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

AUTHORS

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 RNA accessibility. The utility of HybriSeq is demonstrated in characterizing cell-to-cell heterogeneities of a panel of 95 cell-cycle-related genes and the probe-probe heterogeneity within a single transcript.

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 ranges 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 poly adenylated 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 requires 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 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 (Figure 1A), ligating these split probes hybridized to the RNA to ensure specificity (Figure 1B), ligating a unique cell barcode to the hybridized probes via 2 rounds of split-pool barcoding followed by an indexed PCR (Figure 1C), and sequencing the ligated probe-barcodes (Figure 1D), 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.

MATERIAL AND 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. (20). 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 table S1, S2, S3, S4).

Right side probes were 5' phosphorated with T4 Polynucleotide Kinase (NEB). 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 min 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 min 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 min 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 min at room temperature under gentle agitation. Cells were centrifuged for 3 min 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 10min at 37 ºC under gentle agitation. Cells were centrifuged for 3 min at 500g at 4ºC. Cells were resuspended in 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 time at 37 ºC for 5min. Cells were washed in ligation buffer (1X T4 DNA Ligase Reaction Buffer (NEB), 0.4 mM ATP, 40 U/mL Rnasin) and then resuspended in ligation buffer plus 2 µM SplintR Ligase (NEB). Cells were incubated for 1h at 37 ºC with gentle agitation.

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). 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 cell can be added.

Cell barcoding

After probe ligation cells were counted and added to the ligase buffer (1X T4 DNA Ligase Reaction Buffer (NEB), 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 min. 25 µL of barcode wash buffer (50 mM EDTA, 0.5% Tween 20) was added to each well and incubated for 10 min. Cells from all 48 wells were pooled into a single 5ml low bind Eppendorf tube. Cells were centrifuged for 3 min 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 min. 25 µL of barcode wash buffer was added to each well and incubated for 10 min. Cells from all 48 wells were pooled into a single 5ml low bind Eppendorf tube. Cells were centrifuged for 3 min 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) cells were passed through a 40µm strainer and counted. Cell concentration was normalized to 23,000 cells/mL in 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 and Sequencing

Round 3 PCR reactions were centrifuged at full speed for 1 min 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 20min) 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.

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 of target 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 was determined. To measure the % ligated barcode in supplementary figure S3 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 were 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 that 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 analysis.

Mixing/ Single-Cell Purity Experiment

Two HEK293 cell lines, each containing a specific transcript (mNeonGreen$_{1-10}$ (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 or pseudo bulk clusters. The standard deviation was calculated for cell populations for each probe.

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 underlaying 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 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.

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 (Supplementary Figure S1A). We found that ssDNA probes have non-negligible nonspecific binding to the cells, which can contribute to background signal (Supplementary Figure S1B). We tried to improve the specificity by releasing hybridized but not nonspecifically bound probes using RNAes H digestion of the cells (Supplementary Figure S1C), but the signal/background ratio was still low even after optimizing hybridization conditions (Supplementary Figure S1D) (see Supplementary Notes). 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 (Figure 1A, B). Bulk level qPCR measurements of ligated probes in cells showed that with ligation it is possible to saturate the probe signal from a high expression transcript (Supplementary Figure S2B-C) and achieve a specificity > 99% (Supplementary Figure S2B) (see supplemental notes).

To enable single cell analysis, we adapted the split-pooling method (11) to uniquely label the probes in individual cells with cell specific barcodes. 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 (Figure 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 Figure S3B). 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 Figure S3C-F). 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 mNeonGreen$_{1-10}$ (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 Figure S4A) (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% (Figure 1E) (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 (Figure 1G). 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 specific hybridization of ssDNA to transcripts and from the ligation of two adjacent probes hybridized. To evaluate the specificity of HybriSeq we looked at 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 302 UMIs per cell, and GFP probes gave >1,000,000-fold higher signal then nonspecific ligation events with a median 869 UMIs per cell. The average number of nonspecific UMIs per cell was 0.00023. This result suggests that HybriSeq is highly specific.

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 Transcripts per million (TPM) in

HEK293 cells (15). Using this set of probes, we performed HybriSeq for an asynchronous population of HEK293 cells. The resulted bulk expression values correlate well with published bulk RNA-Seq data (15). (r = 0.7) (Figure 1F). UMI counts per cell for probes targeting the same transcript correlate well, with 72% of same transcript probe pairs having a Pearson's R > 0.8 (Supplementary Figure S4B). 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. The use of 3-4 probes per transcript was optimal, while less precise results were seen when 1-2 probes were sampled (Supplementary Figure S4C).

Next, we quantify the relationship between measurement noise and probe number using a simple mathematical model. In many high throughput scRNAseq methods, an individual transcript frequently "dropped out" in the digital gene expression count, making the measurement of lowly expressed transcripts excessively noisy. This issue is in part the result of the nature in which transcripts are sampled by a single priming event at the poly-A tale of transcripts (Supplementary Figure S5B), followed by losses in subsequent reverse transcription and capturing steps. With only one chance to detect a transcript, the probability of detecting that transcript becomes a binomial trial with exactly two outcomes (detected and not detected). The use of multiple probes in HybriSeq, on the other hand, serves as a linear amplification of the transcript before the lossy detection. We approximate the detection of a specific transcript as a Bernoulli trial and modeled with Poisson sampling. In this case, the signal to noise ratio (SNR) in a typical scRNAseq measurement is approximately the square root of the product of the molecules present and the efficiency of capture (Supplementary Figure S5A). Applying this model with the best detection efficiency reported of 45% and a SNR threshold of 2 the lowest number of molecules reliably detected is 8. With the more typical detection efficiency of 10% this number is closer to 40 molecules (Supplementary Figure S5D). Now for the same model with a linear amplification factor as in HybriSeq (Supplementary Figure S5C) 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, consistent with our subsampling analysis Moreover, near single-molecule sensitivity can be achieved when >10 probes are used (Supplementary Figure S5D)

To test our model, we constructed a set of probes completely tiling six transcripts with an expression level from 15-165 TPM in HEK293 cells (Supplementary Figure S5E, Supplementary table S4). 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 system. Our model predicts that for a given transcript/cell value and efficiency of capture the number of UMIs/transcript will increase in a linear fashion with respect to the number of probes subsampled from the measurement (Figure 2A), the standard deviation of the expression (UMIs/transcript/unique probes) will fall off 1/square root of the probe number (Figure 2B), and the SNR will increase as a function of the square root of the probe number (Figure 2C). We observed for all probed transcripts that our simple model explains the trends in the SNR (Figure 2D-F). For all but one (NEFH) of the transcripts tested we were able to achieve a SNR > 2 with fewer than 6 probes (Supplementary Figure S6A-E).

Our transcript tiling results also reveal probe-to-probe variabilities that cannot be explained by CG content nor probe specificity (Supplementary Figure S6F-G). In particular, we observed that certain probes are underrepresented in the sequencing readout relative to the average probe number for a specific transcript or hardly represented at all (Figure 2G-H). For EIF2S2 (ENSG00000125977) the 3′ half of the transcript has very few UMIs associated with it (Figure 2G). We found that this region is mostly composed of the 3′ untranslated region (UTR). While not as pronounced, this depletion is also seen in GHITM (ENSG00000165678) (Figure 2H) and NEFH (ENSG00000100285 (Figure 2I). In contrast, SCAF8, ARL5B, and MARVELD1 showed much more uniform probe occupancy throughout the length of the transcript (Figure 2J-L). Within the cell population profiled, we also observed elevated cell-to-cell heterogeneity in occupancy for a subset of probes (Figure 2M- N). These differences in probe occupancy may be attributed to differentially regulated RNA processing, RNA-protein interactions, secondary structures, etc.

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 (Figure 3A). 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 (Figure 3B). When transcripts are grouped together based on known association to one of the cell cycle phases, their scaled expression shows a clear transcriptional program (Figure 3C). 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 (Figure 3D). 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 (Figure 3E). The proportion of each cell type was similar to a previously published single-cell transcriptome analysis of HEK293 cells (16) when only genes with HybriSeq probes were considered (Supplementary Figure SA-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 Figure S8, S9) and the pattern of co-expression in the scaled expression profile is much clearer when grouped by G1, S, G2M clusters (Figure 3F). 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 transcript accessibility of regulatory RNA or used to understand cis- and trans- RNA interactions regulating translation. The distinctive features of HybriSeq lies in its ability to accurately quantify RNA expression and accessibility 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 for HybriSeq is that probe hybridization and cell barcoding require multiple rounds of washes as well 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 probe-to-probe variabilities that cannot be explained by CG content or specificity for transcripts not known to be alternatively spliced in the cells used. For some transcripts, 3′-UTR targeted probes showed lower abundance than those targeting the rest of the transcript. It is known that the UTR of transcripts can be highly structured and interact with regulatory proteins (17, 18, 19). Therefore, RNA-protein interaction, cis- and trans- RNA interactions, and overall molecule accessibility might partially explain these differences in probe reads. Further 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 underlaying biology pertinent to gene expression regulation and cell-to-cell heterogeneity. In the case of transcripts with alternative splicing, such analysis can still 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.

ACKNOWLEDGEMENTS

We thank Eric Chow for valuable discussions.

FUNDING

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.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

REFERENCES

1. David Osumi-Sutherland, Chuan Xu, Maria Keays, Adam P Levine, Peter V Kharchenko, Aviv Regev, Ed Lein, Sarah A Teichmann (2021) Cell type ontologies of the Human Cell Atlas. Nat Cell Biol. 11, 1129-1135.
2. Madalee G. Wulf, Sean Maguire, Paul Humbert, Nan Dai, Yanxia Bei, Nicole M. Nichols, Ivan R. Corrêa Jr., Shengxi Guan (2019) 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.
3. Xiannian Zhang, Tianqi Li, Feng Liu, Yaqi Chen, Jiacheng Yao, Zeyao Li, Yanyi Huang, Jianbin Wang (2019) Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol. Cell.,73,1,130-142.
4. Johannes W. Bagnoli, Christoph Ziegenhain, Aleksandar Janjic, Lucas E. Wange, Beate Vieth, Swati Parekh, Johanna Geuder, Ines Hellmann & Wolfgang Enard (2018) Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. Nat. Commun., 9, 2937.
5. Valentine Svensson, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic & Sarah A Teichmann (2017) Power analysis of single-cell RNA-sequencing experiments. Nat. Methods., 14, 381–387.
6. Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J. McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau et al. (2020) Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol., 38, 747–755.
7. Guiping Wang, Jeffrey R. Moffitt & Xiaowei Zhuang. (2018) Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. Sci. Rep., 8, 4847.

8. Jamie L. Marshall, Benjamin R. Doughty, Vidya Subramanian, Philine Guckelberger, Qingbo Wang, Linlin M. Chen, Samuel G. Rodriques, Kaite Zhang, Charles P. Fulco, Joseph Nasser et al. (2020) . Proc. Natl. Acad. Sci. U. S. A. HyPR-seq: Single-cell quantification of chosen RNAs via hybridization and sequencing of DNA probes., 117, (52), 33404-33413.
9. Ryan McNulty, Duluxan Sritharan, Seong Ho Pahng, Jeffrey P. Meisch, Shichen Liu, Melanie A. Brennan, Gerda Saxer, Sahand Hormoz & Adam Z. Rosenthal. (2023) Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. Nat. Microbiol., 8, 934–945.
10. Amanda Janesick, Robert Shelansky, Andrew Gottscho, Florian Wagner, Morgane Rouault, Ghezal Beliakoff, Michelli Faria de Oliveira, Andrew Kohlway, Jawad Abousoud, Carolyn Morrison (2022) High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. Biorxiv https://doi.org/10.1101/2022.10.06.510405
11. Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen et al. (2018) Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science., 360 (6385), 176-182.
12. Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen et al. (2020) Massively multiplex chemical transcriptomics at single-cell resolution. Science. 367 (6473) 45-51.
13. Joel J Credle , Christopher Y Itoh , Tiezheng Yuan , Rajni Sharma, Erick R Scott, Rachael E Workman, Yunfan Fan, Franck Housseau, Nicolas J Llosa, W Robert Bell et al. (2017) Multiplexed analysis of fixed tissue RNA using Ligation in situ Hybridization. Nucleic Acids Res., 45,(14), e128.
14. Siyu Feng, Sayaka Sekine, Veronica Pessino, Han Li, Manuel D. Leonetti & Bo Huang (2017) Improved split fluorescent proteins for endogenous protein labeling. Nat. Commun. 8, 370.

15. Max Karlsson, Cheng Zhang, Loren Méar, Wen Zhong, Andreas Digre, Borbala Katona, Evelina Sjöstedt, Lynn Butle, Jacob Odeberg, Philip Dusart. (2021). A single-cell type transcriptomics map of human tissues. Sci Adv., 7, (31).

16. Vuong Tran, Efthymia Papalexi, Sarah Schroeder, Grace Kim, Ajay Sapre, Joey Pangallo, Alex Sova, Peter Matulich, Lauren Kenyon, Zeynep Sayar. (2022) High sensitivity single cell RNA sequencing with split pool barcoding. Biorxiv. https://doi.org/10.1101/2022.08.27.505512

17. Binyamin D. Berkovits & Christine Mayr. (2015) Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. Nature., 522, 363–367.

18. Shih-Han Lee, Christine Mayr. (2019) Gain of Additional BIRC3 Protein Functions through 3′-UTR-Mediated Protein Complex Formation. Mol Cell. 74, (4), 701-712.e9.

19. Christine Mayr. (2017) Regulation by 3′-Untranslated Regions. Annu. Rev. Genet. 51, 171-194.

20. Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, Xiaowei Zhuang. (2016) High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. Proc. Natl. Acad. Sci. U. S. A., 113,(39), 11046-11051.

TABLE AND FIGURES LEGENDS

Figure 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 barcode via split and pool method. Rounds 1 and 2 barcodes are ligated, and round 3 barcodes are added via PCR. D) Sequence structure for the resulted library. E) Sequencing results of 1:1 mixed HEK293 cells stably expressing either mNG or GFP. F) 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. G) 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.

Figure 2. Quantitative accuracy of HybriSeq. A) Model of UMIs/cell with *n* probes, detection efficiency $e = 0.2$, and $1 < N < 8$ transcripts/cell B) Model of the measurement standard deviation for expression. C) Model of the measurement SNR. Blue horizontal line denotes SNR = 2. D) Experimental obtained values for average MARVELD1 UMIs/cell for different subsampled number of probes, *n*. Error bars represent the standard deviation associated with each measurement obtained by sampling unique probes used to calculate average UMIs/cell. E) Experimental obtained values for the standard deviation in expression measured with *n* probes. Green line represents the fitted model from (Sup Fig 5A) with the addition of a baseline variance to account for non-measurement associated heterogeneity. F) Experimental obtained values for the measurement SNR measured with *n* probes n. Error bars represent the standard deviation associated with each measurement obtained by sampling the unique probes used to calculate average SNR. Green line represents the fitted model from D and E. Blue horizontal line denotes SNR = 2. G-L) black: 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 blue line is the average for each probe across all cells. Green line is the location of UTRs. M) UMAP of HEK293 cell in which only EIF2S2 probes were considered colored by Leiden clustered. N) Traces of pseudo bulk clusters from M. Error bars represent the standard deviation associated with each probe in that specific cluster.

Figure 3. Analyzing cell cycle regulated transcripts using HybriSeq. A) UMAP of cell cycle transcripts in HEK293 cells measured with HybriSeq colored by Leiden cluster. B) Heat map of scaled expression values for the top 6 differentially expressed genes for each cluster. C) Heat map of scaled expression values for gene groped together by association with specific phases of the cell cycle and Leiden cluster. D) UMAP in A colored based on Phase score calculated from all genes. E) UMAP in D colored based on binned phase score into G1, S, G2-M. F) Heat map of scaled expression values for gene groped buy by association with specific phases of the cell cycle and binned phase score in E.

Figure 1

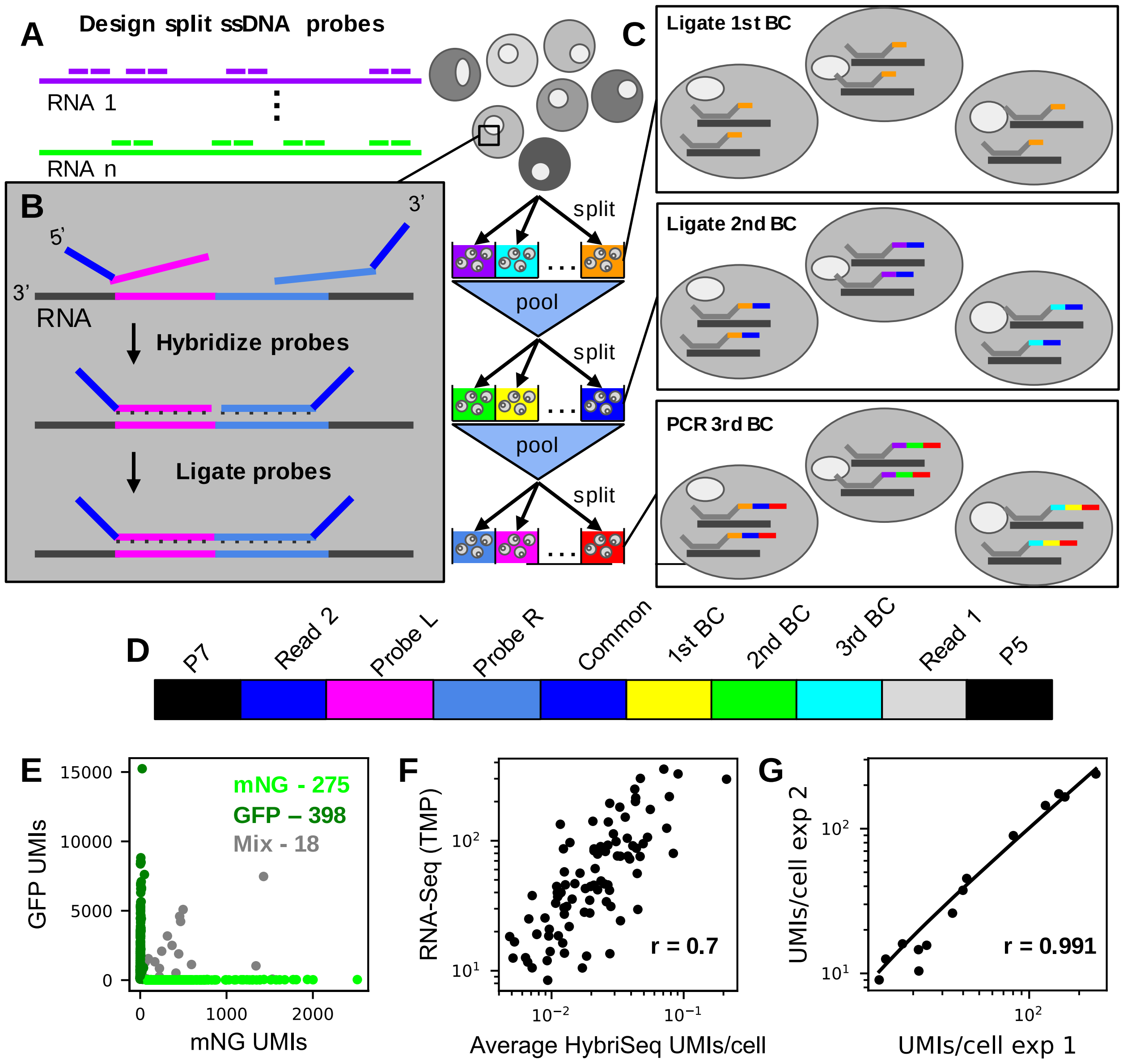

A
Design split ssDNA probes
RNA 1
RNA n
B
5'
3'
3'
RNA
Hybridize probes
Ligate probes
split
pool
split
pool
split
C
Ligate 1st BC
Ligate 2nd BC
PCR 3rd BC
D
P7
Read 2
Probe L
Probe R
Common
1st BC
2nd BC
3rd BC
Read 1
P5
E
mNG - 275
GFP – 398
Mix - 18
GFP UMIs
15000
10000
5000
0
0
1000
2000
mNG UMIs
F
RNA-Seq (TMP)
10^2
10^1
10^{-2}
10^{-1}
r = 0.7
Average HybriSeq UMIs/cell
G
UMIs/cell exp 2
10^2
10^1
10^2
r = 0.991
UMIs/cell exp 1

Figure 2

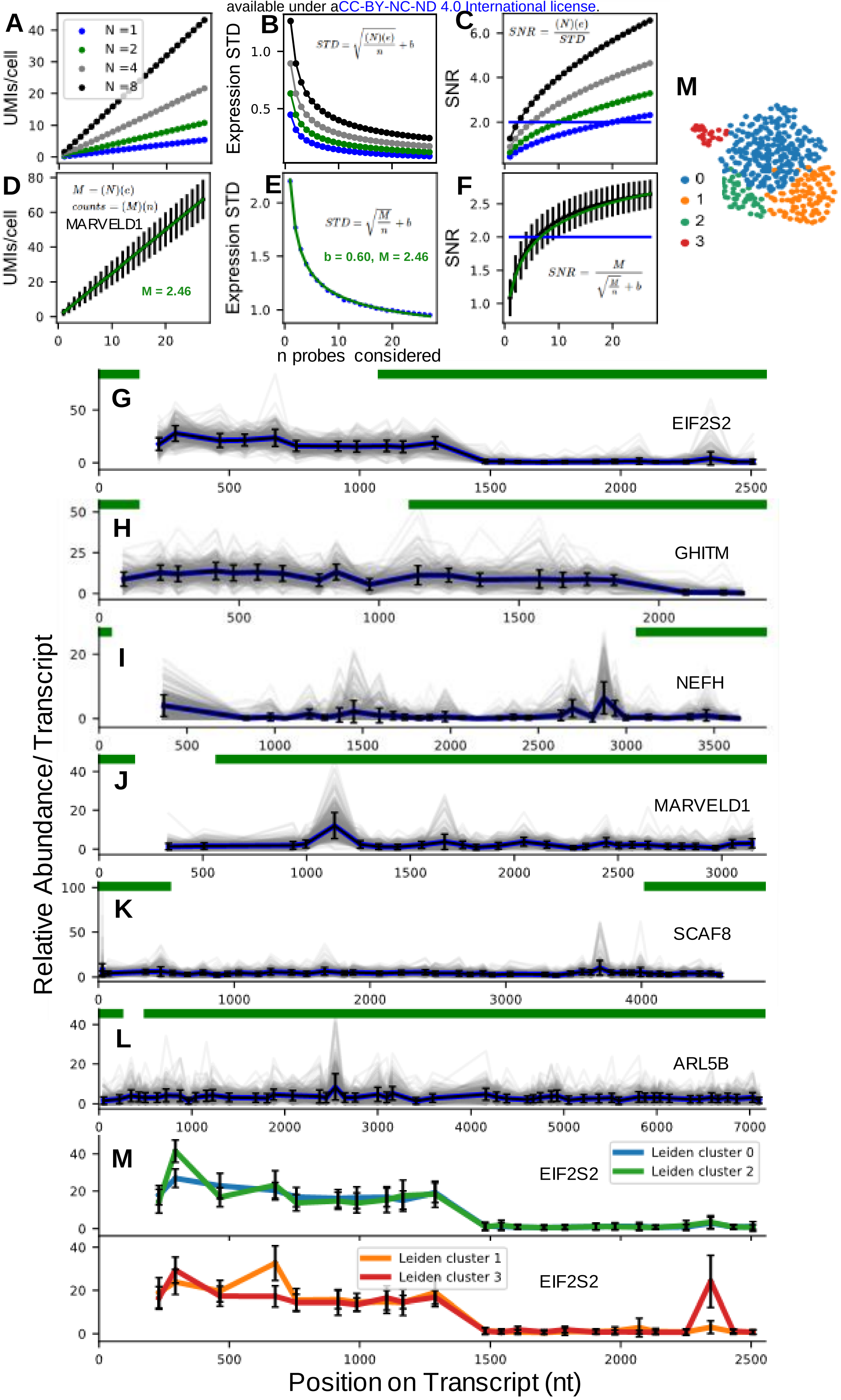
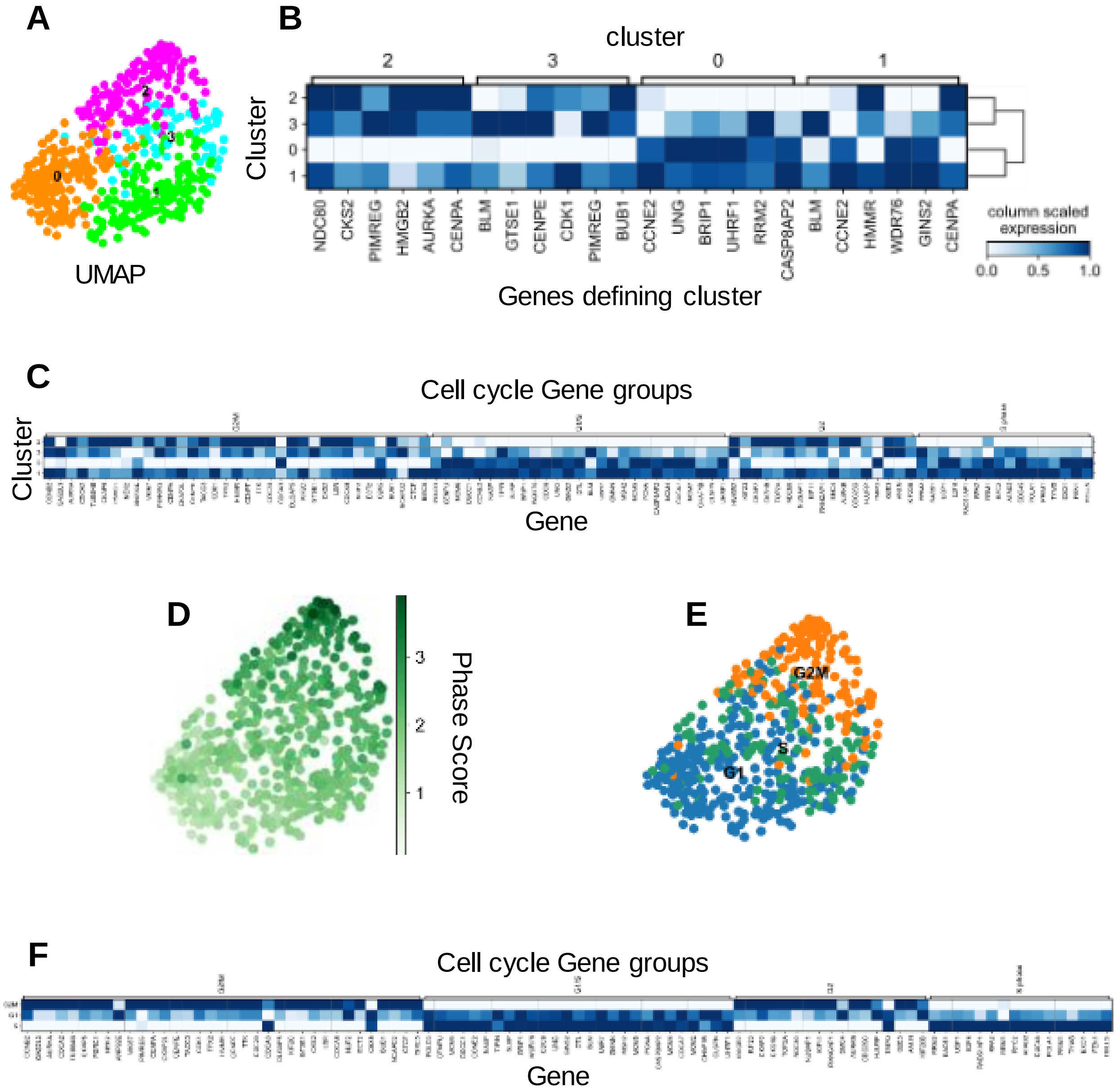

A
UMIs/cell
N =1
N =2
N =4
N =8
B
Expression STD
$STD = \sqrt{\frac{(N)(e)}{n}} + b$
C
SNR
$SNR = \frac{(N)(e)}{STD}$
D
UMIs/cell
$M = (N)(e)$
$counts = (M)(n)$
MARVELD1
M = 2.46
E
Expression STD
$STD = \sqrt{\frac{M}{n}} + b$
b = 0.60, M = 2.46
F
SNR
$SNR = \frac{M}{\sqrt{\frac{M}{n}} + b}$
n probes considered
M
0
1
2
3
G
EIF2S2
H
GHITM
I
NEFH
J
MARVELD1
K
SCAF8
L
ARL5B
M
EIF2S2
Leiden cluster 0
Leiden cluster 2
EIF2S2
Leiden cluster 1
Leiden cluster 3
Relative Abundance/ Transcript
Position on Transcript (nt)

Figure 3


A
2
3
0
1
UMAP
B
cluster
2
3
0
1
Cluster
2
3
0
1
NDC80
CKS2
PIMREG
HMGB2
AURKA
CENPA
BLM
GTSE1
CENPE
CDK1
PIMREG
BUB1
CCNE2
UNG
BRIP1
UHRF1
RRM2
CASP8AP2
BLM
CCNE2
HMMR
WDR76
GINS2
CENPA
Genes defining cluster
column scaled expression
0.0
0.5
1.0
C
Cell cycle Gene groups
Cluster
Gene
D
3
2
1
Phase Score
E
G2M
S
G1
F
Cell cycle Gene groups
Gene