Gauss-power mixing distributions comprehensively describe stochastic variations in RNA-seq data

Akinori Awazu[1,4,*], Takahiro Tanabe[1], Mari Kamitani[2], Ayumi Tezuka[2], Atsushi J. Nagano[2,3]

[1]Department of Mathematical and Life Sciences, Hiroshima University, Kagamiyama 1-3-1, Higashi-Hiroshima, Hiroshima 739-8526, Japan

[2]Research Institute for Food and Agriculture, Ryukoku University, Yokotani 1-5, Seta Ohe-cho, Otsu, Shiga 520-2194, Japan

[3]Faculty of Agriculture, Ryukoku University, Yokatani 1-5, Seta, Ohe-cho, Otsu-shi, Shiga 520–2194, Japan

[4]Research Center for Mathematics on Chromatin Live Dynamics, Hiroshima University, Kagamiyama 1-3-1, Higashi-Hiroshima, Hiroshima 739-8526, Japan

Abstract

Motivation: Gene expression levels exhibit stochastic variations among genetically identical organisms under the same environmental conditions. In many recent transcriptome analyses based on RNA sequencing (RNA-seq), variations in gene expression levels among replicates were assumed to follow a negative binomial distribution although the physiological basis of this assumption remain unclear.

Results: In this study, RNA-seq data were obtained from *Arabidopsis thaliana* under eight conditions (21–27 replicates), and the characteristics of gene-dependent distribution profiles of gene expression levels were analyzed. For *A. thaliana and Saccharomyces cerevisiae,* the distribution profiles could be described by a Gauss-power mixing distribution derived from a simple model of a stochastic transcriptional network containing a feedback loop. The distribution profiles of gene expression levels were roughly classified as Gaussian, power law-like containing a long tail, and mixed. The fitting function predicted that gene expression levels with long-tailed distributions would be strongly influenced by feedback regulation. Thus, the features of gene expression levels are correlated with their functions, with the levels of essential genes tending to follow a Gaussian distribution and those of genes encoding nucleic acid-binding proteins and transcription factors exhibiting long-tailed distributions.

Availability: Fastq files of RNA-seq experiments were deposited into the DNA Data Bank of Japan Sequence Read Archive as accession no. DRA005887. Quantified expression data are available in supplementary information.

Contact: awa@hiroshima-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Stochastic variations in gene expression—known as gene expression noise or phenotype fluctuation—have been observed among individuals in a genetically identical population under the same environmental conditions (Elowitz et al., 2002; Furusawa et al., 2005; Golding et al., 2005; Kaern et al., 2005; Newman et al., 2006; Chang et al., 2008; Konishi et al., 2008; Taniguchi et al., 2010; So et al., 2011; Silander et al., 2012; Woods, 2014). Such variations are thought to be important for maintaining the pluripotency of embryonic stem cells, cell fate decisions, and cellular differentiation in multicellular organisms (Mitsui et al.,

2003; Kaneko, 2006; Kalmar et al., 2009; Ochiai et al., 2014). In rice, genes related to stress responses exhibited larger variations than those involved in other processes (Nagano et al., 2012). Furthermore, recent studies in *Escherichia coli*, the budding yeast *Saccharomyces cerevisiae*, and *Arabidopsis thaliana* have reported that the magnitude of gene expression noise is positively correlated with plasticity—i.e., the variation in expression levels due to mutation or environmental change (Sato et al., 2003; Blake et al., 2003; Landry et al., 2007; Choi and Kim, 2008; Choi and Kim, 2009; Tirosh and Barkai, 2008; Lehner, 2010; Lehner and Kaneko, 2011; Bajic and Poyatos, 2012 ; Singh, 2013, Hirao et al., 2015).

Recent gene expression analyses with sufficiently large replicates have shown that in organisms as diverse as *E. coli* and mammals, fluctuations in protein expression level for a given gene follow a log-normal distribution (Sato et al., 2003; Furusawa et al., 2005; Chang et al., 2008; Konishi et al., 2008). The closely related Frechet distribution was also proposed to describe variations in gene expression levels in *E. coli* and *S. cerevisiae* (Salmann et al., 2012). On the other hand, mathematical modeling of protein expression in *E. coli* suggested that such variations were more closely approximated by a gamma distribution, which is often considered as log-normal (Friedman et al., 2006; Taniguchi et al., 2010).

In many recent high-throughput RNA sequencing (RNA-seq) studies (Mortazavi et al., 2008; Nagalakshmi et al., 2010), variations in gene expression (transcription) levels among replicates were assumed to follow a negative binomial (NB) distribution (Marioni et al., 2008; Robinson et al, 2008; Rapaport et al., 2013; Gierlin'ski et al., 2015; Schurch et al., 2016). An analysis of RNA-seq data from a two-condition, 48-replicate experiment using *S. cerevisiae* revealed that variations in expression levels for each gene conformed to both log-normal and NB distributions (Gierlin'ski et al., 2015). Beta-binomial and Benford distributions have been proposed for fitting gene expression data obtained by RNA-seq (Smith et al., 2016; Karthik et al., 2016). However, the physiological basis of these distributions and the significance of associated parameters remain unclear. Furthermore, it is not known whether such model distributions are applicable to any genes in any organism, especially multicellular organisms.

Gene expression noise in plants has been investigated in rice and *Arabidopsis* (Nagano et al., 2012; Shen et al., 2012, Hirao et al., 2015). However, recent studies were based on transcriptome data from experiments with few replicates (Maruyama-Nakashita et al., 2005; Nemhauser et al., 2006; Kilian et al., 2007; Goda et al., 2008; Less and Galili, 2008; Wittenberg et al., 2012), which limited the inferences that could be made regarding the distribution characteristics of gene expression levels. In the present study, we analyzed RNA-seq data for *A. thaliana* under eight conditions (21–26 replicates) to determine

distribution profiles of gene expression noise (phenotype fluctuation) among individuals in a homogeneous population. We fitted the distribution profiles with a novel distribution function that we termed the Gauss-power (G-P) mixing distribution, which was derived from a simple stochastic transcriptional network model containing a feedback loop. The expression of genes showing a long-tail distribution was strongly influenced by a feedback mechanism; moreover, variations in gene expression levels were correlated with average expression levels and gene function.

2 Results

2.1 Analysis of Arabidopsis RNA-seq data

RNA-seq data from 7- and 22-day-old *Arabidopsis* shoots cultured under a 12:12-h light/dark cycle were obtained 1, 7, 13, and 19 h after the lights were turned on. There were 21 to 27 replicates for each condition. In total, 189 individual plants were analyzed by RNA-seq. We obtained 8.4 million reads on average; 1 sample with fewer than 1 million reads mapped to genes was omitted from subsequent analyses. The expression level of each gene was defined as the number of reads mapped to each gene per 1 million reads (Table S1). For each condition, we examined the distribution profiles of expression levels of ~10,000 genes (Table 1) whose expression levels could be regarded as stationary (see Materials and Methods).

Table 1. Number of replicates in the RNA-seq experiment and number of analyzed genes for each condition of *Arabidopsis*

Condition	Age (days)	Time after light (h)	Replicate	No. of analyzed genes
7-1	7	1	21	10,499
7-7	7	7	21	9287
7-13	7	13	25	10,760
7-19	7	19	24	10,735
22-1	22	1	22	9619
22-7	22	7	24	12,109
22-13	22	13	24	12,810
22-19	22	19	27	11,338

2.2 Cluster analysis of rank-expression level distribution (RED) profiles of genes

RED profiles were obtained for each gene under each condition (harvest time and age of plant) (Fig. 1). For most genes, the REDs showed typical profiles but were very noisy.

However, it is expected that groups of genes whose expression is similarly regulated would have similar RED profiles. We performed a cluster analysis to extract the essential properties of the RED profiles for each gene (see Materials and Methods).

For each condition, 12–15 clusters were obtained from normalized RED profiles (Tables S2 and S3), representing the relationship between standardized expression levels for mean = 0 and standard deviation = 1 and rescaled rank from 0 to 1. The average values of standardized expression levels of genes belonging to the same cluster were expected to reflect the essential features of their RED profiles (Fig. 1).

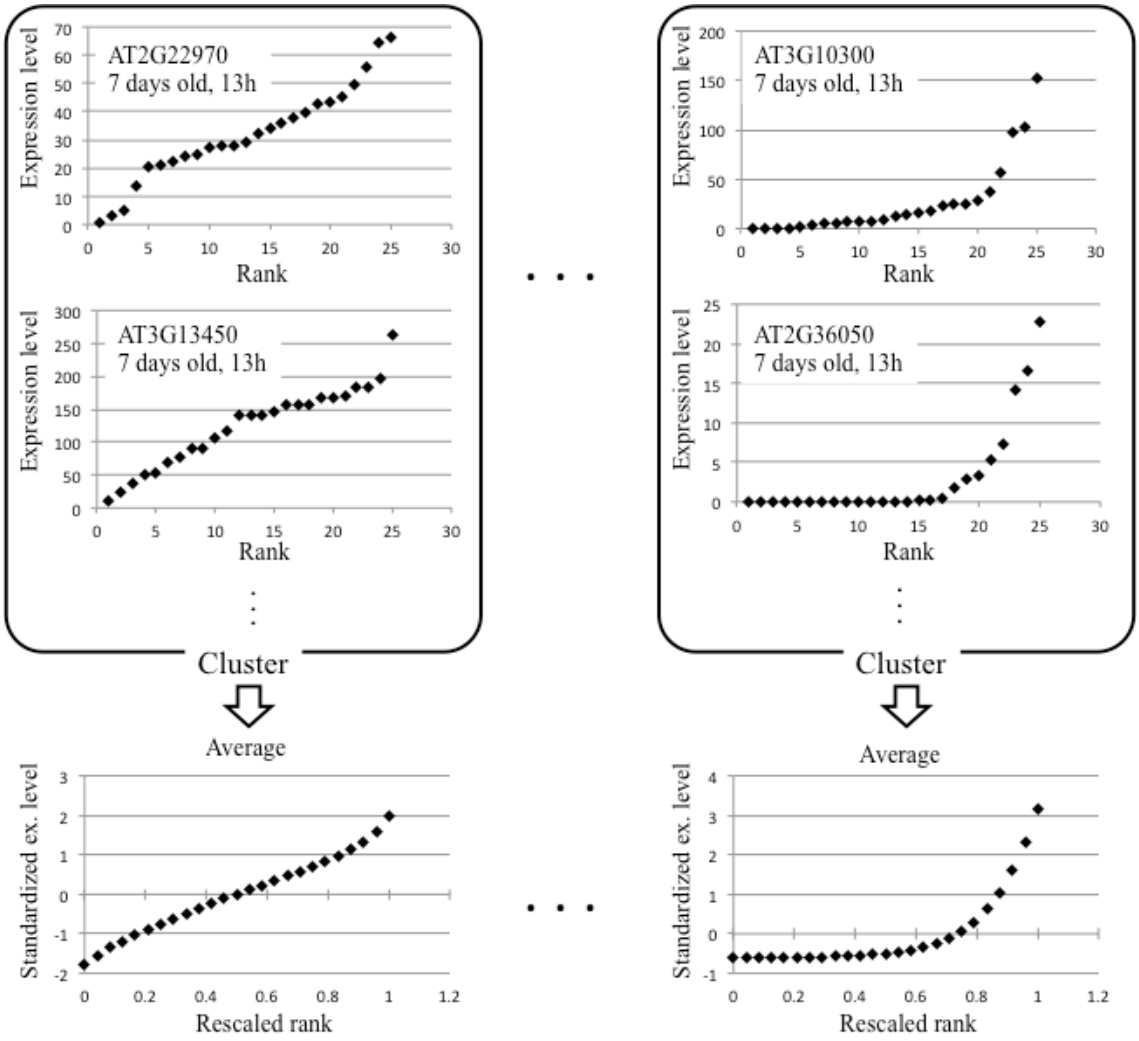

Figure 1. RED profiles of genes and average normalized RED profiles of their respective clusters. (Upper) RED profiles of indicated genes under specific conditions. As examples, results obtained from RNA-seq data at 13 h for 7-day-old *Arabidopsis* are shown. Similar RED profiles were grouped by cluster analysis. (Lower) Average normalized RED profiles for genes belonging to specific clusters.

2.3 Inferences on probability density distribution profiles of gene expression levels

Since gene expression levels are non-negative, we analyzed normalized RED functions that were shifted such that the minimum value on the vertical axis was assumed to be 0 (referred to as shifted normalized RED function) instead of the previous normalized RED functions. The value of the vertical axis of this function represented the rescaled expression level. The inverse function of the shifted normalized RED function provided the probability distribution of rescaled expression levels (Fig. 2), whose derivative yielded the profiles of probability density distribution function of rescaled expression levels (PDL) for each cluster (Figs. 2 and S1–S8). The derivative of this function was estimated by differential approximation (see supplementary information S1). PDL profiles obtained from the clusters showed variable shape, including Gaussian and power law-like distributions.

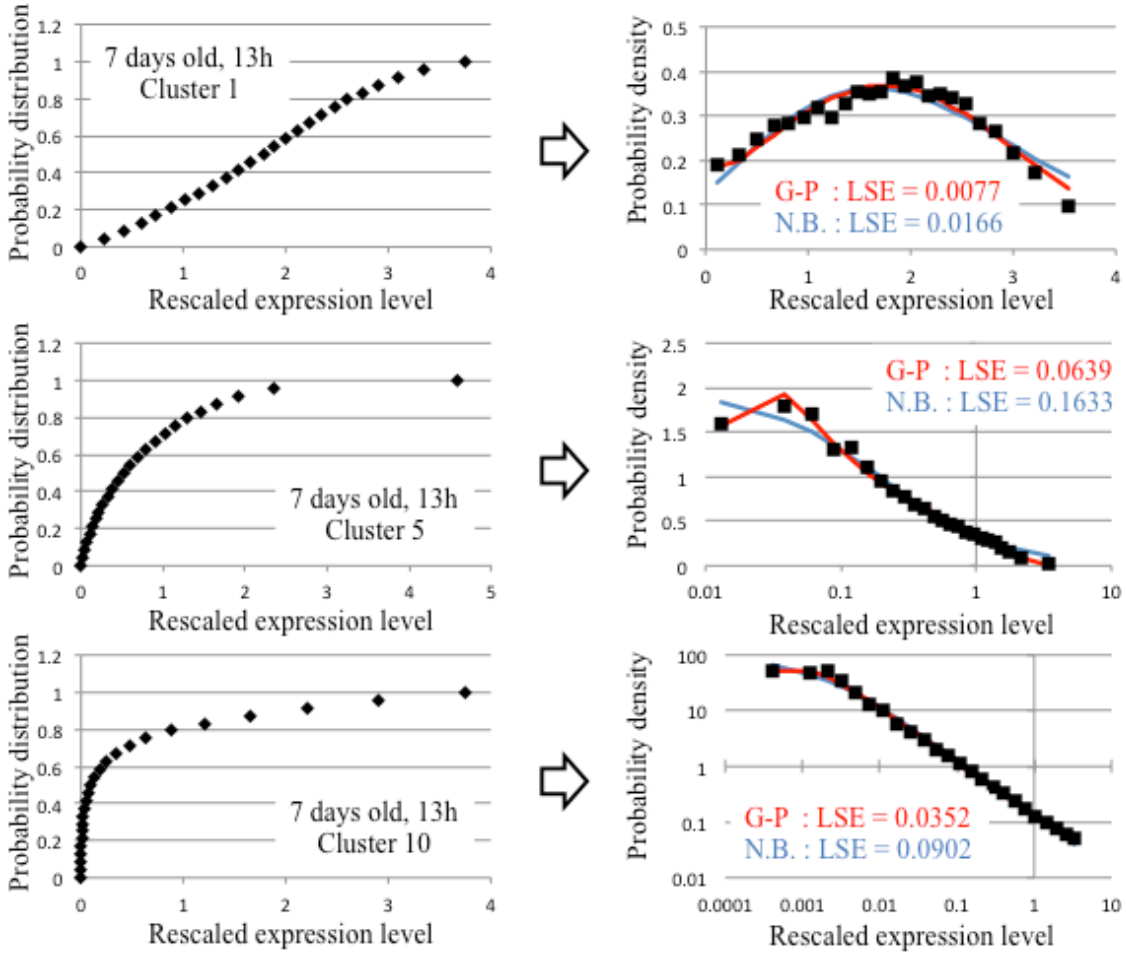

Figure 2. Examples of probability distributions and PDL profiles for indicated clusters. Examples of profiles of probability distribution of rescaled expression levels (left) and of PDL profiles (right) for three clusters. Red and blue represent curves fitted with the G-P and

NB distribution functions, respectively. Least square error was estimated for the fitting curves.

2.4 G-P mixing distribution

The distribution profiles of gene expression levels were systematically classified based on the following mathematical model. A novel distribution function, which we refer to as G-P distribution, is described by equation 1.

$$P(x) = A\frac{K+x}{fx}x^{\frac{2gK-K^2}{f^2}}e^{-\frac{1}{f^2}\left[\frac{gK^2}{x}+(2K-g)x+\frac{x^2}{2}\right]} \qquad (1)$$

This equation is a fitting function of the distribution function of expression level x of the gene of interest X; the parameter A is a normalized coefficient; and f, g, and K are constants whose physiological significance is described below.

In general, gene expression levels are increased by activation and decreased by inhibition of upstream genes in a gene regulatory network. Furthermore, temporal changes in gene expression levels are directly or indirectly influenced by genes themselves, since the gene regulatory network includes many positive and negative feedback loops that are activated in a stochastic manner based on fluctuations in gene expression. Thus, a simplified model of the temporal change in the expression level x of gene X influenced by upstream genes and feedback regulation (Fig. 3) is given by equation 2:

$$\frac{dx}{dt} = G + R(t) + \frac{Fx}{K+x}\eta(t) - Cx \qquad (2)$$

where $R(t)$ and $\eta(t)$ are assumed to be Gaussian white noise with $\langle R(t)\rangle = \langle \eta(t)\rangle = 0$, $\langle R(t)R(t')\rangle = 2D\delta(t-t')$, and $\langle \eta(t)\eta(t')\rangle = 2\delta(t-t')$; the parameters G, K, F, and C are ~[average activation rate of X by upstream genes], ~[average expression level of X required to induce maximum expression of downstream genes], ~[magnitude of feedback effects], and [degradation rate of X], respectively; and P(x) is the steady-state probability distribution for simplified cases where $D \to 0$ where $g = G/C$ and $f = F/C$ (see supplementary information S2).

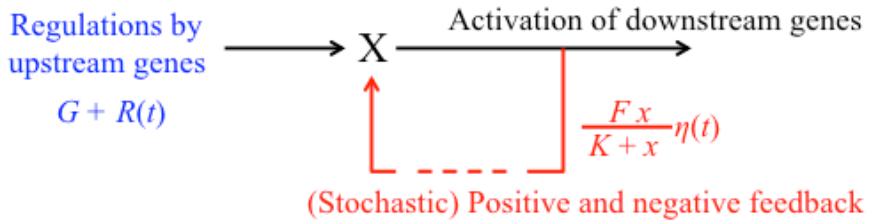

Figure 3. Illustration of a gene network model that fits a G-P distribution function. Gene *X* is regulated by upstream genes and by stochastic feedback.

2.5 Fitting of PDL profiles with the G-P mixing distribution

PDL profiles of each cluster under each condition were fitted with the G-P mixing distribution and (generalized) NB distribution according to equation 3:

$$N(x) = B\frac{\Gamma(sx+k+1)}{\Gamma(k)\Gamma(sx+1)}Q^k(1-Q)^{sx} \qquad (3)$$

where $\Gamma(r)$ is the gamma function and B, s, k, and Q are fitting parameters. Note that the parameter s—which is usually equal to 1—contributes to the generalization for various scales of x.

The characteristics of PDL profiles for some clusters can be extracted from plots with a linear scale axis; however, it is more difficult to extract those of profiles with much larger maximum values and exhibit power law-like profiles, for which log-log plots seem more suitable when the maximum PDL value is greater than 3. In order to extract their detailed characteristics, PDL profiles were fitted using a typical least squares method for maximum PDL values < 3; PDL fitting parameters were chosen so as to minimize the sum of squared errors between log[PDL] and log[fitting functions] when the maximum PDL value was > 3. The results suggest that the G-P distribution has a least square error that is equal to or smaller than that of the NB distribution for PDL profiles of most clusters (Fig. 2 and Table S3). Therefore, in subsequent analyses the PDL profiles were classified according to a G-P distribution.

2.6 Classification of PDL profiles

When PDL profiles of each cluster were fitted to the G-P distribution function, some had $K = 0$, indicating that they were Gaussian, while others had $K >> g$, indicating that they were closer to a power law distribution. PDL profiles were classified as one of three types: Gaussian ($K = 0$), power law-like ($K >> g$), or mixed ($K \approx g$) (Table S3). Here, $f >> g$ was

also obtained when $K >> g$, indicating that when the influence of feedback effects are large relative to the other mechanisms regulating gene expression, gene expression levels exhibit a long-tailed power law-like distribution.

Even for the same gene, PDL profiles varied depending on plant age and harvest time (Table S2). The ratio of occurrence of Gaussian, mixed, and power law-like distributions at four time points in younger plants (7 days old) was ~ 3:3:4, while that of older plants (22 days old) was ~45:29:26. High average expression levels were more frequently associated with a Gaussian as compared to a power law-like distribution; average expression levels and peak value of the frequency distribution were higher for the former than for the latter (Fig. 4).

Gene function was also correlated with PDL profiles (Tables 2 and S4). For example, More than half of "essential genes" (Meinke et al., 2008)showed Gaussian PDL distribution at four time points of young plants and old plants. Furthermore, genes encoding important intracellular components and organelles and those associated with electron transport in metabolic pathways tended to show Gaussian PDL distributions. On the other hand, genes encoding transcription factors and nucleic acid-binding proteins mostly exhibited power law-like and mixed PDL distributions.

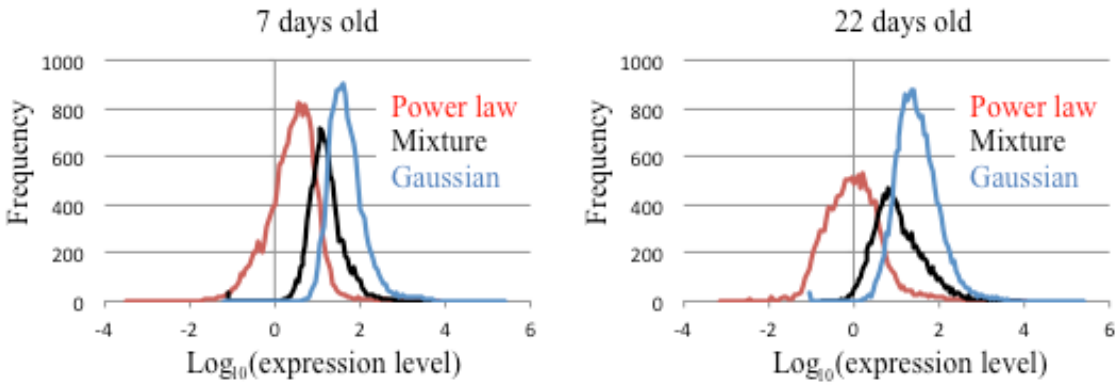

Figure 4. Frequency distributions of average log gene expression levels of gene groups exhibiting distinct PDL profiles. Frequency distributions of average log gene expression levels in cases of Gaussian (blue), mixed (black), and power law-like (red) distributions in 7-day-old (left) and 22-day-old (right) *Arabidopsis*. Differences in average log gene expression levels between Gaussian and mixed, and between mixed and power law-like distributions were significant ($P < 0.01$, t test) at both plant ages.

Table 2. Relationships between gene groups classified according to function and ratio of occurrences of PDL profiles

	7 days old			22 days old		
	Power	Mixture	Gauss	Power	Mixture	Gauss
All genes	0.3933771	0.3024152	0.3042077	0.2619016	0.2909582	0.4471401
Essential genes	0.1693291	0.2875399	0.543131	0.0659755	0.2535344	0.6804901
GO annotated genes	0.2790398	0.3198598	0.4011004	0.178941	0.281477	0.5395821
(GO slim)						
mitochondria	0.4243039	0.2896921	0.2860041	0.2368027	0.3154038	0.4477935
extracellular	0.4165517	0.2796552	0.3037931	0.331464	0.2795159	0.3890201
plastid	0.1217661	0.2738875	0.6043463	0.0647375	0.2341885	0.701074
cytosol	0.1690476	0.3201058	0.5108466	0.0983607	0.2673578	0.6342816
ribosome	0.1813602	0.2934509	0.5251889	0.0989331	0.2735209	0.6275461
transcription factor activity	0.444956	0.2791963	0.2758476	0.307574	0.2956555	0.3967705
nucleic acid binding	0.416722	0.3045786	0.2786994	0.3371711	0.2560307	0.4067982
other molecular functions	0.4045853	0.2683749	0.3270398	0.2707215	0.3035182	0.4257603
structural molecule activity	0.1688009	0.2991851	0.532014	0.08829	0.2732342	0.6384758
electron transport or energy pathways	0.1330049	0.2594417	0.6075534	0.0806452	0.233871	0.6854839

3 Discussion

More than 20 replicates of *A. thaliana* gene expression data at four harvest times of 7- and 22-day-old shoots were obtained and the PDL profiles of each gene were analyzed. Most profiles could be fitted by the G-P distribution. There were three typical PDL profiles; namely, Gaussian, power law-like, and mixed.

The G-P distribution suggested that the various types of PDL profile were highly influenced by network topology, particularly a feedback loop regulating gene expression; for instance, gene groups showing a power law-like distribution were frequently influenced by a feedback mechanism, while this was rare for those exhibiting a Gaussian distribution. Furthermore, the PDL profiles of genes were correlated with their average expression levels and functions; gene groups classified as being essential for survival tended to exhibit Gaussian distributions, whereas those encoding transcription factors and nucleic acid-binding proteins mostly followed a non-Gaussian (i.e., power law-like or mixed) distribution. Furthermore, the expression levels of many genes classified as “unknown” exhibited power law-like distributions (Table S4), suggesting that their expression is predominantly modulated by feedback loops.

PDL profiles of gene expression levels were inferred from publicly available RNA-seq data derived from 48-replicate experiments of *S. cerevisiae* (Gierlin ́ski et al., 2015) in the same manner as in the present study. The G-P as well as the NB distribution function fit the PDL profiles of *S. cerevisiae* (Fig. S9). However, long-tailed power law-like PDL

profiles were not observed, unlike for *Arabidopsis* genes.

Even when the analysis was performed using 24-replicate data randomly selected from the 48-replicate dataset, the results were qualitatively similar to those described above, except that the number of clusters differed (Fig. S10). Although the number of replicates in the present study was smaller than that used in the earlier report, our results reflect the essential properties of the PDL profiles of *Arabidopsis* genes and are expected to apply to a larger number of replicates.

This study mainly focused on the steady-state probability distributions of gene expression levels. However, many *Arabidopsis* genes are regulated by circadian rhythm. Future studies must therefore address the extent to which the present model can be generalized to dynamic situations. Furthermore, intermittent temporal changes were observed in the expression of genes following a power law-like distribution, suggesting a transcriptional burst mechanism (Taniguchi et al., 2010; So et al., 2011; Munsky et al., 2012; Sanchez et al., 2013; Jones et al., 2014; Fujita et al., 2016) for genes whose expression is predominantly influenced by feedback. Such dynamic features of gene expression warrant more detailed analysis.

4 Materials and Methods

4.1 Plant growth conditions and RNA-seq

Seeds of *A. thaliana* (accession Col-0) were sown on Murashige and Skoog medium with 0.5% gellan gum. After incubation for 2 days at 4°C in dark, the seeds were cultivated at 22°C on a 12:12-h light/dark cycle. The whole aerial part of plants 7 or 22 days after germination was collected 1, 7, 13, and 19 h after the start of light period and immediately frozen in liquid nitrogen and stored on −20°C until RNA extraction. Each individual plant was used as a sample for RNA-seq. Total RNA was extracted with the Maxwell 16 LEV Plant RNA kit (Promega, Madison, WI, USA). RNA-seq library preparation was performed as previously described (Nagano *et al.*, 2015); seven lanes of single-end 50-bp sequencing of the library were analyzed using the Hiseq2000 and HiSeq2500 systems (Illumina, San Diego, CA, USA). Sequences were pre-processed, mapped, and quantified according to a previously described pipeline (Kamitani *et al.*, 2016). Fastq files were deposited into the DNA Data Bank of Japan Sequence Read Archive as accession no. DRA005887.

4.2 Analysis of gene expression level variation bias

Owing to technical limitations, there was a time lag of several to 10 min during the harvesting of *Arabidopsis* leaf samples, potentially introducing a bias in the expression levels

of some genes with respect to harvest time. In order to evaluate the variation bias in gene expression levels under each condition, we calculated the average gene expression levels from half of the samples harvested at early time points and half harvested at late time points. Gene expression level was regarded as stationary (unbiased) if the P value in the t test was > 0.2.

4.3 Cluster analysis

k-Means cluster analysis using R software (http://www.r-project.org) was performed for normalized RED profiles. The number of clusters was selected so as to minimize the Bayesian information criterion.

4.4 Data sources for gene classification

To classify each gene, the Gene Ontology Slim classification list was obtained from TAIR (http://www.arabidopsis.org). Data on essential genes were obtained from the SeedGenes Project (http://www.seedgenes.org/GeneList) (Meinke et al., 2008).

Acknowledgments

The authors thank F. Kobayashi for assistance in RNA preparation, and H. Kudoh and S. Takada for helpful discussions. This research was partly supported by the Platform Project for Support in Japan Agency for Medical Research and Development (to A.A.); a Grant-in-Aid for Scientific Research on Innovative Areas “Integrated Analysis of Strategies for Plant Survival and Growth in Response to Global Environmental Changes” from the MEXT of Japan (no. 25119718 to A.A.); a Grant-in-Aid for Scientific Research on Innovative Areas “Initiative for High-Dimensional Data-Driven Science through Deepening of Sparse Modeling” from the MEXT of Japan (no. 26120525 to A.A.); MEXT KAKENHI Grant Number 17K05614 (to A.A.); and Japan Science and Technology Agency Core Research for Evolutional Science and Technology (no. JPMJCR15O2 to A. J. N.); and Japan Society for the Promotion of Science KAKENHI (nos. JP16H06171 and JP16H01473 to A. J. N.).

References

Bajic D., Poyatos J.F, (2012) *Balancing Noise and Plasticity in Eukaryotic Gene Expression*, BMC Genom. **343**, 1–11.

Blake W.J., Kaern M., Cantor C.R, Collins J.J., (2003) *Noise in Eukaryotic Gene Expression*, Nature **422**, 633–637.

Chang H.H., Hemberg M., Barahona M., Ingber D.E., Huang S., (2008) *Transcriptome-Wide Noise Controls Lineage Choice in Mammalian Progenitor Cells*, Nature **453**, 544–547.

Choi J.K., Kim Y.J., (2008) *Epigenetic Regulation and the Variability of Gene Expression*, Nat. Genet. **40**, 141–147.

Choi J.K., Kim Y.J., (2009) *Intrinsic Variability of Gene Expression Encoded in Nucleosome Positioning Sequences*, Nat. Genet. **41**, 498–503.

Elowitz M.B., Levine A.D., Siggia E.D., Swain P.S., (2002) *Stochastic Gene Expression in a Single Cell*, Science **297**, 1183–1186.

Friedman N., Cai L., Xie X.S., (2006) *Linking Stochastic Dynamics to Population Distribution: an Analytical Framework of Gene Expression*, Phys. Rev. Lett. **97**, 168302.

Fujita K., Iwaki M., Yanagida T., (2016) *Transcriptional Bursting is Intrinsically Caused by Interplay Between RNA Polymerases on DNA*, Nat. Comm. **7**, 13788.

Furusawa C., Suzuki S., Kashiwagi A., Yomo T., Kaneko K., (2005) *Ubiquity of Log-Normal Distributions in Intra-Cellular Reaction Dynamic*s, Biophysics **1**, 25–31.

Gierliński M., Cole C., Schofield P., Schurch N.J., Sherstnev A., Singh V., Wrobel N., Gharbi K., Simpson G., Owen-HughesT., et al., (2015) *Statistical Models for RNA-Seq Data Derived from a Two-Condition 48-Replicate Experiment*, Bioinformatics **31**, 3625-3630.

Goda H., Sasaki E., Akiyama K., Maruyama-Nakashita A., Nakabayashi K., Li W., Ogawa M., Yamauchi Y., Preston J., Aoki K., et al., (2008) *The AtGenExpress Hormone and Chemical Treatment Data Set: Experimental Design, Data Evaluation, Model Data Analysis and Data Access*, Plant J. **55**, 526–542.

Golding I., Paulsson J., Zawilski S.M., Cox E.C., (2005) *Real-Time Kinetics of Gene Activity in Individual Bacteria*, Cell, **123**, 1025–1036.

Hirao K., Nagano A.J., Awazu A., (2015) *Noise–Plasticity Correlations of Gene Expression in the Multicellular Organism Arabidopsis Thaliana*, J. Theo. Biol. 387, 13–22.

Jones D.L., Brewster R.C., Phillips R., (2014) *Promoter Architecture Dictates Cell-to-Cell Variability in Gene Expression*, Science **346**, 1533–1536.

Kaern M., Elston T.C., Blake W.J., Collins J.J., (2005) *Stochasticity in Gene Expression: From Theories to Phenotypes*, Nat. Rev. Genet. **6**, 451–464.

Kalmar T., Lim C., Hayward P., Munoz-Desclazo S., Nichols J., Garcia-Ojalvo J., Martinez Arias A., (2009) *Regulated Fluctuations in Nanog Expression Mediate Cell Fate Decisions in Embryonic Stem Cells*, PLoS Biol. **7**, e1000149.

Kamitani M., Nagano A.J., Honjo M.N., Kudoh H., (2016) *RNA-Seq Reveals Virus-Virus and Virus-Plant Interactions in Nature*, FEMS Microbiol. Ecol. **92**, fiw176.

Kaneko K., (2006) *Life: An Introduction to Complex Systems Biology*. Springer, Berlin, Heidelberg.

Karthik D., Stelzer G., Gershanov S., Baranes D., Salmon-Divon M., (2016) *Elucidating Tissue Specific Genes Using the Benford Distribution*, BMC Genomics 17, 595.

Kilian J., Whitehead D., Horak J., Wanke D., Weinl S., Batistic O., D'Angelo C., Bornberg-Bauer E., Kudla J., Harter K., (2007) *The AtGenExpress Global Stress Expression Data Set: Protocols, Evaluation and Model Data Analysis of UV-B Light, Drought and Cold Stress Responses*, Plant J. **50**, 347–363.

Konishi T., Konishi F., Takasaki S., Inoue K., Nakayama K., Konagaya A., (2008) *Coincidence Between Transcriptome Analyses on Different Microarray Platforms Using a Parametric Framework*, PLoS One **3**, e3555.

Landry C.R., Lemos B., Rifkin S.A., Dickinson W.J., Hartl D.L., (2007) *Genetic Properties Influencing the Evolvability of Gene Expression*, Science **317**, 118–121.

Lehner B., (2010) *Conflict Between Noise and Plasticity in Yeast*, PLoS Genet. **6**, e1001185.

Lehner B., Kaneko K., (2011) *Fluctuation and Response in Biology*, Cell. Mol. Life Sci. **68**, 1005–1010.

Less H., Galili G., (2008) *Principal Transcriptional Programs Regulating Plant Amino Acid Metabolism in Response to Abiotic Stresses*, Plant Physiol. **147**, 316–330.

Marioni J.C., Mason C.E., Mane S.M., Stephens M., Gilad Y., (2008) *RNA-Seq: an Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays*, Genome Res. **18**, 1509–1517.

Maruyama-Nakashita A., Nakamura Y., Watanabe-Takahashi A., Inoue E., Yamaya T., Takahashi H., (2005) *Identification of a Novel Cis-Acting Element Conferring Sulfur Deficiency Response in Arabidopsis Roots*, Plant J. **42**, 305–314.

Meinke D., Muralla R., Sweeney C., Dickerman A., (2008) *Identifying Essential Genes in Arabidopsis Thaliana*, Trends Plant Sci. **13**, 483–491.

Mitsui K., Tokuzawa T., Itoh H., Segawa K., Murakami M., Takahashi K., Maruyama M., Maeda M., Yamanaka S., (2003) *The Homeoprotein Nanog is Required for Maintenance of Pluripotency in Mouse Epiblast and ES Cells*, Cell, **113**, 631–642.

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B., (2008) *Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq*, Nat. Methods **5**, 621–628.

Munsky B., Neuert G., van Oudenaarden A., (2012) *Using Gene Expression Noise to Understand Gene Regulation*, Science **336**, 183–187.

Nagalakshmi U., Waern K., Snyder M., (2010) *RNA-Seq: a Method for Comprehensive Transcriptome Analysis*, Curr. Protoc. Mol. Biol. Chapter 4, Unit 4, **11**, 11–13.

Nagano A.J., Honjo N.H., Mihara M., Sato M., Kudoh H., (2015) *Detection of Plant Viruses in Natural Environments by Using RNA-Seq*, Methods Mol. Biol. **1236**, 89–98.

Nagano A.J., Sato Y., Mihara M., Antonio B.A., Motoyama R., Itoh H., Nagamura Y., Izawa T., (2012) *Deciphering and Prediction of Transcriptome Dynamics Under Fluctuating Field Conditions*, Cell **151**, 1358–1369.

Nemhauser J.L., Hong F.X., Chory J., (2006) *Different Plant Hormones Regulate Similar Processes Through Largely Nonoverlapping Transcriptional Responses*, Cell **126**, 467–475.

Newman J.R., Ghaemmaghami S., Ihmels J., Breslow D.K., Noble M., DeRisi J.L., Weissman J.S., (2006) *Single-Cell Proteomic Analysis of S. cerevisiae Reveals the Architecture of Biological Noise*, Nature **441**, 840–846.

Ochiai H., Sugawara T., Sakuma T., Yamamoto T., (2014) *Stochastic Promoter Activation Affects Nanog Expression Variability in Mouse Embryonic Stem Cells*, Sci. Rep. **4**, 7125.

Rapaport F., Khanin R., Liang Y., Pirun M., Krek A., Zumbo P., Mason C.E., Socci N.D., Betel D., (2013) *Comprehensive Evaluation of Differential Gene Expression Analysis Methods for RNA-Seq Data*, Genome Biol. **14**, R95.

Robinson M.D., Smyth G.K., (2008) *Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data*, Biostatistics **9**, 321–332.

Salman H., Brenner N., Tung C., Elyahu N., Stolovicki E., Moore L., Libchaber A., Braun E., (2012) *Universal Protein Fluctuations in Populations of Microorganisms*, Phys. Rev. Lett. **108**, 238105.

Sanchez A., Golding I., (2013) *Genetic Determinants and Cellular Constraints in Noisy Gene Expression*, Science **342**, 1188–1193.

Sato K., Ito Y., Yomo T., Kaneko K., (2003) *On the Relation Between Fluctuation and Response in Biological Systems*, Proc. Natl. Acad. Sci. USA **100**, 14086–14090.

Schurch N.J., Schofield P., Gierliński M, Cole C., Sherstnev A., Singh V., Wrobel N., Gharbi K., Simpson G., Owen-Hughes T., et al., (2016) *How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?*, RNA **22**, 839-851.

Shen X., Pettersson M., Rönnegård L., Carlborg Ö., (2012) *Inheritance Beyond Plain Heritability: Variance-Controlling Genes in Arabidopsis Thaliana*, PLoS Genet. **8**, e1002839.

Silander O.K., Nikolic N., Zaslaver A., Bren A., Kikoin A., Alon U., M. Ackermann, A (2012) *Genome-Wide Analysis of Promoter-Mediated Phenotypic Noise in Escherichia Coli*, PLoS Genet. **8**, e1002443.

Singh G.P, *Coupling Between Noise and Plasticity in E. Coli*, G3 (Bethesda) **3**, 2115 (2013).
Smith G.R., Birtwistle M.R., (2016) *A Mechanistic Beta-Binomial Probability Model for mRNA Sequencing Data*, PLoS One **11**, e0157828.

So L.H., Ghosh A., Zong C., Sepulveda LA., Segev R., Golding I., (2011) *General Properties of Transcriptional Time Series in Escherichia Coli*, Nat. Genet. **43**, 554–560.

Taniguchi Y., Choi P.J., Li G.W., Chen H., Babu M., Hearn J., Emili A., Xie X.S., (2010) *Quantifying E. Coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells*, Science **329**, 533–538.

Tirosh I., Barkai N., (2008) *Two Strategies for Gene Regulation by Promoter Nucleosomes*. Genome Res. **18**, 1084–109.

Wittenberg T.A., Tzin V., Angelovici R., Less H., Galili G., (2012) *Deciphering Energy-Associated Gene Networks Operating in the Response of Arabidopsis Plants to Stress and Nutritional Cues*, Plant J. 70, 954–966.

Woods H.A., (2014) *Mosaic Physiology from Developmental Noise: Within-Organism Physiological Diversity as an Alternative to Phenotypic Plasticity and Phenotypic Flexibility*, J. Exp. Biol. **217**, 35–45.