JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 18, Number 11, 2011
© Mary Ann Liebert, Inc.
Pp. 1723–1741
DOI: 10.1089/cmb.2011.0189

Bacterial Community Reconstruction Using Compressed Sensing

*AMNON AMIR[1] and *OR ZUK[2]

ABSTRACT

Bacteria are the unseen majority on our planet, with millions of species and comprising most of the living protoplasm. We propose a novel approach for reconstruction of the composition of an unknown mixture of bacteria using a single Sanger-sequencing reaction of the mixture. Our method is based on compressive sensing theory, which deals with reconstruction of a sparse signal using a small number of measurements. Utilizing the fact that in many cases each bacterial community is comprised of a small subset of all known bacterial species, we show the feasibility of this approach for determining the composition of a bacterial mixture. Using simulations, we show that sequencing a few hundred base-pairs of the 16S rRNA gene sequence may provide enough information for reconstruction of mixtures containing tens of species, out of tens of thousands, even in the presence of realistic measurement noise. Finally, we show initial promising results when applying our method for the reconstruction of a toy experimental mixture with five species. Our approach may have a potential for a simple and efficient way for identifying bacterial species compositions in biological samples. All supplementary data and the MATLAB code are available at www.broadinstitute.org/ ~orzuk/publications/BCS/.

Key words: algorithms, genomics, machine learning, sequence analysis, sequences, statistics.

1. INTRODUCTION

Microorganisms are present almost everywhere on earth. The population of bacteria found in most natural environments consists of multiple species, mutually affecting each other, and creating complex ecological systems (Keller and Zengler, 2004). In the human body, the number of bacterial cells is over an order of magnitude larger than the number of human cells (Savage, 1977), with typically several hundred species identified in a given sample taken from humans (e.g., over 400 species were characterized in the human gut [Eckburg et al., 2005] while Sears [2005] estimates a higher number of 500–1000, and 500–600 species were found in the oral cavity [Dewhirst et al., 2008; Paster et al., 2001]). Changes in the human bacterial community composition are associated with physical condition and may indicate (Mager et al., 2005)—as well as cause or prevent—various microbial diseases (Guarner and Malagelada, 2003). In a

[1] Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel.
[2] Broad Institute, MIT and Harvard, Cambridge, Massachusetts.
*These authors made an equal contribution to this article.

broader aspect, the study of bacterial communities is a highly active field of research (Medini et al., 2008; Wooley et al., 2010), with goals ranging from understanding the interactions between microorganisms and their plant (Singh et al., 2004) or mammalian (Faith et al., 2011; Muegge et al., 2011) hosts, to spatial, temporal, and meteorological effects on the composition and diversity of microorganisms in urban aerosols (Brodie et al., 2007) and marine environments (Rusch et al., 2007).

Identification of the bacteria present in a given sample is not simple, and technical limitations impede large scale quantitative surveys of bacterial community compositions. Since the vast majority of bacterial species are non-amenable to standard laboratory cultivation procedures (Amann et al., 1995), culture-independent methods are needed. The golden standard of microbial population analysis has been cloning and direct Sanger sequencing of the ribosomal 16S subunit gene (16S rRNA) (Hugenholtz, 2002). However, since each 16S rRNA sequence is sampled randomly from the mixture, the sensitivity of this method is determined by the number of sequencing reactions, and therefore tens to hundreds of sequencing reactions are required for each sample analyzed. A modification of this method for identification of small mixtures of bacteria using a single Sanger sequence has been suggested (Kommedal et al., 2008) and showed promising results when reconstructing mixtures of 2–3 bacteria from a given database of ∼260 human pathogen sequences.

Recently, DNA microarray-based methods (Gentry et al., 2006) and identification via next generation sequencing (Hamady and Knight, 2009) have been used for bacterial community reconstruction. In microarray-based methods, such as the Affymetrix PhyloChip platform (Brodie et al., 2007), the sample 16S rRNA is hybridized with short probes aimed at identification of known microbes at various taxonomy levels. While being more sensitive and cheaper than standard cloning and sequencing techniques, each bacterial mixture sample still needs to be hybridized against a microarray, thus the cost of such methods limit their use for wide-scale studies. Methods based on next generation sequencing obtain a very large number of reads of a short hyper-variable region of the 16S rRNA gene (Armougom and Raoult, 2008; Dethlefsen et al., 2008; Hamady et al., 2008). Usage of such methods, combined with DNA barcoding, enables high-throughput identification of bacterial communities, and can potentially detect species present at very low frequencies. However, since such sequencing methods are limited to relatively short read lengths (typically a few dozens and at most a few hundred bases in each sequence), species identification is not straightforward. In practice, identification using current methods is nonunique and limited in resolution, with reliable identification typically up to the genus level (Huse et al., 2008). Improving resolution depends on obtaining longer read lengths, which is currently technologically challenging, and/or developing novel analytical methods which utilize the (possibly limited) information from each read to allow in aggregate a better separation between the species.

In this work we suggest a novel experimental and computational approach for sequencing-based profiling of bacterial communities (Fig. 1). We demonstrate our method using a single Sanger sequencing reaction

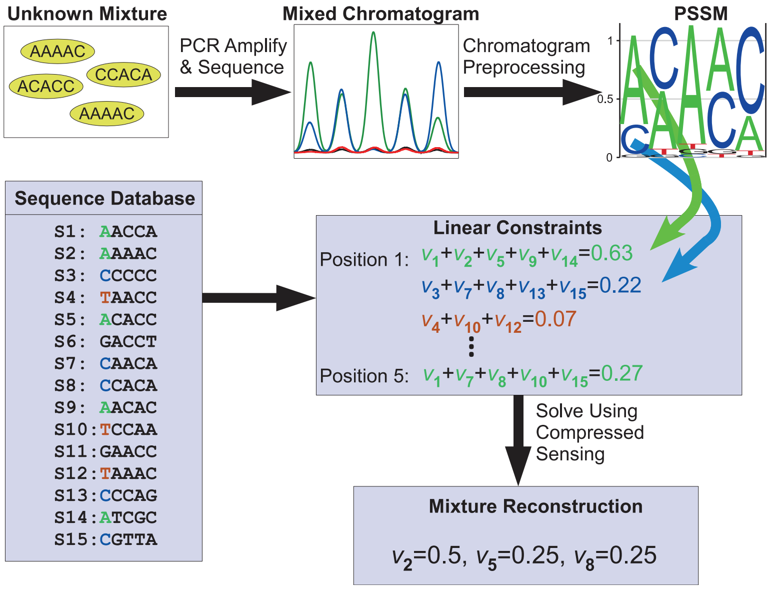

FIG. 1. Schematics of the proposed BCS reconstruction method. The 16S rRNA gene is PCR-amplified from the mixture and then subjected to Sanger sequencing. The resulting chromatogram is preprocessed to create the Position Specific Score Matrix (PSSM). For each sequence position, four linear mixture equations are derived from the 16S rRNA sequence database, with v_i denoting the frequency of sequence i in the mixture, and the frequency sum taken from the experimental PSSM. These linear constraints are used as input to the **CS** algorithm, which returns the sparsest set of bacteria recreating the observed PSSM.

for a bacterial mixture, which results in a linear combination of the constituent sequences. Using this mixed chromatogram as linear constraints, the sequences which constitute the original mixture are selected using a Compressed Sensing (**CS**) framework.

Compressed Sensing (**CS**) (Candes, 2006; Donoho, 2006a) is an emerging field of research, based on statistics and optimization, with a wide variety of applications. The goal of **CS** is recovery of a signal from a small number of measurements, by exploiting the fact that many natural signals are in fact sparse when represented at a certain appropriate basis. **CS** designs sampling techniques that condense the information of a compressible signal into a small amount of data. This offers the possibility of performing fewer measurements than previously appreciated, thus lowering costs and simplifying data-acquisition methods for various types of signals in many distantly related fields such as magnetic resonance imaging (Lustig et al., 2007), single pixel camera (Duarte et al., 2008), geophysics (Lin and Herrmann, 2007), and astronomy (Bobin et al., 2008). Recently, **CS** has been applied to various problems in computational biology, for example, for pooling designs for re-sequencing experiments (Erlich et al., 2010; Shental et al., 2010), for drug-screenings (Kainkaryam and Woolf, 2009), and for designing multiplexed DNA microarrays (Dai et al., 2009), where each spot is a combination of several different probes.

The classical **CS** problem is solving the under-determined linear system,

$$\mathscr{A}\mathbf{v} = \mathbf{b} \tag{1}$$

where $\mathbf{v} = (v_1, \ldots, v_N)$ is the vector of unknown variables, $\mathscr{A}$ is the *sensing* matrix, often called also the *mixing* matrix, and $\mathbf{b} = (b_1, \ldots, b_k)$ are the measured values of the k equations, with the number of variables N far greater than the number of equations k. Without further information, $\mathbf{v}$ cannot be reconstructed uniquely since the system is under-determined. Here one uses an additional sparsity assumption on the solution—by assuming that the solution vector $\mathbf{v}$ has at most s non-zero entries, for some $s \ll N$. According to the **CS** theory, when the matrix $\mathscr{A}$ satisfies certain conditions, one can find the sparsest solution uniquely by using a number of equations, $k = O(s \log(N/s))$, which is only logarithmic in the number of unknowns N, instead of a linear number (N) needed for general solution of a linear system. One notable such sufficient condition on the matrix $\mathscr{A}$ is the Restricted Isometry Property (RIP) (Candes and Tao, 2005; Candes et al., 2006). Briefly, RIP for a matrix $\mathscr{A}$ means that any subset of $2s$ columns of $\mathscr{A}$ is ''almost orthogonal'' (although since $k < N$, the columns cannot be perfectly orthogonal). This property makes the matrix $\mathscr{A}$ ''invertible'' for sparse vectors v with sparsity s, and allows accurate recovery of $\mathbf{v}$ from eq. (1) (Candes and Tao, 2005; Candes et al., 2006).

In this article, we show an efficient application of a single Sanger-sequencing for bacterial communities reconstruction using **CS**. The sparsity assumption is fulfilled by noting that although numerous species of bacteria have been characterized and are present on earth, at a given sample typically only a small fraction of them are present at significant levels.

The proposed Bacterial Compressed Sensing (**BCS**) algorithm uses as inputs a database of known 16S rRNA sequences and a single Sanger-sequence of the unknown mixture, and returns the sparse set of bacteria present in the mixture and their predicted frequencies. We show a successful reconstruction of simulated mixtures containing dozens of bacterial species out of a database of tens of thousands, using realistic biological parameters. In addition, we demonstrate the applicability of our method for a real sequencing experiment using a toy mixture of five bacterial species.

2. METHODS

2.1. The BCS algorithm

In the Bacterial Community Reconstruction Problem, we are given a bacterial mixture of unknown composition. In addition, we have at hand a database of the orthologous genomic sequences for a specific known gene, which is assumed to be present in a large number of bacterial species (in our case, the gene used was the 16S rRNA gene). Our purpose is to reconstruct the identity of species present in the mixture, as well as their frequencies, where the assumption is that the sequences for the gene in all or the vast majority of species present in the mixture are available in the database. The input to the reconstruction algorithm is the measured Sanger sequence of the gene in the mixture (Fig. 1). Since Sanger sequencing proceeds independently for each DNA molecule present in the sample, the sequence chromatogram of the mixture corresponds to the linear combination of the constituent sequences, where the linear coefficients are proportional to the abundance of each species in the mixture.

Let N be the number of known bacterial species present in our database. Each bacterial population is characterized by a vector $\mathbf{v} = (v_1, \ldots, v_N)$ of frequencies of the different species. Denote by $s = \|\mathbf{v}\|_{\ell 0}$ the number of species present in the sample, where $\| \cdot \|_{\ell 0}$ is the ℓ_0 norm which simply counts the number of non-zero elements of a vector $||\mathbf{v}||_{\ell 0} = \sum_i 1_{\{v_i \neq 0\}}$. While the total number of known species N is usually very large (in our case, on the order of tens to hundreds of thousands), a typical bacterial community consists of a small subset of the species, and therefore in a given sample, $s \ll N$, and $\mathbf{v}$ is a sparse vector. We denote the database sequences by a matrix S, where S_{ij} is the j'th nucleotide in the orthologous sequence of the i'th species ($i = 1, .., N, j = 1, .., k$).

We represent the results of the mixture Sanger sequencing as a $4 \times k$ Position-specific-Score-Matrix (PSSM)

$$P = \begin{pmatrix} a_1 & a_2 & \cdots & a_k \\ c_1 & c_2 & \cdots & c_k \\ g_1 & g_2 & \cdots & g_k \\ t_1 & t_2 & \cdots & t_k \end{pmatrix} = \begin{pmatrix} \mathbf{a} \\ \mathbf{c} \\ \mathbf{g} \\ \mathbf{t} \end{pmatrix} \tag{2}$$

where $\mathbf{p}_j = (a_j, c_j, g_j, t_j)^t$ is a column vector representing the observed frequencies at sequence position j of the four nucleotides, with $a_j, c_j, g_j, t_j \geq 0$.

Each position in the mixed sequence gives information about the bacterial composition of the mixture. For example, if at a certain position j, the frequency a_j of "A" in the mixed sequence is 0, and assuming no measurement noise is present, it follows that all bacteria which have "A" at the j'th position of their orthologous gene are not present in the mixture, and their corresponding frequencies in the solution vector must be zero. More generally, the frequency of each nucleotide at a given position j gives a linear constraint on the mixture:

$$\sum_{i=1}^{N} v_i 1_{\{S_{ij} = \text{``}A\text{''}\}} = a_j \tag{3}$$

and similarly for the nucleotides "C", "G", and "T". We next define the $k \times N$ mixture matrix A for the nucleotide "A",

$$A_{ij} = \begin{cases} 1 & S_{ij} = \text{``}A\text{''} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and similarly for the nucleotides "C", "G", and "T". The constraints given by the sequencing reaction can therefore be expressed in matrix form as:

$$A\mathbf{v} = \mathbf{a}, C\mathbf{v} = \mathbf{c}, G\mathbf{v} = \mathbf{g}, T\mathbf{v} = \mathbf{t} \tag{5}$$

The crucial assumption we make in order to cope with the insufficiency of information is the sparsity of the vector $\mathbf{v}$, which reflects the fact that only a small number of species are present in the mixture. We therefore seek a sparse solution for the set of equations (5). **CS** theory shows that under certain conditions on the mixture matrix and the number of measurements, the sparse solution can be recovered uniquely by solving the following minimization problem (Candes and Tao, 2006; Donoho, 2006b; Tropp, 2006),

$$\mathbf{v}^* = \underset{\mathbf{v}}{argmin}\, ||\mathbf{v}||_{\ell 1} = \underset{\mathbf{v}}{argmin} \sum_{i=1}^{N} |v_i| \quad s.t. \quad A\mathbf{v} = \mathbf{a}, C\mathbf{v} = \mathbf{c}, G\mathbf{v} = \mathbf{g}, T\mathbf{v} = \mathbf{t} \tag{6}$$

which is a convex optimization problem whose solution can be obtained in polynomial time. The above formulation requires our measurements to be precisely equal to their expected value based on the species frequency and the linearity assumption for the measured chromatogram. This description ignores the effects of noise, which is typically encountered in practice, on the reconstruction. Measurements of the signal mixtures suffer from various types of noise and biases. Fortunately, the **CS** paradigm is known to be robust to measurement noise (Candes and Tao, 2007; Candes et al., 2006). One can cope with noise by enabling a trade-off between sparsity and accuracy in the reconstruction merit function, which in our case is formulated as,

$$\mathbf{v}^* = \underset{\mathbf{v}}{argmin}\frac{1}{2}\left(||\mathbf{a}-A\mathbf{v}||^2_{\ell 2}+||\mathbf{c}-C\mathbf{v}||^2_{\ell 2}+||\mathbf{g}-G\mathbf{v}||^2_{\ell 2}+||\mathbf{t}-T\mathbf{v}||^2_{\ell 2}\right)+\tau||\mathbf{v}||_{\ell 1} \quad (7)$$

This problem represents a more general form of eq. (6) and accounts for noise in the measurement process. This is utilized by insertion of an ℓ2 quadratic error term. The parameter τ determines the relative weight of the error term vs. the sparsity promoting term, with higher levels of τ leading to a sparser solution. Many algorithms which enable an efficient solution of problem (7) are available, and we have chosen the widely used GPSR algorithm (Figueiredo et al., 2007). The error tolerance parameter was set to $\tau = 10$ for the simulated mixture reconstruction, and $\tau = 100$ for the reconstruction of the experimental mixture. These values achieved a rather sparse solution in most cases (a few species reconstructed with frequencies above zero), while still giving a good sensitivity—our ability to identify correctly species present in the mixture is not compromised significantly. The performance of the algorithm was quite robust to the specific value of τ used, and therefore further optimization of the results by fine tuning τ was not followed in this study.

2.2. *Ribosomal DNA database*

16S rRNA gene sequences were obtained from grenegenes (greengenes.lbl.gov) using database version 06-2007 (DeSantis et al., 2006), which contains approximately 136, 000 chimera checked full-length sequences. Sequences were reverse complemented and aligned with primer 1510R (Gao et al., 2007), resulting in approximately 42,000 sequences matching the primer sequence (with up to six mismatches with the primer). Out of this set, sequences with up to two base-pair difference with another sequence in the database were removed, resulting in $N = 18,747$ unique sequences which were used in this study. This last step was used for two purposes: first, to unite closely related species and enable a coarser identification of species in the mixture (since the information provided by the sequencing may not suffice to distinguish between very close species), and second, to reduce input size to the GPSR algorithm, thus making the **CS** problem more computationally feasible.

We manually added the sequence of *Enterococcus faecalis* (ATCC no. 19433) to the unique sequences list, as it was used in the experimental mixture but did not appear in the database (closest database species has 32 different positions).

2.3. *Experimental mixture reconstruction*

2.3.1. Sample preparation. We used the following strains for the experimental reconstruction: *Escherichia coli W3110, Vibrio fischeri, Staphylococcus epidermidis* (ATCC no. 12228), *Enterococcus faecalis* (ATCC no. 19433), and *Photobacterium leiognathi*. We obtained the 16S rRNA gene from each bacterial strain by boiling for one minute followed by 40 cycles of PCR amplification. Primers used for the PCR were the universal primers 8F and 1510R (Gao et al., 2007), amplifying positions 8–1513 of the *E. coli* 16S rRNA,

8F: 5′-AGAGTTTGATYMTGGCTCAG
1510R: 5′-TACGGYTACCTTGTTACGACTT

For mixture preparation and sequencing, we mixed together equal amounts of DNA from each bacterial 16S rRNA gene, and then sequenced them using an ABI3730 DNA Analyzer (Applied Biosystems, USA) with the 1510R primer.

2.3.2. Preprocessing steps. The input to the **BCS** algorithm is a $4 \times k$ PSSM $(\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{g}, \boldsymbol{t})^t$ of the mixture. However, obtaining this PSSM from an experimental mixture is not trivial. The output of a Sanger-sequencing reaction is a chromatogram, which describes the fluorescence of the four terminal nucleotides as a function of sequence position. In classical single-species sequencing, each peak in the chromatogram corresponds to a single nucleotide in the sequence. Identifying the peaks becomes more complicated when sequencing a mixture of different sequences. It has been previously shown (Bowling et al., 1991; Nickerson et al., 1997) that chromatogram peak height and position depend on the local sequence of nucleotides preceding a given nucleotide. Therefore, when performing Sanger sequencing of a mixture of multiple DNA sequences, the peaks of the constituent sequences may lose their coherence, making it nearly impossible to determine where the chromatogram peaks are located.

We therefore opted for a slightly different approach for chromatogram preprocessing, which does not depend on identifying the peak for each nucleotide. Rather, we bin the chromatogram into constant sized bins, and use the total intensity of each of the four nucleotides in each bin to construct the PSSM used as input to the **BCS** (see Fig. 9A below). A similar process is applied to each sequence in the 16S rRNA database. In order to correct for local-sequence effects, statistics were collected for local-sequence dependence of peak height and position. Similar statistics are used to obtain quality scores for single-sequence chromatogram base-calling in the Phred algorithm (Ewing and Green, 1998; Ewing et al., 1998). By utilizing these statistics, we predict the chromatogram for each sequence in the database, which is then binned and results in a PSSM for the single sequence.

Figure 10 below shows the inherent variation in chromatogram peak distances and heights, a substantial part of it is due to local sequence context, while Figure 11 below shows that considering these local sequence effects significantly improves our estimation of chromatogram peak locations and heights. The database of predicted PSSMs is then used to construct the mixing matrices A, C, G, T participating in the **BCS** problem representation (see eq. (7) and Fig. 9B below). We give further details on the chromatogram and database preprocessing steps in Appendices A and B, respectively.

3. RESULTS

3.1. Simulation results

In order to asses the performance of the proposed **BCS** reconstruction algorithm, random subsets of species from the greengene database (DeSantis et al., 2006) were selected. Within these subsets, the relative frequencies of each species were drawn at random from a uniform frequency distribution normalized to sum to one (results for a different, power-law frequency distribution, are shown later), and the mixture Sanger-sequence PSSM was calculated. This PSSM was then used as the input for the **BCS** algorithm, which returned the frequencies of database sequences predicted to participate in the mixture (Fig. 1).

A sample of a random mixture of 10 sequences, and a part of the corresponding mixed sequence PSSM, are shown in Figure 2A,B, respectively. Results of the **BCS** reconstruction using a 500 bp long sequence are shown in Figure 2C. The **BCS** algorithm successfully identified all of the species present in the original mixture, as well as several false positives (species not present in the original mixture). The largest false positive frequency was 0.01, with a total fraction of 0.04 false positives. In order to quantify the performance of the **BCS** algorithm, we used two main measures: RMSE and recall/precision. RMSE is the Root-Mean Squared-Error between the original mixture vector and the reconstructed vector, defined as $RMSE(\mathbf{v}, \mathbf{v}^*) = ||\mathbf{v} - \mathbf{v}^*||_{\ell 2} = (\sum_{i=1}^{N}(v_i - v_i^*)^2)^{1/2}$. This measure accounts both for the presence or absence of species in the mixture, as well as their frequencies. In the example shown in Figure 2, the RMSE score of the reconstruction was 0.03. As another measure, we have recorded the *recall*, defined as the fraction of species present in the original vector **v**, which were also present in the reconstructed vector **v*** (this is also known as sensitivity), and the *precision*, defined as the fraction of species present in the reconstructed

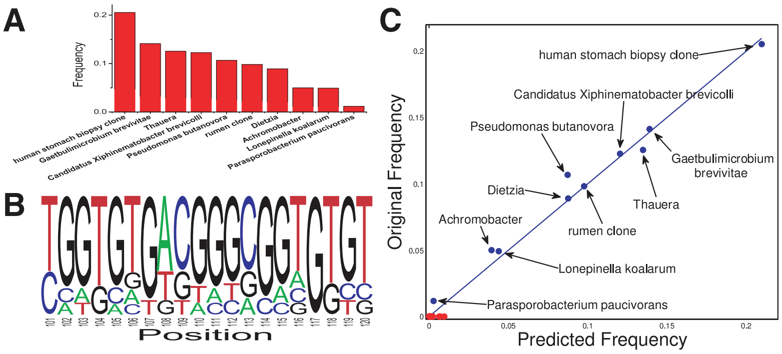

FIG. 2. Sample reconstruction of a simulated mixture. **(A)**. Frequencies and species for a simulated random mixture of $s = 10$ sequences. Species were randomly selected from the 16S rRNA database, with frequencies generated from a uniform distribution. **(B).** A 20 nucleotide sample region of the PSSM for the mixture in (A). **(C).** True vs. predicted frequencies for a sample **BCS** reconstruction for the mixture in (A) using $k = 500$ bases of the simulated mixture. Red circles denote species returned by the **BCS** algorithm which are not present in the original mixture.

vector **v***, which were also present in the original mixture vector **v**. Since the predicted frequency is a continuous variable, whereas the recall/precision relies on a binary categorization, a minimal threshold for calling a species present in the reconstructed mixture was used before calculating the recall/precision scores.

3.1.1. Coherence of database sequences. As detailed in the previous sections, the **CS** theory requires the columns of the mixing matrix to be incoherent, that is, close to orthogonal (e.g., satisfy the RIP condition [Candes, 2008]), in order to allow successful reconstruction using a small number of measurements. In our case, this cannot be achieved, as we were given the sequences determining the mixing matrix and cannot control them.

It has previously been shown (Ben-Haim et al., 2010, Tropp, 2006) that a computationally feasible method for assessing the information content of the mixing matrix is the mutual coherence, defined as the maximal coherence (inner product) between two columns of the mixing matrix. We therefore analyzed the empirical coherence distribution of sequences present in the current database. Even though the sequences are orthologous and thus quite similar, insertions and deletions came to our aid, as they bring similar sequences to being out of phase (e.g., even a deletion of a single base from a sequence, reduces its correlation with a copy of itself from one to a number typically much lower).

The distribution of coherence values for random pairs of database species is shown in Figure 3. While most correlations are centered around 0.25, there exists a small fraction of highly correlated sequences, with 0.005 of the sequence pairs showing a correlation above 0.8, and a maximal correlation value of 0.998. This high mutual coherence value places a limit on the reconstruction performance in the worst case, when such a sequence is present in the mixture. Since the database contains another highly similar sequence, distinguishing between these two is very difficult, and therefore the **CS** reconstruction cannot guarantee complete accuracy. However, given that such similar sequences are typically of closely related species (thus not being able to distinguish between them may be considered acceptable) and since most of the sequences show near random coherence, the reconstruction in most of the cases may still require only a small number of measurements (which translates into a small number of nucleotides read in the sequencing).

3.1.2. Effect of sequence length. To determine the typical sequence length required for reconstruction, we tested the **BCS** algorithm performance using different sequence lengths. In Figure 4A (black line), we plot the reconstruction RMSE for random mixtures of 10 species. To enable faster running times, each simulation used a random subset of $N = 5000$ sequences from the sequence database for mixture generation and reconstruction. It is shown in Figure 4A that using longer sequence lengths results in a larger number of linear constraints and therefore higher accuracy, with ~300 nucleotides sufficing for accurate reconstruction of a mixture of 10 sequences. The large standard deviation is due to a small

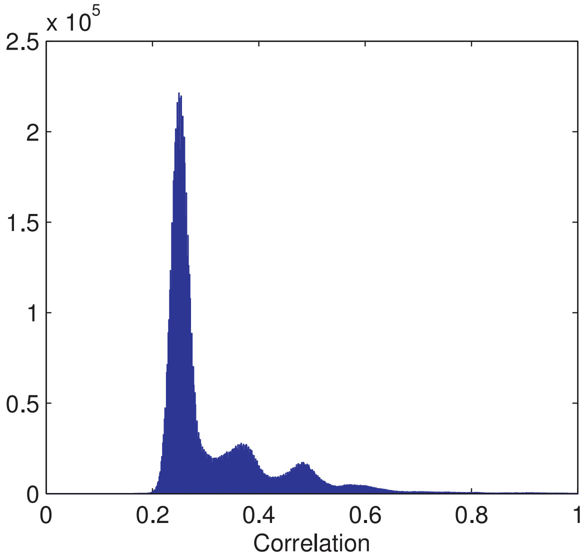

FIG. 3. Coherence distribution of the 16S rRNA sequences. Coherence (inner product) of 10^7 16S rRNA vector pairs chosen randomly from the sequence database (~5.7% of all possible pairs). As each column of the mixture matrix is a binary vector with 1/4 of the coordinates being one, the dot product between two randomly generated vectors is expected to be ~0.25. While most 16S rRNA database pairs exhibit a coherence around 0.25, many pairs exhibit significantly higher correlations, with a few (~0.5%) even exceeding 0.8 (see inset).

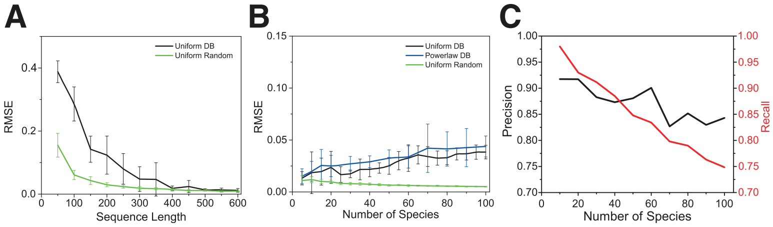

FIG. 4. Reconstruction of simulated mixtures. **(A).** Effect of sequence length on reconstruction performance. RMSE between the original and reconstructed frequency vectors for uniformly distributed random mixtures of $s = 10$ species from the 16S rRNA database (black) or randomly generated sequences (green). Error bars denote the standard deviation derived from 20 simulations. **(B).** Dependence of reconstruction performance on number of species in the mixture. Simulation is similar to (A) but using a fixed sequence length ($k = 500$) and varying the number of species in the mixture. Blue line shows reconstruction performance on a mixture with power-law distributed species frequencies ($v_i \sim i^{-1}$). **(C).** Recall (fraction of sequences in the mixtures identified, shown in red) and precision (fraction of incorrect sequences identified, shown in black) of the **BCS** reconstruction of uniformly distributed database mixtures shown as black line in (B). The minimal reconstructed frequency for a species to be declared as present in the mixture was set to 0.25%.

probability of selection of a similar but incorrect sequence in the reconstruction, which leads to a high RMSE. Due to a cumulative drift in the chromatogram peak position prediction, typical usable experimental chromatogram lengths are in the order of $k \sim 500$ bases rather than the ~ 1000 bases usually obtained when sequencing a single species.

In order to asses the effect of similarites between the database sequences (which leads to high coherence of the mixing matrix columns) on the performance of the **BCS** algorithm, a similar mixture simulation was performed using a database of random nucleotide sequences (i.e., each sequence was composed of i.i.d. nucleotides with 0.25 probability for ''A,'' ''C,'' ''G,'' or ''T''). Using a mixing matrix derived from these random sequences, the **BCS** algorithm showed better performance (green line in Fig. 4A), with ~ 100 nucleotides sufficing for a similar RMSE as that obtained for the 16S rRNA database using 300 nucleotides.

3.1.3. Effect of number of species. For a fixed value of $k = 500$ nucleotides per sequencing run, the effect of the number of species present in the mixture on reconstruction performance is shown in Figure 4B,C. Even on a mixture of 100 species, the reconstruction showed an average RMSE less than 0.04, with the highest false positive reconstructed frequency (i.e., frequency for species not present in the original mixture) being less than 0.01. Using a minimal frequency threshold of 0.0025 for calling a species present in the reconstruction, the **BCS** algorithm shows an average recall of 0.75 and a precision of 0.85. Therefore, while the sequence database did not perform as well as random sequences, the 16S rRNA sequences exhibit enough variation to enable a successful reconstruction of mixtures of tens of species with a small percent of errors.

The frequencies of species in a biologically relevant mixture need not be uniformly distributed. For example, the frequency of species found on the human skin (Gao et al., 2007) were shown to resemble a power-law distribution. We therefore tested the performance of the **BCS** reconstruction on a similar power-law distribution of species frequencies with with $v_i \sim i^{-1}$. Performance on such a power-law mixture is similar to the uniformly distibuted mixture (blue and green lines in Fig. 4B, respectively) in terms of the RMSE. A sample power-law mixture and reconstruction are shown in Figure 5A,B. The recall/precision of the **BCS** algorithm on such mixtures (Fig. 5C) is similar to the uniform distribution for mixtures containing up to 50 species, with degrading performance on larger mixtures, due to the long tail of low frequency species.

3.1.4. Effect of noise on BCS solution. Experimental Sanger sequencing chromatograms contain inherent noise, and we cannot expect to obtain exact measurements in practice. We therefore turned to study the effect of noise on the accuracy of the **BCS** reconstruction algorithm. Measurement noise was modeled as additive i.i.d. Gaussian noise $z_{ij} \sim N(0, \sigma^2)$ applied to each nucleotide read at each position. Noise is compensated for by the insertion of the $\ell 2$ norm into the minimization problem (see eq. (7)), where the factor τ determines the balance between sparsity and error-tolerance of the solution. The effect of added random i.i.d. Gaussian noise to each nucleotide measurement is shown in Figure 6. The reconstruction performance slowly degrades with added noise both for the real 16S rRNA and the random sequence database.

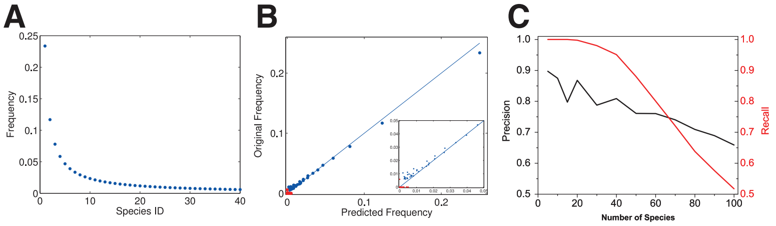

FIG. 5. Sample reconstruction of a power-law mixture. **(A)**. Sorted frequency distribution of 40 random species following a power-law distribution with frequencies $v_i \sim i^{-1}$, $i = 1,.., 40$. **(B).** True vs. predicted frequencies for a sample **BCS** reconstruction for the mixture in (A) using $k = 500$ bases of the simulated mixture. Red circles denote species returned by the **BCS** algorithm which are not present in the original mixture. **(C).** Average precision (black) and recall (red) for the reconstruction of simulated mixtures with power-law distributed frequencies as in (A). The minimal reconstructed frequency for a species to be declared as present in the mixture was set to 0.17%.

Using a noise standard deviation of $\sigma = 0.15$ (which is the approximate experimental noise level) and sequencing 500 nucleotides, the reconstruction performance as a function of the number of species in the mixture is shown in Figure 7. Under this noise level, the **BCS** algorithm reconstructed a mixture of 40 sequences with an average RMSE of 0.07 (Fig. 7B), compared to $\sim$0.02 when no noise is present (Fig. 4B). By using a minimal frequency threshold of 0.006 for the predicted mixture, **BCS** showed a recall (sensitivity) of $\sim$0.7, with a precision of $\sim$0.7 (Fig. 7B), attained under realistic noise levels. To conclude, we have observed that the addition of noise leads to a graceful degradation in the reconstruction performance, and one can still achieve accurate reconstruction with realistic noise levels.

3.2. *Reconstruction of an experimental mixture*

While these simulations show promising results, they are based on correctly converting the experimentally measured chromatogram to the PSSM used as input to the **BCS** algorithm (Fig. 1). A major problem in this conversion is the large variability in the peak heights and positions observed in Sanger sequencing chromatograms (see Fig. 10 below). It has been previously shown that a large part of this variability stems from local sequence effects on the polymerase activity (Lipshutz et al., 1994). In order to overcome this problem, we utilize the fact that both peak position and height are local sequence dependent, in order to accurately predict the chromatograms of the sequences present in the 16S rRNA database. The **CS** problem is then stated in terms of reconstruction of the measured chromatogram using a sparse subset of predicted chromatograms for the 16S rRNA database. This is achieved by binning both the predicted chromatograms and the measured mixture chromatogram into constant sized bins, and applying the **BCS** algorithm on these bins (Fig. 9).

We tested the feasibility of the **BCS** algorithm on experimental data by reconstructing a simple bacterial population using a single Sanger sequencing chromatogram. We used a mixture of five different bacteria: *Escherichia coli W3110, Vibrio fischeri, Staphylococcus epidermidis, Enterococcus faecalis*, and *Photobacterium leiognathi*. A sample of the measured chromatogram is shown in Figure 8A (solid lines). The **BCS** algorithm relies on accurate prediction of the chromatograms of each known database 16S rRNA sequence. In order to asses the accuracy of these predictions, Figure 8A shows a part of the predicted chromatogram of the mixture (dotted lines) which shows similar peak positions and heights to the ones experimentally measured (solid lines). The sequence position dependency of the prediction error is shown in Figure 8B. On the region of bins 125–700, the prediction shows high accuracy, with an average root

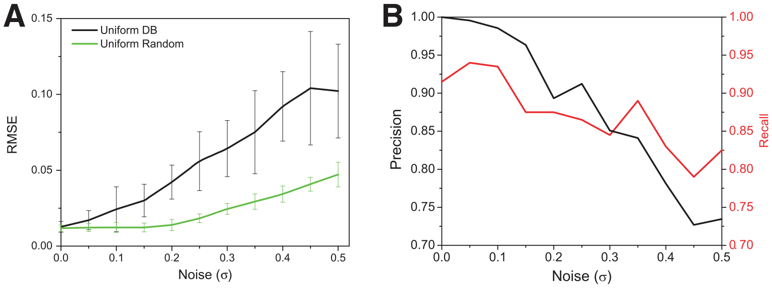

FIG. 6. Effect of noise on reconstruction. **(A).** Reconstruction RMSE of mixtures of $s = 10$ sequences of length $k = 500$ from the 16S rRNA sequence database (black) or random sequences (green), with Gaussian noise added to the chromatogram. **(B).** Recall (red) and precision (black) of the 16S rRNA database mixture reconstruction shown in (A).

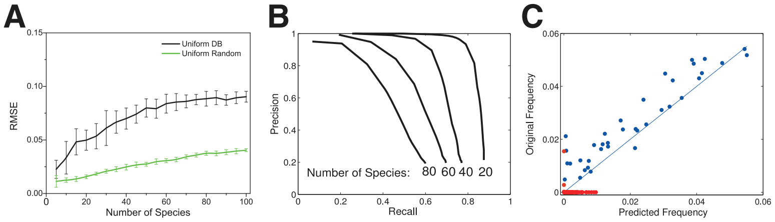

FIG. 7. Reconstruction with experimental noise level. **(A).** Reconstruction RMSE as a function of number of species present in the mixture. Frequencies were sampled from a uniform distribution. Noise is set to $\sigma = 0.15$. Sequence length is set to $k = 500$. Black and green lines represent 16S rRNA and random sequences respectively. **(B).** Recall vs. precision curves for different number of 16S rRNA sequences as in (A) obtained by varying the minimal inclusion frequency threshold. **(C).** Sample reconstruction of $s = 40$ 16S rRNA sequences from (A).

square error of 0.08. The loss of accuracy at longer sequence positions stems from a cumulative drift in predicted peak positions, as well as reduced measurement accuracy. We therefore used the region of bins 125–700 for the **BCS** reconstruction.

Results of the reconstruction are shown in Figure 8C. The algorithm successfully identifies three of the five bacteria (*Vibrio fischeri, Enterococcus faecalis*, and *Photobacterium leiognathi*). Out of the two remaining strains, one (*Staphylococcus epidermidis*) is identified at the genus level, and the other (*Escherichia coli*) is mistakenly identified as *Salmonella enterica*. While *Escherichia coli* and *Salmonella enterica* show a sequence difference in 33 bases over the PCR amplified region, only two bases are different in the region used for the **BCS** reconstruction, and thus the *Escherichia coli* sequence was removed in the database preprocessing stage. When this sequence is manually added to the database (in addition to the *Salmonella enterica* sequence), the **BCS** algorithm correctly identifies the presence of *Escherichia coli* rather than *Salmonella enterica* in the mixture. Another strain identified in the reconstruction—the Kennedy Space Center clone KSC6-79—is highly similar in sequence (differs in five bases over the region tested) to the sequence of *Staphylococcus epidermidis* used in the mixture.

4. DISCUSSION

In this work, we have proposed a framework for identifying and quantifying the presence of bacterial species in a given population using information from a single sequencing reaction. Simulation results with

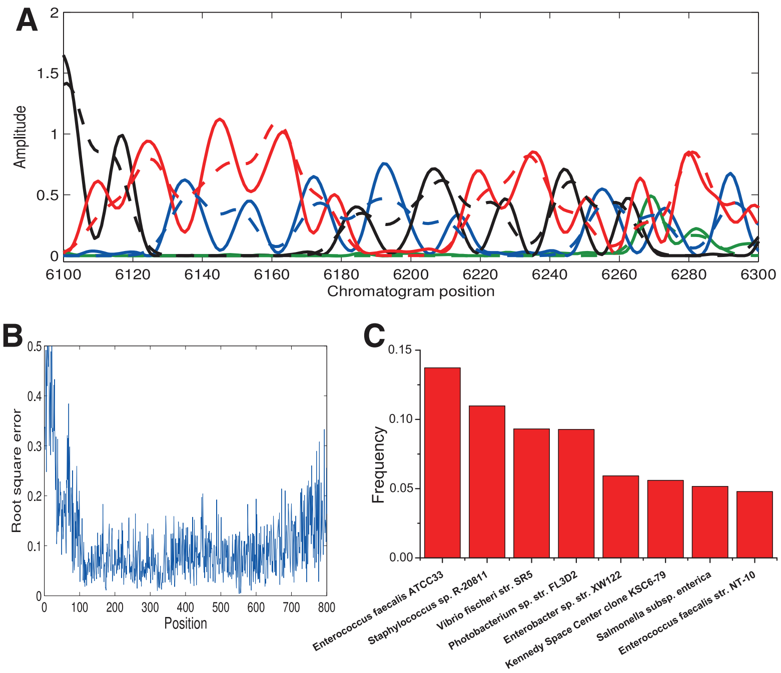

FIG. 8. Reconstruction of an experimental mixture. **(A).** Sample region of the mixed chromatogram (solid lines). 16S rRNA from five bacteria was extracted and mixed at equal proportions. Dotted lines show the local-sequence corrected prediction of the chromatogram using the known mixture sequences. **(B).** Square root distance between the predicted and measured chromatograms shown in (A) as a function of bin position, representing nucleotide position in the sequence. Prediction error was low for sequence positions ~100–700. **(C).** Reconstruction results using the **BCS** algorithm. Runtime was ~20 minutes on a standard PC. Shown are the 8 most frequent species. Original strains were: *Escherichia coli, Vibrio fischeri, Staphylococcus epidermidis, Enterococcus faecalis*, and *Photobacterium leiognathi* (each with 20% frequency).

noise levels comparable to the measured noise in chromatograms obtained experimentally for real sequence indicate that our method can reconstruct mixtures of tens of species. When not enough information is present in the sequence (e.g., when the number of sequences present in the mixture is large), performance of the reconstruction algorithm decays gracefully, and still retains detection of the prominent species.

In order to test the applicability of the **BCS** algorithm to real experimental data, we performed a reconstruction of a toy mixture containing five bacterial species. Results of the sample reconstruction (identification of three out of five species at the strain level, and the additional two at the genus level, when the *E. coli* sequence is included in the 16S rRNA database; see Section 3.2) indicate that with appropriate chromatogram preprocessing, **BCS** can be applied to experimental mixtures. However, further optimization of the sequencing and preprocessing is required in order to obtain more accurate results.

The amount of information needed for identifying the species present in the mixture is logarithmic in the database size (Candes, 2006; Donoho, 2006a), as long as the number of the species present in the mixture is kept constant. Therefore, a single sequencing reaction with hundreds of bases may in principle provide sufficient information for unique reconstruction even when the database contains millions of different sequences. Compressed Sensing enables the use of such information redundancy through the use of linear mixtures of the sample. However, coherence between the columns of the reconstruction matrix may hinder the reconstruction performance. The mixtures are dictated by the sequences in the database, which are exhibit a complicated dependence structure resulting from the phylogenetic relationships of both the species and the 16S rRNA gene. Since the mixing matrix is built using each sequence in the database separately, our method does not rely on correct alignment of the database sequences. While two sequences which differ in a few nucleotides have high coherence and clearly do not contribute to RIP, even a single insertion or deletion completely brings the two sequences to being ''out of phase,'' thus making it easier to distinguish between them using **CS** (provided that the insertion/deletion did not occur to close to the end of the sequenced region). In this case, a species actually present in the mixture is likely to appear in the solution vector with high frequency, whereas sequences of similar species that are different by one or a few insertion or deletion events will violate the linear constrains present in our optimization criteria, and are not likely to ''fool'' the reconstruction algorithm.

While limited to the identification of species with known 16S rRNA sequences, the **BCS** approach may enable low cost simple comparative studies of bacterial population composition in a large number of samples. The performance of our method (or any other method used for species identification) depends on the inherent inter-species variation in the sequenced region. At the most extreme scenario, if two species are completely identical at the 16S rRNA locus, no method will be able to distinguish between them based on this locus alone. In the simulations we presented, we defined a species reconstruction to be accurate when having up to 2 nucleotide difference from the original sequence. Since sequence lengths used were typically around 500bp, the reconstruction sequence accuracy was $<0.4\%$. Average sequence differences within genus has been previously measured to be approximately 3%, whereas within species is approximately 2% (Yarza et al., 2008). Therefore, one can interpret our simulation's performance as measuring reconstruction at sub-species resolution. However, there are a few cases of species with identical or nearly identical 16S rRNA sequences, and therefore these species cannot be discriminated based on 16S rRNA alone. Sequencing of additional loci (such as in the MLST database [Maiden et al., 1998]) is likely to achieve higher reconstruction resolution. Our proposed method can easily be extended to more than one sequencing reaction per mixture, whether they come from the same region or distinct regions, by simply joining all sequencing results as linear constraints. This increases the amount of information available for our reconstruction algorithm, which may enable both to overcome experimental noise present in each sequencing reaction, and to distinguish between closely related species more accurately and at a higher resolution.

5. APPENDIX

A: Chromatogram Preprocessing

The purpose of the chromatogram preprocessing scheme is to convert the raw measured chromatogram data to a PSSM representing the frequency of each base at each position along the sequence in the mixture (Fig. 9A). We provide below a formal algorithm sketch for this step, followed by a more detailed description:

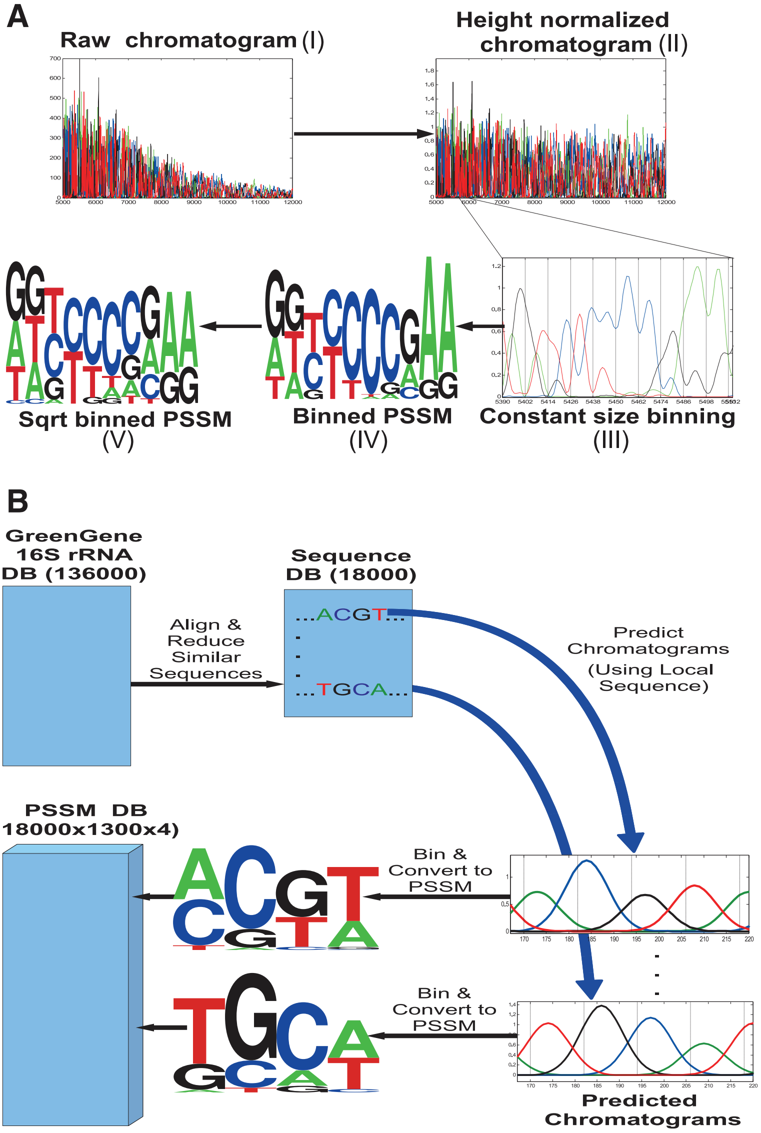

FIG. 9. Preprocessing steps. **(A)**. Preprocessing of the experimental chromatogram. The result of the Sanger-sequencing of a bacterial mixture (I) is normalized by division with a ~1000 pixel total intensity running average to compensate for the peak amplitude decrease. The resulting chromatogram (II) is binned into constant sized bins (sample section shown in III), and the resulting PSSM (sample section shown in IV) is further square-root transformed to obtain the final experimental PSSM (sample section shown in V). **(B).** Preprocessing of the 16S rRNA sequence database. Sequences are first aligned and similar sequences are removed. Then, a predicted chromatogram is generated for each sequence in the database, based on local sequence statistics collected from a training set. Finally, the predicted chromatograms are binned into constant sized binned and the resulting PSSMs are further square-root transformed similarly to (A), to produce the final PSSMs which are stored in the database.

Algorithm 1: Chromatogram preprocessing

Input: $(\mathfrak{a}, \mathfrak{c}, \mathfrak{g}, \mathfrak{t})$ - four fluorescent trace vectors (such as from an .abi or .scf file).
Output: $P = (\mathbf{a}, \mathbf{c}, \mathbf{g}, \mathbf{t})^t$—a PSSM representing nucleotide frequencies

1. Normalize the chromatogram amplitude:

$$\mathfrak{a}_p = \frac{50 \cdot 12 \cdot \mathfrak{a}_p}{\sum_{q=-25\cdot 12}^{25\cdot 12} (\mathfrak{a}_{p+q} + \mathfrak{c}_{p+q} + \mathfrak{g}_{p+q} + \mathfrak{t}_{p+q})} \tag{8}$$

and similarly for $\mathfrak{c}_p$, $\mathfrak{g}_p$ and $\mathfrak{t}_p$.

2. bin into constant sized bins, and apply a square root transformation:

$$a_j = \sqrt{\sum_{p=12j}^{12j+11} \mathfrak{a}_p}, \quad j = 1 \ldots k \tag{9}$$

and similarly for c_j,g_j,t_j.

The input to the chromatogram preprocessing is the measured chromatogram, consisting of four fluorescent trace vectors *a, c, g, t,* where for example a_p represents the signal intensity for nucleotide ''*A*'' at the *p*'s position along the chromatogram, where each position is represented by one pixel in the chromatogram image. The value *p* corresponds roughly to the timing of the sequencing reaction, with a resolution of approximately a dozen points per nucleotide, thus *p* runs from 1 to $\sim 12k$ (a few thousand points in a typical chromatogram—for example, 6900 points in the experimental chromatogram described in Section 3, corresponding to approximately 575 base-pairs).

In a typical Sanger sequencing reaction, the chromatogram peak heights decrease as the position *p* becomes higher (nucleotides further in the sequence which were sequenced later in the sequencing reaction) due to depletion of the dideoxynucleotides. To overcome this long-scale decrease in signal amplitude, prior to the binning step, the amplitude at each position was normalized by division with average total peak height in a ~ 50 base-pair (bp) region around each position (see step 1 in the algorithm description below).

The resulting vectors after the normalization step are binned into constant sized bins (12 pixels per bin), and the sum of intensity values of each bin is computed for the four different nucleotides. Then, we take square root of this sum for the four different nucleotides for the *i*'th bin as the *i*'th column in the output $4 \times k$ PSSM. The square root is used rather than the sum as this was shown to decrease the effect of large outliers. The resulting $4 \times k$ PSSM is used as input to the **BCS** reconstruction.

B: Database Preprocessing

The purpose of the Database Preprocessing scheme is to produce predicted PSSMs for all 16S rRNA sequences in the database (Fig. 9B). For each sequence S_i in the database sequences we compute a PSSM $P_i = (\mathbf{a}_i, \mathbf{c}_i, \mathbf{g}_i, \mathbf{t}_i)^t$. These predicted PSSMs are then used in the **BCS** reconstruction algorithm described in Section 2.1 as ''basis vectors.''

We use a generative model-based approach for simulating the measured PSSMs obtained from an (hypothetical) measured chromatogram for each sequence in the database. The model generates, as an intermediate stage, a predicted chromatogram for the input sequence. This chromatogram is then further processed to obtain the predicted PSSM. The model captures the relations between the sequence of nucleotides comprising a DNA molecule and the chromatogram charts obtained when sequencing such a molecule. The main factor affecting the chromatogram shape is local-sequence context, which we modeled by a 5th order Markov Chain. We fitted the model parameters in a preliminary step by using a training set of sequences with experimentally available chromatograms. We describe this preliminary step in the next section. Next, in the Database PSSMs Generation Step Section, we describe the database preprocessing step performed once the model parameters are fully specified.

Preliminary step: Compute local-sequence adjusted chromatogram statistics. The preliminary step fits model parameters representing context-specific peak height and width, as well as a sequence-position dependent correction factor β used to model variation in peak position. We give a formal algorithm sketch followed by a detailed description:

Algorithm 2: Compute local-adjusted chromatogram statistics

Input: A set of training chromatograms given as $(\mathfrak{a}_i, \mathfrak{c}_i, \mathfrak{g}_i, \mathfrak{t}_i)$ - four fluorescent trace vectors for the *i*-th sequence in the training set.

Output: *H*, *D* - tables of size $4^6 = 4196$ of context-specific peak heights and distances, respectively. β - position-dependent peak-peak distance parameter.

1. Determine S' - the set of nucleotide sequences of the input chromatograms, where S'_i determined by applying the standard ABI base-caller on the *i*-th input chromatogram.
2. Determine chromatogram peak positions $p_{i,j}$ for each sequence *i* and position along the sequence *j* using the standard ABI base-caller. Determine chromatogram peak heights $h_{i,j}$ as the peak height of the trace corresponding to the base $S'_{i,j}$ returned by the ABI base-caller at position $p_{i,j}$.
3. Normalize peak heights $h_{i,j}$ by applying local height correction similarly to step 1 of the chromatogram preprocessing algorithm (see Algorithm 1).
4. Compute context-specific peak height averages: for a given k-mer $\alpha = (\alpha_1, \ldots, \alpha_6)$, compute the averaged peak heights of all the occurrences of α as a k-mer in all training set sequences:

$$\frac{H(\alpha) = \sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, \ldots, \alpha_6 = S'_{i,j}\}} h_{i,j}}{\sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, \ldots, \alpha_6 = S'_{i,j}\}}} \tag{10}$$

5. Compute the relative peak-peak distance for each position,

$$d_{i,j} = \frac{p_{i,j} - p_{i,j-1}}{\sum_{j=2}^{k} p_{i,j} - p_{i,j-1}}. \tag{11}$$

6. Compute context-specific peak distance averages $D(\alpha)$ for each k-mer α by measuring the average relative peak-peak distance between current and previous peaks:

$$D(\alpha) = \frac{\sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, \ldots, \alpha_6 = S'_{i,j}\}} d_{i,j}}{\sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, \ldots, \alpha_6 = S'_{i,j}\}}} \tag{12}$$

7. Fit a position-based linear model for peak distance $d_{i,j}$:

$$d_{i,j} = \gamma + \beta j \tag{13}$$

and output the linear coefficient β.

In the course of the Sanger sequencing process, both the polymerase specificity for incorporating deoxynucleotides over dideoxynucleotides and the fragment mobility depend on sequence local to the incorporation point. Therefore for each nucleotide in the DNA fragment being sequenced, its corresponding chromatogram peak height and position are affected by the preceding nucleotides (Lipshutz et al., 1994). In order to predict and correct for the effect of local sequence context on the resulting chromatogram, we collected statistics from a training set S' of 1000 sequencing runs performed on an ABI3730 machine. Runs were randomly selected from experiments submitted for sequencing in the Weizmann Institute sequencing unit by various labs. The average length of the runs was approximately 800 base-pairs, providing in total chromatogram statistics for $\sim$800,000 nucleotides. Chromatogram heights $h_{i,j}$ were normalized to overcome the long-scale amplitude decrease (as described in Appendix A).

We have modeled the local sequence context by looking at the 5 nucleotides preceding each nucleotide, giving us $4^6 = 4096$ different unique 6-mers, each representing a possible nucleotide and the five nucleotides preceding it. For each unique 6-mer, the fitting step searches for all of its occurrences in S', and averages the peak height and position data of the last nucleotide over all such occurrences in S'. We have used 6-mers as this gives the maximal context length for which we had sufficient statistics to collect for each 6-mer. Approximately 200 instances were available per 6-mer on average, with a minimal number of 16 instances for one 6-mer (it is possible that a smaller local sequence context is sufficient for accurate prediction of chromatogram heights).

The resulting final peak height and distance tables H and D, respectively, each of size $4096(= 4^6)$, are available at www.broadinstitute.org/norzuk/publications/BCS/. While the average height and position were 1 (as was ensured by our normalizations), there was significant variability in height and position according to sequence context, with height values H typically in the range of $\sim$0.5–1.3 and position values P in the range of $\sim$0.8–1.2 (Fig. 10). We observed an additional sequence-independent change in peak-peak distance in the chromatograms studied, where distance between consecutive peaks increases as we move further along the chromatogram. We accounted for this by fitting an additional linear model based only on the nucleotide position along the sequence according to eq. (13), giving an additional parameter of $\beta = 0.00036$ representing increase in peak-peak distance with each position - we used this parameter later to predict the resulting chromatograms (see eq. (15) in Algorithm 3).

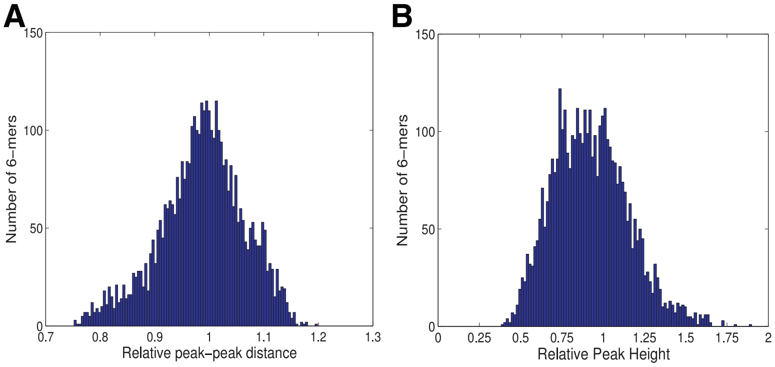

FIG. 10. Local-sequence effect on chromatogram peak height and position. **(A).** Distribution of average normalized peak-peak distances for the 4096 sequence 6-mers. **(B).** Distribution of normalized peak heights for the 4096 sequence 6-mers. Both distributions show a rather wide spread around one, showing that local sequence context has a significant effect on peak height and position.

Database PSSMs Generation step: Generate a database of predicted PSSMs. This step generates predicted PSSMs for the N sequences in the 16S rRNA database S. It uses the model parameters from the training set described in the previous Section. The sequence input S and model parameters are used to determine peak heights and positions and thus compute a set of N chromatograms of the form (a, c, g, t), one for each 16S rRNA gene sequence in S. We then further process these chromatograms to get predicted PSSMs. This step is illustrated in Figure 9B. We give the details next in Algorithm 3:

Algorithm 3: Compute PSSM from chromatogram

Input: S - a set of 16S rRNA sequences from the database with maximal length denoted by k. H, D - context-specific chromatogram peak height and distance tables. β - position-dependent peak distance parameter.

Output: A set of PSSMs $P_i = (\mathbf{a}_i, \mathbf{c}_i, \mathbf{g}_i, \mathbf{t}_i)^t$, one for each sequence S_i in the database.

1. For every nucleotide $S_{i,j}$ in the database, estimate it's chromatogram peak height:

$$a_{i,j} = H(L_{i,j}) \tag{14}$$

where $L_{i,j} = (S_{i,j-5}, \ldots, S_{i,j})$ denotes the local 6-mer sequence context of nucleotide j in the i-th gene sequence ($L_{i,j} \in 1 \ldots 4^6$).

2. For every nucleotide $S_{i,j}$ in the database, estimate it's chromatogram peak position as:

$$b_{i,j} = b_{i,j-1} + D(L_{i,j}) + \beta \cdot j \tag{15}$$

3. For every nucleotide $S_{i,j}$ in the database, create a corresponding peak in the chromatogram using a Gaussian peak function:

$$f_{i,j}(x) = a_{i,j} e^{-\frac{(x-b_{i,j})^2}{2c^2}} \tag{16}$$

where x is sampled in a range [0, k] at a resolution of 1/12 thus giving $12k$ different x values $x_1, \ldots, x_{12k}$ and their corresponding $f_{i,j}$ values.

4. For each sequence compute the four chromatogram trace vectors. The trace vector for nucleotide ''A'' for the i-th sequence is computed as:

$$a_{i,p} = \sum_j f_{i,j}(x_p) 1_{\{S_{i,j} = \text{'}A\text{'}\}} \tag{17}$$

and similarly for the other three nucleotides ('C','G','T').

5. Bin trace vectors to obtain final PSSMs: The four predicted chromatograms trace vectors ($\mathfrak{a}_i$, $\mathfrak{c}_i$, $\mathfrak{g}_i$, $\mathfrak{t}_i$) are binned using a constant bin size of 1 and transformed via square root, according to the chromatogram preprocessing step in eq. (9), to give a PSSM $P_i = (\mathbf{a}_i, \mathbf{c}_i, \mathbf{g}_i, \mathbf{t}_i)^t$ for each 16S rRNA sequence S_i in the database.

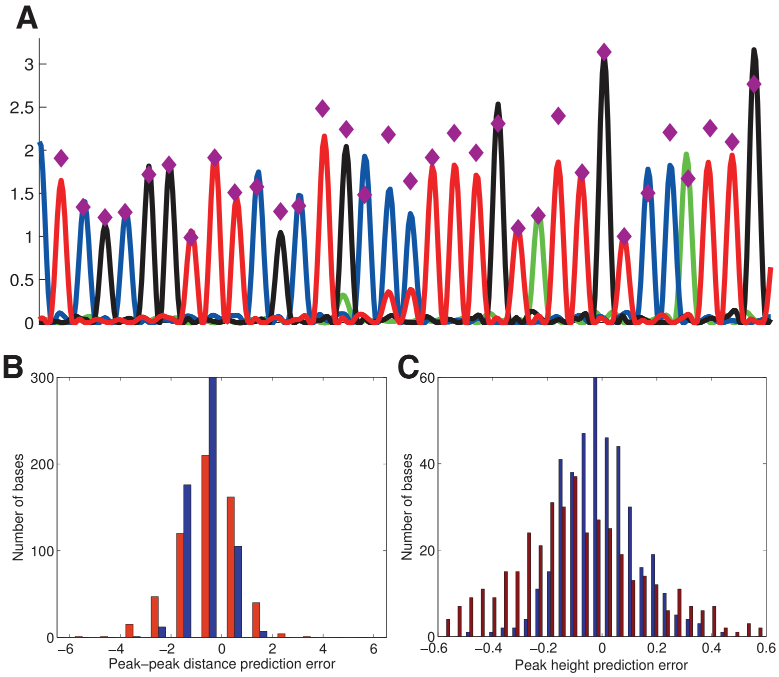

FIG. 11. Correcting for local sequence effects on chromatogram peak heights and positions. **(A).** Sample sequenced chromatogram and prediction (magenta circles) of peak heights and positions based on local (6-mer) sequence. **(B).** Distribution of peak-peak distance differences between predicted and measured peak positions before (red) and after (blue) correction for local sequence effects. The average peak-peak distance is ~12 pixels. **(C).** Distribution of distance between predicted and measured peak heights before (red) and after (blue) correction for local sequence effects. Employing local sequence context improves both height and positions predictions.

The database preprocessing step was applied for a database sequence matrix S, comprised of 18, 747 unique 16S rRNA gene sequences of average length 1,480 bases (see Section 2.2). The output is a set of PSSMs $P_i = (\mathbf{a}_i, \mathbf{c}_i, \mathbf{g}_i, \mathbf{t}_i)^t$, one for each sequence S_i in the database. The database processing scheme is applied only once to the database and the predicted PSSMs are stored and can be used for any new mixture sample obtained. It is applied to each sequence in the database independently.

We generated a chromatogram trace for a given 16S rRNA gene sequence by modeling each peak as a Gaussian centered at the peak position and with height equal to the peak height. The widths of the chromatogram Gaussian peaks were approximated using a constant peak width obtained by setting $c = 0.4$.

Each $f_{i,j}$ was evaluated for x values equally spaced in the entire sequence range $[0,k]$, but has a non-negligible contribution to the entire chromatogram only in the vicinity of the nucleotide position $b_{i,j}$, as is ensured by the Gaussian decay. A resolution of 1/12 was used as it corresponds roughly to the number of pixels available for a single nucleotide in real chromatograms. A chromatogram was generated for each 16S rRNA gene sequence by summing the values of obtained $f_{i,j}$ over all nucleotides.

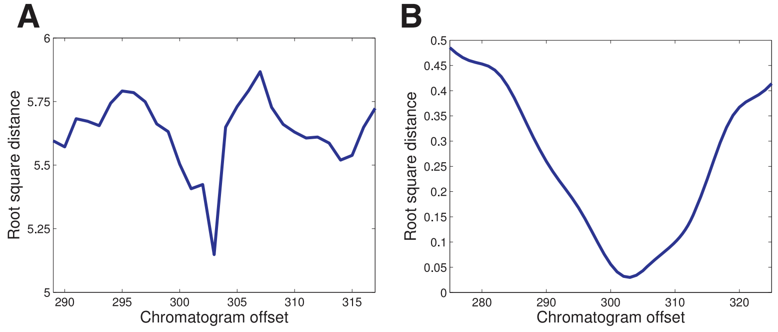

FIG. 12. Determination of chromatogram offset. **(A).** Root square distance between measured chromatogram and the chromatogram predicted from the **BCS** reconstruction. Minimal value is obtained when position 1 in the measured chromatogram is aligned to position 304 in the database. **(B).** Root square distance between measured chromatogram and the chromatogram predicted using the known composition of the five species in the mixture. Minimal value is obtained when position 1 in the measured chromatogram is aligned to position 304 in the database.

Alignment of predicted and measured chromatograms. Sanger-sequencing chromatograms display an initial region ($\sim$100 bases), which is highly noisy and therefore unusable. We are therefore faced with the problem of correctly aligning the initial bin position in the measured chromatogram and the bin positions of the predicted chromatograms. This was solved by trying the **BCS** reconstruction for different initial bin offsets in the measured chromatogram, and selecting for the offset with the lowest reconstruction root square distance (Fig. 12A). This reconstruction root square distance is calculated as the difference between the measured chromatogram and the predicted chromatogram based the reconstructed species frequencies. To verify the validity of this criterion, we also compared the average distance between the measured chromatogram and the predicted mixture chromatogram obtained using the known mixture composition (Fig. 12B), using various offsets for the measured chromatogram binning. Both methods obtained an identical offset, which was used in the reconstruction.

ACKNOWLEDGMENTS

We thank Amit Singer, Yonina Eldar, Gidi Lazovski, and Noam Shental for useful discussions, Eytan Domany for critical reading of the manuscript, Joel Stavans for supporting this research, and Chaime Priluski for assistance with chromatogram peak prediction data.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Amann, R.I., Ludwig, W., and Schleifer, K.H. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59, 143–169, 1995.

Armougom, F., and Raoult, D. 2008. Use of pyrosequencing and DNA barcodes to monitor variations in firmicutes and bacteroidetes communities in the gut microbiota of obese humans. *BMC Genomics* 9, 576.

Ben-Haim, Z., Eldar, Y.C., and Elad, M. 2010. Coherence-based performance guarantees for estimating a sparse vector under random noise. *Signal Process. IEEE Trans.* 58, 5030–5043.

Bobin, J., Starck, J.L., and Ottensamer, R. 2008. Compressed sensing in astronomy. *J. Sel. Topics Signal Process.* 2, 718–726.

Bowling, J.M., Bruner, K.L., Cmarik, J.L., and Tibbetts, C. 1991. Neighboring nucleotide interactions during DNA sequencing gel electrophoresis. *Nucleic Acids Res.* 19, 3089.

Brodie, E.L., DeSantis, T.Z., Parker, J.P.M., et al. 2007. Urban aerosols harbor diverse and dynamic bacterial populations. *Proc. Nat. Acad. Sci. USA* 104, 299–304.

Candes, E.J. 2006. Compressive sampling. *Proc. Int. Cong. Math.* 1433–1452.

Candes, E.J. 2008. The restricted isometry property and its implications for compressed sensing. *Comp. Rendus Acad. Sci.* 346, 589–592.

Candes, E.J., and Tao, T. 2005. Decoding by linear programming. *IEEE Trans. Inform, Theor.* 51, 4203–4215.

Candes, E.J., and Tao, T. 2006. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theor.* 52, 5406–5425.

Candes, E.J., and Tao, T. 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* 35, 2313–2351.

Candes, E.J., Romberg, J.K., and Tao, T. 2006. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.* 59, 1207–1223.

Dai, W., Sheikh, M.A., Milenkovic, O., et al. 2009. Compressive sensing DNA microarrays. *EURASIP J. Bioinform. Syst. Biol.* 162824.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* 72, 5069.

Dethlefsen, L., Huse, S., Sogin, M.L., et al. 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* 6, e280.

Dewhirst, F.E., Izard, J., Paster, B.J., et al. 2008. The human oral microbiome database. Journal vol. pp. 5.

Donoho, D.L. 2006a. Compressed sensing. *IEEE Trans. Inform. Theor.* 52, 1289–1306.

Donoho, D.L. 2006b. For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution. *Commun. Pure Appl. Math.* 59, 797–829.

Duarte, M., Davenport, M., Takhar, D. et al. 2008. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.* 25, 83–91.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., et al. 2005. Diversity of the human intestinal microbial flora. *Science* 308, 1635–1638.

Erlich, Y., Gordon, A., Brand, M., et al. 2010. Compressed genotyping. *IEEE Trans. Inform. Theor.* 56, 706–723.

Ewing, B., and Green, P. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* 8, 186–194.

Ewing, B., Hillier, L.D., Wendl, M.C., et al. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res.* 8, 175–185.

Faith, J.J., McNulty, N.P., Rey, F.E., et al. 2011. Predicting a human gut microbiota's response to diet in gnotobiotic mice. *Science* 333, 101–104.

Figueiredo, M.A.T., Nowak, R.D., and Wright, S.J. 2007. Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Topics Signal Process.* 1, 586–597.

Gao, Z., Tseng, C., Pei, Z., et al. 2007. Molecular analysis of human forearm superficial skin bacterial biota. *Proc. Natl. Acad. Sci. USA* 104, 2927–2932.

Gentry, T., Wickham, G., Schadt, C., et al. 2006. Microarray applications in microbial ecology research. *Microb. Ecol.* 52, 159–175.

Guarner, F., and Malagelada, J.R. 2003. Gut flora in health and disease. *Lancet* 361, 512–519.

Hamady, M., and Knight, R. 2009. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res.* 19, 1141–1152.

Hamady, M., Walker, J.J., Harris, J.K., et al. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods* 5, 235–237.

Hugenholtz, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* 3, reviews0003.8.

Huse, S.M., Dethlefsen, L., Huber, J.A., et al. 2008. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet.* 4, e1000255.

Kainkaryam, R.M., and Woolf, P.J. 2009. Pooling in high-throughput drug screening. *Curr. Opin. Drug Discov. Dev.* 12, 339.

Keller, M., and Zengler, K. 2004. Tapping into microbial diversity. *Nat. Rev. Microbiol.* 2, 141–150.

Kommedal, O., Karlsen, B., and Sabo, O. 2008. Analysis of mixed sequencing chromatograms and its application in direct 16S rDNA sequencing of poly-microbial samples. *J. Clin. Microbiol.* 46, 3766–3771.

Lin, T.T., and Herrmann, F.J. 2007. Compressed wavefield extrapolation. *Geophysics* 72, SM77–SM93.

Lipshutz, R.J., Taverner, F., Hennessy, K., et al. 1994. DNA sequence confidence estimation. *Genomics* 19, 417–424.

Lustig, M., Donoho, D.L., and Pauly, J.M. 2007. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Magnet. Reson. Med.* 58, 1182–1195.

Mager, D.L., Haffajee, A.D., Devlin, P.M., et al. 2005. The salivary microbiota as a diagnostic indicator of oral cancer: a descriptive, non-randomized study of cancer-free and oral squamous cell carcinoma subjects. *J. Trans. Med.* 3, 27.

Maiden, M.C.J., Bygraves, J.A., Feil, E., et al. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA* 95, 3140–3145.

Medini, D., Serruto, D., Parkhill, J., et al. 2008. Microbiology in the post-genomic era. *Nat. Rev. Microbiol.* 6, 419–430.

Muegge, B.D., Kuczynski, J., Knights, D., et al. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974.

Nickerson, D.A., Tobe, V.O., and Taylor, S.L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25, 2745–2751.

Paster, B.J., Boches, S.K., Galvin, J.L., et al. 2001. Bacterial diversity in human subgingival plaque. *J. Bacteriol.* 183, 3770–3783.

Rusch, D.B., Halpern, A.L., Sutton, G., et al. 2007. The sorcerer ii global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 5, e77.

Savage, D.C. 1977. Microbial ecology of the gastrointestinal tract. *Annu. Rev. Microbiol.* 31, 107–133.

Sears, C.L. 2005. A dynamic partnership: celebrating our gut flora. *Anaerobe* 11, 247–251.

Shental, N., Amir, A., and Zuk, O. 2010. Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Res.* 38, e179.

Singh, B.K., Millard, P., Whiteley, A.S., et al. 2004. Unravelling rhizosphere-microbial interactions: opportunities and limitations. *Trends Microbiol.* 12, 386–393.

Tropp, J.A. 2006. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theor.* 52, 1030–1051.

Wooley, J.C., Godzik, A., and Friedberg, I. 2010. A primer on metagenomics. *PLoS Comput. Biol.* 6, e1000667.

Yarza, P., Richter, M., Peplies, J., et al. 2008. The all-species living tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. *Syst. Appl. Microbiol.* 31, 241–250.

Address correspondence to:
Dr. Or Zuk
Broad Institute
MIT and Harvard
Cambridge, MA 02142

E-mail: orzuk@broadinstitute.org