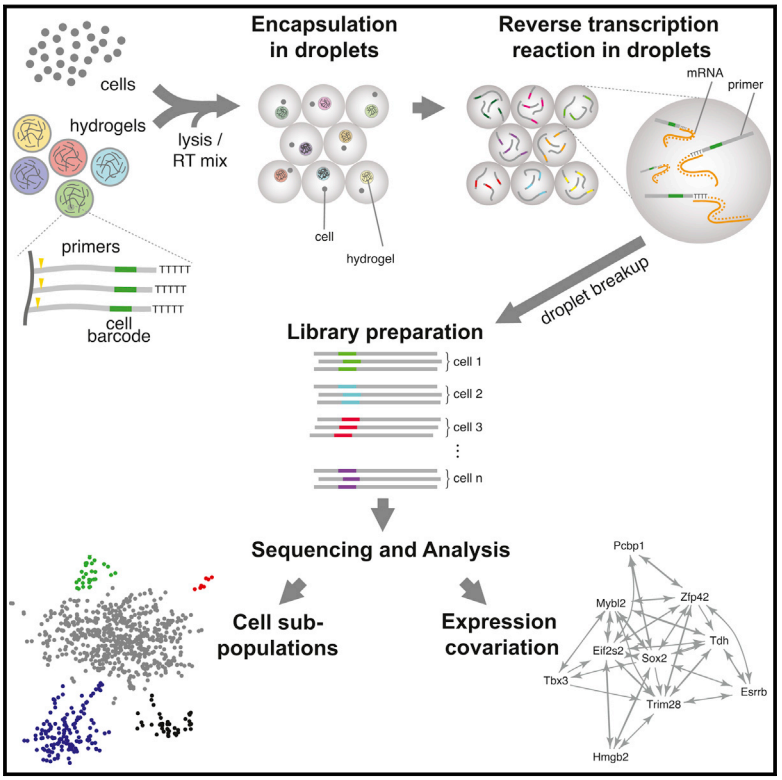
Resource

Cell

Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells

Graphical Abstract



Highlights

- Cells are captured and barcoded in nanolitre droplets with high capture efficiency
- Each drop hosts a hydrogel carrying photocleavable combinatorially barcoded primers
- mRNA of thousands of mouse embryonic stem and differentiating cells are sequenced
- Single-cell heterogeneity reveals population structure and gene regulatory linkages

Authors

Allon M. Klein, Linas Mazutis, ..., David A. Weitz, Marc W. Kirschner

Correspondence

weitz@seas.harvard.edu (D.A.W.), marc@hms.harvard.edu (M.W.K.)

In Brief

Capturing single cells along with a set of uniquely barcoded primers in tiny droplets enables single-cell transcriptomics of a large number of cells in a heterogeneous population. Applying this analysis to mouse embryonic stem cells reveals their population structure, gene expression relationships, and the heterogeneous onset of differentiation.

Accession Numbers

GSE65525

Klein et al., 2015, Cell *161*, 1187–1201
May 21, 2015
http://dx.doi.org/10.1016/j.cell.2015.04.044

CellPress

Resource

Cell

Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells

Allon M. Klein,[1,6] Linas Mazutis,[2,3,6] Ilke Akartuna,[2,6] Naren Tallapragada,[1] Adrian Veres,[1,4,5] Victor Li,[1] Leonid Peshkin,[1] David A. Weitz,[2,*] and Marc W. Kirschner[1,*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2]School of Engineering and Applied Sciences (SEAS), Harvard University, Cambridge, MA 02138, USA
[3]Vilnius University Institute of Biotechnology, Vilnius LT-02241, Lithuania
[4]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[5]Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA
[6]Co-first author
*Correspondence: weitz@seas.harvard.edu (D.A.W.), marc@hms.harvard.edu (M.W.K.)
http://dx.doi.org/10.1016/j.cell.2015.04.044

SUMMARY

It has long been the dream of biologists to map gene expression at the single-cell level. With such data one might track heterogeneous cell sub-populations, and infer regulatory relationships between genes and pathways. Recently, RNA sequencing has achieved single-cell resolution. What is limiting is an effective way to routinely isolate and process large numbers of individual cells for quantitative in-depth sequencing. We have developed a high-throughput droplet-microfluidic approach for barcoding the RNA from thousands of individual cells for subsequent analysis by next-generation sequencing. The method shows a surprisingly low noise profile and is readily adaptable to other sequencing-based assays. We analyzed mouse embryonic stem cells, revealing in detail the population structure and the heterogeneous onset of differentiation after leukemia inhibitory factor (LIF) withdrawal. The reproducibility of these high-throughput single-cell data allowed us to deconstruct cell populations and infer gene expression relationships.

INTRODUCTION

Much of the physiology of metazoans is reflected in the temporal and spatial variation of gene expression among constituent cells. Some variation is stable and has helped us to define both adult cell types and many intermediate cell types in development (Hemberger et al., 2009). Other variation results from dynamic physiological events such as the cell cycle, changes in cell microenvironment, development, aging, and infection (Loewer and Lahav, 2011). Still other expression changes appear to be stochastic in nature (Paulsson, 2005; Swain et al., 2002) and may have important consequences (Losick and Desplan, 2008). To understand gene expression in development and physiology, biologists would ideally like to map changes in RNA levels, protein levels, and post-translational modifications in every cell. Analysis at the single-cell level has until a decade ago principally been through in situ hybridization for RNA, immunostaining for proteins, or more recently with fluorescent chimeric proteins. These methods allow only a few genes to be monitored in each experiment, however. More recently, pioneering work (e.g., Chiang and Melton, 2003; Phillips and Eberwine, 1996) has made possible global transcriptional profiling at the single cell level, though the number of cells is often limited. Although an RNA inventory at the single-cell level does not offer a complete picture of the state of the cell, it can provide important insights into cellular heterogeneity and collective fluctuations in gene expression, as well as crucial information about the presence of distinct cell subpopulations in normal and diseased tissues. There is also hope that gene expression correlations within cell populations can be used to derive lineage structures (Qiu et al., 2011) and pathway structures de novo by reverse engineering (He et al., 2009).

Modern methods for RNA sequence analysis (RNA-seq) can quantify the abundance of RNA molecules in a population of cells with great sensitivity. After considerable effort, these methods have been harnessed to analyze RNA content in single cells. What is needed now are effective ways to isolate and process large numbers of individual cells for in-depth RNA sequencing and to do so with quantitative precision. This requires cell isolation under uniform conditions, preferably with minimal cell loss, especially in the case of clinical samples. The requirements for the number of cells, the depth of coverage, and the accuracy of measurements will depend on experimental considerations, including factors such as the difficulty of obtaining material, the complexity of the cell population, and the extent to which cells are diversified in gene expression space. The depth of coverage necessary is hard to predict a priori, but the existence of rare cell types in populations of interest, such as occult tumor cells or tissue stem cell sub-populations (Simons and Clevers, 2011), combined with independent drivers of heterogeneity such as cell-cycle and stochastic effects, suggests that analyzing large numbers of cells will be necessary.

The challenges of single-cell RNA-seq are easy to appreciate. Measurement accuracy is highly sensitive to the efficiency of its enzymatic steps, and the need for amplification from single cells risks introducing considerable errors. There are major obstacles to parallel processing of thousands of cells and to handling small

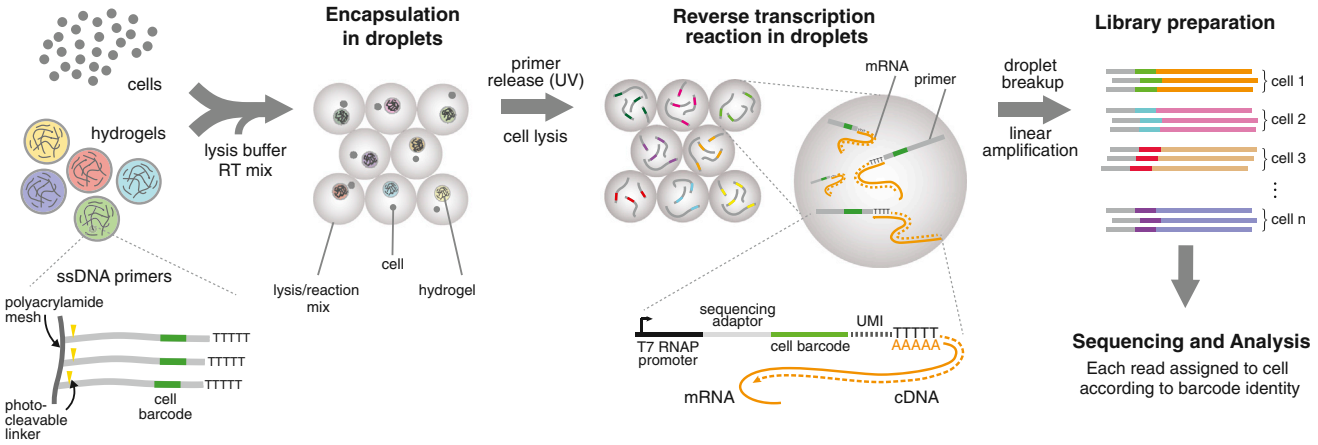


Figure 1. A Platform for DNA Barcoding Thousands of Cells

Cells are encapsulated into droplets with lysis buffer, reverse-transcription mix, and hydrogel microspheres carrying barcoded primers. After encapsulation primers are released. cDNA in each droplet is tagged with a barcode during reverse transcription. Droplets are then broken and material from all cells is linearly amplified before sequencing. UMI = unique molecular identifier.

samples of cells efficiently so that nearly every cell is measured. Microfluidics has emerged as a promising technology for single-cell studies with the potential to address these challenges (Lecault et al., 2012; Wu et al., 2014). Microfluidic chips containing hundreds of valves can trap, lyse, and assay biomolecules from single cells with higher precision and often with better efficiencies than microtiter plates (Streets et al., 2014; Wu et al., 2014). For RNA sequencing of single cells, reduced reaction volumes improve the yields of cDNA and reduce technical variability (Islam et al., 2014; Wu et al., 2014). Yet the number of single cells that can be currently processed with microfluidic chips remains at ~70–90 cells per run, so analyzing large numbers of cells is difficult, and may take so much time that the cells are no longer viable. Moreover, capture efficiency of cells into microfluidic chambers is low, a potential issue for rare or clinical samples. An alternative is the use of microfluidic droplets suspended in carrier oil (Guo et al., 2012; Teh et al., 2008). Cells can be compartmentalized into droplets and assayed for different bio-molecules (Mazutis et al., 2013), their genes amplified (Eastburn et al., 2013), and droplets sorted at high-throughput rates (Agresti et al., 2010). Unlike conventional plates or valve-based microfluidics, droplets are intrinsically scalable: the number of reaction "chambers" is not limited, and capture efficiencies are high since all cells in a sample volume can in principle be captured in droplets.

We exploited droplet microfluidics to develop a technique for indexing thousands of individual cells for RNA sequencing, which we term inDrop (*in*dexing *drop*lets) RNA sequencing. Another droplet-based RNA-seq technology is also described in this issue (Macosko et al., 2015, this issue). Our method has a theoretical capacity to barcode tens of thousands of cells in a single run. Here, we use hundreds to thousands of cells per run, since sequencing depth and cost becomes limiting for us at very high cell counts. We evaluated inDrop sequencing by profiling mouse embryonic stem (ES) cells before and after leukemia inhibitory factor (LIF) withdrawal. A total of over 10,000 barcoded cells and controls were profiled, with ~3,000 ES and differentiating cells sequenced at greater depth for subsequent analysis. Our analysis identifies rare sub-populations expressing markers of distinct lineages that would be difficult to find by profiling a few hundred cells. We show that key pluripotency factors fluctuate in a correlated manner across the entire ES cell population, and we explore whether fluctuations might associate gene products with the pluripotent state. Upon differentiation, we observe dramatic changes in the correlation structure of gene expression, resulting from asynchronous inactivation of pluripotency factors, and the emergence of novel cell states. Altogether, our results showcase the potential of droplet methods to deconstruct large populations of cells and to infer gene expression relationships within a single experiment.

RESULTS

A Microfluidic Platform for Droplet Barcoding and Analysis of Single Cells

The inDrop platform encapsulates cells into droplets with lysis buffer, reverse transcription (RT) reagents, and barcoded oligonucleotide primers (Figure 1). mRNA released from each lysed cell remains trapped in the same droplet and is barcoded during synthesis of cDNA. After barcoding, material from all cells is combined by breaking the droplets, and the cDNA library is sequenced using established methods (CEL-seq) (Hashimshony et al., 2012; Jaitin et al., 2014). The major challenge is to ensure that each droplet carries primers encoding a different barcode. We synthesized a library of barcoded hydrogel microspheres (BHMs) that are co-encapsulated with cells (Figure 2 and S1). Each BHM carries ~10^9 covalently coupled, photo-releasable primers encoding one of 147,456 barcodes, and the pool size could be increased in a straightforward manner. The current pool size allows randomly labeling 3,000 cells with 99% unique labeling (Supplemental Experimental Procedures); many more cells can be processed by splitting a large emulsion into separate tubes.

To barcode the cells, we developed a microfluidic device with four inlets for the BHMs, cells, RT/lysis reagents, and carrier oil; and one outlet port for droplet collection (Figure 3 and S2). The device generates monodisperse droplets that can be varied in the range of 1–5 nl at a rate of ~10–100 drops per second, simultaneously mixing aliquots from the inlets (Figures 3A–3C; Movies S1 and S2). The flow of deformable hydrogels inside the chip can be synchronized due to their close packing and regular release, allowing nearly 100% hydrogel droplet occupancy (Abate et al., 2009). Thus cells arriving into droplets will nearly always be co-encapsulated with barcoded primers. Due to the large cross-section of the microfluidics channel (60 × 80 μm^2), there is no cell size bias in capture. In typical conditions, cells occupy only 10% of droplets, so two-cell events are rare (Figure 3D), and cell aggregates are minimized by passing cells through a strainer or by FACS. Droplets must contain at least one cell and one gel to produce a barcoded library for sequencing; we observed that over 90% of these productive droplets contained exactly one cell and one gel (Figure 3E). After cell and BHM encapsulation, primers are photo-released by UV exposure, a step critical for efficient RT (Figures 1 and 3F).

With this system, we captured cells at a rate of 4,000–12,000/hour, or 2,000–3,000 cells barcoded for every 100 μl of emulsion (Figure 3G). As the cost of sequencing drops, higher scales may become routine.

Validation of Random Barcoding and Droplet Integrity

We tested droplet integrity by barcoding a ~50:50 mixture of mouse ES and human K562 erythroleukemia cells (Figure 4A). In this test each barcode should associate entirely with either mouse or human transcripts; only two-cell events would lead to the appearance of barcodes with mixed profiles. Figure 4A shows that indeed 96% of barcodes mapped to either the mouse or human transcriptome with more than 99% purity. This already low error rate (~4%) could be further reduced by dilution of the cell suspensions, or by sorting singlet droplets (Baret et al., 2009). However, the presence of rare two-cell events does not obscure rare cell sub-populations, since even if 10% of cells are in doublets, then 90% of rare cells will be found as singlets. This is demonstrated later for ES cells, where we found a rare cell type representing <1% of the population.

We also tested that cell barcodes were randomly sampled from the intended pool of possible barcodes. A comparison of barcode identities across a total of 11,085 control droplets consistently showed excellent agreement with random sampling (Figure S3A).

Baseline Technical Noise for inDrops

Two major sources of technical noise in single-cell RNA-seq are variability between cells in mRNA capture efficiency, and the intrinsic sampling noise resulting from capturing finite numbers of mRNA transcripts in each cell. The CEL-seq protocol has been reported to have a capture efficiency of ~3% (Grün et al., 2014) or less (Jaitin et al., 2014), and a variability in capture efficiency of ~25% for pure RNA controls and ~50% for cells (coefficients of variation between samples) when performed in microtiter plates (Grün et al., 2014). Technical noise can also arise during library amplification, but this is mostly eliminated through the use of random unique molecular identifier (UMI) sequences, allowing bioinformatic removal of duplicated reads (Fu et al., 2011; Islam et al., 2014).

An ideal test of technical noise would compare two identical cells, but unfortunately there are no cells where one can assert that the abundance of transcripts would be equal. To test technical noise in our system, we analyzed a control sample of purified total RNA diluted to single-cell concentration (10 pg per droplet), mixed with ERCC RNA spike-in controls of known concentration (Baker et al., 2005) (Figure 4B). We processed 953 droplets with an average of 30 × 10^3 (±21%) UMI-filtered mapped (UMIFM) reads per droplet (Figure 4B), and low sequencing redundancy (averaging 2.3 reads/molecule; Figure S3E). Each droplet gave 5–15 × 10^3 unique gene symbols (25,209 detected in total), correlating strongly with UMIFM counts (Figure 4C). The method showed an excellent linear readout of the ERCC spike-in input concentration (Figure 4D) down to concentrations of 0.5 molecules/droplet on average; below that limit, we tended to over-count transcripts, a bias seen previously (Grün et al., 2014; Hashimshony et al., 2012).

Another measure of method performance is its sensitivity, i.e., the likelihood of detecting an expressed gene. The sensitivity was almost entirely explained by binomial sampling statistics (Figure 4E; Supplemental Experimental Procedures), and thus depends on transcript abundance and the capture efficiency, measured from the ERCC spike-ins to be 7.1% (Figure 4D). With this efficiency, sensitivity was 50% when 10 transcripts were present, and >95% when >45 transcripts were present (Figure 4E). The sensitivity and capture efficiency are lower than those estimated for another single-cell transcriptomics protocol (~20%) (Picelli et al., 2014) but are higher than those reported for CEL-seq (3.4%) (Grün et al., 2014; Hashimshony et al., 2012). Moreover, the low sequencing redundancy suggests that deeper sequencing may further increase efficiency and thus sensitivity.

In accuracy, the method showed very low levels of technical noise, assessed by comparing the coefficient of variation (CV = SD/mean) of each gene across the cell population to its mean abundance. In a system limited only by sampling noise, all genes should obey CV = $(\text{mean})^{-1/2}$. Technical noise can lead to dispersion around this curve, and to a minimum "baseline" CV. After normalization, 99.5% of detected genes were consistent with the power law, with a baseline technical noise of <10% ($n = 25{,}209$; $p > 0.01$ χ^2 test, no multiple hypothesis correction) (Figure 4F). To our knowledge, this noise profile is among the cleanest obtained for single-cell data to date, although the sampling noise is still high (see comparisons in Figure S3H). Consistent with the low noise profile, the mean, and CV values for genes measured in cells (see below) correlated well with results measured by single-molecule fluorescent in situ hybridization (Figure S3 with data from Grün et al., 2014; Pearson correlation $R = 0.92$ for mean, and $R = 0.90$ for CV).

Noise Modeling of Single-Cell Data

Before analyzing cells, we developed a technical noise model of the effects of low sampling efficiency of transcripts and of the effects of cell-to-cell variation (noise) in efficiency. Low efficiency and noise in efficiency affect both the observed cell-to-cell variability of gene expression, and the observed covariation of gene

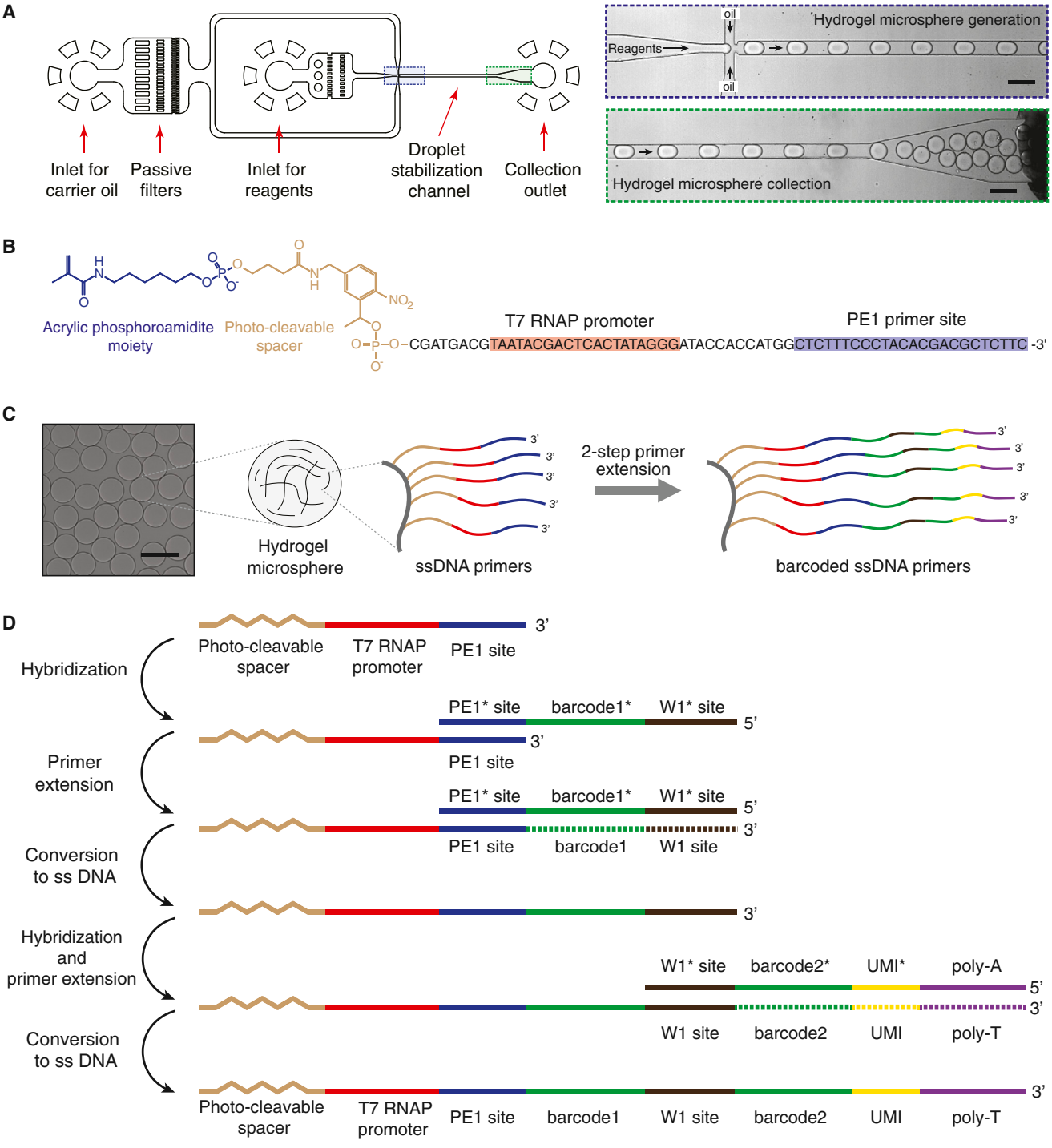


Figure 2. Barcoding Hydrogel Microsphere Synthesis

(A) Microfluidic preparation of hydrogel microspheres containing a common DNA. Scale bars 100 μm.

(B) The common DNA primer: acrylic phosphoroamidite moiety (blue), photo-cleavable spacer (green), T7 RNA polymerase promoter sequence (red), and sequencing primer (blue).

(legend continued on next page)

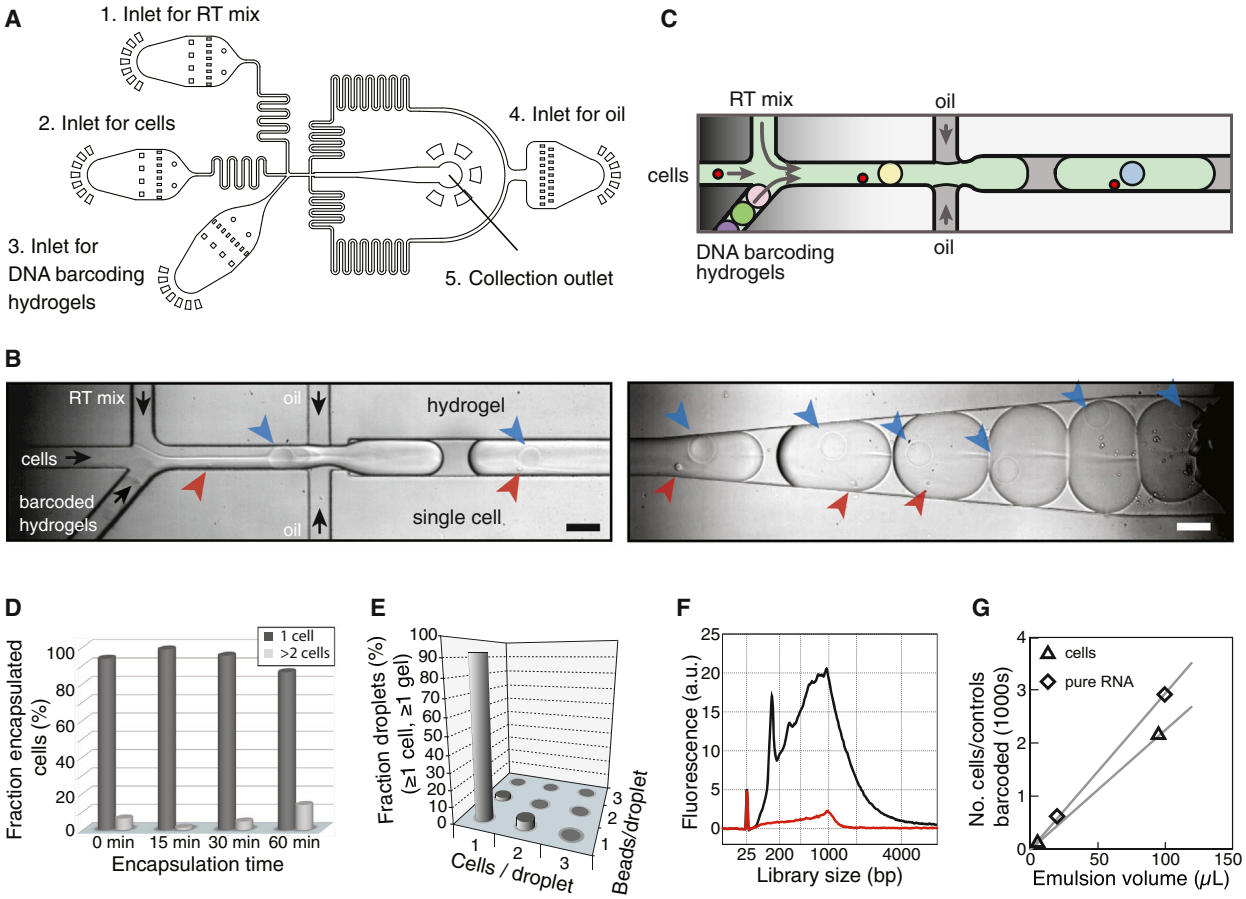


Figure 3. A Droplet Barcoding Device

(A) Microfluidic device design, see also Figure S2.

(B and C) Snapshots of encapsulation (left) and collection (right) modules, see also Movies S1 and S2. Arrows indicate cells (red), hydrogels (blue), and flow direction (black). Scale bars 100 μm.

(D) Droplet occupancy over time.

(E) Cell and hydrogel co-encapsulation statistics showing a high 1:1 cell:hydrogel correspondence.

(F) BioAnalyzer traces showing dependence of library abundance on primer photo-release.

(G) Number of cells/controls as a function of collection volume.

expression. We derived relationships between biological and observed quantities for the CVs of gene abundances across cells, gene Fano Factors (variance/mean), and pairwise correlations between genes (Figure 4G and Theory section of Supplemental Information). The Fano Factor is commonly used to measure noisy gene expression and yet is very sensitive to the efficiency β (Equation 2): even without technical noise, only genes with a Fano Factor $F \gtrsim 1/\beta$ will be noticeably variable in inDrops or other methods for single-cell analysis. The addition of technical noise introduces a "baseline" CV (Brennecke et al., 2013; Grün et al., 2014), and spuriously amplifies true biological variation (Equation 1). Low sampling efficiencies also dampen correlations between gene pairs in a predictable manner, setting an expectation to find relatively weak but nevertheless statistically significant correlations in our data (Equations 2 and 3). These results provide a basis for formally controlling for noise in single-cell measurements.

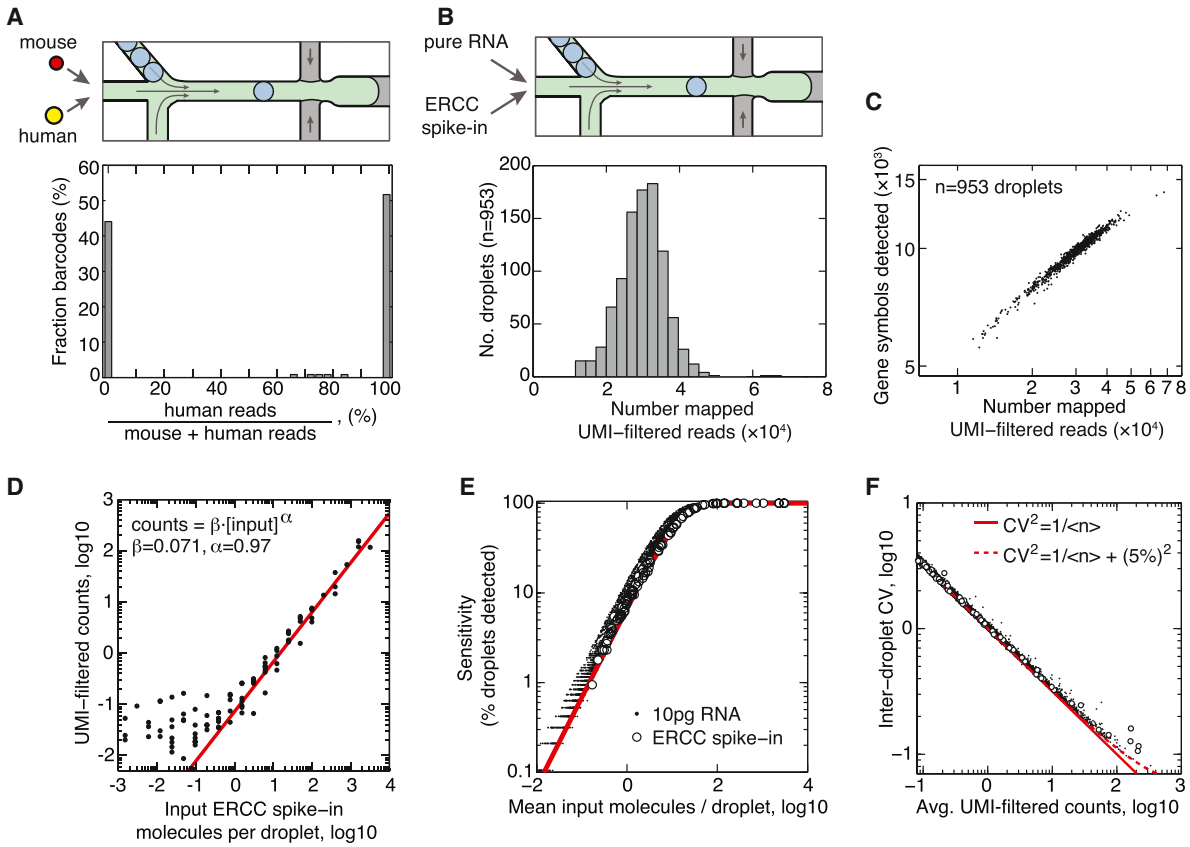
Single-Cell Profiling of Mouse ES Cells

Single-cell transcriptomics can distinguish cell types of distinct lineages even with very low sequencing depths (Pollen et al., 2014). What is less clear is the type of information that can be determined from studying a relatively uniform population subject to stochastic fluctuations. To explore this, we chose to study

(C and D) Method for combinatorial barcoding of the microspheres. * = reverse complement sequence.

(E) The fully assembled primer: T7 promoter (red), sequencing primer (blue), barcodes (green), synthesis adaptor (dark brown), UMI (yellow) and poly-T primer (purple).

See also Figure S1.



G

1) Above-Poisson noise: $CV^2 - \frac{1}{\mu} = \left(CV^2_{\text{bio}} - \frac{1}{\mu_{\text{bio}}}\right)\left(1 + CV^2_{\text{method}}\right) + CV^2_{\text{method}}$

observed | biological variability | noise inflation | background

2) Fano factor: $F = \beta F_{\text{bio}}(1 + CV^2_{\text{method}}) + (1 - \beta) + CV^2_{\text{method}}(\mu - \beta)$

sampling diminution | background

3) Gene correlations: $\underbrace{\text{corr}_{i,j}}_{\text{observed}} \simeq \underbrace{\text{corr}^{(\text{bio})}_{i,j}}_{\text{biological}} \underbrace{\sqrt{(1 - F_i^{-1})(1 - F_j^{-1})}}_{\text{sampling diminution}}$

Figure 4. Technical Noise in Droplet Barcoding

(A) Droplet integrity control: mouse and human cells are co-encapsulated to allow unambiguous identification of barcodes shared across multiple cells; 4% of barcodes share mixed mouse/human reads.

(B) inDrops technical control schematic, and histogram of UMI-filtered mapped (UMIFM) reads per droplet.

(C) Unique gene symbols detected as a function of UMIFM reads per droplet.

(D) Mean UMIFM reads for spike-in molecules are linearly related to their input concentration, with a capture efficiency β = 7.1%.

(E) Method sensitivity *S* as a function of input RNA abundance; red curve is the sensitivity limit of binomial sampling ($S = 1 - e^{-\beta x}$).

(F) CV-mean plot of pure RNA after normalization. Data points correspond to individual gene symbols; solid curve is the binomial sampling noise limit. For abundant transcripts, droplet-to-droplet variability in method efficiency β sets a baseline CV (dashed curve: $CV_\beta = 5\%$), see also Figure S3.

(G) Relationships between observed and biological values of gene CVs, Fano Factors and correlations, showing how low efficiency dampens Fano Factors (Equation 2) and weakens correlations (Equation 3).

mouse ES cells maintained in serum. These cells exhibit well-characterized fluctuations but are still uniform compared to differentiated cell types and thus pose a challenge for single cell-sequencing.

Previous studies have indicated that ES cells are heterogeneous in gene expression (Guo et al., 2010; Hayashi et al., 2008; MacArthur et al., 2012; Martinez Arias and Brickman, 2011; Ohnishi et al., 2014; Singer et al., 2014; Torres-Padilla

and Chambers, 2014; Yan et al., 2013). Other studies, which sorted ES cells into populations expressing high or low levels of the pluripotency factors *Nanog* (Chambers et al., 2007; Kalmar et al., 2009), *Rex1/Zfp42* (Singer et al., 2014; Toyooka et al., 2008), and *Stella/Dppa3* (Hayashi et al., 2008), have suggested that ES cells fluctuate infrequently between two metastable epigenetic states corresponding to a pluripotent inner cell mass (ICM)-like state, and an epiblast-like state poised to differentiate. These pluripotency factors were found to correlate with the expression of the epigenetic modifier Dnmt3b and its regulator Prdm14, and with global differences in chromatin methylation (Singer et al., 2014; Yamaji et al., 2013). Evidence suggests that other sources of heterogeneity also exist in the ES cell population: fluctuations in the Primitive Endoderm (PrEn) marker *Hex*, for example, associate with a bias toward PrEn fate upon differentiation (Canham et al., 2010); fluctuations in Hes1 bias differentiation into Epiblast sub-lineages (Kobayashi et al., 2009); and rare expression of other markers (*Zscan4, Eif1a* and others) associate with a totipotent state with access to extra-embryonic fates (Macfarlan et al., 2012). Whether these multiple fate biases result from dynamic fluctuations of transcription factors or represent stable cell states is not known.

To test inDrop sequencing, we harvested different numbers of cells at different sequencing depths for each of the ES cell runs. We collected 935 ES cells for deep sequencing and two further samples of 2,509 and 3,447 cells from a single dish as technical replicates. We further sampled 145, 302, and 2,160 cells after 2 days after LIF withdrawal; 683 cells after 4 days; and 169 and 799 cells after 7 days. The average number of reads per cell ranged up to 208×10^3 and the average UMIFM counts up to 29×10^3 (Table S1). Technical replicates showed very high reproducibility (Pearson correlation of CVs $R>0.98$, Figure 5A, inset); as did biological replicates ($R = 0.98$), whereas differentiating cells showed distinct expression profiles (Figure S4; $R = 0.94$; 732 genes differentially expressed at more than 2-fold, see Table S2). The capture efficiency β, estimated from comparing UMIFM counts to smFISH results (Figure S3), was slightly lower (4.5%) than for pure RNA.

Heterogeneous Sub-populations of ES Cell Origin

For the 935 ES cells, we identified 2,044 significantly variable genes (Table S3, Figures 5A and 5B) (10% FDR, statistical test in Supplemental Experimental Procedures) expressed at a level of at least 5 UMIFM counts in at least one cell. The set of variable genes was enriched for annotations of metabolism and transcriptional regulation, and for targets of transcription factors associated with pluripotency (Sp1, Elk1, Nrf1, Myc, Max, Tcf3, Lef1), including transcription factors that directly interact with *Pou5f1* and *Sox2* promoter regions (Gao et al., 2013) (Gabpa, Jun, Yy1, Atf3) (Table S3, $10^{-120}<p<10^{-10}$). Among the variable genes, we found pluripotency factors previously reported to fluctuate in ES cells (*Nanog*, *Rex1/Zfp42*, *Dppa5a*, *Sox2*, *Esrrb*) but, notably, the most highly variable genes included known markers of PrEn fate (*Col4a1/2*, *Lama1/b1*, *Sox17*, *Sparc*), markers of Epiblast fate (*Krt8*, *Krt18*, *S100a6*), and epigenetic regulators of the ES cell state (*Dnmt3b*). The vast majority of genes showed very low noise profiles, consistent with Poisson statistics (e.g., *Ttn*, Figure 5B). We evaluated the above-Poisson noise, defined as $\eta = CV^2\text{-}1/\mu$ (μ being the mean UMIFM count), for a select panel of genes (Figure 5C) and found it to be in qualitative agreement with previous reports (Grün et al., 2014; Singer et al., 2014). Unlike the CV or the Fano Factor, η is expected to scale linearly with its true biological value even for low sampling efficiencies (Figure 4G, Equation 1).

To test the idea that ES cells exhibit heterogeneity between a pluripotent ICM-like state and a more differentiated epiblast-like state, we contrasted the expression of candidate pluripotency and differentiation markers in single ES cells. Gene pair correlations (Figure 5D) at first appear consistent with a discrete two-state view, since both the epiblast marker *Krt8* and the PrEn marker *Col4a1* were expressed only in cells low for *Pou5f1* (shown) and other pluripotency markers (Figure S6A). Also in agreement with previous studies (Toyooka et al., 2008), the differentiation-prone state was rare. The correlations also confirmed other known regulatory interactions in ES cells, for example *Sox2*, a known negative target of BMP signaling, was anti-correlated with the BMP target *Id1*. What was more surprising was the finding that multiple pluripotency factors (*Nanog*, *Trim28*, *Esrrb*, *Sox2*, *Klf4*, *Zfp42*) fluctuated in tandem across the bulk of the cell population, but not all pluripotency factors did so (*Oct4/Pou5f1*) (Figure 5D and Figure S6). These observations are not explained by a simple two-state model (Singer et al., 2014), since pluripotency factor levels are not determined only by differentiation state. *Oct4/Pou5f1* instead correlated strongly with cyclin D3 (Figure 5D and Figure S5A), but not other cyclins, suggesting fluctuations of unknown origin.

What then is the structure of the ES cell population? We conducted a principal component analysis (PCA) of the ES cell population for the highly variable genes (Figures 5E and 5F; sensitivity analysis in Figure S5B; gene selection and normalization in Supplemental Experimental Procedures). PCA reveals multiple non-trivial dimensions of heterogeneity (12 dimensions with 95% confidence) (Figure 5E), which are not explained by independent fluctuations in each gene (Marčenko and Pastur, 1967; Plerou et al., 2002). Inspection of the first four principal components, and the principal genes contributing to these components (Figures 5F and S5), revealed the presence of at least three small but distinct cell sub-populations: one rare population (6/935 cells) expressed very low levels of pluripotency markers and high levels of PrEn markers (Niakan et al., 2010); a second cell population (15/935 cells) expressed high levels of *Krt8*, *Krt18*, *S100a6*, *Sfn* and other markers of the epiblast lineage. The third population represented a seemingly uncharacterized state, marked by expression of heat shock proteins *Hsp90*, *Hspa5*, and other ER components such as the disulphide isomerase *Pdia6*. These sub-populations expressed low levels of pluripotency factors, suggesting they are biased toward differentiation or have already exited the pluripotent state. The latter population could also reflect stressed cells.

PCA analysis is a powerful tool for visualizing cell populations that can be fractionated with just two or three principal axes of gene expression. However, when more than three non-trivial principal components exist, PCA alone is not sufficient for dimensionality reduction of high-dimensional data. Using genes identified from PCA, we used t-distributed Stochastic Neighbor Embedding (t-SNE) (Amir et al., 2013; Van der Maaten and

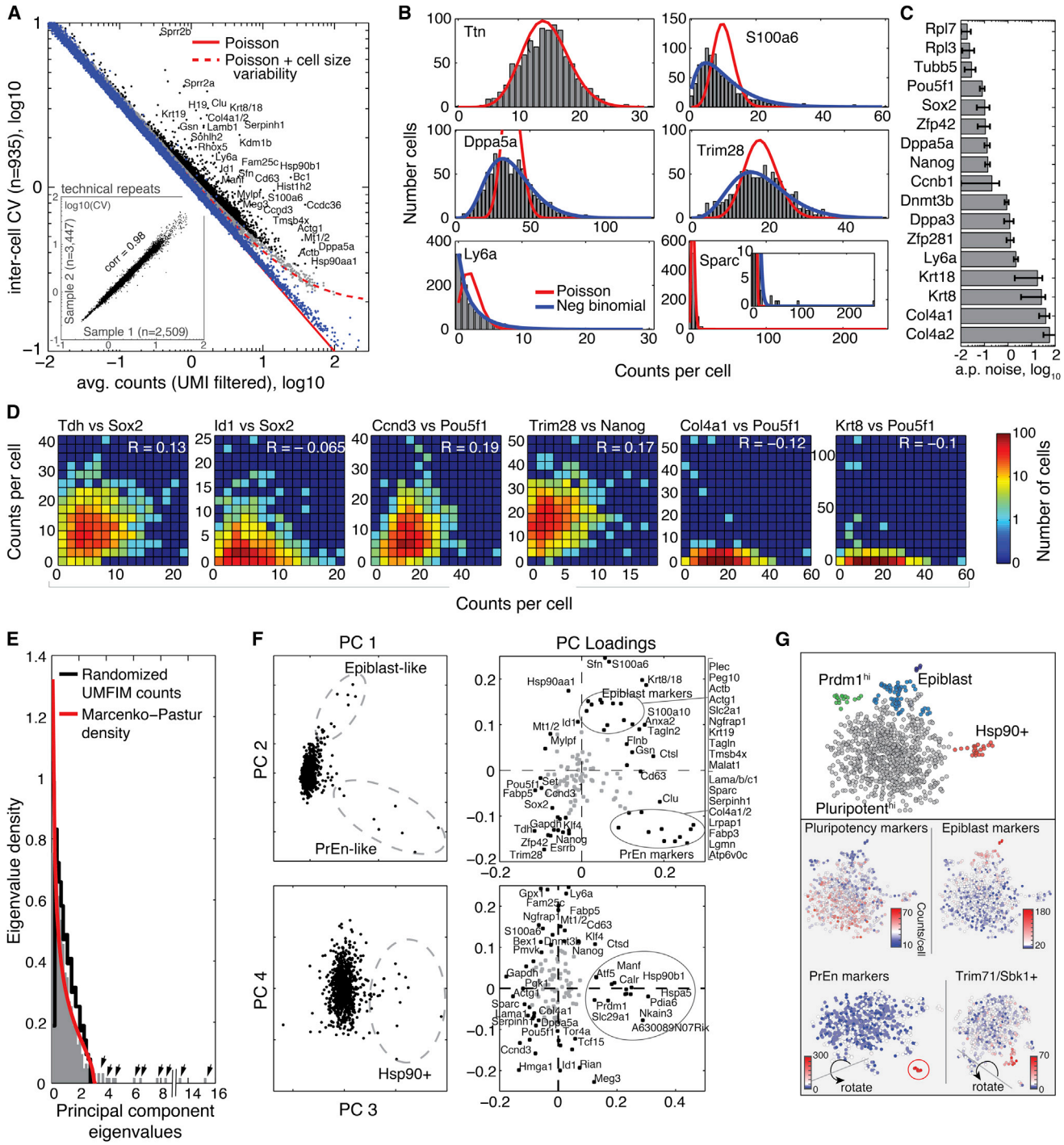


Figure 5. inDrop Sequencing Reveals ES Cell Population Structure

(A) CV-mean plot of the ES cell transcriptome. Pure RNA control (blue); genes significantly more variable than control (black). Solid and dashed curves are as in Figure 4F (variability in cell size = 20%, see Theory Equation S4 in Supplemental Information). Inset: gene CVs of two technical replicate cell populations (total n = 5,956 cells), see also Figure S4.

(B) Illustrative transcript counts showing low (*Ttn*), moderate (*Trim28, Ly6a, Dppa5a*) and high (*Sparc, S100a6*) expression variability; curve fits are Poisson (red) and Negative Binomial (blue) distributions.

(C) Above-Poisson (a.p.) noise, (CV^2-1/mean) of pluripotency differentiation markers. Error bars = SEM.

(D) Co-expression plots recapitulating known and novel gene expression relationships (see main text).

(legend continued on next page)

Hinton, 2008) to further reduce dimensionality (Figure 5G and Figures S5C–S5L) (see Supplemental Experimental Procedures). A continuum of states from high pluripotency to low pluripotency emerged, with several outlier populations at the population fringes. These included the three populations found by PCA, but also two additional fringe sub-populations characterized respectively by high expression of *Prdm1/Blimp1* and *Lin41/Trim71* (Figures S5I–S5L). The first of these expressed moderate levels of the pluripotency factors, while the second expressed low levels. Thus, while we found evidence of ES cells occupying an epiblast-like state as previously reported, and indeed found evidence for collective fluctuations between ICM to epiblast-like states (Figure 5G and Figure S5), these fluctuations do not describe the full range of heterogeneity in the ES cell population.

Functional Signatures in Gene Expression Covariation

In complex mixtures of cells, correlations of gene expression patterns could arise from differences between mature cell lineages. In a population of a single cell type such as the ES cell population studied here, however, fluctuations in cell state might reveal functional dependencies among genes.

To test whether expression covariation might contain regulatory information, we explored the covariation partners of known pluripotency factors using a topological network analysis scheme, similar to approaches developed for comparing multiple bulk samples (Li and Horvath, 2007) (Figure 6A; algorithm in Supplemental Experimental Procedures; sensitivity analysis of the method in Figure S6A). This scheme identifies the set of genes most closely correlated with a given gene (or genes) of interest, and which also most closely correlate with each other. Given the sensitivity of correlations to sampling efficiency (Figure 4G, Equation 3), we reasoned that a method based on correlation network topology would be more robust than using correlation magnitude. Remarkably, the network analysis strongly enriched for pluripotency factors: of the 20 nearest neighbors of *Nanog*, ten are documented pluripotency factors, three more are associated with pluripotency, and one (*Slc2a3*) is syntenic with *Nanog* (Scerbo et al., 2014). Only one gene (*Rbpj*) is dispensable for pluripotency (Oka et al., 1995). The analysis revealed a network of correlated pluripotency factors (Figures S6B), with multiple pluripotency factors neighboring the same previously uncharacterized genes (Supplemental Experimental Procedures and Figure S6C). It is tempting to predict that at least some of these genes are also involved in maintaining the pluripotent state. For *Sox2*, the entire neighborhood consisted of factors directly or indirectly associated with pluripotency (Figure 6C).

The same analysis may provide insight into other biological pathways, although pathways seemingly independent of ES cell biology had no meaningful topological network associations. This suggests that gene correlation networks in single-cell data capture the fluctuations most specific to the biology of the cells being studied but could be harnessed to study other pathways through weak experimental perturbations.

Cell-Cycle Transcriptional Oscillations in ES Cells Are Weak Compared to Somatic Cells

When the network analysis was applied to cyclin B, we found very few neighboring genes (Figure 6C), raising the question of why single-cell data do not reveal broader evidence of cell-cycle-dependent transcription in ES cells. Previous studies have argued for an absence of ES cell-cycle-dependent transcription (White and Dalton, 2005). Cyclins (except cyclin B) are expressed uniformly throughout the cell cycle (Faast et al., 2004; Stead et al., 2002), and the activity of the E2F family of transcription factors, which normally oscillates in somatic cells, is also constitutive in ES cells (Stead et al., 2002). ES cells have a very short cell cycle of ~8–10 hr, with ~80% of cells in S phase (White and Dalton, 2005), and almost no G1 and G2 phases, so that cell-cycle-dependent transcription could be difficult to detect.

We tested whether unperturbed ES cell data showed evidence of cell-cycle transcriptional variation. As a control, we applied inDrops to human K562 erythroleukemia lymphoblasts (n = 239 cells, average 27 × 10^3 UMIFM counts per cell), and focused on 44 transcripts previously categorized to a particular cell-cycle phase (Whitfield et al., 2002). A hierarchical clustering of these genes ordered them across the K562 cell cycle, with anti-correlations between early and late cell-cycle genes (Figure 6E). When the same analysis was repeated for the ES cell population, we found correlations between the cell-cycle genes were extremely weak and only clustered a subset of G2/M genes (Figure 6F). These results confirm that ES cells lack strong cell-cycle oscillations in mRNA abundance, but they do show evidence of limited G2/M phase-specific transcription.

Population Dynamics of Differentiating ES Cells

Upon LIF withdrawal, ES cells differentiate by a poorly characterized process, leading to the formation of predominantly epiblast lineages. In our single-cell analysis, following unguided differentiation by LIF withdrawal (Nishikawa et al., 1998), the differentiating ES cell population underwent significant changes in population structure, qualitatively seen by hierarchical clustering cells (Figure 7A). As validation, and to dissect the changes in the cell population, we first inspected selected pluripotency factors and differentiation markers (Figures 7B and 7C and Table S2). As seen in bulk assays, the average expression of *Zfp42* and *Esrrb* levels dropped rapidly; *Pou5f1* and *Sox2* dropped gradually; the epiblast marker *Krt8* increased steadily; and *Otx2*, one of the earliest transcription factors initiating differentiation from the ICM to the epiblast state, transiently increased by day 2 and then decreased (Yang et al., 2014). The average gene expression was not however representative of individual cells: some cells failed to express epiblast markers and a fraction of these expressed pluripotency factors at undifferentiated levels even 7 days after LIF withdrawal,

(E) The eigenvalue distribution of cell principal components (PC) reveals the number of non-trivial PCs detectable in the data (arrows), compared to eigenvalue distribution of randomized data (black) and to the Marcenko-Pastur distribution for a random matrix (red).

(F) The first four ES cell PCs and their coefficients, revealing three outlier populations.

(G) ES cell tSNE map revealing an axis of pluripotency-to-differentiation with fringe sub-populations at different points on the differentiation axis (see also Figure S6). Top: sub-populations visible in one projection. Bottom: cells colored by abundance of specified gene sets (see Table S4).

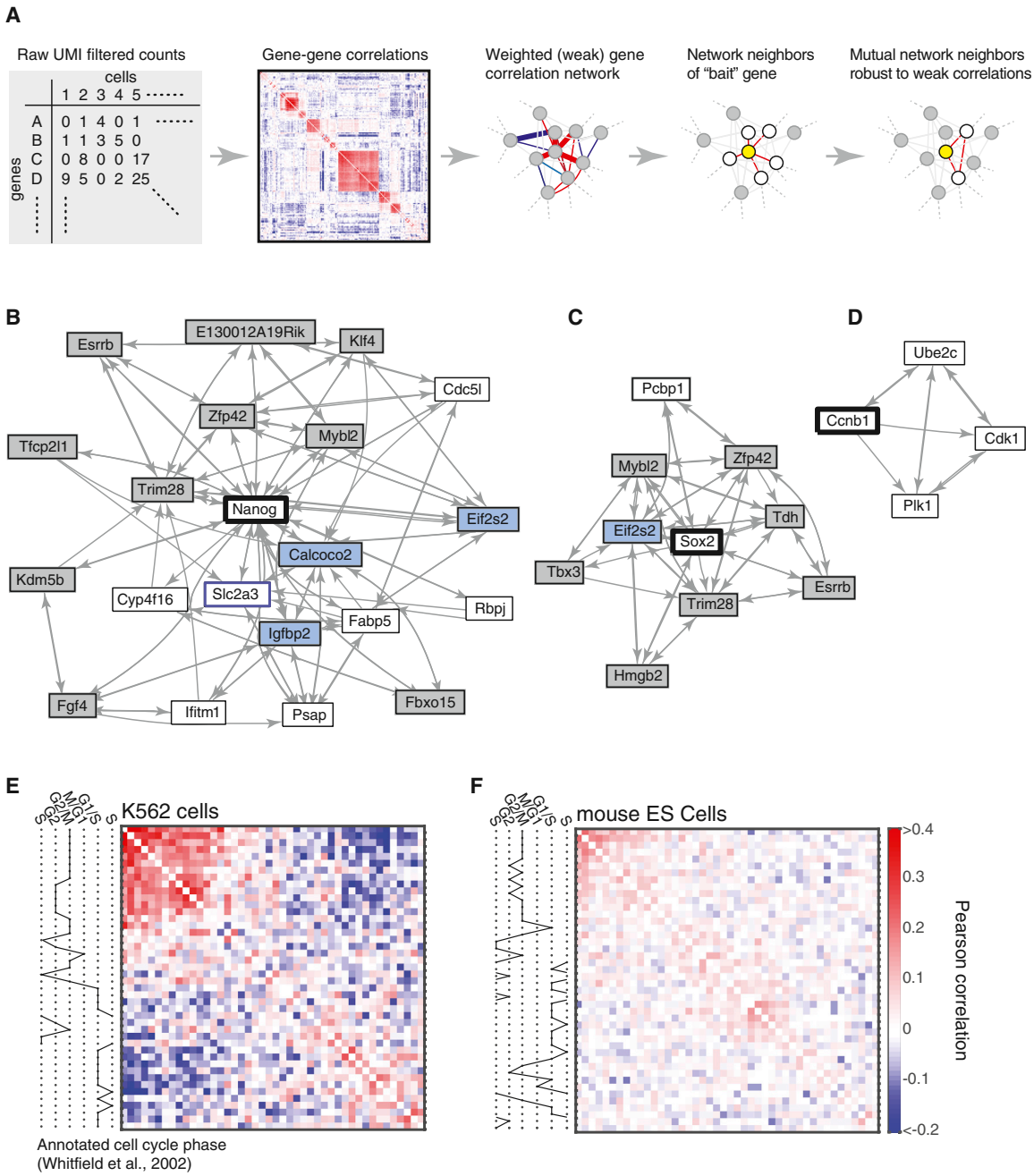


Figure 6. Regulatory Information Preserved in Gene Correlations

(A) A strategy for inferring robust gene associations from cell-to-cell variability with weak and/or highly connected gene correlations, see also Figure S6.

(B–D) Gene neighborhoods of *Nanog*, *Sox2*, and *Cyclin B*. Grey boxes mark validated pluripotency factors; blue boxes mark factors previously associated with a pluripotent state.

(E and F) Correlations of 44 cell-cycle-regulated transcripts in a somatic cell line (K562) and in mouse ES cells shows a loss of cell-cycle-dependent transcription in ES cells (gene names in Figure S6). Genes are ordered by hierarchical clustering. Color scale applies to (E and F).

(Figure 7C). This trend was supported by a PCA analysis of cells from all time points (Figure 7D; see Supplemental Experimental Procedures for gene selection and normalization), showing that after 7 days, 5% (n = 799) of cells overlapped with the ES cell population. The greatest temporal heterogeneity was evident at 4 days post-LIF, with cells spread broadly along the first principal component between the ES cell and differentiating state. The PCA analysis also revealed a metabolic signature (GO annotation: Cellular Metabolic Process, $p = 1.4 \times 10^{-8}$) consistent with the changes occurring upon differentiation (Yanes et al., 2010).

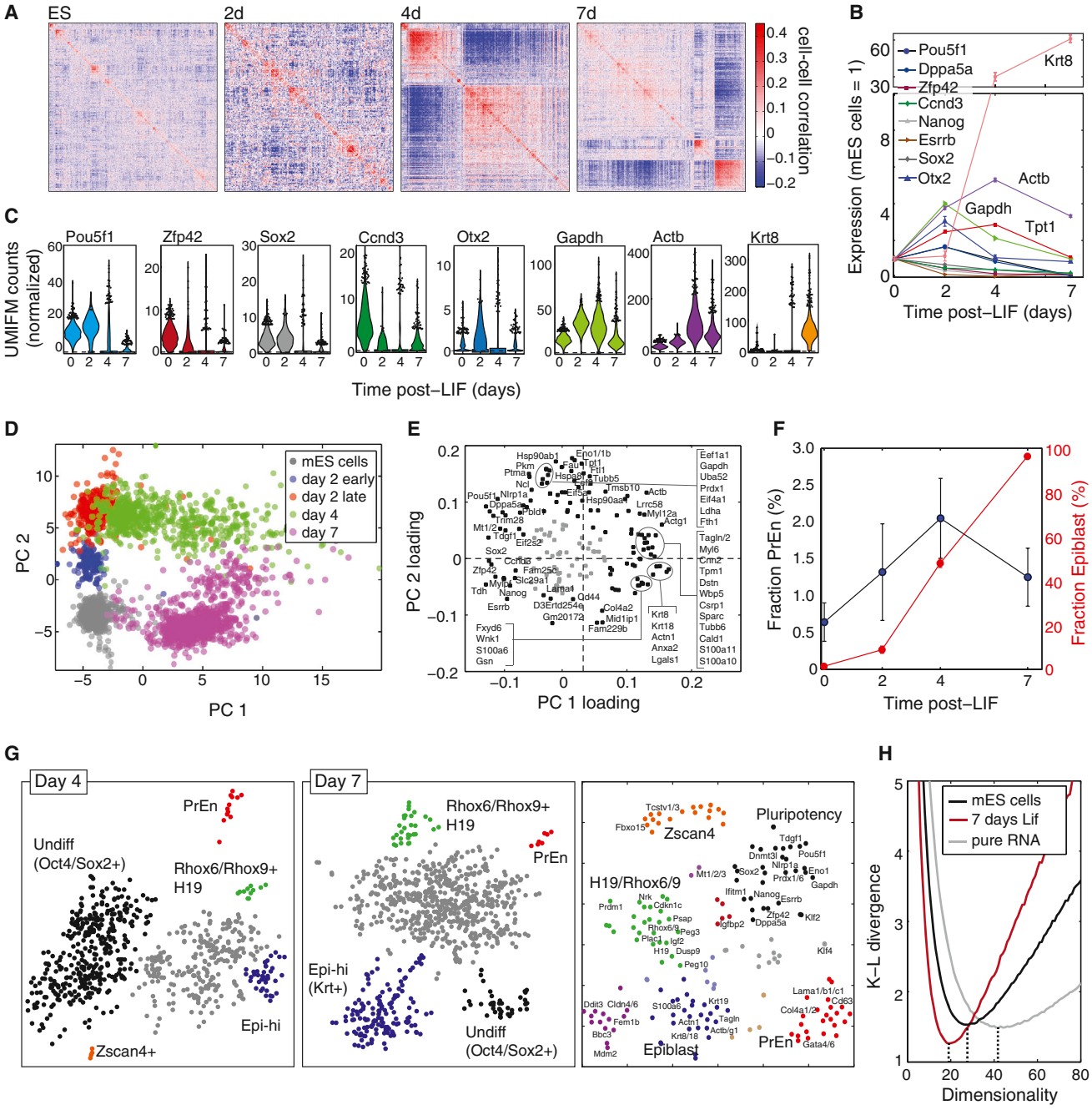


Figure 7. Heterogeneity in Differentiating ES Cells

(A) Changes in global population structure after LIF withdrawal seen by hierarchically clustering cell-cell correlations over highly variable genes.

(B and C) Average (B) and distribution (C) of gene expression after LIF withdrawal; violin plots in (C) indicate the fraction of cells expressing a given number of counts; points show top 5% of cells. Error bars = SEM.

(D and E) First two PCs of 3,034 cells showing asynchrony in differentiation.

(F) Epiblast and PrEn cell fractions as a function of time. Error bars = SEM.

(G) tSNE maps of differentiating ES cells, and of genes (right) reveal putative population markers (see also Figure S7 and Table S4).

(H) Intrinsic dimensionality of gene expression variability in ES cells and following LIF withdrawal, showing a smaller fluctuation sub-space during differentiation. The pure RNA control lacks correlations and displays a maximal fluctuation sub-space.

In addition to heterogeneity due to asynchrony, we visualized population structure by t-SNE and found distinct sub-populations, not all of which mapped to known cell types (Figure 7G; sub-population markers tabulated in Table S4). tSNE of genes over the cells revealed clusters of genes marking distinct sub-populations (Figure 7G, right and Figure S7). At 2 and 4 days post-LIF withdrawal, we identified cells expressing *Zscan4* and *Tcstv1/3*, previously identified as rare totipotent cells expressing markers of the 2-cell stage (Macfarlan et al., 2012). At 4 and 7 days, a population emerged expressing maternally imprinted genes (*H19*, *Rhox6/9*, *Peg10*, *Cdkn1*, and others), suggesting widespread DNA demethylation, possibly in early primordial germ cells. In addition, resident PrEn cells were seen at all time points (Figures 7F and 7G) but failed to expand. In sum, the analysis exposes temporal heterogeneity in differentiation and distinct ES cell fates.

Refinement of Gene Expression upon Differentiation

Our results allow testing suggestions that ES cells are characterized by promiscuous gene expression that becomes refined upon differentiation (Golan-Mashiach et al., 2005; Wardle and Smith, 2004). If so, differentiating cells should become confined to tighter domains in gene expression "space" than ES cells, as measured by the number of independent dimensions over which cells can be found. We evaluated the intrinsic dimensionality of the distribution of ES cells and differentiating cells in gene expression space using the method by (Kégl, 2002). Supporting the refinement hypothesis, we found that intrinsic dimensionality decreased after differentiation (Figure 7H). Thus, ES gene expression fluctuations are weakly coupled compared to the more coherent differences following LIF withdrawal.

DISCUSSION

We report here a platform for single-cell capture, barcoding, and transcriptome profiling, without physical limitations on the number of cells that can be processed. The method captures the majority of cells in a sample, has rapid collection times and has low technical noise. Such a method is suitable for small clinical samples including from tumors and tissue microbiopsies, and opens up the possibility of routinely identifying cell types, even if rare, based on gene expression. This type of data is also valuable for identifying putative regulatory links between genes, by exploiting natural variation between individual cells. We gave simple examples of such inference, but this type of data lends itself to more formal reverse engineering.

We have developed the droplet platform initially for whole-transcriptome RNA sequencing; however, the technology is highly flexible and should be readily adaptable to other applications requiring barcoding of RNA/DNA molecules. Our initial implementation of the method made use of a very simple droplet microfluidic chip, consisting of just a single flow-focusing junction. Future versions of the platform might take further advantage of droplet technology for multi-step reactions, or select target cells by sorting droplets on-chip (Guo et al., 2012).

The method in its current form still suffers some limitations. The major technical drawback we encountered was the mRNA capture efficiency of ~7%, which has only recently become robustly quantifiable using UMI-based filtering (Fu et al., 2011; Islam et al., 2014). Although higher than for several previously published methods, the efficiency is nonetheless too low to allow reliable detection in every cell of genes with transcript abundances lower than 20–50 transcripts. The method is therefore most reliable for profiling medium to highly abundant components of cells, missing some key transcriptional regulators, although we were able to detect almost all mouse transcription factors (1,350 out of 1,405) in a subset of cells, with the key ES cell transcription factors (*Pou5f1*, *Sox2*, *Zfp42*, and 44 other transcription factors) detected in over 90% of all cells. This is a general problem affecting single-cell RNA sequencing, which will require improved cell lytic approaches or optimized enzymatic reactions in library preparation. A second drawback of the method is the random barcoding strategy, which does not allow individual cell identities (marked by shape, size, lineage or location) to be associated with a given barcode.

Despite these limitations, the current method can provide important data addressing many biological problems. This is illustrated by the challenging problem of ES cell heterogeneity and its dynamics during early differentiation. ES cells are not divided into large sub-populations of distinct cell types, and therefore analysis of their heterogeneity requires a sensitive method. Our analysis showed that, in the presence of serum and LIF, fluctuations in *Oct4/Pou5f1* are decoupled from other pluripotency factors. We also found sub-populations of Epiblast and PrEn lineages, and other less well characterized ES cell sub-populations. This heterogeneity may reflect reversible fluctuations, or cells undergoing irreversible differentiation. The unbiased identification of small cell sub-populations requires the scale enabled by droplet methods.

EXPERIMENTAL PROCEDURES

Microfluidic Operation

The microfluidic device (80 μm deep) was manufactured by soft lithography following standard protocols (Supplemental Experimental Procedures). During operation, cells, RT/lysis mix, and collection tubes were kept on ice. Flow rates were 100 μl/hr for cell suspension, 100 μl/hr for RT/lysis mix, 10–20 μl/hr for BHMs, and 90 μl/hr for carrier oil to produce 4 nl drops. BHMs were washed 3× and concentrated by centrifugation 2× at 5krcf, then loaded directly into tubing for injection into the device. Cells were loaded at 50k–100k/ml in 16%v/v Optiprep (Sigma), and maintained in suspension using a microstir bar placed in the syringe. The carrier oil was HFE-7500 fluorinated fluid (3M) with 0.75% (w/w) EA surfactant (RAN Biotechnologies). See Supplemental Experimental Procedures for BHM synthesis, buffer compositions, equipment, and detailed microfluidic protocols.

Library Preparation

After cell encapsulation primers were released by 8 min UV exposure (365 nm at ~10 mW/cm^2, UVP B-100 lamp) while on ice. The emulsion was incubated at 50°C for 2 hr, then 15 min at 70°C, then on ice. The emulsion was split into aliquots of 100–3,500 cells and demulsified by adding 0.2X 20% (v/v) perfluorooctanol, 80% (v/v) HFE-7500 and brief centrifugation. Broken droplets were stored at −20C and processed as per CEL-SEQ protocol, see Supplemental Experimental Procedures.

Tissue Culture

IB10 ES cells are a line derived from the mouse 129/Ola strain (subcloned from E14), kindly provided by Dr. Eva Thomas. Cells were maintained on flasks pre-coated with gelatin at density ~3 × 10^5 cells/ml. ES media contained phenol red free DMEM (GIBCO), 15%v/v fetal bovine serum (GIBCO),

2 mM L-glutamine, 1 × MEM non-essential amino acids (GIBCO), 1%v/v penicillin-streptomycin antibiotics, 110 μM β-mercaptoethanol, 100 μM sodium pyruvate. ESC base media was supplemented with 1,000 U/ml LIF. See Supplemental Experimental Procedures for dissociation protocol and K562 cell culture.

Data Analysis

See Supplemental Experimental Procedures for custom bioinformatics, count normalization, method sensitivity, identification of highly variable genes, PCA and tSNE, and network neighborhood analysis.

ACCESSION NUMBERS

The accession number for the raw sequence data and processed UMIFM counts reported in this paper is Gene Expression Omnibus: GSE65525.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, Supplemental Theory, seven figures, five tables, and two movies and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.04.044.

AUTHOR CONTRIBUTIONS

A.M.K., L.M., I.A. and M.W.K. conceived the method; L.M. developed the microfluidic device; A.M.K., L.M., I.A. performed experiments; V.L. supervised ES cell culture; A.M.K. and A.V. wrote the UMI filtering pipeline; N.T. and A.M.K. developed and performed statistical noise analysis; A.M.K. and L.P. developed and performed cell population and dimensionality analysis; A.M.K., L.M. and M.W.K. wrote the manuscript. D.A.W. and M.W.K. supervised the study. All authors read and commented on the manuscript.

ACKNOWLEDGMENTS

We thank Mira Guo for guidance in hydrogel synthesis; Diego Jaitin for guidance on the CEL-SEQ/MARS-SEQ protocol; Clarissa Scholes and Angela DePace for feedback and for creating Figures 1 and 3; Rebecca Ward for help in editing. This study was supported by NIH SCAP Grant R21DK098818. A.M.K. holds a Career Award at the Scientific Interface from the Burroughs-Wellcome Fund; L.M. holds a Marie Curie International Outgoing Fellowship (300121); A.V. is supported by the HSCI Medical Scientist Training Fellowship and the Harvard Presidential Scholars Fund. L.P. is supported by NIH Grant 5R01HD073104-03. Work in the M.W.K. lab was supported by NIH (R01 GM026875, R01 GM103785, R01 HD073104), and in the D.A.W. lab by NSF (DMR-1310266), the Harvard Materials Research Science and Engineering Center (DMR-1420570), DARPA (HR0011-11-C-0093), and NIH (P01HL120839). A.M.K., L.M., I.A., D.A.W. and M.W.K. have submitted patent applications (US62/065,348, US62/066,188, US62/072,944) for the work described.

Received: November 9, 2014
Revised: February 23, 2015
Accepted: April 20, 2015
Published: May 21, 2015

REFERENCES

Abate, A.R., Chen, C.H., Agresti, J.J., and Weitz, D.A. (2009). Beating Poisson encapsulation statistics using close-packed ordering. Lab Chip *9*, 2628–2631.

Agresti, J.J., Antipov, E., Abate, A.R., Ahn, K., Rowat, A.C., Baret, J.-C., Marquez, M., Klibanov, A.M., Griffiths, A.D., and Weitz, D.A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. Proc. Natl. Acad. Sci. USA *107*, 4004–4009.

Amir, A.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545–552.

Baker, S.C., Bauer, S.R., Beyer, R.P., Brenton, J.D., Bromley, B., Burrill, J., Causton, H., Conley, M.P., Elespuru, R., Fero, M., et al.; External RNA Controls Consortium (2005). The External RNA Controls Consortium: a progress report. Nat. Methods *2*, 731–734.

Baret, J.C., Miller, O.J., Taly, V., Ryckelynck, M., El-Harrak, A., Frenz, L., Rick, C., Samuels, M.L., Hutchison, J.B., Agresti, J.J., et al. (2009). Fluorescence-activated droplet sorting (FADS): efficient microfluidic cell sorting based on enzymatic activity. Lab Chip *9*, 1850–1858.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., and Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

Canham, M.A., Sharov, A.A., Ko, M.S., and Brickman, J.M. (2010). Functional heterogeneity of embryonic stem cells revealed through translational amplification of an early endodermal transcript. PLoS Biol. *8*, e1000379.

Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., and Smith, A. (2007). Nanog safeguards pluripotency and mediates germline development. Nature *450*, 1230–1234.

Chiang, M.-K., and Melton, D.A. (2003). Single-cell transcript analysis of pancreas development. Dev. Cell *4*, 383–393.

Eastburn, D.J., Sciambi, A., and Abate, A.R. (2013). Ultrahigh-throughput Mammalian single-cell reverse-transcriptase polymerase chain reaction in microfluidic drops. Anal. Chem. *85*, 8016–8021.

Faast, R., White, J., Cartwright, P., Crocker, L., Sarcevic, B., and Dalton, S. (2004). Cdk6-cyclin D3 activity in murine ES cells is resistant to inhibition by p16(INK4a). Oncogene *23*, 491–502.

Fu, G.K., Hu, J., Wang, P.H., and Fodor, S.P. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc. Natl. Acad. Sci. USA *108*, 9026–9031.

Gao, F., Wei, Z., An, W., Wang, K., and Lu, W. (2013). The interactomes of POU5F1 and SOX2 enhancers in human embryonic stem cells. Sci Rep. *3*, 1588.

Golan-Mashiach, M., Dazard, J.E., Gerecht-Nir, S., Amariglio, N., Fisher, T., Jacob-Hirsch, J., Bielorai, B., Osenberg, S., Barad, O., Getz, G., et al. (2005). Design principle of gene expression used by human stem cells: implication for pluripotency. FASEB J. *19*, 147–149.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell *18*, 675–685.

Guo, M.T., Rotem, A., Heyman, J.A., and Weitz, D.A. (2012). Droplet microfluidics for high-throughput biological assays. Lab Chip *12*, 2146–2155.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. *2*, 666–673.

Hayashi, K., Lopes, S.M., Tang, F., and Surani, M.A. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell Stem Cell *3*, 391–401.

He, F., Balling, R., and Zeng, A.-P. (2009). Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. J. Biotechnol. *144*, 190–203.

Hemberger, M., Dean, W., and Reik, W. (2009). Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. Nat. Rev. Mol. Cell Biol. *10*, 526–537.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods *11*, 163–166.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively

parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol. *7*, e1000149.

Kégl, B. (2002). Intrinsic dimension estimation using packing numbers. Paper presented at: Advances in neural information processing systems.

Kobayashi, T., Mizuno, H., Imayoshi, I., Furusawa, C., Shirahige, K., and Kageyama, R. (2009). The cyclic gene Hes1 contributes to diverse differentiation responses of embryonic stem cells. Genes Dev. *23*, 1870–1875.

Lecault, V., White, A.K., Singhal, A., and Hansen, C.L. (2012). Microfluidic single cell analysis: from promise to practice. Curr. Opin. Chem. Biol. *16*, 381–390.

Li, A., and Horvath, S. (2007). Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics *23*, 222–231.

Loewer, A., and Lahav, G. (2011). We are all individuals: causes and consequences of non-genetic heterogeneity in mammalian cells. Curr. Opin. Genet. Dev. *21*, 753–758.

Losick, R., and Desplan, C. (2008). Stochasticity and cell fate. Science *320*, 65–68.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, this issue, 1202–1214.

MacArthur, B.D., Sevilla, A., Lenz, M., Müller, F.J., Schuldt, B.M., Schuppert, A.A., Ridden, S.J., Stumpf, P.S., Fidalgo, M., Ma'ayan, A., et al. (2012). Nanog-dependent feedback loops regulate murine embryonic stem cell heterogeneity. Nat. Cell Biol. *14*, 1139–1147.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Marčenko, V.A., and Pastur, L.A. (1967). Dros. Inf. Serv.TRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. Mathematics of the USSR-Sbornik *1*, 457–483.

Martinez Arias, A., and Brickman, J.M. (2011). Gene expression heterogeneities in embryonic stem cell populations: origin and function. Curr. Opin. Cell Biol. *23*, 650–656.

Mazutis, L., Gilbert, J., Ung, W.L., Weitz, D.A., Griffiths, A.D., and Heyman, J.A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. Nat. Protoc. *8*, 870–891.

Niakan, K.K., Ji, H., Maehr, R., Vokes, S.A., Rodolfa, K.T., Sherwood, R.I., Yamaki, M., Dimos, J.T., Chen, A.E., Melton, D.A., et al. (2010). Sox17 promotes differentiation in mouse embryonic stem cells by directly regulating extraembryonic gene expression and indirectly antagonizing self-renewal. Genes Dev. *24*, 312–326.

Nishikawa, S.I., Nishikawa, S., Hirashima, M., Matsuyoshi, N., and Kodama, H. (1998). Progressive lineage analysis by cell sorting and culture identifies FLK1+VE-cadherin+ cells at a diverging point of endothelial and hemopoietic lineages. Development *125*, 1747–1757.

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A.K., Araúzo-Bravo, M.J., Saitou, M., Hadjantonakis, A.K., and Hiiragi, T. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nat. Cell Biol. *16*, 27–37.

Oka, C., Nakano, T., Wakeham, A., de la Pompa, J.L., Mori, C., Sakai, T., Okazaki, S., Kawaichi, M., Shiota, K., Mak, T.W., and Honjo, T. (1995). Disruption of the mouse RBP-J kappa gene results in early embryonic death. Development *121*, 3291–3301.

Paulsson, J. (2005). Models of stochastic gene expression. Phys. Life Rev. *2*, 157–175.

Phillips, J., and Eberwine, J.H. (1996). Antisense RNA Amplification: A Linear Amplification Method for Analyzing the mRNA Population from Single Living Cells. Methods *10*, 283–288.

Picelli, S., Faridani, O.R., Björklund, A.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Plerou, V., Gopikrishnan, P., Rosenow, B., Amaral, L.A.N., Guhr, T., and Stanley, H.E. (2002). Random matrix approach to cross correlations in financial data. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *65*, 066126.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. *32*, 1053–1058.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Jr., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. *29*, 886–891.

Scerbo, P., Markov, G.V., Vivien, C., Kodjabachian, L., Demeneix, B., Coen, L., and Girardot, F. (2014). On the origin and evolutionary history of NANOG. PLoS ONE *9*, e85104.

Simons, B.D., and Clevers, H. (2011). Strategies for homeostatic stem cell self-renewal in adult tissues. Cell *145*, 851–862.

Singer, Z.S., Yong, J., Tischler, J., Hackett, J.A., Altinok, A., Surani, M.A., Cai, L., and Elowitz, M.B. (2014). Dynamic heterogeneity and DNA methylation in embryonic stem cells. Mol. Cell *55*, 319–331.

Stead, E., White, J., Faast, R., Conn, S., Goldstone, S., Rathjen, J., Dhingra, U., Rathjen, P., Walker, D., and Dalton, S. (2002). Pluripotent cell division cycles are driven by ectopic Cdk2, cyclin A/E and E2F activities. Oncogene *21*, 8320–8333.

Streets, A.M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., Yang, L., Fu, Y., Zhao, L., Tang, F., and Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. Proc. Natl. Acad. Sci. USA *111*, 7048–7053.

Swain, P.S., Elowitz, M.B., and Siggia, E.D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. USA *99*, 12795–12800.

Teh, S.Y., Lin, R., Hung, L.H., and Lee, A.P. (2008). Droplet microfluidics. Lab Chip *8*, 198–220.

Torres-Padilla, M.E., and Chambers, I. (2014). Transcription factor heterogeneity in pluripotent stem cells: a stochastic advantage. Development *141*, 2173–2181.

Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K., and Niwa, H. (2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. Development *135*, 909–918.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 85.

Wardle, F.C., and Smith, J.C. (2004). Refinement of gene expression patterns in the early Xenopus embryo. Development *131*, 4687–4696.

White, J., and Dalton, S. (2005). Cell cycle control of embryonic stem cells. Stem Cell Rev. *1*, 131–138.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., and Botstein, D. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell *13*, 1977–2000.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., and Quake, S.R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods *11*, 41–46.

Yamaji, M., Ueda, J., Hayashi, K., Ohta, H., Yabuta, Y., Kurimoto, K., Nakato, R., Yamada, Y., Shirahige, K., and Saitou, M. (2013). PRDM14 ensures naive pluripotency through dual regulation of signaling and epigenetic pathways in mouse embryonic stem cells. Cell Stem Cell *12*, 368–382.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. *20*, 1131–1139.

Yanes, O., Clark, J., Wong, D.M., Patti, G.J., Sánchez-Ruiz, A., Benton, H.P., Trauger, S.A., Desponts, C., Ding, S., and Siuzdak, G. (2010). Metabolic oxidation regulates embryonic stem cell differentiation. Nat. Chem. Biol. *6*, 411–417.

Yang, S.-H., Kalkan, T., Morissroe, C., Marks, H., Stunnenberg, H., Smith, A., and Sharrocks, A.D. (2014). Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency. Cell Rep. *7*, 1968–1981.

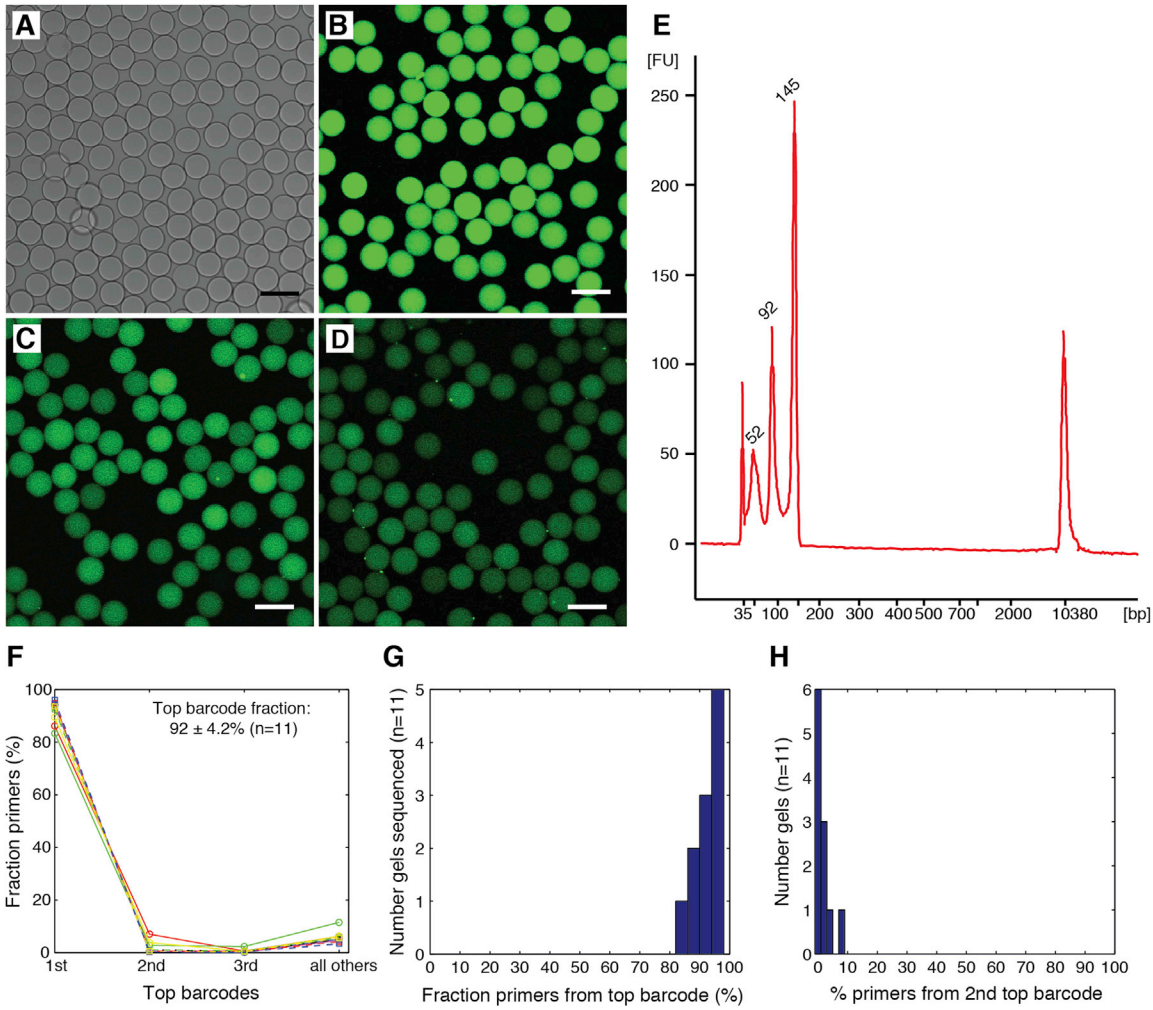


Figure S1. Quantification of DNA Primers Incorporated into Barcoded Hydrogel Microspheres, Related to Figure 2

(A–D) Imaging of BHMs post-synthesis, showing a bright field image of BHMs 63 μm in size (A), and fluorescent confocal imaging after hybridization with complimentary DNA probes targeting PE1 sequence (B), W1 sequence (C) and polyT sequence (D). Scale bars, 100 μm. E) BioAnalyzer electropherogram of DNA primers after photo-cleavage from BHMs, showing the presence of full-length barcodes (largest peaks), as well as synthesis intermediates (two smaller peaks). Peaks at 35 and 10,380 base pairs are gel migration markers. Numbers above the peaks indicate theoretical fragment size in base pairs, but these are not accurate for the single stranded DNA products. Note that fluorescence is proportional to (length × quantity), so it is not an direct measure of relative abundance between the three peaks. (F–H), Results from deep sequencing primers from 11 individual BHMs. F) Rank plot of barcode abundances on each gel; G,H) histograms of the fraction occupied on each BHM by the most-abundant and second-most abundant barcodes detailed in (G) and (H). Perfect synthesis would result in 100% occupied by the top barcode, and 0% by all other barcodes. We instead observe that an average of ~92% of all primers attached to each BHM carried the same dominant barcode.

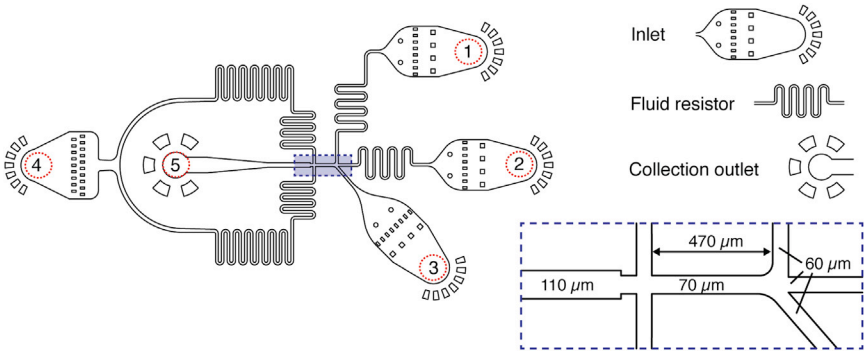
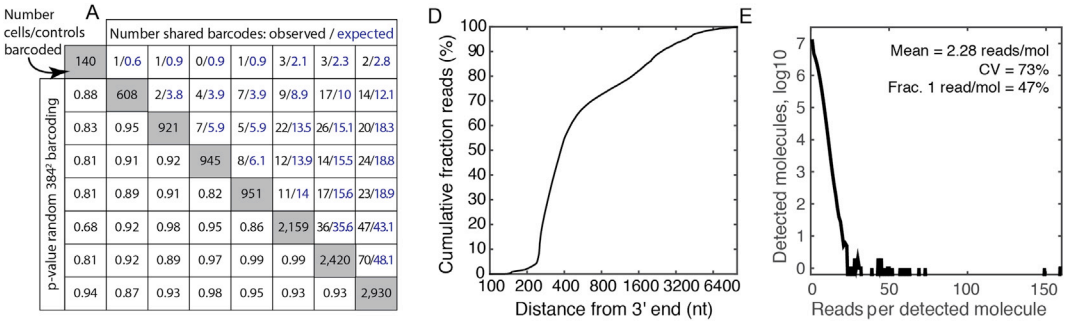


Figure S2. Dimensions of Droplet Microfluidics Device for Cell-Hydrogel-RT/Lysis Mix Co-encapsulation, Related to Figure 3

The device consists of three inlets for RT and lysis reagent mix (1), cell suspension (2), DNA barcoding beads (3) and one inlet for the continuous phase (4). The fluid resistors incorporated into device damp fluctuations arising due to mechanical instabilities of syringe pumps. The aliquot samples are brought together via 60 μm wide channels into the main 70 μm wide channel where they flow laminarly before being encapsulated into droplets at the flow-focusing junction (dashed box). Droplets are collected at the outlet (5) in form of an emulsion. See also Movies S1 and S2.



Study	Mean UMI-filtered transcripts per sample (1000s)	Baseline method noise (after normalization) [a]	Multiplicative Poisson noise [b]	Noise dispersion [c]	Nominal efficiency [f]	Sampling noise (CV), e.g. for mean = 40 transcripts [g]
Current	29	<10%[d]	<5%	0.07	7%	60%
Streets et al., PNAS 2014	N/A (no UMIs)	<25%[d]	N/A (no UMIs)	1.4	>95%	<16%
Islam et al., Nat. Methods 2013	143	60%[e]	300%	2.9	50%	22%
Grun et al., Nat. Methods 2014	77	25%[d]	<5%	0.97	3.4%	86%
Jaitin et al Science 2014	0.75	<10%[e]	<5%	0.7	<1%	>160%

Figure S3. Random Barcoding and Unique Molecular Identifier Filtering, Related to Figure 4

(A) Pair-wise tests of random barcoding for eight inDrop sequencing runs covering between 140–2,930 cells or pure RNA control droplets. Upper triangle shows the observed (black) and expected (blue) number of shared barcodes for each pair of runs with 384^2 random barcoding. Lower triangle shows p values assuming uniform random barcoding from a pool of 384^2 barcodes, which predicts that the observed number of shared barcodes should be hypergeometrically distributed about the expected value. The *p* values have not been corrected for multiple hypothesis testing. (B,C) Comparison of mean and CV counts between inDrop sequencing and single-molecule FISH (smFISH) in mouse ES cells. smFISH data from (Grün et al., 2014); original smFISH data kindly provided by Dominic Grun. (D) Cumulative distribution of mapped read distances from 3′ end of transcripts. (E-G) UMI filtering. (E) Histogram showing the number of reads per original mRNA molecule, defined by a unique cell barcode, mapped gene symbol and UMI. (F,G), Log-log plots of the inter-cell CV (SD/mean) as a function of the mean transcript abundance for genes detected in the mES cell population, without UMI filtering (F), and following UMI filtering (G). Each data point corresponds to a single gene symbol. (H) Plots and table comparing method technical performance of single-cell transcriptomics methods applied to pure RNA or to ERCC spike-ins, for several published methods. The CV versus mean plots were generated using the processed gene expression data provided by the authors of each publication, and filtered to exclude outlier cells as per the author instructions in each paper. Table footnotes: **a**: droplet-to-droplet variability in efficiency [see also CV_β in Figure 2G Equation (1)]. **b**: Multiplicative amplification of sampling (i.e., Poisson) noise for control/spike-in mRNA, defined as ε, such that control mRNA lie on the curve $CV^2 = (1+\varepsilon)/\text{mean}$. **c**: Dispersion around the sampling limit for control transcripts (indicated by the double-arrow in schematic), measured as the CV of the Fano Factors of all detected control samples. **d**: For whole transcriptome (pure RNA control samples). **e**: For ERCC spike-in mRNA only. **f**: As stated by the authors in each paper. **g**: Evaluated as $CV^2 = 1/[n\beta]$, where β is the nominal efficiency stated by the authors in each paper, and n = 40 is the test number of average transcripts.

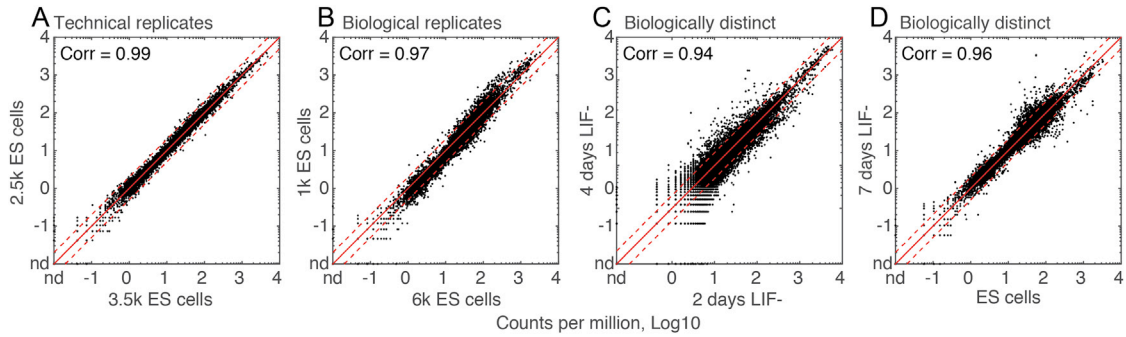


Figure S4. Technical Reproducibility of inDrop Single-Cell Sequencing, Related to Figure 5

Comparison of pooled single-cell data across technical and biological replicates, and across time points. The final UMI-filtered counts are sub-sampled to equalize the sequencing depth of each pair of samples. (A) Technical replicates correspond to two emulsion tubes collected from the same culture plate of ES cells, with 2.5k and 3.5k ES cells respectively in each sample. (B) Biological replicates, corresponding to ES cells collected on different days from different thawed aliquots of ES cells, and processed with different synthesis batches of barcoded hydrogel microspheres (BHMs). (C,D) Biologically distinct samples, comparing different time points post-LIF withdrawal. Table S2 gives a list of differentially expressed genes in the pooled data.

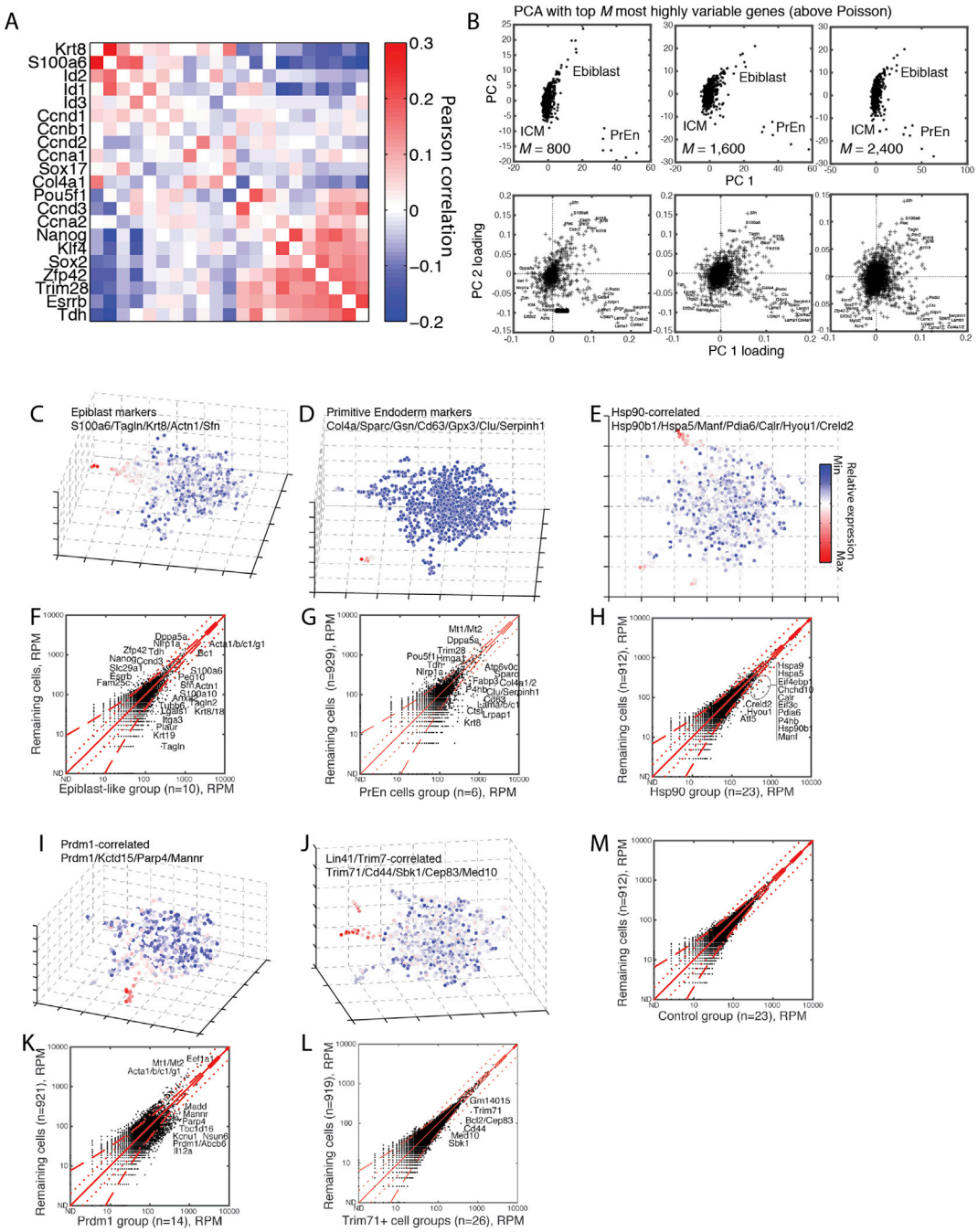
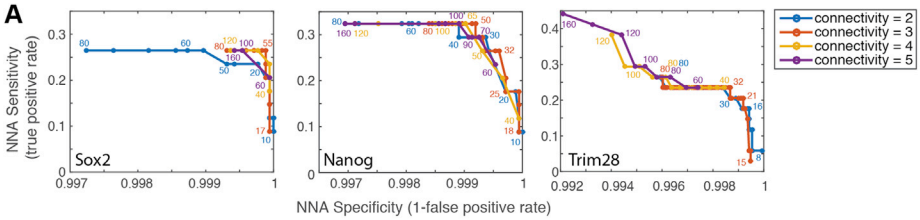
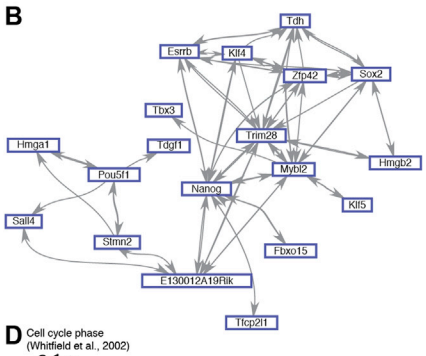


Figure S5. Structure of the mES Cell Population, Related to Figure 5

(A) Pairwise correlations of selected genes across 935 mES cells. The correlations reported here are as observed with no correction for sampling noise, and are therefore weak as expected due to the low sampling efficiency β (cf. Figure 4G, Equation (3) in the main text), and the underlying biological correlations are likely to be significantly stronger than those measured here. B) Sensitivity analysis of PCA to the number of genes selected for PCA (see Supplemental Experimental Procedures), showing the same population structure in Figure 5F is obtained using increasing numbers of variable genes. Top row shows cells projected onto the first two principal components; bottom row shows gene loadings. (C-E,I,J) Projections of a 3-dimensional tSNE map of the ES cell population reveals distinct cell sub-populations; the cells in each panel are colored according to the aggregate expression of the specified markers. (F-H,K,L) Differential gene expression plots of the pooled cells in each cell sub-population, compared to the remaining ES cells. Dotted lines indicate 2-fold differences in expression; dashed lines denote 95% confidence intervals for Poisson sampling statistics. Gene expression is normalized to UMIFM reads per million (RPM). (M) Control plot showing absence of differential gene expression for a randomly selected set of cells.





C

Gene name	No. pluri. neighbors	q value	Pluripotency gene correlation neighbors
Eif2s2	8	3.16E-21	Nanog, Sox2, Klf4, Trim28, Zfp42, Mybl2, Tdh, Hmgb2
Ptma	8	1.64E-19	Nanog, Pou5f1, Esrrb, Trim28, Zfp42, Tdh, Hmgb2, Hmga1
Ccnd3	7	3.12E-17	Pou5f1, Klf5, Mybl2, Tdh, E130012A19Rik, Hmga1, Stmn2
Asns	6	2.41E-15	Nanog, Klf4, Tfcp2l1, Trim28, Zfp42, Mybl2
Gapdh	6	3.10E-13	Pou5f1, Trim28, Mybl2, E130012A19Rik, Hmga1, Stmn2
Gm12504	5	4.89E-11	Nanog, Trim28, Zfp42, Mybl2, Tdh
H2afz	5	5.30E-11	Esrrb, Trim28, Tdh, Hmgb2, E130012A19Rik
Npm1	5	2.12E-10	Trim28, Tdh, Hmgb2, E130012A19Rik, Hmga1
Hsp90ab1	5	4.49E-10	Pou5f1, Trim28, Tdh, Hmgb2, Hmga1
Wdr5	4	1.33E-08	Klf4, Zfp42, Mybl2, Tdh
Rnf7	4	5.32E-08	Trim28, Zfp42, Mybl2, Tdh
Gm13826	4	2.68E-07	Trim28, Tdh, Hmga1, Stmn2
Pkm	4	4.60E-07	Pou5f1, Trim28, Tdh, Hmga1
Erf	3	2.22E-05	Nanog, Esrrb, Tdh
Fabp5	3	2.97E-05	Nanog, Fbxo15, Hmgb2
Pa2g4	3	7.87E-05	Trim28, Tdh, E130012A19Rik
Akr1b3	3	1.20E-04	Pou5f1, Hmga1, Stmn2
Pcbp1	2	1.15E-03	Sox2, Zfp42
Pmvk	2	2.30E-03	Klf5, Mybl2
Slc2a3	2	2.76E-03	Nanog, Tfcp2l1
Acsf2	2	4.89E-03	Sall4, E130012A19Rik
Bcas2	2	4.89E-03	Sall4, E130012A19Rik
Tyw1	2	6.04E-03	Sall4, Stmn2
Pla2g2d	2	6.90E-03	Klf4, Esrrb
Ubald2	2	1.15E-02	Klf4, Mybl2
1700001K19Rik	2	1.71E-02	Stat3, Fbxo15
Nalcn	2	1.71E-02	Stat3, Fbxo15
Scarna2	2	1.71E-02	Stat3, Fbxo15
Igfbp2	2	1.84E-02	Nanog, Mybl2
Snrpn	2	1.95E-02	Mybl2, E130012A19Rik
Snurf	2	1.95E-02	Mybl2, E130012A19Rik
Fbl	2	2.41E-02	Esrrb, Tdh
Hmces	2	2.41E-02	Esrrb, Tdh
Atp5b	2	2.76E-02	Esrrb, Trim28
Ifitm2	2	2.76E-02	Esrrb, Trim28
Mcm3	2	3.42E-02	Tdh, E130012A19Rik
Aamp	2	3.91E-02	Trim28, E130012A19Rik
Marcksl1	2	3.91E-02	Trim28, Hmgb2
Oaz1	2	3.91E-02	Trim28, Hmgb2
Dppa5a	2	4.35E-02	Pou5f1, Stmn2
Gm6083	2	4.35E-02	Pou5f1, Stmn2
Gabarapl2	2	4.60E-02	Trim28, Mybl2
Tmsb10	2	4.83E-02	Trim28, Stmn2

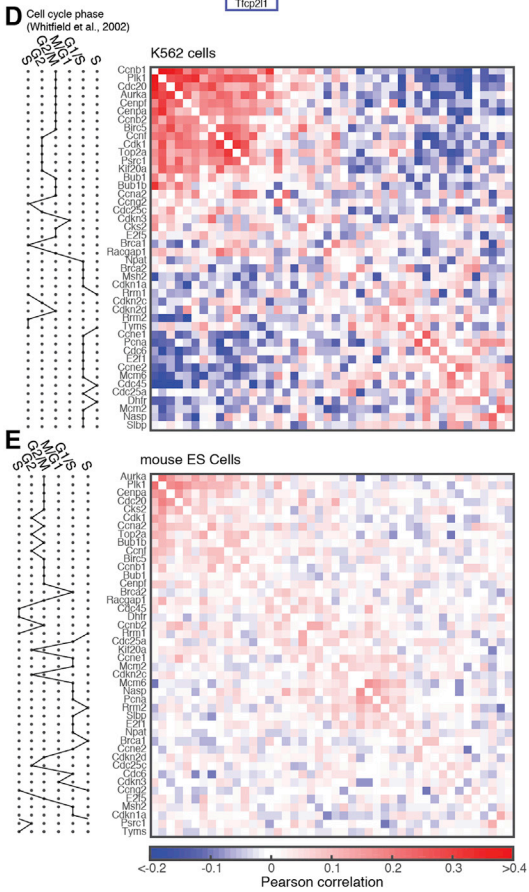


Figure S6. Pluripotency Network Neighborhood Analysis (NNA) and Comparison of Transcriptional Signatures of Somatic and ES Cell Cycles, Related to Figure 6

(A–C) NNA Analysis (see Supplemental Experimental Procedures). A) ROC curves (Sensitivity versus Specificity) for the NNA of Nanog, Sox2 and Trim28 with respect to variations in the NNA parameters *N* and *X*. B) Mutual NNA neighbors among a curated list of established pluripotency genes. C) List of 43 genes that are not established pluripotency factors but are found to be NNA neighbors of at least two pluripotency genes, with re-occurrence q-value < 0.05 corrected for multiple hypothesis testing (see Supplemental Experimental Procedures). D-E) Enlarged versions of Figures 6E and 6F with gene names.

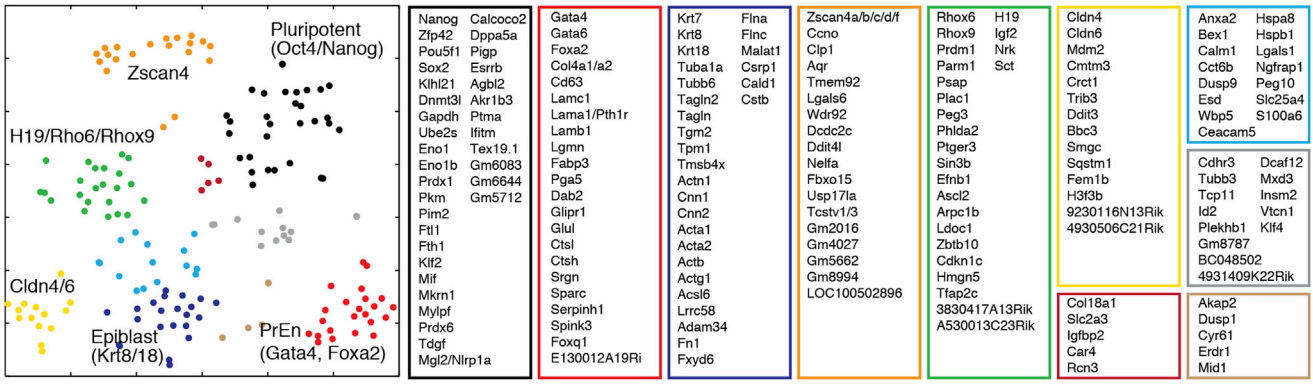


Figure S7. tSNE Map of Principal Genes at 4 Days Post-LIF Withdrawal, Related to Figure 7

This figure reproduces Figure 5G of the main text with full gene annotation.

Cell

Supplemental Information

Droplet Barcoding for Single Cell Transcriptomics Applied to Embryonic Stem Cells

Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner

Table of Contents

A. Supplementary Figure Legends

Figure S1. Quantification of DNA Primers Incorporated into Barcoded Hydrogel microspHeres (BHMs), related to Figure 2

A-D) Imaging of BHMs post-synthesis, showing a bright field image of BHMs 63 µm in size (A), and fluorescent confocal imaging after hybridization with complimentary DNA probes targeting PE1 sequence (B), W1 sequence (C) and polyT sequence (D). Scale bars, 100 µm. E) BioAnalyzer electropherogram of DNA primers after photo-cleavage from BHMs, showing the presence of full-length barcodes (largest peaks), as well as synthesis intermediates (two smaller peaks). Peaks at 35 and 10380 base pairs are gel migration markers. Numbers above the peaks indicate theoretical fragment size in base pairs, but these are not accurate for the single stranded DNA products. Note that fluorescence is proportional to (length x quantity), so it is not an direct measure of relative abundance between the three peaks. (F-H), Results from deep sequencing primers from 11 individual BHMs. F) Rank plot of barcode abundances on each gel; G,H) histograms of the fraction occupied on each BHM by the most-abundant and second-most abundant barcodes detailed in (G) and (H). Perfect synthesis would result in 100% occupied by the top barcode, and 0% by all other barcodes. We instead observe that an average of ~92% of all primers attached to each BHM carried the same dominant barcode.

Figure S2. Dimensions of Droplet Microfluidics Device for Cell-Hydrogel-RT/Lysis Mix co-encapsulation, related to Figure 3

The device consists of three inlets for RT and lysis reagent mix (1), cell suspension (2), DNA barcoding beads (3) and one inlet for the continuous phase (4). The fluid resistors incorporated into device damp fluctuations arising due to mechanical instabilities of syringe pumps. The aliquot samples are brought together via 60 µm wide channels into the main 70µm wide channel where they flow laminarly before being encapsulated into droplets at the flow-focusing junction (dashed box). Droplets are collected at the outlet (5) in form of an emulsion. See also **Supplementary Movies S1 and S2**.

Figure S3. Random Barcoding and Unique Molecular Identifier (UMIs) Filtering, related to Figure 4 A) Pair-wise tests of random barcoding for eight inDrop sequencing runs covering between 140-2,930 cells or pure RNA control droplets. Upper triangle shows the observed (black) and expected (blue) number of shared barcodes for each pair of runs with 384^2 random barcoding. Lower triangle shows *p*-values assuming uniform random barcoding from a pool of 384^2 barcodes, which predicts that the observed number of shared barcodes should be hypergeometrically distributed about the expected value. The *p* values have not been corrected for multiple hypothesis testing. B,C) Comparison of mean and CV counts between inDrop sequencing and single molecule FISH (smFISH) in mouse ES cells. smFISH data from (Grun et al., 2014); original smFISH data kindly provided by Dominic Grun. D) Cumulative distribution of mapped read distances from 3' end of transcripts. E-G) UMI filtering. E) Histogram showing the number of reads per original mRNA molecule, defined by a unique cell barcode, mapped gene symbol and UMI. F,G), Log-log plots of the inter-cell CV (standard deviation/mean) as a function of the mean transcript abundance for genes detected in the mES cell population, without UMI filtering (F), and following UMI filtering (G). Each data point corresponds to a single gene symbol. H) Plots and table comparing method technical performance of single cell transcriptomics methods applied to pure RNA or to ERCC spike-ins, for several published methods. The CV vs mean plots were generated using the processed gene expression data provided by the authors of each publication, and filtered to exclude outlier cells as per the author instructions in each paper. Table footnotes:

a: droplet-to-droplet variability in efficiency [see also CV_β in Fig. 2G Eq. (1)].

b: Multiplicative amplification of sampling (i.e. Poisson) noise for control/spike-in mRNA, defined as ε, such that control mRNA lie on the curve $CV^2=(1+\varepsilon)/\text{mean}$.

c: Dispersion around the sampling limit for control transcripts (indicated by the double-arrow in schematic), measured as the CV of the Fano Factors of all detected control samples.
d: For whole transcriptome (pure RNA control samples).
e: For ERCC spike-in mRNA only.
f: As stated by the authors in each paper.
g: Evaluated as $CV^2=1/[n\beta]$, where β is the nominal efficiency stated by the authors in each paper, and $n=40$ is the test number of average transcripts.

Figure S4. Technical Reproducibility of inDrop Single Cell Sequencing, related to Figure 5

Comparison of pooled single cell data across technical and biological replicates, and across time points. The final UMI-filtered counts are sub-sampled to equalize the sequencing depth of each pair of samples. A) Technical replicates correspond to two emulsion tubes collected from the same culture plate of ES cells, with 2.5k and 3.5k ES cells respectively in each sample. B) Biological replicates, corresponding to ES cells collected on different days from different thawed aliquots of ES cells, and processed with different synthesis batches of barcoded hydrogel microspheres (BHMs). C,D) Biologically distinct samples, comparing different time points post-LIF withdrawal. **Table S2** gives a list of differentially expressed genes in the pooled data.

Figure S5. Structure of the mES Cell Population, related to Figure 5

A) Pairwise correlations of selected genes across 935 mES cells. The correlations reported here are as observed with no correction for sampling noise, and are therefore weak as expected due to the low sampling efficiency β (cf. **Fig. 4G**, Eq. (3) in the main text), and the underlying biological correlations are likely to be significantly stronger than those measured here. B) Sensitivity analysis of PCA to the number of genes selected for PCA (see **Extended Methods**), showing the same population structure in **Fig. 5F** is obtained using increasing numbers of variable genes. Top row shows cells projected onto the first two principal components; bottom row shows gene loadings. C-E,I,J) Projections of a 3-dimensional tSNE map of the ES cell population reveals distinct cell sub-populations; the cells in each panel are colored according to the aggregate expression of the specified markers. F-H,K,L) Differential gene expression plots of the pooled cells in each cell sub-population, compared to the remaining ES cells. Dotted lines indicate two-fold differences in expression; dashed lines denote 95% confidence intervals for Poisson sampling statistics. Gene expression is normalized to UMIFM reads per million (RPM). M) Control plot showing absence of differential gene expression for a randomly selected set of cells.

Figure S6. Pluripotency Network Neighborhood Analysis (NNA) and Comparison of Transcriptional Signatures of Somatic and ES Cell Cycles, related to Figure 6

A-C) NNA Analysis (see Extended Methods). A) ROC curves (Sensitivity vs Specificity) for the NNA of Nanog, Sox2 and Trim28 with respect to variations in the NNA parameters N and X. B) Mutual NNA neighbors among a curated list of established pluripotency genes. C) List of 43 genes that are not established pluripotency factors but are found to be NNA neighbors of at least two pluripotency genes, with re-occurrence q-value < 0.05 corrected for multiple hypothesis testing (see Extended Methods). D-E) Enlarged versions of **Fig. 6E-F** with gene names.

Figure S7. tSNE map of principal genes at 4 days post-LIF withdrawal, related to Figure 7

This figure reproduces Fig. 5G of the main text with full gene annotation.

B. Supplemental Movie Legends

Movie S1. Slow Motion Movie of Droplet Encapsulation Nozzle Combining Cells, Barcoded Hydrogel Microspheres and RT Lysis Buffer, related to Figure 3

Time elapsed indicated on the bottom. Scale as per **Fig. S2.**

Movie S2. Droplet Collection Channel Downstream of Encapsulation, related to Figure 3

Time elapsed indicated on the bottom.

C. Supplemental tables and table legends

Table S1. Sequencing Run Statistics, related to Figures 4-7

Sample (days post-LIF)	Emulsion volume (µL)	Platform	Total library reads (unfiltered)	Number cells / barcodes	Average filtered reads/cell	UMI Filtered Mapped (UMIFM) counts/cell	
						Average	Coeff. of variation (CV)
Pure RNA control	16	HiSeq 2500	166,031,332	953	89,116	32,490	21%
mES LIF+	40	NextSeq	413,138,104	935	199,193	29,511	36%
mES day 2, (early)	6	HiSeq 2500	119,859,024	145	119,386	20,609	35%
mES day 2, (late)	26	MiSeq	17,660,550	303	38,788	8,441	36%
mES day 4	40	MiSeq	11,557,428	683	10,237	4,661	43%
mES day 7	8	HiSeq 2500	92,805,168	169	153,035	27,065	38%
mES day 7	40	NextSeq	250,187,951	799	208,231	26,216	55%
mES day 2, early	95	HiSeq 2500	33,751,186	2,168	4,987	2,608	42%

Table S4. Cell Sub-populations of ES cells, related to Figures 5 and 7

The "Clustering Index" is the mean Silhoutte score (1) of all cells in each group evaluated according to a correlation distance metric (see Methods). Index values are bounded between -1 (unclustered) to 1 (highly isolated cluster). An index value near 0 typically describes a cohesive cluster that is not well separated from others.

Data set	**Cell group description**	**Cell group size (number / total)**	**Clustering index (-1<x<1)**	**Top high-expressed genes (compared to population average)**
mES day 0	Primitive endoderm-like	6/935	0.76	*Gsn, Col4a1/2, Serpinh1, Lama1/b1/c1, Sparc, Srgn, P4ha2, Lrpap1, Podxl, Ctsl, S100a10, Pgk1, Slc2a3, Tfpi, Amn, Fbp2, Gpx3, Man2c1os, Lpar3, Cd63*
	Epiblast-like	40/935 (all) 6/935 (*Krt8*-high)	0.20 (all) 0.55 (*Krt8*-high)	*Actg1, Anxa2, Krt8/18/19, Plaur, Cnn1, Tagln, Plin2, Flnc, Tinagl1, Slc2a1, Fam160b2, Mmab, Sfn, Plec, S100a6, Flnb, Ngfrap1*
	Hsp90-high	10/935	0.47	*Atf5, Calr, Hsp90b1, Hspa5, Manf, Pdia6, Creld2, Hyou1, Derl3, Prph, Chchd10*
	Prdm1-high	13/935	0.47	*Prdm1, Baat, Nsun6, Parp4, Srgn, Ssh1*
	Trim71-high	12/935	0.31	*Trim71, Cd44, Med10, Myo15, Bcl2/Cep83, Kdm1b, Sbk1, Csf1r Psg18/20, Prss44*
4,7 days post-LIF withdrawal	*H19/Rhox6/9*+	14/683, 10/899	0.36, 0.20	*H19, Igf2, Rhox6/9, Fabp3, Igfbp2, Sct, Vgf, Pmp22, Rhox5, Itm2a, Rhox5, 1700001F09Rik, Peg10*
	Pluripotency-high	21/683, 31/899	0.16, 0.15	*Trim28, Tex19.1, Tdh, Tdgf1, Spry4, Sox2, Psors1c2, Pou5f1, Phc1, Ogfod3, Mylpf, Mt1/2, Mkrn1, Mkm1, L1td1, Kcnj14, Gad1, G3bp2, Dnmt3l, Cdh16, Nlrp1a, 4930526L06Rik, 3110021A11Rik*
	Zscan4-hi	4/683, 0/899	0.45, N/A	*Zscan4a/c/d/f, Fbxo15, Tcstv1/3, Dazl, Calcoco2, Mylpf, Dcdc2c, Lmx1a, Ddit4l, Aqr, Clp1, Tmem92, Usp17la, 2310039L15Rik, B020031M17Rik, Gm4027, Gm20767, Gm7102, Gm8994*
	Primitive endoderm-like	7/683, 4/899	0.55, 0.65	*Gata6, Amn, Cd63, Ctsl, Col4a1/2, Lama1/b1/c1, Upp1, Sparc, P4ha2, Serpinh1, Fst, Lrpap1, P4hb, Ctsh, Clu, Epas1, Pga5*

Table S5. List of DNA Oligonucleotides, related to Experimental Procedures

1. BHM synthesis:	
Hydrogel-incorporated DNA primer	5'-/5Acryd/iSpPC/CGATGACG TAATACGACTCACTATAGGG ATACCACCATGG CTCTTTCCCTACACGACGCTCTTC-3'
barcode 1 (W1*-bc1-PE1*)	5'-AAGGCGTCACAAGCAATCACTC 10987654321 AGATCGGAAGAGCGTCGTGTAGGGAAAGAG-3'
Barcode 2/UMI (T19V*-UMI-bc2-W1*):	5'-BAAAAAAAAAAAAAAAAAA NNNNNN 87654321 AAGGCGTCACAAGCAATCACTC-3'
FAM-PE1*	/56-FAM/AGATCGGAAGAGCGTCGTGTAGGGAAAGAG
FAM-W1*	/56-FAM/AAGGCGTCACAAGCAATCACTC
FAM-A20	/56-FAM/AAAAAAAAAAAAAAAAAAAA
Fully assembled DNA primers:	CGATGACG TAATACGACTCACTATAGGG ATACCACCATGG CTCTTTCCCTACACGACGCTCTTCCGATCT 1234567890 GAGTGATTGCTTGTGACGCCTT 12345678 NNNNNN TTTTTTTTTTTTTTTTTTV
2. Library preparation as per ref. (2):	
RNA ligation:	/5Phos/AGATCGGAAGAGCGGTTCAGCAGGAATGCC/3SpC3/
2nd RT primer:	GTCTCGGCATTCCTGCTGAAC
PCR enrichment primers:	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTA CACGA CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCT GCTGAAC

E. Supplemental Experimental Procedures

Microfluidic Device Design and Operation

Design. The design of the microfluidics device used in this work is indicated in **Figure S2** and integrates several features. As described in the main text, it contains four inlets for, i) barcoded hydrogel microspheres (BHMs), ii) cell suspension, iii) reverse transcription (RT) and lysis reagent mix and iv) carrier oil, and one outlet port for droplet collection. To reduce flow fluctuations potentially arising due to mechanics of syringe pumps, we incorporated fluid resistors in the form of serpentine channels, while passive filters at each inlet prevent channels from clogging. The device consists of two junctions, one for bringing the three aqueous inputs together, and a second junction for sample encapsulation, where aqueous and oil phases meet and droplet generation occurs. To stabilize drops against coalescence we use 0.75% (w/w) EA-surfactant (RAN Biotechnologies Inc.,) dissolved in HFE-7500 (3M) fluorinated fluid. The dimensions of the microfluidic channels were chosen after optimization to maximize the number of BHM and cell co-encapsulation events. The width (60 µm) of the BHM injection channel is designed such as that the BHMs (63 µm in diameter) passing through this channel become slightly squeezed thus facilitating their close packing and arrangement into a single-file. The BHMs entering into the main channel (70 µm wide) can then move freely downstream the flow before being encapsulated into individual droplets. Because of their close packing, the arrival of BHMs becomes highly regular allowing nearly 100% loading of single-bead per droplet. This feature is one of the most important characteristics of the developed chip ensuring that i) almost each cell encapsulated into a droplet is exposed to one barcoded primer, and ii) there is a minimal loss of non-barcoded-cells.

Soft lithography. The microfluidic device with rectangular microfluidic channels 80 µm deep was manufactured following a protocol reported recently (3). Briefly, a 3" size silicon wafer was coated with SU-8 3050 photoresist (MicroChem) at uniform 80 µm film thickness, baked at 65ºC for 20 min and exposed to 365 nm UV light for 40 s (at ~8 mW cm^2) through the mask having a corresponding design indicated in **Figure S2** and baked for 5 min at 95ºC. The un-polymerized photoresist was dissolved with propylene glycol monomethyl ether acetate, silicon wafer rinsed with isopropanol and dried on a 95ºC hot plate for 1 min. The PDMS base and cross-linker (Dow Corning) was mixed at 10:1 ratio and ~ 30 mL poured into the Petri dish containing a developed silicon wafer, degassed and incubated overnight at 65ºC. The PDMS layer was then peeled-off and inlet-outlet ports were punched with a 1.2 mm biopsy punch (Harris Uni Core). The patterned side of PDMS was then treated with oxygen plasma and bounded to the clean glass slide. The micro-channels were treated with water repellent Aquapel (PPG Industries) and the device was then used in the experiments.

Microfluidic device operation. During device operation, cell suspension and RT/lysis mix were cooled with ice-cold jackets, and droplets were collected into a single 1.5 mL tube (DNA LoBind, Eppendorf) placed on an ice-cold rack (IsoTherm System, Eppendorf). To prevent water loss from the droplets due to evaporation during RT incubation, we placed 200 µL of mineral oil layer (Sigma) on top of the emulsion. Throughout the experiments we used flow rates at 100 µL/hr for cell suspension, 100 µL/hr for RT/lysis mix, 10-20 µL/hr for BHMs and 80 µL/hr for carrier oil to produce 4 nL drops at a frequency of 15 droplets per second. Each aqueous phase was injected into the microfluidic device via polyethylene tubing (ID 0.38 x OD 1.09 mm, BB31695-PE/2) connected to a needle of a sterile 1 mL syringe (Braun) placed on a syringe pump (Harvard Apparatus, PC2 70-2226).

Loading barcoded hydrogel microspheres (BHMs) into the microfluidic device. After synthesis, BHMs were stored in $T_{10}E_{10}T_{0.1}$ buffer containing 10mM Tris-HCl [pH 8.0], 10mM EDTA, 0.1% (v/v) Tween-20. Before loading onto the microfluidic chip, BHMs were washed in $T_{10}E_{0.1}T_{0.1}$

buffer containing 10mM Tris-HCl [pH 8.0], 0.1mM EDTA and 0.1% (v/v) Tween-20, and then resuspended in 1X RT buffer (Invitrogen Superscript III buffer) supplemented with 0.5% (v/v) IGEPAL CA-630 and concentrated by centrifugation at 5000 rcf for 2 min. After removal of the supernatant BHMs were concentrated for a second time to achieve a close packing and eventually loaded directly into tubing connected to a oil-filled syringe for injection into the microfluidic device. The composition of BHM sample prior to concentration was 100 µL pre-concentrated BHMs in $T_{10}E_{0.1}T_{0.1}$, 20 µL 10% (v/v) IGEPAL CA-630, 40 µL 5X First-Strand buffer and 40 µL nuclease-free water (total aliquot volume 200 µL).

Cell preparation and injection. The cell encapsulation process relies on random arrival of cells into the device. To minimize two or more cells from entering the same drop, we encapsulated cells with an average occupancy of 1 cell in 5-10 droplets, by diluting cell suspensions to ~50-100,000 cells/mL. To prevent cell sedimentation in the syringe or other pars of the system, we suspended cells in 1X PBS buffer with 16% (v/v) density gradient solution Optiprep (Sigma). We typically used 20,000 cells suspended in 160 µL 0.5X PBS (17-516F, Lonza), 32 µL Optiprep (1114542, Axis-Shield) and 8 µL 1% (v/v) BSA (B14, Thermo Scientific), in a total volume 200 µL. Cells were maintained in suspension using a micro-stir bar placed in the syringe, and rotated using a magnet attached to a rotating motor.

Reverse transcription/lysis mix. The RT/lysis mix consisted of 25 µL 5X First-Strand buffer (18080-044 Life Technologies), 9 µL 10% (v/v) IGEPAL CA-630 (#18896 Sigma), 6 µL 25 mM dNTPs (Enzymatics N2050L), 10µL 0.1M DTT (#18080-044, Life Technologies), 15 µL 1M Tris-HCl [pH 8.0] (51238 Lonza), 10µL Murine RNase inhibitor (M0314, NEB), 15 µL SuperScript III RT enzyme (200 U/µL, #18080-044, Life Technologies) and 60 µL nuclease-free water (AM9937 Ambion), having a total volume 150 µL.

Surfactant and carrier oil used for production of droplets. The carrier oil was HFE-7500 fluorinated fluid (3M) with 0.75% (w/w) EA surfactant (RAN Biotechnologies). EA-surfactant is a tri-block copolymer having an average molecular weight of ~13.000 g mol^{-1}. It has two perfluoropolyether tails (M_W ~6.000 g mol^{-1}) connected via poly(ethylene)glycol (M_W ~600 g mol^{-1}) head group. The surfactant is highly soluble in fluorinated fluids and nearly insoluble in the aqueous phase providing equilibrium interfacial tension of ~ 2 mN/m.

Barcoding inside droplets. After cell encapsulation primers were released from the BHMs by exposing the tube containing the emulsion droplets to UV light (365 nm at ~10 mW/cm^2, BlackRay Xenon Lamp) while on ice. Next, the tube was heated to 50ºC and incubated for 2 hours to allow cDNA synthesis to occur and then terminated by heating for 15 min at 70ºC. The emulsion was then cooled on ice for 1 min and demulsified by adding 1 volume of PFO solution (20% (v/v) perfluorooctanol and 80% (v/v) HFE-7500). The aqueous phase from the broken droplets was transferred into a separate DNA Lo-Bind tube (Eppendorf) and processed as per the CEL-SEQ protocol with modifications described in the library preparation section.

Synthesis and Quality Control of Barcoded Hydrogel Microspheres

BHM Synthesis. BHM synthesis relies on microfluidic emulsification of acrylamide:bis-acrylamide solution supplemented with acrydite-modified DNA primer, which is incorporated into the hydrogel mesh upon acrylamide polymerization. After polymerization, the BHMs are released from droplets, washed several times and processed by split-pool synthesis for combinatorial barcoding. Below we outline the detailed protocol of performing such hydrogel bead synthesis followed by combinatorial barcoding.

BHM synthesis begins by emulsifying gel precursor solution into 63 µm size droplets using the microfludic chip indicated in **Figure 2.** The composition of the dispersed phase is 10 mM Tris-HCl [pH 7.6], 1 mM EDTA, 15 mM NaCl containing 6.2% (v/v) acrylamide, 0.18% (v/v) bis-acrylamide, 0.3% (w/v) ammonium persulfate and 50 µM acrydite-modified DNA primer (IDT,

see **Figure 2** for sequence). As a continuous phase we use fluorinated fluid HFE-7500 carrying 0.4% (v/v) TEMED and 1.5% (w/w) EA-surfactant. The flow rates are 400 µL/hr for the aqueous phase and 900 µL/hr for the oil phase. Droplets were collected into a 1.5 mL tube under 200 µL mineral oil and incubated at 65ºC for 12 hours to allow polymerization of beads to occur. The resulting solidified beads are washed twice with 1 mL of 20% (v/v) 1H,1H,2H,2H-perfluorooctanol (B20156, Alfa Aesar) in HFE-7500 oil and twice with 1 mL of 1% (v/v) Span 80 (S6760, Sigma) in hexane (BDH1129-4LP, VWR) with 0.5-1 min incubation between each step and finally centrifuged at 5000 rcf for 30s. After final centrifugation the hexane phase is aspirated and the resulting BHM pellet is dissolved in 1 mL of TEBST buffer (10 mM Tris-HCl [pH 8.0], 137 mM NaCl, 2.7 mM KCl, 10 mM EDTA and 0.1% (v/v) Triton X-100). To remove traces of hexane, the beads are washed three times with 1 mL TEBST buffer at 5000 rcf for 30 s and finally resuspended in 1 mL TEBST buffer and stored at +4ºC. According to previous reports (4), these BHMs contain pores ~ 100 nm in size. In addition, the beads having elastic modulus of ~ 1 kPa (5) are squishy, which allows to pack them into a concentrated gel mass without loosing their integrity.

BHM split-pool combinatorial barcoding. To prepare barcoded primers on the hydrogel microspheres, we used the two-step enzymatic extension reaction summarized in **Figure 2**. To begin, we first pre-loaded a 384-well plate with 9 µL of 15 µM primer 5'-W1*-bc1-PE1* encoding the first-half of a barcode (where 'bc1' indicates a unique sequence for each well, see also **Table S5** for nucleotide sequence information). We then added 6 µL of reaction mix containing ~40,000 hydrogel beads (carrying 5'-Ac-PC-T7p-PE1 primer), 2.5x isothermal amplification buffer (NEB) and 0.85 mM dNTP (Enzymatics) into each well (accounting ~10^7 beads in total). After denaturation at 85ºC for 2 min and hybridization at 60ºC for 20 min we added 5 µL of *Bst* enzyme mix (1.8U of *Bst* 2.0 and 0.3 mM dNTP in 1X isothermal amplification buffer) giving to a final volume in each well of 20 µL. After incubation at 60ºC for 60 min, the reaction was stopped by adding 20 µL of stop buffer into each well (100mM KCl, 10mM Tris-HCl [pH 8.0], 50mM EDTA, 0.1% (v/v) Tween-20) and incubated on ice for 30 min to ensure that EDTA chelates magnesium ions and inactivates *Bst* enzyme. Next, the beads were collected into a 50 mL Falcon tube, centrifuged at 1000 rcf for 2 min and washed three times with 50 mL of stop buffer containing 10 mM EDTA. To remove the second strand we suspended gels in 20 mL of 150 mM NaOH, 0.5% (v/v) Brij 35P and washed twice with 10 mL of 100 mM NaOH, 0.5% (v/v) Brij 35P. The alkaline solution was then neutralized with buffer 100 mM NaCl, 100 mM Tris-HCl [pH 8.0], 10 mM EDTA, 0.1% (v/v) Tween-20 and washed once in 10 mL $T_{10}E_{10}T_{0.1}$ buffer (10 mM Tris-HCl [pH 8.0], 10 mM EDTA, 0.1% (v/v) Tween-20) and twice in 10 mL $T_{10}E_{0.1}T_{0.1}$ buffer (10 mM Tris-HCl [pH 8.0], 0.1 mM EDTA, 0.1% (v/v) Tween-20) and finally beads were suspended in 1.3 mL of $T_{10}E_{0.1}T_{0.1}$ buffer.

For the second barcoding step we prepared a second 384-microtiter plate pre-loaded with 9 µL of 15 µM primer 5'-T19V*-UMI-bc2-W1* (where 'bc2' indicates a unique sequence for each well and UMI is a random hexanucleotide, see also **Table S5** for sequence information), and repeated the procedure as for the first 384-well plate.

Quantification of ssDNA primers on the beads. To quantify the amount of the ssDNA primers per BHM, we performed fluorescence *in situ* hybridization (FISH) with complimentary DNA probes targeting the un-extended DNA "stub" (PE1), the barcoded primer after one extension step (W1) and the primer after two extension steps ($T_{19}V$) (see **Table S5** for sequence information). Hybridization was performed in a 40 µL volume at room temperature for 20 min by suspending ~4000 DNA-barcoding beads in hybridization buffer (1M KCl, 5 mM Tris-HCl [pH 8.0], 5 mM EDTA, 0.05% (v/v) Tween-20) together with 10 µM FAM-labeled probe. The high salt concentration was necessary to avoid melting of the probe targeting $T_{19}V$ (dA_{20}-FAM), which has weak binding even at room temperature. We validated the absence of background fluorescence in microspheres lacking DNA primers. After incubation, beads were washed three

times with 1.4 mL hybridization buffer, re-suspended in 40 µL and fluorescence intensity was recorded under confocal microscope (Leica). The average fluorescence intensity of beads with PE1*-FAM, W1*-FAM and dA20-FAM was 2286 ± 271, 1165 ± 160 and 718 ± 145, respectively (**Figure S1A-D**). This corresponds to incorporation efficiencies of ~50% for W1/PE1 and 60% for polyT/W1, which gives the final efficiency of 31% or ~15 µM of fully barcoded ssDNA primers per bead. Accounting of the BHM volume, this equals ~ 10^9 copies of fully extended ssDNA primers per single bead.

To validate the release of primers from the hydrogel mesh we suspended ~4000 beads in 20 µL $T_{10}E_{0.1}T_{0.1}$ buffer (10mM Tris pH 8.0, 0.1mM EDTA, 0.1% Tween 20) and exposed them to UV light (365 nm at ~10 mW/cm^2) for 8 min. A gel electropherogram of 1 µL of supernatant using a BioAnalyzer High Sensitivity DNA Analysis Kit (Agilent Technologies) confirmed the presence of three DNA bands (**Figure S1E**), which is in agreement with FISH results from above.

Single-molecule sequencing of primers from single BHMs. To test the composition of BHMs after synthesis, we randomly picked 11 BHMs and sequenced them using the Illumina MiSeq sequencing platform. For this purpose, the BHMs were first hybridized to a fluorescent FISH probe (PE1-FAM) as described above, and were manually picked using a dissection microscope (Nikon) under fluorescent illumination and transferred into 0.2 mL PCR tubes pre-filled with 5 uL DNA Suspension (DS) buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA). The tubes were then exposed to UV light (~ 10 mW/cm^2) for 15 min while keeping them on ice. After UV exposure, 0.5 µL of 5 µM PE2-(barcode)$_n$-A19 primer (herein, *n* represents 10 different barcodes) was added to the tube and mixed with 4.5 µL of *Bst* 2.0 ready-to-use reaction solution. The samples having 10 µL final volume were then incubated at room temperature for 10 min, inactivated for 3 min at 95ºC and cooled down on ice. Next, 20 µL of master mix containing 50% (v/v) Kapa HiFi HotStart ready mix (2X, KK2601), 15% (v/v) PE1/PE2 primers, and 35% (v/v) nuclease-free water were added into each tube, and DNA was amplified with PCR (95°C for 5 min, 30 cycles at 98°C for 20 s, 60°C for 15 s, 72°C for 30 s, and final step at 72°C for 5 min). The size of the PCR products was assessed by gel-electrophoresis, purified with GenElute PCR CleanUp Kit (Na1020-1KT, Sigma) and all samples diluted down to 10 ng/uL. In the final step all samples were pooled together and sequenced using MiSeq Illumina platform by following manufacturer recommendations. Sequencing results are presented in **Figure S1F-H**.

Limits on the Number of Cells per Single Sequencing Run

For a large pool of barcoded hydrogel microspheres (BHMs), each carrying one of *N* barcodes, what is the maximum number of cells that can be captured before two or cells will carry the same barcode? This question is akin to the so-called birthday problem, with barcodes analogous to days of the year, and BHMs analogous to the people in a room. The expected number of observed barcodes from sampling *n* BHMs is $n_{obs} = N\left(1 - e^{-n/N}\right)$. Thus, the expected multi-barcoding error, defined as the fraction of cells carrying the same barcode, is approximately $f_{err}=1-n_{obs}/n$. The error becomes large when n~N, so in practice the number of sampled cells must be much smaller than the number of barcodes, i.e. $n<<N$, and therefore the limit of obtaining barcoded single-cells is $f_{err}\approx n/2N$. The number of barcoded single-cells *n* depends on the tolerated error, for example,allowing for an error of less than f_{err}=1% sets an upper limit $n=N/50$. Thus, for the value of $N=384$[2] which corresponds to two 384-well plates in our experiment, a 1% multiple-barcoding error arises at the limit n=2,949 cells. All of our libraries were prepared with fewer than this number of cells and therefore have negligible multi-barcoding errors. To barcode a larger number of cells, emulsions may be split into pools of <3,000 cells and prepared into libraries carrying an additional library barcode. Alternatively, BHMs may be synthesized using more than 2 rounds of split pool synthesis to exponentially increase the initial barcode pool *N*.

Cell Cultures

The mouse embryonic stem (ES) cells are a line derived from the mouse 129/Ola strain called IB10 (subcloned from E14). They were kindly provided by Dr. Eva Thomas (University of Wuerzburg, Germany) from their paper (Thomas et al., Cell Reprogram 2012). The ES cells were maintained in ESC base media inside culture flasks pre-coated with gelatin at 37°C in 5% CO_2 and 60-80% humidity at density ~3 × 10^5 cells/mL. The ESC media contained phenol red free DMEM (Gibco), supplemented with 15% (v/v) fetal bovine serum (Gibco), 2 mM L-glutamine, 1x MEM non-essential amino acids (Gibco), 1% (v/v) penicillin-streptomycin antibiotics, 110 µM β-mercaptoethanol, 100 µM sodium pyruvate. In the undifferentiated state the ESC base media was supplemented with Leukemia Inhibitory Factor (LIF) at final concentration 1000 U/mL and for unguided mES differentiation the media was without LIF. Within 2 days of LIF withdrawal the culture experienced significant morphological changes indicating the differentiation of mES cells.

Prior to ES cell encapsulation the flask was washed with 1x PBS (without Mg^{2+} and Ca^{2+} ions) and treated with 1x trypsin/EDTA solution for 3 min at 37ºC. The trypsin was quenched by adding equal volume of ESC base media. Detached cells were centrifuged at 260*g* for 3 min and re-suspended in ~ 3 mL fresh ESC base media. After passing through a 40 µm size strainer, cells were counted with hemocytometer and diluted in 0.5x PBS supplemented with 0.04% (v/v) BSA and 16% (v/v) OptiPrep solution to obtain desirable amount of cells (typically 20.000 cells in 200 µL). The suspension was transferred into 1 mL syringe containing a micro-stir bar, and connected to microfluidics device and injected at 100 µL/hr flow rate as described above.

The K-562 cell line (ATCC, CCL-243) was maintained in DMEM supplemented with 10% (v/v) fetal bovine serum and 1% (v/v) penicillin-streptomycin at 37°C in 5% CO_2 and 60-80% humidity atmosphere, at density ~3 × 10^5 cells ml^{-1}. For encapsulation experiments K-562 cells were pelleted at 200g and then prepared as described for ES cells. After scoring the number of cells, K562 cells were mixed with mES cells at ratio 1:1.

DNA-Library Preparation

Library preparation was based on the modified CEL-Seq/MARS-Seq protocol in (2) with minor modifications. The workflow of DNA library preparation can be summarized as follows: RT → ExoI → SPRI purification (SPRIP) → SSS → SPRIP → T7 in vitro transcription linear amplification → SPRIP → RNA Fragmentation → SPRIP → primer ligation → RT → library enrichment PCR.

Referring to the detailed protocol in (2), the following modifications were made to the protocol: the RT primer included the P5/PE1 adaptor while the ligation primer includes the P7/PE2 adaptor, a flipped orientation to that in (2); prior to ExoI treatment, the aqueous phase from broken droplets was centrifuged at 4°C for 15 minutes at 14krcf to pellet cell debris and gels; during ExoI treatment, 10U HinFI were added to digest primer dimers that may have formed during the RT reaction; the original DNAse digestion step was omitted after linear amplification; after linear amplification the resulting amplified RNA libraries were analyzed on an Agilent BioAnalyzer before proceeding; before primer ligation, the samples were treated with Shrimp Alkaline Phosphatase for 30 minutes. The number of final PCR cycles required for final library enrichment PCR ranged from 10-13 cycles. The remaining steps are otherwise unchanged.

Sequencing and Data Filtering

Paired-end sequencing was performed on Illumina MiSeq, HiSeq 2500 and NextSeq machines as detailed in **Table S1**. Read 1 was used to obtain the sample barcode and UMI sequences; read 2 was then mapped to a reference transcriptome as described below. The reads were first filtered based on presence in read 1 of two sample barcode components separated by the W1 adaptor sequence (see **Figure 2** and **Table S5**). Read 2 was then trimmed using Trimomatic (6) (version 0.30; parameters: LEADING:28 SLIDINGWINDOW:4:20 MINLEN:19). Barcodes for each read were matched against a list of the 384^2 pre-determined barcodes, and errors of up to two nucleotides mismatch were corrected. Reads with a barcode separated by more than two nucleotides from the reference list were discarded. The reads were then split into barcode-specific files for mapping and UMI filtering.

Bioinformatic Processing from Raw Reads to UMI-Filtered Counts

Sequence analysis pipelines intended for bulk (non-single cell) applications map ambiguous reads probabilistically in a manner that can spuriously couple otherwise independently expressed genes. This problem may be particularly acute for 3'-sequencing of single cells since UTR regions can be similar across multiple genes; and in relatively uniform cell populations such as ES cells, which are characterized by a wide network of weak gene expression couplings that become comparable to those generated spuriously. One could discard all ambiguous reads (e.g. (Jaitin et al., 2014)), but this approach may affect genes non-uniformly depending on the uniqueness of the transcript tails. We overcame the read-mapping problem by writing a bioinformatic pipeline that tracks mapping ambiguities and facilitates their elimination in downstream analysis, as detailed below.

Reads split into barcode-specific files were aligned using Bowtie (version 0.12.0, parameters: -n 1 –l 15 –e 300 –m 200 –best –strata -a) to the mouse transcriptome. We also re-processed the data sets with different bowtie parameter sets without changing the qualitative results of the analysis, although the number of UMIFM reads varied between ~14-30k depending on parameter choice. The reference transcriptome was built using all annotated transcripts (extended with a 125bp poly-A tail) from the UCSC mm10 genome assembly for mouse ES cells, and the hg19 assembly for human K562 cells. We used a custom Python and PySAM script to process mapped reads into counts of UMI-filtered transcripts per gene. Alignments from bowtie were filtered in the following way: (1) for each read, we retained at most one alignment per gene, across all isoforms, by choosing the alignment closest to the end of the transcript. (2) If a read aligned to multiple genes, we excluded any alignments more than 400 bp away from the end of the transcript; this is motivated by the strong 3' bias of the CEL-SEQ method. This step results in approximately 5% increase in the number of final UMIFM reads obtained, as compared to simply discarding any ambiguous read. (3) We then excluded reads mapping to more than 10 genes, and (4) we performed a UMI filtering step described in the following paragraph. Finally, (5) if a read still aligned to more than 2 genes after UMI filtering, we excluded the read altogether. This step results in a very modest (<1%) increase in the final UMIFM reads obtained, as compared to simply discarding any ambiguous reads. However, by performing this final step, we were able to report for each gene any other genes from which it could not be distinguished in at least one read; this allowed us to exclude spurious correlations in our downstream analysis resulting from mapping ambiguities. This turned out to be an important step to avoid correlating genes based on sequence similarity as well as based on expression. We confirmed the robustness of the pipeline to this final step by re-processing the data with a maximum of 1-4 allowed alignments per read. After steps (1-5) were carried out separately for each sample, the resulting gene expression tables were concatenated and loaded into MATLAB for analysis.

UMI filtering (step 4 above) was carried out as follows. Each distinct UMI is associated with a set of genes through the set of reads carrying the UMI. For each UMI, we identified the minimal set of genes that can account for the full set of reads with this UMI. This problem is known as the 'Hitting Set Problem' (or 'Set Cover Problem') (7). We applied a greedy algorithm (8) to obtain the most parsimonious gene set for each UMI. For this final gene set, we kept only one read per gene per UMI. With this approach, some subsets of genes may still be undistinguishable from each other because they are supported by the same set of ambiguously aligned reads. Step (5) in the previous paragraph was thus used to eliminate ambiguous reads beyond a predetermined threshold. To illustrate the UMI filtering step, consider a single UMI present in two reads, the first aligning to genes A and B and the second aligning to genes B and C. Although neither read aligns unambiguously, gene B alone can explain the presence of both reads and thus the alignments to genes A and C are discarded, and just one of the two reads is kept for gene B.

UMI-Filtered Count Normalization

Prior to normalization, the variation in the total UMI-filtered mapped (UMIFM) counts per sample barcode was 21% to 55% (coefficient of variation), see **Table S1**. The CV appeared to grow during differentiation, suggesting that some of the variation in total UMIFM counts arose from differences in cell size rather than in variation in RT efficiency. This was consistent with an increase in the inferred value of $CV_{1/N}$ over time from the normalized counts, as described in the **Theory Supplement**. We normalized all counts by total-count normalization, i.e. the normalized counts for gene *j* in cell *i* is given in terms of the un-normalized counts, $m_{i,j}$, as $\hat{m}_{i,j} = m_{i,j}\bar{M}/M_i$, where $M_i = \sum_j m_{i,j}$ and $\bar{M}$ is the average of M_i over all cells. The effects of normalization on the gene CVs and correlations are provided in the **Theory Supplement**, see Eq. (S4).

Relating Method Sensitivity to Method Efficiency

This section derives the form of the sensitivity curve (solid red line) in **Figure 4E**, predicted for a case where the only limitation to detection is the capture efficiency, β, which is assumed to be uniform across all gene transcripts and droplets. All other possible effectors of sensitivity, such as sequence-specific or length-specific biases, are assumed negligible. The excellent fit reinforces these assumptions. With these assumptions, if *n* is the number of transcripts for a given gene in a given droplet, then the probability of detecting zero transcripts for a gene in this droplet is $p_0(n) = (1-\beta)^n$. The sensitivity *S* is then obtained by marginalizing $p_0(n)$ over the distribution of *n*, which, in the case of the pure RNA sample, is Poisson-distributed about a mean value $\bar{n}$. One obtains $S(\bar{n}) = 1 - \sum_{n=0}^{\infty} p_0(n)\,\mathrm{Poiss}[n;\bar{n}]$, giving,

$$S(\bar{n}) = 1 - e^{-\bar{n}\beta}$$

which is the curve plotted in **Figure 4E**, with the value of β measured from **Figure 4D**. Since $p_0(n)$ is an exponential, this curve can also be identified as one minus the moment generating function (MGF) of the Poisson distribution, $MGF(q) = e^{-\bar{n}(1-e^{-q})}$, evaluated at $q = -\log(1-\beta)$. The quality of the fit demands that variations in β between droplets be small, which is consistent with the low CV in the total counts, *M*, between control droplets, evaluated as CV_M=21%. For non-control samples, the input distribution for each gene is no longer a Poisson distribution, and the detection frequency $S(\bar{n})$ is instead different for each gene and, under the assumptions given here, is related to the MGF of the underlying gene expression distribution evaluated at $q = -\log(1-\beta) \approx \beta$.

Identification of Highly Variable Genes

The procedure is described in section S-III of the Theory Supplement. Briefly, genes were assigned a test statistic *v* defined in Eq. (S13) of the Theory Supplement, and the statistic was compared to that obtained from the pure RNA sample. This test statistic is derived from **Figure 4G**, Eq. (1), and corrects for variability due to sampling noise and technical noise in measurement. Because this test does not account for noise arising from bioinformatic analysis -- for example, differences in read mapping given different parameter sets -- we repeated the test for multiple bioinformatic parameter sets (see above), and then took the intersection of all sets as a final list of highly variable genes that is insensitive to the choice of sequence mapping parameters. Results are provided in **Table S3**.

Selection and Filtering of Principal Gene Sets for PCA and tSNE Analysis

Since each gene carries intrinsic sampling noise that is uncorrelated to other genes, it is expected that for whole-transcriptome data, a large fraction of the variability observed across all genes will not be explained by the top principal components. For the same reasons, differences between cell sub-populations may appear weak if a large number of "bystander" genes (which vary little between populations) are included in evaluating cell-cell correlations. To overcome these sampling limitations, we analyzed the ES cell population structure using only a sub-set of genes chosen to reflect known ES cell biology while also reporting on the most variable genes at each time point. The general strategy for selecting appropriate gene sets for **Figures 5F,G** and **7G** was as follows: (1) for each time point, we included the top 200 most variable genes as determined by the *v*-score (**Theory supplement** and **Table S3**), which is closely related to the gene Fano Factors; these genes were complemented with a curated list of genes implicated in ES cell biology. We also repeated the PCA analysis without curated genes, and including increasing numbers of variable genes that have at least 5 UMIFM counts in at least one cell (800, 1,600 and 2,400 genes tested). We found no change to the population structure across the first principal components (**Figure S5B**), and found that the majority of the curated genes were included in these lists. (2) We then performed a preliminary principal component analysis (PCA) on the cell population, using the initial gene set, and used the results to select only principal genes, i.e. genes contributing to non-random principal components (PCs) as shown in **Figure 5E**. The principal genes are those with the highest loading coefficients for each non-random PC. (3) For each gene *g* in the set, the set was then re-expanded to include five additional genes that correlated most strongly with *g*. Finally, (4) to avoid bioinformatic ambiguities from distorting downstream analysis, we grouped together the counts from any genes in the final set that were reported to map ambiguously to the same reads. For this final step we made use of the list of ambiguous mapping partners for each gene, generated by the custom bioinformatics pipeline described above. For example, since a number of reads mapped ambiguously to both *Col4a1* and *Col4a2*, the counts from these two genes were summed before further analysis to yield a single composite gene *Col4a1/Col4a2*, thus eliminating any structure arising from the strong correlation between these two genes. The final gene set derived at the end of step (4) was used for subsequent PCA and tSNE analysis at each time point.

PCA and tSNE Analysis of Population Structure

tSNE was carried out using the MATLAB implementation by van der Maaten and Hinton, http://lvdmaaten.github.io/tsne/#implementations. Both PCA and tSNE of cells (**Figures 5F,G, and 7G**, and **Figure S5B**) was carried out on total count-normalized UMIFM counts, after z-score standardization of gene expression (setting the standard deviation to unity). tSNE of

genes (**Figure 7G**, right panel, and **Figure S7**) was performed on total count-normalized UMIFM counts without z-score standardization.

As a control, repeating the PCA/tSNE analysis after row-randomizing the data revealed no cell sub-populations.

PCA across Multiple Time Points

To visualize temporal heterogeneity in differentiation (**Figure 7D**), we performed PCA on cells grouped from all time points using the top 200 most highly variable genes from the ES cell data set and day 7 post-LIF withdrawal. To this list we added the 200 genes with the greatest differential expression between ES and day 7 data (**Figure S4 and Table S2**). No correction was performed to ensure an equal number of cells per time point. UMIFM counts were normalized to give the same average total counts per cell per time point. Because all time points contain Primitive Endoderm, ICM and Epiblast cells, a simple PCA on the cell populations reveals common sub-populations rather than dynamics. Thus, to ensure that the first two principal components visualized the temporal trajectory of the cells, prior to PCA we excluded cells for which any of the ten nearest neighbors belonged to a different time point. Following PCA, the excluded cells were then restored after calculating the principal component projections to generate **Figure 7D**.

Network Neighborhood Analysis

Algorithm. The network neighborhood analysis (NNA) requires a distance metric between gene pairs, a "bait" gene, and two parameters defining the neighborhood size N and the minimal connectivity X. We used the distance metric d=[1-(Pearson correlation)], where the correlation is taken over all cells. From d, an unweighted, directed network that constitutes the "network neighborhood of gene G_0" is constructed as follows: (1) for the given gene G_0 of interest, a directed edge is introduced between G_0 and each of the N genes with the smallest distance to G_0, which we define as a set G_1. Note that when N is much smaller than the number of genes in the system, the set G_1 is positively correlated to G_0, since most gene-gene correlations are very near zero. Thus, no cut-off in the correlation values is required for this analysis. (2) To each member of the set G_1 are similarly added additional directed edges to N closest genes, together forming a set G_2. Note that G_2 may include G_0 and members of G_1. The resulting preliminary network has (N+1)*N directed edges in total, and up to 1+(N+1)*N vertices representing G_0, G_1 and G_2. (3) The network is trimmed iteratively by removing any vertex that has fewer than X incoming edges. The final network is the X-connected network neighborhood of gene G_0. If the network is not empty, it consists of at least $X+1$ genes including: the gene G_0; some members of G_1 that are also nearest neighbors of at least X-1 other members of G_1; and some members of G_2 that are the nearest neighbors of at least X members of G_1.

Sensitivity and specificity as a function of NNA parameters. Robustness of NNA to spurious correlations was tested by row-randomizing the UMIFM count data. Randomized data yielded empty neighborhoods for all genes tested with all parameters. Randomization also revealed that the magnitude of correlations between NNA neighbors could not be explained by chance alone (p-value < $2 \cdot 10^4$ concluded from $2 \cdot 10^4$ randomization trials).

To test the sensitivity of NNA to the choice of (N, X), we generated an ROC curve for three test genes, Nanog, Trim28 and Sox2, for a range of parameters (N, X) (**Figure S6A**). The ROC curve is almost unchanged for X=2-5, although it shifts to higher N as X increases. To generate the ROC curves, one requires a true positive gene set composed of validated pluripotency factors expected to associate with Nanog and Sox2. This is problematic since no such set can

be complete based on existing literature, and also because Nanog and Sox2 may have meaningful associations outside of the pluripotency network. We curated a list of 33 pluripotency factors, not limiting ourselves to transcription factors *(9)*: *Pou5f1, Nanog, Sox2, Esrrb, Zic3, Sall4, Klf4, Klf5, Zfp143, Nr0b1, Tfcp2l1, Tbx3, Tcl1, Myc, Zfx, Cnot3, Trim28, Smad1, Stat3, Zfp42, Nacc1, Zfp281, Zic3, Tdh, Calcoco2, Mybl2, Kdm5b, Fbxo15, E130012A19Rik, Hmgb2, Hmga1, Tdgf1, Stmn2.* This list is likely to be incomplete but provides a basic training set. We assumed that any gene not in the true positive set is a negative hit, and we excluded all genes that could not be found in at least one cell at a level of >5 UMIFM counts, in order to avoid scoring genes with insufficient coverage. This negative set constituted ~17,000 genes. This is likely to underestimate sensitivity and specificity, since several genes associating with Nanog and Sox2 are associated with pluripotency (e.g. *Dppa5a*), but are not in the positive set as their requirement in pluripotency has not been demonstrated. As seen in **Figure S6A,** for this test sensitivity was ~30% and specificity > 99.9%. For the neighborhoods plotted in **Figure 6B-D**, we selected N=50, X=3.

Pluripotency correlation neighborhood. To assess the significance of novel factors associating the pluripotency network, a network neighborhood was constructed for each of the curated pluripotency genes (PGs) (using N=50, X=3), and non-pluripotency genes that were neighbors of at least two PGs were identified (**Figure 6B-D** shows adjacency between PGs only). A p-value giving the significance for the re-occurrence of these PG-neighboring genes was evaluated for the null hypothesis that a given candidate neighbors two or more PGs by chance (a variant of the Fisher exact test), and corrected for multiple hypothesis testing. This test is informal, because each of the PGs does not represent an independent sampling event, as the PGs are themselves correlated. The corrected (informal) p-value is $q \approx z_1 * \ldots * z_k / M^{k-1}$, where z_i is the number of non-PF neighbors of the i-th PG, k is the number of PGs for which the candidate is found to be a neighbor, and M~17,000 is the number of genes acting as the background pool. The 43 genes with q<0.05 are given in **Figure S6C**.

E. Theory Supplement

Here we derive Eqs. (1)-(3) in Fig. 2G of the main text, showing how the main sources of technical noise encountered in inDrop RNA sequencing and other transcriptomic methods affect gene expression variability and correlations observed in single cell sequencing data. We develop a statistical test based on the results presented here to identify highly variable genes.

E.1 A model for technical noise

Let n_i be the number of mRNA molecules in a given cell that correspond to a particular gene i, with a distribution $P_{\rm bio}(n_i)$ across all cells being analyzed. Also, let m_i be the number of UMI-filtered mapped (UMIFM) sequencing reads mapping to the same gene i, with distribution $P_{\rm obs}(m_i)$ obtained from the sequencing run. Ultimately we want to use the distribution of UMIFM reads $P_{\rm obs}(m_i)$ to infer properties of $P_{\rm bio}(n_i)$ such as its average $E[n_i]$, its variance $\mathrm{Var}(n_i)$, its coefficient of variation $\mathrm{CV}_i = \sqrt{\mathrm{Var}(n_i)}/E[n_i]$, and its modality. We also want to infer properties of the joint distributions of multiple genes – for example, the strength of the correlation between gene i and gene j, which we can calculate from the pairwise distributions $P_{\rm bio}(n_i, n_j)$. The challenge in developing a model of noise in single cell transcriptomics is to explain how the joint distribution of the UMIFM read counts of all genes $P_{\rm obs}(\{m_i\})$ relates the joint distribution of the transcript counts of these genes $P_{\rm bio}(\{n_i\})$. From $P_{\rm obs}(\{m_i\})$ we can extract any marginal distribution of interest.

To relate the set of read counts across all genes $\{m_i\}$ to the set of transcript counts $\{n_i\}$, we start with the chain rule,

$$P_{\rm obs}(\{m_i\}) = \sum_n P_{\rm bio}(\{n_i\}) \prod_i Q_i(m_i|n_i), \tag{S1}$$

where $\{Q_i(m_i|n_i)\}$ are the conditional probability distributions for observing $\{m_i\}$ UMIFM reads given $\{n_i\}$ transcripts. The extent to which $P_{\rm obs}(\{m_i\})$ reflects the biology depends on the structure of $\{Q_i(m_i|n_i)\}$. Note that, implicit in our notation for Q_i, we have assumed that transcripts are sampled independently of one another within each droplet, although there may be variation between droplets. This assumption is supported by the excellent fit of the sensitivity curve (Fig. 2E) assuming independent and random sampling of mRNA transcripts (see Supplementary Methods section on Sensitivity). Thus the number of UMIFM reads m_i for a given gene should depend only on the actual number of transcripts n_i for that gene and not on the number of UMIFM reads m_j or transcripts n_j for any other gene.

To construct $Q_i(m_i|n_i)$, we consider the type of noise apparent in our system. Previous studies (Grun et al., 2014; Brennecke et al., 2013; Islam et al., 2014) assumed that sequenced transcripts are sampled from the pool of all transcripts in a cell according to Poisson statistics, i.e. through random sampling with replacement. These studies were motivated by the observation that the majority of genes in a sample follow a Poisson noise relationship, $\mathrm{CV}^2 \sim 1/(\mathrm{mean})$, with a baseline additive technical noise. Here, we make a subtle correction to these studies by noting that sampling of mRNA transcripts occurs *without* replacement, giving rise to a Binomial, not Poisson, distribution. A Binomial sampling is expected when sequencing depth is sufficient such that the number of reads per cell is limited only by the capture efficiency of the transcripts; if however the number of sequenced reads is too low to capture the full complexity of the RNA-Seq library (for

example, if the number of cells sequenced in a single run is very high), the limited number of reads per cell gives rise to a Hypergeometric distribution instead of a Binomial, but with similar results. With this correction we are able to show from first principles how a baseline noise arises, and that it not only additively complements the gene CVs but also multiplicatively amplifies existing biological variation.

The choice of a Binomial distribution is motivated by the observation that the number of UMIFM reads m_i for a given gene from a given cell can only under-estimate the true number of transcripts n_i. Thus m_i is drawn from a Binomial distribution characterized by a sampling efficiency β corresponding to the probability of any individual mRNA molecule being sampled. Since β can fluctuate between droplets independently of m_i, Q_i is the Binomial distribution marginalized over fluctuations in β, viz.

$$Q_i(m_i|n_i) = \int d\beta\, \xi(\beta) Bi(m_i; n_i, \beta), \tag{S2}$$

where $Bi(m_i; n_i, \beta)$ is the Binomial distribution,

$$Bi(m_i; n_i, \beta) = \binom{n_i}{m_i} \beta^{m_i} (1-\beta)^{n_i - m_i},$$

and $\xi(\beta)$ is the distribution across droplets of sampling efficiencies β. Note that Eqs. (S1) and (S2) assume that all transcripts within the same droplet are sampled with the same efficiency β, which may not be true (for example) if some transcripts within the same cell are more or less accessible to primer capture than others. These equations also ignore other sources of noise such as ambiguities in mapping UMIFM reads to genes and rare events in which two cells are present in the same droplet. Despite these limitations, from Eqs. (S1) and (S2) one may derive predictions for the observed gene CVs and correlations that agree well with trends seen in the data, and which provide an intuitive explanation for sources of variation in the data.

Having laid out the basic structure and assumptions of our noise model, we now use them to relate variability and correlations in the number of UMIFM reads for genes to those properties of the actual number of transcripts. We also explore how normalizing the data affects our results.

E.2 Expressions for noise in single cell transcriptomics

From the noise model in section E.1, we find that technical noise amplifies existing biological variation of a gene's abundance across cells and weakens correlations between genes. Here we formalize these intuitive behaviors through equations that relate the biological CV and pairwise correlation strength with their experimentally observable counterparts.

E.2.1 Technical noise amplifies above-Poisson biological variation

Equations (S3) and (S4) below present the key relationships describing the observed CV of gene expression. The first equation holds for unnormalized data; the second equation refers to data normalized by the total counts per cell (as defined below), with the normalized UMIFM counts denoted as $\hat{m}$. The normalization procedure reduces technical noise in the efficiency β, but it undesirably inflates the CV estimates for each gene by the cell-to-cell variability in total mRNA

content $N = \sum_i n_i$, which may reflect fluctuations in cell size or cell cycle stage. Eq. (S4) is accurate for genes whose transcript abundances are independent of the total number of transcripts in a cell, N, an assumption that is almost certainly incorrect for genes that correlate strongly with the cell cycle. In this section we drop the subscript i from all equations since they apply to genes individually rather than jointly.

No normalization:

$$\mathrm{CV}_m^2 - \frac{1}{E[m]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1 + \mathrm{CV}_\beta^2\right) + \mathrm{CV}_\beta^2 \tag{S3}$$

Total count normalization:

$$\mathrm{CV}_{\hat{m}}^2 - (1 + \mathrm{CV}_M^2)(1 + \mathrm{CV}_{1/N}^2)\frac{1}{E[\hat{m}]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1 + \mathrm{CV}_{1/N}^2\right) + \mathrm{CV}_{1/N}^2 \tag{S4}$$

Technical noise is represented in both Eqs. (S3) and (S4) by variability CV_β in the sampling efficiency of the method. Eq. (S4) includes, as we would expect, variability CV_M across cells or control droplets in the total number of UMIFM counts, $M = \sum_i m_i$. Note that M and CV_M are empirical quantities that can be calculated directly from the data. Eq. (S4) also captures variability in the total number N of mRNA transcripts originally present in those cells, in the form of $\mathrm{CV}_{1/N}$.

We begin the derivation of Eqs. (S3) and (S4) the same way. Both equations follow from the Laws of Total Expectation and Total Variance applied to the conditional means and variances of (normalized) read counts m ($\hat{m}$), conditioned on the actual number of transcripts n and the sampling efficiency β. For Binomial sampling, these conditional moments are as follows:

$$\begin{aligned} E[m|n,\beta] &= \beta n \\ \mathsf{Var}(m|n,\beta) &= \beta(1-\beta)n. \end{aligned}$$

We now calculate the unconditional moments $E[m]$ and $\mathsf{Var}(m)$ in terms of these conditional ones using the Laws of Total Expectation and Total Variance:

$$E[m] = E_{n,\beta}[E[m|n,\beta]], \tag{S5}$$

$$\mathsf{Var}(m) = E_{n,\beta}[\mathsf{Var}(m|n,\beta)] + \mathsf{Var}_{n,\beta}(E[m|n,\beta]), \tag{S6}$$

where $E_{n,\beta}[g(n,\beta)] = E_n[E_\beta[g(n,\beta)]]$ is the expected value of a function $g(n,\beta)$ over the distributions of n and β, and $Var_{n,\beta}(g) = E_{n,\beta}[g^2(n,\beta)] - E_{n,\beta}^2[g(n,\beta)]$. We obtain:

$$\begin{aligned} E[m] &= E[\beta]E[n] \\ \mathsf{Var}(m) &= E[\beta]E[n] - E[\beta^2]E[n] + E[\beta^2]E[n^2] - E[\beta]^2E[n]^2 \end{aligned}$$

We arrive at Eq. (S3) by evaluating $\mathrm{CV}_m^2 = \mathsf{Var}(m)/E[m]^2$ and simplifying the result using the identity $\frac{E[\beta^2]}{E[\beta]^2} = 1 + \mathrm{CV}_\beta^2$.

Next we turn to total count normalization [Eq. (S4)]. In total count normalization, normalized read counts $\hat{m}$ are calculated in each droplet as follows:

$$\hat{m} \equiv m\frac{E[M]}{M}$$

where $M = \sum_i m_i$ is the total number of UMIFM reads (i.e., counts or library size) for a given cell, and $E[M]$ is the average of those totals across all cells. Because each transcript count m_i is binomially distributed conditional on β and n_i, M is binomially distributed conditional on β and $N = \sum_i n_i$. Since N is large, we say that M in each individual droplet is approximately its conditional expectation $E[M|\beta, N] = \beta N$. With this approximation, and taking β and N to be independent,

$$\hat{m} = m\frac{E[\beta]}{\beta}\frac{E[N]}{N} = m\frac{E[\beta]}{\beta}R \tag{S7}$$

To simplify subsequent algebra we define the random variable $R \equiv E[N]/N$. Proceeding as before, the conditional moments for $\hat{m}$ are

$$\begin{aligned}
E[\hat{m}|n, \beta, R] &= \left(\frac{E[\beta]}{\beta}R\right)E[m|n, \beta] = E[\beta]Rn \\
\mathsf{Var}[\hat{m}|n, \beta, R] &= \left(\frac{E[\beta]^2}{\beta^2}R^2\right)\mathsf{Var}(m|n, \beta) = E[\beta]^2(\beta^{-1} - 1)R^2 n.
\end{aligned}$$

The quantity $\mathsf{Var}[\hat{m}|n, \beta, R]$ depends on β^{-1}. Using Equation (S6) we conclude that the unconditional variance will depend on the inverse moment $E\left[\beta^{-1}\right]$, which we can approximate using a power series expansion,

$$\begin{aligned}
E\left[\beta^{-1}\right] &= \frac{1}{E[\beta]}E\left[\frac{1}{1 + (\beta - E[\beta])/E[\beta]}\right] \\
&= \frac{1}{E[\beta]}E\left[\sum_{k=0}^{\infty}(-1)^k\frac{(\beta - E[\beta])^k}{E[\beta]}\right] \\
&= \frac{1}{E[\beta]}\left(1 + \mathrm{CV}_\beta^2\right) + O\left(\frac{E[(\beta - E[\beta])^3]}{E[\beta]^3}\right).
\end{aligned}$$

In the final line above, we note that the quadratic term in the power series is CV_β^2. The higher order terms depend on the third and higher mean-normalized central moments of β, which we can safely ignore if the noise in β is small. Armed with this approximation for $E\left[\beta^{-1}\right]$, we find that

$$E[\hat{m}] = E[\beta]E[R]E[n] \tag{S8}$$

$$\begin{aligned}
\mathsf{Var}(\hat{m}) \approx\ & E[\beta]E[R^2]E[n]\left(1 + \mathrm{CV}_\beta^2\right) - E[\beta]^2E[R^2]E[n] \\
& + E[\beta]^2E[R^2]E[n^2] - E[\beta]^2E[R]^2E[n]^2.
\end{aligned} \tag{S9}$$

Now dividing Eq. (S9) by the square of Eq. (S8) gives

$$\mathrm{CV}_{\hat{m}}^2 - \frac{E[R](1 + \mathrm{CV}_R^2)(1 + \mathrm{CV}_\beta^2)}{E[\hat{m}]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1 + \mathrm{CV}_R^2\right) + \mathrm{CV}_R^2$$

To recover Eq. (S4), we make use of the following consequences of the equalities $R = E[N]/N$ and $M = \beta N$. First, we note that $\mathrm{CV}_R^2 = \mathrm{CV}_{1/N}^2$. Second, we can repurpose our power series expansion above for N instead of β and see that $E[R] = E[N]E[N^{-1}] \approx 1 + \mathrm{CV}_N^2$. Finally, since β and N are independent, we can say that $(1 + \mathrm{CV}_N^2)(1 + \mathrm{CV}_\beta^2) = (1 + \mathrm{CV}_{\beta N}^2) = 1 + \mathrm{CV}_M^2$.

E.2.2 Technical noise weakens observed gene-gene correlations

Technical noise may either weaken pairwise correlations between genes, or spuriously generate correlations through normalization. If two genes are sampled unevenly, their relationship in the

sample may look quite different from their relationship in the original pool. Moreover, correlation is sensitive to scale – two low-abundance genes are much more likely to seem uncorrelated than two highly abundant genes. The equation we develop here helps us understand more formally how sampling and noise in sampling weaken correlations that we observe between genes through their UMIFM read counts $\mathrm{corr}(\hat{m}_i, \hat{m}_j)$. Here we consider only the case of total count normalization. We begin with the definition of the correlation coefficient,

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \frac{\mathrm{Cov}(\hat{m}_i, \hat{m}_j)}{\sqrt{\mathsf{Var}(\hat{m}_i)\mathsf{Var}(\hat{m}_j)}},$$

and rewrite this expression in terms of CVs:

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \frac{\mathrm{Cov}(\hat{m}_i, \hat{m}_j)}{E[\hat{m}_i]E[\hat{m}_j]} \frac{1}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}} = \frac{\widetilde{C}(\hat{m}_i, \hat{m}_j)}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}},$$

where $\widetilde{C}$ is the normalized covariance. The connection between $\mathrm{corr}(\hat{m}_i, \hat{m}_j)$ and $\mathrm{corr}(n_i, n_j)$ becomes apparent once we realize that

$$\widetilde{C}(\hat{m}_i, \hat{m}_j) = (1 + CV_{1/N}^2)\widetilde{C}(n_i, n_j) + CV_{1/N}^2\,, \tag{S10}$$

which follows from the fact that $E[\hat{m}_i\hat{m}_j] = E[\beta]^2 E[n_i n_j]E[R^2]$. We are reminded that normalization by a noisy quantity (in this case $1/N$) can spuriously inflate positive covariances, and eliminate weak negative covariances (or inflate them if $\widetilde{C}(n_i, n_j) < -1$). From Eq. (S10) it follows that

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \mathrm{corr}(n_i, n_j)\frac{\mathrm{CV}_{n_i}\mathrm{CV}_{n_j}}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}}(1 + CV_{1/N}^2) + \frac{CV_{1/N}^2}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}}. \tag{S11}$$

To develop an intuition for the effects of sampling on gene-gene correlations, we assume that the variability between droplets in total counts $\mathrm{CV}_{1/N}$ is small, as is the case for undifferentiated ES cells. Then, using Eq. (S4) to relate $\mathrm{CV}_{n_{i,j}}$ to $\mathrm{CV}_{\hat{m}_{i,j}}$, Eq. (S11) becomes,

$$\begin{aligned} \mathrm{corr}(\hat{m}_i, \hat{m}_j) &= \mathrm{corr}(n_i, n_j)\alpha_i\alpha_j\,, \\ \alpha_{k\in\{i,j\}} &= \sqrt{\left(1 - \frac{1 + \mathrm{CV}_M^2 - E[\beta]}{F_{\hat{m}_k}}\right)} \end{aligned} \tag{S12}$$

where $F_{\hat{m}} = \mathsf{Var}(\hat{m})/E[\hat{m}]$ is the expected value of the *observed* gene Fano factor. To obtain Eq. (S12), we make use of the relationship

$$\frac{\mathrm{CV}_n}{\mathrm{CV}_{\hat{m}}} = \sqrt{\frac{\mathsf{Var}(n)}{E[n]}\frac{E[\hat{m}]}{\mathsf{Var}(\hat{m})}\frac{E[\hat{m}]}{E[n]}} = \sqrt{\frac{F_n}{F_{\hat{m}}}E[\beta]E[R]},$$

and then relate F_n and $F_{\hat{m}}$ by multiplying Eq. (S4) by $E[\hat{m}]$.

Note that the degree to which technical noise dampens the correlation between genes i and j is sensitive to the mean expression levels of both genes and to the sampling efficiency through the Fano factors. Since sampling efficiency is low, $E[\beta] \ll 1$, and we can approximate Eq. (S12) in terms of observable quantities only as

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) \approx \mathrm{corr}(n_i, n_j)\sqrt{\left[1 - F_{\hat{m}_i}^{-1}(1 + \mathrm{CV}_M^2)\right]\left[1 - F_{\hat{m}_j}^{-1}(1 + \mathrm{CV}_M^2)\right]},$$

giving Eq. (3) in Fig. 2G of the main text.

E.3 Identifying highly variable genes

A key goal of our data analysis is to identify genes whose expression in a population of cells is highly variable. More precisely, we wish to identify genes whose abundances are significantly over-dispersed relative to a Poisson distribution, which would result from uniform, non-fluctuating expression of transcripts in all cells. In this analysis, we use a test statistic that, at any given mean gene expression level, gives more weight to genes whose CV is many times larger than that of a Poisson random variable with the same mean. Based on Eq. (S4), a reasonable proposal for a test statistic, v, is:

$$v = \frac{\mathrm{CV}^2_{\hat{m}}}{\left(1+\mathrm{CV}^2_M\right)\left(1+\mathrm{CV}^2_{1/N}\right)\Big/ E[\hat{m}]+\mathrm{CV}^2_{1/N}} \tag{S13}$$

By defining v in this way, we make concrete precisely what we do when we identify outliers by eye on a plot of genes' CV versus mean abundance such as Fig. 2F. The additive constant noise term $\mathrm{CV}^2_{1/N}$ keeps us from identifying a gene as highly variable in a population of cells if we can attribute much of that variability to differences in cell size. We infer $\mathrm{CV}^2_{1/N}$ from the data; for the ES cell data, $\mathrm{CV}_{1/N}$ ranges from $\sim 20\%$ on Day 0 to $\sim 35\%$ on Day 7 post-Lif withdrawal. For our RNA controls $\mathrm{CV}_{1/N}$ is much smaller – typically on the order of 5%, consistent with $\mathrm{CV}_{1/N}$ describing variability in total mRNA content per droplet. For both cells and RNA controls we calculate CV_M directly from the data; its values for the ES cell data are given in Table S1. The test statistic proposed here is similar to that proposed previously in (Brennecke et al., 2013), but with two key differences. First, here there is just one parameter to be inferred from the data $CV^2_{1/N}$, not two; second, we tested and found that the empirical distribution of v is not a χ^2 distribution as proposed in that study.

To develop a test for variability, we need a null distribution that describes the possible spread in v given that a gene's counts across cells are actually Poisson-distributed. For this purpose one may calculate v for a set of pure RNA controls, allowing for different values of $\mathrm{CV}_{1/N}$ and CV_M in each sample. One can then compute a p-value for each ES cell gene by comparing its v-score to the reference distribution, and thus test how many genes are significantly variable using Benjamini and Hochberg's method to control the false discovery rate (FDR).

F. Supplemental References

1. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20(0):53-65.
2. Jaitin DA*, et al.* (2014) Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* 343(6172):776-779.
3. Mazutis L*, et al.* (2013) Single-cell analysis and sorting using droplet-based microfluidics. *Nat Protoc* 8(5):870-891.
4. Holmes DL & Stellwagen NC (1991) Estimation of Polyacrylamide-Gel Pore-Size from Ferguson Plots of Normal and Anomalously Migrating DNA Fragments .1. Gels Containing 3-Percent N,N'-Methylenebisacrylamide. *Electrophoresis* 12(4):253-263.
5. Tse JR & Engler AJ (2010) Preparation of hydrogel substrates with tunable mechanical properties. *Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.]* Chapter 10:Unit 10 16.
6. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
7. Lawler EL (1966) Covering Problems: Duality Relations and a New Method of Solution. *SIAM Journal on Applied Mathematics* 14(5):1115-1132.
8. Chvatal V (1979) A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* 4(3):233-235.
9. Xue K, Ng JH, & Ng HH (2011) Mapping the networks for pluripotency. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 366(1575):2238-2246.