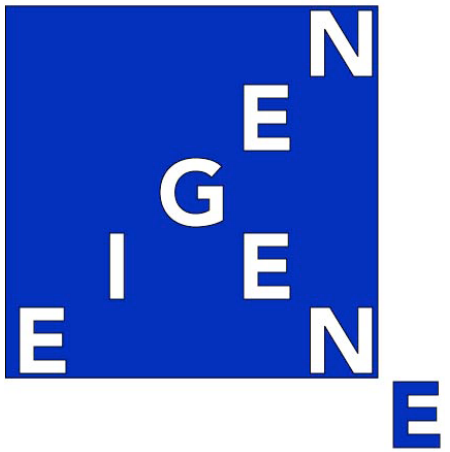
Cancer Diagnostics and Prognostics from Comparative Spectral Decompositions of Patient-Matched Genomic Profiles

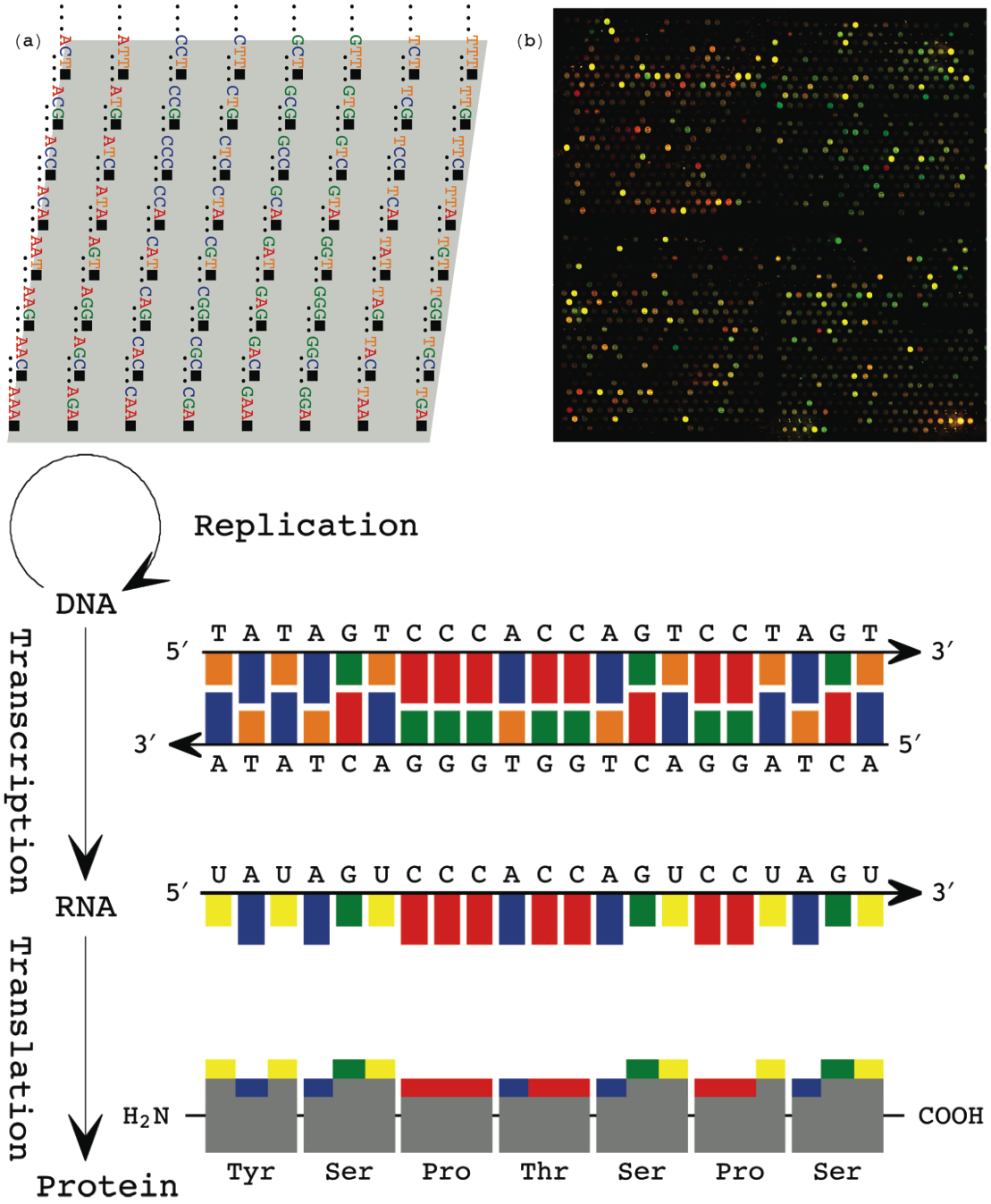
Orly Alter

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

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



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.

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

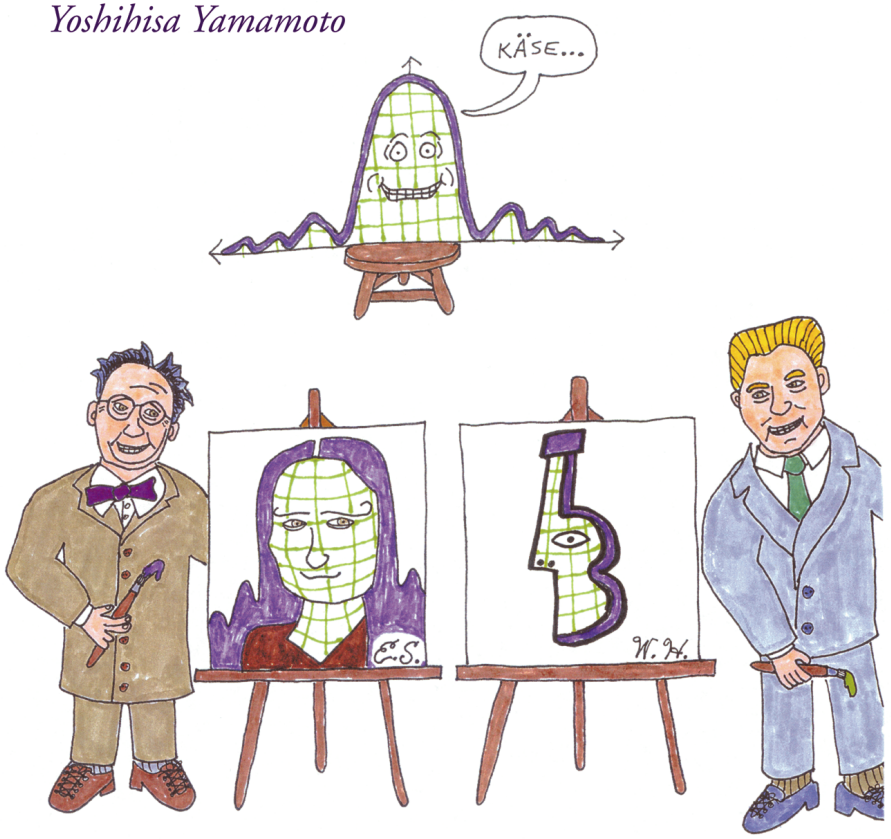
ALTER
YAMAMOTO

Quantum Measurement of a Single System

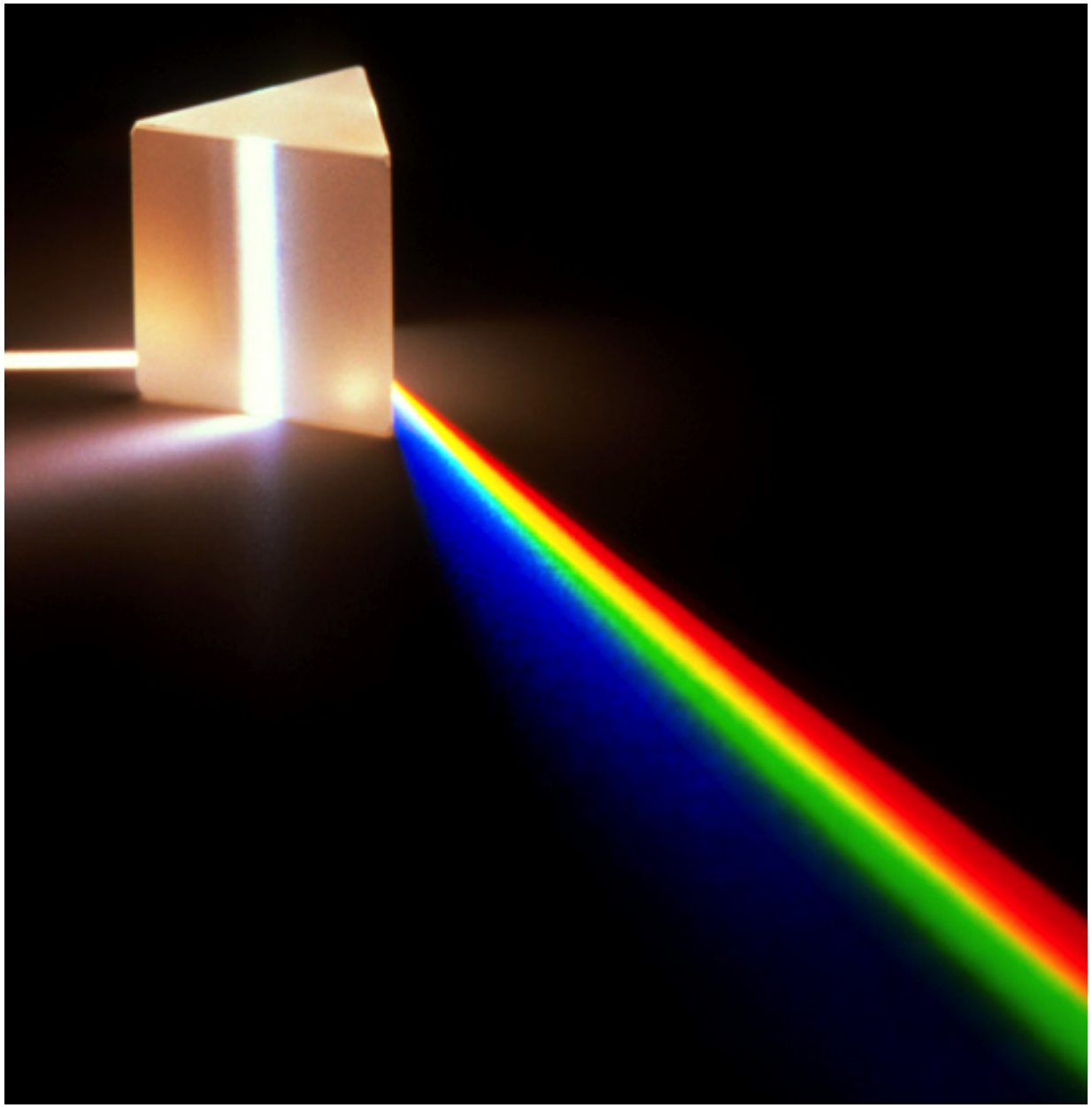
WILEY
INTER-
SCIENCE

Quantum Measurement of a Single System

Orly Alter
Yoshihisa Yamamoto



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;

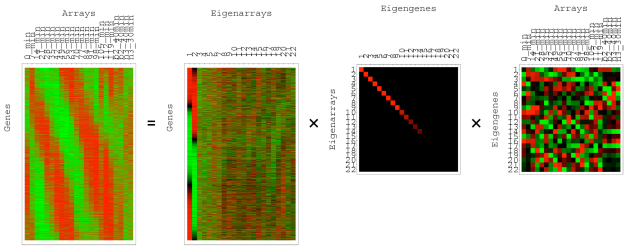
Predict global mechanisms that govern the activity of DNA and RNA.

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).

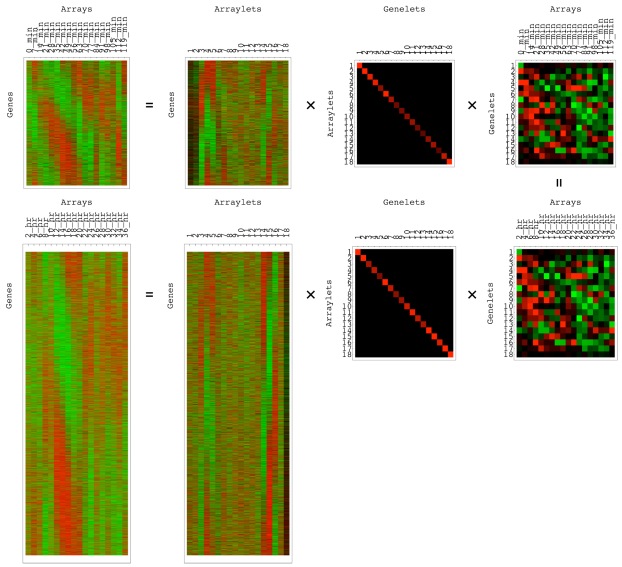


"Eigengenes" and "eigenarrays" → cellular processes and states in one dataset.

Eigenvalue Decomposition

Comparative GSVD

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

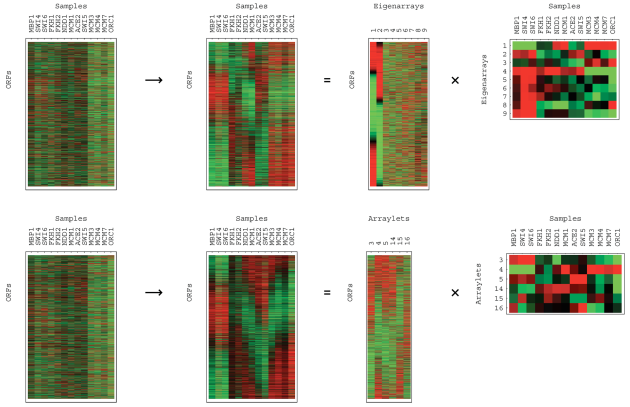


"Genelets" and "arraylets" → phenomena exclusive to one of, or common to two datasets.

Generalized Eigenvalue Decomposition

Integrative Pseudoinverse

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

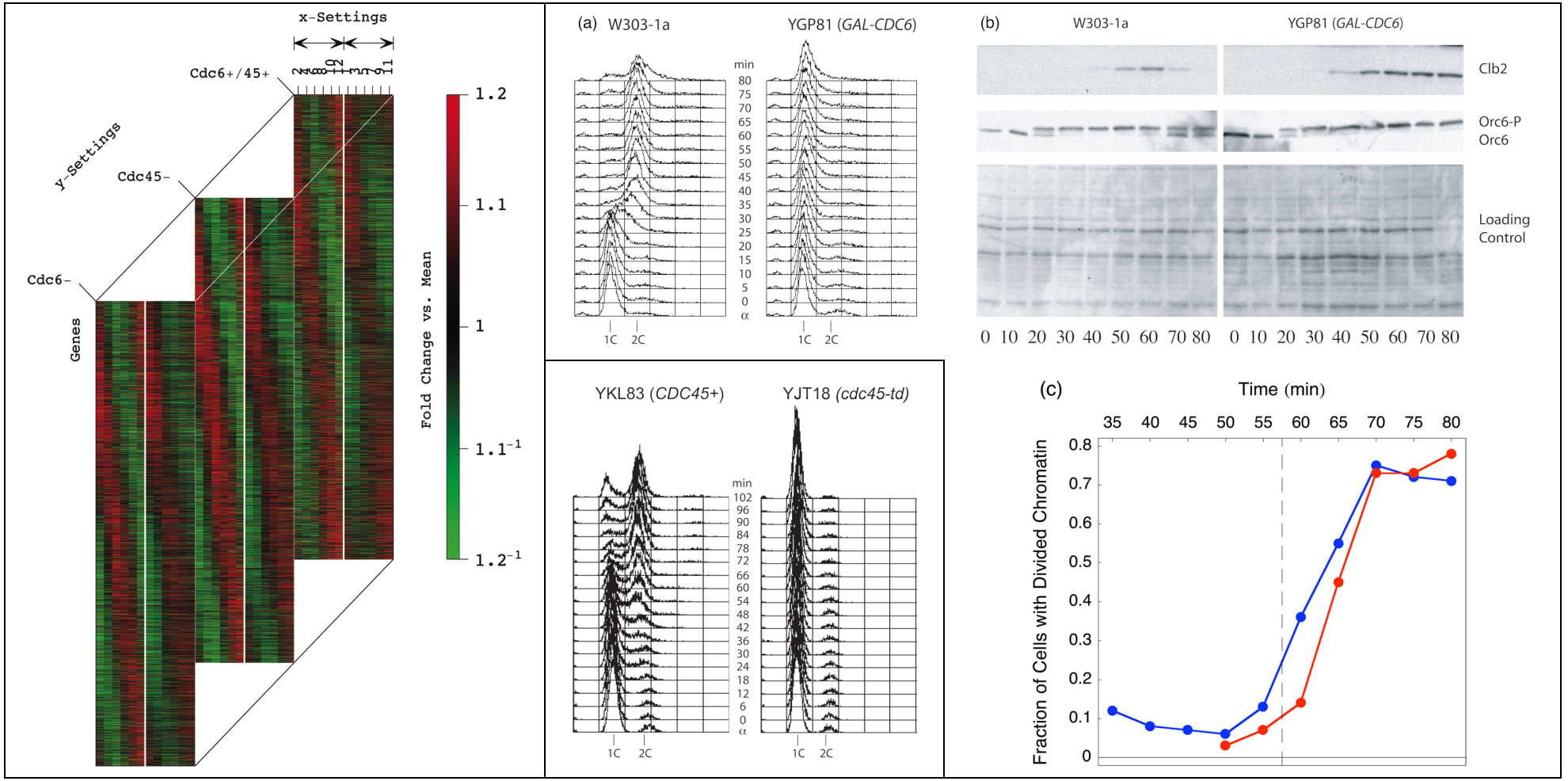


"Pseudoinverse correlation" → causal coordination between two datasets.

Inverse Projection

Effects of DNA Replication on RNA Expression: Experimental Verification of a Computationally Predicted Mode of Regulation

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



Matrix and tensor modeling of large-scale molecular biological data can be used to correctly predict previously unknown cellular mechanisms.

HOSVD for Integrative Analysis of a High-Dimensional Dataset

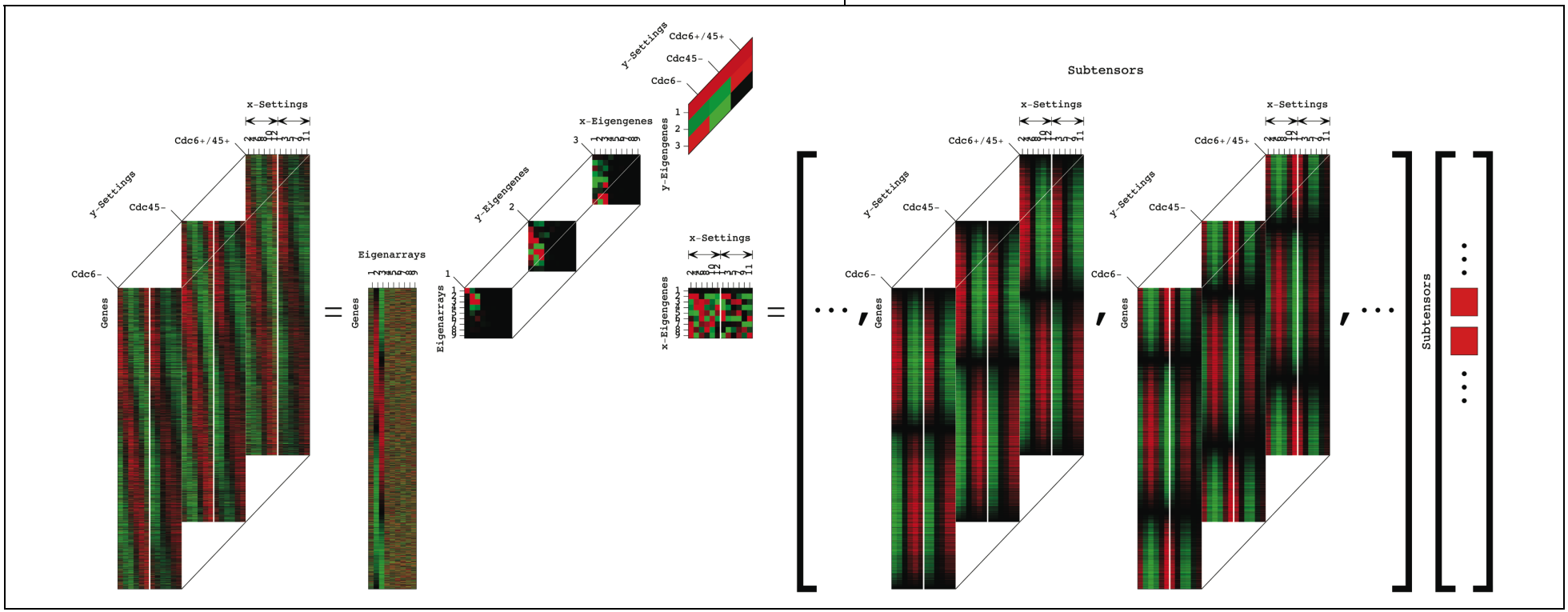
Omberg, Golub & Alter, *PNAS* 104, 18371 (2007); http://alterlab.org/HOSVD/

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,

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



De Lathauwer, De Moor & Vandewalle, *SIMAX* 21, 1253 (2000).

HOSVD for Integrative Analysis of a High-Dimensional Dataset

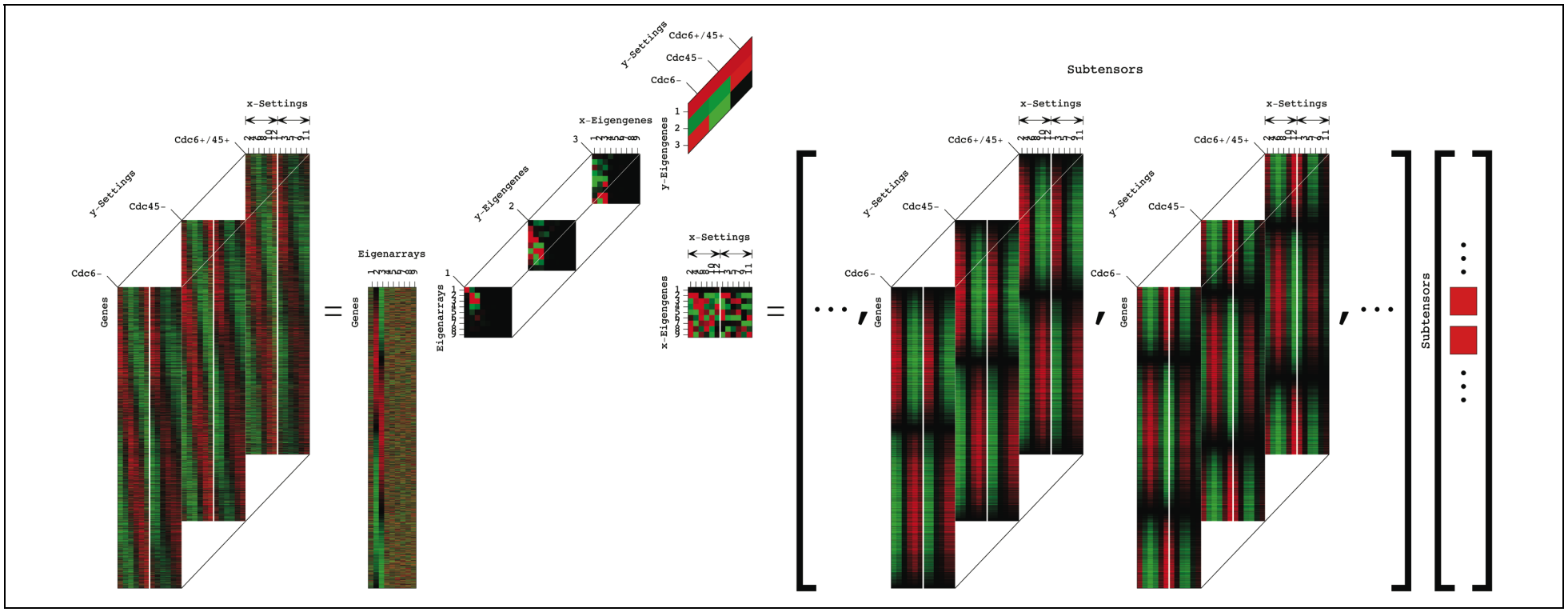
Omberg, Golub & Alter, *PNAS* 104, 18371 (2007); http://alterlab.org/HOSVD/

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

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

A "degenerate subtensor space rotation" gives one unique subtensor,

$$\mathcal{R}_{a+k,b,c}\mathcal{S}(a+k,b,c) = \mathcal{R}_{abc}\mathcal{S}(a,b,c) + \mathcal{R}_{kbc}\mathcal{S}(k,b,c).$$

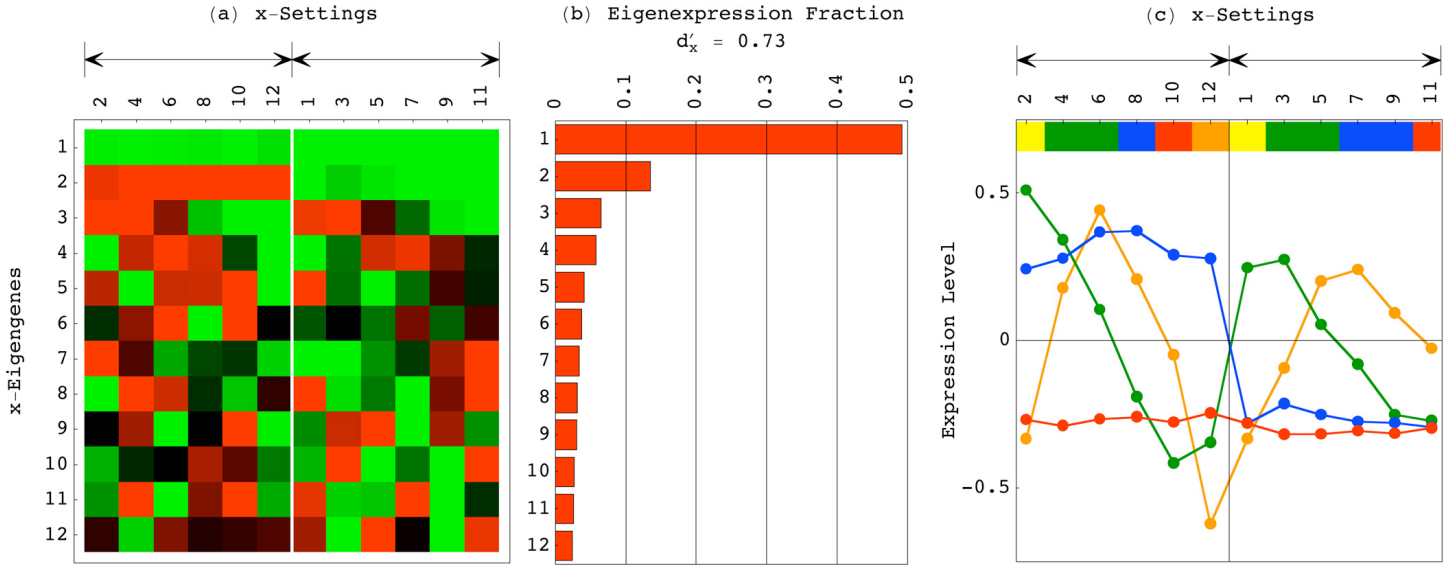


De Lathauwer, De Moor & Vandewalle, *SIMAX* 21, 1253 (2000).

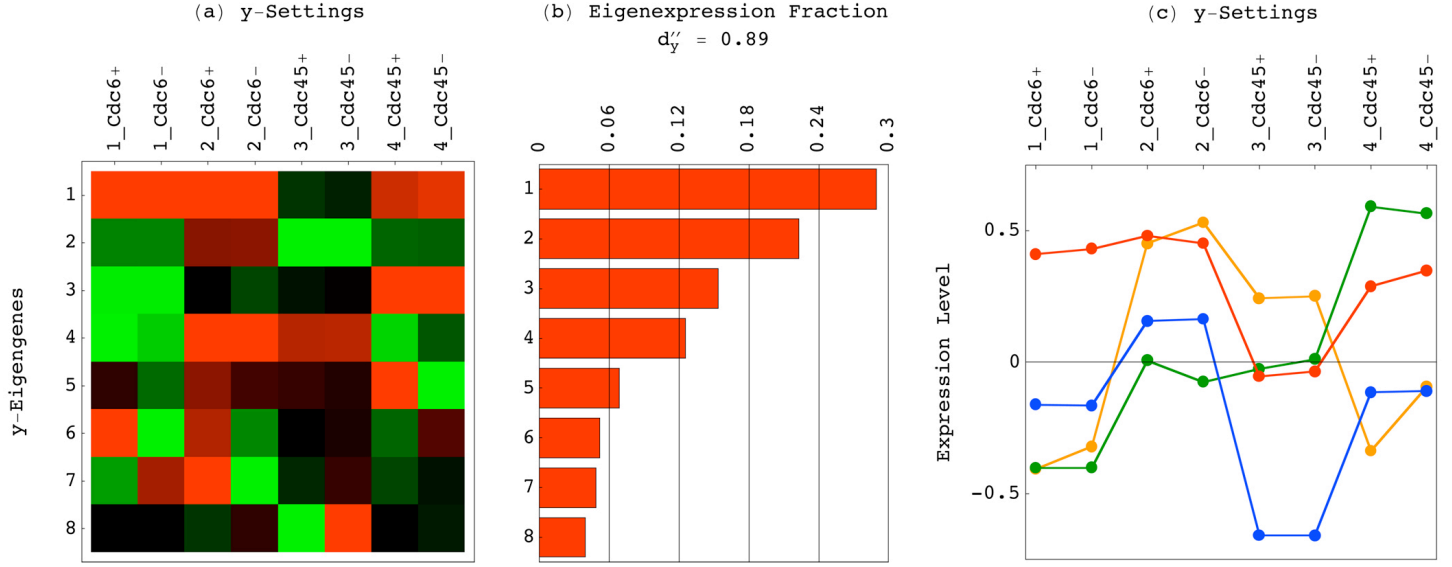
HOSVD Detection and Removal of Artifacts

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

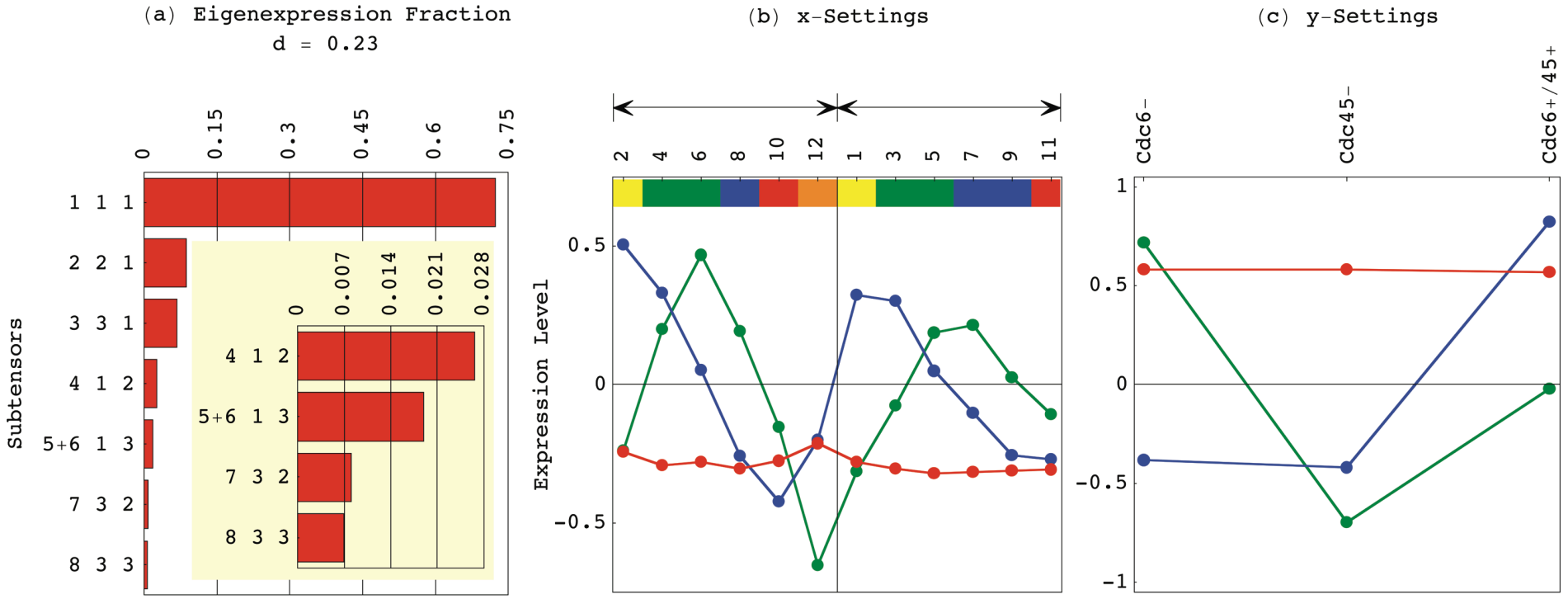


Culture batch, microarray platform and protocols



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).

3,3,1	7%	>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

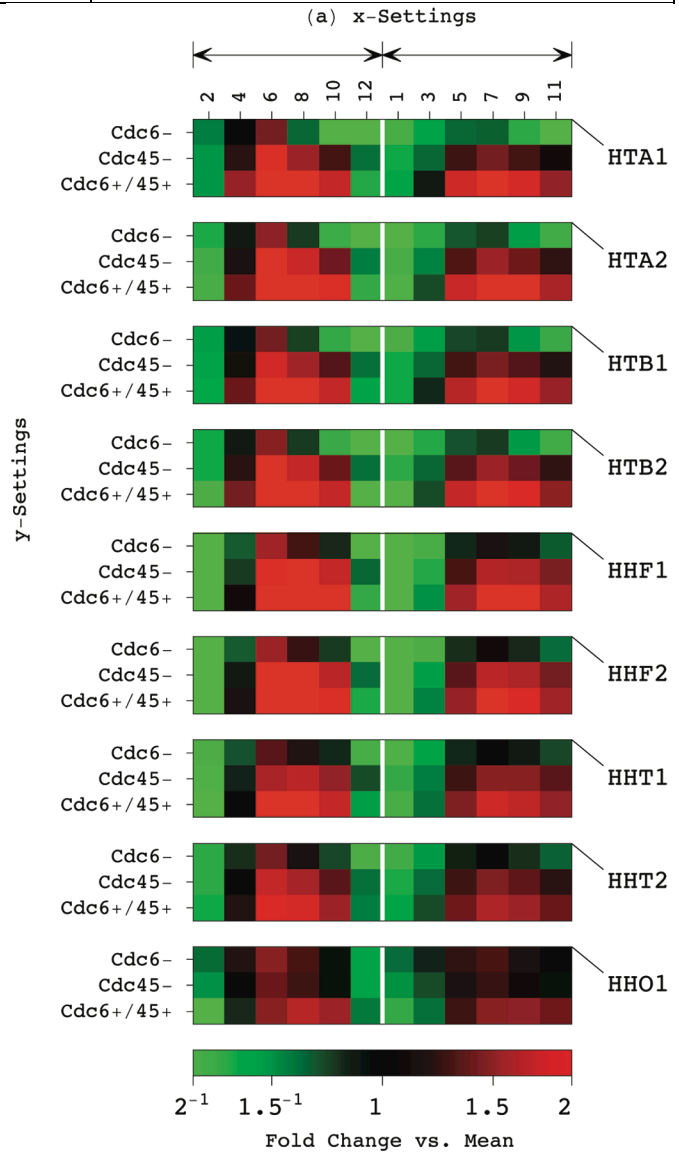
4,1,2	2.7%	>0	↑ ARSs 3’ $\sim 10^{-2}$	↓ histones $< 10^{-12}$
7,3,2	0.8%	>0	↑ histones $< 5 \cdot 10^{-4}$	

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.

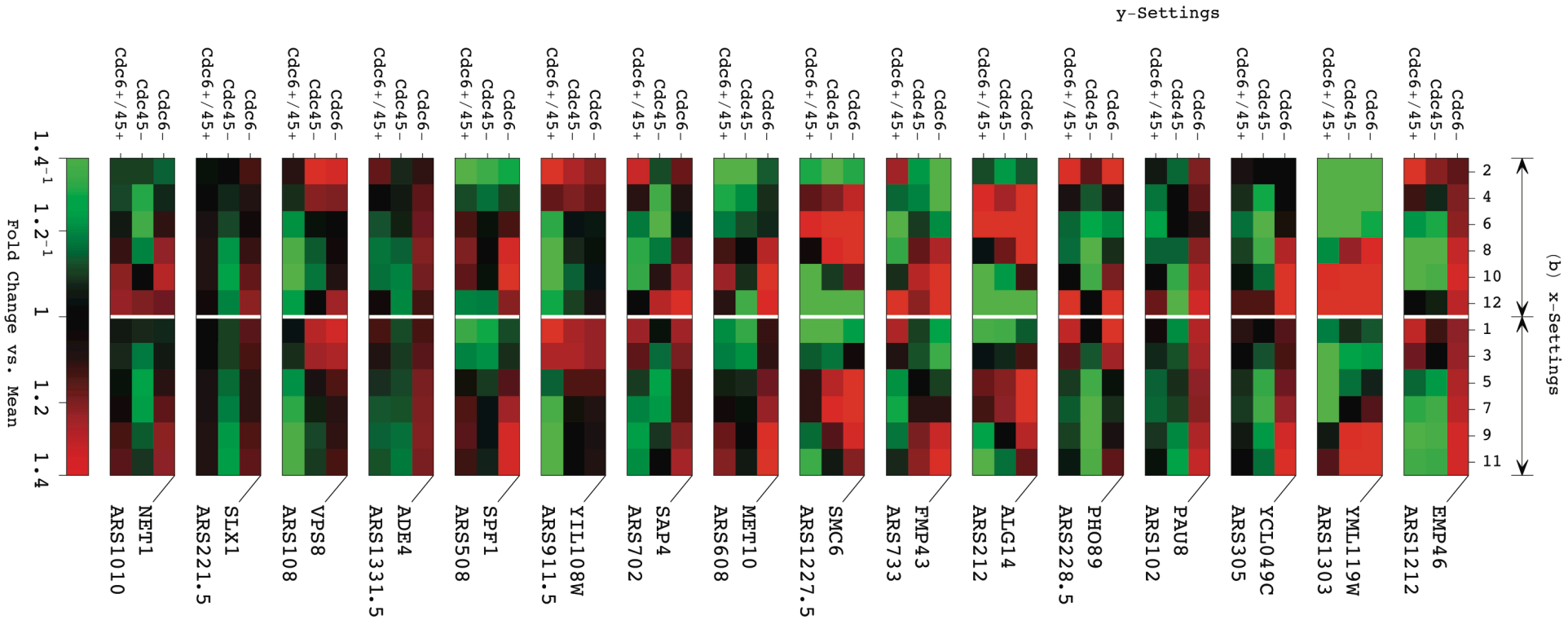
Lycan, Osley & Hereford, *MCB* 7, 614 (1987).



Origin Binding-Dependent Perturbations

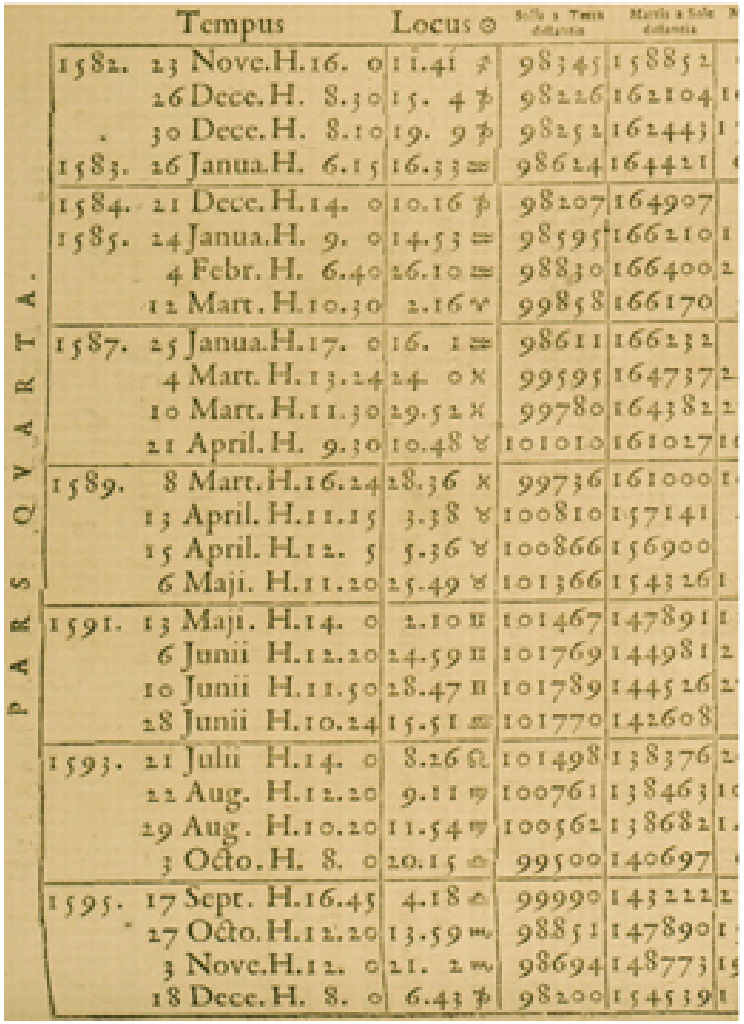
5+6,1,3	1.9%	>0	↑ histones $<2\cdot10^{-8}$	↓ ARSs 3' $<2\cdot10^{-3}$
8,3,3	0.7%	>0		↓ ARSs 3' $<7\cdot10^{-4}$

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.



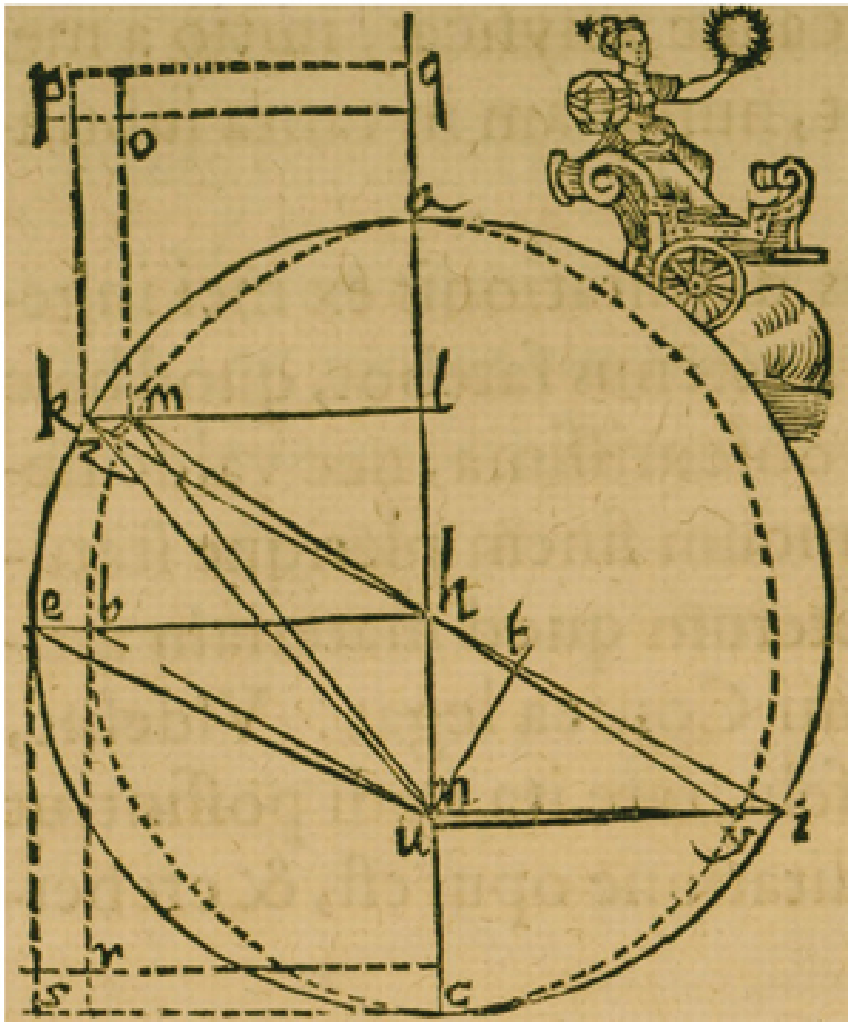
Patterns Underlie Principles of Nature: Global Correlations to Causal Coordination

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

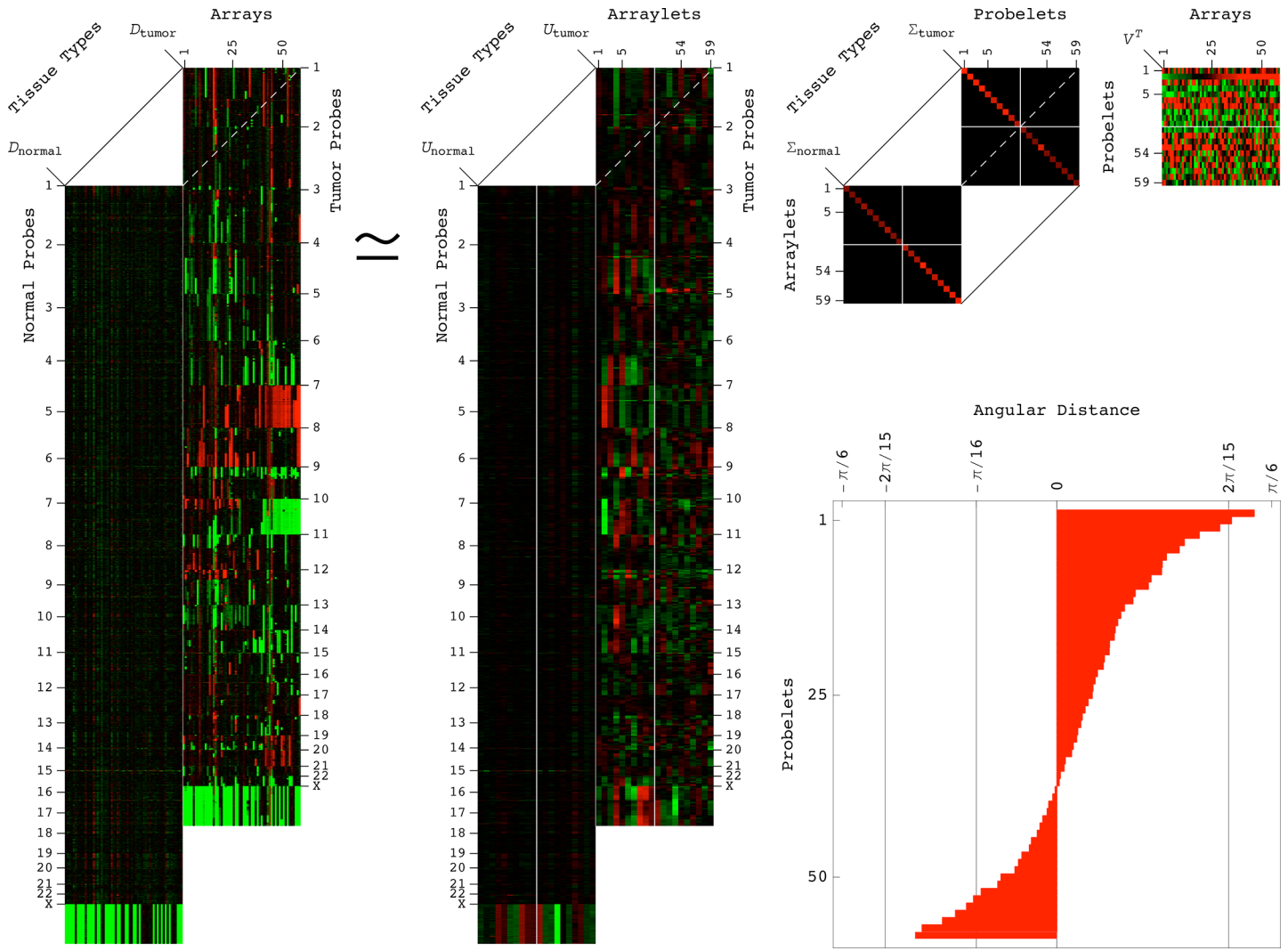


Kepler's discovery of his first law of planetary motion from mathematical modeling of Brahe's astronomical data.

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

GSVD Modeling of Patient-Matched LGA Tumor and Normal Genomic Profiles

Aiello & Alter, *PLoS One* 11, e0164546 (2016); http://alterlab.org/astrocytoma_prognosis/



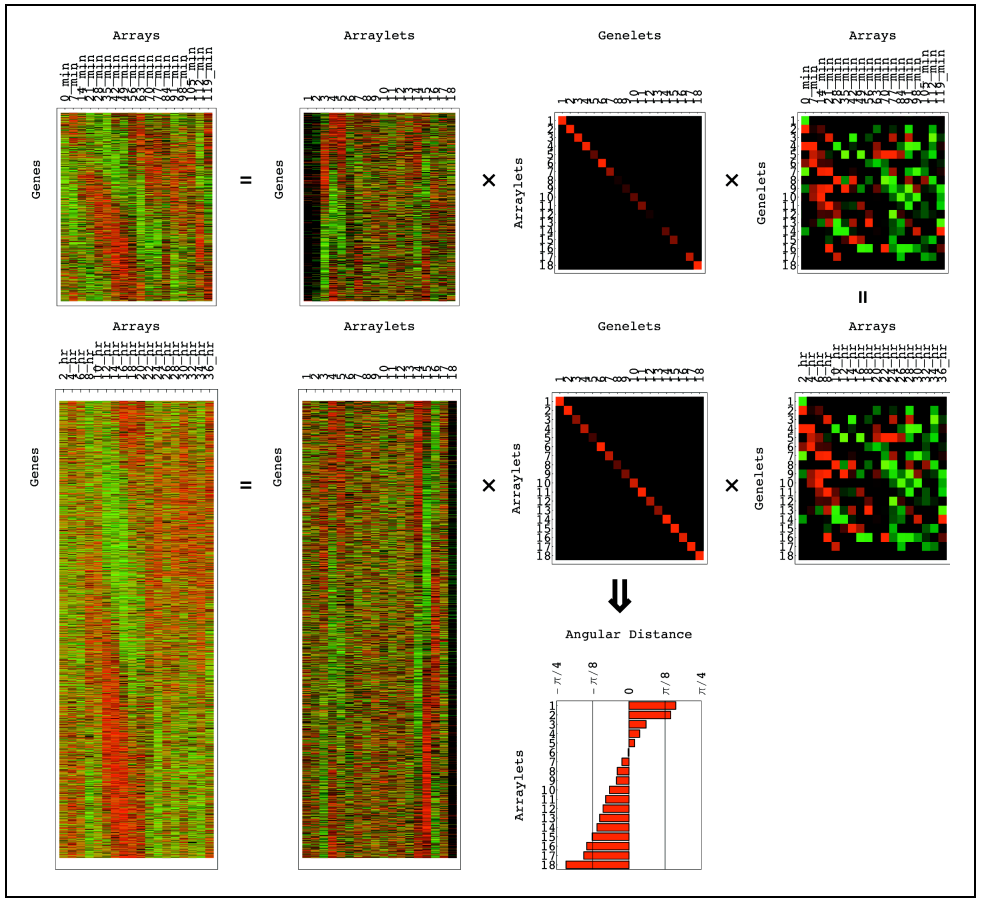
The GSVD, formulated as a comparative spectral decomposition, can create a single model from, i.e., simultaneously find similarities and dissimilarities across two column-matched but row-independent matrices.

GSVD for Comparative Analysis of Two Different Two-Dimensional Datasets

Alter, Brown & Botstein, *PNAS* 100, 3351 (2003); http://alterlab.org/GSVD/

The GSVD simultaneously separates the two datasets into paired weighted sums of outer products, of each normalized right basis vector, or a "probelet" (a pattern of variation across the patients), which is identical for both datasets, combined with one of the two corresponding orthonormal left basis vectors, or "arraylets" (the tumor- and normal-specific patterns of variation across the genome),

$$D_i = U_i \Sigma_i V^T = \sum_{n=1}^{N} \sigma_{i,n} u_{i,n} \otimes v_n^T, \quad i = 1,2.$$

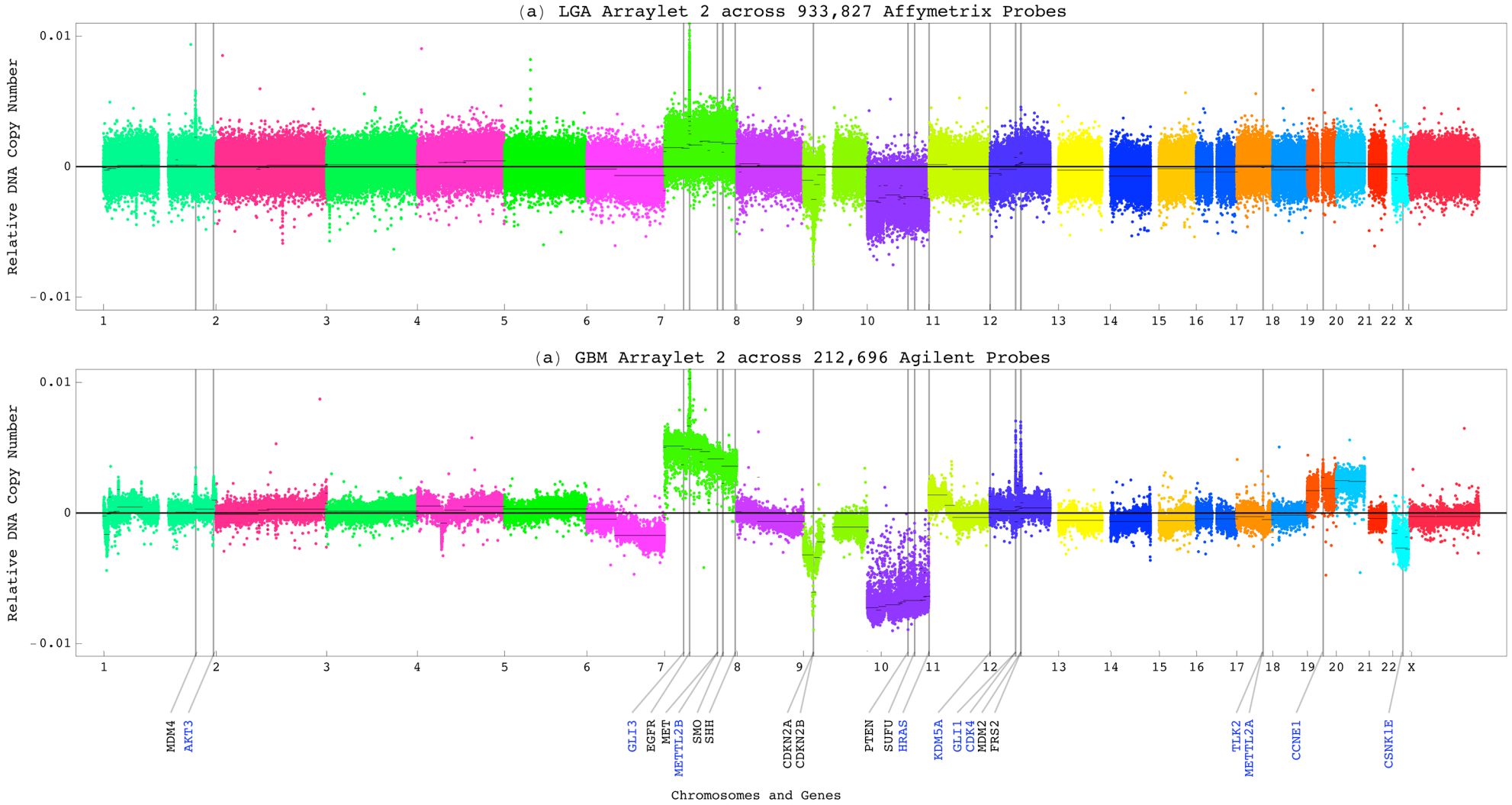


The significance of a probelet and its corresponding arraylet in one dataset relative to the second is defined by the "angular distance,"

$$-\pi/4 \le \arctan(\sigma_{1,n}/\sigma_{2,n}) - \pi/4 \le \pi/4.$$

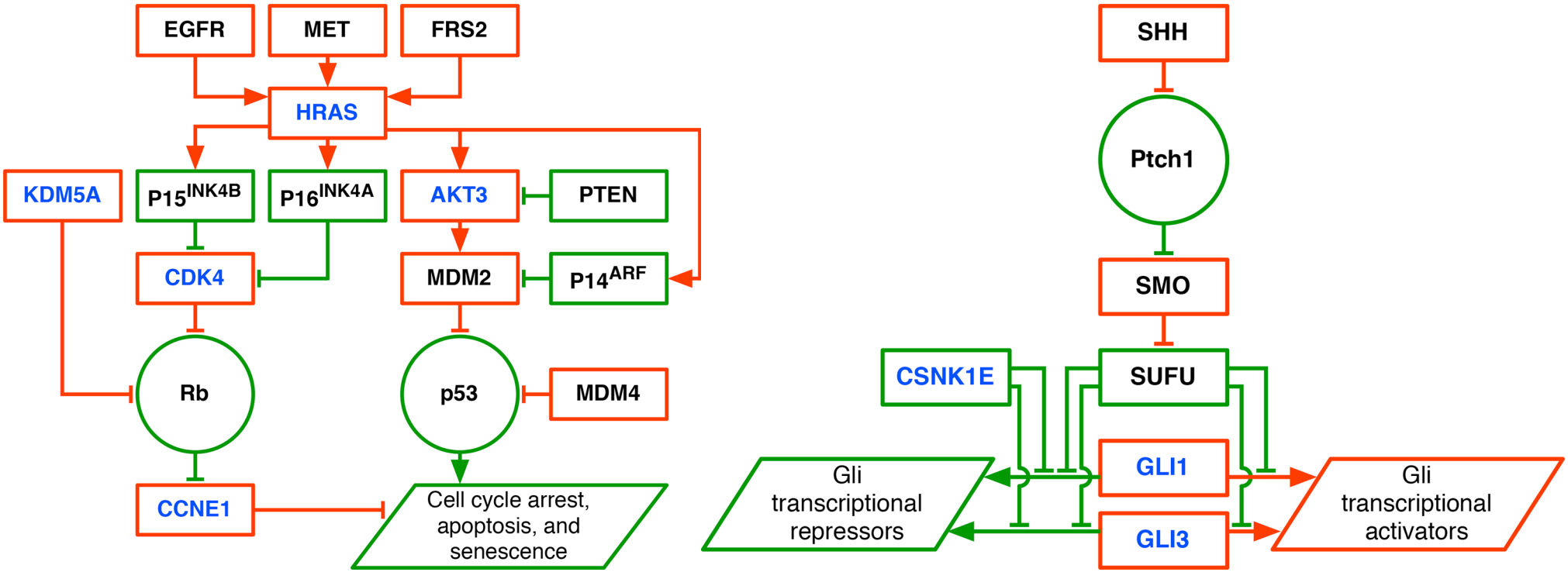
Van Loan, *SINUM* 13, 76 (1976); Paige & Saunders, *SINUM* 18, 398 (1981);
Van Loan, *Numer Math* 46, 479 (1985).

Global Pattern of LGA Tumor-Exclusive Copy-Number Alterations (CNAs) Encompassed in that Uncovered in GBM



Lee,* Alpert,* Sankaranarayanan & Alter, *PLoS One* 7, e30098 (2012);
Aiello, Ponnapalli & Alter (in preparation).

GBM-Specific CNAs Encode for Enhanced Transformation and Proliferation in GBM Relative to LGA

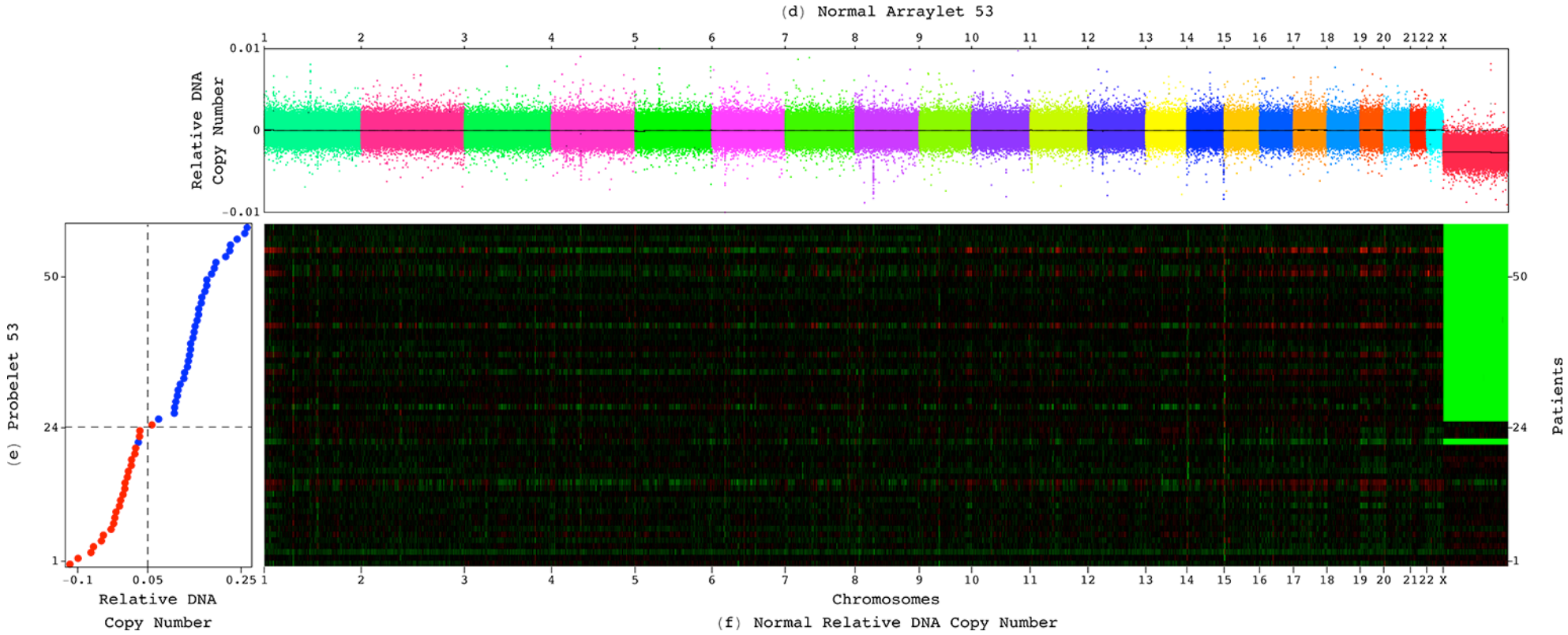


The rat sarcoma virus (Ras) and Hedgehog (Hh) signaling pathways.

Karnoub & Weinberg, *Nat Rev Mol Cell Biol* 9, 517 (2008);
Rohatgi & Scott, *Nat Cell Biol* 9, 1005 (2007).

The GSVD Separates the LGA Pattern from Common Sources of Biological Variation

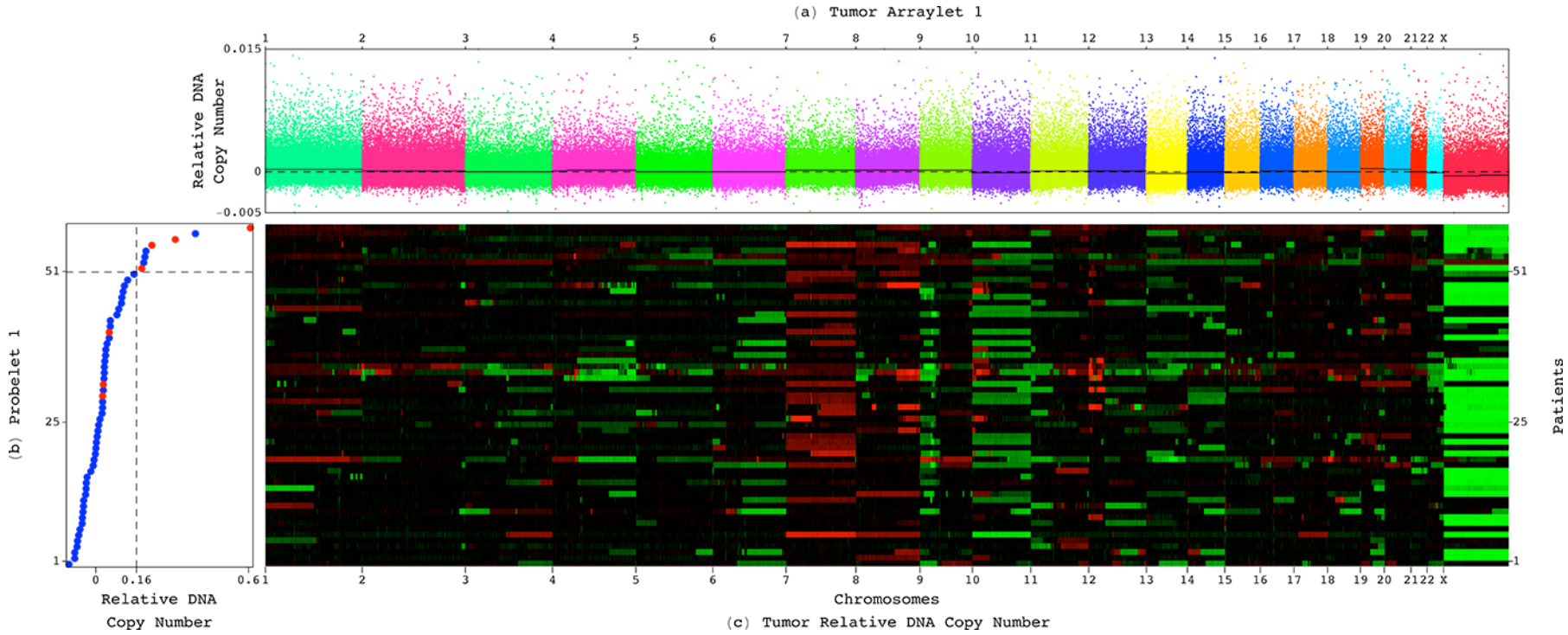
The GSVD identifies copy-number variations (CNVs) that occur in the normal human genome and are preserved in the LGA tumors, e.g., male-specific X chromosome deletion, without a-priori knowledge.

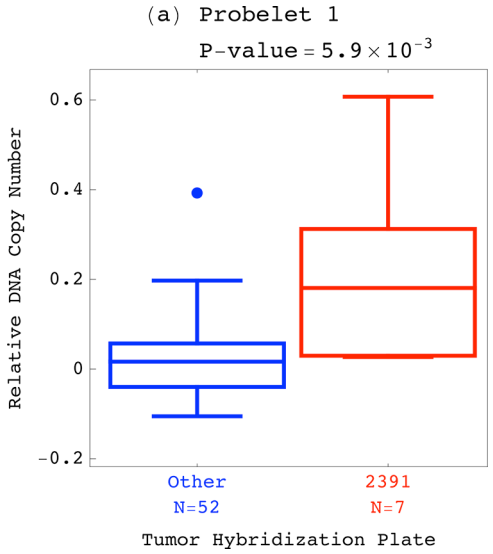


Patients' gender is correctly identified also where the TCGA database entries and the copy-number gender assignments are in discrepancy.

NHGRI's Interest in 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

The GSVD Separates the LGA Pattern from Exclusive Sources of Experimental Variation

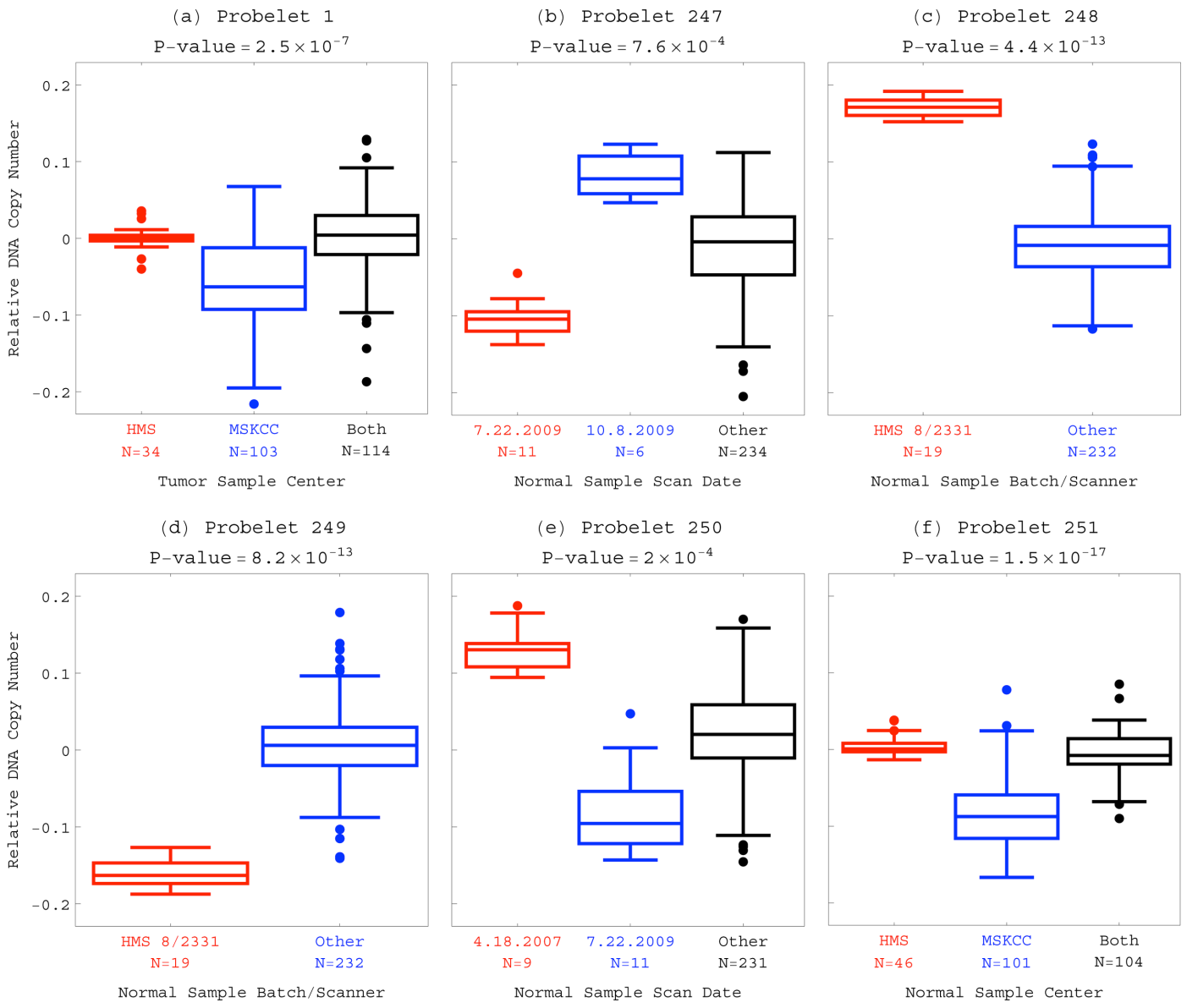




The GSVD identifies experimental variations, e.g., in tissue batch, genomic center, hybridization plate batch, without a-priori knowledge.

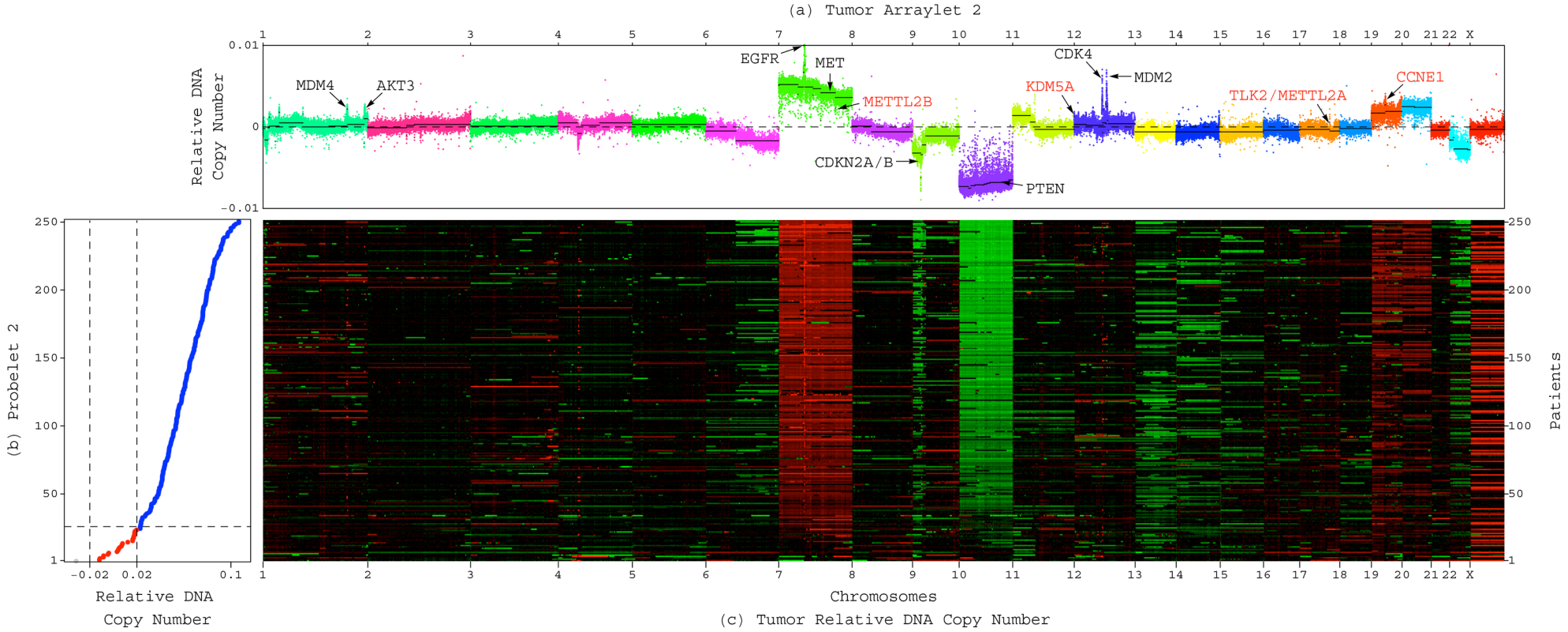
Experimental Variations Exclusive to the Tumor or Normal Profiles

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



Global Pattern of GBM Tumor-Exclusive CNAs Predicts Drug Targets

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

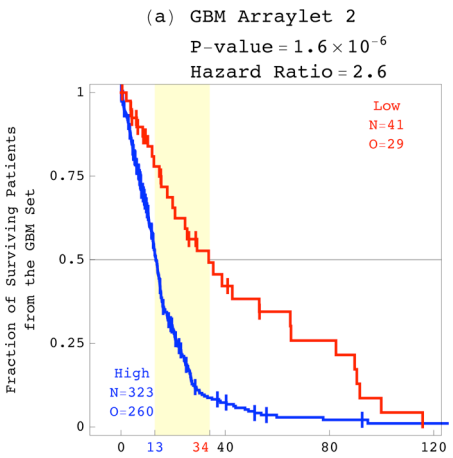


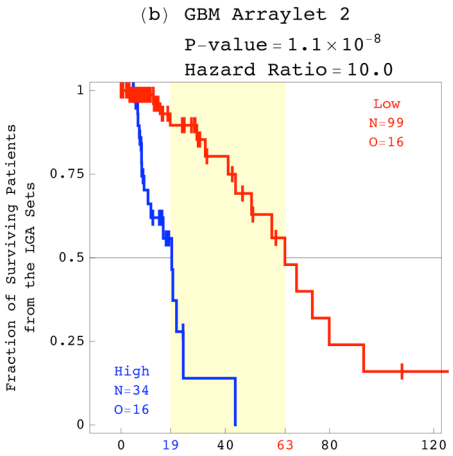
The pattern includes most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of the patients, including biochemically putative drug targets: the cell cycle-regulated serine/threonine kinase-encoding *TLK2*, and the tRNA methyltransferase-like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

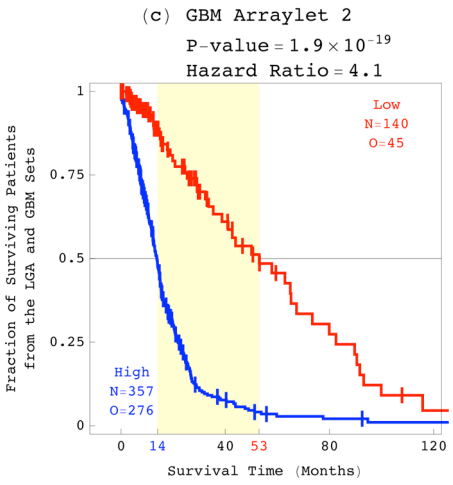
Platform-Independent Genomic Predictor of Astrocytoma Outcome

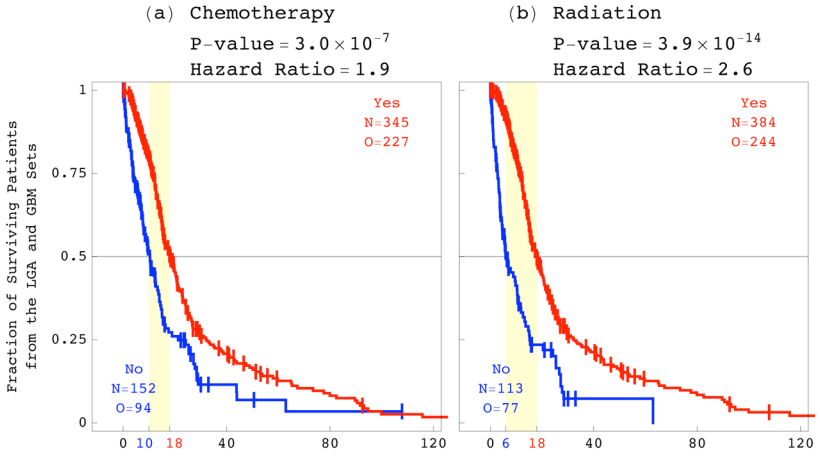
Aiello & Alter, *PLoS One* 11, e0164546 (2016);
http://alterlab.org/astrocytoma_prognosis/

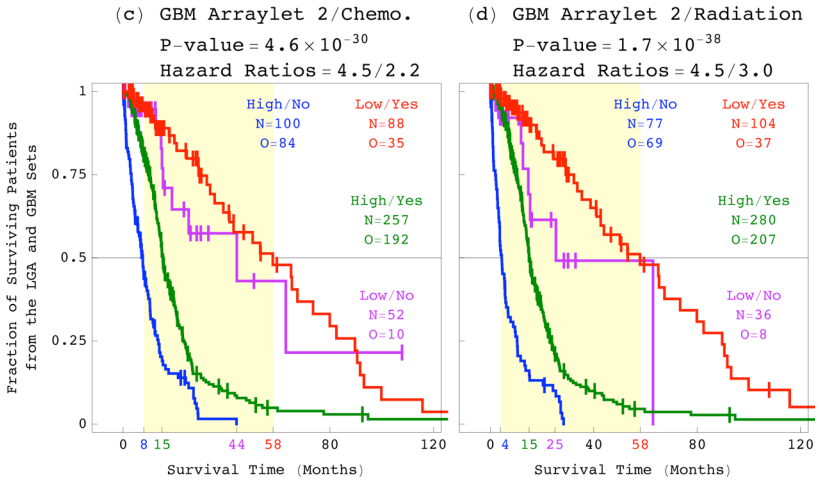
The LGA pattern is correlated with a patient's survival and response to treatment. The GBM pattern identifies among the LGA patients a subtype, statistically indistinguishable from that among the GBM patients, where the CNA genotype is correlated with a one-year survival phenotype.



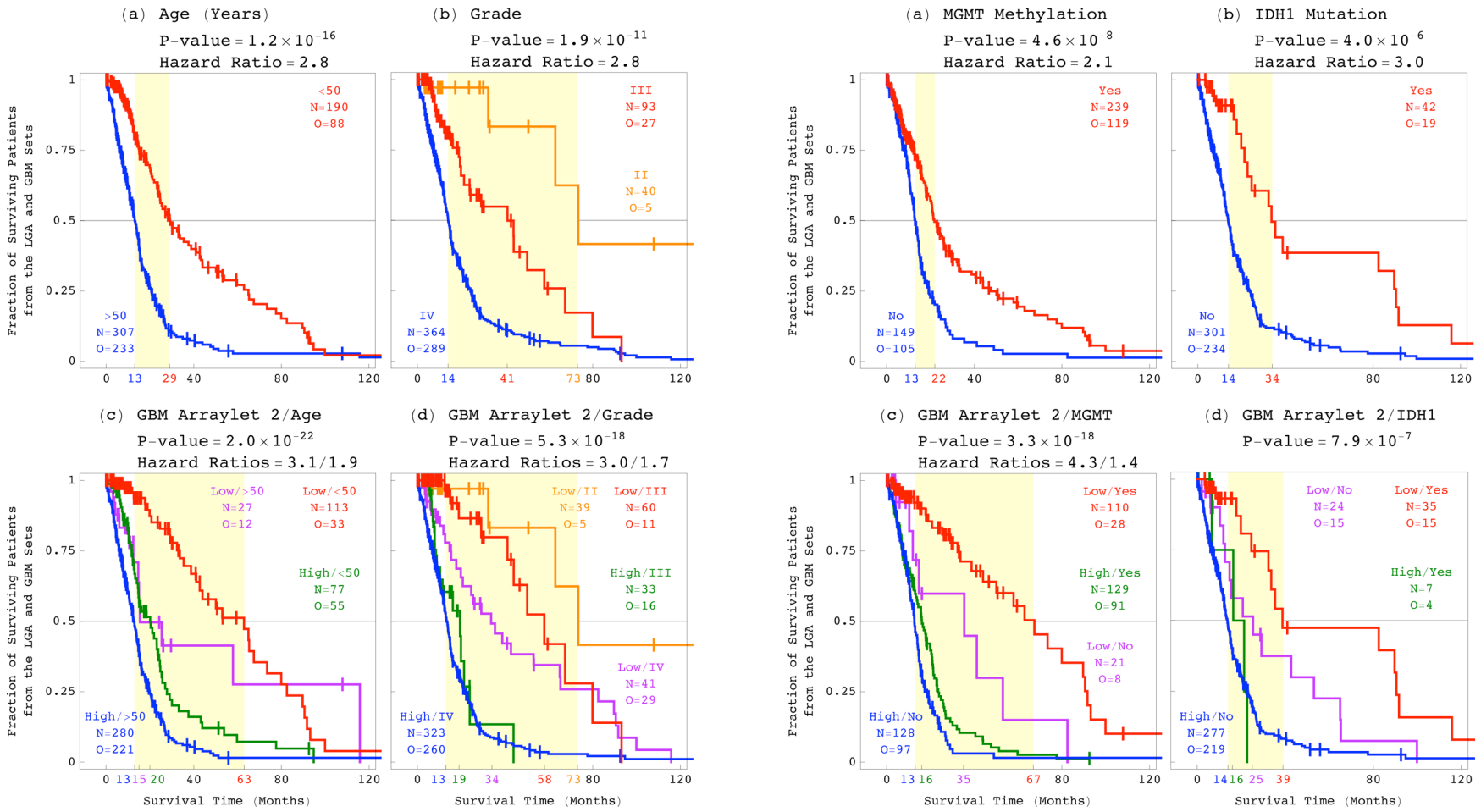








Statistically Better Than, and Independent of Age, Grade, and Laboratory Tests



Recurring DNA CNAs were observed in astrocytoma tumors' genomes for decades, however, copy-number subtypes predictive of patients' outcomes were not identified before, despite the growing number of datasets recording different aspects of the disease, and due to a need for frameworks that can simultaneously find similarities and dissimilarities across the datasets.

Tensor GSVD Modeling of Patient- and Platform-Matched OV Tumor and Normal Genomic Profiles

Sankaranarayanan,* Schomay,* Aiello & Alter,
PLoS One 10, e121396 (2015);
http://alterlab.org/OV_prognosis/

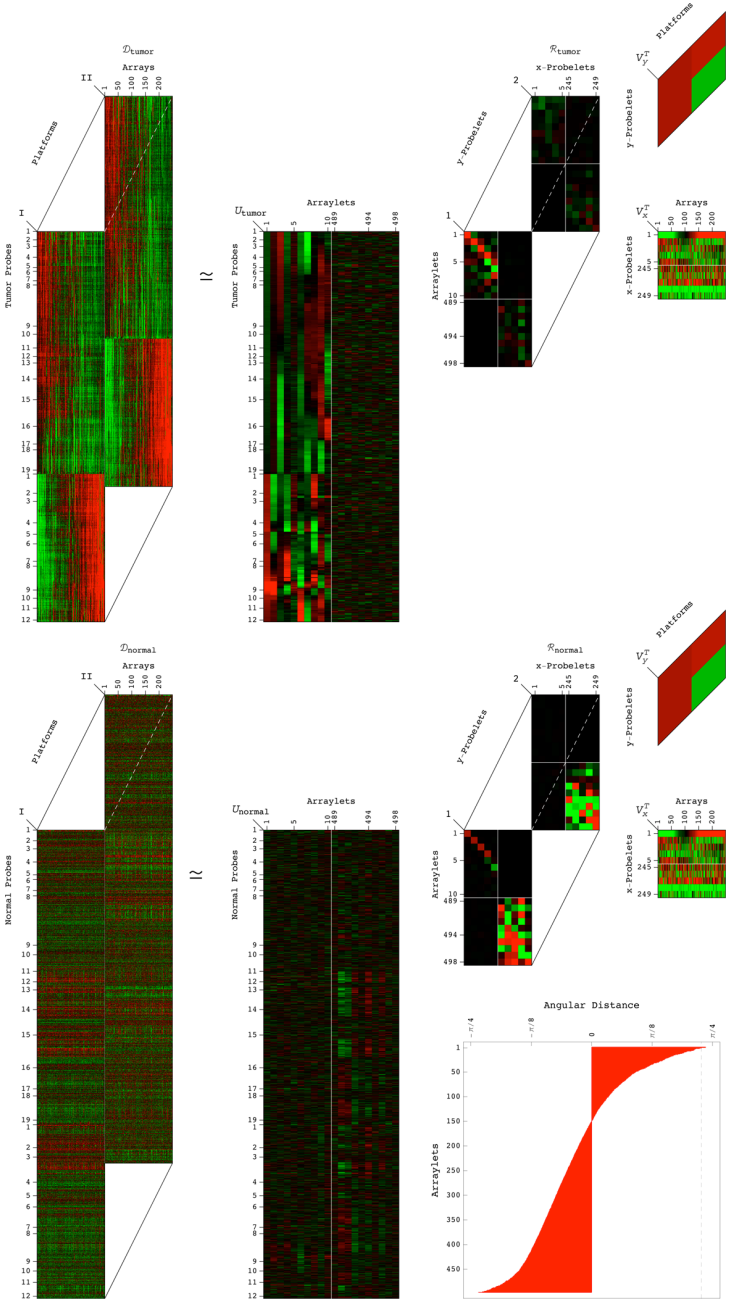
$$\mathcal{D}_i = \mathcal{R}_i \times_a U_i \times_b V_x \times_c V_y$$

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

$$\mathcal{S}_i(a,b,c) = u_{i,a} \otimes v_{x,b}^T \otimes v_{y,c}^T,$$

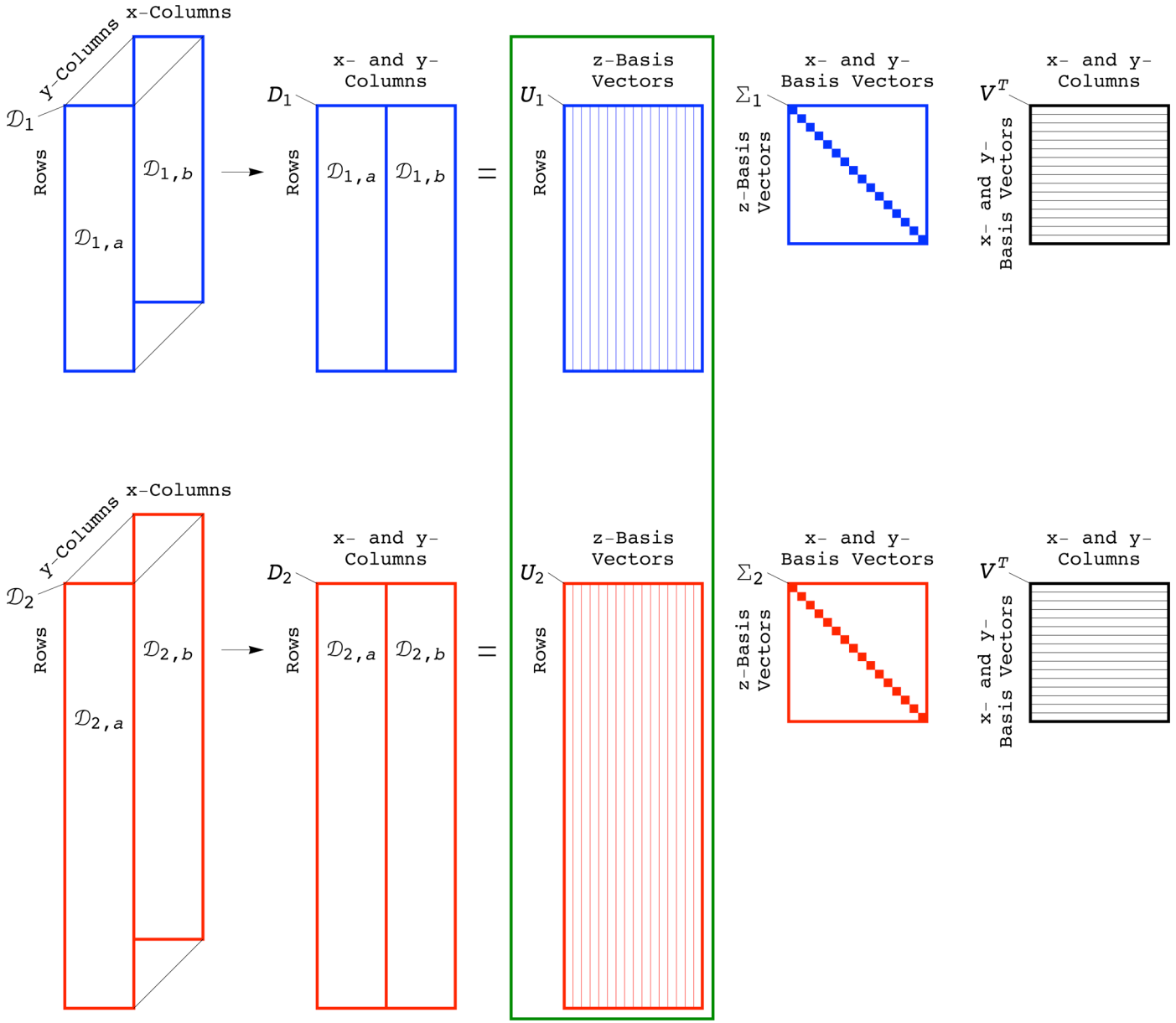
$$i = 1, 2.$$

This novel, exact decomposition extends the GSVD and the tensor HOSVD from a decomposition of either two column-matched matrices or one tensor, respectively, to a decomposition of two order-matched, column-matched, and row-independent tensors.



Tensor GSVD for Comparative Analysis of Two Different High-Dimensional Datasets

Sankaranarayanan,* Schomay,* Aiello & Alter, *PLoS One* 10, e121396 (2015);
Schomay, Aiello & Alter (in preparation).



Tensor GSVD for Comparative Analysis of Two Different High-Dimensional Datasets

Sankaranarayanan,* Schomay,* Aiello & Alter, *PLoS One* 10, e121396 (2015);
Schomay, Aiello & Alter (in preparation).

The mathematical properties of the tensor GSVD allow interpreting its variables and operations in terms of the similar as well as dissimilar, e.g., biomedical reality between the datasets.

Supplementary Lemma 1:

The tensor GSVD exists for two tensors of the same order since it is constructed from the GSVDs of the tensors unfolded into full column rank matrices.

Supplementary Lemma 2:

The tensor GSVD has the same uniqueness properties as the GSVD.

Supplementary Corollary 1:

The tensor GSVD of two second-order tensors reduces to the GSVD of the corresponding matrices.

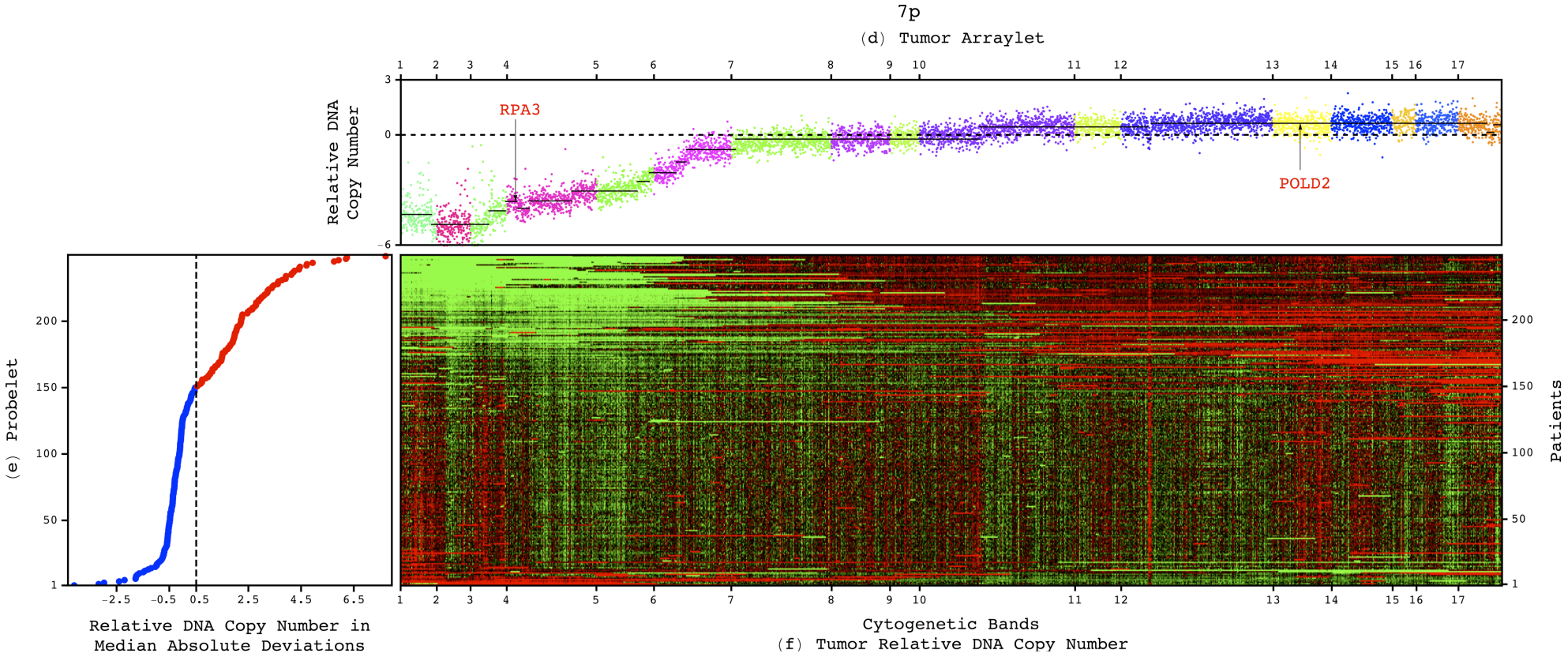
Supplementary Theorem 1:

The tensor GSVD of the tensor, which row mode unfolding gives the identify matrix, and a tensor of the same column dimensions reduces to the HOSVD of the tensor.

Theorem 1:

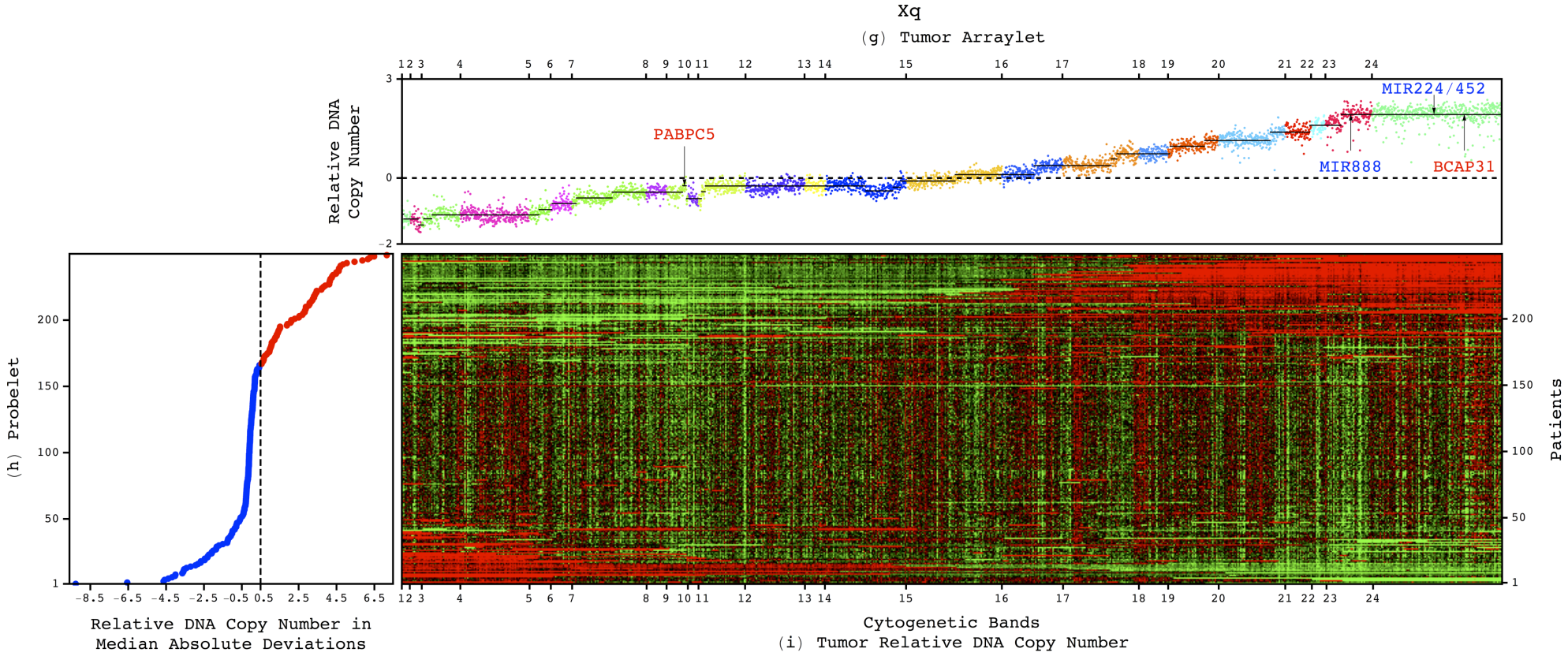
The tensor GSVD angular distance equals that of the row mode GSVD.

Chromosome Arm-Wide Patterns of OV Tumor-Exclusive Platform-Consistent CNAs



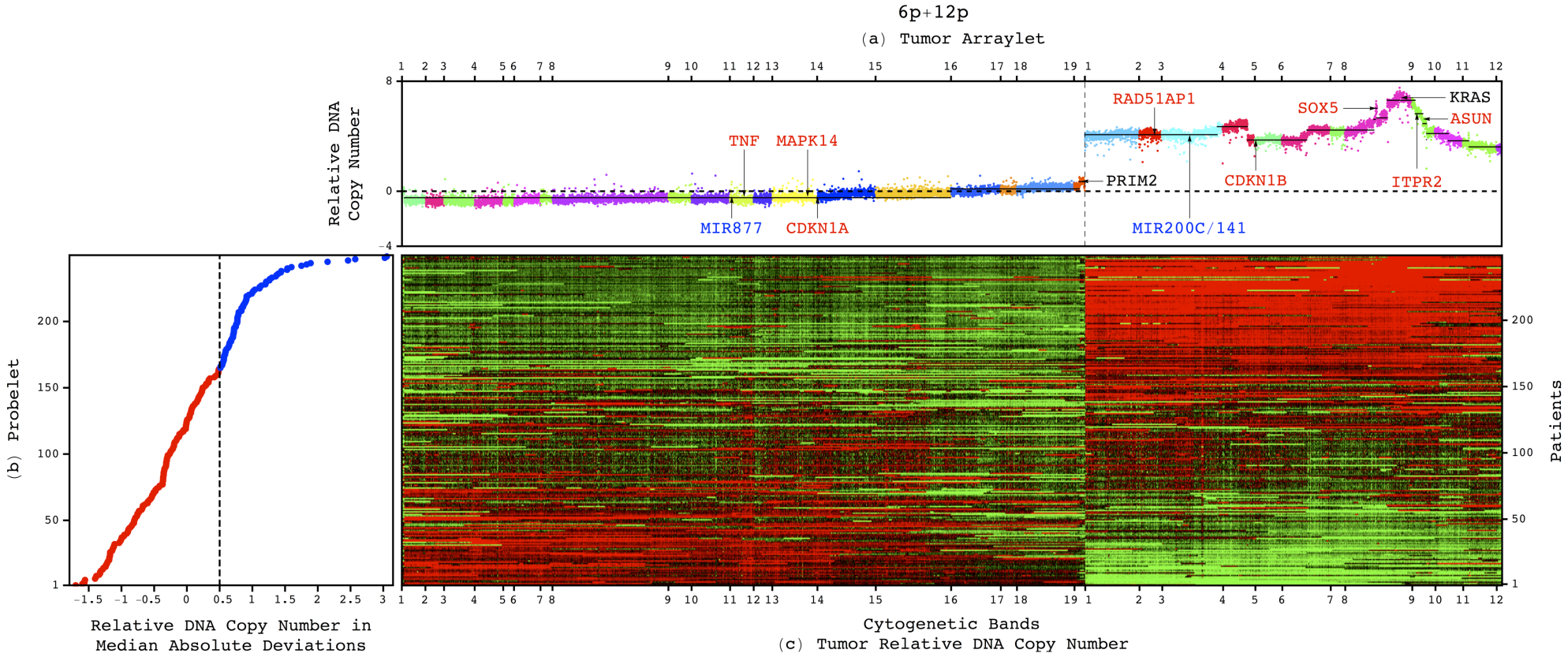
Co-occurring deletion and underexpression of *RPA3*, and amplification and overexpression of *POLD2* on 7p are correlated with DNA double-strand break repair via homologous recombination during replication, i.e., DNA stability, and a longer survival.

Chromosome Arm-Wide Patterns of OV Tumor-Exclusive Platform-Consistent CNAs



Co-occurring deletion of *PABPC5*, and amplification and overexpression of *BCAP31* on Xq are correlated with a cellular immune response, and a longer survival time.

Chromosome Arm-Wide Patterns of OV Tumor-Exclusive Platform-Consistent CNAs Encoding for Cell Transformation



Loss of the p21-encoding *CDKN1A* and the p38-encoding *MAPK14* on 6p, and gain of *KRAS* on 12p, combined but not separately, can lead to transformation of human normal to tumor cells. There exist drugs that interact with *CDKN1A*, *MAPK14*, and *RAD51AP1*.

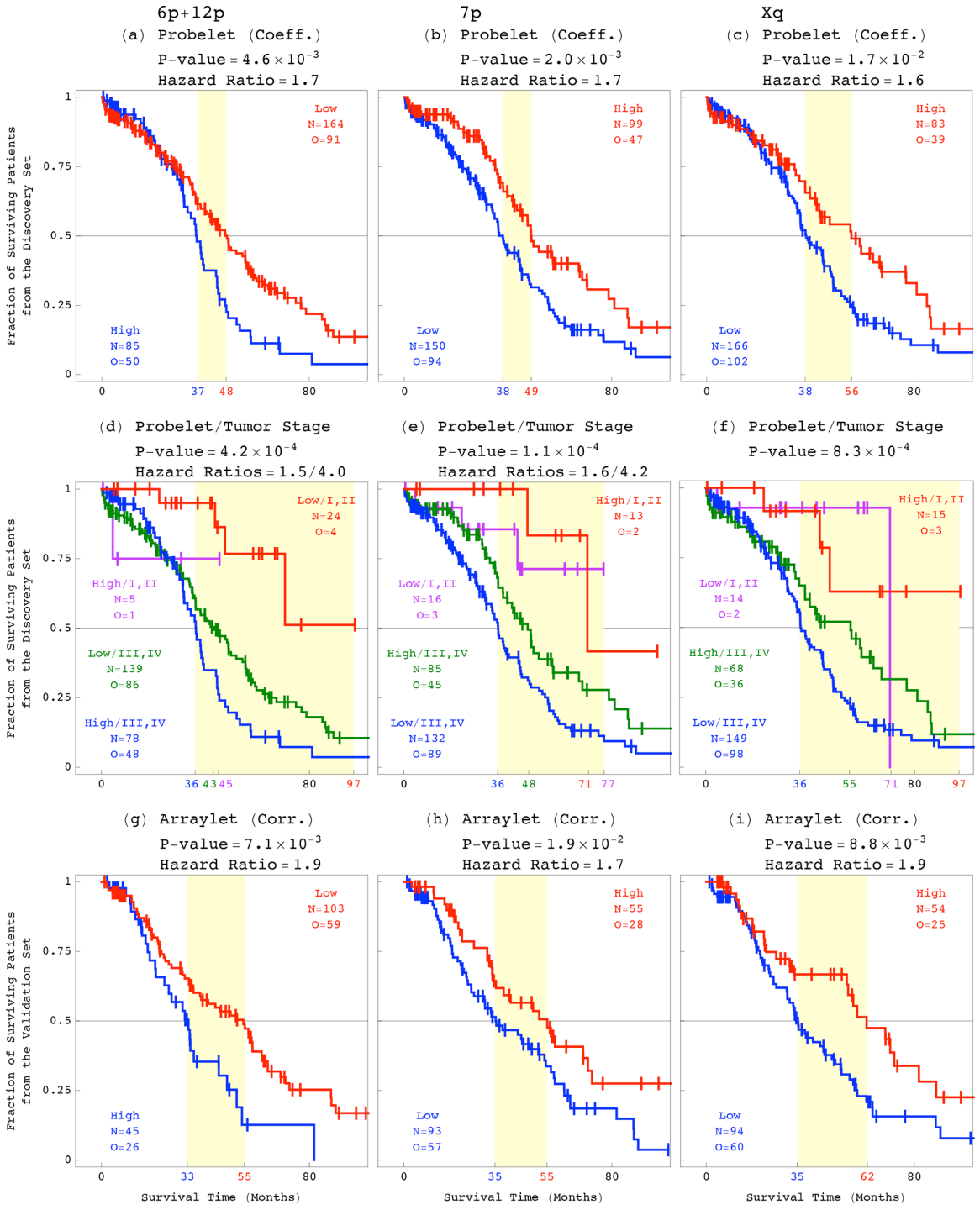
Hahn, Counter, Lundberg, Beijersbergen, Brooks & Weinberg, *Nature* 400, 464 (1999).

Genomic Predictors of OV Survival

Chromosome arm-wide patterns are correlated with, and possibly causally related to, ovarian cancer survival.

Despite recent large-scale profiling efforts, the best prognostic indicator of OV prior to the discovery of these patterns was the tumor's age at diagnosis.

The patterns are independent of stage, and combined with stage make better predictors than stage alone.

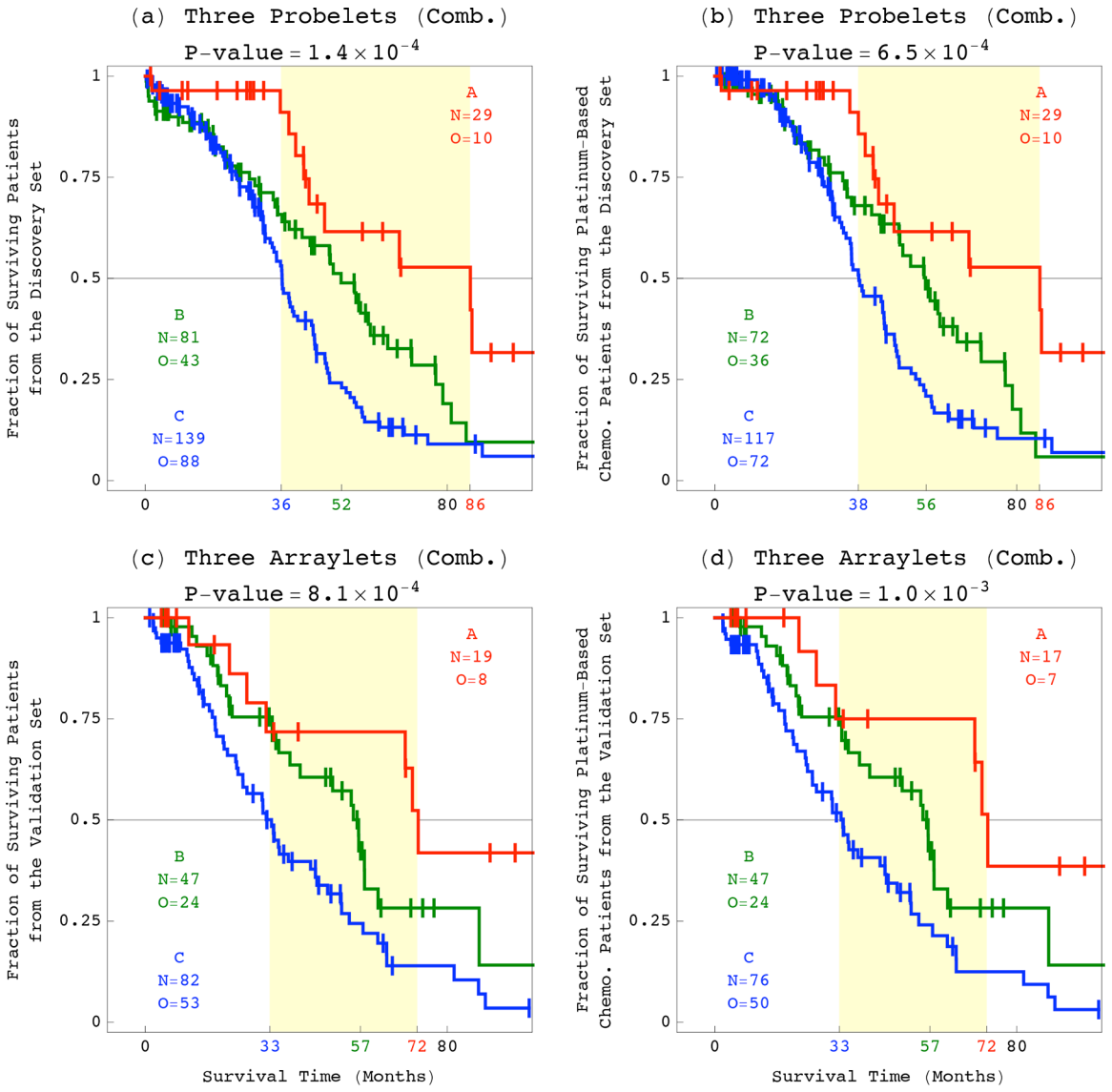


Genomic Predictors of OV Response to Platinum-Based Chemotherapy

~25% primary OV tumors are resistant, and most recurrent OV tumors develop resistance to platinum, the first-line treatment for >30 years.

There exist drugs for resistant tumors, but no pathology laboratory diagnostic exists that distinguishes between resistant and sensitive tumors before the treatment.

Multi-drug resistance can make the selection of the first drug the difference between success and failure.

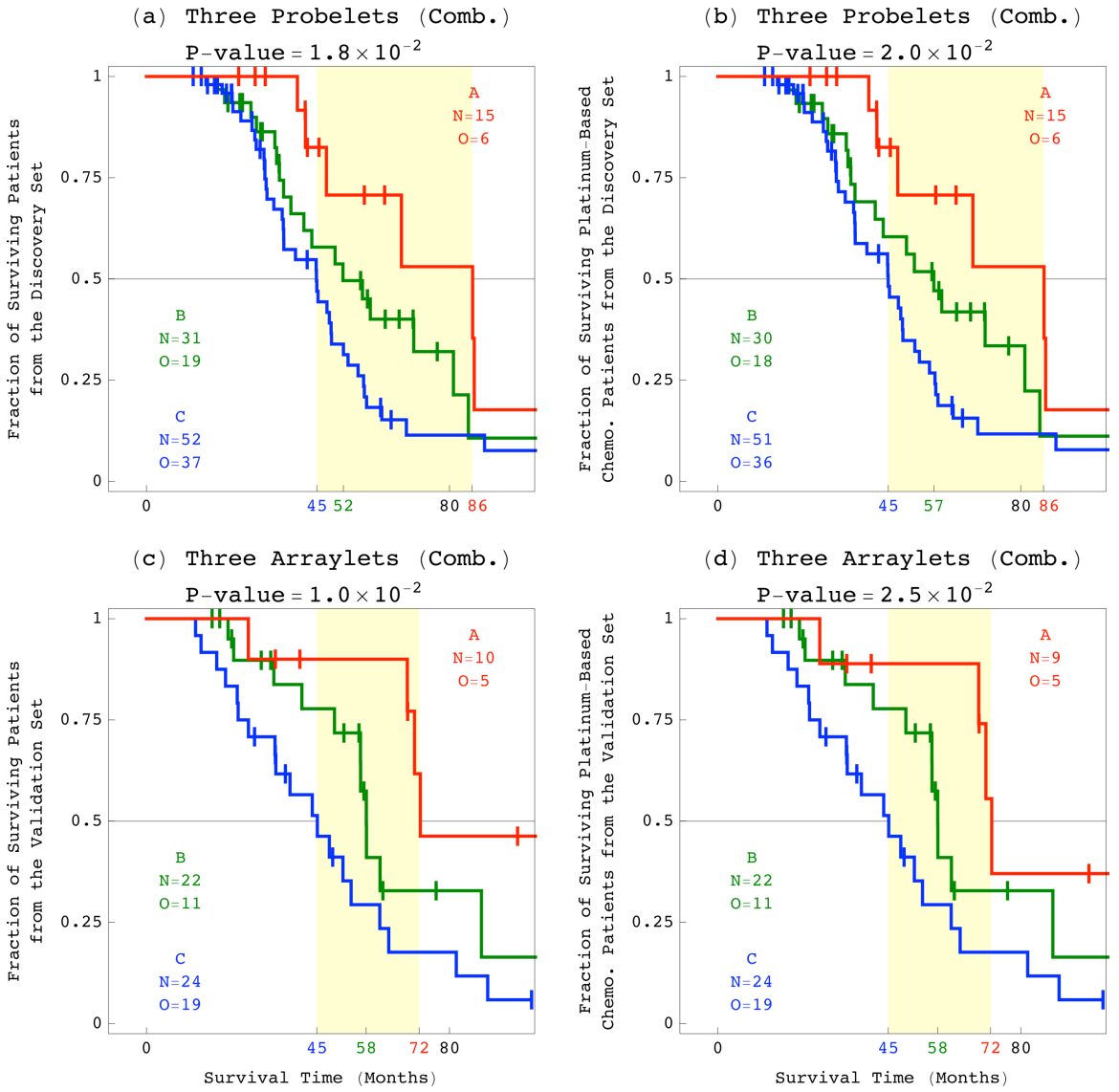


Genomic Predictors of OV Outcome Throughout the Course of the Disease

Correlated with survival and response to platinum throughout the course of the disease.

Independent of the primary tumor's stage and other indicators that are determined during treatment, i.e., the residual disease after surgery, the outcome of subsequent therapy, and the neoplasm status.

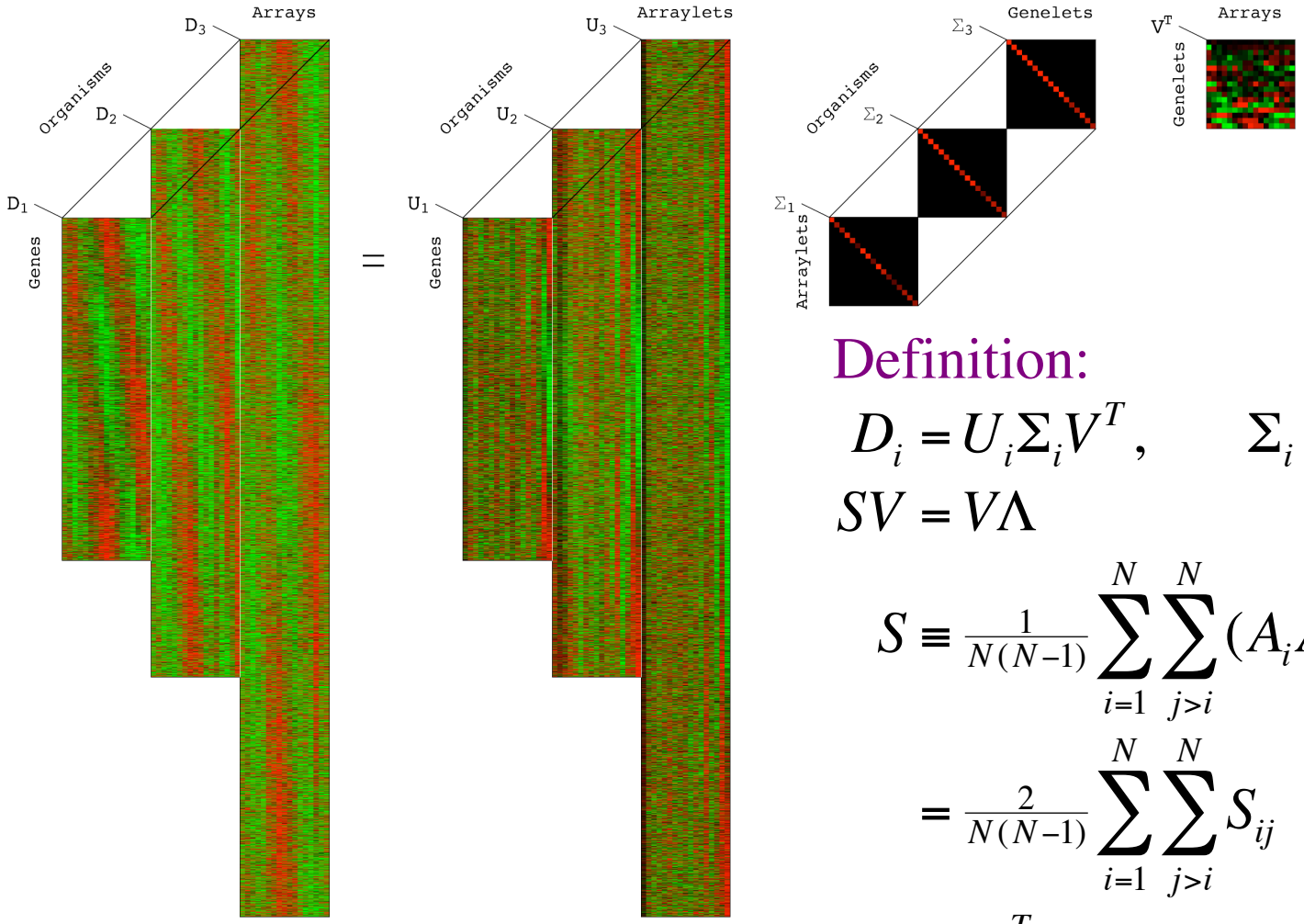
Independent of the time interval to tumor recurrence or progression.



Alter, *American Association for Cancer Research (AACR) Special Conference on Advances in Ovarian Cancer Research: Exploiting Vulnerabilities* (Orlando, FL, October 17–20, 2015). Philadelphia, PA: AACR (January 2016); http://dx.doi.org/10.1158/1557-3265.OVCA15-A60

HO GSVD for Comparative Analysis of Multiple Two-Dimensional Datasets

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



Definition:

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

$$SV = V\Lambda$$

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

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

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

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

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.

HO GSVD for Comparative Analysis of Multiple Two-Dimensional Datasets

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.

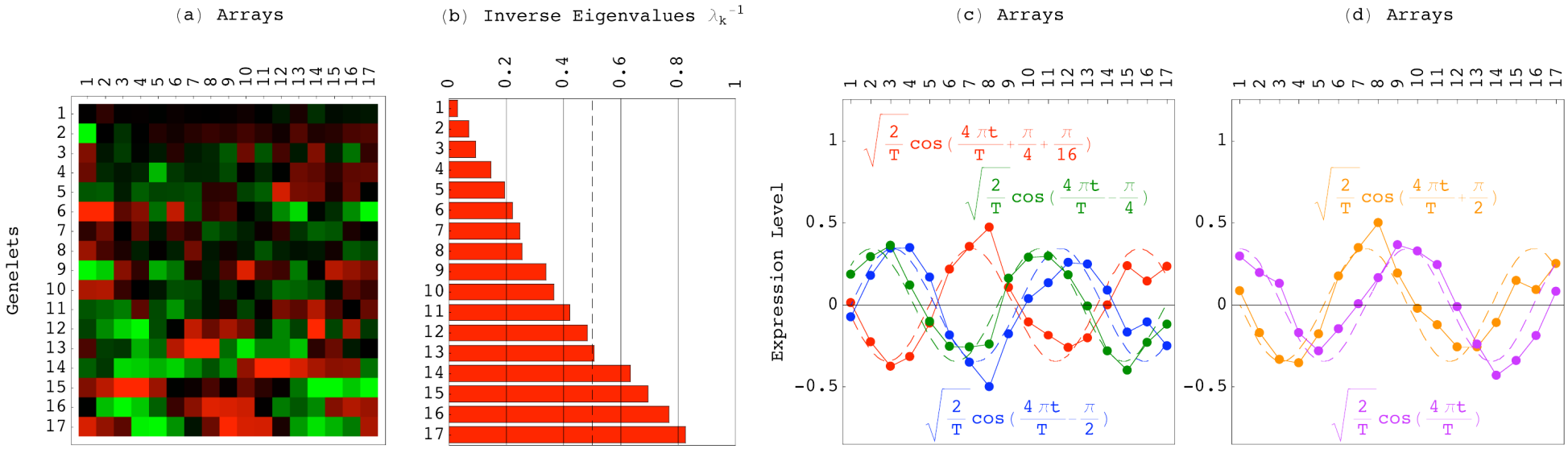
Supplementary Theorem 6 and Conjecture 1:
A role in iterative approximation algorithms.

Mathematical variables → biological reality

Genelets of almost equal significance in all datasets

→ processes common to all genomes:

Approximately Common HO GSVD Subspace



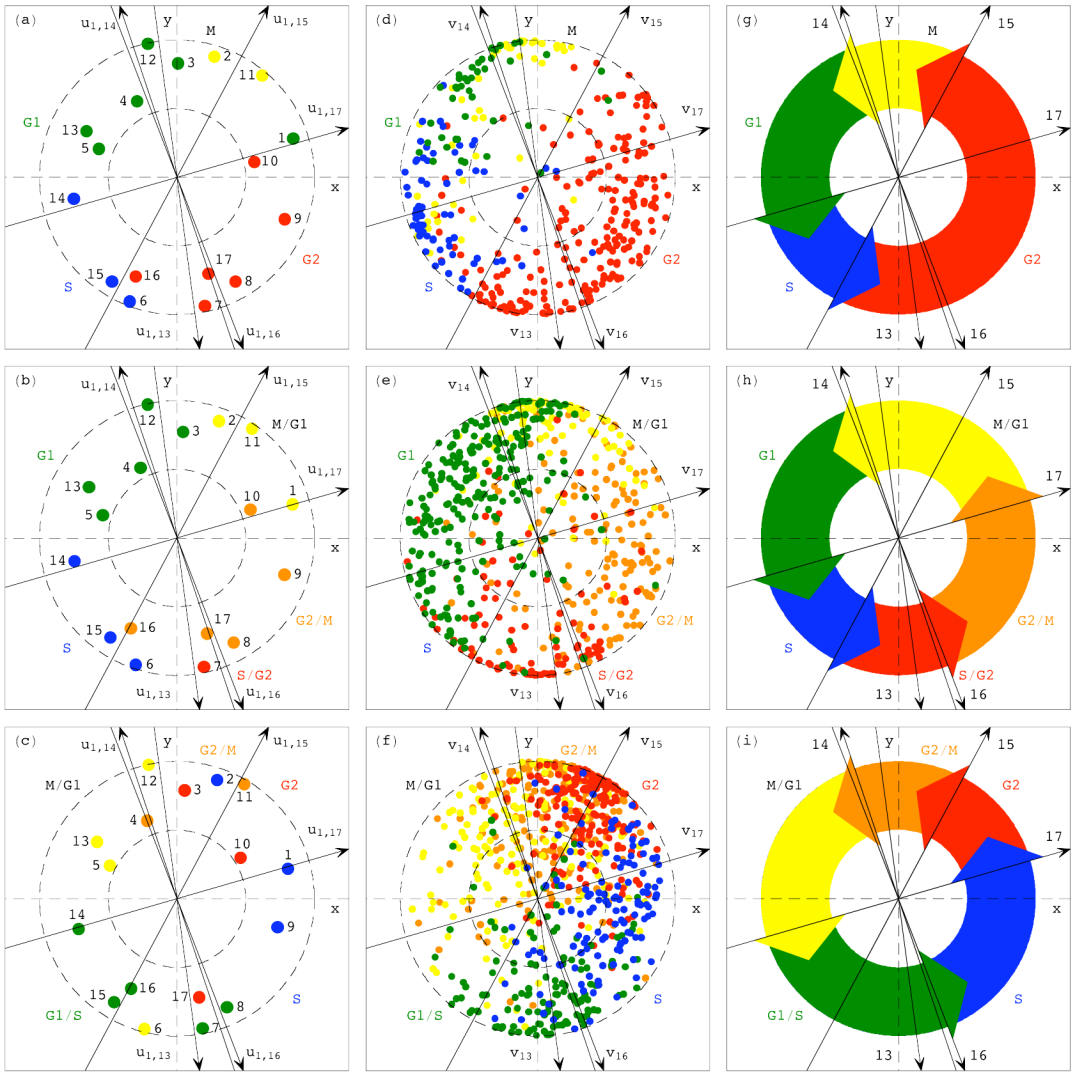
In a comparison of global cell cycle mRNA expression from *S. pombe*, *S. cerevisiae* and human, the approximately common HO GSVD subspace represents the cell cycle mRNA expression oscillations, which are similar among the datasets.

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

Mathematical operations → biological reality

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).

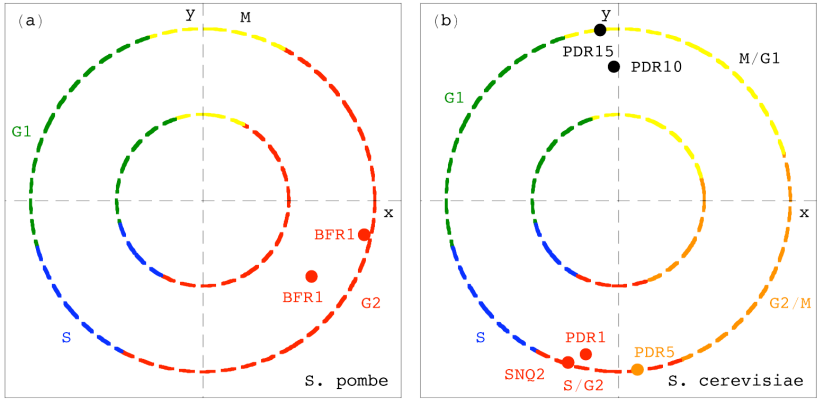
Saccharomyces cerevisiae

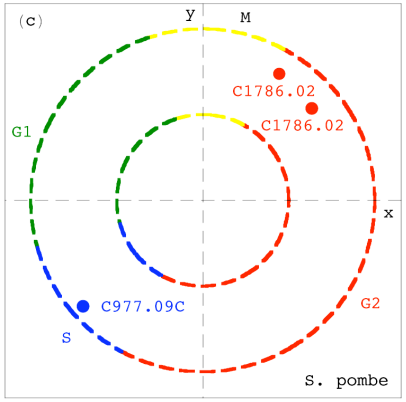
Spellman et al., *MBC* 9, 3273 (1998).

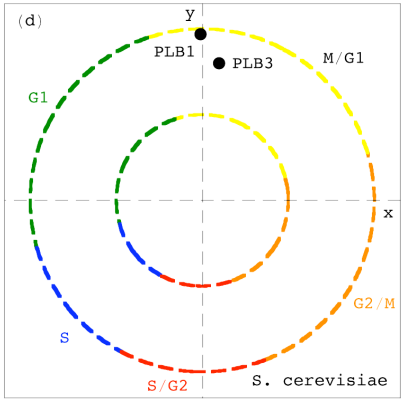
Human

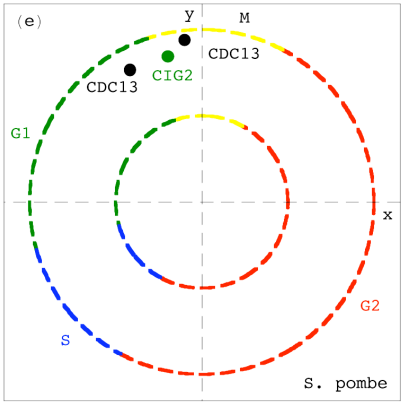
Whitfield et al., *MBC* 13, 1977 (2002).

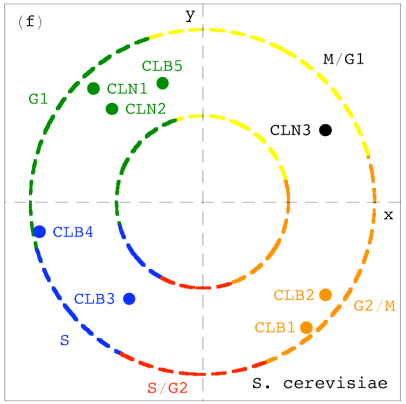
Simultaneous Classification Independent of Sequence Similarity

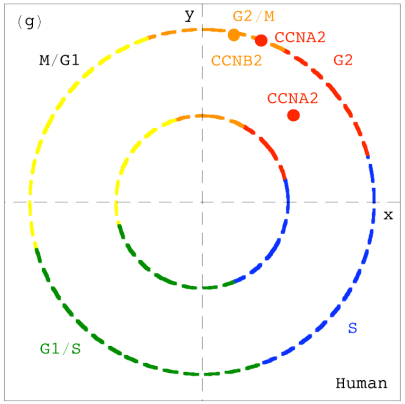












Genes of highly conserved sequences across the three organisms but significantly different cell cycle peak times are correctly classified.

ABC Transporter Superfamily Genes

Phospholipase B-Encoding Genes and B Cyclin-Encoding Genes

Patterns Underlie Principles of Nature: Statistics to Processes

→ Brownian motion.

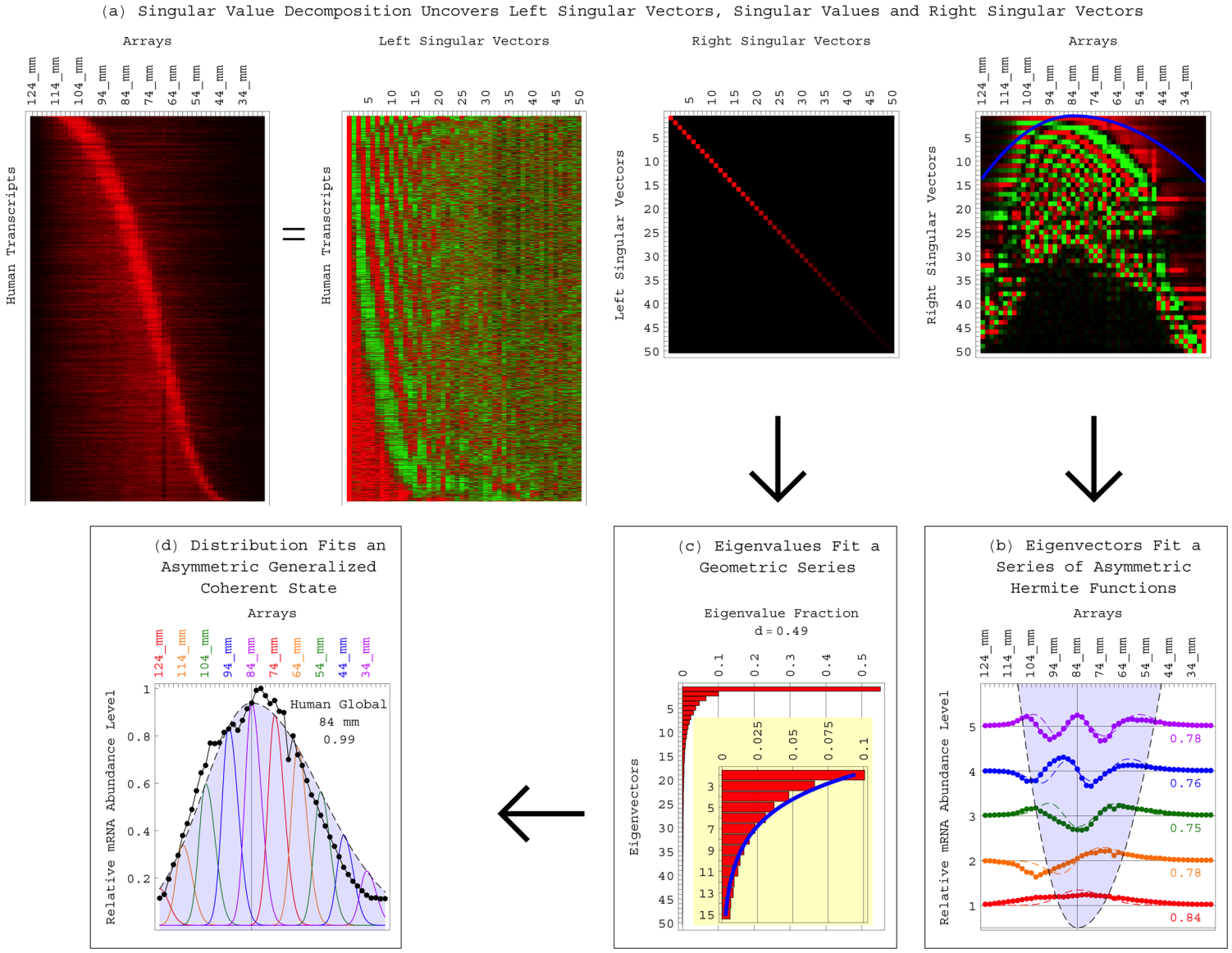
Einstein, *Ann Phys* 17, 549 (1905).

→ Bacterial sensitivity and resistance to viruses.

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

SVD Identifies Transcript Length Distribution Functions from DNA Microarray Data

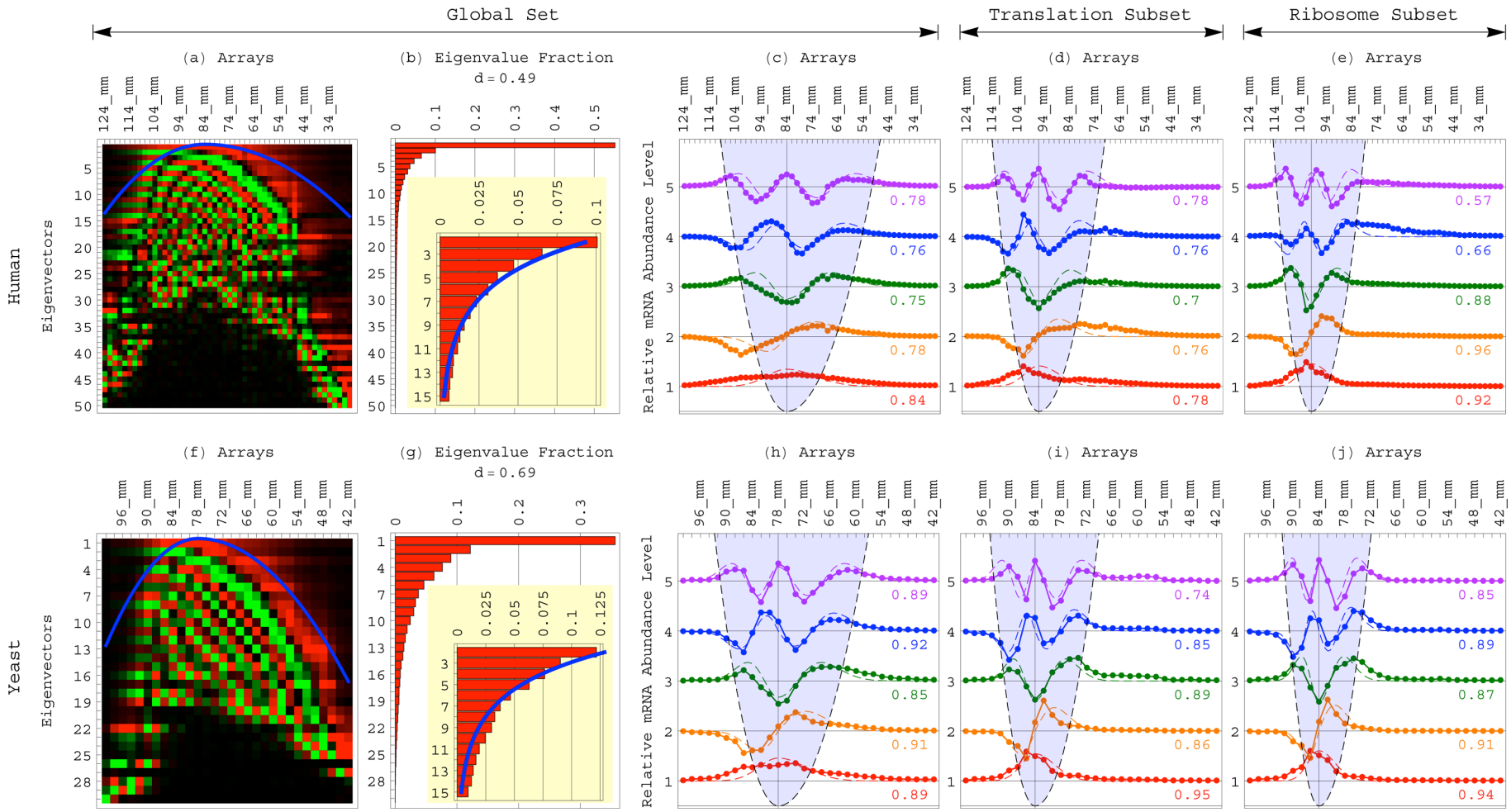
Bertagnolli, Drake, Tennessen & Alter, *PLoS One* 8, e78913 (2013);
http://alterlab.org/GBM_metabolism/



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

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

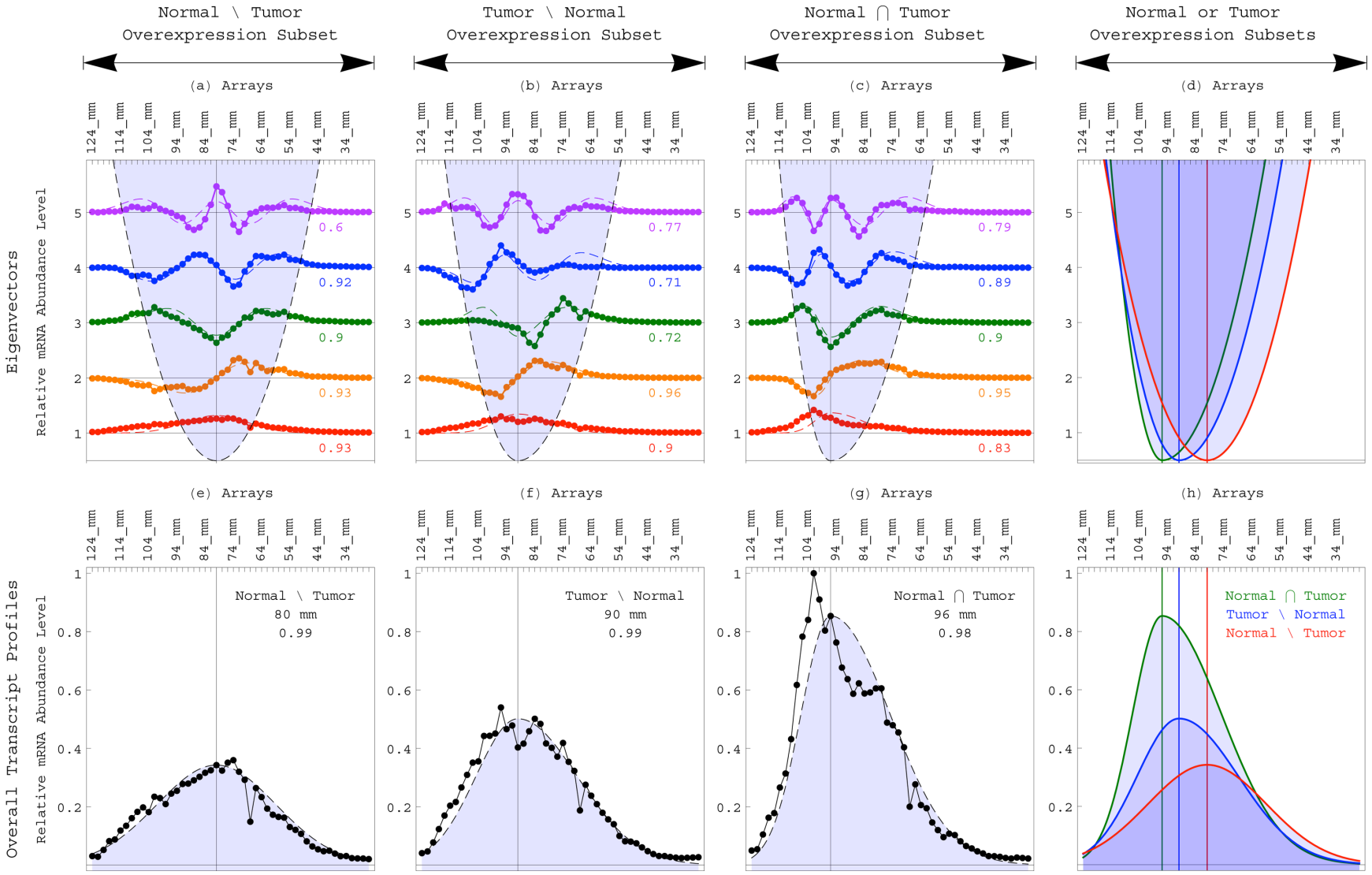
Drake & Alter, *Rao Conference at the Interface between Statistics and the Sciences* (Hyderabad, India, December 30, 2009 – January 2, 2010), Rao Best Poster Prize.



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 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.

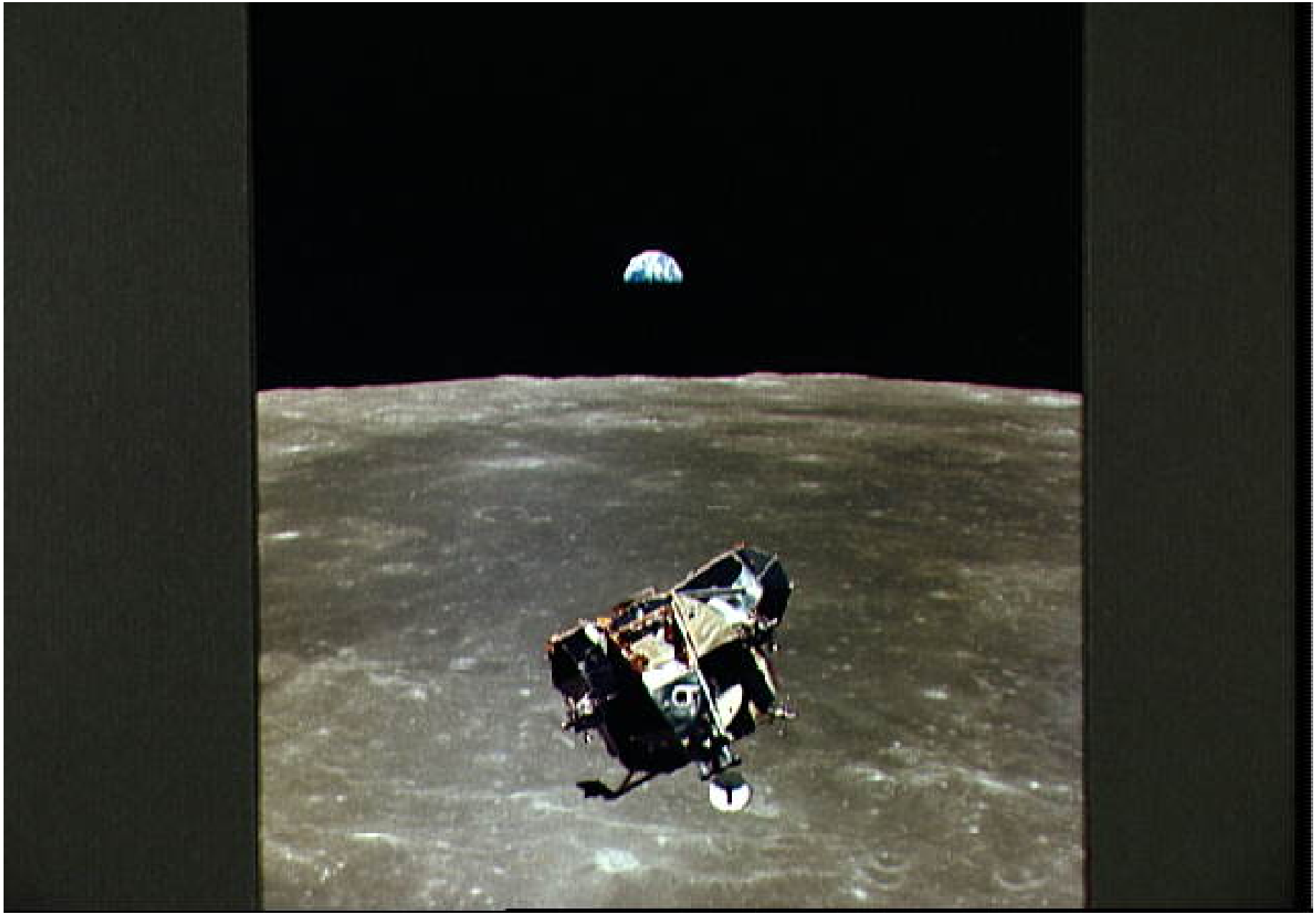
Alter & Golub, *PNAS* 103, 11828 (2006).

→ 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).

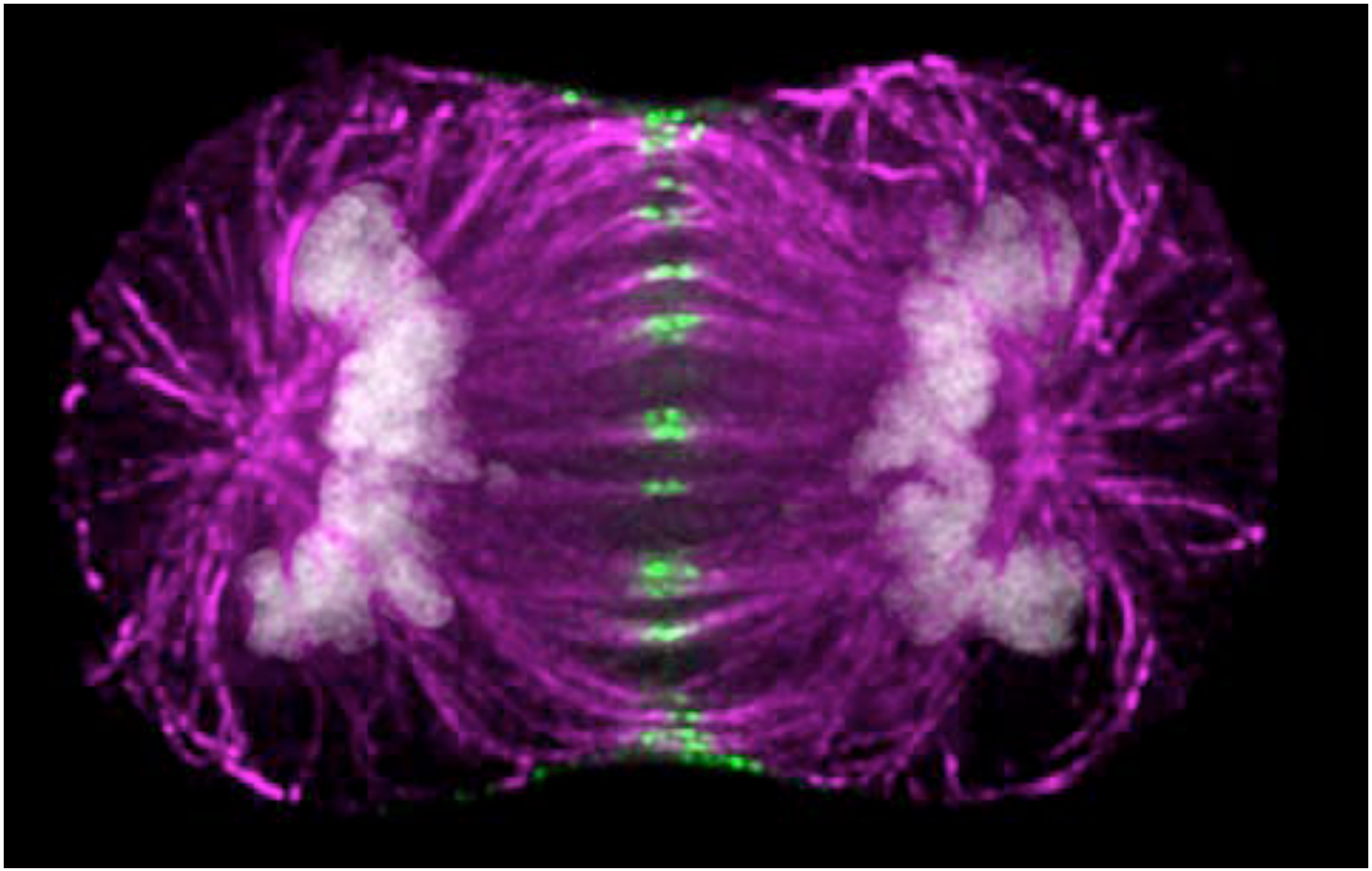
The interplay between mathematical modeling and experimental measurement is at the basis of the "effectiveness of mathematics" in physics.

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.

NCI U01 CA-202144

Multi-Tensor Decompositions for Personalized Cancer Diagnostics and Prognostics

http://physics.cancer.gov/network/UniversityofUtah.aspx;
http://alterlab.org/physics_of_cancer/

Co-Investigators:

Heidi A. Hanson
Utah Population Database, Utah

Randy L. Jensen
Neurosurgery, Utah

Cheryl A. Palmer
Pathology, Utah

Carl T. Wittwer
Pathology, Utah

Consultants:

Roger A. Horn
Mathematics, Utah

Commentators:

Matthew P. Scott
Developmental Biology, Stanford

Robert A. Weinberg
Biology, MIT

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 HG-000038

NHGRI R01 HG-004302

NSF CAREER DMS-0847173

NCI U01 CA-202144

Thank you!

Ph.D. Students:

Katherine A. Aiello, BE

Theodore E. Schomay, BE

Cody A. Maughan, BE

B.S. Students:

Kaitlin J. McLean, BM Physics

K99 Postdoc Alumni:

Jason M. Tennessen, Genetics

Ph.D. Alumni:

Preethi Sankaranarayanan, BE

Chaitanya Muralidhara, CMB

Sri Priya Ponnapalli, ECE

Larsson Omberg, Physics

Kayta Kobayashi, Pharmacy

B.S. Alumni:

Nicolas M. Bertagnolli, Math

Justin A. Drake, BME & SSC

Andrew M. Gross, BME & SSC

Joel R. Meyerson, BME & Gov

Physics-Inspired Multi-Tensor Decompositions

Create a single coherent model from multiple high-dimensional datasets. By using the complex structure of the datasets, rather than simplifying them as is commonly done, the frameworks can:

→ detect and remove experimental artifacts or batch effects;
→ identify and separate the biologically similar from the dissimilar;
→ uncover previously unknown phenomena.

Generalize the SVD from a single two-dimensional dataset to multiple three- and higher-dimensional datasets. The SVD underlies:

→ theoretical physics;
→ recommendation systems, e.g., the Netflix challenge;
→ Google's PageRank algorithm.

Find what others miss, and outperform algorithms that:

→ are sensitive to artifacts (e.g., hierarchical clustering);
→ require a-priori knowledge (e.g., analysis of variance);
→ require data modifications (e.g., Bayesian statistics or topological data analysis);
→ vary the single-dataset SVD (e.g., independent component analysis or randomized decompositions).

Nielsen, West, Linn, Alter et al., *Lancet* 359, 1301 (2002).

The SVD is also used for the stable computation of principal component analysis (PCA).

The SVD is Different Than PCA

→ **PCA assumes preprocessing of the data, which limits the data interpretation** (e.g., the SVD of a dataset can identify the probability distribution function that is sampled by the dataset with no a-priori assumptions; PCA cannot).

Alter & Golub, *PNAS* 103, 11828 (2006);
Cadima & Jolliffe, *Pak J Statist* 25, 473 (2009);
Bertagnolli, Drake, Tennessen & Alter, *PLoS One* 8, e78913 (2013).

→ PCA identifies patterns across the columns separately from patterns across the rows; **the SVD simultaneously computes the corresponding sets of patterns across the rows and columns, ensuring consistent data interpretation**.

Alter, Brown & Botstein, *PNAS* 97, 10101 (2000);
Fellenberg, Hauser, Brors, Neutzner, Hoheisel & Vingron, *PNAS* 98, 10781 (2001).

→ PCA, as it is programmed in most computational packages, is limited to classifying the data based upon the two or three patterns that capture most of the information in the data (e.g., variance in the case of column centering); **the SVD maintains all data patterns, and not just for data classification**.