Cell PRESS
Open ACCESS
Cell Reports
Resource

CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification

Tamar Hashimshony,[1,2] Florian Wagner,[1,2] Noa Sher,[1,2] and Itai Yanai[1,*]
[1]Department of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel
[2]These authors contributed equally to this work
*Correspondence: yanai@technion.ac.il
http://dx.doi.org/10.1016/j.celrep.2012.08.003

SUMMARY

High-throughput sequencing has allowed for unprecedented detail in gene expression analyses, yet its efficient application to single cells is challenged by the small starting amounts of RNA. We have developed CEL-Seq, a method for overcoming this limitation by barcoding and pooling samples before linearly amplifying mRNA with the use of one round of in vitro transcription. We show that CEL-Seq gives more reproducible, linear, and sensitive results than a PCR-based amplification method. We demonstrate the power of this method by studying early *C. elegans* embryonic development at single-cell resolution. Differential distribution of transcripts between sister cells is seen as early as the two-cell stage embryo, and zygotic expression in the somatic cell lineages is enriched for transcription factors. The robust transcriptome quantifications enabled by CEL-Seq will be useful for transcriptomic analyses of complex tissues containing populations of diverse cell types.

INTRODUCTION

For many biological questions a single-cell-level description of gene regulation is advantageous to cell populations (Tang et al., 2011; Wang and Bodovitz, 2010). Microscopy, FACS, or real-time PCR-based methods can provide a single-cell aspect to experiments but are able to assay only a handful of genes at a time. High-throughput technologies such as microarrays and RNA-Seq provide a full view of the expression of all genes but are limited by the amount of RNA needed for analysis. This can be solved by adding an RNA amplification step, either by exponential PCR-based amplification or linear in vitro transcription (IVT) amplification (Eberwine et al., 1992). With PCR practically any RNA starting amount can be employed, simply by adding additional cycles, thereby allowing analysis at the single-cell level. However, efforts for linear amplification of RNA from single cells have been challenged by IVT's lower bound of ~400 pg total RNA as input material for a single round of amplification. Therefore, to date, IVT has not been efficiently used for amplification of RNA from single cells (Tang et al., 2011).

With the dramatically decreasing costs of sequencing, RNA-Seq (Wang et al., 2009) has emerged as the preferred method for transcriptomic analyses, overtaking microarrays, providing an imperative for any transcriptomic method to be adapted for RNA-Seq. A PCR-based amplification protocol has been used in combination with SOLID sequencing (Tang et al., 2009). Recently, the PCR-based method has been extended to include a multiplexing step for the amplification of multiple cells in parallel, allowing for high-throughput analysis, and uses the Illumina sequencing platform (Islam et al., 2011). In comparison, IVT of single cells has been described before; however, it is labor intensive (Eberwine et al., 1992), requiring three rounds of amplification (~5 days work/cell) and has not been adapted for multiplexed sequencing. These considerations have hitherto prevented the higher quality possible with IVT from being adapted for single-cell RNA-Seq.

Here, we present CEL-Seq (Cell Expression by Linear amplification and Sequencing), a protocol that meets the demand of linear amplification by IVT for sufficient material by pooling barcoded samples, therefore allowing the efficient linear amplification of RNA from single cells and their analysis by sequencing. We compare the performance of our method on two mammalian cell types to that of a PCR-based approach and use spike-ins to establish CEL-Seq's exact reproducibility and sensitivity at very low amounts of input RNA. Finally, we apply our protocol to study sister cells from early *C. elegans* embryos, and demonstrate that CEL-Seq's high performance can be used to reliably distinguish between cell types, even in cases where only subtle biological differences are present.

RESULTS

CEL-Seq Performs Multiplexed Single-Cell Transcriptomics by Linear Amplification

The CEL-Seq method begins with a single-cell reverse-transcription reaction using a primer designed with an anchored polyT, a unique barcode, the 5′ Illumina sequencing adaptor, and a T7 promoter (Figure 1A; see Experimental Procedures for details). Next, second-strand synthesis is performed and then the cDNA samples are pooled and consequently comprise sufficient template material for an IVT reaction. The amplified RNA is then subjected to directional RNA library preparation. The RNA is fragmented to a size distribution appropriate for sequencing, the Illumina 3′ adaptor is added by ligation, RNA is reverse transcribed to DNA, and the 3′-most fragments that

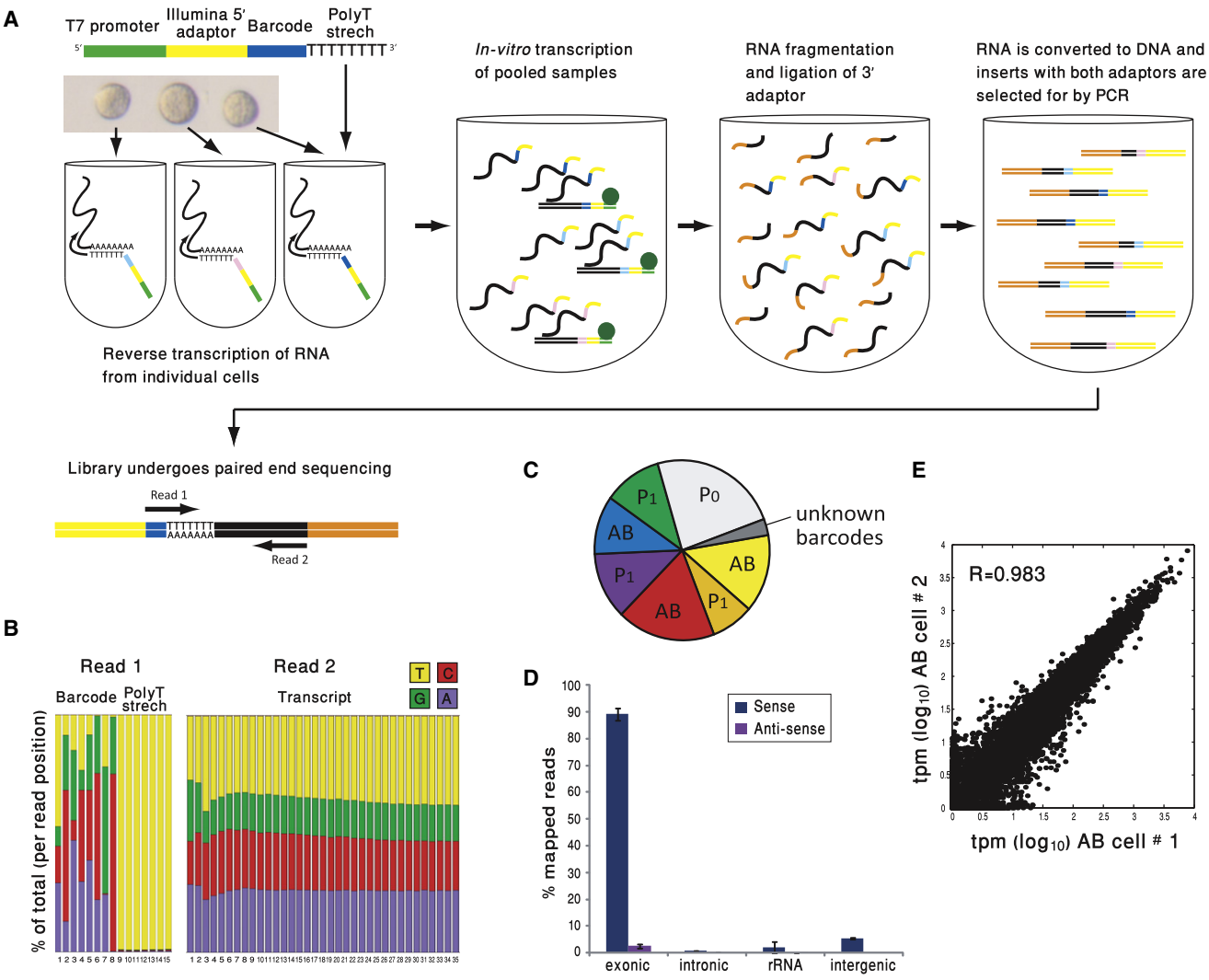

Figure 1. The CEL-Seq Method

(A) Individual cells are added to tubes, each with a uniquely bar-coded primer for reverse transcription. After second-strand synthesis, the reactions are pooled for IVT. The amplified RNA is then fragmented and purified before entry into a modified version of the Illumina directional RNA protocol, the molecules with both Illumina adaptors are selected, and the DNA library is sequenced with paired-end reads.

(B) Nucleotide distribution in the sequenced paired-end reads. Each nucleotide position is represented by one column, with the first base on the left.

(C) Barcode distribution of one IVT reaction after demultiplexing. The cells from three two-cell stage *C. elegans* embryos (denoted P_1 and AB) and a single one-cell stage embryo (denoted P_0) were amplified together in a single multiplexed IVT reaction.

(D) Distribution of the reads mapping to the *C. elegans* genome in the six AB/P_1 cells. Error bars indicate the SD.

(E) Correlation between biological AB replicates.

See also Figure S1C.

contain both Illumina adaptors and a barcode are selected. The resulting library undergoes paired-end sequencing, where the first read recovers the barcode, whereas the second identifies the mRNA transcript (Figure 1A). Thus, by multiplexing CEL-Seq takes advantage of the different input requirements of the reverse-transcription and IVT reactions to obtain sufficient RNA from single cells for a single round of linear amplification.

As an initial test, we applied CEL-Seq to individual cells isolated from three two-cell *C. elegans* embryos (Table S1). On average, 95.5% of the filtered reads had a barcode located precisely at the beginning of the first read, invariably followed by a polyT stretch (Figure 1B). Barcodes from all six samples were represented, indicating the success of the individual single-cell reverse-transcription reactions (Figure 1C). We mapped reads to the *C. elegans* genome and found that 91.7% stemmed from mRNA. Only 2.0% stemmed from ribosomal RNA (rRNA), demonstrating CEL-Seq's specificity for polyadenylated transcripts. This was also supported by the extremely low-detected expression levels of core histone mRNAs, which are thought to be highly expressed yet mostly nonpolyadenylated

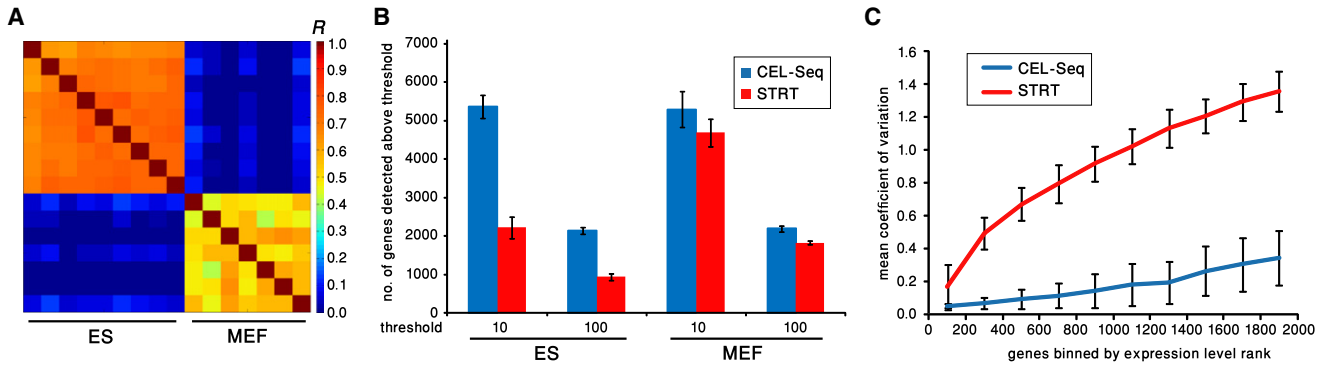

Figure 2. Benchmarking of CEL-Seq on Mouse ES and MEF Cells

(A) CEL-Seq Pearson's correlation coefficients among the ES and MEF cells (on $\log_{10}$ tpm values), computed as previously described by Islam et al. (2011), on the 1,000 genes most highly expressed in ES cells and 1,000 genes most highly expressed in MEF cells (1,385 genes).

(B) Mean number of genes detected above two thresholds (10 and 100 tpm) across the ES and MEF cell types using CEL-Seq and the "STRT" PCR-based method (Islam et al., 2011). Error bars indicate 95% confidence intervals.

(C) Reproducibility according to expression level. For each gene the coefficient of variation was computed across the $\log_{10}$ tpm values in the ES cells for the STRT and CEL-Seq methods. The genes were then ranked by expression level in bins of 200 from high to low. For each bin the mean and SD of the coefficients of variation of the genes are shown.

See also Figure S2 for additional analyses.

(Figure S1A). Finally, RNase treatment of cells did not produce amplified RNA indicating the specificity of the method to RNA.

CEL-Seq is highly strand specific because 97.2% of exonic reads exhibited sense orientation (Figure 1D). The reads mapped exclusively to the 3′ end of transcripts (Figure S1B), which was expected because CEL-Seq only retains the 3′-most fragments of transcripts (Figure 1A). Some reads mapped to intergenic sequences, but manual inspection of the aligned reads revealed that this can likely be explained by incomplete 3′ UTR annotations (data not shown). Expression levels were then estimated by counting all reads mapping to each gene, and normalized to give the read count in transcripts per million (tpm; see Experimental Procedures). The expression levels of all genes (henceforth, transcriptome) across biological replicates showed an average correlation of R = 0.979 (Figures 1E and S1C). A negative control—starting with the growth media lacking a cell—resulted in very few reads (Figure S1D).

CEL-Seq Outperforms a PCR-Based Multiplexed RNA-Seq Method

We next sought to compare the performance of our protocol to the STRT method, a previously introduced PCR-based multiplexed RNA-Seq method by Islam et al. (2011). We thus applied CEL-Seq to the cell types compared in this previous study by Islam et al. (2011) and determined the transcriptomes of nine mouse embryonic stem (ES) cells and seven mouse embryonic fibroblasts (MEFs) (Table S1). When comparing the distribution of expression levels of each single-cell transcriptome across methods and cell type, CEL-Seq shows more reproducible distributions of expression (Figure S2A). We found that CEL-Seq produced higher correlations for ES cells and distinguished between cell types more clearly, when examining the highly expressed genes in either cell type (Figures 2A and S2B). Furthermore, CEL-Seq detected significantly more genes in the ES cells (Figure 2B). In order to ensure a fair comparison between the two methods, we quantified reproducibility for each method separately, using the same criteria as previously described by Islam et al. (2011), and found that with CEL-Seq, significantly lower noise was detected across biological replicates for both cell types tested (Figures 2C and S2C). Finally, principal component analysis based on CEL-Seq data better distinguishes between cell types than the corresponding STRT data (Figure S2D). Biological variation between replicates is a confounding factor when trying to establish the performance of a method. We therefore also compared the two methods based on expression levels of exogenously introduced RNA (see below) and found that CEL-Seq provided more reproducible measurements (Figure S2E).

CEL-Seq Is Highly Sensitive and Reproducible

In order to determine CEL-Seq's sensitivity and reproducibility, we analyzed different amounts of purified *C. elegans* RNA from mixed embryonic stages, eliminating biological variability present between single cells and allowing us to use different dilutions of the same RNA. We prepared stepwise dilutions, from 40 pg of total RNA down to levels representative of mammalian single cells (~5 pg). In parallel we sequenced a 1 ng RNA sample from the same preparation for use as a reference. In half of the samples, we added exogenous "carrier" RNA to test whether the overall amount of RNA in a reverse-transcription reaction affects the efficiency of the reverse-transcription step and found that it did not: the number of reads that mapped to the *C. elegans* genome depended only on the amount of *C. elegans* RNA present in a given sample (Figure S3A). Each sample also contained a set of 92 spike-in RNAs with defined concentrations, spanning more than five orders of magnitude (Baker et al., 2005), which showed a linear response across the entire detection range (R^2 = 0.87 ± 0.04 for the 10 pg samples; Figures 3A and S3B).

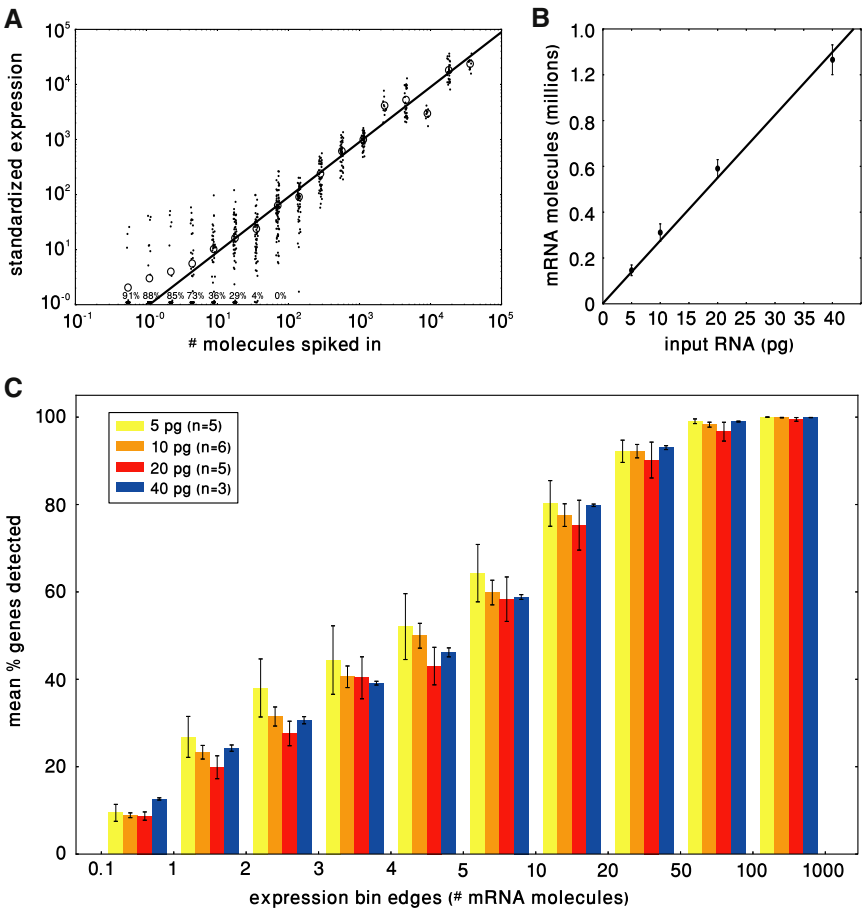

Figure 3. Sensitivity and Reproducibility of CEL-Seq

(A) CEL-Seq achieves a linear response over the entire detection range. The plot indicates the 92 ERCC (Baker et al., 2005) spike-in levels from six 10 pg replicate samples. Averages for each of the 17 groups of spike-ins with the same nominal concentration are shown as larger circles. For each sample the spike-ins were normalized such that the average expression of the 12th spike-in group containing 1,000 molecules was set to 1,000. The fraction of spike-ins in each group without detected expression is shown at the bottom of the figure. The line indicates an idealized linear relationship. Systematic deviations from the line for some spike-ins with high concentrations are likely due to differences in G/C content.

(B) *C. elegans* transcript counts per sample based on linear regression of spike-in expression levels (5 pg, n = 5; 10 pg, n = 6; 20 pg, n = 5; 40 pg, n = 3). The number of molecules calculated is directly proportional to the amount of input RNA. Error bars indicate the SD.

(C) CEL-Seq sensitivity. For different amounts of input RNA, the bars indicate the percentage of genes detected (at least one read) as a function of their absolute copy number, calculated based on the 1 ng reference sample. Error bars indicate the SDs.

See Figure S3 for additional analyses.

We next invoked the spike-ins to assess CEL-Seq's reproducibility and sensitivity based upon absolute molecule counts. These calculations indicated that 10 pg of the *C. elegans* RNA contained approximately 280,000 mRNA molecules (Figure 3B). Using that number, and the relative expression level of each gene in the reference sample, we calculated the expected number of transcripts per gene in each of our dilution series and binned the genes in each sample according to this number. Comparison of the different dilutions showed that CEL-Seq's absolute sensitivity does not decrease with smaller starting amounts of RNA (Figures 3C and S3C). For example 50% of all genes with four to five copies and virtually all genes present in more than 50 copies per cell were detected, independent of the total amount of RNA analyzed (Figure 3C). Similarly, the absolute reproducibility of the method is also not compromised by smaller starting amounts (Figure S3C). In summary, CEL-Seq's performance in estimating a transcript's abundance depends solely on its absolute copy number in the sample. Thus, whereas the IVT imposes a minimal starting amount of ~400 pg total RNA for sufficient yield in a single round, the reverse-transcription reaction does not have this restriction. CEL-Seq thus exploits this difference by pooling reverse-transcription reactions together to arrive at the IVT threshold.

We also assessed the required sequencing depth for accurate transcriptomic data using CEL-Seq. We created 12 technical replicates of 20 pg *C. elegans* RNA. We simulated different sequencing depths and found that at one million reads, 91% of the genes with an average expression >100 tpm had an expression level within 20% of that of the averaged expression and that sequencing deeper did not further improve reproducibility. A total of 250,000 reads produced results that were already sufficient for good quantification and very similar to an order of magnitude higher sequencing depth (Figure S3D).

Transcriptomic Analysis of Single Cells in the Early *C. elegans* Embryo Identifies Differential and New Expression

We next asked whether CEL-Seq can be used to identify biological differences between closely related cells in the *C. elegans* embryo. *C. elegans* embryonic development begins with unequal cleavages producing founder cells—termed blastomeres—some of which are depicted in Figure 4B. The AB and P_1 sister blastomeres were examined 10 min after cell division. Examining their transcriptomes, we detected 17 genes with a mean 2-fold difference showing significantly different expression ($p < 0.05$, t test, FDR-corrected; Figure 4A). The most highly expressed of these is *mex-3*, whose differential expression is supported by previous work by Draper et al. (1996). Shuffling the groupings of the triplicates resulted in no differentially expressed genes. Next, we sought to proceed in developmental time, using CEL-Seq to further characterize the early embryonic transcriptome. We collected cells by comparing the germ

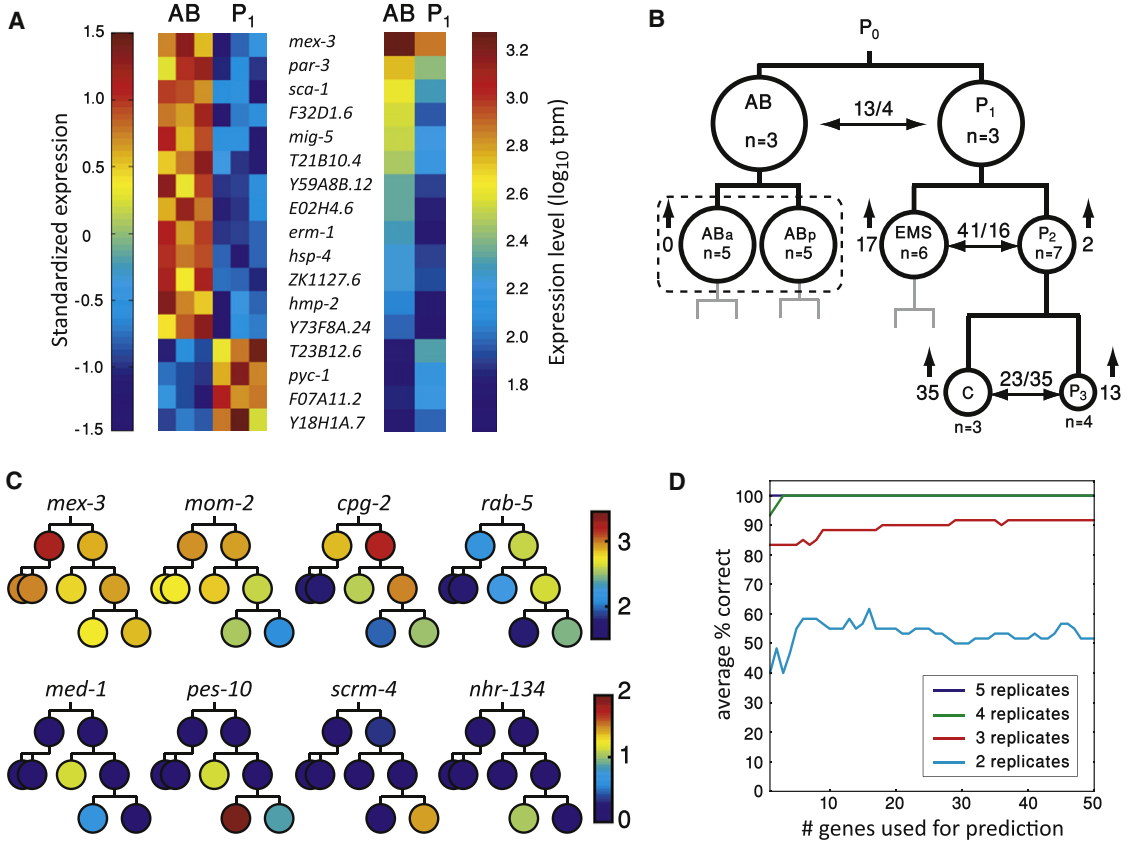

Figure 4. Dissecting the Early *C. elegans* Embryo with CEL-Seq

(A) Differential expression analysis in the two-cell stage blastomeres, AB and P_1. A t test was made for the 137 genes for which there was expression >100 tpm and at least 2-fold change between the means of the triplicates. The 17 genes with $p < 0.05$ (FDR corrected) are shown along with the mean expression (right) and standardized expression of triplicates on the left (mean subtracted and SD divided).

(B) The blastomeres examined in this study are abstracted in the cell lineage. The number of new transcripts is indicated under the vertical arrows. The two-sided arrows indicate the number of differentially expressed genes in the anterior versus the posterior blastomeres, respectively. A list of all differentially expressed and new transcripts is in Tables S2 and S3, respectively.

(C) Gene expression levels (log_{10} tpm; see color scale on right) for the indicated genes; cell lineage is as in (B).

(D) Classification of the AB and P_1 blastomeres. For the indicated number of replicates used in the training data, the performance of the machine-learning classifier (see Experimental Procedures) in assigning blastomere identities is shown as a function of the number of genes included for prediction.

See Figure S4 for additional analyses.

lineage to the somatic cells: P_1 was allowed to divide to its two daughter cells, P_2 and EMS, which were analyzed, and similarly, P_2 was allowed to divide to P_3 and C, which were then analyzed. We also collected the AB daughter cells. We assayed for differential distribution of transcripts in these additional cells similarly to the AB/P_1 analysis (Table S2), as well as asking which genes are newly expressed by comparing with the mother cell (Figure 4B; Table S3). The differential distribution between AB/P_1 and EMS/P_2 has also been tested by microarray and regular RNA-Seq, in both cases using pooled cells, showing good correlation (Figure S4A). New transcription in a particular cell type was scored when the transcript's concentration exceeded 10 tpm and was <1 tpm in the mother cell (median across replicates). We found that in the four-cell stage, only the EMS cell has considerable expression of new transcripts. Among these are *med-1* and *pes-10*, known to be newly expressed in this cell (Maduro et al., 2001; Seydoux et al., 1996) (Figure 4C). Interestingly, newly expressed EMS genes are enriched for transcription factors (4 out of 17; $p < 10^{-3}$, hypergeometric distribution): *med-1* and three ccch genes—*ccch-2*, *F38C2.7*, and *Y116A8C.19*—not previously known to be expressed this early in the embryo.

The C and P_3 cells are born at the eight-cell stage. Again, the P-lineage cell has new expression of fewer genes than its somatic sister. The C cell expresses more new genes than EMS, suggesting that additional repressors are removed with the progression of development. Importantly, the C and EMS newly transcribed genes share a highly significant overlap (ten genes; $p < 10^{-27}$, hypergeometric distribution; see Figure S4B). Out of the 35 C genes, 6 are transcription factors (three times more than expected; $p < 10^{-3}$, hypergeometric distribution). This suggests a somatic program evenly kick started upon divergence from P-lineage transcriptional repression, likely by PIE-1 (Seydoux et al., 1996), and a crucial role for specific transcription factors in early development. Finally, we found early expression

of 13 genes in P_3 (Table S3); of these, one—*scrm-4*—is a target of *deps-1*, a P granule-associated protein (Spike et al., 2008), suggesting that CEL-Seq is indeed detecting expression in a cell previously thought to be transcriptionally inert (Seydoux et al., 1996).

Finally, we were interested in whether transcriptomic profiles obtained with CEL-Seq can be used to computationally predict the identity of unknown cells, as can be the case when analyzing cells from complex tissues. We reasoned that distinguishing between the closely related *C. elegans* sister blastomeres would pose a veritable challenge and therefore constitute a suitable test case. In order to achieve additional statistical power, we analyzed AB and P_1 blastomeres from six more embryos. We built a classifier that uses a training set of sister blastomeres to choose a specified number of maximally informative genes (Experimental Procedures). We then used this information to predict the likely identities of new pairs of sister blastomeres. Surprisingly, in assessing the performance of our classifier using cross-validation (Figures 4D and S4C), we found that for both the two- and the eight-cell stage sisters, predictions can be made with an average 80%–90% success rate based on data from only three embryos. Using four or five replicates, the success rate increased to 90%–100%. For EMS and P_2 the classifier still did far better than guessing, but the lack of high success rates even with five replicates indicated that a systematic bias or outlier effect was present in the data. Still, the good success in highly similar blastomeres underscores the potential of our method for future transcriptomic applications.

DISCUSSION

Single-cell transcriptomics is poised to revolutionize biological research and medical practice (Tang et al., 2011; Wang and Bodovitz, 2010), allowing for unbiased and comprehensive cell characterization. It is acknowledged, that whenever possible, linear amplification by IVT is preferable to exponential amplification by PCR (Tang et al., 2011). Here, we describe a protocol that combines the power of linear amplification by IVT, with a pooling procedure that allows the efficient analysis of many samples in parallel. We presented evidence that CEL-Seq is a sensitive, accurate, and reproducible single-cell transcriptomics method. We tested CEL-Seq on mammalian cells and nematode embryonic blastomeres, made extensive use of a suite of spike-ins, and established the exact sensitivity and reproducibility of the method using extremely low amounts of purified RNA. Here, we review the advantages and limitations of the method and finally consider CEL-Seq's possible applications.

CEL-Seq's key advantage over other protocols (Islam et al., 2011; Tang et al., 2009) arises from its ability to harness the power of IVT, providing both multiplexing and reproducibility. By pooling many samples to a single IVT, a single round of amplification is sufficient, and CEL-Seq provides significantly reduced hands-on time both for the amplification and downstream processing, allowing for the preparation of dozens of samples for sequencing within 2–3 days. CEL-Seq makes use of commercially available kits for the amplification and sequencing library preparation; only the bar-coded primers and a few enzymes need to be obtained separately. This makes for the cost-effectiveness of CEL-Seq. Furthermore, the protocol is complete from start to end: the amplification, library construction, and downstream bioinformatics analysis are seamlessly connected. The amplification can be done for up to 50 cells (samples) per day by a single person. The bar-coded primers are simple to create and manufacture. Because the barcode is 8 bp in length and can be longer, the number of samples that may have unique barcodes in a given IVT is essentially unlimited. These converge to a single sample ready to interface with library preparation enabling the library preparation for ~10 of these, or ~500 cells. The library construction kit is a standard Illumina kit that itself provides an additional barcode for each library such that multiple IVTs can be analyzed together on the same sequencing lane. Finally, we provide our analysis pipeline ready for integration within the Galaxy framework, such that expression values can easily be obtained within a matter of hours.

In addition to working with single cells, CEL-Seq comes with several other desirable properties, such as strand specificity (>98% of exonic reads come from the sense strand) and barcoding efficiency (>96% of the reads contain barcodes). After amplification, the CEL-Seq protocol selects for the single 3′-most fragment of each transcript. In contrast to virtually all other RNA-Seq methods, this greatly simplifies the estimation of expression levels because no normalization by gene length is necessary. Thus, it will be of interest to invoke CEL-Seq whenever RNA amplification is necessary, even when not working with single cells.

We found that CEL-Seq outperformed STRT, a previously introduced PCR-based multiplexed single-cell RNA-Seq method (Islam et al., 2011), in terms of robustness, sensitivity, and reproducibility, and suffered from significantly less technical noise. We note that the two methods examine different ends of the mRNA transcript: 5′ for STRT, but 3′ for CEL-Seq. In addition there may have been unavoidable differences in culturing conditions of the cell types analyzed. However, several lines of evidence indicate that it is unlikely that these aspects account for the observed performance differences: (1) expression levels were not compared directly across the two methods, (2) significant differences were found for both cell types, and (3) methods were also compared based on spike-ins, which are not affected by these factors.

The limitations of CEL-Seq fall into the categories of specificity to mRNA, 3′ bias, and sensitivity to small copy numbers. CEL-Seq does not detect miRNAs and other nonpolyadenylated transcripts. This can be seen as an advantage because the bar-coded transcripts are largely depleted of rRNA (<2%), which increases the efficiency of the sequenced reads to measure mRNA levels. Due to its strong 3′ bias, the method is severely limited in its ability to distinguish alternative splice forms. Another aspect of the 3′ localization of the reads is that in species with genomes that are not well annotated, the reads will map to unannotated 3′ UTRs. This could be remedied to some extent by artificially extending transcript annotations beyond the annotated 3′ end. Finally, a crucial issue with any single-cell gene expression method is the sensitivity to the detection of lowly expressed genes. We have calculated that if the transcript is at five copies, there is a 50% chance of its identification by CEL-Seq. Relative to RNA-Seq of pooled samples, this may seem less

sensitive; however, pooling effectively increases the copy number that we have shown to be the important parameter for detection. Nevertheless, CEL-Seq on single cells will capture variation in the expression levels among cells.

EXPERIMENTAL PROCEDURES

Single-Cell Isolation

C. elegans blastomeres were isolated as previously described by Edgar (1995). Mammalian cells were obtained by trypsin treatment of adherent cells; see also the Extended Experimental Procedures. Individual cells (or media without a cell for negative control) were transferred with a micropipette into a 0.5 μl drop of egg salts or PBS for *C. elegans* blastomeres or mammalian cells, respectively, placed on the cap of a 0.5 ml LoBind Eppendorf tube, excess liquid was aspirated off, and frozen in liquid nitrogen. Samples were stored at −80°C.

CEL-Seq Primer Design

The reverse-transcription primer was designed with an anchored polyT, a unique barcode, the 5′ Illumina adaptor, and a T7 promoter. The T7 promoter sequence was as previously described by Baugh et al. (2001). The Illumina 5′ adaptor sequence was as used in the Illumina small RNA kit. The barcodes were of length eight and designed in groups of four, such that the first five nucleotides will have equal representation of all four nucleotides to allow for template generation and crosstalk corrections that are based on the first four nucleotides read in the Illumina platform. The barcodes were designed such that each pair is different by at least two nucleotides, so that a single sequencing error will not produce the wrong barcode. All used primers are described in the detailed Extended Experimental Procedures.

Linear mRNA Amplification

Ambion's MessageAmp II aRNA Kit (AM1751) was used with the following modifications. The polyT primer was replaced with the CEL-Seq primer. The reverse-transcription reaction was performed at one-tenth volume, with 5 ng of primer per reaction. A total of 0.2 μl of the primer mixed with 1 μl of water or 1 μl of a 1:500,000 dilution of the ERCC spike-in kit (a total of 1.2 μl) was added directly to the lid of the Eppendorf tube where the cell was frozen, and incubated at 70°C for 10 min (with the lid of the thermal cycler heated to 70°C). The sample was spun to the bottom of the tube midincubation. After the second-strand synthesis, samples were pooled and cleaned on a single column before proceeding to the IVT reaction at two-fifths volume for 13 hr. RNA was fragmented (one-fifth volume of 200 mM Tris-acetate [pH 8.1], 500 mM KOAc, 150 mM MgOAc added) for 3 min at 94°C, and the reaction was stopped by placing on ice and the addition of one-tenth volume of 0.5 M EDTA, followed by RNA cleanup. The RNA quality and yield were assayed using a Bioanalyzer (Agilent).

Library Construction and Sequencing

Illumina's directional RNA sequencing protocol was used with the following modifications. A total of 5 ng of RNA was used as input. The mRNA pull-down and fragmentation steps were skipped because amplified RNA represents only mRNA sequences and was already fragmented. Only the 3′ Illumina adaptor was ligated—diluted 1:5 prior to ligation to obtain the appropriate molar ratio with the reduced amount of RNA. A total of 12 cycles of PCR was performed with an elongation time of 30 s. Libraries were sequenced on the Illumina HiSeq2000 according to standard protocols. Paired-end sequencing was performed, reading at least 15 bases for read 1, and 50 bases for read 2, and the Illumina barcode when needed.

Expression Analysis Pipeline

Transcript abundances were obtained from the sequencing data using custom scripts organized into a multistep, paralleled computational pipeline within the Galaxy framework (Giardine et al., 2005). Briefly, after trimming and filtering, the paired-end reads were demultiplexed based on the first eight bases of the first read. For each sample, reads were mapped to the *C. elegans* reference genome (WS230; http://www.wormbase.org), counted using htseq-count (http://www-huber.embl.de/users/anders/HTSeq), and normalized by dividing by the total number of counted reads and multiplying with 10^6. Because CEL-Seq retains one fragment per transcript, this procedure yields the estimated gene expression levels in tpm. Absolute copy numbers were obtained by first performing least-squares linear regression on the spike-in values. The resulting factor was used to convert the tpm values to mRNA copy numbers (scripts available at yanailab.technion.ac.il).

Classification of Blastomere Identities

A machine-learning classifier was devised to predict the identities of pairs of sister blastomeres, and its performance was assessed using cross-validation. Briefly, for a given number of samples to be used as training data, and a given number of genes to be used in the prediction (*G*), the classifier first ranks genes according to the significance of their differential expression in the training data (using a paired t test). The *G*-most different genes are selected, and their mean differences between the classes are normalized by their SD, to yield a reference vector. For a new set of sister blastomeres, the score for the two possible classifications is calculated as the Euclidean distance between the differences normalized by the SD from the training data, and the reference vector. The predicted classification is chosen according to the smaller distance. The number of possible combinations of data sets to choose for the training step is N choose k, where k is the number of embryos to be used, and N is the total number of embryos analyzed. For each k, all such possibilities are tested, and the average success rate is reported.

ACCESSION NUMBERS

The NCBI SRA accession number for the sequence data reported in this paper is SRP014672.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, four figures, three tables, and detailed protocol of the CEL-Seq method and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2012.08.003.

LICENSING INFORMATION

ACKNOWLEDGMENTS

This work was supported by the Israel Science Foundation, an FP7 IRG grant and the Lorry I. Lokey Interdisciplinary Center for Life Sciences and Engineering at the Technion. We acknowledge the Gepstein and Aberdam laboratories at the Technion for assistance with the mammalian cells. We also thank Daniel Glikman for a critical reading and advice. T.H., N.S., and I.Y. conceived the method. N.S. led the development of the method. T.H. isolated the blastomeres and mouse cells, performed the validation with other methods, and managed the DNA sequencing. F.W. developed the pipeline to derive the gene expression levels and performed the quality controls. T.H., F.W., and I.Y. analyzed the data. All authors contributed to the experimental designs and the writing of the manuscript.

Received: June 12, 2012
Revised: July 18, 2012
Accepted: August 3, 2012
Published online: August 30, 2012

REFERENCES

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.

Baugh, L.R., Hill, A.A., Brown, E.L., and Hunter, C.P. (2001). Quantitative analysis of mRNA amplification by in vitro transcription. Nucleic Acids Res. *29*, E29.

Draper, B.W., Mello, C.C., Bowerman, B., Hardin, J., and Priess, J.R. (1996). MEX-3 is a KH domain protein that regulates blastomere identity in early *C. elegans* embryos. Cell *87*, 205–216.

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., Zettel, M., and Coleman, P. (1992). Analysis of gene expression in single live neurons. Proc. Natl. Acad. Sci. USA *89*, 3010–3014.

Edgar, L.G. (1995). Blastomere culture and analysis. Methods Cell Biol. *48*, 303–321.

Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome Res. *15*, 1451–1455.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. *21*, 1160–1167.

Maduro, M.F., Meneghini, M.D., Bowerman, B., Broitman-Maduro, G., and Rothman, J.H. (2001). Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in *C. elegans*. Mol. Cell *7*, 475–485.

Seydoux, G., Mello, C.C., Pettitt, J., Wood, W.B., Priess, J.R., and Fire, A. (1996). Repression of gene expression in the embryonic germ lineage of *C. elegans*. Nature *382*, 713–716.

Spike, C.A., Bader, J., Reinke, V., and Strome, S. (2008). DEPS-1 promotes P-granule assembly and RNA interference in *C. elegans* germ cells. Development *135*, 983–993.

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

Tang, F., Lao, K., and Surani, M.A. (2011). Development and applications of single-cell transcriptome analysis. Nat. Methods *8* (4, Suppl), S6–S11.

Wang, D., and Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. Trends Biotechnol. *28*, 281–290.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.