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

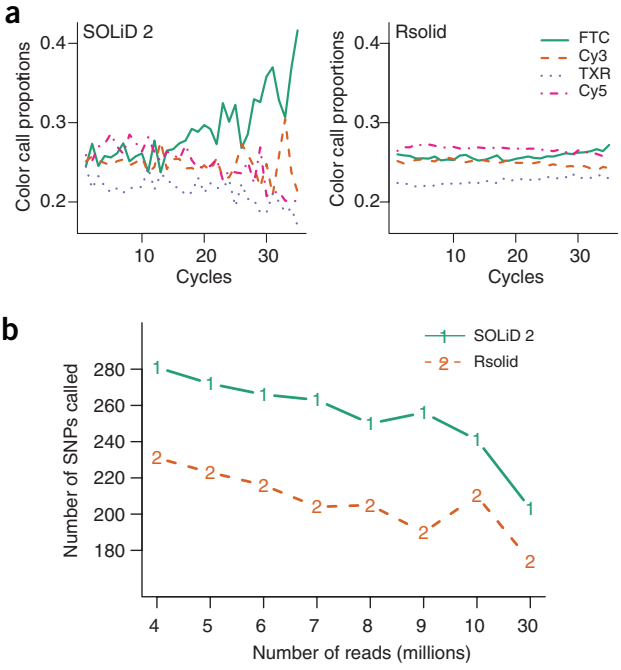

Figure 1 | Effect of normalization on color proportions and SNP calling. (**a**) Color proportions in sample of *E. coli* genomic DNA on each sequencing cycle. Color calls as reported by the SOLiD 2 system (left) and after normalization by Rsolid (right). FTX, TXR, Cy3 and Cy5 are dyes used by SOLiD. (**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).

Table 1 | Improvement in accuracy and mapping

Metric	*E. coli*	*H. sapiens*
Total mapped reads	+6.35%	+4.43%
Perfectly mapped reads	+12.97%	+7.21%
Uniquely mapped reads	+6.37%	+7.24%
Overall errors per mapped read	−4.92%	−1.82%
Valid adjacent errors per mapped read	−6.42%	−2.20%

These samples were processed in two different laboratories, with independent library preparations and sequencing machines.

intensity distributions while taking into account dinucleotide frequencies (**Supplementary Methods**). This ensures that the intensity distributions across channels are comparable in later cycles, as they are in earlier cycles (**Supplementary Fig. 1**), removing a substantial amount of the color-call bias seen in later cycles (**Fig. 1a**).

We report results on *E. coli* and *Homo sapiens* genomic DNA samples, processed in independent laboratories and machines. Mapping and accuracy statistics were substantially improved after normalization (**Table 1** and **Supplementary Tables 1–4**). We observed a 2–6% reduction in the rate of valid adjacent color errors, which is particularly important because SOLiD's two-base encoding cannot correct for this type of error. This manifests as a substantial reduction in the number of false positive SNP calls made (**Fig. 1b**).

The recently released SOLiD 3plus system, using smaller beads, produces as much as 10 times more intensity data. Data from this new system have similar statistical properties (data not shown) and our normalization algorithm scaled as expected (**Supplementary Methods**). Whereas increased yield is less of a concern with the new system, the need for accurate color calls is still of paramount importance in applications such as genotyping, bisulfite sequencing and assembly, in which single base-resolution data are required. There is still a need for accurate color calls to avoid false positive variant calls even at high coverage (**Fig. 1b**).

Rsolid is publically available from http://rafalab.jhsph.edu/Rsolid and runs on the R computing environment, making it straightforward to include in existing data pipelines.

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

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Hao Wu, Rafael A Irizarry & Héctor Corrada Bravo

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
e-mail: rafa@jhu.edu or hcorrada@gmail.com

1. McKernan, K.J., *et al. Genome Res.* **19,** 1527–1541 (2009).
2. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
3. Bravo, H.C. & Irizarry, R.A. *Biometrics* advance online publication, 13 November 2009 (doi:10.1111/j.1541-0420.2009.01353.x).
4. Quinlan, A.R., Stewart, D.A., Stromberg, M.P. & Marth, G. *Nat. Methods* **5**, 179–181 (2008).
5. Kao, W.C., Stevens, K. & Song, Y.S. *Genome Res.* **19**, 1884–1895 (2009).
6. Hayden, E.C. *Nature* **451**, 378–379 (2008).