Discovery of Mechanisms and Prognosis of Cancers from Matrix and Tensor Modeling of Large-Scale Molecular Biological Data

Orly Alter

**Departments of Bioengineering and Human Genetics, and
Scientific Computing and Imaging (SCI) Institute,
University of Utah, and
Utah Science, Technology and Research (USTAR) Initiative**

**orly@sci.utah.edu
http://alterlab.org/**

A groundbreaking look at the nature of quantum mechanics

With new technologies permitting the observation and manipulation of single quantum systems, the quantum theory of measurement is fast becoming a subject of experimental investigation in laboratories worldwide. This original new work addresses open fundamental questions in quantum mechanics in light of these experimental developments.

Using a novel analytical approach developed by the authors, *Quantum Measurement of a Single System* provides answers to three long-standing questions that have been debated by such thinkers as Bohr, Einstein, Heisenberg, and Schrödinger. It establishes the quantum theoretical limits to information obtained in the measurement of a single system on the quantum wavefunction of the system, the time evolution of the quantum observables associated with the system, and the classical potentials or forces which shape this time evolution. The technological relevance of the theory is also demonstrated through examples from atomic physics, quantum optics, and mesoscopic physics.

Suitable for professionals, students, or readers with a general interest in quantum mechanics, the book features recent formulations as well as humorous illustrations of the basic concepts of quantum measurement. Researchers in physics and engineering will find *Quantum Measurement of a Single System* a timely guide to one of the most stimulating fields of science today.

ORLY ALTER, PhD, is currently a postdoctoral fellow in the Department of Genetics at Stanford University. **YOSHIHISA YAMAMOTO, PhD,** is a professor in the Departments of Applied Physics and Electrical Engineering at Stanford University. He is currently the director of the ICORP Quantum Entanglement Project of the Japanese Science and Technology (JST) Corporation. While they collaborated on the research presented in this book, Yamamoto was the director of the ERATO Quantum Fluctuation Project of JST, and Alter was a doctoral student at the Department of Applied Physics at Stanford. She was selected as a finalist for the American Physical Society Award for Outstanding Doctoral Thesis Research in Atomic, Molecular or Optical Physics for 1998 for this work.

Cover Illustration: David B. Oberman

ISBN 0-471-28308-8
90000
9 780471 283089

WILEY-INTERSCIENCE
John Wiley & Sons, Inc.
Scientific, Technical, and Medical Division
605 Third Avenue, New York, N.Y. 10158-0012
New York • Chichester • Weinheim
Brisbane • Singapore • Toronto

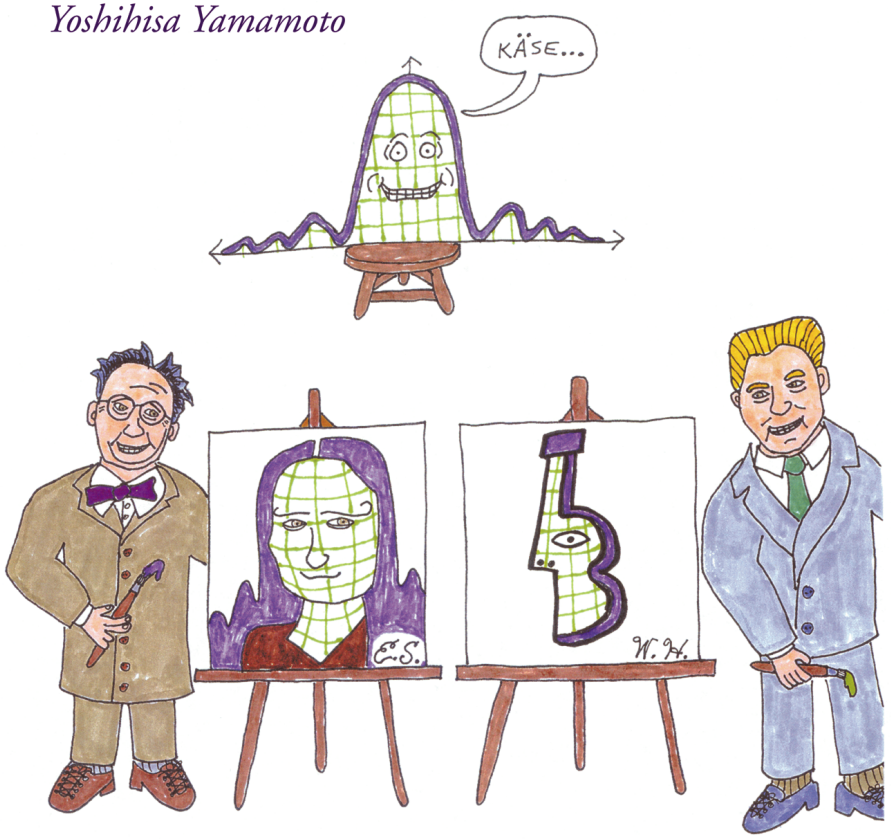
ALTER
YAMAMOTO

Quantum Measurement of a Single System

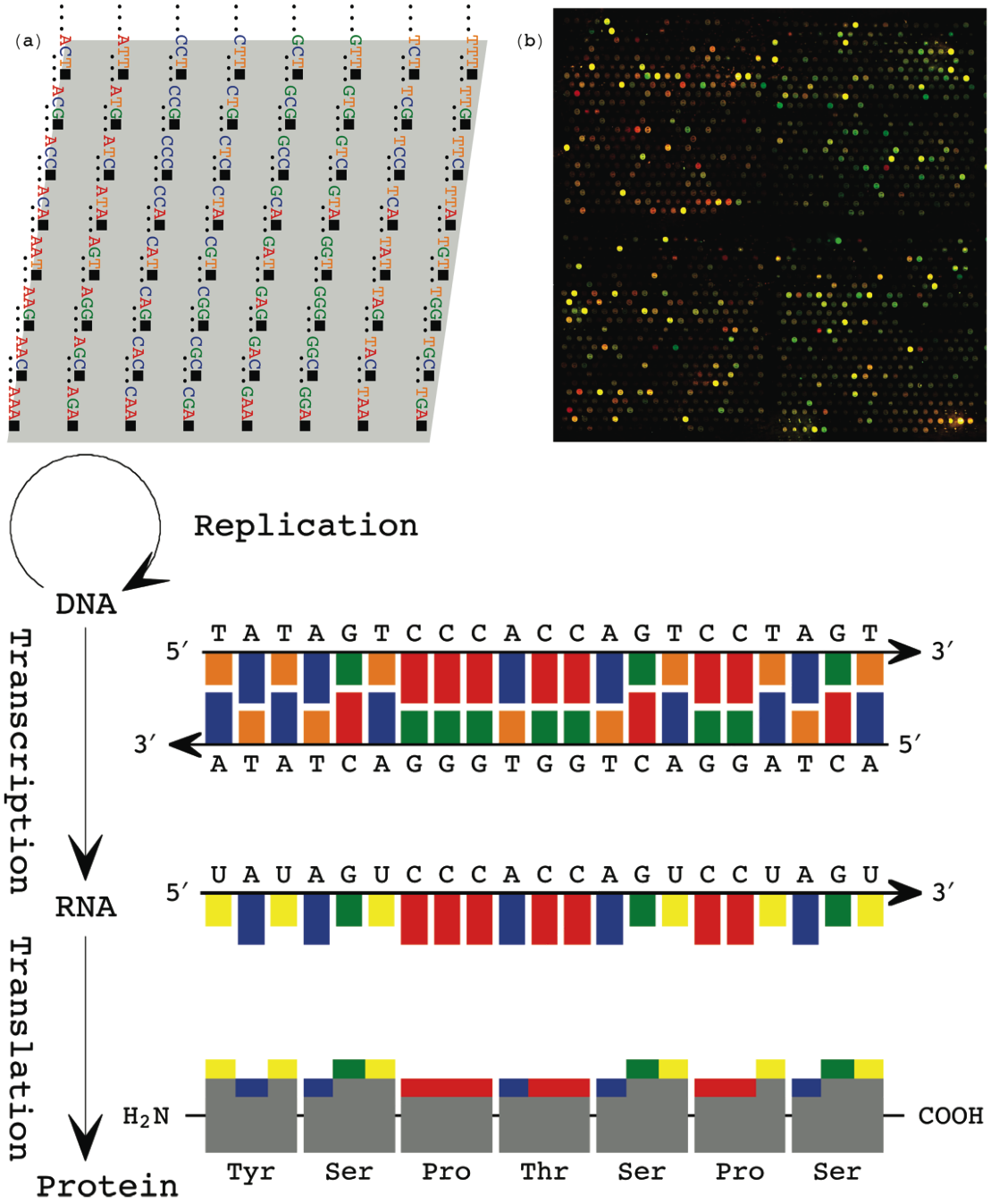
WILEY
INTER-
SCIENCE

Quantum Measurement of a Single System

Orly Alter
Yoshihisa Yamamoto

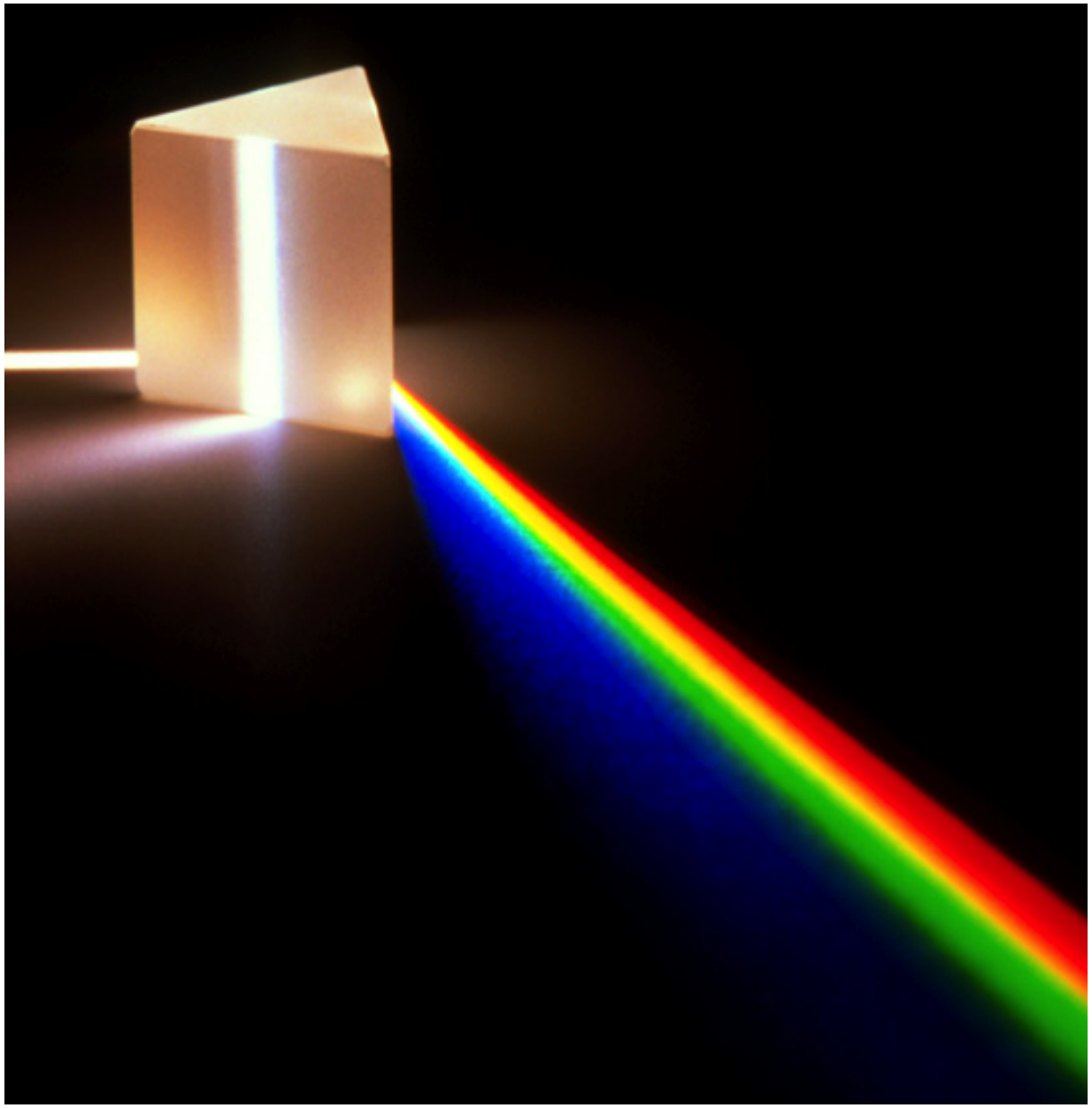


High-Throughput Biotechnologies Record Global Signals



DNA microarrays, e.g., rely on hybridization to record the complete genomic signals that guide the progression of cellular processes, such as abundance levels of DNA, RNA, and DNA- and RNA-bound proteins on a genomic scale.

Global Mathematical Vocabulary for Molecular Biological Discovery



Develop generalizations of the matrix and tensor decompositions that underlie the theoretical description of the physical world;

Create models that compare and integrate different types of large-scale molecular biological data;

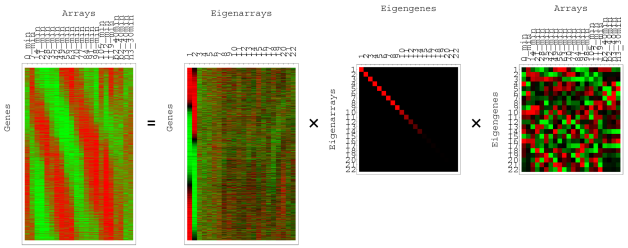
Predict global mechanisms that govern the activity of DNA and RNA.

Physics-Inspired Matrix (and Tensor) Models

Mathematical frameworks for the description of the data, in which the mathematical variables and operations might represent biological reality.

SVD

Alter, Brown & Botstein,
PNAS 97, 10101 (2000).

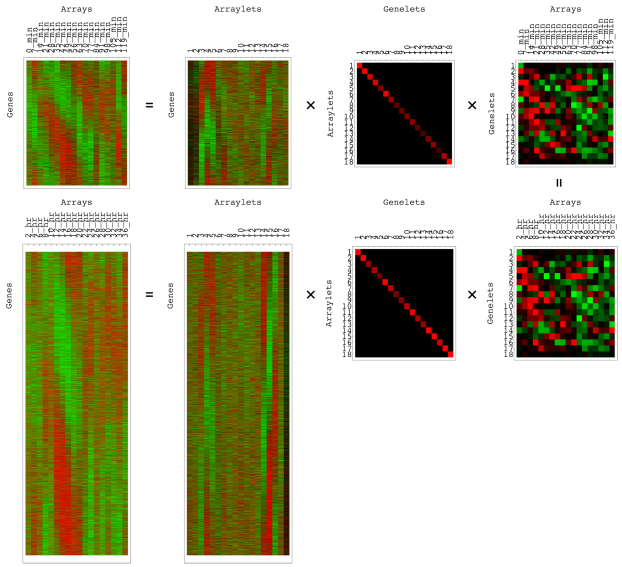


"Eigengenes" and "eigenarrays" → cellular processes and states in a single dataset.

Eigenvalue Decomposition

Comparative GSVD

Alter, Brown & Botstein,
PNAS 100, 3351 (2003).

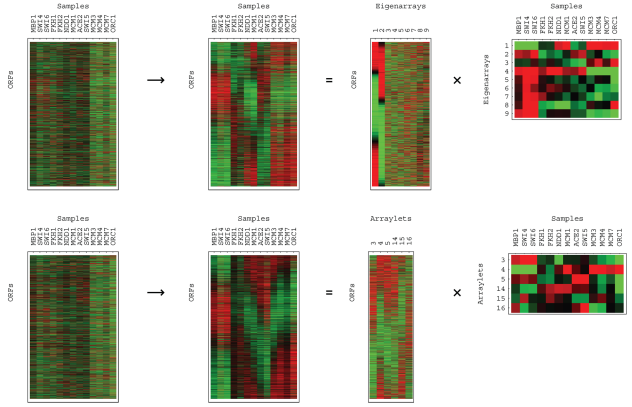


"Genelets" and "arraylets" → phenomena exclusive to one of, or common to two datasets.

Generalized Eigenvalue Decomposition

Integrative Pseudoinverse

Alter & Golub,
PNAS 101, 16577 (2004).

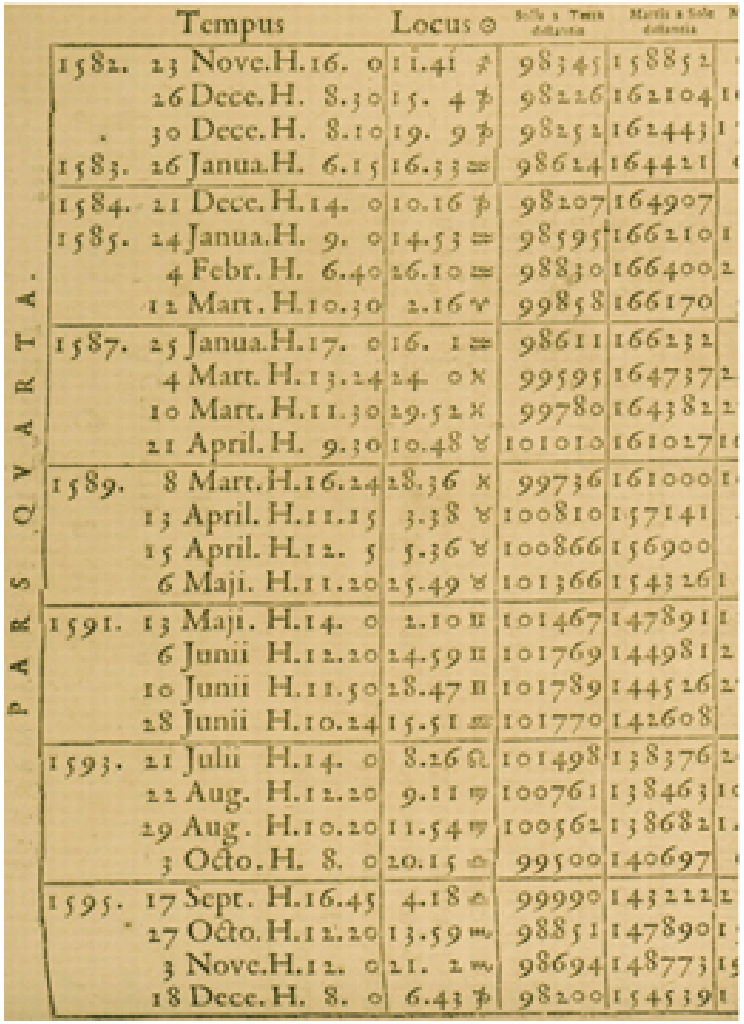


"Pseudoinverse correlation" → causal coordination between two datasets.

Inverse Projection

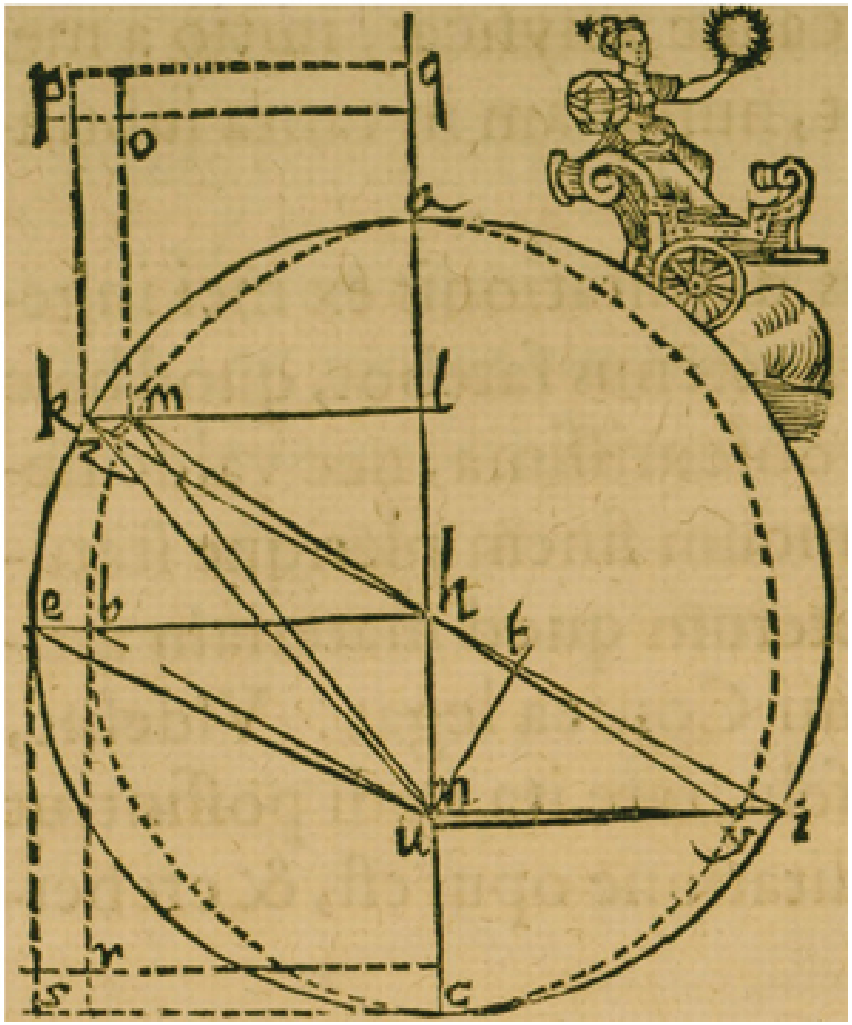
Patterns Underlie Principles of Nature: Global Correlations to Causal Coordination

Alter, *PNAS* 103, 16063 (2006);
Alter, in *Microarray Data Analysis: Methods and Applications* (Humana Press, 2007), pp. 17–59.



PARS QVARTA.

Tempus		Locus ☉		
1582.	23 Nove. H. 16. 0	11. 41 ♐	98345	158852
	26 Dece. H. 8. 30	15. 4 ♑	98226	162104
	30 Dece. H. 8. 10	19. 9 ♑	98252	162443
1583.	26 Janua. H. 6. 15	16. 33 ♒	98624	164421
1584.	21 Dece. H. 14. 0	10. 16 ♑	98207	164907
1585.	24 Janua. H. 9. 0	14. 53 ♒	98595	166210
	4 Febr. H. 6. 40	26. 10 ♒	98830	166400
	12 Mart. H. 10. 30	2. 16 ♈	99858	166170
1587.	25 Janua. H. 17. 0	16. 1 ♒	98611	166232
	4 Mart. H. 13. 24	24. 0 ♓	99595	164737
	10 Mart. H. 11. 30	29. 52 ♓	99780	164382
	21 April. H. 9. 30	10. 48 ♉	101010	161027
1589.	8 Mart. H. 16. 24	28. 36 ♓	99736	161000
	13 April. H. 11. 15	3. 38 ♉	100810	157141
	15 April. H. 12. 5	5. 36 ♉	100866	156900
	6 Maji. H. 11. 20	25. 49 ♉	101366	154326
1591.	13 Maji. H. 14. 0	2. 10 ♊	101467	147891
	6 Junii H. 12. 20	24. 59 ♊	101769	144981
	10 Junii H. 11. 50	28. 47 ♊	101789	144526
	28 Junii H. 10. 24	15. 51 ♋	101770	142608
1593.	21 Julii H. 14. 0	8. 26 ♌	101498	138376
	22 Aug. H. 12. 20	9. 11 ♍	100761	138463
	29 Aug. H. 10. 20	11. 54 ♍	100562	138682
	3 Octo. H. 8. 0	20. 15 ♎	99500	140697
1595.	17 Sept. H. 16. 45	4. 18 ♎	99990	143222
	27 Octo. H. 12. 20	13. 59 ♏	98851	147890
	3 Nove. H. 12. 0	21. 2 ♏	98694	148773
	18 Dece. H. 8. 0	6. 43 ♑	98200	154519



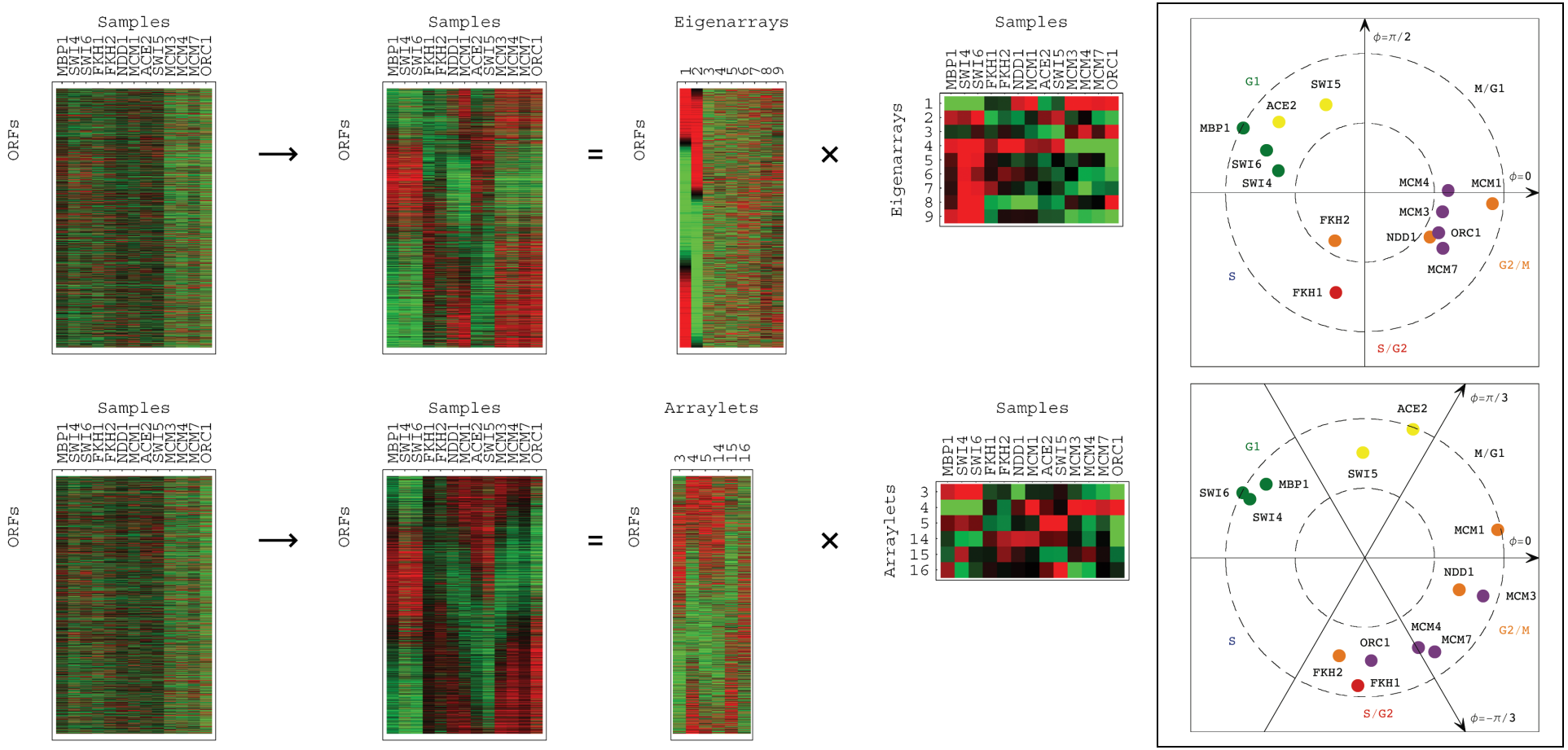
Kepler's discovery of his first law of planetary motion from mathematical modeling of Brahe's astronomical data.

Kepler, *Astronomia Nova* (Voegelinus, Heidelberg, 1609).

Integrative Pseudoinverse Projection Predicts a Global Mode of Genetic Regulation

Alter & Golub, *PNAS* 101, 16577 (2004); http://alterlab.org/pseudoinverse/

Alter, Golub, Brown & Botstein, *Miami Nature Biotechnology Winter Symposium: Cell Cycle, Chromosomes and Cancer* (January 31 – February 4, 2004, Miami Beach, FL).



DNA binding of replication initiation proteins is correlated with minimum expression of adjacent genes during the cell cycle stage G1.

Simon et al., *Cell* 106, 697 (2001); Wyrick et al., *Science* 294, 2397 (2001).

Novel Correlation Between DNA Replication and RNA Expression Might Be Due to a Previously Unknown Mechanism of Regulation

→ Replication initiation requires binding at replication origins during G1.
Diffley, Cocker, Dowell, & Rowley, *Cell* 78, 303 (1994).

→ Replication initiation proteins are involved with transcriptional silencing at the yeast mating loci.
Micklem et al., *Nature* 366, 87 (1993).

Either one of two previously unknown modes of regulation might underlie this correlation:

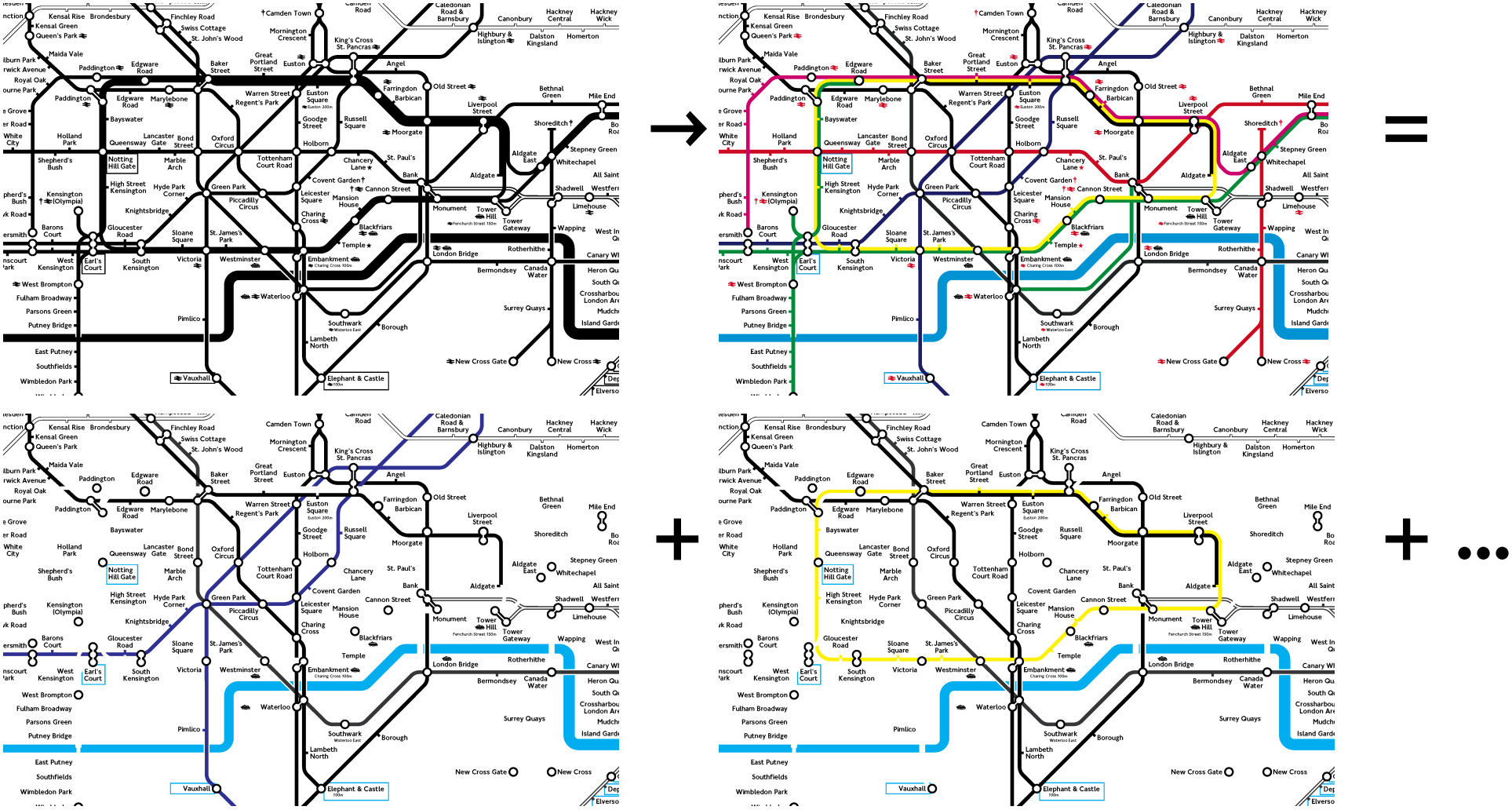
→ Replication might regulate transcription:
Binding at origins might interfere with adjacent gene expression.

→ Transcription might regulate replication:
G1 gene expression might reduce the efficiency of adjacent origins.
Donato, Chung & Tye, *PLoS Genet*. 2, E141 (2006);
Snyder, Sapolsky & Davis, *MCB* 8, 2184 (1988).

This demonstrates that a data-driven mathematical model of DNA microarray data can be used to predict a cellular mechanism of regulation that is truly on a genome scale.

Networks are Tensors of "Subnetworks"

Alter & Golub, *PNAS* 102, 17559 (2005);
http://alterlab.org/network_decomposition/



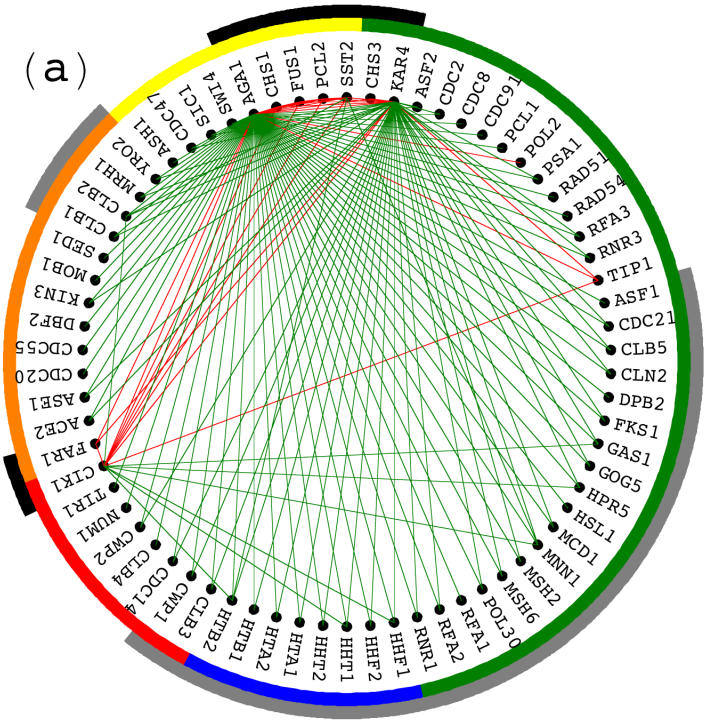
The relations among the activities of genes, not only the activities of the genes alone, are known to be pathway-dependent, i.e., conditioned by the biological and experimental settings in which they are observed.

Math Variables → Biology

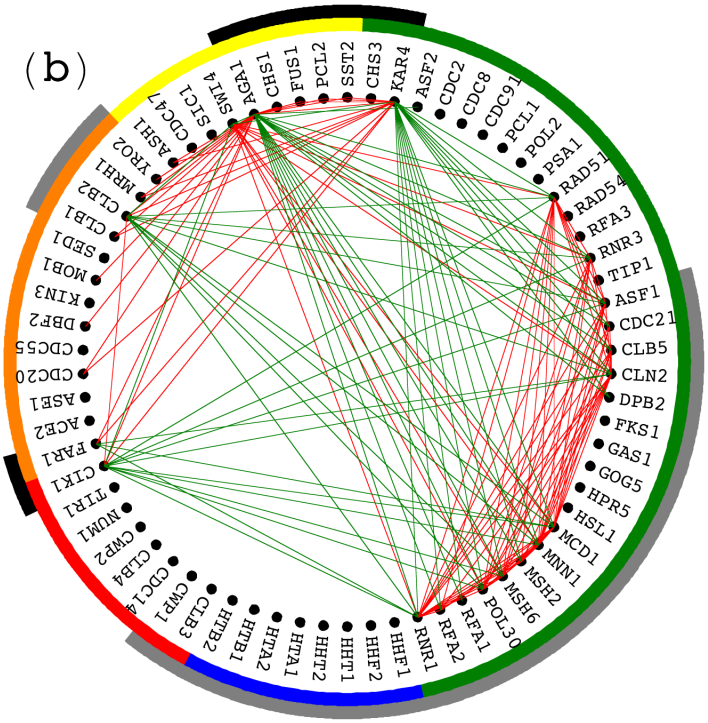
Significant EVD subnetworks → functionally independent pathways:

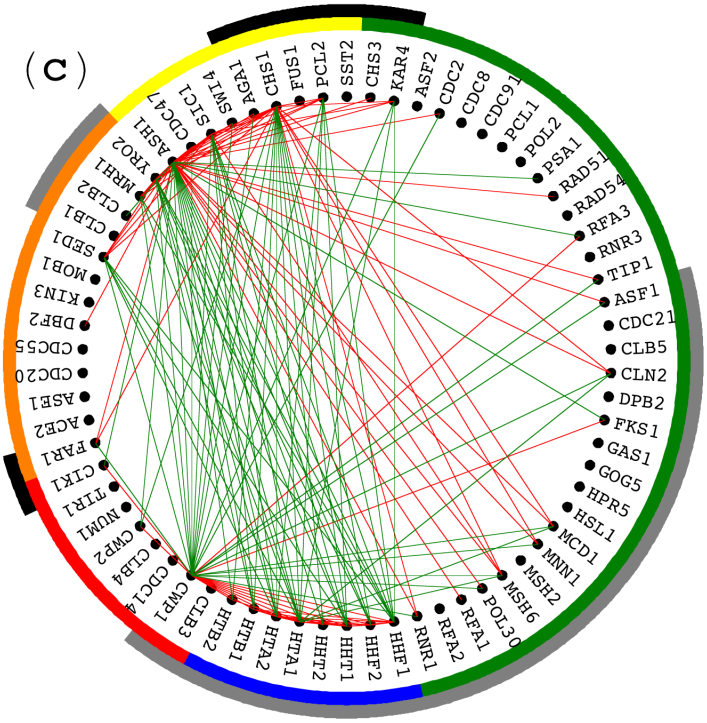
Pheromone Signaling Pathway

KAR4 || *CIK1*



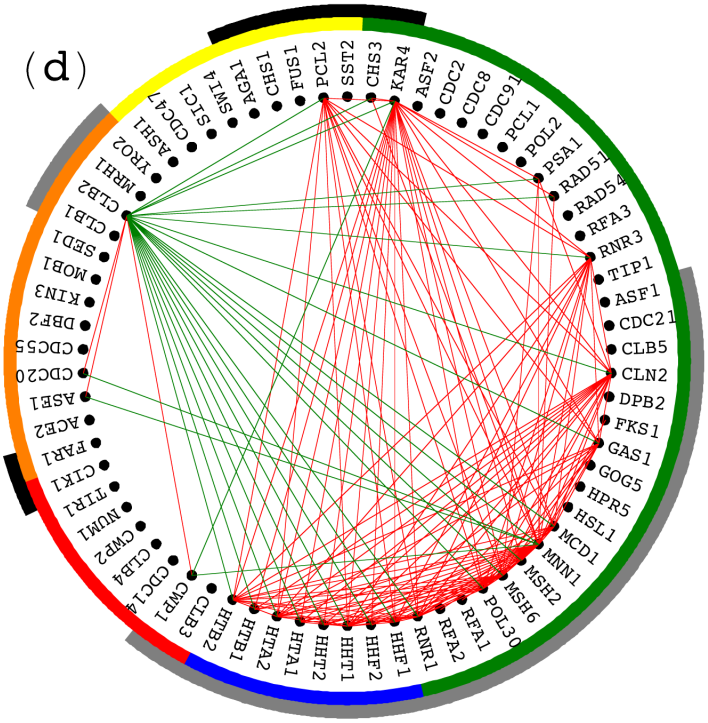
Pheromone Arrest Exit & G_1 Entry





Cell Cycle S ↔ M

KAR4 || −*CIK1*



Cell Cycle G_1 ↔ G_2

Integrative Higher-Order SVD Predicts an Equivalent Global Mode of Genetic Regulation

Omberg, Golub & Alter, *PNAS* 104, 18371 (2007);
http://alterlab.org/HOSVD/

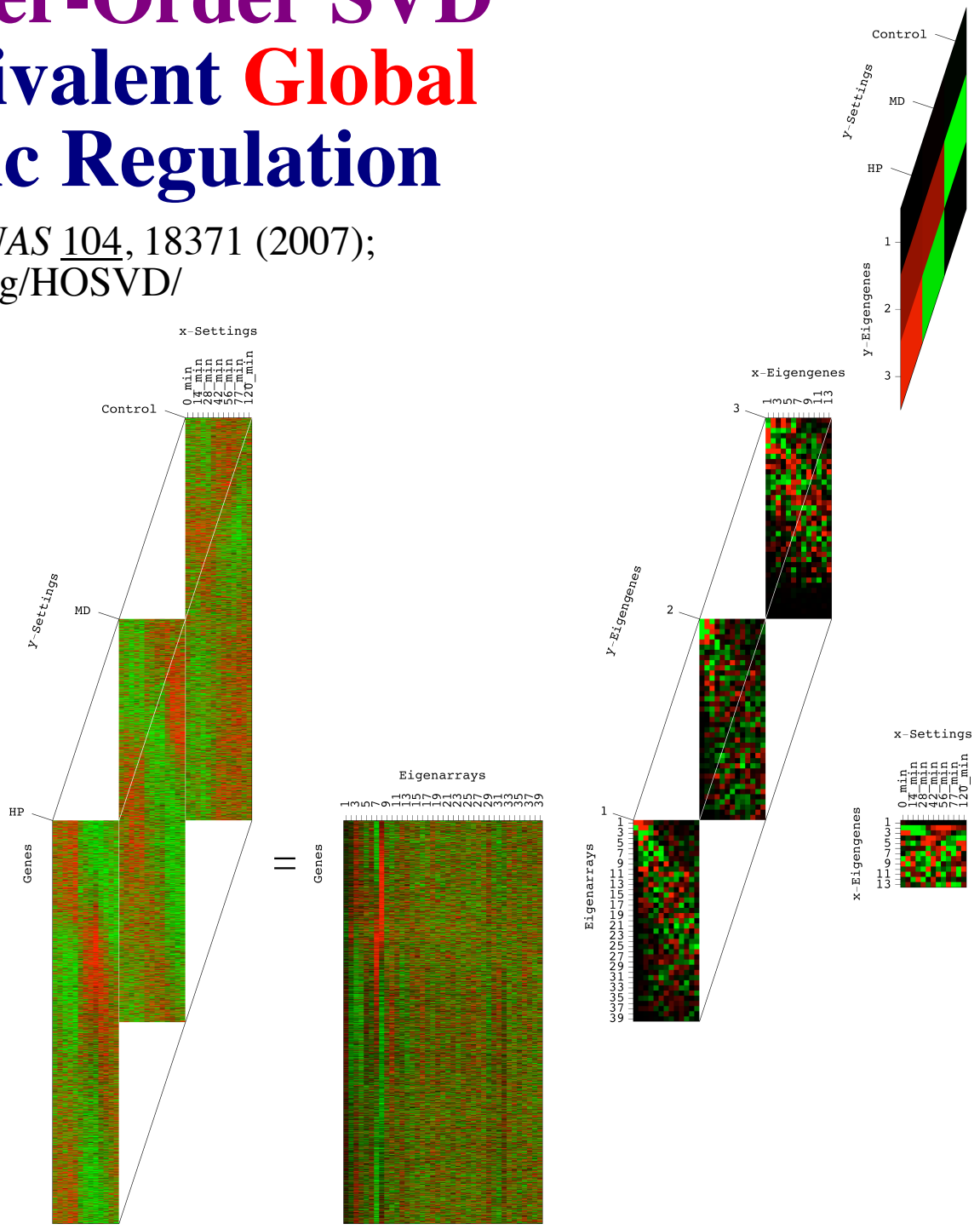
This HOSVD is computed from each SVD of the data tensor unfolded along all axes perpendicular to one given axis,

$$\mathcal{T} = \mathcal{R} \times_a U \times_b V_x \times_c V_y$$

De Lathauwer, De Moor & Vandewalle, *SIMAX* 21, 1253 (2000);
Kolda, *SIMAX* 23, 243 (2001);
Zhang & Golub, *SIMAX* 23, 543 (2001).

mRNA expression from cell cycle time courses under different conditions of oxidative stress

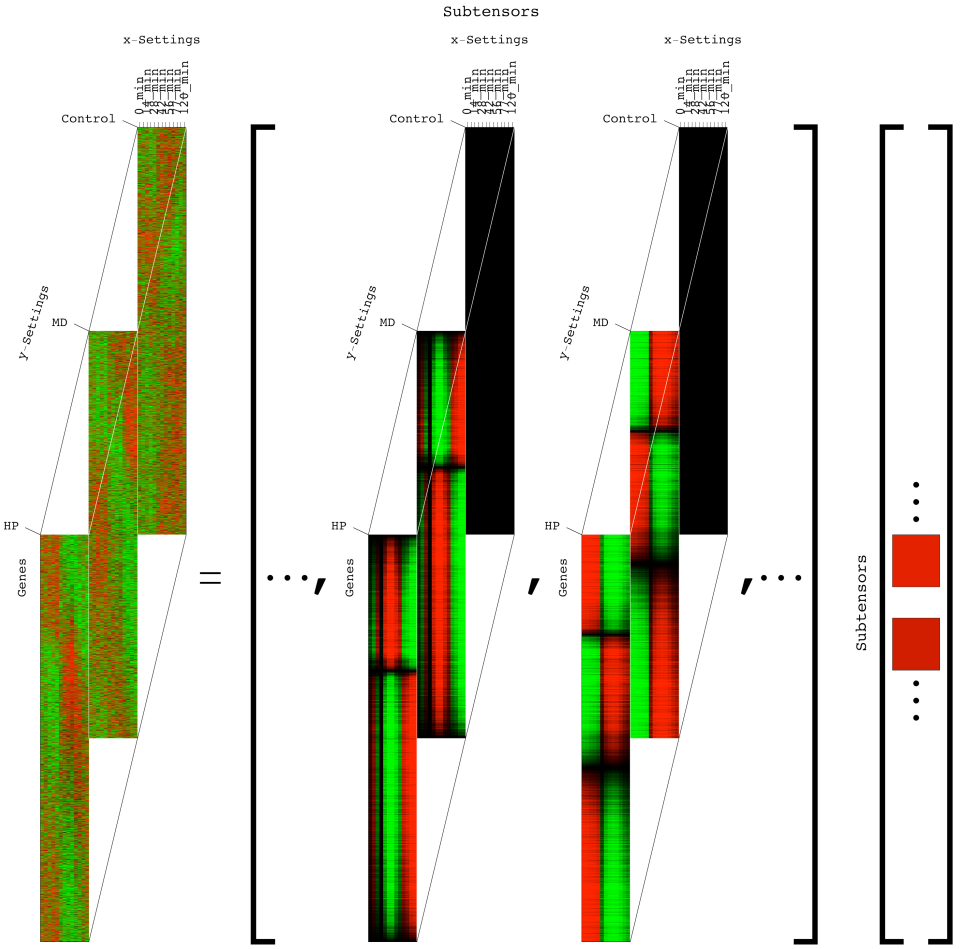
Shapira, Segal & Botstein, *MBC* 15, 5659 (2004);
Spellman et al., *MBC* 9, 3273 (1998).



Mathematical Reformulation of the HOSVD

The data tensor is a superposition of all rank-1 "subtensors," i.e., outer products of an eigenarray, an *x*- and a *y*-eigengene,

$$\mathcal{T} \equiv \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{R}_{abc} \mathcal{S}(a,b,c).$$



The significance of a subtensor is defined by the corresponding "fraction," computed from the higher-order singular values,

$$\mathcal{P}_{abc} \equiv \mathcal{R}_{abc}^2 \Big/ \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{R}_{abc}^2 .$$

The complexity of the data tensor is defined by the "normalized entropy,"

$$0 \le d = \frac{-1}{2\log(LM)} \sum_{a=1}^{LM} \sum_{b=1}^{L} \sum_{c=1}^{M} \mathcal{P}_{abc} \log(\mathcal{P}_{abc}) \le 1$$

"Degenerate subspace rotation" replaces two subtensors $|\mathcal{R}_{abc}| = |\mathcal{R}_{kbc}|$ with a unique rank-1 subtensor $\mathcal{R}_{a+k,b,c}\mathcal{S}(a+k,b,c) = \mathcal{R}_{abc}\mathcal{S}(a,b,c) + \mathcal{R}_{kbc}\mathcal{S}(k,b,c).$

Math Variables → Biology

Significant subtensors → independent biological programs or experimental phenomena:

$\mathcal{S}(k,l,m)$	$\mathcal{P}_{klm}$	$\mathcal{R}_{klm}$
1,1,1	70%	>0

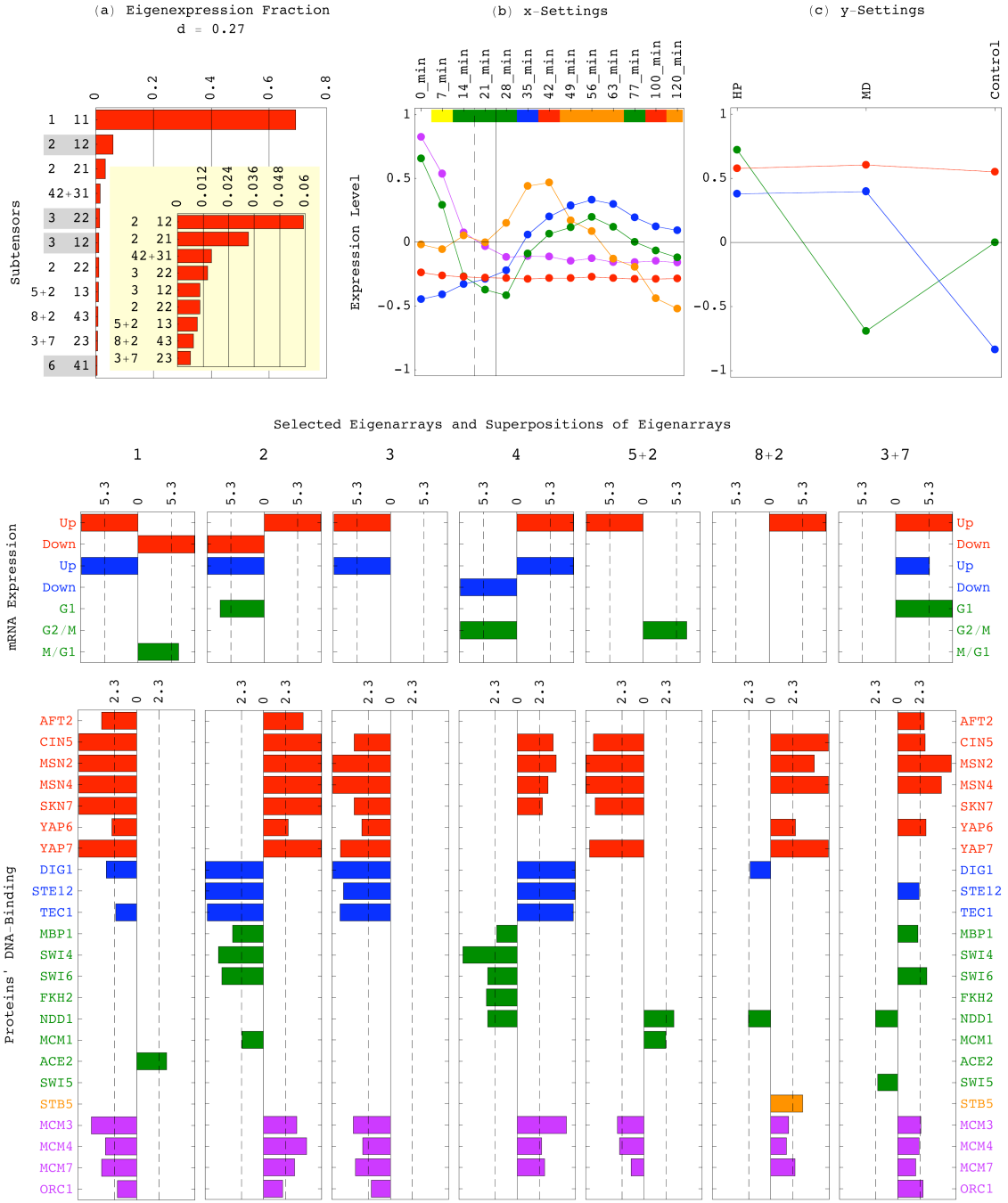
Steady State

2,1,2	6%	<0
2,2,1	3.3%	>0
2,2,2	1%	>0

Oxidative Stress in Time and Across Conditions

4,2+3,1	1.6%	>0
3,2,2	1.4%	<0
3,1,2	1%	<0

Pheromone Responses



Math Operations → Biology

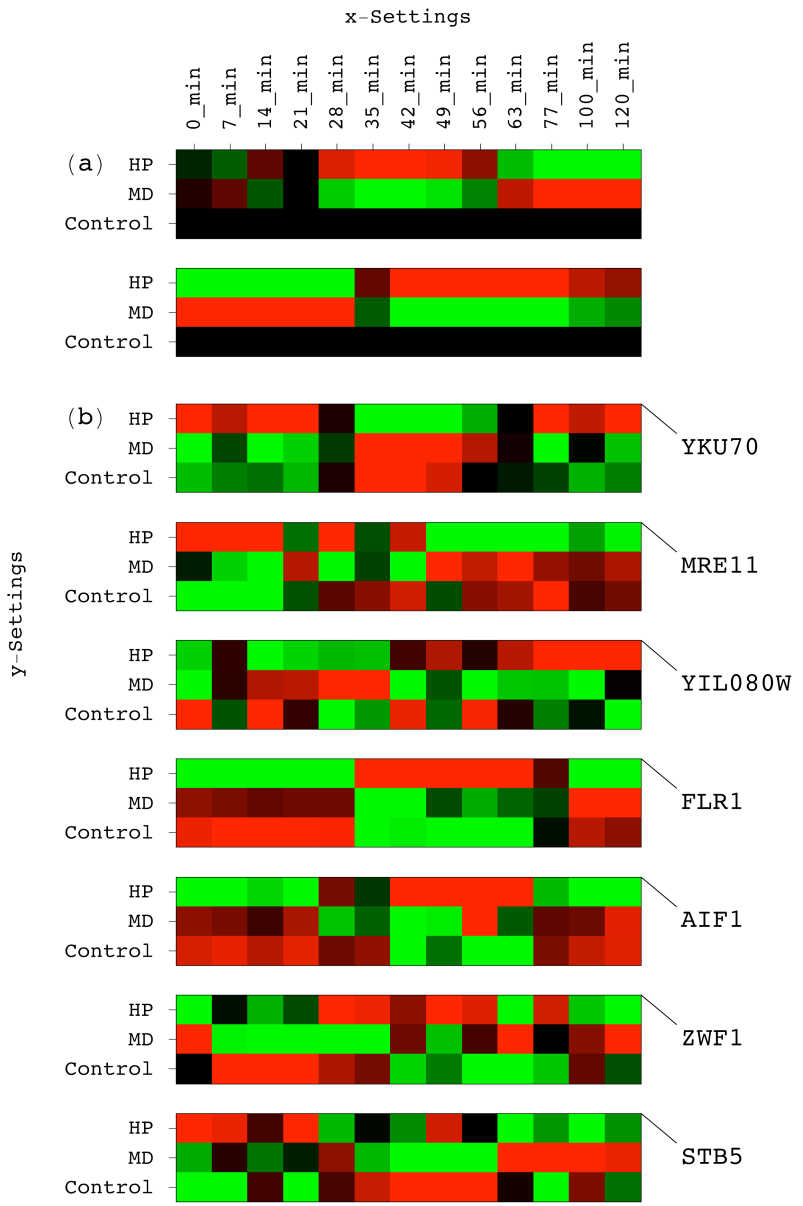
$\mathcal{S}(k,l,m)$	$\mathcal{P}_{klm}$	$\mathcal{R}_{klm}$
5+2,1,3	0.9%	>0
8+2,4,3	0.75%	>0
3+7,2,3	0.6%	>0

HP vs. MD-Induced Expression

Flattery-O'Brien & Dawes,
J. Biol. Chem. 273, 8564 (1998).

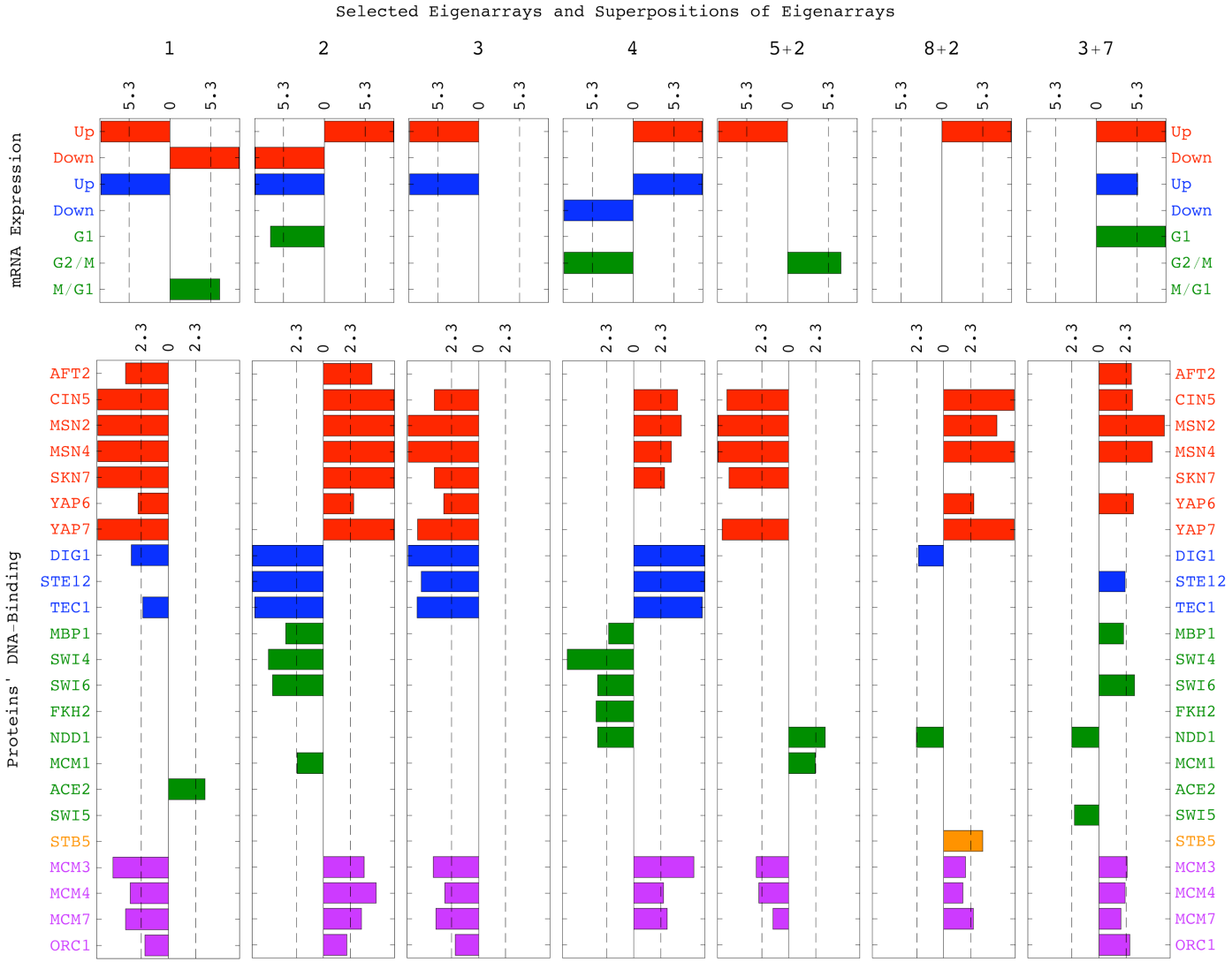
Classification identifies genes significant in terms of the information that they capture in each subtensor → global picture of time-dependence of HP vs. MD-induced expression:

The conserved genes *YKU70*, *MRE11*, *AIF1* and *ZWF1*, and the processes of retrotransposition, apoptosis and the oxidative pentose phosphate cycle that they are involved in, play significant, yet previously unrecognized, roles in the differential effects of HP and MD on cell cycle progression.



Equivalent Global Correlation Between DNA Replication and RNA Expression is Revealed

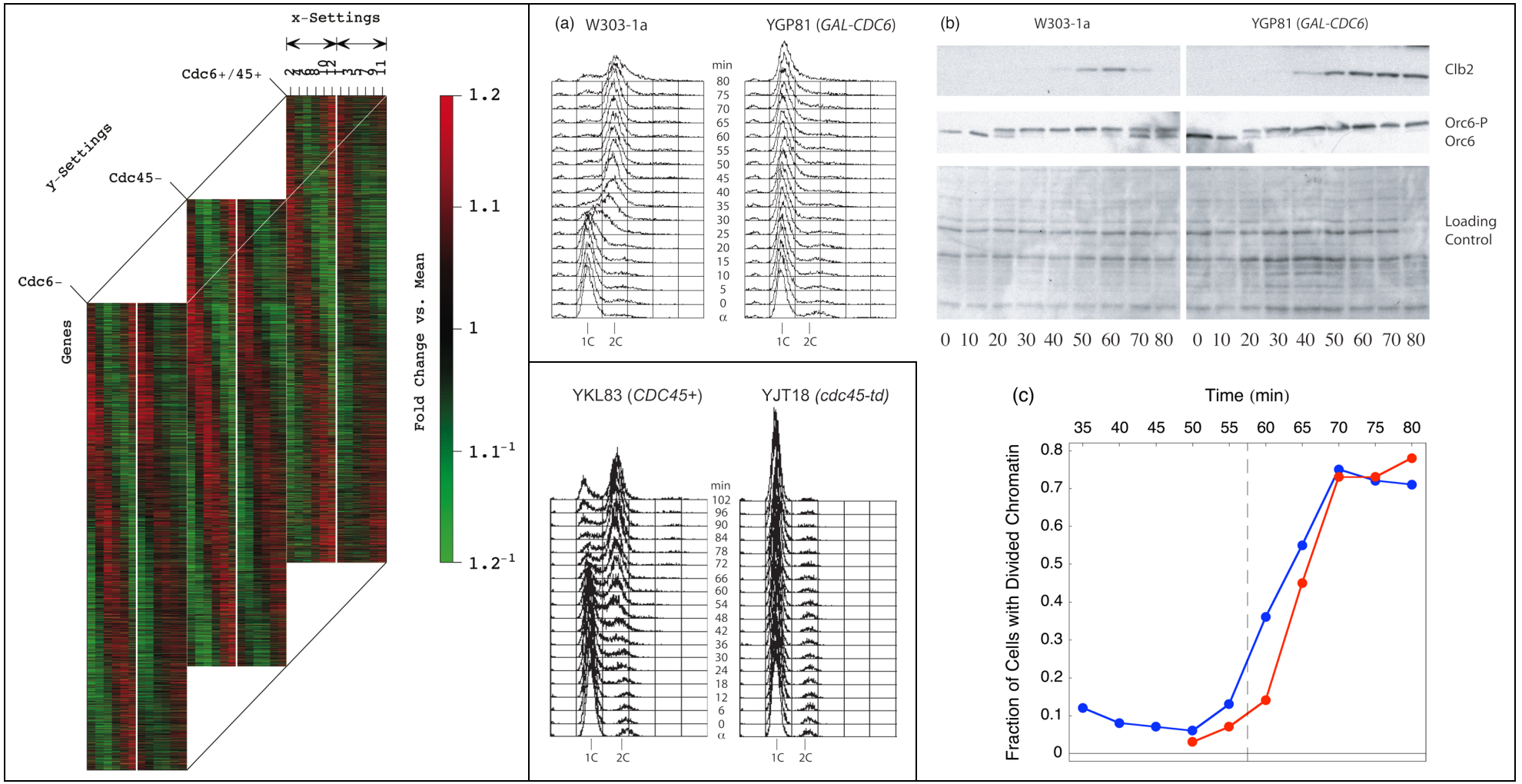
Overexpression of the binding targets of the replication initiation proteins correlates with that of oxidative stress activators-bound genes.



Cocker et al., *Nature* 379, 180 (1996); Blanchard et al., *MBC* 13, 1536 (2002).

Analysis of Synchronized Cdc6⁻/45⁻ Cultures where DNA Replication Initiation is Prevented without Delaying Cell Cycle Progression

Omberg, Meyerson, Kobayashi, Drury, Diffley & Alter, *MSB* 5, 312 (2009);
http://alterlab.org/verification_of_prediction/

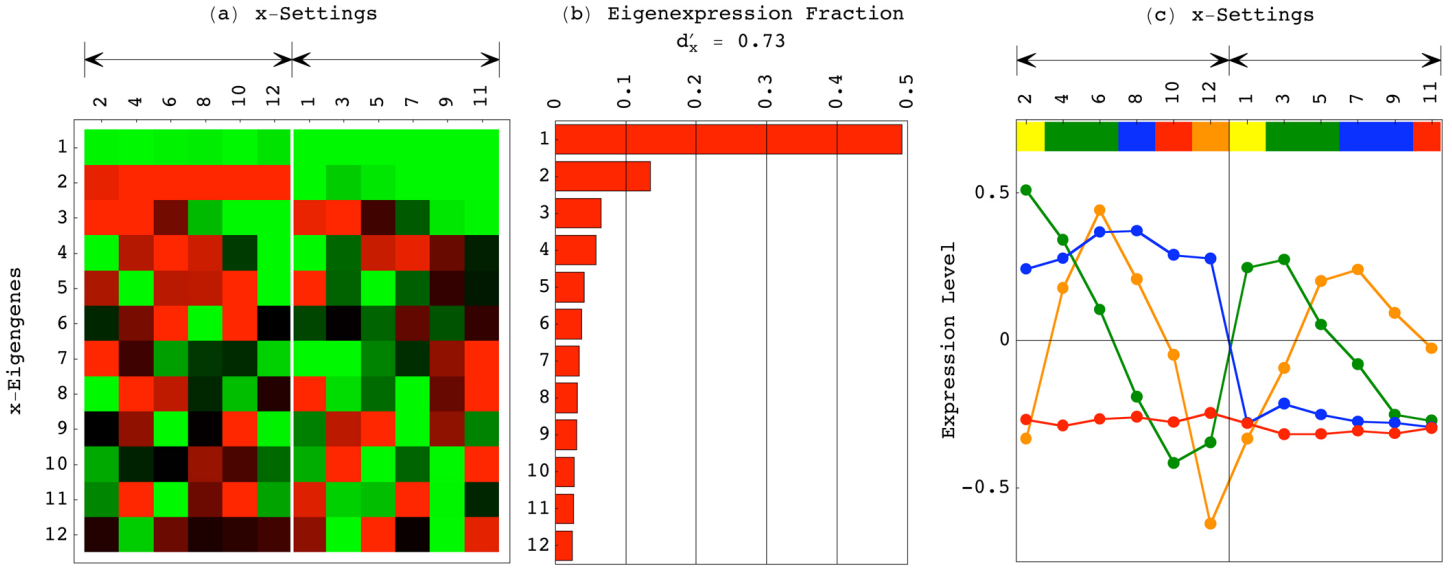


Gerke, Chen & Cohen, *Genetics* 174, 985 (2006).

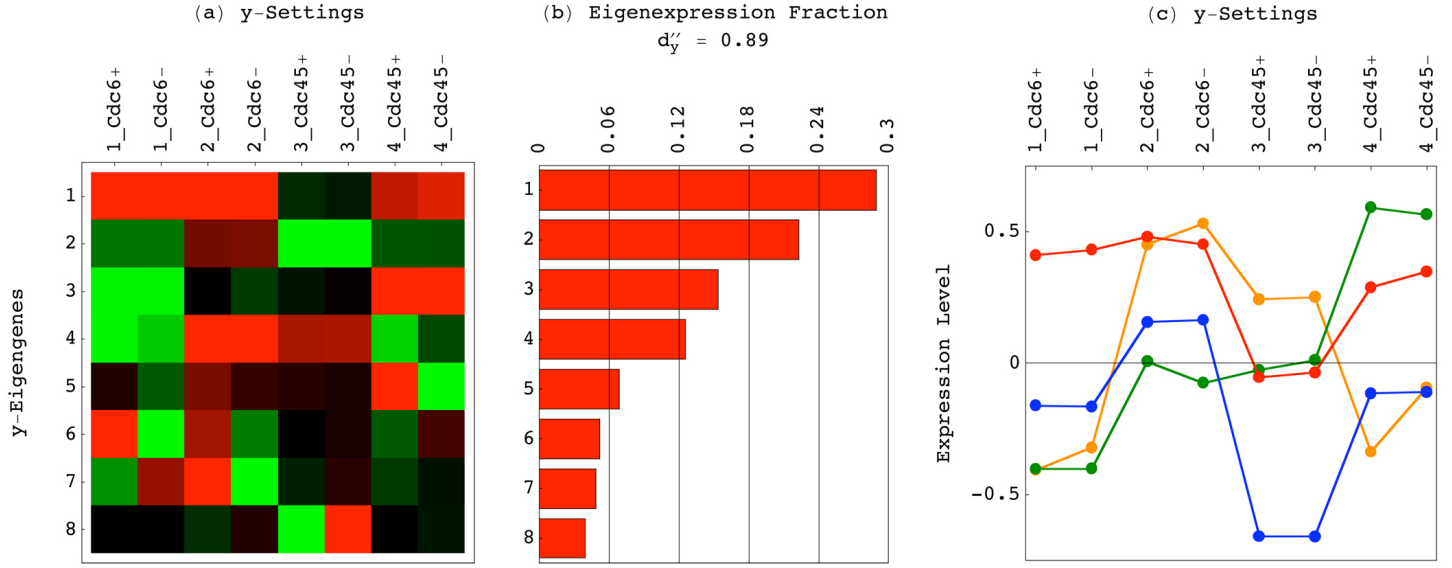
HOSVD Detection and Removal of Artifacts

Reconstructing the data tensor of 4,270 genes × 12 time points, or *x*-settings × 8 time courses, or *y*-settings, filtering out "*x*-eigengenes" and "*y*-eigengenes" that represent experimental artifacts.

Batch-of-hybridization

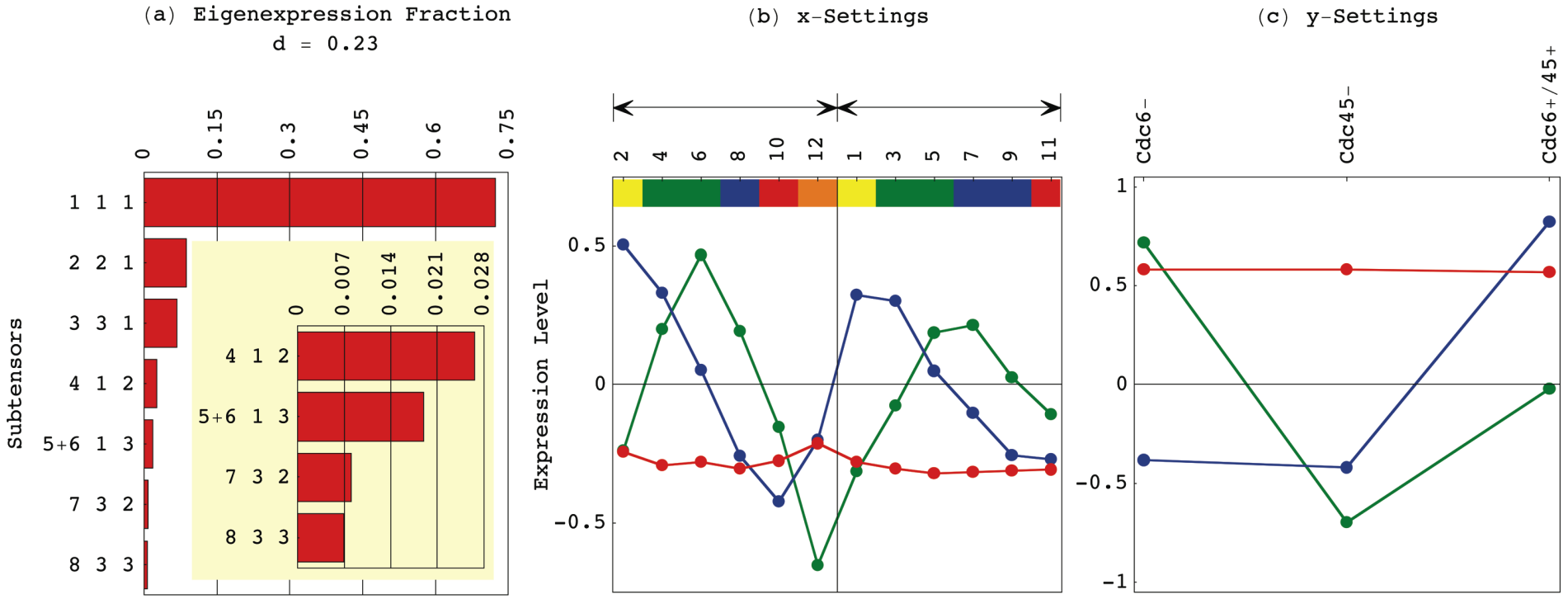


Culture batch, microarray platform and protocols



Swinnen, Van Huffel, Van Loven & Jacobs, *Med Biol Eng Comput* 38, 297 (2000).

Uncovering Effects of Replication and Origin Activity on mRNA Expression with HOSVD



$\mathcal{S}(k,l,m)$	$\mathcal{P}_{klm}$	$\mathcal{R}_{klm}$
1,1,1	72%	>0

Steady State

First, ~88% of mRNA expression is independent of DNA replication.

Orlando et al., *Nature* 453, 944 (2008).

2,2,1	9%	>0	↑ M/G1	$<2\cdot10^{-33}$	↓ S/G2	$<7\cdot10^{-16}$
3,3,1	7%	>0	↑ G1/S	$<2\cdot10^{-77}$	↓ G2/M	$<3\cdot10^{-36}$

Unperturbed Cell Cycle

Replication-Dependent Perturbations

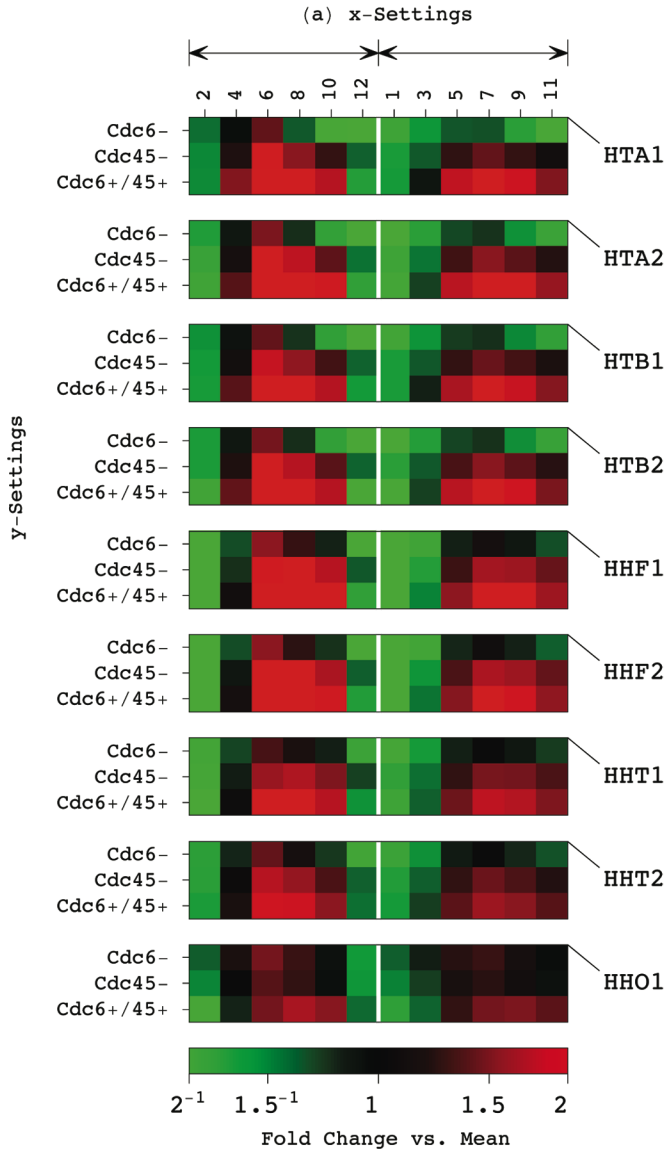
4,1,2	2.7%	>0	↑ ARSs 3’ $\sim 10^{-2}$	↓ histones $< 10^{-12}$
7,3,2	0.8%	>0	↑ histones $< 5 \cdot 10^{-4}$	

DNA replication increases time-averaged and G1/S expression of histones.

Histones are overexpressed in the control relative to the Cdc6$^-$ condition, and to a lesser extent also relative to the Cdc45$^-$ condition (a *P*-value $\sim 2 \cdot 10^{-15}$).

Second, the requirement of DNA replication for efficient histone gene expression is independent of conditions that elicit DNA damage checkpoint responses.

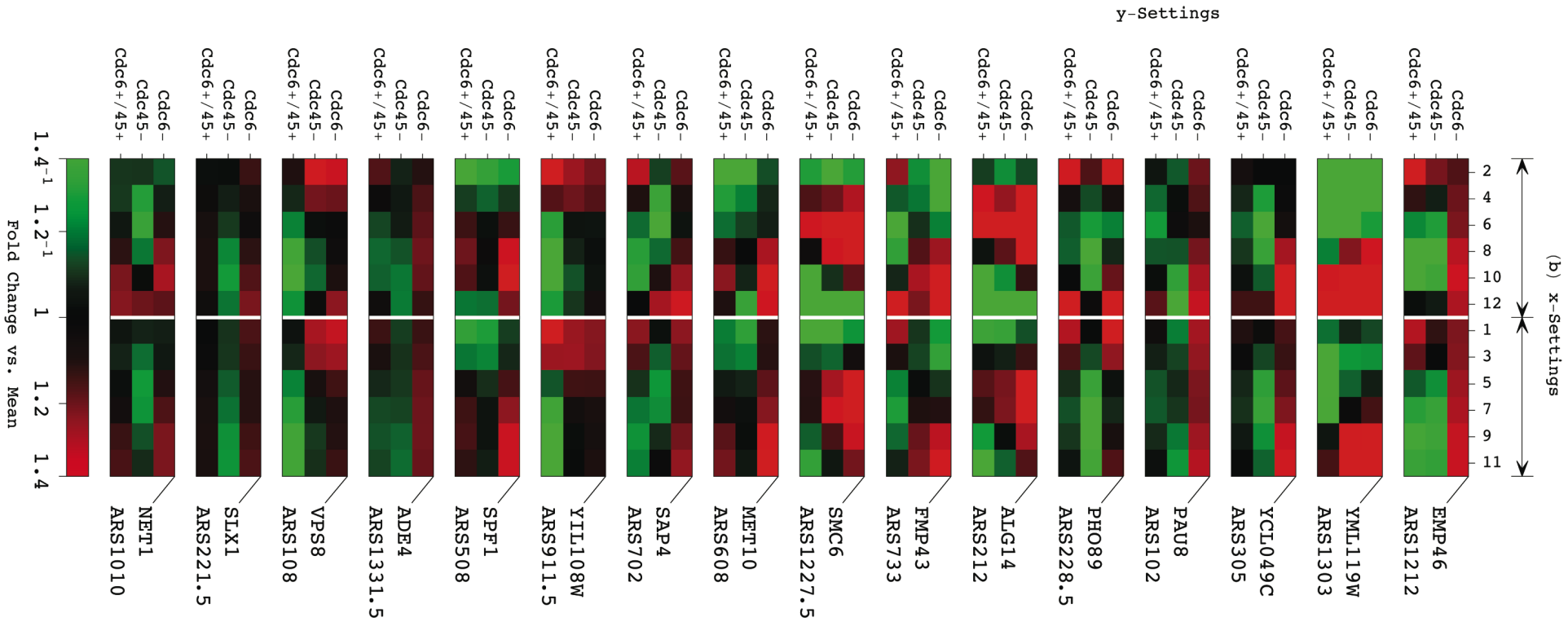
Lycan, Osley & Hereford, *MCB* 7, 614 (1987).



Origin Binding-Dependent Perturbations

5+6,1,3	1.9%	>0	↑ histones $<2\cdot10^{-8}$	↓ ARSs 3' $<2\cdot10^{-3}$
8,3,3	0.7%	>0		↓ ARSs 3' $<7\cdot10^{-4}$

Origin binding decreases time-averaged and G2/M expression of genes with ARSs near their 3' ends. These genes are overexpressed in the Cdc6$^-$ relative to the Cdc45$^-$ condition, and to a lesser extent also relative to the control (a *P*-value $<4\cdot10^{-7}$) → Third, origin licensing decreases expression of genes with origins near their 3' ends, revealing that downstream origins can regulate the expression of upstream genes.



Experimental Verification of the Computationally Predicted Mechanism

Omberg, Meyerson, Kobayashi, Drury, Diffley & Alter, *MSB* 5, 312 (2009);
http://dx.doi.org/10.1038/msb.2009.70

→ These experimental results reveal that downstream origins can regulate the expression of upstream genes.

→ These experimental results verify the computationally predicted mechanism of regulation that correlates binding of the licensing proteins Mcm2–7 with reduced expression of adjacent genes during the cell cycle stage G1.

Alter & Golub, *PNAS* 101, 16577 (2004);
Alter, Golub, Brown & Botstein, *Proc MNBWS* 15 (2004).

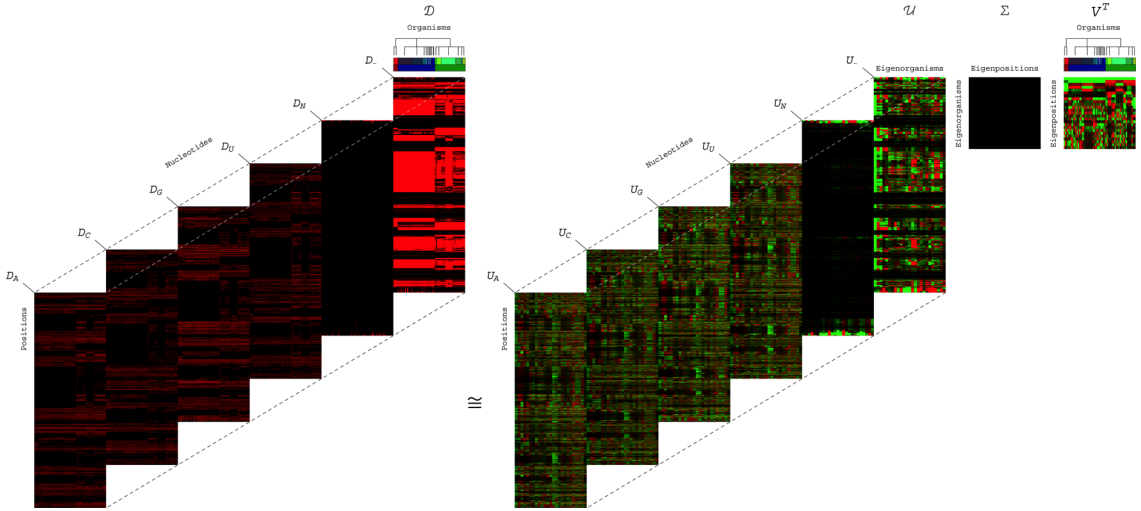
→ These experimental results are also in agreement with the equivalent correlation between overexpression of binding targets of Mcm2–7 and expression in response to oxidative stress.

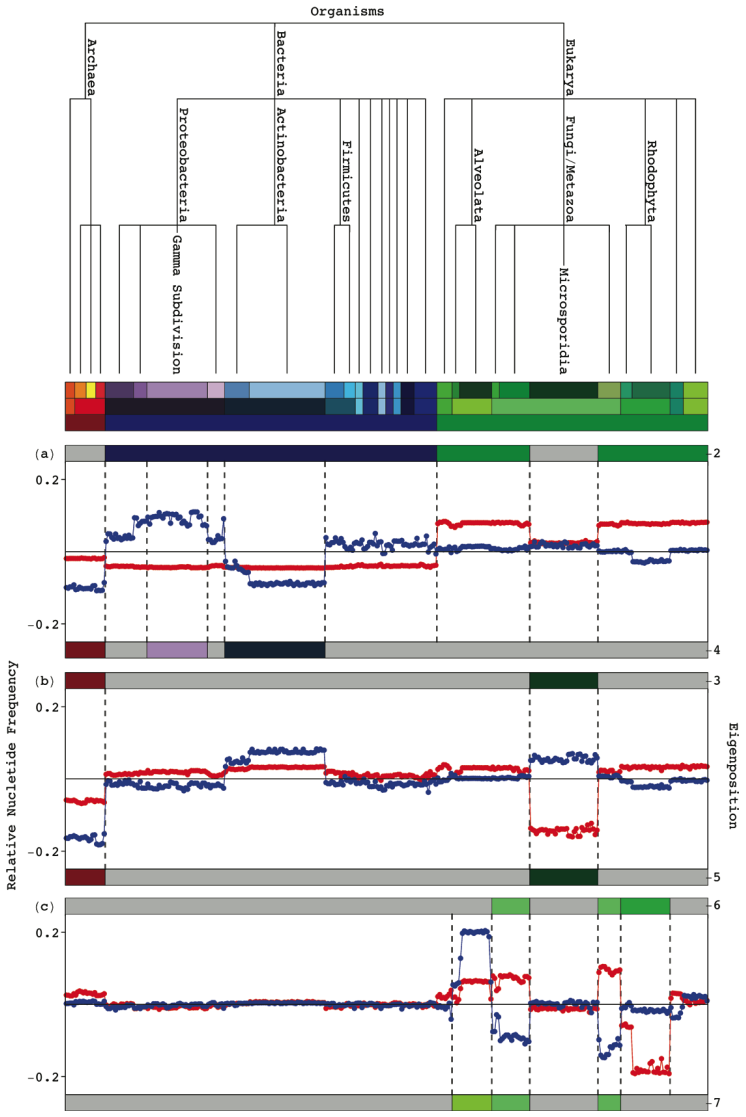
Omberg, Golub & Alter, *PNAS* 104, 18371 (2007);
Cocker, Piatti, Santocanale, Nasmyth & Diffley, *Nature* 379, 180 (1996);
Blanchard et al., *MBC* 13, 1536 (2002).

→ This demonstrates that mathematical modeling of DNA microarray data can be used to correctly predict biological mechanisms.

Mode-1 HOSVD Predicts Evolutionary Convergence and Divergence Modes and Correlations with Structural Motifs in rRNA

Muralidhara, Gross, Gutell & Alter, *PLoS One* 6, e18768 (2011); http://alterlab.org/rRNA/

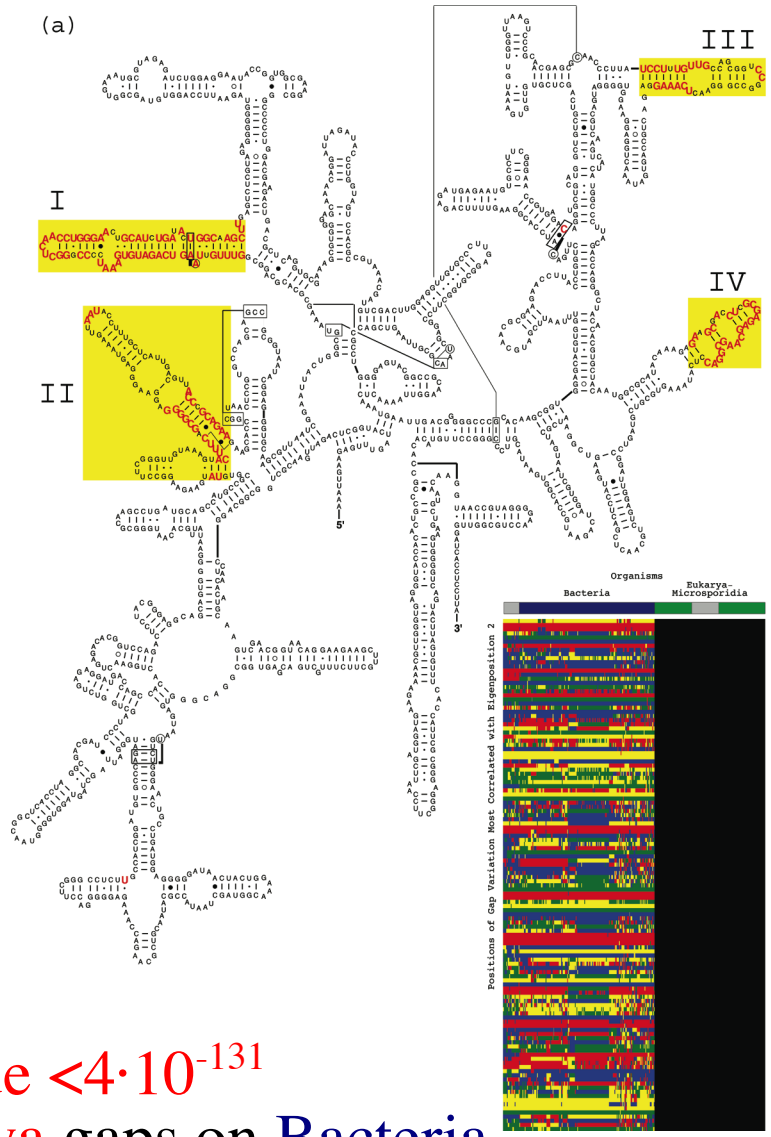


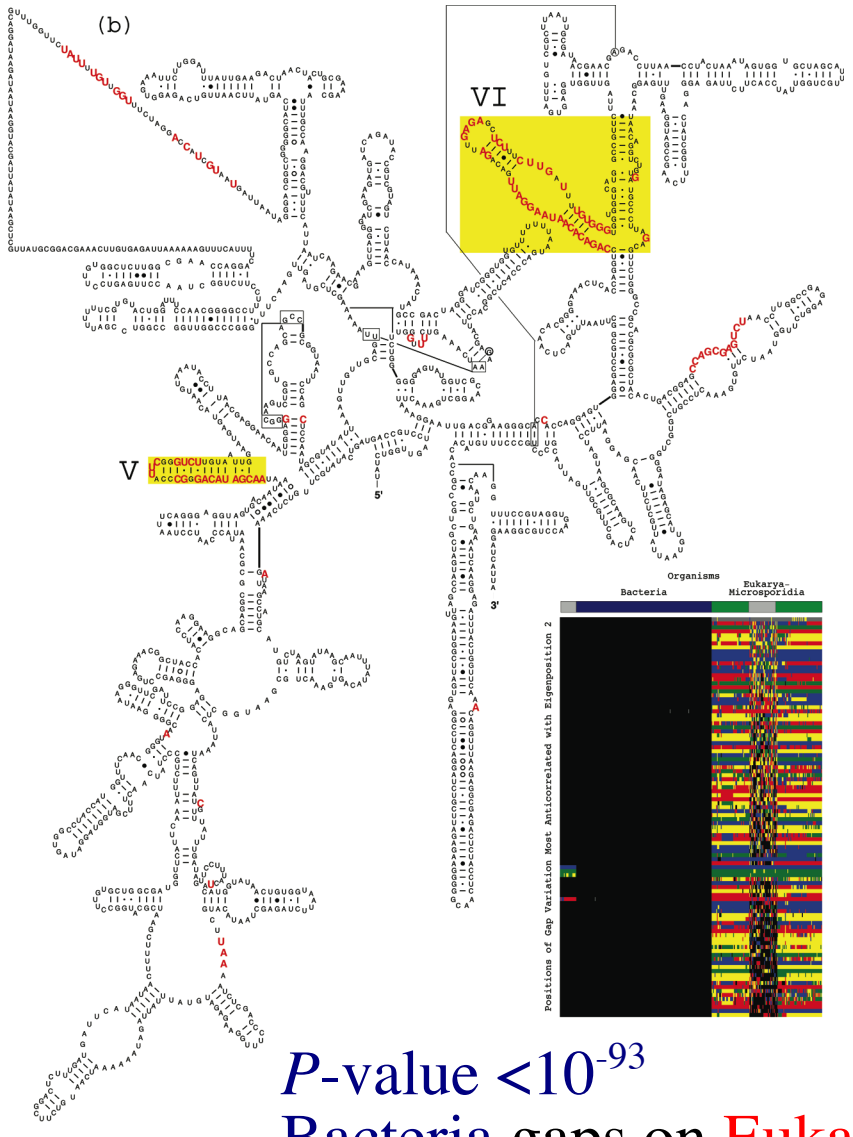


Even on the level of a single rRNA molecule, an organism's evolution is composed of multiple pathways due to concurrent forces that act independently upon different rRNA degrees of freedom.

Mode-1 HOSVD uncovers patterns of similar and dissimilar nucleotide frequency variation across the taxonomic groups, consistent between 16S and 23S rRNAs.

Corresponding Nucleotide-Specific Variations Across the Positions Map Out Known and New Insertions and Deletions of Substructure





P-value $<4 \cdot 10^{-131}$
Eukarya gaps on Bacteria

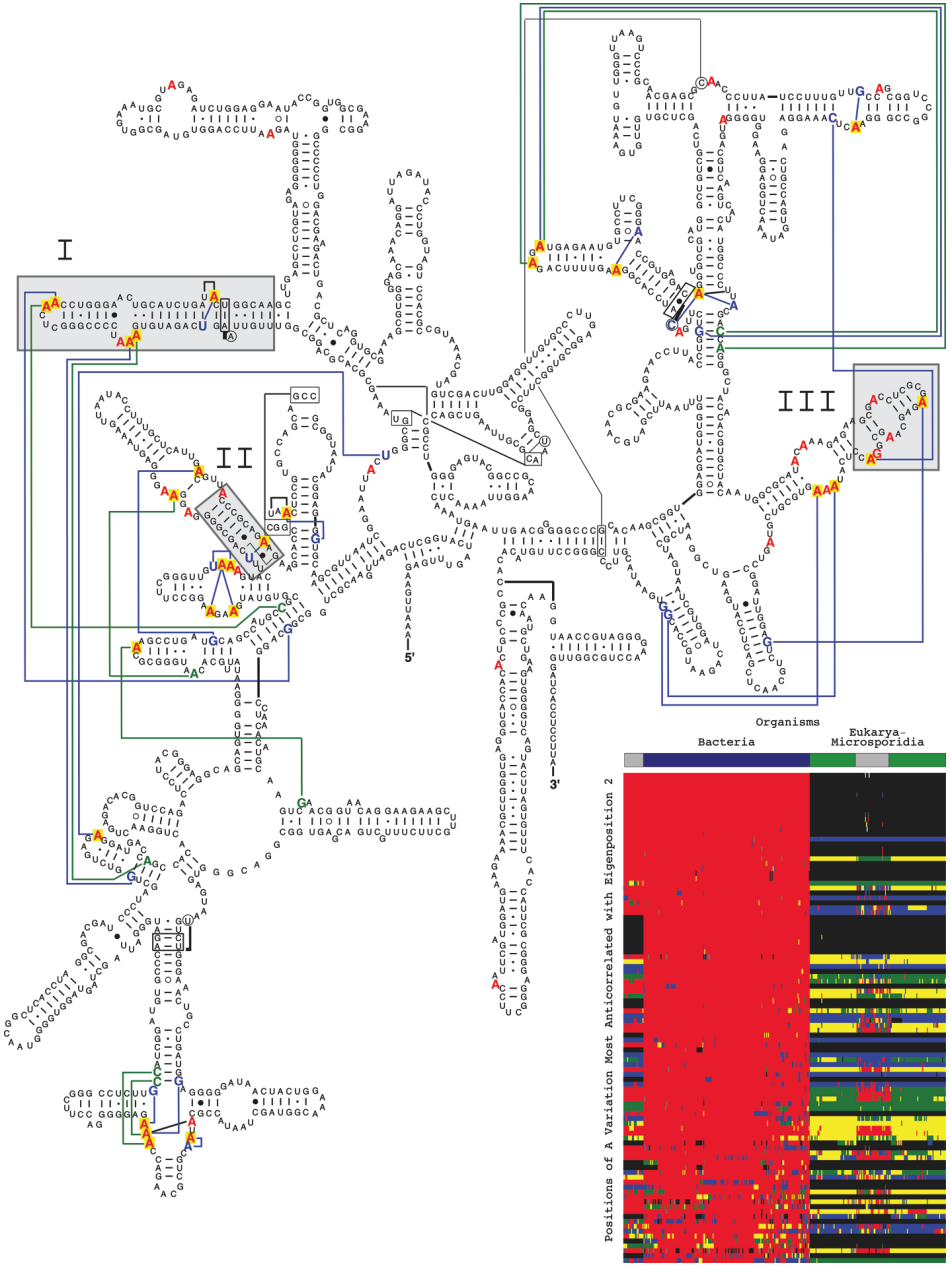
P-value $<10^{-93}$
Bacteria gaps on Eukarya

Nucleotide-Specific Variations Across the Positions Enriched in Unpaired Adenosines

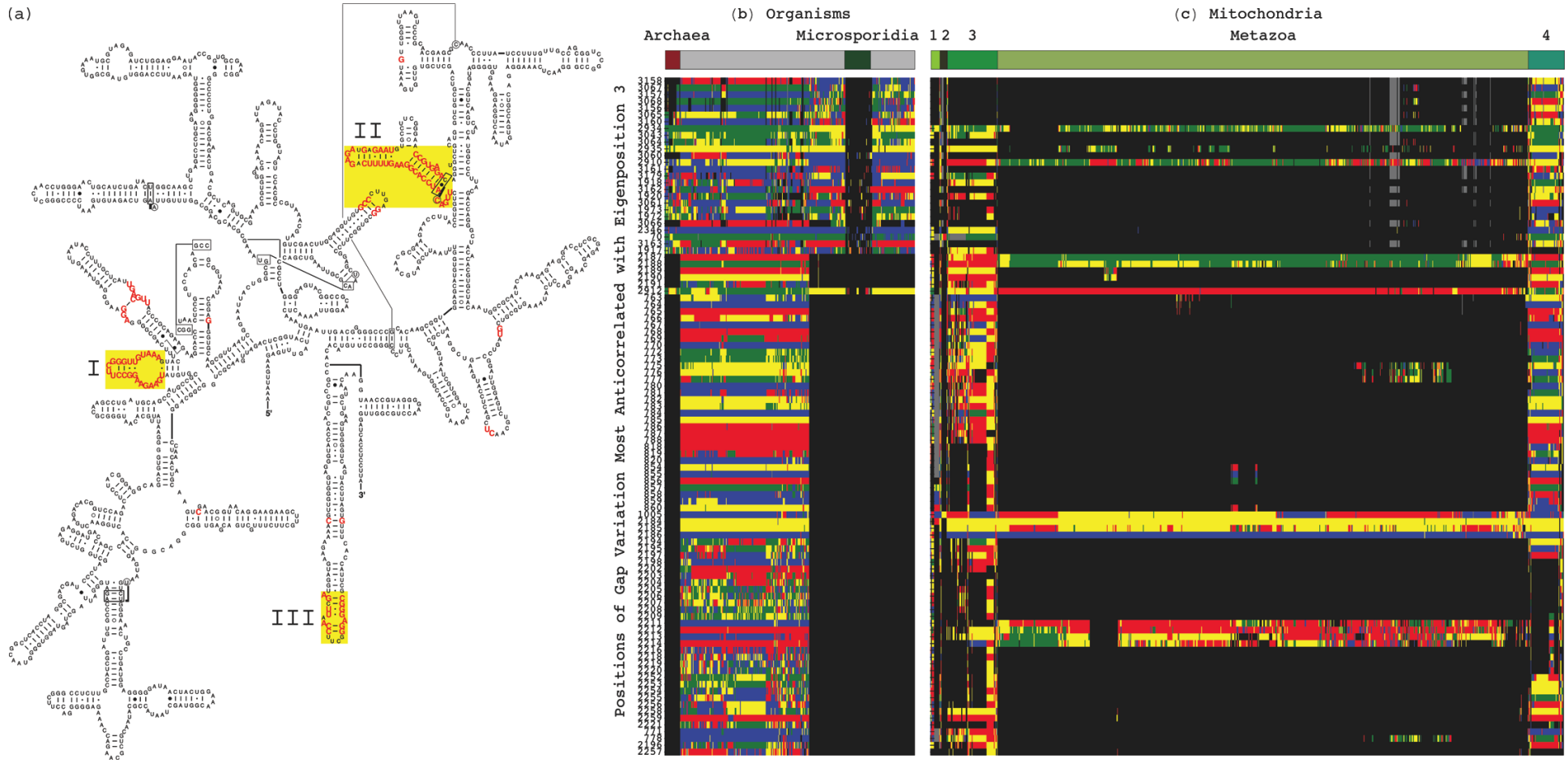
Adenosines, unpaired in the rRNA secondary structure, participate in tertiary structure interactions and are involved in rRNA folding and function.

All 50 unpaired adenosines conserved exclusively in Bacteria are significant in differentiating Bacteria from Eukarya (a *P*-value ~10^{-82}).

The crystal structure of the bacterium *Thermus thermophilus* reveals that 28 of these are involved in tertiary interactions.



Two Novel Coexsiting Subgenic Relationships between Archaea and Microsporidia



Archaea and Microsporidia share gaps that map out substructures and are enriched in unpaired adenosines, also missing in Metazoan Mitochondria, relative to Bacteria. Deletions and insertions of substructures and unpaired adenosines distinguish Archaea from Microsporidia 16S and 23S rRNAs.

Higher-Order GSVD for Comparison of mRNA Expression from Multiple Organisms

Ponnapalli, Saunders, Van Loan & Alter, *PLoS One* 6, e28072 (2011); http://alterlab.org/HO_GSVD/

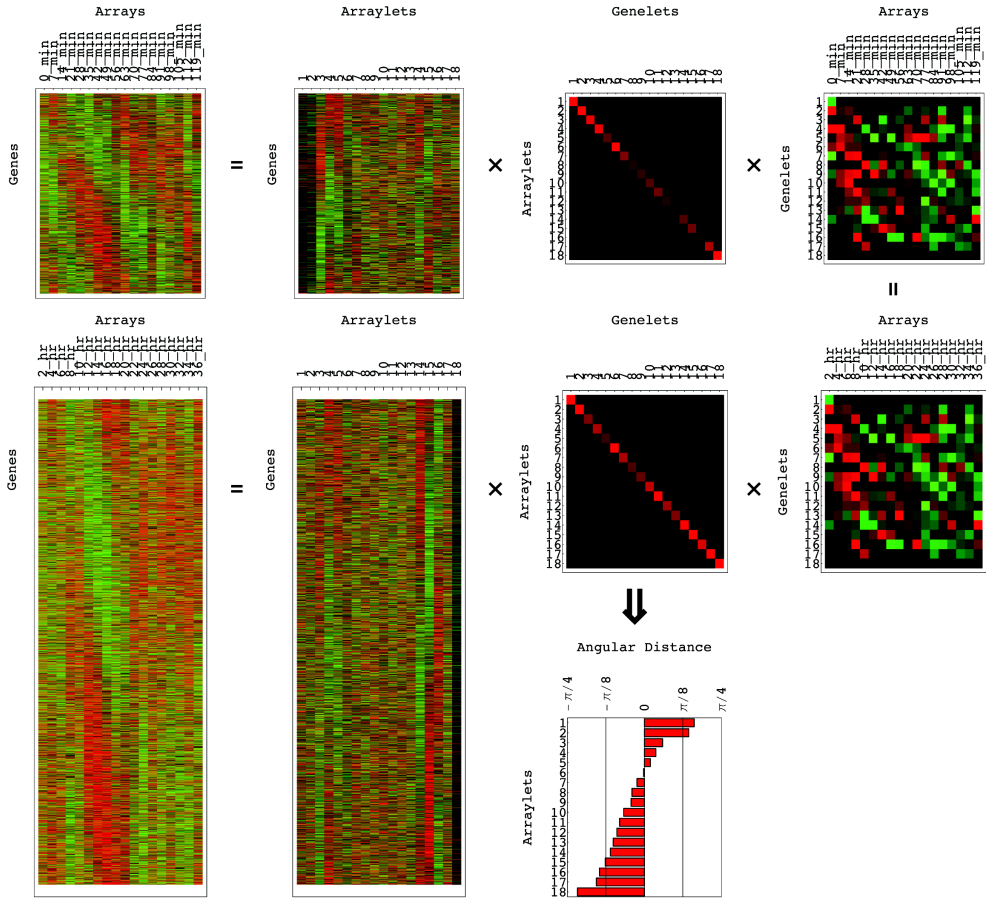
Ponnapalli, Golub & Alter, *Stanford University and Yahoo! Research Workshop on Algorithms for Modern Massive Datasets* (June 21–24, 2006, Stanford, CA).

The number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing in many areas of science.

This is accompanied by a need for mathematical frameworks that can compare multiple large-scale matrices with different row dimensions.

The only such framework to date, the GSVD, is limited to two matrices.

Yeast Spellman et al. *MBC* 9, 3273 (1998).



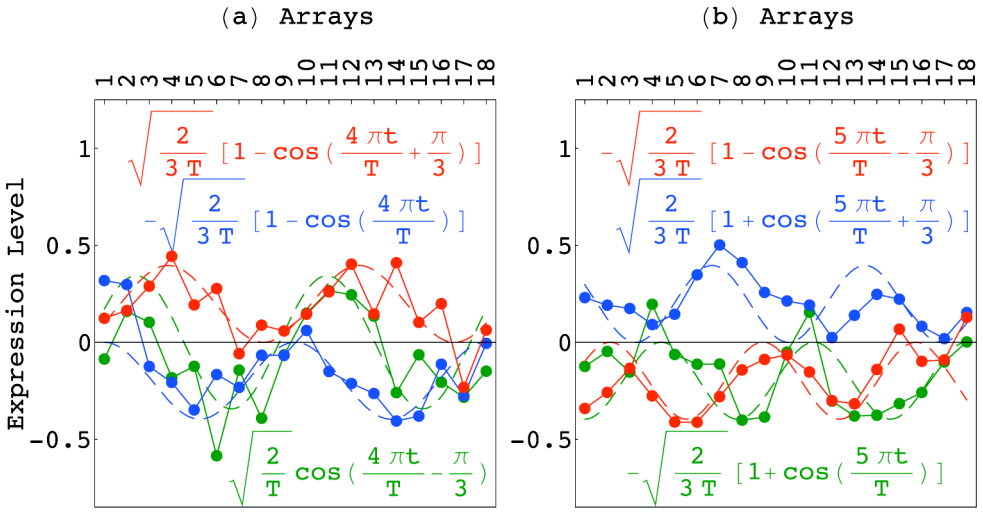
Human Whitfield et al. *MBC* 13, 1977 (2002).

Alter, Brown & Botstein, *PNAS* 100, 3351 (2003).

Math Variables → Biology

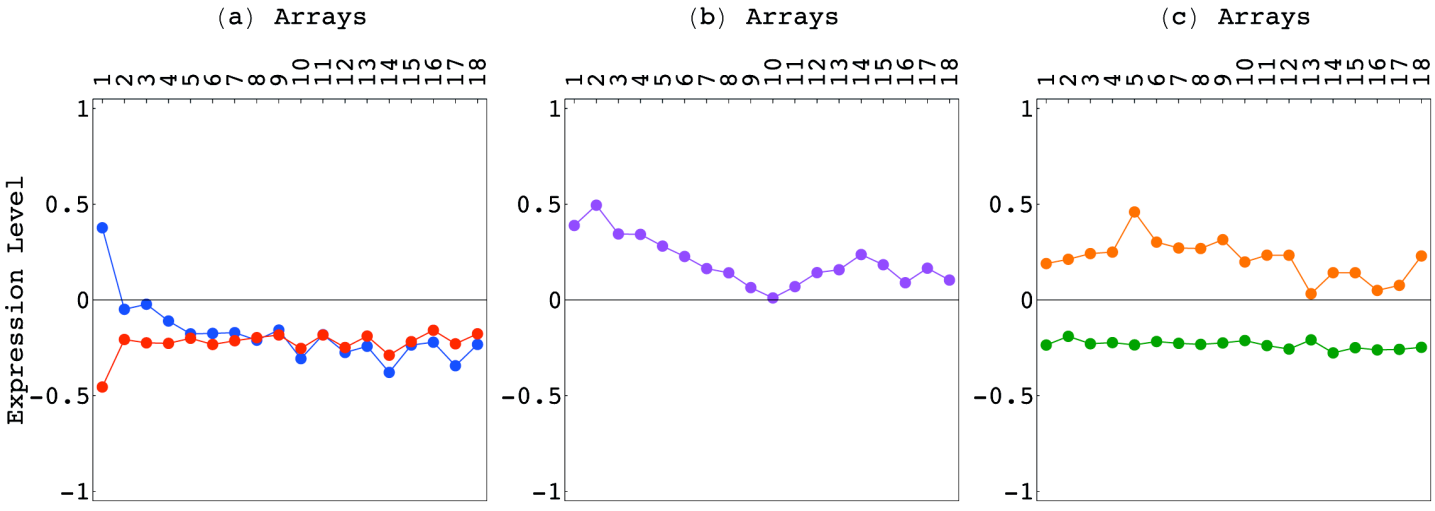
Genelets of almost equal significance in both datasets → processes common to both genomes:

Common Cell Cycle Subspace



Genelets of almost no significance in one dataset relative to the other → genome exclusive processes:

Exclusive Synchronization Responses Subspaces



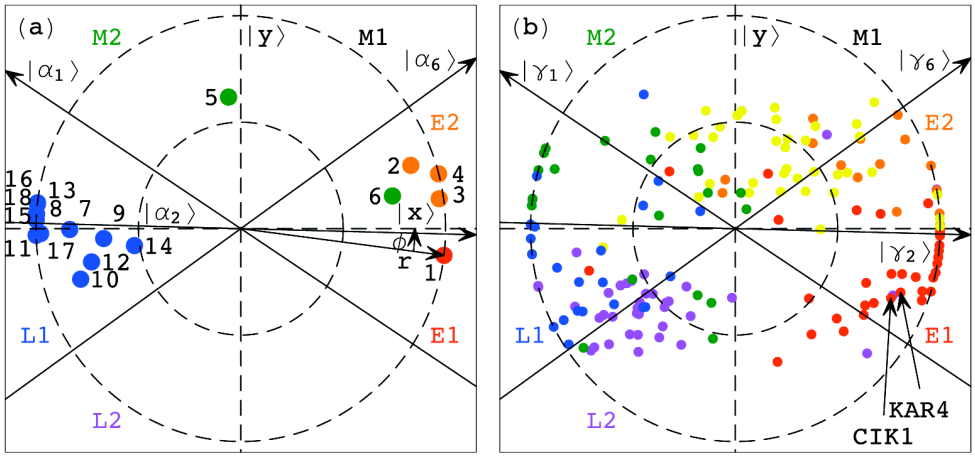
← *Saccharomyces cerevisiae* Human →

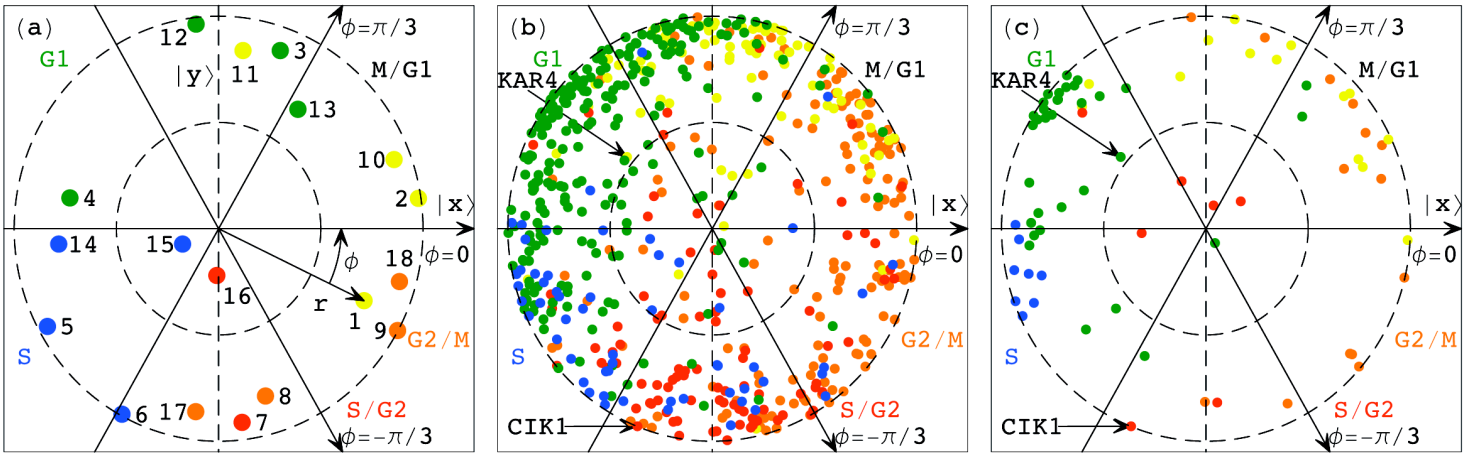
Math Operations → Biology

Data reconstruction in two subspaces → experimental observation of differential expression of a genome in the two cellular programs these subspaces represent:

Differential Expression in Yeast During Mating and Cell Cycle

Pheromone Synchronization Response Subspace: KAR4 is required for CIK1 induction during mating



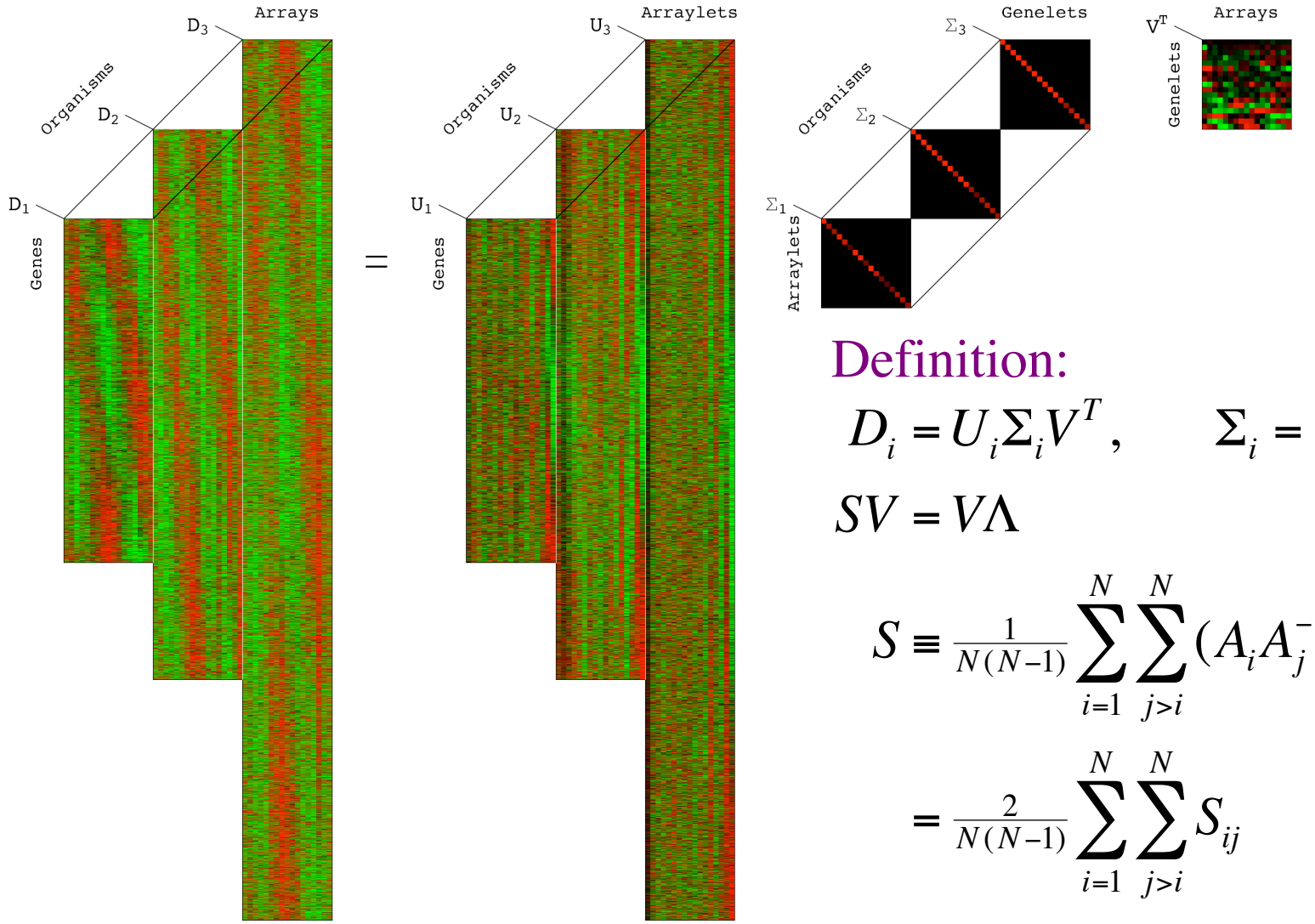


Common Cell Cycle Subspace: Mitotic expression of CIK1 during S/G2 is independent of KAR4

Kurihara, Stewart, Gammie & Rose, *MCB* 16, 3990 (1996).

Mathematical Definition of a Novel HO GSVD

Ponnapalli, Golub & Alter, *Stanford University and Yahoo! Research Workshop on Algorithms for Modern Massive Datasets* (June 21–24, 2006, Stanford, CA).



Definition:

$$D_i = U_i \Sigma_i V^T, \qquad \Sigma_i = \mathrm{diag}(\sigma_{i,k})$$

$$SV = V\Lambda$$

$$S \equiv \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i}^{N} (A_i A_j^{-1} + A_j A_i^{-1})$$

$$= \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i}^{N} S_{ij}$$

Assumption: $D_i \in \mathcal{R}^{m_i \times n}$

$$A_i = D_i^T D_i, \qquad S_{ij} = \tfrac{1}{2}(A_i A_j^{-1} + A_j A_i^{-1})$$

The matrix V, identical in all factorizations, is obtained from the balanced eigensystem of S, which does not depend upon the ordering of D_i.

Mathematical Properties of the HO GSVD

Ponnapalli, Saunders, Van Loan & Alter, *PLoS One* 6, e28072 (2011); http://alterlab.org/HO_GSVD/

This exact decomposition extends to higher orders all of the mathematical properties of the GSVD except for complete orthogonality of U_i for all i.

Supplementary Theorems 1–5:
For N=2, our HO GSVD leads algebraically to the GSVD.

Theorem 1: S has n independent eigenvectors, and the eigenvectors and eigenvalues of S are real.

Theorem 2: The eigenvalues of S satisfy $\lambda_k \geq 1$.

Theorem 3: **The common HO GSVD subspace.** An eigenvalue satisfies $\lambda_k=1$ if and only if the corresponding right basis vector v_k is of equal significance in all matrices D_i and D_j, i.e., $\sigma_{i,k} / \sigma_{j,k}=1$ for all i and j, and the corresponding left basis vector $u_{i,k}$ is orthonormal to all other left basis vectors in U_i for all i.

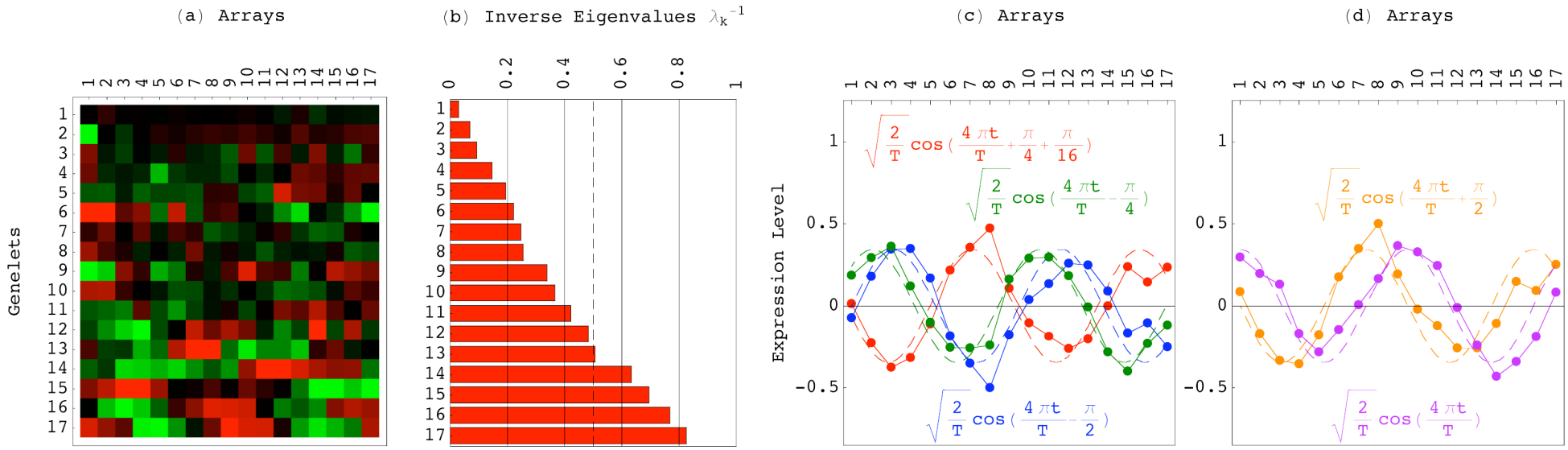
Corollary 1: An eigenvalue satisfies $\lambda_k=1$ if and only if the corresponding right basis vector v_k is a generalized singular vector of all pairwise GSVD factorizations of the matrices D_i and D_j with equal corresponding generalized singular values for all for all i and j.

Supplementary Theorem 6 and Conjecture 1:
A role in iterative approximation algorithms.

Math Variables → Biology

Genelets of almost equal significance in all datasets →
processes common to all genomes:

Approximately Common HO GSVD Subspace



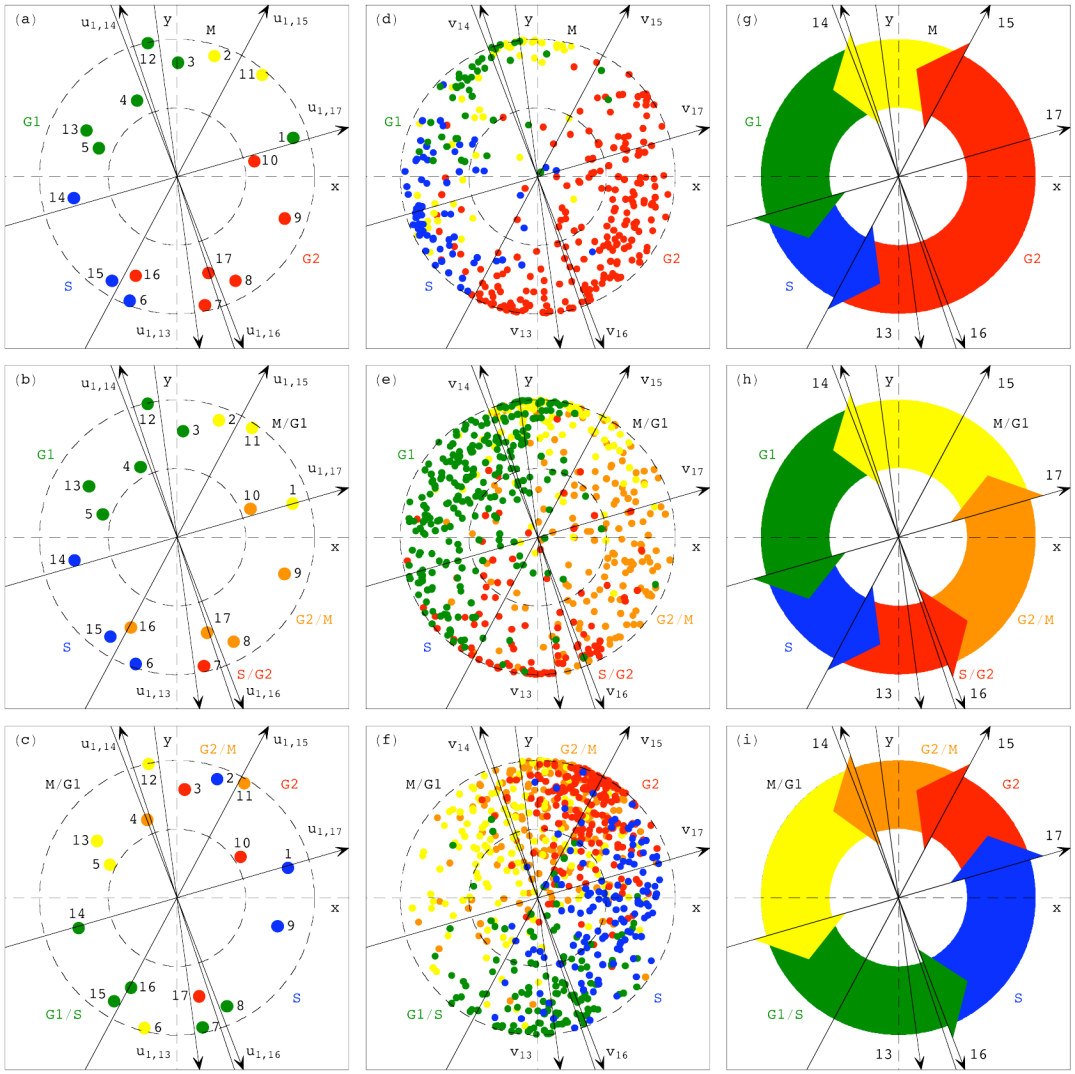
In a comparison of global cell cycle mRNA expression from *S. pombe*, *S. cerevisiae* and human, the approximately common HO GSVD subspace represents the cell cycle mRNA expression oscillations, which are similar among the datasets.

Simultaneous reconstruction in the common subspace, therefore, removes the experimental artifacts, which are dissimilar, from the datasets.

Math Operations → Biology

Simultaneous classification in the common HO GSVD subspace → biological similarity in the regulation of the cellular programs that are conserved across the species:

Common Cell Cycle Subspace



Schizosaccharomyces pombe

Rustici et al. *Nat. Genet.* 36, 809 (2004).

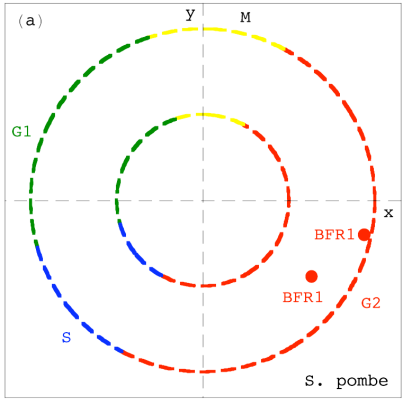
Saccharomyces cerevisiae

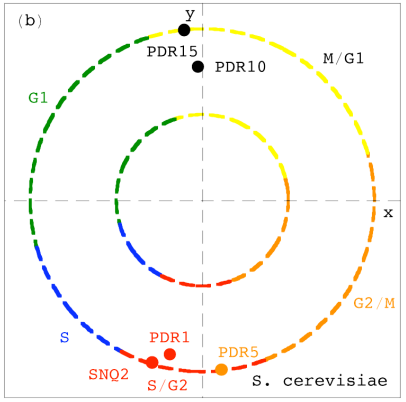
Spellman et al. *MBC* 9, 3273 (1998).

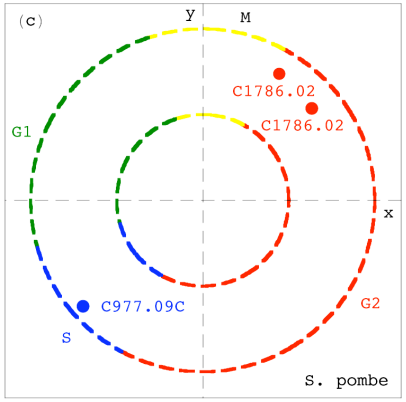
Human

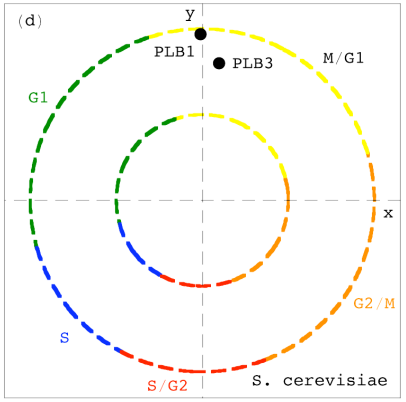
Whitfield et al. *MBC* 13, 1977 (2002).

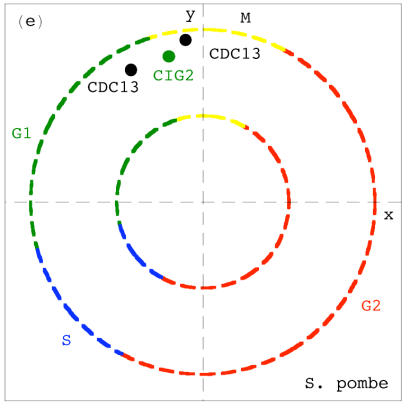
Simultaneous Classification Independent of Sequence Similarity

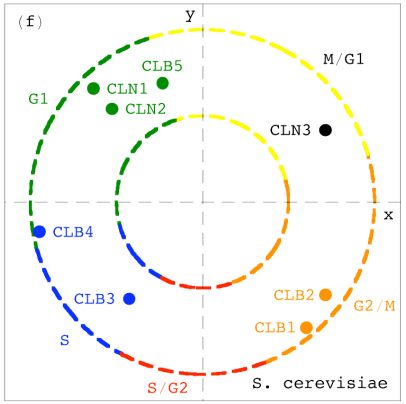


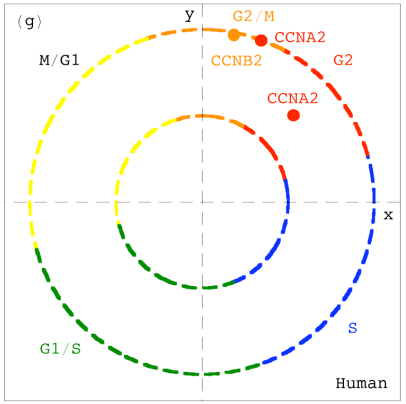












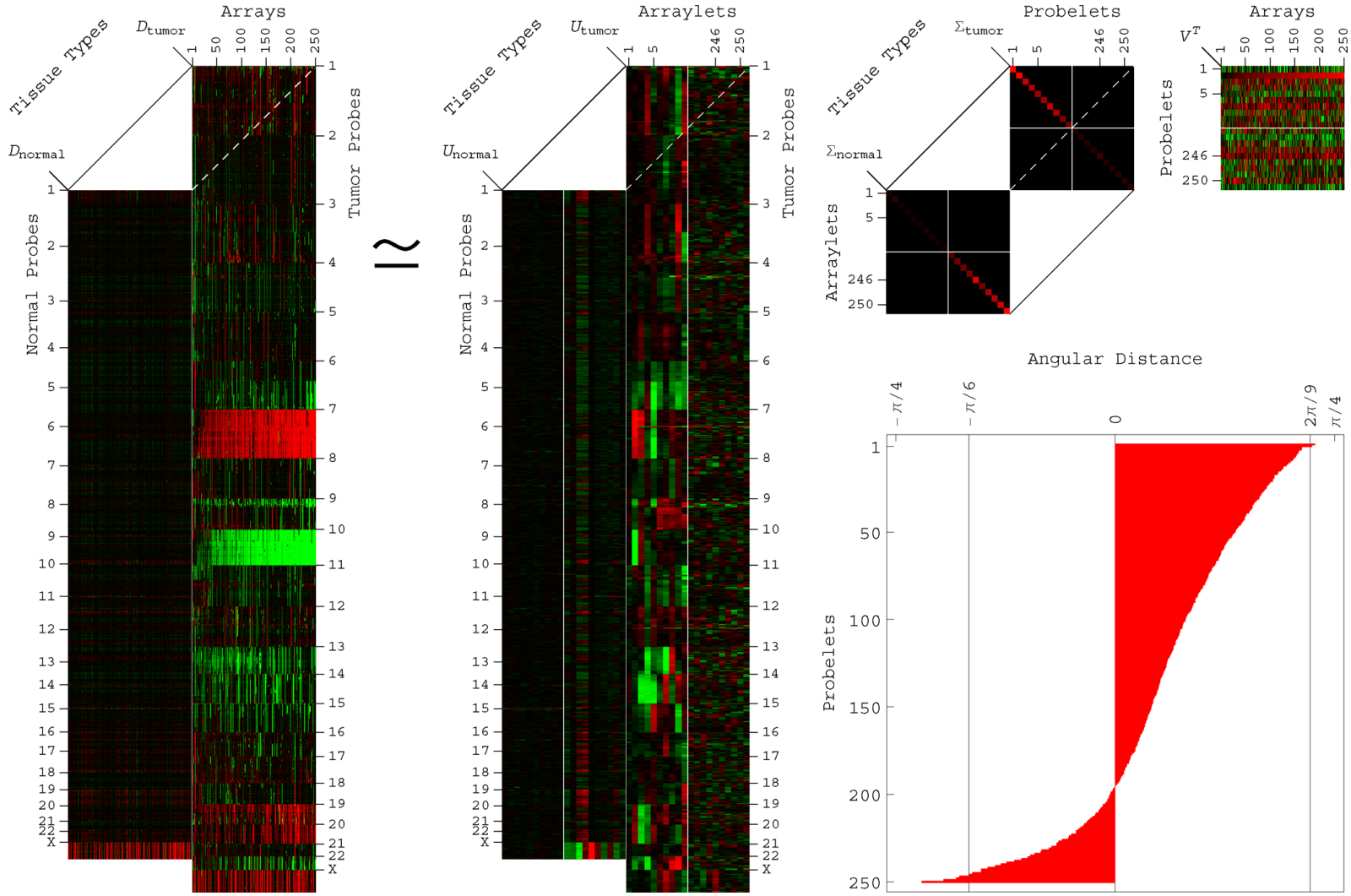
Genes of highly conserved sequences across the three organisms but significantly different cell cycle peak times are correctly classified.

ABC Transporter Superfamily Genes

Phospholipase B-Encoding Genes and B Cyclin-Encoding Genes

GSVD for Comparison of Patient-Matched Tumor and Normal Genomic Profiles

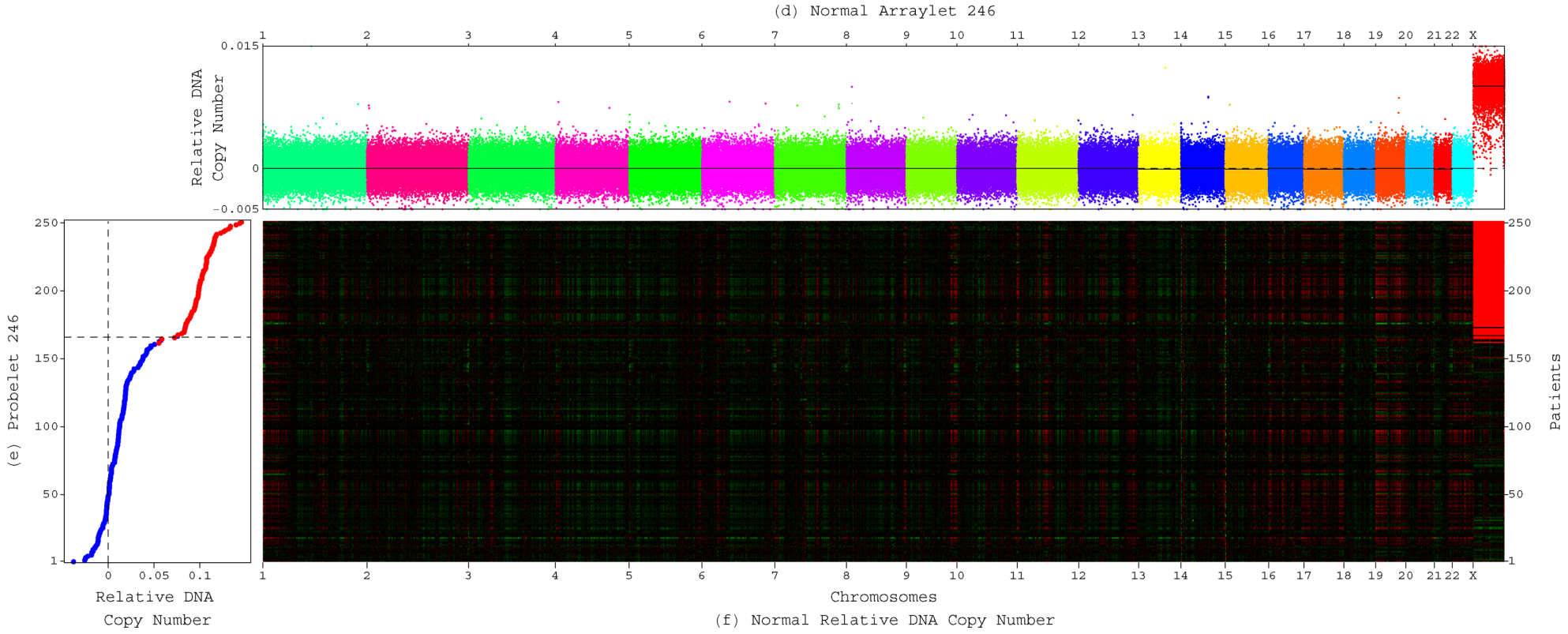
Lee,* Alpert,* Sankaranarayanan & Alter, *PLoS One* 7, e30098 (2012);
http://alterlab.org/GBM_prognosis/



The number of large-scale datasets recording multiple aspects of a single phenomenon is increasing in many areas, e.g., personalized medicine.

Copy-Number Variations (CNVs) Common to the GBM Tumor and Normal Brain

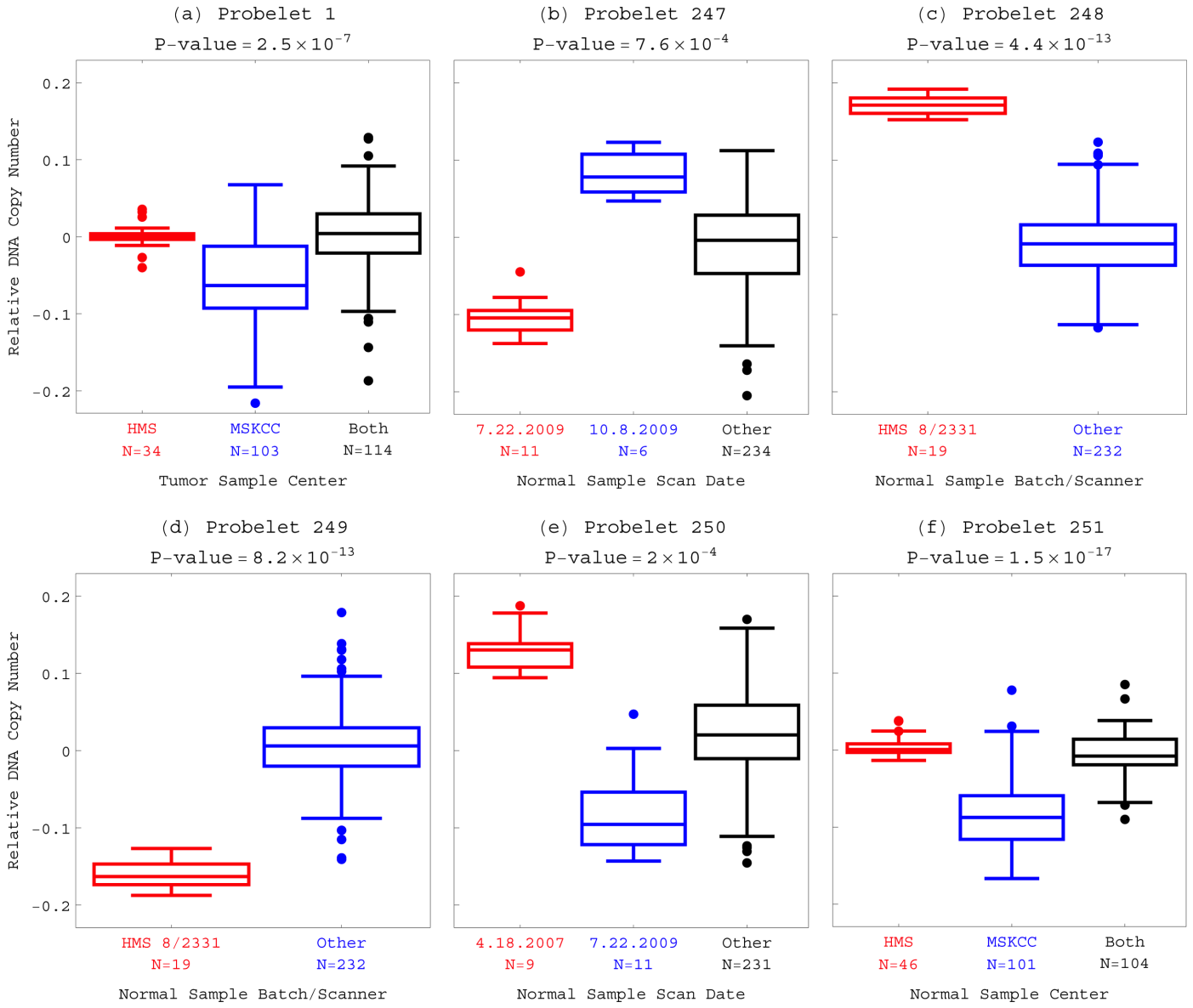
GSVD identifies CNVs that occur in the normal human genome and are preserved in the GBM tumors, e.g., female-specific X chromosome amplification, without a-priori knowledge of these variations.



Notice of the National Human Genome Research Institute's Interest in Receiving Applications to Analyze and Develop Methods for X Chromosome Genome-wide Association (GWA) Data; http://grants.nih.gov/grants/guide/notice-files/NOT-HG-11-021.html

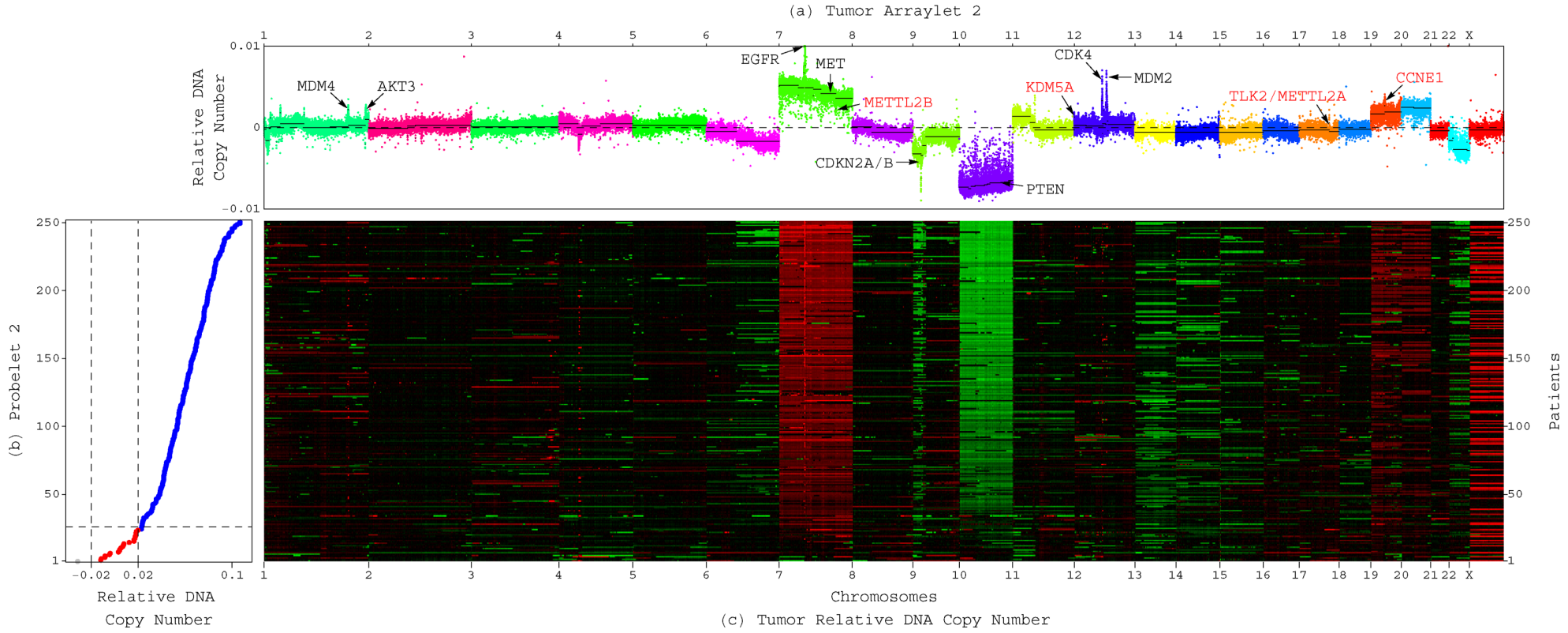
Experimental Variations Exclusive to the Tumor or Normal Profiles

GSVD identifies experimental variations, e.g., in tissue batch, genomic center, hybridization date and scanner.



Global Pattern of Tumor-Exclusive Aberrations Predicts Drug Targets

Lee & Alter, *60th Annual Meeting of the ASHG* (Washington, DC, November 2–6, 2010).



The pattern includes most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of the patients: the biochemically putative drug target, cell cycle-regulated serine/threonine kinase-encoding *TLK2*, the tRNA methyltransferase *METTL2A*, and the cyclin E1-encoding *CCNE1*.

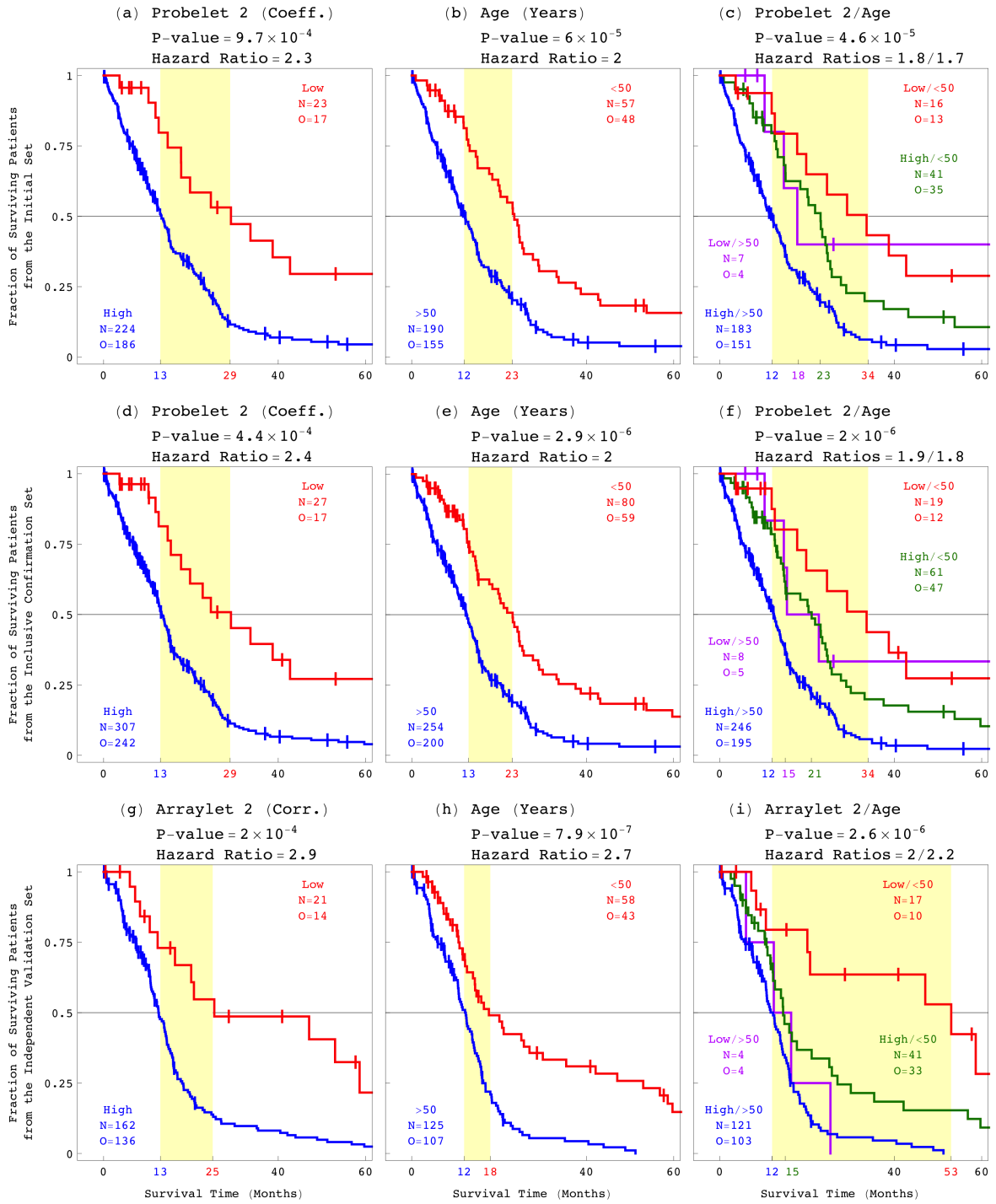
Global Predictor of GBM Survival

The global pattern is correlated with, and possibly causally related to, brain cancer survival.

The GBM survival phenotype is the outcome of its global genotype.

Despite recent large-scale profiling efforts, the best prognostic predictor of GBM prior to the discovery of this pattern was the patient's age at diagnosis.

The pattern is independent of age, and combined with age, makes a predictor better than age alone.



Patterns Underlie Principles of Nature: Statistics to Processes

→ Brownian motion.

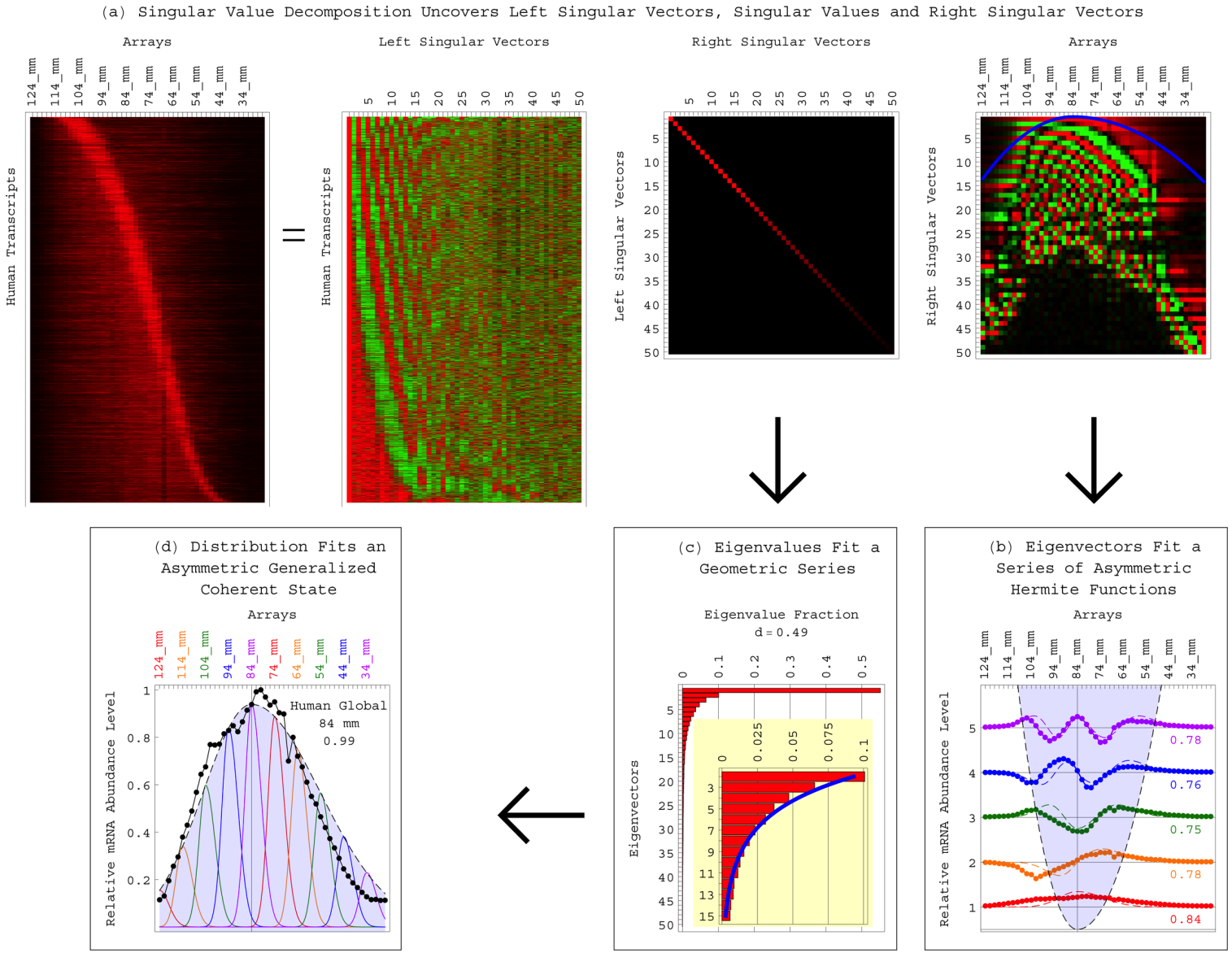
Einstein, *Ann Phys* 17, 549 (1905).

→ Bacterial sensitivity and resistance to viruses.

Luria & Delbrück, *Genetics* 28, 491 (1943).

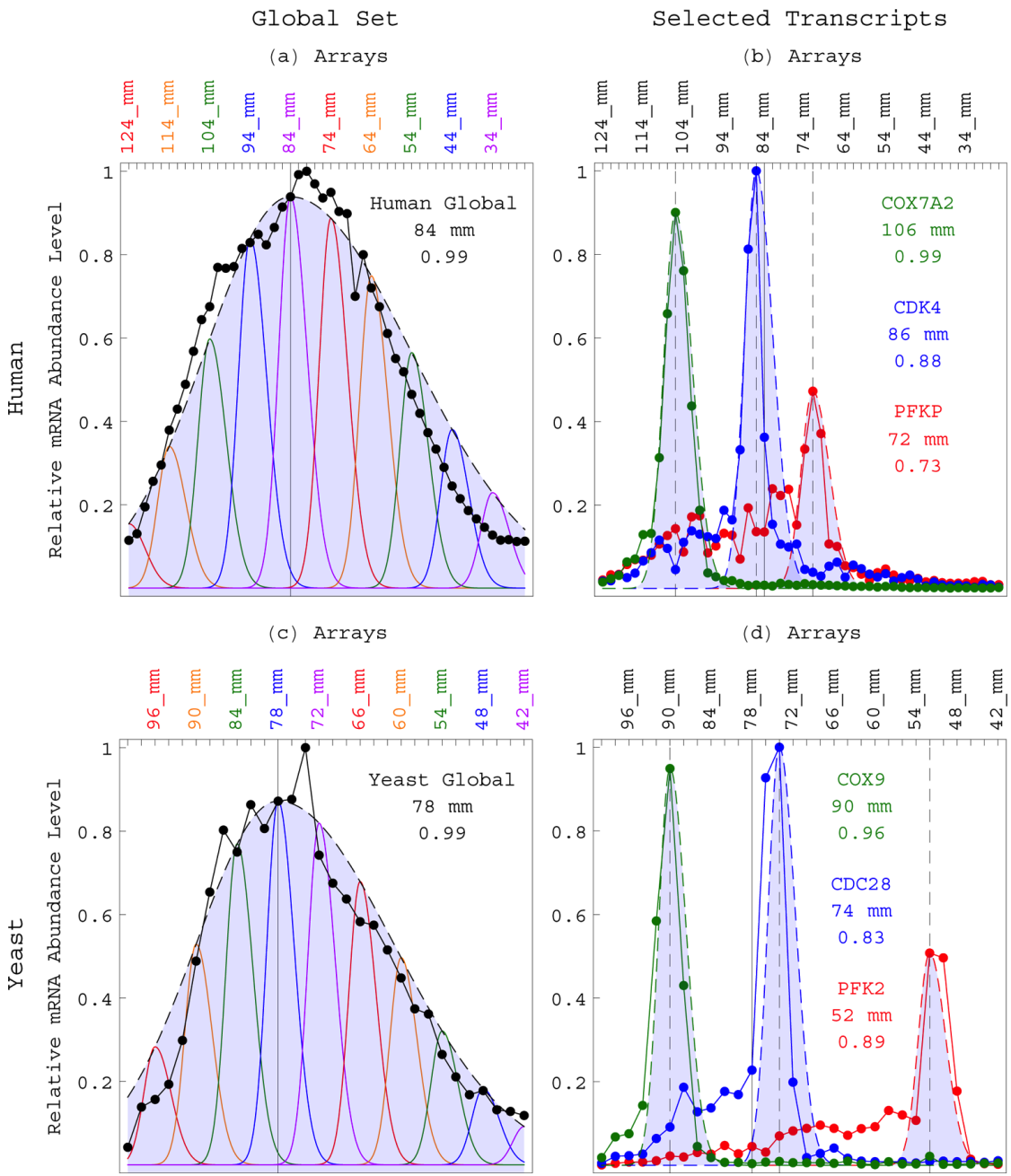
SVD Identifies Transcript Length Distribution Functions from DNA Microarray Data

Alter & Golub, *PNAS* 103, 11828 (2006); http://alterlab.org/harmonic_oscillator/



Hurowitz et al., *PLoS One* 2, e460 (2007); Hurowitz & Brown, *Genome Biology* 5, R2 (2003).

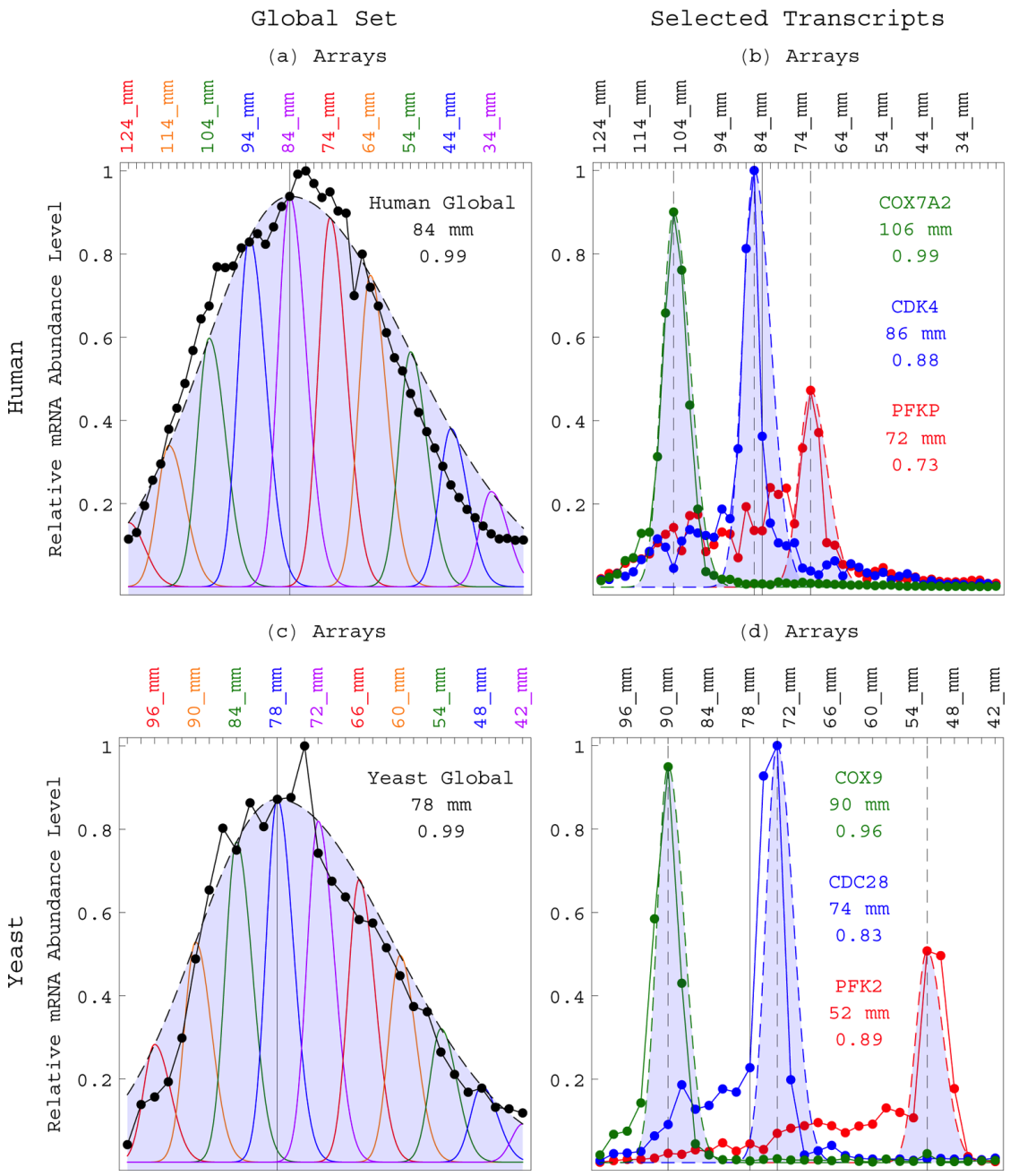
Transcript Length Distribution Functions are "Asymmetric" Coherent States



→ The profile of a single transcript fits an asymmetric Gaussian.

→ The distribution of the peaks of the transcript profiles fits an asymmetric Gaussian.

Transcript Length Distribution Functions are "Asymmetric" Coherent States



Prediction:

The asymmetry of the profile of a single transcript might be due to an asymmetry in the Brownian motion or thermal broadening of a moving rather than a stationary band of identical transcripts.

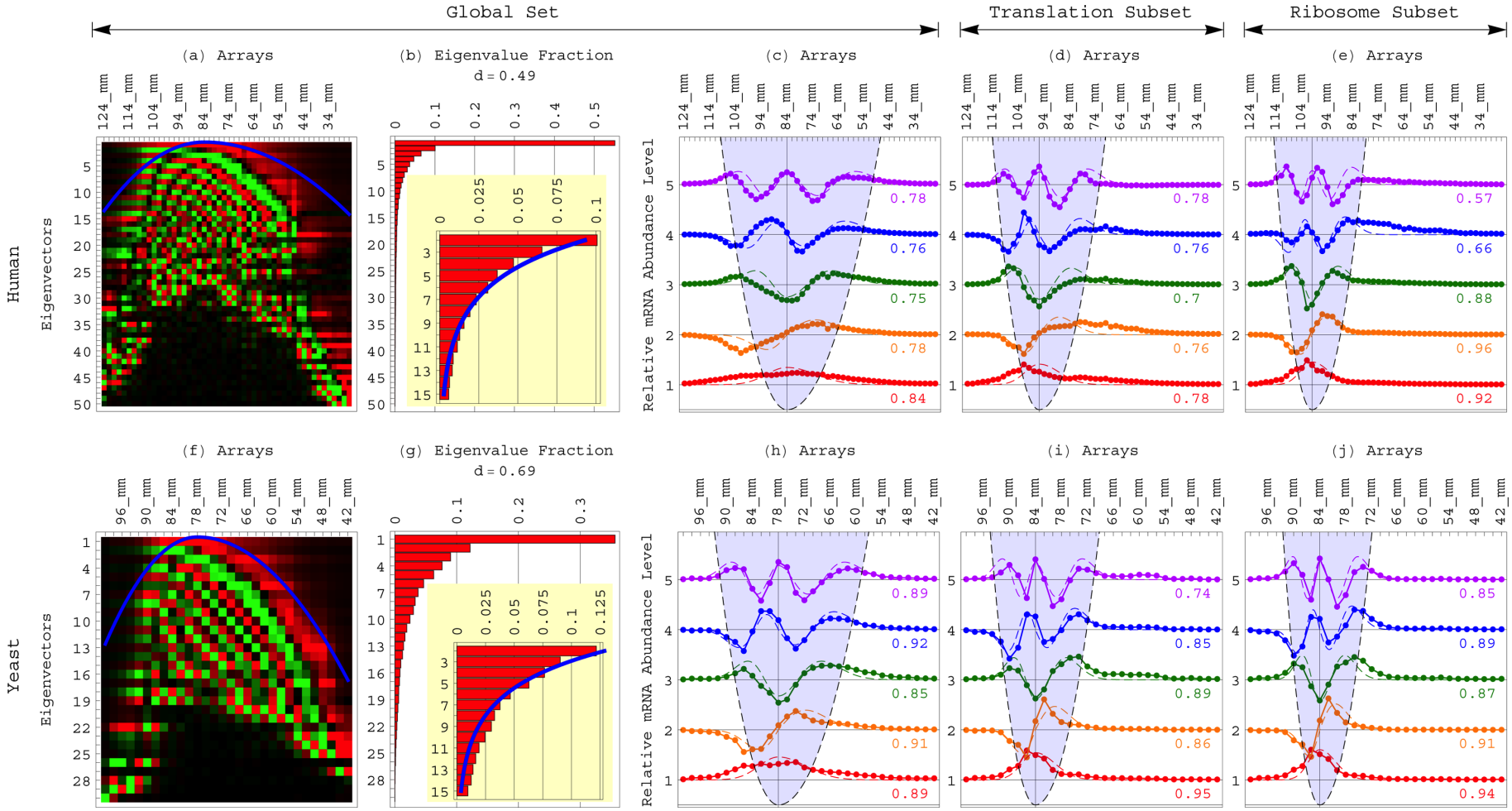
→ Modeling of genomic data can be used to predict physical principles.

Hypothesis:

Two competing evolutionary forces determine transcript lengths in the manner of the restoring force of the harmonic oscillator.

Conserved Relations between a Gene's Metabolic Ontology and its Transcript's Length

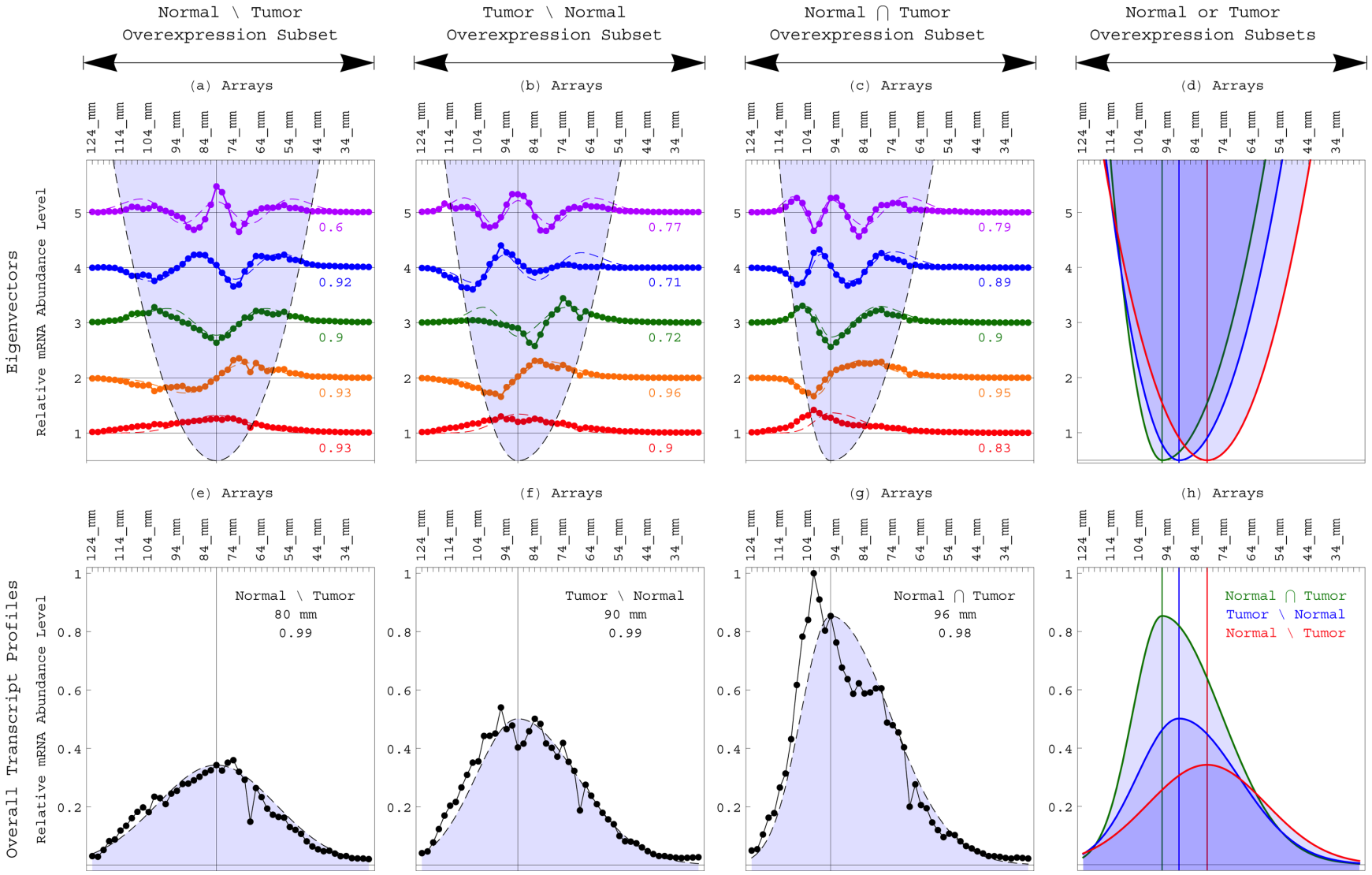
Drake & Alter, *Rao Conference at the Interface between Statistics and the Sciences* (December 30, 2009 – January 2, 2010, Hyderabad, India), Rao Best Poster Prize.



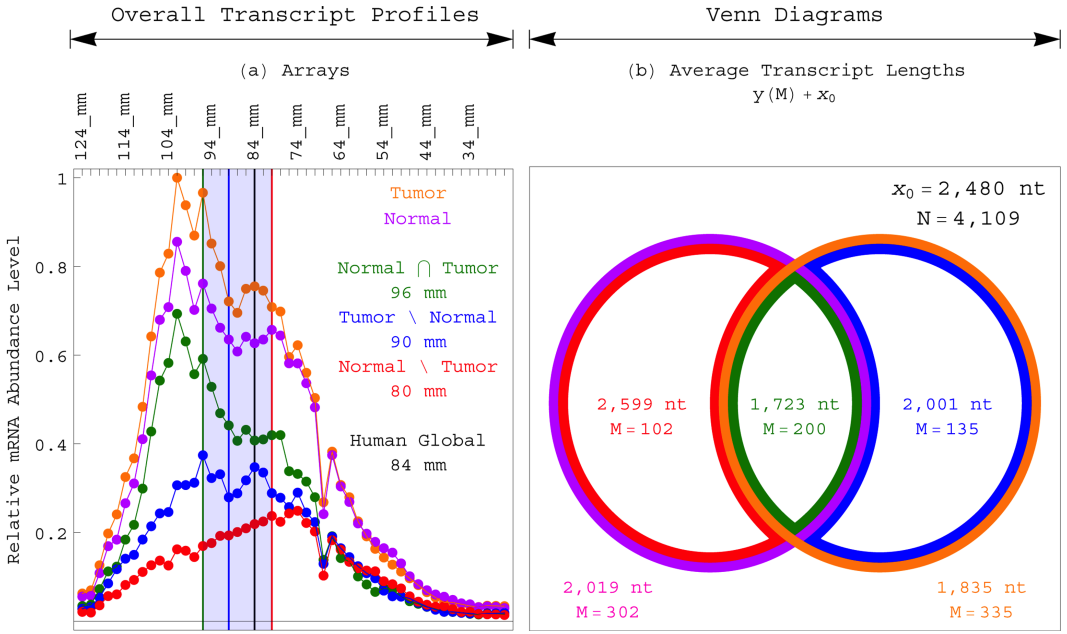
Transcripts involved in protein synthesis or mitochondrial metabolism are significantly shorter than typical, and in particular, significantly shorter than those involved in glucose metabolism.

GBM Tumors Maintain Normal Brain Overexpression of Short Transcripts but Suppress Longer, Normally Overexpressed Ones

Bertagnolli, Drake, Tennessen & Alter, *PLoS One* 8, e78913 (2013);
http://alterlab.org/GBM_metabolism/.



Global Relations among Transcript Length, Cellular Metabolism and Tumor Development



GBM tumors maintain normal brain overexpression of short transcripts, involved in protein synthesis and mitochondrial metabolism, but suppress longer, normally overexpressed transcripts, involved in glucose metabolism and brain activity.

Overexpression Subset	Gene Ontology	Global Transcript Set				Global Gene Set			
		a	B	b	P-value	a	B	b	P-value
Normal ∩ Tumor	Translation	200	178	36	4.4×10^{-14}	204	380	64	6.0×10^{-46}
	Ribosome		78	28	4.0×10^{-18}		155	52	7.1×10^{-54}
	Respiratory ETC		55	21	1.9×10^{-14}		89	22	1.1×10^{-19}
	MRCC I		25	9	1.3×10^{-6}		34	6	2.4×10^{-5}
	COX Activity		14	9	2.1×10^{-9}		20	8	8.3×10^{-10}
Normal	Glucose Metabolic Process	302	100	17	8.2×10^{-4}	309	187	14	4.7×10^{-4}
	Glycolysis		29	9	1.5×10^{-4}		59	6	4.6×10^{-3}
Normal \ Tumor	Neuron Projection	102	259	22	2.0×10^{-7}	105	534	24	4.3×10^{-11}
	Synaptic Transmission		238	19	4.0×10^{-6}		535	26	9.5×10^{-13}

Global Mode for Tumor and Normal Cells to Differentially Regulate Metabolism in a Transcript Length-Dependent Manner

Hanahan & Weinberg, *Cell* 100, 57 (2000);
Shermoen & O'Farrell, *Cell* 67, 303 (1991).

→ This shows that the functioning of a cell can be inferred from the lengths of over- and underexpressed genes, independent of the sequences of the genes.

→ A previous hypothesis from mathematical modeling of evolutionary forces that act upon transcript length in the manner of the restoring force of the harmonic oscillator is supported.

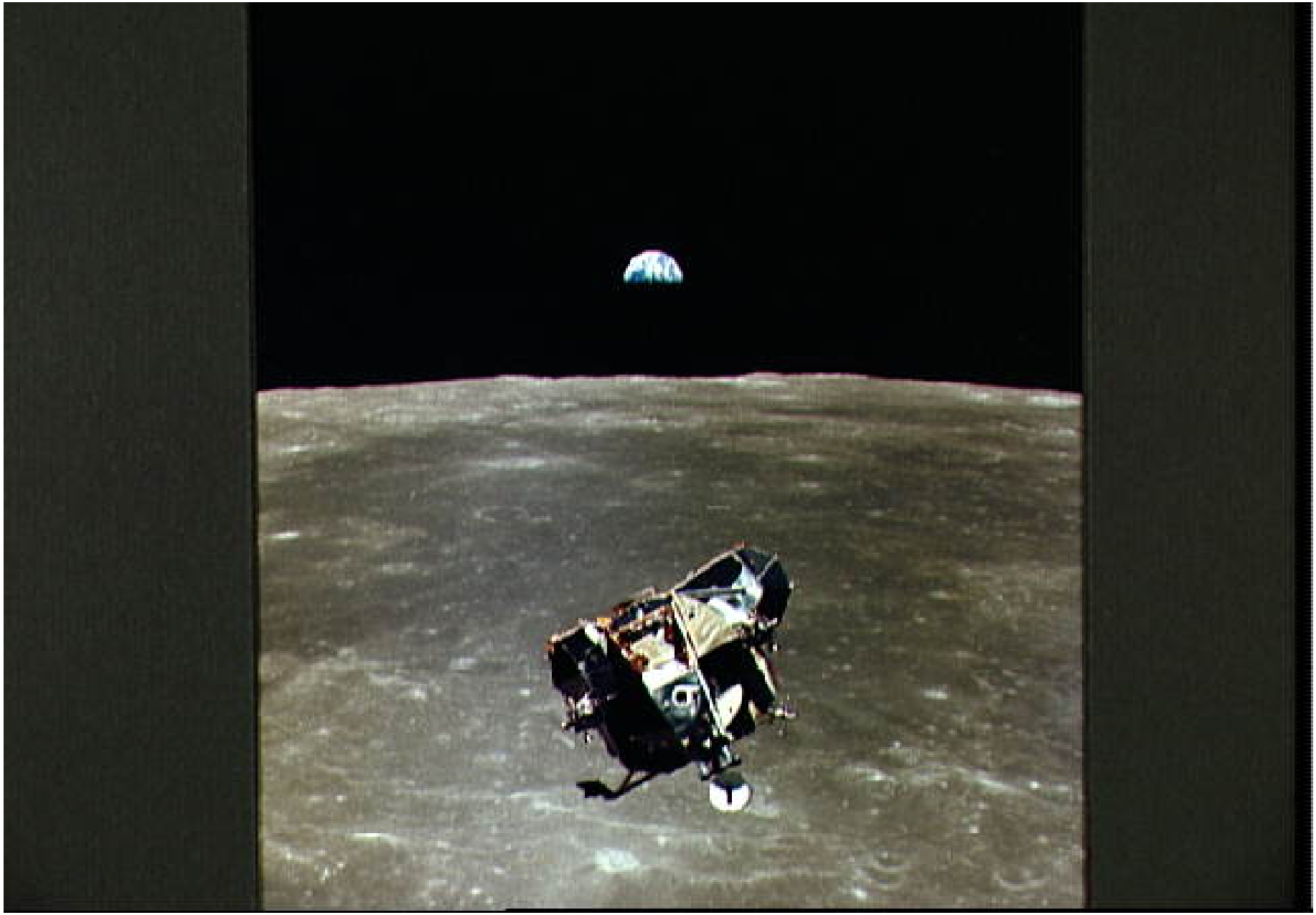
Alter & Golub, *PNAS* 103, 11828 (2006).

→ A previous prediction of asymmetry in the gel electrophoresis thermal broadening (or Brownian motion) of a moving, rather than a stationary, band of identical mRNA molecules is also supported.

Duke & Viovy, *Phys Rev Lett* 68, 542 (1992);
Slater, *Electrophoresis* 14, 1 (1993);
Tinland, Pernodet & Pluen, *Biopolymers* 46, 201 (1998).

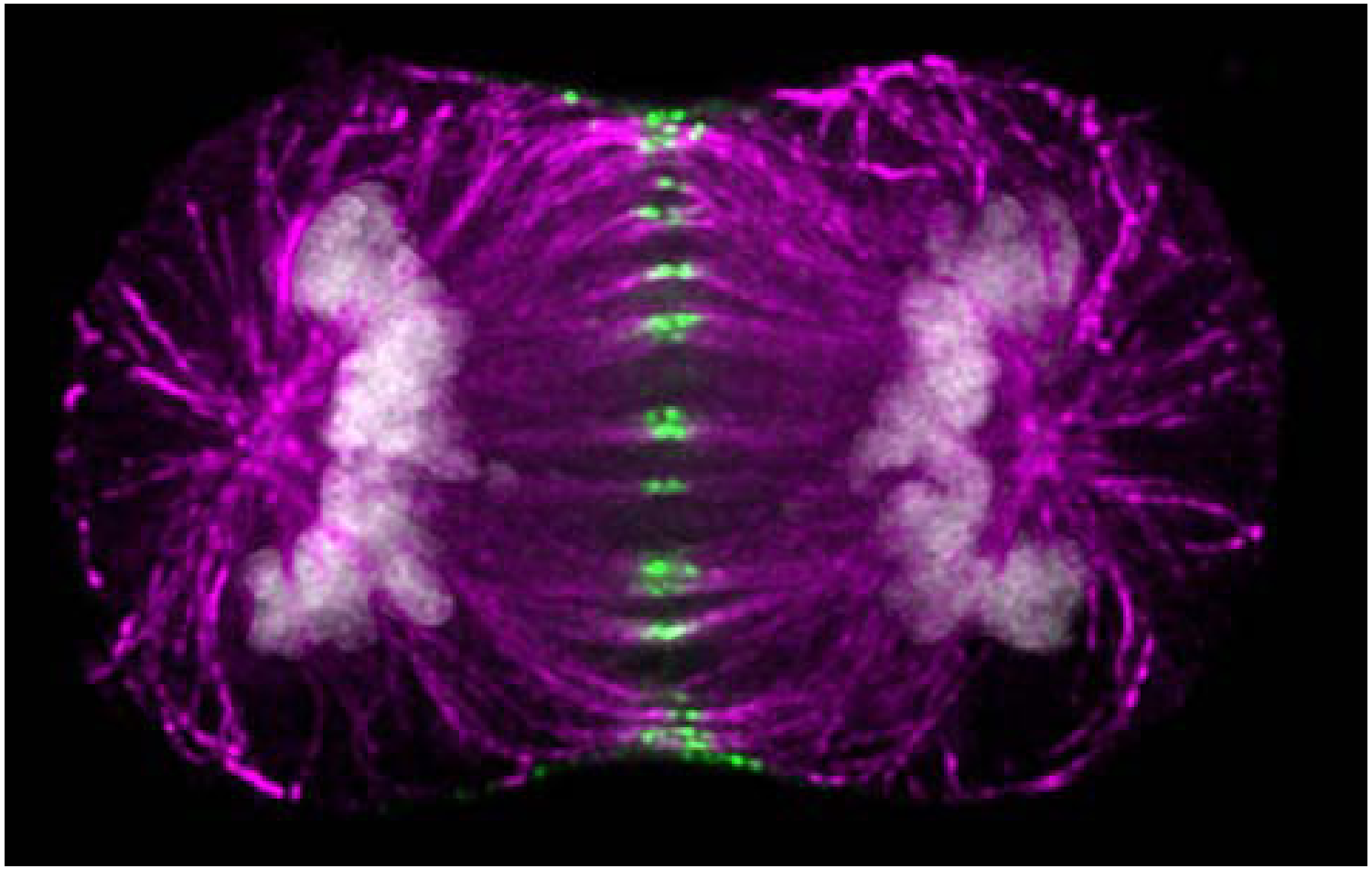
The interplay between mathematical modeling and experimental measurement is at the basis of the "effectiveness of mathematics" in physics.

Wigner, *Commun Pure Appl Math* 13, 1 (1960).



Mathematical modeling of large-scale molecular biological data can lead beyond classification of genes and cellular samples to the discovery and ultimately also control of molecular biological mechanisms.

Alter, *PNAS* 103, 16063 (2006).



Andrews & Swedlow, Nikon Small World (2002).

Our models bring physicians a step closer to one day being able to predict and control the progression of cancers as readily as NASA engineers plot the trajectories of spacecraft today.

Collaborators:

John F. X. Diffley
Cancer Research UK, London

Michael A. Saunders
Operations Research, Stanford

Charles F. Van Loan
Computer Science, Cornell

David Botstein
Genomics, Princeton

Patrick O. Brown
Biochemistry, Stanford

Gene H. Golub
Computer Science, Stanford

Support:

NHGRI K01 Award HG000038

NHGRI R01 HG004302

NSF CAREER DMS-0847173

Thank you!!!

Ph.D. Students:

Katherine A. Aiello, BE

Preethi Sankaranarayanan, BE

Theodore E. Schomay, BE

Undergraduate:

Nicolas M. Bertagnolli, Math & BE

K99/R00 Postdoc:

Jason M. Tennessen, Genetics

Ph.D. Alumni:

Kayta Kobayashi, Pharmacy

Chaitanya Muralidhara, CMB

Larsson Omberg, Physics

Sri Priya Ponnapalli, ECE

B.S. Alumni:

Justin A. Drake, BME & SSC

Andrew M. Gross, BME &SSC

Joel R. Meyerson, BME & Gov