Processing and modeling genome-wide expression data using singular value decomposition

Orly Alter[a], Patrick O. Brown[b], and David Botstein[a]

Departments of [a]Genetics and [b]Biochemistry, Stanford University, Stanford, CA 94305

ABSTRACT

We describe the use of singular value decomposition in transforming genome-wide expression data from genes × arrays space to reduced diagonalized "eigengenes" × "eigenarrays" space, where the eigengenes (or eigenarrays) are unique orthonormal superpositions of the genes (or arrays). Normalizing the data by filtering out the eigengenes (and eigenarrays) that are inferred to represent additive or multiplicative noise, experimental artifacts, or even irrelevant biological processes enables meaningful comparison of the expression of different genes across different arrays in different experiments. Sorting the data according to the eigengenes and eigenarrays gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. After normalization and sorting, the significant eigengenes and eigenarrays can be associated with observed genome-wide effects of regulators, or with measured samples, in which these regulators are overactive or underactive, respectively.

Keywords: Singular value decomposition (SVD); Genome-wide expression data; DNA microarrays.

INTRODUCTION

DNA microarray technology (1,2) and genome sequencing have advanced to the point that it is now possible to monitor gene expression levels on a genomic scale (e.g., 3). These new data promise to enhance fundamental understanding of life on the molecular level, from regulation of gene expression and gene function to cellular mechanisms, and may prove useful in medical diagnosis, treatment and drug design. Analysis of these new data requires mathematical tools that are adaptable to the large quantities of data, while reducing the complexity of the data to make them comprehensible. Analysis so far has been limited to identification of genes and arrays with similar expression patterns using clustering methods (e.g., 4–9).

We describe the use of singular value decomposition (SVD) (10), also known as Karhunen-Loève expansion (11) and as principal-component analysis (12), in analyzing genome-wide expression data (13). SVD is a linear transformation of the expression data from the genes × arrays space to the reduced "eigengenes" × "eigenarrays" space. In this space the data are diagonalized, such that each eigengene is expressed only in the corresponding eigenarray, with the corresponding "eigenexpression" level indicating their relative significance. The eigengenes and eigenarrays are unique, and therefore also data-driven, orthonormal superpositions of the genes and arrays, respectively.

We show that several significant eigengenes and the corresponding eigenarrays capture most of the expression information, thus allowing for dimension reduction and estimation of missing data. Decorrelation of the eigengenes (and eigenarrays) suggests the possibility that some of the significant eigengenes (and corresponding eigenarrays) represent independent processes, biological or experimental (and corresponding cellular states), which contribute to the overall expression. Decoupling of the eigengenes (and eigenarrays) allows filtering out any one of these processes from the data without eliminating genes or arrays. Normalizing the data by filtering out the eigengenes (and the corresponding eigenarrays) that are inferred to represent additive or multiplicative noise, experimental artifacts or even irrelevant biological processes, enables meaningful comparison of the expression of different genes across different arrays in different experiments. Such normalization may improve any further analysis of the expression data. Sorting the data according to the correlations of the genes (and arrays) with eigengenes (and eigenarrays) gives a global picture of the dynamics of gene expression, in which individual genes and arrays appear to be classified into groups of similar regulation and function, or similar cellular state and biological phenotype, respectively. These groups of genes (or arrays) are not defined by overall similarity in expression, but only by similarity in the expression of any chosen subset of eigengenes (or eigenarrays).

We also show that upon comparing two or more similar experiments, with a regulator being overactive or underactive in one but normally expressed in the others, the expression pattern of one of the significant eigengenes may be

Microarrays: Optical Technologies and Informatics, Michael L. Bittner, Yidong Chen,
Andreas N. Dorsel, Edward R. Dougherty, Editors, Proceedings of SPIE Vol. 4266 (2001)

correlated with the expression patterns of this regulator and its targets. This eigengene, therefore, can be associated with the observed genome-wide effect of the regulator. The expression pattern of the corresponding eigenarray is correlated with the expression patterns observed in samples in which the regulator is overactive or underactive. This eigenarray, therefore, can be associated with these samples. We conclude that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning.

MATHEMATICAL FRAMEWORK: SINGULAR VALUE DECOMPOSITION

The relative expression levels of N genes of a model organism, which may constitute almost the entire genome of this organism, in a single sample, are probed simultaneously by a single microarray. A series of M arrays, which are almost identical physically, probes the genome-wide expression levels in M different samples, i.e., under M different experimental conditions. Let the matrix $\hat{e}$, of size N-genes $\times$ M-arrays, tabulate the full expression data. Each element of $\hat{e}$ satisfies $\langle n|\hat{e}|m\rangle \equiv e_{nm}$ for all $1 \leq n \leq N$ and $1 \leq m \leq M$, where e_{nm} is the relative expression level of the nth gene in the mth sample as measured by the mth array.* The vector in the nth row of the matrix $\hat{e}$, $\langle g_n| \equiv \langle n|\hat{e}$, lists the relative expression of the nth gene across the different samples which correspond to the different arrays. The vector in the mth column of the matrix $\hat{e}$, $|a_m\rangle \equiv \hat{e}|m\rangle$, lists the genome-wide relative expression measured by the mth array.

SVD is then linear transformation of the expression data from the N-genes $\times$ M-arrays space to the reduced L-"eigenarrays" $\times$ L-"eigengenes" space, where $L = \min\{M, N\}$ (see Fig. 13 in supplemental material at http://genome-www.stanford.edu/SVD/),

$$\hat{e} \quad = \quad \hat{u}\,\hat{\epsilon}\,\hat{v}^T. \tag{1}$$

In this space the data are represented by the diagonal nonnegative matrix $\hat{\epsilon}$, of size L-eigengenes $\times$ L-eigenarrays, which satisfies $\langle k|\hat{\epsilon}|l\rangle \equiv \epsilon_l \delta_{kl} \geq 0$ for all $1 \leq k, l \leq L$, such that the lth eigengene is expressed only in the corresponding lth eigenarray, with the corresponding "eigenexpression" level ϵ_l. Therefore, the expression of each eigengene (or eigenarray) is decoupled from that of all other eigengenes (or eigenarrays). The "probability of eigenexpression"

$$p_l = \epsilon_l^2 / \sum_{k=1}^{L} \epsilon_k^2, \tag{2}$$

indicates the relative significance of the lth eigengene and eigenarray in terms of the fraction of the overall expression that they capture. Assume also that the eigenexpression levels are arranged in decreasing order of significance, such that $\epsilon_1 \geq \epsilon_2 \geq \ldots \geq \epsilon_L \geq 0$. The "normalized Shannon entropy" of a dataset,

$$0 \leq \; d = \frac{-1}{\log(L)} \sum_{k=1}^{L} p_k \log(p_k) \; \leq 1, \tag{3}$$

measures the complexity of the data from the distribution of the overall expression between the different eigengenes (and eigenarrays), where $d = 0$ corresponds to an ordered and redundant dataset in which all expression is captured by a single eigengene (and eigenarray), and $d = 1$ corresponds to a disordered and random dataset where all eigengenes (and eigenarrays) are equally expressed.

The transformation matrices $\hat{u}$ and $\hat{v}^T$ define the N-genes $\times$ L-eigenarrays and the L-eigengenes $\times$ M-arrays basis sets, respectively. The vector in the lth row of the matrix $\hat{v}^T$, $\langle \gamma_l| \equiv \langle l|\hat{v}^T$, lists the expression of the lth eigengene across the different arrays. The vector in the lth column of the matrix $\hat{u}$, $|\alpha_l\rangle \equiv \hat{u}|l\rangle$, lists the genome-wide expression in the lth eigenarray. The eigengenes and eigenarrays are orthonormal superpositions of the genes and arrays, such that the transformation matrices $\hat{u}$ and $\hat{v}$ are both orthogonal

$$\hat{u}^T\hat{u} = \hat{v}^T\hat{v} = \hat{I}, \tag{4}$$

where $\hat{I}$ is the identity matrix. Therefore, the expression of each eigengene (or eigenarray) is not only decoupled but also decorrelated from that of all other eigengenes (or eigenarrays). The eigengenes and eigenarrays are unique,

*In this manuscript, $\hat{m}$ denotes a matrix, $|v\rangle$ denotes a column vector and $\langle u|$ denotes a row vector, such that $\hat{m}|v\rangle$, $\langle u|\hat{m}$ and $\langle u|v\rangle$ all denote inner products and $|v\rangle\langle u|$ denotes an outer product.

except in degenerate subspaces, defined by subsets of equal eigenexpression levels, and except for a phase factor of ± 1, such that each eigengene (or eigenarray) captures both parallel and antiparallel gene (or array) expression patterns. Therefore, SVD is data-driven, except in degenerate subspaces.

SVD Calculation. According to Eqs. (1) and (4), the M-arrays $\times$ M-arrays symmetric correlation matrix $\hat{a} = \hat{e}^T\hat{e} = \hat{v}\hat{\epsilon}^2\hat{v}^T$ is represented in the L-eigengenes $\times$ L-eigengenes space by the diagonal matrix $\hat{\epsilon}^2$. The N-genes $\times$ N-genes correlation matrix $\hat{g} = \hat{e}\hat{e}^T = \hat{u}\hat{\epsilon}^2\hat{u}^T$ is represented in the L-eigenarrays $\times$ L-eigenarrays space also by $\hat{\epsilon}^2$, where for $L = \min\{M, N\} = M$, $\hat{g}$ has a null subspace of at least $N - M$ null eigenvalues. We calculate the SVD of a dataset $\hat{e}$, with $M \ll N$, by diagonalizing $\hat{a}$, and then projecting the resulting $\hat{v}$ and $\hat{\epsilon}$, onto $\hat{e}$ to obtain $\hat{u} = \hat{e}\hat{v}\hat{\epsilon}^{-1}$.

Missing Data Estimation. The high entropy of an expression dataset, where the most significant eigengenes (and eigenarrays) capture most of the expression information, suggests the possibility of dimension reduction, where the eigenexpression levels corresponding to the least significant eigengenes (and eigenarrays) are approximated to be zero. The inference that these significant eigengenes (and eigenarrays) represent independent processes (or cellular states) (13, and see also 14–17) suggests the possibility of using the expression patterns of these eigengenes for meaningful estimation of missing data. For the nth gene $|g_n\rangle$, with missing data in $M' < M$ of the arrays, we estimate the missing expression level in the mth array $\langle m|g_n\rangle$ to be a superposition of the expression levels of the $L' < M - M'$ significant eigengenes $\{|\gamma_l'\rangle\}$ in the mth array, as calculated for the subset of $N' < N$ genes with no missing data in any of the M arrays. The coefficients of this superposition are determined by the expansion of the expression of the nth gene across all $M - M'$ arrays with no missing data, $|g_n\rangle_{M'}$, in the subspace spanned by the expression patterns of the significant eigengenes across the same $M - M'$ arrays, $\{|\gamma_l'\rangle_{M'}\}$, such that $\langle m|g_n\rangle \rightarrow \sum_{l=1}^{L'} \langle m|\gamma_l'\rangle \ {}_{M'}\langle\beta_l'|g_n\rangle_{M'}$ where ${}_{M'}\langle\beta_k'|\gamma_l'\rangle_{M'} = \delta_{kl}$, i.e., $\{\ {}_{M'}\langle\beta_l'|\}$ span the $L' \times (M - M')$ subspace $(\hat{v}_{M'}')^\dagger$ that is pseudoinverse to the $(M - M') \times L'$ subspace $\hat{v}_{M'}'$, which is spanned by $\{|\gamma_l'\rangle_{M'}\}$. We use the SVD of $\hat{v}_{M'}' \equiv \hat{U}\hat{\omega}\hat{V}^T$ to calculate the pseudoinverse $(\hat{v}_{M'}')^\dagger = \hat{V}\hat{\omega}^{-1}\hat{U}^T$, that according to Eq. (4) satisfies $(\hat{v}_{M'}')^\dagger\hat{v}_{M'}' = \hat{I}$, where $\hat{V}\hat{V}^T = \hat{I}$ for all $L' < M - M'$. Assuming that the L' significant eigengenes as calculated for the N' genes with no missing data are meaningful patterns for missing data estimation, we expect these eigenegenes and corresponding probablities of eigenexpression to be similar to those calculated for the full dataset of N genes after the missing data are estimated.

Pattern Inference. The decorrelation of the eigengenes (and eigenarrays) suggests the possibility that some of the eigengenes (and eigenarrays) represent independent regulatory programs or processes (and corresponding cellular states), which contribute to the overall expression. We infer that an eigengene $|\gamma_l\rangle$ represents a regulatory program or process from its expression pattern across all arrays, when this pattern is biologically interpretable. This inference may be supported by a corresponding coherent biological theme reflected in the functions of the genes, whose expression patterns correlate or anticorrelate with the pattern of this eigengene. With this we assume that the corresponding eigenarray $|\alpha_l\rangle$ (which lists the amplitude of this eigengene pattern in the expression of each gene $|g_n\rangle$ relative to all other genes $\langle n|\alpha_l\rangle = \langle g_n|\gamma_l\rangle/\epsilon_l$) represents the cellular state which corresponds to this process. We infer that the eigenarray $|\alpha_l\rangle$ represents a cellular state from the arrays whose expression patterns correlate or anticorrelate with the pattern of this eigenarray. Upon sorting of the genes, this inference may be supported by the expression pattern of this eigenarray across all genes, when this pattern is biologically interpretable.

Data Filtering and Normalization. The decoupling of the eigengenes and eigenarrays allows filtering the data without eliminating genes or arrays from the dataset. We filter out any of the eigengenes $|\gamma_l\rangle$ (and the corresponding eigenarray $|\alpha_l\rangle$) $\hat{e} \rightarrow \hat{e} - \epsilon_l|\alpha_l\rangle\langle\gamma_l|$, by substituting zero for the eigenexpression level $\epsilon_l = 0$ in the diagonal matrix $\hat{\epsilon}$ and reconstructing the data according to Eq. (1). We normalize the data by filtering out those eigengenes (and eigenarrays) which are inferred to represent additive or multiplicative noise or experimental artifacts.

Degenerate Subspace Rotation. The uniqueness of the eigengenes and eigenarrays does not hold in a degenerate subspace, defined by equal eigenexpression levels. We approximate significant similar eigenexpression levels $\epsilon_l \approx \epsilon_{l+1} \approx \ldots \approx \epsilon_m$ with $\epsilon_l = \ldots = \epsilon_m = \sqrt{\sum_{k=l}^{m} \epsilon_k^2/(m - l + 1)}$. Therefore, Eqs. (1)–(4) remain valid upon rotation of the corresponding eigengenes $\{(|\gamma_l\rangle, \ldots, |\gamma_m\rangle) \rightarrow \hat{R}(|\gamma_l\rangle, \ldots, |\gamma_m\rangle)\}$, and eigenarrays $\{(|\alpha_l\rangle, \ldots, |\alpha_m\rangle) \rightarrow \hat{R}(|\alpha_l\rangle, \ldots, |\alpha_m\rangle)\}$, for all orthogonal $\hat{R}$, $\hat{R}^T\hat{R} = \hat{I}$. We choose a unique rotation $\hat{R}$ by subjecting the rotated eigengenes to $m - l$ constraints, such that these constrained eigengenes may be advantageous in interpreting and presenting the expression data.

Data Sorting. Inferring that eigengenes (and eigenarrays) represent independent processes (and cellular states) allows sorting the data by similarity in the expression of any chosen subset of these eigengenes (and eigenarrays), rather than by overall similarity in expression (13, and see also 14–17). Given two eigengenes $|\gamma_k\rangle$ and $|\gamma_l\rangle$ (or

eigenarrays $|\alpha_k\rangle$ and $|\alpha_l\rangle$), we plot the correlation of $|\gamma_k\rangle$ with each gene $|g_n\rangle$, $\langle\gamma_k|g_n\rangle/\langle g_n|g_n\rangle$ (or $|\alpha_k\rangle$ with each array $|a_m\rangle$) along the y-axis, vs. that of $|\gamma_l\rangle$ (or $|\alpha_l\rangle$) along the x-axis. In this plot, the distance of each gene (or array) from the origin is its amplitude of expression in the subspace spanned by $|\gamma_k\rangle$ and $|\gamma_l\rangle$ (or $|\alpha_k\rangle$ and $|\alpha_l\rangle$), relative to its overall expression $r_n \equiv \langle g_n|g_n\rangle^{-1}\sqrt{|\langle\gamma_k|g_n\rangle|^2 + |\langle\gamma_l|g_n\rangle|^2}$ (or $r_m \equiv \langle a_m|a_m\rangle^{-1}\sqrt{|\langle\alpha_k|a_m\rangle|^2 + |\langle\alpha_l|a_m\rangle|^2}$). The angular distance of each gene (or array) from the x-axis is its phase in the transition from the expression pattern $|\gamma_l\rangle$ to $|\gamma_k\rangle$ and back to $|\gamma_l\rangle$ (or $|\alpha_l\rangle$ to $|\alpha_k\rangle$ and back to $|\alpha_l\rangle$) $\tan\phi_n \equiv \langle\gamma_k|g_n\rangle/\langle\gamma_l|g_n\rangle$, (or $\tan\phi_m \equiv \langle\alpha_k|a_n\rangle/\langle\alpha_l|a_m\rangle$). We sort the genes (or arrays) according to ϕ_n (or ϕ_m).

BIOLOGICAL DATA ANALYSIS: ELUTRIATION-SYNCHRONIZED CELL CYCLE

Spellman et al. (3) monitored genome-wide mRNA levels, for 6113 ORFs of the budding yeast *Saccharomyces cerevisiae* simultaneously, over approximately one cell cycle period, $T \approx 390$m, in a yeast culture synchronized by elutriation, relative to a reference mRNA from an asynchronous yeast culture, at 30m intervals. The elutriation dataset we analyze (see supplemental dataset and Mathematica notebook at http://genome-www.stanford.edu/SVD/) tabulates the measured ratios of gene expression levels for the $N = 6018$ genes, $N' = 5986$ of which with no missing data in the $M = 14$ arrays, corresponding to the 14 measured time points, and 32 with no missing data in 13 of the 14 arrays. Of these genes, 791 were classified by Spellman et al. as cell cycle regulated and 103 were classified as cell cycle regulated by traditional methods.

Missing Data Estimation. We use the $L' = 4$ most significant eigengene patterns, as calculated for the subset of 5986 genes with no missing data in the 14 arrays (13), in order to estimate the missing data in the remaining 32 genes, with no missing data in 13 of the 14 arrays. We find that these eigenegenes and corresponding probablities of eigenexpression are similar to those calculated for the full dataset of 6018 genes after the missing data are estimated.

Pattern Inference. Consider the 14 eigengenes of the full elutriation dataset. The 1st and most significant eigengene $|\gamma_1\rangle$, which describes time-invariant relative expression during the cell cycle (Fig. 14*a* in supplemental material at http://genome-www.stanford.edu/SVD/), captures more than 90% of the overall relative expression in this experiment (Fig. 14*b*). The entropy of the dataset, therefore, is small $d = 0.14 \ll 1$. This suggests that the underlying processes are manifested by weak perturbations of a steady state of expression. This also suggests that time-invariant additive constants due to uncontrolled experimental variables may be superimposed on the data. We infer that $|\gamma_1\rangle$ represents experimental additive constants superimposed on a steady gene expression state, and assume that $|\alpha_1\rangle$ represents the corresponding steady cellular state.

The 2nd, 3rd and 4th eigengenes, which show oscillations during the cell cycle (Fig. 2*c*), capture about 3%, 1% and 0.5% of the overall relative expression, respectively. The time variation of $|\gamma_3\rangle$ fits a normalized sine function of period T, $\sqrt{2/T}\sin(2\pi t/T)$. We infer that $|\gamma_3\rangle$ represents expression oscillation, which is consistent with gene expression oscillations during a cell cycle, and assume that $|\alpha_3\rangle$ represents the corresponding cellular state. The time variations of the 2nd and 4th eigengenes fit a cosine function of period T with $\sqrt{1/2}$ the amplitude of a normalized cosine with this period, $\sqrt{1/T}\cos 2\pi t/T$. However, while $|\gamma_2\rangle$ shows decreasing expression on transition from $t = 0$ to 30m, $|\gamma_4\rangle$ shows increasing expression. We infer that $|\gamma_2\rangle$ and $|\gamma_4\rangle$ represent initial transient increase and decrease in expression in response to the elutriation, respectively, superimposed on expression oscillation during the cell cycle, and assume that $|\alpha_2\rangle$ and $|\alpha_4\rangle$ represent the corresponding cellular states.

Data Filtering and Normalization. We filter out the 1st eigengene and eigenarray of the elutriation dataset, $\hat{e} \to \hat{e}_C = \hat{e} - \epsilon_1|\alpha_1\rangle\langle\gamma_1|$, removing the steady state of expression. Each of the elements of the dataset $\hat{e}_C$, $\langle n|\hat{e}_C|m\rangle \equiv e_{C,nm}$, is the difference of the measured expression of the nth gene in the mth array from the steady-state levels of expression for these gene and array as calculated by SVD. Therefore, $e^2_{C,nm}$ is the variance in the measured expression of the nth gene in the mth array. Let $\hat{e}_{LV}$ tabulate the natural logarithm of the variances in elutriation expression, such that each element of $\hat{e}_{LV}$ satisfies $\langle n|\hat{e}_{LV}|m\rangle \equiv \log(e^2_{C,nm})$ for all $1 \le n \le N$ and $1 \le m \le M$, and consider the eigengenes of $\hat{e}_{LV}$ (Fig. 15*a* in supplemental material at http://genome-www.stanford.edu/SVD/). The 1st eigengene $|\gamma_1\rangle_{LV}$, which captures more than 80% of the overall information in this dataset (Fig. 15*b*), describes a weak initial transient increase superimposed on a time-invariant scale of expression variance. The initial transient increase in the scale of expression variance may be a response to the elutriation. The time-invariant scale of expression variance suggests that a steady scale of experimental as well as biological uncertainty is associated with the expression data. This also suggests that time-invariant multiplicative constants due to uncontrolled experimental variables may be superimposed on the data. We filter out $|\gamma_1\rangle_{LV}$, removing the steady scale of expression variance, $\hat{e}_{LV} \to \hat{e}_{CLV} = \hat{e}_{LV} - \epsilon_{1,LV}|\alpha_1\rangle_{LV\ LV}\langle\gamma_1|$.

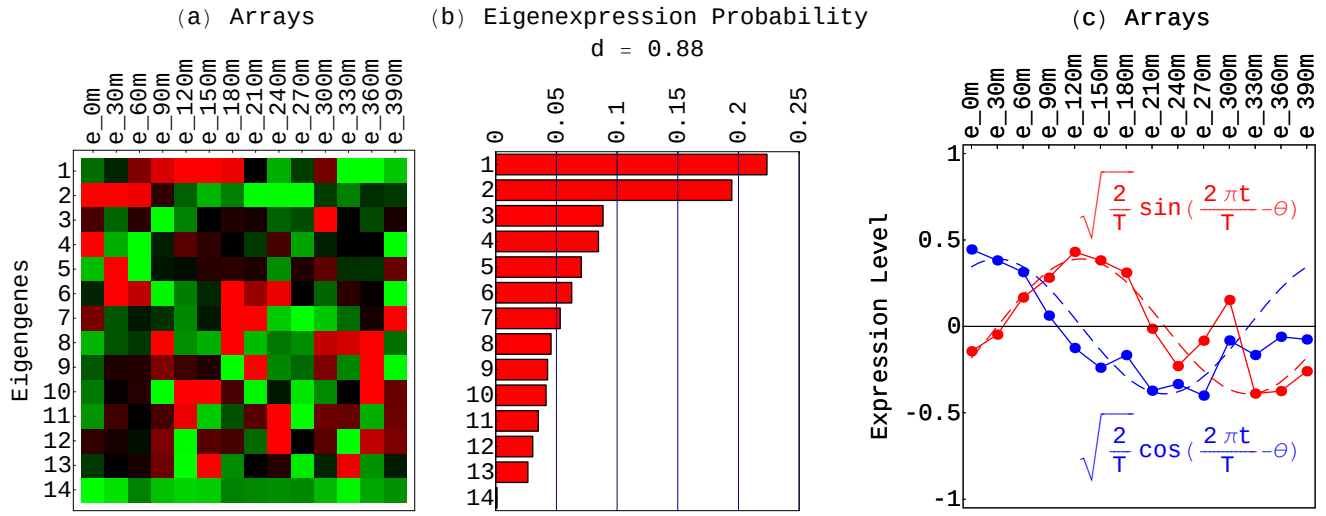

Figure 1. Normalized elutriation eigengenes. (a) Raster display of $\hat{v}_N^T$, the expression of 14 eigengenes in 14 arrays, with overexpression (red), no change in expression (black), and underexpression (green) around the steady-state level of expression. (b) Bar chart of the probabilities of eigenexpression, showing $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ capture about 20% of the overall normalized expression each, and a high entropy of $d = 0.88$. (c) Line-joined graphs of the expression levels of $|\gamma_1\rangle_N$ (red) and $|\gamma_2\rangle_N$ (blue) in the 14 arrays, fit dashed graphs of normalized sine (red) and cosine (blue) of period $T = 390$m and phase $\theta = 2\pi/13$, respectively.

The normalized elutriation dataset $\hat{e}_N$, where each of its elements satisfies $\langle n|\hat{e}_N|m\rangle \equiv \text{sign}(e_{C,nm})\sqrt{\exp(e_{CLV,nm})}$, tabulates for each gene and array expression patterns that are approximately centered at the steady-state expression level (i.e., of approximately zero arithmetic means), with variances which are approximately normalized by the steady scale of expression variance (i.e., of approximately unit geometric means). The 1st and 2nd eigengenes, $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, of $\hat{e}_N$ (Fig. 1a), which are of similar significance, capture together more than 40% of the overall normalized expression (Fig. 1b). The time variations of $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ fit normalized sine and cosine functions of period T and initial phase $\theta \approx 2\pi/13$, $\sqrt{2/T}\sin(2\pi t/T - \theta)$ and $\sqrt{2/T}\cos(2\pi t/T - \theta)$, respectively (Fig. 1$c$). We infer that $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ represent cell cycle expression oscillations, and assume that the corresponding eigenarrays $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ represent the corresponding cell cycle cellular states. Upon sorting of the genes (and arrays) according to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ (and $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$), the initial phase $\theta \approx 2\pi/13$ can be interpreted as a delay of 30m between the start of the experiment and that of the cell cycle stage G1. The decay to zero in the time variation of $|\gamma_2\rangle_N$ at $t = 360$ and 390m can be interpreted as dephasing in time of the initially synchronized yeast culture.

Data Sorting. Consider the normalized expression of the 14 elutriation arrays $\{|a_m\rangle\}$ in the subspace spanned by $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, which is assumed to represent approximately all cell cycle cellular states (Fig. 2a). All arrays have at least 25% of their normalized expression in this subspace, with their distances from the origin satisfying $0.5 \leq r_m < 1$, except for the 11th array $|a_{11}\rangle$. This suggests that $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ are sufficient to approximate the elutriation array expression. The sorting of the arrays according to their phases $\{\phi_m\}$, which describes the transition from the expression pattern $|\alpha_2\rangle_N$ to $|\alpha_1\rangle_N$ and back to $|\alpha_2\rangle_N$, gives an array order that is similar to that of the cell cycle time points measured by the arrays, an order which describes the progress of the cell cycle expression from the M/G1 stage through G1, S, S/G2, and G2/M and back to M/G1. Since $|\alpha_1\rangle_N$ is correlated with the arrays $|a_4\rangle$, $|a_5\rangle$, $|a_6\rangle$, and $|a_7\rangle$ and is anticorrelated with $|a_{13}\rangle$ and $|a_{14}\rangle$, we associate $|\alpha_1\rangle_N$ with the cell cycle cellular state of transition from G1 to S, and $-|\alpha_1\rangle_N$ with the transition from G2/M to M/G1. Similarly, $|\alpha_2\rangle_N$ is correlated with $|a_2\rangle$ and $|a_3\rangle$, and therefore we associate $|\alpha_2\rangle_N$ with the transition from M/G1 to G1. Also, $|\alpha_2\rangle_N$ is anticorrelated with $|a_8\rangle$ and $|a_{10}\rangle$, and therefore we associate $-|\alpha_2\rangle_N$ with the transition from S to S/G2. With these associations the phase of $|a_1\rangle$, $\phi_1 = -\theta \approx -2\pi/13$, corresponds to the 30m delay between the start of the experiment and that of the cell cycle stage G1, which is also present in the inferred cell cycle expression oscillations $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$.

Consider also the expression of the 6018 genes $\{|g_n\rangle\}$ in the subspace spanned by $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, which is inferred to represent approximately all cell cycle expression oscillations. One may expect that genes that have almost all of their normalized expression in this subspace with $r_n \approx 1$ are cell cycle regulated, and that genes that have almost no

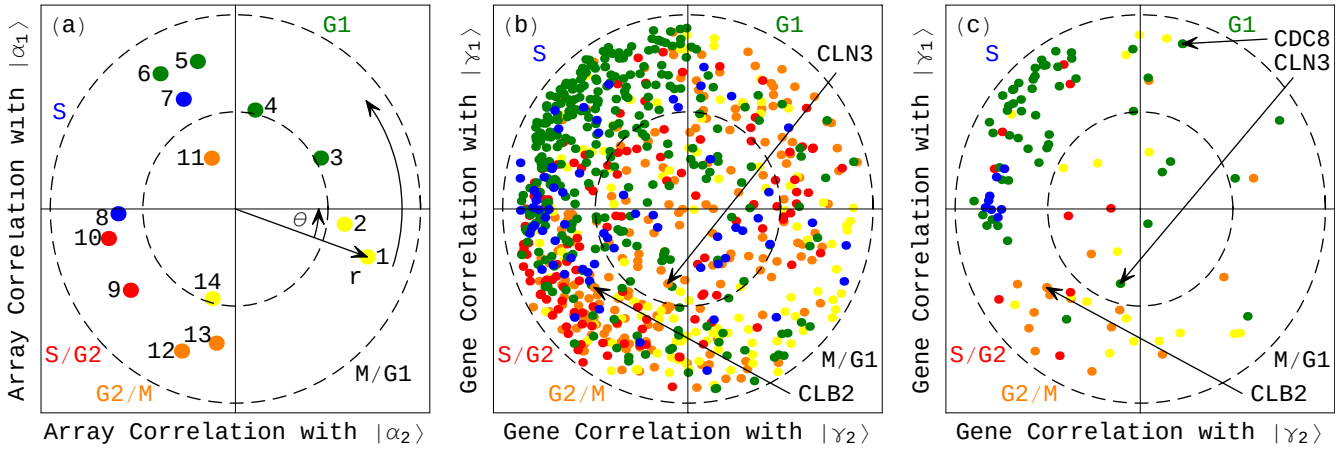

Figure 2. Normalized elutriation expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_N$ along the y-axis vs. that with $|\alpha_2\rangle_N$ along the x-axis, color-coded according to the classification of the arrays into the 5 cell cycle stages, M/G1 (yellow), G1 (green), S (blue), S/G2 (red) and G2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 791 cell cycle regulated genes, color-coded according to the classification by Spellman et al. (3). (*c*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 103 cell cycle regulated genes, color-coded according to the traditional classification (3).

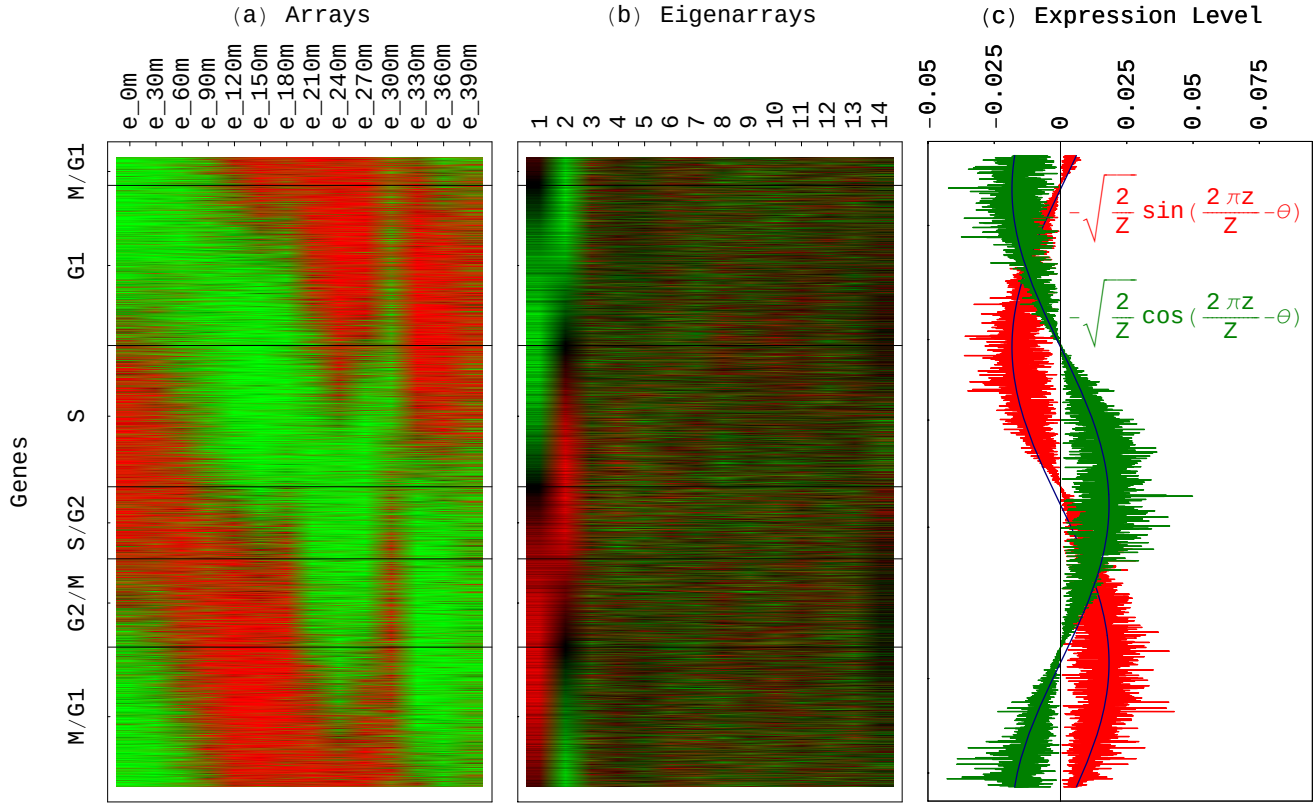

Figure 3. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized elutriation. (*a*) Normalized elutriation expression of the sorted 6018 genes in the 14 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z = N - 1 = 6017$ and phase $\theta \approx 2\pi/13$ (blue), respectively.

expression in this subspace with $r_n \approx 0$, are not regulated by the cell cycle at all. Indeed, of the 791 genes that were classified by Spellman et al. (3) as cell cycle regulated, 646 have more than 25% of their normalized expression in this subspace (Fig. 2*b*). Also, 92 of the 103 genes that were classified as cell cycle regulated by traditional methods (including, for example, *CDC8*, which was not classified by Spellman et al. as cell cycle regulated) have more than 25% of their normalized expression in this subspace (Fig. 2*c*). We sort all 6018 genes according to their phases $\{\phi_n\}$, to describe the transition from the expression pattern $|\gamma_2\rangle_N$ to that of $|\gamma_1\rangle_N$ and back to $|\gamma_2\rangle_N$, starting at $\phi_1 \approx -2\pi/13$. One may expect this to order the genes according to the stages in the cell cycle in which their expression patterns peak. However, for the 791 cell cycle regulated genes this sorting gives a classification of the genes into the 5 cell cycle stages, which is somewhat different than the classification by Spellman et al. Similarly, for the 103 cell cycle regulated genes this sorting gives a classification which is somewhat different than the traditional classification. This may be due to the poor quality of the elutriation expression data, as synchronization by elutriation was not very effective in this experiment. For the *CDC15*- and α factor-synchronized yeast cell cycle expression there is much better agreement between the two classifications (see below, Figs. 7 and 11).

With all 6018 genes sorted, the gene variations of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ fit normalized sine and cosine functions of period $Z \equiv N - 1 = 6017$ and initial phase $\theta \approx 2\pi/13$, $-\sqrt{2/Z}\sin(2\pi z/Z - \theta)$ and $-\sqrt{2/Z}\cos(2\pi z/Z - \theta)$, respectively, where $z \equiv n - 1$ (Fig. 3*b* and *c*). The sorted and normalized elutriation expression approximately fits a traveling wave of expression, varying sinusoidally across both genes and arrays, such that the expression of the nth gene in the mth array satisfies $\langle n|\hat{e}_N|m\rangle \propto -2\cos[2\pi(t/T - z/Z)]/\sqrt{ZT}$ (Fig. 3*a*).

BIOLOGICAL DATA ANALYSIS: *CDC15*-SYNCHRONIZED CELL CYCLE

Spellman et al. (3) also monitored genome-wide mRNA levels, for 6113 ORFs simultaneously, over approximately two and a half cell cycle periods, in a yeast culture synchronized by *CDC15*, relative to a reference mRNA from an asynchronous yeast culture, at 10m intervals (except at t = 20, 40, 60, 260, and 280m) for T = 290m. The arrays corresponding to samples from t = 10 to 290m at 20m intervals were all hybridized on one day, while these corresponding to samples from t = 80 to 240m at 20m intervals were all hybridized on another day. Experimental artifacts due to array hybridization appear to vary from one day to another, and therefore these two groups of arrays could be thought of as corresponding to two different experiments. The dataset for these two *CDC15* experiments we analyze (see supplemental dataset and Mathematica notebook at http://genome-www.stanford.edu/SVD/) tabulates the ratios of gene expression levels for the N = 6026 genes, N' = 5611 of which with no missing data in the the M = 24 arrays, and 415 with no missing data in at least 22 of the 24 arrays. Of these genes, 791 were classified by Spellman et al. as cell cycle regulated and 102 were classified as cell cycle regulated by traditional methods.

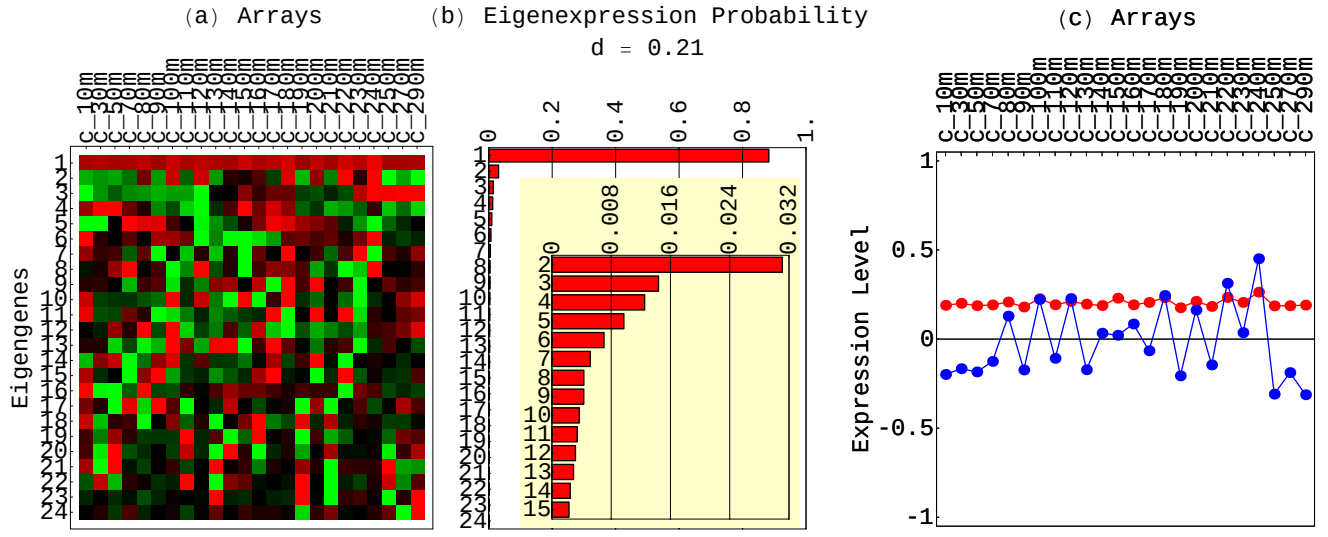

Figure 4. *CDC15* eigengenes. (*a*) Raster display of $\hat{v}^T$, the expression of 24 eigengenes in 24 arrays. (*b*) Bar chart of the probabilities of eigenexpression showing about 90% of the overall relative expression in $|\gamma_1\rangle$, and about 3% in $|\gamma_2\rangle$. (*c*) Line-joined graphs of the expression levels of $|\gamma_1\rangle$ (red) and $|\gamma_2\rangle$ (blue) in the 24 arrays; $|\gamma_1\rangle$ is inferred to represent the steady-state expression and $|\gamma_2\rangle$ is inferred to represent additive artifacts due to array hybridization.

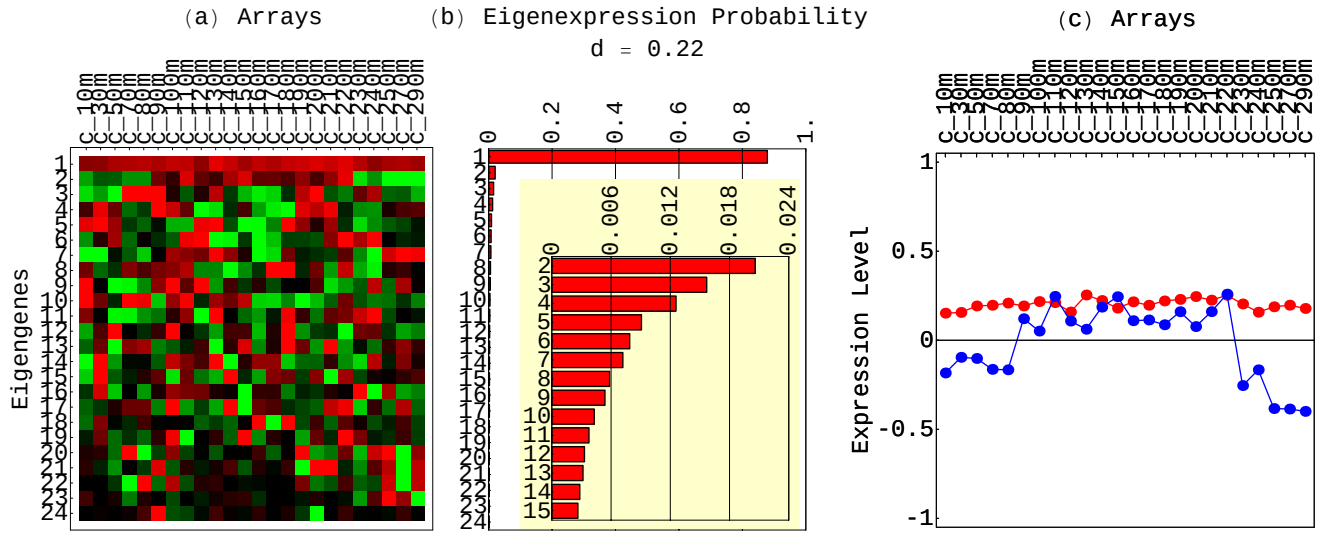

Figure 5. Eigengenes of the natural logarithm of the variances in *CDC15* expression $\hat{e}_{LV}$. (*a*) Raster display of $\hat{v}^T_{LV}$. (*b*) $|\gamma_1\rangle_{LV}$ captures about 90% and $|\gamma_2\rangle_{LV}$ captures about 2% of the overall expression information in this dataset. (*c*) Line-joined graphs of the expression levels of $|\gamma_1\rangle$ (red) and $|\gamma_2\rangle$ (blue); $|\gamma_1\rangle_{LV}$ is inferred to represent the steady scale of expression variance and $|\gamma_2\rangle_{LV}$ is inferred to represent multiplicative artifacts due to array hybridization.

Missing Data Estimation. We use the $L' = 6$ most significant eigengene patterns, as calculated for the subset of 5611 genes with no missing data in the 24 arrays, in order to estimate the missing data in the remaining $N - N' = 415$ genes, 335 of which with no missing data in 23 of the 24 arrays, and 80 with no missing data in 22 of the 24 arrays. We find that these eigenegenes and corresponding probablities of eigenexpression are similar to those calculated for the full dataset of 6026 genes after the missing data are estimated.

Pattern Inference and Data Filtering and Normalization. Consider the 24 eigengenes of the full *CDC15* dataset (Fig. 4*a*). The entropy of this dataset $d = 0.21$ (Fig. 4*b*) is higher than that of the elutriation dataset, because the *CDC15* dataset combines two experiments, and therefore is less redundant. The 1st eigengene $|\gamma_1\rangle$, which captures about 90% of the overall relative expression, descibes time (and experiment) invariant relative expression (Fig. 4*c*). We infer that $|\gamma_1\rangle$ represents the steady-state expression, and filter it out, approximately centering the dataset. The 2nd eigengene $|\gamma_2\rangle$, which captures about 3% of the overall relative expression, describes expression which is approximately invariant within each experiment, but varies between the experiments (Fig. 4*c*). This suggests that time-invariant additive constants due to uncontrolled variables in the process of array hybridization may be superimposed on the data. We infer that $|\gamma_2\rangle$ represents such additive artifacts due to array hybridization, and filter it out. The eigengenes $|\gamma_l\rangle$ for $7 \leq l \leq L = 24$ all show rapidly varying expression during the cell cycle (Fig. 4*a*), and we filter them out as well, such that $\hat{e} \rightarrow \hat{e}_C = \hat{e} - \sum_{l=1}^{2} \varepsilon_l |\alpha_l\rangle\langle\gamma_l| - \sum_{l=7}^{24} \varepsilon_l |\alpha_l\rangle\langle\gamma_l| = \sum_{l=3}^{6} \varepsilon_l |\alpha_l\rangle\langle\gamma_l|$. The remaining eigengenes $|\gamma_l\rangle$ for $3 \leq l \leq 6$ show expression oscillations of about two and a half periods during the cell cycle (Fig. 16 in supplemental material at http://genome-www.stanford.edu/SVD/). We infer that these eigengenes represent cell cycle expression oscillations and assume that the corresponding eigenarrays represent the corresponding cell cycle cellular states.

Let $\hat{e}_{LV}$ tabulate the natural logarithm of the variances in expression in the *CDC15* experiments, such that each element of $\hat{e}_{LV}$ satisfies $\langle n|\hat{e}_{LV}|m\rangle \equiv \log(e^2_{C,nm})$, and consider the eigengenes of $\hat{e}_{LV}$ (Fig. 5*a*). The 1st eigengene $|\gamma_1\rangle_{LV}$, which captures about 90% of the overall information in this dataset (Fig. 5*b*), describes a time (and experiment) invariant scale of expression variance (Fig. 5*c*), and we filter it out. The 2nd eigengene $|\gamma_2\rangle_{LV}$, which captures about 2% of the overall information in this dataset (Fig. 5*b*), describes a scale of expression variance which is approximately invariant within each experiment, but varies between the experiments (Fig. 5*c*). This suggests that time-invariant multiplicative constants due to uncontrolled variables in the process of array hybridization may be superimposed on the data. We infer that $|\gamma_2\rangle$ represents such multiplicative artifacts due to array hybridization, and filter it out, $\hat{e}_{LV} \rightarrow \hat{e}_{CLV} = \hat{e}_{LV} - \sum_{l=1}^{2} \varepsilon_{l,LV} |\alpha_l\rangle_{LV\ LV}\langle\gamma_l|$.

The normalized *CDC15* dataset $\hat{e}_N$, where each of its elements satisfies $\langle n|\hat{e}_N|m\rangle \equiv \text{sign}(e_{CS,nm})\sqrt{\exp(e_{CLV,nm})}$,

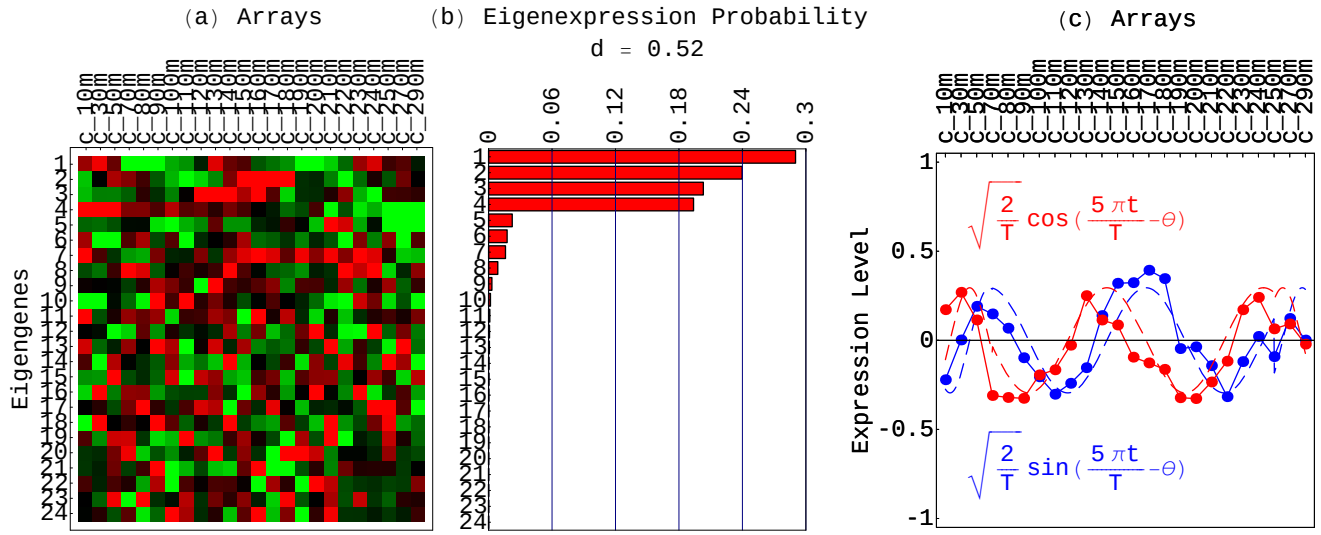

Figure 6. Normalized *CDC15* eigengenes. (*a*) Raster display of $\hat{v}_N^T$, the expression of 24 eigengenes in 24 arrays. (*b*) $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ capture about 30% and 24% of the overall normalized expression, respectively. (*c*) Line-joined graphs of the expression levels of $|\gamma_1\rangle_N$ (red) and $|\gamma_2\rangle_N$ (blue) in the 24 arrays, fit dashed graphs of normalized sine (red) and cosine (blue) of period $2T/5 \approx 120$m and phase $\theta \approx \pi/4$, respectively.

tabulates for each gene and array expression patterns that are approximately centered at the steady-state expression level (i.e., of approximately zero arithmetic means), with variances which are approximately normalized by the steady scale of expression variance (i.e., of approximately unit geometric means). Note that the entropy of $\hat{e}_N$ is $d = 0.52$, suggesting that $\hat{e}_N$ is less redundant (and maybe more informative) than $\hat{e}$. The eigengenes $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ (Fig. 6*a*), which are of similar significance, capture together about 50% of the overall normalized expression (Fig. 6*b*). Their time variations fit normalized sine and cosine functions of two and a half $2T/5 \approx 120$m periods and initial phase $\theta \approx \pi/4$, $\sqrt{2/T}\cos(5\pi t/T - \theta)$ and $\sqrt{2/T}\sin(5\pi t/T - \theta)$, respectively (Fig. 6*c*). We infer that $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ represent cell cycle expression oscillations, and assume that the corresponding eigenarrays $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ represent the corresponding cell cycle cellular states. Upon sorting of the genes (and arrays) according to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ (and $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$), the initial phase $\theta \approx \pi/4$ can be interpreted as a delay of 30m between the start of the experiment and that of the cell cycle stage G1.

Data Sorting. Consider the normalized expression of the 24 *CDC15* arrays $\{|a_m\rangle\}$ in the subspace spanned by $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ (Fig. 7*a*). All arrays have at least 25% of their normalized expression in this subspace, except for $|a_{11}\rangle$, $|a_{22}\rangle$, $|a_{23}\rangle$, and $|a_{24}\rangle$, which measure $t = 140$, 250, 270, and 290m in the cell cycle. This can be interpreted as dephasing in time of the initially synchronized yeast culture. The sorting of the arrays according to their phases $\{\phi_m\}$, which describes the transition from the expression pattern $|\alpha_1\rangle_N$ to $|\alpha_2\rangle_N$ and back to $|\alpha_1\rangle_N$, gives an array order that is similar to that of the cell cycle time points measured by the arrays, an order which describes the progress of the cell cycle expression from the M/G1 stage through G1, S, S/G2, and G2/M and back to M/G1 about two and a half times. Even though the arrays measure equally spaced cell cycle time points, the arrays are not equally spaced, suggesting that the initially synchronized culture dephases in time, or that the cell cycle may not progress linearly in time. Since $|\alpha_1\rangle_N$ is correlated with the arrays $|a_2\rangle$ and $|a_3\rangle$ and is anticorrelated with $|a_5\rangle$ and $|a_6\rangle$, we associate $|\alpha_1\rangle_N$ with the cell cycle cellular state of transition from M/G1 to G1, and $-|\alpha_1\rangle_N$ with the transition from S to S/G2. Similarly, $|\alpha_2\rangle_N$ is correlated with $|a_{12}\rangle$, $|a_{13}\rangle$, and $|a_{14}\rangle$, and therefore we associate $|\alpha_2\rangle_N$ with the transition from G1 to S. Also, $|\alpha_2\rangle_N$ is anticorrelated with $|a_8\rangle$, and $|a_9\rangle$, and therefore we associate $-|\alpha_2\rangle_N$ with the transition from G2/M to M/G1. With these associations the phase of $|a_1\rangle$, $\phi_1 = -\theta \approx -\pi/4$, corresponds to the 30m delay between the start of the experiment and that of the cell cycle stage G1, which is also present in the inferred cell cycle expression oscillations $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$.

Consider also the expression of the 6026 genes $\{|g_n\rangle\}$ in the subspace spanned by $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$. Of the 791 genes that were classified by Spellman et al. (3) as cell cycle regulated, 769 have more than 25% of their normalized expression in this subspace (Fig. 7*b*). Also, 98 of the 102 genes that were classified as cell cycle regulated by traditional

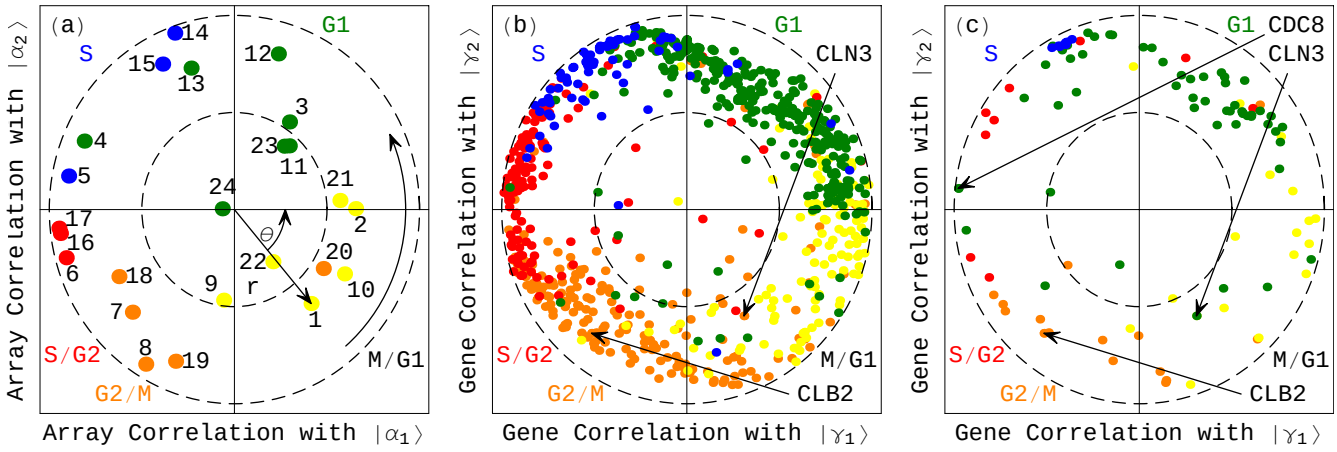

Figure 7. Normalized *CDC15* expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_2\rangle_N$ along the y-axis vs. that with $|\alpha_1\rangle_N$ along the x-axis, color-coded according to the classification of the arrays into the 5 cell cycle stages, M/G1 (yellow), G1 (green), S (blue), S/G2 (red) and G2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_2\rangle_N$ vs. that with $|\gamma_1\rangle_N$, for 791 cell cycle regulated genes, color-coded according to the classification by Spellman et al. (3). (*c*) Correlation of each gene with $|\gamma_2\rangle_N$ vs. that with $|\gamma_1\rangle_N$, for 102 cell cycle regulated genes, color-coded according to the traditional classification (3).

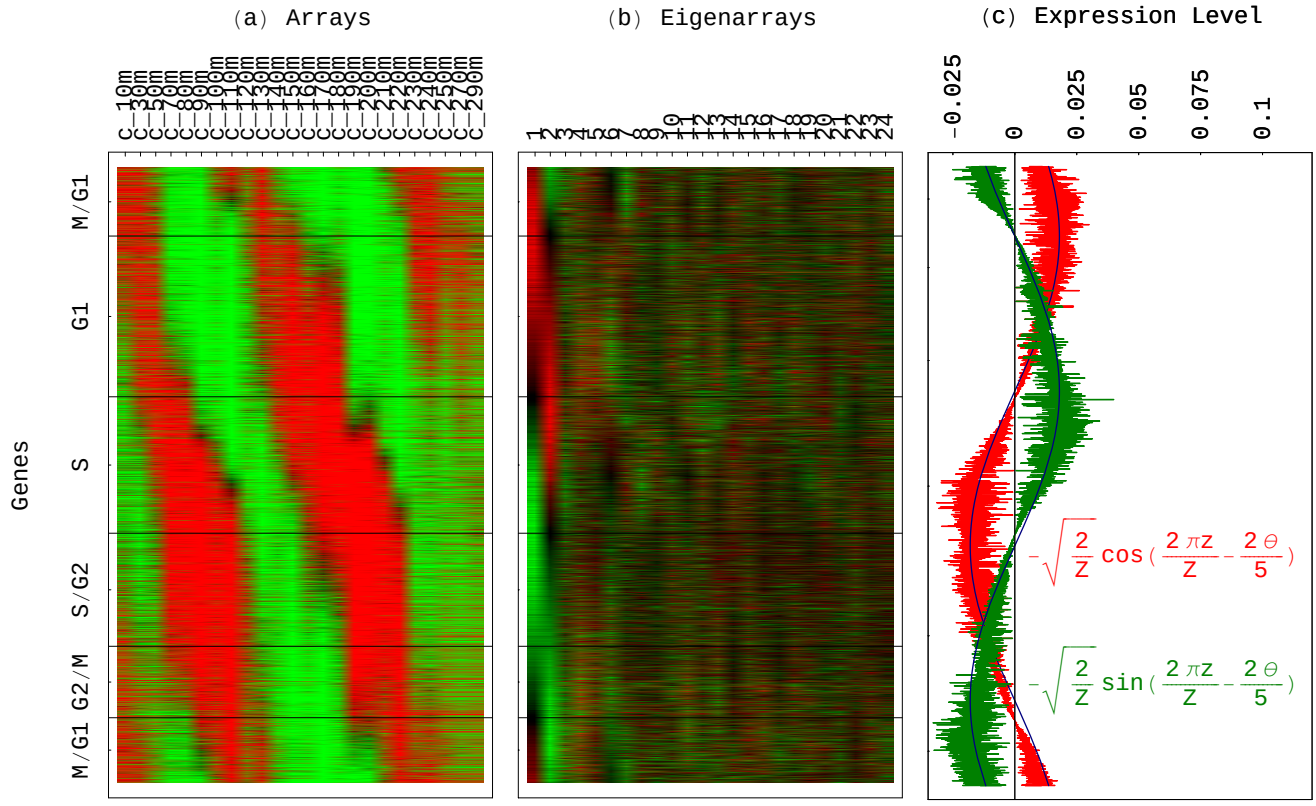

Figure 8. Genes sorted by relative correlation with $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$ of normalized *CDC15*. (*a*) Normalized *CDC15* expression of the sorted 6026 genes in the 24 arrays, showing traveling wave of expression. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$, the eigenarrays corresponding to $|\gamma_1\rangle_N$ and $|\gamma_2\rangle_N$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_N$ (red) and $|\alpha_2\rangle_N$ (green) fit normalized sine and cosine functions of period $Z = N - 1 = 6025$ and phase $2\theta/5 \approx \pi/10$ (blue), respectively.

methods (including, for example, *CDC8*, which was not classified by Spellman et al. as cell cycle regulated) have more than 25% of their normalized expression in this subspace (Fig. 7*c*). We sort all 6026 genes according to their phases $\{\phi_n\}$, to describe the transition from the expression pattern $|\gamma_1\rangle_N$ to that of $|\gamma_2\rangle_N$ and back to $|\gamma_1\rangle_N$, starting at $\phi_1 \approx -\pi/4$, ordering the genes according to the stages in the cell cycle in which their expression patterns peak, and describing the progress of the cell cycle along the genes. For the 791 cell cycle regulated genes this sorting gives a classification of the genes into the 5 cell cycle stages, which is in good agreement with the classification by Spellman et al. Similarly, for the 103 cell cycle regulated genes this sorting gives a classification which is in good agreement with the traditional classification.

With all 6026 genes sorted, the gene variations of $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ fit normalized sine and cosine functions of period $Z \equiv N-1 = 6025$ and initial phase $2\theta/5 \approx \pi/10$, $-\sqrt{2/Z}\sin(2\pi z/Z - 2\theta/5)$ and $-\sqrt{2/Z}\cos(2\pi z/Z - 2\theta/5)$, respectively, where $z \equiv n-1$ (Fig. 8*b* and *c*). The sorted and normalized *CDC15* expression fits approximately a traveling wave of expression, varying sinusoidally across both genes and arrays, such that the expression of the nth gene in the mth array satisfies $\langle n|\hat{e}_N|m\rangle \propto -2\cos(5\pi t/T - 2\pi z/Z - 3\theta/5)/\sqrt{ZT}$ (Fig. 8*a*).

BIOLOGICAL DATA ANALYSIS: α FACTOR-SYNCHRONIZED CELL CYCLE AND *CLB2* AND *CLN3* OVERACTIVATIONS

Spellman et al. (3) also monitored genome-wide mRNA levels, for 6113 yeast ORFs simultaneuosly, over approximately two cell cycle periods, in a yeast culture synchronized by α factor, relative to a reference mRNA from an asynchronous yeast culture, at 7m intervals for 119m. They also measured, in two independent experiments, mRNA levels of yeast strain cultures with overactivated *CLB2*, which encodes a G2/M cyclin, both at $t = 40$m relative to their levels at the start of overactivation at $t = 0$. Two additional independent experiments measured mRNA levels of strain cultures with overactivated *CLN3*, which encodes a G1/S cyclin, at $t = 30$ and 40m relative to their levels at the start of overactivation at $t = 0$. The dataset for the α factor, *CLB2* and *CLN3* experiments we analyze (see supplemental dataset and Mathematica notebook at http://genome-www.stanford.edu/SVD/) tabulates the ratios of gene expression levels for the $N = 6013$ genes, $N' = 5035$ of which with no missing data in the the $M = 22$ arrays, and 978 with no missing data in at least 19 of the 22 arrays. Of these genes, 790 were classified by Spellman et al. as cell cycle regulated and 103 were classified as cell cycle regulated by traditional methods.

Missing Data Estimation. We use the $L' = 11$ most significant eigengene patterns, as calculated for the subset of 5035 genes with no missing data in the 22 arrays (13), in order to estimate the missing data in the remaining $N - N' = 978$ genes, 835 of which with no missing data in 21 of the 22 arrays, 127 with no missing data in 20 of the 22 arrays, and 16 with no missing data in 19 of the 22 arrays. We find that these eigenegenes and corresponding probablities of eigenexpression are similar to those calculated for the full dataset of 6013 genes after the missing data are estimated.

Pattern Inference and Data Filtering and Normalization. Consider the 22 eigengenes of the full α factor, *CLB2*, and *CLN3* dataset (Fig. 17*a* in supplemental material at http://genome-www.stanford.edu/SVD/). The entropy of this dataset $d = 0.29$ (Fig. 17*b*) is higher than those of the elutriation and the *CDC15* datasets, because the α factor, *CLB2*, and *CLN3* dataset combines three experiments, and therefore is less redundant. The 1st eigengene $|\gamma_1\rangle$, which captures about 80% of the overall relative expression, descibes time (and experiment) invariant relative expression (Fig. 17*c*). We infer that $|\gamma_1\rangle$ represents the steady-state expression, and filter it out, approximately centering the dataset. The eigengenes $|\gamma_2\rangle$ and $|\gamma_3\rangle$ describe initial transient increase and decrease, respectively, in relative expression superimposed over time-invariant relative expression during the cell cycle. We infer that $|\gamma_2\rangle$ and $|\gamma_3\rangle$ represent the responses to synchronization by α factor, and filter them out. The eigengenes $|\gamma_{10}\rangle$ and $|\gamma_l\rangle$ for $12 \le l \le L = 22$ all show rapidly varying expression during the cell cycle and steady-state expression in the *CLB2*- and *CLN3*-overactive arrays (Fig. 17*a*), and we filter them out as well, such that $\hat{e} \to \hat{e}_C = \hat{e} - \sum_{l=1}^{3}\varepsilon_l|\alpha_l\rangle\langle\gamma_l| - \varepsilon_{10}|\alpha_{10}\rangle\langle\gamma_{10}| - \sum_{l=12}^{22}\varepsilon_l|\alpha_l\rangle\langle\gamma_l| = \sum_{l=4}^{9}\varepsilon_l|\alpha_l\rangle\langle\gamma_l| + \varepsilon_{11}|\alpha_{11}\rangle\langle\gamma_{11}|$. The remaining eigengenes $|\gamma_l\rangle$ for $4 \le l \le 9$ and $|\gamma_{11}\rangle$ show expression oscillations of two periods during the α factor-synchronized cell cycle, from $t = 7$ to 119m for a duration of $T = 112$m (Fig. 17*d*–*f*), suggesting a delay of 7m from the start of the experiment to the start of the progress of a cell cycle of period $T/2 = 66$m in the initially arrested culture. We infer that these eigengenes represent cell cycle expression oscillations and assume that the corresponding eigenarrays represent the corresponding cell cycle cellular states.

Let $\hat{e}_{LV}$ tabulate the natural logarithm of the variances in expression in the α factor, *CLB2*, and *CLN3* experiments, such that each element of $\hat{e}_{LV}$ satisfies $\langle n|\hat{e}_{LV}|m\rangle \equiv \log(e^2_{C,nm})$, and consider the eigengenes of $\hat{e}_{LV}$

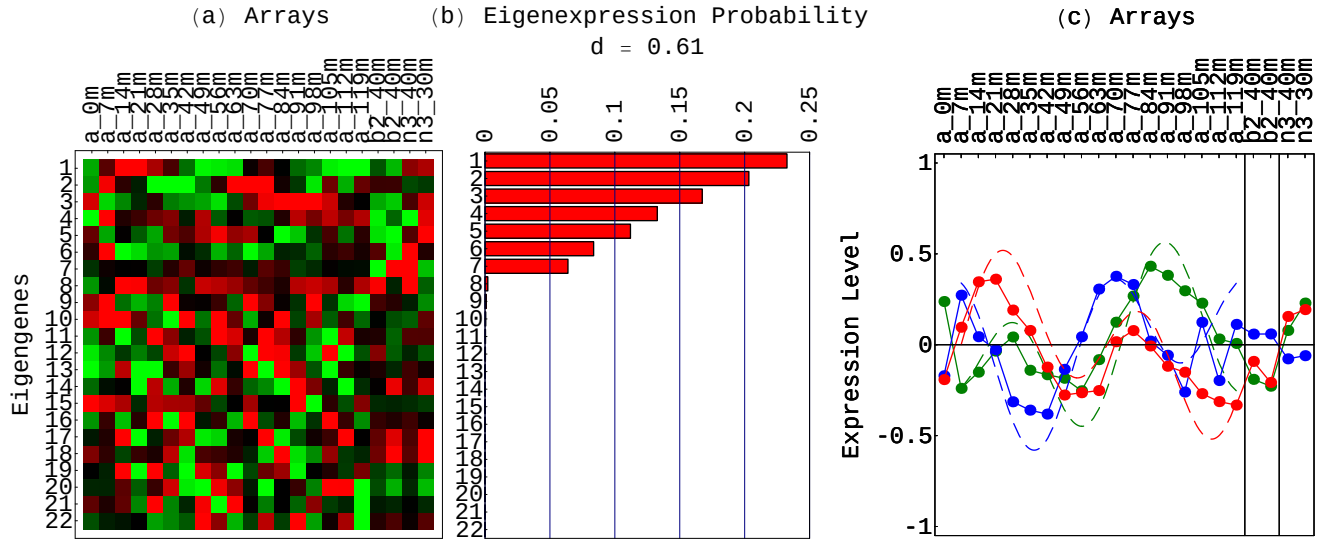

Figure 9. Normalized α factor, *CLB2*, and *CLN3* eigengenes. (a) Raster display of $\hat{v}_N^T$. (b) $|\gamma_1\rangle_N$, $|\gamma_2\rangle_N$, and $|\gamma_3\rangle_N$ capture more than, about and less than 20% of the overall normalized expression, respectively, and span an approximately degenerate subspace. (c) Line-joined graphs of the expression levels of $|\gamma_1\rangle_N$ (red), $|\gamma_2\rangle_N$ (blue), and $|\gamma_3\rangle_N$ (green) fit dashed graphs of periodic functions with period $T/2 = 66$m superimposed on periodic functions with period $T = 112$m from $t = 7$ to $t = 119$m during the cell cycle.

(Fig. 18a in supplemental material at http://genome-www.stanford.edu/SVD/). The 1st eigengene $|\gamma_1\rangle_{LV}$, which captures about 90% of the overall information in this dataset (Fig. 18b), describes a time (and experiment) invariant scale of expression variance, and we filter it out, $\hat{e}_{LV} \to \hat{e}_{CLV} = \hat{e}_{LV} - \varepsilon_{1,LV}|\alpha_1\rangle_{LV\,LV}\langle\gamma_1|$. The normalized α factor, *CLB2*, and *CLN3* dataset $\hat{e}_N$, where each of its elements satisfies $\langle n|\hat{e}_N|m\rangle \equiv \text{sign}(e_{CS,nm})\sqrt{\exp(e_{CLV,nm})}$, tabulates for each gene and array expression patterns that are approximately centered at the steady-state expression level (i.e., of approximately zero arithmetic means), with variances which are approximately normalized by the steady scale of expression variance (i.e., of approximately unit geometric means). Note that the entropy of $\hat{e}_N$ is $d = 0.61$, suggesting that $\hat{e}_N$ is less redundant (and maybe more informative) than $\hat{e}$. The eigengenes $|\gamma_1\rangle_N$, $|\gamma_2\rangle_N$, and $|\gamma_3\rangle_N$ of $\hat{e}_N$ (Fig. 9a), which are of similar significance, capture together about 60% of the overall normalized expression (Fig. 9b). The time variations of $|\gamma_1\rangle_N$, $|\gamma_2\rangle_N$, and $|\gamma_3\rangle_N$ fit normalized sine and cosine functions of two $T/2 = 66$m periods superimposed on normalized sine function of one $T = 112$m period from $t = 7$ to 119m during the cell cycle, $\sqrt{2/T}\sin(4\pi t/T) + \sqrt{1/T}\sin(2\pi t/T)$, $\sqrt{2/T}\cos(4\pi t/T) - \sqrt{1/T}\sin(2\pi t/T)$, and $\sqrt{2/T}\sin(4\pi t/T - \pi/4) - \sqrt{1/T}\sin(2\pi t/T)$, respectively (Fig. 9$c$). While $|\gamma_1\rangle_N$ and $|\gamma_3\rangle_N$ describe underexpression in both *CLB2*-overactive arrays $|a_{19}\rangle$ and $|a_{20}\rangle$, and overexpression in both *CLN3*-overactive arrays $|a_{21}\rangle$ and $|a_{22}\rangle$, $|\gamma_2\rangle$ describes the antiparallel expression pattern of overexpression in $|a_{19}\rangle$ and $|a_{20}\rangle$ and underexpression in $|a_{21}\rangle$ and $|a_{22}\rangle$.

Degenerate Subspace Rotation. We approximate the eigenexpression levels $\varepsilon_{1,N} \approx \varepsilon_{2,N} \approx \varepsilon_{3,N}$ with $\varepsilon_{1,RN} = \varepsilon_{2,RN} = \varepsilon_{3,RN} = \sqrt{(\varepsilon_{1,N}^2 + \varepsilon_{2,N}^2 + \varepsilon_{3,N}^2)/3}$. First, we rotate the eigengenes $|\gamma_1\rangle_N \to \hat{R}_1|\gamma_1\rangle_N = \cos\rho_1|\gamma_1\rangle_N + \sin\rho_1|\gamma_2\rangle_N$, and $|\gamma_2\rangle_N \to \hat{R}_1|\gamma_2\rangle_N = -\sin\rho_1|\gamma_1\rangle_N + \cos\rho_1|\gamma_2\rangle_N$, and corresponding eigenarrays $|\alpha_1\rangle_N \to \hat{R}_1|\alpha_1\rangle_N = \cos\rho_1|\alpha_1\rangle_N + \sin\rho_1|\alpha_2\rangle_N$, and $|\alpha_2\rangle_N \to \hat{R}_1|\alpha_2\rangle_N = -\sin\rho_1|\alpha_1\rangle_N + \cos\rho_1|\alpha_2\rangle_N$. Requiring the rotated 2nd eigengene $\hat{R}_1|\gamma_2\rangle_N$ to describe equal expression in the *CLB2*-overactive array $|a_{20}\rangle$ and the *CLN3*-overactive array $|a_{21}\rangle$, both measured at $t = 40$m after the start of overactivation, such that ${}_N\langle a_{20}|\hat{R}_1|\gamma_2\rangle_N = {}_N\langle a_{21}|\hat{R}_1|\gamma_2\rangle_N$, gives the unique rotation angle $\rho_1 \approx -\pi/8$. Second, we rotate the eigengenes $\hat{R}_1|\gamma_1\rangle_N \to \hat{R}_2\hat{R}_1|\gamma_1\rangle_N = \cos\rho_2\hat{R}_1|\gamma_1\rangle_N + \sin\rho_2|\gamma_3\rangle_N$, and $|\gamma_3\rangle_N \to \hat{R}_2|\gamma_3\rangle_N = \sin\rho_2\hat{R}_1|\gamma_1\rangle_N - \cos\rho_2|\gamma_3\rangle_N$, and corresponding eigenarrays $\hat{R}_1|\alpha_1\rangle_N \to \hat{R}_2\hat{R}_1|\alpha_1\rangle_N = \cos\rho_2\hat{R}_1|\alpha_1\rangle_N + \sin\rho_2|\alpha_3\rangle_N$, $|\alpha_3\rangle_N \to \hat{R}_2|\alpha_3\rangle_N = \sin\rho_2\hat{R}_1|\alpha_1\rangle_N - \cos\rho_2|\alpha_3\rangle_N$. Requiring the rotated 3rd eigengene $\hat{R}_2|\gamma_3\rangle_N$ to describe equal expression in the arrays $|a_{20}\rangle$ and $|a_{21}\rangle$, such that ${}_N\langle a_{20}|\hat{R}_2|\gamma_3\rangle_N = {}_N\langle a_{21}|\hat{R}_2|\gamma_3\rangle_N$, gives the unique rotation angle $\rho_2 \approx \pi/5$.

After these rotations (Fig. 10c), the time variations of $|\gamma_1\rangle_{RN} = \hat{R}_2\hat{R}_1|\gamma_1\rangle_N$ and $|\gamma_2\rangle_{RN} = \hat{R}_1|\gamma_2\rangle_N$ fit normalized

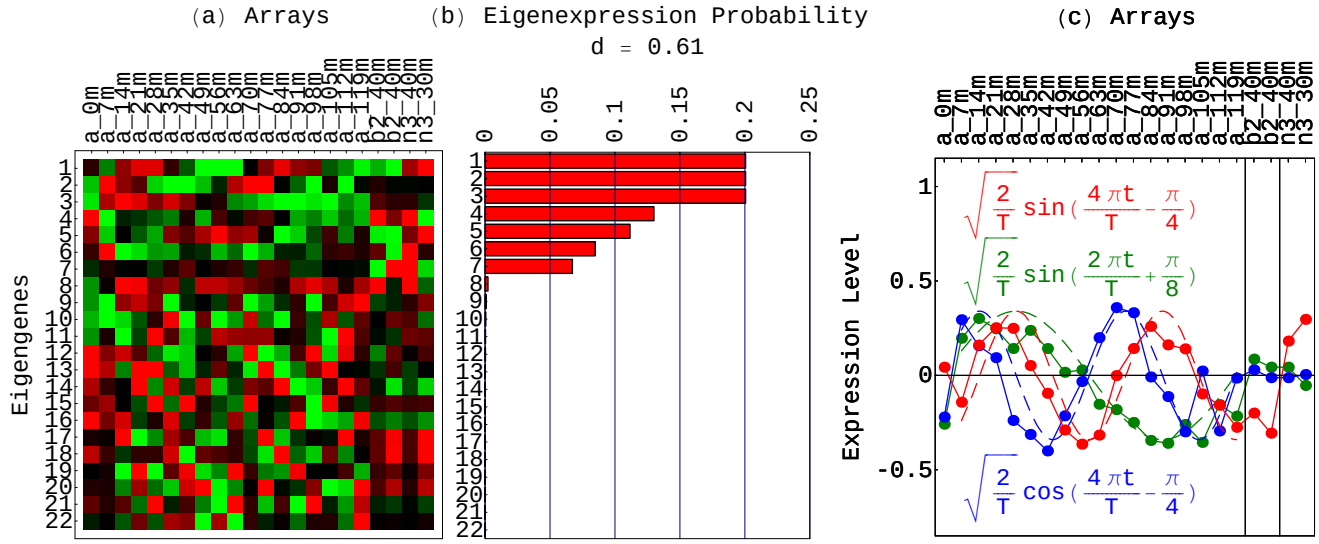

Figure 10. Rotated normalized α factor, *CLB2*, and *CLN3* eigengenes. (a) Raster display of $\hat{v}^T_{RN}$ after rotation, where $|\gamma_1\rangle_{RN} = \hat{R}_2\hat{R}_1|\gamma_1\rangle_N$, $|\gamma_2\rangle_{RN} = \hat{R}_1|\gamma_2\rangle_N$, and $|\gamma_3\rangle_{RN} = \hat{R}_2|\gamma_3\rangle_N$. ($b$) $|\gamma_1\rangle_{RN}$, $|\gamma_2\rangle_{RN}$ and $|\gamma_3\rangle_{RN}$ capture 20% of the overall normalized expression each. (c) Expression levels of $|\gamma_1\rangle_{RN}$ (red) and $|\gamma_2\rangle_{RN}$ (blue) fit dashed graphs of normalized sine (red) and cosine (blue) of period $T/2 = 66$m and phase $\pi/4$, respectively, and $|\gamma_3\rangle_{RN}$ (green) fits dashed graph of normalized sine of period $T = 112$m and phase $-\pi/8$, from $t = 7$ to $t = 119$m during the cell cycle.

sine and cosine functions of two $T/2 = 66$m periods during the cell cycle, from $t = 7$ to 119m, and initial phase $\theta \approx \pi/4$, $\sqrt{2/T}\sin(4\pi t/T - \theta)$, and $\sqrt{2/T}\cos(4\pi t/T - \theta)$, respectively. Upon sorting of the genes (and arrays) according to $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ (and $|\alpha_1\rangle_{RN} = \hat{R}_2\hat{R}_1|\alpha_1\rangle_N$ and $|\alpha_2\rangle_{RN} = \hat{R}_1|\alpha_2\rangle_N$), the initial phase $\theta \approx \pi/4$ can be interpreted as a delay of 7m between the start of the cell cycle progres at $t = 7$m and that of the cell cycle stage G1 at $t = 14$m. The time variation of $|\gamma_3\rangle_{RN} = \hat{R}_2|\gamma_3\rangle_N$ fits normalized sine function of one $T = 112$m period from $t = 7$ to 119m and initial phase $-\theta/2 \approx -\pi/8$, $\sqrt{2/T}\sin(2\pi t/T + \theta/2)$, suggesting differences in expression during the first cell cycle 66m period and the successive 66m period, which may be due to dephasing of the initially synchronized yeast culture. While $|\gamma_2\rangle_{RN}$ and $|\gamma_3\rangle_{RN}$ describe steady-state expression in the *CLB2*- and *CLN3*-overactive arrays, $|\gamma_1\rangle_{RN}$ describes underexpression in the *CLB2*-overactive arrays and overexpression in the *CLN3*-overactive arrays. We, therefore, infer that $|\gamma_1\rangle_{RN}$ represents cell cycle expression oscillations that are *CLB2*- and *CLN3*-dependent, whereas $|\gamma_2\rangle_{RN}$ represents cell cycle expression oscillations that are *CLB2*- and *CLN3*-independent, and $|\gamma_3\rangle_{RN}$ represents variations in the cell cycle expression from the first period to the second, which also appear to be *CLB2*- and *CLN3*-independent. We assume that $|\alpha_1\rangle_{RN}$, $|\alpha_2\rangle_{RN}$, and $|\alpha_3\rangle_{RN}$ represent the corresponding cellular states. With this we also infer that the subspace spanned by $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ approximately represents all cell cycle expression oscillations, and we assume that the subspace spanned by $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ approximately represents all cell cycle cellular states.

Data Sorting. Consider the normalized expression of the 22 α factor, *CLB2*, and *CLN3* arrays $\{|a_m\rangle\}$ in the subspace spanned by $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ (Fig. 11a). All arrays have at least 25% of their normalized expression in this subspace, except for $|a_1\rangle$, $|a_{14}\rangle$, and $|a_{16}\rangle$, which measure $t = 0$, 91, and 105m in the cell cycle, respectively, and $|a_{19}\rangle$, and $|a_{21}\rangle$, which measure *CLB2*- and *CLN3*-overactivation at $t = 40$m after the start of overactivation, respectively. The sorting of the arrays according to their phases $\{\phi_m\}$, which describes the transition from the expression pattern $|\alpha_2\rangle_{RN}$ to $|\alpha_1\rangle_{RN}$ and back to $|\alpha_2\rangle_{RN}$, gives an array order that is similar to that of the cell cycle time points measured by the arrays, an order that describes the progress of the cell cycle expression from the M/G1 stage through G1, S, S/G2, and G2/M and back to M/G1 twice. Even though the arrays measure equally spaced cell cycle time points, the arrays are not equally spaced, suggesting that the initially synchronized culture dephases in time, or that the cell cycle may not progress linearly in time.

Since $|\alpha_1\rangle_{RN}$ is correlated with the arrays $|a_{13}\rangle$ and $|a_{14}\rangle$, as well as with $|a_{21}\rangle$ and $|a_{22}\rangle$, which measure the *CLN3*-overactive samples, we associate $|\alpha_1\rangle_{RN}$ with the cell cycle cellular state of transition from G1 to S, which is simulated by *CLN3* overactivation. Also, $|\alpha_1\rangle_{RN}$ is anticorrelated with the arrays $|a_9\rangle$ and $|a_{10}\rangle$, as well as with

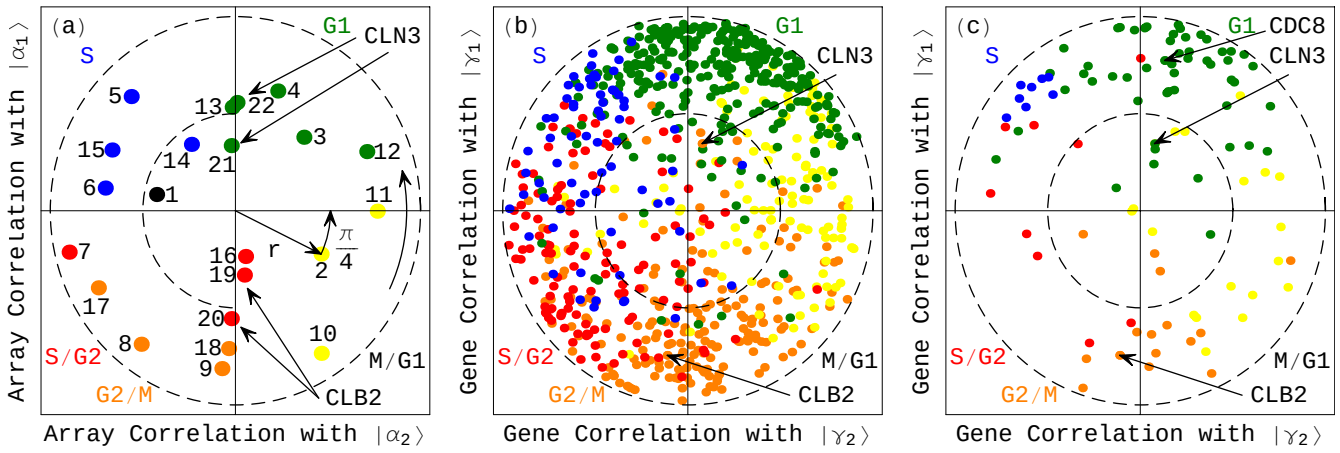

Figure 11. Rotated normalized α factor, *CLB2*, and *CLN3* expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_{RN}$ along the y-axis vs. that with $|\alpha_2\rangle_{RN}$ along the x-axis, color-coded according to the classification of the arrays into the 5 cell cycle stages, M/G1 (yellow), G1 (green), S (blue), S/G2 (red) and G2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 790 cell cycle regulated genes, color-coded according to the classification by Spellman et al. (3). (*c*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 103 cell cycle regulated genes, color-coded according to the traditional classification (3).

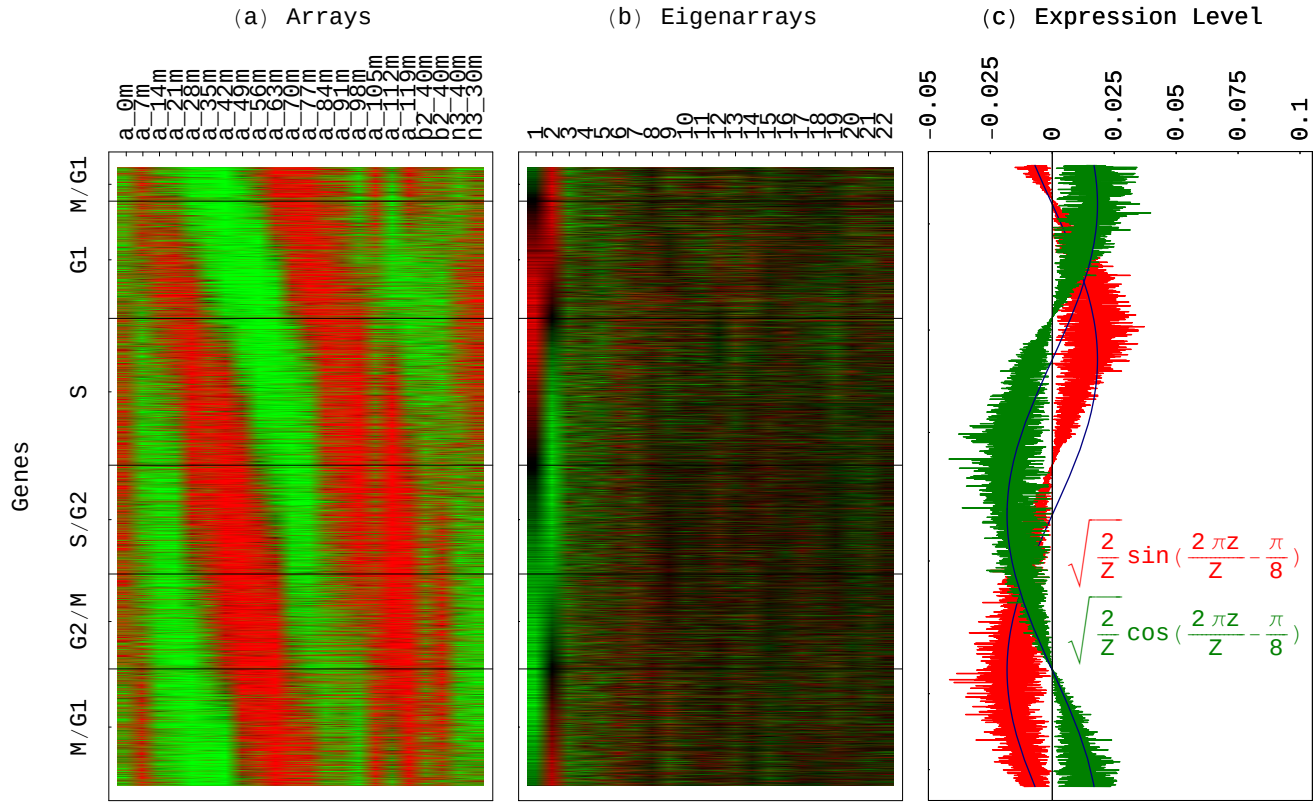

Figure 12. Genes sorted by relative correlation with $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$ of rotated normalized α factor, *CLB2*, and *CLN3*. (*a*) Normalized expression of the sorted 6013 genes in the 22 arrays, showing traveling wave of expression from $t = 0$ to 119m during the cell cycle and standing waves of expression in the *CLB2*- and *CLN3*-overactive arrays. (*b*) Eigenarrays expression; the expression of $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$, the eigenarrays corresponding to $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$, displays the sorting. (*c*) Expression levels of $|\alpha_1\rangle_{RN}$ (red) and $|\alpha_2\rangle_{RN}$ (green) fit normalized sine and cosine functions of period $Z = N - 1 = 6012$ and phase $\pi/8$ (blue), respectively.

$|a_{19}\rangle$ and $|a_{20}\rangle$, which measure the *CLB2*-overactive samples. We associate $-|\alpha_1\rangle_{RN}$ with the cellular transition from G2/M to M/G1, which is simulated by *CLB2* overactivation. Similarly, $|\alpha_2\rangle_{RN}$ appears to be correlated with $|a_2\rangle$, $|a_3\rangle$, $|a_{11}\rangle$, and $|a_{12}\rangle$, anticorrelated with $|a_6\rangle$, $|a_7\rangle$, $|a_{15}\rangle$, and $|a_{17}\rangle$, and uncorrelated with $|a_{19}\rangle$, $|a_{20}\rangle$, $|a_{21}\rangle$, or $|a_{22}\rangle$. We therefore associate $|\alpha_2\rangle_{RN}$ with the cellular transition from M/G1 to G1 (which appears to be *CLB2*- and *CLN3*-independent), and $-|\alpha_2\rangle_{RN}$ with the cellular transition from S to S/G2 (which also appears to be *CLB2*- and *CLN3*-independent). With these associations the phase of $|a_1\rangle$, $\phi_1 \approx -\pi/4$, corresponds to the 7m delay between the start of the progress of the cell cycle at $t = 7$m and the start of the cell cycle stage G1 at $t = 14$m, which is also present in the inferred cell cycle expression oscillations $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$.

Consider also the expression of the 6013 genes in the subspace spanned by $|\gamma_1\rangle_{RN}$ and $|\gamma_2\rangle_{RN}$. Of the 790 genes that were classified by Spellman et al. (3) as cell cycle regulated, 683 have more than 25% of their normalized expression in this subspace (Fig. 11*b*). Also, 88 of the 103 genes that were classified as cell cycle regulated by traditional methods (including, for example, *CDC8*, which was not classified by Spellman et al. as cell cycle regulated) have more than 25% of their normalized expression in this subspace (Fig. 11*c*). We sort all 6013 genes according to their phases $\{\phi_n\}$, to describe the transition from the expression pattern $|\gamma_2\rangle_{RN}$ to that of $|\gamma_1\rangle_{RN}$ and back to $|\gamma_2\rangle_{RN}$, starting at $\phi_1 \approx -\pi/4$, ordering the genes according to the stages in the cell cycle in which their patterns of expression peak, and describing the progress of the cell cycle along the genes. For the 790 cell cycle regulated genes this sorting gives a classification of the genes into the 5 cell cycle stages, which is in good agreement with the classification by Spellman et al. Similarly, for the 103 cell cycle regulated genes this sorting gives a classification of the genes, which is in good agreement with the traditional classification. For example, classification of *CLN3*, which uses the elutriation- and *CDC15*-synchronized yeast cell cycle expression data, suggests that the expression of *CLN3* peaks in the cell cycle stage G2/M (see above, Figs. 2 and 7), and is in agreement with the classification by Spellman et al. However, classification of *CLN3*, which uses the α factor-synchronized yeast cell cycle expression data together with the expression data for *CLB2*- and *CLN3*-overactivation, suggests that the expression of *CLN3* peaks late in the cell cycle stage G1 (Fig. 11). This classification may be in better agreement with the classification by traditional methods, which suggests that *CLN3* peaks early in the cell cycle stage G1.

Since $|\gamma_1\rangle_{RN}$ is correlated with genes that peak late in the cell cycle stage G1 and early in S, among them *CLN3*, we associate $|\gamma_1\rangle_{RN}$ with the cell cycle expression oscillations that start at the transition from G1 to S, and are dependent on *CLN3*, which encodes a G1/S cyclin. Also, $|\gamma_1\rangle_{RN}$ is anticorrelated with genes that peak late in G2/M and early in M/G1, among them *CLB2*, and therefore we associate $-|\gamma_1\rangle_{RN}$ with the oscillations that start at the transition from G2/M to M/G1 and are dependent on *CLB2*, which encodes a G2/M cyclin. Similarly, $|\gamma_2\rangle_{RN}$ is correlated with genes that peak late in M/G1 and early in G1, anticorrelated with genes that peak late in S and early in S/G2, and uncorrelated with *CLB2* and *CLN3*. We, therefore, associate $|\gamma_2\rangle_{RN}$ with the oscillations that start at the transition from M/G1 to G1 (and appear to be *CLB2*- and *CLN3*-independent), and $-|\gamma_2\rangle_{RN}$ with the oscillations that start at the transition from S to S/G2 (and appear to be *CLB2*- and *CLN3*-independent).

With all 6013 genes sorted the gene variations of $|\alpha_1\rangle_{RN}$ and $|\alpha_2\rangle_{RN}$ fit normalized sine and cosine functions of period $Z \equiv N - 1 = 6012$ and initial phase $\pi/8$, $\sqrt{2/Z}\sin(2\pi z/Z - \pi/8)$, and $\sqrt{2/Z}\cos(2\pi z/Z - \pi/8)$, respectively, where $z \equiv n - 1$ (Fig. 12*b* and *c*). The normalized and sorted cell cycle expression approximately fits a traveling wave, varying sinusoidally across both genes and arrays, such that the expression of the nth gene in the mth array satisfies $\langle n|\hat{e}_N|m\rangle \propto 2\cos(4\pi t/T - 2\pi z/Z - \pi/8)/\sqrt{ZT}$. The normalized and sorted expression in the *CLB2*- and *CLN3*-overactive arrays approximately fits standing waves, constant across the arrays and varying sinusoidally across genes only, which appear similar to $-|\alpha_1\rangle_{RN}$ and $|\alpha_1\rangle_{RN}$, respectively (Fig. 12*a*).

CONCLUSIONS

We conclude that SVD provides a useful mathematical framework for processing and modeling genome-wide expression data, in which both the mathematical variables and operations may be assigned biological meaning. Normalizing the data by filtering out the eigengenes (and eigenarrays) that are inferred to represent additive or multiplicative noise, experimental artifacts, or even irrelevant biological processes enables meaningful comparison of the expression of different genes across different arrays in different experiments. Sorting the data according to the eigengenes (and eigenarrays) gives a global picture of the dynamics of gene expression, in which individual genes (and arrays) appear to be classified into groups of similar regulation and function (or similar cellular state and biological phenotype). Upon comparing two or more similar experiments, with a regulator being overactive or underactive in one but normally expressed in the others, one of the significant eigengenes may be correlated with the expression patterns of

this regulator and its targets, while the corresponding eigenarray is correlated with the expression patterns observed in samples in which the regulator is overactive or underactive.

ACKNOWLEDGEMENTS

We thank S. Kim for insightful discussions, G. Sherlock for technical assistance and careful reading, and J. Doyle and P. Green for thoughtful reviews of this work. This work was supported by a grant from the National Cancer Institute (National Institutes of Health, CA77097). O. A. is an Alfred P. Sloan and U. S. Department of Energy Postdoctoral Fellow in Computational Molecular Biology, and a National Human Genome Research Institute Individual Mentored Research Scientist Development Awardee in Genomic Research and Analysis (National Institutes of Health, 1 K01 HG00038-01). P. O. B. is an Associate Investigator of the Howard Hughes Medical Institute.

REFERENCES

1. S. P. Fodor, R. P. Rava, X. C. Huang, A. C. Pease, C. P. Holmes, and C. L. Adams, "Multiplexed biochemical assays with biological chips," *Nature* **364**, pp. 555–556, 1993.
2. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science* **270**, pp. 467–470, 1995.
3. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell* **9**, pp. 3273–3297, 1998.
4. F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nat. Biotechnol.* **16**, pp. 939–945, 1998.
5. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA* **95**, pp. 14863–14868, 1998.
6. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA* **96**, pp. 6745–6750, 1999.
7. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub, "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proc. Natl. Acad. Sci. USA* **96**, pp. 2907–2912, 1999.
8. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat. Genet.* **22**, pp. 281–285, 1999.
9. M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, Jr., and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci. USA* **97**, pp. 262–267, 2000.
10. G. H. Golub, and C. F. Van Loan, *Matrix Computation, 3rd ed.*, Johns Hopkins University Press, Baltimore, 1996.
11. S. G. Mallat, *A Wavelet Tour of Signal Processing, 2nd ed.*, Academic Press, San Diego, 1999.
12. T. W. Anderson, *Introduction to Multivariate Statistical Analysis, 2nd ed.*, Wiley, New York, 1984.
13. O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proc. Natl. Acad. Sci. USA* **97**, pp. 10101–10106, 2000.
14. X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi, "Large-scale temporal gene expression mapping of central nervous system development," *Proc. Natl. Acad. Sci. USA* **95**, pp. 334–339, 1998.
15. S. G. Hilsenbeck, W. E. Friedrichs, R. Schiff, P. O'Connell, R. K. Hansen, C. K. Osborne, S. A. Fuqua, "Statistical analysis of array expression data as applied to the problem of tamoxifen resistance," *J. Natl. Cancer Inst.* **91** pp. 453-459, 1999.
16. S. Raychaudhuri, J. M. Stuart, and R. B. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series," in *Biocomputing 2000*, R. B. Altman, K. Lauderdale, A. K. Dunker, L. Hunter, and T. E. Klein, eds., *Proc. Pac. Symp. Biocomput. 2000*, pp. 455–466, World Scientific, Singapore, 2000.
17. N. S. Holter, M. Mitra, A. Maritan, M. Cieplak, J. R. Banavar, and N. V. Fedoroff, "Fundamental patterns underlying gene expression profiles: simplicity from complexity," *Proc. Natl. Acad. Sci. USA* **97**, pp. 8409–8414, 2000.