

Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex

Micah Hamady¹, Jeffrey J Walker², J Kirk Harris³, Nicholas J Gold² & Rob Knight⁴

We constructed error-correcting DNA barcodes that allow one run of a massively parallel pyrosequencer to process up to 1,544 samples simultaneously. Using these barcodes we processed bacterial 16S rRNA gene sequences representing microbial communities in 286 environmental samples, corrected 92% of sample assignment errors, and thus characterized nearly as many 16S rRNA genes as have been sequenced to date by Sanger sequencing.

Pyrosequencing¹ has the potential to revolutionize many sequencing efforts, including assessments of microbial community diversity^{2–4}. It eliminates the laborious step of producing clone libraries and generates hundreds of thousands of sequences in a single run. Two factors limit culture-independent 16S rRNA-based analysis of microbial community composition⁵ through pyrosequencing: each individual run is expensive, and splitting a single plate across multiple runs is difficult. One solution is barcoding, in which one adds a unique tag to each primer before PCR amplification^{6–8}. Because each sample is amplified with a known tagged primer, we can sequence an equimolar mixture of PCR-amplified DNA from each sample and assign sequences to samples based on the unique barcodes. This technique can be used to process as many as 25 samples in a single pyrosequencing run⁸.

Existing barcoding methods have limits both in the number of unique barcodes they use and in their ability to detect sequencing errors that change sample assignments (this robustness is especially important for sample assignment because the 5' end of the read is somewhat more error-prone⁹). We have developed a new set of barcodes based on error-correcting codes¹⁰, which are widely useful in devices ranging from cell phones to compact disc players. In this study, we chose a class of error-correcting codes called Hamming codes, which use a minimum amount of redundancy and are simple to implement using standard linear algebra techniques.

Other encoding schemes, such as Golay codes, may also prove useful. Briefly, Hamming codes, like all error-correcting codes, use the principle of redundancy and add redundant parity bits to transmit data over a noisy medium. Here we encoded sample identifiers with redundant parity bits, and 'transmitted' these sample identifiers as codewords. We encoded each base using 2 bits and used 8 bases for each codeword; hence we transmitted 16-bit codewords. Hamming codes use only a subset of the possible codewords, choosing those that lie at the center of multidimensional spheres (hyperspheres) in a binary subspace. Single bit errors fall within hyperspheres associated with each codeword, and thus we can correct them (**Fig. 1a**), but double bit errors do not, and thus we can detect but not correct them.

Let n be the total number of bits in the codeword, and k be the number of encoded bits of information. Hamming codes use $n - k$ bits of redundancy, and because not all 2^n possible codewords are used, there are 2^k valid, error-correcting codewords that form a k -dimensional subspace. The Hamming distance is the number of bits that differ between two vectors in this subspace, and the relevant parameter for error correction is the minimum Hamming distance. Let t be the radius of a sphere in this subspace where we can correct any change within this sphere. The error-correcting capability is the largest radius such that all Hamming spheres are disjoint: $t = \text{floor}((d_{\min} - 1) / 2)$, where d_{\min} is the minimum Hamming distance (**Fig. 1**). Thus, the minimum Hamming distance between codewords needed to correct a single error is 3.

Here we used Hamming codes to encode sample identifiers as DNA translations of each binary codeword using 2 bits per base. Thus, our 8-base codewords ($n = 16$) used 11 bits for sample identifiers ($k = 11$) and 5 bits of redundancy ($n - k = 5$). There were thus $2^{11} (= 2,048)$ possible 8-base codewords (for comparison, 4-base barcodes can encode up to 16 codewords, and 16-base barcodes can encode up to 67 million codewords, so the technique is readily scalable). To pick our maximal set of 1,544 barcodes (**Supplementary Data 1** online), we chose an encoding scheme for the four bases (A, T, C, G) that resulted in the largest number of valid 'candidate' barcodes. We then filtered these barcodes to optimize PCR and sequencing performance using the following criteria: G+C content of 40–60%, no consecutive triples of the same base, and no perfect self-complementarity or complementarity between the 8-base barcode and the primer. We wrote the decoding software in Python, based on an existing description of Hamming codes¹⁰ (see **Supplementary Data 2** online for a decoding example using this software).

To test these barcodes, we determined the bacterial composition of 286 environmental samples by PCR-amplifying, sequencing

¹Department of Computer Science, UCB 430 and ²Department of Molecular, Cellular and Developmental Biology, UCB 347, University of Colorado, Boulder, Colorado 80309, USA. ³Department of Pediatrics, University of Colorado Denver and Health Sciences Center, Children's Hospital, 13123 E. 16th Avenue Box B395, Aurora, Colorado 80045, USA. ⁴Department of Chemistry and Biochemistry, UCB 215, University of Colorado, Boulder, Colorado 80309, USA. Correspondence should be addressed to R.K. (rob.knight@colorado.edu).

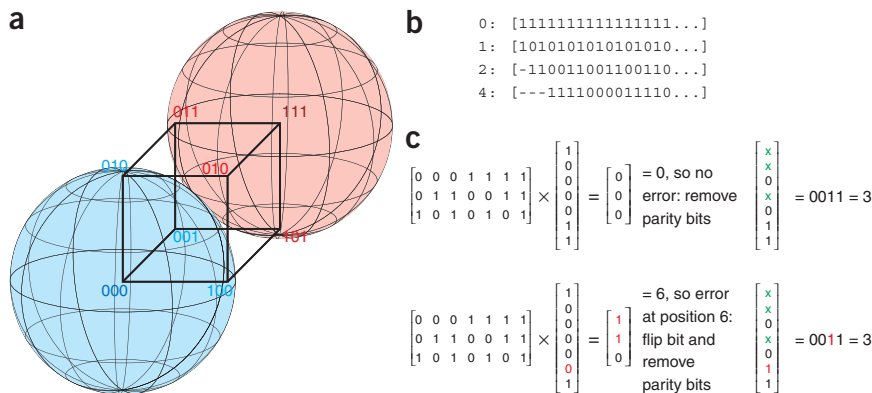


Figure 1 | Operation of Hamming error-correcting codes. **(a)** Consider a hypersphere centered at 000: any single-bit error (010, 001 and 100) falls within a radius of 1, and we can thus correct it. Likewise with the hypersphere centered at 111. **(b)** Regions of a codeword of length 16 (or longer) checked by parity bits at positions 0, 1, 2 and 4: bits that each position checks are marked with 1. **(c)** Example of decoding a 'received' codeword containing the binary value of 3 (0011) ($n = 7$, $k = 4$): the first case contains no errors; the second contains a single-bit error at position 6 that we can detect and correct.

and analyzing 681,688 16S rRNA gene sequences from a single sequencing run of the 454 Life Sciences Genome Sequencer FLX (Roche). We used 286 of the 1,544 candidate codewords to synthesize barcoded PCR primers to use in PCR to amplify a region (27F–338R) of the 16S rRNA gene that is optimal for phylogenetic analysis from pyrosequencing reads¹¹.

We extracted total DNA from samples of human lung, river water, the Guerrero Negro microbial mat, particles filtered from air and hot spring water using a modified bead-beating solvent extraction¹². For each sample, we amplified the 16S rRNA gene using a composite forward primer and a reverse primer containing a unique 8-base barcode used to tag each PCR product (**Supplementary Data 1**).

We performed four independent PCRs (**Supplementary Methods** online) for each of 286 samples, along with a no-template (water) negative control. For each sample, we combined the four replicate PCR products, purified them with Ampure magnetic purification beads (Agencourt), quantified them using the Quant-iT PicoGreen dsDNA Assay kit (Invitrogen) and a fluorospectrometer (Nanodrop ND3300). Then we created a master DNA pool by combining these 286 purified products in equimolar ratios to create a master DNA pool to a final concentration of 21.5 ng/μl. We sent this pool for pyrosequencing with primer A at 454 Life Sciences as described^{1,2}. After removal of low-quality sequences and trimming of primer sequences, 437,544 sequences remained, each representing ~240–280 bases of 16S rRNA sequence. We based the quality determination of each sequencing read on criteria previously described⁹.

We assigned each remaining sequence to a sample based on the barcodes, picked operational taxonomic units (OTUs) at 96% identity, aligned one sequence representing each of the 25,351 OTUs (in comparison, a recent study of 202 globally diverse environments identified only 21,752 OTUs at the 97% level¹³), built a phylogenetic tree, and clustered the samples based on their similarities in bacterial phylogenetic diversity with UniFrac^{14,15} (**Supplementary Methods**). The clustering (**Fig. 2**) correlated perfectly with sample type: all the lung samples clustered together, as did all the North American river samples, the microbial mat

samples, air samples, hot spring samples and two African river samples. The distribution of major lineages was as expected (**Supplementary Fig. 1** online). We analyzed 19 DNA samples in triplicate with three independent barcode primers, and in each case the replicate samples clustered together in the UniFrac analysis: of a total of 61 replicate samples, all but one pair clustered together (data not shown). This suggests that these barcoded primers amplified equivalently in PCR. We found that 1,345 sequences (0.3%) had decoding errors, of which 1,241 (92.2%) could be corrected to valid barcodes.

These results demonstrate that we could use the tagged barcoding strategy to obtain sequences from hundreds of samples in a single sequencing run and to perform phylogenetic analyses of microbial communities from pyrosequencing data. Here we

analyzed nearly as many 16S rRNAs as the total number determined to date by Sanger sequencing (although the tagged sequences from this study were much shorter, averaging only ~270 nucleotides of the ~1,500 nucleotides of the 16S rRNA gene). Our approach also provided several important advantages over other approaches. First, we could detect and correct errors in the barcodes, and could estimate the total error rate and eliminate possible mis-assignment. Second, our barcodes could encode more samples than using the four-nucleotide approach⁷, but required only 8 nucleotides rather than 20 or 44 (refs. 8 and 6, respectively), which is important when read lengths are limited. Third, we tagged only one

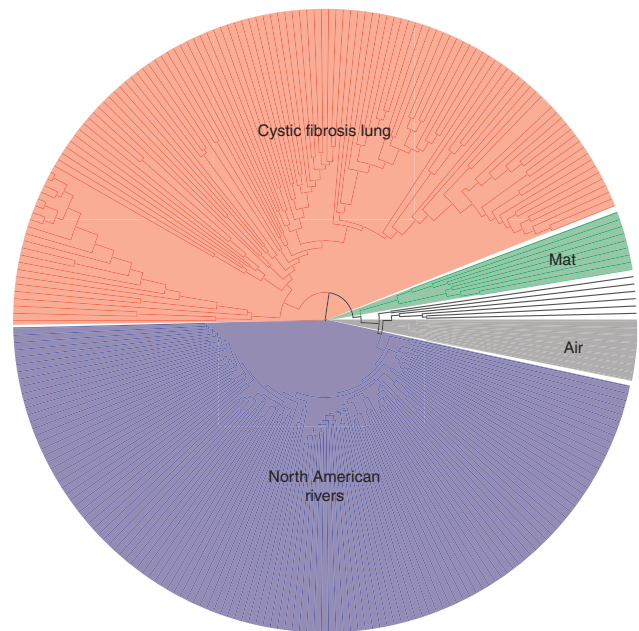


Figure 2 | UniFrac clustering by community was essentially perfect with sequences from pyrosequencing. Samples from cystic fibrosis lung, Guerrero Negro microbial mat, air and North American rivers cluster by environment type.

end of the sequence rather than both ends as in references 6 and 8. This feature is especially important for 16S rRNA sequencing, because variation in the length of variable regions in different species may preclude the second tag from being read. Our strategy, including alternative encoding schemes, should be useful for many applications: although there is a tradeoff between the number of samples per run and the number of sequences per sample, the average of ~1,500 sequences/sample in this study exceeded the number of sequences collected in all but the largest Sanger sequencing studies¹³. The combination of error-correcting barcodes and massively parallel sequencing will rapidly revolutionize our understanding of microbial habitats located throughout our biosphere as well as those associated with our human bodies.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank N. Pace, L. Gold and F. Accurso for support and encouragement, J.I. Gordon and R. Bushman for helpful discussions, and R. Ley, C. Lozupone and D. McDonald for feedback on the manuscript. This work was supported in part by the US National Institutes of Health–University of Colorado at Boulder Molecular Biophysics Training Program (T32GM065103), and grants from the Cystic Fibrosis Foundation and National Institutes of Health (U01 HL081335-01, P01DK078669).

AUTHOR CONTRIBUTIONS

M.H. and R.K. designed and implemented the analyses, and wrote the manuscript. J.J.W., J.K.H. and N.J.G. generated the 454 dataset.

Published online at <http://www.nature.com/naturemethods/>
Reprints and permissions information is available online at
<http://npg.nature.com/reprintsandpermissions>

1. Margulies, M. *et al.* *Nature* **437**, 376–380 (2005).
2. Sogin, M.L. *et al.* *Proc. Natl. Acad. Sci. USA* **103**, 12115–12120 (2006).
3. Huber, J.A. *et al.* *Science* **318**, 97–100 (2007).
4. Roesch, L.F.W. *et al.* *ISME J.* **1**, 283–290 (2007).
5. Pace, N.R. *Science* **276**, 734–740 (1997).
6. Binladen, J. *et al.* *PLoS ONE* **2**, e197 (2007).
7. Hoffmann, C. *et al.* *Nucleic Acids Res.* **35**, e91 (2007).
8. Parameswaran, P. *et al.* *Nucleic Acids Res.* **35**, e130 (2007).
9. Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L. & Welch, D.M. *Genome Biol.* **8**, R143 (2007).
10. Morelos-Zaragoza, R.H. *The Art of Error-Correcting Coding* (John Wiley & Sons, Hoboken, New Jersey, 2006).
11. Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D. & Knight, R. *Nucleic Acids Res.* **35**, e120 (2007).
12. Dojka, M.A., Hugenholtz, P., Haack, S.K. & Pace, N.R. *Appl. Environ. Microbiol.* **64**, 3869–3877 (1998).
13. Lozupone, C.A. & Knight, R. *Proc. Natl. Acad. Sci. USA* **104**, 11436–11440 (2007).
14. Lozupone, C., Hamady, M. & Knight, R. *BMC Bioinformatics* **7**, 371 (2006).
15. Lozupone, C. & Knight, R. *Appl. Environ. Microbiol.* **71**, 8228–8235 (2005).