Electronic Supplementary Material (ESI) for Molecular BioSystems

Gene expression scaled by distance to the genome replication site

Ying *et al*

Supporting Information

Contents

Supplementary note I

FCM data processing

Population analysis was performed by flow cytometry. The flow data sets were analyzed using custom-designed scripts written in R as previously described[1, 2]. In the analysis, the data sets were processed through several filtering steps, a total of four steps from the raw data (Fig. S1). To show the effects of each processing step, the distributions or density plots of three examples are presented, as follows: MDS42 (left panels), MDS42*ΔleuC*::*Ptet_gfp* (type II, middle panels), and MDS42*ΔleuC*::*gfp* (type I, right panels). These three constructs enabled autofluorescence of *E. coli* cells and relatively high and low levels of *gfp* expression in cells, respectively. First, the raw data sets for the fluorescence beads, which were loaded to calculate the cell concentration, were gated appropriately. In addition, the events that occurred at the bottom or top of the detector's range, which could be due to systematic errors, were eliminated (A, arrows). The data were then gated with a narrow FSC window at the peak of the FSC histogram to reduce cell size variability (B, blue lines). The resultant distributions of cellular GFP abundance (*i.e.*, GFP FI) still contained the high and low extreme outliers of green fluorescence intensity (C, arrows). To prevent these unreliable rare events, one percent of the total cells representing both the highest and the lowest values of fluorescence intensity were removed (D). Finally, the autofluorescence (D, distributions in gray) was subtracted, yielding the true distributions of fluorescence intensity (E). The mean values of GFP FI were calculated accordingly. A total of 6 to 8 measurements (GFP FI distributions) were used for each genetic construct to acquire the mean GFP FI values shown in Fig. 2B (Table S3).

Supplementary note II

Genome replication models

The linear regression for the expression level (*f*), as a function of the distance from *oriC* (*x*), can be described by the following equation (Eq. 1.1). In the present study, the value of *a*/*b* was experimentally determined to be –3.3 to –2.6×10^{-7} bp^{-1}.

$$f(x) = ax + b \qquad \text{(Eq. 1.1)}$$

To verify whether this experimentally determined value was reasonable or reliable, we constructed a mathematical model, presented in Eq. 2.1, in which the expression level (*f*) is simply determined by the promoter activity (*p*) multiplied by the gene dosage (*d*). The promoter activity, which can be considered the expression efficiency, is proportional to the protein abundance per single gene per cell.

$$f = p \cdot d \qquad \text{(Eq. 2.1)}$$

The gene dosage (*d*) scales by the distance from *oriC* (*x*), which depends on the number of replication forks on chromosome. In particular, the gene dosage (*d*) scales down from the copy number (*c*) at the replication initiation site (*x* = 0) to 1 copy at the termination region (*x* = *G*/2). That is, the value of *d* varies from 1 to *c*. Here, *c* depends on the mode of replication, for instance, the dual-fork replication for rapid growth and the single-fork replication for slow growth. Considering a simple linear case, the gene dosage (*d*) is proportional to the distance from *oriC*, as the following equation described.

$$d = (c-1)\left(1 - \frac{x}{\frac{1}{2}G}\right) + 1 = c - \left(\frac{c-1}{\frac{1}{2}G}\right)x \qquad \text{(Eq. 2.2)}$$

Here, G represents the genome size. The first term, c, represents the gene dosage at the replication initiation site (x = 0), and the second term, $\left(\frac{c-1}{\frac{1}{2}G}\right)x$, represents how much the gene dosage decreases with x. The maximal decrease of the gene dosage is (c–1), when $x = G/2$. Introducing Eq. 2.2 into Eq. 2.1 results in the following equation, which is identical to Eq. 1.2 in the main text.

$$f = p\left\{c - \left(\frac{c-1}{\frac{1}{2}G}\right)x\right\} \qquad \text{(Eq. 2.3)}$$

Comparing Eq. 1.1 with Eq. 2.3, the following relations among the parameters used in the linear model were drawn.

$$a = -p\left(\frac{c-1}{\frac{1}{2}G}\right) \qquad \text{(Eq. 2.4)}$$
$$b = pc$$

Dual-fork replication model

According to the bidirectional dual-fork replication model[3], the gene dosage (d) scales down from 4 copies (c = 4) at the replication initiation site (x = 0) to 1 copy at the termination region ($x = G/2$), that is, the value of d varies from 1 to 4. G represents the genome size, which was 3.98×10^6 bps for strain MDS42[5]. By introducing this relationship into the mathematical model (Eqs. 2.2 and 2.3), the expression level can be rewritten in Eq. 2.5.

$$d = 3\left(1 - \frac{x}{\frac{1}{2}G}\right) + 1 \qquad \text{(Eq. 2.5)}$$

$$f(x) = p\left\{3\left(1-\frac{x}{\frac{1}{2}G}\right)+1\right\} = -\frac{6p}{G}x + 4p \quad \text{(Eq. 2.6)}$$

Consequently, the following equation (Eq. 2.7.1) is obtained by either comparing the linear relationship (Eq. 2.6) with the linear regression (Eq. 1.1) or directly from Eq. 2.4. The following theoretical value was calculated in accordance with that relationship (Fig. 2C, bottom gray line).

$$\frac{a}{b} = -\frac{3}{2G} = -3.77\times10^{-7}\ \text{bps}^{-1} \quad \text{(Eq. 2.7.1)}$$

Single-fork replication model

According to the same analytical procedure (Eqs. 2.5 and 2.6), the scaling down of the gene dosage (*d*) from 2 (*c* = 2) to 1 copy in the single-fork replication model[4] yields the following theoretical value (Fig. 2C, upper gray line).

$$\frac{a}{b} = -\frac{1}{2G} = -1.26\times10^{-7}\ \text{bps}^{-1} \quad \text{(Eq. 2.7.2)}$$

Supplementary note III

Parallelism and equivalence in the linear regressions of scaled expression

The linear regression of the expression level (*f*) as a function of the distance from *oriC* (*x*) can be described by the following equation (Eq. 1.1), where *x* represents the distance from *oriC*.

$$f(x) = ax + b \quad \text{(Eq. 1.1)}$$

According to this equation, the relationships between the raw expression level and the distance from *oriC* were described by the following formulas (Fig. S2, upper):

Type II $f(x) = -4.03\times10^{-6}x + 12.9$ (Eq. 1.1.1)

Type III, no IPTG $f(x) = -1.05\times10^{-6}x + 3.8$ (Eq. 1.1.2)

Type III, 10 μM IPTG $f(x) = -5.53\times10^{-7}x+2.2$ (Eq. 1.1.3)

First, the equality (equivalence) of these three regression coefficients[10] was tested. Similar to the analysis of covariance (ANCOVA), the null hypothesis was that the slopes (coefficients) of these regression lines were the same (H_0), while the alternative was they were not the same (H_1). The statistical analysis (3 groups, 37 gene expression levels in each group (Fig.1B)) yielded an F value of 68.44, given by the ratio of the regression mean square to the error mean square (*MSR/MSE*). According to the *F* distribution (degrees of freedom d1=3-1=2, d2=3×37-2×3-22=83, 22 is the number of missing values), the difference in coefficients was highly significant ($p=2\times10^{-16}$), as can be observed from the plots (Fig. S2, upper) without statistical tests. Thus, these regression lines are not parallel, and there were interactions among the groups (data sets).

Second, when the plots were normalized by the intercept *b*, scaled expression levels with similar slopes were acquired (Fig.S2, bottom):

Type II $a/b = -3.28\times10^{-7} = k_1$

Type III, IPTG 0 μM $a/b = -2.79\times10^{-7} = k_2$

Type III, IPTG 10 μM $a/b = -2.55\times10^{-7} = k_3$

Consequently, the linear regression formulas (Eq. 1.1.1-1.1.3) could be rewritten as the following equations (Eq. 1.1.4-1.1.6). Note that the linearity of the chromosomal location-mediated gene dosage effect is maintained by scaling.

Type II $g(x) = k_1x+1$ (Eq. 1.1.4)

Type III, no IPTG $g(x) = k_2x+1$ (Eq. 1.1.5)

Type III, 10 μM IPTG $g(x) = k_3x+1$ (Eq. 1.1.6)

The equality test of the regression slopes was performed on the scaled slopes as described above (*MSR/MSE* = 1.14), and the statistical significance (*p*-value) was as low as 0.325. This means that

the null hypothesis, that is, that the slopes (k_1, k_2 *and* k_3) are the same, could not be rejected. In other words, the coefficients of regressions were equivalent.

Supplementary note IV

Experimental and analytical details

Genetic constructs

The *E. coli* strain MDS42[5] was purchased from Scarab Genomics (Madison, WI, USA) and used as the host strain. The Red system was used to replace the target gene with the reporter gene, *gfp*, either with or without the universal promoter P_{tet}. The fast-maturing *gfp* used is identical to the reporter gene used in our previous studies. Both the P_{tet}-containing and GFP-only sequences were PCR-amplified (Table S1) from the pBRgalKGR plasmid (pBR322 derivative), which contained the P_{tet}-*gfp*-P_{kan}-*Km* cassette downstream of a terminator sequence for genome replacement, as previously described[6]. The manipulation of the homologous recombination and the phenotypic and genetic verification of the transformants (positive colonies) were performed as described previously[6]. Subsequently, a plasmid carrying a repressor protein, TetR, for controlling the P_{tet}-derived gene expression was introduced into each construct. The plasmid (pBRTetRCm) was constructed by simply removing the *rfp* sequence from a previously reported plasmid (pBRintC series) in which the *rfp*, *tetR*, and *cat* genes are under the control of the *lac* promoter (P_{lac}) and its operator[6]. The final collection of constructs is summarized in Table S2.

Cell culture

Cell culture was performed using M63 minimal medium supplemented with 19 amino acids at a final concentration of 0.2 mM each and tyrosine at a concentration of 0.05 mM. The cells carrying the plasmid (TetR) were cultured in the presence of the necessary antibiotics, 50 μg/mL ampicillin,

and the inducer IPTG as indicated. The cell cultures were controlled during the early logarithmic growth phase, and the final cell concentrations were approximately 10^7 cells/mL. The cell concentration was measured by flow cytometry. All cell cultures were transferred for several passages (days) to acquire multiple measurements of the growth rate and GFP abundance. The growth rates were calculated based on the initial and final cell concentrations, as described previously[9]. A total of six to eight replicate cultures were used for each genetic construct (Table S2).

Flow cytometry

The GFP expression (fluorescence intensity) and relative cell size were evaluated using a flow cytometer (FACSCanto™II; Becton Dickinson) equipped with a 488-nm argon laser and a 515–545-nm emission filter (GFP). The following PMT voltage settings were applied: forward scatter (FSC), 280; side scatter (SSC), 400; GFP, 600. The flow rate for the sample measurements was set to low. The cell samples, which were mixed with known concentrations of fluorescent beads (3 μm Fluoresbrite YG Microspheres; Polysciences), were loaded to calculate the cell concentration. The flow data sets were analyzed using custom-designed scripts written in R as previously described[1, 2]. Examples are shown in Fig. S1. The data processing procedure is described in detail in Supplementary note I.

Omics data and statistical analyses

The proteome data sets reporting the mean protein abundance were from Taniguchi *et al.*[7] and Ishihama *et al.*[8], respectively. The transcriptome data sets reporting the gene expression profiling of MDS42 and MG1655 were from Ying *et al.*[9] (GEO Series accession number GSE33212). The mean values of seven replicates of exponential growth at 37°C were applied. The gene names and the chromosomal positions (distances from *oriC*) were based on the following publicly deposited

genome sequences: MDS42, DDBJ accession ID AP012306; MG1655, GenBank accession ID U00096; MC4100, GenBank accession ID NC012759. Statistical analyses were performed using the software package R. Pearson's correlation coefficient was used to assess the association between gene expression levels and the distance from *oriC*. The *p*-value indicates the significance of the correlation by testing the null hypothesis of no linear relationship between expression level and chromosomal location.

Supplementary references

1. Y. Shimizu, S. Tsuru, Y. Ito, B. W. Ying and T. Yomo, *PLoS One*, 2011, **6**, e23953.
2. S. Tsuru, J. Ichinose, A. Kashiwagi, B. W. Ying, K. Kaneko and T. Yomo, *Phys Biol*, 2009, **6**, 036015.
3. H. Yoshikawa, A. O'Sullivan and N. Sueoka, *Proc Natl Acad Sci U S A*, 1964, **52**, 973-980.
4. S. Cooper and C. E. Helmstetter, *J Mol Biol*, 1968, **31**, 519-540.
5. G. Posfai, G. Plunkett, 3rd, T. Feher, D. Frisch, G. M. Keil, K. Umenhoffer, V. Kolisnychenko, B. Stahl, S. S. Sharma, M. de Arruda, V. Burland, S. W. Harcum and F. R. Blattner, *Science*, 2006, **312**, 1044-1046.
6. B. W. Ying, Y. Ito, Y. Shimizu and T. Yomo, *J Biosci Bioeng*, 2010, **110**, 529-536.
7. Y. Taniguchi, P. J. Choi, G. W. Li, H. Chen, M. Babu, J. Hearn, A. Emili and X. S. Xie, *Science*, 2010, **329**, 533-538.
8. Y. Ishihama, T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner and D. Frishman, *BMC Genomics*, 2008, **9**, 102.
9. B. W. Ying, S. Seno, F. Kaneko, H. Matsuda and T. Yomo, *BMC Genomics*, 2013, **14**, 25.
10. J.H. McDonald. *Handbook of Biological Statistics*, 2nd ed. Sparky House Publishing, Baltimore, Maryland, 2009.

Supplementary figures and figure legends

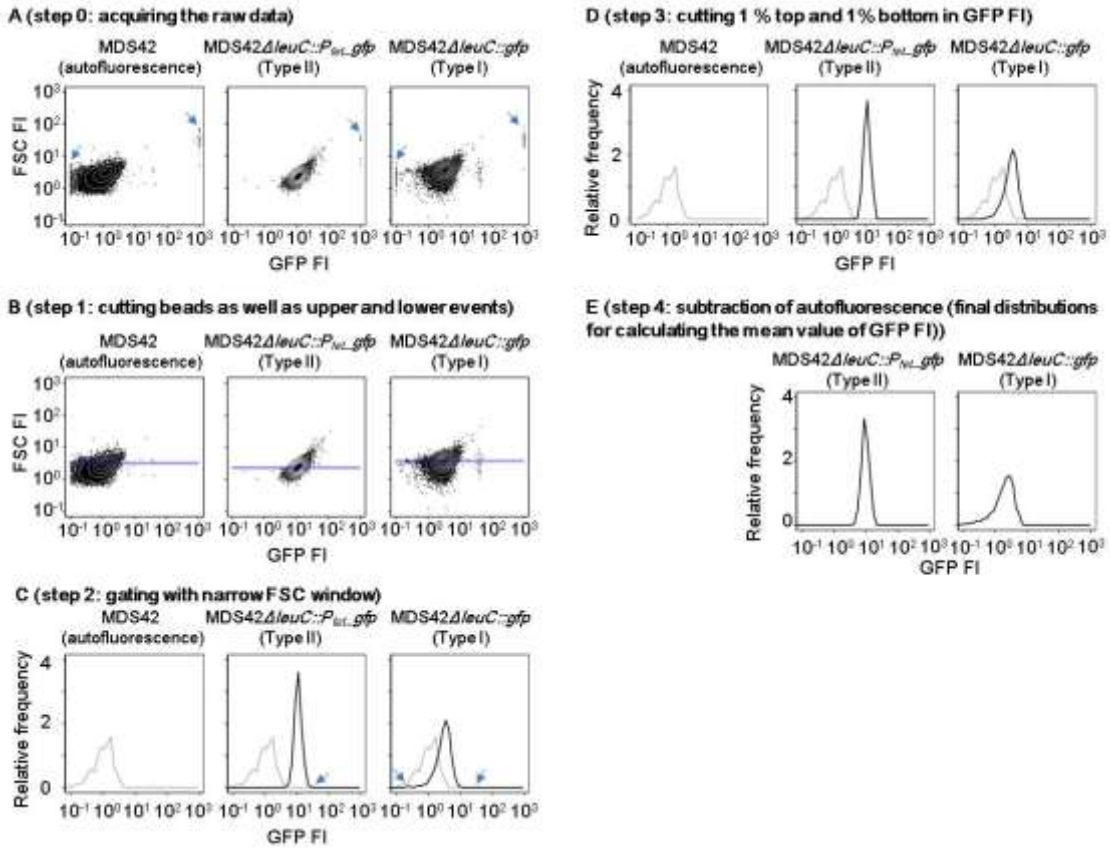

Figure S1. FCM data processing. Examples of the data processing, which comprised four steps, are illustrated.

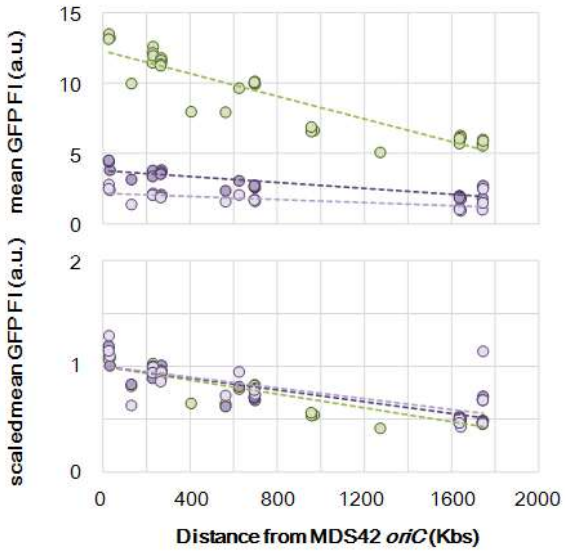

Figure S2 Chromosomal location dependency scaled by genome position. The upper and lower panels represent the raw and scaled gene expression levels (mean GFP FI), respectively. The type II (pistachio) and type III constructs (lilac, colors in light and dark indicate the growth conditions in the presence and absence of 10 μM IPTG, respectively) are shown. The correlation coefficients (R^2) of the linear regressions (upper panel) were 0.85 (green), 0.75 (dark lilac) and 0.51 (light lilac), and those of the corresponding exponential regressions were 0.87, 0.76 and 0.56, respectively.

Supplementary tables

Table S1. Primers. Primers used for genome replacement and genetic confirmation of the constructs.

Table S2. Strains and clones. The genetically engineered *E. coli* cells comprising 3 types of genetic designs (types I, II and III) used in this study are summarized. The growth rates within the early exponential growth phase are indicated. μ1–4, μ_ave, and μ_sd indicate the growth rates of the repeated cell cultures, the average growth rate of the corresponding construct, and the standard deviation of the growth rate, respectively.

Table S3. The mean values of GFP FI of the three types of genetic constructs. Data sets of the average GFP abundance (mean) and the standard deviation (sd) of repeated experiments are shown. The data sets were used to generate the plots in Fig. 2B.