Genetic vehicle comprising majority of lowly expressed genes guide cell fate decision

Masa Tsuchiya[1,2#*], Vincent Piras[1,2#], Masaru Tomita[1,2], Alessandro Giuliani[3] & Kumar Selvarajoo[1,2#*]

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka, 997-0035, Japan.

[2] Systems Biology Program, School of Media and Governance, Keio University, Fujisawa, 252-8520, Japan.

[3] Istituto Superiore di Sanita', Environment and Health Department, Viale Regina Elena 299, 00161, Rome, Italy.

#Equal contributions

*To whom correspondence should be addressed:

Masa Tsuchiya, Email: tsuchiya@ttck.keio.ac.jp, Tel/Fax: +81-235-29-0829

Kumar Selvarajoo, Email: kumar@ttck.keio.ac.jp, Tel/Fax: +81-235-29-0830

Abstract

Cells remarkably take a specific differentiation path among the multiple possibilities that can arise through the multi-dimensional regulation of genome activities. Such deterministic processes suggest the existence of cellular attractors. However, the origins and drivers of the attractors still remain elusive. Here we analyzed the temporal neutrophil differentiation microarray data for two different stimuli, dimethyl sulfoxide (DMSO) and all-trans-retinoic acid (atRA), and expressed their collective dynamics by temporal Pearson correlation and mutual information coordinates. We constructed ensemble of the genes which showed reduction of correlation fluctuations following the inverse square root law. Evaluating their temporal probability density distributions resulted in the emergence of distinct high density localizations from non-localized low density spread distributions, forming attractor cores for both atRA and DMSO. These attractor cores overlapped, pointing to the existence of a neutrophil cell fate attractor. Notably, we found the localizations of correlation distributions for the majority of lowly expressed genes (LEGs) ensembles overlapped with the whole genome attractor cores, while the remaining genomes' localizations did not overlap. Therefore, we postulate the existence of genetic vehicle, made up mainly of LEGs, for the guidance of cell fate.

Keywords: Cell fate decision, attractor, low gene expression, nonlinear correlation, neutrophil

Introduction

Cell fate decision involves reprogramming of a precursor cell into the differentiated cell state. It is intriguing to observe a specific path chosen by the cell to take among the several possibilities that can arise from the regulation and response of multitudes of molecules during differentiation. Understanding how a deterministic process emerges from network of complex molecular interactions is not only the most relevant problem of biology, but also an important cross-disciplinary interest.

Huang *et al.* showed the convergence of human pro-myelocytic leukaemia HL-60 cells' differentiation into neutrophils by the action of two different reagents, DMSO and atRA (Huang *et al.* 2005), by the analysis of 2773 highly expressed genes so as to establish multidimensional "attractor" states in biology (Waddington 1957; Kauffman 1979; Kauffman 1993; Huang *et al.* 2009). However, it is unclear whether only highly expressed genes are relevant for the neutrophil cell fate. Recently, we have shown that ensembles of lowly expressed genes play an important role in yielding global response to lipopolysaccharide (LPS) stimulated macrophages. Basically, by reducing noise of an ensemble of lowly expressed genes following the inverse square root law, we showed local and global effects on genome-wide expression; local effect being the pro-inflammatory response of a small number of highly expressed genes, while global response being the collective activation of diverse processes comprising the rest of the lowly expressed genes

(Tsuchiya *et al.* 2009*a*; Tsuchiya *et al.* 2009*b*).

In this paper, we analyzed the *entire* microarray data of HL-60 cells (Huang *et al.* 2005), using RMA normalization known to produce robust reproducible results for the entire range of expression units (Bolstad *et al.* 2003). To uncover the significance of global response guiding cell fate decision, we used Pearson (linear) correlation and mutual information (nonlinear correlation) metrics and show, for the first time, that the lowly expressed portion of the genome is responsible for guiding cell fate decision. This finding is potentially revolutionary in biology where, until now, relevant mechanisms are dissected and described following only the highly expressed portion of genome.

Evaluating ensembles of gene expression dynamics in Pearson correlation and mutual information space

To begin, we tackled the issue of noise at single-gene level by investigating the ensemble property of the gene population (Tsuchiya *et al.* 2009*a*; Tsuchiya *et al.* 2009*b*). We expressed the system collective dynamics by means of temporal Pearson correlation and mutual information acting as coordinates to trace the *N*-dimensional expression vectors' trajectories for each stimulus (atRA and

DMSO). Pearson correlations $r_v(V(t_i),V(t_0)) = \frac{\boldsymbol{V}(t_i)\cdot\boldsymbol{V}(t_0)}{|\boldsymbol{V}(t_i)||\boldsymbol{V}(t_0)|}$ assess linear relationship between two vectors $\boldsymbol{V}(t_i)$ and $\boldsymbol{V}(t_0)$ in time, while nonlinear dependency is checked by mutual information $I(V(t_i),V(t_0)) = -\sum_{x\in V(t_i)} p_i(x)\ln(p_i(x)) - \sum_{y\in V(t_0)} p_0(y)\ln(p_0(y)) + \sum_{x\in V(t_i),y\in V(t_0)} p(x,y)\ln(p(x,y)) - \varepsilon$ (Steuer *et al*. 2002; Steuer *et al*. 2004), where the joint probability distribution function $p(x,y)$, and marginal probability distribution functions, $p_i(x)$ at t_i and $p_0(y)$ at t_0 are estimated by means of an histogram-based approach by discretizing gene expression into K=10 bins (Steuer *et al*. 2002). Note: Due to the discretization, mutual information I incurs systematic error ε (Steuer *et al*. 2002). Since randomly ordered data should destroy correlations, we expect I to be close to zero, therefore, we calculated the minimum I for 100 random permutations of gene deviation vectors ($\boldsymbol{V}(t_i)$). However, we found a positive value for minimum I instead of zero, and so subtracted this minimum positive value from final I. Furthermore, for comparing I of atRA and DMSO response, we used normalized $\hat{I}(V(t_i),V(t_0)) = \frac{I(V(t_i);V(t_0))}{I(V(t_0);V(t_0))}$ and called $\hat{I}$ as I throughout the text.

Each stimulus's dynamic genome expression activity (partial and whole) is defined by N-dimensional gene deviation-from-average vectors at time t_i (i = 0,1,…,M),

$V(t_i) = (v_1(t_i), v_2(t_i), .., v_j(t_i), .. v_N(t_i))$, where $v_j(t_i) = x_j(t_i) - \bar{x}_j$ is gene deviation of j^{th} gene at t_i, $x_j(t_i)$ is the corresponding gene expression value at t_i, and $\bar{x}_j$ is the average expression of j^{th} gene over M+1 discrete time points. We modified typical Pearson r to be r_v by subtracting their average expression value, $\bar{x}_j$, over M + 1 time points instead of subtracting the mean of whole genome expression from each gene expression at each time point (Tsuchiya *et al.* 2009*b*). This index thus measures the temporal correlation of genome-wide expression deviations from their average values so as to allow discriminating gene expressions with different amplification but similar temporal profiles. In our study, N = 12625 genes/ORFs and M = 12 (where t = 0, 2, 4, 8, 12, 18, 24, 48, 72, 96, 120, 144, 168h). For simplicity we include ORFs as genes.

Whole genome attractors and their cores determined by localization of probability density distribution of correlations

Attractor basins in dynamic non-linear systems can be determined when long time series are available. However, since we only have limited time data, we described attractor core rather than basin. The attractor core is defined as the region where the density of r_v and I of atRA and DMSO localizes.

To investigate the localization of the temporal probability density distributions of the correlation space (r_v vs I), we randomly grouped genes from whole genome into different ensembles (n = 10, 50, 100, 200, 500, 1000) and found distributions of r_v and I for 100 groups transited from scattered/spread to concentrated/focused ones for $n_t \geq 100$, where $n_t \approx \sqrt{N}$ (see figure S1 in supplementary material). This result shows standard deviations of r_v and I distributions at each time point are reduced by $\alpha / \sqrt{n}$ law at increasing n. Using this ensemble of genes, temporal probability density distributions show the emergence of distinct high density localizations (forming attractor core for $n > 50$) from non-localized low density spread distribution (for $n < 50$) , especially after 48hrs (figure 1*a* and 1*b*). Thus, we chose $n_t = 200$ for plotting the probability density distributions, with good resolution, of r_v and I for each time point of the gene expression data, $\{\boldsymbol{V}(t_0),..,\boldsymbol{V}(t_M)\}$.

Notably, attractor core of atRA and DMSO overlapped with the convergence of the corresponding whole genome trajectories of r_v and I, pointing to the existence of a neutrophil cell fate attractor (figure 1*c* and 1*d*), largely independent of the stimulus. We verified that the standard deviation of the superposition of r_v and I distributions scales with $\alpha / \sqrt{n}$ (see figure S2 in supplementary material). Note that for the other localization of r_v and I, it coincided with the whole

genome trajectory loops indicating intermediary cell differentiation states (Chang *et al.* 2006) (figure 1*d*).

The role of groups of lowly expressed genes in the development of the cell fate attractor cores shown by the reduction of correlation fluctuations

We grouped genes according to temporal gene deviations to investigate the role of lowly expressed genes. Standard deviation (σ_j) of temporal expressions of j^{th} gene, $\{x_j(t_0),..,x_j(t_M)\}$, were evaluated and sorted from the whole genome vector $\boldsymbol{V}(t_0)$ from highest to lowest, where the sorted whole genome vector at t_0 is represented by $\boldsymbol{S}(t_0) = (s_1(t_0), s_2(t_0),.., s_j(t_0),..s_N(t_0))$. For subsequent time points (i = 1,..,12), we retained the sorting order of genes as of t_0. Note: We used σ instead of coefficient of variation ($CV = \sigma/\mu$) as we are dealing with the dynamical motions of genes, rather than normalized motions as often used in conventional approaches. Nevertheless, we compared *CV*s with σ and found a linear relation between them (figure S3 in supplementary material), ruling out any possible trivial scale effect as explanation of our results.

We sequentially split $\boldsymbol{S}(t_0)$ into smaller groups of n genes (n = 1, 10, 50, 100, 200, 500,

1000). For the k^{th} group vector at t_0, containing n genes, we defined $G_k(t_0) = \{g_1(t_0),.., g_j(t_0),.., g_n(t_0)\}$, where $g_j(t_0) = s_{(k-1)n+j}(t_0)$, with group's average standard deviation $\sigma_k = 1/n \sum_{j=1}^{n} \sigma_{g_j(t_0)}$. Then, we monitored the temporal evolution relative to each group $G_k(t_0)$ by plotting the average values of $G_k(t_0)$ vs $G_k(t_i)$ (i = 1,..,12) (figure 2*a* and 2*b*). The distribution of the average values in the plot transited from scatter to asymptotic emergent curves at $n_t \cong \sqrt{N}$ with the reduction of fluctuation in terms of distance from curves obeying the $\alpha / \sqrt{n}$ law (figure 2*c* and 2*d*). Thus, we chose n_t = 200 genes, with acceptably good resolution, as the basic genome element to follow the cell population trajectories toward the attractor. For whole genome we, therefore, obtained 63 genome elements of 200 genes, totaling 12600 genes, the remaining 25 genes with very low σ were discarded.

It is worth noting that the distribution of the number of time points of these basic genome elements $G_k(t_0)$ reaching the attractor core is non-continuous in terms of their σ values (figure 3*a*). Evaluating only these genome elements entering the attractor cores is not sufficient since

genome elements reaching or not reaching attractor cores are probabilistic. This is seen when a genome element not entering the core can potentially change its destiny, if a single gene shift from it is made sequentially according to σ values, resulting in a maximum of 199 additional genome elements within which several genome elements enter the attractor core (see $G_7(t_0)$ for atRA and $G_{13}(t_0)$ for DMSO of figure 3*a*). That is, even though a genome element does not enter the core, it may still possess probability to enter the attractor core. Thus, to improve this, we removed genome elements that enter the attractor core (see green bars in figure 3) and shifted the remaining genomes by 50 genes from the highest σ values to create new genome elements (see figure S4 in supplementary material). We repeated this procedure until no green set is found.

Majority of genome elements comprising of lowly expressed genes entering attractor cores

For atRA, the distribution of time points reaching attractor core shows distinctively genome elements comprising mainly of lowly expressed genes (LEGs) with $0.15 < \sigma < 0.49$, that enter attractor core (green sets), while the remaining (brown sets) do not enter (figure 3*a*, lower panels). Next, we evaluated the trajectories of all 63 genome elements and compared each with the whole genome trajectory, in terms of the Euclidean distance of r_v and I. We observed most genome

elements that enter the attractor core (green sets with $0.27 < \sigma < 0.37$) are close to minimum distance with whole genome trajectory (figure 3*b*), indicating this set possesses a scalable whole genome response. For DMSO response, we observed green set possessing larger distributions including higher σ, i.e., $0.14 < \sigma < 0.83$. However, the green sets predominantly consisting of LEGs possess minimum Euclidean distance (figure 3*b*).

The distribution difference of genome elements reaching attractor core between atRA and DMSO could be due to the mechanism specificity of the response to the stimulus. Notwithstanding the independence of the biological end-point (neutrophil differentiation) from the different stimuli, DMSO is known to exert its biological action by the activation of key transcription factors such as NF- B (Lee *et al.* 2005), thereby inducing several highly expressed (high σ) NF- B - dependent genes, whereas atRA penetrates the nucleus and directly remodels chromatin structure (Klein *et al.* 2000). The fact that atRA, unlike DMSO, does not induce any known master gene with high expression levels leading into the attractor core (figure 3*a*), perhaps indicates the non-instructive signaling for the induction of LEGs related to small-amplitude DNA motions.

To further investigate the role played by the green sets, we assessed the localization of r_v and I distributions for atRA and DMSO responses, and found their localizations overlapped at the neutrophil cell fate attractor core (figure 4*a*, left panels). This corresponds to the fact the two stimuli elicit the same biological end-point, the generation of a mature neutrophil cell,

correspondent to the attractor core. However, the removal of green sets from whole genomes shows the localizations for atRA and DMSO of the remaining portion of the genome do not overlap indicating these (largely highly expressed) genes fail to demonstrate convergence and consequently the reach of the neutrophil attractor state (figure 4*a*, right panels). Moreover, the corresponding trajectories of both atRA and DMSO converged for green set (V), but not for the rest of genome elements (see W-V, figure 4*a*). Thus, these results point to the role of 'green-set' as the "vehicle" collectively driving whole genome trajectory into the attractor core.

Conclusion

In summary, although LEGs are often considered as noisy and insignificant especially in microarray experiments, here we demonstrated the collective motion of genome elements containing predominantly of LEGs mirrors the entire genome trajectory toward the final attractor state thus indicating LEGs as possible candidates for effectively driving cells toward their attractor states. Since the dynamics of gene expression is connected with the dynamics of chromatin structural changes (Komarova & Soffer 2005; Yakushevich 2007), finding the underlying mechanisms for the motion of LEGs might decipher fluctuations in chromatin dynamics that determines cell fate decision. In this regard, we note that the temporal average expression profiles of LEGs show damped oscillation (figure 4*b*). This damped oscillation of LEGs' expressions may

suggest collective dynamics of small-amplitude DNA fluctuations within chromatin structure that diminishes after cell differentiation is completed. It is worth stressing the difference between atRA and DMSO stimulated dynamics: the first one representing a pure 'democratic' process with no evidence of 'dominant players' while the second appears as a 'mixed' type process (Bar-Yam *et al.* 2009) with a local initial intervention of few high expressed genes followed by the collective motion of LEGs. It will be interesting to know how the concerted motion of LEGs, together with well-known master instructive genes, such as Yamanaka factors (Takahashi & Yamanaka 2006), drives the differentiation of pluripotent stem cells as well as other biological processes that could acquire a completely different perspective under the proposed model.

Acknowledgements

We thank Midori Hashimoto and Shohei Sonohara for stimulating discussions. We dedicate this work to our late friends and mother (of KS), Lilly Pillay Selvarajoo, who passed away while this manuscript was in preparation. We appreciate our family and system immunology group members at IAB, Keio University. This study is funded by CREST of Japan, Science and Technology, Tsuruoka City and Yamagata prefecture governments, Global COE Program of Keio University and JASSO.

References

Bar-Yam, Y., Harmon, D. & de Bivort, B. 2009 Attractors and democratic dynamics. *Science* **20**, 1016-1017.

Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. 2003 A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193

Chang, H. H., Oh, P. Y., Ingber, D. E. & Huang, S. 2006 Multistable and multistep dynamics in neutrophil differentiation. *BMC Cell. Biol.* **7**, 11

Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. 2005 Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* **94**, 128701.

Huang, S. 2009. Reprogramming cell fates: reconciling rarity with robustness. *Bioessays* **31**, 546-560

Kauffman, S. A. 1969 Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *J. Theor. Biol.* **22**, 437-467.

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

Klein, E. S., Wang, J. W., Khalifa, B., Gavigan, S. A. & Chandraratna R. A. S. 2000 Recruitment of nuclear corepressor and coactivator to the retinoic acid receptor by retinoid ligands. *J. Biol. Chem.* **275**, 19401-19408.

Komarova, N. & Soffer, A. 2005 Nonlinear waves in double-stranded DNA. *Bull. Math. Biol.* **67**, 701-718.

Lee, Y. R., Shim, H. J., Yu, H. N., Song, E. K., Park, J., Kwon, K. B., Park, J. W., Rho, H. W., Park, B. H., Han, M. K. & Kim, J. S. 2005 Dimethylsulfoxide induces upregulation of tumor suppressor protein PTEN through nuclear factor- B activation in HL-60 cells. *Leuk. Res.* **29**, 401-405.

Steuer, R., Kurths, J., Daub, C. O., Weise, J. & Selbig, J. 2002 The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics* **18**, S231-240.

Steuer, R., Daub, C. O., Selbig, J. & Kurths, J. 2004 In *Proceedings of the 27th annual conference of the Gfkl.* (eds D. Baier & K. D. Wernecke), p.p.81-90. Springer Verlag.

Takahashi, K., and Yamanaka, S. 2006 Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663-676.

Tsuchiya, M., Selvarajoo, K., Piras, V., Tomita, M. & Giuliani, A. 2009*a Physica A* **388**, 1738-1746.

Tsuchiya, M., Piras, V., Choi, S., Akira, S., Tomita, M., Giuliani, A. & Selvarajoo, K. 2009*b* Emergent Genome-Wide Control in Wildtype and Genetically Mutated Lipopolysaccarides-Stimulated Macrophages. *PLoS ONE* **4**, e4905.

Venables, W. N. & Ripley, B. D. 2002 *Modern Applied Statistics with S.* Springer Verlag.

Waddington, C. H. 1957 *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical*

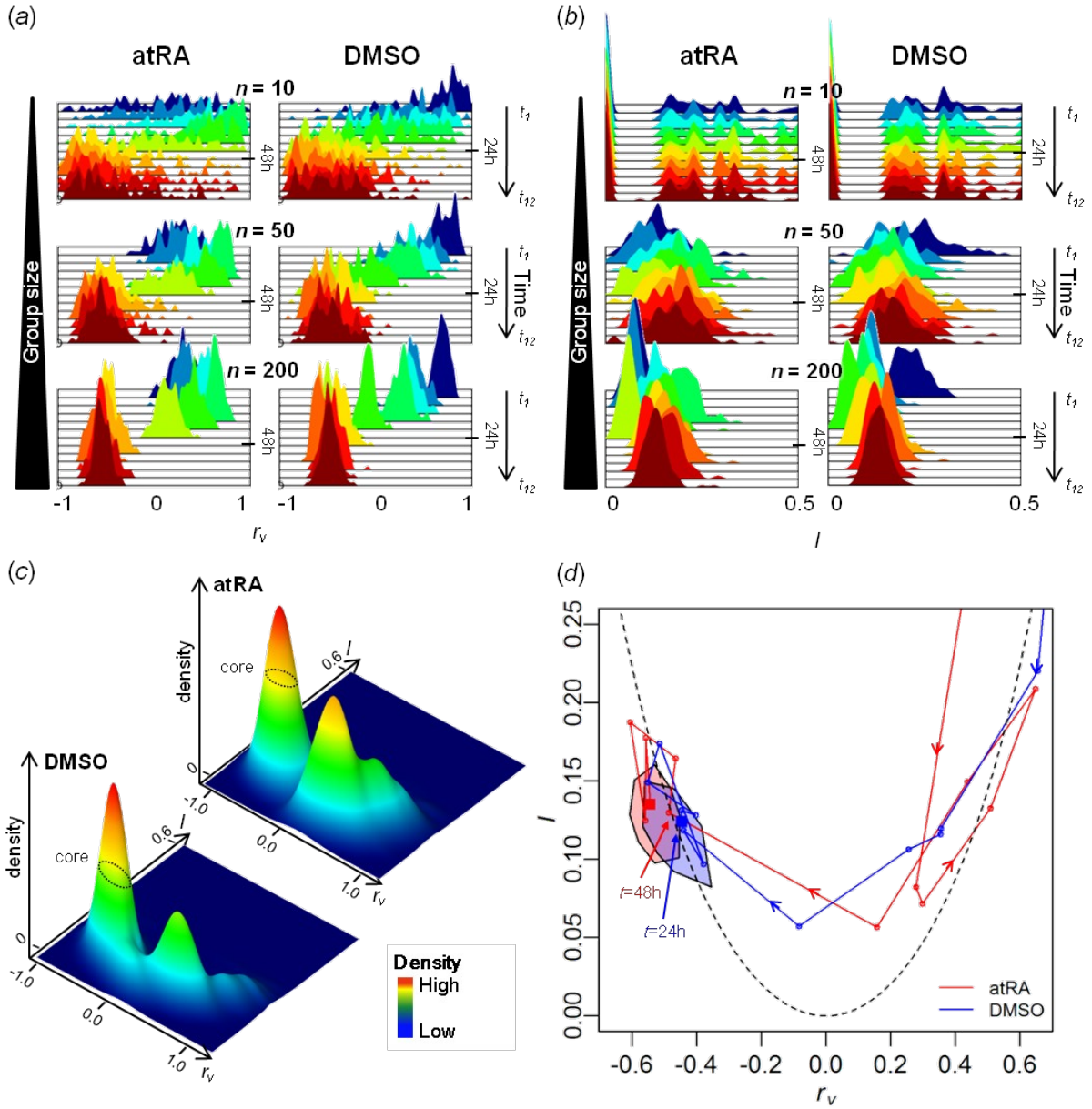
Biology. Macmillan, New York.

Yakushevich, L. V. 2007 Is DNA a nonlinear dynamical system where solitary conformational waves are possible? *J. Biosci.* **26**, 305-313.

Figure Captions

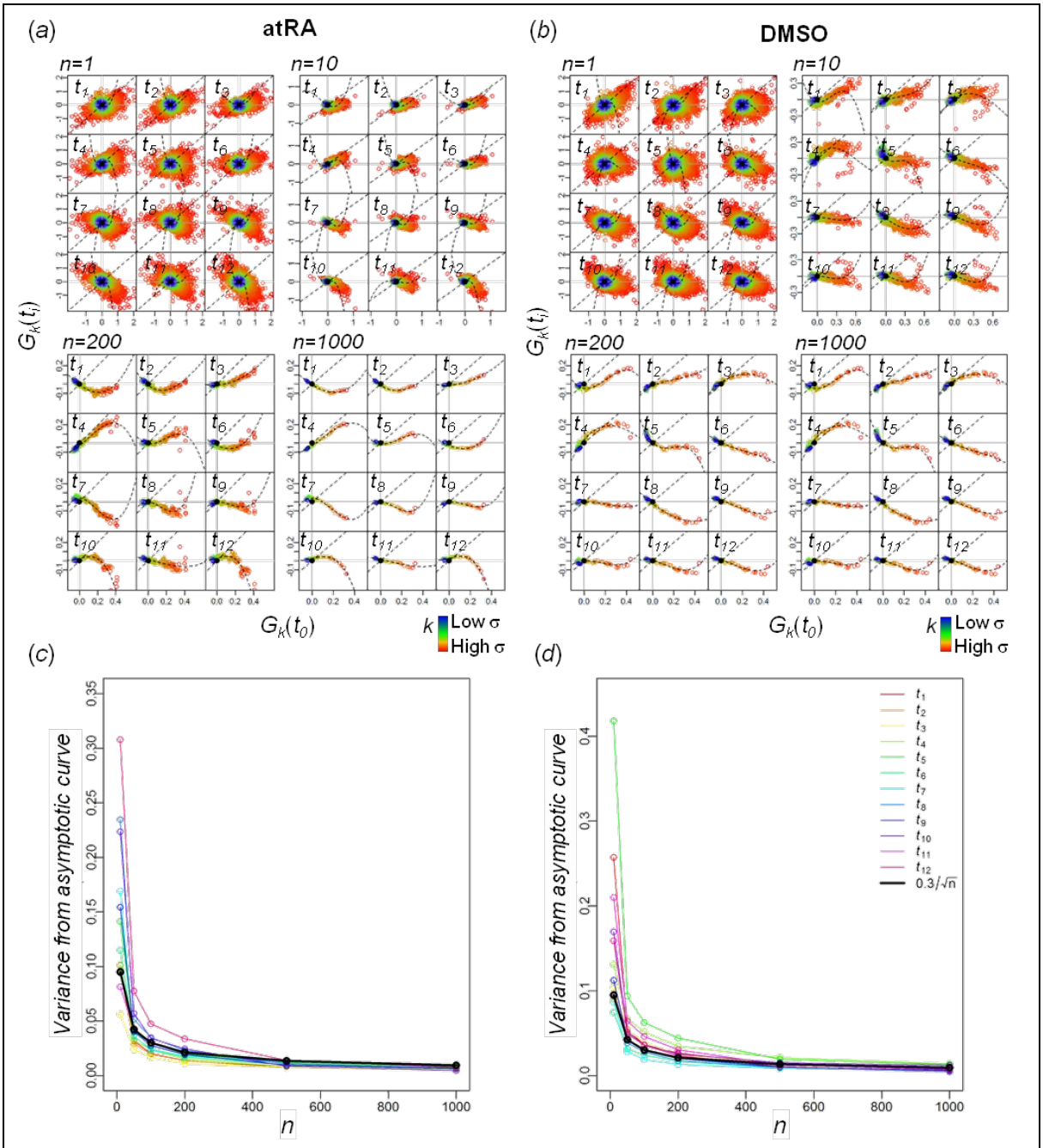
Figure 1. Determination of whole genome attractors and cores for atRA and DMSO response. Temporal probability density distributions of (*a*) r_v and (*b*) I for atRA (left panel) and DMSO (right panel) for n = 10, 50, 200 groups of genes. As n increases the distributions transit from scattered to localized at $r_v \cong 0.55$ for $t \geq 48$h (atRA) and $r_v \cong 0.45$ for $t \geq 24$h (DMSO), where bandwidth of Gaussian kernel is given by 0.02 and 0.01 for r_v and I distributions, respectively. (*c*) 3D plot of the superposition of the probability density distributions of r_v and I over all time points. Attractor core of atRA (top panel) and DMSO (bottom panel), indicated by dotted line, is determined by selecting maximum slope (inflexion curve, i.e. the highest acceleration) of density change d over change of r_v and I, i.e. $\nabla d(r_v, I) = \left(\frac{\partial d}{\partial r_v}, \frac{\partial d}{\partial I} \right)$ (i.e. $\|\nabla d\|$ is maximum). Note that the probability density is discretized on lattice using the MASS R library (two-dimensional kernel

density estimation (Venables *et al.* 2002)) and its change is estimated in 8 adjacent directions (octagon) of each lattice. We repeated this process 30 times to obtain average attractor core. We called average attractor core simply as the attractor core. (*d*) Whole genome trajectories of r_v and I for atRA and DMSO are represented by taking the average of 100 trajectories of 200 (n_t) randomly chosen genes from whole genomes for each time (t = 0, 2, 4, 8, 12, 18, 24, 48, 72, 96, 120, 144, 168h). Filled polygons indicate attractor cores from (b). Dotted curve represents $I = -\log\sqrt{1-r_v^2}$ indicating linear correlation of I estimated by Gaussian distributions.



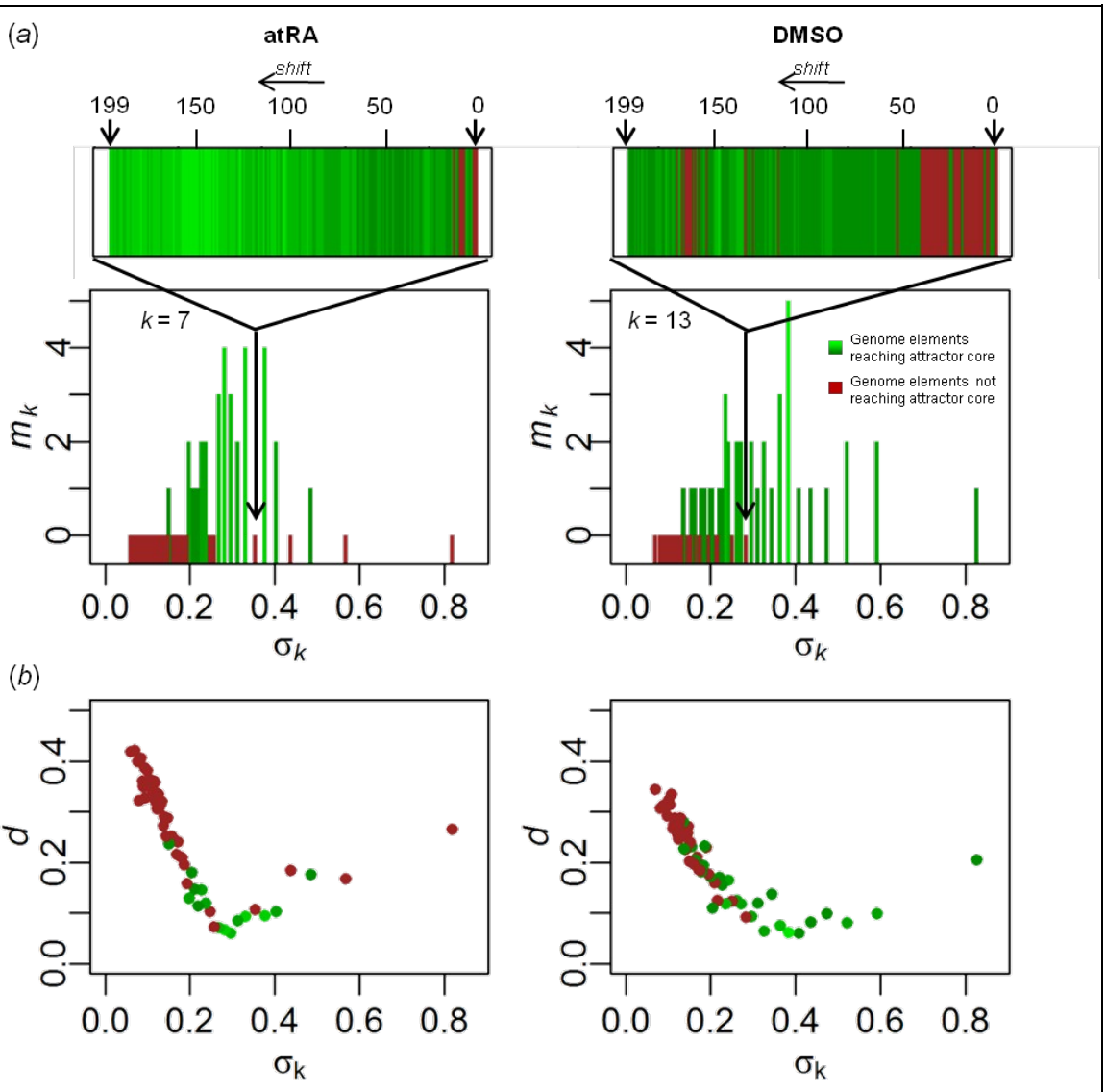
Tsuchiya et al.
Figure 1

Figure 2. Transition from scatter to asymptotic emergent curves when plotting $G_k(t_0)$ vs $G_k(t_i)$. Plot of average values $G_k(t_0)$ vs $G_k(t_i)$ (i = 0,..,12) for the grouping of n = 1, 10, 200, 1000. The emerging asymptotic curve, which is approximated for $n \geq 1000$, is determined by nonlinear least squares fitting with cubic polynomial, (*a*) atRA and (*b*) DMSO. Each time points' variance of the difference in y-axis between the asymptotic curve and $G_k(t_i)$ (colored lines) obeys the $\alpha/\sqrt{n}$ law (thick black line) where $\alpha \cong 0.3$ with transition occurring around $n_t \approx \sqrt{N}$ for (*c*) atRA and (*d*) DMSO.



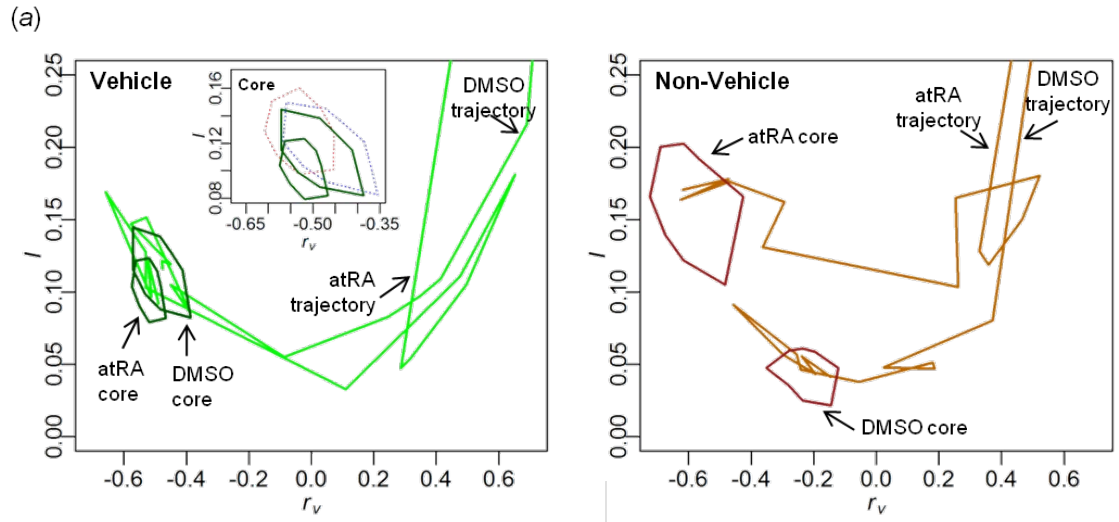
Tsuchiya et al.
Figure 2

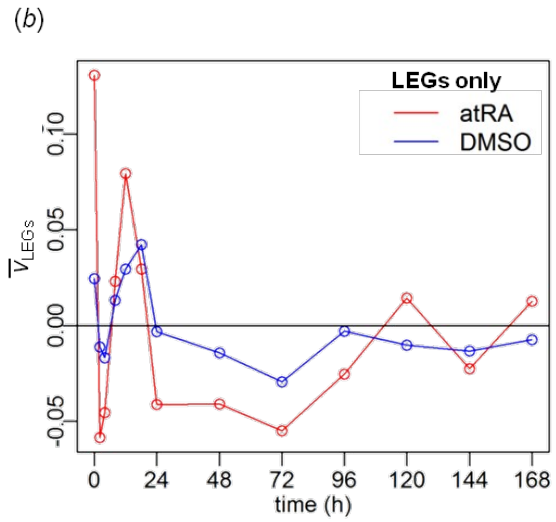
Figure 3. Genetic vehicle, comprising mainly lowly expressed genes, guide neutrophil cell fate. (*a*) Upper panels: 199 additional genome elements generated from k^{th} genome elements (k = 7 (atRA) and k = 13 (DMSO)) by sequential single gene shift according to σ values. Green shows genome elements entering attractor core, while brown shows elements that do not. Lower panels: Distribution of number of time points m_k reaching attractor core for k^{th} genome element, sorted by the corresponding standard deviation σ_k for atRA (right panel) and DMSO responses (left panel) for the first iteration (see figure S4 in supplementary material for further iterations). Elements with at least one time point reaching attractor cores are shown in green (vehicle) and the remaining elements in brown. (*b*) Euclidean distance *d*, between the whole genome's trajectory and each genome element's trajectory. Green and brown indicate genome elements reaching and not reaching attractor core, respectively. The shaded green is according to m_k values, the higher with lighter tone.



Tsuchiya et al.
Figure 3

Figure 4. The role of lowly expressed genes. (*a*) Left panel, overlapping attractor cores of atRA and DMSO occurs for genetic vehicle (V, green sets in maintext, left panel), but does not occur after removal of genetic vehicle from each whole genome (W-V, right panel). Corresponding trajectories of both atRA and DMSO (as indicated by arrows) converges for V, but not for W-V. Insert in left panel shows overlapping attractor cores of atRA and DMSO for the whole genomes (indicated by red and blue dotted polygons respectively). (*b*) Damping oscillations of the average gene deviation of lowly expressed genes (LEGs), $\bar{v}(t_i) = \frac{1}{m} \sum_{j \in LEGs}^{m} v_j(t_i)$, with time ($i$ = 0,..,12), where $v_j(t_i)$ is gene deviation value at t_i, and m is the number of LEGs (i.e. green set for atRA, green set with $\sigma < 0.52$* for DMSO), where m = 4600 (atRA) and 7699 (DMSO). *Maximum σ obtained for atRA green set genes.





Tsuchiya et al.
Figure 4

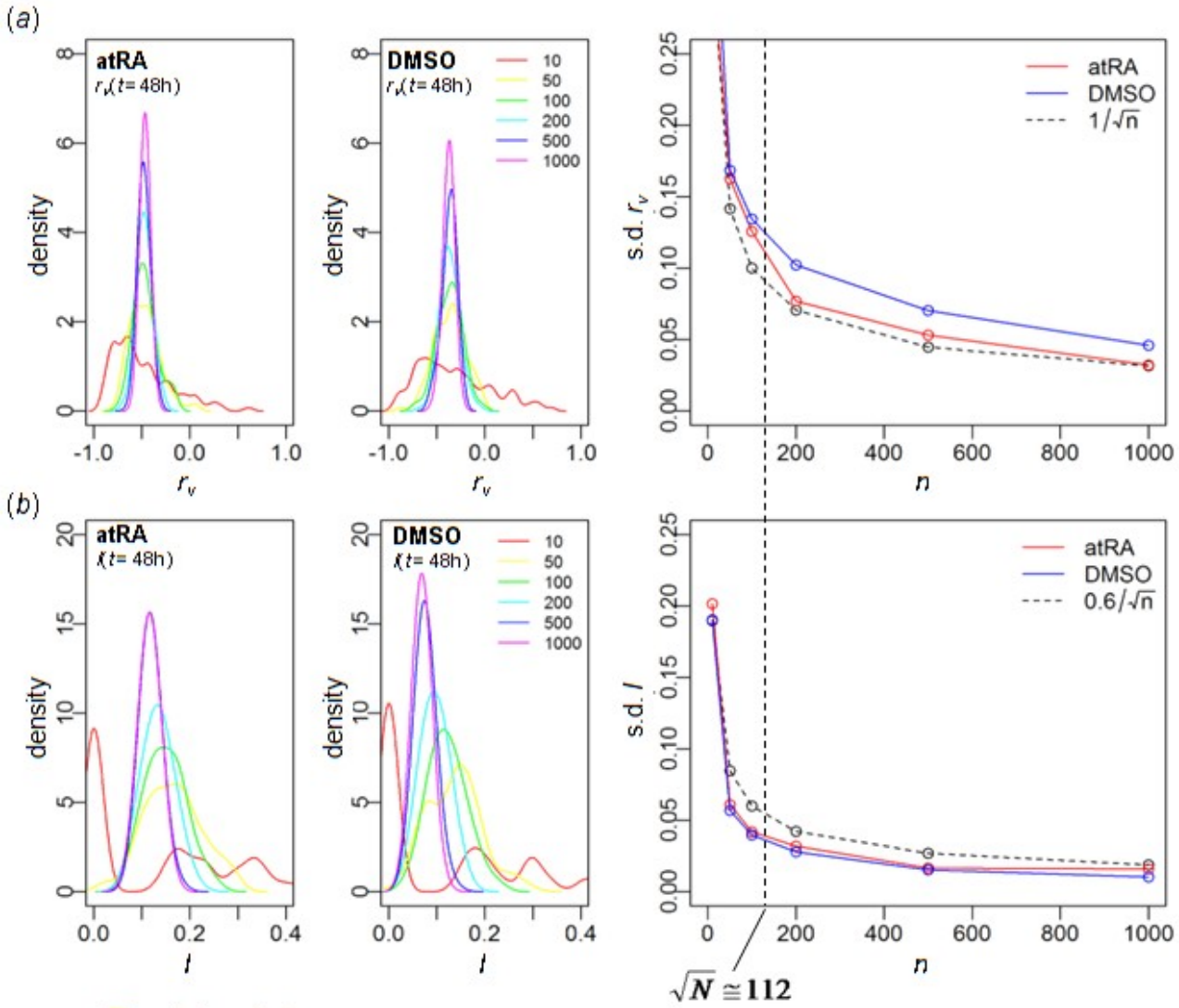
Supplementary Material

Supplementary Figures

Supplementary Figures

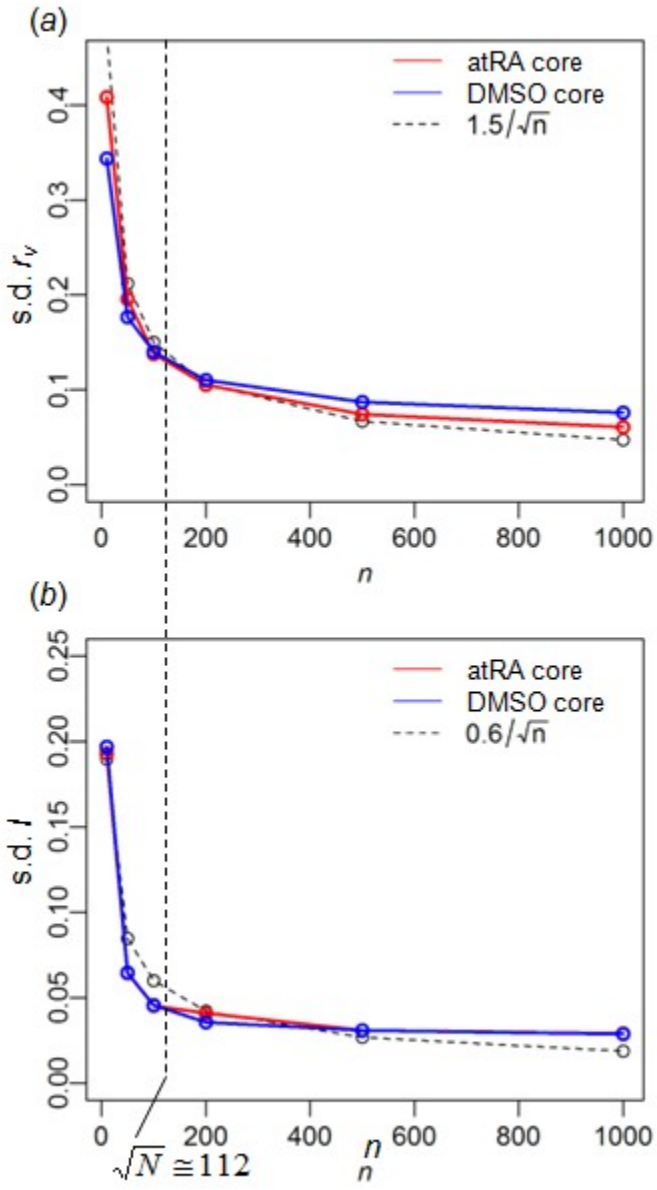
Figure S1. Transition from scattered to smooth distributions of r_v and I. Distributions of (a) r_v and (b) I for groups of n randomly chosen genes from whole genome (n = 10, 50, 100, 200, 500, 1000), estimated by Gaussian kernel with 100 repeats at t = 48h, left panel for atRA response and right panel for DMSO (similar profiles are obtained for all time points). Standard deviation of the distributions of r_v and I (right panels of (a) and (b)) at t = 48h decreases as n increases, following a

$\alpha / \sqrt{n}$ law, $\alpha \cong 1$ for r_v and $\alpha \cong 0.6$ for I.



Tsuchiya et al.
Figure S1

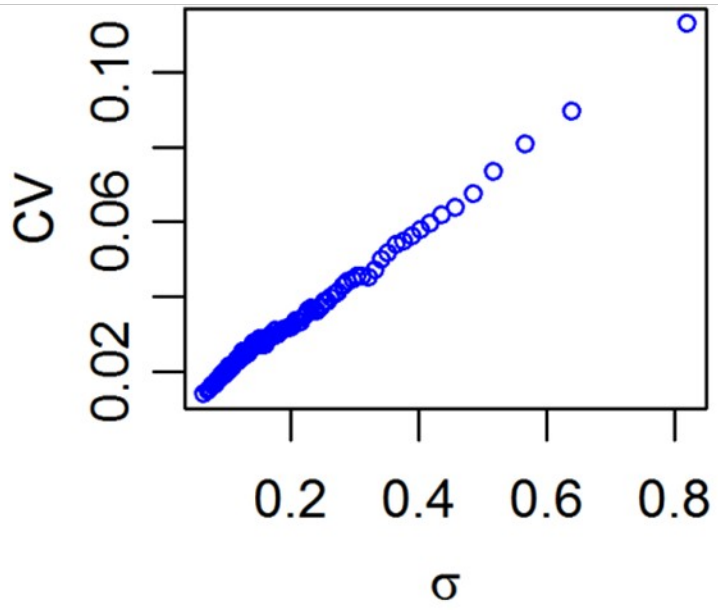
Figure S2. Standard deviation of the attractor cores follows $\alpha/\sqrt{n}$ law. Standard deviation of the superposition of (a) r_v and (b) I at attractor core for atRA (red line) and DMSO (blue line) decreases as n increases, following $\alpha/\sqrt{n}$ law where = 1.5 (for r_v) and 0.6 (for I) .



Tsuchiya et al.
Figure S2

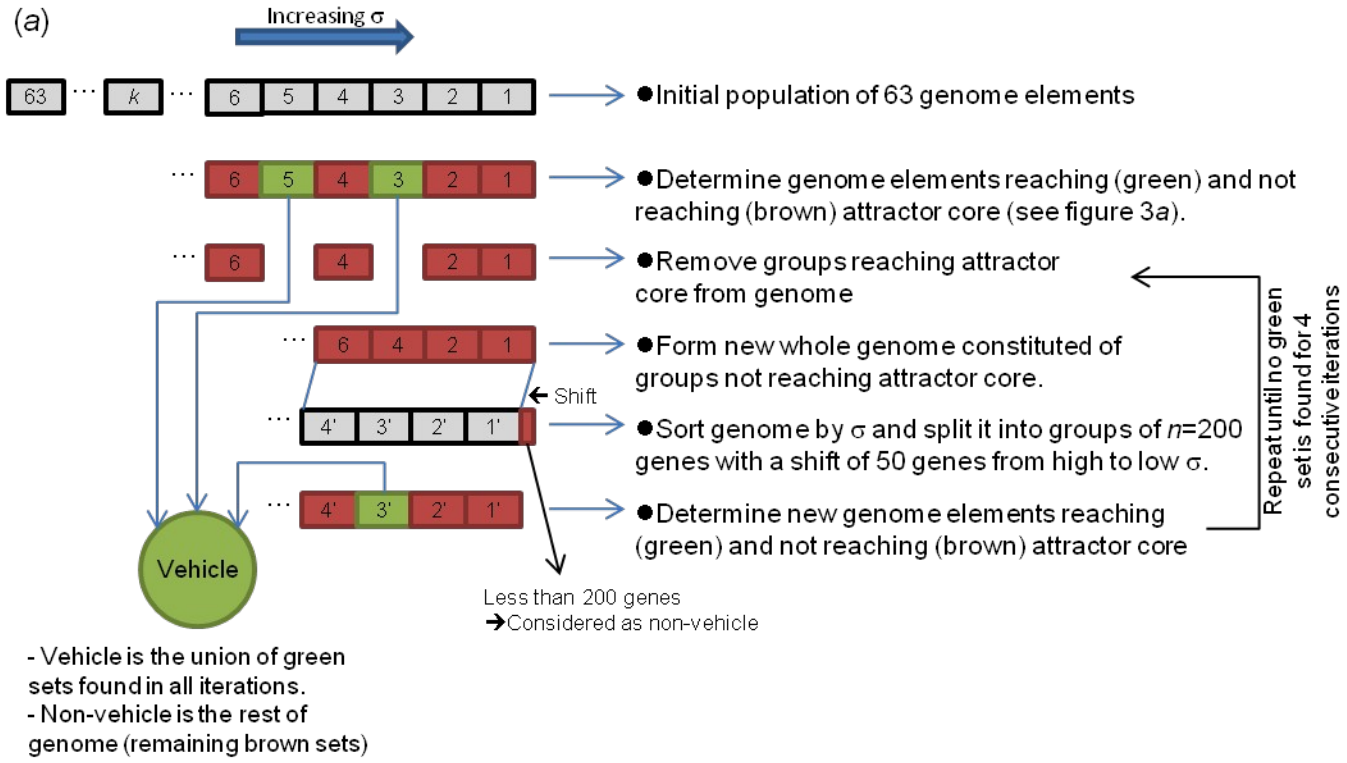
Figure S3. Comparing CV with σ. Standard deviation, σ vs. CV for genome elements of $n_t = 200$

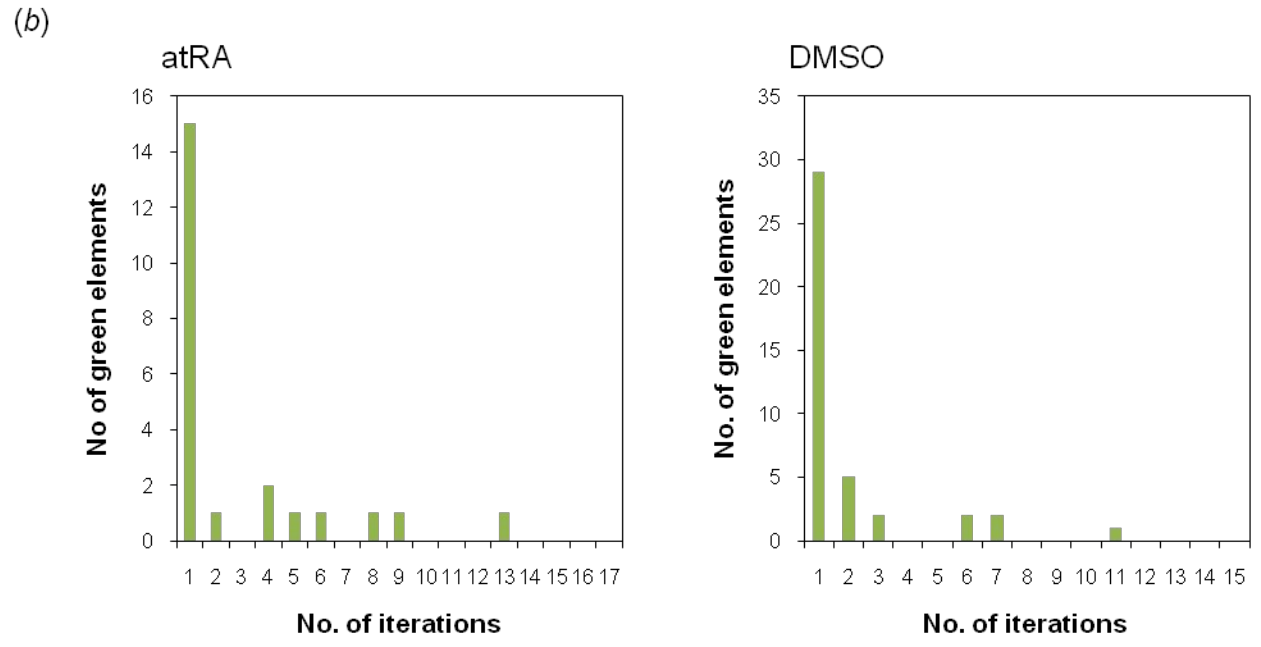
genes sorted by σ.



Tsuchiya et al.
Figure S3

Figure S4. Identifying genome elements that form genetic vehicle. (a) Schematic iterative procedure to determine genetic vehicle. (b) Number of genome elements reaching attractor with respect to iterations.





Tsuchiya et al.
Figure S4