Image-based transcriptomics in thousands of single human cells at single-molecule resolution

Nico Battich[1–3], Thomas Stoeger[1–3] & Lucas Pelkmans[1]

Fluorescence *in situ* hybridization (FISH) is widely used to obtain information about transcript copy number and subcellular localization in single cells. However, current approaches do not readily scale to the analysis of whole transcriptomes. Here we show that branched DNA technology combined with automated liquid handling, high-content imaging and quantitative image analysis allows highly reproducible quantification of transcript abundance in thousands of single cells at single-molecule resolution. In addition, it allows extraction of a multivariate feature set quantifying subcellular patterning and spatial properties of transcripts and their cell-to-cell variability. This has multiple implications for the functional interpretation of cell-to-cell variability in gene expression and enables the unbiased identification of functionally relevant *in situ* signatures of the transcriptome without the need for perturbations. Because this method can be incorporated in a wide variety of high-throughput image-based approaches, we expect it to be broadly applicable.

Large-scale transcriptomics with microarrays or RNA-seq is usually applied on a population of RNA molecules pooled from a large number of cells[1–4]. Although sequencing of single-cell transcriptomes has been performed[5–8], current approaches work reliably only for abundant RNAs[9], are feasible for only a small number of single cells and do not reveal the subcellular localization of transcripts.

FISH may overcome this, but it is not an automated large-scale approach. Using branched DNA (bDNA) technology, we applied single-molecule FISH (sm-FISH) to automated large-scale experiments. bDNA sm-FISH allows the use of one standard protocol and automation with high-throughput liquid-handling equipment and high-resolution screening microscopes. In conjunction with high-performance computing, bDNA sm-FISH enables the large-scale multivariate profiling of RNA transcript abundance as well as subcellular localization and patterning in thousands of single human cells per transcript with single-molecule sensitivity.

RESULTS

bDNA allows accurate single-molecule RNA measurements

In bDNA FISH, for which reagents are available from Advanced Cell Diagnostics and Affymetrix, multiple pairs of primary probes hybridize to two consecutive regions of 20–30 nucleotides at multiple positions along the transcript. Each primary probe pair jointly provides a hybridization site for a preamplifier probe, which hybridizes multiple amplifier probes that allow binding of a large number of labeled probes[10–14] (**Fig. 1a**). This contrasts with the most widely used sm-FISH approach, o-nuc sm-FISH, which employs oligonucleotides labeled with 1–5 fluorophores and lacks a signal-amplification step[15,16] (**Fig. 1b**). Consequently, o-nuc sm-FISH required a 600-times-longer exposure and a 100-times-greater camera gain than bDNA FISH to generate images with discernible spots for endogenous *MYC* mRNA in HeLa cells using a 100×/1.49–numerical aperture (NA) oil-immersion objective and electron-multiplying charge-coupled device (EMCCD) cameras (**Fig. 1c**,**d** and **Supplementary Fig. 1a**,**b**). With these different settings for exposure and gain, both approaches resulted in similar spot counts: 191.0 ± 66.4 (mean ± s.d.) spots per cell for bDNA FISH (n = 28 cells) and 189.0 ± 61.0 spots per cell for o-nuc sm-FISH (n = 20 cells). Under equal imaging conditions, bDNA spots were 100 times brighter than o-nuc spots (**Fig. 1e**–**g** and **Supplementary Fig. 1**), resulting in a signal-to-noise ratio that was at least 2–3 times higher than that of o-nuc sm-FISH (**Fig. 1h** and **Supplementary Note 1**). Furthermore, by labeling the same transcript with two different probe set types (**Supplementary Fig. 2a**–**c** and **Supplementary Note 2**), 80.8% (n = 4,703 spots) of *KIF11* transcripts and 84.58% (n = 2,979 spots) of *ERBB2* transcripts labeled with type 1 probe sets were also labeled with type 6 probe sets, which are similar accuracies to that reported for o-nuc sm-FISH[16]. Thus, bDNA FISH and o-nuc sm-FISH detected comparable numbers of discrete spots in single cells with a similar accuracy, but bDNA FISH yielded brighter spots with a better signal-to-noise ratio.

bDNA sm-FISH allows high-throughput RNA measurements

We next used a fully automated confocal microscope to image large fields of cells with a 40×/0.95-NA air objective and scientific

[1]Faculty of Sciences, Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]Systems Biology PhD program, Life Science Zurich Graduate School, ETH Zurich and University of Zurich, Zurich, Switzerland. [3]These authors contributed equally to this work. Correspondence should be addressed to L.P. (lucas.pelkmans@imls.uzh.ch).

RECEIVED 21 MARCH; ACCEPTED 28 AUGUST; PUBLISHED ONLINE 6 OCTOBER 2013; DOI:10.1038/NMETH.2657

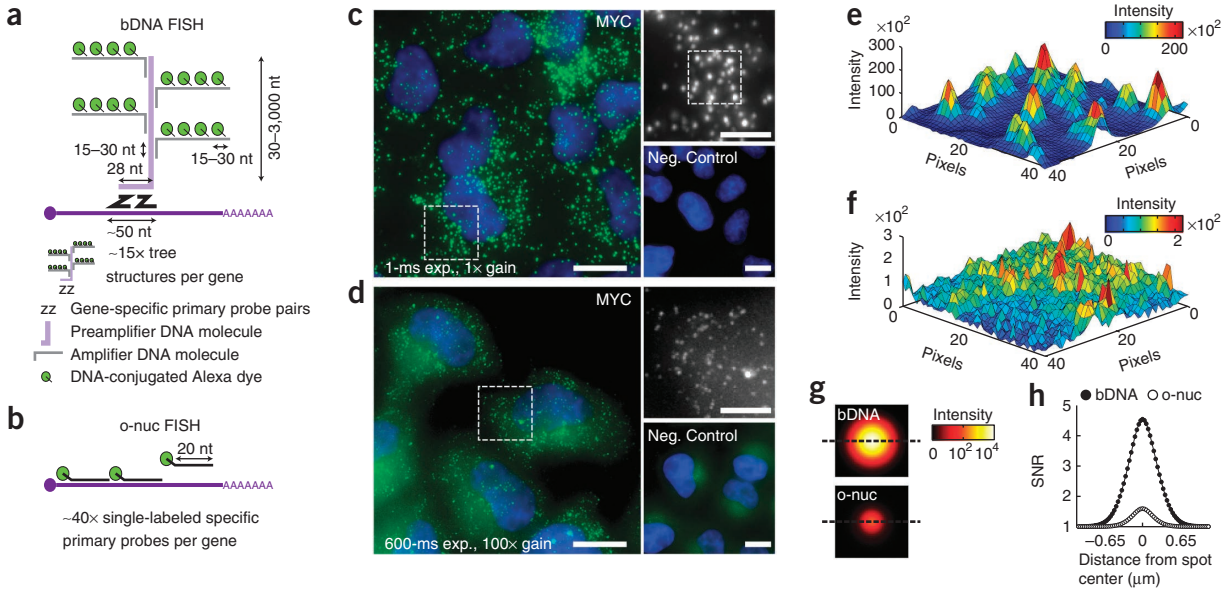

Figure 1 | bDNA FISH results in bright spots with high signal-to-noise ratio (SNR). (**a**) The bDNA FISH technique. Gene-specific primary probe pairs hybridize to the targeted RNA; tree-like structures composed of preamplifiers, amplifiers and labeled probes can be built onto these pairs, leading to signal amplification. nt, nucleotides. (**b**) The o-nuc FISH technique. The primary probes are directly labeled with a single fluorophore. (**c**) sm-FISH of endogenous MYC in HeLa cells with the bDNA method (green). Images were taken on an epifluorescence microscope using a 100×-magnification oil-immersion objective (NA = 1.49) and a back-illuminated EMCCD camera. The negative control with no primary probe pairs is also shown (bottom right). Cell nuclei are stained with DAPI (blue). Scale bars (**c**,**d**), 13 μm (overview images) and 5 μm (insets). (**d**) As in **c** but using o-nuc sm-FISH. (**e**) Intensity profile of the marked region in the top right subpanel of **c** after extracellular background subtraction. (**f**) As in **e** but for the area marked in **Supplementary Figure 1c**, the settings for which were an exposure time of 1 ms and a camera gain of 1. (**g**) Mean-modeled spots at subpixel resolution for bDNA sm-FISH and o-nuc sm-FISH after local background subtraction using a 1-ms exposure time and camera gain set to 1 (n = 100 detected spots). Dashed lines mark the spot equator. (**h**) SNR (**Supplementary Note 1**) along the equator line of the modeled subpixel spots after extracellular background subtraction; n = 100 detected spots.

complementary metal-oxide semiconductor (sCMOS) cameras. We performed FISH against the endogenous transcripts of *ERBB2*, *MYC* and *TFRC* in ~10^4 HeLa cells per gene in a 384-well plate format. Spots could be observed for each gene with the bDNA method only (**Supplementary Fig. 3**), and this method generated a highly reproducible mean number of spots per cell (23.41 ± 0.47, 203.01 ± 8.02 and 187.93 ± 6.88 for *ERBB2*, *MYC* and *TFRC*, respectively; n = 4 wells; **Supplementary Table 1**). Notably, the median number of spots per cell detected for *MYC* was comparable to that obtained with bDNA (P = 0.54, Mann-Whitney-Wilcoxon test) and o-nuc sm-FISH (P = 0.52, Mann-Whitney-Wilcoxon test) using 100×/1.49-NA magnification and EMCCD cameras.

To confirm that the spots were specific for *ERBB2*, *MYC* and *TFRC*, we performed gene silencing with RNAi. The spot-count reduction observed was strong and comparable to that determined from qPCR measurements (**Supplementary Fig. 3** and **Supplementary Table 1**). Furthermore, probe pairs against the *Escherichia coli* gene *dapB* showed a false positive rate of 0.44 ± 1.0 mean spots per cell (n = 21,094 cells). To test nuclear accessibility of the bDNA probes, we performed bDNA FISH against the nuclear-localized *SNORD3* transcripts and found no signal in the nucleus (**Supplementary Fig. 4a**,**b**). Although acetic acid in the fixation buffer[17] increased the nuclear signal for *SNORD3* and *HPRT1*, it reduced cytoplasmic spots (**Supplementary Fig. 4b**,**c**) leading to inaccurate measurements of the mature mRNA for *HPRT1* (ref. 18).

Next we tested the number of primary probe pairs that ensures that each transcript in the cytoplasm is detected by the signal of at least one primary probe pair. For both *ERBB2* and *HPRT1*, ten primary probe pairs allowed a detection of more than 80%, and 15 primary probe pairs allowed a detection of more than 90%, of the maximum number of detectable transcripts (**Supplementary Fig. 5a**,**b**). Single-cell distributions of spots per cell and their Fano factors, i.e., variance divided by mean spots per cell, also stabilized from ten primary probe pairs onwards (**Supplementary Fig. 5c**–**e**).

We then evaluated the single-spot detection accuracy of high-throughput bDNA FISH in single cells (**Supplementary Fig. 6**). The single-cell correlations of spot counts per cell for *KIF11* and *ERBB2* transcripts labeled simultaneously with two probe sets of different color (**Supplementary Fig. 3a**) were 0.976 and 0.836, respectively (Pearson correlation; **Supplementary Fig. 6a**,**b**). We estimated that for *KIF11*, 2.5% of the total cell-to-cell variability was of technical origin, whereas for *ERBB2* this was 21.8% (**Supplementary Fig. 6**). The higher fraction of technical variance in single-cell measurements for *ERBB2* was likely due to its lower expression (24.16 ± 14.55 spots per cell, n = 10,524 cells) compared to *KIF11* (73.23 ± 52.01 spots per cell, n = 10,223 cells). Thus, bDNA FISH with 15 primary probe pairs is suitable for sensitive, specific and reproducible high-throughput transcript quantification in 384-well plates at single-molecule and single-cell resolution for both low-and high-abundance transcripts.

Experimental and image-analysis pipeline

To assess the feasibility of applying our approach at the genome scale, we constructed a library of bDNA probes in 384-well

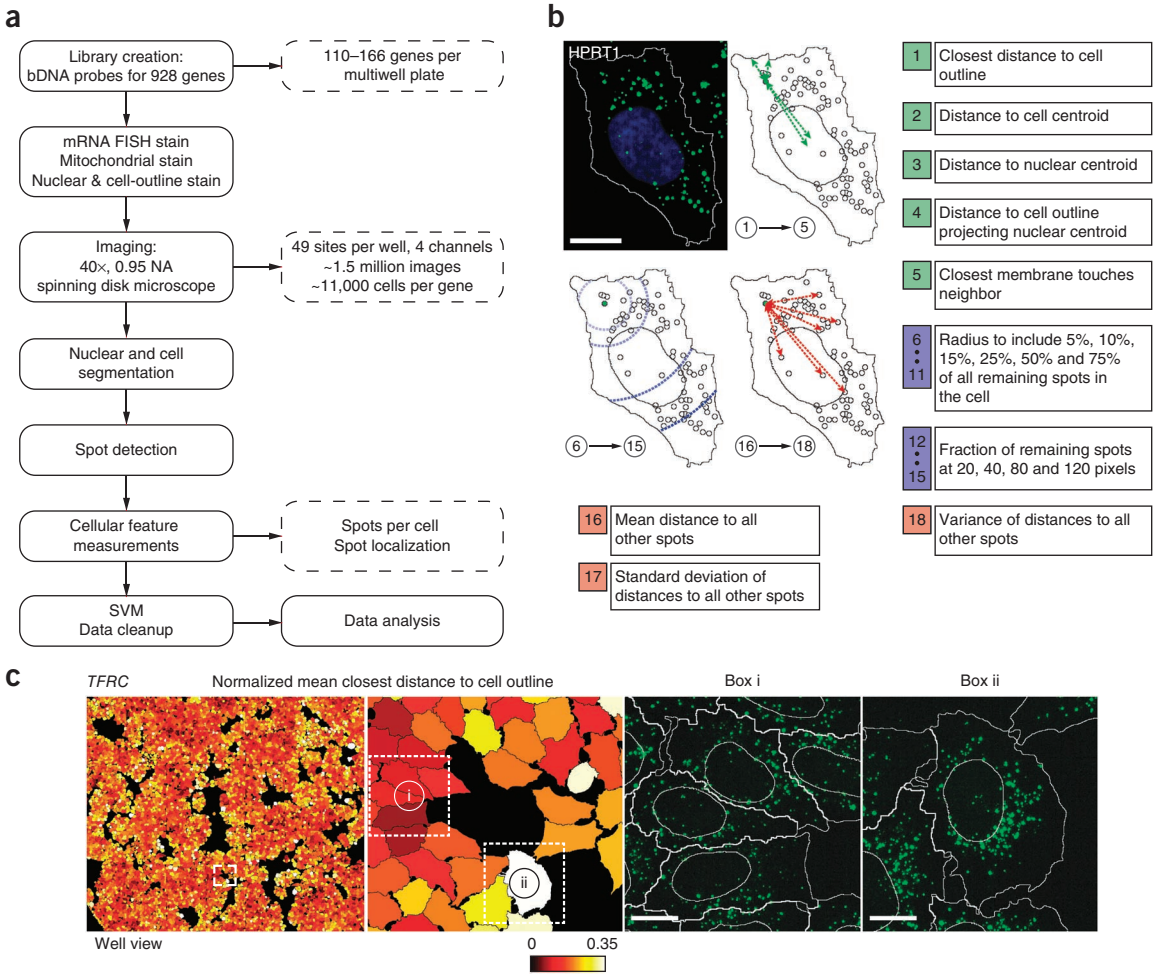

Figure 2 | Image-based transcriptomics pipeline. (**a**) Primary probes for 928 genes were plated within the center 180 wells of 384-well plates. 384-well plates containing cells were then stained in parallel with the bDNA sm-FISH reagents, MitoTracker to stain mitochondria, DAPI to stain nuclei and a protein-reactive fluorescent dye to stain whole cells. Plates were imaged at 40× magnification. Images were analyzed using CellProfiler and a custom spot-detection algorithm. Supervised machine learning (SVM) was applied to ensure high data quality by eliminating undesired phenotypes and segmentation and staining artifacts. (**b**) Features extracted to describe spot localization in single cells. Features 1–5 map every spot with respect to the cell and the nucleus. Features 6–18 map a spot relative to all other spots in the cell. (**c**) Mean closest distance to the cell outline (divided by the square root of the cell area in pixels) of all spots in a cell for *TFRC* transcripts in a population of cells. Green, bDNA sm-FISH; blue, DAPI (cell nucleus). Scale bars, 13 µm.

plates targeting 928 human genes involved in basic cellular functions, cancer, signaling, endocytosis and metabolism (**Supplementary Table 2**). In addition, we modified existing algorithms[16,19–21] to create a robust high-throughput spot-detection pipeline (**Supplementary Note 3**, **Supplementary Fig. 7** and **Supplementary Software**). We automated the experimental protocol using a liquid-handling platform (**Supplementary Protocol**), and image analysis[19] and supervised machine learning data cleanup[22] were submitted to high-performance computing using iBRAIN[23]. As a proof of principle, we performed two independent biological replicates of *in situ* transcriptomics in an unperturbed HeLa cell line (**Fig. 2a**).

We acquired confocal images in ten *z* planes, with a step size of 1 µm, covering the full cellular height at 49 sites in each well. Because two-dimensional spot detection on maximum-intensity projections of *z* stacks yielded virtually identical numbers of spots per cell as three-dimensional spot detection (**Supplementary Fig. 7i**), we performed all our quantifications on projected *z* stacks. We obtained 18 primary spot features that reflect the relative localization of each spot in a single cell, with respect to both the cell and other spots (**Fig. 2b**,**c**). To give an impression of the information contained in one such feature, we depicted the single-cell values for mean closest distance of detected spots to the cell outline for the transcript *TFRC* (transferrin receptor 1) in a segmented population of cells (**Fig. 2c**).

High-throughput quantitative image-based transcriptomics

The mean number of spots per cell for each gene was highly reproducible between the two biological replicates (Pearson correlation of 0.989; **Fig. 3a** and **Supplementary Table 3**). In addition, the absolute gene expression level of control genes across plates was very similar (**Supplementary Fig. 8**). When comparing distributions of single-cell spot counts of each gene with the negative control *dapB* (**Fig. 3b**), we found that 857 of 928 gene transcripts contained a significant fraction of cells with spot counts higher than those for *dapB* ($P < 10^{-4}$ for both replicates; **Supplementary Note 4** and **Fig. 3c**). These 857 detected genes displayed a bimodal distribution of low- and high-expressed transcripts with

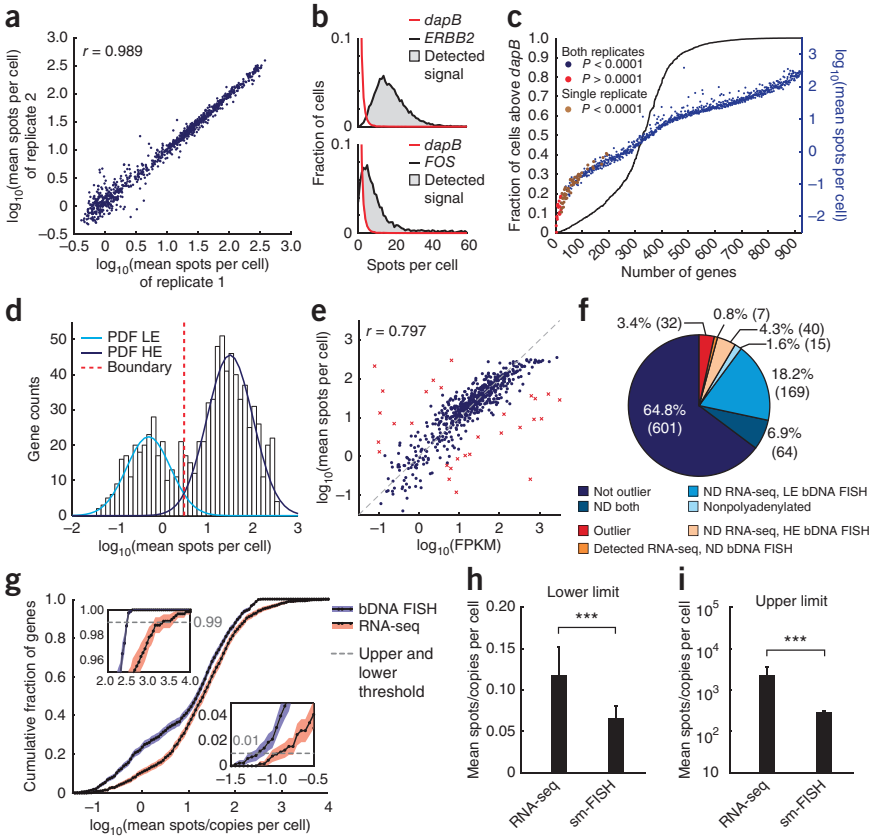

Figure 3 | Image-based transcriptomics is reproducible, sensitive and comparable to RNA-seq. (**a**) $\log_{10}$(mean spots per cell) correlation of biological replicates. The Pearson correlation is shown. (**b**) Relative distribution of *dapB* compared to those of two examples, *ERBB2* and *FOS* (replicate 1). The gray area is the fraction of cells above background (or detected signal) for a given gene. (**c**) Fraction of cells above background (black line) and corrected mean expression level (data points) in $\log_{10}$(mean spots per cell) for each gene; $n = 500$ bootstraps. Colors indicate whether the fraction of cells above background reached significance ($P < 1.0 \times 10^{-4}$) in none (red), one (brown) or both replicates (blue). (**d**) Distribution of corrected $\log_{10}$(mean spots per cell) for blue data points in **c** (857 genes). Solid lines indicate the probability density function (PDF) of low-expressed (LE) and high-expressed (HE) genes. The dashed line is the estimated boundary between the two classes (3.01 spots per cell). (**e**) Correlation of RNA-seq, $\log_{10}$(fragments per kilobase of exon model per million mapped reads (FPKM)), and high-throughput bDNA sm-FISH $\log_{10}$(mean spots per cell). Outliers are shown in red. *r* is the Pearson correlation before outlier rejection. The dashed line is a guide for the eye. (**f**) Detailed comparison of transcript detection for RNA-seq and high-throughput bDNA sm-FISH. ND, not detected. (**g**) Cumulative fraction of genes as a function of the expression level in $\log_{10}$(mean spots/copies per cell) for RNA-seq and high-throughput bDNA sm-FISH. Smaller panels show the estimation of lower and upper limits of detection (dashed lines). Shaded areas represent the s.d. of 1,000 bootstraps. (**h**,**i**) Quantification of the lower (**h**) and upper (**i**) limits of detection for RNA-seq and high-throughput bDNA sm-FISH. Error bars, s.d. of 1,000 bootstraps. $***P < 1.0 \times 10^{-3}$.

a boundary between them at 3.01 ± 0.50 mean spots per cell ($n = 1{,}000$ bootstrapped samples; **Fig. 3d**, **Supplementary Fig. 9** and **Supplementary Note 4**). Such a boundary was previously estimated at a lower value[24].

Notably, the correlation between mean spot count per cell and transcript abundance measured with RNA-seq (**Supplementary Fig. 10**) was 0.797 (Pearson correlation) or 0.842 (Spearman correlation) (**Fig. 3e**), which increased to 0.917 (Pearson correlation) or 0.915 (Spearman correlation) after outlier rejection. For 71.7% of transcripts, both methods either detected a signal at similar levels (64.8%) or did not detect a signal (6.9%; **Fig. 3f**). 22.5% of transcripts were detected only by bDNA sm-FISH (18.2% as low-expressed transcripts), whereas 0.8% of transcripts were detected only by RNA-seq (0.54% as low-expressed transcripts, not shown). Of the remaining 5% of transcripts, 1.6% were detected only by bDNA sm-FISH because they were nonpolyadenylated, whereas 3.4% were detected by both methods but their levels did not correlate (**Fig. 3e**). Comparing the detection sensitivity and dynamic range of high-throughput bDNA sm-FISH with RNA-seq revealed that at the lower limit of detection, high-throughput bDNA sm-FISH was more sensitive than RNA-seq (0.066 ± 0.015 and 0.118 ± 0.034 spots/copies per cell, respectively; $n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**h**). At the upper limit of detection, high-throughput bDNA sm-FISH showed a ceiling effect at 288.98 ± 18.24 spots/copies per cell at the mean level ($n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**i**). At the single-cell level, however, we obtained spot counts higher than 1,500 (for example, for 18S1–18S5 RNA, *CYTB* and *GAPDH*; not shown). For RNA-seq, the upper limit of detection for the genes in our library was 2,262.41 ± 1,278.74 copies per cell ($n = 1{,}000$ bootstrapped samples; **Fig. 3g**,**h**).

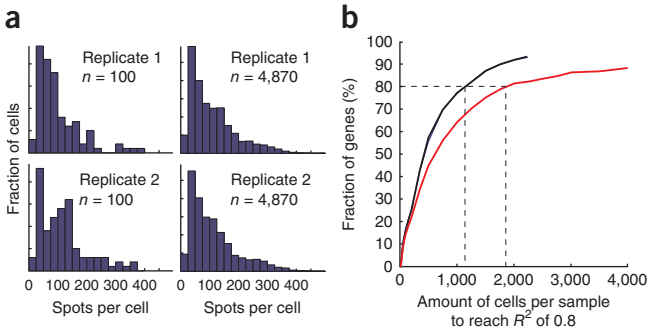

Figure 4 | Minimum number of cells required for reproducible single-cell distributions of transcript abundance. (**a**) Example distributions of transcript abundance of the cell cycle–associated gene *PLK1* in two biological replicate measurements using a sample size of 100 single cells or a sample of 4,870 single cells. If only 100 cells are sampled, the distributions of single-cell spot counts (at a bin size of 25 spots) are dissimilar. (**b**) Number of cells that need to be sampled to reach a coefficient of determination (R^2) of 0.8 between single-cell spot count distributions within (black and blue lines) or between (red line) the two replicates. Dashed lines indicate requirement of cells for 80% of all genes.

Thus, high-throughput bDNA sm-FISH generates highly reproducible results and is a quantitative method for large-scale transcriptomics with high sensitivity that rivals RNA-seq for low-abundance transcripts.

Requirements for reproducible single-cell distributions

Most studies on cell-to-cell variability in RNA transcript copy number have so far relied on the quantification of, at maximum, several hundred single cells[24–27]. However, it is unclear how many

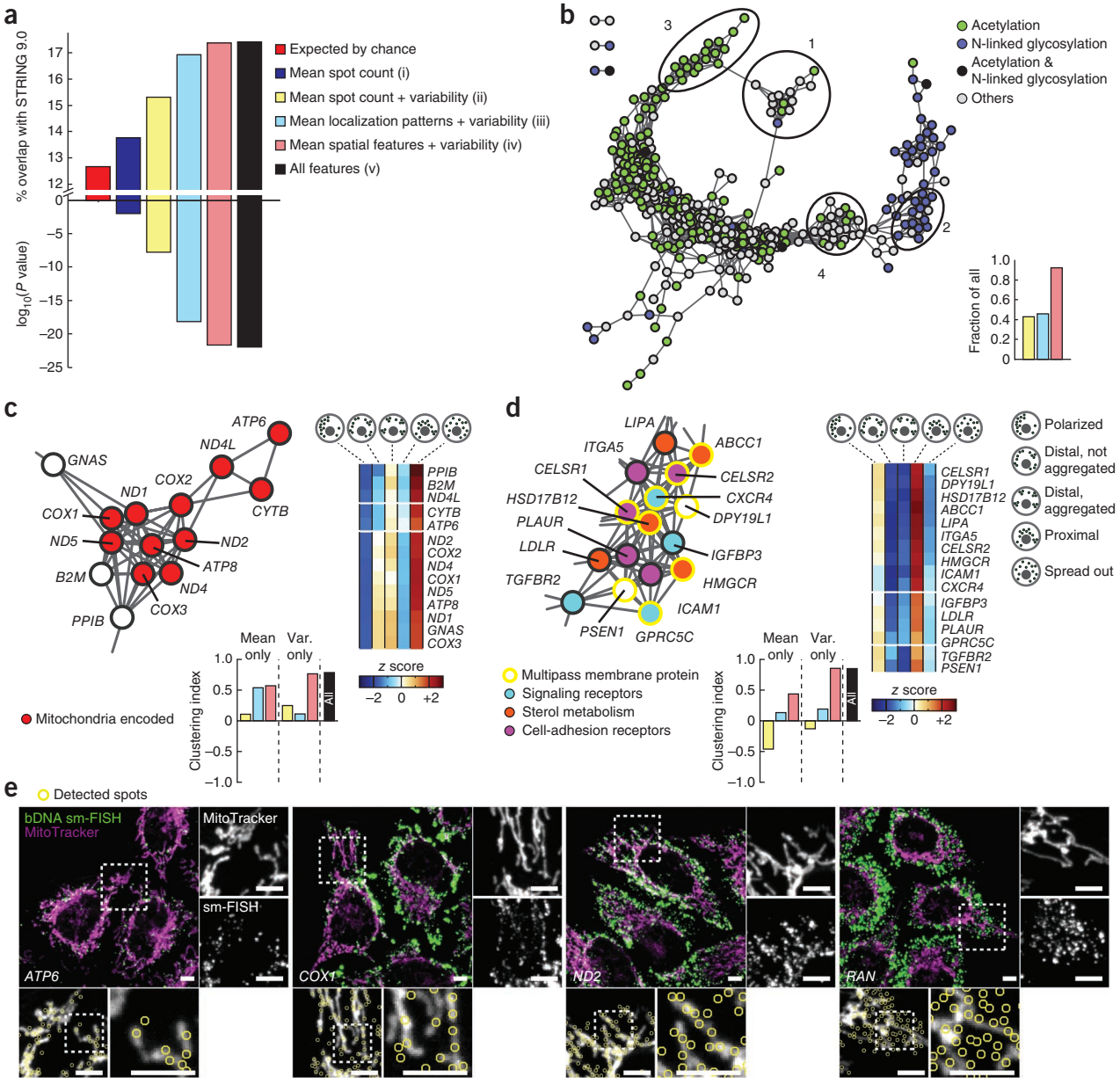

Figure 5 | Quantitative signatures of the *in situ* transcriptome. (**a**) Overlap of the 5% smallest pairwise gene-gene distances with known gene interactions in STRING 9.0 and their respective *P* values. Data are shown for five different sets of features: (i) mean RNA spot count per cell (blue); (ii) mean RNA spot count per cell and features of its distribution (variability) (yellow); (iii) mean single-cell classification of localization patterns per gene and features of the classification distributions (light blue; see **Supplementary Fig. 12**); (iv) mean spatial features of spots per gene and features of their distributions (salmon); and (v) the combination of all extracted features (black). See also **Supplementary Figure 14b.** (**b**) Gene network (4,873 edges) obtained with the 5% smallest gene-gene distances derived from the combination of all features (black bar in **a**). Only connected genes are shown (96.8% of included genes). Node colors indicate genes encoding acetylated proteins (green), N-linked glycosylated proteins (blue), those that undergo both modifications (black) and others (gray). The bar graph indicates the fraction of edges that are also retrieved with three specific feature subsets, subsets ii–iv in **a**; color-coding as in **a**. Subregions in the network correspond to **c** (subregion 1), **d** (subregion 2) and **Supplementary Figure 15d,e** (subregions 3 and 4). (**c**,**d**) Subregion 1, a tight cluster of genes encoded in the mitochondrial genome (red nodes, **c**); and subregion 2, a tight cluster of genes encoding cell-adhesion receptors (purple nodes, **d**), signaling receptors (light blue nodes, **d**) or proteins involved in sterol metabolism (orange nodes, **d**). Subregion 2 contains multipass membrane proteins (yellow-outlined nodes, **d**). *z*-scored mean classification distributions of cells for all five main types of single-cell spot localization patterns (specified at right) are shown as clustered heat maps. Bar graphs indicate the clustering index for three specific feature subsets, subsets ii–iv in **a**; color-coding as in **a**. (**e**) Subcellular localization of transcripts from the mitochondria-encoded genes *ATP6*, *COX1* and *ND2*, as well as of the transcripts from *RAN* (which does not cluster in subregion 1), with respect to MitoTracker. Yellow circles are detected spots. Scale bars, 5 µm.

cells must be sampled to obtain reproducible single-cell spot count distributions. We therefore compared random samples of increasing number of single cells for each gene to a second sample from the same cell population and a sample derived from the biological replicate (**Fig. 4a** and **Supplementary Fig. 11a**). Across all tested genes, 100 cells sufficed to obtain reproducible measurements of the mean, variance and Fano factor (Pearson correlation of 0.997, 0.951 and 0.910, respectively; **Supplementary Fig. 11b**). However, the third, fourth and fifth central moments required 215, 274 and 1,764 single cells, respectively, to obtain a Pearson correlation of 0.75 (**Supplementary Fig. 11c**). Furthermore, correlating whole spot count distributions revealed that at least 1,100 single cells were required to reach a high coefficient of determination ($R^2 = 0.8$) for 80% of the genes when different samples from the same cell population were compared, and 1,800 cells were required for samples coming from different biological replicates (**Fig. 4b** and **Supplementary Fig. 11d–g**). Thus, for most genes in a nonsynchronized unperturbed HeLa cell line, at least 1,000 single cells must be sampled to obtain reproducible single-cell distributions of transcript copy number.

Quantitative signatures of the *in situ* transcriptome

Finally, we wrote algorithms to harness the multivariate feature set quantifying subcellular localization and patterning of single transcripts within thousands of single cells. The first algorithm performs unsupervised clustering of all single cells to identify the main types of subcellular spot localization patterns, aiding biological interpretability (**Supplementary Fig. 12**, **Supplementary Note 5** and **Supplementary Software**). In the generated data set, this algorithm revealed five main types of single-cell patterns: a polarized distribution, distal distribution, distal and aggregated distribution, proximal (perinuclear) distribution and spread-out distribution of spots (**Supplementary Fig. 13**). The second algorithm maximizes the information contained within the multivariate feature set by computing additional features describing the variability of the spot count per cell and the spatial distribution of spots (**Supplementary Fig. 14**, **Supplementary Table 4** and **Supplementary Note 6**).

We then tested various combinations of the information obtained from these two algorithms to evaluate their ability to cluster genes that are functionally associated in a database of known and predicted protein interactions (STRING v.9.0; ref. 28) (**Supplementary Fig. 14**). This analysis revealed that quantitative information about subcellular patterns and spatial properties of transcripts and their variability across single cells were more powerful at identifying functional interactions than were features of mean spot count and its variability (**Fig. 5a** and **Supplementary Fig. 15a**,**b**).

We next created a network (**Fig. 5b**) from the top 5% of calculated gene-gene distances on the basis of their similarity in transcript features. The majority of edges in this network could be derived from spatial features and their variability (**Fig. 5b** and **Supplementary Fig. 15c**). Globally, genes that encode acetylated proteins translated in the cytosol separated from genes that encode N-linked glycosylated proteins translated at the endoplasmic reticulum (ER). Extensive subclustering within these two domains indicated that our feature set revealed details of subcellular patterning of transcripts and its cell-to-cell variability with functional relevance beyond general differences in translation sites.

A specific isolated region in the network (**Fig. 5b**) was formed by a tight subcluster of 11 of the 13 mRNA-coding genes encoded by mitochondria and showed an enrichment of cells with a spread-out and a distal distribution of transcripts (**Fig. 5c**). This subcluster was distinguished from its immediate surrounding by features of subcellular patterning as well as of spatial properties and their variability (**Fig. 5c**), whereas features of transcript abundance and variability did not contribute to this subclustering. Further analysis revealed that whereas transcripts of *ATP6*, *COX1* and *ND2* localized to stained mitochondria, transcripts of *RAN*, which is not part of this cluster (although it was nearby in the network), fell outside of the stained mitochondria (**Fig. 5e**).

The region of the network consisting of genes encoding N-linked glycosylated proteins also showed subclustering. One of these subclusters (**Fig. 5d**) consisted of genes encoding proteins involved in sterol metabolism and cell adhesion, and signaling receptors. The majority of these were multipass membrane proteins. This subcluster displayed an enrichment for cells with a perinuclear distribution of transcripts (**Fig. 5d**), a result consistent with localization to the ER[28]. The subcluster was distinguished from its immediate surroundings in the network by spatial properties and their variability, suggesting localization at specific subdomains of the ER. Features of transcript abundance alone would prevent this subclustering (**Fig. 5d**). Also, the region in the network enriched for acetylated proteins displayed further subclustering (**Fig. 5b**), with one subcluster of genes encoding ribosomal proteins and proteins involved in glycolysis and energy production and another subcluster of genes encoding proteins involved in endocytosis and ubiquitination (**Supplementary Fig. 15d**,**e**).

Taken together, the extracted feature set contains multiple types of information about specific *in situ* signatures of the transcriptome. In particular, features of subcellular localization and patterning and their variability allow the unbiased identification of functional interactions between genes without the need for any perturbation or costaining.

DISCUSSION

We have demonstrated the feasibility of large-scale image-based transcriptomics by applying sm-FISH in an automated high-throughput manner in human tissue culture cells, achieving comparable results to RNA-seq at the mean expression level. Most of the bDNA sm-FISH reagents used in this study were produced by Affymetrix upon our request and have since become available to other customers, thereby making our approach readily accessible. Currently, bDNA sm-FISH shows limited detection of nuclear transcripts, has less dynamic range than RNA-seq for high-abundance transcripts and may, for a few transcripts, obtain aberrant readouts. Another limitation is the small number of different transcripts that can be quantified in the same single cell compared to that by single-cell RNA-seq, which can achieve quantification of more than 6,000 transcripts per cell[8]. However, the bDNA signal amplification tree may allow extensive barcoding, which could be exploited for single-cell multiplexing in the near future[29–31].

High-throughput bDNA sm-FISH scales dramatically better than single-cell RNA-seq in the number of single cells that can be measured within the same sample[8], which is important for reproducible measurements of cell-to-cell variability in RNA transcript

abundance. It is also more sensitive than single-cell RNA-seq for low-abundance transcripts, reveals absolute copy numbers and allows the quantification of multivariate features of transcript patterning within and across thousands of single cells. Our analysis of these features revealed that shared properties of the variability in subcellular transcript localization across unperturbed single cells outperform cell-to-cell variability in transcript abundance in retrieving functional associations between genes.

Further development in the types of analysis shown here combined with perturbation experiments will increase the power of this approach. We expect that high-throughput bDNA sm-FISH will find broad applications as it can be directly included in various image-based approaches. This will enable a more direct examination of the causal links between molecular and phenotypic cell-to-cell variability.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to acknowledge B. Snijder and Y. Yakimovich for help with computational analysis and infrastructure, J. Patterson for assistance, Q. Nguyen and S. Lai from Affymetrix for helpful comments on experimental procedures, J. Ellenberg (European Molecular Biology Laboratory) for reagents, and all members of the lab for useful comments on the manuscript. L.P. acknowledges financial support for this project from SystemsX.ch, the European Union, University of Zurich and University of Zurich Research Priority Program in Systems Biology and Functional Genomics.

AUTHOR CONTRIBUTIONS

L.P. initiated the study. N.B., T.S. and L.P. designed and analyzed the experiments and wrote the manuscript. N.B. and T.S. performed the experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21**, 33–37 (1999).
2. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
3. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
4. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
5. Tang, F. *et al.* mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
6. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
7. Ramsköld, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
8. Shalek, A.K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).
9. Goetz, J.J. & Trimarchi, J.M. Transcriptome sequencing of single cells with Smart-Seq. *Nat. Biotechnol.* **30**, 763–765 (2012).
10. Lau, J.Y. *et al.* Significance of serum hepatitis C virus RNA levels in chronic hepatitis C. *Lancet* **341**, 1501–1504 (1993).
11. Kern, D. *et al.* An enhanced-sensitivity branched-DNA assay for quantification of human immunodeficiency virus type 1 RNA in plasma. *J. Clin. Microbiol.* **34**, 3196–3202 (1996).
12. Player, A.N., Shen, L.P., Kenny, D., Antao, V.P. & Kolberg, J.A. Single-copy gene detection using branched DNA (bDNA) in situ hybridization. *J. Histochem. Cytochem.* **49**, 603–612 (2001).
13. Kenny, D., Shen, L. & Kolberg, J.A. Detection of viral infection and gene expression in clinical tissue specimens using branched DNA (bDNA) *in situ* hybridization. *J. Histochem. Cytochem.* **50**, 1219–1227 (2002).
14. Ma, X.-J., Wu, X. & Luo, Y. Biomarkers for differentiating melanoma from benign nevus in the skin. US patent application 20120071343 (2012).
15. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. Visualization of single RNA transcripts *in situ*. *Science* **280**, 585–590 (1998).
16. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
17. Chartrand, P., Bertrand, E., Singer, R.H. & Long, R.M. Sensitive and high-resolution detection of RNA *in situ*. *Methods Enzymol.* **318**, 493–506 (2000).
18. Bhatt, D.M. *et al.* Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell* **150**, 279–290 (2012).
19. Carpenter, A.E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
20. Raj, A. & Tyagi, S. Detection of individual endogenous RNA transcripts *in situ* using multiple singly labeled probes. *Methods Enzymol.* **472**, 365–386 (2010).
21. So, L.H. *et al.* General properties of transcriptional time series in *Escherichia coli*. *Nat. Genet.* **43**, 554–560 (2011).
22. Rämö, P., Sacher, R., Snijder, B., Begemann, B. & Pelkmans, L. CellClassifier: supervised learning of cellular phenotypes. *Bioinformatics* **25**, 3028–3030 (2009).
23. Snijder, B. *et al.* Single-cell analysis of population context advances RNAi screening at multiple levels. *Mol. Syst. Biol.* **8**, 579 (2012).
24. Hebenstreit, D. *et al.* RNA sequencing reveals two major classes of gene expression levels in metazoan cells. *Mol. Syst. Biol.* **7**, 497 (2011).
25. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
26. Trcek, T., Larson, D., Moldón, A., Query, C. & Singer, R. Single-molecule mRNA decay measurements reveal promoter- regulated mRNA stability in yeast. *Cell* **147**, 1484–1497 (2011).
27. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
28. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39**, D561–D568 (2011).
29. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* **9**, 743–748 (2012).
30. Nguyen, Q.N., Lipshutz, R.J. & Ma, Y. Methods of labeling cells, labeled cells, and uses thereof. US patent application 20120178081 (2012).
31. Levesque, M.J. & Raj, A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods* **10**, 246–248 (2013).

ONLINE METHODS

Cell culture. HeLa Kyoto cells were kindly provided by J. Ellenberg (EMBL, Heidelberg). Cells were tested for identity by karyotyping and tested for absence of mycoplasma before use. Culturing was done in DMEM (Gibco) supplemented with 10% FCS and glutamine (complete medium). Seeding was at a density of 700 cells per well when using 384-well plates (Greiner) and 10,000 cells per well when using a LabTek chambered #1.0 borosilicate coverglass system of eight chambers. Cells were incubated for 3 d at 37 °C, 95% humidity and 5% CO_2. For image-based transcriptomics, a full cell culture was regrown from a single cell in six passages, after which cells were harvested, frozen and kept at −80 °C until use. Only cells imaged at 100× magnification were grown in LabTek chambers.

Microscopy. For high-magnification oil-immersion imaging, we used a Nikon Eclipse Ti inverted fluorescence microscope, an Apo TIRF 100× objective (Nikon) of 1.49 NA and an EMCCD camera (ImageEM 1K C9100-14, Hamamatsu). High-throughput *in situ* transcriptomics imaging, was done with an automated spinning disk microscope from Yokogawa (CellVoyager 7000), with an enhanced CSU-X1 spinning disk (Microlens-enhanced dual Nipkow disk confocal scanner, wide view type), a 40× Olympus objective of 0.95 NA, and a Neo sCMOS cameras (Andor, 2,560 × 2,160 pixels), acquiring 49 sites per well and ten *z* planes per site spanning 9 µm (**Supplementary Table 5**). The number of *z* planes was chosen so that every spot was visible in at least two planes as described previously[20]. The primary probe pair saturation curves were measured with an ImageXpress Micro fluorescence microscope (Molecular Devices), a Plan Apo 40× objective (Nikon) of 0.95 NA and a CoolSNAP HQ camera.

Oligonucleotide single-molecule RNA FISH. Quasar 570–labeled oligonucleotide Stellaris FISH RNA probes targeting *TFRC*, *MYC* and *ERBB2* mRNA were obtained from Biosearch Technologies. Probe hybridization was performed as indicated by the manufacturer.

Branched DNA single-molecule RNA FISH. All gene-specific primary probe pairs, amplification systems and custom-designed probes for measurement of saturation curves and double-labeling experiments were purchased from Affymetrix upon specific request and have since been made commercially available. Experiments were performed following the **Supplementary Protocol**. In the signal-saturation experiments, 15 individual primary probe pairs targeting *ERBB2* and *HPRT* were acquired from Affymetrix. Probe pairs were then combined in such a way to generate 30 primary probe-pair combinations per gene spanning a range of 1–15 targeted sites per gene. bDNA sm-FISH was then performed as described in the **Supplementary Protocol**. For acetic acid experiments, glacial acetic acid was added at the required [v/v]% to the fixation solution (4% paraformaldehyde in PBS).

Calculation of signal-to-noise ratios. Spot detection of 100×-magnification images was done as described in the **Supplementary Note 2**, although no spot bias correction was applied. Calculation of the signal-to-noise ratio is described in **Supplementary Note 1.**

Library construction. The final library was composed of probes against 925 human genes of general interest (**Supplementary Table 2**), probes against three positive-control genes (*ERBB2*, *HPRT* and *ACTB*) covering a wide range of expression levels and probes against a bacterial gene (*dapB*) as negative control. The library was mostly composed of QuantiGene View RNA type I primary probe pairs, although some genes were targeted with QuantiGene View RNA types VI, VIII or X. Primary probes for all genes were then organized in six 384-well plates according to plate layout in the **Supplementary Protocol**. Aliquots in such plates were diluted 1:5 and then 1:10 to arrive at the working concentration of primary probe sets.

siRNA gene knockdown. Validated siRNAs targeting *ERBB2* (SI02223571, Hs_ERBB2_14), *MYC* (SI00300902, Hs_MYC_5) and *TFRC* (SI00301896, Hs_TFRC_5) were obtained from Qiagen. Reverse transfection was done using Lipofectamine2000 (Invitrogen) according to the manufacturer's specifications. Cells were fixed 3 d after transfection for bDNA sm-FISH.

Quantitative reverse-transcription PCR. RNA was extracted with the RNeasy mini kit including the optional on-column DNA digestion (Qiagen) and reverse transcribed with oligo(dT) primers using the Transcriptor High Fidelity cDNA Synthesis kit (Roche) according to the manufacturers' protocols. Real-time PCR was done with a Mesa Green qPCR Mastermix Plus for Sybr Assay (Eurogentec) with the following primers. hs_TFRC_fwd: catttgtgagggatctgaacca; hs_TFRC_rev: cgagcagaatacagccactgtaa; hs_ERBB2_fwd: agaccatgtccgggaaaacc; hs_ERBB2_rev: caggtagc tcatccccttgg; hs_MYC_fwd: cgactctgaggaggaacaagaa; hs_MYC_rev: actctgacctttgccaggag; hs_TBP_fwd: gcccgaaacgccgaatata; hs_TBP_rev: cgtggctctcttatcctcatga; hs_EEF1A1_fwd: agcaaaaa tgacccaccaatg; hs_EEF1A1_rev: ggcctggatggttcaggata.

Image analysis. All images were analyzed with the image analysis software CellProfiler[19]. Methods required for this study were implemented in Matlab and, when possible, as new CellProfiler modules (see **Supplementary Software**). Nuclei were segmented using images from the 4,6-diamidino-2-phenylindole (DAPI) staining. The cell outlines were then identified using the watershed algorithm. Spot detection was carried out as described in **Supplementary Note 2**. Standard CellProfiler features for intensity, size and texture were then extracted for nuclei and cells. For data cleanup, we applied supervised machine learning with CellClassifier[22,23] to exclude cells showing segmentation problems or aberrant staining, undergoing mitosis or being multinucleated. Computational image analysis was done using the Brutus high-performance computing cluster (ETH Zurich) and the computational task manager iBRAIN[23]. All modules and source code developed for this project can be downloaded at https://github.com/pelkmanslab/.

RNA-seq. Total RNA was extracted from cell lysates using the RNeasy mini kit (Qiagen) with on-column digestion of DNA as specified by the manufacturer. Transcriptome sequencing was performed by LC Sciences. Briefly, RNA quality was assessed using the RNA 600 LabChip (Agilent). Sample preparation was done using the TruSeq RNA Sample Prep Kit v.2 (RS-122-2001, Illumina) as specified by the manufacturer. Enrichment for polyadenylated

doi:10.1038/nmeth.2657

RNA was done using poly(T) beads, and cDNA was obtained from random primers after RNA fragmentation. Sequencing was done using a HiSeq 2000 sequencer from Illumina. Read alignment was done using Bowtie v.0.12.7 (ref. 32) against the human genome (hg19), and FPKM values were generated using TopHat v.1.3.2 (ref. 33) and Cufflinks v.1.3.0 (ref. 34). The FPKM value for a given gene was derived by adding all FPKM values assigned to all transcripts of the gene (**Supplementary Table 6**). For both RNA-seq replicates we obtained ~1.1×10^8 mappable reads, of which ~1.01×10^8 were mapped to exons, ~5.6×10^7 reads mapped to spanning exons and ~8.8×10^6 reads mapped to introns.

Estimation of boundary between low- and high-expressed transcripts. A Gaussian mixture model of corrected mean spots per cell was learned assuming two distributions representing the low- and high-expressed transcripts, respectively. Modeling was done using Matlab. The boundary between the two distributions was set where the probability of being low expressed or high expressed given a mean spot number per cell was equal, i.e., $P(w_1|x) = P(w_2|x)$, where w_1 and w_2 are the low- and high-expressed gene classes, respectively, and x represents a given mean spot number per cell. The computation of the boundary was bootstrapped 1,000 times with replacement. Then the mean boundary value and its s.d. were calculated.

Outlier detection in bDNA sm-FISH vs. RNA-seq comparison. Calculation of the fraction of cells with spot counts above background and mean spot per cell correction was performed according to **Supplementary Note 4**. The correlation plot obtained from $\log_{10}$(FPKM), and corrected $\log_{10}$(spots per cell) was regressed using robust LOESS with the Computational Statistics Matlab library[35]. The shortest distance to the regression line was measured for every gene, and outliers were defined as those points whose distance was bigger than two times the s.d. of all distances.

Calculations of upper and lower detection limits. Conversion of $\log_{10}$(FPKM) to $\log_{10}$(spots per cell) was done by linear regression and extrapolation with the 601 genes whose expression agreed between RNA-seq and high-throughput bDNA sm-FISH. Regression was done with the Matlab Statistics Toolbox "regress" function. Cumulative fractions were calculated by 1,000 bootstrap random samples of 301 genes without replacement, and upper and lower limits of detection were set to the 0.99 and 0.01 cumulative fractions. P values were calculated using a two-sample t-test.

Estimation of the minimal amount of cells required for reproducible cell-to-cell variability. For those transcripts whose mean spot count per cell agreed well with RNA-seq with uncorrected spot counts (612 genes, not shown), we randomly sampled an increasing equal number of cells from each of the two biological replicate experiments. We then calculated the distribution of single cells in each of the samples from zero spots per cell to the highest number of spots per cell, using a bin size of one spot. The Pearson correlation between two distributions within a replicate or between the two replicates was then measured. The procedure was bootstrapped 100 times, and correlation values were computed for every gene and every sampling size, from which the R^2 was then computed. We calculated the Pearson correlation of distribution mean, variance, Fano factor and central moments over all genes at each sampling size for two distributions sampled (i) within a replicate or (ii) between the two replicates. The procedure was bootstrapped 100 times.

Estimation of percentage of genes with highly reproducible multivariate transcript readouts. For each multivariate readout of each gene, its mean ranked distance to its replicate was obtained. This was done by comparing the Euclidean distance of a given gene to all genes of the replicate assay in a given feature space. The ranked distance to the replicate of the same gene was determined. To account for each gene being tested twice, we used for each the mean distance of each replicate. A readout of a gene was defined as highly reproducible when its replicate readout was within the 5% closest distances, unless otherwise specified.

Network construction from the 5% smallest gene-gene distances. Feature selection as described in **Supplementary Note 6** was performed for genes whose mean raw spot count in both replicate assays was at least ten spots per cell and whose raw spot count correlated well with RNA-seq counts (442 genes). For each set of features from individual repetitions of feature selection, Euclidean distances between genes were calculated on features normalized by z-scoring over all genes included in the feature selection. To account for differences in the total amount of features and, thus, the absolute Euclidean distance between individual rounds of elimination, we normalized the Euclidean distances by taking the square root of the square of the Euclidean distances divided by the number of features. The normalized distances at this point were averaged over all 60 iterations for each starting feature set to obtain a mean dissimilarity matrix. Next, gene-gene distances were defined as the Euclidean distance between genes using the mean dissimilarity matrix and then ranked with the smallest distance obtaining rank 1 while excluding similarities of a gene to itself. For network analysis, the top 5% ranking gene-gene distances for every feature set were used.

Calculation of the clustering index for network nodes in subregions. Networks built from different feature sets after selection were used for the calculation of the clustering index between nodes G (genes) belonging to the four subregions of interest (**Fig. 5b**). For every network, edges connecting the G nodes were defined in two categories: k edges that connect G nodes to other G nodes, i.e., these edges connect genes that are within a given subregion, or q edges that connect G nodes to other nodes in the network, i.e., genes that are outside the given subregion. Then the clustering index I is given by the following expression:

$$I = \frac{\sum_{i=1}^{n_g} \frac{(K_i - Q_i)\,\mathrm{sgn}(K_i)}{K_i + Q_i}}{n_g}$$

where n_g is the number of G nodes in the given subregion, K_i is the number of k edges connecting a given G_i node, Q_i is the number of q edges connecting a given G_i node, and $\mathrm{sgn}(K_i)$ is a sign function whose value is 0 when $K_i = 0$ and 1 when $K_i > 0$. Thus the clustering index will be positive if the given G nodes have on average more k edges than q edges.

Calculations of enrichments in STRING 9.0. The reference data set was obtained from the STRING 9.0 database (http://string-db.org/). For each gene-gene distance, the presence of a reported interaction in STRING was determined. For every feature set, the overlap was defined as the fraction of gene-gene distances that was present in STRING 9.0 and whose distance was smaller than the indicated distance. *P* values are given by the hypergeometric probability density function and are the sum of the *P* values of all possibilities that yield at least the observed amount of overlapping gene-gene interactions.

Network analysis. Network analysis and automated force-directed visualization was performed using Cytoscape[36]. Heat maps displaying clustered fractions of cells of the five main types of single-cell spot localization patterns for the example network subregions in **Figure 5c**,**d** and **Supplementary Figure 15d**,**e** were derived from the *z*-scored means of the classification distributions for every pattern type. (**Supplementary Note 6**). Hierarchical clustering using a Euclidean distance and average linkage was performed in Matlab.

Statistical analysis. The bootstrapped samples obtained from calculation of fraction of cells above background (**Supplementary Note 4**) for every replicate gene was compared to the distributions of fractions expected by random using the Mann-Whitney-Wilcoxon test implemented in Matlab. The *P* values obtained were corrected for multiple testing using the Holm-Bonferroni method. To identify genes with fractions of cells above background, we set a conservative significance value of $P = 10^{-4}$.

32. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
34. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
35. Martinez, W.L. & Martinez, A.R. *Computational Statistics Handbook with MATLAB* 2nd edn. (CRC Press, 2008).
36. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

doi:10.1038/nmeth.2657