonas, G2M05 Photobacterium
D: D3R06 Colwellia, E3M10 Cellulophaga, G2R14 Vibrio, I2M14 Marinobacterium
E: D2R05 Alteromonadaceae, E3R09 Winogradskyella, A1R03 Shewanella, D2R04 Rhodobacteraceae
F: G2M18 Saccharospirillaceae, I3M06 Marinobacterium, C2R09 Paracoccus, I2M19 Marinobacterium
G: C3R15 Flavobacteriaceae, E3R01 Tenacibaculum, B3M02 Psychromonas, G2R10 Vibrio

**Preparing mixtures of 16S amplicons and sequencing.** For two-, four-, and seven-strain mixtures, genomic DNA was extracted as previously reported[15]. 16S genes were amplified with 27F (AGAGTTTGATCMTGGCTCAG) and 1492R (TACGGYTACCTTGTTACGACTT) universal primers, as follows:

| Reagent | Volume |
|---|---|
| ddH2O | 23.5µL |
| 5X HF Buffer | 10 µL |
| dNTPs (10mM) | 1 µL |
| 27F primer (3uM) | 5 µL |
| 1492R primer (3uM) | 5 µL |
| Phusion polymerase | 0.5 µL |
| Genomic DNA | 5 µL |
| Total | 50 µL |

PCR cycle conditions were as follows:

| Step | Temperature | Duration |
|---|---|---|
| Initial denaturation | 98°C | 30 seconds |
| Amplification (30 cycles) | 98°C | 30 seconds |
| | 50°C | 30 seconds |
| | 72°C | 90 seconds |
| Final extension | 72°C | 10 minutes |

For experiments shown in Figs. 2 and 3, we ran six PCRs for each strain to ensure that we had sufficient amplicon DNA to prepare all the mixtures. We pooled each set of six reactions into a single tube, then SPRI-cleaned the products. We estimated DNA concentrations via Nanodrop, and subsequently diluted all samples to 3 ng/µL (concentration measurements required subsequent computational correction, see Materials and Methods). We added 5 µL of 27F primer at 15 µM to 40 µL of amplicon at 3 ng/µL, to yield 45 µL with a primer concentration of 1.6 µM and a template concentration of 2.6 ng/µL. These concentrations are what Genewiz recommends for Sanger sequencing (https://www.genewiz.com/Public/Resources/Sample-Submission-Guidelines/Sanger-Sequencing-Sample-Submission-Guidelines/Sample-Preparation#sanger-sequence, accessed 2018 Apr 2). We split the 45 µL of sample into three separate plates, each with 15 µL of sample per well, and submitted each plate on a different day over the course of one week.

Sequencing was performed by Genewiz as a drop-off service for $6/sample (<48 samples) or $4/sample (>48 samples). We routinely received results within 24 hours of submitting our samples. Our ABIF file metadata suggests Genewiz sequencing was performed on a 3730xl DNA Analyzer, using BigDyeV3.

**Processing ABIF files.** We used the 'sangerseqR' Bioconductor package (Hill et al., 2014) in R (R Core Team, 2017) to read in ABIF (.ab1) files. In ABIF files, there are two types of data we considered using: “raw” fluorescence traces, and “processed” data. While the details of the processing method are not available, the process appears to involve baseline subtraction, low-pass filtering, and an unknown temporal adjustment. Attempts to use the “raw” traces were stymied by the poor temporal alignment of the traces and required searching a much larger range of alignment parameters, yielding a significantly slower analysis. Before analysis, we additionally normalized the amplitudes of all reference files such that the mean amplitude was one over the region to be used for alignment.

**Algorithm for fitting mixed electropherograms.** We initially tried optimizing Equation (1) via the Nelder-Mead algorithm (also called downhill simplex) but found that this method tended to yield solutions that were very dependent on starting estimates (suggesting many local minima). Instead, we adopted an approach in which we determine the warping parameters for one strain at a time. To determine the warping parameters for a single strain (“aligning a single strain”), we use the dynamic programming approach pioneered in correlation-optimized warping(Nielsen et al., 1998). In brief, we first calculate the sum of squared errors over a 2D grid of values for parameters $b_1$ and $b_2$ (the boundaries of the first warping segment). (Note that to do so, we rapidly calculate optimal $f_i$ values for each $(b_1, b_2)$ pair using non-negative least squares.) For each possible value of $b_2$, we only keep track of the best value of $b_1$ and its corresponding error. We then repeat that same process, but this time evaluating the error for a grid of possible values for $b_2$ and $b_3$ (the boundaries of the second segment). For every $b_2$-$b_3$ pairing, we calculate the error over that segment, plus the lowest possible error for that value of $b_2$ over any value of $b_1$. We then record the lowest error obtainable for any given value of $b_3$, and the corresponding $b_2$ that yields that optimum. We repeat this process for all five warping segments. This process does not necessarily yield a globally optimal solution because, for the errors for each segment to be additive, we allowed each segment to have its own amplitude parameter $f_i$. However, ultimately this parameter $f_{i,}$ must be the same for all segments. We thus globally refine parameter estimates by minimizing equation (1) directly via Nelder-Mead, starting from the ($b_1$-$b_6$) estimates obtained as described above (which are usually very near the final optimum).

To align multiple strains, we sequentially align strains one after another. We first align each strain individually, identify the one that yields the greatest improvement in the fit, and fix that strain’s alignment parameters ($b_1$-$b_6$). We then repeat that process for the remaining strains, always greedily fixing the parameters of whichever strain yields the greatest improvement in the fit (reduction in squared error). Our rationale was that this would ensure that we always fit the majority component of the mixture before fitting the minority components.

More formally, our algorithm is as follows:

```
Inputs:
•  $Y[t,c]$ - Mixed electropherogram matrix
•  $X_1[t,c], X_2[t,c], \ldots, X_n[t,c]$ - Individual reference electropherogram matrices ($n$ is number of strains)

Run:
1. Initialize $F = \emptyset$, the indices of strains for which the alignment parameters have been fixed.
2. Initialize $U = \{1, ..., n\}$, the indices of strains for which the alignment parameters have not yet been fixed.
3. Initialize $A$ as an empty $n \times 6$ matrix for the alignment parameters.
4. While $U \neq \emptyset$
    1. For every strain index $i$ in $U$, find a warping of $X_i$ via dynamic programming that yields the greatest reduction in equation (1), conditional on all strains with fixed alignments $X_f$ for $f \in F$.
    2. Identify which strain $j$, yields the most improvement to the fit.
    3. Fine-tune the alignment parameters of strain $j$ by downhill simplex.
    4. Record strain $j$'s alignment parameters in row $j$ of matrix $A$.
    5. Add strain $j$ to $F$, remove it from $U$.
5. Fine-tune all alignment parameters simultaneously by optimizing over $A$ via downhill simplex, starting from $A$ (the best individual alignments obtained in step 4).
```

On a HP laptop with an Intel Core i7-7500U CPU and 16 GB RAM, our fitting approach took on average 10.8 seconds for a two-strain mixture, and 138 seconds to fit a seven-strain mixture. Notably, the computation time is expected to be at worst quadratic in the number of strains that could potentially be in the mixture.

**Accounting for concentration errors.** Before mixing amplicons to make mock communities, we attempted to normalize all amplicon stock solutions to the same concentration, as measured by Nanodrop. However, the Nanodrop has limited precision, and as such we have fit a model to correct for remaining non-uniformity in stock solution concentration. For all 120 amplicon mixture samples in Figs. 2 and 3, we fit a model in which the amplicon concentration of each strain *i* was $C_i$ times greater than expected. We then calculate the mixture fractions that would have resulted had the concentrations been equal.

$$f_{i,corrected} = \frac{\frac{f_{i,observed}}{C_i}}{\sum_{j=1}^{n} \frac{f_{j,observed}}{C_j}}$$

We estimated $C_i$ values by finding the $C_i$ that minimized the mean squared difference between the corrected fractions and the known fractions at which the amplicon stock solutions were volumetrically mixed. Arbitrarily, $C_A$ was set to 1, and all $C_i$ values were defined relative to that (therefore our model has seven free parameters). Concentration error estimates were as follows: $C_B$=5.35 (not reliable due to poor estimates of $f_{observed}$), $C_C$=1.10, $C_D$=0.80, $C_E$=0.66, $C_F$=0.91, $C_G$=0.86, $C_H$=0.75.

**Model community dynamics and Illumina sequencing.**
Communities used in Fig. 5 contained non-overlapping sets of four marine isolates. We grew communities at room temperature while shaking in 200uL of 2216 Marine Broth media in a 96-well deep well plate. Communities were diluted 100-fold and transferred to new plates every 24 hours and sampled after two weeks for DNA extraction and sequencing. DNA extraction was performed using a Epicentre MasterPure Kit.

Illumina 16S V4-V5 library preparation and sequencing were performed at Integrated Microbiome Resource (IMR) on an Illumina MiSeq (paired-end, 300-basepair reads). For Sanger sequencing, samples were PCR-amplified using identical conditions as described for the amplicon mixtures (above), but in 25 µL volumes instead of 50 µL.

**Analysis of Illumina sequencing data.** Illumina sequencing of model communities resulted in 514,524 reads (average per sample of 64,316 reads, range over all samples of 46,633-72,095 reads). Paired-end reads were merged with vsearch –fastq_mergepairs (10 mismatches allowed in overlap region) and trimmed of primer sequences with cutadapt 1.16. We estimated read counts for each isolate by assigning trimmed and merged reads that perfectly matched a known isolate 16S rRNA V4-V5 sequence to that isolate. Reads that did not match a known isolate were discarded. Fractional abundances were estimated by dividing the read count for each isolate in a sample by the read counts for all isolates in that sample.

# Supplemental Figures and Tables

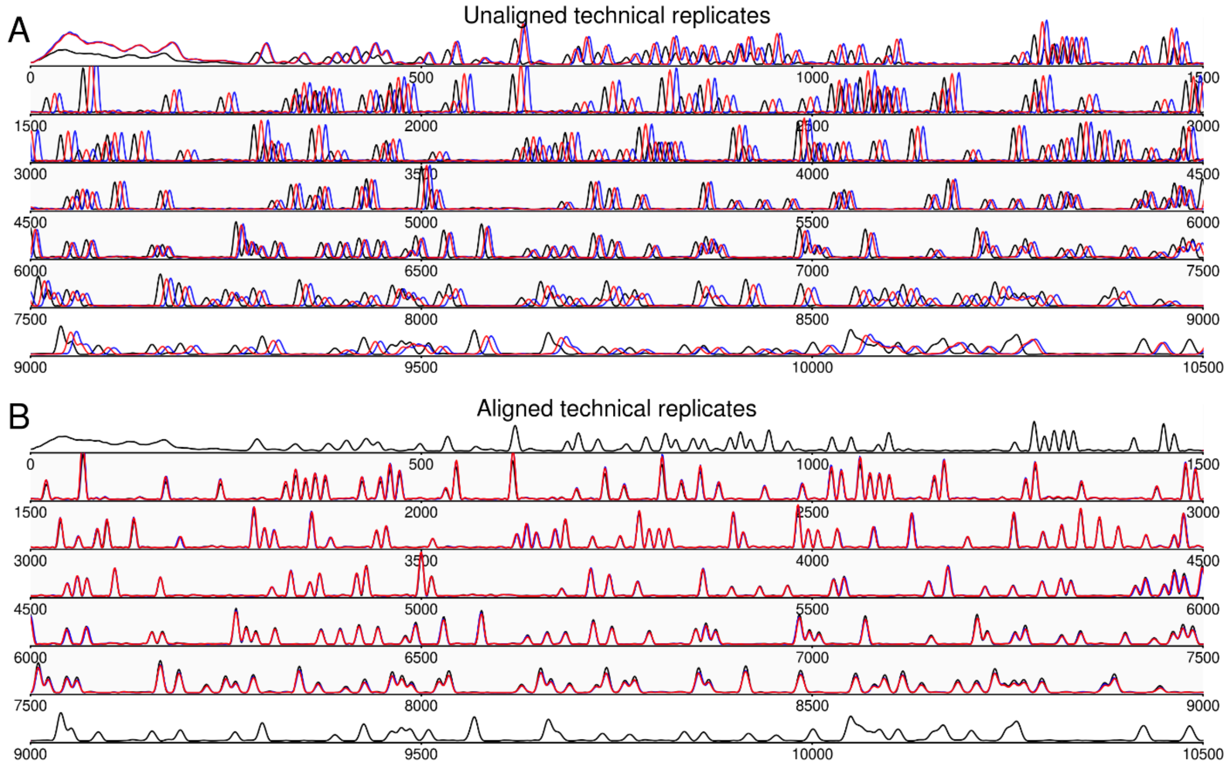

**Fig S1. Aligning via time-warping can correct for temporal variability among technical replicates.** Related to Figure 1**.**

(A) Three technical replicates (black, red and blue) for a sample of 16S DNA (showing only a single fluorescence channel for clarity).

(B) Technical replicates two and three aligned to technical replicate one, over indices 1500-9000, covering ~630 bases.

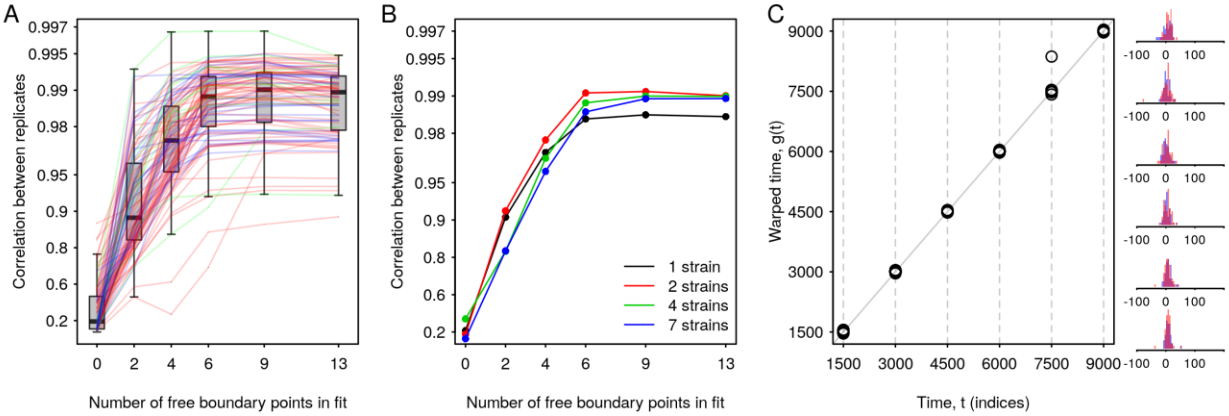

**Fig S2. Determining optimal warping flexibility and typical range of warping parameters.** Related to Figure 1.

(A) A six-parameter alignment yields good fits without unnecessary degrees of freedom. Fit quality was quantified as post-alignment Pearson correlation between technical replicates. With less than six boundary parameters, fits can be improved by increasing the warping flexibility, but beyond six parameters improvements are minimal. Lines indicate individual replicates and are colored according to the legend shown in (B).

(B) Same data as in (A) but averaged for each type of mixture. The similarity between lines shows that the required flexibility of the warping function does not depend on the sample type.

(C) Warping functions estimated from fitting each sample to its two technical replicates (N=48 samples; 8 single-strain samples; 30 two-strain samples; 5 four-strain samples; 5 seven-strain samples). At right are histograms of offsets for each boundary (how many time indices the boundary was moved in the alignment). Pink bars show alignment parameters for replicate 2, blue bars show alignment parameters for replicate 3, suggesting no day-specific effects.

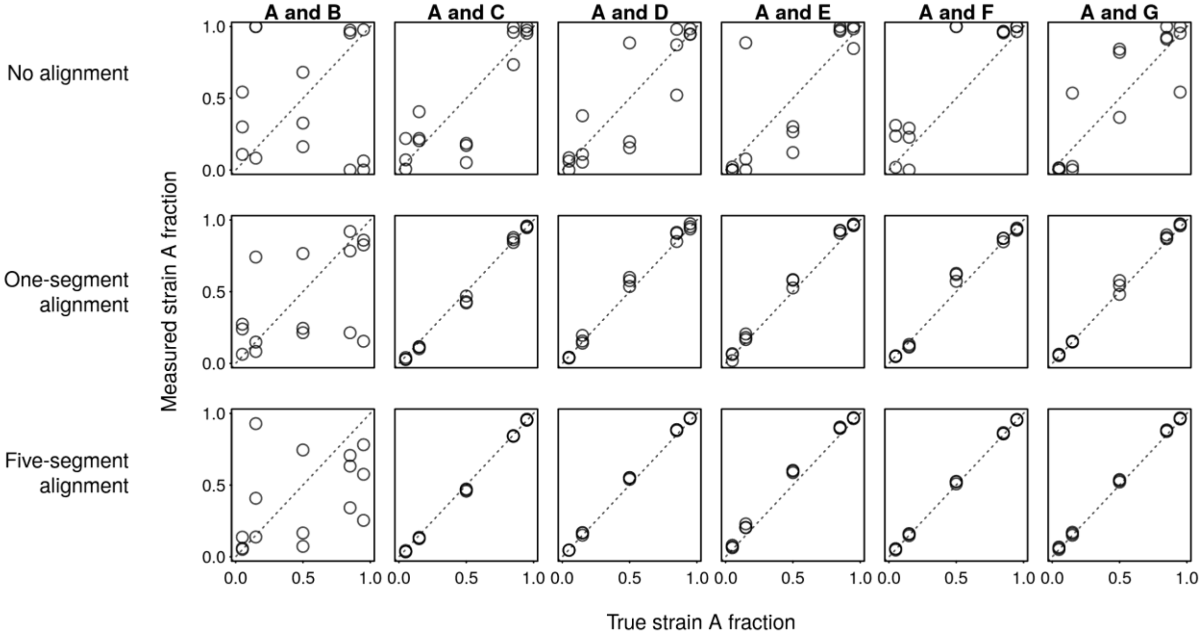

**Fig S3. Flexible alignment is necessary to quantify community composition accurately.** Related to figure 2. Without alignment, estimates of the fractions of each component are poor (top row). With 2-parameter alignment it yields more accurate results (middle row), but still much less precise than those obtained with 6-parameter alignment (bottom row). Data in this figure is not corrected for error in stock solution concentrations.

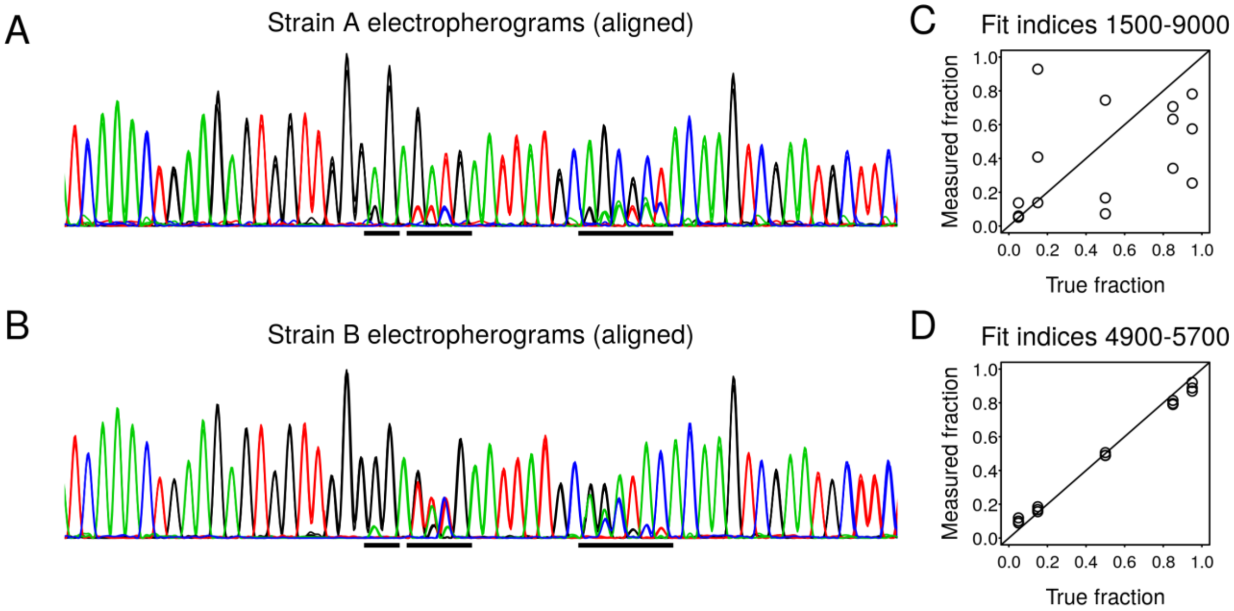

**Fig S4. Strains A and B have similar electropherograms, but can be quantified by CASEU when restricting the analysis to the differing region.** Related to figure 2.

**(A)** Overlay of aligned triplicate electropherograms of strain A, spanning electropherogram indices 4900-5700 (total electropherogram length is typically ~13000 indices, fit region is indices 1500-9000 for all other analyses in this paper). Colors correspond to the four fluorescence channels.

(B) Overlay of electropherograms of strain B, aligned to an arbitrarily-chosen replicate of strain A. Underlined stretches of ~6-7 bases are the positions at which these electropherograms differ. Over the remainder of the fit region, the electropherograms yield identical sequences.

(C) Measured fractional abundances of strain in mixtures of strains A and B (same as lower left panel of Fig S3), fitting the full electropherogram region as used throughout the rest of the paper, approximately 630 bases.

(D) Same as (C), but restricting the fitting to roughly the region in which the two genes differ (the region shown in (A) and (B)).

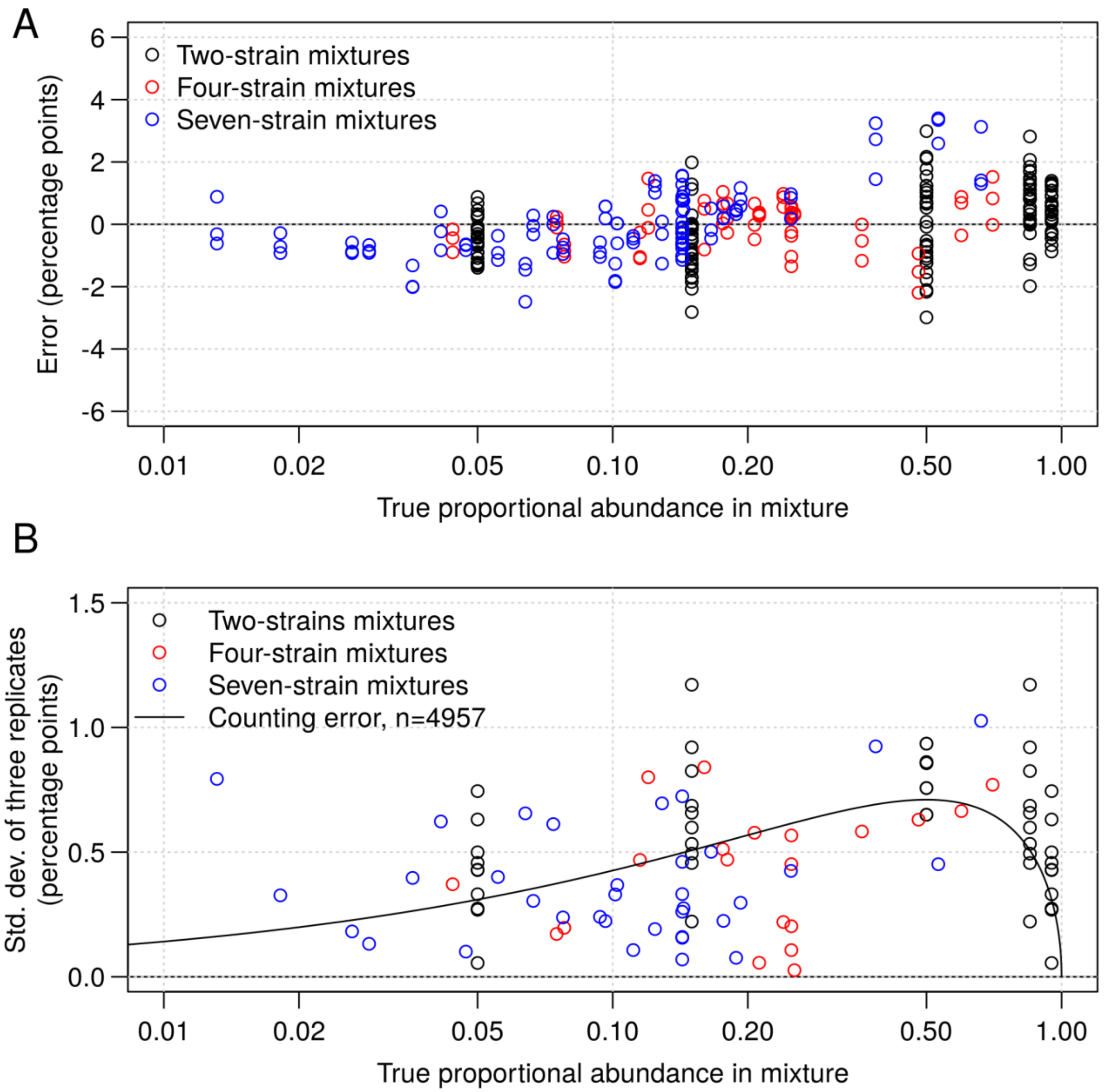

**Fig S5. CASEU error magnitude is only weakly dependent on strain abundance.** Related to Figures 2 and 3.

(A) Errors are similar in magnitude regardless of a strain's abundance, though there is some bias in the seven-strain mixtures at higher proportions (blue circles).

(B) Standard deviation of abundances calculated from triplicate Sanger sequencing measurements are generally around ~0.5% and are comparable to those that would be expected from counting based methods like next-generation sequencing or plate counts with *n*=4957 counts (best fit to $\sigma_{binomial} = \sqrt{\frac{p(1-p)}{n}}$).

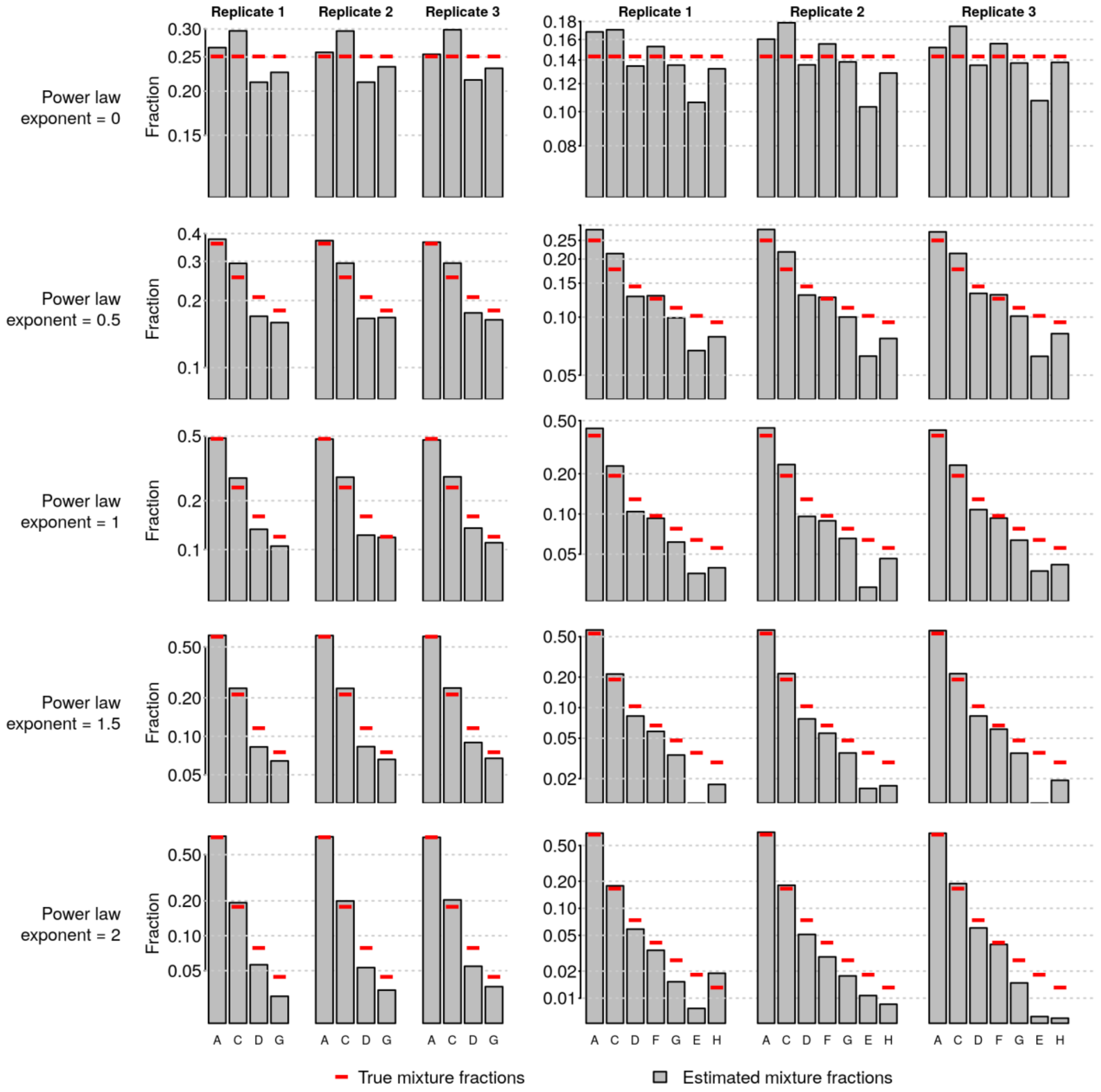

**Fig. S6. Without correcting for concentration errors in stock solutions, four- and seven-strain mixtures are consistent across replicates but biased.** Related to Figure 3. Data are the same as shown in Fig. 3 but without correcting for stock solution concentration errors. Solid bars are CASEU measurements after accounting for stock concentration error, whereas red lines show true mixture proportions based on power law distributions. In power law distributions, the abundance of the i$^{\text{th}}$ most abundant strain is proportional to $\frac{1}{i^{\alpha}}$ where α is the power law exponent.

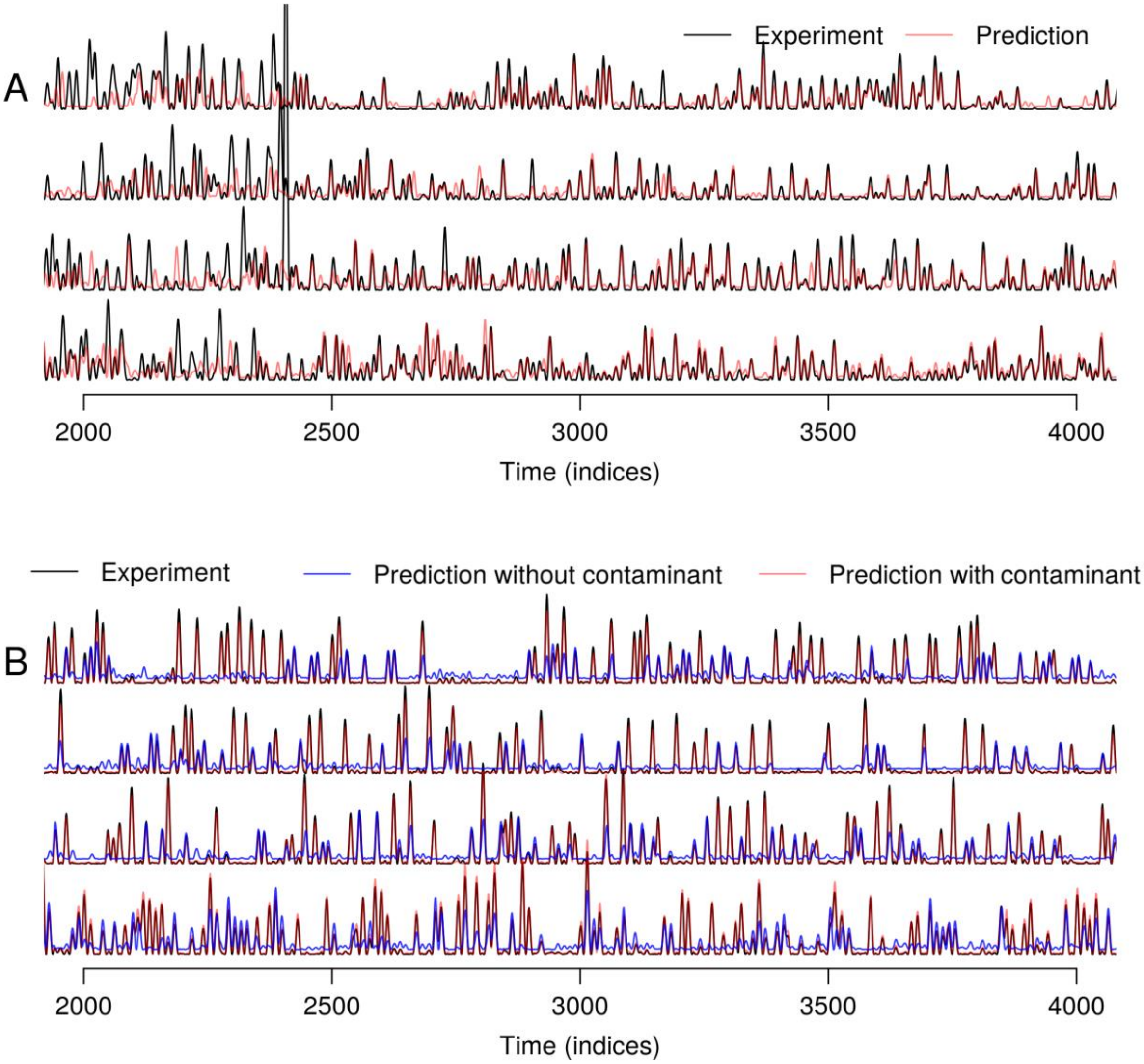

**Fig S7. Assessing CASEU fits to detect sample preparation and/or sequencing errors.** Related to Figure 5.

(A) The CASEU fit to a sample (red lines) does not accurately reproduce the observed mixed electropherogram (black lines). Each of the four traces shows a single fluorescence channel of the electropherogram. The best fit remained poor even when excluding the spike around t=2400.

(B) The CASEU fit to a contaminated community (blue) did not accurately reproduce the observed mixed electropherogram (black). After including the contaminating strain (identified by Illumina sequencing), the CASEU fit (red) reproduces the observed electropherogram.

E) Histogram of correlations between experiment and prediction for the model communities (n=39), showing two clear outliers (the communities shown in (A) and (C)).

```
Strain A    11    NNNNNNNNTANNNNTGNNNGTCGAGCGGAACGACAACATTGAATCTTCG
Strain B    11    GNNNNNNNTANNCNTGCAG-TCGAGCGGAACGACACTAACAATCCTTCG
Strain C    11    NNGGCNNNNACACATGCAG-TCGAGCGGAACGAGAATAG-----CTT--
                   *   *** *    **    ************* *  *      ***

Strain A    61    GAGGATTTGTTGGGCGTCGAGCGGCGGACGGGTGAGTAATGCCTAGGAA
Strain B    61    GGTGCGTTAATGGGCGTCGAGCGGCGGACGGGTGAGTAATGCCTAGGAA
Strain C    61    -----GCTATTCGGCGTCGAGCGGCGGACGGGTGAGTAATGCCTGGGAA
                         *  * ******************************** ****
```

**Table S1. Strain C possesses a deletion near the beginning of the gene relative to strains A and B.** Related to Figure 2. Truncated multiple sequence alignment (using Clustal 2.1) of sequences for strains A, B and C. Outside of this region, no gaps were found in the alignment. The gap in strain C begins roughly 50 bases after the end of the sequencing primer (27F) and leads to an offset of 12 bases for the remainder of the electropherogram.