QIIME allows analysis of high-throughput community sequencing data

To the Editor: High-throughput sequencing is revolutionizing microbial ecology studies. Efforts like the Human Microbiome Projects[1] and the US National Ecological Observatory Network[2] are helping us to understand the role of microbial diversity in habitats within our own bodies and throughout the planet.

Pyrosequencing using error-correcting, sample-specific barcodes allows hundreds of communities to be analyzed simultaneously in multiplex[3]. Integrating information from thousands of samples, including those obtained from time series, can reveal large-scale patterns that were inaccessible with lower-throughput sequencing methods. However, a major barrier to achieving such insights has been the lack of software that can handle these increasingly massive datasets. Although tools exist to perform library demultiplexing and taxonomy assignment[4,5], tools for downstream analyses are scarce.

Here we describe 'quantitative insights into microbial ecology' (QIIME; prounounced 'chime'), an open-source software pipeline built using the PyCogent toolkit[6], to address the problem of taking sequencing data from raw sequences to interpretation and database deposition. QIIME, available at http://qiime.sourceforge.net/, supports a wide range of microbial community analyses and visualizations that have been central to several recent high-profile studies, including network analysis, histograms of within- or between-sample diversity and analysis of whether 'core' sets of organisms are consistently represented in certain habitats. QIIME also provides graphical displays that allow users to interact with the data. Our implementation is highly modular and makes extensive use of unit testing to ensure the accuracy of results. This modularity allows alternative components for functionalities such as choosing operational taxonomic units (OTUs), sequence alignment, inferring phylogenetic trees and phylogenetic and taxon-based analysis of diversity within and between samples (including incorporation of third-party applications for many steps) to be easily integrated and benchmarked against one another (**Supplementary Fig. 1**).

We applied the QIIME workflow to a combined analysis of previously collected data (see **Supplementary Discussion**) for distal gut bacterial communities from conventionally raised mice, adult

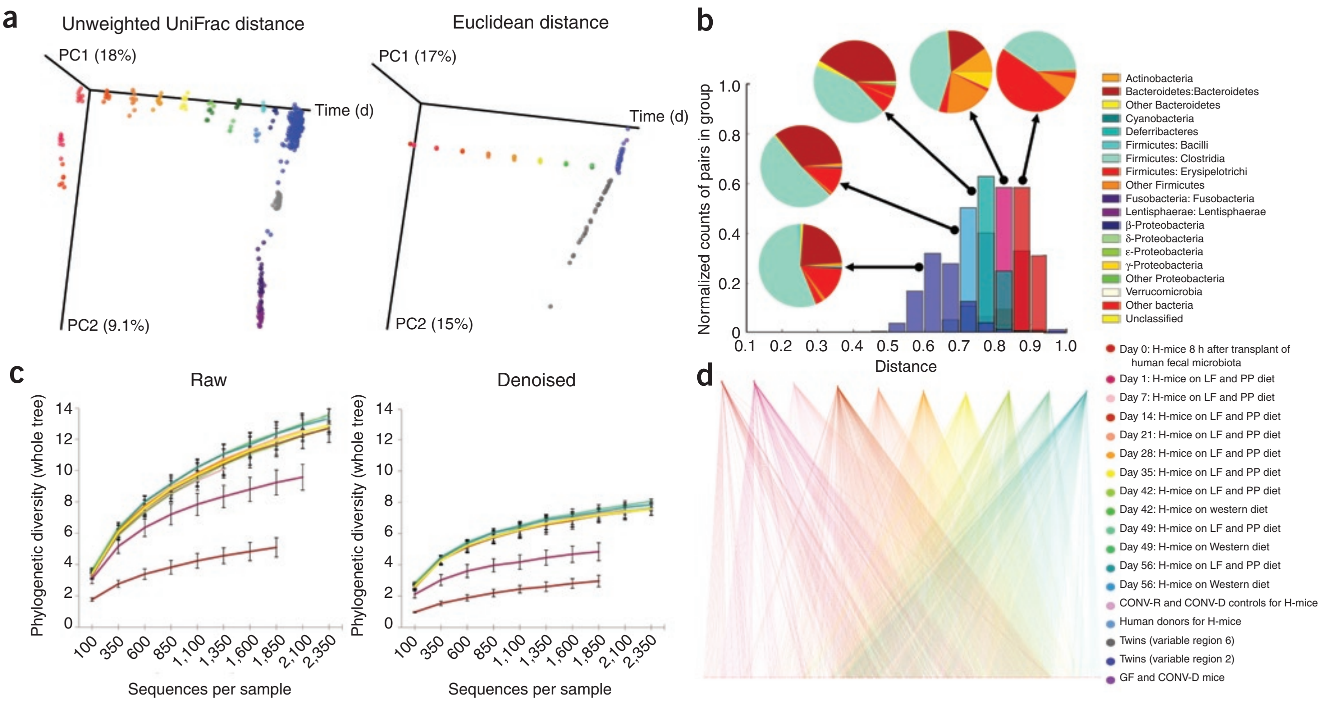

Figure 1 | QIIME analyses of the distal gut microbiotas of conventionally raised and conventionalized mice, gnotobiotic mice colonized with a human fecal gut microbiota (H-mice), and human adult mono- and dizygotic twins. (**a**) Principal coordinates analysis plots for mice, H-mice and twins. Colors correspond to separate samples by species and time point, and are consistent throughout the panels. (**b**) Unweighted UniFrac distance histograms between the data for fecal microbiota of human twins; human donors for the H-mice study; day 56 post-transplant H-mice on a low-fat (LF) and plant polysaccharide–rich (PP) diet; day 1 H-mice (LF and PP diet); and day 0 H-mice. Taxonomic classifications are presented at the class level. (**c**) Alpha diversity rarefaction plots of phylogenetic diversity for the H-mice samples. (**d**) OTU network connectivity of H-mice time series data. CONV-D, conventionalized mice; CONV-R, conventionally raised mice; and GF, germ-free mice.

human monozygotic and dizygotic twins and their mothers, and a time series study of adult germ-free mice after they received human fecal microbiota (**Fig. 1**, **Supplementary Table 1** and **Supplementary Discussion**). This analysis combined ten full 454 FLX runs and one partial run, totalling 3.8 million bacterial 16S rRNA sequences from previously published studies, including reads from different regions of the 16S rRNA gene.

QIIME is thus a robust platform for combining heterogeneous experimental datasets and for rapidly obtaining new insights about various microbial communities. Because QIIME scales to millions of sequences and can be used on platforms from laptops to high-performance computing clusters, we expect it to keep pace with advances in sequencing technology and to facilitate characterization of microbial community patterns ranging from normal variations to pathological disturbances in many human, animal and other environmental ecosystems.

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

ACKNOWLEDGMENTS
We thank our collaborators for their helpful suggestions on features, documentation and the manuscript, and our funding agencies for their commitment to open-source software. This work was supported in part by Howard Hughes Medical Institute and grants from the Crohn's and Colitis Foundation of America, the German Academic Exchange Service, the Bill and Melinda Gates Foundation, the Colorado Center for Biofuels and Biorefining and the US National Institutes of Health (DK78669, GM65103, GM8759, HG4872 and its ARRA supplement, HG4866, DK83981 and LM9451).

COMPETING FINANCIAL INTERESTS
The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

J Gregory Caporaso[1,12], Justin Kuczynski[2,12], Jesse Stombaugh[1,12], Kyle Bittinger[3], Frederic D Bushman[3], Elizabeth K Costello[1], Noah Fierer[4], Antonio Gonzalez Peña[5], Julia K Goodrich[5], Jeffrey I Gordon[6], Gavin A Huttley[7], Scott T Kelley[8], Dan Knights[5], Jeremy E Koenig[9], Ruth E Ley[9], Catherine A Lozupone[1], Daniel McDonald[1], Brian D Muegge[6], Meg Pirrung[1], Jens Reeder[1], Joel R Sevinsky[10], Peter J Turnbaugh[6], William A Walters[2], Jeremy Widmann[1], Tanya Yatsunenko[6], Jesse Zaneveld[2] & Rob Knight[1,11]

[1]Department of Chemistry and Biochemistry, University of Colorado, Boulder, Colorado, USA. [2]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, Colorado, USA. [3]Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [4]Cooperative Institute for Research in Environmental Sciences and Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado, USA. [5]Department of Computer Science, University of Colorado, Boulder, Colorado, USA. [6]Center for Genome Sciences, Washington University School of Medicine, St. Louis, Missouri, USA. [7]Computational Genomics Laboratory, John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia. [8]Department of Biology, San Diego State University, San Diego, California, USA. [9]Department of Microbiology, Cornell University, Ithaca, New York, USA. [10]Luca Technologies, Golden, Colorado, USA. [11]Howard Hughes Medical Institute, Boulder, Colorado, USA. [12]These authors contributed equally to this work.
e-mail: rob.knight@colorado.edu

PUBLISHED ONLINE 11 APRIL 2010; DOI:10.1038/NMETH.F.303

1. National Institutes of Health Human Microbiome Project Working Group *et al.* *Genome Res.* **19**, 2317–2323 (2009).
2. Hopkin, M. *Nature* **444**, 420–421 (2006).
3. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. & Knight, R. *Nat. Methods* **5**, 235–237 (2008).
4. Cole, J.R. *et al. Nucleic Acids Res.* **37**, D141–D145 (2009).
5. Schloss, P.D. *et al. Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
6. Knight, R. *et al. Genome Biol.* **8**, R171 (2007).

Intensity normalization improves color calling in SOLiD sequencing

To the Editor: Applied Biosystems' SOLiD system[1] is a commonly used massively parallel DNA sequencing platform for applications from genotyping and structural variation analysis[1] to transcriptome quantification and reconstruction[2]. Like other sequencing technologies, it measures fluorescence intensities from dye-labeled molecules to determine the sequence of DNA fragments. Ultimately, sequences are determined by complicated statistical manipulations of noisy intensity measurements, and systematic biases may mislead downstream analysis[3]. Several proposed methods improve base calling and quality metrics for other sequencing technologies[3–5], and we now present Rsolid, software implementing an intensity normalization strategy for the SOLiD platform that substantially improves yield and accuracy at small computational costs (6% increase in total matches, 13% increase in perfect matches, 5% reduced error rate and a substantial reduction in false positive single-nucleotide polymorphism (SNP) calls in an *Escherichia coli* genomic DNA sample).

In the SOLiD system, the proportions of color calls across sequencing cycles are extremely variable (**Fig. 1a**), even though they should be equal across sequencing cycles and proportional to the dinucleotide content of the library (**Supplementary Methods**). This bias can be traced to the fluorescence intensity measurements used to make the color calls (**Supplementary Fig. 1**). The distributions of intensities are similar across channels in early sequencing cycles, but a color bias starts to appear in later cycles. The Rsolid method uses a simple and computationally efficient procedure to normalize the color-channel

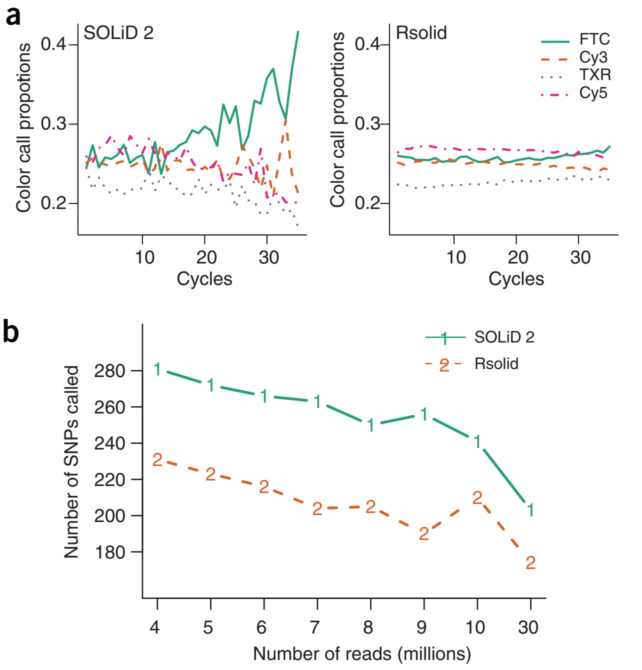

Figure 1 | Effect of normalization on color proportions and SNP calling. (**a**) Color proportions in sample of *E. coli* genomic DNA on each sequencing cycle. Color calls as reported by the SOLiD 2 system (left) and after normalization by Rsolid (right). FTX, TXR, Cy3 and Cy5 are dyes used by SOLiD. (**b**) Number of false positive SNPs called in *E. coli* at various coverage. After normalization, fewer SNPs were called even at high coverage (30 M reads correspond to ~100-fold coverage).