

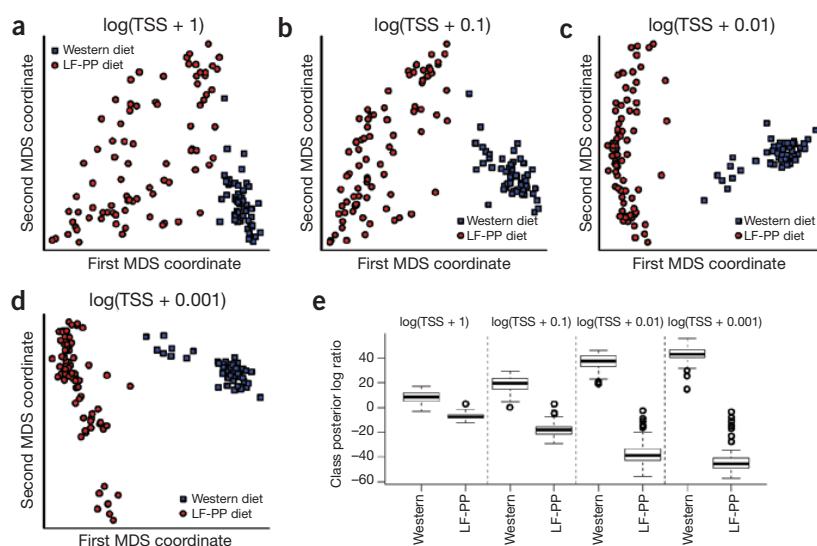


Figure 1 | Effect of pseudocount choice on total-sum scaling (TSS) normalization. (a–d) Clustering analysis using TSS with pseudocounts exponentially decreasing from 1 to 0.001. (e) Corresponding class posterior log ratios for each pseudocount value.

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 paper².

The purpose of the analysis reported in Figure 1 of our paper² 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 paper²) 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}

¹Graduate program in Applied Mathematics, Statistics, and Scientific Computation, University of Maryland, College Park, Maryland, USA. ²Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. ³Department 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).

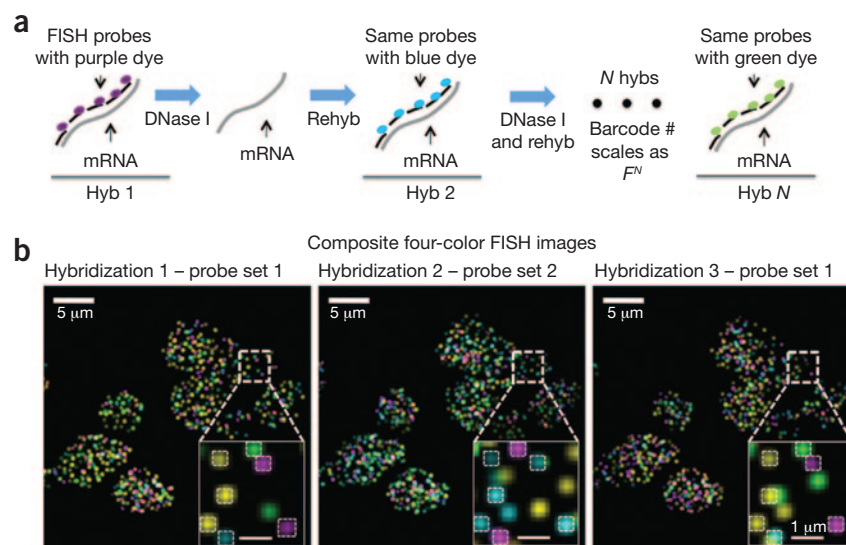
Single-cell *in situ* RNA profiling by sequential hybridization

To the Editor: In our previous paper, Lubeck and Cai¹, 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 ($88.5\% \pm 11.0\%$ efficiency ($\pm$ standard deviation); **Supplementary Fig. 2** and **Supplementary Note**). The remaining signal was photobleached (**Supplementary Fig. 3**). Even after six hybridizations, mRNAs were observed at $70.9\% \pm 21.8\%$ of the original intensity (**Supplementary Fig. 4**). We observed that $77.9\% \pm 5.6\%$ 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 polymorphisms³, as well as chromosome loci⁴, in single cells. In combination with our previous report of super-resolution FISH¹, 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 Zhiyentayev¹, Mubhij Ahmad¹ & Long Cai¹

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

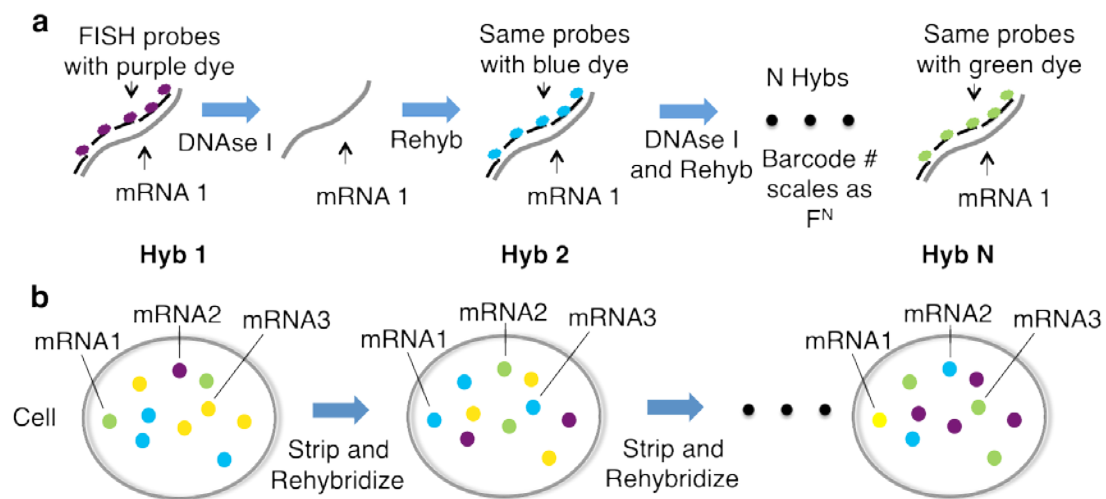
1. Lubeck, E. & Cai, L. *Nat. Methods* **9**, 743–748 (2012).
2. Ke, R. *et al. Nat. Methods* **10**, 857–860 (2013).
3. Levesque, M.J., Ginart, P., Wei, Y. & Raj, A. *Nat. Methods* **10**, 865–867 (2013).
4. Levesque, M.J. & Raj, A. *Nat. Methods* **10**, 246–248 (2013).

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 MutationTaster¹, 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.

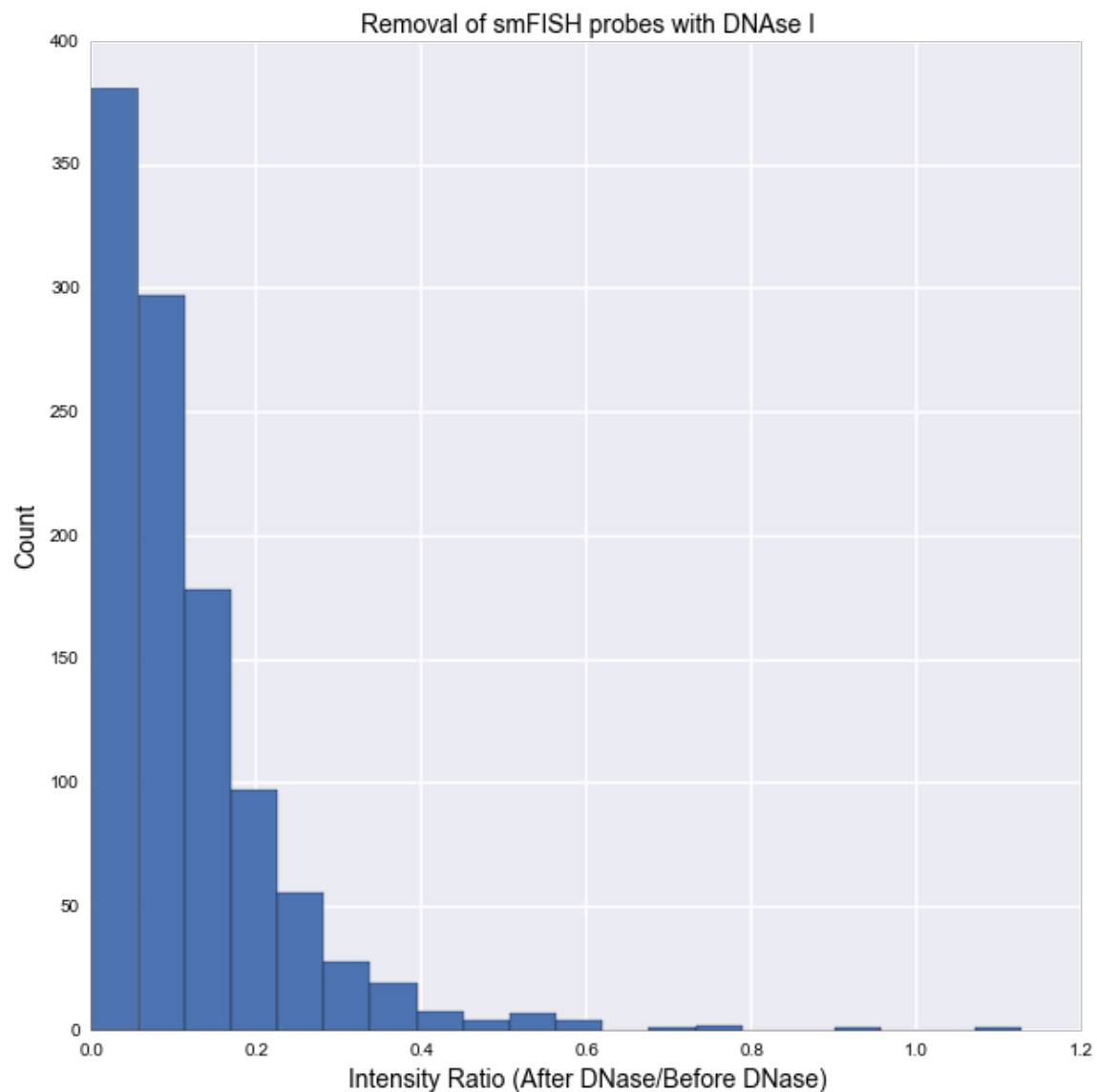
MutationTaster2 includes all publicly available single-nucleotide polymorphisms (SNPs) and indels from the 1000 Genomes Project² (hereafter referred to as 1000G) as well as known disease variants from ClinVar³ and HGMD Public⁴. Alterations found more than four times in the homozygous state in 1000G or in HapMap⁵ 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 project⁶ and JASPAR⁷, and score the evolutionary conservation around DNA variants (**Supplementary Methods**). To reduce the number of false positive splice-site

Supplementary Figure 1



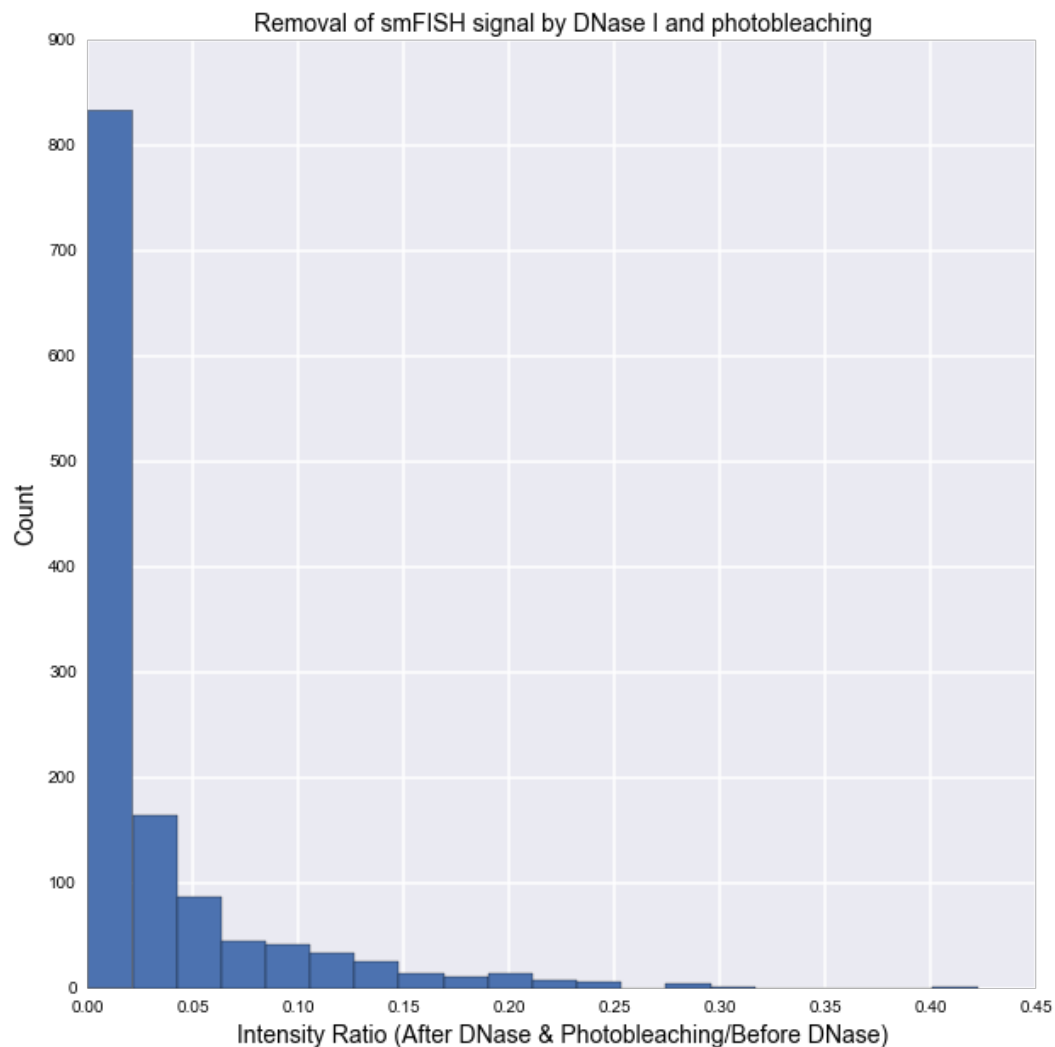
(A) Schematic of sequential hybridization and barcoding. **(B)** Schematic of the FISH images of the cell. In each round of hybridization, the same spots are detected, but the dye associated with the transcript changes. The identity of the mRNA is encoded in the temporal sequence of dyes hybridized.

Supplementary Figure 2



DNase I efficiently removes smFISH probes bound to mRNA. Spots were imaged before and after a 4 hour DNase I treatment in anti-bleaching buffer. The mean, median and STD of the intensity ratio after treatment are 11.5%, 8.3% and 11%. The ratio of the spot intensities after and before DNase I treatment is plotted for each spot. $n = 1084$ spots.

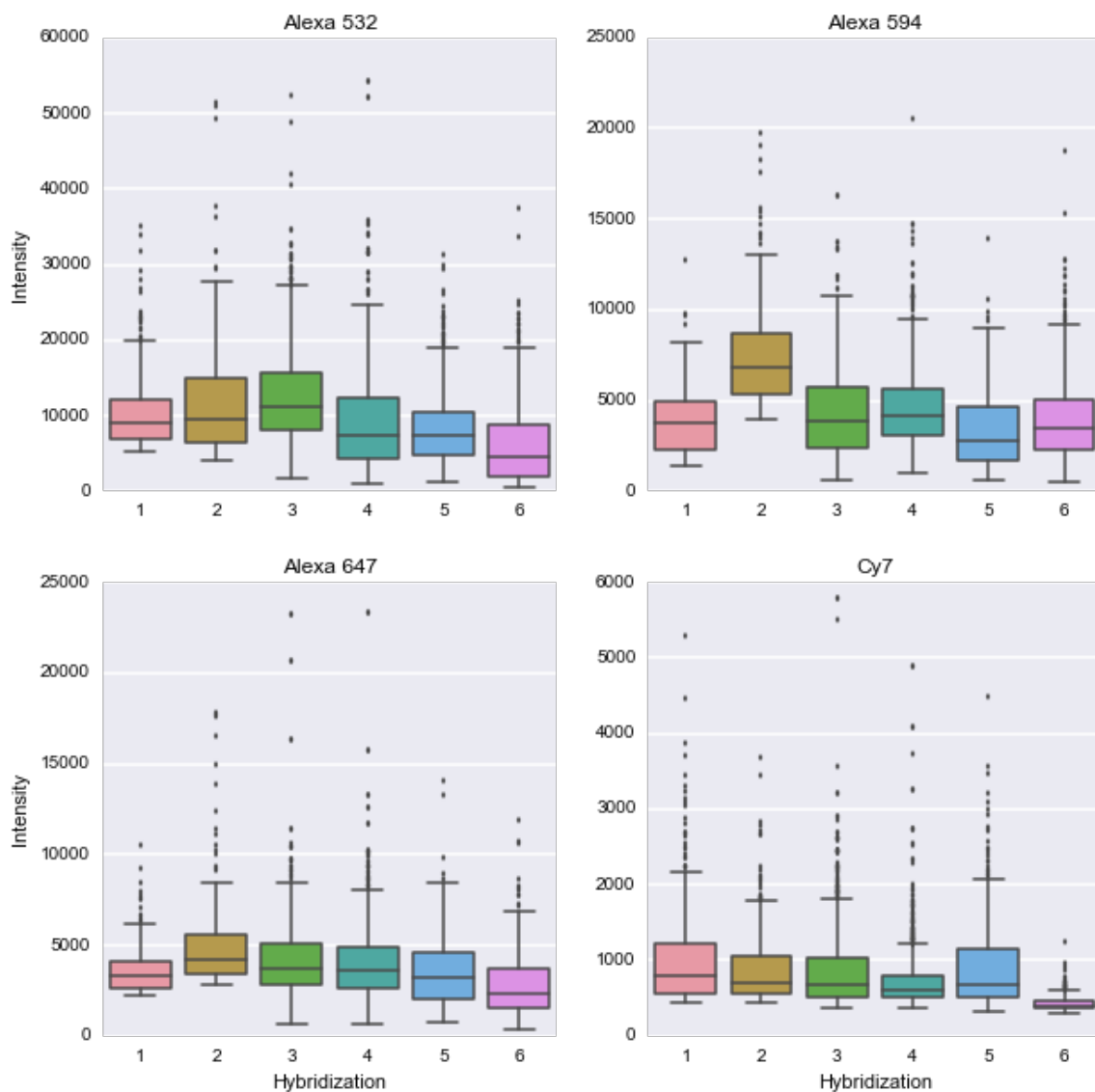
Supplementary Figure 3



Photobleaching removes residual intensity following DNase I treatment. Spots were bleached by 10 seconds of excitation following a 4 hour DNase I treatment. The mean, median and STD of the intensity ratio after bleaching are 0.03%, 0.01% and 0.049%. The ratio of the spot intensities after and before DNase I treatment is plotted for each spot. n = 1286 spots.

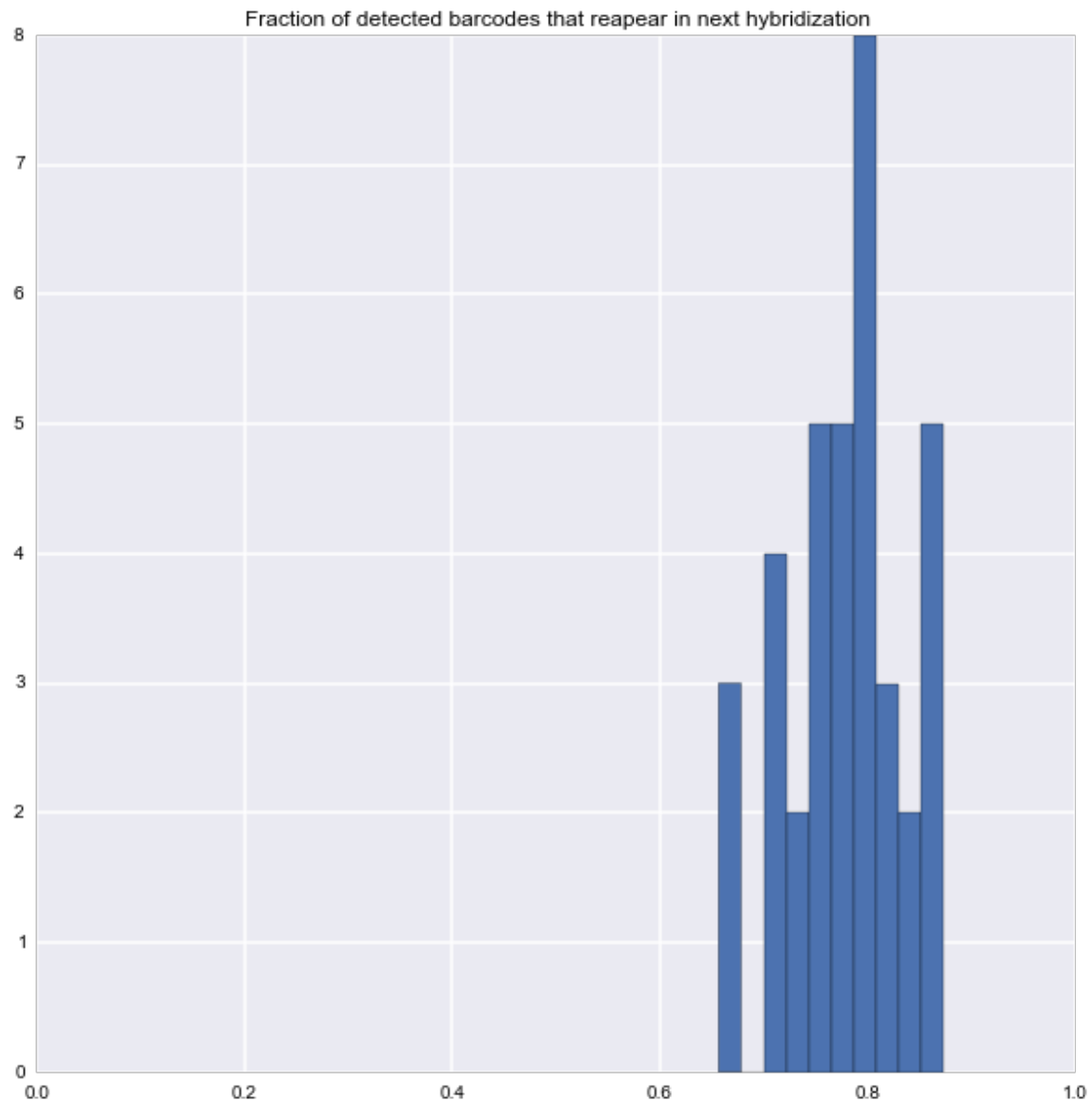
Supplementary Figure 4

Intensity v. Hybridization



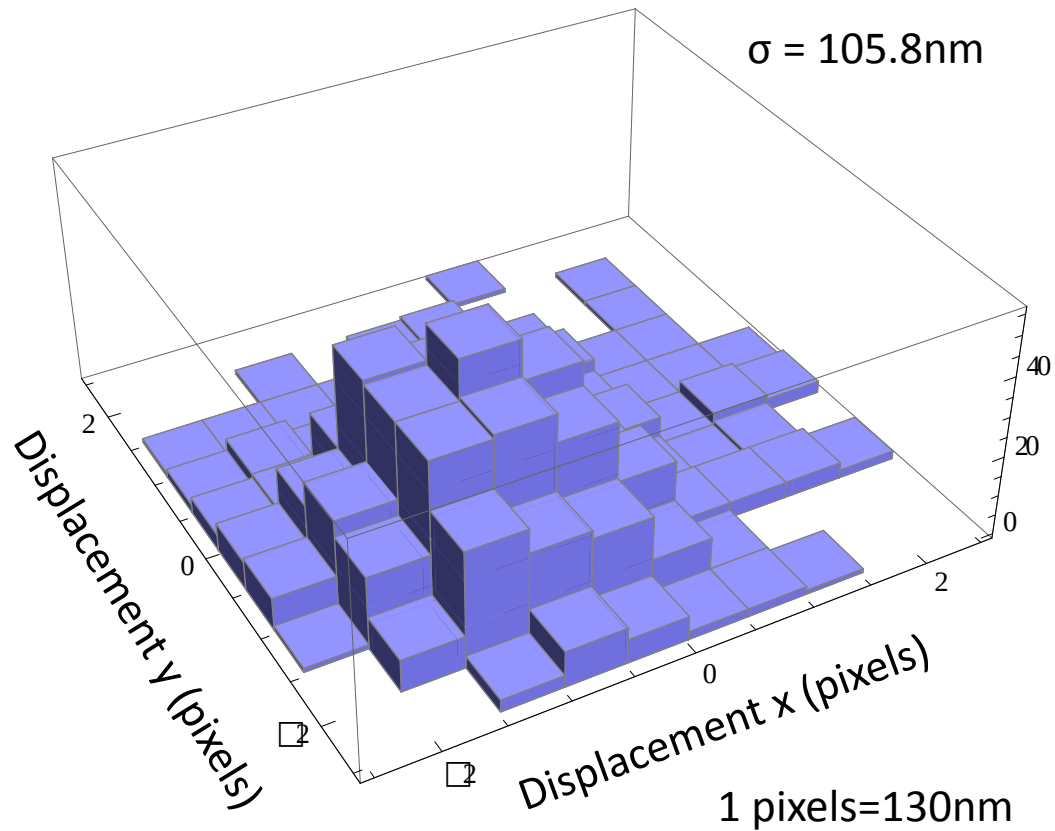
mRNAs are stable over multiple rounds of re-hybridization. The intensity distributions of smFISH spots are plotted over 6 hybridizations. Two hybridizations were repeated 3 times to make 6 total hybridizations. Spots were identified by their co-localization with spots in the next identical hybridization. For each boxplot the number of spots counted was between 191 and 1337.

Supplementary Figure 5



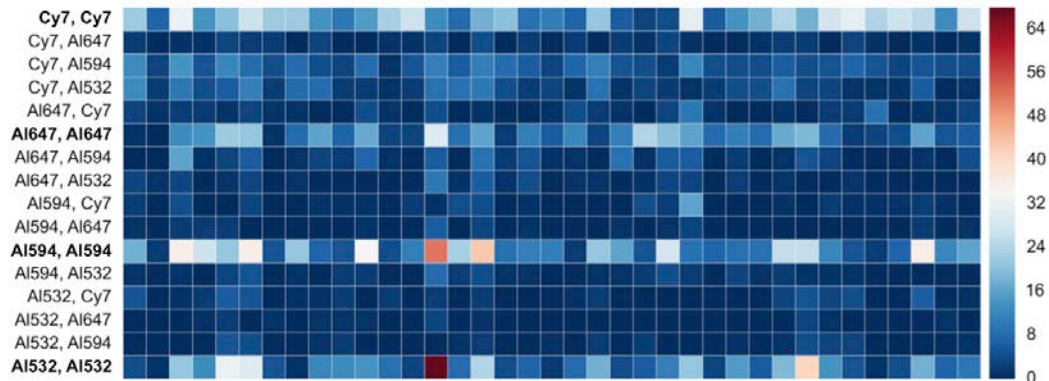
Fraction of barcodes identified from first two rounds of hybridization that reoccur in following round of hybridization per cell. Barcodes were identified by co-localization through all three hybridizations. $77.9 \pm 5.6\%$ of barcodes reoccur. $n = 37$ cells.

Supplementary Figure 6



Point-wise displacement between FISH points in Hybridizations 1 and 3. FISH dots in the Cy5 images in Hybridization 1 and 3 are extracted, fitted with 2D Gaussians. The point-wise displacements are shown in the 3D histogram. The standard deviation is 105.8nm, indicating that mRNAs can be localized to 100nm between 2 rounds of hybridizations. $n = 1199$ spots.

Supplementary Figure 7



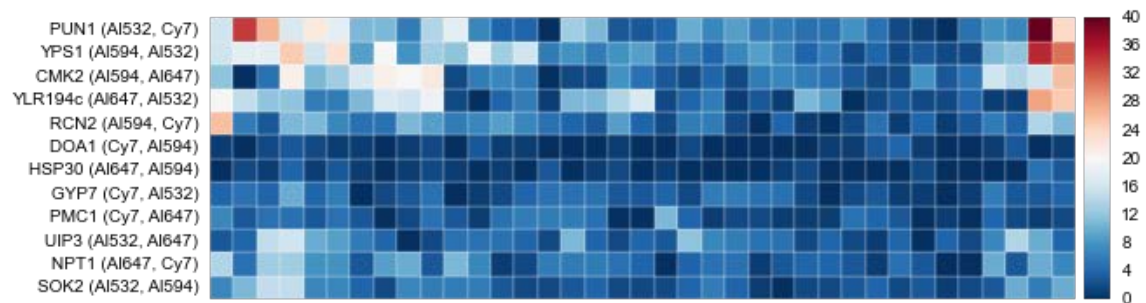
Barcodes identified between repeat hybridizations of the same probe set (hybridization 1 and 3). Barcodes were identified by co-localization between the hybridizations. Each column corresponds to an individual cell. Each row corresponds to a specific barcode identified between hybridization 1 and 3. Bolded row names correspond to repeated color barcodes that should co-localize between hybridization 1 and 3. Non-bolded row names correspond to false positive barcodes. For example, a large number of barcodes are detected for (Alexa 532, Alexa 532), indicating co-localization of spots in the Alexa 532 channels. $n = 37$ cells.

Al532 = Alexa 532

Al594 = Alexa 594

Al647 = Alexa 647 Al647 = Alexa 647

Supplementary Figure 8



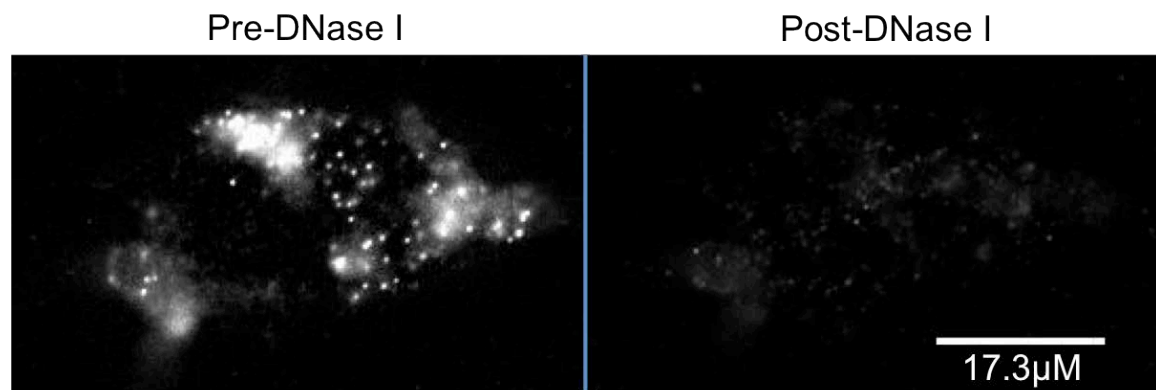
Single cell mRNA levels from barcode extraction. Barcodes were identified by co-localization between hybridizations 1 and 2. Each column corresponds to an individual cell. n = 37 cells.

Al532 = Alexa 532

Al594 = Alexa 594

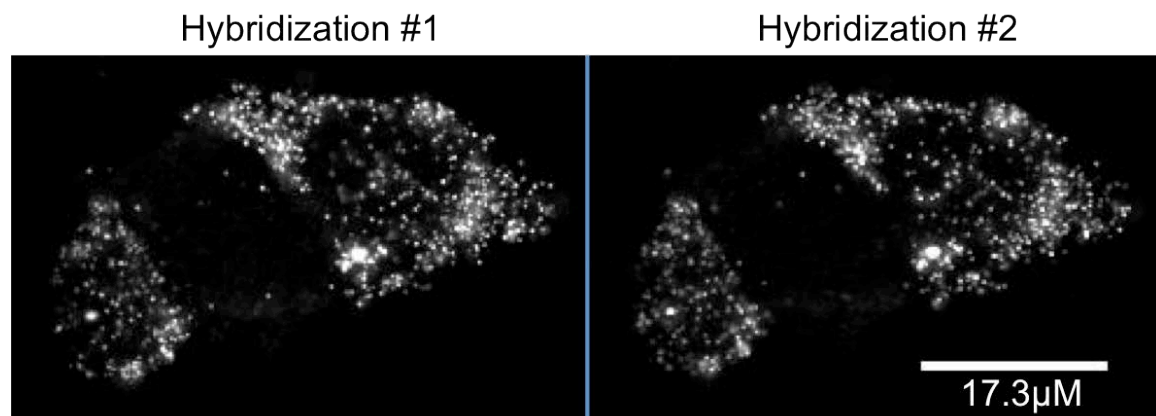
Al647 = Alexa 647

Supplementary Figure 9



DNase I stripping of Nanog Alexa 647 probes in mouse embryonic stem cells (mESCs). Forty-eight probes targeting Nanog were hybridized in mESCs. Probes were stripped off by 30 minutes of DNase I incubation at a concentration of 3 Units/ μ L.

Supplementary Figure 10



Re-Hybridization of Nanog mRNA in Mouse Embryonic Stem Cells (mESCs). Probes were stripped off by 30 minutes of DNase I incubation at a concentration of 3 Units/ μ L. Nanog Alexa 647 probes were re-hybridized for 12 hours and imaged. Images are 2D maximum projections created from z stacks of 11 images taken every 1.5 μ m.

Supplementary Methods

Sample Preparation

MDN1-GFP yeast cells were grown in YPD supplemented with 50mM CaCl₂ to OD 0.3. Cells were fixed in 1% Formaldehyde 5% Acetic Acid for 5 minutes, rinsed 3X in Buffer B and spheroplasted for 1 hour at 30°C. Cells were stored in 70% EtOH at -20°C for up to two weeks.

Coverslips were prepared by sonicating 3X with alternating solutions of 1M NaOH and 100% EtOH followed by a final round of sonication in acetone. A 2% solution of (3-Aminopropyl) triethoxysilane(Sigma 440140) was prepared in acetone and the cleaned coverslips were immediately submerged in it for two minutes. Amine-modified coverslips were rinsed and stored in ultra pure water at room temperature.

Fixed yeast cells were pre-treated with a 0.5U/uL solution of DNase I (Roche 04716728001) for 30 minutes at 23°C. Following treatment, yeast cells were adhered to coated coverslips by physically compressing a dilute solution of yeast between two amine-modified coverslips. The coverslips were then carefully peeled apart and immediately submerged in a 1% Formaldehyde solution for 2.5 minutes. Following fixation coverslips were dried and a flow cell was constructed by adhering an adhesive coated flow cell to the coverslip (GraceBio Labs SA84-0.5-SecureSeal). FluoSphere 365nm fluorescent beads were added to the coverslip to measure drift over multiple hybridizations(Life F8805). Flow cells were stored at 4°C covered with parafilm.

Probe Preparation

Probes were prepared according to the method in our previous paper¹. For each gene, 24 probes are used. All 24 probes for each set of genes are coupled to one of the four dyes used, Alexa 532, 594, Cy5 and Cy7.

Hybridization

Flow cells were hybridized at a concentration of 2nM/probe overnight in a hybridization buffer of 10% Dextran Sulfate (Sigma D8906), 10% Formamide, 2X SSC. Following hybridization, samples were washed in a 30% Formamide, 0.1% Triton-X 100 buffer pre-heated to 37°C before adding to room temperature samples for 10 minutes. Samples were washed several times with 2X SSC to remove diffusing probes.

Imaging

Samples were immersed in an anti-bleaching buffer² consisting of 20mM Tris-HCL, 50mM NaCl, 0.8% Glucose, Saturated Trolox(Sigma: 53188-07-1), Pyranose

oxidase(Sigma P4234) at an OD_{405nm} of 0.05, and catalase at a dilution of 1/1000 (Sigma:9001-05-2).

Probe Displacement

Following imaging, cells were washed in DNase I buffer (Roche) and allowed to sit in 0.5U/uL DNase I (Roche) for 4 hours. To inhibit DNase cells were washed 2X with 30% Formamide, 0.1% Triton-X 100, 2X SSC. Following DNase treatment cells were imaged once more in anti-bleaching buffer to determine DNase I probe stripping rates. To remove remaining probe signal samples were bleached with 10 seconds of excitation in all imaging channels and imaged once more with standard excitation times to record residual signal.

Re-hybridization

Samples were re-hybridized on the microscope according to the previously outlined conditions. Samples were covered with parafilm during hybridization on the scope to prevent evaporation.

Up to six rounds of hybridizations were carried out on the same sample. Each round of hybridization took place overnight on the microscope, with DNase treatment and imaging occurring during the day. In the iterative hybridization scheme applied in this correspondence, we used two rounds of hybridization to barcode the mRNAs. The barcode scheme was then repeated, such that hyb1 and hyb3 were performed using the same probes, while hyb2 and hyb4 were done with another set of probes. The co-localization between hyb1 and hyb3 gave a calibration for transcripts that were detected, while hyb1 and hyb2 yielded the barcoding data.

Data Analysis

Data analysis was carried out with ImageJ, Python and Matlab. Since the sample drifted during the experiments, the raw images were aligned using cross-correlation based registration method that was determined from the DAPI channel of each imaging position. The drift-correction was then propagated to the other 4 color channels corresponding to the same position. The images were then deconvolved to decrease the overlap between adjacent FISH spots. We note that we rarely observed spots overlap in individual channels, but spots in different channels can overlap in their point spread functions (PSFs) when the images are overlaid. Therefore, we processed the raw data based on an iterative Lucy-Richardson algorithm³⁻⁴. The PSF of the microscope was estimated by averaging the measured bead images (~200 nm diameter) in the DAPI channel of the microscope. Using this measured point spread function with the Lucy-Richardson algorithm, we performed maximum-likelihood estimation of fluorescent emitter distribution in the FISH images after computing this process over ~20 iterations.

The output of this deconvolution method provides resolved FISH data and increases the barcode assignment fidelity.

Dots corresponding to FISH signals in the images were identified using a local maximum function⁵. Dots below a threshold were discarded for further analysis. The value of the threshold was determined by optimizing the co-localization between hyb1 and hyb3 images, which were hybridized with the same probe sets. The maximum intensity pixel for each PSF was used as a proxy for the location of that mRNA molecule.

The barcodes were extracted automatically from the dots corresponding to mRNAs in hyb1 and hyb2. The algorithm calculated the pairwise distances between each point identified in hyb1 with all the points identified in hyb2. For each point in hyb1, the closest neighbor in hyb2 was identified. If that distance were 0 or 1 pixel and the closest neighbor of the point in hyb2 were also the original point in hyb1, then the barcode pair was confirmed. The symmetrical nearest neighbor requirements decreased the false assignment of barcodes. To reduce false positives in cy7, we required points detected in hyb 1 cy7 to reappear in hyb 3 in cy7.

Supplementary Note

For this publication we chose to strip off probes with DNase I due to its low cost and rapid activity. In principal any method that removes probes from mRNA and leaves it intact could be used in our barcoding approach. Other available methods include strand-displacement⁶ and high temperature or formamide washes. Compared to these other methods DNase I has the advantage of not requiring probe redesigns from standard smFISH probes, and not perturbing the sample with harsh washes.

We saw a rapid loss of DAPI signal from dsDNA within seconds, while smFISH probes took a substantially longer period of time (10s of minutes) to be degraded. Although the efficiency of DNase I probe removal is low relative to the dsDNA cleavage rate, the process is still observed in a short amount of time.

In our experiments we see that 11.5% of the fluorescent signal remains on mRNA after DNase I treatment. The remaining signal is reduced almost to zero by bleaching. We note that our re-hybridization technique would not work without the DNase I treatment. If probes were still bound to mRNA but bleached, the following rounds of hybridization would not work. Photobleaching is not necessary for barcoding, but does simplify the process by removing residual signal that might give false positives in further rounds of barcoding.

Some of the 11.5% of residual probes bound to mRNA may inhibit further rounds of hybridization. We note that residual probes are not significantly inhibiting progressive rounds of hybridization as our data shows a minor drop in hybridization efficiency for 5 hybridizations.

References

1. Lubeck, E. & Cai, L. *Nat. Methods* **9**, 743–48 (2012).
2. Swoboda, M. *ACS Nano* **6**, 6364–69 (2012).
3. Lucy, L.B. *The Astronomical Journal*. **79**, 745 (1974).
4. Richardson, W.H. *J. Opt. Soc. Am.* **62**, 55-59 (1972).
5. Shaffer, S. M. *PLoS ONE*, **8**, e75120 (2013).
6. Duose, D. Y. *Nucleic Acids Research*, **40**, 3289-3298 (2012).