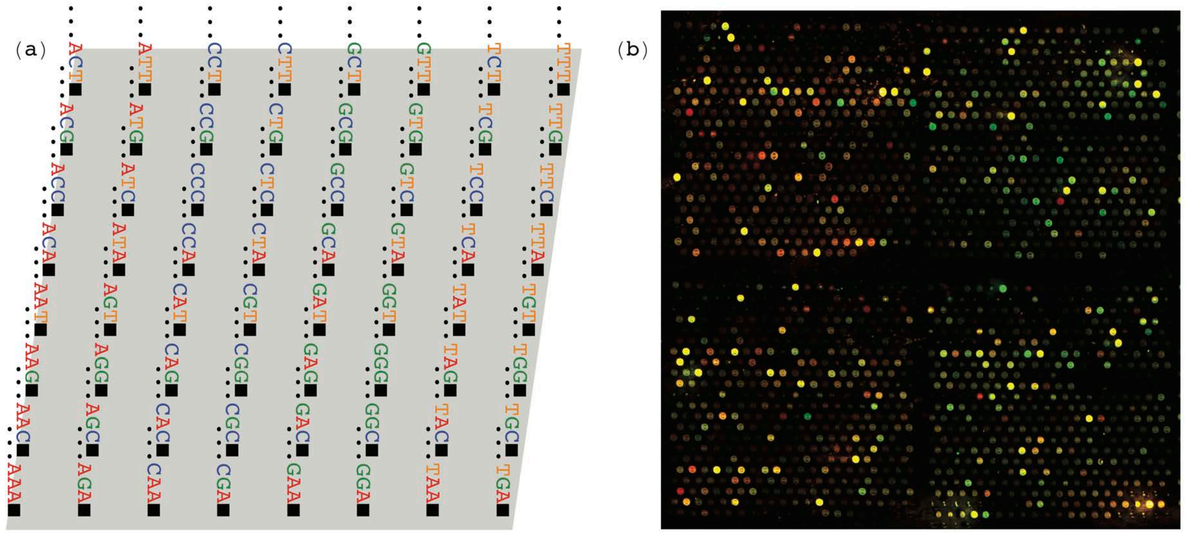
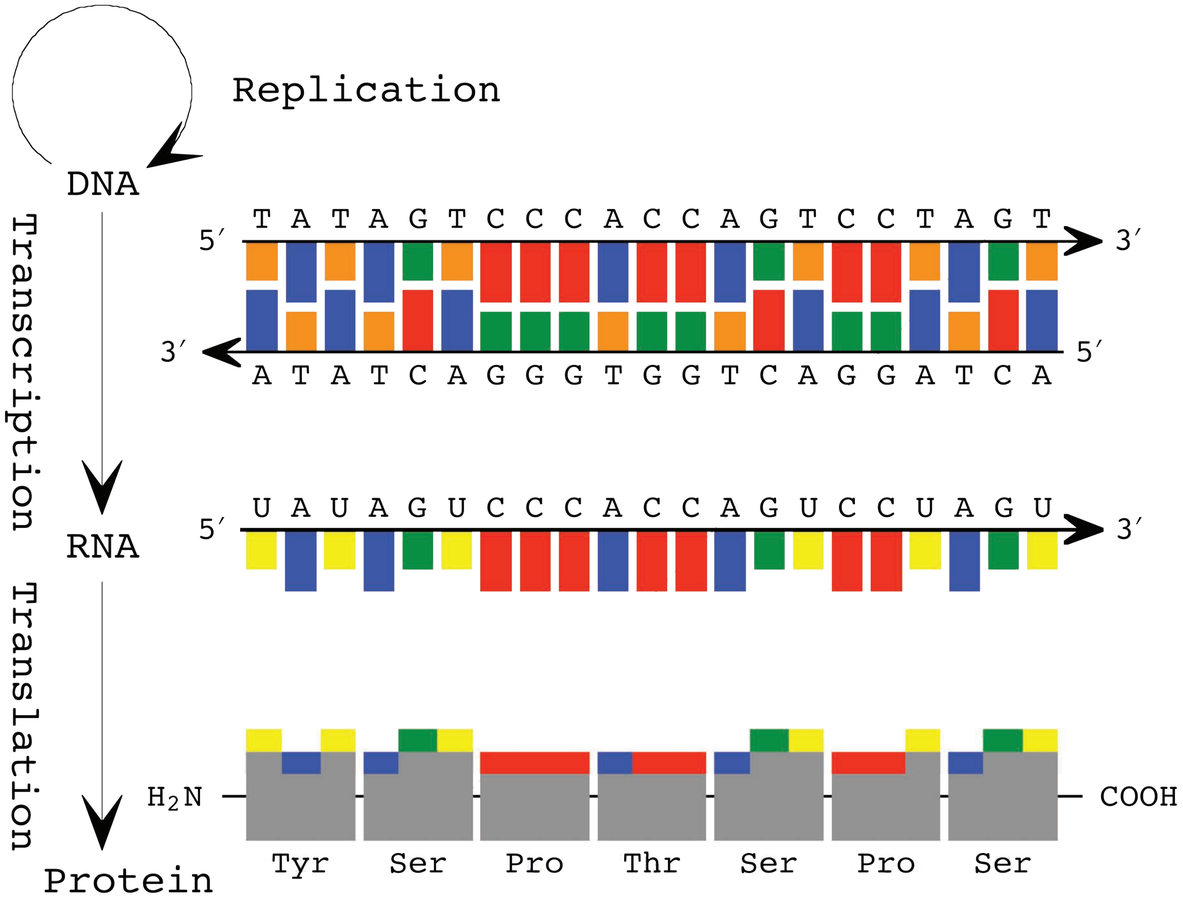
Multi-Tensor Decompositions for Personalized Medicine

Orly Alter

USTAR Associate Professor of Bioengineering and Human Genetics,
Scientific Computing and Imaging Institute and Huntsman Cancer Institute,
University of Utah
orly@sci.utah.edu
https://alterlab.org/

Omic Technologies Record Global Signals





The adoption of omic technologies in the cancer clinic is giving rise to an increasing number of large-scale high-dimensional datasets recording multiple patient-matched aspects of the disease.

Collins & Hamburg, *N Engl J Med* 369, 2369 (2013).

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

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

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

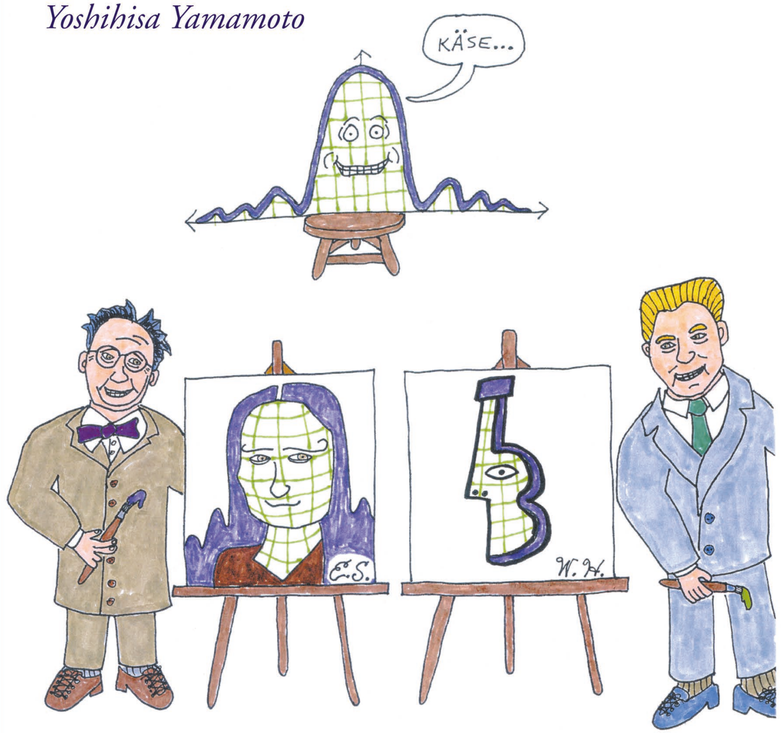
ALTER
YAMAMOTO

Quantum Measurement of a Single System

WILEY INTER-SCIENCE

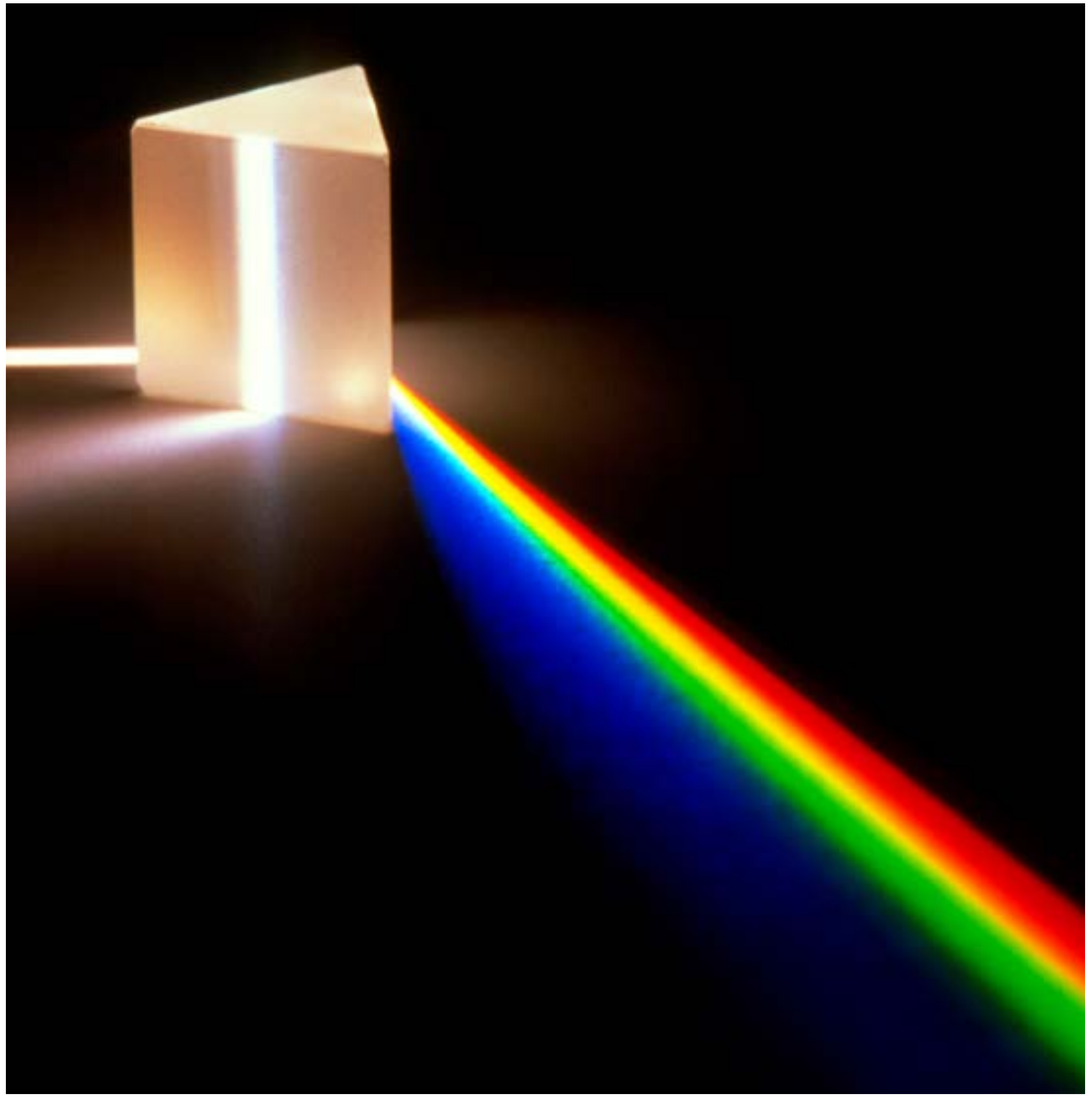
Quantum Measurement of a Single System

Orly Alter
Yoshihisa Yamamoto



Global Mathematical Vocabulary for Molecular Biological Discovery

The **singular value decomposition (SVD)** underlies the theoretical description of the physical world.



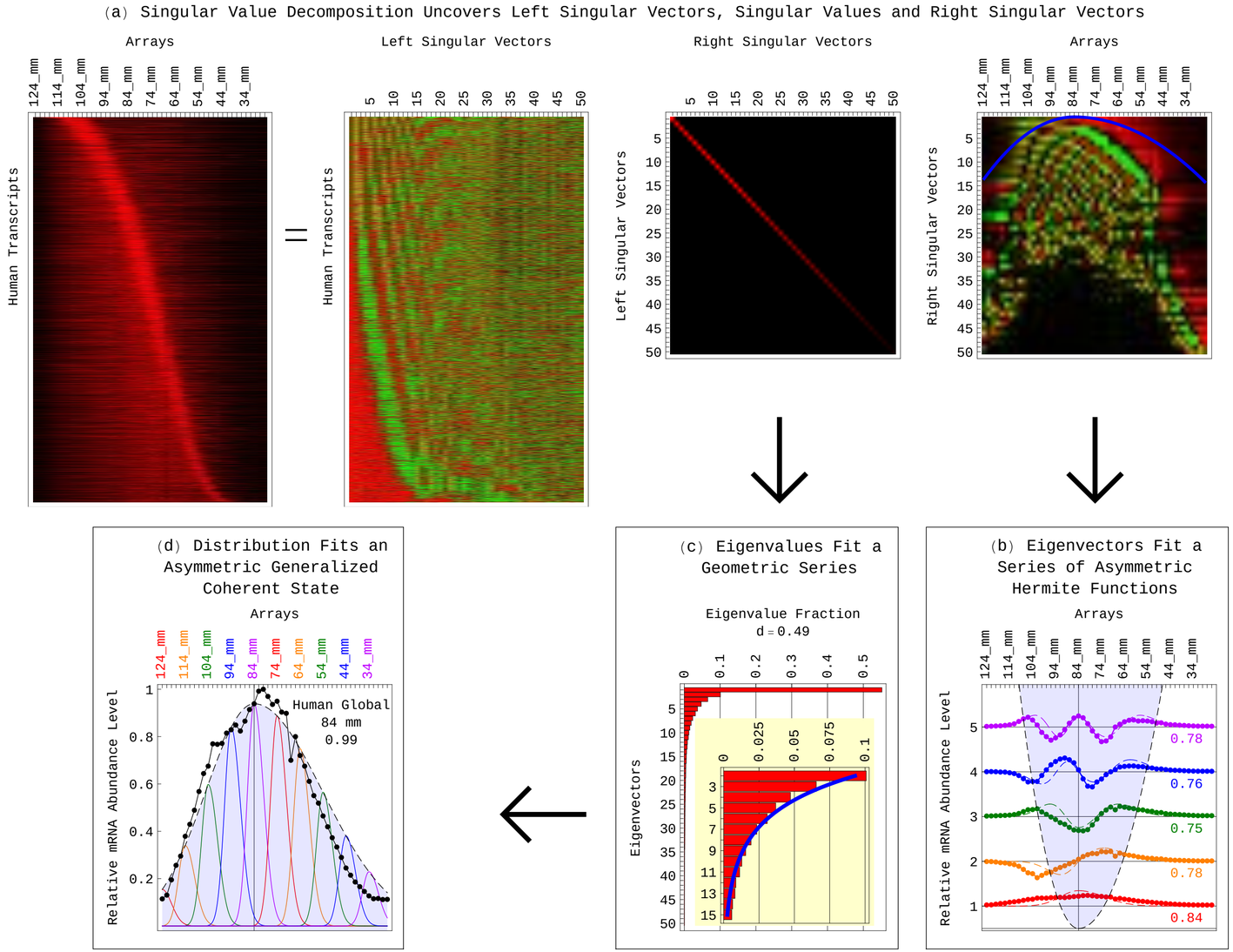
Alter & Yamamoto, *Quantum Measurement of a Single System*. New York, NY: Wiley-Interscience (2001); https://doi.org.10.1002/9783527617128

Global Mathematical Vocabulary for Molecular Biological Discovery

The **singular value decomposition (SVD)** underlies the theoretical description of the physical world, and possibly also the molecular biological world.

Generalizations of the SVD can be formulated that integrate and compare different data types.

The computations of the SVD and its generalizations scale with data sizes.



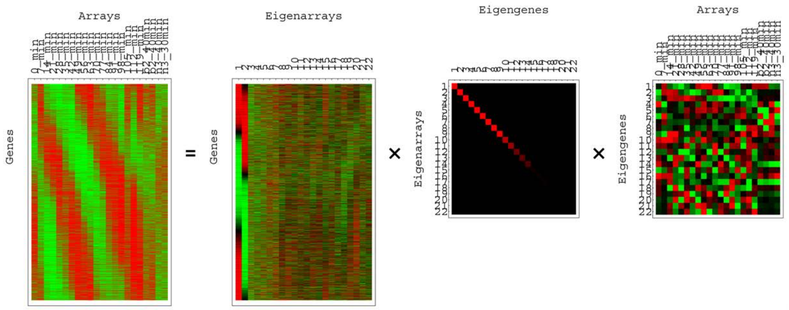
Alter & Golub, *PNAS* 103, 11828 (2006); https://alterlab.org/harmonic_oscillator/

Physics-Inspired Matrix and Tensor Models

The SVD and its generalizations are interpretable in terms of the known biology and batch effects that underlie, i.e., compose, the data.

SVD

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

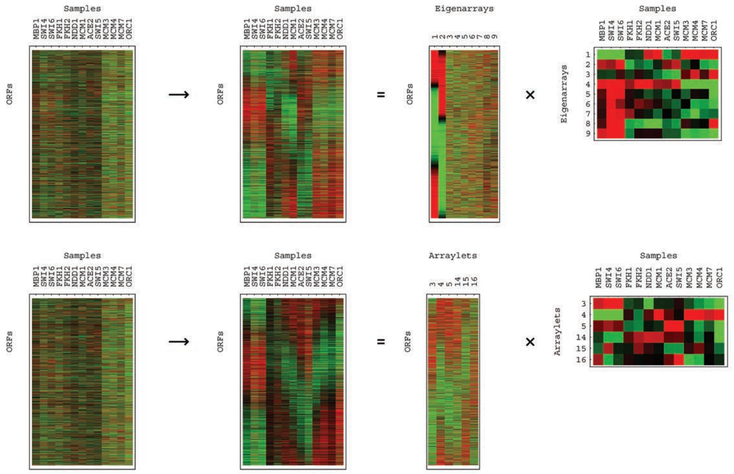


"Eigengenes" and "eigenarrays" → cellular processes and states in one dataset.

Eigenvalue Decomposition

Integrative Pseudoinverse

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

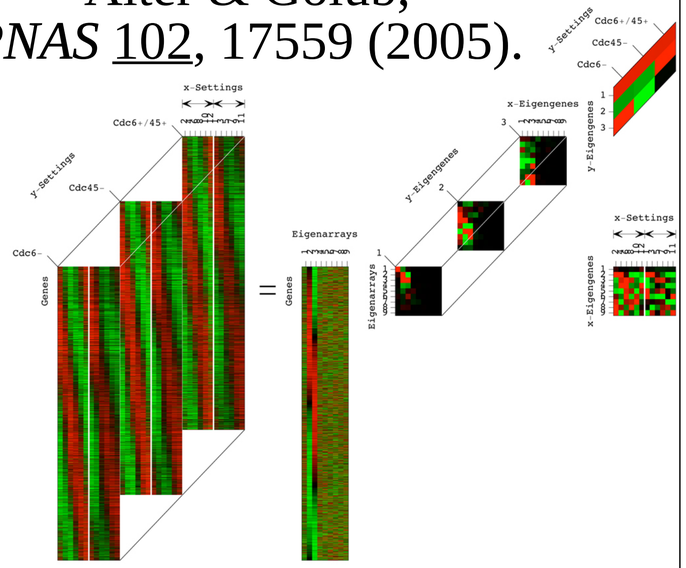


"Pseudoinverse correlation" → causal coordination between two datasets.

Inverse Projection

Integrative Tensor SVD

Alter & Golub,
PNAS 102, 17559 (2005).



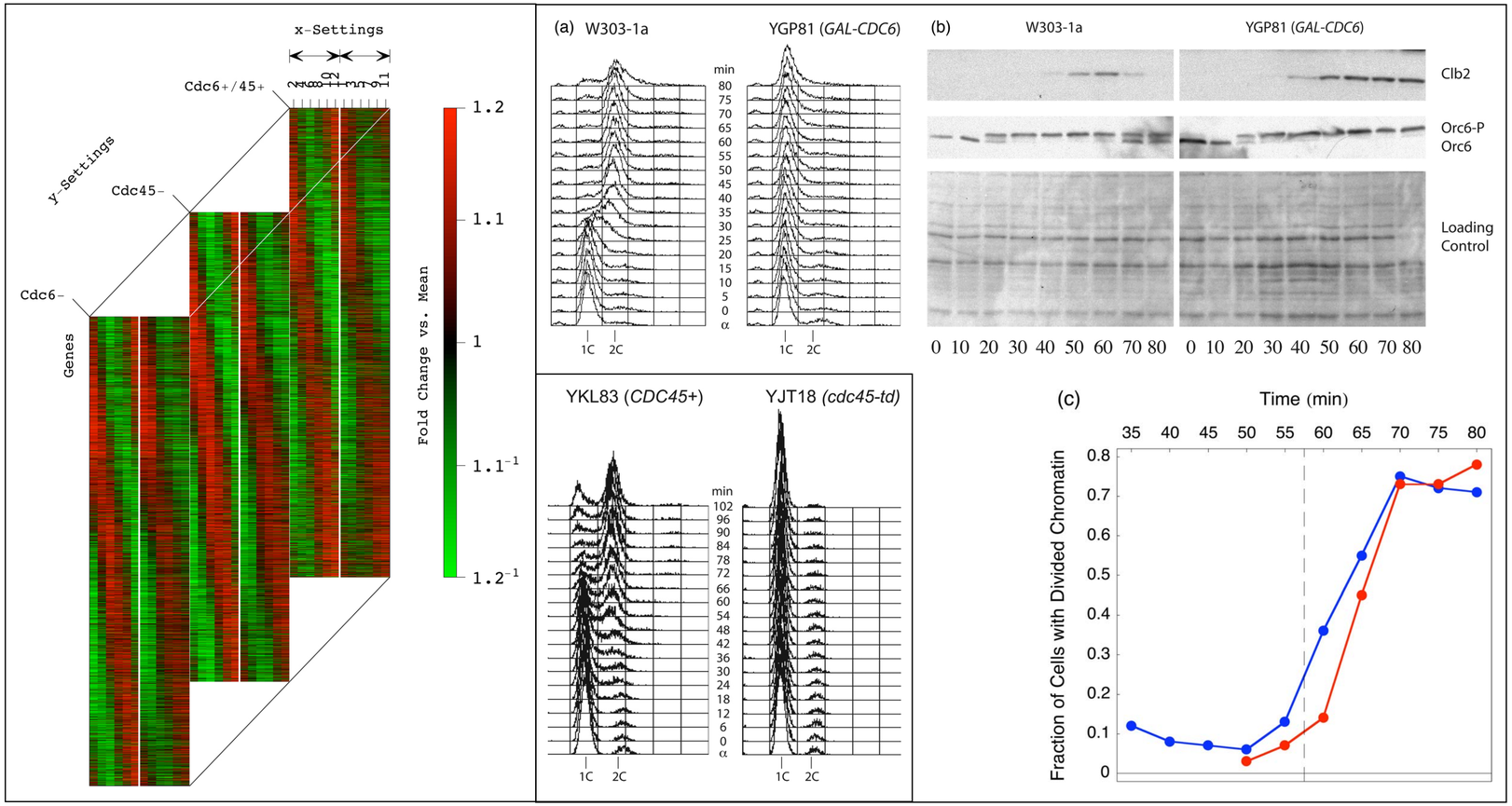
"x-" and "y-Eigengenes" and "eigenarrays" → interrelations among the processes and states of one higher-order dataset.

Experimental Verification of a Computationally Predicted Cellular Mechanism for DNA Replication to Affect RNA Expression

Omberg, Meyerson, Kobayashi, Drury, Diffley & Alter, *MSB* 5, 312 (2009); https://alterlab.org/verification_of_prediction/; Omberg, Golub & Alter, *PNAS* 104, 18371 (2007).

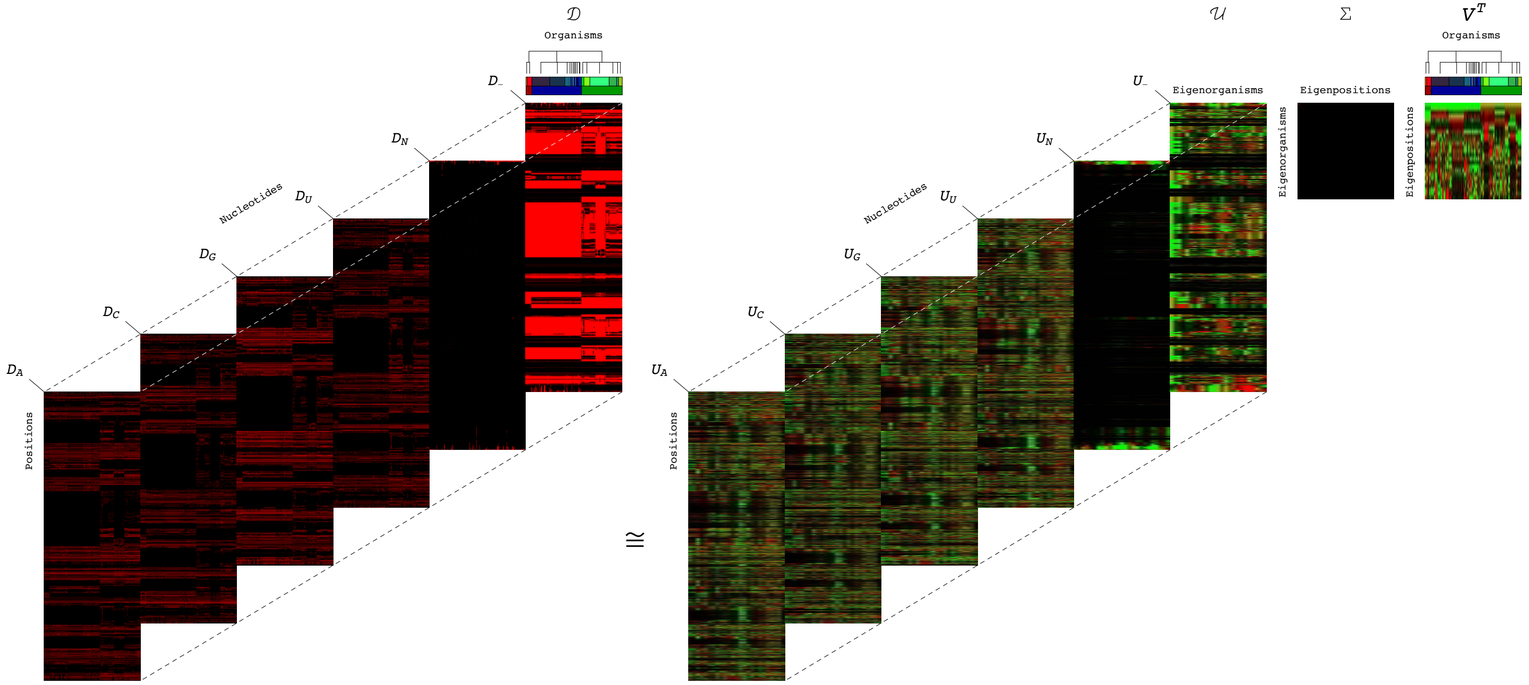
The SVD and its generalizations can **correctly** predict previously unknown and experimentally verifiable **global** mechanisms.

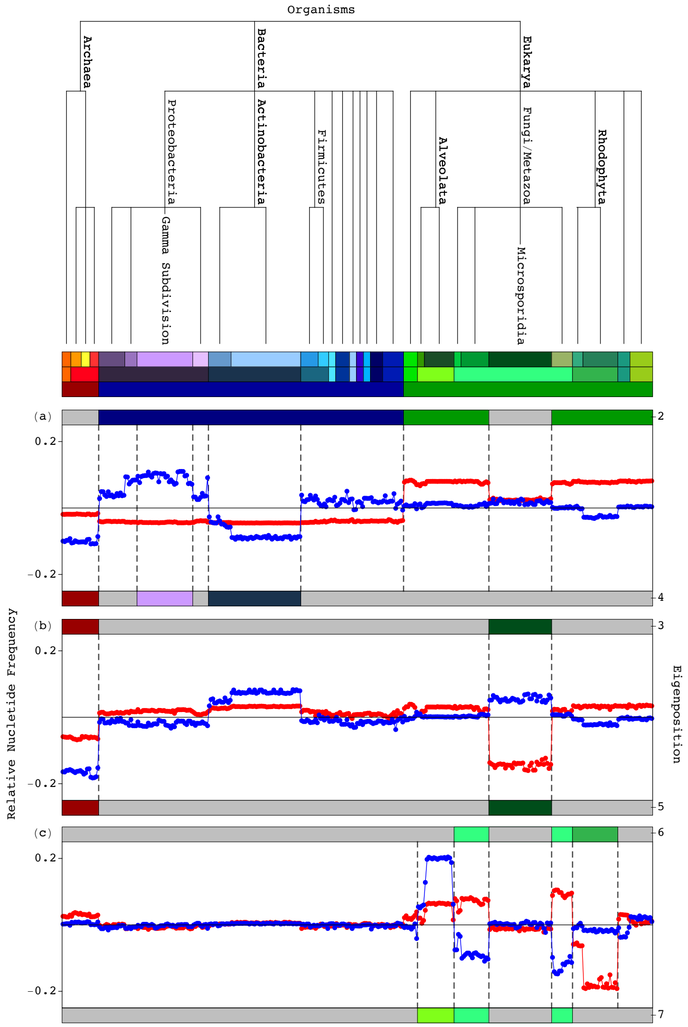
Replication origin licensing decreases the expression of genes with origins near their 3' ends, revealing that downstream origins can regulate the expression of upstream genes.



Computationally Predicted Evolutionary Mechanisms of Convergence and Divergence of Substructures in rRNA

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





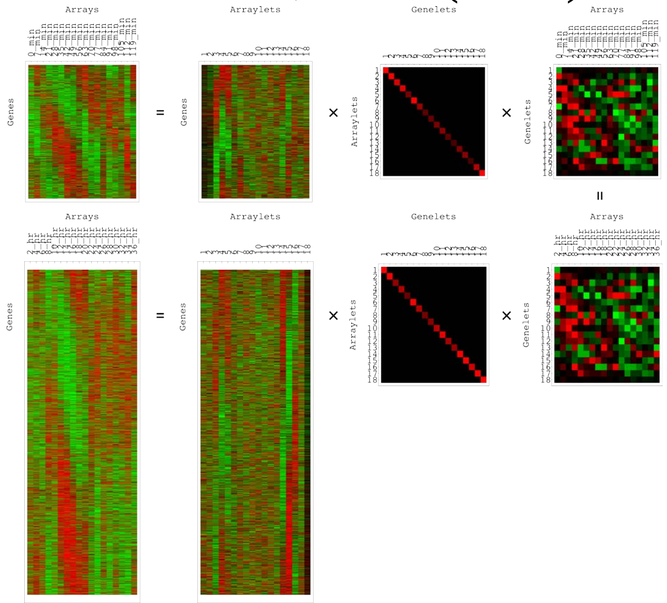
The tensor SVD discovers nucleotide variations across the taxonomic groups, consistent among the 16S, 23S, and 5S ribosomal RNAs (rRNAs), that map out known and new insertions and deletions of secondary substructures enriched in, e.g., unpaired adenosines (As) that form tertiary interactions.

Comparative Multi-Tensor Decompositions

Can discover novel accurate, precise, and actionable genotype-phenotype relationships, relevant to populations based upon small cohorts, that can be validated in clinical trials.

GSVD

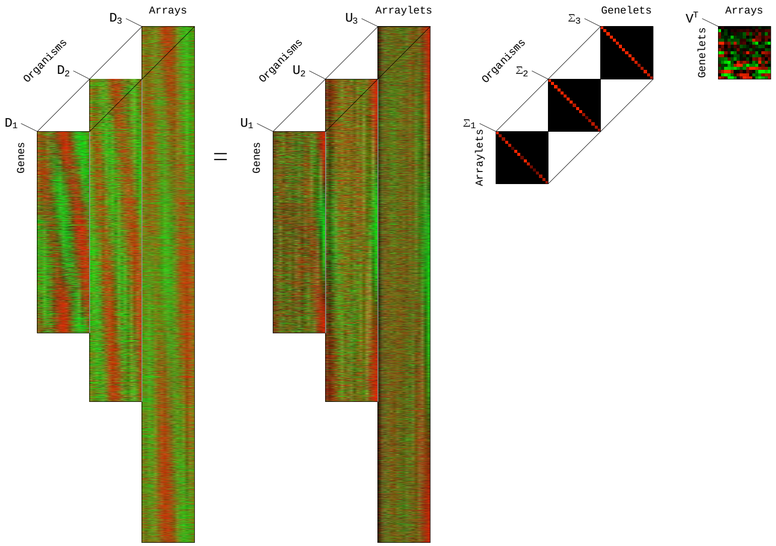
Alter, Brown & Botstein, *PNAS* 100, 3351 (2003).



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

HO GSVD

Ponnapalli, Saunders, Van Loan & Alter, *PLoS One* 6, e28072 (2011).

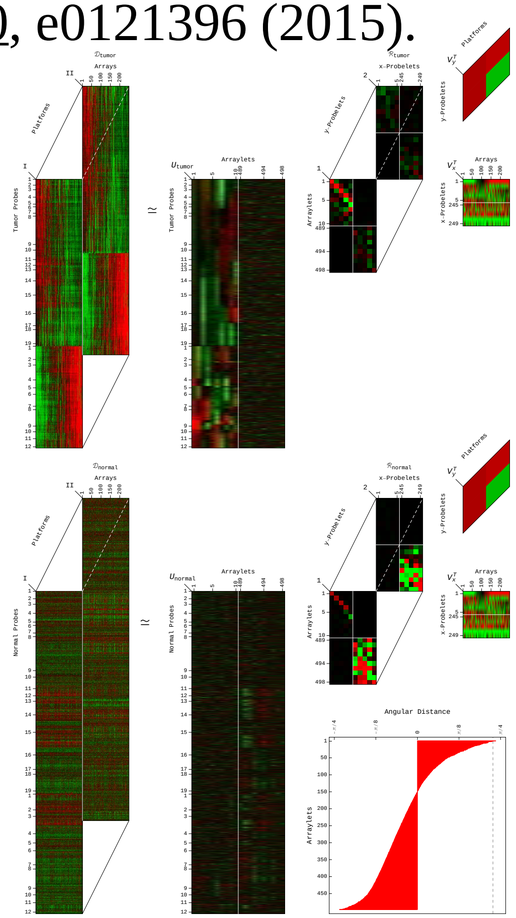


"Genelets" and "arraylets" → phenomena exclusive to one or more of, or common to multiple, datasets.

Tensor GSVD

Sankaranarayanan, et int., Alter, *PLoS One* 10, e0121396 (2015).

"Subtensors" → exclusive or common between two higher-order datasets, consistent or varying across orders.



Retrospective Clinical Trial Experimentally Validates Glioblastoma Genome-Wide Pattern of DNA Copy-Number Alterations Predictor of Survival

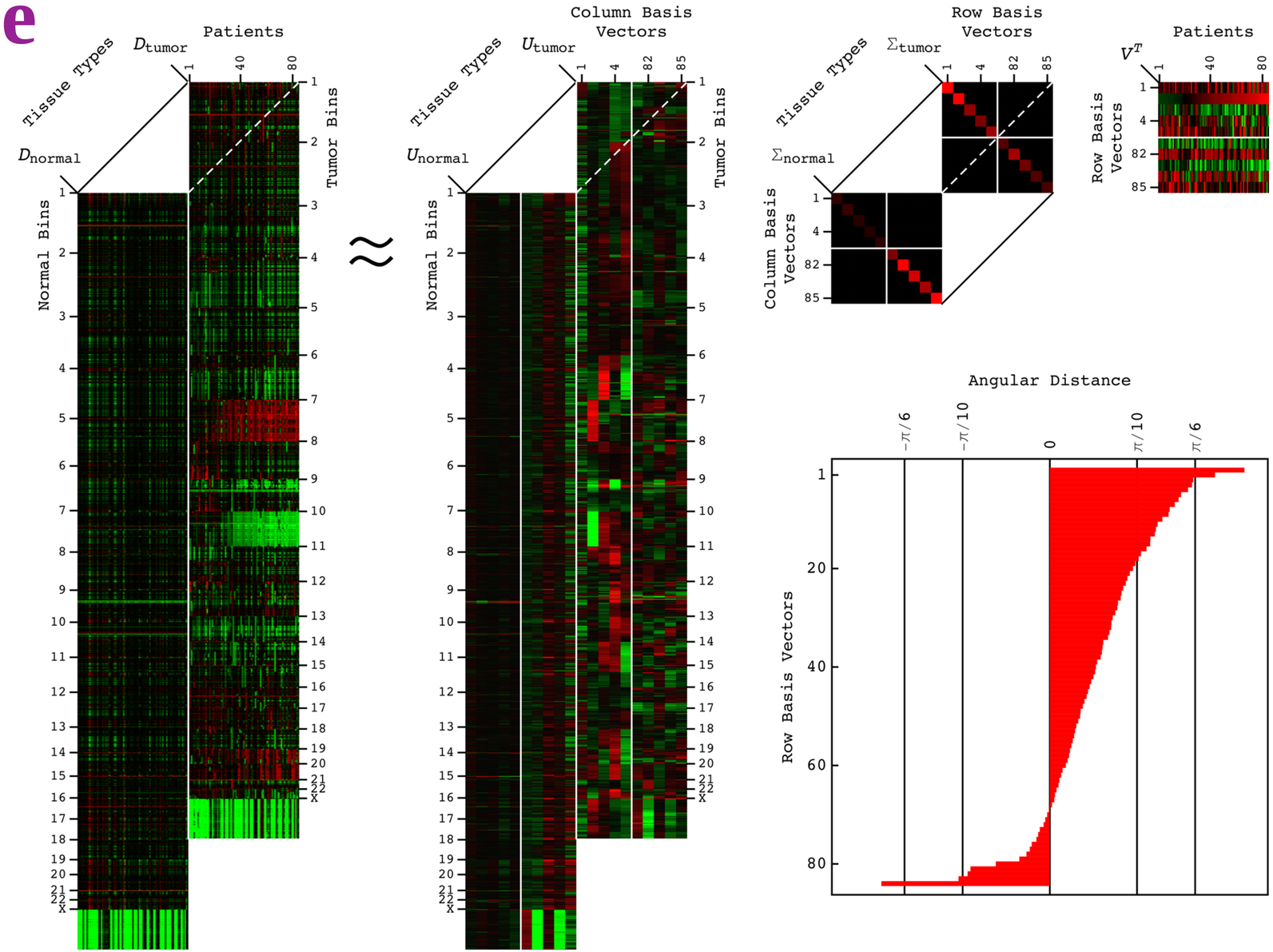
Ponnapalli, Bradley, Devine, Bowen, Coppens, Leraas, Milash, Li, Luo, Qiu, Wu, Yang, Wittwer, Palmer, Jensen, Gastier-Foster, Hanson, Barnholtz-Sloan & Alter, *APL Bioeng* 4, 026106 (2020); https://alterlab.org/GBM_retrospective_clinical_trial/

→ For 70 years, the best indicator of a patient's survival has been age at diagnosis.
Netsky et al., *J Neurosurg* 7, 269 (1950); Curran Jr., et int., Nelson, *J Natl Cancer Inst* 85, 704 (1993); Gorlia, et int., Stupp, *Lancet Oncol* 9, 29 (2008).

→ Recurring DNA copy-number alterations (CNAs) have been recognized as a hallmark of cancer for over a century and have been observed in glioblastoma (GBM) tumors.
Boveri, *Concerning the Origin of Malignant Tumours*. Jena, Germany: Gustav Fischer Verlag (1914).

→ Repeated previous attempts to associate a GBM tumor's DNA CNAs with a patient's outcome failed, including previous studies of data from the Cancer Genome Atlas that (TCGA) used other methods.
Weber, et int., Cremer, *Lab Invest* 74, 108 (1996); Wiltshire, et int., Bigner, *Neuro Oncol* 2, 164 (2000); Misra, et int., Feuerstein, *Clin Cancer Res* 11, 2907 (2005).

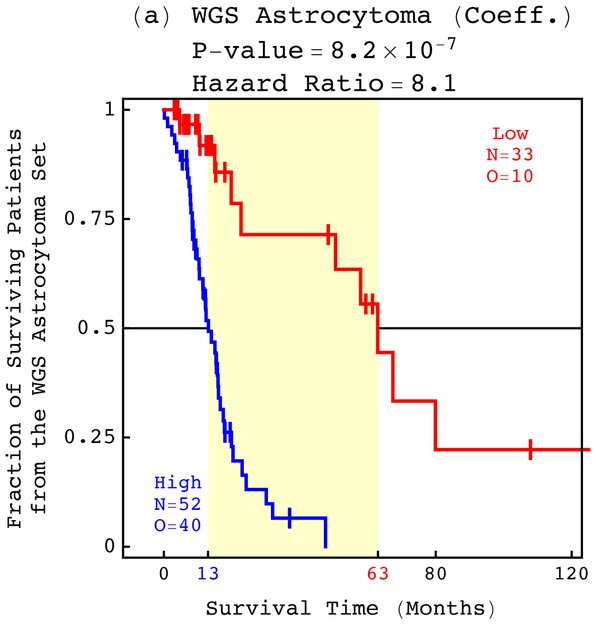
Mathematically Universal Biologically Consistent Genotype-Phenotype

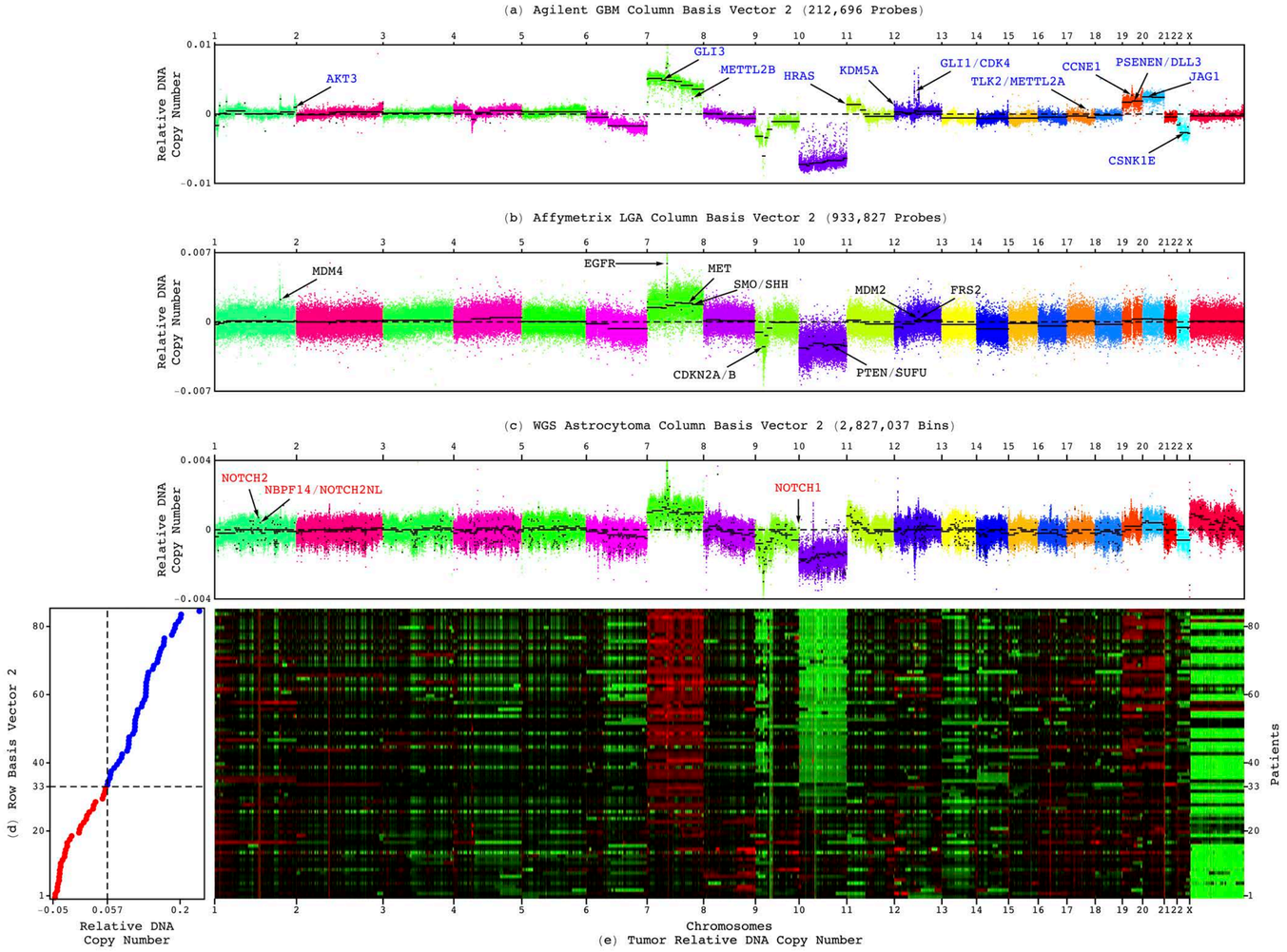


Lee, et int., Alter, *PLoS One* 7, e30098 (2012); Aiello & Alter, *PLoS One* 11, e0164546 (2016); Aiello, Ponnapalli & Alter, *APL Bioeng* 2, Special Topic: Bioengineering of Cancer 031909 (2018).

Invariably Uncovered by and Only by the Generalized SVD (GSVD)

Like the Agilent GBM and Affymetrix lower-grade astrocytoma patterns, the whole-genome sequencing (WGS) astrocytoma pattern is **correlated with a shorter, roughly one-year median survival time**.



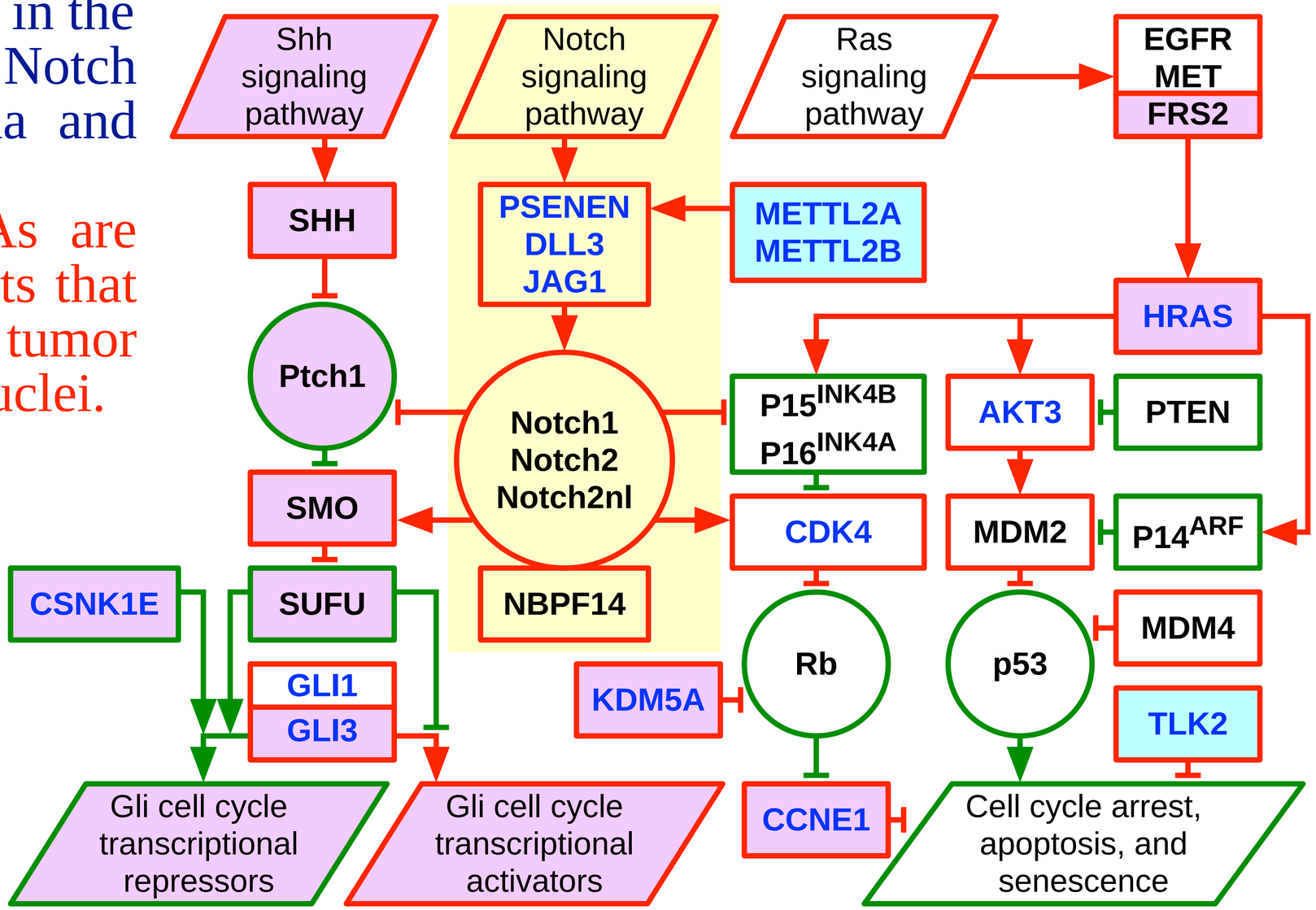


TCGA Research Network, *Nature* 455, 1061 (2008); *N Engl J Med* 372, 2481 (2015).

Encodes for Transformation via Ras, Shh, and Notch

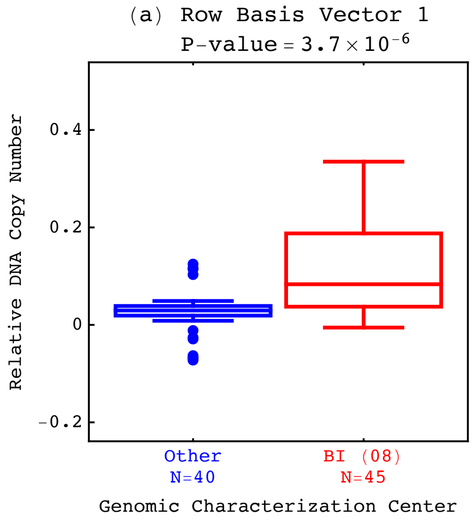
Includes most CNAs known in GBM, e.g., in the rat sarcoma (Ras) pathway, and as many previously unrecognized, e.g., in the sonic hedgehog (Shh) and Notch pathways in medulloblastoma and neuroblastoma.

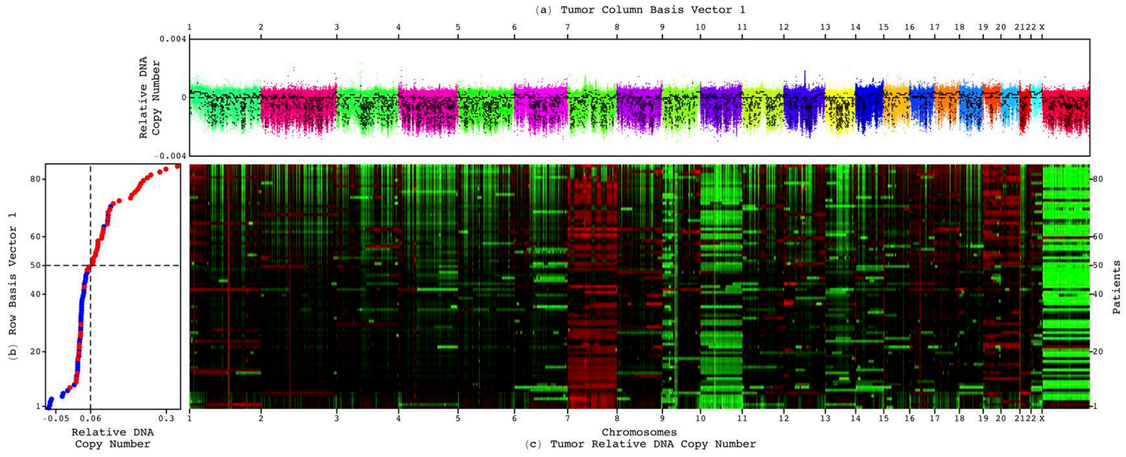
Some of these natural CNAs are analogous to artificial elements that transform human normal into tumor cells with grossly polyploid nuclei.

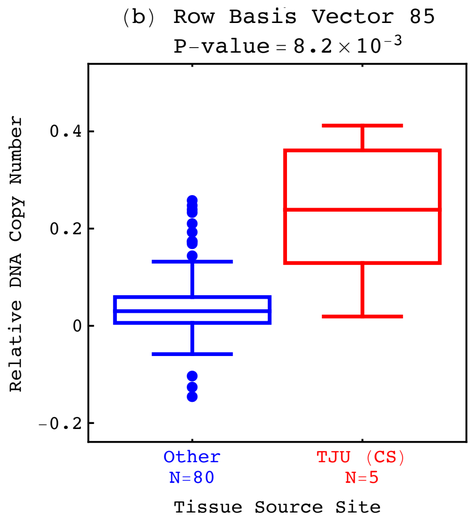


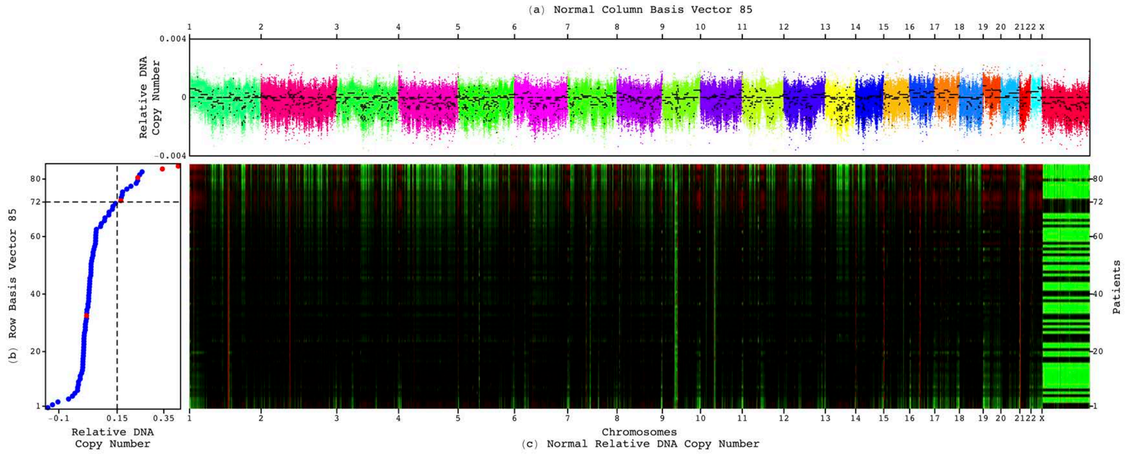
Waldman, et int., Vogelstein, *Nature* 381, 713 (1996); Hahn, et int., Weinberg, *Nature* 400, 464 (1999); Irianto, et int., Discher, *Curr Biol* 27, 210 (2017); Mukherjee, et int., Fonkem, *J Neuropathol Exp Neurol* 79, 562 (2020).

GSVD Blind Separation of WGS Batch Effects





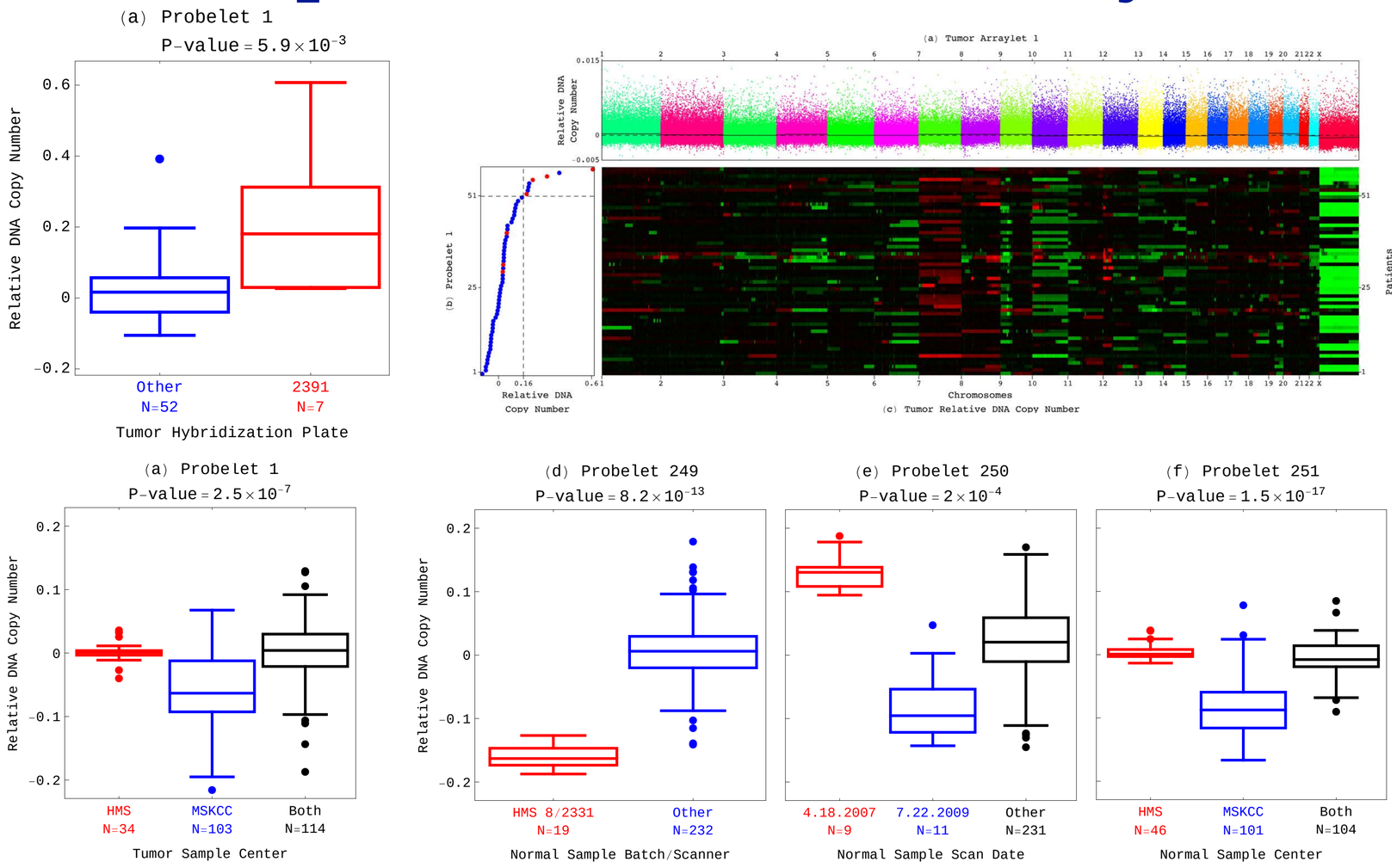




Guanine-cytosine (GC) content effects vary in magnitude between batches.

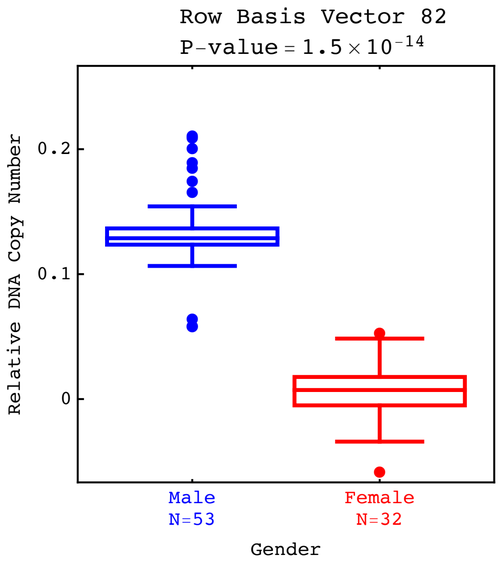
Roberts, Carneiro & Schatz, *Genome Biol* 14, 405 (2013).

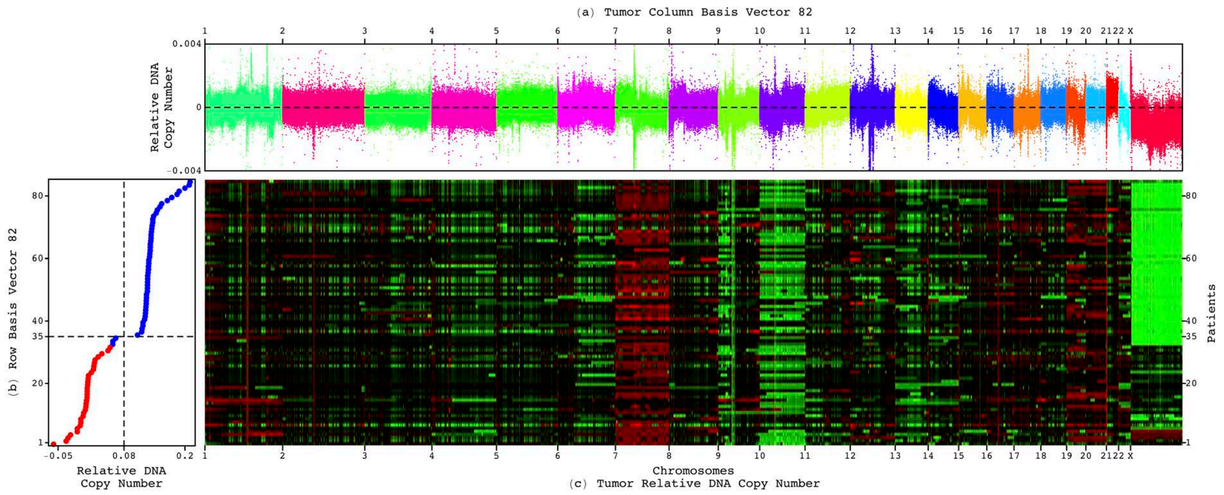
GSVD Blind Separation of Microarray Batch Effects

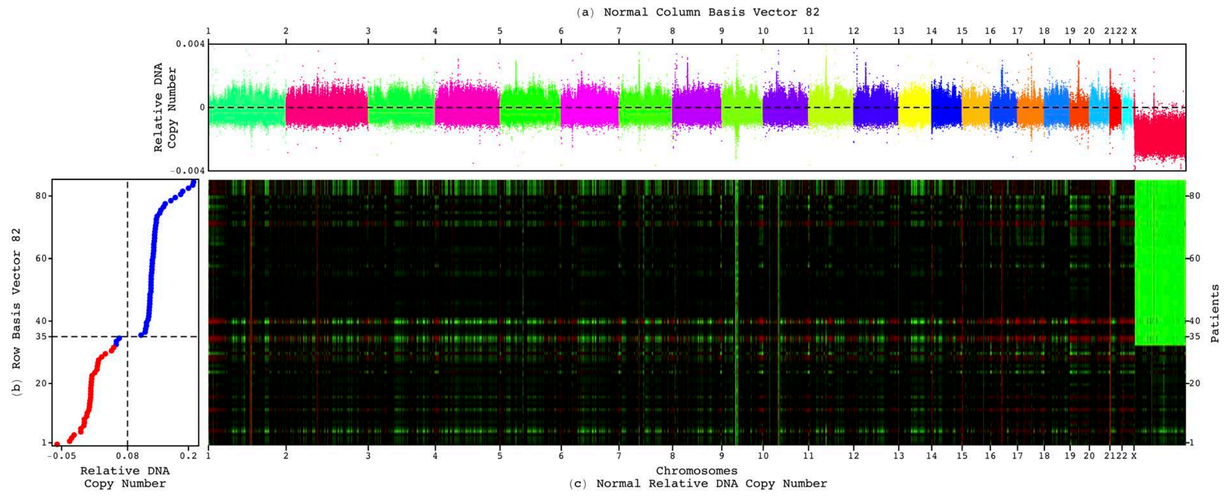


WGS with Affymetrix single-nucleotide polymorphism (SNP) and Agilent comparative genomic hybridization (CGH) microarrays represent the main genomic profiling technologies.

GSVD Blind Separation of Normal Variations





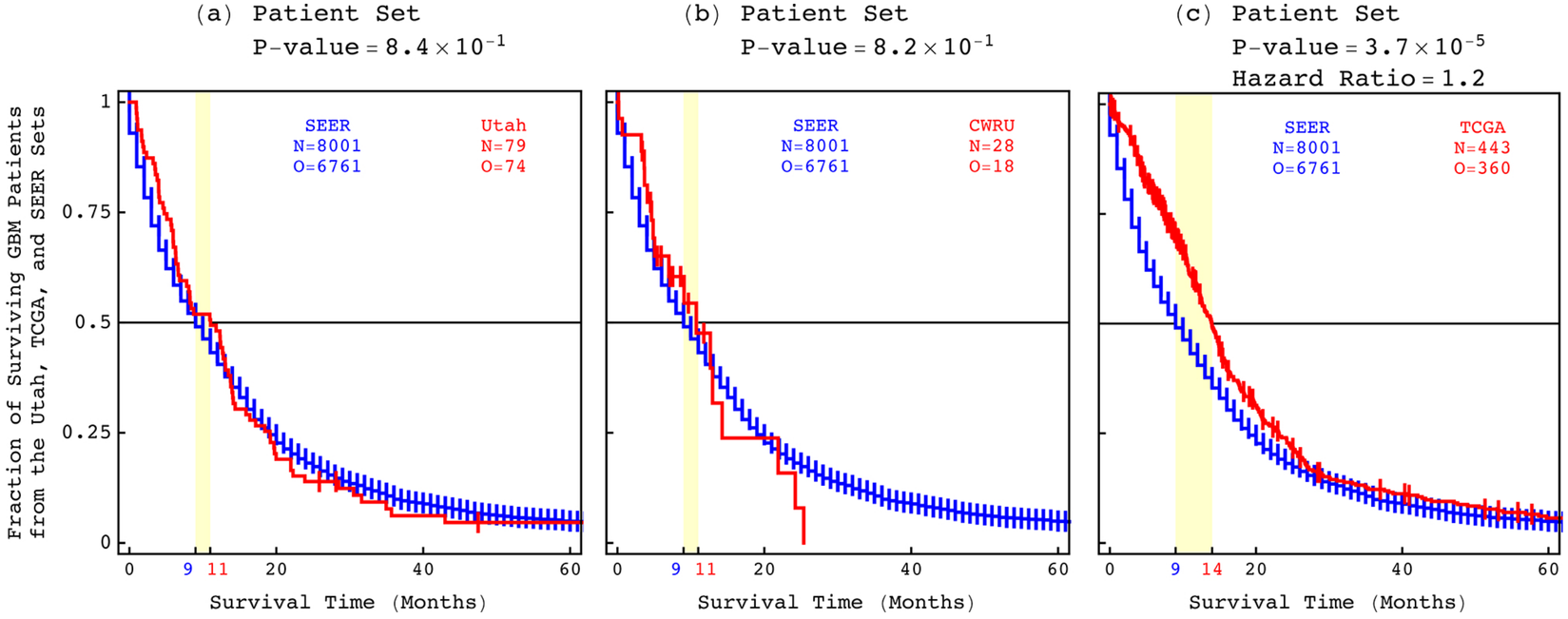


The normal male-specific X chromosome deletion is conserved in the tumors.

TCGA gender labels were corrected.

https://grants.nih.gov/grants/guide/notice-files/NOT-HG-11-021.html

The Utah set of 79 Patients is Statistically Representative of the U.S. Adult GBM Population



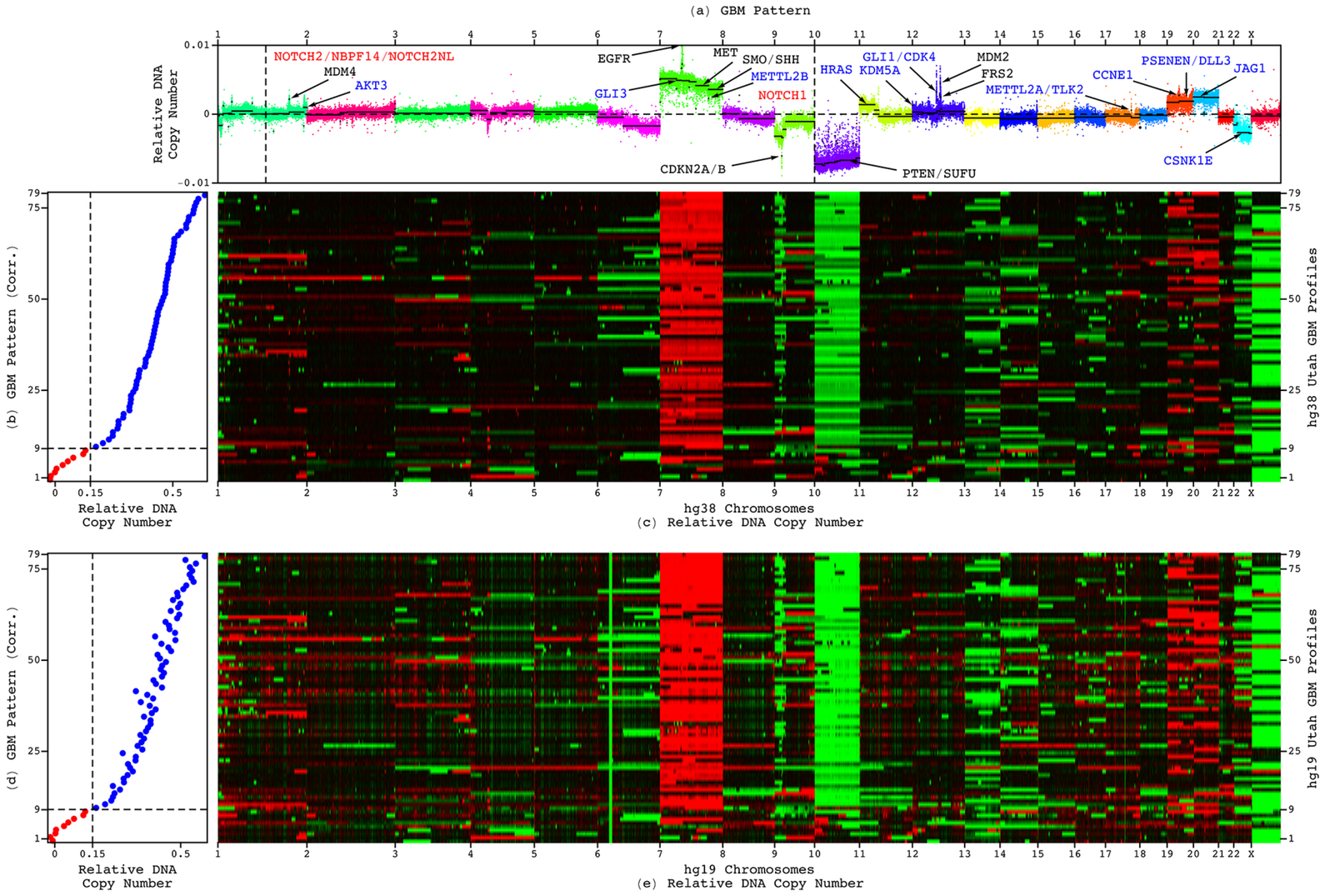
Phenotype		Group	Utah Set	CWRU Set	TCGA Set	SEER Set	Utah vs. SEER χ^2 P-Value	CWRU vs. SEER χ^2 P-Value	TCGA vs. SEER χ^2 P-Value
Normal	Sex	Female	27	12	169	3331	1.8×10^{-1}	9.0×10^{-1}	1.5×10^{-1}
		Male	52	16	274	4670			
	Race	Not White	5	3	40	959	1.2×10^{-1}	8.4×10^{-1}	1.1×10^{-1}
		White	74	25	386	7042			
	Ethnicity	Hispanic	2	0	11	507	1.7×10^{-1}	1.7×10^{-1}	4.4×10^{-3}
		Not Hispanic	77	28	382	7494			
Disease	Age (Years)	<50	10	3	117	1164	6.4×10^{-1}	5.7×10^{-1}	1.2×10^{-11}
		≥50	69	25	326	6837			

Profiling Technology and Reference Human Genome Affect <1% of the Classifications

Experimental batch effects normally reduce the reproducibility, i.e., precision, of classifications based upon between one to a few hundred genomic loci by >30%.

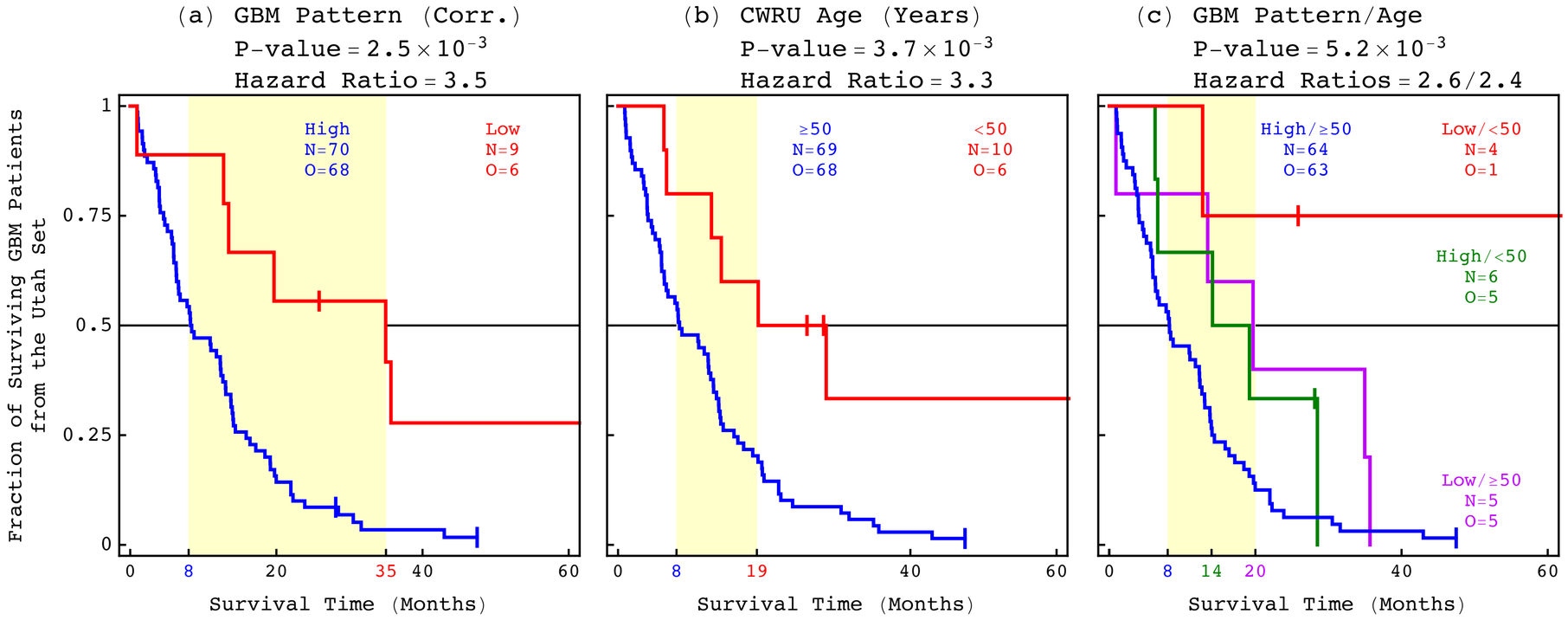
Pinto, et int., Feuk, *Nat Biotechnol* 29, 512 (2011).

Intratumor heterogeneity affects ≈11% of the classifications.



With a 2.25-year Kaplan–Meier (KM) median survival difference, a 3.5 univariate Cox hazard ratio, and a 0.78 concordance index, i.e., accuracy.

Statistically Better Than and Independent of the Best Other Indicator, i.e., Age



In general as well as in patients who receive treatments, i.e., chemotherapy and radiation.

Independent of chemotherapy and radiation and the post-surgical resection metrics, i.e., the Karnofsky performance score and the percent primary tumor resection.

Statistically Better Than and Independent of the Best Other Indicator, i.e., Age

Greater median survival differences, univariate hazard ratios and concordance indices, i.e., accuracies, and lower log-rank and Wald *P*-values as well as Akaike information criterion (AIC) values.

Bivariate hazard ratios within 95% confidence intervals of univariate ratios.

Patients (Number)	Predictor	KM Group	Median (Months)	Log-Rank *P*-Value	Cox Model	Hazard Ratio	95% Confidence Interval	Wald *P*-Value	Concordance Index
79	GBM Pattern (Corr.)	Low	35	2.5×10^{-3}	Univariate	3.5	1.5– 8.2	4.3×10^{-3}	0.78
		High	8						
	CWRU Age (Years)	<50	19	3.7×10^{-3}		3.3	1.4– 7.5	6.1×10^{-3}	0.78
		≥50	8						
	GBM Pattern (Corr.)	Low/<50		5.2×10^{-3}	Bivariate	2.6	1.1– 6.3	3.4×10^{-2}	0.76
	CWRU Age (Years)	Low/≥50	20			2.4	1.0– 5.7	5.0×10^{-2}	
		High/<50	14						
		High/≥50	8						
47	GBM Pattern (Corr.)	Low	36	4.8×10^{-3}	Univariate	4.1	1.4– 11.8	8.7×10^{-3}	0.86
		High	14						
	CWRU Age (Years)	<50	28	7.9×10^{-3}		3.7	1.3– 10.5	1.3×10^{-2}	0.81
		≥50	14						
59	GBM Pattern (Corr.)	Low	36	2.0×10^{-3}		4.6	1.6– 13.0	4.3×10^{-3}	0.88
		High	12						
	CWRU Age (Years)	<50	19	1.7×10^{-2}		2.7	1.2– 6.4	2.2×10^{-2}	0.72
		≥50	13						
75	GBM Pattern (Corr.)	Low	35	2.1×10^{-3}		4.0	1.6– 10.1	4.0×10^{-3}	0.79
		High	8						
	CWRU Chemotherapy	Yes	14	1.0×10^{-12}		6.0	3.5– 10.4	7.0×10^{-11}	0.93
		No	4						
	CWRU Radiation	Yes	13	4.4×10^{-15}		10.8	5.3– 22.1	7.6×10^{-11}	0.93
		No	2						
74/75	GBM Pattern (Corr.)	Low/Yes	36	1.5×10^{-12}	Bivariate	3.8	1.3– 10.8	1.4×10^{-2}	0.91
	CWRU Chemotherapy	High/Yes	14			5.0	2.9– 8.6	9.0×10^{-9}	
		High/No	4						
	GBM Pattern (Corr.)	Low/Yes	36	4.2×10^{-15}		4.6	1.6– 13.0	4.4×10^{-3}	0.92
	CWRU Radiation	High/Yes	12			9.3	4.5– 19.3	1.6×10^{-9}	
		High/No	3						
52	GBM Pattern (Corr.)	Low	35	8.1×10^{-3}	Univariate	3.2	1.3– 7.9	1.1×10^{-2}	0.74
		High	12						
	CWRU Karnofsky Score	≥60	16	7.3×10^{-7}		5.8	2.7– 12.4	8.0×10^{-6}	0.87
		<60	5						
51/52	GBM Pattern (Corr.)	Low/≥60	35	1.3×10^{-6}	Bivariate	3.5	1.3– 9.4	1.3×10^{-2}	0.83
	CWRU Karnofsky Score	High/≥60	14			4.5	2.0– 9.9	2.0×10^{-4}	
		High/<60	5						
28	GBM Pattern (Corr.)	Low	35	6.1×10^{-3}	Univariate	10.6	1.4– 80.9	2.3×10^{-2}	0.97
		High	6						
	CWRU Percent Resection	≥30	12	3.0×10^{-2}		2.8	1.1– 7.6	3.9×10^{-2}	0.73
		<30	5						
	GBM Pattern (Corr.)	Low/≥30	35	4.5×10^{-3}	Bivariate	9.8	1.3– 76.2	2.9×10^{-2}	0.80
	CWRU Percent Resection	High/≥30	8						
		High/<30	5						

Statistically Better Than and Independent of the Best Other Indicator, i.e., Age in the TCGA Set of 443 Patients

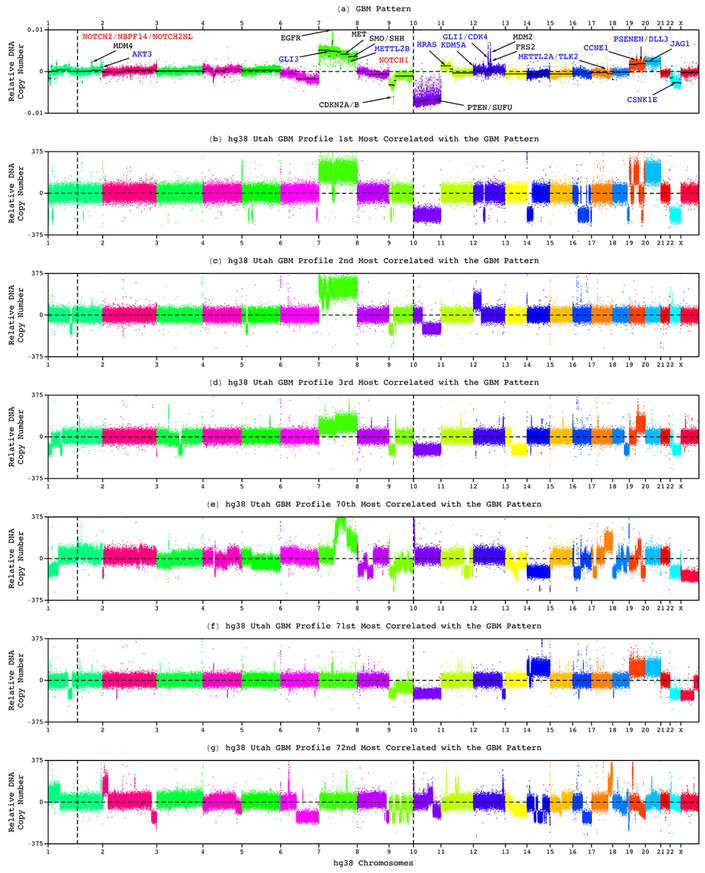
Better than and independent of the existing pathology laboratory tests, i.e., for *MGMT* promoter methylation and *IDH1* mutation, as well as better than *TERT* gene expression.

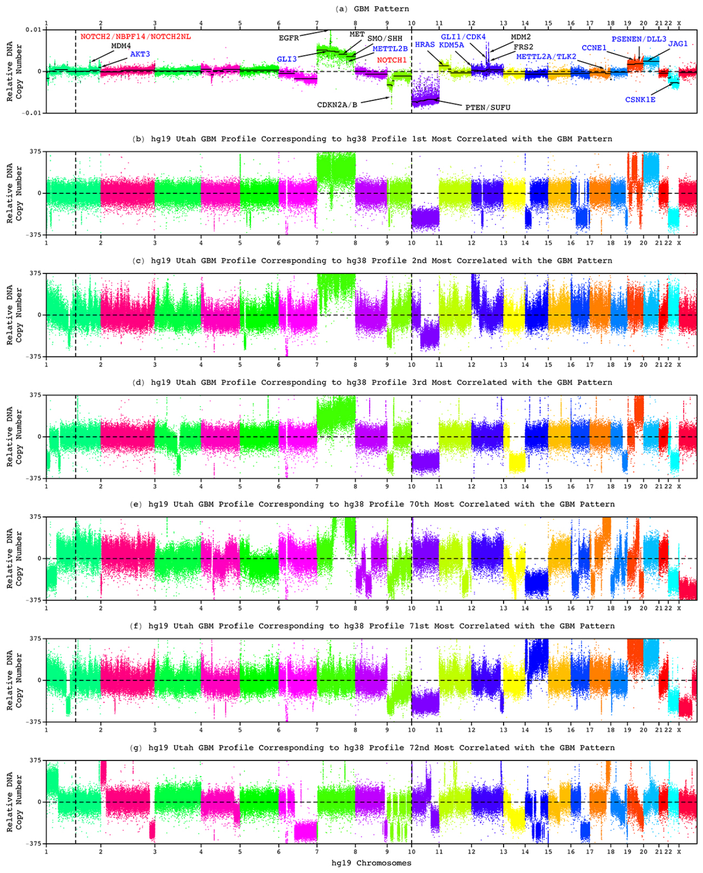
Before progressing to GBM standard of care, *MGMT*, *IDH1*, and *TERT*, have already been used as indicators of survival and *MGMT* also as an indicator of response to alkalyting agents in other types of cancer.

Hegi et al., *NEJM* 352, 997 (2005); Hotta, et int., Ikenaga, *J Neurooncol* 21, 135 (1994); Parsons, et int., Kinzler, *Science* 321, 1807 (2008); Nguyen, et int., Lai, *Neuro Oncol* 19, 394 (2017); Batchelor & Louis, in *UpToDate*, eds. Loeffler & Wen. Waltham, MA: Wolters Kluwer (2018).

Patients (Number)	Predictor	KM Group	Median (Months)	Log-Rank *P*-Value	Cox Model	Hazard Ratio	95% Confidence Interval	Wald *P*-Value	Concordance Index
443	GBM Pattern (Corr.)	Low	29	1.4×10^{-6}	Univariate	2.5	1.7– 3.6	2.9×10^{-6}	0.75
		High	13						
	TCGA Age (Years)	<50	21	4.9×10^{-10}		2.2	1.7– 2.8	1.1×10^{-9}	0.72
		≥50	12						
	GBM Pattern (Corr.)	Low/<50	39	2.1×10^{-10}	Bivariate	1.9	1.3– 2.8	1.7×10^{-3}	0.71
	TCGA Age (Years)	High/<50	18			1.9	1.4– 2.4	1.9×10^{-6}	
		Low/≥50	16						
		High/≥50	12						
284	GBM Pattern (Corr.)	Low	50	3.2×10^{-4}	Univariate	2.5	1.5– 4.1	4.9×10^{-4}	0.73
		High	15						
	TCGA Age (Years)	<50	22	3.3×10^{-5}		1.9	1.4– 2.6	4.3×10^{-5}	0.64
		≥50	15						
	GBM Pattern (Corr.)	Low/<50	50	5.5×10^{-5}	Bivariate	1.8	1.1– 3.2	2.9×10^{-2}	0.64
	TCGA Age (Years)	High/<50	18			1.6	1.1– 2.2	6.8×10^{-3}	
		High/≥50	15						
		Low/≥50	6						
327	GBM Pattern (Corr.)	Low	39	2.8×10^{-5}	Univariate	2.4	1.6– 3.8	4.8×10^{-5}	0.73
		High	15						
	TCGA Age (Years)	<50	23	6.7×10^{-6}		1.9	1.4– 2.5	8.9×10^{-6}	0.66
		≥50	15						
	GBM Pattern (Corr.)	Low/<50	47	2.7×10^{-6}	Bivariate	2.0	1.3– 3.1	3.0×10^{-3}	0.65
	TCGA Age (Years)	High/<50	20			1.6	1.2– 2.1	1.7×10^{-3}	
		Low/≥50	16						
		High/≥50	15						
255	GBM Pattern (Corr.)	Low	34	3.2×10^{-4}	Univariate	2.5	1.5– 4.2	5.0×10^{-4}	0.73
		High	13						
	TCGA *MGMT* Methylation	Yes	18	4.7×10^{-3}		1.5	1.1– 2.0	5.0×10^{-3}	0.58
		No	13						
	GBM Pattern (Corr.)	Low/Yes	39	3.0×10^{-4}	Bivariate	2.4	1.4– 4.0	1.1×10^{-3}	0.61
	TCGA *MGMT* Methylation	Low/No	16			1.4	1.1– 1.9	1.7×10^{-2}	
		High/Yes	16						
		High/No	12						
329	GBM Pattern (Corr.)	Low	34	5.5×10^{-5}	Univariate	2.4	1.6– 3.8	9.2×10^{-5}	0.77
		High	14						
	TCGA *IDH1* Mutation	Yes	35	2.6×10^{-4}		2.4	1.5– 3.9	4.0×10^{-4}	0.78
		No	14						
	GBM Pattern (Corr.)	Low/Yes	39	9.6×10^{-5}	Bivariate	1.9	1.2– 3.1	8.4×10^{-3}	0.77
	TCGA *IDH1* Mutation	Low/No	25			1.7	1.0– 3.0	4.1×10^{-2}	
		High/Yes	24						
		High/No	14						
107	GBM Pattern (Corr.)	Low	34	1.8×10^{-2}	Univariate	3.7	1.2– 12.1	2.7×10^{-2}	0.90
		High	13						
	TCGA *TERT* Expression	No	25	9.5×10^{-3}		2.5	1.2– 5.1	1.2×10^{-2}	0.77
		Yes	13						
	GBM Pattern (Corr.)	Low/No	34	2.4×10^{-2}	Bivariate				0.80
	TCGA *TERT* Expression	Low/Yes	25						
		High/No	24						
		High/Yes	12						
335	GBM Pattern (Corr.)	Low	34	3.0×10^{-5}	Univariate	2.4	1.6– 3.6	4.9×10^{-5}	0.72
		High	14						
	TCGA Karnofsky Score	≥60	15	1.0×10^{-5}		4.1	2.1– 8.1	4.7×10^{-5}	0.81
		<60	5						
	GBM Pattern (Corr.)	Low/≥60	34	1.6×10^{-8}	Bivariate	2.3	1.5– 3.5	6.9×10^{-5}	0.74
	TCGA Karnofsky Score	High/≥60	15			3.8	1.9– 7.6	1.1×10^{-4}	
		High/<60	5						

A Patient's Survival is the Outcome of Their Tumor's Whole Genome, i.e., Genetic Background





Chromosome 10 deletion, chromosome 7 amplification, and chromosome arm 9p deletion, appear in the tumor genomes of some but not all 70 patients with high and, separately, some but not all nine patients with low correlations of their tumor profiles with the pattern.

Personalized Prognostics, Diagnostics, and Therapeutics with a Genome-Wide Predictor of Survival

This is the first predictor that encompasses the whole tumor genome.

→ The prognostic classification can help manage, e.g., GBM pseudoprogression: **validate Clinical Laboratory Improvement Amendments (CLIA) -certified and College of American Pathologists (CAP) -accredited, i.e., technical clinical WGS**.
Patel, et int., Ellingson, *J Neurooncol* 139, 399 (2018); Walter, et int., Czernin, *J Nucl Med* 53, 393 (2012); Rehm, et int., Lyon, *Genet Med* 15, 733 (2013).

→ The diagnostic classification could help therapies progress to regulatory approval. Even if a drug targets just one gene, the patient's response depends on the whole genome. Only one new drug has advanced from trials to care in 40 years: **Predict long survivors in a personalized vaccine trial that are not explained by other factors**.
Threadgill, et int., Magnuson, *Science* 269, 230 (1995); Rich, et int., Friedman, *J Clin Oncol* 22, 133 (2004); Grossman & Ellsworth, *J Clin Oncol* 34, e13522 (2016); Liau, et int., Bosch, *J Transl Med* 16, 142 (2018).

→ The therapeutic predictions, of previously unrecognized targets that are correlated with survival, e.g., the druggable *METLL2A/B* and *TLK2*, could lead to new drugs: **Preclinically test in GBM**.
Kim, et int., Wang, *Mol Cancer Res* 14, 920 (2016); Zhang, Guo & Boulianne, *Gene* 280, 135 (2001); Hayden Gephart, et int., Scott, *J Neurooncol* 115, 61 (2013).

Proof of Principle that the Multi-Tensor Decompositions are Uniquely Suited for Discovering Accurate, Precise, and Actionable Genotype-Phenotype Relationships Relevant to the Population Based upon Small Cohorts

They have overcome three distinct challenges that other methods had not.

- → **They found consistent patterns across whole genomes, which have 3B nucleotides.** Minimally preprocessed datasets with no feature engineering account for robustness to perturbations and are possible because of the computational scalability.
- → **They did that across the tumor and the matching normal genomes simultaneously.** By using the complex structure of the datasets rather than simplifying or standardizing them as is commonly done, they can separate patterns which occur only in the tumor genomes from those that occur in the genomes of normal cells in the body and variations caused by experimental inconsistencies.
- → **They did so in small cohorts of patients, ~100, that are typical in clinical trials.** The structure of the datasets accounts for sensitivity to relationships in small discovery sets and imbalanced validation sets of large genomic profiles, and is possible because of the mathematical formulation as frameworks of blind source separation (BSS).

High-Performance Computing

Protected cloud environments facilitate this work by providing on-demand access to storage and compute.

→ This enables implementing numerical analysis algorithms at scale on massive and dense datasets. The size of the astrocytoma WGS raw binary alignment map (BAM) files, e.g., is >20T bytes.

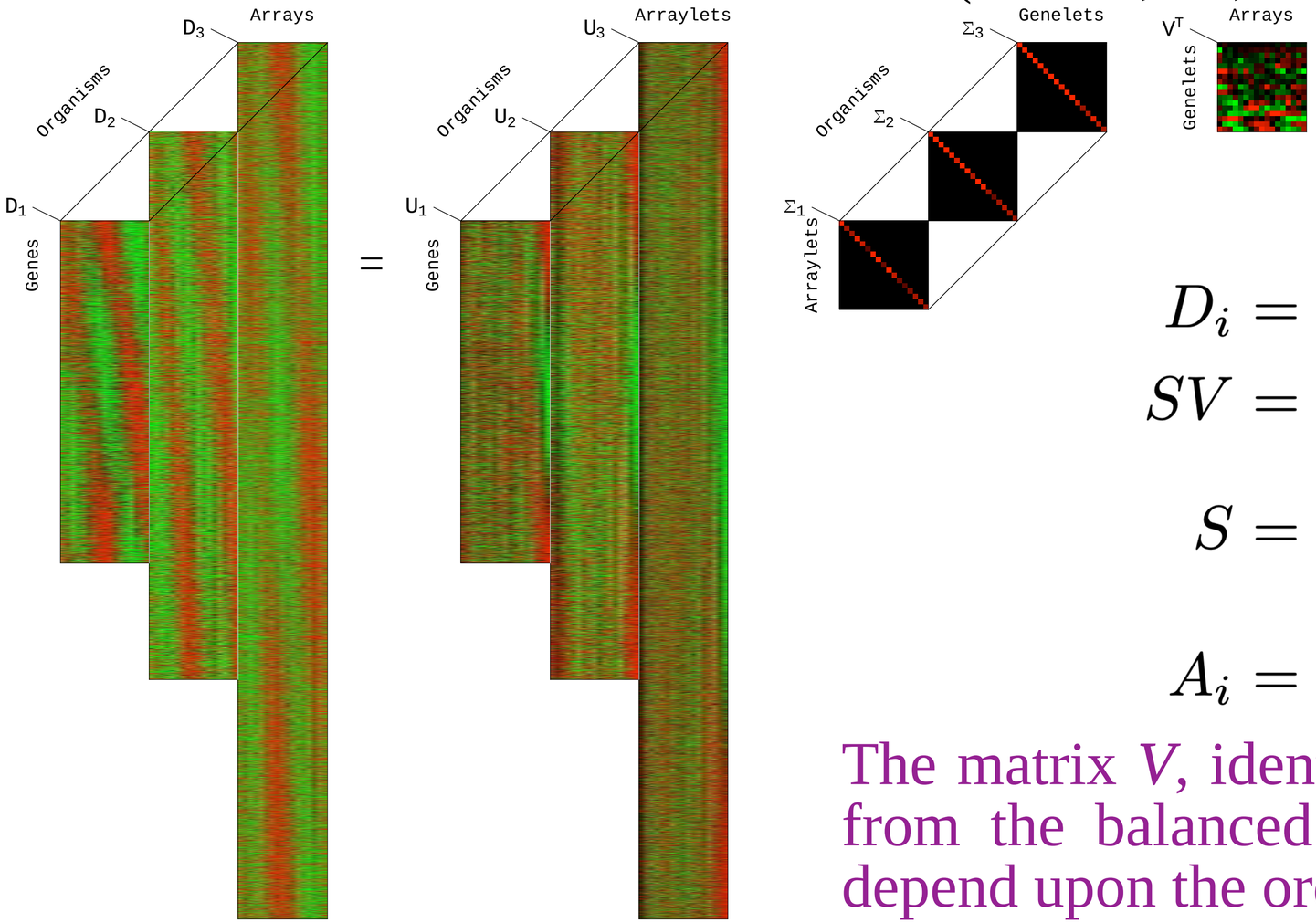
→ This enables testing the robustness of the results to hundreds or more perturbations to the datasets, e.g., due to changes in the preprocessing of the BAM files. In generating the astrocytoma read-count profiles, the changes to the preprocessing of the raw BAM files included varying the bin sizes in the range of 100–2.5K nucleotides.

To generate the profiles at the 3B nucleotide-level resolution, a 5T-byte random-access memory (RAM) may be needed.

→ This in turn enables us to prognostically and diagnostically classify, e.g., GBM patients based upon the genome-wide pattern of DNA CNAs with >99% reproducibility, i.e., precision, among profiling technologies, and with ≥75% concordance, i.e., accuracy, and therapeutically predict previously unrecognized GBM drug targets.

Higher-Order GSVD for Comparative Analysis of Multiple Second-Order Datasets

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



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

$$SV = V\Lambda,$$

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

$$A_i = D_i^T D_i, \quad i = 1, 2, \ldots, N.$$

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.

Higher-Order GSVD for Comparative Analysis of Multiple Second-Order Datasets

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

The exact HO GSVD directly extends to multiple matrices all mathematical properties of the GSVD except for complete orthogonality of U_i for all i.

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

Theorem 1: S has n independent eigenvectors, and its eigenvalues 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: $\lambda_k=1$ 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.

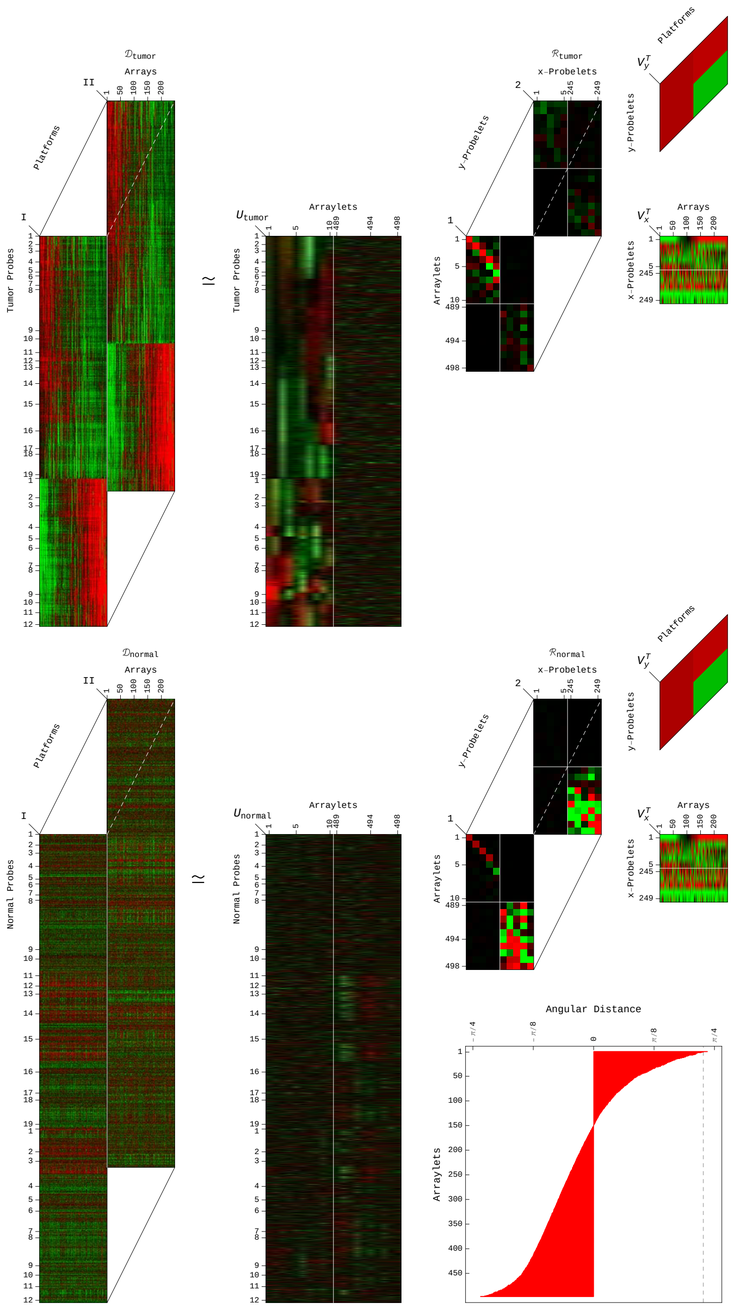
Mathematically Universal
Biologically Consistent
Global Genotype-Phenotype

Bradley, Aiello, Ponnapalli,* Hanson* & Alter, *APL Bioeng* 3, 036104 (2019); https://alterlab.org/adenocarcinomas_genotype-phenotype/

New tumors, e.g., metastasis, are the leading cause of death from lung, uterine, and ovarian adenocarcinomas, where most patients experience progression-free survival after the primary treatment.

Yet, no indicator existed that predicts **the benefit of platinum in terms of overall survival** past the primary treatment.

6p+12p primary tumor's genotypes predictive of the patient's overall survival phenotypes, **in general as well as following platinum treatment of the primary tumor, and throughout the course of the disease**, were discovered by the GSVD and tensor GSVD.



Tensor HO GSVD for Comparative Analysis of Multiple Higher-Order Datasets

Ponnapalli & Alter (in preparation);
Ponnapalli, Saunders, Van Loan & Alter, *PLoS One* 6, e28072 (2011).

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

$$D_i = (\ldots, \mathcal{D}_{i,:lm}, \ldots)$$

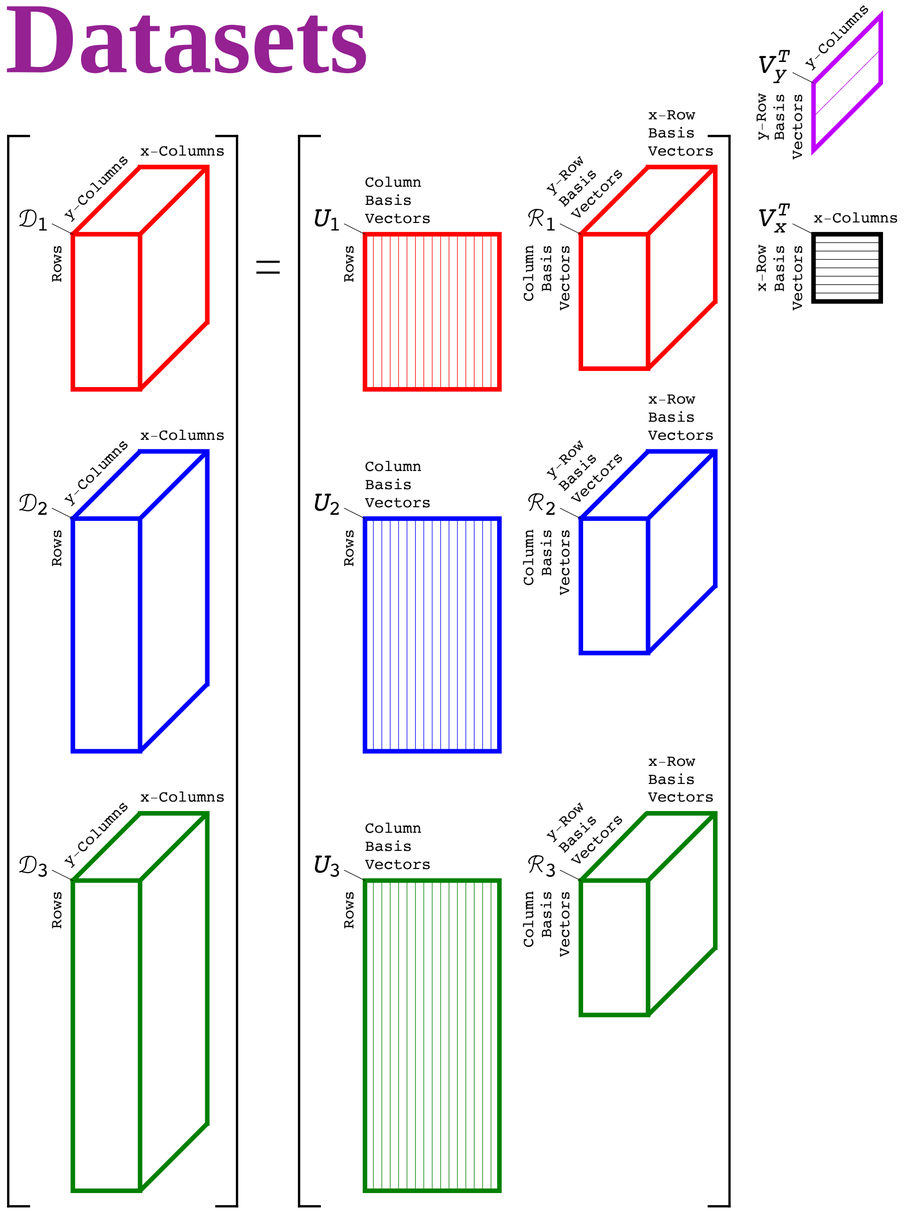
$$= U_i \Sigma_i V^T,$$

$$\mathcal{R}_i = \mathcal{D}_i \times_a U_i^\dagger \times_b V_x^{-T} \times_c V_y^{-T},$$

$$U_i U_i^\dagger = D_i D_i^\dagger \neq I,$$

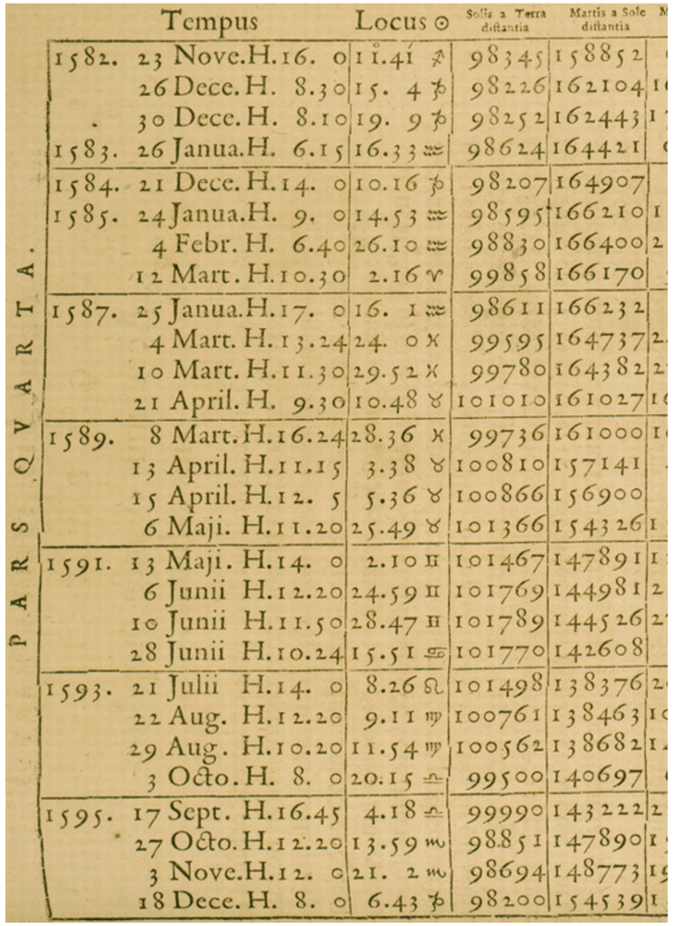
$$U_i U_i^\dagger D_i = D_i D_i^\dagger D_i = D_i,$$

$$i = 1, 2, \ldots, N.$$

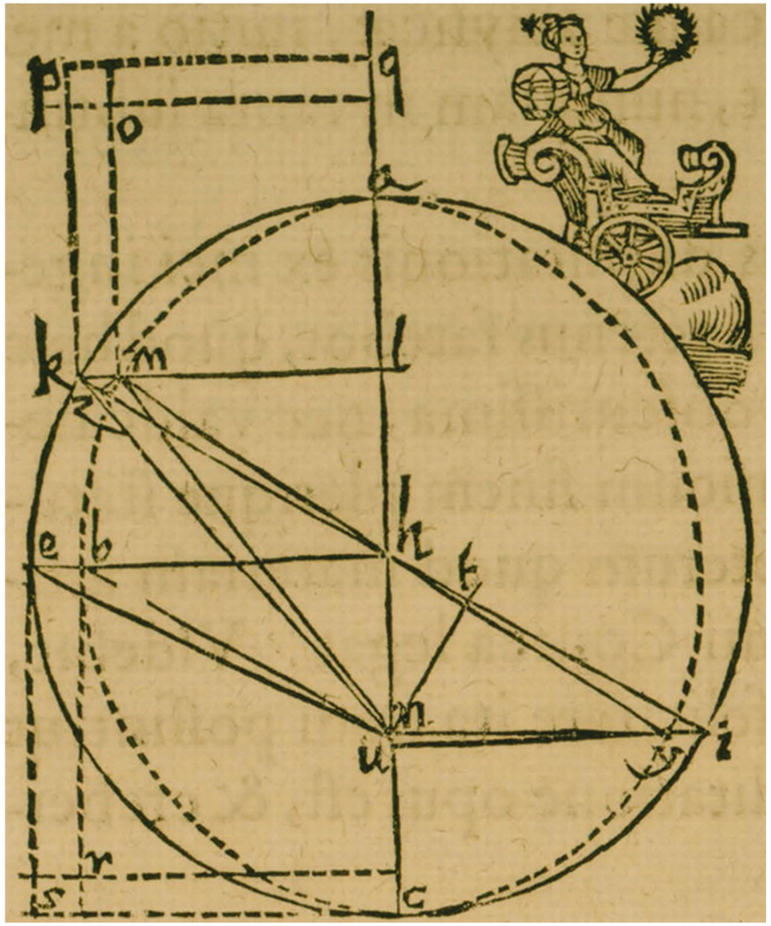


Correlations to Causal Coordination:
Global Patterns Underlie Principles of Nature

Alter, *PNAS* 103, 16063 (2006);
Alter, in *Microarray Data Analysis: Methods and Applications*. New York, NY: Humana Press (2007).



	Tempus	Locus ☉	Solis a Terra distantia	Martis a Sole distantia
PARS QVARTA.				
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	154539



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

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

Statistics to Processes:
Global Patterns Underlie Principles of Nature

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

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

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

→ Brownian motion.
Einstein, *Ann Phys* 17, 549 (1905).

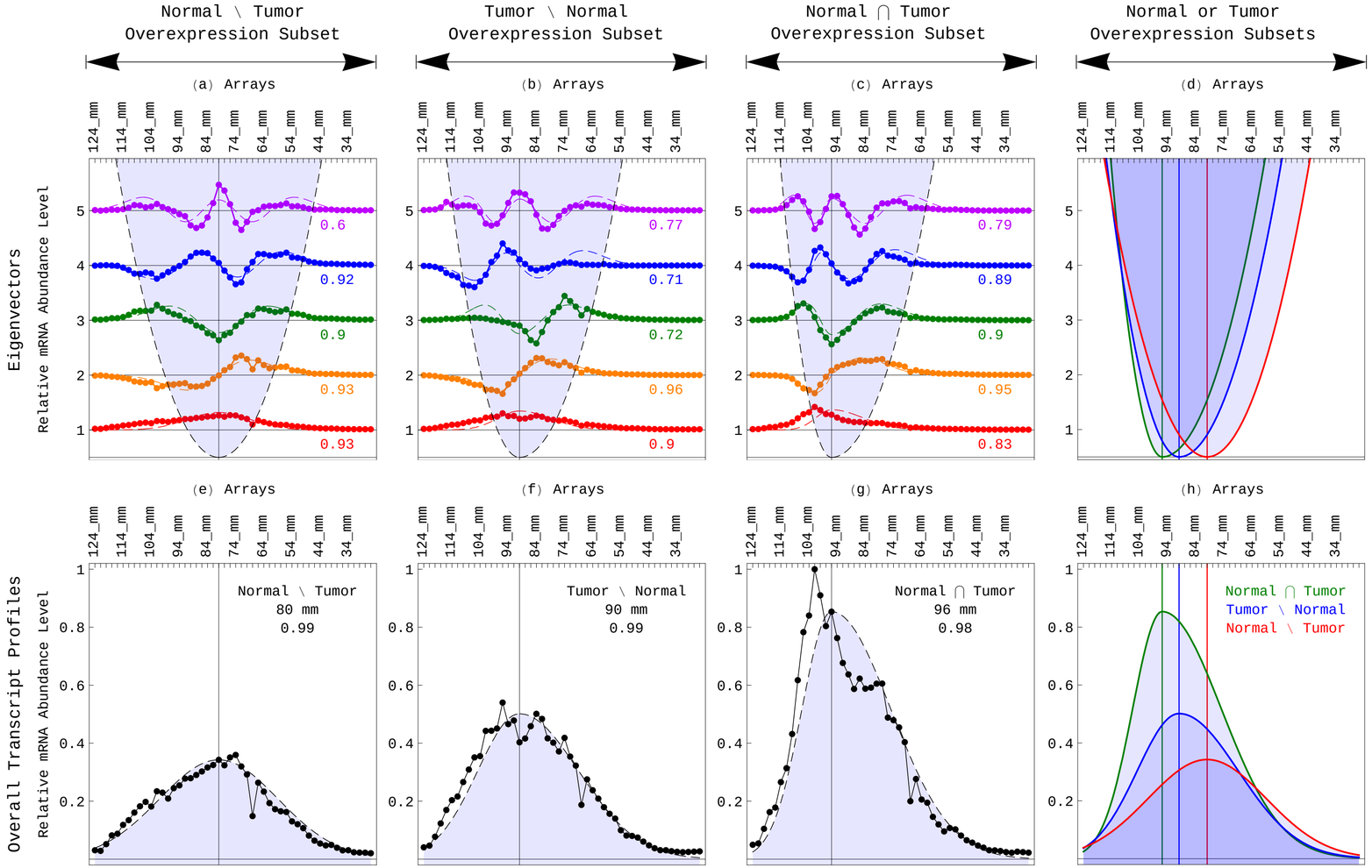
→ Bacterial sensitivity and resistance to viruses.
Luria & Delbrück, *Genetics* 28, 491 (1943).

Computationally Predicted Physical Mechanism for Cells to Differentially Regulate Metabolism in a Transcript Length-Dependent Manner

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

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.

Oh et al., *Nat Struct Mol Biol* 24, 993 (2017).

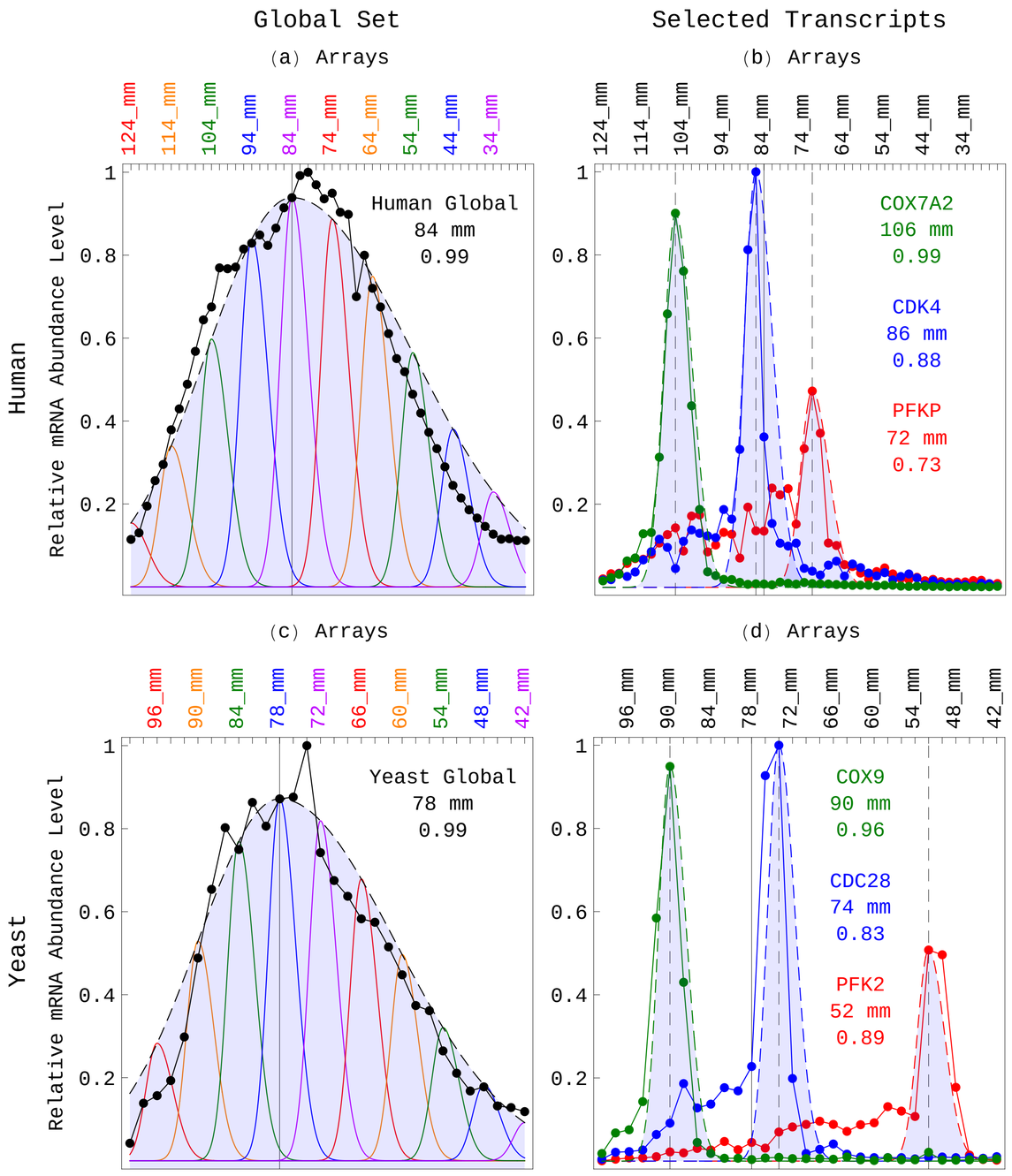


Computationally Predicted Physical Mechanisms for Asymmetric Broadening of a Moving Transcript Band in Gel Electrophoresis and Asymmetric Evolutionary Restoring-Like Forces Acting on Transcript Length

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

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

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



Multi-Tensor Decompositions for Personalized Medicine

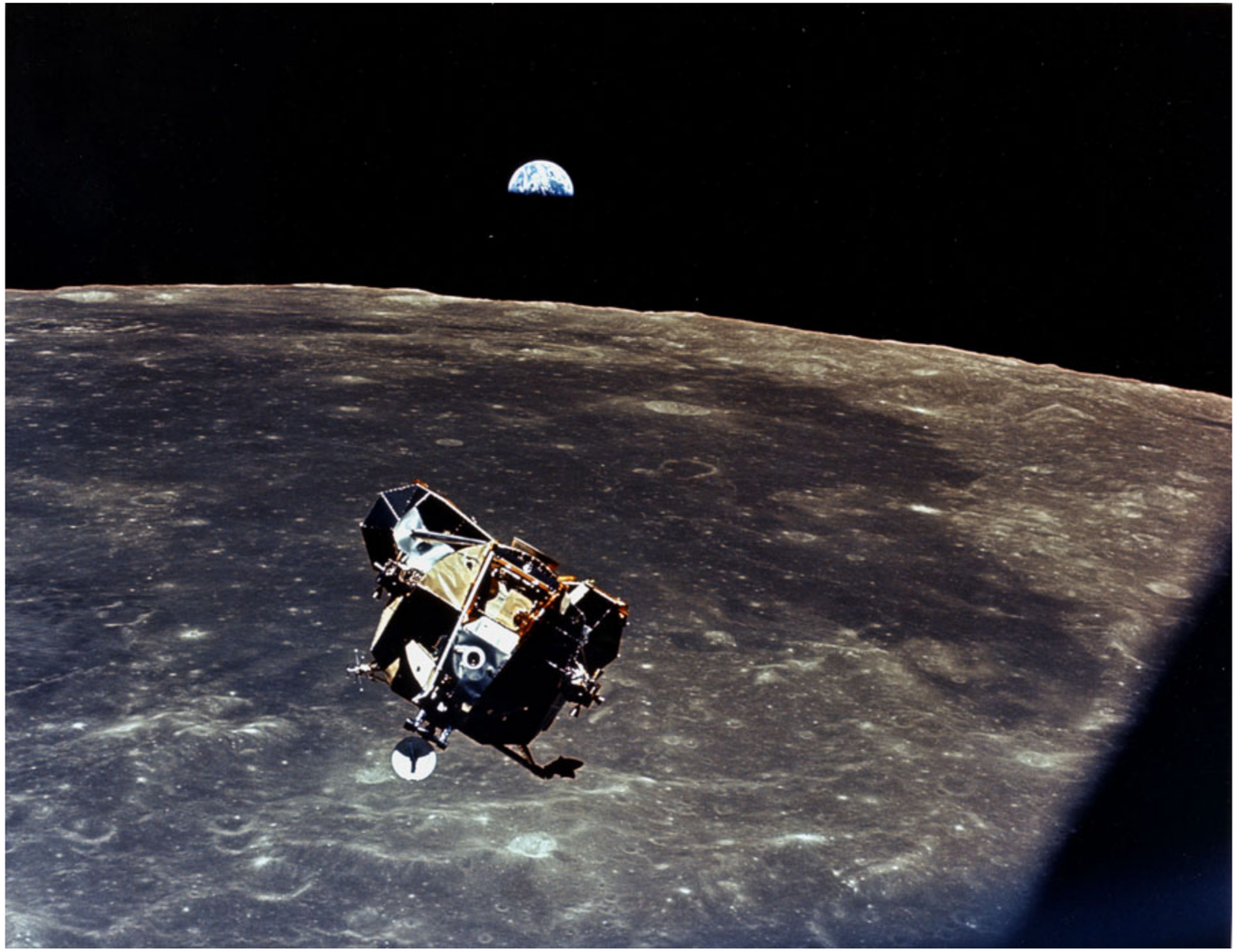
? Scale up mathematics.
Edelman & Wang, *SIMAX* 41, 1826 (2020).

? Scale up modeling.
Palacios-Flores, et int., Palacios, *PNAS* 118, e2025192118 (2021).

? Scale up experiments.
Durbin, et int., Qi, *AACR Special Conference on Advances in Pediatric Cancer Research* (Montreal, Canada, September 17–20, 2019). Philadelphia, PA: AACR (2020).

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

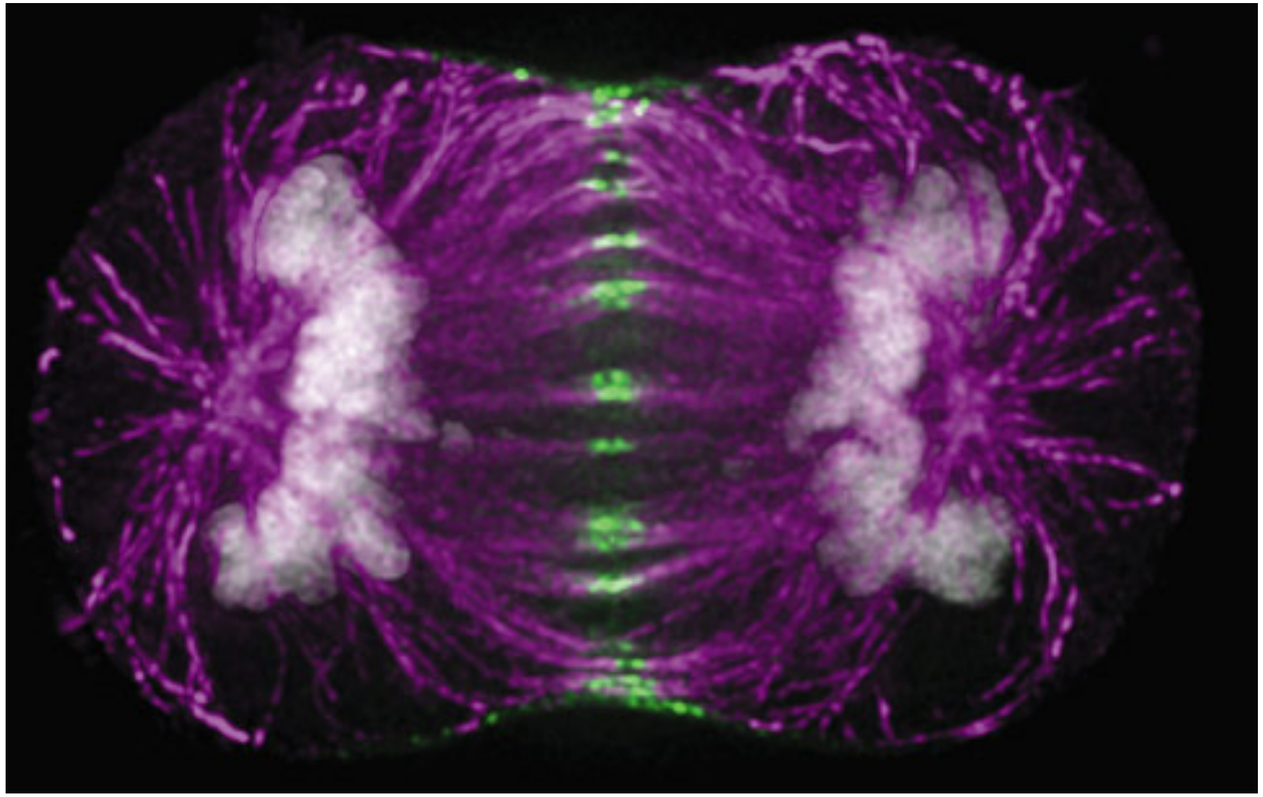
Wigner, *Commun Pure Appl Math* 13, 1 (1960).



Collins, *NASA Apollo 11* (1969).

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 mechanisms that govern the activity of DNA and RNA.

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:

Jill S. Barnholtz-Sloan
Population Health Sciences, CWRU

Jay Bowen & Julie M. Gastier-Foster
Laboratory Medicine, NCH

Randy L. Jensen
Neurosurgery, Utah

Cheryl A. Palmer
Pathology, Utah

Carl T. Wittwer
Pathology, Utah

Huanming Yang
Medical Genetics, BGI

John F. X. Diffley
Cancer Research UK, London

Gene H. Golub
Computer Science, Stanford

Michael A. Saunders
Operations Research, Stanford

Charles F. Van Loan
Computer Science, Cornell

Yoshihisa Yamamoto
Applied Physics, Stanford

Senior Research Affiliate & Ph.D. Alumna:

Sri Priya Ponnapalli, ECE

K99 Postdoctoral Alumnus:

Jason M. Tennessen, Genetics

Ph.D. Alumni:

Larsson Omberg, Physics
Kayta Kobayashi, Pharmacy
Chaitanya Muralidhara, CMB

B.Sc. Alumni:

Joel R. Meyerson, BME & Gov
Andrew M. Gross, BME & SSC
Justin A. Drake, BME & SSC
Nicolas M. Bertagnolli, Mathematics

Support:

CDC 75D301 21C11016
NCI U01 CA-202144
https://physics.cancer.gov/network/UniversityofUtah.aspx
NSF CAREER DMS-0847173
NHGRI K01 HG-000038 & R01 HG-004302
DOE FG03 99ER62836

Thank you!

Physics-Inspired Multi-Tensor Decompositions

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

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

Directly 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., PageRank and the Netflix challenge.

Physics-Inspired Multi-Tensor Decompositions

Find what other methods miss, and outperform methods that:

→ require large amounts of training data (e.g., deep learning);

→ require training and are sensitive to imbalanced class representations (e.g., supervised learning);

→ require data quantization and are sensitive to cutoff selections (e.g., Bayesian statistics and topological data analysis);

→ vary the one-dataset SVD and are, therefore, not exact or unique, rather than use the complex structure of the data (e.g., independent component analysis, sparse and nonnegative factorizations, and randomized decompositions);

→ are unsupervised but require data cleaning and are sensitive to artifacts and batch effects (e.g., hierarchical clustering);

→ are supervised and require a-priori knowledge (e.g., analysis of variance).

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

The SVD is used for the stable computation of principal component analysis (PCA).
Ringnér, *Nat Biotechnol* 26, 303 (2008).

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).

Cadima & Jolliffe, *Pak J Statist* 25, 473 (2009);
Muralidhara, Gross, Gutell & Alter, *PLoS One* 6, e18768 (2011); https://alterlab.org/rRNA/

→ 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 et al., in *Microarrays: Optical Technologies and Informatics*. Bellingham, WA: International Society for Optics and Photonics (SPIE) (2001); https://alterlab.org/SVD/
Fellenberg, et int., 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**.

There are nontrivial connections between the GSVD and canonical correlations analysis (CCA).

The GSVD is Different Than CCA

Ewerbring & Luk, *J Comput Appl Math* 27, 37 (1989).

De Clercq et al., *IEEE Trans Biomed Eng* 53, 2583 (2006);
Rustandi, Just & Mitchell, in *MICCAI 2009* (London, UK, September 20–24, 2009);
Tenenhaus & Tenenhaus, *Psychometrika* 76, 257 (2011).