Supporting Online Material for

Attractors and Democratic Dynamics

Yaneer Bar-Yam,* Dion Harmon, Benjamin de Bivort

*E-mail: yaneer@necsi.edu

Published 20 February 2009, *Science* **323**, 1016 (2009)
DOI: 10.1126/science.1163225

This PDF file includes:

Global Patterns of Gene Expression

and

Master Regulators:

Supporting Online Text

for

Attractors and Democratic Dynamics

Yaneer Bar-Yam[1], Dion Harmon[1], and Benjamin de Bivort[1,2]

[1]New England Complex Systems Institute,

24 Mt. Auburn St., Cambridge, Massachusetts 02138

[2] Rowland Institute at Harvard University,

100 Edwin Land Blvd., Cambridge, MA 02139

Genome Wide Dynamics and Control: Archetypes for Collective Behavior

Differences in perspectives about regulatory structures are also related to differences in concept and representation of dynamical processes. Describing individual gene effects begins with identifying individual genes, mechanisms of gene interactions, and pathways of gene products. Describing attractors involves characterizing the convergence of transcriptome wide cell states where majorities of genes determine behavior rather than any one. The difference between low dimensional (few variable) and high dimensional (many variable) collective dynamics is central. Characterizing cellular regulatory networks more generally as distributed control systems where individual genes can exert strong influence requires bridging these two views.

We describe a framework in which individual genes and collective states can be considered together to evaluate their mutual influence. The difficulty we overcome is the contrast in the quantities needed to describe the two different pictures. What is needed are analogs of control coefficients, which have been used to study the impact of individual catalysts on system metabolic flows (*S1*).

Identifying such coefficients requires a measure of differences of collective states. While Pearsons correlation might be used (*S2*), in order to define a fundamentally justifiable measure we identify an archetype, e.g. a representation of a particular cell type, using expression values of all genes $\{e_i^\alpha\}$, where i is the gene index, and α is a cell type label. These values may be taken as a representative member, or a mean over a population of cells of the same type. When a cell has that type, individual expression values may deviate from the archetype values. However, when considered over all genes, the deviations are bounded.

We measure the deviation of a gene expression value e_i relative to an archetype, normalized by the expected deviation over a reference population of cells,

$$d_i^\alpha = \left(\log(e_i) - \log(e_i^\alpha)\right)/\sigma_i. \tag{1}$$

We use logarithms of expression values to obtain better-behaved distributions. The normalization σ_i is chosen to establish a common range of values for d_i^α and can be set to the standard deviation over the reference cell population of $\log(e_i)$, i.e. a population of cells that are of a particular phenotype (σ_i may have additional labels to identify the reference

population). If the population is not large we can approximate $\sigma_i = \sigma$ by considering all genes together in taking the standard deviation, or combining multiple phenotype populations. The proximity of an arbitrary state to the archetype, the conformity or conformance, is given by

$$m^\alpha = \frac{1}{N} \sum_i \mathrm{f}(d_i^\alpha), \tag{2}$$

where f is a function that is 1 for values close to zero, i.e. when a gene expression level is proximate to the archetypical value, and goes to zero as it deviates therefrom. N is the number of genes. The purpose of this function is to prevent individual gene deviations from determining the distance, which should instead depend on whether or not many expression levels are close to archetypical values. We use

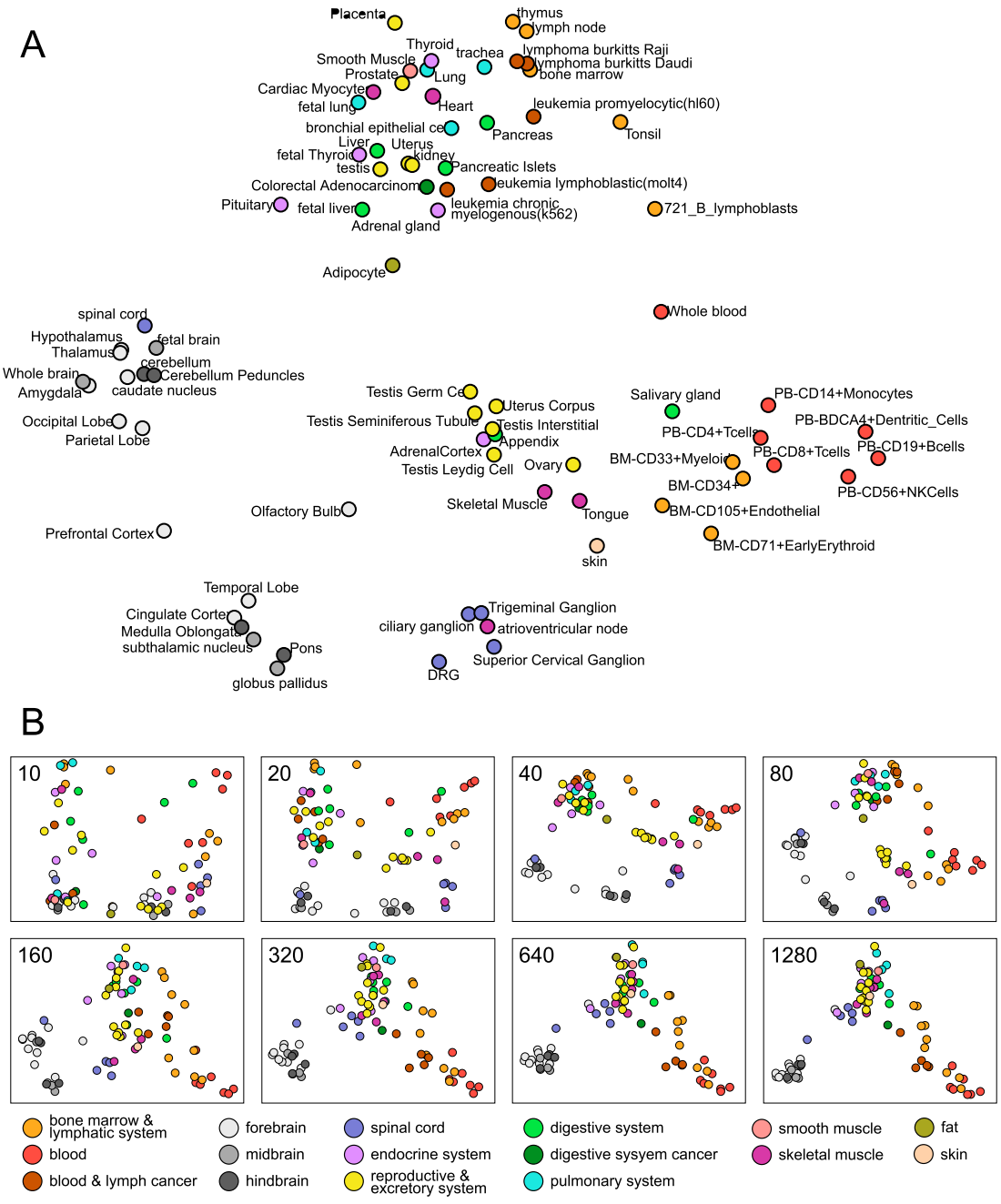
$$m^\alpha = \frac{1}{N} \sum_i \left(1 - \tanh((d_i^\alpha)^2)\right). \tag{3}$$

Control coefficients are specified by the rate of change of the collective displacement $\hat{m}^\alpha = (1 - m^\alpha)$ with respect to the control parameter, measured logarithmically—i.e. the exponent of a power law relationship. Thus we define control or sensitivity coefficients as

$$c_i = \frac{e_i}{\hat{m}^\alpha} \frac{d\hat{m}^\alpha}{e_i} = \frac{d \log(\hat{m}^\alpha)}{d \log(e_i)}. \tag{4}$$

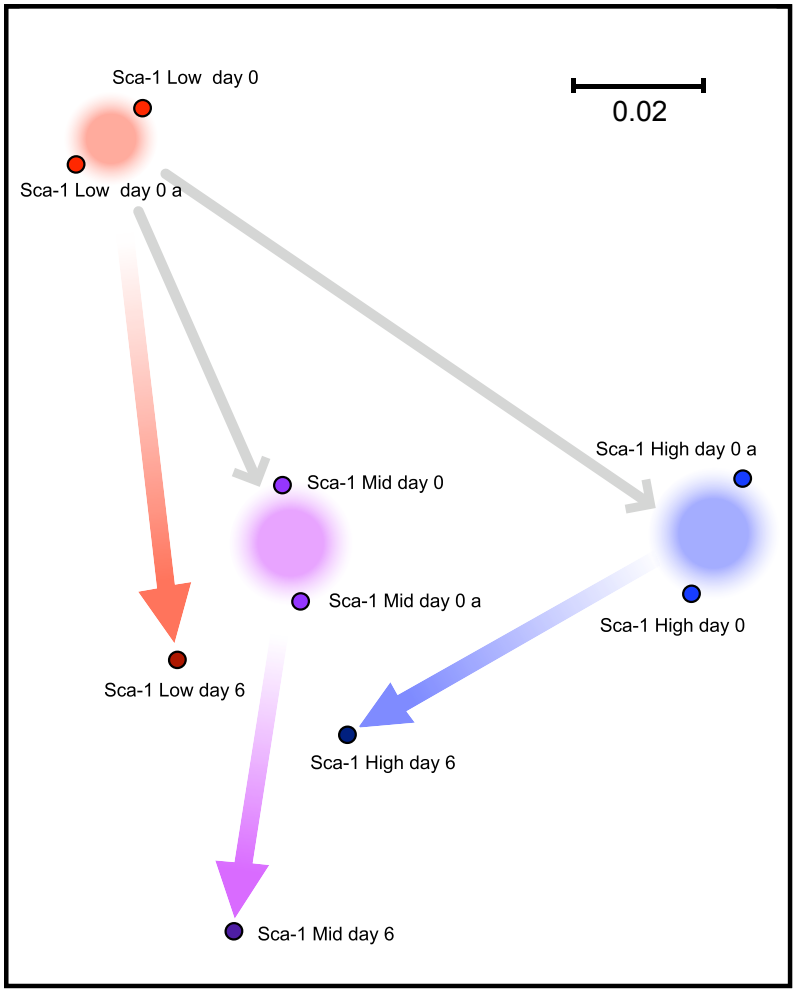
As an example, we evaluated this for Sca-1 in the data of Chang et al. (*S2*), after performing a number of tests (see Supplementary Figure 2), and found a control coefficient of 0.52, consistent with the coupling found between this gene and collective behavior in that paper. This is obtained despite the cells having a high concentration of Sca-1 protein on the cell surface having lower mRNA expression compared to cells with intermediate protein concentrations, presumably due to feedback. We also obtained many other control coefficients from the same data, including 2.20 for Sfpi1, and -.59 for Gata1, lineage-specific transcription factors involved in stem-cell differentiation. Given the nature of the data, this indicates correlation relative to the chosen states and not causation, and the latter can also be characterized by relevant experiments.

Supplementary Figure 1



Principal component analysis reveals transcriptome attractors across tissue types. A) The transcriptional profiles of 79 human tissue and tumor cell types (*S3*) fall into several clusters when they are plotted in the two dimensions that draw the greatest distinctions among tissue types when considering the 80 genes with most varying expression levels. The dimensions of maximum variation are obtained by principal component analysis (*S4*). Tissues are color coded by category. B) As in A for analyses done with the n most varying genes as indicated. The tissue type clusters approach their final conformation when hundreds of genes are considered.

Supplementary Figure 2



Collective dynamics of cell types. Dots represent high dimensional cell states from Chang et al. (*S2*), with two replicates each of cultures distinguished by low (red), mid (purple) and high (blue) concentrations of the surface marker Sca-1, and subsequent convergence of these cultures after 6-days. An optimized two-dimensional embedding of distances given by $\hat{m}^\alpha$ (see supporting online text) is shown, with scale bar in units of $\hat{m}^\alpha$. Control coefficients can be calculated as the ratio of the $\log(\hat{m}^\alpha)$ (grey arrows) change to the log of individual gene expression value change using as reference the Sca-1 low 0 day state.

References

S1. H. Kacser, J. A. Burns, The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65 (1973).

S2. H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, S. Huang, *Nature* **453**, 544 (2008).

S3. A. I. Su et al., *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6062 (2004).

S4. I. Joffe, *Principal Component Analysis* (Wiley, Hoboken, NJ, 2005).