BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 17 2013, pages 2096–2102
doi:10.1093/bioinformatics/btt336

Sequence analysis

Advance Access publication June 20, 2013

Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing

David Koslicki[1,*], Simon Foucart[2] and Gail Rosen[3]

[1]Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43201, USA and [2]Department of Mathematics and [3]Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Many metagenomic studies compare hundreds to thousands of environmental and health-related samples by extracting and sequencing their 16S rRNA amplicons and measuring their similarity using beta-diversity metrics. However, one of the first steps—to classify the operational taxonomic units within the sample—can be a computationally time-consuming task because most methods rely on computing the taxonomic assignment of each individual read out of tens to hundreds of thousands of reads.

Results: We introduce Quikr: a QUadratic, K-mer–based, Iterative, Reconstruction method, which computes a vector of taxonomic assignments and their proportions in the sample using an optimization technique motivated from the mathematical theory of compressive sensing. On both simulated and actual biological data, we demonstrate that Quikr typically has less error and is typically orders of magnitude faster than the most commonly used taxonomic assignment technique (the Ribosomal Database Project's Naïve Bayesian Classifier). Furthermore, the technique is shown to be unaffected by the presence of chimeras, thereby allowing for the circumvention of the time-intensive step of chimera filtering.

Availability: The Quikr computational package (in MATLAB, Octave, Python and C) for the Linux and Mac platforms is available at http://sourceforge.net/projects/quikr/.

Contact: koslicki.1@mbi.osu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 27, 2013; revised on June 5, 2013; accepted on June 6, 2013

1 INTRODUCTION

Reconstructing the taxonomic composition of a bacterial community taken from an environmental sample (be it a human-associated, ocean, or soil sample) is critical for understanding the role that such a community might play in affecting change in that environment. A popular reconstruction approach (Cole *et al*., 2009; Jumpstart Group, 2012; Lan *et al*., 2012; Wang and Zhang, 2011, Wang *et al*., 2007) is to use 16S rRNA amplicon sequencing (like Roche's 454 technology) to produce many (∼400 000–1 000 000) moderate-length (∼400–700 bp) reads of specific variable regions of the 16S rRNA gene and then individually classify these reads using a custom database with BLAST or in a Bayesian framework like the Ribosomal Database Project's (RDP) Naïve Bayesian Classifier (NBC) (Wang *et al*., 2007). RDP's NBC is widely used owing to its speed but it can still take several days to assign millions of reads on a desktop computer, thereby alienating users who do not have access to large computer clusters.

We introduce a method that enables desktop analysis: we take a novel approach by reconstructing all taxonomic concentrations of a bacterial community simultaneously (as opposed to read-by-read classification). This allows for orders of magnitude decrease in execution time while maintaining comparable (and often better) reconstruction fidelity. This method, based on ideas from compressive sensing, was inspired by and tangentially related to (Amir and Zuk, 2011) wherein sparsity-promoting algorithms were used to analyze mixtures of dye-terminator reads resulting from Sanger sequencing. Here, however, we take a *k*-mer–based approach that is designed for high-throughput sequencing technologies. This is similar in spirit to the *k*-mer–based approach in (Meinicke *et al*., 2011) but herein we use a distribution estimation procedure based on compressive sensing. Put briefly, our method measures the frequency of *k*-mers (for a fixed $k \sim 6$) in a database of 16S rRNA genes for known bacteria, calculates the frequency of *k*-mers in the given sample, and then reconstructs the concentrations of the bacteria in the sample by solving an underdetermined system of linear equations under a sparsity assumption. To solve this system, we employ MATLAB's (MATLAB, 2012) iterative implementation of typical nonnegative least squares and hence we refer to this method as *Quikr*: QUadratic, Iterative, *K*-mer–based Reconstruction. We point out that Quikr has not yet been optimized for performance but still demonstrates orders of magnitude speed improvement over RDP's NBC.

2 METHODS

2.1 *k*-mer training matrix

The training step consists of converting an input database of 16S rRNA sequences into a *k-mer training matrix*. For a fixed *k*-mer size, we calculate the frequency of each *k*-mer in each database sequence. Hence, given a database of 16S rRNA sequences $D = \{d_1, \ldots, d_M\}$, the $(i,j)^{\text{th}}$ entry of the *k*-mer training matrix $A^{(k)}$ is the frequency of the i^{th} *k*-mer (in lexicographic order) in the j^{th} sequence d_j.

Herein, we consider two different databases of 16S rRNA sequences. The first database, D_{small}, is the same as the training database for RDP's NBC version 7. This database consists of 10 046 sequences and will allow for direct comparison of Quikr to RDP's NBC.

*To whom correspondence should be addressed.

The second database, D_{large}, consists of the 275 727 sequences that remained after applying TaxCollector (Giongo *et al.*, 2010) to the entire RDP 16S rRNA database 10.28. Applying TaxCollector had the net effect of labeling each sequence with taxonomic information obtained from NCBI (Benson *et al.*, 2009; Sayers *et al.*, 2009), discarding duplicate sequences and discarding sequences that were missing genus labels. Training the RDP's NBC with the database D_{large} would lead to prohibitively long classification times (>17 min per read on a 2.0 GHz Intel E7-4820 processor) and so demonstrates how Quikr can incorporate much more known information than RDP's NBC.

Forming the k-mer training matrix for D_{small} and D_{large} took ~15 s and 15 min, respectively, on a 2.0 GHz Intel E7-4820 processor.

2.2 Sample k-mer frequencies

Given a sample dataset of 16S rRNA reads, we calculate the frequency of all k-mers in the entire sample. We refer to this vector $s^{(k)}$ as the *sample k-mer frequency vector*. Note that the calculation of $s^{(k)}$ is an easily parallelizable problem that can be computed efficiently in an online fashion.

2.3 Sparsity promoting quadratic optimization

We assume that the given environmental sample only contains bacteria that exist in the database $D = \{d_1, \ldots, d_M\}$ being used. Hence, we can represent the composition of the sample as a vector x with non-negative entries summing to one (i.e. a probability vector) where x_i is the concentration of the organism with 16S rRNA sequence d_i. However, as will be demonstrated in section 3.5, the Quikr method still performs well when the sample *does* contain novel bacteria not in the database being used.

We consider the idealized situation, in which sample noise and errors introduced by short reads are ignored. The problem at hand is then to reconstruct the bacterial concentrations x by solving the underdetermined linear system

$$A^{(k)}x = s^{(k)}, \tag{1}$$

Under the plausible assumption that relatively few bacteria from the database D are actually present in the given sample (that is, x is a sparse vector), we can solve equation (1) by modifying some techniques from compressive sensing. We use a variant of basis-pursuit denoising (Chen *et al.*, 1998), which reduces to a non-negative least squares problem. The details regarding this sparsity promoting, iterative, quadratic optimization procedure are contained in the Supplementary Material.

Occasionally, Quikr experiences convergence issues. However, as detailed in the Supplementary Material, filtering out the shortest sequences from a given sample solved this issue in every situation we encountered.

2.4 Reconstruction metrics

There are a variety of metrics used in the literature to asses bacterial community reconstruction fidelity (for example see Amir and Zuk, 2011; Clemente *et al.*, 2011; Rosen *et al.*, 2008, Segata *et al.*, 2012 and Wang *et al.*, 2007). We denote the *actual* and *predicted* concentrations of the bacteria as probability vectors x and x^*, respectively. The reconstruction metric primarily used herein is the ℓ_1 distance between x and x^*: $||x - x^*||_{\ell_1}$. This quantity takes values between 0 and 2 (with perfect reconstruction being $||x - x^*||_{\ell_1} = 0$) and is commonly referred to as 'total error' (as it is the total of the absolute errors). We also use precision, sensitivity, specificity and accuracy; these error metrics vary between 0 and 1 (with higher values reflecting better reconstruction fidelity). The definitions of these quantities are contained in the Supplementary Material. Note that the correlation between x and x^* is not an effective reconstruction metric because the sparsity of x and x^* and the high number of true negatives typically make $\text{corr}(x, x^*) := x^\top x^*/(||x||_2||x^*||_2)$ too close to the optimal value 1.

The term *reconstruction fidelity* will be used to communicate generically how well x^* approximates x.

2.5 Simulated data

To test the performance of the Quikr method, the shotgun/amplicon read simulator Grinder (Angly *et al.*, 2012) was used to generate a large variety of simulated 454 pyrosequencing datasets. These datasets were designed to mimic reads generated by Roche's GS FLX and FLX+ amplicon systems, so read-length distributions were set to be normally distributed with a mean of 400 or 700 bp and a standard deviation of 50 or 100 bp. The primers B27F, B357F and BU968F were chosen to target the V1-V3, V3-V5 and V6-V9 variable regions, respectively. Only forward primers were used because amplicon sequencing allows for filtering on sequencing direction. Three different diversity values were chosen to be 10^2, 10^3 and 10^4, and abundance was modeled by a power-law or exponential distribution with parameters 0.705 and 1, respectively. Because most sequencing errors in these systems are due to homopolymer errors, such errors were modeled by using Balzer's model (Balzer *et al.*, 2010). Chimera percentages were set at 0, 10 and 30%. Since only amplicon sequencing is considered, no copy or length bias was employed.

In all, 216 different simulated datasets were generated with over 172 million reads, resulting in over 78 billion bases.

2.6 Mock communities

To benchmark the Quikr method on real biological data, we examined the mock microbial communities developed in (Haas *et al.*, 2011). These communities contain staggered concentrations of 16S rRNA genes for each of 21 different organisms that span a diverse range of properties (GC content, genome size, etc.). This mock microbial community was then sequenced independently at four different institutions with primers designed to target the V1-V3, V3-V5 and V6-V9 variable regions resulting in 12 different 454 datasets with an average read length of 439 bp and standard deviation 38 bp. Detail regarding the precise conditions under which this data was obtained appears in (Haas *et al.*, 2011, pages 499–500).

2.7 Human microbiome data

To further benchmark the Quikr method on real biological data, we applied the Quikr method to the Human Microbiome Project's (HMP; The Human Microbiome Project Consortium, 2012) trimmed sequences resulting from SRA study id SRP002395. This dataset consists of approximately 72 million reads over 5034 samples targeting the V1–V3 and V6–V9 variables regions.

3 RESULTS

3.1 Speed comparison

We performed all benchmarks against RDP's NBC because this is considered to be the fastest 16S rRNA classifier to date (Liu *et al.*, 2008). Figure 1 shows a log-log plot of the number of reads analyzed versus time for RDP's NBC version 10.28 with training set 7 (this is the same as database D_{small}, see section 2.1) and Quikr with $k = 6$ using the database D_{small}. Note the significant improvement in speed: it takes Quikr well less than 1 min to analyze over 1 million reads. While RDP's NBC computational complexity in the number of reads N is approximately $\mathcal{O}(N)$, on this data Quikr is approximately $\mathcal{O}(N^{1/5})$.

3.2 Simulated data results

The Quikr method was applied to all 216 simulated datasets using k-mer sizes in the range $k = 1, \ldots, 6$ for both databases

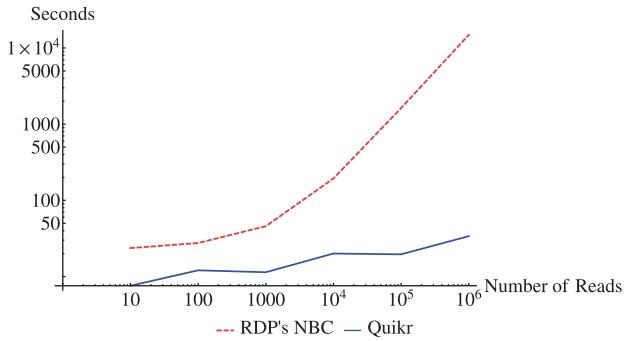

Fig. 1. Log–Log plot of number of reads versus time (in seconds) for both RDP's NBC and Quikr

D_{small} and D_{large}. We observed that at the genus level the mean ℓ_1 error decreased roughly linearly (linear regression $R^2 = 0.953$) as a function of k-mer size. However, the total algorithm time increased exponentially. This behavior is to be expected owing to the exponential increase in number of k-mers as a function of k. These patterns were observed at all taxonomic ranks with both training databases. We recommend using the k-mer size $k = 6$, as this provides a good trade-off between reconstruction fidelity and execution time.

For comparison purposes, we also classified the simulated data using the popular RDP's NBC (Wang *et al.*, 2007) version 10.28 with training set 7 (this is the same training data as the database D_{small}). Figure 2 compares the timing, mean ℓ_1 error at various taxonomic ranks, as well as precision, sensitivity, specificity and accuracy at the genus level between Quikr (using k-mer size $k = 6$) and RDP's NBC.

As part (a) of Figure 2 shows, Quikr is orders of magnitude faster than RDP's NBC no matter which training database is used. Indeed, using D_{large}, Quikr took an average of 1730 s per dataset (or 520 reads per second). Using D_{small}, Quikr took an average of only 26.4 s per dataset (or 34 091 reads per second). Compare this with RDP's NBC taking an average of 23 978 s per dataset (or 38 reads per s).

Part (b) in Figure 2 demonstrates that both methods show an increase in mean ℓ_1 error as one moves to lower taxonomic ranks. At the genus level and using the training database D_{large}, Quikr shows a 46.5% improvement in ℓ_1 error over RDP's NBC. Using the training database D_{small}, Quikr has comparable error to RDP's NBC down to the family level. Using this smaller database, Quikr results in more error than RDP's NBC at the genus level.

Part (c) in Figure 2 shows that when using D_{large}, Quikr has comparable specificity and accuracy, and only slightly lower averages for precision and sensitivity when compared with RDP's NBC at the genus level. This pattern continues when using the database D_{small} except here Quikr is much less sensitive than RDP's NBC but shows comparable precision, specificity and accuracy.

These results demonstrate that when using the training database D_{small}, Quikr is an extremely fast method that gives a good high-level characterization of a given sample. When using the training database D_{large}, Quikr is a fast and accurate classification method even down to the genus level.

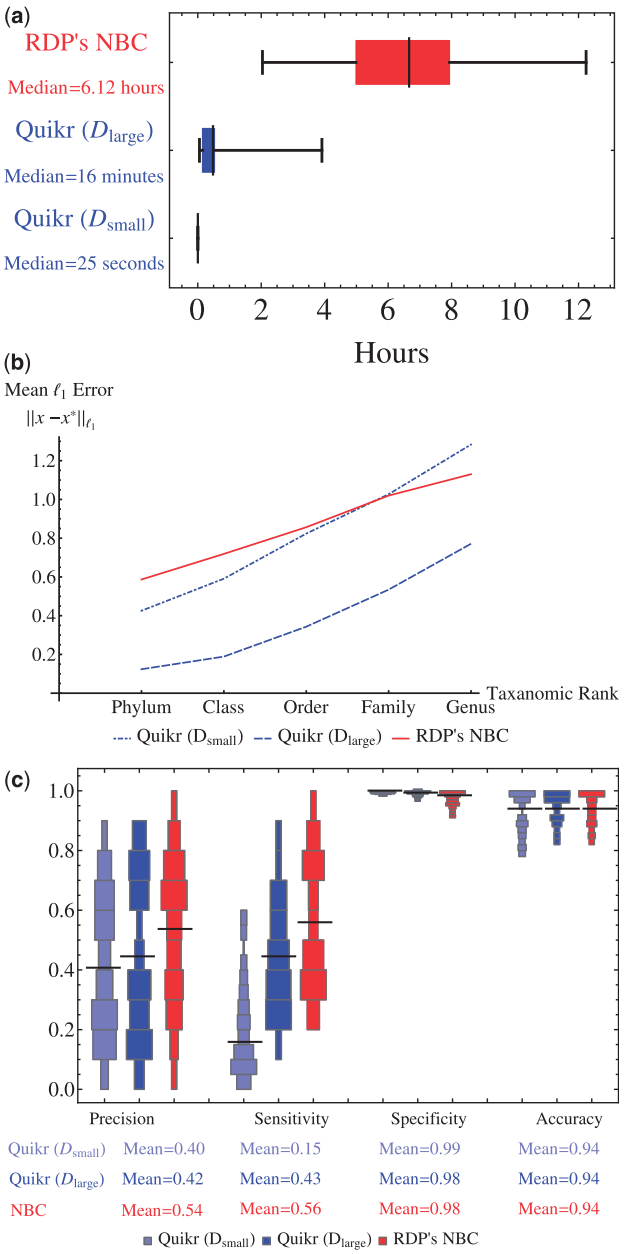

Fig. 2. Comparison of Quikr to RDP's NBC on simulated data. Throughout, RDP's NBC version 10.28 with training set 7 was used. (**a**) Algorithm execution time for RDP's NBC and Quikr trained using D_{large} and D_{small}. Whiskers denote range of the data, vertical black bars designate the median and the boxes demarcate quantiles. (**b**) ℓ_1 error averaged over all 216 simulated datasets versus taxonomic rank for RDP's NBC and Quikr trained using D_{small} and D_{large}. (**c**) Histogram densities for other error metrics at the genus level for RDP's NBC and when Quikr was trained using D_{small} and D_{large}. The horizontal black bars represent the mean

3.3 Mock communities results

We analyzed the 12 mock communities with the Quikr method for k-mer size $k = 6$ with both training databases D_{large} and D_{small}, as well as the RDP's NBC version 10.28 with training set 7 (which is the same as database D_{small}). Figure 3 compares

the timing, mean ℓ_1 error at various taxonomic ranks, as well as the remaining error metrics at the genus level between Quikr and RDP's NBC. Similarly to the simulated data in section 3.2, with training database D_{large}, Quikr is on average much faster than RDP's NBC, and significantly faster when using the training database D_{small} [(see part (a) of Fig. 3)]. As part (b) of Figure 3 shows, the ℓ_1 errors of both methods are comparable. Furthermore, when using the training database D_{large}, Quikr has less error than RDP's NBC at the genus level. Lastly, when Quikr uses the training database D_{large}, both methods have comparable precision, sensitivity, specificity and accuracy (note Quikr is slightly more precise, specific and accurate). When using D_{small}, Quikr is significantly less sensitive than RDP's NBC, but the other error metrics give similar values.

Figure 4 shows the consensus/mean predicted phyla over all 12 mock communities for both the Quikr method (using database D_{small}) and RDP's NBC. The correlation between predicted and actual concentrations for Quikr is 0.9724 versus RDP's NBC correlation of 0.9700. The concentrations for each phyla as predicted by Quikr are closer on average to the actual concentrations than that of RDP's NBC at the cost of a false-positive phylum of Tenericutes. However, the mock communities contained 18S rRNA of *Candida albicans* and the *k*-mer frequencies of this species is closer to the average *k*-mer frequencies of the Tenericutes than any other phyla. Future plans for Quikr include developing a measure for novel taxa so as to address this issue of potential false positives.

This demonstrates again that when using the training database D_{small}, Quikr is an extremely fast method that gives a good high-level characterization of a given sample. When using the training database D_{large}, Quikr is a fast and very accurate classification technique.

3.4 HMP results

To demonstrate that Quikr is fit for utilization on a desktop computer, we analyzed the 5034 samples of HMP data on an iMac with a 3.4 GHz Intel i-7 processor. Using the default training database D_{small} (which corresponds to RDP's training set 7), Quikr took 7.6 h to analyze the entire HMP data set. Retraining with the Greengenes (DeSantis *et al.*, 2006) 91%-OTU database of 5878 sequences, Quikr took only 4.8 h to analyze the entire HMP dataset. The results of analyzing the HMP data with the Grenegenes database were then analyzed in QIIME (Caporaso *et al.*, 2010) to produce a PCoA plot, which is included in Figure 5. This plot can be compared with Figure 1a in (Koren *et al.*, 2013) reproduced here as Figure 5c. To quantify the variability of a particular category (body site in this case), QIIME (Caporaso *et al.*, 2010) includes several methods, such as Adonis and ANOSIM, which can asses the statistical significance of groupings in a PCoA plot as well as indicate how much of the variation is explained by such groupings. Because we did not have the file that generated Figure 5c and are only demonstrating the concept of applying Quikr for generating fast PCoA plots, we did not investigate further such quantitative comparisons. The results are qualitatively similar enough in their clustering and distinguishing of body sites to conclude that Quikr is effective in facilitating the transformation of raw reads into an accurate PCoA plot in less than a workday on a typical scientist's desktop computer.

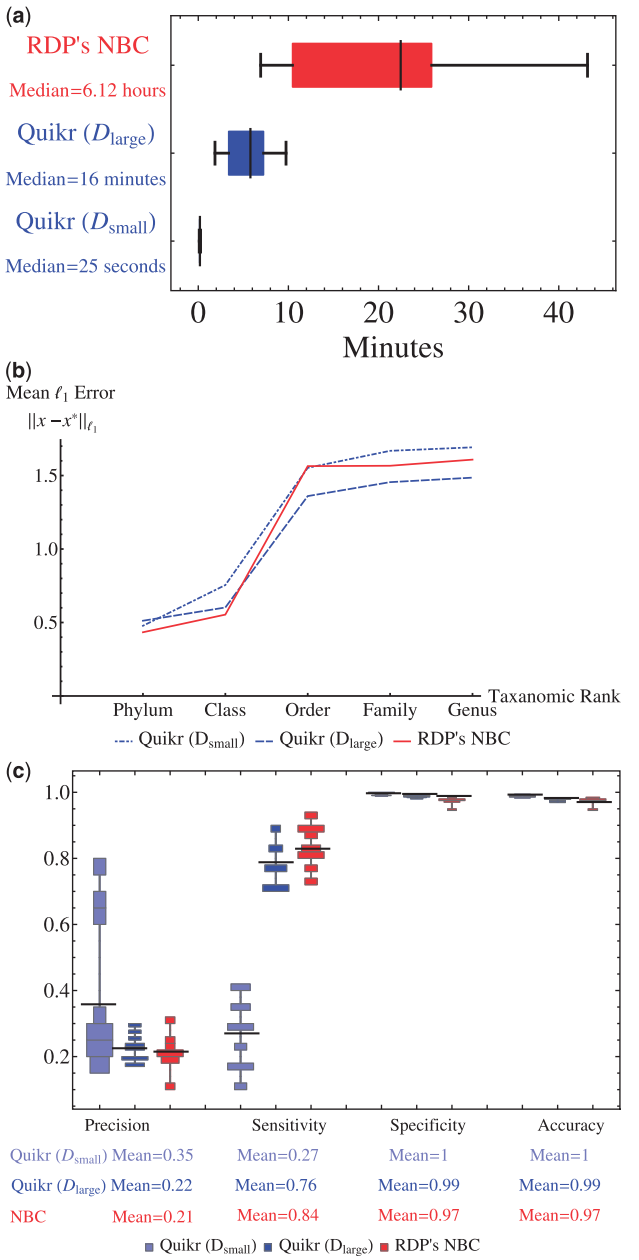

Fig. 3. Comparison of Quikr to RDP's NBC using the mock communities. Throughout, RDP's NBC version 10.28 with training set 7 was used. (**a**) Algorithm execution time for RDP's NBC and Quikr trained using D_{large} and D_{small}. Whiskers denote range of the data, vertical black bars designate the median and the boxes demarcate quantiles. (**b**) ℓ_1 error averaged over all the mock communities versus taxonomic rank for RDP's NBC and for Quikr trained using D_{small} and D_{large}. (**c**) Histogram densities for other error metrics at the genus level for RDP's NBC and Quikr trained using D_{small} and D_{large}. Horizontal black bars represent the mean

3.5 Cross-validation

To gauge how well the Quikr method will perform when the given sample contains 16S rRNA not in the database (simulating novelty), we performed a 5-fold cross-validation. Throughout the

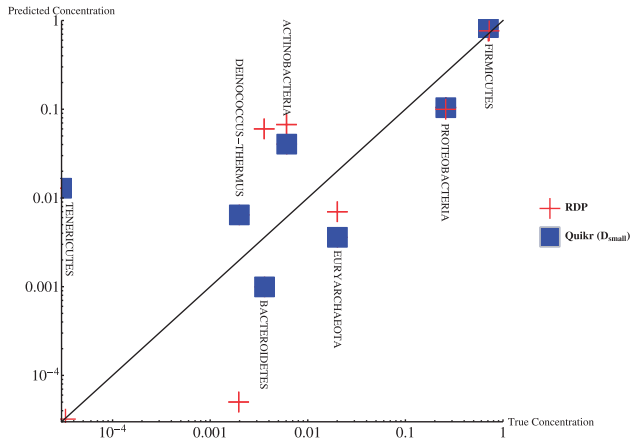

Fig. 4. Actual phyla concentration versus consensus predicted concentration (mean over all 12 samples) for the reconstruction of the mock communities via RDP's NBC and Quikr (with D_{small}). The mock communities contained 18S rRNA for *C.albicans* whose *k*-mer frequency vector was closest to the mean Tenericute *k*-mer frequency vector than any other phyla

cross-validation, the *k*-mer size was fixed at $k = 6$. The database D_{large} described in section 2.1 was partitioned into five disjoint sets and one-fifth was set aside as testing data with the remaining four-fifth used to form a new *k*-mer matrix as in section 2.1. Grinder (Angly *et al.*, 2012) parameters were then chosen to generate a test sample from the testing data. In particular, these parameters were chosen as follows: primers targeting the V1-V3 variable regions, read lengths normally distributed with mean 400 bp and standard deviation 50 bp, 800 000 total reads, exponential abundance model, diversity of 100 species, homopolymer error model as in Balzer (Balzer *et al.*, 2010) and 10% chimera percentage. The mean of each reconstruction metrics was then taken over the choice of which one-fifth was the testing data. Lastly, an average was taken over 10 iterates of this procedure. RDP's NBC was also used to classify the test samples.

Table 1 summarizes the results of this procedure for the ℓ_1 error metric. Because Quikr has a smaller mean ℓ_1 error and tighter variance, this demonstrates that even if the given sample contains novel sequences not present in the database, the Quikr method will still give high reconstruction fidelity down to the genus level. Similar results were observed for the remaining error metrics.

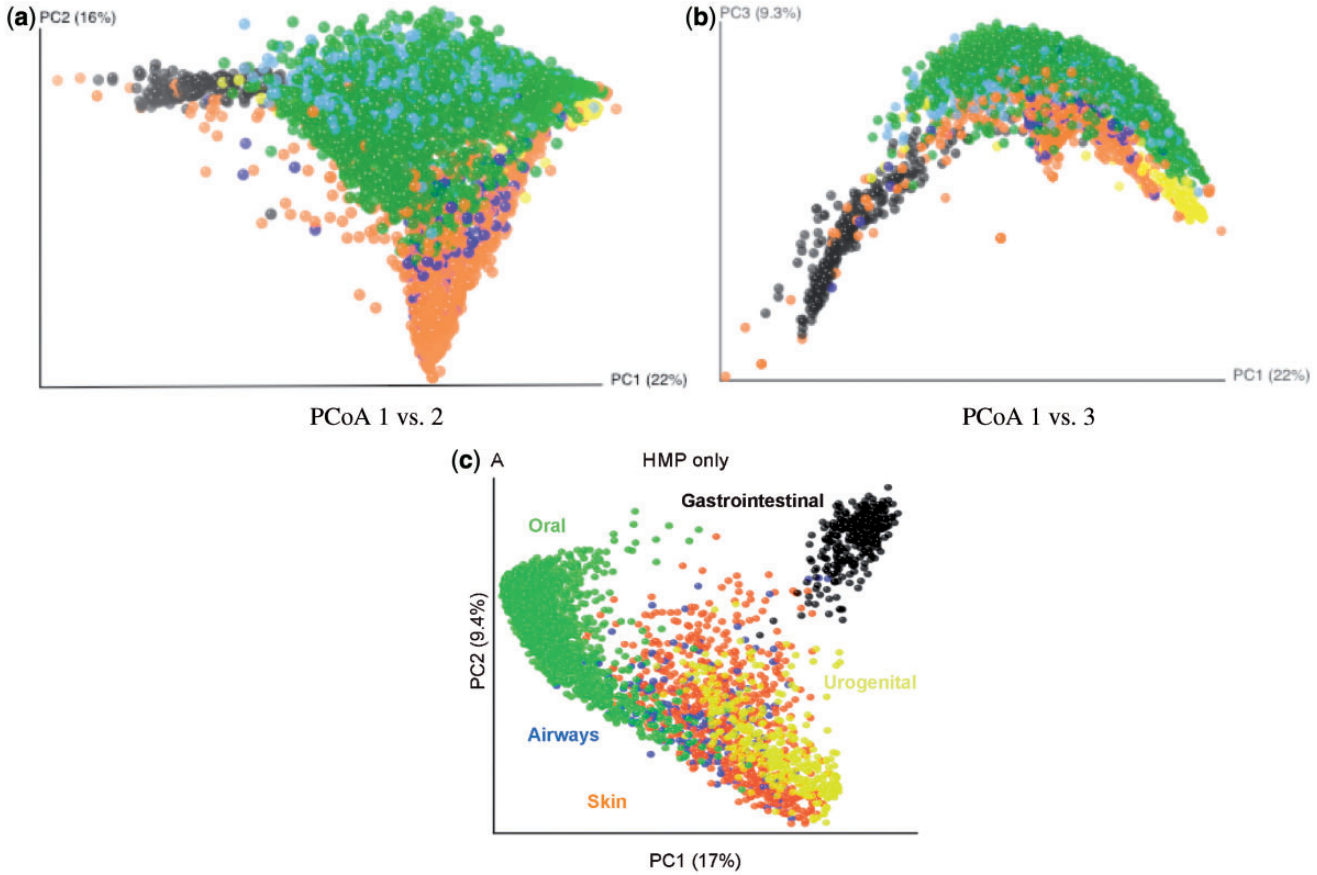

Fig. 5. (**a**, **b**) QIIME (weighted Unifrac) analysis using the Greengenes 91% OTU database, which took ~6 h for Quikr + QIIME complete analysis. Color legend: gut (black), oral (green), throat (light blue), skin (orange), nasal (bold blue) and urogenital (yellow). (**c**) Figure 1a from (Koren *et al.*, 2013)

Table 1. Results of 10 iterates of the 5-fold cross-validation procedure at the genus level (smaller values are better)

	Quikr	RDP's NBC
Mean ℓ_1 error $\pm$ variance	0.835 ± 0.00354	1.209 ± 0.0792

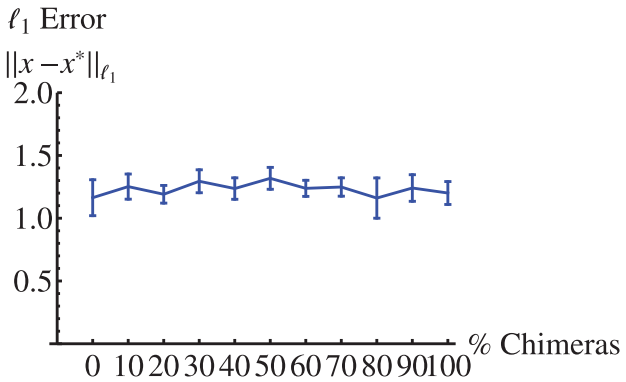

Fig. 6. Mean ℓ_1 error at the genus level for the Quikr method versus percentage of chimeras. Error bars depict standard deviation over 10 simulations

3.6 Chimeras

The presence of chimeras in an amplicon sample can significantly affect downstream analysis when using classification algorithms such as Bayesian classifiers (Ashelford *et al.*, 2005), and is possibly the culprit for overestimates of the so-called 'rare biosphere' (Edgar *et al.*, 2011). Identifying and removing chimeras is a computationally intensive and only partially solved problem (Edgar *et al.*, 2011, Haas *et al.*, 2011; Huber *et al.*, 2004; Quince *et al.*, 2011). It is therefore a significant advantage of the Quikr method that it is completely unaffected by the presence of chimeras. Quikr's unaffectedness by chimeras is due to the k-mer frequency of a chimera being well-estimated by the weighted sum of the k-mer frequencies of the constituent sequences that generated the chimera.

To present experimental evidence of this invariance, we selected Grinder (Angly *et al.*, 2012) parameters to be the same as in section 3.5, but varied the percentage of chimeras from 0 to 100% in 10% increments, with 10 simulations being performed at each increment. An ANOVA analysis resulted in $P = 0.927$, hence there is no statistically significant evidence that the slope of a linear regression deviates from zero. Figure 6 illustrates this fact by plotting the mean ℓ_1 error and standard deviation over the 10 simulations versus percent chimeras. Hence, it can be concluded that it is unnecessary to filter for chimeras before using the Quikr method.

4 DISCUSSION

Quikr represents a new paradigm in algorithms for bacterial community reconstruction. By leveraging ideas from compressive sensing, an entire sample can be analyzed quickly and accurately. Depending on how it is trained, Quikr can be used as either an extremely rapid, almost constant time, high-level community profiling tool or else (using a larger training database) a fast, extremely accurate technique. Besides improvements in speed, other advantages include the ability to use massive training databases (like D_{large}) that would be much too large for standard techniques (like RDP's NBC). Furthermore, Quikr is unaffected by the presence of chimeras, so the time-consuming chimera-removal step in standard analytic pipelines can be completely circumvented.

ACKNOWLEDGEMENTS

The authors thank Chris Cramer for writing the k-mer counting portion of the code and J. Calvin Morrison of Drexel University for the Python and C implementations. This work was initiated when D.K. was with Drexel University.

Funding: NSF (DMS-1120622); NSF (0931642 to D.K. in part).

Conflict of Interest: none declared.

REFERENCES

Amir,A. and Zuk,O. (2011) Bacterial community reconstruction using compressed sensing. *J. Comput. Biol.*, **18**, 1723–1741.

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **61**, 1–8.

Ashelford,K.E. *et al.* (2005) At Least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Biol.*, **71**, 7724–7736.

Balzer,S. *et al.* (2010) Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics (Oxford, England)*, **26**, i420–i425.

Benson,D.A. *et al.* (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Chen,S.S. *et al.* (1998) Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**, 33–61.

Clemente,J.C. *et al.* (2011) Flexible taxonomic assignment of ambiguous sequencing reads. *BMC Bioinformatics*, **12**, 8.

Cole,J.R. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.

DeSantis,T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environm. Microbiol.*, **75**, 5069–5072.

Edgar,R.C. *et al.* (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)*, **27**, 2194–2000.

Giongo,A. *et al.* (2010) TaxCollector: modifying current 16S rRNA databases for the rapid classification at six taxonomic levels. *Diversity*, **2**, 1015–1025.

Haas,B.J. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res.*, **21**, 494–504.

Huber,T. *et al.* (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics (Oxford, England)*, **20**, 2317–2319.

Jumpstart Consortium Human Microbiome Project Data Generation Working Group. (2012) Evaluation of 16S rRNA-based community profiling for human microbiome research. *PLoS One*, **7**, e39315.

Koren,O. *et al.* (2013) A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Comput. Biol.*, **9**, e1002863.

Lan,Y. *et al.* (2012) Using the RDP classifier to predict taxonomic novelty and reduce the search space for finding novel organisms. *PLoS One*, **7**, e32491.

Liu,Z. *et al.* (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.*, **38**, e120.

MATLAB. (2012) *The MathWorks, Inc.* Natick, MA, USA.

Meinicke,P. *et al.* (2011) Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics*, **27**, 1–7.

Quince,C. *et al.* (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 38.

Rosen,G. *et al.* (2008) Metagenome fragment classification using N-mer frequency profiles. *Adv. Bioinformatics*, **2008**, 205969.

Sayers,E.W. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.

Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–817.

The Human Microbiome Project Consortium. (2012) A framework for human microbiome research. *Nature*, **486**, 215–221.

Wang,C. and Zhang,D. (2011) A novel compression tool for efficient storage of genome resequencing data. *Nucleic Acids Res.*, **39**, 5–10.

Wang,Q. *et al.* (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, **73**, 5261–5267.