Bacterial Community Reconstruction Using Compressed Sensing - Supplementary Information

1 Ribosomal DNA Database

16S rRNA gene sequences were obtained from grenegenes (`greengenes.lbl.gov`) using database version 06-2007 [2], which contains approximately 136000 chimera checked full length sequences. Sequences were reverse complemented and aligned with primer 1510R [5], resulting in approximately 42000 sequences matching the primer sequence (with up to 6 mismatches with the primer). Out of this set, sequences with up to 2 base-pair difference with another sequence in the database were removed, resulting in $N = 18747$ unique sequences which were used in this study. This last step was used in order to reduce the size of the input to the GPSR algorithm, thus enabling solution of the **CS** problem using a standard PC.

We manually added the sequence of *Enterococcus faecalis* (ATCC # 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 Experimental Mixture Reconstruction

2.1 Sample Preparation.

Strains used for the experimental reconstruction were: *Escherichia coli W3110, Vibrio fischeri, Staphylococcus epidermidis* (ATCC # 12228), *Enterococcus faecalis* (ATCC # 19433) and *Photobacterium leiognathi*. The 16S rRNA gene was obtained 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 [5], amplifying positions 8-1513 of the *E. coli* 16S rRNA:

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

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

2.2 Preprocessing Steps.

The input to the **BCS** algorithm is a $4 \times k$ PSSM $(\mathbf{a}, \mathbf{c}, \mathbf{g}, \mathbf{t})^t$ 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 (see e.g. [1, 7]) 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 preprocessing of the chromatogram, which does not depend on identifying the peak for each nucleotide. Rather, the chromatogram is binned into constant sized bins, and the total intensity of each of the four nucleotides in each bin is used to construct the PSSM used as input to the **BCS** (see Figure S1A). 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 [3, 4]. 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. This database of predicted PSSMs is then used to construct the mixing matrices A, C, G, T participating in the **BCS** problem representation (see eq. 6 in the main text and Figure S1B). Details on the chromatogram and database preprocessing steps are given in sections 3 and 4, respectively.

3 Chromatogram Preprocessing

In order to apply the **BCS** algorithm on the experimentally measured mixture chromatogram, several preprocessing steps are required. The purpose of the chromatogram preprocessing step is to convert the measured chromatogram to a PSSM representing the frequency of each base at each position in the mixture (see Figure S1A). We provide below a formal algorithm sketch for this step, followed by a more detailed description:

Algorithm: 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:

$$\mathfrak{a}_p = \frac{50 \cdot 12 \cdot \mathfrak{a}_p}{\sum_{q=-25\cdot 12}^{25\cdot 12}(\mathfrak{a}_{p+q} + \mathfrak{c}_{p+q} + \mathfrak{g}_{p+q} + \mathfrak{t}_{p+q})} \tag{1}$$

   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:

$$a_j = \sqrt{\sum_{p=12j}^{12j+11} \mathfrak{a}_p}\,, \quad j = 1...k \tag{2}$$

   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 $\mathfrak{a}, \mathfrak{c}, \mathfrak{g}, \mathfrak{t}$, where for example $\mathfrak{a}_p$ 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 the Results section, 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 $\sim$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.

4 Database Preprocessing

The purpose of the Database Preprocessing scheme is to produce predicted PSSMs for all 16S rRNA sequences in the database (see Figure S1B). 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 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 essentially captures the relations between the sequence of a DNA molecule and the chromatogram charts obtained when sequencing such a molecule. The main factor effecting the chromatogram shape is local-sequence context, which is modeled by a 5-th order Markov Chain. The model parameters are fitted in a preliminary step by using a training set of sequences with experimentally available chromatograms. This preliminary step is described in the next section. In the following section, we describe the database preprocessing step performed once the model parameters are fully specified.

4.1 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: 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 position 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 previous section.)

4. Compute context-specific peak height averages: for a given k-mer $\alpha = (\alpha_1, .., \alpha_6)$, compute the averaged peak heights of all the occurrences of α as a k-mer in all training set sequences:

$$H(\alpha) = \frac{\sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, .., \alpha_6 = S'_{i,j}\}} h_{i,j}}{\sum_{i,j} 1_{\{\alpha_1 = S'_{i,j-5}, .., \alpha_6 = S'_{i,j}\}}} \tag{3}$$

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

$$d_{i,j} = \frac{p_{i,j} - p_{i,j-1}}{\sum_{j=2}^{k} p_{i,j} - p_{i,j-1}}. \tag{4}$$

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

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

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

$$d_{i,j} = \gamma + \beta j \tag{6}$$

   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 [6]. 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 the Chromatogram Preprocessing section).

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 5 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 in the supplementary data files. 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 $\sim 0.5-1.3$ and position values P in the range $\sim 0.8-1.2$ (see Figure S2). An additional sequence-independent non-linearity in the peak position was observed in the chromatograms studied, where distance between consecutive peak increases as we move further along the chromatogram. This was accounted for by fitting an additional linear model based only on the nucleotide position along the sequence, giving an additional parameter of $\beta = 0.00036$ representing increase in peak-peak distance with each position (see next section in eq. (8)).

4.2 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 $(\mathfrak{a}, \mathfrak{c}, \mathfrak{g}, \mathfrak{t})$, one for each 16S rRNA gene sequence in S. These chromatograms are then further processed to get predicted PSSMs. This step is illustrated in Figure S1B, and is described below.

Algorithm: Compute local-adjusted chromatogram statistics

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:

$$a_{i,j} = H(L_{i,j}) \tag{7}$$

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

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

$$b_{i,j} = b_{i,j-1} + D(L_{i,j}) + \beta \cdot j \tag{8}$$

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

$$f_{i,j}(x) = a_{i,j} e^{-\frac{(x-b_{i,j})^2}{2c^2}} \tag{9}$$

where x is sampled in a range $[0, k]$ at a resolution of $1/12$ thus giving $12k$ different x values $x_1, .., x_{12k}$ 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:

$$\mathfrak{a}_{i,p} = \sum_j f_{i,j}(x_p) 1_{\{S_{i,j}='A'\}} \tag{10}$$

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. (2), 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.

The database preprocessing step was applied for a database sequence matrix S, comprised of 18747 unique 16S rRNA gene sequences of average length 1480 bases (see the Ribosomal DNA Database section of the Materials and Methods section). 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.

4.2.1 Alignment of Predicted and Measured Chromatograms

Sanger-sequencing chromatograms display an initial region (∼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 (see Figure S5A). 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 (see Figure S5B), using various offsets for the measured chromatogram binning. Both methods obtained an identical offset, which was used in the reconstruction.

References

1. Bowling, J., Bruner, K., Cmarik, J., Tibbetts, C.: Neighboring nucleotide interactions during DNA sequencing gel electrophoresis. Nucleic acids research 19(11), 3089 (1991)

2. DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and environmental microbiology 72(7), 5069 (2006)

3. Ewing, B., Green, P.: Base-calling of automated sequencer traces usingPhred. II. error probabilities. Genome research 8(3), 186 (1998)

4. Ewing, B., Hillier, L., Wendl, M., Green, P.: Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment. Genome research 8(3), 175 (1998)

5. Gao, Z., Tseng, C., Pei, Z., Blaser, M.: Molecular analysis of human forearm superficial skin bacterial biota. Proceedings of the National Academy of Sciences 104(8), 2927 (2007)

6. Lipshutz, R., Taverner, F., Hennessy, K., Hartzell, G., Davis, R.: DNA sequence confidence estimation. Genomics 19(3), 417–424 (Feb 1994)

7. Nickerson, D., Tobe, V., Taylor, S.: PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Research 25(14), 2745 (1997)

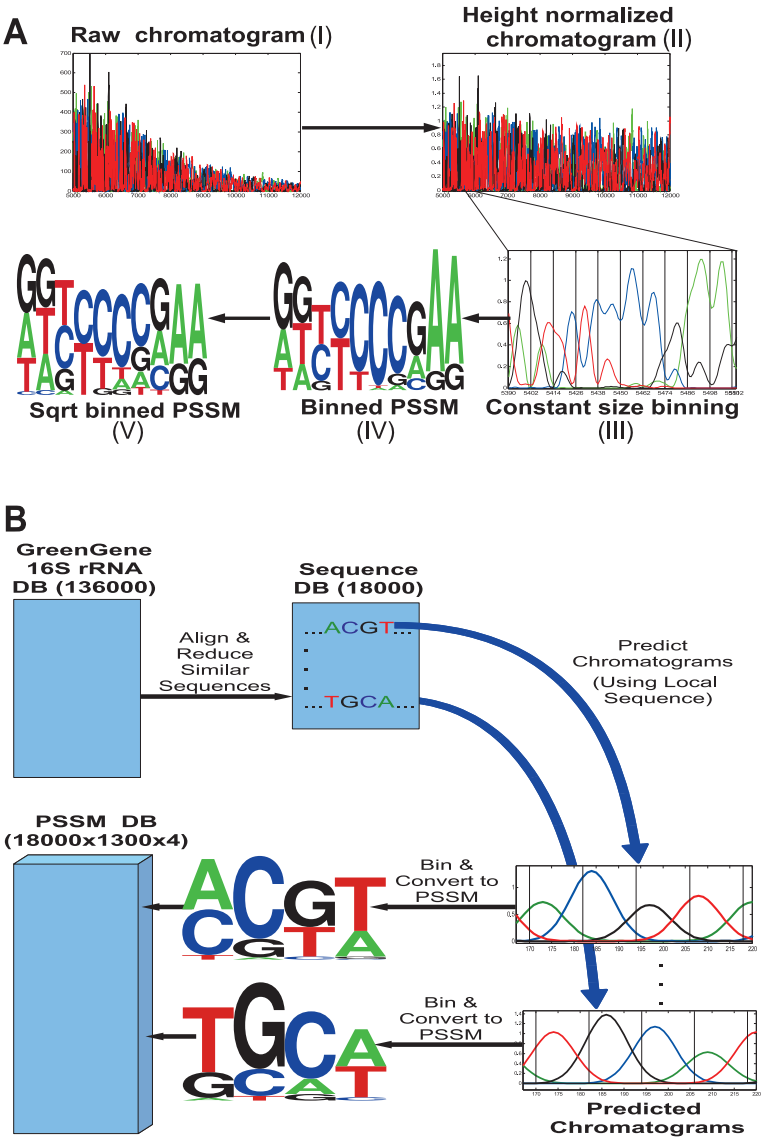

Figure 1. (Supplementary) Preprocessing steps A. Preprocessing of the experimental chromatogram. The result of the Sanger-sequencing of a bacterial mixture (I) is normalized by division with a $\sim$1000 pixel total intensity running average to compensate for the peak amplitude decrease. The resulting 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.

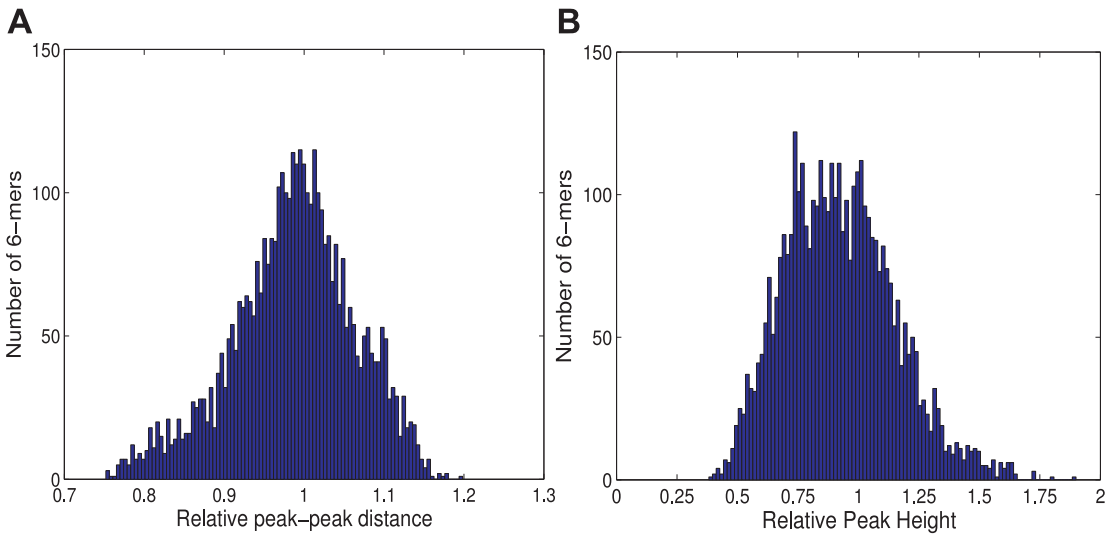

Figure 2. (Supplementary) Local-sequence effect on chromatogram peak height and position. A. Distribution of average normalized peak-peak distances for the 4096 sequence 6-mers. **B.** Distribution of normalized peak heights for the 4096 sequence 6-mers. Both distributions show a rather wide spread around one, showing that local sequence context has a significant effect on peak height and position.

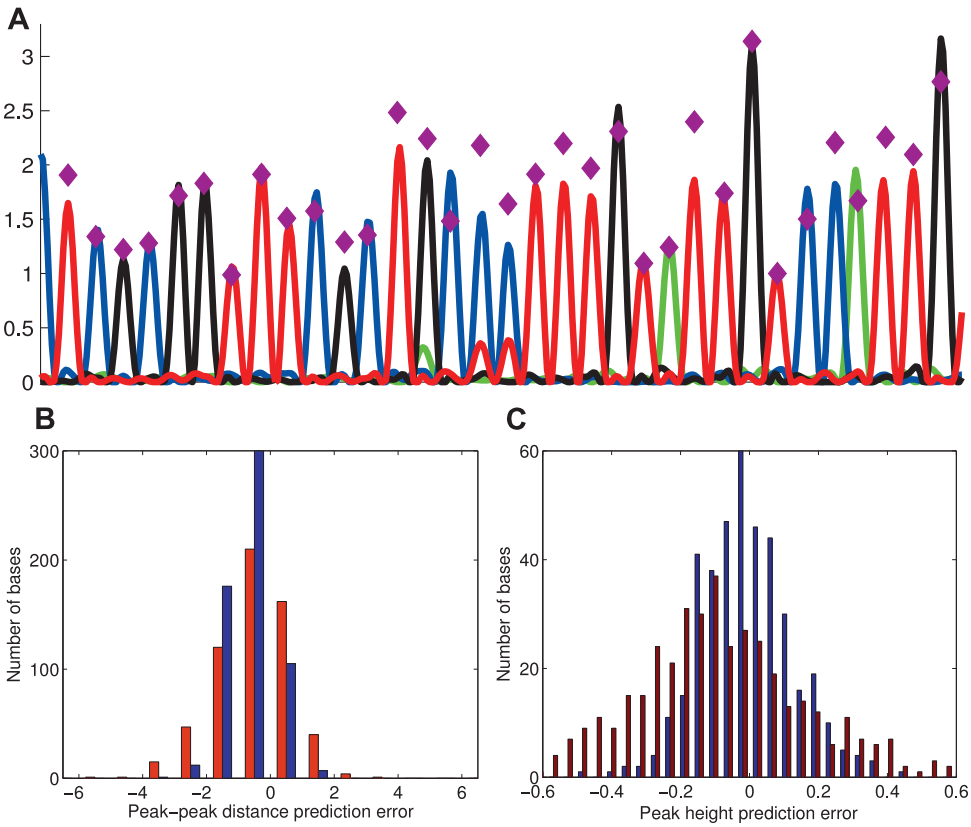

Figure 3. (Supplementary) Effect of local sequence on chromatogram peak heights and positions. A. Sample sequenced chromatogram and prediction (magenta circles) of peak heights and positions based on local (6-mer) sequence. **B.** Distribution of peak-peak distance differences between predicted and measured 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.

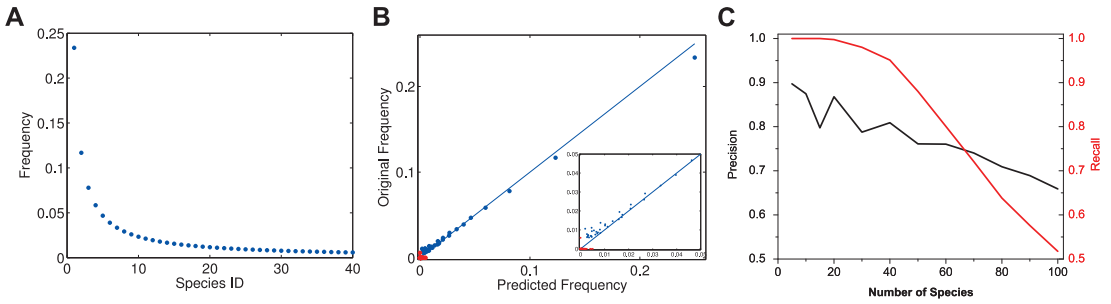

Figure 4. (Supplementary) Sample reconstruction of a power-law mixture. A. Sorted frequency distribution of 40 random species following a power-law distribution with frequencies $v_i \sim i^{-1}, i = 1, .., 40$. **B.** True vs. predicted frequencies for a sample **BCS** reconstruction for the mixture 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%.

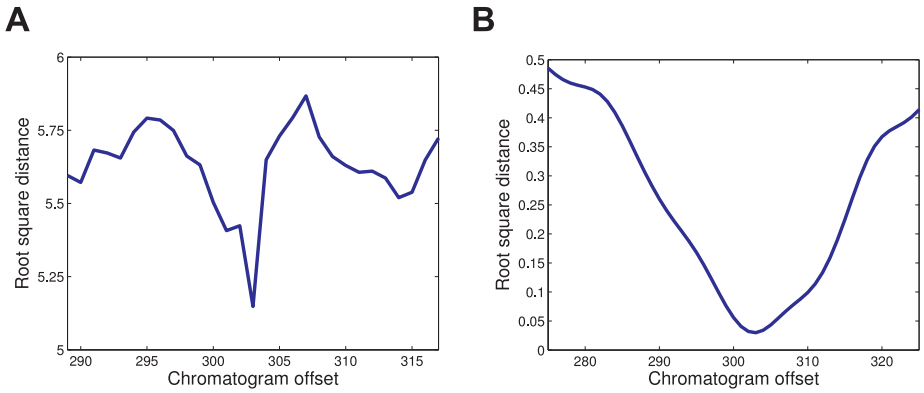

Figure 5. (Supplementary) Determination of chromatogram offset. A. Root square distance between measured chromatogram and the chromatogram predicted from the **BCS** reconstruction. Minimal value is obtained when position 1 in the measured chromatogram is aligned to position 304 in the database. **B.** 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.