APPENDIX

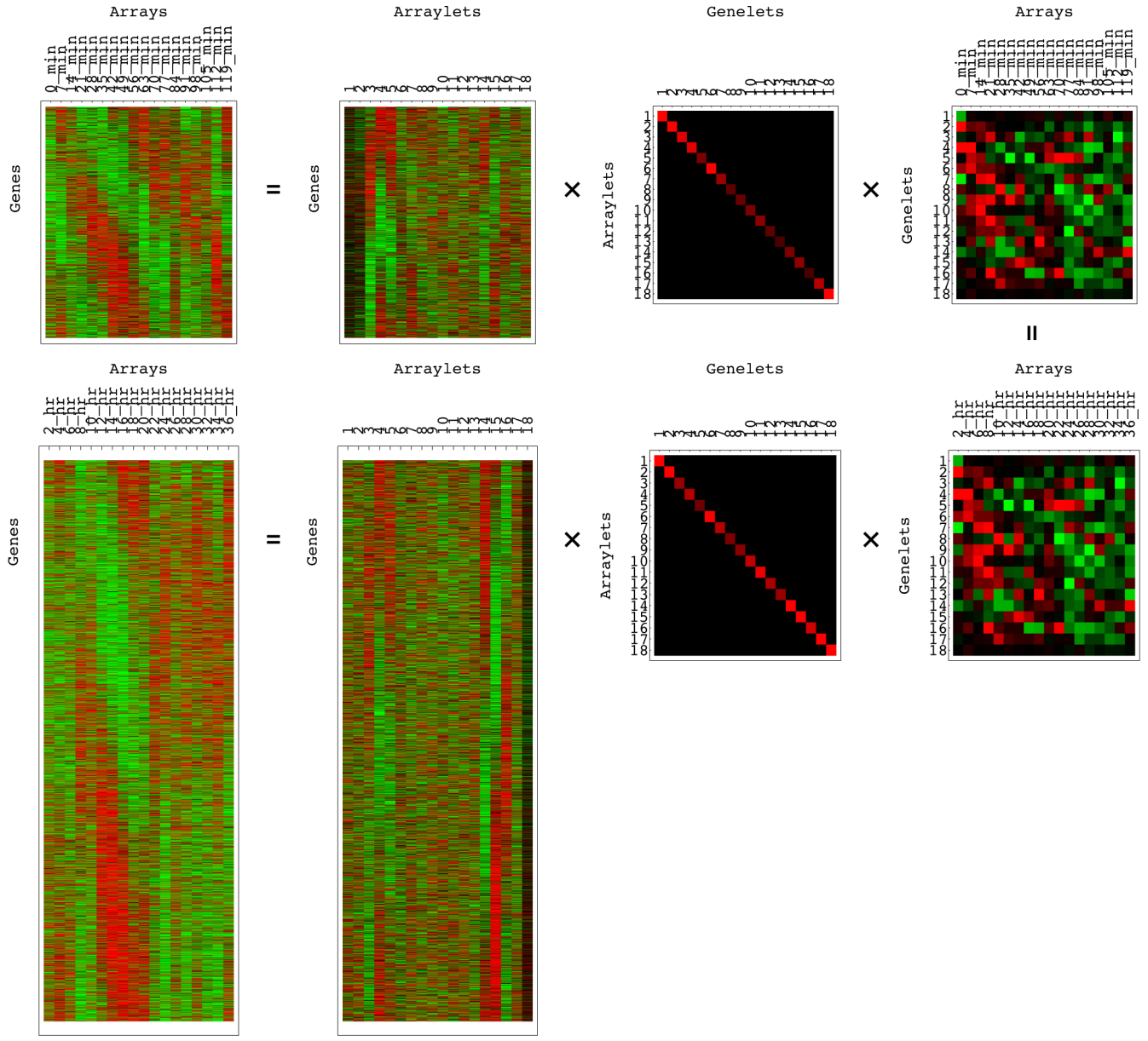

Fig. 5. Generalized singular value decomposition (GSVD) of the yeast and human cell-cycle expression data sets. Shown is a raster display of $\hat{e}_1 = \hat{u}_1\,\hat{\epsilon}_1\,\hat{x}^{-1}$ and $\hat{e}_2 = \hat{u}_2\,\hat{\epsilon}_2\,\hat{x}^{-1}$ with overexpression (red), no change in expression (black), and underexpression (green) centered at gene- and array-invariant expression, showing linear transformation of the yeast and human data from the $4,523$-genes and $12,056$-genes $\times$ 18-arrays spaces to the reduced diagonalized 18-arraylets $\times$ 18-genelets spaces using the $4,523$-genes and $12,056$-genes $\times$ 18-arraylets basis sets and the 18-genelets $\times$ 18-arrays shared basis set.

Significance of the Genelets and the Corresponding Arraylets

The relative significance of the genelet $\langle\gamma_m|$, i.e., the significance of $\langle\gamma_m|$ in the first data set as compared to its significance in the second, is determined by the ratio of the expression information captured by this genelet and its corresponding arraylet $|\alpha_{1,m}\rangle$ in the first data set to the expression information captured by this genelet and its corresponding arraylet $|\alpha_{1,m}\rangle$ in the second data set, $\epsilon_{1,m}/\epsilon_{2,m}$. The two data sets are independent of each other and are uncorrelated. Therefore, their projections onto the genelet $\langle\gamma_m|$ can be thought of as orthogonal vectors of the magnitudes $\epsilon_{1,m}$ and $\epsilon_{2,m}$, respectively,

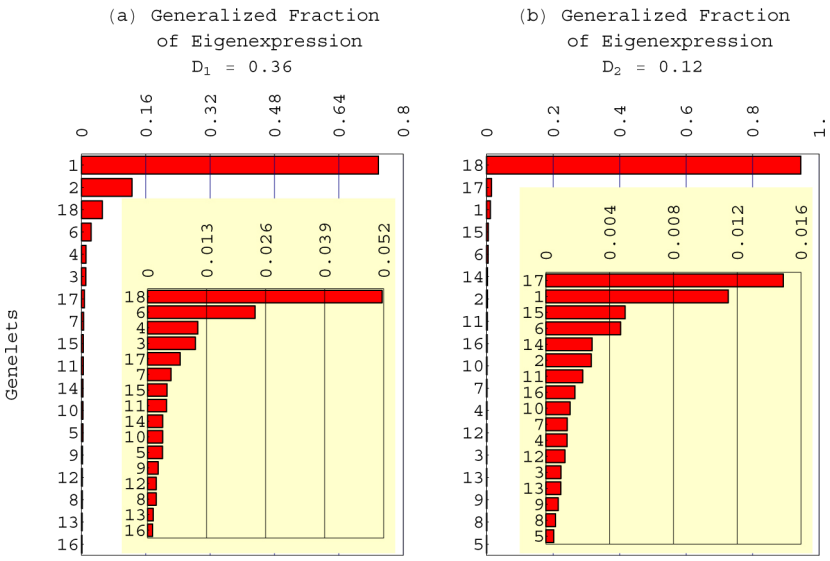

Fig. 6. Bar charts of the generalized fractions of eigenexpression of the genelets. (*a*) Yeast generalized fractions of eigenexpression showing $\langle\gamma_1|$ and $\langle\gamma_2|$ capture ≈70% and 15% of the overall yeast expression, respectively, and a generalized entropy of $D_1 = 0.36$. (*b*) Human generalized fractions of eigenexpression, showing $\langle\gamma_{18}|$ capture >90% of the overall human expression, and a generalized entropy of $D_2 = 0.12$.

where $\vec{\epsilon}_{1,m} \perp \vec{\epsilon}_{2,m}$. The total projection of both data sets onto the genelet $\langle\gamma_m|$ can be thought of as the vector sum of these two orthogonal vectors, $\vec{\epsilon}_m = \vec{\epsilon}_{1,m} + \vec{\epsilon}_{2,m}$. The angle between the total projection $\vec{\epsilon}_m$ and that of the second data set $\vec{\epsilon}_{2,m}$, $\arctan(\epsilon_{1,m}/\epsilon_{2,m})$, measures the extent to which the total projection $\vec{\epsilon}_m$ lies in the direction of the projection of either data set, $\vec{\epsilon}_{1,m}$ or $\vec{\epsilon}_{2,m}$, and quantifies the relative contribution of each data set to this total projection.

The angular distance $\theta_m = \arctan(\epsilon_{1,m}/\epsilon_{2,m}) - \pi/4$ of Eq. **2** is an antisymmetric measure for the relative significance of $\langle\gamma_m|$, such that upon the exchange of the first and second data sets the angular distance changes its sign, i.e., $\theta_m \to -\theta_m = \arctan(\epsilon_{2,m}/\epsilon_{1,m}) - \pi/4$.

The decorrelation of both sets of arraylets, where $\langle\alpha_{i,k}|\alpha_{i,m}\rangle = \delta_{km}$ for all $1 \le k, m \le M$ and $i = 1, 2$, allows determination of the significance of the genelet $\langle\gamma_m|$ relative to all other genelets in each data set separately. In analogy to the "fractions of eigenexpression" in the SVD of each data set (refs. 1–3; see also refs. 4–7 and 8,9), the "generalized fractions of eigenexpression" of each data set,

$$P_{i,m} = \epsilon_{i,m}^2 / \sum_{k=1}^{M} \epsilon_{i,k}^2, \quad \text{where } i = 1, 2, \tag{4}$$

indicate the significance of each genelet, $\langle\gamma_m|$, and its corresponding arraylet in this data set, either $|\alpha_{1,m}\rangle$ or $|\alpha_{2,m}\rangle$, in terms of the fraction of the overall expression information that they capture in this data set alone. Note that the generalized fractions of eigenexpression $P_{1,m}$ and $P_{2,m}$ can be thought of as the probability that a gene of the first or second data set, respectively, expresses the mth genelet, and at the same time also the probability that an array of the first or second data set, respectively, expresses the corresponding mth arraylet of this data set.

The "generalized normalized Shannon entropy" of each data set,

$$0 \le \; D_i = \frac{-1}{\log(M)} \sum_{k=1}^{M} P_{i,k} \log(P_{i,k}) \; \le 1, \quad \text{where } i = 1, 2, \tag{5}$$

measures the complexity of the data from the distribution of the overall expression among the different genelets and corresponding arraylets, where $D_i = 0$ corresponds to an ordered and redundant data set in which all expression is captured by a single genelet and its arraylet in this data set, and $D_i = 1$ corresponds to a disordered and random data set where all genelets and arraylets of this data set are equally expressed.

Significance of the Genelets in the Yeast and Human Data Sets. According to their generalized fractions of eigenexpression (Fig. 6), the genelets $\langle\gamma_1|$, $\langle\gamma_2|$, $\langle\gamma_{18}|$, $\langle\gamma_6|$, $\langle\gamma_4|$, $\langle\gamma_3|$, and $\langle\gamma_{17}|$ are significant in the yeast data set. The genelets $\langle\gamma_{18}|$, $\langle\gamma_{17}|$, $\langle\gamma_1|$, $\langle\gamma_{15}|$, $\langle\gamma_6|$, $\langle\gamma_{14}|$, and $\langle\gamma_2|$ are significant in the human data set. All other genelets are neither significant in the yeast data set, where they capture <1% of the overall yeast expression information each, nor in the human data set, where they capture <0.2% of the overall human expression information each.

According to their angular distances (Fig. 1), the genelets $\langle\gamma_1|$ and $\langle\gamma_2|$ are highly significant in the yeast

data relative to the human data, with $\theta_1, \theta_2 > \pi/7$; $\langle\gamma_3|$ and $\langle\gamma_4|$ are almost equally significant in both data sets with slightly higher significance in the yeast data than in the human data, with $0 < \theta_3, \theta_4 < \pi/16$; $\langle\gamma_6|$ is equally significant in both data sets, with $\theta_6 \sim 0$; $\langle\gamma_{14}|$ and $\langle\gamma_{15}|$ are almost equally significant in both data sets with slightly higher significance in the human data than in the yeast data, with $-\pi/6 < \theta_{14}, \theta_{15} < 0$; and $\langle\gamma_{17}|$ and $\langle\gamma_{18}|$ are highly significant in the human data relative to the yeast data, with $\theta_{17}, \theta_{18} < -\pi/6$.

The genelet $\langle\gamma_5|$ is almost degenerate with the genelets $\langle\gamma_4|$ and $\langle\gamma_6|$, where the corresponding angular distances are similar, $\theta_4 \sim \theta_5 \sim \theta_6$. The genelet $\langle\gamma_{16}|$ is almost degenerate with the genelets $\langle\gamma_{15}|$ and $\langle\gamma_{17}|$, where $\theta_{15} \sim \theta_{16} \sim \theta_{17}$. Therefore, of the 18 genelets and their two sets of 18 corresponding arraylets, we consider the genelets $\langle\gamma_1|$ and $\langle\gamma_2|$ and their corresponding arraylets in the yeast data set, $|\alpha_{1,1}\rangle$ and $|\alpha_{1,2}\rangle$, the genelets $\langle\gamma_3|$, $\langle\gamma_4|$, $\langle\gamma_5|$, $\langle\gamma_6|$, $\langle\gamma_{14}|$, $\langle\gamma_{15}|$, and $\langle\gamma_{16}|$ and their corresponding arraylets in both the yeast and the human data sets, and the genelets $\langle\gamma_{17}|$ and $\langle\gamma_{18}|$ and their corresponding arraylets in the human data set, $|\alpha_{2,17}\rangle$ and $|\alpha_{2,18}\rangle$.

Probabilistic Significance of the Associations of the Genelets and the Corresponding Arraylets with Biological Processes

Sorting the genes of either data set $\hat{e}_i$, where $i = 1, 2$, according to their coefficients of the genelet $\langle\gamma_m|$ in the GSVD expansion, as these coefficients are listed in the corresponding arraylet of this data set $|\alpha_{i,m}\rangle$, allows an examination of the annotations of the group of genes with largest positive coefficients (i.e., largest contributions from $\langle\gamma_m|$ that are parallel to this genelet) and the group with the largest negative coefficients (i.e., largest contributions from $\langle\gamma_m|$ that are antiparallel to this genelet). A coherent biological theme may be reflected in the annotations of either one of these two groups of genes. We associate the genelet $\langle\gamma_m|$ with the regulatory program or biological process that corresponds to these biological themes. We also assume that the corresponding arraylet $|\alpha_{i,m}\rangle$ represents the cellular state of this biological process in the data set $\hat{e}_i$.

To evaluate the significance of the association of the genelet $\langle\gamma_m|$ with this biological process, we estimate the probability that such a coherent biological theme may be reflected in the annotations of a random unordered group of genes, which is selected from the group of all genes in either data set without replacements (10). Consider a data set of N_i genes, where $K_i \leq N_i$ of them are annotated to be of a particular molecular function or biological process, and where $k_i \leq K_i$ of these genes are in the group of the $n_i \leq N_i$ genes with the largest either positive or negative coefficients of the genelet $\langle\gamma_m|$. The probability that a random group of n_i genes that are selected without replacements from the group of N_i genes includes at least k_i genes out of the group of K_i of the given annotation, i.e., the P-value of a given association by annotation, that is calculated assuming hypergeometric probability distribution of the k_i and K_i groups of annotations among the n_i and N_i groups of genes, is

$$P(k_i; n_i, N_i, K_i) = \binom{N_i}{n_i}^{-1} \sum_{k=k_i}^{n_i} \binom{K_i}{k} \binom{N_i - K_i}{n_i - k}, \quad \text{where } i = 1, 2. \tag{6}$$

The binomial coefficient

$$\binom{N_i}{n_i} \equiv \frac{N_i!}{n_i!(N_i - n_i)!}, \quad \text{where } i = 1, 2, \tag{7}$$

counts the number of distinguishable possibilities for selecting without replacements an unordered group of n_i genes from a group of N_i genes. Similarly, the binomial coefficient $\binom{K_i}{k}$ counts the number of distinguishable possibilities for selecting without replacements an unordered group of k genes of a given annotation from the group of K_i genes of this annotation in the data set, and $\binom{N_i - K_i}{n_i - k}$ counts the number of distinguishable possibilities for selecting without replacements the remaining unordered group of $n_i - k$ genes of any different annotation from the group of $N_i - K_i$ genes of any different annotation in the data set. The lower is the probability for a random occurance of the given coherent biological theme in the annotations of genes with the largest contributions from the genelet $\langle\gamma_m|$, the higher is the probability that this genelet represents the corresponding biological process, and the higher is the probabilistic significance of the association of the genelet with this biological process.

Probabilistic Significance of the Association of the Genelets with the Yeast and Human Cell Cycles. Sorting the six yeast and six human arraylets that are associated with the yeast and human cell-cycle cellular states, respectively (see Data Sets 9 and 10, which are published as supporting information on the PNAS web site, pnas.org, and also at https://alterlab.org/GSVD/), the groups of genes with largest either positive or negative coefficients of the corresponding genelets (i.e., largest either parallel or antiparallel contributions from these genelets), appear to include approximately $n_i = 100$ genes each, where $i = 1, 2$ (Fig. 7).

For each group of 100 yeast genes, we count the number of cell cycle-regulated genes that peak in each one of the yeast cell-cycle stages, M/G_1, G_1, S, S/G_2, and G_2/M, and the number of genes that are not cell cycle-regulated according to the microarray classification of Spellman *et al.* (ref. 11; see Data Set 2) and, separately, also the traditional classification (see Data Set 3) and calculate the corresponding P-values. For each group of 100 human genes, we count the number of cell cycle-regulated genes

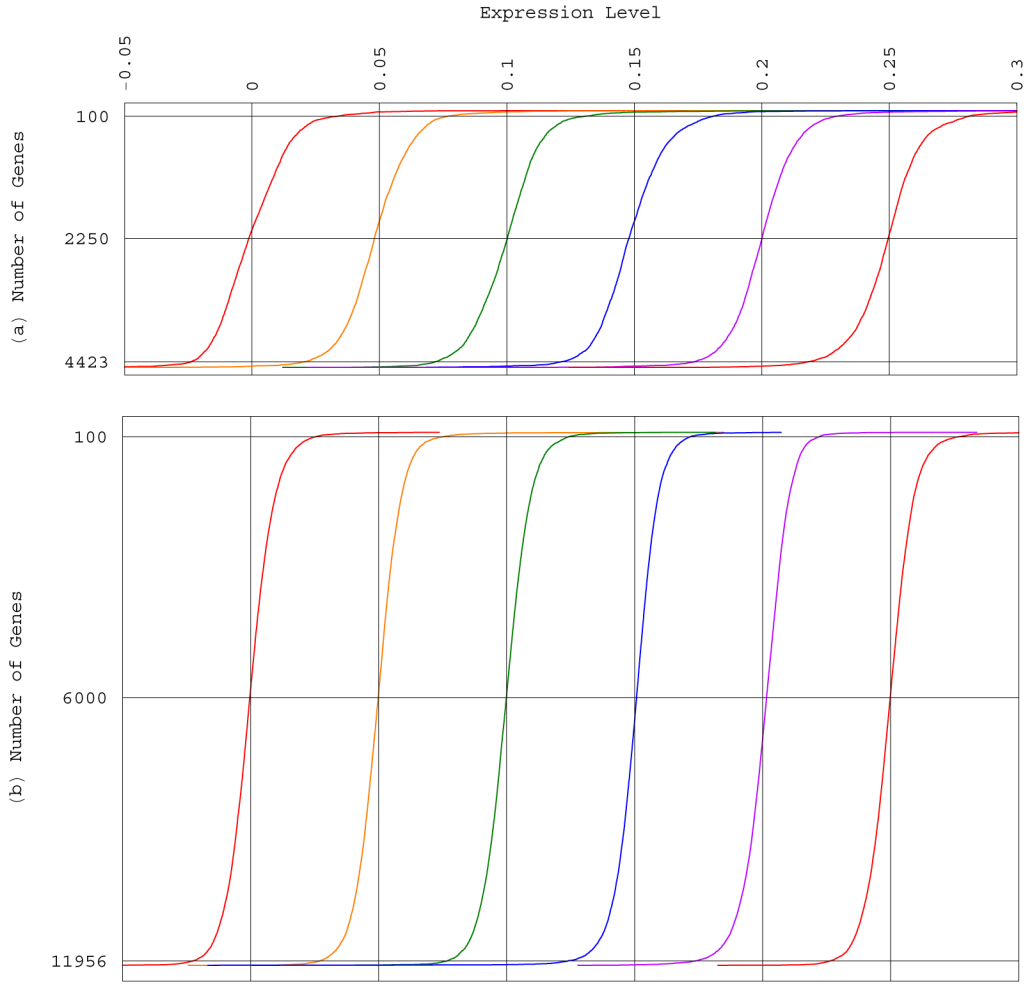

Fig. 7. Sorted yeast and human cell-cycle arraylets. (*a*) Yeast cell-cycle arraylets $|\alpha_{1,3}\rangle$ (red), $|\alpha_{1,4}\rangle$ (orange), $|\alpha_{1,5}\rangle$ (green), $|\alpha_{1,14}\rangle$ (blue), $|\alpha_{1,15}\rangle$ (violet), and $|\alpha_{1,16}\rangle$ (red), each sorted according to their expression levels, include $\approx$100 genes with largest parallel or antiparallel contributions from the corresponding cell-cycle genelets, $\langle\gamma_3|$, $\langle\gamma_4|$, $\langle\gamma_5|$, $\langle\gamma_{14}|$, $\langle\gamma_{15}|$, and $\langle\gamma_{16}|$. (*b*) Human cell-cycle arraylets $|\alpha_{2,3}\rangle$ (red), $|\alpha_{2,4}\rangle$ (orange), $|\alpha_{2,5}\rangle$ (green), $|\alpha_{2,14}\rangle$ (blue), $|\alpha_{2,15}\rangle$ (violet), and $|\alpha_{2,16}\rangle$ (red), each sorted according to their expression levels, also include $\approx$100 genes with largest parallel or antiparallel contributions from the corresponding cell-cycle genelets.

that peak in each one of the human cell-cycle stages, S, G_2, G_2/M, M/G_1, and G_1/S, and the number of genes that are not cell cycle-regulated according to the microarray classification of Whitfield *et al.* (ref. 12; see Data Set 6) and, separately, also the traditional classification (see Data Set 7) and calculate the corresponding *P*-values. The cell-cycle stage with the smallest *P*-value, as calculated for the 100 genes with largest either positive or negative contributions from a given genelet, is then the most likely either parallel or antiparallel, respectively, association for this genelet and its corresponding arraylet (Table 1).

The probabilistic significance of these associations by annotations, estimated using combinatorics, is high: For each genelet and corresponding arraylet in each data set, at least one of the four *P*-values, following either the microarray or the traditional classification, for either parallel or antiparallel association, is smaller than 0.01. Most of the *P*-values are orders of magnitude smaller than 0.01. Almost all parallel and antiparallel associations of each genelet and its corresponding arraylet in either data set are consistently half of a cell-cycle period apart. For example, following the microarray classifications, the genelet $\langle\gamma_{15}|$ is associated in parallel with the yeast cell-cycle stage S/G_2 and the human cell-cycle stage G_1/S; $\langle\gamma_{15}|$ is also associated in antiparallel with the yeast and human stages that are half of a cell-cycle period apart, respectively, i.e., the yeast cell-cycle stage G_1 and the human cell-cycle stage G_2. Almost all yeast and human associations following the microarray classifications by Spellman *et al.* and Whitfield *et al.*, respectively, are consistent with the associations following the traditional

	Classification	Data set	Genelet and arraylet	Most likely parallel association	P-value of parallel association	Most likely antiparallel association	P-value of antiparallel association
a	Yeast	Microarray	3	G_1	2.1×10^{-49}	G_2/M	1.6×10^{-18}
			4	G_2/M	2.9×10^{-15}	G_1	1.1×10^{-36}
			5	M/G_1	1.3×10^{-36}	S/G_2	7.2×10^{-8}
			14	G_2/M	8.8×10^{-8}	G_1	2.6×10^{-13}
			15	S/G_2	5.9×10^{-7}	G_1	3.3×10^{-14}
			16	M/G_1	6.6×10^{-9}	S	7.5×10^{-3}
b		Traditional	3	G_1	1.7×10^{-12}	G_2/M	1.9×10^{-4}
			4	M/G_1	8.2×10^{-6}	G_1	2.6×10^{-22}
			5	M/G_1	1.2×10^{-10}	S	5.4×10^{-4}
			14	G_2/M	1.9×10^{-4}	G_1	2.2×10^{-8}
			15	G_2/M	3.2×10^{-3}	G_1	5.4×10^{-14}
			16	M/G_1	2.6×10^{-7}	S	1.5×10^{-5}
c	Human	Microarray	3	G_2/M	5.6×10^{-3}	G_1/S	7.9×10^{-2}
			4	S	8.2×10^{-21}	M/G_1	1.3×10^{-3}
			5	G_2/M	6.9×10^{-8}	G_1/S	6.4×10^{-5}
			14	G_2	3.3×10^{-34}	M/G_1	1.3×10^{-3}
			15	G_1/S	4.0×10^{-37}	G_2	1.9×10^{-37}
			16	G_2/M	2.0×10^{-33}	G_1/S	3.0×10^{-10}
d		Traditional	3	G_2/M	1.0×10^{-1}	S	3.2×10^{-3}
			4	S	1.5×10^{-8}	G_1/S	7.6×10^{-3}
			5	G_2	4.9×10^{-2}	G_1/S	1.2×10^{-1}
			14	G_2	6.6×10^{-8}	None	5.4×10^{-1}
			15	G_1/S	2.1×10^{-13}	G_2/M	2.1×10^{-14}
			16	G_2/M	9.0×10^{-17}	S	1.1×10^{-5}

Table 1. Parallel and antiparallel associations of genelets and corresponding arraylets, in the yeast data set according to the microarray classification of Spellman *et al.* (*a*) and the traditional classification (*b*) and in the human data set according to the microarray classification of Whitfield *et al.* (*c*) and the traditional classification (*d*).

classifications. For example, the genelet $\langle\gamma_3|$ is associated in parallel with the yeast cell-cycle stage G_1 and the human cell-cycle stage G_2/M, following both the microarray and traditional classifications.

Most of these associations are in agreement with the expression patterns of the genelets (Fig. 2), taking into account the initial synchronization of the yeast data in the cell-cycle stage M/G_1 and that of the human data in S. For example, following the traditional classifications, the 0-phase genelet $\langle\gamma_4|$ is associated in parallel with the yeast cell-cycle stage M/G_1, and both 0-phase genelets $\langle\gamma_4|$ and and $-\langle\gamma_{16}|$ are associated in parallel with the human cell-cycle stage S. Most of these associations are also in agreement with the simultaneous comparative classification of the yeast and human cell-cycle regulated genes (Fig. 3). These simultaneous associations therefore outline a correspondence between the groups of yeast genes and those of human genes. For example, yeast genes that peak at M/G_1 correspond to human genes that peak at S, the cell-cycle stages in which the yeast and human cultures are synchronized initially, respectively.

Missing Data Estimation Using SVD

The relative expression level of the nth gene in the mth sample as measured by the mth array in either the yeast or the human data set is nullified and presumed missing when it is deemed invalid, e.g., when the ratio of the measured expression signal to the measured background signal of either the synchronized culture or the asynchronous reference culture is >1.5 for the yeast data or 3.5 for the human data. We estimate these missing data in each data set by using SVD (2).

SVD is linear transformation of each expression data set from the N_i-genes $\times$ M-arrays space to the reduced L_i-"eigenarrays" $\times$ L_i-"eigengenes" space, where $L_i = \min\{M, N_i\}$,

$$\hat{e}_i \;=\; \hat{u}_i\,\hat{\epsilon}_i\,\hat{v}_i^T, \quad \text{where } i = 1, 2. \tag{8}$$

In this space the data set $\hat{e}_i$ is represented by the diagonal nonnegative matrix $\hat{\epsilon}_i$, of size L_i-eigengenes $\times$ L_i-eigenarrays, which satisfies $\langle k|\hat{\epsilon}_i|l\rangle \equiv \epsilon_{i,l}\delta_{kl} \geq 0$ for all $1 \leq k, l \leq L_i$, such that the lth eigengene is expressed only in the corresponding lth eigenarray, with the corresponding "eigenexpression" level $\epsilon_{i,l}$. Therefore, the

expression of each eigengene and eigenarray is decoupled from that of all other eigengenes or eigenarrays, respectively. The "fraction of eigenexpression"

$$p_{i,l} = \epsilon_{i,l}^2 / \sum_{k=1}^{L_i} \epsilon_{i,k}^2, \quad \text{where } i = 1, 2, \tag{9}$$

indicates the relative significance of the lth eigengene and eigenarray in terms of the fraction of the overall expression that they capture in the data set $\hat{e}_i$. Note that the fractions of eigenexpression $p_{i,l}$ can be thought of as the probability that a gene of the $\hat{e}_i$ data set expresses the lth eigengene of this data set and at the same time also the probability that an array of this data set expresses the corresponding lth arraylet. Assume also that the eigenexpression levels are arranged in decreasing order of significance such that $\epsilon_{i,1} \geq \epsilon_{i,2} \geq \ldots \geq \epsilon_{i,L_i} \geq 0$. The "normalized Shannon entropy" of each data set,

$$0 \leq \ d_i = \frac{-1}{\log(L_i)} \sum_{k=1}^{L_i} p_{i,k} \log(p_{i,k}) \ \leq 1, \quad \text{where } i = 1, 2, \tag{10}$$

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

The transformation matrices $\hat{u}_i$ and $\hat{v}_i^T$ define the N_i-genes $\times$ L_i-eigenarrays and the L_i-eigengenes $\times$ M-arrays basis sets, respectively. The vector in the lth row of each matrix $\hat{v}_i^T$ lists the expression of the lth eigengene of the data set $\hat{e}_i$ across the different arrays. The vector in the lth column of each matrix $\hat{u}_i$ lists the genome-wide expression in the lth eigenarray of the data set $\hat{e}_i$. The eigengenes and eigenarrays are orthonormal superpositions of the genes and arrays such that the transformation matrices $\hat{u}_i$ and $\hat{v}_i$ are both orthogonal

$$\hat{u}_i^T \hat{u}_i = \hat{v}_i^T \hat{v}_i = \hat{I}, \quad \text{where } i = 1, 2, \tag{11}$$

and where $\hat{I}$ is the identity matrix. Therefore, the expression of each eigengene and eigenarray is not only decoupled but also decorrelated from that of all other eigengenes or eigenarrays, respectively. The eigengenes and eigenarrays are unique except in degenerate subspaces, defined by subsets of equal eigenexpression levels, and except for a phase factor of ± 1, because each eigengene and eigenarray captures both parallel and antiparallel gene- or array-expression patterns, respectively. Therefore, SVD is data-driven except in degenerate subspaces.

SVD Calculation. According to Eqs. **8** and **11**, the M-arrays $\times$ M-arrays symmetric correlation matrix of each data set $\hat{a}_i = \hat{e}_i^T \hat{e}_i = \hat{v}_i \hat{\epsilon}_i^2 \hat{v}_i^T$ is represented in the L_i-eigengenes $\times$ L_i-eigengenes space by the diagonal matrix $\hat{\epsilon}_i^2$. The N_i-genes $\times$ N_i-genes symmetric correlation matrix of each data set $\hat{g}_i = \hat{e}_i \hat{e}_i^T = \hat{u}_i \hat{\epsilon}_i^2 \hat{u}_i^T$ is represented in the L_i-eigenarrays $\times$ L_i-eigenarrays space also by $\hat{\epsilon}_i^2$, where for $L_i = \min\{M, N_i\} = M$, $\hat{g}_i$ has a null subspace of at least $N_i - M$ null eigenvalues. In theory it is possible to calculate the SVD of each data set $\hat{e}_i$, with $M < N_i$, by diagonalizing $\hat{a}_i$ and then projecting the resulting $\hat{v}_i$ and $\hat{\epsilon}_i$ onto $\hat{e}_i$ to obtain $\hat{u}_i = \hat{e}_i \hat{v}_i \hat{\epsilon}_i^{-1}$. In practice we avoid computing the correlation matrices $\hat{a}_i$ and use the numerically robust SVD algorithm (8).

Calculation of Missing Data Estimates. The distribution of expression information among the eigengenes and eigenarrays of an expression data set, such that several 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 0. The inference that these significant eigengenes and eigenarrays represent independent processes or cellular states, respectively (refs. 1–3; see also refs. 4–7), suggests the possibility of using the expression patterns of these eigengenes for meaningful estimation of missing data.

In each data set, for the nth gene $|g_{i,n}\rangle$, with missing data in $M_i' < M$ of the arrays, we estimate the missing expression level in the mth array $\langle m|g_{i,n}\rangle$ to be a superposition of the expression levels of the $L_i' < M - M_i'$ significant eigengenes $\{|\gamma_{i,l}'\rangle\}$ in the mth array as calculated for the subset of $N_i' < N_i$ 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 $M - M_i'$ arrays with no missing data, $|g_{i,n}\rangle_{M_i'}$, in the subspace spanned by the expression patterns of the significant eigengenes across the same $M - M_i'$ arrays, $\{|\gamma_{i,l}'\rangle_{M_i'}\}$, such that $\langle m|g_{i,n}\rangle \rightarrow \sum_{l=1}^{L_i'} \langle m|\gamma_{i,l}'\rangle \ {}_{M_i'}\langle \beta_{i,l}'|g_{i,n}\rangle_{M_i'}$, where ${}_{M_i'}\langle \beta_{i,k}'|\gamma_{i,l}'\rangle_{M_i'} = \delta_{kl}$, i.e., $\{ {}_{M_i'}\langle \beta_{i,l}'|\}$ span the $L_i' \times (M - M_i')$ subspace $(\hat{v}_{M_i'}')^\dagger$ that is pseudoinverse to the $(M - M_i') \times L_i'$ subspace $\hat{v}_{M_i'}'$, which is spanned by $\{|\gamma_{i,l}'\rangle_{M_i'}\}$. We use the SVD of $\hat{v}_{M_i'}' \equiv \hat{U}_i \hat{\omega}_i \hat{V}_i^T$ to calculate the pseudoinverse $(\hat{v}_{M_i'}')^\dagger = \hat{V}_i \hat{\omega}_i^{-1} \hat{U}_i^T$, which according to Eq. **11** satisfies $(\hat{v}_{M_i'}')^\dagger \hat{v}_{M_i'}' = \hat{I}$, where $\hat{V}_i \hat{V}_i^T = \hat{I}$ for all $L_i' < M - M_i'$. Assuming that the L_i' 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 data set of N_i genes after the missing data are estimated.

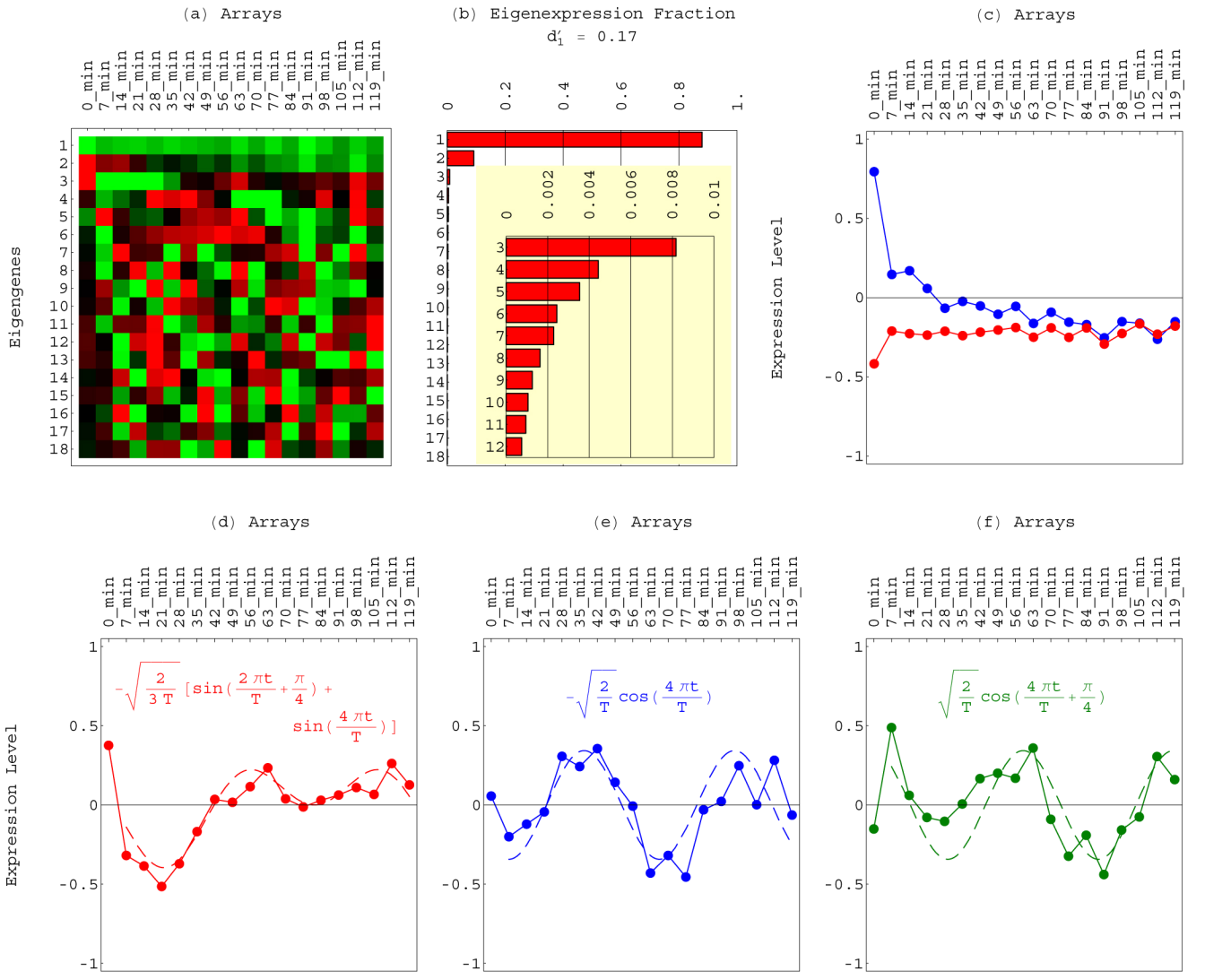

Fig. 8. Yeast eigengenes as calculated for the $N'_1 = 2,698$ yeast genes with no missing data in the 18 arrays. (*a*) Raster display of $(\hat{v}'_1)^T$, the expression of 18 eigengenes in 18 arrays with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene. (*b*) Bar chart of the fractions of eigenexpressions showing ≈90% of the overall relative expression in the first eigengene, 1% in the second, 0.8% in the third, and 0.4% in both the fourth and fifth eigengenes, and a low entropy of $d'_1 = 0.17 \ll 1$. (*c*) Line-joined graphs of the expression levels of the first (red) and second (blue) eigengenes in the 18 arrays. (*d*) Line-joined graph of the expression levels of the third eigengene (red) fits a dashed graph of a sine of two periods superimposed on a sine of one period (red). (*e* and *f*) Line-joined graphs of the expression levels of the fourth (blue) and fifth (green) eigengenes fit dashed graphs of normalized cosines of two periods and initial phases of 0 (blue) and $\pi/4$ (green).

Estimation of the Missing Yeast Data. The data set for the yeast experiments we analyze (see Data Set 1) tabulates the ratios of gene-expression levels for the $N_1 = 4,523$ genes, $N'_1 = 2,698$ of which with no missing data in the $M_1 = 18$ arrays and 1,825 with no missing data in at least 15 of the 18 arrays. We use the $L'_1 = 5$ most significant eigengene patterns as calculated for the subset of 2,698 genes with no missing data in the 18 arrays in order to estimate the missing data in the remaining 1,825 genes, with no missing data in 15 of the 18 arrays. We find that these eigenegenes and corresponding fractions of eigenexpression (Fig. 8) are similar to those calculated for the full data set of 4,523 genes after the missing data are estimated (Fig. 9), suggesting that the five most significant eigengenes, as calculated for the 2,698 genes with no missing data, are meaningful patterns for estimating the missing data in the yeast data set.

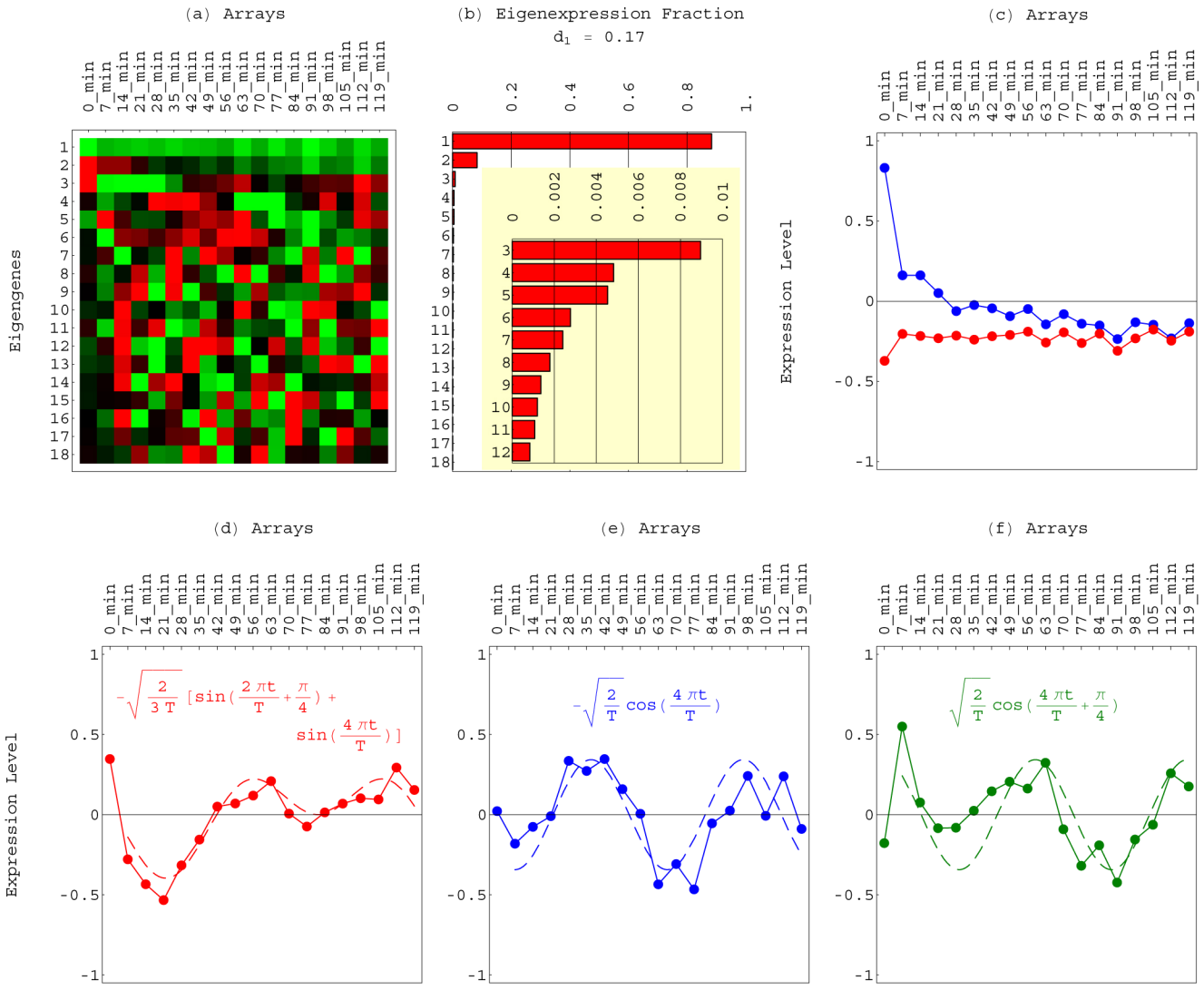

Fig. 9. Yeast eigengenes as calculated for the $N_1 = 4,523$ yeast genes after missing data estimation. (*a*) Raster display of $\hat{v}_1^T$, the expression of 18 eigengenes in 18 arrays with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene. (*b*) Bar chart of the fractions of eigenexpressions showing ≈90% of the overall relative expression in the first eigengene, 1% in the second, 0.9% in the third, and 0.5% in both the fourth and fifth eigengenes, and a low entropy of $d_1 = 0.17 \ll 1$. (*c*) Line-joined graphs of the expression levels of the first (red) and second (blue) eigengenes in the 18 arrays. (*d*) Line-joined graph of the expression levels of the third eigengene (red) fits a dashed graph of a sine of two periods superimposed on a sine of one period (red). (*e* and *f*) Line-joined graphs of the expression levels of the fourth (blue) and fifth (green) eigengenes fit dashed graphs of normalized cosines of two periods and initial phases of 0 (blue) and $\pi/4$ (green).

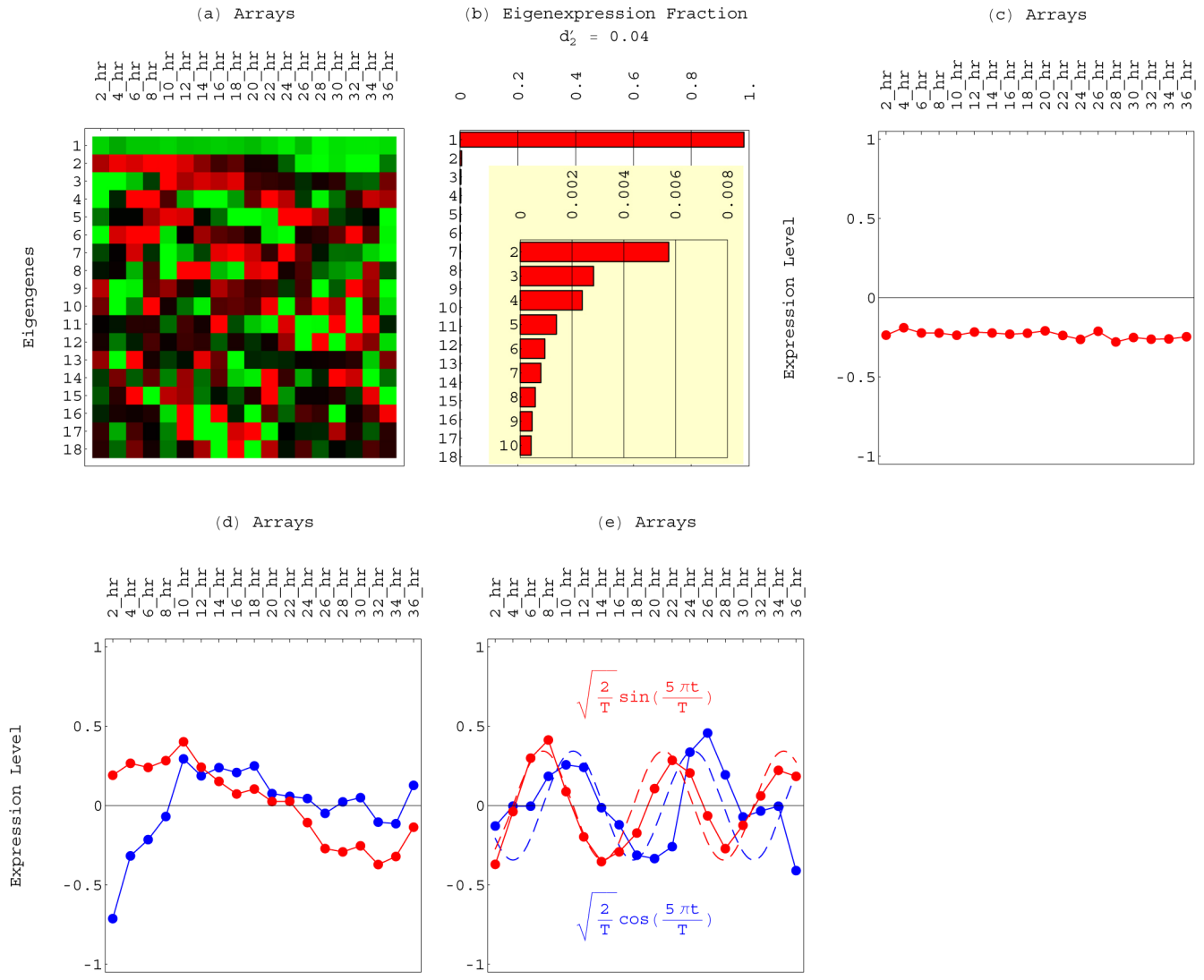

Fig. 10. Human eigengenes as calculated for the $N_2' = 4,360$ human genes with no missing data in the 18 arrays. (*a*) Raster display of $(\hat{v}_2')^T$, the expression of 18 eigengenes in 18 arrays with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene. (*b*) Bar chart of the fractions of eigenexpressions showing >95% of the overall relative expression in the first eigengene, ≈0.6% in the second, 0.3% in the third and fourth, and 0.1% in the fifth eigengene, and a very low entropy of $d_2' = 0.04 \ll 1$. (*c*) Line-joined graph of the expression levels of the first eigengene (red) in the 18 arrays. (*d*) Line-joined graphs of the expression levels of the second (red) and third (blue) eigengenes. (*e*) Line-joined graphs of the expression levels of the fourth (red) and fifth (blue) eigengenes fit dashed graphs of normalized sine (red) and cosine (blue) of two and a half periods.

Estimation of the Missing Human Data. The data set for the human experiments we analyze (see Data Set 5) tabulates the ratios of gene expression levels for the $N_2 = 12,056$ genes, $N_2' = 4,360$ of which with no missing data in the $M_2 = 18$ arrays and 7,696 with no missing data in at least 15 of the 18 arrays. We use the $L_2' = 5$ most significant eigengene patterns as calculated for the subset of 4,360 genes with no missing data in the 18 arrays in order to estimate the missing data in the remaining 7,696 genes, with no missing data in 15 of the 18 arrays. We find that these eigenegenes and corresponding fractions of eigenexpression (Fig. 10) are similar to those calculated for the full data set of 12,056 genes after the missing data are estimated (Fig. 11), suggesting that the five most significant eigengenes as calculated for the 4,360 genes with no missing data are meaningful patterns for estimating the missing data in the human data set.

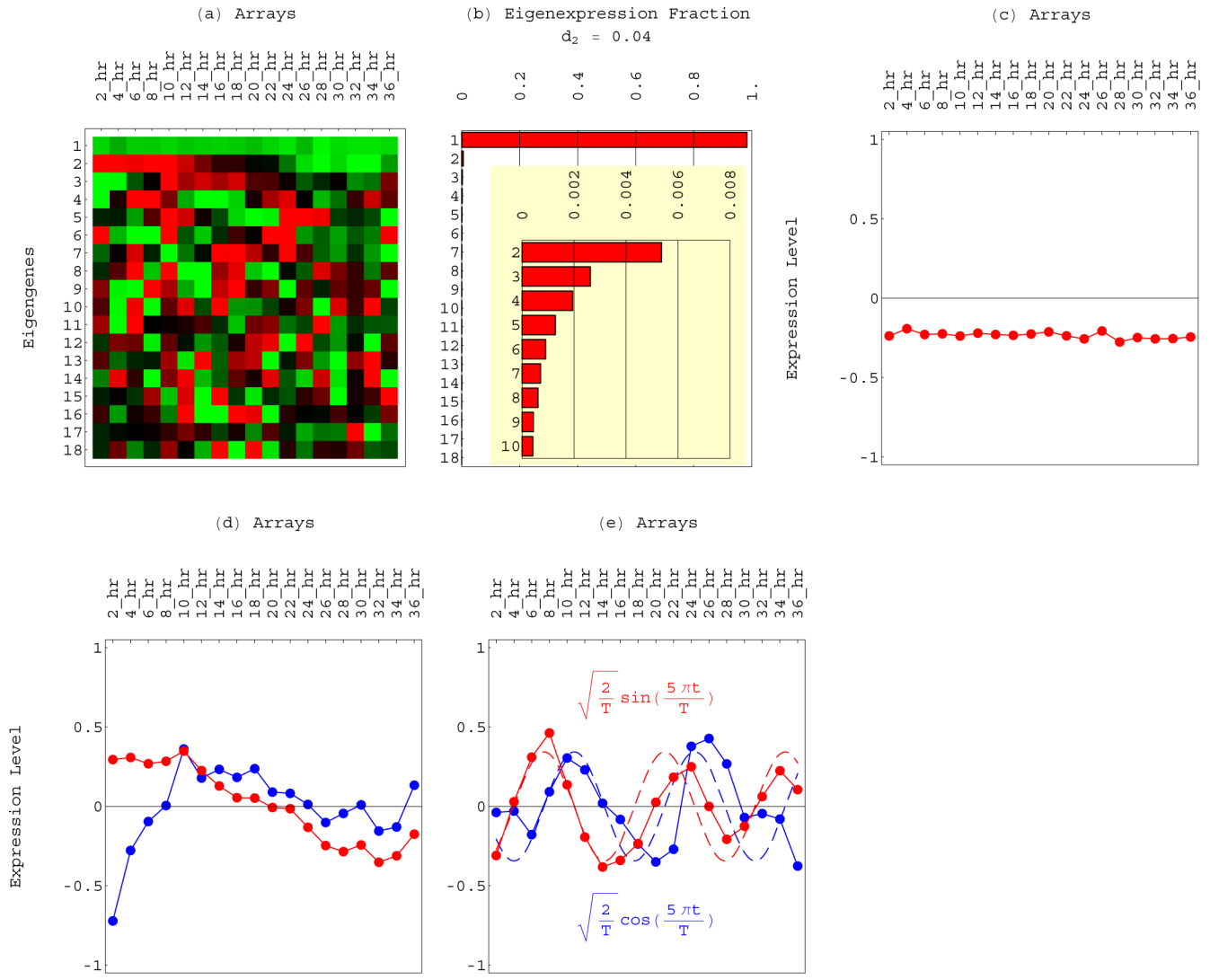

Fig. 11. Human eigengenes as calculated for the $N_2 = 12,056$ human genes after missing data estimation. (*a*) Raster display of $\hat{v}_2^T$, the expression of 18 eigengenes in 18 arrays, with overexpression (red), no change in expression (black), and underexpression (green) around the steady state, which can be associated with the first eigengene. (*b*) Bar chart of the fractions of eigenexpressions showing >95% of the overall relative expression in the first eigengene, ≈0.5% in the second, 0.3% in the third, 0.2% in the fourth, and 0.1% in the fifth eigengene, and a very low entropy of $d_2 = 0.04 \ll 1$. (*c*) Line-joined graph of the expression levels of the first eigengene (red) in the 18 arrays. (*d*) Line-joined graphs of the expression levels of the second (red) and third (blue) eigengenes. (*e*) Line-joined graphs of the expression levels of the fourth (red) and fifth (blue) eigengenes fit dashed graphs of normalized sine (red) and cosine (blue) of two and a half periods.

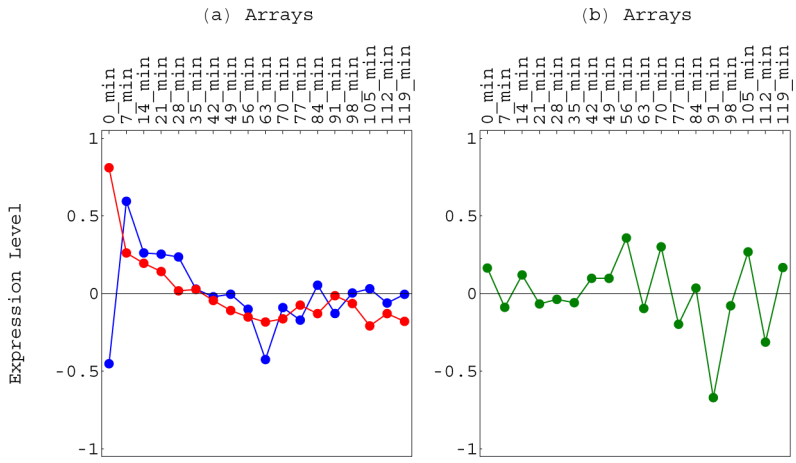

Fig. 12. Yeast pheromone-response subspace projected from the three-dimensional genelets subspace onto two-dimensional least-squares-approximated subspace by using SVD. (*a*) Line-joined graphs of the expression levels of the two orthonormal vectors $|x\rangle$ (red) and $|y\rangle$ (blue), which least-squares-approximate the genelets that are inferred to span the exclusive yeast pheromone response $\langle\gamma_1|$, $\langle\gamma_2|$ and $\langle\gamma_6|$. (*b*) Line-joined graph of the expression levels of the third orthonormal vector $|z\rangle$ (green).

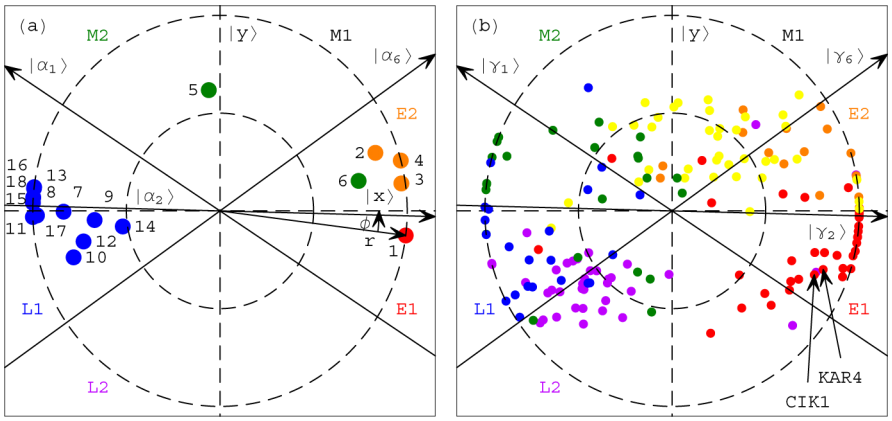

Fig. 13. Yeast expression reconstructed in the three-dimensional pheromone-response subspaces approximated by two-dimensional subspaces. (*a*) Yeast array expression projected onto $|y\rangle$ along the y-axis vs. that onto $|x\rangle$ along the x-axis and color-coded according to classification of the arrays into six stages in the pheromone response and transition into the cell cycle: early E_1 (red) and E_2 (orange), middle M_1 (yellow) and M_2 (green), and late stages in the time course L_1 (blue) and L_2 (violet). The dashed unit and half-unit circles outline 100% and 50% of added up (rather than cancelled out) contributions of the three arraylets to the overall projected expression. The arrows describe the projections of the arraylets $|\alpha_{1,1}\rangle$, $|\alpha_{1,2}\rangle$, and $|\alpha_{1,6}\rangle$. (*b*) Yeast gene expression of 172 (see Data Set 4) genes projected onto $|y\rangle$ along the y-axis vs. that onto $|x\rangle$ along the x-axis and color-coded according to the traditional understanding of this program (13,15): Genes that peak in E_1 are known to be involved in α-factor response, mating, adaptaion to mating signal, and cell-cycle arrest; E_2, filamentous and pseudohyphal growths and cell polarity, M_1, ATP synthesis; M_2, chromatin modeling; L_1, chromatin binding and architecture; and L_2, phosphate and iron transport.

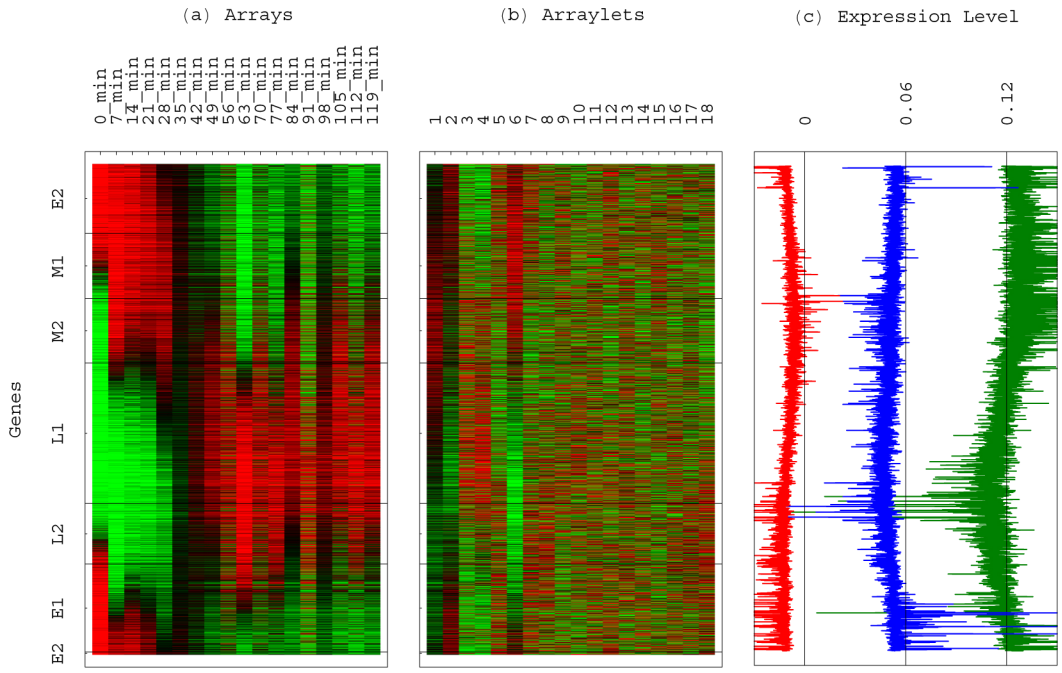

Fig. 14. Yeast expression reconstructed in the three-dimensional pheromone-response subspace with genes sorted according to their phases in the two-dimensional subspace that approximate them. (*a*) Yeast expression of the sorted 4,523 genes in the 18 arrays centered at their gene- and array-invariant levels. (*b*) Yeast expression of the sorted 4,523 genes in the 18 arraylets centered at their array-invariant levels. The expression of the arraylets $|\alpha_{1,1}\rangle$, $|\alpha_{1,2}\rangle$, and $|\alpha_{1,6}\rangle$ displays the sorting. (*c*) Yeast pheromone-response arraylets expression levels $|\alpha_{1,1}\rangle$ (red), $|\alpha_{1,2}\rangle$ (blue), and $|\alpha_{1,6}\rangle$ (green).

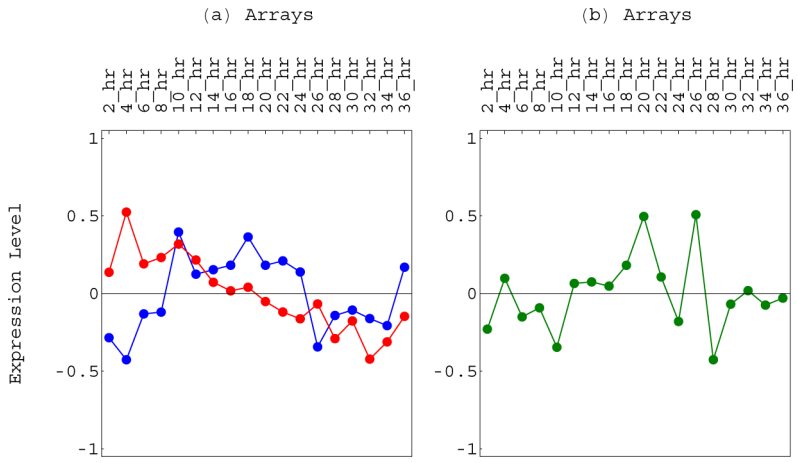

Fig. 15. Human synchronization stress-response subspace projected from the three-dimensional genelets subspace onto two-dimensional least-squares-approximated subspace by using SVD. (*a*) Line-joined graphs of the expression levels of the two orthonormal vectors $|x\rangle$ (red) and $|y\rangle$ (blue), which least-squares-approximate the genelets that are inferred to span the exclusive human synchronization stress response $\langle\gamma_6|$, $\langle\gamma_{17}|$, and $\langle\gamma_{18}|$. (*b*) Line-joined graph of the expression levels of the third orthonormal vector $|z\rangle$ (green).

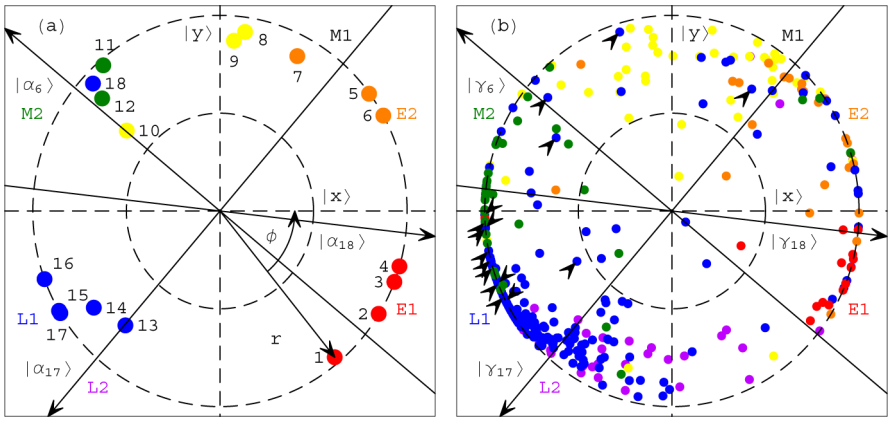

Fig. 16. Human expression reconstructed in the three-dimensional synchronization stress-response subspaces approximated by two-dimensional subspaces. (*a*) Human array expression projected onto $|y\rangle$ along the y-axis vs. that onto $|x\rangle$ along the x-axis and color-coded according to classification of the arrays into six stages in the stress response and transition into the cell cycle: early E_1 (red) and E_2 (orange), middle M_1 (yellow) and M_2 (green), and late stages in the time course L_1 (blue) and L_2 (violet). The dashed unit and half-unit circles outline 100% and 50% of added up (rather than cancelled out) contributions of the three arraylets to the overall projected expression. The arrows describe the projections of the arraylets $|\alpha_{2,6}\rangle$, $|\alpha_{2,17}\rangle$, and $|\alpha_{2,18}\rangle$. (*b*) Human gene expression of 348 genes (see Data Set 8) projected onto $|y\rangle$ along the y-axis vs. that onto $|x\rangle$ along the x-axis and color-coded according to the current understanding of this program (12,14): Genes that peak in E_1 include genes that are known to be involved in serum response; E_2, phosphastes and ubiquilins; M_1, interleukins, integrins, and nexins; M_2, MCMs; L_1, ribosomal proteins, histones, and tubulins; and L_2, cytochromes, golgins, and NADHs. The arrows point to 16 human histones that were not classified by Whitfield *et al.* as cell-cycle regulated based on their overall expression.

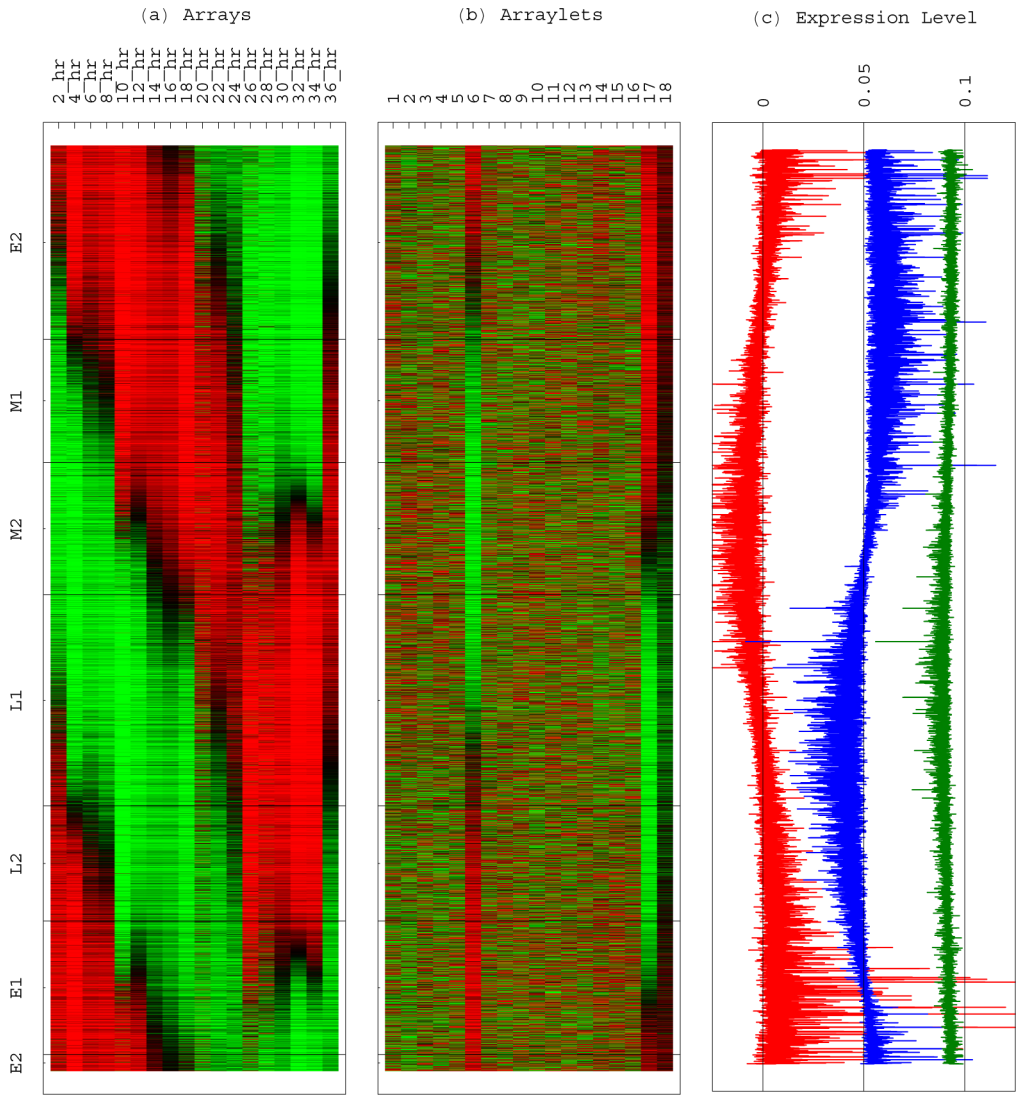

Fig. 17. Human expression reconstructed in the three-dimensional synchronization stress-response subspace with genes sorted according to their phases in the two-dimensional subspace that approximate them. (*a*) Human expression of the sorted 12,056 genes in the 18 arrays centered at their gene- and array-invariant levels. (*b*) Human expression of the sorted 12,056 genes in the 18 arraylets centered at their array-invariant levels. The expression of the arraylets $|\alpha_{2,6}\rangle$, $|\alpha_{2,17}\rangle$, and $|\alpha_{2,18}\rangle$ displays the sorting. (*c*) Human synchronization stress-response arraylets expression levels: $|\alpha_{2,6}\rangle$ (red), $|\alpha_{2,17}\rangle$ (blue), and $|\alpha_{2,18}\rangle$ (green).

[1] Alter, O., Brown, P. O. & Botstein, D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 10101–10106.
[2] Alter, O., Brown, P. O. & Botstein, D. (2001) in *Microarrays: Optical Technologies and Informatics*, eds. Bittner, M. L., Chen, Y., Dorsel, A. N. & Dougherty, E. R. (International Society for Optical Engineers, Bellingham, Washington) Vol. 4266, p. 186.
[3] Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Ferro, M., Sherlock, G., Pollack, J. R., Brown, P. O., Botstein, D. & van de Rijn, M. (2002) *Lancet* **359**, 1301–1307.
[4] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L. & Somogyi, R. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 334–339.
[5] Hilsenbeck, S. G., Friedrichs, W. E., Schiff, R., O'Connell, P., Hansen, R. K., Osborne, C. K. & Fuqua, S. A. (1999) *J. Natl. Cancer Inst.* **91**, 453–459.
[6] Raychaudhuri, S., Stuart, J. M. & Altman, R. B. (2000) in *Proceedings of the Pacific Symposium of Biocomputing*, eds. Altman, R. B., Lauderdale, K., Dunker, A. K., Hunter, L. & Klein T. E., (World Scientific, Singapore), p. 455.
[7] Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R. & Fedoroff, N. V. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 8409–8414.
[8] Golub, G. H. & Van Loan, C. F. (1996) *Matrix Computation* (Johns Hopkins University Press, Baltimore), 3rd Ed.
[9] Paige, C. C. & Saunders, M. A. (1981) *SIAM. J. Numer. Anal. USA* **18**, 398–405.
[10] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. (1999) *Nat. Genet.* **22**, 281–285.
[11] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. & Futcher, B. (1998) *Mol. Biol. Cell* **9**, 3273–3297.
[12] Whitfield, M. L., Sherlock, G., Saldanha, A., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O. & Botstein, D. (2002) *Mol. Biol. Cell* **13**, 1977–2000.
[13] Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D. & Cherry, J. M. (2002) *Nucleic Acids Res.* **30**, 69–72.
[14] Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. & Cherry, J. M. (2001) *Nucleic Acids Res.* **29**, 152–155.
[15] Kurihara, L. J., Stewart, B. G., Gammie, A. E. & Rose, M. D. (1996) *Mol. Cell. Biol.* **16**, 3990–4002.