Appendix

Part 1. Checklist

Experiment Design

- **Type of experiment:** A time course of dedifferentiation of wild-type *Dictyostelium discoideum* compared to a common standard to study the gene expression profile during dedifferentiation.
- **Experimental factors:** The controlled variables are the developmental stage at which the cells were disaggregated and the time of dedifferentiation at which the cells were collected for RNA extraction. We analyzed a time course of each dedifferentiation process from three different stages [Aggregation (Agg): 0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8 hours; Finger: 0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 hours; Mexican hat (Mex): 0, 0.5, 1, 2, 4, 6, 8, 12, 16, 20, 24, 28 hours].
- **The number of hybridizations performed in the experiment:** Agg: 3 x 2; Slug: 6 x 2; Mex: 8 x 2 (for details see Table 1 "Hybridization design").
- **The type of reference used for the hybridizations:** A pool of equal portions of RNA samples prepared from vegetative cells and from 5 developmental times (0, 3, 6, 12, 17, 24 hours). This reference was used in all of our previous experiments and thus allowed comparison between dedifferentiation and development.
- **Hybridization design:** Each RNA sample was hybridized to an array containing 7,744 targets, each target printed in duplicate. Each sample was co-hybridized with the common standard (for details see Table 1 "Hybridization design").
- **Quality control steps taken:**
  1) Each RNA sample was quantified by absorbance measurement (A_{260}) and tested by Northern blot analysis to test RNA integrity and to verify the absorbance measurement.
  2) The common reference served as an internal quality control for labeling and hybridization.
  3) Each chip contains duplicate targets.
  4) Threshold intensity values were used to estimate whether to accept or reject chips. These values were determined based on our previous studies (1).
- **URL of any supplemental websites or database accession numbers:**
  1) cDNA clones from the *Dictyostelium* cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html).
  2) Genomic DNA clones from the *Dictyostelium* Genome Project at Baylor College of Medicine (http://dictygenome.bcm.tmc.edu/).

Samples used, extract preparation, and labeling

- **The origin of the biological sample and its characteristics:** *Dictyostelium discoideum* strain AX2 (see text for references).

- **Manipulation of biological samples and protocols used:** Cells were grown in HL5 liquid broth. Log-phase cells were washed free of nutrients, dispersed on nitrocellulose filters at 3.5 × 10^6 cells per cm^2, and developed at 22°C. At each stage, aggregation (Agg), fingers (Slug), or Mexican hat (Mex), multicellular structures were harvested by filtration through 77-µm nylon mesh, resuspended in 20 mM potassium phosphate, pH 6.4/20 mM EDTA and dissociated by repeated pipetting. Dissociated cells were collected by filtration through 32-µm nylon mesh, resuspended in HL5 at 1-2 × 10^6 cells per ml, and shaken at 200 rpm at 22°C to induce dedifferentiation process.

- **Protocol for preparing the hybridization extract:** At each time point, 1 × 10^8 cells were collected and resuspended in 1 ml TRIzol reagent (Life Technologies), and total RNA was extracted according to the manufacturer's protocol. The RNA was resuspended in 20 mM Mops, pH 7.0, and concentration were determined by spectrophotometory and verified by Northern blot analysis.

- **Labeling protocol(s):** Total RNA (10 µg) was mixed with 1 µg of labeled $(dT)_{18}$ primer in 13 µl of water, incubated at 70°C for 10 min and on ice for 2-5 min. Reaction buffer (GIBCO/BRL), 0.1 M DTT, 0.5 mM of each dNTP, and 200 units of Superscript II (GIBCO/BRL) were added and the reaction was incubated at 42°C for 2 hours. Experimental RNA samples were labeled with Cy5 primers and the reference RNA sample, a pool of RNA samples from different developmental stages, was labeled with Cy3. All the primers were HPLC-purified by the manufacturer (Operon Technologies). Reverse transcription reactions were terminated with 0.1 M EDTA. RNA was degraded by adding 0.3 M NaOH and incubating at 60°C for 20 min. The reaction was neutralized with 0.4 M Tris·HCl, pH 7.6. Labeled cDNA was purified by ethanol precipitation, resuspended in 10 µl of distilled deionized water, and mixed with 130 µl of PerfectHyb Plus Hybridization buffer (Sigma H 7033).

- **External controls (spikes):** None

- **Hybridization procedures and parameters:** Labeled cDNA was resuspended in 10 µl of distilled deionized water, mixed with 130 µl of PerfectHyb Plus Hybridization buffer, and hybridized to arrays containing 7,744 hybridization targets after heat treatment (95°C, 2 min) using a GeneTAC hybridization station (Genomic Solutions) according to the manufacturer's recommended protocol as indicated below.

- **The protocol and conditions used during hybridization, blocking and washing:** Labeled cDNA was hybridized to Arrays using a GeneTAC automatic hybridization station (Genomic Solutions) for 2 hours at 65°C. Arrays were sequentially washed with three solutions, 2× SSC/0.5% SDS, 0.5× SSC/0.5% SDS, and 0.1× SSC, for 30 sec at room temperature twice each.

- **Measurement data and specifications:** The arrays were scanned with a ScanArray5000 scanner (GSI Lumonics) according to the manufacturer's recommended protocol. PMT and Laser settings were not recorded.

- **The quantitation based on the images:** All images were processed with the GLEAMS software package (NuTec Sciences). Additional information about the software and its performance are given in Part 2 below.

- **Type of scanning hardware and software used:** a ScanArray5000 scanner (GSI Lumonics) and ScanArray Microarray Acquisition system (GSI Lumonics).
- **A description of the measurements produced by the image-analysis software and a description of which measurements were used in the analysis:** Adaptive morphological detection method (per manufacturers' specifications).
- **The complete output of the image analysis *before* data selection and transformation (spot quantitation matrices):** All the quantitation files are in the folder "Quant," which is available at http://dictygenome.org/supplement/gadi/pnas_0306983101/Katoh_supplement.zip. See Table 2 "Quantitation data map" for detail.
- **Data selection and transformation procedures:** See Part 2 below.
- **Final gene expression data table(s) used by the authors to make their conclusions *after* data selection and transformation (gene expression data matrices):** All the normalized data are in the folder "Norm," which is available at http://dictygenome.org/supplement/gadi/pnas_0306983101/Katoh_supplement.zip. We provide normalized data on all of the array targets and on the selected data presented in Figs. 2 and 3 of the manuscript. See Table 3 "Normalized data map" for detail.

Array Design

- **General array design, including the platform type, surface, and coating specifications:** A spotted glass array. The slides were coated with 3-glycidoxypropyltrimethoxysilane (Aldrich). Target DNA was printed on the activated glass slides on 200-µm centers with a Cartesian Pixsys5500 robot using Chipmaker II pins (Tele-chem Intl.). The full array has 8 × 4 subarrays, each containing 22 × 22 dots. Meta column: 4 Meta Row: 4; Column: 22 Row: 22; in duplicate (4 × 8-22 × 22). See Table 4 "Array Map" for detail.
- **For each reporter, its type:** 5,655 cDNA clones from the *Dictyostelium* cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html); 987 cDNA clones were selected from a low-redundancy screen of a lambda library and a plasmid library of cDNA from late developmental stages and from vegetative and early developmental stages of AX4 cells, respectively; 647 genomic DNA clones from the *Dictyostelium* Genome Project at Baylor College of Medicine (http://dictygenome.bcm.tmc.edu/) were selected as long open reading frames that matched published protein sequences; and 96 clones were from miscellaneous sources. Sequences were compared to public databases and annotated (1). The degree of redundancy is less than 20%. The array also contained 198 control targets that were made from the *Dictyostelium* ribosomal 17S RNA gene, histone H1, actin8, and *mhcK* as well as control targets from yeast genes and "no DNA" controls. Altogether, the array contained 7,744 targets. The entire array was printed in duplicate.
- **The source of the reporter molecules:** 5,655 cDNA clones from the *Dictyostelium* cDNA project (http://www.csm.biol.tsukuba.ac.jp/cDNAproject.html); 987 cDNA clones were selected from a low-redundancy screen of a lambda library and a plasmid library of cDNA from late developmental stages and from vegetative and early developmental stages of AX4 cells, respectively; 647 genomic DNA clones from the *Dictyostelium* Genome Project at Baylor College of Medicine (http://dictygenome.bcm.tmc.edu/) were selected.

- **The method of reporter preparation:** DNA targets were amplified from plasmids by PCR with common oligonucleotides and their size was verified by gel electrophoresis. PCR products were purified by precipitation with 50% isopropyl alcohol/0.3 M sodium acetate, pH 5.2, washed once with 70% ethanol, dissolved in water, and adjusted to 800 mM sodium chloride 200 mM sodium phosphate, pH 10.5.
- **The spotting protocols used, including the array substrate, the spotting buffer, and any postprinting processing, including cross-linking:** Glass slides (Gold Seal Products, VWR) were washed by sonication in acetone for 10 min, rinsed twice in distilled water, immersed in 1 M NaOH for 10 min, washed in distilled water, and immersed for 3 min in 3% (vol/vol) 3-glycidoxypropyltrimethoxysilane (Aldrich) made in 95% ethanol/acetic acid, pH 5.0. The slides were washed in 100% ethanol, dried with nitrogen gas, and baked for at least 2 hours at 100°C. Target DNA in 800 mM sodium chloride/200 mM sodium phosphate, pH 10.5, solution was printed on the activated glass slides on 200-µm centers with a Cartesian Pixsys5500 robot using Chipmaker II pins (Tele-chem Intl.). Arrays were stored desiccated in the dark.
- **Any additional treatment performed prior to hybridization:** The arrays were treated with hot water for 1 min at 100°C for target denaturation prior to hybridization.

Part 2. Quantitation, Normalization, and Analysis

We performed three steps of data processing:

(A) Quantitation: to convert the scanned microarray images to quantitative values.

(B) Pre-processing: to normalize the quantitative values of single-microarray experiments.

(C) Processing: to normalize the data across multiple chips.

Finally, we used clustering methods (D) to compare the different biological experiments and to visualize the data. These steps are explained in detail below.

(A) Data quantitation (using GLEAMS)

Quantitation of microarray image files was performed using the GLEAMS (NuTec Sciences) software. GLEAMS detects and quantifies the expression spots automatically in a batch mode for high throughput of parallel data processing. The batch auto-alignment is based on a method requiring only knowledge of the number of rows and columns of dots in the array. Each spot is quantified by taking a 5% trimmed mean of the pixel values within the spot in each channel.

(B) Data preprocessing (single-chip normalization)

After quantitation, data files were passed through a single-chip normalization procedure. The role of normalization is to correct for spatial and intensity artifacts, to estimate the variability of replicate log-ratios, and to bring the data to a common measurement scale to allow for subsequent multiarray comparisons. The normalization was implemented in six consecutive steps: 1, data rejection/thresholding; 2, quantile adjustment of single-channel values; 3, bias adjustment; 4, averaging of on-chip replicates; 5, by-signal-size variance estimation; and 6, scaling of the final values using the estimated by-signal-size variance. The process is explained in detail below:

Step 1: Data rejection/thresholding. Thresholding was performed to reject low-quality spots prior to any other step in the normalization. Thresholding removed spots with low signal intensity, high background, or abnormal pixel area. The methodology was to estimate the univariate distribution of spot characteristics on each array by using the observed single channel intensities, background, and spot pixel area from values in the quantitation files. The fit was determined by using the robust L2E estimation procedure (2). The fitting returned a mean and variance value for each characteristic. Then, for each spot, standardized quantities are calculated for all three characteristics (intensity, background, and area):

$$Z_{i,character} = \frac{X_{i,character} - \mu_{character}}{\sigma_{character}}$$

where μ and σ are estimated from the robust fitting and *i* indexes the spot. Again, such separate standardized values were formed for each of the characteristics (intensity, background, and area) for each spot. The rule for excluding spots was:

Reject if $|Z_{i,area}| > 3 \vee Z_{i,Ch1} < -3 \vee Z_{i,Ch2} < -3 \vee Z_{i,BG1} > 3 \vee Z_{i,BG2} > 3$

where Ch1 and Ch2 represent intensity and BG1 and BG2 represent background in the respective channels.

Step 2: Quantile adjustment of single channel values. Next, we perform a quantile adjustment on the single-channel data for all the arrays in a given time course. The quantile adjustment made the distribution of intensities the same for each channel in all arrays under consideration. The quantile adjustment was performed on many arrays at the same time (a multiarray procedure). For computational ease and speed we performed this adjustment on individual time courses (a bundle of 10-12 arrays for the experiments in this study).

Step 3: Bias adjustment. After thresholding and quantile adjustment, we calculated log-ratios (base 2) for each spot on each array and removed single-chip biases from the log-ratio data. Two forms of biases have been observed: spatial patterns and intensity-dependent patterns. Such patterns may arise as artifacts of the printing, hybridization, or scanning procedures.

In the following model we consider: two bias (shift) terms to account for spatial [$f(x, y)$] and intensity [$h(s)$] artifacts, an array target-specific term to estimate the "true" expression value for each gene (g), and a statistical error term (e).

$$LR' = f(x, y) + h(s) + g + e$$

LR' is the log-ratio value for individual spots that results from multiarray quantile adjustment of the single-channel values.

$f(x, y)$ is the smooth spatial adjustment of the log-ratio values where x and y are the pixel coordinates of spots on the array. A smoothing parameter of 0.05 was chosen to give the best robust performance according to the criterion of the number of genes that show significant time pattern (ANOVA F test, time treated a factor) after normalization.

$h(s)$ is the smooth intensity adjustment of the log-ratio values. The s value is the sum of the log (base 2) for single-channel intensity values (raw, not quantile-transformed). We used a smoothing parameter of 0.45 that was determined as described above.

g is the gene effect; after subtraction of the spatial and intensity biases, these values are taken to be the by-gene averages of the residual values. Our array has 2-fold replication of features, and we find on-chip error to be very small after spatial and intensity biases are removed.

e is a statistical error term modeled to have a mean of 0 and a variance that is signal-intensity dependent. We fit variance as a function of intensity by using "loess" and a smoothing parameter of 0.45.

These fits were performed in a stepwise manner, first fitting and subtracting the spatial term $f(x,y)$ and then fitting the intensity adjustment $h(s)$ to the residual values. The spatial adjustment made the overall median of the single chip adjusted log-ratios very close to zero; this feature accounts for the overall mean component of the stepwise process

Step 4: Averaging of on-chip replicates. With the exception of highly replicated control spots, our chip contains 2-fold replication of each target. The bias-adjusted log-ratios were calculated for each group of replicate spots on each array. Residuals were then computed for the replicate spots about each by-target mean. The residuals allowed for estimation of the variance of the error term in the stepwise process for spot values on each array. The variance estimates provided a method for scaling the by-spot mean values. Scaling, as has been noted by Yang *et al.* (3), enhances multi-array comparisons.

Step 5: By-signal-size variance estimation. The estimation of error variance for replicate spots proceeded from graphical evidence that the variance of the errors may depend on the brightness of the spots. We therefore fit a nonparametric regression (using the loess procedure in S-Plus) to relate the variance of residuals to the average brightness of the replicate spots. We used a smoothing parameter of 0.4 in this loess step, but we set a ceiling for the highest possible variance at a fixed threshold of 0.5. The units of variance were squared log-ratios on the log (base 2) scale, so that a variance of 0.5 has an interpretation of 0.7071-fold.

Step 6: Scaling of the final values. The last component of the single-chip normalization is to scale the processed log-ratios by the estimated by-signal-size standard deviation. This rescaling step makes use of the on-chip variances for each target. The rescaling adjustment has the consequence of making log-ratios across signal brightness more comparable within the chip, as higher variances for dim spots may inflate these log-ratios due to error alone. The scaling adjustment has the benefit of making multiarray analysis more successful, as the scaled values can make the measurement standardized. Therefore, we divide the estimated *g* value (see step 3) by the square root of the variances to obtain the single-array normalized values.

(C) Data processing [multi-chip normalization using an analysis of covariance (ANCOVA) model]

To minimize the impact of experimental variation we perform the experiment with three kinds of replication (see Table 1). First, we have on-chip replication: printing of the same target material in multiple locations on each array, which allows for the single-chip normalization procedure described in Section B. In addition, we analyze multiple arrays from the same RNA extraction. The use of multiple chips from the same RNA accounts for hybridization variation. Finally, we perform the assay with three or more biological replicates of each dedifferentiation treatment. We integrate both the technical and biological replicates to form normalized dedifferentiation data.

Data were multiarray scaled and combined into multi-experiment sets for all subsequent analysis (4). An ANCOVA model was fit to each time course. The model contains a categorical term for hybridization batch effects and continuous terms for representing each gene as a time function. The smooth curves were fit to a polynomial basis of degree 5 generated by using the poly() function in S-Plus. The result of the analysis is a set of coefficients for each gene at each time point in each dedifferentiation treatment, as well as an across-time mean for each gene. These values are available in the folder "Norm," which is available at http://dictygenome.org/supplement/gadi/pnas_0306983101/Katoh_supplement.zip. One file is given for each of the three dedifferentiation treatments.

Model for time-course fitting (ANCOVA)

$$g(t) = \mu + b + \sum_{i=1}^{1=5} a_i t'^{i} + e$$

The time model considers each gene to have a batch effect *b*, an overall mean, and a time pattern. The model is fit in a stepwise manner, first subtracting the across time-course mean for each batch for each gene. Then, the residual values are fit to a time function by using the aov() function in S-Plus. The *t'* values are the transformed (rotated) orthonormal basis determined by using the poly() function in S-Plus. The use of the orthonormal basis makes downstream clustering of fit coefficients more effective.

D. Data analysis (comparison with developmental data, clustering, and visualization)

To visualize the result and to compare with our previous results from *Dictyostelium* development, we selected a set of approximately 2,000 genes by filtering on the *T*-statistic contrast described in Van Driessche *et al.* (1). Briefly, a contrast score was formed for each gene based on a standardized additive combination of developmental expression values against a set of coefficients determined by a line stretching from –1 to +1 for the 13 time points of the developmental time course (figure 2 in ref. 1). Clustering was performed on the smooth time fit coefficients from the three experiments. We concatenated the three sets of coefficients from each experiment into a 15-element vector of coefficients for each gene. Clustering was performed recursively by *k*-means while varying *k* from 2 to 10. The coefficients from the fits in each treatment were then combined into a single data set, and *k*-means clustering was performed on the smooth curve coefficients. One of the clusters in the *k*-means result showed strong up-regulation at the time of the dedifferentiation transition. This group consists of 270 genes, a number later refined to 259 to account for redundancy in the cDNA library.

Unique scores were determined for each gene to order the genes according to how well they match this pattern, and this ordered set of genes is displayed in the heat maps shown in Fig. 3*A*. The scores in the file "Fig3A" in the folder "after data selection" in "Norm" at http://dictygenome.org/supplement/gadi/pnas_0306983101/Katoh_supplement.zip represent the degree to which a gene matches the mean pattern within each dedifferentiation treatment for the group of 270 genes (see Fig. 3*A* in the printed text).

In addition to the unsupervised analysis, a directed analysis was performed using both the developmental time course and the dedifferentiation time courses. We have generated a filter to select for genes that: (*i*) have a low fit to the developmental consensus pattern (to find dedifferentiation-specific genes); (*ii*) follow the pattern T_{max}(Mex) > T_{max}(Finger) > T_{max}(Agg), where T_{max} is the time of maximal expression and Mex, Finger, and Agg represent the respective developmental stages (to account for the finding that the dedifferentiation timing is directly proportional to the initial developmental time); and (*ii*) whose expression level at the T_{max}(Agg) is greater than their level of expression at the aggregation stage (to find genes that are induced during dedifferentiation). We found a list of 120 genes satisfying this criterion. We examined the significance of this group by (*i*) generating random filters and examining the size of the groups we generated and (*ii*) looking at the within-group variance of randomly generated cohorts of 120 genes. The group we found is significant according to both of these criteria. To test for nonrandomness we generated 10,000 random filters where T_{max}(Agg) was replaced with a random T and measured group size and variance. The filtered group exhibited an unusually large statistic with an estimated tail probability of less than 0.05 for both the size and within-group variance statistics. The values are provided in the file "Fig3B" in the folder "after data selection" in "Norm" at http://dictygenome.org/supplement/gadi/pnas_0306983101/Katoh_supplement.zip, corresponding to Fig. 3*B* in the printed text.

1. Van Driessche, N., Shaw, C., Katoh, M., Morio, T., Sucgang, R., Ibarra, M., Kuwayama, H., Saito, T., Urushihara, H., Maeda, M., Takeuchi, I., Ochiai, H., Eaton, W., Tollett, J., Halter, J., Kuspa, A., Tanaka, Y. & Shaulsky, G. (2002) *Development (Cambridge, U.K.)* **129,** 1543-1552.

2. Scott, D. W. (2001) *Technometrics* **43**, 274-285.

3. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. & Speed, T. P. (2002) *Nucleic Acids Res.* **30**(4), e15. www.stat.berkeley.edu/users/terry/zarray/Html/normspie.html.

4. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. (2003) *Bioinformatics* **19,** 185-193.