Published online March 17, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 5 ***e42***
doi:10.1093/nar/gkl050

An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis

Kazuki Kurimoto[1], Yukihiro Yabuta[1], Yasuhide Ohinata[1], Yukiko Ono[1,4], Kenichiro D. Uno[2], Rikuhiro G. Yamada[3], Hiroki R. Ueda[2,3] and Mitinori Saitou[1,5,6,*]

[1]Laboratory for Mammalian Germ Cell Biology, [2]Functional Genomics Subunit and [3]Laboratory for Systems Biology, Center for Developmental Biology, RIKEN Kobe Institute, 2-2-3 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan, [4]Department of BioScience, Tokyo University of Agriculture, Setagaya-ku, Tokyo 156-8502, Japan, [5]Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency, 4-1-8 Hon-cho, Kawaguchi, Saitama 332-0012, Japan and [6]Laboratory of Molecular Cell Biology and Development, Graduate School of Biostudies, Kyoto University, Oiwake-cho, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan

Received January 17, 2006; Revised February 3, 2006; Accepted February 23, 2006

ABSTRACT

A systems-level understanding of a small but essential population of cells in development or adulthood (e.g. somatic stem cells) requires accurate quantitative monitoring of genome-wide gene expression, ideally from single cells. We report here a strategy to globally amplify mRNAs from single cells for highly quantitative high-density oligonucleotide microarray analysis that combines a small number of directional PCR cycles with subsequent linear amplification. Using this strategy, both the representation of gene expression profiles and reproducibility between individual experiments are unambiguously improved from the original method, along with high coverage and accuracy. The immediate application of this method to single cells in the undifferentiated inner cell masses of mouse blastocysts at embryonic day (E) 3.5 revealed the presence of two populations of cells, one with primitive endoderm (PE) expression and the other with pluripotent epiblast-like gene expression. The genes expressed differentially between these two populations were well preserved in morphologically differentiated PE and epiblast in the embryos one day later (E4.5), demonstrating that the method successfully detects subtle but essential differences in gene expression at the single-cell level among seemingly homogeneous cell populations. This study provides a strategy to analyze biophysical events in medicine as well as in neural, stem cell and developmental biology, where small numbers of distinctive or diseased cells play critical roles.

INTRODUCTION

Accurate quantitative monitoring of gene expression provides an essential first step towards understanding the properties or states of cells of interest. The recent completion of the genome sequencing of many organisms and the development of microarray platforms encompassing whole-genome information have created unprecedented opportunities to perform genome-wide gene expression profiling in various biological contexts, including some disease states, thus contributing to the understanding of life at the systems level (1–3). However, these analyses often require relatively large quantities of total RNA (nanogram to microgram order, at least above the 1000-cell level) as starting materials (4–6). It remains considerably difficult to carry out these analyses from smaller samples, especially from single cells. However, analyses of small numbers of cells are very often required in various aspects of biological study, including developmental biology—where specific subsets of rare cells play essential roles—and stem cell biology in adult tissues. Moreover, the importance of the ability to analyze biological processes at the single-cell level was demonstrated recently in a typical case, where single-cell analysis revealed a signaling network's

*To whom correspondence should be addressed. Tel: +81 78 306 3376; Fax: +81 78 306 3377; Email: saitou@cdb.riken.jp

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© The Author 2006. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.001

essential properties, which population measurements had not been able to reveal (7).

Two major strategies enable the amplification of cDNAs from single cells: exponential amplification (8,9) and multiple-round linear amplification (10–12). Apparently, regarding the general nature of amplification, each method has its advantages and disadvantages compared with the other (13–15), and thus these two strategies are essentially complementary in gene expression profiling. However, especially when mRNAs from single cells are to be amplified, exponential amplification has the advantages of amplification efficiency and methodological simplicity in comparison with multiple-round linear amplification protocols (13,16,17). It should also be noted that a fully validated method for genome-wide microarray analysis of a single-cell transcriptome using multiple-round linear amplification has yet to be described.

The exponential amplification from single cells tags the 3′ ends of RT products with poly(dA) so that the cDNAs can be amplified with a single poly(dT)-tailed primer in a nondirectional manner. Exponential amplification methods amplify a wide range of mRNAs and have been used mainly to clone highly expressed genes (9,18–22), and have been applied recently to both cDNA and oligonucleotide microarrays (13,20,21). However, these methods (13,21) still introduce by-products, including amplified primer concatamers or sequences without polyadenylation signals derived from mispriming on either RNA or genomic DNA template, and they produce both systematic biases and random errors in gene representation (13,23) (see also Results). These flaws need to be reduced in order to achieve a more quantitatively accurate microarray analysis of the transcriptome.

Here, we describe a method to directionally amplify cDNAs highly representatively from single cells using relatively few PCR cycles. The amplified products are universally applicable to oligonucleotide microarrays after isothermal production of cRNAs from the 5′-terminally positioned T7 promoter, using the standard labeling protocol supplied by the manufacturers of prevalent microarray platforms. The amplified products showed high coverage and accuracy, with higher representation and reproducibility compared with those amplified by Brady and Iscove's original method (8), which was utilized by Tietjen *et al*. (21). We amplified cDNAs from single cells of morphologically undifferentiated inner cell masses (ICMs) of mouse blastocysts at embryonic day (E) 3.5 and applied these cDNAs to GeneChip microarrays, which identified the gene expression profiles indicating overt differentiation of the morphologically identical cells toward a fate of either primitive endoderm (PE) or epiblast. This method is generally applicable to many important biological questions that require accurate transcriptome analysis at a single-cell resolution.

MATERIALS AND METHODS

cDNA synthesis from single cells or single-cell level total RNA

Total RNA was purified from ES cells using the RNeasy Mini kit (Qiagen, Hilden, Germany). For the preparation of diluted RNA, we serially diluted the total RNA of ~1000 ng/μl to concentrations of 2.5 ng/μl, 250 pg/μl and 25 pg/μl. Then, 0.4 μl (10 pg) of the final dilution (25 pg/μl) was directly added to the single-cell lysis buffer (see below).

Mouse embryos at E3.5 blastocyst (C57BL/6) were collected in DMEM (Gibco, Gaithersburg, MD) with 0.5% BSA. Embryonic fragments containing ICM were cut out by a glass needle and incubated with 0.05% trypsin and 0.5 mM EDTA for 7 min, followed by dissociation into single cells by a mouth pipette. Dissociated single cells were randomly picked up for single-cell cDNA synthesis. The entire process was performed as quickly as possible in order to minimize the effect of trypsin/EDTA treatment on gene expression.

The cDNA synthesis and exponential amplification procedures were modified from those in previous reports (18,24). Single cells isolated from an embryo, or a single-cell equivalent amount of RNA, were seeded into 0.5 ml thin-walled PCR tubes containing 4.5 μl of cell lysis buffer [1× PCR buffer II (Applied Biosystems, Foster City, CA), 1.5 mM $MgCl_2$ (Applied Biosystems), 0.5% NP40, 5 mM DTT, 0.3 U/μl Prime RNase Inhibitor (Eppendorf, Hamburg, Germany), 0.3 U/μl RNAguard RNase Inhibitor (Amersham Biosciences, Piscataway, NJ), 0.2 ng/μl primer V1$(dT)_{24}$ and 0.05 mM each of dATP, dCTP, dGTP and dTTP], containing an appropriate amounts of spike RNAs (see below). The sequence of the V1 $(dT)_{24}$ primer was 5′-ATATGGATCCGGCGCGCCGTCGACTTTTTTTTTTTTTTTTTTTTTTTT-3′. All the primers described in this paper were purchased from Operon Biotechnology (Huntsville, AL) or Hokkaido System Science (Sapporo, Japan). After 15 s centrifugation, cell lysis was performed at 70°C for 90 s, and the reaction tubes were immediately put on ice for 1 min. A 0.3 μl volume of RT mixture [133.3 U/μl SuperScript III (Invitrogen), 3.33 U/μl RNAguard RNase Inhibitor (Invitrogen, Carlsbad, CA), and 1.1–1.3 μg/μl T4 gene 32 protein (Roche, Basel, Switzerland)] was added to each reaction tube. The reaction mixture was incubated at 50°C for 5 min and heat-inactivated at 70°C for 10 min. The tubes were immediately put on ice for 1 min, and after 15 s centrifugation, 1.0 μl of Exonuclease I mixture [1× Exonuclease I buffer (Takara, Shiga, Japan) and 0.5 U/μl Exonuclease I (Takara)] was added to each tube. The reaction mixture was incubated at 37°C for 30 min and heat-inactivated at 80°C for 25 min. The reaction tubes were then put on ice for 1 min. Next, 6 μl of terminal deoxynucleotidyl transferase (TdT) mixture [1× PCR buffer II, 1.5 mM $MgCl_2$, 3 mM dATP, 0.1 U/μl RNaseH (Invitrogen) and 0.75 U/μl TdT (Invitrogen)] was added to each tube, and the mixture was incubated at 37°C for 15 min followed by heat-inactivation at 70°C for 10 min. The synthesized poly(dA)-tailed RT product in each tube (12 μl) was divided into four 0.2 ml thin-walled PCR tubes (3 μl each). Then, 19 μl of PCR mixture I [1× *ExTaq* buffer, 0.25 mM each of dATP, dCTP, dGTP and dTTP, 0.02 μg/μl primer V3 $(dT)_{24}$, and 0.05 U/μl *ExTaq* Hot Start Version (Takara)] was added to each tube for the first round of PCR: 95°C for 3 min, 50°C for 2 min and 72°C for 3 min. The sequence of V3 $(dT)_{24}$ was 5′-ATATCTCGAGGGCGCGCCGGATCCTTTTTTTTTTTTTTTTTTTTTTT-3′. The tubes were immediately put on ice for 1 min, and 19 μl of PCR mixture II was added, with a composition almost the same as that of PCR buffer I but with primer V1 $(dT)_{24}$ replacing primer V3 $(dT)_{24}$. A drop of mineral oil (Sigma-Aldrich, St Louis, MO) was then added to each tube. A 20-cycle PCR amplification was then

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.002

performed according to the following schedule: 95°C for 30 s, 67°C for 1 min and 72°C for 3 min with a 6 s extension per cycle. The amplified cDNA was purified with a QIAquick PCR purification kit (Qiagen) and dissolved in 50 μl of buffer EB (10 mM Tris–HCl, pH 8.5).

The resultant cDNA was subjected to another amplification step to allocate the T7 promoter sequence at the 5′-terminus. A 49.4 μl volume of PCR mixture III [1× *ExTaq* buffer, 0.25 mM each of dATP, dCTP, dGTP and dTTP, 0.02 μg/μl primer T7-V1 (5′-GGCCAGTGAATTGTAATACGACTCACTATA GGGAGGCGGATATGGATCCGGCGCGCCGTCGAC-3′), 0.02 μg/μl primer V3 $(dT)_{24}$ and 0.05 U/μl *ExTaq* Hot Start Version] was added to each of eight 0.2-ml thin-walled PCR tubes containing 0.63 μl of the 20 cycle amplified cDNA. A nine-cycle amplification was then performed according to the following schedule: 95°C for 5 min 30 s, 64°C for 1 min and 72°C for 5 min 18 s for the first cycle; and 95°C for 30 s, 67°C for 1 min and 72°C for 5 min 18 s with an extension of 6 s per cycle for another eight cycles. The products were mixed together after the reaction, purified with a QIAquick PCR purification kit, and dissolved in 30 μl of buffer EB. The PCR product was purified with 2% agarose gel electrophoresis to remove by-product DNA shorter than 300 bp. The cDNA was extracted from a gel fragment with a QIAquick Gel Extraction kit (Qiagen) and dissolved in 35 μl of buffer EB. A 47.8 μl volume of PCR mixture III was added to each of four 0.2 ml thin-walled PCR tubes containing 2.2 μl of the purified cDNA, and an additional one-cycle PCR (95°C for 5 min 30 s, 67°C for 1 min and 72°C for 16 min) was performed. The products were mixed together after the reaction, purified with the QIAquick PCR purification kit, and dissolved in 30 μl of buffer EB.

To prepare the spike RNAs, *Escherichia coli* cells containing plasmids encoding poly(A)-tailed *Bacillus subtilis lys*, *phe*, *thr*, and *dap* genes were purchased from the American Type Culture Collection (ATCC, Manassas, VA; the ATCC numbers were 87482, 87483, 87484 and 87486, respectively). The sense-strand RNAs were synthesized with the MEGAscript T3 kit (Ambion, Austin, TX) and purified with the RNeasy Mini kit. An appropriate amount of spike RNA mixture was added to the cell lysis buffer and to 5 μg (5×10^5 cells) of total RNA for the microarray experiments, so that the reaction mixture contained poly(A)-tailed *Lys*, *Dap*, *Phe* and *Thr* RNAs at 1000, 100, 20 and 5 copies per cell, respectively.

The single-cell cDNA synthesis with the original protocol was performed exactly as described elsewhere (21).

Quantitative real-time PCR analysis

Quantitative real-time PCR (Q-PCR hereafter) was performed using the 7900 Real Time PCR System (Applied Biosystems) according to the manufacturer's instructions. The sequences of the primers are listed in Table 1. The doubling efficiencies of all primers were measured (Table 1). Using these efficiencies, the threshold cycle (Ct) values were corrected to calculate the expression levels.

Microarray hybridization and data processing

Eight independently amplified cDNA samples and ES cellular total RNA (5 μg in each of eight individual tubes) were subjected to the One-Cycle Target Labeling procedure for biotin labeling by *in vitro* transcription (IVT) (Affymetrix, Santa Clara, CA). The cRNA was subsequently fragmented and hybridized to the GeneChip Mouse Genome 430 2.0 array (Affymetrix) according to the manufacturer's instructions. The microarray image data were processed with the GeneChip Scanner 3000 (Affymetrix) to generate CEL data. The CEL data were then subjected to analysis with dChip software (25), which has the advantage of normalizing and processing multiple datasets simultaneously. Data obtained from the eight nonamplified controls from ES cells, from the eight independently amplified samples from the diluted ES cellular RNA, and from the amplified cDNA samples from 20 single ICM cells were normalized separately within the respective groups, according to the program's default setting. The model-based expression indices (MBEI) were calculated using the PM/MM difference mode with log-2 transformation of signal intensity and truncation of low values to zero. The absolute calls (Present, Marginal and Absent) were calculated by the Affymetrix Microarray Software 5.0 (MAS 5.0) algorithm using the dChip default setting. The expression levels of only the Present probes were considered for all quantitative analyses described below. The GEO accession number for the microarray data is GSE4309.

Calculation of coverage and accuracy

The true positive was defined as probes called Present in at least six of the eight nonamplified controls (see also Results), and the true expression levels were defined as the log-averaged expression levels of the Present probes. The definition of coverage is (the number of truly positive probes detected in amplified samples)/(the number of truly positive probes). The definition of accuracy is (the number of truly positive probes detected in amplified samples)/(the number of probes detected in amplified samples). The expression levels of the amplified and nonamplified samples were divided by the class interval of $2^{0.5}$ (2^0, $2^{0.5}$, 2^1, $2^{1.5}$ ...), where accuracy and coverage were calculated. These expression level bins were also used to analyze the frequency distribution of the detected probes.

Microarray data comparison

The dataset from a previous single-cell microarray study [Supplementary Data set S6 (21)] was used for the comparison between the method applied by Tietjen *et al*. (21) and our own. In that previous study, the expression levels (average difference values; ADV) and absolute calls were generated by the Affymetrix GeneChip System v3.2 algorithm, and the default setting of the data analysis parameters was changed to lower the stringency of the Absolute call in order to compensate for the lack of full-length transcripts in their single-cell cDNAs (21). The expression levels of only the Present probes were considered for all the quantitative analyses. All the data were analyzed with Microsoft Excel. The datasets of the amplified samples (10 pg; $N = 6$) were normalized so that the mean ADVs became equal; the mean ADVs were 3008.6 and 4196.0 before and after the normalization, respectively. The true positive was defined as the Present probes in the undiluted total RNA (tot Hu RNA; $N = 1$). The ADV bins were normalized so that the mean expression levels of the data became equal to that of our present data (see the paragraph below) to compare the probe distributions, coverage and accuracy of

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.003

Table 1. Gene-specific primers used for screening and Q-PCR of single-cell cDNA

Gene	5′ primer	3′ primer	A[a]	B[a]
Aqp8	TTGCGAGAGGGCAGGGATTA	TGTGCATGAATTGGGTTCCA	−0.72	35.13
AV101904	ACAGATTGTGTTGATTGACCCTTCC	CCACCTCTTAGCTAAATTGTCTTGA	−0.97	33.10
BB242234	GGCAATTACGATAAGGAAACCCTTTA	GCAAATAAAGCCACGCCACA	−0.97	33.37
BC017612	GGGTGATGCTGTTGGGATCA	TCAAATCCCATCAAGCACAGAGA	−0.86	33.23
Bcl2l14	TGAGGGACGTGGACACCAGA	TCCCGAAGCCCAGCATTCTA	−1.02	34.83
Bcl7a	GGATGGCGGCTACACATTCC	CCTCTACCAGGTCCTCTGCGATA	−0.93	33.97
c-Myc	AAGGAGAACGGTTCCTTCTGAC	GCTGAAGCTTACAGTCCCAAAG	−0.93	32.42
Cubn	ATCTCTGCCCACGCCATCA	ACCACCAGGCTCTGCCTTCA	−0.99	34.70
Dap	CCAGACCGCGGCCTAATAATG	CGCTTCTTCCACCAGTGCAG	−1.04	35.21
Dnmt1	GGCCATGGCTGACACTAAGCTG	CACCTGCACAGTGGCAGATCTG	−0.93	32.37
Dnmt3a	GACTCGCGTGCAATAACCTTAG	GGTCACTTTCCCTCACTCTGG	−1.03	35.08
Dnmt3b	CTCGCAAGGTGTGGGCTTTTGTAAC	CTGGGCATCTGTCATCTTTGCACC	−1.04	34.88
Dnmt3l	CCAGGGCAGATTTCTTCCTAAGGTC	TGAGCTGCACAGAGGCATCC	−0.93	33.22
Eras	GTAGCTGTGGCTGCTCTGTAG	GATGTCTGTGGTAACTTGGTCG	−0.92	33.79
Esg1	AAGGAGTGCTGAAGCTGGAGG	CAGCTTAACCTGCATCCAGGTC	−0.90	32.06
Ezh2	TCAGGAACCTTGAGTACTGTGG	CTTTGCAGCTGGTGAGAAGGC	−0.86	36.81
Fgf4	AAGGCACCTGCCCTGTTCTG	GGGAGCTAGCTGGCTGAAGAAA	−0.92	32.59
Fgfr2	CCTTCTGCCCGGGTTACACA	GATGCTGGGCTTTTGCCATC	−0.96	34.09
Foxh1	GACCTGCTCTGTGATCTAGAC	ATGCTGTACCAGGAAAGGAGC	−0.92	33.70
fragilis	TGGTCCTCAGCATCCTGATGG	AGGGTGAAGCACTTCAGGACC	−0.96	34.05
G9a	CTTCTTCAGCTCCAGGGACATC	GAATGCTTGCACTTCTCAGAGC	−0.98	35.44
Gapdh	ATGAATACGGCTACAGCAACAGG	CTCTTGCTCAGTGTCCTTGCTG	−1.04	35.51
Gata4	CCTAAACCTTACTGGCCGTAGC	ACAATGTTAACGGGTTGTGGAG	−1.00	33.83
Gata6	CACAGTCCCCGTTCTTTTACTG	GTGGTACAGGCGTCAAGAGTG	−1.04	34.74
Hhex	ACTTGGCTCCCGGTGTCTGTT	TGAGTCAGTTTCCCTGCCTGT	−0.92	31.96
Hnf4a	AGAACCTTTCAGGGTTCAGGAG	GCCACAGAGAGCTCTAGCAAAG	−0.97	34.22
Klf2	TCGAGGCTAGATGCCTTGTGA	AAACGAAGCAGGCGGCAGA	−1.03	34.62
Jak1	ACCGAATGATCAGCTGCATAGC	ACATTGAGATTTCAGGGCAAGG	−1.04	38.39
Lama1	ACAACGGTGCCGGAAGGATA	CGGAATTCCCGTCCACAGTC	−0.80	34.31
Lefty1	TGTCGCTGAATCTGGGCTGAG	TAGCAAAGCCAGTATTGCCTAG	−1.01	34.58
Lys	GCCATATCGGCTCGCAAATC	AACGAATGCCGAAACCTCCTC	−1.04	35.21
nanog	CTTTCACCTATTAAGGTGCTTGC	TGGCATCGGTTCATCATGGTAC	−0.93	34.73
Ndg1	TCAGCAATGGACATTGACTTTCG	CCCAGCATTGCCTATACCAGAGA	−0.97	33.91
nodal	AGCCACTGTCCAGTTCTCCAG	GTGTCTGCCAAGCATACATCTC	−0.88	32.55
Oct4	GATGCTGTGAGCCAAGGCAAG	GGCTCCTGATCAACAGCATCAC	−0.93	34.82
Pdgfra	GTTTCCAGGGCATGGGTGAG	AAGGAAGCACACGGGTGGAC	−1.01	34.02
Pfkl	GAATCTGCGGCTGATGCTGA	CCGGGAACGCAGGAACAGTA	−0.80	32.18
Phe	TGAGCTCTAGGCCCAAAACGAC	TCCGGTTTTAGTCGGACGTG	−1.01	35.21
Prkcz	CGACCAGTCCGAATTTGAAGGCTTTG	TCACAGGCGTGTCCACAACA	−0.81	32.45
Pthr1	GCCCAATGGTGTCACACACG	CCGTCGTCCTTGGGAACTGTC	−0.92	33.64
Rex1	TCCATGGCATAGTTCCAACAG	TAACTGATTTTCTGCCGTATGC	−1.02	33.89
Rhebl1	TGGGCACACAGTATGGCACTC	GGGTGGGTCCTGGGTAATCA	−0.98	33.76
Rhpn2	TGACAGAGTCACGGTTGACACA	CAGGGATAGCGGGCACACTTTC	−0.97	33.72
Runx1	TACTGAGCTGAGCGCCATCG	CCTCCGGGATTCTTCCTGG	−0.86	32.72
Serpinh1	CCCGAGCCCTCTTCAGTCTTC	CATGCACTGTCTCCCAGCTC	−0.81	34.09
Soat1	TGTCTAATGCGTAGTGACTTTCCTTG	GAAGCCAGCTTCTGGAGCCTTA	−0.91	34.04
Sox17	TTCTGTACACTTTAATGAGGCTGTTC	TTGTGGGAAGTGGGATCAAG	−1.05	35.18
Sox2	CATGAGAGCAAGTACTGGCAAG	CCAACGATATCAACCTGCATGG	−0.99	35.63
SpiC	GGCAACCGGAAACCCATGAC	TGGAGAACAGCCTCGCTGAA	−0.94	32.66
Stella	AGGCTCGAAGGAAATGAGTTTG	TCCTAATTCTTCCCGATTTTCG	−1.02	34.27
Thr	GCCGATGCCGTAAAAGCAAG	CAGCTCAGGCACAAGCATCG	−1.05	35.21
Tiar	GATTGTAATTCACCACAGGCTG	GCAGAACTAGAAGACATGTGCC	−0.99	32.79
Tnap	TACTCCACTGGCCTGTCTCTG	GCATCCTTGTAAACATCAGCACG	−1.00	33.26
Tyk2	ACACACACAGGAAGGCCATTTG	AGGGCCAGAGAGGTCAAGACAG	−1.03	34.65
vHnf1	GCAGTATTGTCATGATGGCTCTC	TGCATCAGTTTGTTCGATGATG	−0.85	30.68
Yy1	TTCAGTTACAGAAAGTGGTGCTC	AAATAAGGCTGTGCTTACAGAGC	−0.92	33.64

[a]Log_2 [copy number of the Q-PCR template (10μl scale reaction)] $= A \times \text{Ct} + B$.
Coefficient A represents the $\log_2$-transformed doubling efficiency of the primer set. Coefficient B represents the $\log_2$-transformed copy number of PCR product detectable at the threshold level with SYBR green.

both studies. The Pearson's correlation coefficient (R) was calculated for the pair-wise comparisons of amplified samples ($R^2 = 0.69 \pm 0.06$). The R^2 values were also calculated to compare the log-averaged ADVs of the amplified samples and the undiluted total RNA, and the best value was obtained under the definition of detection in which ⩾1/6 amplified samples are called Present (0.41 for log-averaged ADVs). Outliers in the undiluted sample (M11147_at, U57341_r_at, X01677_f_at, X00351_f_at) were excluded in this analysis because they severely affect the R^2 values ($R^2 = 0.32$ when outliers are included).

To make the data obtained in this study comparable with those in the previous study, we re-calculated the expression levels of our data using the MAS 5.0 algorithm using the Affy

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.004

package of the Bioconductor program suite for the R package (http://www.bioconductor.org) (26), instead of Li and Wang's model-based algorithm (25). The data normalization for the amplified and nonamplified samples ($N = 8$ each) was performed as described for the previous data; the means of ADV before and after the normalization were 3533.2/3512.7 and 1629.2/1634.7, respectively. All the statistics were re-calculated using these data.

Analysis of gene expression profiles of E3.5 blastocysts

The unsupervised clustering and class neighbor analyses of the microarray data from E3.5 blastocysts were performed using GenePattern software (http://www.broad.mit.edu/cancer/software/genepattern/), which performs the signal-to-noise ratio analysis/*T*-test in conjunction with the permutation test to preclude the contribution of any sample variability, including those from methodology and/or biopsy, at high confidence. The analyses were conducted on the 14 128 probes for which at least 6 out of 20 single ICM cells provided Present calls and at least 1 out of 20 samples provided expression levels >20 copies per cell. The expression levels calculated for probes with Absent/Marginal calls were truncated to zero.

To calculate relative gene expression levels, the Ct values obtained with Q-PCR analyses were corrected using the efficiencies of the individual primer pairs quantified either with whole mouse genome (BD Biosciences) or plasmids that contain gene fragments (Table 1). The relative expression levels were further transformed into copy numbers with a calibration line calculated using the spike RNAs included in the reaction mixture ($\log_{10}$[expression level] = 1.05 × $\log_{10}$[copy number] + 4.65).

The Chi-square test for independence was performed to evaluate the association of gene expressions with *Gata4*, which well represents the difference between cluster 1 and cluster 2 determined by the unsupervised clustering and which is restricted to PE at later stages. The expression levels of individual genes measured with Q-PCR were classified into three categories: high (>100 copies per cell), middle (10–100 copies per cell), and low (<10 copies per cell). The Chi-square and *P*-values for independence from *Gata4* expression were calculated based on this classification. Chi-squared was defined as follows: $\chi^2 = \Sigma\Sigma\,(n f_{ij} - f_i f_j)^2 / n f_i f_j$, where i and j represent expression level categories (high, middle or low) of the reference (*Gata4*) and the target gene, respectively; f_i, f_j, and f_{ij} represent the observed frequency of categories i, j and ij, respectively; and n represents the sample number ($n = 24$). The degrees of freedom were defined as $(r - 1) \times (c - 1)$, where r and c represent available numbers of expression level categories of *Gata4* and of the target gene, respectively.

RESULTS

Design and improvement of the new method

To improve the original exponential amplification method from single-cell-level mRNA, we set out to design an amplification strategy that used fewer cycles so as to reduce the risk of introducing bias, and that used two different amplification primers so that the resultant directional cDNAs would be applicable to further isothermal linear amplification from the 5′-terminally allocated T7 promoter (Figure 1C). We isolated total RNA from ES cells (1×10^7) and diluted it to the single-cell level (~10 pg) for use as amplification templates. Using Q-PCR, the representation and reproducibility of amplification were examined by comparing the abundance of gene products in unamplified cDNAs synthesized from undiluted total RNA versus that in amplified cDNAs from the diluted RNA (Figure 1A).

First, to ensure uniform amplification efficiency for all the mRNA species, we set the initial RT reaction to be as short as 5 min, which resulted in PCR products within the range of ~500–1500 bp (not shown). Second, to reduce the risk of distorted representation, as delineated in Figure 1B, we defined the number of cycles needed to obtain a sufficient amount of cDNA for microarray analysis. In an ideal amplification, the amount of cDNA doubles in each cycle. Therefore, single-cell mRNAs (~0.2 pg) could be amplified up to ~200 ng (~10 μg as total RNA) (1×10^6-fold) after a 20-cycle PCR, and thus we used 20 cycles as the initial amplification step for each experiment. Third, we found that both systematic and random errors of amplification depend not only on the number of cycles but also largely on the primer sequences (data not shown). We set up >20 different primer pairs to find an appropriate set of primers. Furthermore, after examining >30 sequences from the amplified products, we found significant by-product contamination; the first primer molecules that were not annealed to mRNAs were tailed with poly(dA) by TdT and amplified efficiently by the subsequent PCR. Therefore, we included Exonuclease I treatment (a 3′–5′ exonuclease) after reverse transcription to specifically degrade the unreacted first primer. This treatment dramatically reduced the amplification of by-product, which was removed by gel purification (Materials and Methods). In examining 40 sequences from the amplified product, we found that none of them contained by-product and that sense–antisense orientation was completely preserved (Figure 2 and Table 2). The BLAST search analysis found that 97.5% of the sequenced cDNAs (39/40) had bona fide transcript ends (Figure 2 and Table 2). The average length of a sequenced PCR product was 854 ± 276 bp (Figure 2 and Table 2), consistent with the observation mentioned above. Considering that the original method resulted in by-product occupying 70% of the amplified cDNA pool (13), the present result shows a clear improvement in amplification. Moreover, the efficiency of the cDNA amplification was also enhanced by the Exonuclease I treatment (data not shown). Finally, to average the intrinsic variation during PCR (23), PCR amplification was performed in four independent tubes, and the products were mixed together after the reaction. With all of these modifications, the final scheme of our method involved first-strand cDNA synthesis by reverse transcription using poly(dT)-tailed V1 primer, removal of unreacted V1 primers by Exonuclease I digestion, the addition of poly(dA) to the 3′ end of the first-strand cDNA by TdT, and second-strand synthesis using another poly(dT)-tailed primer, V3, followed by PCR with the V1$(dT)_{24}$ and V3$(dT)_{24}$ primers in four independent tubes (Figure 1C).

To examine the performance of our method, we compared the representations of 23 genes typically expressed in ES cells at widely differing levels (*Gapdh, Oct4, nanog, Sox2, Ezh2, Tiar, fragilis, Yy1, stella, Tnap, Esg1, Eras, cMyc, Foxh1,*

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.005

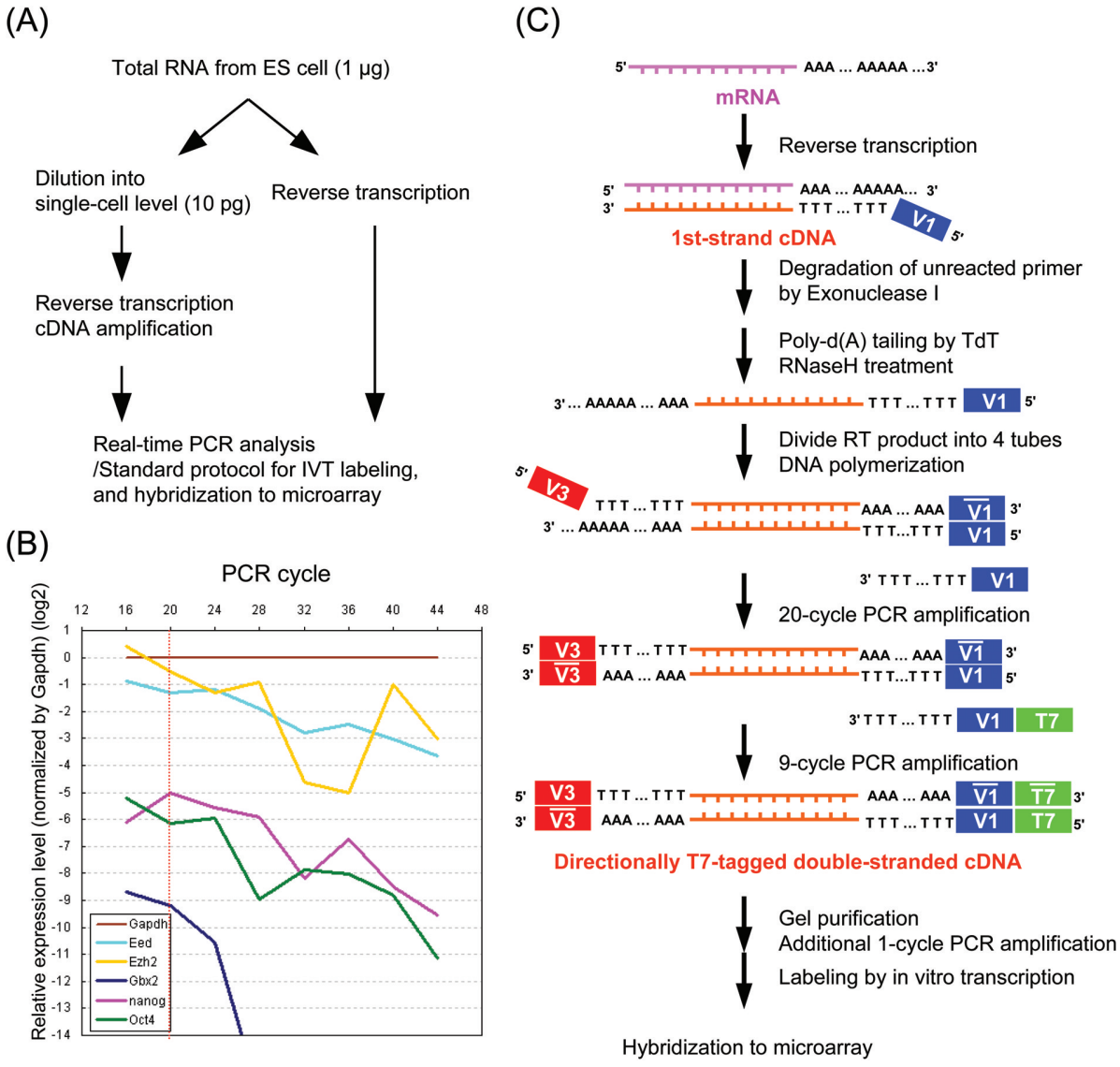


Figure 1. Schematic diagram of the key features of the global cDNA amplification method. (**A**) Evaluation system to verify representation of amplified cDNA from diluted ES cellular RNA by Q-PCR and/or microarray. (**B**) Gene representation distorted during the global PCR. Diluted ES cellular RNA (10 pg) was amplified as described elsewhere (21), and the replicates of amplification were sequentially sampled at 16, 20, 24, 28, 32, 36, 40 and 44 cycles. The expression levels of *Gapdh*, *Eed*, *Ezh2*, *Gbx2*, *nanog* and *Oct4* were measured by Q-PCR, normalized by that of *Gapdh*, and represented with brown, cyan, yellow, blue, pink and green lines, respectively. The averages of four independent experiments are plotted. (**C**) Schematic diagram of cDNA amplification. The mRNA and cDNA are colored pink and orange, respectively. The V1, V3 and T7 promoter sequences are represented by blue, red and green boxes, respectively. The bars above the letters represent the complementary sequences.

Dnmt1, Dnmt3b, G9a, Jak1A, nodal, lefty1, Tyk2, Fgf4, Rex1) and four spike RNAs, which are artificially poly(A)-tailed prokaryotic mRNAs that are often used as controls in microarray experiments (*B.subtilis* Lys, Dap, Phe and Thr RNAs of 1000, 100, 20 and 5 copies per cell, respectively) between 10 independently amplified cDNAs and nonamplified cDNAs (synthesized from 1 µg of total RNA) (Figure 3). We also amplified the diluted RNA using the original method (10 independent amplifications) (24) as described previously (21) and compared the amplification quality between our method and the original one. As shown in Figure 3, with our method, representation was generally preserved in almost all the expression-level ranges, with small variations among experiments, especially for genes with greater than 20 copies per cell. The majority of the genes were plotted between 4-fold difference lines compared with the nonamplified control in each experiment ($R^2 = 0.734$, with 65 and 89% of the genes plotted between 2.0- and 4.0-fold difference lines, respectively). In sharp contrast, although the original method was generally successful on highly expressed genes and detected the genes in all the expression-level ranges at a rate comparable with the new method (88%), it introduced large variations in representation on genes with fewer than a few hundred copies per cell ($R^2 = 0.496$, with 7 and 22% of the genes plotted between 2.0- and 4.0-fold difference lines, respectively). These results indicate that our method amplifies

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.006

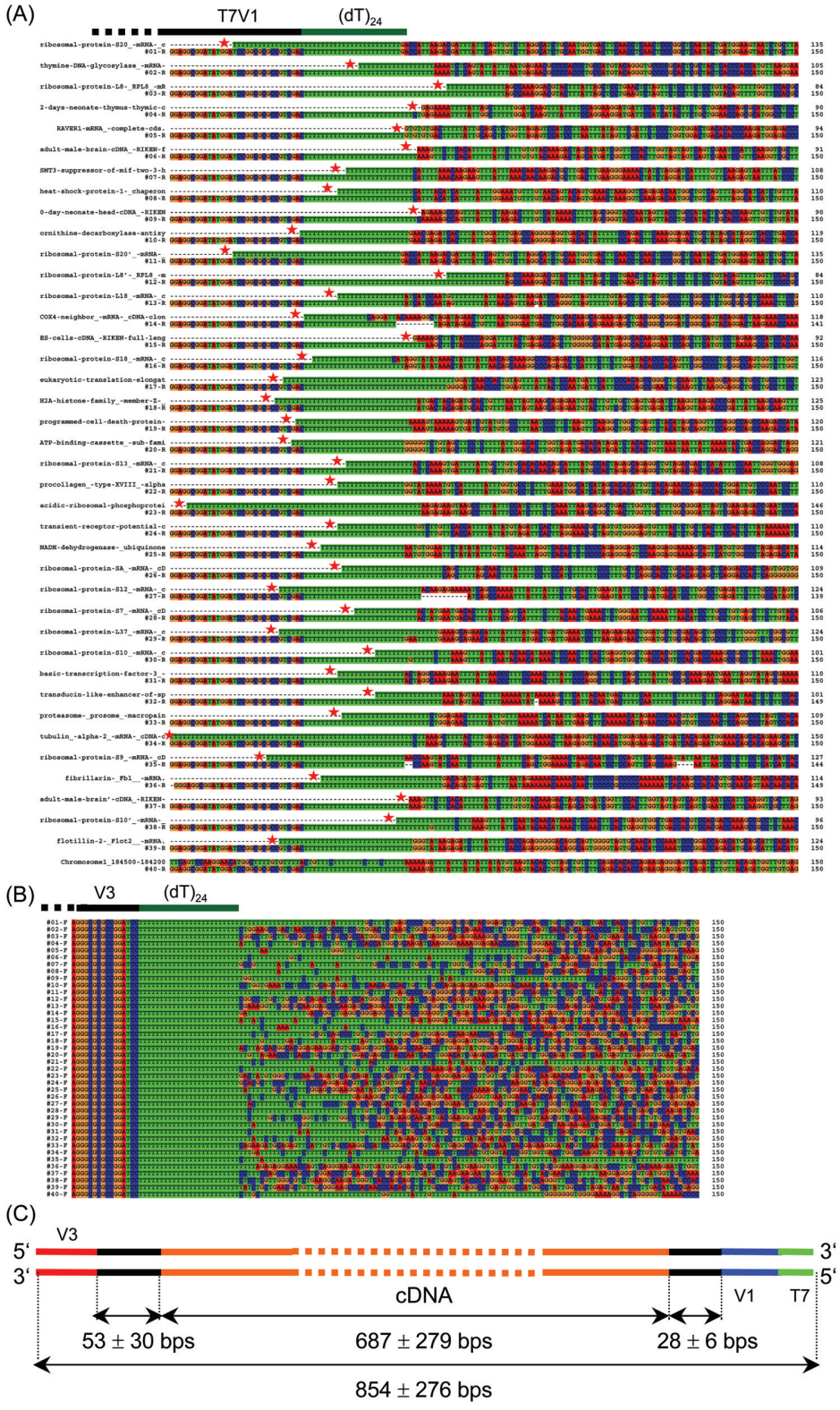


Figure 2. Sequence analysis of amplified cDNAs. (**A**) Sequence of the amplified cDNAs aligned with the complementary sequences of the corresponding 3′ transcript ends. A, C, G and T are represented by letters in red, blue, orange and green boxes, respectively. The cDNA sequences from the NCBI database (upper) and the PCR products (lower) are aligned in a pairwise manner. The transcript ends in cDNA sequences from the NCBI database are indicated by red stars. (**B**) Sequence data on the 5′ ends of the amplified transcripts. The nucleic acids are represented in the same manner as in (A). (**C**) Schematic summary of the PCR products. The average nucleotide lengths measured by the sequencing of the 40 cDNAs are indicated. Blue, red and green bars represent V1 and V3 primer sequences and T7 promoter sequence, respectively. The poly(dA/dT) tracts of variable lengths are represented with black bars. The cDNA body is represented with orange bar.

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.007

Table 2. Annotation of cDNAs amplified from 10 pg of total ES cellular RNA

No.	Annotation	Accession	PCR product length (bp)	cDNA length (bp)
1	*Mus musculus* ribosomal protein S20 (Rps20)	BC090389	665	494
2	*M.musculus* thymine DNA glycosylase	BC085470	1258	1116
3	*M.musculus* ribosomal protein L8 (Rpl8)	U67771	994	835
4	*M.musculus* ribosomal protein L11 (Rpl11)	BC025077	738	588
5	*M.musculus* mRNA for mKIAA1978	AY275472	788	584
6	*M.musculus* ribosomal protein S29	BC024393	481	289
7	*M.musculus* SMT3 suppressor of mif two 3 homolog 2 (yeast) (Sumo2)	BC017522	1152	983
8	*M.musculus* heat shock protein 1 (chaperonin 10) (Hspe1)	BC024385	782	589
9	*M.musculus* 0 day neonate head cDNA, RIKEN full-length enriched library	AK161234	1310	1089
10	*M.musculus* ornithine decarboxylase antizyme (Oaz1)	NM_008753	1103	967
11	*M.musculus* ribosomal protein S20	BC090389	665	496
12	*M.musculus* ribosomal protein L8 (RPL8) mRNA	U67771	995	836
13	*M.musculus* ribosomal protein L18	BC082290	710	575
14	*M.musculus* neighbor of Cox4	BC009103	1105	955
15	*M.musculus* ES cells cDNA, RIKEN full-length enriched library	AK131947	792	611
16	*M.musculus* ribosomal protein S18	BC081458	738	540
17	*M.musculus* eukaryotic translation elongation factor 2 (Eef2)	NM_007907	1652	1490
18	*M.musculus* H2A histone family, member Z	NM_016750	1117	945
19	*M.musculus* programmed cell death protein 11	BC051231	565	428
20	*M.musculus* ATP-binding cassette, sub-family F (GCN20), member 2 (Abcf2)	NM_013853	934	797
21	*M.musculus* ribosomal protein S13	BC090397	754	510
22	*M.musculus* procollagen, type XVIII, alpha 1 (Col18a1)	NM_009929	1400	1230
23	*M.musculus* acidic ribosomal phosphoprotein P0	BC011291	1062	923
24	*M.musculus* transient receptor potential cation channel, subfamily C, member 2	BC067003	947	807
25	*M.musculus* NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 7 (B14.5a)	BC055698	626	483
26	*M.musculus* laminin receptor 1 (ribosomal protein SA)	BC081461	468	297
27	*M.musculus* ribosomal protein S12	BC092044	649	490
28	*M.musculus* ribosomal protein S7	BC002014	837	664
29	*M.musculus* ribosomal protein L37	BC054388	517	354
30	*M.musculus* ribosomal protein S10	BC019725	738	561
31	*M.musculus* basic transcription factor 3	BC080837	1029	859
32	*M.musculus* transducin-like enhancer of split 3, homolog of *Drosophila* E(spl)	BC006672	830	652
33	*M.musculus* proteasome (prosome, macropain) 26S subunit,non-ATPase, 3	BC003197	830	689
34	*M.musculus* tubulin, alpha 2 (Tuba2)	BC108394	616	436
35	*M.musculus* ribosomal protein S9	BC031746	945	697
36	*M.musculus* fibrillarin	NM_007991	997	849
37	*M.musculus* adult male brain cDNA	AK002939	420	284
38	*M.musculus* ribosomal protein S10	BC019725	722	575
39	*M.musculus* flotillin 2 (Flot2)	NM_008028	860	722
40	*M.musculus* Chromosome 1	AC107762	388	179

cDNAs more representatively and reproducibly than the original protocol.

Evaluation of representation in amplified cDNAs by microarray analysis

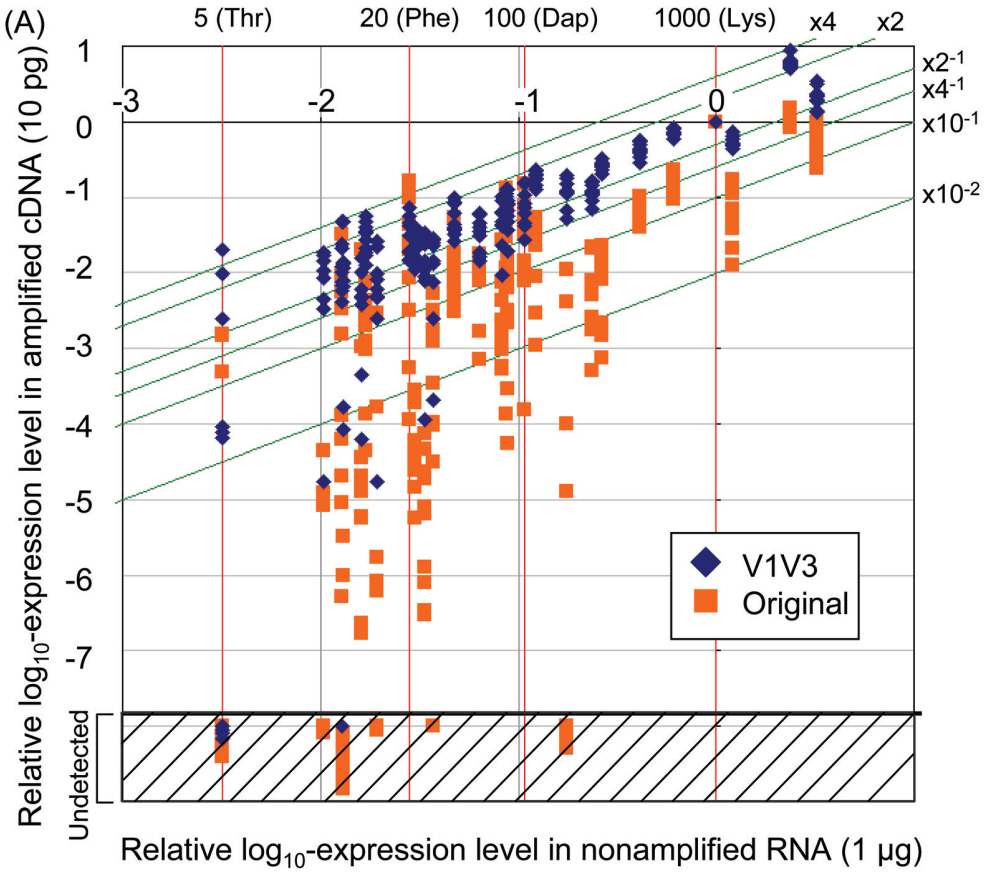
To examine the quality of the amplified products using microarray analysis, the T7 promoter was allocated to the 5′-terminal end of the amplified cDNA with additional PCR using T7-V1 and V3 primer pairs (Materials and Methods). The cRNAs were synthesized by IVT from eight independently amplified cDNA samples and applied to Affymetrix GeneChip Mouse Genome 430 2.0 microarray slides. We also synthesized eight labeled cRNA samples, each from 5 μg of undiluted total RNA, and hybridized them to the microarray as nonamplified controls.

The normalized scatter plots (Figure 4A) showed a good correlation between the two independently amplified samples ($R^2 = 0.89 \pm 0.01$), with 64 and 83% of the probe sets called Present plotted between 2.0- and 3.5-fold difference lines, respectively. The log-averaged expression levels of the eight nonamplified controls and the eight independently amplified samples were also considerably correlated ($R^2 = 0.70$) (Figure 4B), with 84% of the probe sets detected in both (see below for our definition of detection) plotted between the 3.5-fold difference lines. Furthermore, the averaged log-expression levels of the spike RNAs (1000, 100, 20 and 5 copies per cell) were highly proportional to their copy numbers ($R^2 = 0.996$) (Figure 4F). The *Lys*, *Dap* and *Phe* RNAs were detected in all eight amplified samples, while the *Thr* RNA was detected in three of those samples, indicating that this method can quantify genes expressed at 5–1000 copies per cell with appropriate sample numbers (see also below). Importantly, genes estimated to express >20 copies per cell exhibited more faithful amplification: 72% of the detected probe sets showed <2-fold standard deviation (SD) values of the log-expression levels. These results indicate that our amplification method preserves the representation of gene expression in the original mRNA with high reproducibility.

Next, we assessed the relationship between expression level and its ranking in the amplified products. In many different cell types and many organisms, from yeast to human, the relative expression levels and their rankings among observed genes have been shown to exhibit a power-law distribution with an exponent close to −1, (Zipf's law) (27). Both the nonamplified and amplified samples were shown to obey this law, with plot

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.008



(B)

	<2	<4	<10	<100	Detected	R^2
V1V3	65% (168/260)	89% (232/260)	93% (242/260)	95% (248/260)	98% (255/260)	0.734
AL1	7% (17/260)	22% (56/260)	41% (106/260)	62% (162/260)	88% (229/260)	0.496

Figure 3. Performance of the new cDNA amplification method. (**A**) Representative and reproducible amplification of single-cell cDNAs by the new method (V1V3: blue squares) compared with the original one (orange squares). The previous method was performed exactly as described in a preceding single-cell microarray study (21). Genes known to be expressed in ES cells and the spike RNAs (*Lys*, *Dap*, *Phe* and *Thr* for 1000, 100, 20 and 5 copies per cell, respectively) are examined: From higher expression levels, *Ezh2*, *Gapdh*, *Oct4*, *Lys*, *Esg1*, *Sox2*, *Rex1*, *nanog*, *G9a*, *Dnmt3b*, *Dap*, *lefty1*, *fragilis*, *Dnmt1*, *Fgf4*, *Eras*, *Yy1*, *cMyc*, *nodal*, *Phe*, *Foxh1*, *Tiar*, *Jak1A*, *Tnap*, *stella*, *Tyk2*, *Thr*. The log expression level of each gene is measured by Q-PCR and normalized with that of the spike RNA *Lys* (1000 copies per cell). Results of 10 independent amplifications (10 pg) are plotted against the nonamplified control (1 µg). The undetected genes are plotted in the shaded region. Red lines indicate the expression levels of the spike RNA in the nonamplified control. Green lines indicate fold differences from the nonamplified control. (**B**) Statistical comparison of the present (V1V3) and original (AL1) methods. The frequencies of the probes within the indicated fold differences from the nonamplified control are shown. The frequencies and the R^2 values of the total detected probes are also shown. The population parameter is 260, as the number of examined genes except for Lys (used for normalization) is 26 and the sample number is 10.

shapes very similar to those in previously reported results and with slopes close to −1 (−0.84 and −0.95, respectively) (Figure 4C and D) (27). In addition, the differences in expression level rankings between the nonamplified and amplified samples were within 2.5-fold for the majority of the detected probes (85%) (Figure 4E). These data demonstrate another aspect of representation preserved by our method.

Detection ability of the new single-cell microarray method

Next, we calculated the coverage (the rate of true positives detected in the amplified samples) and accuracy (the rate of detected probes that are true positives) of a single amplified sample using our method (Figure 5A and B) (for the definitions of coverage and accuracy see Materials and Methods). For these analyses, the true positive was defined as the probes called Present (MAS 5.0) in at least six of the eight nonamplified controls (yellow crosses in Figure 5C); the probes called Present reproducibly in pair-wise comparisons (open square in Figure 5C) showed similar frequency distributions. Under this definition, about 45% of all probes (20 317) were assumed to be true positives.

Coverage of the single amplified samples as a function of expression level was plotted based on the definition of true positive (the blue squares in Figure 5A). Coverage was highly

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.009

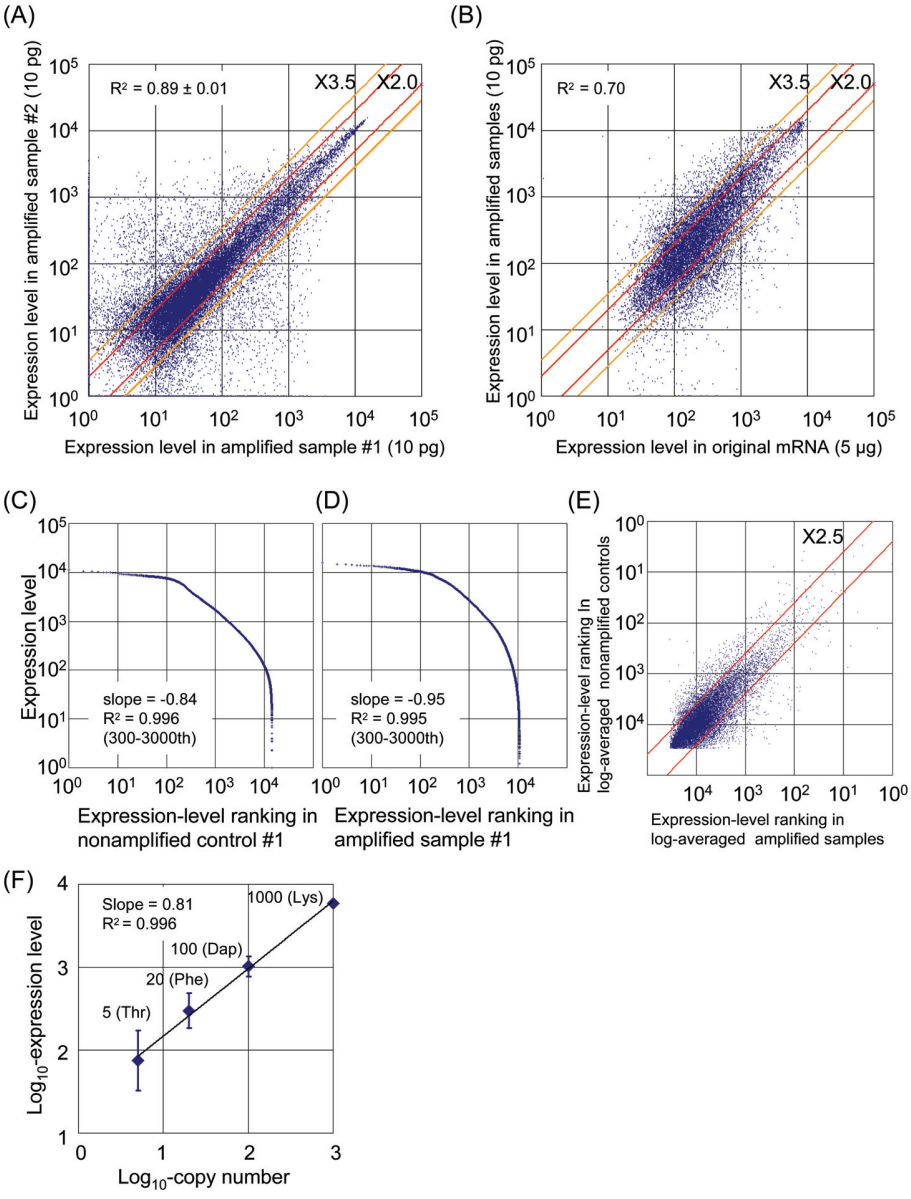


Figure 4. Performance of the single-cell-level microarray using the new method. (**A**) Scatter plots of data obtained from two independently amplified samples from 10 pg ES cellular RNA. Expression levels of all probes are plotted. R^2 values were calculated for probes detected reproducibly in pair-wise comparisons. (**B**) Scatter plots of data obtained from nonamplified (5 µg total RNA) and amplified samples. The log-averaged expression levels of probes detected in both are plotted. The 2.0- and 3.5-fold differences are represented by red and yellow lines, respectively, in (A) and (B). (C and D) Relationship between expression levels and their ranking in total RNA from ES cells (**C**) and an amplified cDNA (**D**). (**E**) Scatter plots of expression level ranking between amplified and nonamplified samples. The red lines represent 2.5-fold differences. (**F**) Expression levels of amplified spike RNAs proportional to their copy numbers (the probe set IDs are AFFX-LysX-3_at, AFFX-DapX-3_at, AFFX-PheX-3_at and AFFX-ThrX-3_at). The log-transformed expression levels were averaged and plotted, with the bars representing SD.

dependent on the expression level. Importantly, however, the vast majority of true positives in the expression level ranges greater than 20 and 5 copies per cell were successfully detected in the single amplified samples (94 and 84%, respectively). This demonstrated that our method detects most genes that are expressed at considerable levels. Accuracy, plotted in a similar manner (blue squares in Figure 5B), showed that 97 and 93% of the detected probes were truly positive in the expression level range greater than 20 and 5 copies per cell, respectively.

Next, we performed a combined analysis of eight independently amplified samples. First, detection by multiple amplified samples should be defined appropriately ('definition of detection', hereafter); most stringently, it can be defined as probes called Present in all eight amplified samples. Least stringently, it can be defined as probes called Present in at least one of the eight amplified samples (colored crosses in Figure 5D). In comparison with the single-sample analysis, in the multiple-sample analysis the coverage was improved under the definitions of detection where ⩾1 to ⩾5 of the eight amplified

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.010

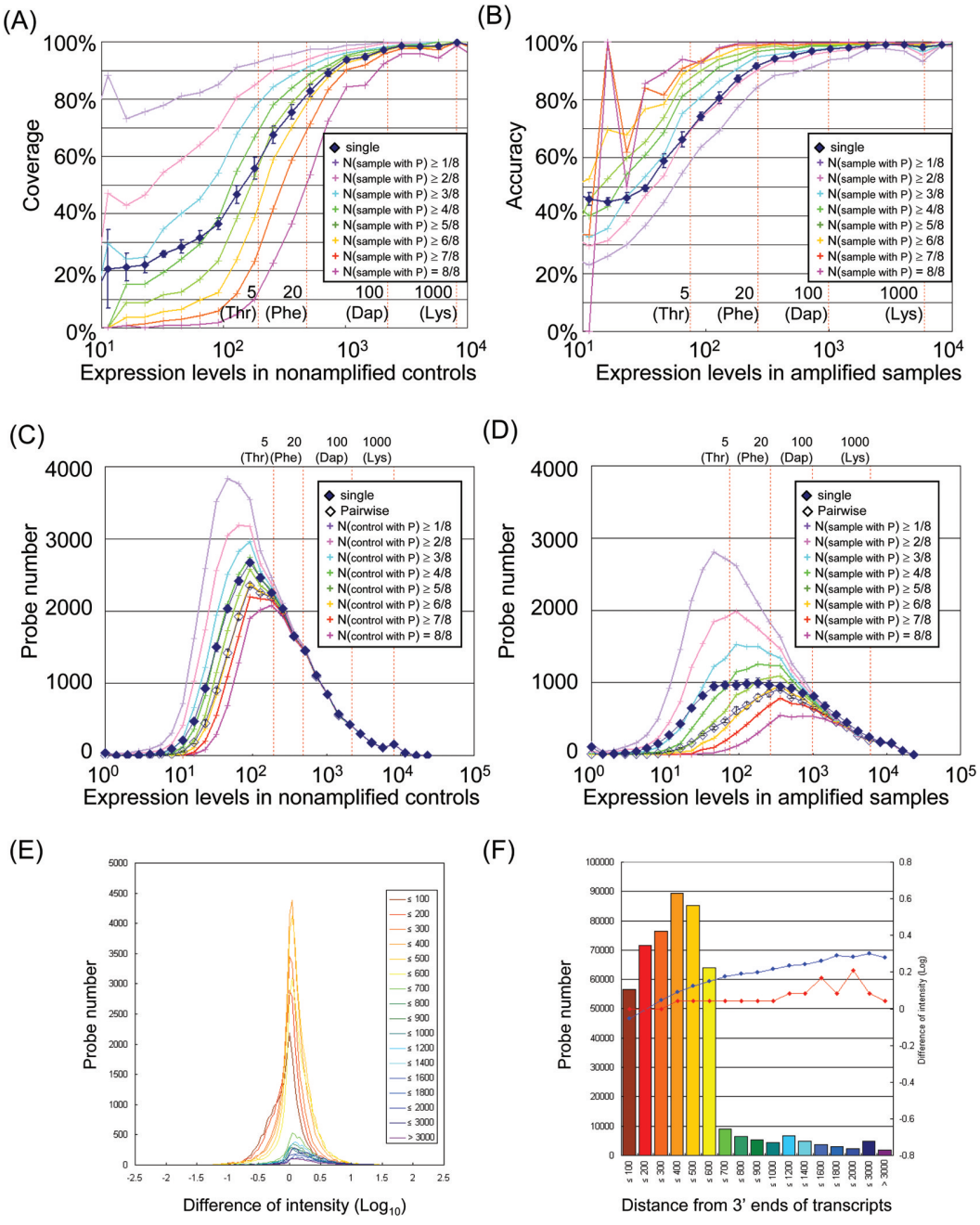


Figure 5. Detection ability of single-cell microarray using the new method. (**A**) Coverage of the amplified samples, plotted against the expression level in the original RNA. The blue squares represent the means of coverage in single-sample analyses, with bars representing SDs. The results of multiple-sample analyses under the definitions of detection where ⩾1–8 of the 8 amplified samples are called Present are represented by the crosses colored with purple, light pink, cyan, green, light green, yellow, red and hot pink, respectively. (**B**) Accuracy of the amplified samples as a function of expression level. The representation code is the same as in (A). (**C**) Frequency distribution of probes detected in the nonamplified controls as a function of expression level. (**D**) Frequency distribution of probes detected in the amplified samples as a function of expression level. The closed blue squares represent the means of the frequency of the Present probes in single nonamplified controls (C) and amplified samples (D), respectively, while the open blue squares represent the means of probes called Present reproducibly in pair-wise comparisons. The color code in the multiple sample analyses is similar to that in (A), corresponding to the definitions of true positive (C) and detection (D), respectively. The expression levels of the spike RNAs in the nonamplified controls (A and C) and amplified samples (B and D) are represented by red dashed lines. (**E**) Position effects of probe locations on signal intensities. The probes (individual probes, not probe sets) in the Affymetrix GeneChip Mouse Genome 430 2.0 array were classified according to the distance from the probe location to the 3′ ends of the transcripts. The histograms of the probes located within 600 bp from the 3′ ends are represented by warm colors (red/yellow), while those beyond 600 bp are represented by cold colors (blue/green). The probe frequencies were plotted against the difference in intensity between nonamplified controls and amplified samples ($\log_{10}$ transformed). These plots were generated from probes called Present. The probe locations were determined using the EnsEMBL transcript database or, for probes not contained in the EMBL database, using the EST data provided by Affymetrix. (**F**) Frequency distribution of probes against the distances from the 3′ ends of the transcripts. The total number of probes on the array in each location category is represented by a bar. The color code is the same as in (B). The blue circles and red squares represent the averages and peaks of intensity difference. Note that both are roughly constant relative to the probe location, with shifts of <2-fold ($\approx 10^{0.3}$).

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.011

samples are called Present (Figure 5A). On the other hand, the accuracy was improved under definitions of detection where ⩾3 to 8 of the eight amplified samples are called Present (Figure 5B). Therefore, we consider that the most appropriate definition of detection may be that at least three of the eight amplified samples are called Present (cyan crosses in Figure 5A and B). Under this definition, about 39% of all probes (17 425) were assumed to have been detected. Importantly, accuracy was high when this definition of detection was used. Coverage and accuracy were 92 and 94% in the expression level range >5 copies per cell, respectively, and both of them are 97% in the expression level range >20 copies per cell.

Interestingly, coverage was by far more dependent than accuracy on the definition of detection, especially in lower expression levels (Figure 5A and B). This suggests that the unreproducible reduction of Present calls from nonamplified controls observed at low expression levels of amplified samples (filled and open squares in Figure 5D) was due to random detection failures (false negatives), possibly caused in the dilution process of the RNA and/or in the initial RT reaction, rather than to the effect of nonspecific amplification and/or hybridization (false positives). It has often been pointed out that the three-prime restriction of the RT products may result in the failure to detect some probes. Although we observed this limitation in some of the detected probes (data not shown), it seems to occur only in a minority of them (Figure 5E and F) and therefore would not be a major drawback of the new methodology. As determined by sequencing, the average length of amplified cDNA bodies without poly(dA) tract and primer sequences on both ends was 687 ± 279 bp (Figure 2C and Table 2), while the majority of the probes (89%) were located within 600 bp of the 3′-terminus of the transcripts (Figure 5F). Collectively, these results clearly demonstrate the high reliability of our single-cell level microarray method.

We also statistically compared the representation, reproducibility, coverage and accuracy of our microarray data with those by the original method published by Tietjen *et al*. (21) and found that, consistent with our Q-PCR examination (Figure 3A), representation, reproducibility and accuracy were improved with our method (for detailed comparison see Materials and Methods, Discussion and Figure 7).

More sensitive identification of transcriptional differences among different single-cell populations is therefore possible with our method, with appropriate statistical analyses including that of signal-to-noise ratio/*T*-test in conjunction with the permutation test. This combination of tests precludes the contribution of any sample variability, including that from the methodology, at high confidence (see Materials and Methods and below).

Application of the method to single cells in blastocysts of mouse embryos

The ICM of the early blastocyst at E3.5, a source of ES cell derivation (28,29), is a morphologically homogeneous population of undifferentiated pluripotent cells that give rise to all embryonic lineages (30). A recent study, however, has started to show that a subset of cells in the ICM exhibits a sign of differentiation into PE as early as E3.5, which is evidenced by lineage-tracing experiments and the localized expression of *Gata6*, a critical transcription factor for PE differentiation (31). To assess the power of our single-cell microarray method, we went on to analyze the transcriptomes of single cells in the ICM at E3.5. We isolated blastocysts at E3.5 and dissociated the ICM into single cells by trypsin-EDTA treatment. To prepare cDNA samples, we then randomly picked a total of 55 single cells. The synthesized cDNAs were screened by gene-specific PCR using *Oct4* and *Cdx2* to remove trophectoderm cells (32,33), and 50 cells were identified as *Oct4*-positive and *Cdx2*-negative. The expression levels of *Oct4* in these cells were found to be almost uniform by Q-PCR (Figure 6E). Therefore, 20 randomly selected samples were applied to the GeneChip Mouse Genome 430 2.0 array, and were analyzed by unsupervised hierarchical clustering (34) using GenePattern software (35), which performs the signal-to-noise ratio analysis/*T*-test in conjunction with the permutation test to preclude the contribution of any sample variability, including that from the methodology, at high confidence. These analyses revealed that these samples can be classified roughly into two groups: cluster 1, consisting of 9 cells, and cluster 2, consisting of 11 (Figure 6A). The class neighbor analysis (35) identified the genes that are differentially expressed between the two clusters (Figure 6B and Supplementary Table S1), and the class predictors revealed that the genes related to the epiblast (*nanog* and *Fgf4*) (36–38) and the PE (*Gata4*, *Gata6* and *cubilin*) (31,39,40) each belong to either one cluster or the other. The signal-to-noise scores calculated by the class neighbor algorithm were greater than the 5% level in the histograms of the scores generated by random permutations ($N = 100$) corresponding to each rank (35), for all of the top 100 genes for each cluster. The epiblast-related gene *Fgf4* and the pluripotent marker *nanog* were up-regulated in cluster 1, and were associated with genes such as transcription factors (*Sox2*, *c-Myc*, *Klf2*, *SpiC*), signal transduction factors (*Prkcz, Trio, Rhebl1* and *Rhpn2*), and apoptosis-associated genes (*Bcl2l14*, *Bcl7a* and *Ndg1*) (Figure 6B and C). On the other hand, the endoderm-related genes were up-regulated in cluster 2 in association with many other genes, including transcription factors (*Sox17*, *Runx1*), cell surface receptors (*Fgfr2*, *Pdgfr2* and *Pthr1*), a basement membrane component (*Lama1*), and *de novo* DNA methyltransferases (*Dnmt3a* and *Dnmt3l*) (Figure 6B and C). We confirmed the expression of these differentially expressed genes by Q-PCR (Figure 6C). Furthermore, we examined the expression levels of key genes in all 50 samples and found that the trend we identified by the microarray experiments using 20 samples was essentially preserved (Figure 6E); the rate of *Gata4*-positive cells in 20 samples used in microarray analysis [65% (13/20)] was comparable with that in all the isolated 50 ICM cells [60% (30/50)] (see also below), suggesting that the randomly chosen 20 samples are a suitable representative of the population of ICM cells. These findings suggest that morphologically indistinguishable ICM cells with similar *Oct4* expression levels do indeed undergo bi-directional differentiation toward either a PE or epiblast fate even by E3.5, with robust expression of lineage-determining transcription factors (the expression levels of *Gata4* and *Gata6*, the most upstream regulator for PE differentiation, are as high as a few hundred copies per cell, and those for *nanog* and *Klf2* are as high as a thousand copies per epiblast cell).

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.012

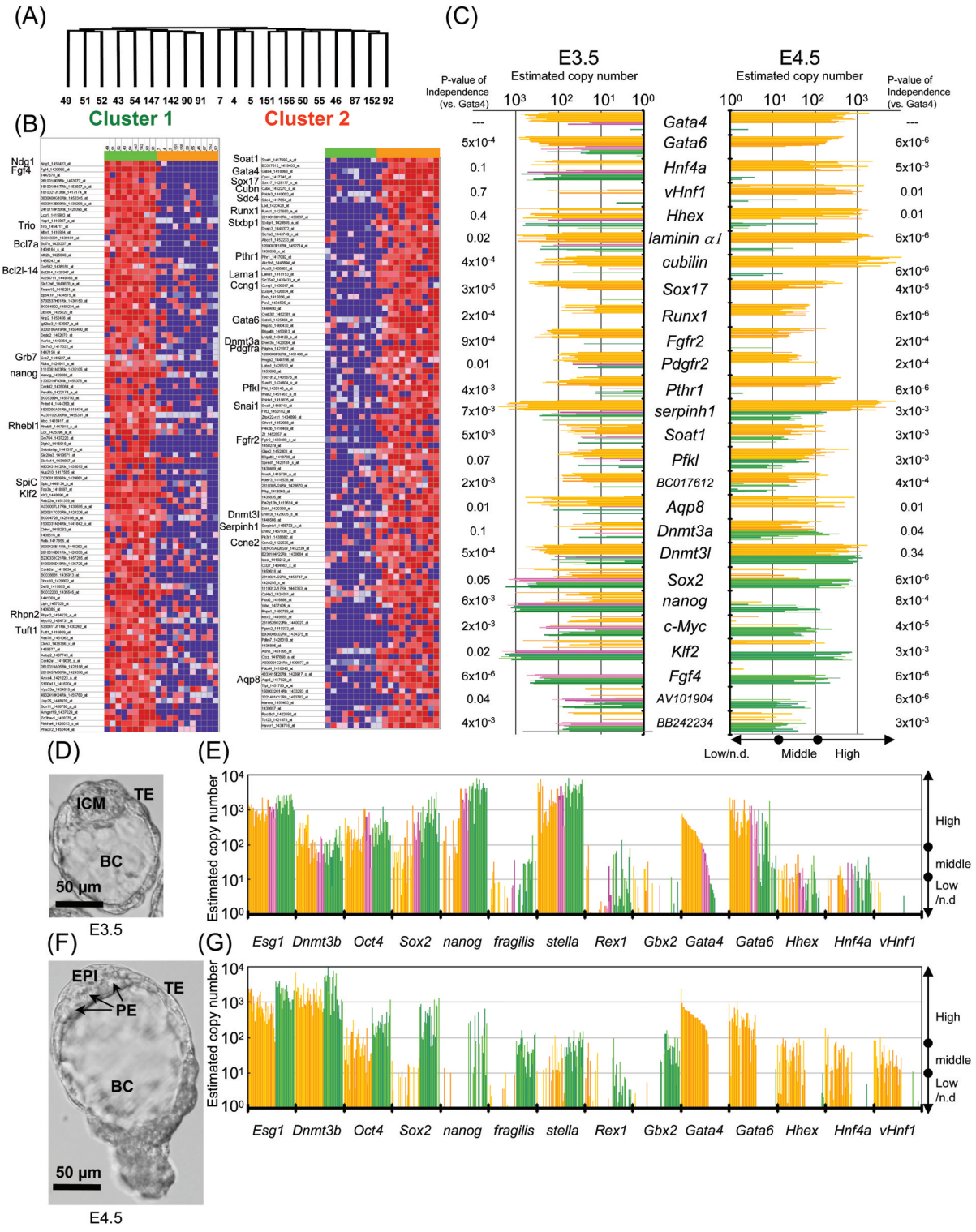


Figure 6. Direct application of the newly developed method to single ICM cells from mouse E3.5 blastocyst reveals the presence of two distinct cell populations. (**A**) Hierarchical clustering of single ICM cells. (**B**) Heat map representation of differentially expressed genes (top 100). The expression levels are color-coded from red (high) to blue (low). The expression levels are normalized in the lows. (**C**) The correlation of gene expression is preserved between E3.5 and E4.5. The copy numbers of expressed genes were estimated with Q-PCR. Orange, pink and green bars represent high, middle and low/non-detectable expression of *Gata4*, respectively. *P*-values of the Chi-square test for independence from *Gata4* expression are indicated. (D and F) Blastocysts at E3.5 (**D**) and E4.5 (**F**). The typical embryos used for single-cell experiments are shown. (E and G) Expression levels of key genes related to PE and epiblast at E3.5 (**E**) and E4.5 (**G**). All of the single-cell samples of ICMs are shown. The representation code is the same as in (C).

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.013

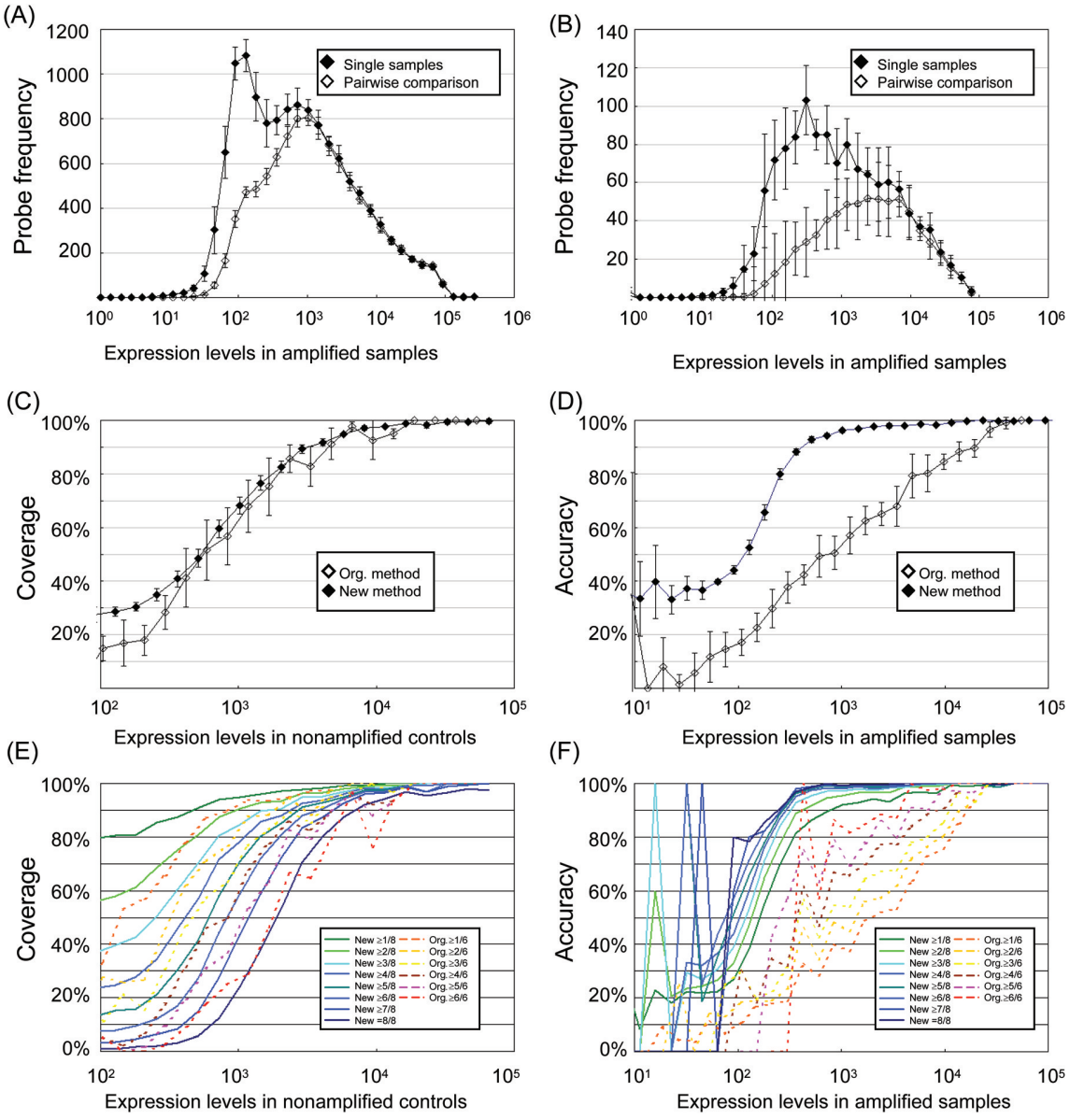


Figure 7. Statistical comparison between the new and original (Org.) methods. (A and B) Frequency distribution of Present call in the new (**A**) and original (**B**) methods. The data of the original method were obtained from a previous microarray study using GeneChip [Supplementary Data set S6 (21)]. The closed squares represent the average frequencies of the probes (mean ± SD) called Present in single samples. The open squares represent the average frequency (mean ± SD) of the probes called Present reproducibly in pair-wise comparisons. The expression levels are normalized as described in the text. (**C**) Coverage in single sample analyses. The closed and open squares represent coverage of the new and original methods, respectively, with the bars representing SD. (**D**) Accuracy in single-sample analyses. The representation manner is the same as in (C). (**E**) Coverage in multiple sample analyses. Data from the new method are represented by solid lines in cold colors (blue/green). Data from the original method are represented by dashed lines in warm colors (red/yellow). Each line represents the indicated definition of detection. (**F**) Accuracy in multiple-sample analyses. The representation manner is the same as in (E).

To confirm this observation, we next examined whether or not genes identified as differentially expressed are preserved in morphologically differentiated PE and epiblast at E4.5. We prepared cDNAs from 65 single cells of dissociated ICMs at E4.5, of which 61 were identified as ICM cells (*Oct4*-positive, *Cdx2*-negative). Among the ICM cells, we found that 57% (35/61) express both *Gata4* and *Gata6* as PE and that 39% (24/61) are *Gata4*- and *Gata6*-negative and *Oct4*-positive as epiblast, while only 3% (2/61) showed disassociated expression of *Gata4* and *Gata6* (Figure 6G). The *Gata4*- and *Gata6*-positive cells at E4.5 almost exclusively expressed other PE markers, including *Hhex*, *Hnf4a* and *vHnf1* (41–44). This is in contrast with the expression at E3.5, when *Hhex* and *Hnf4a* were expressed in both the endoderm- and epiblast-like populations while *vHnf1* showed no detectable expression in either population (Figure 6C). Furthermore, the PE markers were much more closely associated with each other at E4.5 than at E3.5, as shown in Figure 6C, E and G, since 32% (16/50) of the isolated ICM cells at E3.5 showed *Gata6* expression with low or undetectable expression of *Gata4* [60% (30/50) and 90% (45/50) of the isolated ICM cells at E3.5 showed *Gata4* and *Gata6* expression,

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.014

respectively, (Figure 6E and G)]. These results indicated that PE is indeed well established at E4.5.

Using Q-PCR, we examined the expression of 32 genes that were differentially expressed between clusters 1 and 2 (16 genes each) in E4.5 PE and epiblast using Q-PCR. As shown in Figure 6C, the correlation of gene expression between the E3.5 and E4.5 samples was well conserved. Of the 16 genes up-regulated in cluster 2 at E3.5 (*Gata4*, *Gata6*, *Cubn*, *Lama1*, *Sox17*, *Aqp8*, *BC017612*, *Fgfr2*, *Pdgfra*, *Pfkl*, *Pthr1*, *Soat1*, *Runx1*, *Dnmt3l*, *Dnmt3a, Serpinh1*), all but 2 (*Dnmt3a* and *Dnmt3l*) preserved the correlation at E4.5 (Figure 6C, orange bars). This strongly indicated that PE-like cells at E3.5 are indeed the precursors of established PE at E4.5. On the other hand, of the 16 genes up-regulated in cluster 1 (*Fgf4*, *Sox2*, *nanog*, *Klf2*, *c-Myc*, *SpiC*, *Rhpn2*, *Rhebl1*, *Trio*, *Prkcz*, *Bcl2l14*, *Bcl7a*, *Ndg1*, *Tuft1*, *BB242234*, *AV101904*) 7 preserved the correlation in E4.5 epiblast cells (Figure 6C, green bars). As with the endoderm-related genes, all these genes were more closely associated with each other and disassociated from PE markers at E4.5 than at E3.5. The *P*-values of the Chi-square test for independence (Materials and Methods) indicated that the observed coherent gene expressions were statistically significant (Figure 6C). Remarkably, the expression of epiblast marker *Sox2* was highly up-regulated in the *Gata4*- and *Gata6*-negative cells; it was expressed in both, with only a marginal difference between expression patterns at E3.5. Moreover, the expression of pluripotency marker *nanog* was considerably decreased at E4.5, although the *nanog* expression was restricted to the *Sox2*-positive and *Gata4*- and *Gata6*-negative cells. This observation is consistent with a previous study showing that *nanog* expression is high in the ICMs of early blastocysts, while in later blastocysts it decreases in ICMs and is excluded from the PE (38). Moreover, it was intriguing that, by the hierarchical clustering, all the ICM cells at E3.5 were defined as being closer to each other than to ES cells (data not shown), suggesting that the PE- and epiblast-like populations at E3.5 are basically very close to each other. Collectively, these observations demonstrate that the single-cell microarray technology can identify the very onset of biologically relevant quantitative differences in gene expression among single cells and can identify genes that may work at the most upstream step of the PE differentiation from ICM cells.

DISCUSSION

Improvement of single-cell microarray experiments

Reliable technology for the precise monitoring of global gene expression at the single-cell level has been sought in various areas of biological sciences. The method we presented consists of relatively few cycles of directional exponential amplification, followed by isothermal linear amplification. The representation and reproducibility of the cDNAs amplified by this method were superior to those amplified by the original method, as shown by the Q-PCR analysis (Figure 3). The microarray data obtained with GeneChip Mouse Genome 430 2.0 arrays were highly consistent with this result. The analyses of the absolute detection calls also showed high coverage and accuracy with our method.

To date, successful single-cell-transcript analyses based on exponential amplification have been reported: modifications of Brady and Iscove's original method (13,20,21); an arbitrarily-primed PCR method (45); and a combination of SMART-PCR and two-round IVT (46). Two of them, using modified Brady and Iscove's methods, have been well characterized (13,21).

We describe here some of the statistical differences between the microarray data obtained with our method and that obtained with the original one (8,21), in both the qualitative (coverage and accuracy) and quantitative (representation and reproducibility) aspects (Figure 7). For this comparison, we used the microarray data in the previous study [Supplementary Data set S6 from Tietjen *et al*. (21)] in which repeatedly sampled diluted (10 pg) and undiluted total RNA from human glioblastoma cell culture were used, exactly as in the case of our ES cell sample analyses (see Materials and Methods for the normalization between Tietjen's and our data). The frequency distributions of the detected probes are shown in Figure 7A and B. For the qualitative comparison, we calculated coverage and accuracy based on the definitions described above. In single-sample analyses, while the coverage obtained by our method is comparable with that obtained by the previous analysis, the accuracy is unambiguously improved in the new method (Figure 7C and D), consistent with the Q-PCR analysis (Figure 3B). In multiple-sample analyses, on the other hand, the difference between the new and original methods is more remarkable; the least stringent definition of detection in our method ($\geqslant$1 of 8 amplified samples call Present, blue line in Figure 7F) still shows an accuracy comparable with that achieved under the most stringent definition of detection in the previous study (all six of the amplified samples call Present, dashed orange line in Figure 7F). These results suggest the qualitative improvement achieved with our method.

For quantitative aspects, our single-cell microarray showed higher correlations between independently amplified samples and between nonamplified and amplified samples than those from the previous study (R^2 values of the former were 0.89 versus 0.69, R^2 values of the latter 0.69 were versus 0.41, respectively), suggesting improvements in both representation and reproducibility. These comparisons are highly consistent with our extensive Q-PCR comparison of the quality of amplified samples between the original and the new methods (Figure 3).

Our protocol includes Exonuclease I treatment to degrade un-reacted first-strand synthesis primer. This treatment resulted in an efficient PCR with a good amount of amplified product obtained after relatively few PCR cycles and avoiding the risks of distorted gene representation. Directional amplification generated the sense-antisense orientation of the PCR products, enabling the direct applicability of amplified cDNA to high-density oligonucleotide microarrays according to the standard labeling protocols. These factors may have contributed to the improvement of the quality of the microarray data with our method.

Advantage of exponential amplification over linear amplification in single-cell analyses

Although a full statistical validation has yet to be described properly, there are some reports of single-cell microarray analyses using linear amplification protocols (12,47–49).

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.015

Regarding the general nature of amplification, the exponential and linear amplifications are essentially complementary to each other in gene expression profiling (13–15). However, when applied to single-cell-level samples, exponential amplification has clear advantages over linear amplification. First, owing to the limited amplification capacity of the latter, the existing two-round IVT labeling procedures provide $\approx 10^6$-fold or less amplification (50). This amount is slightly insufficient for one oligonucleotide microarray experiment according to the standard protocols recommended by, for example, the Affymetrix GeneChip series and the Code link platforms, although future improvements in the IVT and/or microarray reagents or protocols might overcome this shortcoming. On the other hand, the exponential methods, including ours, can provide abundant cDNA products. This advantage means that researchers can use the cDNAs obtained by a single experiment for a wide range of analyses of single-cell transcriptomes—including the screening of amplified single-cell cDNAs by marker-gene-specific PCR, detailed Q-PCR analyses on specific genes, verification of amplification quality before microarray applications, and a few dozen experiments on various microarrays. This advantage also indicates that the exponential amplification protocol offers better cost-performance.

Moreover, the two-round linear amplification procedure requires purification of total RNA from single cells at the onset of cDNA synthesis, as well as complicated processes including repeated DNA/RNA synthesis and purification (10–12). The RNA purification from single cells may lead to the loss of mRNAs expressed at low levels. In contrast, our PCR-based method does not require the single-cell-level RNA isolation and is conducted in a single tube without buffer exchanges; it is completed within a few hours and enables multiple samples to be processed in one experiment. These were advantages of the original method (8,9). In addition, linear amplification produces RNA, while DNA is the final product in exponential amplification. DNA is more suitable for use as a PCR template for quality checking and screening cells using marker gene expressions. DNA is also better than RNA for long-term storage because it is chemically more stable. These advantages would make exponential amplification more suitable for some practical situations where repeated samplings of a number of single-cell cDNAs are required.

Application of the single-cell microarray method to early mouse embryogenesis

We applied this newly developed method to single ICM cells from E3.5 mouse blastocysts, which differentiate into either PE or epiblast cells within one day. Since there are so few ICM cells (no more than 20), all the transcriptome analyses in this stage have been performed with hundreds of whole embryos, including trophectoderm cells (51–54), which provide expression profiles of the mixtures of embryonic and extra-embryonic components.

Using the improved single-cell method, we performed genome-wide analysis of single ICM cells for the first time. We found that the morphologically homogeneous ICM cells have, as early as E3.5, at least two populations of cells, one with PE- and the other with epiblast-like gene expression, which is consistent with a previous report showing the distinct expression of one transcription factor in embryos at this stage (31). We discovered genes associated with either population, and found that many of them showed preserved and higher correlations to the well-differentiated PE and/or epiblast cells of E4.5 embryos. This demonstrated that morphologically indistinguishable ICM cells do indeed undergo differentiation toward either PE or epiblast fate already by E3.5. It will be interesting to examine how far we can trace back the origin of blastomeres that show distinct gene expression profiles in preimplantation embryos. At the same time, such an experiment may reveal a distinct set of genes critical for lineage allocation.

In conclusion, we have developed a highly quantitative and reproducible cDNA amplification method from single cells and have effectively demonstrated its applicability to practical biological questions of interest. As shown with one example here, this method is applicable to a wide variety of biological/biophysical questions that require resolution at the single-cell level.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Miss Junko Nishio for her excellent help in the microarray experiments, including IVT labeling and hybridization, and Dr Minoru S.H. Ko for critical reading of the manuscript. This study was supported in part by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and by a PRESTO project grant from Japan Science and Technology Agency (M.S.). Funding to pay the Open Access publication charges for this article was provided by RIKEN.

Conflict of interest statement. None declared.

REFERENCES

1. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
2. Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.
3. Westerhoff,H.V. and Palsson,B.O. (2004) The evolution of molecular biology into systems biology. *Nat. Biotechnol.*, **22**, 1249–1252.
4. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
5. 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.
6. Wang,E., Miller,L.D., Ohnmacht,G.A., Liu,E.T. and Marincola,F.M. (2000) High-fidelity mRNA amplification for gene profiling. *Nat. Biotechnol.*, **18**, 457–459.
7. Korobkova,E., Emonet,T., Vilar,J.M., Shimizu,T.S. and Cluzel,P. (2004) From molecular noise to behavioural variability in a single bacterium. *Nature*, **428**, 574–578.
8. Brady,G., Barbara,M. and Iscove,N. (1990) Representative *in vitro* cDNA amplification from individual hemopoietic cells and colonies. *Methods Mol. Cell. Biol.*, **2**, 17–25.
9. Dulac,C. and Axel,R. (1995) A novel family of genes encoding putative pheromone receptors in mammals. *Cell*, **83**, 195–206.

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.016

10. Van Gelder,R.N., von Zastrow,M.E., Yool,A., Dement,W.C., Barchas,J.D. and Eberwine,J.H. (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl Acad. Sci. USA*, **87**, 1663–1667.
11. 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.
12. Kamme,F., Salunga,R., Yu,J., Tran,D.T., Zhu,J., Luo,L., Bittner,A., Guo,H.Q., Miller,N., Wan,J. *et al*. (2003) Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J. Neurosci*., **23**, 3607–3615.
13. Iscove,N.N., Barbara,M., Gu,M., Gibson,M., Modi,C. and Winegarden,N. (2002) Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat. Biotechnol*., **20**, 940–943.
14. Klur,S., Toy,K., Williams,M.P. and Certa,U. (2004) Evaluation of procedures for amplification of small-size samples for hybridization on microarrays. *Genomics*, **83**, 508–517.
15. Ji,W., Zhou,W., Gregg,K., Lindpaintner,K. and Davis,S. (2004) A method for gene expression analysis by oligonucleotide arrays from minute biological materials. *Anal. Biochem*., **331**, 329–339.
16. Ko,M.S.H., Ko,S.B.H., Takahashi,N., Nishiguchi,K. and Abe,K. (1990) Unbiased amplification of a highly complex mixture of DNA fragments by 'lone linker'-tagged PCR. *Nucleic Acids Res*., **18**, 4293–4294.
17. Petalidis,L., Bhattacharyya,S., Morris,G.A., Collins,V.P., Freeman,T.C. and Lyons,P.A. (2003) Global amplification of mRNA by template-switching PCR: linearity and application to microarray analysis. *Nucleic Acids Res*., **31**, e142.
18. Saito,H., Kubota,M., Roberts,R.W., Chi,Q. and Matsunami,H. (2004) RTP family members induce functional expression of mammalian odorant receptors. *Cell*, **119**, 679–691.
19. Saitou,M., Barton,S.C. and Surani,M.A. (2002) A molecular programme for the specification of germ cell fate in mice. *Nature*, **418**, 293–300.
20. Chiang,M.K. and Melton,D.A. (2003) Single-cell transcript analysis of pancreas development. *Dev. Cell*, **4**, 383–393.
21. Tietjen,I., Rihel,J.M., Cao,Y., Koentges,G., Zakhary,L. and Dulac,C. (2003) Single-cell transcriptional analysis of neuronal progenitors. *Neuron*, **38**, 161–175.
22. Brady,G., Billia,F., Knox,J., Hoang,T., Kirsch,I.R., Voura,E.B., Hawley,R.G., Cumming,R., Buchwald,M. and Siminovitch,K. (1995) Analysis of gene expression in a complex differentiation hierarchy by global amplification of cDNA from single cells. *Curr. Biol*., **5**, 909–922.
23. Brail,L.H., Jang,A., Billia,F., Iscove,N.N., Klamut,H.J. and Hill,R.P. (1999) Gene expression in individual cells: analysis using global single cell reverse transcription polymerase chain reaction (GSC RT-PCR). *Mutat. Res*., **406**, 45–54.
24. Brady,G. and Iscove,N.N. (1993) Construction of cDNA libraries from single cells. *Methods Enzymol*., **225**, 611–623.
25. Li,C. and Wong,W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
26. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
27. Furusawa,C. and Kaneko,K. (2003) Zipf's Low in Gene Expression. *Phys. Rev. Lett*., **90**, 088102.
28. Evans,M.J. and Kaufman,M.H. (1981) Establishment in culture of pluripotential cells from mouse embryos. *Nature*, **292**, 154–156.
29. Martin,G.R. (1981) Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl Acad. Sci. USA*, **78**, 7634–7638.
30. Gardner,R.L. and Rossant,J. (1979) Investigation of the fate of 4-5 day post-coitum mouse inner cell mass cells by blastocyst injection. *J. Embryol. Exp. Morphol*., **52**, 141–152.
31. Rossant,J., Chazaud,C. and Yamanaka,Y. (2003) Lineage allocation and asymmetries in the early mouse embryo. *Philos. Trans. R Soc. Lond. B Biol. Sci*., **358**, 1341–1348; discussion 1349.
32. Nichols,J., Zevnik,B., Anastassiadis,K., Niwa,H., Klewe-Nebenius,D., Chambers,I., Scholer,H. and Smith,A. (1998) Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell*, **95**, 379–391.
33. Beck,F., Erler,T., Russell,A. and James,R. (1995) Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Dev. Dyn*., **204**, 219–227.
34. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
35. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
36. Niswander,L. and Martin,G.R. (1992) Fgf-4 expression during gastrulation, myogenesis, limb and tooth development in the mouse. *Development*, **114**, 755–768.
37. Mitsui,K., Tokuzawa,Y., Itoh,H., Segawa,K., Murakami,M., Takahashi,K., Maruyama,M., Maeda,M. and Yamanaka,S. (2003) The Homeoprotein Nanog Is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells. *Cell*, **113**, 631–642.
38. Chambers,I., Colby,D., Robertson,M., Nichols,J., Lee,S., Tweedie,S. and Smith,A. (2003) Functional expression cloning of nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell*, **113**, 643–655.
39. Arceci,R.J., King,A.A., Simon,M.C., Orkin,S.H. and Wilson,D.B. (1993) Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. *Mol. Cell. Biol*., **13**, 2235–2246.
40. Assemat,E., Vinot,S., Gofflot,F., Linsel-Nitschke,P., Illien,F., Chatelet,F., Verroust,P., Louvet-Vallee,S., Rinninger,F. and Kozyraki,R. (2005) Expression and role of cubilin in the internalization of nutrients during the peri-implantation development of the rodent embryo. *Biol. Reprod*., **72**, 1079–1086.
41. Barbacci,E., Reber,M., Ott,M.O., Breillat,C., Huetz,F. and Cereghini,S. (1999) Variant hepatocyte nuclear factor 1 is required for visceral endoderm specification. *Development*, **126**, 4795–4805.
42. Duncan,S.A., Manova,K., Chen,W.S., Hoodless,P., Weinstein,D.C., Bachvarova,R.F. and Darnell,J.E.,Jr (1994) Expression of transcription factor HNF-4 in the extraembryonic endoderm, gut, and nephrogenic tissue of the developing mouse embryo: HNF-4 is a marker for primary endoderm in the implanting blastocyst. *Proc. Natl Acad. Sci. USA*, **91**, 7598–7602.
43. Thomas,P.Q., Brown,A. and Beddington,R.S. (1998) Hex: a homeobox gene revealing peri-implantation asymmetry in the mouse embryo and an early transient marker of endothelial cell precursors. *Development*, **125**, 85–94.
44. Goldin,S.N. and Papaioannou,V.E. (2003) Paracrine action of FGF4 during periimplantation development maintains trophectoderm and primitive endoderm. *Genesis*, **36**, 40–47.
45. Brandt,S., Kloska,S., Altmann,T. and Kehr,J. (2002) Using array hybridization to monitor gene expression at the single cell level. *J. Exp. Bot*., **53**, 2315–2323.
46. Gustincich,S., Contini,M., Gariboldi,M., Puopolo,M., Kadota,K., Bono,H., LeMieux,J., Walsh,P., Carninci,P., Hayashizaki,Y. *et al*. (2004) Gene discovery in genetically labeled single dopaminergic neurons of the retina. *Proc. Natl Acad. Sci. USA*, **101**, 5069–5074.
47. Seshi,B., Kumar,S. and King,D. (2003) Multilineage gene expression in human bone marrow stromal cells as evidenced by single-cell microarray analysis. *Blood Cells Mol. Dis*., **31**, 268–285.
48. Chow,N., Cox,C., Callahan,L.M., Weimer,J.M., Guo,L. and Coleman,P.D. (1998) Expression profiles of multiple genes in single neurons of Alzheimer's disease. *Proc. Natl Acad. Sci. USA*, **95**, 9620–9625.
49. Eberwine,J., Kacharmina,J.E., Andrews,C., Miyashiro,K., McIntosh,T., Becker,K., Barrett,T., Hinkle,D., Dent,G. and Marciano,P. (2001) mRna expression analysis of tissue sections and single cells. *J. Neurosci*., **21**, 8310–8314.
50. Eberwine,J. (2001) Single-cell molecular biology. *Nature Neurosci*., **4**, 1155–1156.
51. Zeng,F. and Schultz,R.M. (2005) RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo. *Dev. Biol*., **283**, 40–57.
52. Zeng,F., Baldwin,D.A. and Schultz,R.M. (2004) Transcript profiling during preimplantation mouse development. *Dev. Biol*., **272**, 483–496.
53. Wang,Q.T., Piotrowska,K., Ciemerych,M.A., Milenkovic,L., Scott,M.P., Davis,R.W. and Zernicka-Goetz,M. (2004) A genome-wide study of gene activity reveals developmental signaling pathways in the preimplantation mouse embryo. *Dev. Cell*, **6**, 133–144.
54. Hamatani,T., Carter,M.G., Sharov,A.A. and Ko,M.S.H. (2004) Dynamics of global gene expression changes during mouse preimplantation development. *Dev. Cell*, **6**, 117–131.

Downloaded from https://academic.oup.com/nar/article/34/5/e42/1146394 by Weil Gotshal and Manges LLP (Global) user on 15 March 2023

PARSE EX1037.017