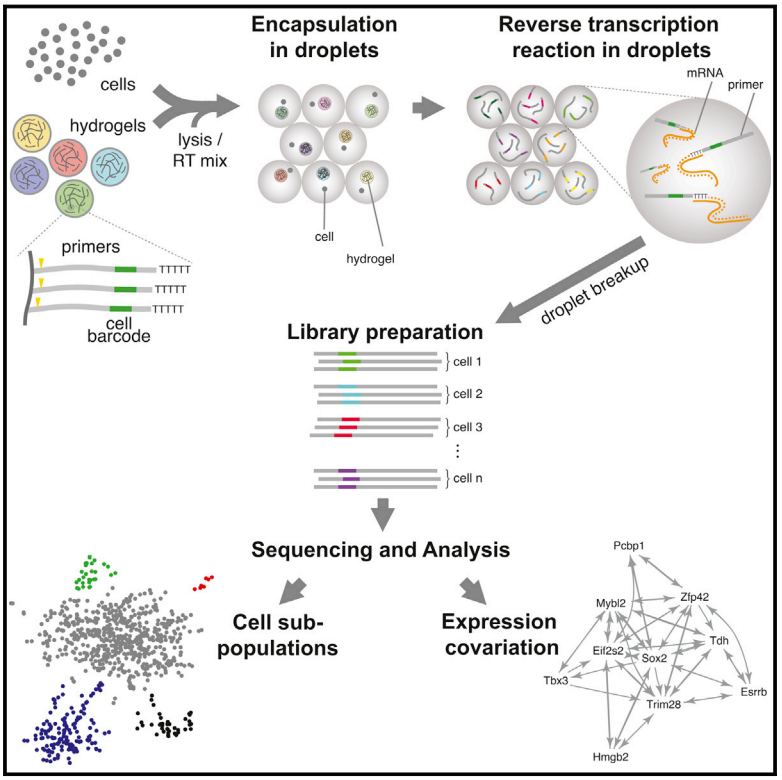
Resource

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

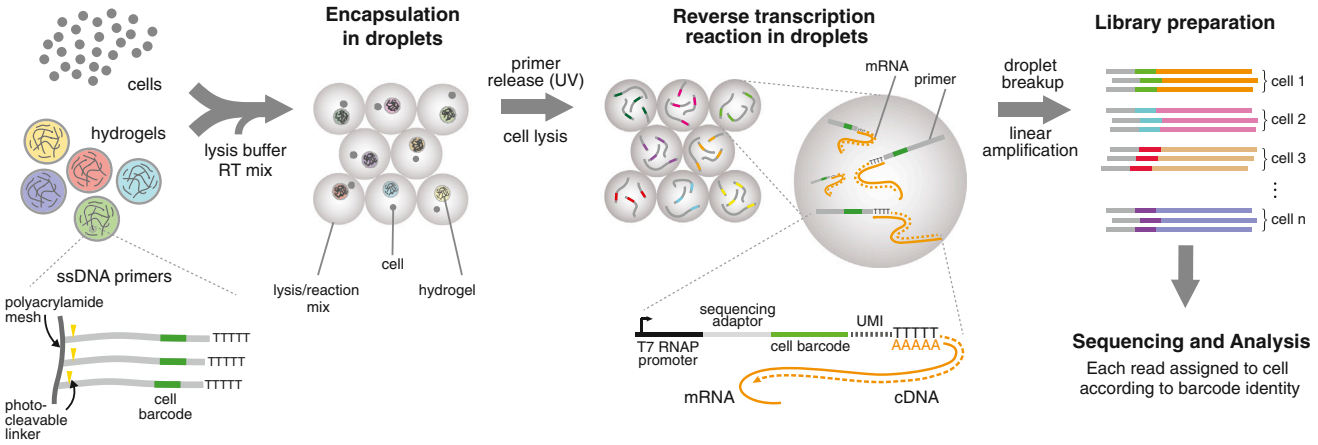


Figure 1. A Platform for DNA Barcoding Thousands of Cells

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

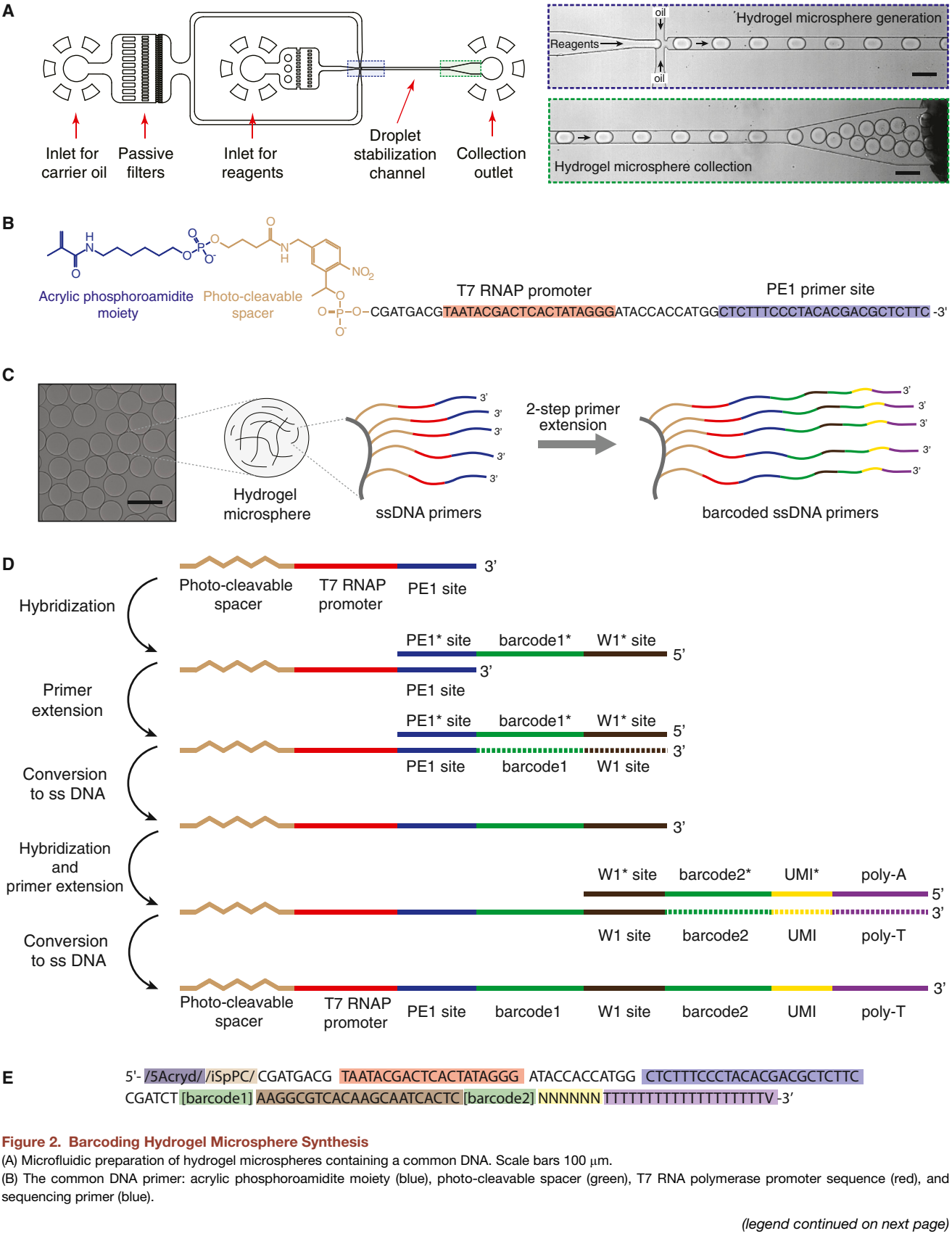


Figure 2. Barcoding Hydrogel Microsphere Synthesis

(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)

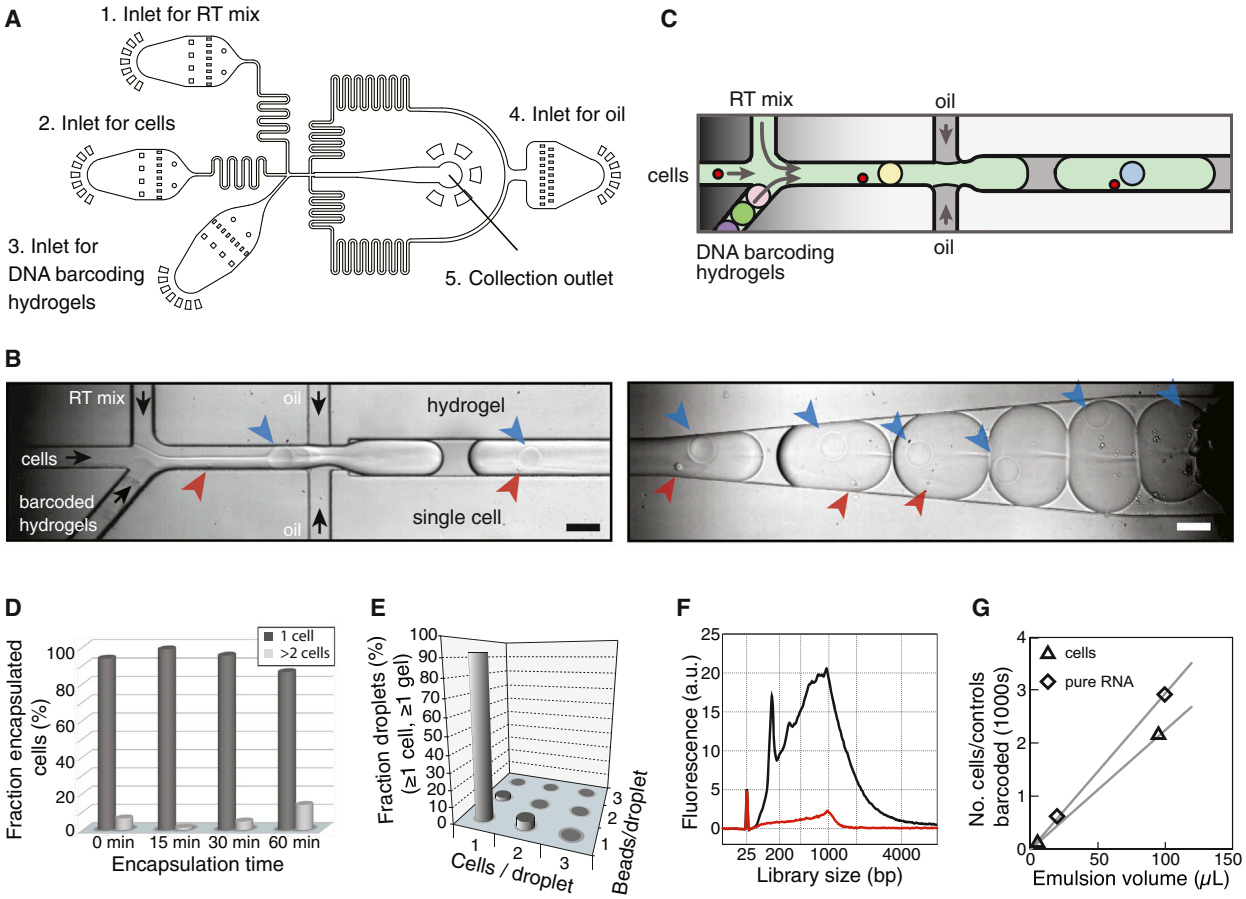


Figure 3. A Droplet Barcoding Device

(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.

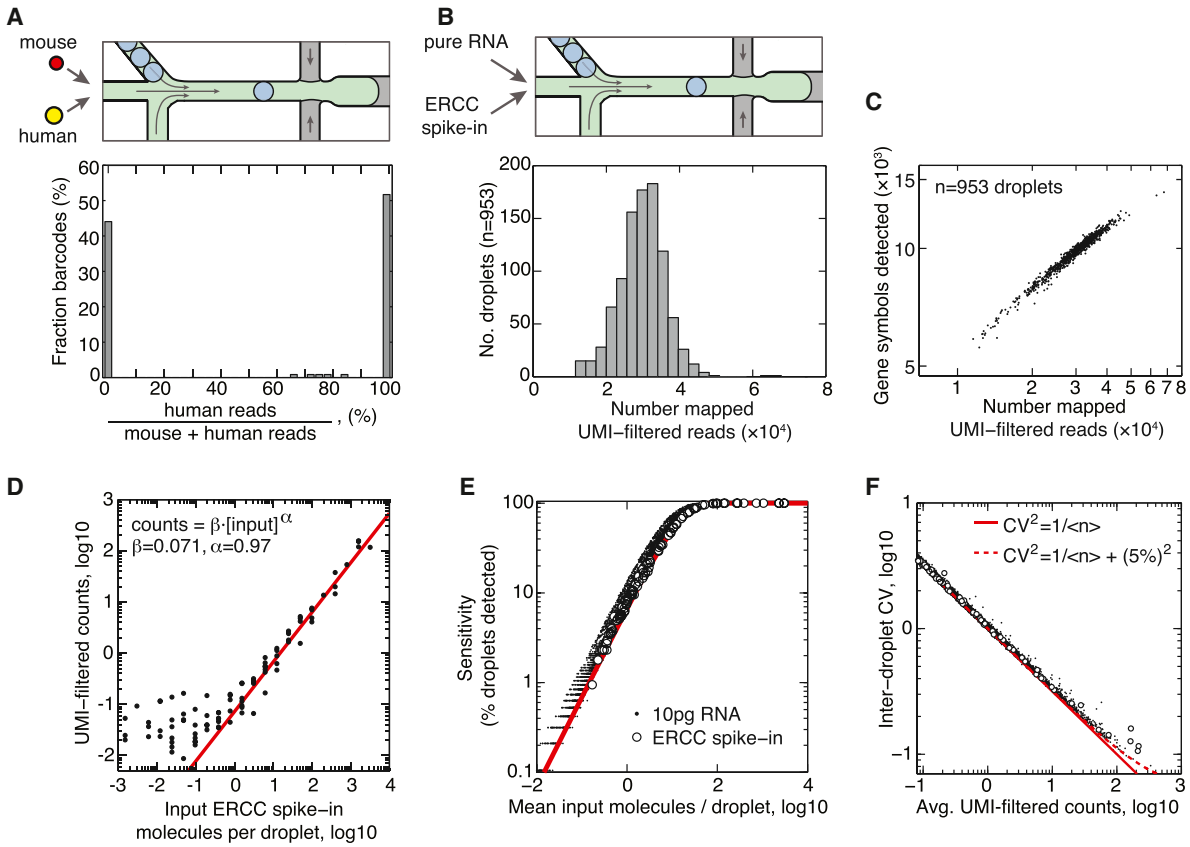
Single-Cell Profiling of Mouse ES Cells

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: $\text{corr}_{i,j} \simeq \text{corr}^{(\text{bio})}_{i,j} \sqrt{(1 - F_i^{-1})(1 - F_j^{-1})}$

observed | biological | 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

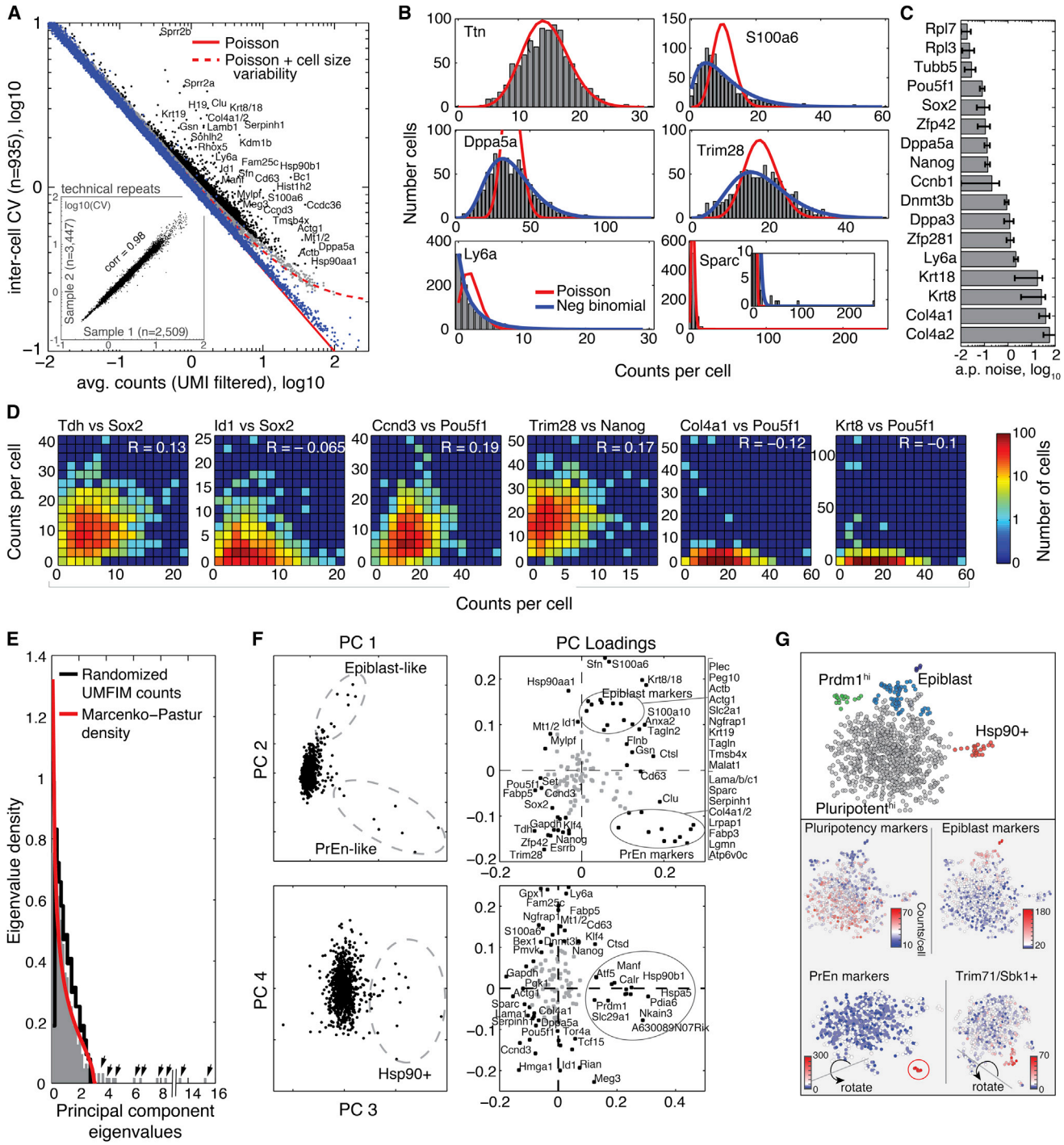


Figure 5. inDrop Sequencing Reveals ES Cell Population Structure

(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).

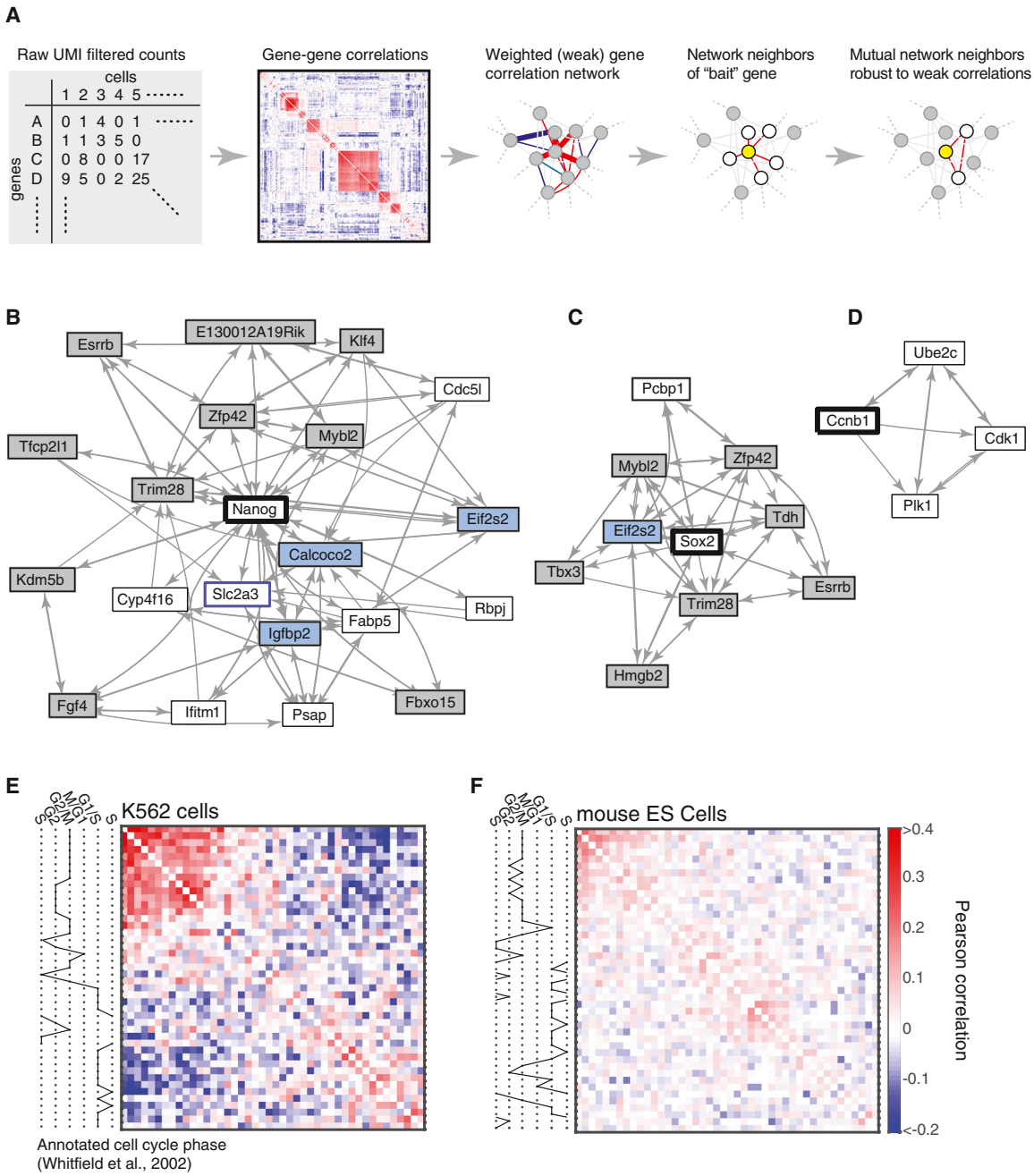


Figure 6. Regulatory Information Preserved in Gene Correlations

(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).

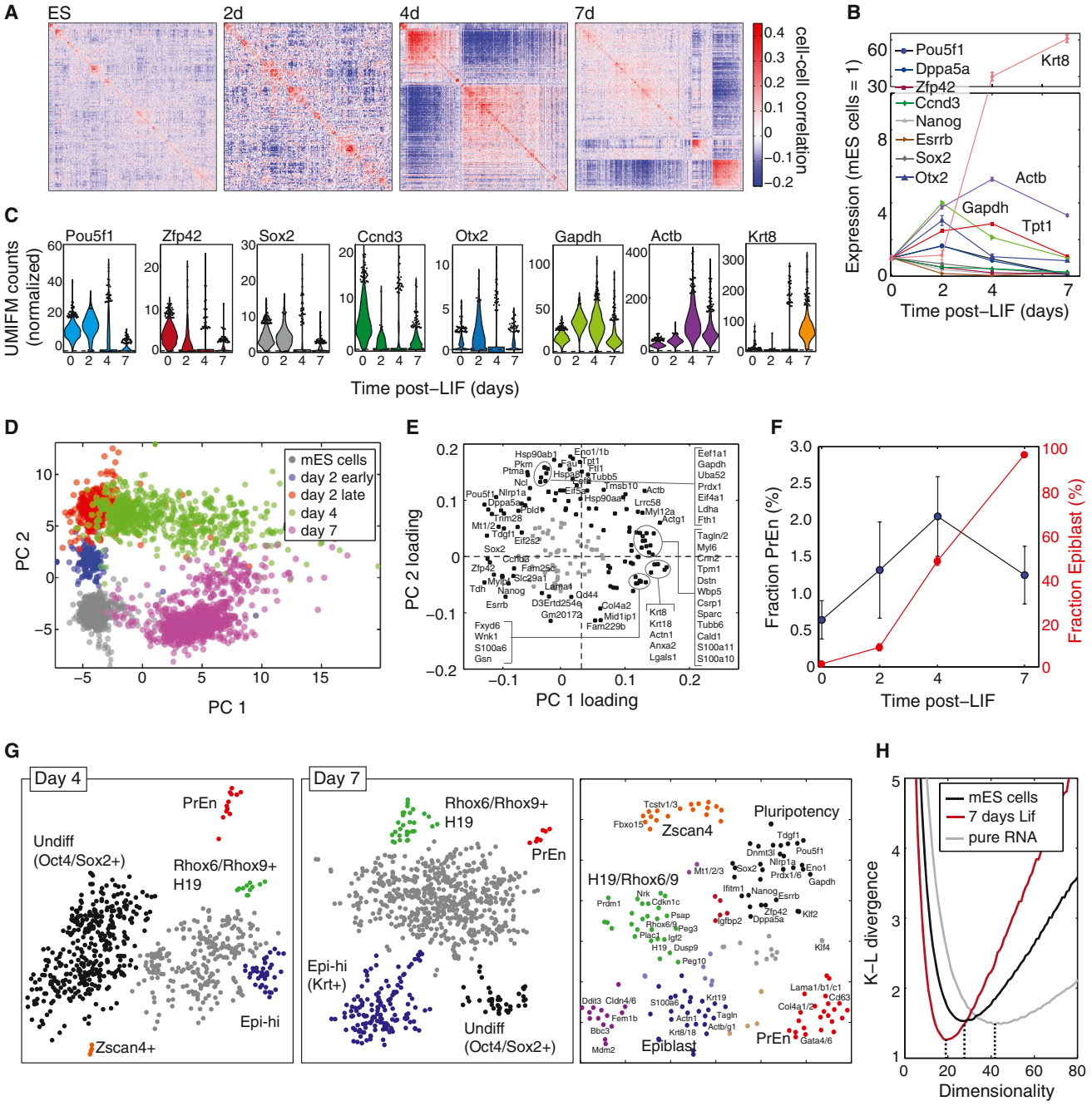


Figure 7. Heterogeneity in Differentiating ES Cells

(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.

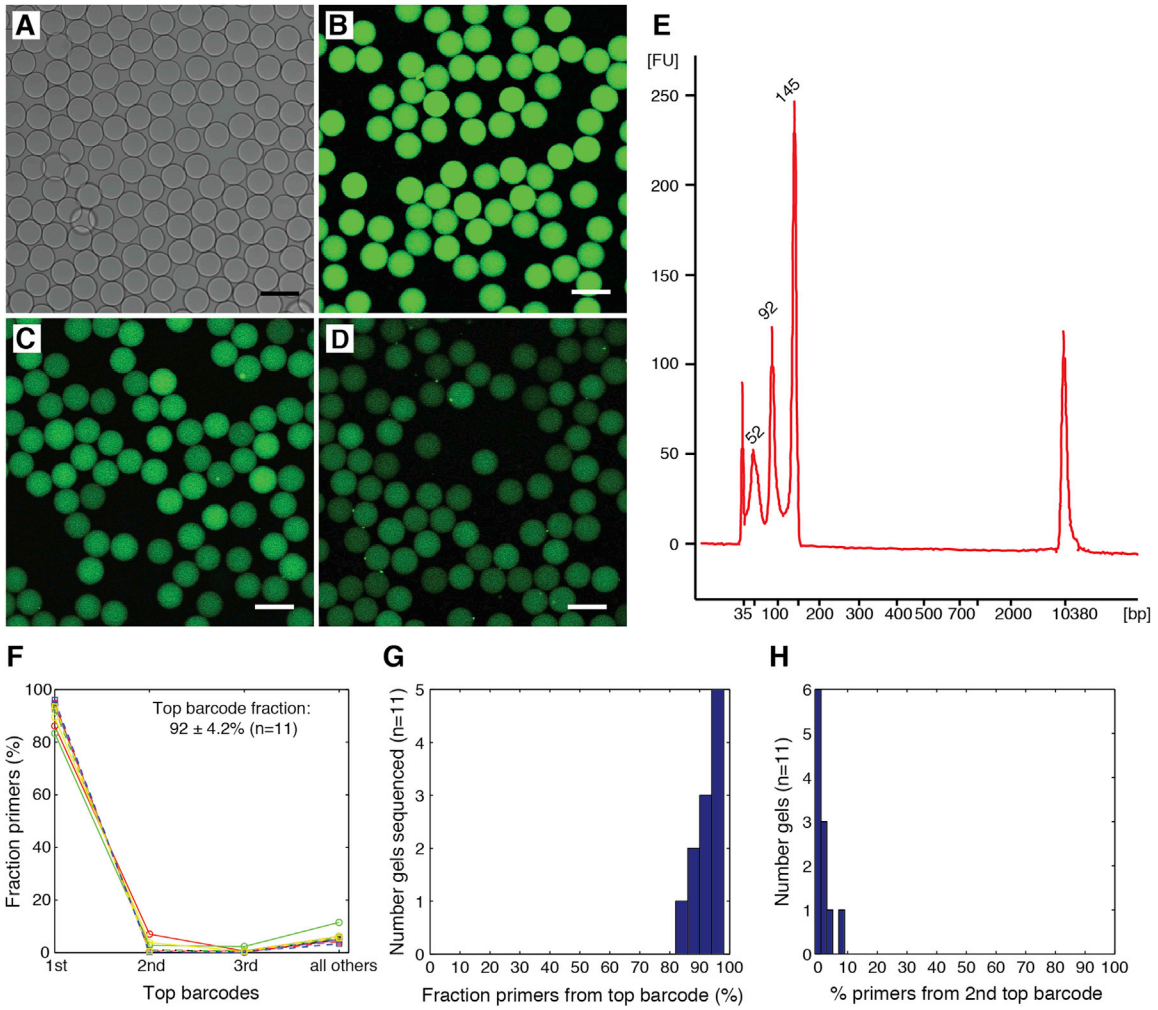


Figure S1. Quantification of DNA Primers Incorporated into Barcoded Hydrogel Microspheres, Related to Figure 2

(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.

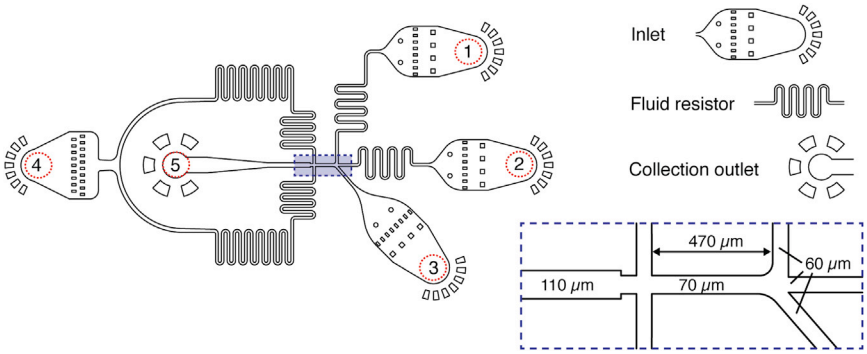
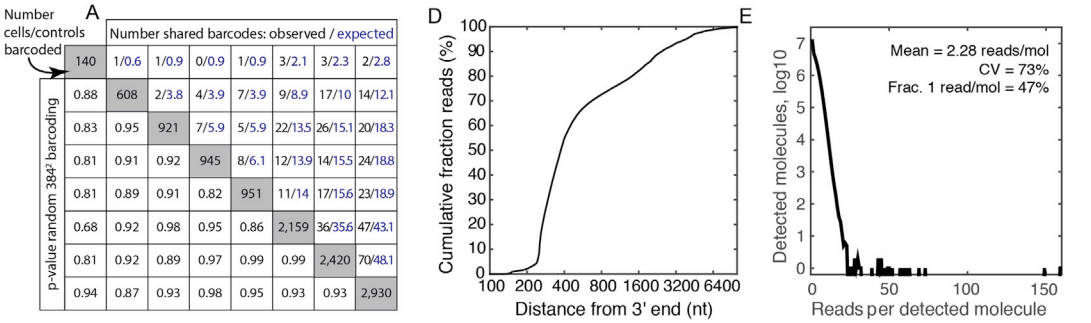


Figure S2. Dimensions of Droplet Microfluidics Device for Cell-Hydrogel-RT/Lysis Mix Co-encapsulation, Related to Figure 3

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.

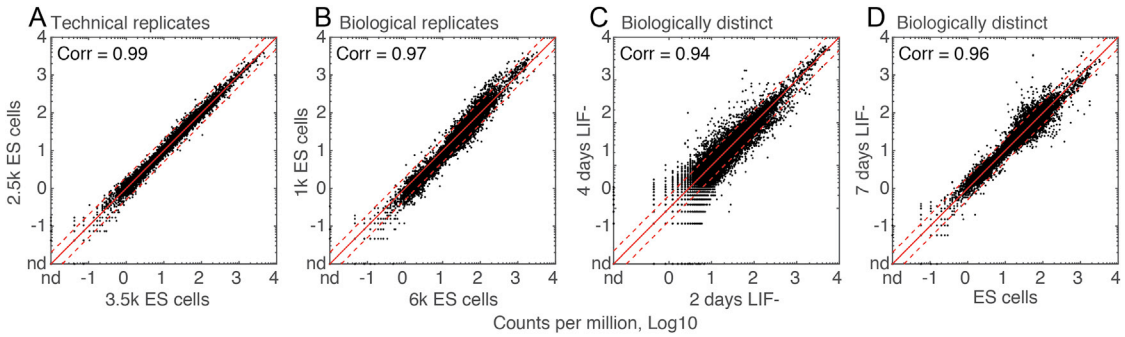


Figure S4. Technical Reproducibility of inDrop Single-Cell Sequencing, Related to Figure 5

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.

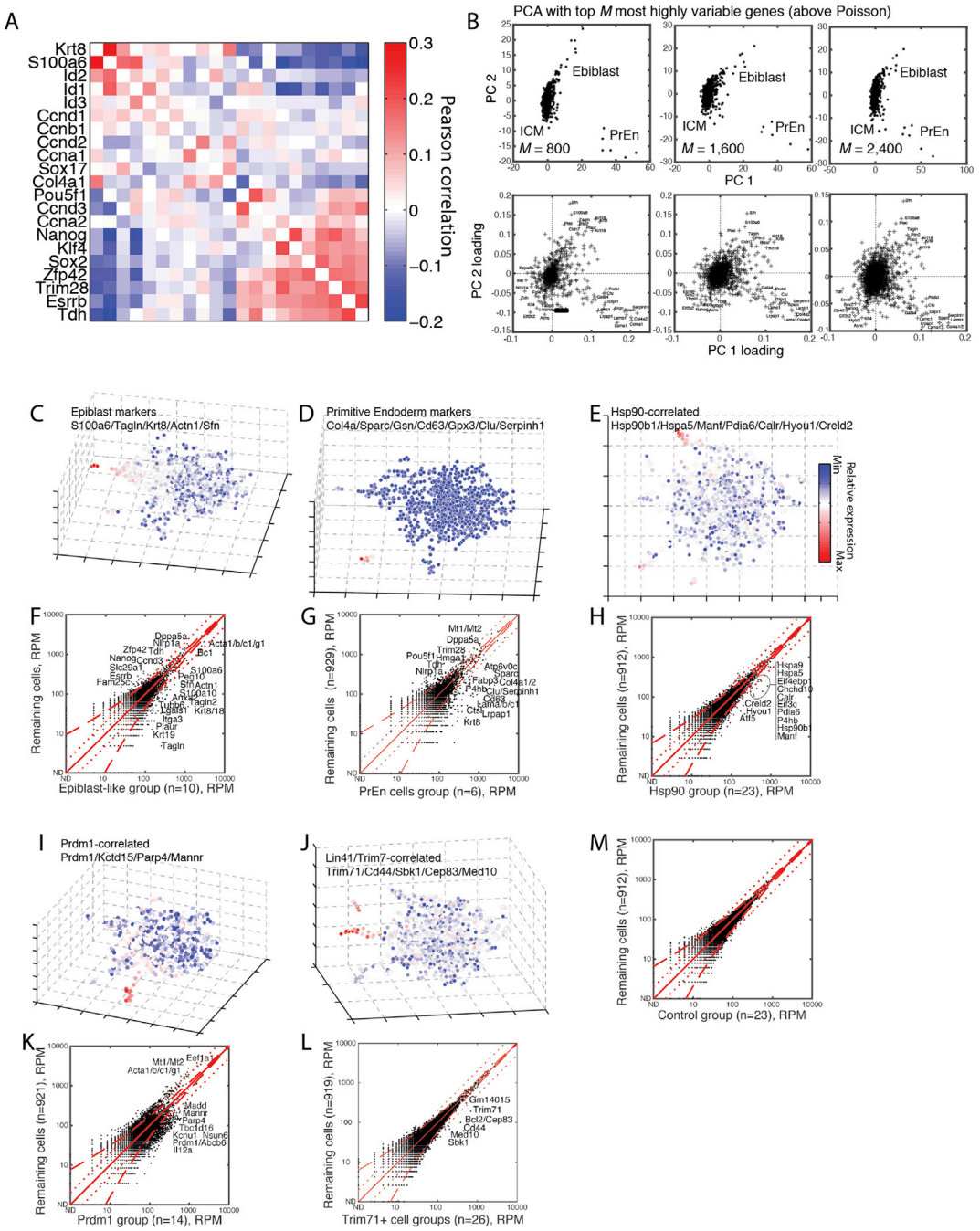
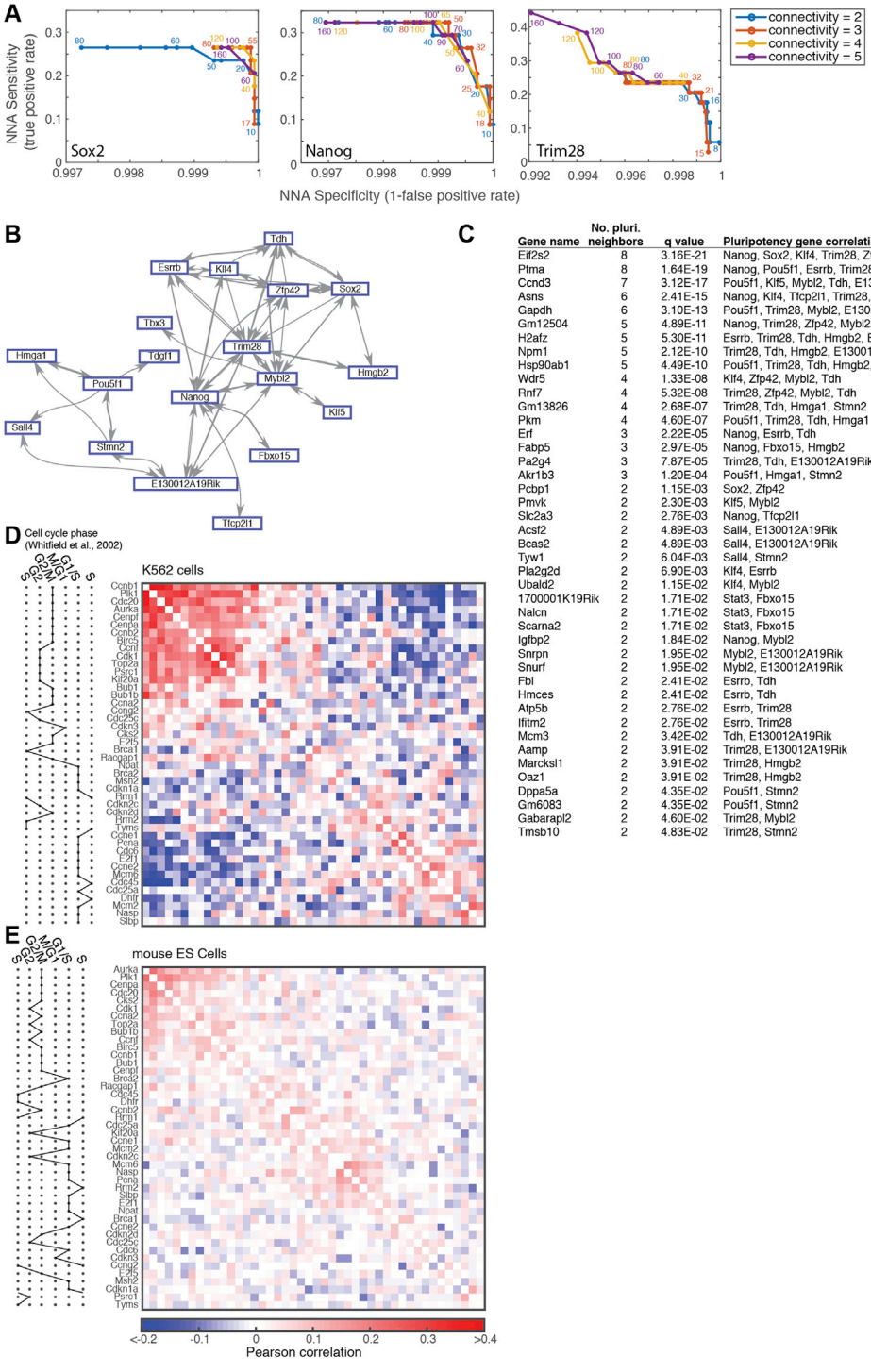


Figure S5. Structure of the mES Cell Population, Related to Figure 5

(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.



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.

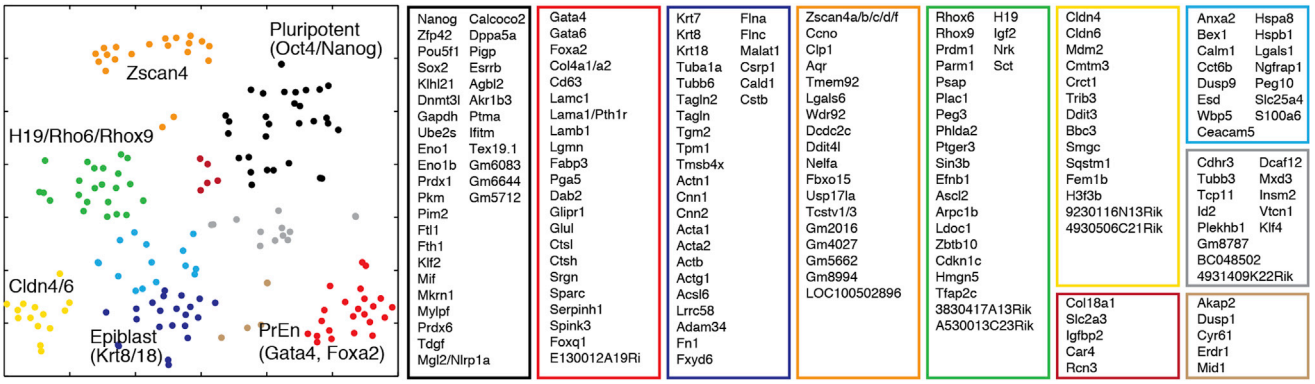


Figure S7. tSNE Map of Principal Genes at 4 Days Post-LIF Withdrawal, Related to Figure 7

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