Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq

Saiful Islam,[1,4] Una Kjällquist,[1,4] Annalena Moliner,[2] Pawel Zajac,[1] Jian-Bing Fan,[3] Peter Lönnerberg,[1] and Sten Linnarsson[1,5]

[1]*Laboratory for Molecular Neurobiology, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77 Stockholm, Sweden;* [2]*Department of Neuroscience, Karolinska Institutet, SE-171 77 Stockholm, Sweden;* [3]*Illumina Inc., San Diego, California 92121, USA*

Our understanding of the development and maintenance of tissues has been greatly aided by large-scale gene expression analysis. However, tissues are invariably complex, and expression analysis of a tissue confounds the true expression patterns of its constituent cell types. Here we describe a novel strategy to access such complex samples. Single-cell RNA-seq expression profiles were generated, and clustered to form a two-dimensional cell map onto which expression data were projected. The resulting cell map integrates three levels of organization: the whole population of cells, the functionally distinct subpopulations it contains, and the single cells themselves—all without need for known markers to classify cell types. The feasibility of the strategy was demonstrated by analyzing the transcriptomes of 85 single cells of two distinct types. We believe this strategy will enable the unbiased discovery and analysis of naturally occurring cell types during development, adult physiology, and disease.

[Supplemental material is available for this article. The microarray data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE29087.]

Comprehensive gene expression profiling was first made practical by microarrays, which enabled the study of thousands of genes in tens of samples. Microarrays have two major shortcomings: They are limited to known genes, and they have limited sensitivity and dynamic range. RNA sequencing (RNA-seq) overcomes these problems by sequencing RNA directly (Ozsolak et al. 2009) or after reverse-transcription to cDNA (Cloonan et al. 2008; Mortazavi et al. 2008; Wang et al. 2008). Quantitation is based simply on hit counts, with great sensitivity and nearly unlimited dynamic range.

Tissues are rarely homogeneous, however, and therefore any expression profile based on a tissue sample will blend the true expression profiles of its constituent cells. One way of getting around this problem would be to analyze single cells instead of cell populations, and indeed, single-cell methods have been developed for both microarrays (Kurimoto et al. 2006; Esumi et al. 2008) and, recently, RNA-seq (Tang et al. 2009). These methods are suitable for the analysis of small numbers of single cells and, in particular, may be used to study cells that are difficult to obtain in large numbers, such as oocytes and the cells of the early embryo.

However, single-cell transcriptomics must confront two great challenges. First, markers suitable for the prospective isolation of defined cell populations are not available for every cell type, reflecting the fact that few cell types are clearly defined in molecular terms. Second, transcript abundances vary greatly from cell to cell. For example, beta actin (*Actb*) mRNA content varies more than three orders of magnitude between pancreatic islets cells (Bengtsson et al. 2005). Similar results have been reported for RNA polymerase II (Raj et al. 2006), human *GAPDH* (Warren et al. 2006; Lagunavicius et al. 2009), *SPI1* (also known as *PU.1*) (Warren et al. 2006), and *TBP*, *B2M*, *SDHA*, and *EEF1G* mRNAs (Taniguchi et al. 2009) and at present seems to be a common feature of the transcriptome. Most of the variation may be intrinsic, caused by burst-like stochastic activation of transcription, where brief episodes of mRNA synthesis lasting a few minutes are separated by periods of transcriptional silence of similar duration (Chubb et al. 2006). As a consequence, a random sample of cells would show great variation in their content of particular mRNAs, ranging from those cells that have just undergone a burst, to those that have nearly completely degraded their mRNA; this has been directly observed for RNA polymerase II transcription in situ using a fluorescent probe targeting the 52-copy repeat in that gene (Raj et al. 2006).

Recently, the power of single-cell analysis for unbiased cell-type classification was demonstrated in an experiment based on single-cell Q-PCR (Guo et al. 2010). By sampling not just a few, but large numbers of, single cells and by focusing in particular on transcription factors known to be relevant, the investigators were able to correctly classify the three cell types known to be present in the mouse blastocyst. However, since Q-PCR is limited to small numbers of genes, it is not feasible to survey, for example, the entire set of transcription factors. Thus there is a need for a method to access the entire transcriptomes of large numbers of single cells.

Here we describe single-cell tagged reverse transcription (STRT), a highly multiplexed method for single-cell RNA-seq on the Illumina platform. We prepare barcoded cDNA libraries from 96 single cells and analyze them by sequencing. From each transcript, a single read is obtained, corresponding to a template-switching site located preferentially at the 5′ end of the mRNA. We then use similarity of expression patterns to build an in silico map of cells and how they are related. This way, both single-cell detail and cell type–specific population averages are available and can be studied without the mixing of data from unrelated cell types. Importantly, both known and novel factors specifically expressed in a cell type can be analyzed, since the resulting data set comprises the entire transcriptome of each cell.

[4]These authors contributed equally to this work.
[5]Corresponding author.
E-mail sten.linnarsson@ki.se.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.110882.110.

21:1160–1167 © 2011 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/11; www.genome.org

Results

Single-cell tagged reverse transcription

In brief, each sample was prepared by picking single cells into the wells of a 96-well PCR plate preloaded with lysis buffer and then by adding reverse transcription reagents to generate a first-strand cDNA. Eight synthetic mRNAs were added to each well as internal controls. To incorporate a well-specific (and hence cell-specific) barcode, we exploited the reverse transcriptase template-switching mechanism (Schmidt and Mueller 1999) whereby a helper oligo directs the incorporation of a specific sequence at the 3′ end of the cDNA molecule (Fig. 1A). A different helper oligo was used in each well, with distinct six-base barcodes and a universal primer sequence. After cDNA synthesis, the 96 reactions were pooled, purified, and amplified by single-primer PCR in a single tube. Cell-to-cell amplification bias was thus reduced, and the number of PCR cycles could be kept low. The amplified samples were then adapted for Illumina sequencing. We named the procedure STRT. For details, see Methods.

Here we report data from 92 single cells collected from two different mouse cell types: embryonic stem cells (ES R1) (Wood et al. 1993) and embryonic fibroblasts (MEFs; as a control, we separately prepared 96 wells with 10 pg per well of a human brain reference RNA, henceforth called RefRNA). We obtained 110 million raw reads on five sequencing lanes on an Illumina Genome Analyzer IIx. Reads lacking a proper barcode, mostly caused by errors in sample preparation or sequencing, were removed. Of the remaining 82 million reads, 80% could be placed on the mouse genome allowing for up to two sequencing errors, resulting in hits to 13,879 annotated genes and 940 repeat families. The number of mapped reads was reduced by 99% in negative control wells, confirming that observed signals originated from bona fide cDNA synthesis in positive wells. The remaining misassigned reads may have been generated in part by sequencing errors, or oligonucleotide synthesis errors, in the barcode. Mapped reads were then classified as illustrated in Figure 1B.

The background from, for example, genomic DNA contamination or unspliced pre-mRNA, judged by hits to intronic sequence, was low (0.1 reads per million per kilobase [RPKM]), as clearly seen in Figure 2A and Supplemental Figure 1 (note the paucity of reads on reverse strand and in introns). To quantify this, we determined the number of hits to the exons, introns, 1000-bp flanking regions, and splice junctions of each gene (Fig. 2C). Fifty-five percent of all reads mapped to exons, and 8% mapped to splice junctions. The remaining reads mapped to introns (9%), upstream/downstream regions (1% each), and repeats (10%). Since introns and repeats span much more of the genome than do exons, we then normalized for the total feature length and total number of reads (Fig. 2D). We found 9.5 RPKM in exonic sequence but only 0.1 RPKM in introns, indicating great specificity for expressed mRNA and rejection of genomic DNA and unspliced pre-mRNA. Reads aligning to repeats were dominated by the B2 family of short interspersed elements, which are known to generate polyadenylated transcripts (Borodulina and Kramerov 2008) and are abundantly expressed in many tissues including the early embryo (Taylor and Piko 1987).

Strand information is often required to properly assign reads to transcriptional units, since genes frequently overlap on opposite strands and since the 5′ and 3′ UTRs are often incorrectly annotated. STRT preserved strand information, as shown by the ratio of sense to antisense reads on exons (115-fold) and splice junctions (293-fold). We found an elevated density of reads upstream of and downstream from genes (Fig. 2D), suggesting the frequent presence of neighboring genes in these regions, and we found evidence of 505 pairs of expressed genes with exons overlapping in opposite orientation.

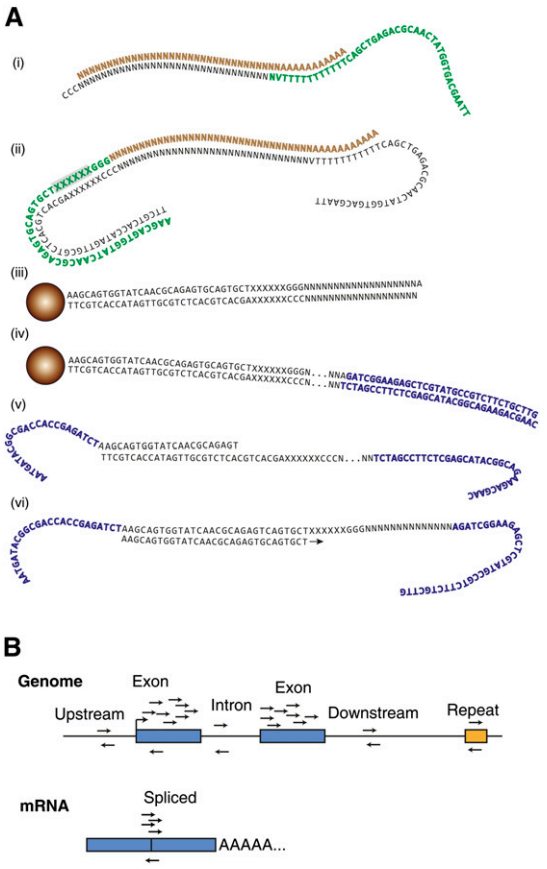

Figure 1. Single-cell tagged reverse transcription (STRT). (*A*) Overview of the method, illustrating the main steps in sample preparation: (*i*) mRNA (brown) is reverse transcribed using a tailed oligo-dT primer (green), generating a first-strand cDNA with 3-6 added cytosines; (*ii*) a helper oligo (green) causes template-switching and thereby introduces a barcode (shaded) and a primer sequence into the cDNA; (*iii*) the product is amplified by single-primer PCR exploiting the template-suppression effect and is then immobilized on beads, fragmented, and A-tailed; (*iv*) the Illumina P2 adapter (blue) is ligated to the free end; (*v*) the P1 adapter is introduced in the library PCR step, using a primer tailed with the P1 sequence (blue); and (*vi*) the final library is sequenced from the P1 side using a custom primer. Each read (arrow) begins by the barcode, followed by three to six Cs, followed by the mRNA insert. (*B*) Illustration of read mapping and annotation, for a two-exon gene. Reads mapping to the sense strand of exons, as well as to splice junctions, were counted toward the expression of the gene. Reads mapping upstream of, downstream from, or in introns were counted for quality control purposes, and antisense hits were used to judge the background level.

Assessing the length of single-cell cDNA

Initial experiments showed that heating during lysis caused partial degradation of RNA, leading to frequent hotspots of template switching. Omitting the heating step and optimizing reverse transcription resulted in a majority of full-length cDNAs. To confirm this observation, we examined the set of eight synthetic mRNAs added to each well. We found that more than 85% of all reads occurred within the first 5% of the length of the RNA, nearly all of which

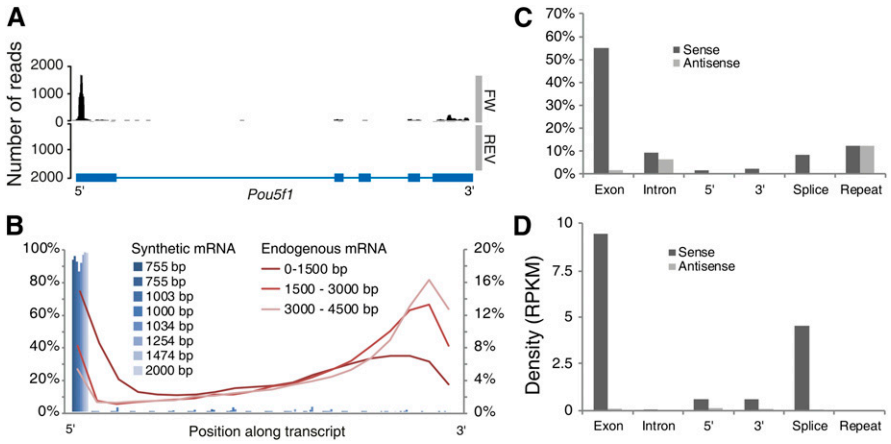

Figure 2. Read distribution. (*A*) Example of reads mapped to both strands of the 5-kb *Pou5f1* locus, shown as a coverage plot. The gene structure is shown in blue below the graph. Most reads aligned near the 5′ end of the gene. (*B*) Density of reads as a function of the position along the transcript, in 5% length bins. The figure shows eight synthetic mRNA (blue bars) and averages for all genes categorized by transcript length as indicated. (*C*) Read-mapping statistics, showing the fraction of all mapped reads that overlapped each type of annotation (cf. Fig. 1B). The vertical scale shows the percentage of all reads that mapped to exons, introns, splice junctions, 1000 bp upstream of and 1000 bp downstream from transcriptional units, and known repeats. In each case, the black bar shows reads mapped in the sense orientation, and the gray bar shows reads mapped in antisense. Repeats were not directionally annotated and therefore were hit equally on both strands. (*D*) The same statistics as in *C* but normalized for the total length of each feature class, expressed as RPKM. This shows more clearly the level of enrichment of exons versus introns, demonstrating good specificity for mRNA and rejection of genomic DNA and/or inspliced intronic RNA.

were placed within a few bases of the known 5′ end of the transcript (Fig. 2B). We conclude that our protocol had reliably identified the true 5′ end of transcripts, at least up to 2.0 kb in length (corresponding to the longest control mRNA). However, surprisingly we found that hits on endogenous mRNA showed a bimodal distribution, with a sharp 5′ spike and a broader distribution of hits approaching the 3′ end (Fig. 2B; as exemplified by *Pou5f1* in Fig. 2A and several genes in Supplemental Fig. 1). In contrast to synthetic mRNA, endogenous mRNAs of similar length showed only 16% reads at their 5′ end. In addition, longer transcripts showed a progressively larger proportion of hits to their 3′ UTR. The same effect was observed for RefRNA (data not shown). Not all genes followed the average pattern: We found instances of very long transcripts almost exclusively represented by 5′ reads, for example, the 7-kb *Malat1* noncoding RNA; yet some short genes, for example, *Actb*, apparently lacked 5′ reads (Supplemental Fig. 1). The origin of apparently truncated mRNAs warrants further investigation.

If the distribution of reads on endogenous genes were due to a failure to complete cDNA synthesis, we would expect that quantification of long transcripts may be affected. Indeed, we found a correlation between length and expression level: The most abundant genes were <2 kb on average, whereas intermediate and low-expressed genes were >2.0 kb (Supplemental Fig. 2a). However, this effect was also seen in microarray data from the same cell types (Supplemental Fig. 2b) and may be explained by natural selection for compact genes among abundant and universally expressed genes (Eisenberg and Levanon 2003). We conclude that STRT did not suffer a significant quantitative bias against long transcripts.

To determine the depth of sequencing required to sample most of the available complexity, we studied the "new discovery" rate as a function of read depth. As shown in Supplemental Figure 3a, the sample reported here was not sequenced to saturation and appeared to contain at least 10 million distinct molecules. This leads to a rough lower-bound estimate of 100,000 mRNA molecules per well, not far from the 241,000 overall average found by normalizing against control mRNA (see below).

In contrast, the rate of discovery of distinct genes diminished more rapidly, and most were detected within the first 10 million reads (Supplemental Fig. 3b,c).

Absolute mRNA abundance in single cells

An advantage of single-cell analysis is that the number of cells (i.e., one) in each sample is known with certainty. If internal mRNA controls are added in known amounts, it is therefore possible to obtain an estimate of the total number of mRNA molecules present in each well.

By using this approach, we estimate that 10 pg RefRNA contained 103,000 mRNA molecules per well on average (Fig 3A), while an average of 241,000 mRNA molecules were present in wells containing cells (Fig. 3B). Only 812 molecules were found on average in negative control wells. Interestingly, there was a sharp distinction between cell types. While we found just 22,000 molecules per well in ES cells, there were 505,000 on average in MEFs, suggesting that the latter cell type contained about 20-fold more mRNA. The same difference was observed for ribosomal RNA (small subunit), suggesting that ES cells contain overall less RNA than MEFs.

ES cells were karyotypically normal (data not shown), and MEFs were primary cells, arguing against any difference in ploidy. The most abundant endogenous ribonucleases (RNases K, 4, and 2b) were all significantly more abundant in MEFs than ES cells and were uncorrelated with total molecule number, arguing against active degradation in ES cells. All cells were picked on the same occasion and in the same reagents and were amplified in the same tube, ruling out batch effects.

It is possible that ES cells express fewer genes or, alternatively, that cell lysis was significantly less effective in ES cells than MEFs. To distinguish these hypotheses, we compared the relative expression levels of the 20 highest expressed genes in each cell type. If the difference in mRNA abundance was due to incomplete lysis, we would expect a random loss of mRNA molecules, and hence the top genes in ES and MEF cells would account for a similar percentage of all mRNA. However, we found instead a striking difference: While in MEFs the top 20 genes account for 15% of all mRNA, they accounted for 39% in ES cells. Similarly, the top thousand genes in ES cells ranged from 1.4–2709 molecules per cell, whereas for MEFs the range was 77–7044 molecules per cell (in negative controls, the range was 0.05–4.6 molecules). We conclude that the most likely explanation for the smaller number of mRNA molecules in ES cells is that they in fact express fewer genes. Extending this analysis to all genes shows that, indeed, a smaller number of genes in ES cells explains a larger fraction of mRNA molecules, compared with both MEFs and RefRNA (Fig. 3C).

To validate the finding, we extracted RNA in bulk from a known number of cells, and found a 5.5-fold difference in total

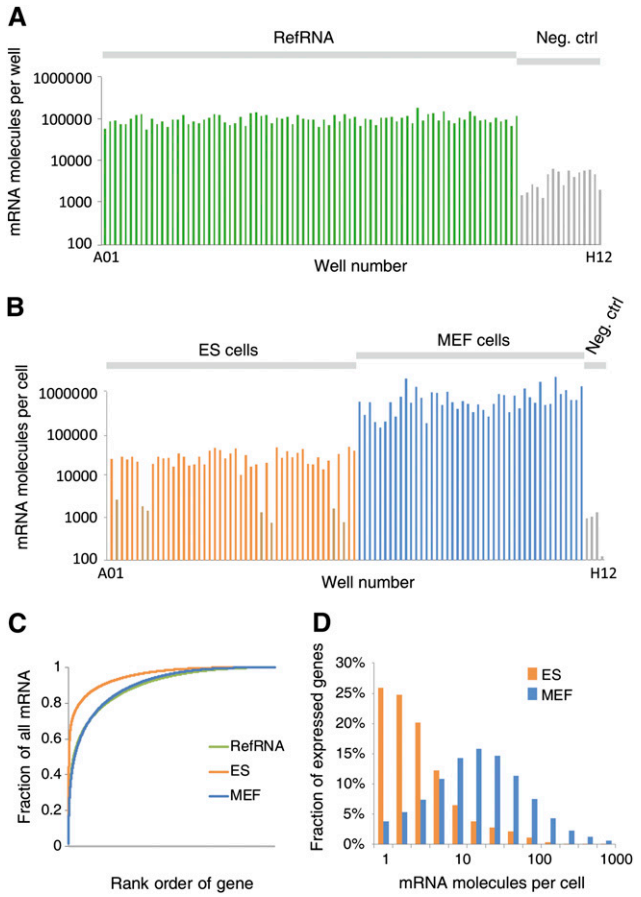

Figure 3. Number of mRNA molecules detected per cell. Approximately 2500 molecules of eight synthetic control mRNAs were spiked into each well. Using the number of reads mapped to synthetic mRNA as a normalizing factor, we converted the raw read counts from each well to an absolute number of mRNA molecules. The figures show the molecule count for each cell ordered by position on the reaction plate. (*A*) Molecule counts obtained from brain reference total RNA at 10 pg per well. The average observed was 103,000 molecules per well (negative controls: 4300 per well). (*B*) Molecule counts obtained from cells. A total of 48 ES cells, 44 MEF cells, and four empty wells were included. The overall average was 241,000 per cell (negative controls: 841 per cell). Seven wells apparently failed (molecule numbers similar to the negative controls, shown in pale orange), and were omitted from further analysis. (*C*) The cumulative fraction of all mRNA as a function of rank order gene expression level. Apparently, a smaller number of genes was expressed, compared with MEFs and RefRNA. (*D*) The distribution of gene copy number across all genes and cells.

RNA content (ES cells: 0.8–1.2 pg per cell; MEFs 4.8–4.9 pg per cell). Thus ES cells express less mRNA, from a smaller number of genes, compared with MEFs.

As a consequence of the differences in total detected molecule number, ES cells and MEFs showed distinct distributions of detected mRNA copy numbers in single cells (Fig. 3D). While the median copy number in MEFs was 15, in ES cells it was just two.

Splice variants

As a result of the nonuniform distribution of reads along each transcript, we could only detect alternative splicing that occurred at specific positions. This made it difficult to perform a systematic analysis of alternative splicing. Nevertheless, 13% of all exon-mapping reads spanned splice junctions and were thus informative of splice patterns; this set included junctions joining adjacent as well as nonadjacent exons in a total of 7339 genes. Examining these reads, we found evidence of 1580 alternatively spliced genes, with an average of 1.9 alternative splice events each.

Quantification of gene expression levels

In order to generate a quantitative measure of gene expression comparable to the commonly used RPKM (Mortazavi et al. 2008), we counted the number of hits to each annotated gene and normalized the data to transcripts per million (t.p.m.). We did not normalize by transcript length (as in the RPKM measure) because a single amplifiable molecule was generated for each input mRNA molecule, irrespective of its length.

Since single-cell cDNA was pooled before amplification, the yields of different cells could not be subsequently normalized. As a consequence, cells were unequally sampled and the limit of detection varied. Whereas highly expressed genes like *Actb* were detected in every cell, the probability of detection decreased for genes expressed at lower levels; the level required to reach 50% detection probability was 29 t.p.m. in MEFs and 103 t.p.m. in ES cells (Fig. 4A).

A scatterplot including the top 1000 genes in ES and MEF cells (1465 genes in total) showed that cell-to-cell variability was high, as expected (Fig. 4B). Surprisingly, the similarity was greater between ES cells than between MEFs, despite the fact that MEFs contained more mRNA and were thus sequenced to greater depth. The correlation coefficient was 0.86 for ES cells and 0.63 for MEFs, compared with 0.86 for RefRNA. However, the greatest differences were observed *between* cell types, where correlation coefficients were lower and groups of cell type–specific genes could be identified. For pairwise scatterplots of eight representative cells, see Supplemental Figure 4.

The reproducibility of expression measurements was assessed using the eight synthetic control mRNAs included in each well. Reproducibility was good when the number of molecules per well was in the range 100–1000, but as the number of molecules per well decreased, mRNA was detected in decreasing numbers of wells. Approximately 10 molecules per well were required to reach 50% detection probability (Fig. 4C). Nevertheless, biological variance exceeded technical variance at all levels of expression (Fig. 4D).

The measured expression levels agreed with those obtained by Q-PCR and to a lesser extent microarray hybridization (Supplemental Fig. 5). In agreement with published reports based on single-cell Q-PCR (Bengtsson et al. 2005), *Actb* mRNA abundance showed an approximately log-normal distribution across cells (Supplemental Fig. 6), shifted toward higher levels in MEFs compared with ES cells. RNA polymerase II (large subunit) was expressed at 12 ± 19 molecules per cell in MEFs, comparable to the 33 ± 79 t.p.m. found by direct detection in situ (Raj et al. 2006), assuming 300,000 transcripts per cell.

Revealing cell type relationships in a two-dimensional cell map

We wished to visualize cell–cell relationships on a two-dimensional map, such that more closely related cells would be located near each other. In this way, we hoped to be able to detect and distinguish cell types based solely on expression data, without relying on preexisting markers. A near-complete separation into distinct cell-type clusters was achieved using a graph-based method (see Methods).

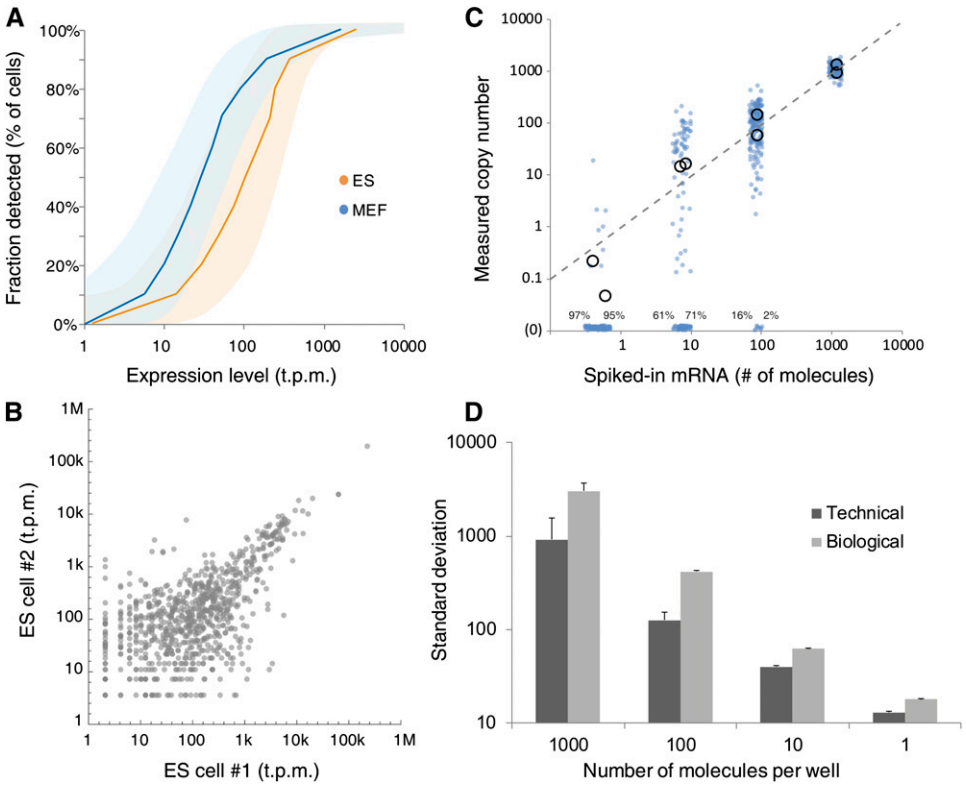

Figure 4. Quantitative accuracy. (*A*) Shows the probability of detection as a function of expression level for ES and MEF cells (shaded areas show 95% intervals). (*B*) Representative single-cell scatterplot showing the set of genes belonging to the top 1000 in ES and MEF cells (1465 in total). (*C*) The measured copy number for each of eight synthetic control mRNAs, across the entire plate. Circles show averages, whereas blue dots show individual data points (jittered for clarity). Zero measurements are shown along the horizontal axis, with percentage zeros indicated. For comparison, the dashed line indicates the ideal 45° slope. (*D*) Comparison of technical variance (based on control RNA) and biological variance (based on the average of genes with expression levels within ±20% of the indicated copy number) across the entire plate. Error bars, 95% confidence intervals; in each case the confidence intervals were nonoverlapping.

Briefly, we constructed a graph with nodes representing cells and with edges representing cell-to-cell similarity of expression pattern. We then used a force-directed layout to project the graph to two dimensions (Fig. 5A), resulting in only a single apparently misplaced cell. A conventional principal component analysis (PCA) revealed the same separation (Fig. 5B), but less distinctly. Again, a single ES cell clustered with the MEFs. These results demonstrate that the single-cell expression profiles contained enough information to distinguish cell types de novo. Both PCA and graph-based analysis clearly distinguished the cell types tested here, but the graph-based method generated more homogenous, well-separated clusters.

We then projected gene expression data onto the map, which provided an easy way to quickly grasp gene expression patterns in both single cells and in the clusters representing cell types (Fig. 5C). A set of well-known ES cell markers (*Dppa5a*, *Sox2*, and *Pou5f1*) were clearly specifically expressed in ES cells, although their expression levels varied widely from cell to cell (note the logarithmic color scale). Similar results were observed for *Sall4*, *Zfp42*, *Zic3*, and *Esrrb*. A few genes important for pluripotency (*Klf4*, *Myc*, and *Klf2*) were more broadly expressed. In contrast, MEFs were characterized by high expression of cytoskeletal and matrix proteins such as *Actb*, *Thbs1*, and *Fn1*. The power of large-scale single-cell analysis was evident in the fact that while not every cell expressed every marker, patterns of gene activity were highly consistent at the cluster level. For example, lower expressed transcription factors characteristic of ES cells went undetected in some individual ES cells, but the overall pattern of expression in the ES cell cluster was unambiguous and consistent with their identity as ES cells. As expression levels decreased, the fraction of expressing cells also decreased, reflecting the stochastic nature of gene expression as well as the sensitivity limits of the method (cf. Fig. 4A,B**)**.

Ranking genes by their differential expression in the two clusters revealed that MEFs were characterized by high expression of cytoskeletal proteins, whereas ES cells were highly enriched for ribosomal proteins. The top 10 MEF-enriched genes were cytoskeletal or matrix proteins (fibronectin 1, thrombospondin 1, beta actin, tropomyosin 1, sparc (also known as osteonectin), thymosin beta 4, collagen 1a2, and vimentin) with only two exceptions (*Malat 1* and *S100a6*). In contrast, the top 10 ES-enriched genes were ribosomal (eight subunits) plus ferritin light chain 1 and 2. In fact, of the top 60 genes enriched in ES cells, 45 were ribosomal subunits, representing half of all known ribosomal proteins. This may be a result of the fact that ES cells expressed fewer genes overall, so that key housekeeping genes were relatively enriched. Reassuringly, the known ES cell markers *Dppa5a*, *Sox2*, and *Pou5f1* were also highly enriched, ranked 11, 84, and 135, respectively.

The cell map representation demonstrated that (1) individual cells showed highly variable expression patterns (mostly due to technical variation), yet their overall pattern of expression was

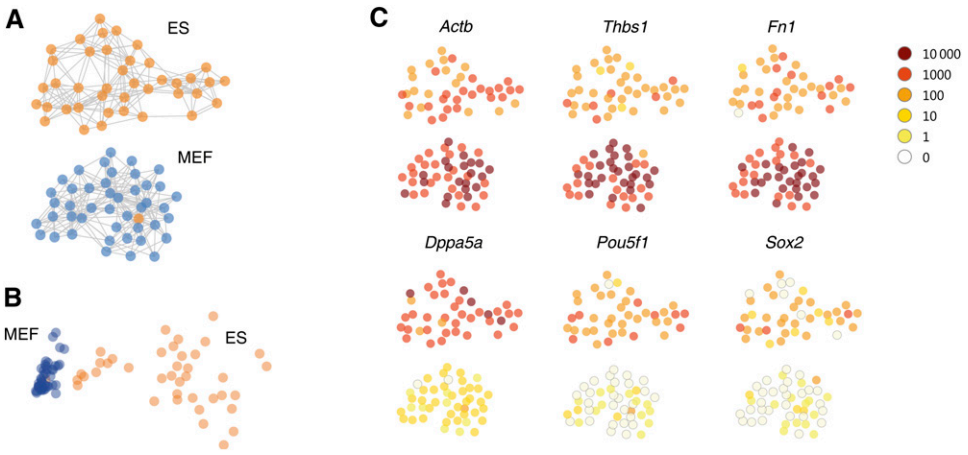

Figure 5. Graph-based visualization ("cell map"). (*A*) Cells, represented by graph nodes (circles) were laid out randomly, and edges (gray lines) were drawn from each cell to the five other cells it was most highly correlated with. Then, a force-directed layout was used to lay out the graph on the plane. In this stage, cells repelled each other uniformly but were held together by edges acting as elastic springs. The resulting visual map was consistent with known cell identities (ES cells in orange, MEFs in blue), with a single apparently misplaced cell. Note the lack of edges connecting the clusters, showing that the graph has separated into disjoint components. (*B*) The same data analyzed by principal component analysis (PCA), again with a single apparently misplaced cell but with less distinct separation by cell type. (*C*) The expression of selected genes is shown on a logarithmic color scale (*inset*, *upper right*). The *top* row shows genes enriched in MEFs, while the *bottom* row shows genes enriched in ES cells and known to be ES cell markers

sufficient to group cells of one type together as a cluster; and (2) once a cluster of cells was formed, representing a distinct cell type, patterns of gene expression at the cluster level were unambiguous. We conclude that shotgun single-cell expression profiling is an efficient strategy to access single-cell expression data in heterogeneous populations of cells.

Discussion

We have shown that large-scale single-cell expression profiling can be used to form cell type–specific clusters. This allows analysis of cell type–specific patterns of gene expression both at the single-cell level and the population level, without the need for known markers or even a prior knowledge that a certain type of cell exists. We propose that this general strategy can be extended to study all kinds of mixed samples, such as specific progenitors active during organogenesis, small populations of stem cells embedded in adult tissues, heterogeneous tumor cell samples, rare circulating tumor cells, and more.

What unites all these disparate scientific lines of inquiry is the need to unmix heterogeneous populations of cells. Currently, unmixing is primarily achieved either by physically isolating cells based on known cell surface markers or by genetically labeling the desired cells so that they can be isolated based on, for example, GFP expression. However, the use of previously known markers precludes the discovery of new cell types and always risks resulting in mixed data if the markers were not truly specific. In contrast, we have shown that cells of distinct types can be unmixed purely in silico, provided that large numbers of single-cell expression profiles are generated. Although in this case clusters did correspond to distinct cell-types, in general the structure of "cell-type space" is unknown. Classical cell types may indeed harbor multiple distinct substates, as exemplified by the fluctuating expression of Hes1 in ES cells (Kobayashi et al. 2009). These substates may or may not be possible to distinguish using large-scale single-cell transcriptome analysis, depending on what proportion of the transcriptome is regulated between the substates. The question of the number of functionally distinct cell types, and substates, and their relationships has hardly begun to be explored.

Importantly, then, a very high throughput, scalable method for single-cell expression profiling was required. RNA-seq has the advantage, over microarrays, of permitting high levels of multiplexing, while generating more specific, sensitive, and accurate expression data. However, sample preparation for RNA-seq is still labor-intensive and fairly expensive. We therefore developed a method to prepare a barcoded single-cell cDNA sample from 96 cells in a single incubation step. As a consequence, 96 cells could be pooled and treated as a single sample throughout the procedure, which greatly increased our throughput and reduced cost. The entire procedure takes 2 d to perform, from 96 living cells to finished samples loaded on the Genome Analyzer. The cost, including all reagents and consumables to generate about 100 million 55-bp reads on an Illumina Genome Analyzer IIx, is approximately $5000 (that is, about $50 per cell; however, as shown in Supplemental Fig. 3, a larger number of reads would be necessary to reach saturation, which will be proportionately more expensive).

A previous report detailed a single-cell RNA-seq method (Tang et al. 2009) for the SOLiD platform. However, that method has so far only been applied to a small number of atypically large cells (a total of seven cells were reported, each 10- to 100-fold larger than most somatic cells), did not maintain strand specificity (thus complicating the analysis of the approximately 3000 overlapping genes in the genome), and required a several-day procedure to prepare each cell for sequencing. In a more recent work, the method was applied to approximately 30 single cells, also from the early embryo (Tang et al. 2010). The lack of internal controls in that method precluded a direct estimate of the total number of mRNA molecules recovered.

Here we chose to generate data on a larger number of single cells, each analyzed at a relatively shallow depth of coverage. This allowed us to produce a cell map with high resolution. In fact, the more cells are added, the more accurate will be the aggregate data obtained from each distinct cell type (cluster) and the better the resolution in the "cell type space." For example, the ES cells here were sampled at 241 000 reads per cell on average, but altogether

9.8 million reads were obtained from the ES cells. Thus after identifying clusters of cells representing a distinct cell type, deeper sequencing data with greater sensitivity were immediately available for that type. Sampling a large number of cells will be especially important when the approach is applied to complex tissues, where some types of cells may be present only in a small minority. In addition, as sequencing costs continue to decrease, the tradeoff between number of cells and number of reads will become less pressing.

An important aspect of STRT is its ability to pinpoint the exact location of the 5′ end of transcripts. This could be used to analyze promoter usage in single cells and, in effect, provides a straightforward method for single-cell CAGE (cap analysis of gene expression). Although we found that endogenous transcripts often were not full length, this was not due to any inherent limitation, since synthetic mRNAs up to 2 kb were nearly completely full length. The discrepancy may be attributed to degradation during cell harvesting and picking, although it is possible that it also reflects the presence of partially degraded mRNA in living cells.

Several aspects of the method could be improved. Technical variation may in part be explained by PCR amplification bias (we used 20 cycles to amplify single-cell cDNA and 12 cycles for the Illumina sample preparation). It will be important to try to further reduce the total amount of amplification. Similarly, the efficiency of converting mRNA into amplifiable cDNA could probably be further increased, which would lead to increased sensitivity (currently, only about 1000–6000 genes were reliably detected in ES cells, 2000–8000 in MEFs, as shown in Supplemental Fig. 3). Another shortcoming of the present method is that it does not span the entire transcript length. This precludes systematic analysis of alternative splicing, at least for exons located away from the 5′ and 3′ UTRs. It will be an important subject of future research to find a way to barcode and amplify fragments representing the entire transcript, while maintaining multiplexing and without also including undesired RNA species such as ribosomal RNA.

We envisage the future use of very large-scale single-cell transcriptional profiling to build a detailed map of naturally occurring cell types, which would give unprecedented access to the genetic machinery active in each type of cell at each stage of development.

Methods

Cell culture and RNA purification

ES R1 cells were cultured as previously described (Moliner et al. 2008). MEFs were prepared as primary cells from mouse embryos and were maintained in DMEM with 10% FBS, 1× penicillin/streptomycin, 1× Glutamax, and 0.05 mM 2-mercaptoethanol. All reagents were from Invitrogen. MAQC Human Brain Reference RNA was purchased from Ambion. Total RNA was prepared from 100,000 and 300,000 counted cells using TRIzol (Invitrogen) according to the manufacturer's instructions and was quantified on an Agilent BioAnalyzer using the RNA 6000 Nano chip. Concentration measurements by Qubit (Invitrogen) yielded similar results.

Single-cell tagged reverse transcription

A cell capture plate (AbGene Thermo-Fast 96 catalog no. 0900) was prepared, containing 5 μL of STRT buffer (20 mM Tris-HCl at pH 8.0, 75 mM KCl, 6 mM $MgCl_2$, 0.02% Tween-20) with 400 nM STRT-V3-T30 (5′-biotin-AAGCAGTGGTATCAACGCAGAGTCGACT$_{30}$VN-3′; this and all other oligos were from Eurofins MWG Operon) and 400 nM STRT-V2-n (5′-AAGCAGTGGTATCAACGCAGAGTGCAG TGCTXXXXXXrGrGrG-3′, where "rG" denotes a riboguanine and "XXXXXX" was a 6-bp barcode) (see Supplemental Table 1). Each well of the capture plate contained a different template-switching helper oligo (STRT-V2-1 through STRT-V2-96) with a distinct barcode.

Cells were dissociated enzymatically using TrypLE Express (Invitrogen), washed, and resuspended in phosphate-buffered saline (PBS). A single cell was collected into each well of a 96-well capture plate using a custom-built semi-automated cell picker, and the plate was immediately frozen on dry ice. When total RNA was analyzed instead, 1 μL of 10 pg/μL was added to each well.

The cell capture plate was thawed, and 5 μL reverse transcription mix (4 mM DTT, 2 mM dNTP, 5 U/μL Superscript II in 100 mM Tris-HCl at pH 8, 375 mM KCl, 0.1% Tween-20, 6 mM $MnCl_2$, 2500 molecules of control mRNA) was added to each well. The synthetic mRNA consisted of eight different in vitro-transcribed mRNAs (Ambion ArrayControl) ranging from 755–2000 bp, in a dilution series (calculated to contain 1180, 1170, 88, 88, 9, 7, 0.6, and 0.4 molecules of each species). The plate was incubated (10°C for 10 min, 42°C for 45 min) to complete reverse transcription and template switching.

To purify the cDNA and remove unreacted primers, 100 μL MyOne carboxylate beads (Invitrogen) was washed twice in 100 μL EBT (10 mM Tris-Cl at pH 8.5, 0.02% Tween-20) and then resuspended in 2 mL 14% PEG-6000 in 0.9 M NaCl, and 20 μL of this mixture was added to each well. Beads from all wells were pooled, washed in 70% ethanol twice, dried, and eluted in 37 μL EBT in a 1.5-mL polyallomer tube (Beckman).

The cDNA was amplified in a single tube in 50 μL of 200 μM dNTP, 200 nM STRT-PCR primer (5′-biotin-AAGCAGTGGTATC AACGCAGAGT-3′), 1× Advantage2 DNA Polymerase Mix (Clontech) in 1× Advantage2 PCR buffer (Clontech) with 1 min at 95°C followed by 20 cycles of 15 sec at 95°C, 30 sec at 65°C, 4 min at 68°C, with heated lid. A 5 μL aliquot was amplified for another five cycles and visualized on a 1.2% agarose E-gel (Invitrogen) to confirm the range of cDNA lengths.

The product was immobilized on MyOne C1 Streptavidin beads. Twenty microliters of beads was washed twice in 50 μL 2× BWT (10 mM Tris HCl at pH 7.5, 1 mM EDTA, 2 M NaCl, 0.02% Tween-20) and added to the remaining 45 μL PCR product. After a 10-min incubation at room temperature, the beads were washed three times in 50 μL 1× BWT and twice in EBT.

Sample preparation for high-throughput sequencing

Amplified cDNA was fragmented by DNase I in the presence of Mn^{2+}, which causes a preference for double-strand breaks. Beads were resuspended in DNase I buffer supplemented with 10 mM $MnCl_2$ and DNase I diluted to 0.0003 U/μL in a total volume of 120 μL for exactly 8 min at room temperature. The reaction was stopped by washing the beads five times in 50 μL EBT.

DNA ends were repaired and A-tailed as follows. Beads were resuspended in 25 μL EBT, and 25 μL NEBNext End Repair reaction mix (New England Biolabs) was added; then the beads were incubated for 30 min at room temperature. The beads were washed twice in EBT, then resuspended in 21 μL EBT; 2.5 μL NEBNext dA tailing buffer and 1.5 μL Klenow exo- (both NEB) were added, and the reaction was incubated for 30 min at 37°C, followed by two washes in 50 μL EBT.

An adapter containing the Illumina P2 sequence (5′-CAAGC AGAAGACGGCATACGAGCTCTTCCGATCT-3′ and 3′-PHO-GTTC GTCTTCTGCCGTATGCTCGAGAAGGCTAG-PHO-5′) was ligated by resuspending the beads in 25 μL EBT and adding 2× ligation mix (2× NEBuffer 4, 2 μM adapter, 1 U/μL T4 DNA ligase, 2 U/μL SalI-HF,

and 2 mM ATP; all reagents from NEB) and incubating for 30 min at 37°C and then washing twice in 25 μL EBT. The SalI-HF enzyme releases 3′ fragments bound to the beads.

The beads were then used as template in a 100 μL PCR reaction (200 μM dNTP, 400 nM each primer of primers 5′-AATGATACGGCGACCACCGAGATCTAAGCAGTGGTATCAACGCAGAGT-3′ and 5′-CAAGCAGAAGACGGCATACGAG-3′, 200 μM dNTP, 0.2 U/μL Phusion polymerase in Phusion HF buffer; all from NEB) with 30 sec at 98°C, followed by 12 cycles of [10 sec at 98°C, 30 sec at 65°C, 30 sec at 72°C] followed by 5 min at 72°C. The product was purified by AmPure XP (Beckman Coulter) and resuspended in 20 μL EBT. Yield was typically 1–3 ng/μL.

The sample was size-selected on a 2% E-gel with SYBR Safe (Invitrogen), recovering the 200 400 bp range by Qiaquick Gel Extraction (replacing the heating step with 15 min of vigorous agitation).

Cluster formation and sequencing-by-synthesis was performed in-house on a Genome Analyzer IIx according to the manufacturer's protocols (Illumina, Inc.).

Mapping, quantification, and visualization

Raw reads were sorted by barcode, trimmed, and mapped to the mouse genome using Bowtie (Langmead et al. 2009). Unmapped reads were discarded. Then, for each annotated feature in the NCBI 37.1 assembly, all mapping reads were counted to generate a raw count for each genomic feature. Finally, the raw reads for each cell were normalized to transcripts per million. A detailed description of the analysis pipeline will be published elsewhere, and the software is available from us upon request.

To visualize cells in a two-dimensional landscape, we first computed all pairwise similarities using the Bray-Curtis distance (because it handled the noise in low-expressed genes well; standard correlation yielded similar results, but with a few more misplaced cells). We then built a similarity graph by letting nodes represent cells, and connecting each cell to its five most similar cells. A force-directed layout was used to project the graph to two dimensions, revealing the internal structure based on cell–cell similarities.

Acknowledgments

We thank A. Metsis and K. Jääger for useful discussions and C. Ibàñez and P. Ernfors for useful comments on the manuscript. This work was supported by Swedish Research Council grant 521-2006-3991 and Swedish Foundation for Strategic Research grant MDB09-0052.

References

Bengtsson M, Stahlberg A, Rorsman P, Kubista M. 2005. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* **15:** 1388–1392.

Borodulina OR, Kramerov DA. 2008. Transcripts synthesized by RNA polymerase III can be polyadenylated in an AAUAAA-dependent manner. *RNA* **14:** 1865–1873.

Chubb JR, Trcek T, Shenoy SM, Singer RH. 2006. Transcriptional pulsing of a developmental gene. *Curr Biol* **16:** 1018–1025.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* **5:** 613–619.

Eisenberg E, Levanon EY. 2003. Human housekeeping genes are compact. *Trends Genet* **19:** 362–365.

Esumi S, Wu SX, Yanagawa Y, Obata K, Sugimoto Y, Tamamaki N. 2008. Method for single-cell microarray analysis and application to gene-expression profiling of GABAergic neuron progenitors. *Neurosci Res* **60:** 439–451.

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

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

Kurimoto K, Yabuta Y, Ohinata Y, Ono Y, Uno KD, Yamada RG, Ueda HR, Saitou M. 2006. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* **34:** e42. doi: 10.1093/nar/gkl050.

Lagunavicius A, Merkiene E, Kiveryte Z, Savaneviciute A, Zimbaite-Ruskuliene V, Radzvilavicius T, Janulaitis A. 2009. Novel application of Phi29 DNA polymerase: RNA detection and analysis in vitro and in situ by target RNA-primed RCA. *RNA* **15:** 765–771.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Moliner A, Enfors P, Ibanez CF, Andang M. 2008. Mouse embryonic stem cell-derived spheres with distinct neurogenic potentials. *Stem Cells Dev* **17:** 233–243.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. *Nature* **461:** 814–818.

Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. 2006. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4:** e309. doi: 10.1371/journal.pbio.0040309.

Schmidt WM, Mueller MW. 1999. CapSelect: a highly sensitive method for 5′ CAP-dependent enrichment of full-length cDNA in PCR-mediated analysis of mRNAs. *Nucleic Acids Res* **27:** e31. doi: 10.1093/nar/27.21.e31.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A et al. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6:** 377–382.

Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA. 2010. Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6:** 468–478.

Taniguchi K, Kajiyama T, Kambara H. 2009. Quantitative analysis of gene expression in a single cell by qPCR. *Nat Methods* **6:** 503–506.

Taylor KD, Piko L. 1987. Patterns of mRNA prevalence and expression of B1 and B2 transcripts in early mouse embryos. *Development* **101:** 877–892.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Warren L, Bryder D, Weissman IL, Quake SR. 2006. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proc Natl Acad Sci* **103:** 17807–17812.

Wood SA, Allen ND, Rossant J, Auerbach A, Nagy A. 1993. Non-injection methods for the production of embryonic stem cell-embryo chimaeras. *Nature* **365:** 87–89.

Received May 27, 2010; accepted in revised form April 19, 2011.