Robust Global Regulations of Gene Expression in Biological Processes:
A Major Driver of Cell Fate Decision Revealed

Masa Tsuchiya[1,2] and Masaru Tomita[1,2]

[1] Institute for Advanced Biosciences, Keio University,
Tsuruoka, 997-0035, Japan.
[2] Sytems Biology Program, School of Media and Governance,
Keio University, Fujisawa, 252-8520, Japan.

tsuchiya@ttck.keio.ac.jp
mt@sfc.keio.ac.jp

Midori Hashimoto

Graduate School of Frontier Science,
The University of Tokyo,
Kashiwa, Chiba, 277-8564, Japan

Midori_84-10a@nenv.k.u-tokyo.ac.jp

Abstract– One of the fundamental mysteries in biology is to grasp how a cell can choose a specific cellular process in coordinating complex multi-molecular interactions through DNA, RNA, proteins and metabolites. We investigated the existence of guiding principles in gene expression dynamics for three distinct processes: i) the innate immune response of macrophages to lipopolysaccharide (LPS) stimulation, ii) interleukin 12 (IL-12) and IL-4 stimulated naive CD4+ T cell differentiation into helper T1 and T2 cells, and iii) HL-60 cell differentiation into neutrophil. We found that for all processes, forming groups of genes reduces gene expression fluctuations, revealing the emergent hidden collective genome-wide (global) expression dynamics and their biological roles. This demonstrates an important role of lowly expressed genes, which have been considered insignificant and noisy, as key players for collective global dynamics. Moreover, in neutrophil differentiation, there are specific gene ensembles (collectively named "genome vehicle") responsible for the cell fate, and the collective motion of lowly and moderately variable genes within the genome vehicle guides whole genome expression dynamics to the neutrophil cell state. Elucidation of biophysical origin for the global guiding principle might provide a breakthrough understanding of how a genome can control cellular processes such as cell fate and differentiation.

Index Terms - gene expression dynamics, global regulation, innate immune, cell differentiation, biological attractor.

I. INTRODUCTION

A large number of molecules constituting of DNA, RNA, proteins and metabolites in a cell interact with each other to execute numerous inter- and intracellular processes in response to environmental changes or physiochemical perturbations. Forming a large degree of interconnectivity, cells control and self-regulate vital cellular processes such as the immune responses, cell cycle, cell division and differentiation. It is truly intriguing to grasp how a specific path during cellular process such as differentiation can be chosen from vast number of combinatorial possibilities that can arise through the complex multi-molecular interactions. The basis for such determinism may stem from the formation of attractor states in cellular processes [1,2]. The concept of an attractor was developed in a dynamical system, where the whole system evolves dynamically towards an attractor set such as a point, a curve, and a manifold. Metaphorically, C. H. Waddington [3] suggested that a cell fate would be determined toward a local minimum (attractor) on epigenetic energy landscape, where a series of "valleys" and "ridges" describe stable cellular states (local minima) and barriers (local maxima) between those states, respectively. The epigenetic landscape is a proposal for the existence of global molecular regulation in cell fate decision.

To grasp global regulation of gene expressions in complex cellular systems, we investigated three distinct processes, i) the innate immune response of macrophages to LPS, ii) IL-12 and IL-4 stimulated naive CD4+ T cell differentiation into helper T1 and T2 cells, and iii) HL-60 cell differentiation into neutrophil cells. In studies of large-scale high-throughput gene expression such as microarray, estimation of signal intensity is difficult because of unspecific binding affinity between probes and targets mRNAs, and especially in the low level expression changes, the effect of noises, compared with specific binding activity,

is likely larger than that for highly variable genes (more than 2-fold change in expression values from t_0).

To overcome the difficulties of dealing with single gene expression noises, we grouped genes of wildtype whole genome into k ensembles, x_i^w ($i = 1,2,\ldots k$), where x_i^w is i^{th} group in wildtype genome, $\bar{x}_i^w$, its average expression value, integer k for N/n, and n is the number of genes in a group; for LPS innate responses of macrophages and naive CD4+ T cell differentiation, groups are made according to expression changes between initial and different times, whereas for HL-60 cell differentiation having enough time-points, grouping is based on expression variation, where average value of each gene expression in time is subtracted from its expression, and genes are sorted by standard deviation in time. We found that average expression value, $\bar{y}_i^d$, for the corresponding group at different time point or different genotype follows an asymptotic curve,

$$F\left(\bar{x}_i^w, \bar{y}_i^d\right) = \text{constant, i.e.,}\ \bar{y}_i^d = f(\bar{x}_i^w), \qquad (1)$$

as k is increased, showing gene expressions of a group fluctuating around the asymptotic curve. This is an evidence of global regulation of gene expressions; otherwise the points, $\{(\bar{x}_i^w, \bar{y}_i^d)\}$, were scattered over the space.

Notably, finding (1) uncovers 1) for one hour, even single knock out of a signaling molecule, myeloid differentiation primary response gene (88) (MyD88), in LPS simulated microphages can switch thousands of up-regulated gene expressions in wildtype to down-regulation, whereas TIR-domain-containing adapter-inducing interferon-β (TRIF) KO down-regulates thousands of gene expressions slightly from the wildtype response [4], 2) for T-cell differentiation, global switching behavior of gene expression between help T1 and T2 differentiations occurs, and 3) for HL-60 cell differentiation, asymptotic gene expression dynamics exhibits oscillating behaviors during differentiation process [5]. Furthermore, elucidating a role of global regulation in neutrophil differentiation, we demonstrated the existence of specific gene ensembles (collectively named "genome vehicle (GV)") responsible for the cell fate. The collective motion of lowly and moderately variable genes within the GV is to guide cell fate of the whole genome from HL60 to neutrophil cells [5].

These results revealed emergent hidden global regulations in genome-wide gene expression, driving orchestrated diverse regulations of biological processes [4-6]. This stems from cooperative effect of genome-wide gene expressions at cell population level. These global regulations of gene expressions do not change genetic code, and hence, the existence of guiding principles in DNA structural changes [7,8] associated with possible genome-wide epigenetic modifications might be suggested.

II. GLOBAL LPS SIMULATED INNATE IMMUNE RESPONSES

To investigate gene expression dynamics in LPS simulated innate immune response on macrophages, we evaluated the correlation structure of Affymetrix mouse expression data for wildtype, MyD88 knockout (KO), TRIF KO, Double KO (DKO) at 0, 1 and 4 hours. Upon LPS binding, TLR4 triggers two major intracellular pathways, the MyD88-dependent and TRIF-dependent pathways to generate innate immune response; Pearson auto-correlation, $r(x,y)$, i.e., angle between two n-dimensional expression vectors, x and y, is calculated:

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}X_i Y_i}{\sqrt{\sum_{i=1}^{n}X_i^2 \sum_{i=1}^{n}Y_i^2}} \qquad (2)$$

$$= \frac{X \cdot Y}{|X||Y|} = \cos\theta,$$

where $X = (x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x})$, $Y = (y_1 - \bar{y}, y_2 - \bar{y}, \ldots, y_n - \bar{y})$, $\bar{x}$ and $\bar{y}$ are mean values of expressions, θ is angle between two expression vectors, and n is 22690. The vector x represents whole genome gene expression of a genotype (wildtype, KOs) at $t = 0$ and y for other time point.

Using (2), our result showed that temporal auto-Pearson correlation on DKO whole genome still has progressive response (Figure 1A). This suggests DKO still possesses LPS response contrary to the current belief, where it was believed that no immune response would occur in case of the double knockout of MyD88 and TRIF molecules [9].
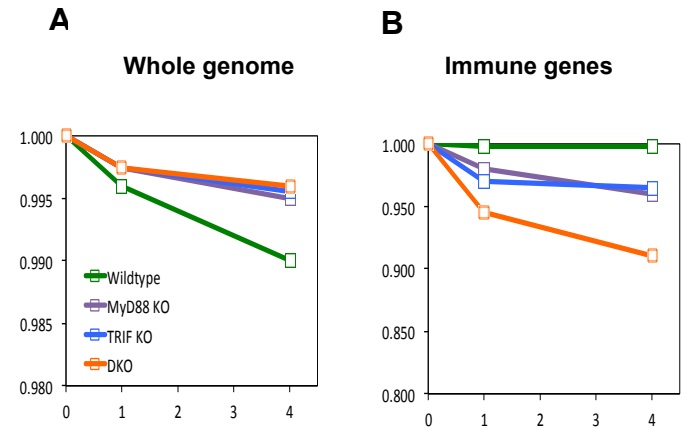

Figure 1. Temporal Pearson auto-correlation: A) The whole genome auto-correlation of all genotypes, comparing with their initial whole genome gene expressions (*t*=0), shows progressive response, correspondent to a progressive displacement of correlation from unity, to LPS stimulation (unit in x-axis is hour; for y-axis, Pearson correlation value). B) DKO has a flat profile as expected, almost perfect correlation between different times is observed due to the lack of any classical immune response to LPS.

Additionally, we confirmed that the DKO auto-correlation response on 157 immune genes has a flat profile (Figure 1B), almost perfect correlation between different times, i.e., the lack of any classical (*local*) immune response to LPS (i.e., two vectors in different time has the same direction, and therefore no response). Thus, as a result, the *global* response on DKO is different from the local innate immune response (see also Figure 2C) [4].
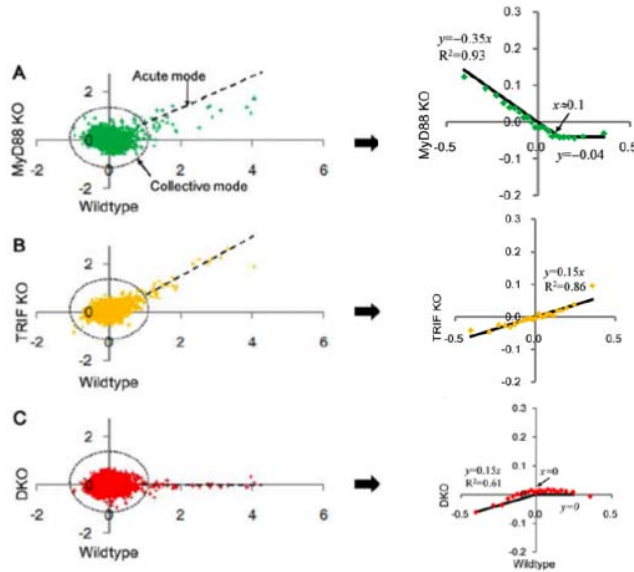

Figure 2. Emergence of asymptotic whole genome collective behaviors in LPS stimulated immune response. Large scatter in collective mode and linear distribution in local mode. Genome-wide single ORFs (left panels) expression changes for 0–1h between genotypes: wildtype (x-axis) vs. A) MyD88KO, B) TRIF KO, C) DKO. Right panels: corresponding plots for average value of group with 1000 genes (ORFs) in the wildtype genome sorted by their expression changes for 0-1h (*x*-axis) vs. the corresponding average value for other genotypes (*y*-axis).

III. EMERGENT GENOME-WIDE CONTROL IN LPS STIMULATED MACROPHAGES

To understand the temporal progress of biphasic (local and global) responses, we investigated the changes of whole genome expression for 0-1h in each genotype. We plotted the expression change (Δx) of single open reading frame (ORF, i.e, gene), where wildtype is for x-axis and other genome types for y-axis (Figure 2, left panel). Scattered areas around zero stem from single gene expression noises. These gene expressions are considered noisy and insignificant for statistical analysis on immune response, since their biophysical origins have yet to be elucidated. To overcome the difficulties of dealing with single gene expression noises,

we grouped genes of wildtype whole genome sorted by from highest to lowest expression changes. Plotting average value of expression for each group against the corresponding average value of other genotype, remarkably, we observed the transition from a large scatter in expression distributions around the origin for single ORFs to smooth linear curves for group of 50 (ORFs) onwards for all genotypes; grouping of expression distribution for 50 ORFs or above forms Gaussian like distribution [4] and the average value of each group follows linear lines (refer to (1)). Interestingly, majority of genes in emergent patterns are lowly expressed genes, indicating these gene expressions arising from random and independent process, Brownian dynamics.

These linear curves reveal the emergence of regulatory signature working at the level of groups of genes. Furthermore, the fluctuations from these asymptotic lines reduce as the grouping size is increased. Notably, for DKO, the result reinforced that DKO possesses genome-wide LPS response through MyD88- and TRIF- independent manner. These emergent linear asymptotic curves can be linearly (see (1)) superposed to decipher the global gene regulatory differential control principle of the transcriptional and mRNA decay machineries between the wildtype and mutant genomes (Figure 3, see details in [4]). These works also have shed new light on innate immune response, providing a significant role for the lowly expressed genes in the diverse collective mode, which are often considered noisy and insignificant in microarray experiments.
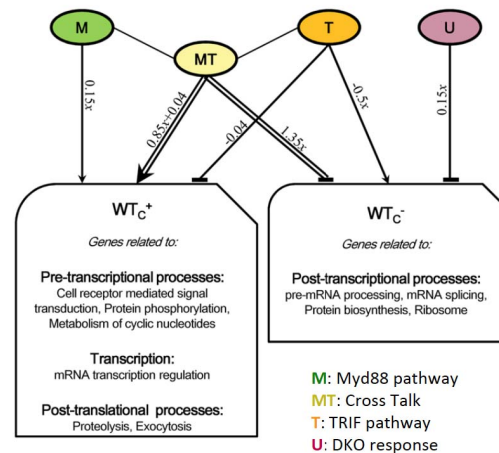

Figure 3. Deciphering gene regulatory mechanisms from the emergent signature. Individual effects of M, T, MT and U on the biological processes regulated in collective mode of WTc+(wildtype upregulated ORFs) and WTc-(wildtype downregulated ORFs). *x* represents average value of expression change of a group (see Fig. 2B) for 0-1h in wildtype LPS response. Arrows indicate activation through dominant transcription and T-shaped lines indicate repression through dominant mRNA decay.

IV. THE DIFFERENTIATION PROCESS OF NAÏVE CD4+ T CELLS

In innate immunity, dendritic cells and macrophages recognize pathogens and present a peptide fragment of pathogen, called antigen, to naïve CD4+T cells differentiating into specific types of helper T cells and trigger acquired immunity. Helper T cells produce certain cytokines such as interferon-gamma (IFN-gamma) and interleukin-4 (IL-4), and activate other immunocompetent cells such as killer T cell and B cell. Interleukin-12 (IL-12) and IL-4 respectively induce differentiation of naïve CD4+ T cells into helper T1 cells (Th1) and helper T2 cells (Th2). Imbalance between Th1 and Th2 responses leads to immune disorders such as autoimmune diseases, allergy and arthritis. Thus, the differentiation process plays an important role as the connector between innate and acquired immunity.

To understand the genome-wide differences between IL-12 and IL-4 stimulated cells, we compare the gene expression change from 0h in IL-12 and IL-4 stimulated cells with control T cells, activating only CD3 at each time point (2, 6 and 48h) (GSE2770 for naïve CD4+ T cells, [10]). To show the difference of the expression change between each stimulated cells and control cells, and plotted the values of single genes and the average values of the groups of 50, 100, 500 and 2000 genes sorted from highest to lowest expression change from 0h at each time points in control cells. Moreover in order to make sure the meaning of the average values, we figured out the distribution of the values in one groups. Plotting gene expressions for single genes showed a scattered plot, however when the average values of groups of 500 genes were plotted, we observed that up-regulated gene groups in IL-12 stimulated cells were down-regulated in IL-4 condition at 0-6h, and that down regulated gene groups in stimulated cells IL-12 were up-regulated in IL-4 condition at 0-48h. This genome-wide switching behavior indicates that the decision of the differentiation process is occurred at whole-genome scale, not only by highly expressed genes.

V. EMERGENCE OF ASYMPTOTIC WHOLE GENOME COLLECTIVE BEHAVIORS IN HL-60 CELL DIFFERENTIATION INTO NEUTROPHIL

To investigate the collective behavior of gene expressions in HL-60 cell differentiation [2], we grouped genes according to their expression variance (gene variation) across time and split the whole genome into p groups, where p is the integer values of N/n for n = 10, 50, 100, 200, 500, 1000 (N = 12625). We plotted the set of mean values of gene expressions for p groups, which show that the groups' mean values transited from scatter to the emergent (temporally oscillating) asymptotic curves at each time point (Figure 4, see details in [5]). Again, the fluctuations from these asymptotic lines reduce following the inverse square root law as the grouping size is increased. This suggests that the temporal genome-wide averaging behavior of collective

expression dynamics exists. We set n = 200 genes since it produced acceptably good resolution.
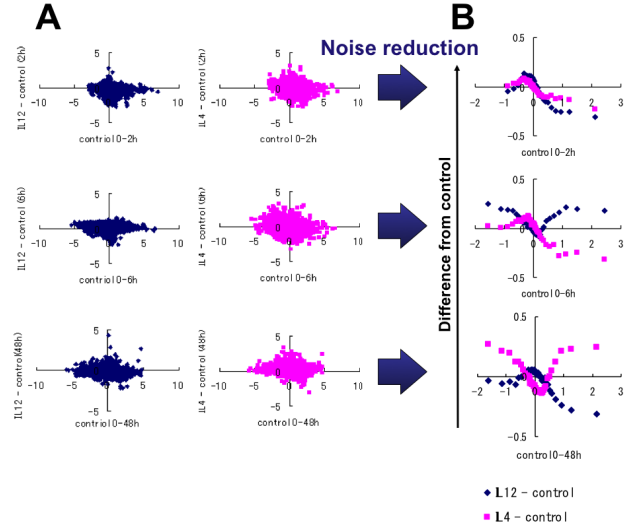

Figure 3: Plot of the average of gene expression change from control. A) Plot of genome-wide expression change between control and, IL-12 (Blue) or IL-4 (Magenta) stimulated cells for y-axis, B) Plot of average values of expression change of groups of 500 genes sorted from highest to lowest expression from 0h to each time point in control cells for x-axis and average values of expression change between control and, IL-12 or IL-4 stimulated cells for y-axis.

VI. THE EXISTENCE OF GENOME VEHICLES AS COLLECTIVE DYNAMICS OF SPECIFIC GENE ENSEMBLES CRUCIAL FOR NEUTROPHIL DIFFERENTIATION

To elucidate further understanding of the significant role of the global response [6,11], we randomly grouped genes from whole genome (N = 12625) into different ensemble sizes (n = 10, 50, 100, 200, 500, 1000) with 100 repeats at each time point (t = 0, 2, 4, 8, 12, 18, 24, 48, 72, 96, 120, 144, 168h) to evaluate the dynamics of correlation distributions of the gene ensembles in the correlation space, (r_v, I), that is defined by Pearson correlation, r_v and mutual information, I of gene variation from the initial time point (see definition of r_v and I in [5]). Plotting average correlation values at each time confirms the formation of Gaussian distribution ([4], due to the central limit theorem). The ensemble size of n = 200, with good resolution, was chosen to evaluate the probability distributions of r_v and I for each time point of the gene expression data.

Utilizing noise reduction by grouping genes, we plotted the probability distributions of r_v and I versus time, and observed that as the ensemble size is increased, the distributions localized to specific points (r_v, I), especially after 48h (Figure 5A). The superposition of the probability distributions (SPD) of r_v and I of atRA and DMSO responses overlap indicating the presence of cell fate attractor, as it

corresponds to the fact the two stimuli elicit the same biological end-point, the generation of a mature neutrophil cell. Regarding the attractor region, we adopted the concept of critical (inflection) points as used in phase transitions in thermodynamic systems to determine the boundary of the neutrophil attractor; unlike continuous dynamics, we are unable to determine the attractor basin for neutrophil differentiation due to the limited temporal data points. Common inflection curves between atRA and DMSO responses, evaluated from the gradients of the SPD for r_v and I, determine the neutrophil attractor (Figure 5A). Tracking the ensembles' trajectories, we noticed that only certain, not all, fall into the attractor in a fractal-like manner [5]. Particularly, the removal of these genome elements from the whole genomes, for both atRA and DMSO responses, no longer have the common region, i.e., destroys the attractor (Figure 5B). This provides evidence for the existence of specific genome elements (named ''genome vehicle'') responsible for the neutrophil attractor. Selective portions of fractal-like gene ensembles are responsible for the neutrophil cell fate decision acting as 'drivers' of reaching a genome-wide characteristic profile.
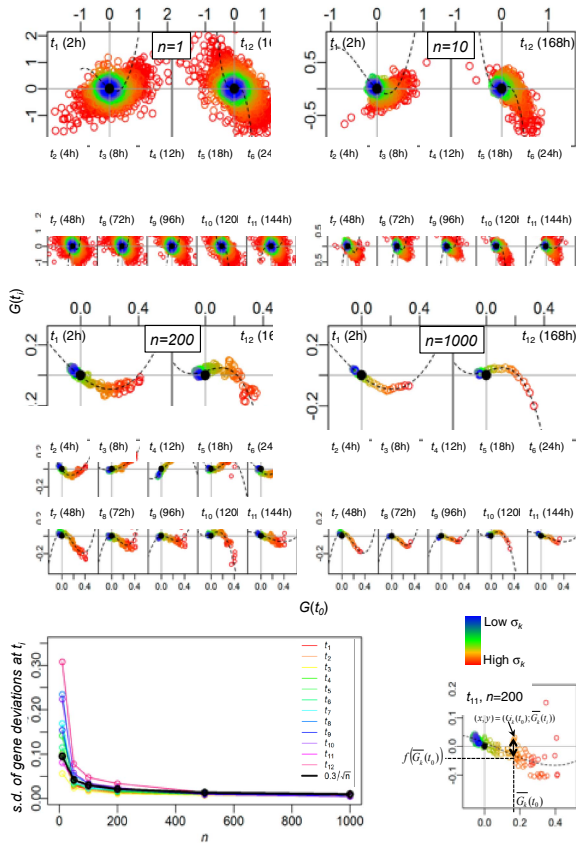
standard deviation of the ranked groups at t_i ($i = 0, 2, 4, 8, 12, 18, 24, 48, 72, 96, 120, 144, 168h$) decreases and mean values of ensembles approach to the asymptotic curve (dashed lines), obeying the inverse square root law (center panel, thick black line) with a transition occurring around $\sqrt{N}$. Note for DMSO response, the similar transition also occurs (not shown, see in [5]).
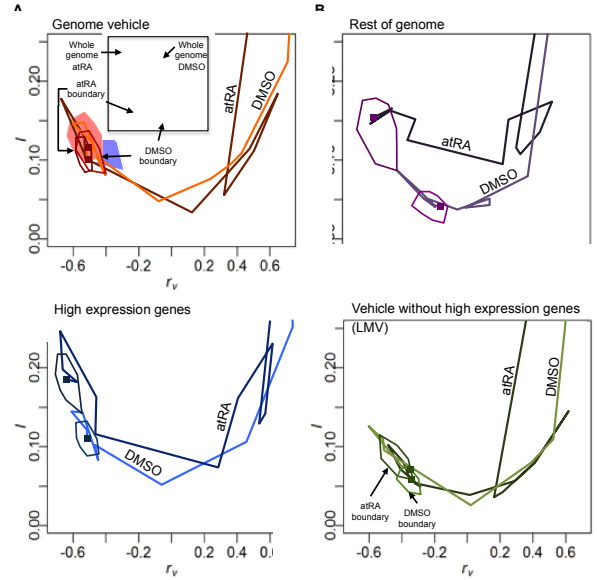

Figure 5. **The loss of the attractor when genome vehicles are removed.** The SPD (superposition of the probability distributions) boundaries (defined by their joining inflection points, see details in [5]) and trajectories for atRA (plain lines with darker tone) and DMSO (lighter tones) responses of (A) genome elements falling into attractor (i.e., genome vehicles, GV) overlap and converge, indicating the formation of neutrophil attractor. (Overlapping SPD boundaries of atRA and DMSO responses of the whole genomes are indicated by red and blue polygons respectively, insert is a zoom of the attractor region, red and blue lines indicate atRA and DMSO SPDs respectively, brow lines indicate GV SPDs), (B) rest of genome elements without genome vehicles do not overlap and converge, (C) high expression genes (2-fold change from t_0 for at least one time point) of the genome vehicles do not overlap and converge, (D) lowly and moderately variable (LMV) genes of the genome vehicles still overlap and converge, retaining the neutrophil attractor. Last time point is represented by a square.

VII. CONCLUSIONS AND DISCUSSIONS

We showed the existence of global regulations of gene expressions in three distinct biological immune processes; innate immune response and immune cell differentiations. Interestingly, we observed that these global regulations are revealed when grouping gene expressions more than 50 genes, especially, in low expressed genes, which are considered noisy and insignificant in microarray experiments. These findings have shed new light on collective behaviors of lowly expressed genes as global regulations. Furthermore, we demonstrated role of the global regulation in the neutrophil differentiation; the

Figure 4. Emergence of asymptotic whole genome collective behaviors in neutrophil differentiation process of HL-60 cells. 12625 genes (N) are grouped according to their variance across time, where G stands for atRA whole genome and the ensemble size is n. The whole genome was sorted from the highest to the lowest standard deviation. As n is increased, the

collective motion of lowly and moderately variable genes within the genome vehicle (GV) guides the cell fate from HL60 to neutrophil cells. The dynamics of gene expression is connected with the dynamics of chromatin structural changes. Therefore, elucidation of the origin of collective behaviors of lowly expressed genes in dynamics of DNA structural transitions [7,8] might provide novel insights for cell fate decisions, especially how well-known master instructive genes, such as Yamanaka factors [12] can drive genomes in differentiation of pluripotent stem cells.

ACKNOWLEDGMENTS

REFERENCES

[1] S.A. Kauffman, *The Origins of Order*, New York: Oxford University Press, 1993.

[2] S. Huang, G. Eichler, Y. Bar-Yam, D.E. Ingber, "Cell fates as high-dimensional attractor states of a complex gene regulatory network", Phys. Rev. Lett., Vol.94, pp128701-5, 2005.

[3] Waddington CH. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*, New York: Macmillan, 1957.

[4] M. Tsuchiya, V. Piras, S. Choi, S. Akira, M. Tomita, A. Giuliani and K. Selvarajoo, "Emergent genome-wide control in wildtype and genetically mutated lipopolysaccarides stimulated macrophages", PLoS ONE , vol.4, e4905, 2009.

[5] M. Tsuchiya, V. Piras, A. Giuliani, M. Tomita and K. Selvarajoo, "Collective dynamics of specific gene ensembles crucial for neutrophil differentiation: the existence of genome vehicles revealed", PLoS ONE, vol.5, e12116, 2010.

[6] M. Tsuchiya, K. Selvarajoo, V. Piras, M. Tomita and A. Giuliani, "Local and Global responses in complex gene regulation networks", Physica A, vol.388, pp. 1738-1746, 2009.

[7] Y. Takenaka, H. Nagahara, H. Kitahata, and K. Yoshikawa, "Large-scale on-off switching of genetic activity mediated by the folding-unfolding transition in a giant DNA molecule: An hypothesis", Phys. Rev. E., vol.77, pp. 031905-10, 2008.

[8] H. Nagahara and K. Yoshikawa, "Large system in a small cell: A hypothetical pathway from a microscopic stochastic process towards robust genetic regulation", Chem. Phys. Lett., vol. 494, pp. 88-94, 2010.

[9] T. Hirotani, M. Yamamoto, Y. Kumagai, et al., "Regulation of lipopolysaccharide-inducible genes by MyD88 and Toll/IL-1 domain containing adaptor inducing IFN-beta", Biochem Biophys Res Commun vol.328, 2005, pp. 383–392.

[10] T. Barrett and R. Edgar, "Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis", Methods in Enzymology, vol.411, pp. 352-369, 2006.

[11] M. Tsuchyia, S. T. Wong, Z. Y. Yeo, et al., "Gene expression waves: cell cycle independent collective dynamics in cultured cells", FEBS J. vol.274, pp. 2878-2886, 2007.

[12] S. Yamanaka and H. M. Blau, "Nuclear reprogramming to a pluripotent state by three approaches", Nature vol.465, pp. 704-710, 2010.