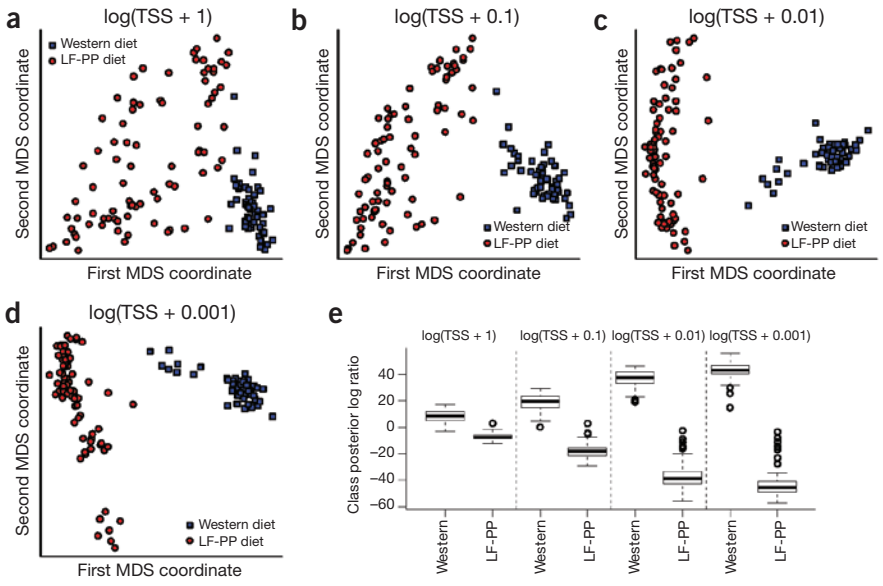

Figure 1 | Effect of pseudocount choice on total-sum scaling (TSS) normalization. (**a**–**d**) Clustering analysis using TSS with pseudocounts exponentially decreasing from 1 to 0.001. (**e**) Corresponding class posterior log ratios for each pseudocount value.

data (for example, 16S metagenomics data). Specifically, they argue that the "conceptually simpler" total-sum scaling (TSS) normalization method should not be dismissed by unfair characterization of its performance. In our figure, log transformation was applied to data normalized with our cumulative-sum scaling (CSS) method but not to TSS-normalized data. However, we stated clearly in the Online Methods section that CSS-normalized data were log transformed, and we showed the performance of log-transformed TSS-normalized data as Supplementary Figure 2 in our paper[2].

The purpose of the analysis reported in Figure 1 of our paper[2] was to compare the preprocessing procedure proposed in our paper (CSS + log transform) to established practices in the field. The prevalent practice in metagenomics is to use TSS, without transformation, as we reported. Although we agree that a thorough study of the impact of normalization is necessary for the field, this was not the purpose of our paper.

We used a continuity correction of 1 when applying a log transform to preserve the sparsity of metagenomic data because this is an important intrinsic characteristic of the data, as opposed to a mere necessity for modeling as suggested by Costea *et al.*[1]. They argue that TSS should not be dismissed on the basis of this analysis because it achieves similar clustering separation after log transformation with a carefully selected pseudocount. We are not the first to describe issues with TSS (for example, see papers in the RNA-seq literature cited from our paper[2]) and provide further evidence that its use in common practice is problematic. Although conceptual simplicity is a good characteristic of a data analysis tool, its use without further validation and testing is unwarranted. Advocating a log transform with a carefully chosen pseudocount is no longer conceptually simple. For instance, as Costea *et al.* show in their Supplementary Figure 2, TSS-normalized data are very sensitive to the choice of pseudocount, as opposed to other normalization procedures (including our CSS method). To complete their analysis, we show the effect of this pseudocount on clustering analysis (Fig. 1).

In general, Costea *et al.*[1] do raise an important point: thorough studies of the effects of preprocessing on downstream analyses (of which clustering and differential abundance testing are two) are sorely needed in the field of large-scale metagenomics studies. This is especially important for TSS normalization, which is still in wide use in the field and which has not been thoroughly analyzed and tested; its adoption is driven by its conceptual simplicity, not by principled analysis. Such a study was beyond the scope of our paper and is insufficiently addressed by a commentary on our paper. We look forward to a thorough treatment of these issues in an independent full publication.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Joseph N Paulson[1,2], Héctor Corrada Bravo[2,3] & Mihai Pop[2,3]

[1]Graduate program in Applied Mathematics, Statistics, and Scientific Computation, University of Maryland, College Park, Maryland, USA. [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [3]Department of Computer Science, University of Maryland, College Park, Maryland, USA.
e-mail: hcorrada@umiacs.umd.edu or mpop@umiacs.umd.edu

1. Costea, P.I., Zeller, G., Sunagawa, S. & Bork, P. *Nat. Methods* **11**, 359 (2014).
2. Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. *Nat. Methods* **10**, 1200–1202 (2013).

Single-cell *in situ* RNA profiling by sequential hybridization

To the Editor: In our previous paper, Lubeck and Cai[1], we used super-resolution microscopy to resolve a large number of mRNAs in single cells. In this Correspondence, we present a sequential barcoding scheme to multiplex different mRNAs.

Here, the mRNAs in cells are barcoded by sequential rounds of hybridization, imaging and probe stripping (**Fig. 1a** and **Supplementary Fig. 1**). As the transcripts are fixed in cells, the corresponding fluorescent spots remain in place during multiple rounds of hybridization and can be aligned to read out a fluorophore sequence. This sequential barcode is designed to uniquely identify an mRNA.

During each round of hybridization, we targeted each transcript with a set of fluorescence *in situ* hybridization (FISH) probes labeled with a single type of fluorophore. We imaged the sample and then treated it with DNase I to remove the FISH probes. In a subsequent round the mRNA was hybridized with the same FISH probes but labeled with a different dye. The number of barcodes available scales as F^N, where F is the number of fluorophores and N is the number of hybridization rounds. For example, with four dyes, eight rounds of hybridization can cover the entire transcriptome ($4^8 = 65{,}536$). As a demonstration, we barcoded 12 genes in single yeast cells with four dyes and two rounds of hybridization ($4^2 = 16$, with

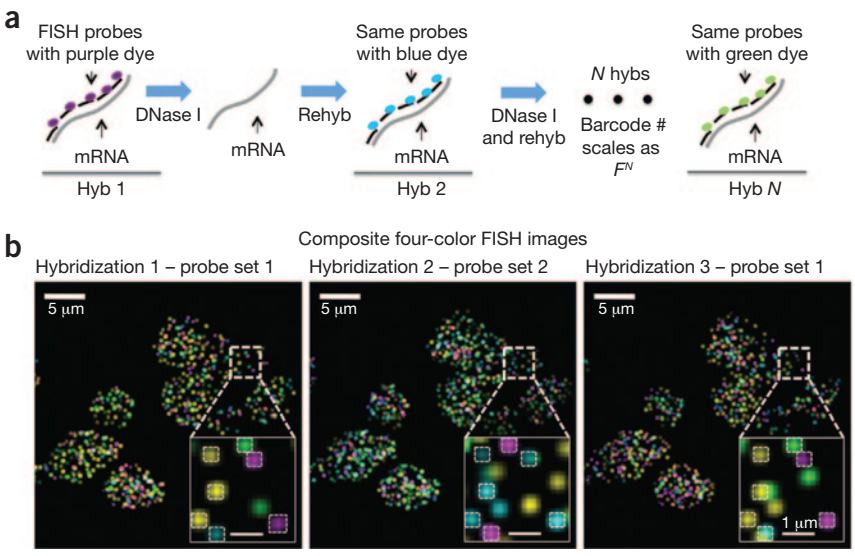

Figure 1 | Sequential barcoding. (**a**) Schematic of sequential barcoding. In each round of hybridization, 24 probes are hybridized on each transcript, imaged and then stripped by DNase I treatment. The same probe sequences are used in different rounds of hybridization (hyb), but probes are coupled to different fluorophores. (**b**) Composite four-color FISH data from three rounds of hybridizations on multiple yeast cells. Twelve genes are encoded by two rounds of hybridization, with the third hybridization using the same probes as hybridization 1. The boxed regions are magnified in the bottom right corner of each image. Spots colocalizing between hybridizations are detected (as outlined in insets) and have their barcodes extracted. Spots without colocalization are due to nonspecific binding of probes in the cell as well as mishybridization. The number of instances of each barcode can be quantified to provide the abundances of the corresponding transcripts in single cells.

four barcodes left out). We first immobilized cells on glass surfaces (**Supplementary Methods**). The DNA probes were hybridized, imaged and then removed by DNase I treatment (88.5% ± 11.0% efficiency (± standard deviation); **Supplementary Fig. 2** and **Supplementary Note**). The remaining signal was photobleached (**Supplementary Fig. 3**). Even after six hybridizations, mRNAs were observed at 70.9% ± 21.8% of the original intensity (**Supplementary Fig. 4**). We observed that 77.9% ± 5.6% of the spots that colocalized in the first two hybridizations also colocalized with the third hybridization (**Fig. 1b** and **Supplementary Figs. 5** and **6**). We quantified the mRNA abundances by counting the occurrence of corresponding barcodes in the cell ($n = 37$ cells; **Supplementary Figs. 7** and **8**). We also show that mRNAs can be stripped and rehybridized efficiently in adherent mammalian cells (**Supplementary Figs. 9** and **10**).

Sequential barcoding has many advantages. First, it scales up quickly; with even two dyes the coding capacity is in principle unlimited. Second, during each hybridization, all available FISH probes against a transcript can be used, thereby increasing the brightness of the FISH signal. Last, barcode readout is robust, enabling full *z* stacks on native samples.

This barcoding scheme is conceptually akin to sequencing transcripts in single cells with FISH. In contrast with the technique used by Ke *et al.*[2], our method takes advantage of the high hybridization efficiency of FISH (>95% of the mRNAs are detected[1,3]) and the fact that base-pair resolution is usually not needed to uniquely identify a transcript. We note that FISH probes can also be designed to resolve a large number of splice isoforms and single-nucleotide polymorphisms[3], as well as chromosome loci[4], in single cells. In combination with our previous report of super-resolution FISH[1], the sequential barcoding method will enable the transcriptome to be directly imaged at single-cell resolution in complex samples such as brain tissue.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2892).

ACKNOWLEDGMENTS
This work is funded by US National Institutes of Health single-cell analysis program award R01HD075605.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Eric Lubeck[1,2], Ahmet F Coskun[1,2], Timur Zhiyentayev[1], Mubhij Ahmad[1] & Long Cai[1]

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California, USA. [2]These authors contributed equally to this work.
e-mail: lcai@caltech.edu

1. Lubeck, E. & Cai, L. *Nat. Methods* **9**, 743–748 (2012).
2. Ke, R. *et al. Nat. Methods* **10**, 857–860 (2013).
3. Levesque, M.J., Ginart, P., Wei, Y. & Raj, A. *Nat. Methods* **10**, 865–867 (2013).
4. Levesque, M.J. & Raj, A. *Nat. Methods* **10**, 246–248 (2013).

MutationTaster2: mutation prediction for the deep-sequencing age

To the Editor: The majority of the gene variants discovered by next-generation sequencing (NGS) projects are either intronic or synonymous. These variants are difficult to interpret because their effects on protein expression and function tend to be less obvious than those of missense or nonsense variants. Here we present MutationTaster2 (http://www.mutationtaster.org/), the latest version of our web-based software MutationTaster[1], which evaluates the pathogenic potential of DNA sequence alterations. It is designed to predict the functional consequences of not only amino acid substitutions but also intronic and synonymous alterations, short insertion and/or deletion (indel) mutations and variants spanning intron-exon borders.

MutationTaster2 includes all publicly available single-nucleotide polymorphisms (SNPs) and indels from the 1000 Genomes Project[2] (hereafter referred to as 1000G) as well as known disease variants from ClinVar[3] and HGMD Public[4]. Alterations found more than four times in the homozygous state in 1000G or in HapMap[5] are automatically regarded as neutral. Variants marked as pathogenic in ClinVar are automatically predicted to be disease causing, and the disease phenotype is displayed. We have integrated tests for regulatory features, including data from the ENCODE project[6] and JASPAR[7], and score the evolutionary conservation around DNA variants (**Supplementary Methods**). To reduce the number of false positive splice-site