HybriSeq: Probe-based Device-free Single-cell RNA Profiling

Daniel Foyt [1], David Brown [2], Shuqin Zhou[2], Bo Huang[2,3,4,*]

[1] UCSF-UC Berkeley Joint Graduate Program in Bioengineering, University of California San Francisco, San Francisco, California, 94143, United States of America

[2] Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, 94143, United States of America

[3] Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California, 94143, United States of America

[4] Chan Zuckerberg Biohub - San Francisco, San Francisco, California, 94158, United States of America

* To whom correspondence should be addressed. Tel1: +1 4154761866; Fax: +1 4155141028; Email: bo.huang@ucsf.edu.

Abstract

We have developed the HybriSeq method for single-cell RNA profiling, which utilizes in situ hybridization of multiple probes for targeted transcripts, followed by split-pool barcoding and sequencing analysis of the probes. We have shown that HybriSeq can achieve high sensitivity for RNA detection with multiple probes and profile differential splicing. The utility of HybriSeq is demonstrated in characterizing cell-to-cell heterogeneities of a panel of 95 cell-cycle-related genes and the detection of misannotated transcripts.

Introduction

With its ability to profile individual transcriptomes of many cells, single-cell RNA sequencing (scRNAseq) has proven to be an invaluable tool in understanding cell-to-cell heterogeneity and gene regulatory networks in complex systems (1). Most scRNAseq methods capture polyadenylated RNA and then use reverse transcription to convert it into double-stranded DNA that is compatible with sequencing reactions (2). Although this approach can analyze mRNAs in an unbiased way, the typical detection efficiencies for individual RNA transcripts range between 5-45% (3, 4, 5), largely caused by the inefficiency of the template-switching reaction during reverse transcription. These inefficiencies are particularly deleterious for detection of low copy number RNA and lead to drop out or noisy measurements making classification of subtle phenotypes difficult with few cells. (6).

In contrast to the low detection efficiency in scRNAseq, single-molecule fluorescence *in situ* hybridization (smFISH) regularly achieves a detection efficiency close to 100% by utilizing multiple probes to probe the target RNA directly (7). Taking this concept, single-cell RNA profiling can also be achieved by sequencing multiple *in situ* hybridization probes for one given transcript to decrease the likelihood of a molecule going undetected and increase the measurement confidence. Indeed, several probe-based single-cell RNA profiling methods have been developed recently, such as HyPR-Seq (8), ProBac-seq (9), and 10X Genomics Chromium Flex protocol (10). Due to their probe-based nature, these methods are inherently targeted, allowing for efficient utilization of sequencing reads, and they are not limited to profiling polyadenylated RNA like many scRNAseq methods. On the other hand, they each have their unique limitations. For instance, their probe chemistry either requires complex oligo hybridization and ligation steps, leading to low probe detection efficiency and high background, or simply relies only on hybridization-based specificity, leading to low specificity. Additionally, all of them use microfluidic partitioning of single cells, which can limit the number of cells profiled

and require costly instrumentation. In contrast, highly scalable methods such as SPLiT-Seq (11) and Sci-Plex (12) can sequence millions of cells by utilizing combinatorial indexing. Combining a probe-based approach with combinatorial indexing can thus enable single-cell RNA profiling with both high sensitivity/efficiency and high throughput.

In pursuit of this goal, we developed **Hybri**dization of probes to RNA targets followed by **Seq**uencing (**HybriSeq**). This method involves *in situ* hybridization of multiple split single-strand DNA (ssDNA) probes to one or many target RNAs in fixed and permeabilized cells (Fig. 1a), ligating these split probes hybridized to the RNA to ensure specificity (Fig. 1b), ligating a unique cell barcode to the hybridized probes via two rounds of split-pool barcoding followed by an indexed PCR (Fig. 1c), and sequencing the ligated probe-barcodes, This method can sensitively detect transcripts in a targeted fashion without the need for microfluidics. We demonstrate the utility of this method by profiling the cell-cell heterogeneity in an asynchronous immortalized cell line.

Results

Development and Validation of HybriSeq

To establish a method for efficient hybridization and recovery of ssDNA probes to target RNAs with low nonspecific binding, we performed *in situ* hybridization in fixed and permeabilized HEK293 cells in suspension and quantified the efficiency and specificity of probe recovery by sequencing. We found that ssDNA probes have non-negligible nonspecific binding to the cells, which can contribute to background signal (Supplementary Fig. 1b). We tried to improve the specificity by releasing hybridized but not nonspecifically bound probes using RNAes H digestion of the cells (Supplementary Fig. 1c), but the signal/background ratio was still low even after optimizing hybridization conditions (Supplementary Figs. 1d-e) (see Supplementary Note 1). Therefore, we adopted a method similar to LISH (13), splitting the probe into two parts and ligating hybridized pairs using SplintR ligase that acts on DNA-RNA hybrids (Figs. 1a-b). Bulk-level qPCR measurements of ligated probes in cells showed that with ligation (Supplementary Fig. 2a), it is possible to saturate the probe signal from a high expression transcript (Supplementary Figs. 2b-c) and achieve a specificity > 99% (Supplementary Fig. 2b) (see Supplementary Note 1).

To enable single-cell analysis, we adapted the split-pooling method (11) to uniquely label the probes in individual cells with cell-specific barcodes (see Supplementary Note 2). In 96-well

plates, hybridized and ligated probes are labeled with well-specific barcodes via ligation on the 3' end in two rounds of split and pool procedures followed by a third round of barcoding by PCR with well specific primers (Fig. 1c). Depending on the path a cell takes through this procedure, all the probes in that cell will have one of 884,736 possible unique cell barcodes (CBC), of which < 5% are utilized to avoid excessive CBC collision. Different from previous split-pooling methods, our main challenge is the mixing of barcodes between cells, which can arise from inefficient ligation before pooling, excess ligatable barcode oligos in subsequent ligation reactions, and priming of incompletely ligated species. We screened a variety of barcode ligation and washing/quenching conditions in bulk with qPCR. We found that long ligation times and high barcode oligo concentrations are needed for efficient barcode ligation (Supplementary Fig. 3b). Additionally, we found that quenching barcode oligos as opposed to blocking linker strands resulted in less barcode hopping and that washing away excess barcodes after each ligation step led to significantly less barcode hopping (Supplementary Figs. 3c-f).

To investigate the performance of HybriSeq at the single-cell level, we designed a set of probes (5-6 probes per transcript) targeting mNeonGreen2$_{1-10}$ (mNG) and GFP$_{1-10}$ (GFP) (Supplementary table S1) transcripts. Using human embryonic kidney 293 (HEK293) cells stably expressing either mNG or GFP at a variable range of expression levels (14) we profiled these transcripts with HybriSeq and sequenced libraries to a median per cell saturation of 74% (3990 reads per cell) and observed a total of 691 cells (Supplementary Fig. 4a) (921 cells expected) and a median of 557 UMIs/cell (Unique Molecular Identifier). To determine the single-cell purity of HybriSeq, we performed a cell mixing experiment of the mNG and GFP cells in equal proportions. We observed that 2.6% of CBC contained multiple probes from both mNG and GFP suggesting a doublet rate of 5.2% (Fig. 1d) (from a 50/50 mix of cells, half of the doublets will arise from two cells with the same CBC). UMI counts per probe per cell for three highly expressed transcripts (Supplementary Table S2) were highly correlated between biological replicates with a Pearson's R > 0.99 (Fig. 1e). A median of 99.6% of reads for each CBC were specific to either mNG or GFP probes. These data suggest HybriSeq libraries have a high level of single-cell purity and reproducibility. This multiple rate is higher than the expected multiple rate of 2.45%. This is most likely due to cell clumping, ambient probes, or RNA leaking from cells. While nonzero this is lower than most droplet-based approaches.

HybriSeq specificity arises from both the specific hybridization of ssDNA to transcripts and from the ligation of two adjacent probes hybridized. To evaluate the specificity of HybriSeq, we

examined reads in the library that contained left-probe and right-probe targeting regions not predicted to be adjacent to each other. We compared the amount of these nonspecific ligation events to the specific and correctly ligated events. mNG probes gave > 400,000-fold higher signal than nonspecific ligation events with a median of 302 UMIs per cell, and GFP probes gave > 1,000,000-fold higher signal than nonspecific ligation events with a median of 869 UMIs per cell. The average number of nonspecific UMIs per cell was 0.00023. This result suggests that HybriSeq is highly specific.

Variability of Probe Detection Efficiency

To characterize the variability of detection efficiency for probes designed using our method, we constructed a set of probes completely tiling 6 transcripts (tiling probes) with expression levels ranging from 15 to 165 Transcripts per Million (TPM) in HEK293 cells (Supplementary Fig. 5a, Supplementary Table 4). These transcripts are expected to only have one isoform expressed that does not have expected variation during the cell cycle, which is the main source of heterogeneity in a monoculture cell culture system. Our transcript tiling results reveal that the vast majority of probes targeting the same transcript have low variabilities in the average detected number of UMIs per cell (Fig. 2). For example, SCAF8, ARL5B, and MARVELD1 showed a relatively uniform probe occupancy throughout the length of the transcript with few outliers (Figs. 2c-e). Surprisingly, for the two transcripts with high expression, we observed that continuous stretches of probes are underrepresented or hardly represented at all, despite the otherwise highly uniform probe representation (Fig. 2a-b). For EIF2S2 (ENSG00000125977) the 3' half of the transcript has very few UMIs associated (Fig. 2a). While not as extensive, this depletion is also seen in GHITM (ENSG00000165678) (Fig. 2b). This depletion cannot be explained by CG content nor probe specificity (Supplementary Fig. 5b-c). We confirmed that this underrepresentation of probes for parts of EIF2S2 and GHITM transcripts is not due to poor probe accessibility or detection efficiency issues by removing RNA from its *in situ* context. We performed *In vitro* bulk hybridization and ligation by extracting total RNA from HEK293 cells, treating RNA with proteinase K to remove protein, hybridizing tiling probes, and capturing RNA with Poly(dT) magnetic beads to enrich mRNA. To rule out cis/trans RNA secondary structures we included conditions in which deprotonated RNA was denatured in the presence of probes. We observed close to no representation of probes from the regions that were underrepresented in our single-cell experiments and little difference between native and denatured RNA (Supplementary Fig. 6a). We suspected that these segments of the RNA, which have relatively few probes associated with them, could be missing entirely. To determine if this was indeed the

case, we examined another dataset that captured full-length polyadenylated RNA sequences. To this end, we analyzed 15 long-read sequencing datasets (Supplementary Table 10) from the Long-read RNA-Seq Genome Annotation Assessment Project (15). We observe little representation of probe-depleted regions in all datasets (Supplementary Fig. 6b-c) and a high fraction of reads truncated before their annotated polyadenylation site (Supplementary Figs. 6b-c), suggesting these transcripts may be misannotated in the reference transcriptome. These results demonstrate that HybriSeq has the potential to reveal intra-transcript biological variabilities, such as those resulting from alternative splicing.

Linear Signal Amplification using Multiple Probes

Having demonstrated that the probe-to-probe variability across the same transcript is low with HybriSeq, we reasoned that multiple probes for the same transcript act to linearly amplify the signal to enhance the signal-to-noise ratio (SNR). Thus, we created a simple mathematical model (Supplementary Fig. 7a) by approximating the detection of a specific transcript or probe as a binomial trial (exactly two possible outcomes: detected and not detected). In the case of a scRNAseq measurement, there is only one chance to capture a transcript (Supplementary Fig. 7b), leading to potential “dropout” and excess noise caused by a single priming event. In this scenario, our model predicts that signal scales with the number of transcripts present and the efficiency of capture, the noise is proportional to the square root of the signal, and the SNR is also proportional to the square root of the signal (Supplementary Fig. 7a). Applying this model with the best detection efficiency reported of 45% and an SNR threshold of 2, the lowest number of RNAs reliably detected is 8. With the more typical detection efficiency of 10%, this number is closer to 40 RNAs/cell (Supplementary Fig. S7D). Now for the same model with a linear amplification factor as in HybriSeq (Supplementary Fig. 7c) and an average detection efficiency for a single probe of 20%, a similar or better lower limit of detection can theoretically be accomplished with > 2 probes (Supplementary Fig. 7d).

To confirm our model’s predictions, we subsampled varying numbers of probes for each transcript from our tiling experiment and calculated the signal (UMIs/cell), noise (signal standard deviation), and SNR as a function of number of probes sampled. Our model predicts that for a given transcript number and capture efficiency, the number of UMIs/transcript will increase in a linear fashion with respect to the number of probes subsampled, the noise will fall off as 1/square root of the number of probes subsampled, and the SNR will increase as a function of the square root of the number of probes subsampled (Supplementary Fig. 8a). We observed for

all probed transcripts that our simple model explains the trends in the signal, noise, and SNR (Supplementary Figs. 8b-g). For all but one of the transcripts tested (NEFH), we were able to achieve an SNR > 2 with fewer than 6 probes (Supplementary Figs. 8b-g). In fact, near single-molecule/cell sensitivity can be achieved when > 20 probes are used (e.g. SCAF8 and ARL5B when summing up all probes (Supplementary Fig. 8a).

Quantitative Accuracy of HybriSeq

To demonstrate the profiling of a panel of RNAs using HybriSeq, we constructed a set of probes targeting 95 transcripts (2-4 probes per target) associated with the cell cycle (Supplementary Table S3). These transcripts range in bulk expected expression of 5-355 TPM in HEK293 cells (16). Using this set of probes, we performed HybriSeq for an asynchronous population of HEK293 cells. The resulting bulk expression values correlate well with published bulk RNA-Seq data (16). (r = 0.7) (Fig. 3a). UMI counts per cell for probes targeting the same transcript correlate well, with 72% of the same transcript probe pairs having a Pearson's R > 0.8 (Supplementary Fig. 4b). To determine the effect on measurement precision if fewer probes per transcript were used, we subsampled the number of probes used to calculate a Pearson correlation coefficient. As expected, the use of 3-4 probes per transcript was optimal, while less precise results were seen when 1-2 probes were sampled (Supplementary Fig. 4c) consistent with our model and tiling experiments.

Probing cell-to-cell heterogeneities with HybriSeq

A monoculture of proliferating cells will have cell-cell transcriptional heterogeneity due to the asynchronous progression through the cell cycle. To demonstrate the ability for HybriSeq to characterize such heterogeneities, we analyzed the HybriSeq data set which probes the 95 cell-cycle-related transcripts. Dimensionality reduction was performed on the cell gene matrix and the resulting UMAP projection was clustered with the Leiden algorithm (Fig. 3b). The transcripts with the most variable expression used to define the Leiden clusters showed groupings of genes with similar expression profiles that are typically associated with a particular phase of the cell cycle (Fig. 3c). When transcripts are grouped together based on known association to one of the cell cycle phases, their scaled expression shows a clear transcriptional program (Fig. 3d). These results suggest that the Leiden clusters represent rough boundaries of cell cycle phases. Because clustering approaches like Leiden are less efficient in assigning a cell state along a continuous axis of variation, we also used an alternative approach by calculating a phase score for each cell based on known cell cycle associated genes (Fig. 3e). Based on the binned phase

scores, clustering the cells into three phases, G1, S, and G2M, shows a more biologically representative clustering than the Leiden clustering (Fig. 3f). The proportion of each cell type was similar to a previously published single-cell transcriptome analysis of HEK293 cells (17) when only genes with HybriSeq probes were considered (Supplementary Figs. 9a-b). Additionally, the expression distribution profiles of cell binned by phase score show a clearer trend compared to the subtle trend seen with Leiden clustering (Supplementary Fig. 10) and the pattern of co-expression in the scaled expression profile is much clearer when grouped by G1, S, G2M clusters (Fig. 3g). Notably, our HybriSeq results were obtained using an Illumina MiSeq V3 and substantially fewer reads than other whole transcriptome methods, demonstrating that HybriSeq is an affordable approach to targeted single-cell RNA profiling.

Discussion

Here, we present HybriSeq, a probe-based, microfluidics-free method to sensitively profile a set of targeted RNA in single cells. HybriSeq provides a unique set of advantages that overcome current limitations in scRNAseq approaches. First, by utilizing many probes per transcript, HybriSeq offers the ability to confidently detect low-expression transcripts by decreasing the measurement noise. Second, because of the targeted and scalable nature of probe-based split-pool methods, HybriSeq can cost-effectively profile specific biology in many cells by only including probes for transcripts of interest, which greatly increases the efficiency of sequencing and reduces the cost. Finally, HybriSeq utilizes a split-pool approach to label cells with unique cell barcodes, which eliminates the need for microfluidic devices used in other probe-based single-cell RNA profiling methods (8,9,10). This feature allows for the use of cost-effective, off-the-shelf reagents and a simple protocol that is accessible to most users. The unique features of HybriSeq unlock possibilities that were once unattainable with conventional scRNAseq methods. For example, HybriSeq could profile cell-cell heterogeneity in splicing patterns. The distinctive features of HybriSeq lie in its ability to accurately quantify RNA expression and intra-transcript variation across diverse transcripts, facilitating the study of cellular transcriptional heterogeneity with heightened sensitivity and resolution.

While powerful in its ability to sensitively detect RNA, the sensitivity of HybriSeq and other probe-based single-cell RNA profiling methods is limited by the length of the RNA molecule being measured, which restricts the total number of probe binding sites. This is the case for all *in situ* hybridization-based approaches and methods utilizing random priming or cDNA fragmentation. For short RNA targets, the number of probes able to hybridize to a transcript

could be small even with reduced probe length. A potential workaround to this problem is to use probes with partially overlapped hybridization target regions, as has been utilized in multiplexed FISH methods (7). Moreover, although probe-based methods are efficient in counting transcript copy numbers, they are not designed to sequence the RNA molecule itself, thus rendering it inappropriate for detecting RNA sequence variants or modifications. Last, a limitation of HybriSeq is that probe hybridization and cell barcoding require multiple rounds of washes as well as multiple ligation steps. Each of these steps is associated with inefficiency that contributes multiplicatively to decreased sensitivity. Increasing the probe number per transcript could, in some cases, compensate for these inefficiencies.

Our transcript tiling results have shown low probe-to-probe variability for the vast majority of probes. For some transcripts, targeted probes showed lower abundance than those targeting the rest of the transcript, which we conclude is due to the absence of this region of the transcript and not lack of probe accessibility. Indeed, we confirmed that this absence of probe binding was due to unannotated alternative polyadenylation sites. While it is known that the UTR of transcripts can be highly structured and interact with regulatory proteins (18, 19, 20) we were not able to see this impact probe binding in the transcripts we tested. Considering that certain probes show higher cell-to-cell variabilities compared to other probes targeting the same transcript, this pattern of enrichment/depletion may indeed be indicative of underlying biology pertinent to gene expression regulation and cell-to-cell heterogeneity. In the case of transcripts with alternative splicing, such analysis can be performed by including probes for introns and across splicing junctions, showcasing the advantage of non-3'-biased detection in HybriSeq. Furthermore, investigation into this phenomenon will also yield useful insights into probe design for FISH-based spatial transcriptomic approaches, which rely on hybridization to make measurements.

Methods

HybriSeq Split probe design

HybriSeq ssDNA probes are composed of five regions split into two probes as follows from 5' to 3':

1. (Left probe) 20 nt priming region which is a partial Illumina Nextara read 2 or a different universal priming region.
2. (Left probe) 30 nt left probe targeting region.

3. (Right probe) 30 nt right probe targeting region. The first two bases are either A or T. SplintR ligase has higher efficiency when C or G are not in the first two bases of the ligation site.
4. (Right probe) 8 nt random UMI sequence
5. (Right probe) 20 nt round 1 ligation handle

A probe design pipeline was adapted from Moffitt et al. (21). With minor changes. For calculating gene and isoform level specificity of probes our pipeline only considers the center 30 nt of the targeting region (last 15 nt left probe + first 15 nt of right probe) and does not directly consider melting temperature as a parameter when selecting probes but considers CG content.

Probes were obtained from IDT (Integrated DNA Technologies) in the 50 nmole oPools format or individually as single probes ordered as DNA oligos (Supplementary Tables S1, S2, S3, S4). Probes design and all analysis use human reference genome GRCh38.

Right side probes were 5' phosphorated with T4 Polynucleotide Kinase (New England Biolabs - NEB ref M0201S). Probes were then column cleaned with ssDNA/RNA Clean & Concentrator (Zymo D7010) and quantified. Left-side probes were added at an equal molar concentration and used in hybridization.

HEK293 Cell culture

HEK293 cells were cultured in DMEM + 10% FBS & 1% Penicillin-Streptomycin. Cells were washed twice with 1x PBS, then detached by incubating 2-5 minutes at room temperature with 3 mL of 0.25% Trypsin. Once cells were detached, they were added to 7 mL of media with 10% FBS. In cell mixing experiments, cells were combined at the desired concentrations at this step.

Fixation

Cells were centrifuged for 3 minutes at 500 g at 4 ºC. Cells were washed in 1 mL of 1X PBS. The cells were then passed through a 40 µm strainer into a 15 mL falcon tube and counted. Cells were centrifuged for 3 minutes at 500 g at 4ºC. Cells were resuspended in 0.5 mL/million cells of 4% freshly prepared formaldehyde solution in 1x PBS. Cells were fixed for 30 minutes at room temperature under gentle agitation. Cells were centrifuged for 3 minutes at 500 g at 4 ºC and washed 2 times in 1x PBS. The cells were then passed through a 40 µm strainer into a 15 mL falcon tube and counted.

Hybridization & Ligation

Cells were resuspended in Hybridization buffer (30% formamide, 1% BSA, 0.5% Tween 20, 2X SSC, 40U/ml RNasin) for 10 minutes at 37 ºC under gentle agitation. Cells were centrifuged for 3 minutes at 500g at 4ºC. Cells were resuspended in a Hybridization buffer with probes at 10nM/probe. Cells were incubated at 37 ºC for 18-24 h with gentle agitation. Cells were then washed in wash buffer (20% formamide, 0.5% Tween 20, 4X SSC, 40 U/mL RNasin) two times at 37 ºC for 5 minutes. Cells were washed in ligation buffer (1X T4 DNA Ligase Reaction Buffer (NEB ref B0202S), 0.4 mM ATP, 40 U/mL RNasin) and then resuspended in ligation buffer plus 2 µM SplintR Ligase (NEB ref M0375S). Cells were incubated for 1h at 37 ºC with gentle agitation.

RNA Isolation

After cells were treated with Trypsin and detached from the plate cells were centrifuged for 3 minutes at 500 g at 4 ºC. Cells were washed in 1 mL of 1X PBS. ~5 million cells were centrifuged for 3 minutes at 500 g at 4 ºC and total RNA was extracted with RNeasy Mini kit (Qiagen ref 74104) according to the manufacturer's protocol. Total RNA was then treated with DNase I (NEB ref M0303L) by combining 10ug of total RNA in 1X DNase I reaction buffer (NEB ref M0303L) with 2 units of DNase I (NEB ref M0303S) and heating at 37 ºC for 15 minutes. DNA-depleted RNA was extracted with an RNeasy Mini kit (Qiagen ref 74104) according to the manufacturer's protocol.

***In vitro* Hybridization, Ligation, and mRNA Capture**

1ug of RNA was diluted in *in vitro* hybridization buffer (final concentration: 30% formamide, 2X SSC, 40U/ml RNasin) with probes (10nM/probe) to a total volume of 25ul. For the denatured condition, RNA was heated at 90 ºC for 5 minutes before being incubated at 37 ºC for 18-24 h. For the native condition, RNA was incubated at 37 ºC for 18-24 h. After hybridization, 0.2mg of Oligo d(T)25 Magnetic Beads (NEB ref S1419S) in 75ul of 20X SSC were added to the RNA. RNA was incubated at 37 ºC for 5 minutes before being cooled to 4 ºC at 0.1 ºC/second to anneal RNA to beads. Tubes containing the bead/RNA hybrids were placed against a strong magnet for 2 minutes until the solution was clear. Beads were washed two times in 20X SSC to remove unbound RNA, and wash solution was removed. 50ul of 1X DNA ligation buffer plus 2 µM SplintR Ligase was added to the Bead/RNA hybrids and incubated for 1 hour at 37 ºC to ligate adjacent RNA-bound probes. 150ul of 20X SSC was added to the ligation mix and incubated for 5 minutes at 25 ºC. Tubes containing the bead/RNA hybrids were placed against a

strong magnet for 2 minutes until the solution was clear. Bead/RNA hybrids were resuspended in 1X RNase H buffer containing 5 units of RNase H (NEB ref M0297S) and incubated at for 1 hour at 37 ºC to release probes into solution. Tubes containing the mixture were placed against a strong magnet for 2 minutes until the solution was clear, and the solution was collected for sequencing library preparation.

Preparing Oligos for Ligations

The first and second barcoding steps consist of a ligation reaction. Each round uses a different set of 96 well barcoding plates. Ligation rounds have a universal linker (Supplementary Table S5) strand with partial complementarity to a second strand containing the unique well-specific barcode sequence added to each well (Supplementary Table S6, S7). These strands were annealed together prior to barcoding to create a DNA molecule with three domains: a 15 nt 5' overhang that is complementary to the 15 nt 3' overhang present on the right-side probe, a well-specific barcode sequence, and a 15 nt 3' overhang complementary to the 5' overhang present on the next barcode molecule to be subsequently ligated. For the second-round barcodes, the 3' overhang acts as a universal priming region to which the third round well-specific primer can anneal and extend in a PCR. Barcode strands (IDT) for the ligation rounds are added to 96 well plates and their 5' ends phosphorylated with T4 Polynucleotide Kinase (NEB ref M0201S). After 5' phosphorylation, equal molar amounts of linker strand are added to each well making the final concertation 5.4 µM. Oligos for ligation are annealed by heating plates to 95 °C for 2 minutes and cooling down to 20 °C at a rate of –0.1 °C per second. For ligation reactions, 2.31 µL of barcode/linker oligos are added to 96 well plates to which cells can be added.

Cell barcoding

After probe ligation cells were counted and added to the ligase buffer (1X T4 DNA Ligase Reaction Buffer (NEB ref B0202S), 0.4 mM ATP, 40 U/mL RNasin, 0.5% Tween 20, 1% BSA, 200,000 U/mL T4 ligase) so that the final volume was 1.1 mL at a 22,000 cells/mL. Cells were passed through a 40 µm strainer. 22.69 µL of cells in ligase buffer were added to each well of 48 wells of a 96-well protein low bind plate, which had 2.31 µL of barcode 1 and linker 1 oligos already in each well. Cells were mixed by gently pipetting up and down. Plates were sealed and incubated at 25 ºC for 2 h. 2 µL of 62.5 µM quenching oligo 1 (Supplementary table S5) were added to each well and mixed by pipetting. Plates were sealed and incubated at 25 ºC for 30 minutes. 25 µL of barcode wash buffer (50 mM EDTA, 0.5% Tween 20) was added to each well and incubated for 10 minutes. Cells from all 48 wells were pooled into a single 5ml low-bind

Eppendorf tube. Cells were centrifuged for 3 minutes at 500 g at 4 ºC. Cells were washed two times in barcode wash buffer (+5 μM quenching oligo 1) and then washed in ligase buffer (+5 μM quenching oligo 1, -T4 ligase). Cells were resuspended in 1.1ml ligase buffer (+5 μM quenching oligo 1) and passed through a 40 μm strainer. 22.69 μL of cells in ligase buffer (+5 μM quenching oligo 1) were added to each well of 48 wells of a 96-well protein low bind plate, which had 2.31 μL of barcode 2 and linker 2 oligos already in each well. Cells were mixed by gently pipetting up and down. Plates were sealed and incubated at 25 ºC for 2 h. 2 μL of 62.5 μM quenching oligo 2 were added to each well and mixed by pipetting. Plates were sealed and incubated at 25 ºC for 30 minutes. 25 μL of barcode wash buffer was added to each well and incubated for 10 minutes. Cells from all 48 wells were pooled into a single 5ml low-bind Eppendorf tube. Cells were centrifuged for 3 minutes at 500 g at 4 ºC. Cells were washed two times in barcode wash buffer (+5 μM each of quenching oligo 1 & 2) and then resuspended in ice-cold 1x ThermoPol reaction buffer (NEB ref B9004S) cells were passed through a 40μm strainer and counted. Cell concentration was normalized to 23,000 cells/mL in a cold ThermoPol reaction buffer. 115 cells were dispensed into 8 wells of a strip tube. 20 μL of PCR solution (1x KAPA HiFi HotStart ReadyMix (final concentration) and forward primer) with well-specific round 3 reverse primers (Supplementary Table S8) added to each well, so that the final concentration of each primer was 0.4 μM. PCR thermocycling was performed as follows: 95 °C for 30 sec, then 20 cycles at 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, followed by a final extension at 72 ºC for 30 seconds.

Library Preparation for Sequencing, Single Cell

Round 3 PCR reactions were centrifuged at full speed for 1 minute to pellet cells. All round 3 PCR reaction solution was removed, pooled, and column purified with the Zymo DNA clean & concentrator kit (Zymo 11-305). Purified libraries were analyzed on an Agilent TapeStation Systems (D1000 kit) to check for the correct size. If the predominate band was the correct size (252 ± 2 bp or 232 ± 2 bp depending if the left probe included a partial read 2 sequence) and was < 90% of the library the purified PCR product was run on a 2% agarose (TBE) electrophoresis gel (200V 20 minutes) and the correct size band was cut out and extracted from the agarose with the Zymo Gel recovery kit (Zymo D4002). We observe that libraries that contained left probes containing the non-read 2 priming regions produced some nonspecific amplification requiring size selection purification. The purified pooled round 3 DNA product was placed into a final limited cycle PCR to add Illumina sequencing adaptors. The adapter addition PCR reaction was as follows: 0.5 ng DNA from pooled round 3 PCR product, 0.4 μM P7 forward

primer, 0.4 μM P5 reverse primer and 1X KAPA HiFi HotStart ReadyMix. PCR thermocycling was performed as follows: 95 °C for 30 sec, then 10 cycles at 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, followed by a final extension at 72 ºC for 30 seconds. The PCR reaction was removed and purified with a 0.8X ratio of SPRI beads to generate an Illumina-compatible sequencing library.

Library Preparation for Sequencing, Bulk

For bulk *in vitro* experiments, 1ul of RNase H digestion product was loaded into a 25 μL 1x KAPA HiFi HotStart ReadyMix PCR reaction with 0.3 μM primers containing partial Illumina sequencing adapters. PCR thermocycling was performed as follows: 95 °C for 30 sec, then 15 cycles at 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, followed by a final extension at 72 ºC for 30 seconds. 1ul of this unpurified PCR reaction was then added to a similar PCR reaction containing the primers for the addition of the full Illumina sequencing adaptors. PCR thermocycling was performed as follows: 95 °C for 30 sec, then 8 cycles at 95 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 30 seconds, followed by a final extension at 72 ºC for 30 seconds. The PCR reaction was removed and purified with a 1.3X ratio of SPRI beads to generate an Illumina-compatible sequencing library.

Illumina Sequencing

15 Pm libraries were sequenced on a MiSeq (Illumina) using a 150 nucleotide (nt) V3 kit in paired-end format. Read 1 (75 nt) covered the cell barcode and read 2 (75 nt) covered the probe and UMI.

RNase H specificity

After non-split probe hybridization and washing, cells were resuspended in RNase H reaction buffer containing 20 U/mL of RNase H enzyme (NEB M0297S). Cells were incubated for one hour at 37 °C with gentle agitation. Released probes were quantified with sequencing or qPCR.

qPCR

qPCR was performed on probes or barcoded probes (Supplementary table S9) released from cells via RNase H release or heat release. Cells and released probes were centrifuged to pellet cells and the supernatant was purified with spin columns and DNA was eluted in 20ul of water (Zymo ssDNA/RNA clean & concentrator). 1 μL of purified samples were loaded into each 20 μL reaction of a qPCR with 0.3 μM primers (Supplementary table S9) according to manufacturer's

instructions using Maxima SYBR Green qPCR Master Mix (Thermo Fisher Ref K0222). Thermocycling and measurements were performed on a QuantStudio™ 5 System with the follows temperatures: 95 ºC for 60 seconds, 45 cycle of 95 ºC for 15 seconds and 60 ºC for 30 seconds (1.6 ºC/second ramp rate) in which the fluorescence was recorded at the end of each cycle. QuantStudio™ Design and Analysis Software v1.4.1 was used to analyze fluorescence signal and calculate C_T values. A standard curve was made by running a dilution series of the arget oligo (ordered from IDT), and the Ct values from this were used as the standard curve from which the concentrations of target oligos in the sample were determined. To measure the % ligated barcode in supplementary Fig. 3 the concentration of ligated species (right side probe + barcode 1) was compared to the concentration of just the right-side probe present in the reaction.

HybriSeq Computational Pipeline

We constructed a pipeline to analyze HybriSeq data by taking raw sequencing reads and constructing a count matrix (counts per probe per cell). Briefly, we identify real barcodes, identify probe-targeting regions with correct ligation, remove duplicates using UMIs, and filter out reads not containing barcodes or probe-targeting regions. Detailed key steps were as follows:

1. From the demultiplexed FASTQs generated by the Illumina analysis software, we filtered out reads not containing common regions contained in barcode one, two, and three in the correct location.
2. To determine the unique barcode, a whitelist of each round of barcode sequences was constructed including barcodes within a hamming distance of two. With this list, barcode sequences for each round of split pool indexing were determined from read 1. From this, a unique cell barcode was constructed. Reads for which no barcode could be found were excluded.
3. To determine the targeting region from read 2, a whitelist of each probe was constructed including probe sequences within a hamming distance of two. With this list, both left- and right-side probe targeting regions were determined. Reads containing targeting regions not predicted to be adjacent were excluded. From read 2 we also extracted the 8 bp simple UMI included on the right-side probe.
4. We constructed a data frame of reads that included the unique cell barcode, probe targeting region, and simple 8bp UMI. We then collapsed duplicate reads by considering

a combined UMI which contained the 8bp simple UMI, the unique cell barcode, and the probe targeting region.

5. We generate a count table of UMIs per probe per unique cell barcode or UMIs per transcript per unique cell barcode.
6. To determine which unique cell barcodes were associated with real cells a threshold for UMIs/cell was calculated by taking 10% of the 99th percentile of the top set of unique cell barcodes equal to the number of expected cells and considering a doublet rate of ~5% or visually setting the threshold at the first knee of the cell rank - UMI plot. We note that when only considering lowly or highly variably expressed transcripts, the inclusion of probes targeting moderately and stably expressed transcripts can help set a threshold.
7. The Scanpy library in Python was used for all standard single-cell analyses.

Mixing/ Single-Cell Purity Experiment

Two HEK293 cell lines, each containing a specific transcript (mNeonGreen2$_{1-10}$ (mNG) and GFP$_{1-10}$ (GFP)) were subjected to the standard HybriSeq protocol. Probes targeting mNG and GFP (Supplementary Table S1) were added to the probe mixture during the hybridization step. At the PFA fixation step, equal concentrations of each cell line were mixed.

Probe Tiling Analysis

Each transcript was analyzed independently only considering probes targeting that transcript. Probe counts for each cell were normalized so that the total sum of all normalized counts in each cell was equal to the median UMIs/cell of the cell population. This was done to account for differences in expression levels between cells. The average relative counts were taken for each probe and plotted as a trace for all cells. The standard deviation was calculated for each probe. For bulk *in vitro* and *in situ* tiling experiments, each transcript was analyzed independently, only considering probes targeting that transcript. Probe counts were normalized so that the total sum of all normalized counts was equal to 200. The standard deviation was calculated for each probe from three independent experiments.

Long-read RNA-Seq Genome Annotation Assessment Project (LRGASP) Data Analysis

15 datasets were selected from the LRGASP by querying untreated/unperturbed polyadenylated long-read human cell line RNA sequencing data generated using the Pacific Biosciences Sequel system, which was replicated and had sufficient sequencing depth (Supplementary Table S10). Binary Alignment Map (BAM) files, excluding unfiltered BAM files, were retrieved from ENCODE

(Supplementary table S10). Sequences that mapped to chr20: 34088309-34112243 (EIF2S2) and chr10:84139482-84154841 (GHITM) were extracted from BAM files using Samtools. For each annotated transcript sequence (EIF2S2: ENST00000374980.3, GHITM: ENST00000372134.6) a list of every 20bp subsequence was created. The number of occurrences of each 20bp subsequence in each sample's extracted alignments was calculated for both transcripts separately and normalized to the total number of alignments in each dataset to generate the fraction of reads which contain each 20bp subsequence. Standard deviation was calculated across the 15 datasets.

Measurement Variability Model and Simulation

To model measurement noise associated with sampling a specific transcript in a cell we started off by making a few assumptions.

- Sampling of a transcript in a cell can be modeled with a Poisson distribution.
- The probability of capturing a transcript or probe is the same for all probes targeting the same transcript or priming events.
- The background signal from random probe ligation is minimal and can be assumed to be negligible.
- Probe binding a transcript does not influence different probes binding the same transcript.
- Probes are hybridized at a saturating concentration.
- The underlying cell-cell heterogeneity can be modeled as a constant value of standard deviation and is not dependent on the number of probes used.
- All cells have the same efficiency of detection for the same transcript.

To model single cell transcript measurement variability:

Let N be the number of specific transcripts in a cell, n be the number of detection chances per transcript in a cell, e be the efficiency at which n is successfully detected, and C be the number of counts or UMIs for a specific transcript. If we assume that N is Poisson, the variability associated with counts C is equal to the mean of C and we define measurement noise as the standard deviation of the measurement C:

$$C = (N)(n)(e), \quad Noise_C = \sigma = \sqrt{(N)(n)(e)} \qquad (1,2)$$

Taking the ratio of the counts C to the noise associated with C we get the signal to noise ratio (SNR)

$$SNR = \sqrt{(N)(n)(e)} \qquad (3)$$

For a population of cells, C will scale linearly with n. If we define expression, M, as C normalized to the number of probes used to make the measurement, expression is given by:

$$M = \frac{C}{n} = \frac{(N)(e)(n)}{n} = (N)(e), \quad Noise_M = \frac{\sqrt{(N)(n)(e)}}{n} + b = \sqrt{\frac{M}{n}} + b \qquad (4,5)$$

Here we assume that that the contribution to noise in the expression measurement from biology is independent of the number of probes used to make the measurements and can be defined as constant b.

The expression SNR is then given by the ratio of M to Noise associated with M:

$$SNR_M = \frac{M}{\sqrt{\frac{M}{n}} + b} \qquad (6)$$

For the simulations in fig 2 we experimentally determine M by taking the slope of the line fit to the UMI/cell – probe number plot. This slope is the number of counts you would expect to gain for each additional probe included in the analysis.

We then non-linearly used least squares to fit the function for Noise of M to the standard deviation of M as a function of the number of probes used to make the measurement keeping M constant from the experimentally determined M and only fitting the model by optimizing b.

Calculation of signal, standard deviation, and SNR for multiple probes

Total probe counts for each cell were normalized so that the total sum of all normalized counts in each cell was equal to the total median UMIs/cell of the cell population. This was done to account for differences in expression levels between cells, as the goal is to gain an understanding of the measurement-associated variation and not necessarily the underlying inherent biological variation. To calculate the average signal or counts, for each number of probes considered (n), a random set of probes was chosen without replacement, and the number UMIs/cell was calculated along with a standard deviation for each n. To calculate the SNR, the ratio of average expression (UMIs/cell/n) to the standard deviation of expression was calculated for all n. This procedure was repeated 10,000 times, randomly sampling the set of probes used to make the measurement and the average and standard deviations of these calculations were plotted.

References

1. Osumi-Sutherland, D., Xu, C., Keays, M., Levine, A.P., Kharchenko, P.V., Regev, A., Lein, E., Teichmann. S.A. Cell type ontologies of the human cell atlas. *Nat Cell Biol*. **11**, 1129-1135 (2021).
2. Wulf, M.G., Maguire, S., Humbert, P., Dai, N., Bei, Y., Nichols, N.M., Corrêa, I.R., Guan, S. Non-templated addition and template switching by moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each otherMMLV-type reverse transcriptase template switching. *J BIOL CHEM*. **48**, 18220-18231 (2019).
3. Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y., Wang, J. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq. Systems. *Mol. Cell*. **73**,1,130-142 (2019).
4. Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J., Hellmann, I. & Enard, W. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun*. **9**, 2937 (2018).
5. Svensson, V., Natarajan, K.N., Ly, L., Miragaia, R.J., Labalette,C., Macaulay, I.C., Cvejic, A., & Teichmann, S.A. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods.* **14**, 381–387(2017).
6. Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, N., Gruen, D., Lau, J.K. and Boutet, S.C. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature biotechnology* ***38***(6), pp.747-755 (2020).
7. Wang, G., Moffitt, J.R. and Zhuang, X. Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Scientific reports* ***8***(1), p.4847 (2018).
8. Marshall, J.L., Doughty, B.R., Subramanian, V., Guckelberger, P., Wang, Q., Chen, L.M., Rodriques, S.G., Zhang, K., Fulco, C.P., Nasser, J. and Grinkevich, E.J. HyPR-seq: single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes. *Proceedings of the National Academy of Sciences* ***117***(52), pp.33404-33413 (2020).
9. McNulty, R., Sritharan, D., Pahng, S.H., Meisch, J.P., Liu, S., Brennan, M.A., Saxer, G., Hormoz, S. and Rosenthal, A.Z. Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. *Nature Microbiology* ***8***(5), pp.934-945 (2023).
10. Janesick, A., Shelansky, R., Gottscho, A., Wagner, F., Rouault, M., Beliakoff, G., de Oliveira, M.F., Kohlway, A., Abousoud, J., Morrison, C. and Drennon, T.Y. High

resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *BioRxiv* pp.2022-10 (2022).

11. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W. and Pun, S.H. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* ***360***(6385), pp.176-182 (2018).
12. Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L. and Zhang, F. Massively multiplex chemical transcriptomics at single-cell resolution. *Science* ***367***(6473), pp.45-51 (2020).
13. Credle, J.J., Itoh, C.Y., Yuan, T., Sharma, R., Scott, E.R., Workman, R.E., Fan, Y., Housseau, F., Llosa, N.J., Bell, W.R. and Miller, H. Multiplexed analysis of fixed tissue RNA using Ligation in situ Hybridization. *Nucleic acids research* ***45***(14), pp.e128-e128 (2017).
14. Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M.D. and Huang, B. Improved split fluorescent proteins for endogenous protein labeling. *Nature communications* ***8***(1), p.370 (2017).
15. Pardo-Palacios, F.J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J.E., De María, M., Adams, M.S., Balderrama-Gutierrez, G. and Behera, A.K. Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv* (2021).
16. Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P. and Edfors, F. A single–cell type transcriptomics map of human tissues. *Science advances* ***7***(31), p.eabh2169 (2021).
17. Tran, V., Papalexi, E., Schroeder, S., Kim, G., Sapre, A., Pangallo, J., Sova, A., Matulich, P., Kenyon, L., Sayar, Z. and Koehler, R. High sensitivity single cell RNA sequencing with split pool barcoding. *Biorxiv* pp.2022-08 (2022).
18. Berkovits, B.D. and Mayr, C. Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature* ***522***(7556), pp.363-367 (2015).
19. Lee, S.H. and Mayr, C. Gain of additional BIRC3 protein functions through 3-UTR-mediated protein complex formation. *Molecular cell* ***74***(4), pp.701-712 (2019).
20. Mayr, C. Regulation by 3′-untranslated regions. *Annual review of genetics* ***51***, pp.171-194 (2017).

21. Moffitt, J.R., Hao, J., Wang, G., Chen, K.H., Babcock, H.P. and Zhuang, X. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* ***113***(39), pp.11046-11051 (2016).

Acknowledgements

We thank Eric Chow for the valuable discussions. This work was partially supported by the National Institutes of Health [DK127421 to B.H.]. Funding for open access charge: National Institutes of Health. B.H. is a Chan Zuckerberg Biohub - San Francisco Investigator.

Author Contributions

D.F. performed the experiments, analyzed the data, and contributed to the interpretation of the results. B.H. conceived the project. D.B. and S.Z. conceived and designed the proof-of-concept experiments. B.H. and D.F wrote the manuscript.

Competing Interests

The authors declare no conflict of interest.

Figure Legends

Fig. 1. Schematic and validation of HybriSeq. **a.** Multiple split probes are designed per transcript of interest. **b.** Hybridization and ligation of split probes. **c.** Labeling probes with unique cell barcodes via the split-and-pool method. Rounds 1 and 2 barcodes are ligated, and round 3 barcodes are added via PCR. **d.** Sequencing results of 1:1 mixed HEK293 cells stably expressing either mNG or GFP. **E.** Scatter plot of average HybriSeq UMIs/HEK293 cell in two independent biological replicates. Each dot represents a single probe for either GAPDH, RPL13A, or ACTB.

Fig. 2. Intra-transcript Probing Variation. a-f. Relative probe counts for probes targeting the specific transcript with standard deviation across all cells plotted for each probe. Gray lines are traces for individual cells, and the blue line is the average for each probe across all cells. The means of average counts for all probes reported to the right with standard deviations. (*) For EIF2S2, only the first 11 out of 24 probes were considered in calculating the mean and standard deviation. For GHITM only the first 17 out of 20 probes were considered in calculating the mean and standard deviation.

Fig. 3. Analyzing cell cycle-regulated transcripts using HybriSeq. a. Scatter plot of cell line matched bulk RNA-Seq TMP number and HybriSeq average UMIs/cell for 95 cell cycle associated genes. Each point represents a transcript measured. **b.** UMAP of cell cycle transcripts in HEK293 cells measured with HybriSeq colored by Leiden cluster. **c.** Heat map of scaled expression values for the top 6 differentially expressed genes for each cluster. **d.** Heat map of scaled expression values for gene grouped together by association with specific phases of the cell cycle and Leiden cluster. **e.** UMAP colored by the Phase score calculated from all genes. **f**. UMAP in e colored based on binned phase score into G1, S, G2-M. **g.** Heat map of scaled expression values for gene groped buy by association with specific phases of the cell cycle and binned phase score in **f**.

Figure 1

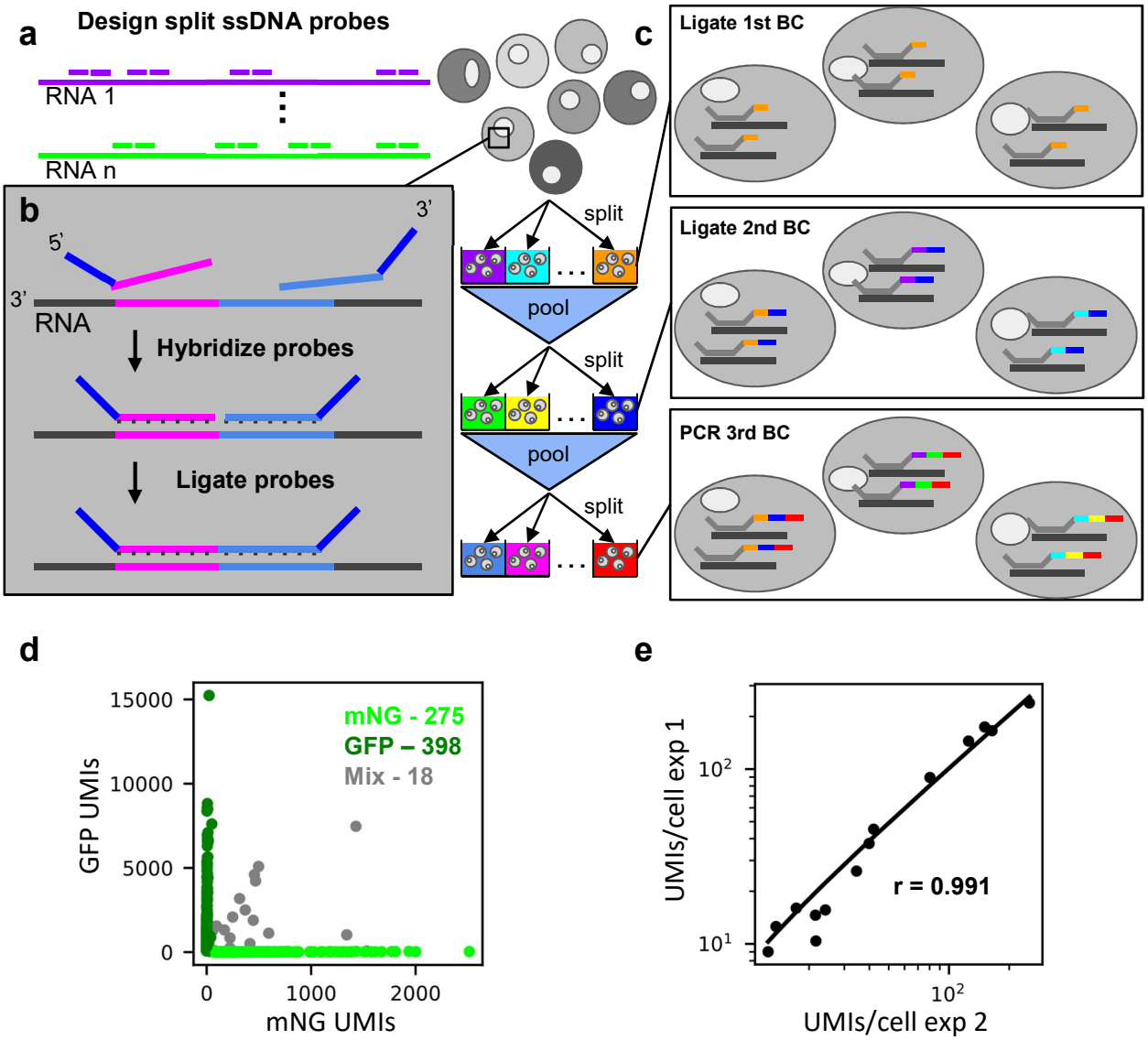

a
Design split ssDNA probes
RNA 1
RNA n
b
5'
3'
3'
RNA
Hybridize probes
Ligate probes
c
split
pool
split
pool
split
Ligate 1st BC
Ligate 2nd BC
PCR 3rd BC
d
mNG - 275
GFP – 398
Mix - 18
GFP UMIs
mNG UMIs
15000
10000
5000
0
0
1000
2000
e
UMIs/cell exp 1
UMIs/cell exp 2
r = 0.991
10^2
10^1
10^2

Figure 2

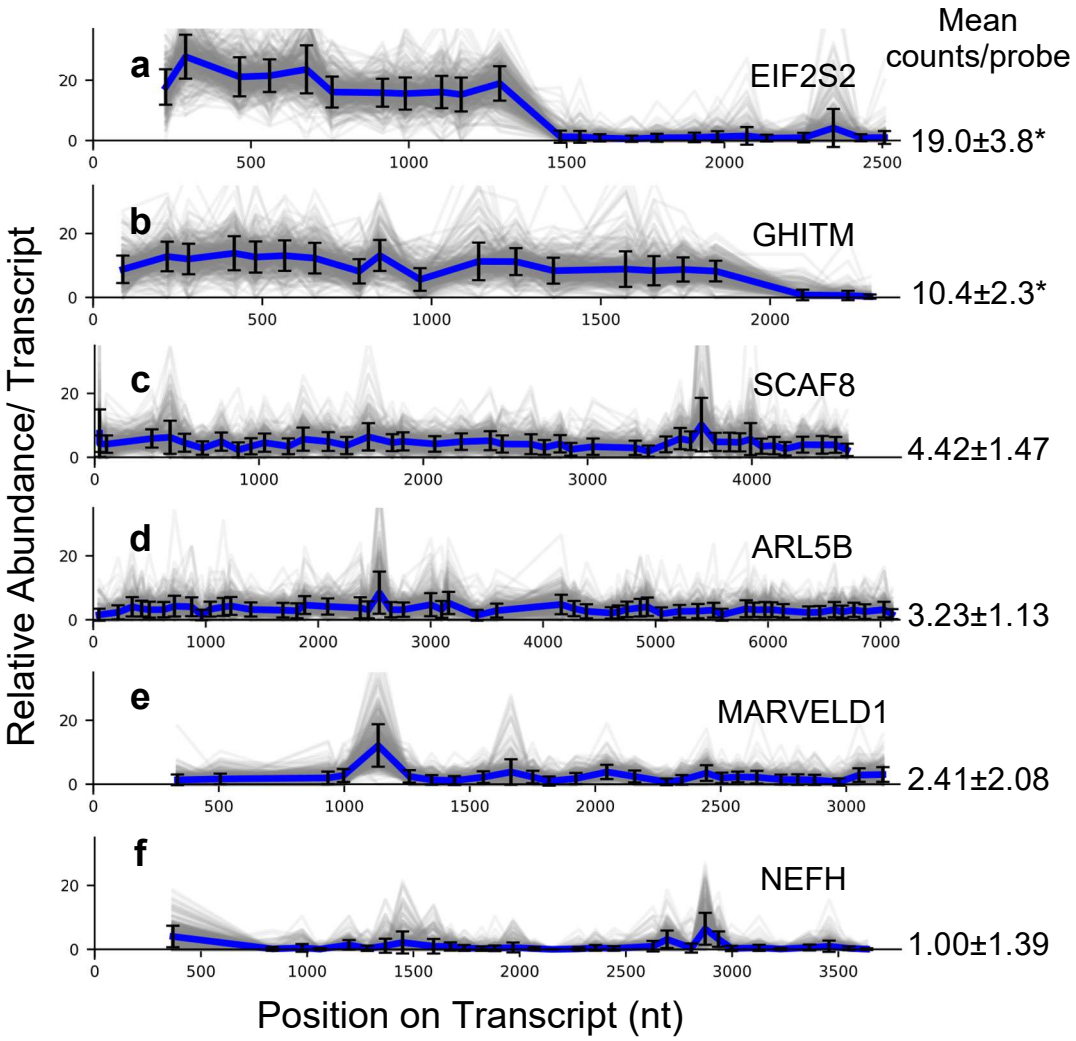
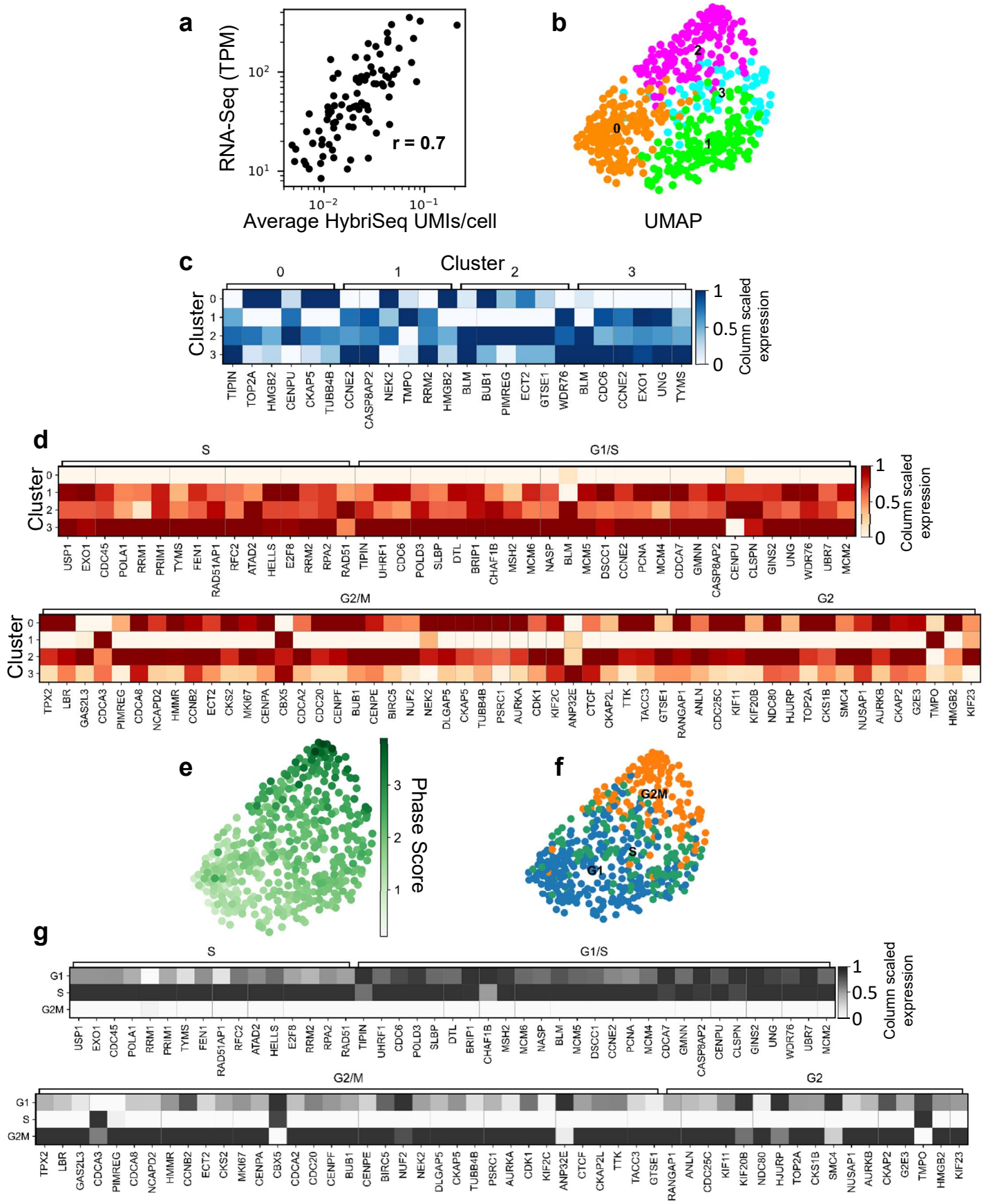

Mean
counts/probe
a
EIF2S2
19.0±3.8*
b
GHITM
10.4±2.3*
c
SCAF8
4.42±1.47
d
ARL5B
3.23±1.13
e
MARVELD1
2.41±2.08
f
NEFH
1.00±1.39
Relative Abundance/ Transcript
Position on Transcript (nt)

Figure 3


a
RNA-Seq (TPM)
r = 0.7
Average HybriSeq UMIs/cell
b
UMAP
c
Cluster
Column scaled expression
TIPIN TOP2A HMGB2 CENPU CKAP5 TUBB4B CCNE2 CASP8AP2 NEK2 TMPO RRM2 HMGB2 BLM BUB1 PIMREG ECT2 GTSE1 WDR76 BLM CDC6 CCNE2 EXO1 UNG TYMS
d
S
G1/S
USP1 EXO1 CDC45 POLA1 RRM1 PRIM1 TYMS FEN1 RAD51AP1 RFC2 ATAD2 HELLS E2F8 RRM2 RPA2 RAD51 TIPIN UHRF1 CDC6 POLD3 SLBP DTL BRIP1 CHAF1B MSH2 MCM6 NASP BLM MCM5 DSCC1 CCNE2 PCNA MCM4 CDCA7 GMNN CASP8AP2 CENPU CLSPN GINS2 UNG WDR76 UBR7 MCM2
G2/M
G2
TPX2 LBR GAS2L3 CDCA3 PIMREG CDCA8 NCAPD2 HMMR CCNB2 ECT2 CKS2 MKI67 CENPA CBX5 CDCA2 CDC20 CENPF BUB1 CENPE BIRC5 NUF2 NEK2 DLGAP5 CKAP5 TUBB4B PSRC1 AURKA CDK1 KIF2C ANP32E CTCF CKAP2L TTK TACC3 GTSE1 RANGAP1 ANLN CDC25C KIF11 KIF20B NDC80 HJURP TOP2A CKS1B SMC4 NUSAP1 AURKB CKAP2 G2E3 TMPO HMGB2 KIF23
e
Phase Score
f
G2M
S
G1
g
G1
S
G2M