

(51) International Patent Classification:  
*G06F 19/10* (2011.01)(21) International Application Number:  
PCT/US2012/054315(22) International Filing Date:  
7 September 2012 (07.09.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
61/533,141 9 September 2011 (09.09.2011) US  
61/553,840 31 October 2011 (31.10.2011) US(71) Applicant (for all designated States except US): **UNIVERSITY OF UTAH RESEARCH FOUNDATION**  
[US/US]; 615 Arapleen Drive, Suite 310, Salt Lake City, UT 84108 (US).

(72) Inventor; and

(75) Inventor/Applicant (for US only): **ALTER, Orly**  
[US/US]; 615 Arapleen Drive, Suite 310, Salt Lake City, UT 84108 (US).(74) Agent: **HILL, James, William**; McDermott Will & Emery  
LLP, 4 Park Plaza, Suite 1700, Irvine, CA 92614-2559 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: GENOMIC TENSOR ANALYSIS FOR MEDICAL ASSESSMENT AND PREDICTION

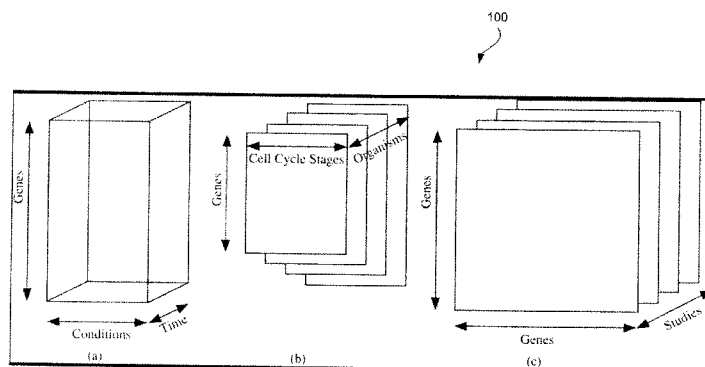


Figure 1

(57) Abstract: Systems and methods are described for medical characterization of biological data. One such method includes applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ ; where the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and determining an indicator of a health parameter of a subject, the determining being based on the eigenvectors and on values, associated with the subject, of the at least two index parameters. In some cases, the eigenvectors of  $A^TA$  are the same as the eigenvectors of  $B^TB$ .

## **GENOMIC TENSOR ANALYSIS FOR MEDICAL ASSESSMENT AND PREDICTION**

### **Cross Reference to Related Applications**

[001] This application claims the benefit of U.S. Provisional Application No. 61/533,141, filed September 9, 2011, and U.S. Provisional Application No. 61/553,840, filed October 31, 2011, each of the foregoing applications is incorporated by reference in its entirety.

### **Government License Rights**

[002] This invention was made with government support under the Utah Science Technology and Research (USTAR) Initiative and the National Human Genome Research Institute (NHGRI) R01 Grant HG-004302. The Government has certain rights to this invention.

### **Technical Field**

[003] The subject technology relates generally to computational medicine and computational biology.

### **Background**

[004] In many areas of science, especially in biotechnology, the number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing. This increase is accompanied by a fundamental need for mathematical frameworks that can compare multiple large-scale matrices with different row dimensions. Some of these areas may involve disease prediction based on biological data related to patient and normal samples.

[005] For example, glioblastoma multiforme (GBM), the most common malignant brain tumor in adults, is characterized by poor prognosis. GBM tumors may exhibit a range of copy-number alterations (CNAs), many of which play roles in the cancer's pathogenesis. Large-scale gene expression and DNA methylation profiling efforts have identified GBM molecular subtypes, distinguished by small numbers of biomarkers. However, the best prognostic predictor for GBM remains the patient's age at diagnosis.

[006] Therefore, there is a need for a more effective method for disease related characterization of biological data. The subject technology provides such characterization.

### **Brief Description of the Drawings**

[007] Figure 1 is a high-level diagram illustrating examples of tensors including biological datasets, according to some embodiments.

[008] Figure 2 is a high-level diagram illustrating a linear transformation of three-dimensional arrays, according to some embodiments.

[009] Figure 3 is a block diagram illustrating a biological data characterization system coupled to a database, according to some embodiments.

[0010] Figure 4 is a flowchart of a method for disease related characterization of biological data, according to some embodiments.

[0011] Figures 5A-5C are diagrams illustrating survival analyses of patients classified GBM-associated chromosome (10, 7, 9p) number changes, according to some embodiments. X-axis: survival time (months); Y-axis: fraction of surviving patients from the initial site. Figure 5A, line 1: No CNA, N=145, O=110; line 2: Loss, N=102, O=85. Figure 5B, line 1: No CNA, N=197, O=167; line 2: Gain, N=50, O=36. Figure 5C, line 1: No CNA, N=219, O=178; line 2: Loss, N=28, O=25.

[0012] Figure 6 is a diagram illustrating genes that are found in chromosomal segment 17:57,851,812-17:57,973,757 of the human genome, according to some embodiments.

[0013] Figure 7 is a diagram illustrating gene that is found in chromosomal segment 7:127,892,509-7:127,947,649 of the human genome, according to some embodiments.

[0014] Figure 8 is a diagram illustrating genes that are found in chromosomal segment 12:33,854-12:264,310 of the human genome, according to some embodiments.

[0015] Figure 9 is a diagram illustrating genes that are found in chromosomal segment 19:33,329,393-19:35-322,055 of the human genome, according to some embodiments.

[0016] Figure 10 is a diagram illustrating survival analyses of an initial set of a number of patients classified by chemotherapy or GSVD and chemotherapy, according to some embodiments. X-axis (all graphs): survival time (months); Y-axis, graphs (a) & (b): Fraction of surviving patients from the initial set; Y-axis, graphs (c) & (d): Fraction of surviving patients from the inclusive confirmation set; Y-axis, graphs (e) & (f): Fraction of surviving patients from the independent validation set. (a) line 1: No, N=49, O=46; line 2: Yes, N=187, O=147. (b) line 1: High/No, N=45, O=42; line 2: High/Yes, N=169, O=135; line 3: Low/No, N=4, O=4; line 4: Low/Yes, N=18, O=12. (c) line 1: No, N=62, O=57; line 2: Yes, N=255, O=188. (d) line 1: High/No, N=58, O=53; line 2: High/Yes, N=233, O=176; line 3: Low/No, N=4, O=4; line 4: Low/Yes, N=22, O=12. (e) line 1: No, N=24, O=22; line 2: Yes, N=130, O=103. (f) line 1: High/No, N=22, O=20; line 2: High/Yes, N=115, O=93; line 3: Low/No, N=2, O=2; line 4: Low/Yes, N=15, O=10.

[0017] Figure 11 is a diagram illustrating a high-order generalized singular value decomposition (HO GSVD) of biological data, according to some embodiments.

[0018] Figure 12 is a diagram illustrating a right basis vector of Figure 4 and mRNA expression oscillations in three organisms, according to some embodiments.

[0019] Figure 13 is a diagram illustrating an HO GSVD reconstruction and classification of a number of mRNA expressions, according to some embodiments.

[0020] Figure 14 is a diagram illustrating simultaneous HO GSVD sequence-independent classification of a number of genes, according to some embodiments.

[0021] Figure 15 is a diagram illustrating simultaneous correlations among the n=17 arraylets in one organism, according to some embodiments.

[0022] Figure 16 is a diagram illustrating three dimensional least squares approximation of the five-dimensional approximately common HO GSVD subspace, according to some embodiments;

[0023] Figure 17 is a diagram illustrating an example of *S. pombe* global mRNA expression reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments.

[0024] Figure 18 is a diagram illustrating an example of *S. cerevisiae* global mRNA expression reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments.

[0025] Figure 19 is a diagram illustrating a human global mRNA expression reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments.

[0026] Figure 20 is a diagram illustrating significant probelets and corresponding tumor and normal arraylets uncovered by GSVD of the patient-matched GBM and normal blood aCGH profiles, according to some embodiments.

[0027] Figure 21 is a diagram illustrating survival analyses of three sets of patients classified by GSVD, age at diagnosis or both, according to some embodiments. X-axis (all graphs): survival time (months); Y-axis, graphs (a)-(c): Fraction of surviving patients from the initial set; Y-axis, graphs (d)-(f): Fraction of surviving patients from the inclusive confirmation set; Y-axis, graphs (g)-(i): Fraction of surviving patients from the independent validation set. (a) line 1: High, N=224, O=186; line 2: Low, N=23, O=17. (b) line 1: >50, N=190, O=155; line 2: <50, N=57, O=48. (c) line 1: High/>50, N=183, O=151; line 2: low/<50, N=16, O=13; line 3: High/<50, N=41, O=35; line 4: Low/>50, N=7, O=4. (d) line 1: High, N=307, O=242; line 2: Low, N=27, O=17. (e) line 1: >50, N=254, O=200; line 2: <50, N=80, O=59. (f) line 1: High/>50, N=246, O=195; line 2: low/<50, N=19, O=12; line 3: High/<50, N=61, O=47; line 4: Low/>50, N=8, O=5. (g) line 1: High, N=162, O=136; line 2: Low, N=21, O=14. (h) line 1: >50, N=125,

O=107; line 2: <50, N=58, O=43. (i) line 1: High/>50, N=121, O=103; line 2: low/<50, N=17, O=10; line 3: High/<50, N=41, O=33; line 4: Low/>50, N=4, O=4.

**[0028]** Figure 22 is a diagram illustrating most significant probelets in tumor and normal data sets, age at diagnosis or both, according to some embodiments.

**[0029]** Figure 23 is a diagram illustrating survival analyses of an initial set of a number of patients classified by GBM-associated chromosome number changes, according to some embodiments.

**[0030]** Figure 24 is a diagram illustrating survival analyses of an initial set of a number of patients classified by copy number changes in selected segments, according to some embodiments. X-axis (all graphs): survival time (months); Y-axis (all graphs): Fraction of surviving patients from the initial set. (a) line 1: No CNA, N=213, O=176; line 2: Gain, N=34, O=27. (b) line 1: No CNA, N=233, O=190; line 2: Gain, N=8, O=7. (c) line 1: Gain, N=148, O=120; line 2: No CNA, N=98, O=82. (d) line 1: No CNA, N=195, O=166; line 2: Gain, N=52, O=37. (e) line 1: No CNA, N=227, O=192; line 2: Gain, N=19, O=11. (f) line 1: Loss, N=128, O=102; line 2: No CNA, N=118, O=100. (g) line 1: No CNA, N=145, O=120; line 2: Loss, N=102, O=83. (h) line 1: No CNA, N=235, O=193; line 2: Gain, N=9, O=7. (i) line 1: No CNA, N=207, O=170; line 2: Gain, N=39, O=32. (j) line 1: No CNA, N=227, O=186; line 2: Gain, N=19, O=17. (k) line 1: No CNA, N=191, O=167; line 2: Gain, N=56, O=36. (l) line 1: No CNA, N=231, O=191; line 2: Gain, N=14, O=11.

**[0031]** Figure 25 is a diagram illustrating survival analyses of an initial set of a number of patients classified by a mutation in one of the genes, according to some embodiments;

**[0032]** Figure 26 is a diagram illustrating a first most tumor-exclusive probelet and a corresponding tumor arraylet uncovered by GSVD of the patient-matched GBM and normal blood aCGH profiles, according to some embodiments.

**[0033]** Figure 27 is a diagram illustrating a normal-exclusive probelet and a corresponding normal arraylet uncovered by GSVD, according to some embodiments.

[0034] Figure 28 is a diagram illustrating another normal-exclusive probelet and a corresponding normal arraylet uncovered by GSVD, according to some embodiments.

[0035] Figure 29 is a diagram illustrating yet another normal-exclusive probelet and a corresponding normal arraylet uncovered by GSVD, according to some embodiments.

[0036] Figure 30 is a diagram illustrating yet another normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments.

[0037] Figure 31 is a diagram illustrating a first most normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments.

[0038] Figure 32 is a diagram illustrating differences in copy numbers among the TCGA annotations associated with the significant probelets, according to some embodiments.

[0039] Figure 33 is a diagram illustrating copy-number distributions of one of the probelet and the corresponding normal arraylet and tumor arraylet, according to some embodiments.

[0040] Figure 34 is a table illustrating proportional hazard models of three sets of patients classified by GSVD, according to some embodiments.

[0041] Figure 35 is a table illustrating enrichment of significant probelets in TCGA annotations, according to some embodiments.

[0042] Figure 36 is a diagram illustrating HO GSVD of biological data related to patient and normal samples, according to some embodiments.

[0043] Figure 37 is a diagram illustrating that the GSVD of two matrices  $D_1$  and  $D_2$  is reformulated as a linear transformation of the two matrices from the two rows x columns spaces to two reduced and diagonalized left basis vectors x right basis vectors spaces, according to some embodiments. The right basis vectors are shared by both datasets. Each right basis vector corresponds to two left basis vectors.

[0044] Figure 38 is a diagram illustrating that the higher-order GSVD (HO GSVD) of three matrices  $D_1$ ,  $D_2$ , and  $D_3$  is a linear transformation of the three matrices from the three rows x columns spaces to three reduced and diagonalized left basis vectors x right basis vectors spaces, according to some embodiments. The right basis vectors are shared by all three datasets. Each right basis vector corresponds to three left basis vectors.

[0045] Figure 39 is a diagram illustrating a higher-order EVD (HOEVD) of the third-order series of the three networks, according to some embodiments.

[0046] Figure 40 is a Table showing the Cox proportional hazard models of the three sets of patients classified by GSVD, chemotherapy or both, according to some embodiments. In each set of patients, the multivariate Cox proportional hazard ratios for GSVD and chemotherapy are similar and do not differ significantly from the corresponding univariate hazard ratios. This means that GSVD and chemotherapy are independent prognostic predictors. The P-values are calculated without adjusting for multiple comparisons.

[0047] Figure 41 is a diagram illustrating the Kaplan-Meier (KM) survival analyses of only the chemotherapy patients from the three sets classified by GSVD, according to some embodiments.

[0048] Figure 42 is a diagram illustrating the KM survival analysis of only the chemotherapy patients in the initial set, classified by a mutation in *IDH1*, according to some embodiments.

[0049] Figure 43 is a diagram illustrating the KM survival analyses of only the chemotherapy patients in the initial set of 251 patients classified by copy number changes in selected segments, according to some embodiments. X-axis (all graphs): survival time (months); Y-axis (all graphs): Fraction of surviving chemotherapy patients from the initial set. (a) line 1: No CNA, N=162, O=128; line 2: Gain, N=25, O=19. (b) line 1: No CNA, N=178, O=139; line 2: Gain, N=5, O=4. (c) line 1: Gain N=109, O=85; line 2: No CNA, N=77, O=61. (d) line 1: No CNA, N=149, O=123; line 2: Gain, N=38, O=24. (e) line 1: No CNA, N=171, O=139; line 2: Gain, N=15, O=8. (f) line 1: Loss, N=96, O=74; line 2: No CNA, N=90, O=72. (g) line 1: No

CNA, N=110, O=86; line 2: Loss, N=77, O=61. (h) line 1: No CNA, N=176, O=138; line 2: Gain, N=9, O=7. (i) line 1: No CNA, N=160, O=126; line 2: Gain, N=27, O=21. (j) line 1: No CNA, N=171, O=134; line 2: Gain, N=15, O=13. (k) line 1: No CNA, N=144, O=123; line 2: Gain, N=43, O=24. (l) line 1: No CNA, N=174, O=138; line 2: Gain, N=12, O=9.

### Summary

**[0050]** Given increasingly large datasets of biological information associated with disease states, there is a need for an enhanced mathematical framework that can assist in disease related characterization of the datasets including providing effective diagnostic and prognostic predictors and treatment plans.

**[0051]** Some embodiments provide systems, computer readable storage media including instructions, and computer-implemented methods, for disease related characterization of biological data.

**[0052]** Some such methods include the following steps: by a processor, applying a decomposition algorithm to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ ; wherein the data comprise indicators, represented in at least one of respective rows and columns of the tensor, of values of at least two index parameters; and determining, based on the eigenvectors and on values, associated with a subject, of the at least two index parameters, an indicator of a health parameter of the subject; wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability and an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

**[0053]** Optionally, the method further comprises outputting said indicator of health parameter along with a medical assessment, such as an assessment of disease risk (e.g., the subject's probability of developing a disease; the presence or the absence of a disease; the actual or predicted onset, progression, severity, or treatment outcome of a disease, etc.). The medical assessment can be informed to either a physician, or the subject. Optionally, appropriate

recommendations can be made (such as a treatment regimen, a preventative treatment regimen, an exercise regimen, a dietary regimen, a life style adjustment etc.) to reduce the risk of developing the disease, or design a treatment regiment that is likely to be effective in treating the disease.

**[0054]** In some embodiments, the index parameters comprise at least two of: patient identifications, tissue type identifications, a health status of one or more patients, a bioactive agent exposure status, an environmental exposure status, a nucleotide sequence copy numbers, DNA sequences, mRNA sequences, mRNA levels, a micro-RNA expression level, a DNA methylation level, a level of binding of proteins to DNA, a level of binding of proteins to RNA, gene product levels, gene product activity levels, a cell cycle status, a biochemical status, imaging data, a treatment status, biomarker levels, or time periods.

**[0055]** In some embodiments, the applying further comprises generating a diagonal matrix of singular values for each of A and B, and wherein the determining is further based on at least one of the diagonal matrices.

**[0056]** In some embodiments, the eigenvectors of  $A^T A$  are the same as the eigenvectors of  $B^T B$ .

**[0057]** The subject technology is embodied by at least the following items:

1. A method, for medical characterization of a subject based on biological data, comprising:

applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^T A$ ,  $BB^T$ , and  $B^T B$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

determining an indicator of a health parameter of a subject, the determining based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

2. The method of item 1, wherein the index parameters comprise at least two of: patient identifications, tissue type identifications, a health status of one or more patients, a bioactive agent exposure status, an environmental exposure status, a nucleotide sequence copy numbers, DNA sequences, mRNA sequences, mRNA levels, a micro-RNA expression level, a DNA methylation level, a level of binding of proteins to DNA, a level of binding of proteins to RNA, gene product levels, gene product activity levels, a cell cycle status, a biochemical status, imaging data, a treatment status, biomarker levels, or time periods.

3. The method of item 1 or 2, wherein the applying further comprises generating a diagonal matrix of singular values of each of A and B, and wherein the determining is further based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^T A$ .

4. The method of any one of items 1-3, wherein the eigenvectors of  $A^T A$  are the same as the eigenvectors of  $B^T B$ .

5. The method of any one of items 1-4, wherein the determining occurs at a first time, and further comprising repeating the determining at a second time to track a course of a health condition of the subject.

6. The method of any one of items 1-5, wherein at least one of the index parameters is measurable by at least one of a DNA microarray, DNA sequencing, a protein microarray, or mass spectrometry.

7. The method of any one of items 1-6, wherein the data comprises chromatin or histone modification, and wherein the data derived from a patient-specific sample including at least one of a normal tissue, a disease-related tissue, or a culture of a patient's cell.

8. The method of any one of items 1-7, wherein the data comprise at least one of magnetic resonance imaging (MRI) data, electrocardiogram (ECG) data, electromyography (EMG) data, or electroencephalogram (EEG) data.

9. The method of any one of items 1-8, wherein the applying substantially removes from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

10. The method of any one of items 1-9, wherein the algorithm decomposes the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

11. The method of any one of items 1-10, wherein the applying classifies the subject into a subgroup of patients based on at least patient-specific genomic data.

12. The method of any one of items 1-11, wherein the applying correlates an outcome of a therapeutic method and a genomic predictor in the data.

13. A system, for medical characterization of a subject based on biological data, comprising:

a processor configured to apply a decomposition algorithm to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

an analysis module configured to determine an indicator of a health parameter of a subject, based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

14. The system of item 13, wherein the processor is further configured to generate a diagonal matrix of singular values of each of submatrices A and B, and wherein the analysis module is further configured to determine the indicator of the health parameter based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^T A$ .

15. The system of item 13 or 14, wherein the analysis module is further configured to determine the indicator of the health parameter at a first time, and to repeat the determination at a second time to track a course of a health condition of the subject.

16. The system of any one of claims 13-15, wherein the processor is further configured to substantially remove from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

17. The system of any one of items 13-16, wherein the processor is further configured to apply the decomposition algorithm to decompose the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

18. The system of any one of items 13-17, wherein the processor is further configured to apply the decomposition algorithm to classify the subject into a subgroup of patients based on at least patient-specific genomic data.

19. The system of any one of items 13-18, wherein the processor is further configured to apply the decomposition algorithm to correlate an outcome of a therapeutic method and a genomic predictor in the data.

20. A non-transitory machine-readable medium comprising instructions that, when executed by one or more processors, perform the following acts:

applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^T A$ ,  $BB^T$ , and  $B^T B$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

determining an indicator of a health parameter of a subject, the determining based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

21. The machine-readable medium of item 20, wherein the applying further comprises generating a diagonal matrix of singular values of each of A and B, and wherein the determining is further based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^T A$ .

22. The machine-readable medium of item 20 or 21, wherein the applying substantially removes from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

23. The machine-readable medium of any one of items 20-22, wherein the algorithm decomposes the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

24. The machine-readable medium of any one of items 20-23, wherein the applying classifies the subject into a subgroup of patients based on at least patient-specific genomic data.

25. The machine-readable medium of any one of items 20-24, wherein the applying correlates an outcome of a therapeutic method and a genomic predictor in the data.

26. A method, for medical characterization of a subject based on biological data, comprising:

(a) applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^T A$ ,  $BB^T$ , and  $B^T B$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters;

(b) determining an indicator of a health parameter of a subject, the determining based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject; and

(c) outputting said indicator of health parameter along with a medical assessment.

### **Description of Embodiments**

**[0058]** Figure 1 is a high-level diagram illustrating examples of tensors 100 including biological datasets, according to some embodiments. In general, a tensor representing a number of biological datasets may comprise an Nth-order tensor including a number of multi-dimensional (e.g., two or three dimensional) matrices. The Nth-order tensor may include a number of biological datasets. Some of the biological datasets may correspond to one or more biological samples. Some of the biological dataset may include a number of biological data arrays, some of which may be associated with one or more subjects. Some examples of biological data that may be represented by a tensor includes tensors (a), (b) and (c) shown in Figure 1. The tensor (a) represents a third order tensor (i.e., a cuboid), in which each dimension (e.g., gene, condition and time) represent a degree of freedom in the cuboid. If unfolded into a matrix, these degrees of freedom may be lost and most of the data included in the tensor may also be lost. However, decomposing the cuboid using a tensor decomposition technique, such as higher-order eigen-value decomposition (HOEVD) or higher-order single value decomposition (HOSVD) may uncover patterns of mRNA expression variations across the genes, the time points and conditions.

**[0059]** In the example tensor (b) the biological datasets are associated with genes and the one or more subjects comprises organisms and data arrays may include cell cycle stages. The

tensor decomposition in this case may allow, for example, integrating global mRNA expressions measured for various organisms, removal of experimental artifacts and identification of significant combinations of patterns of expression variation across the genes, for various organisms and for different cell cycle stages. Similarly, in tensor (c) the biological datasets are associated with a network K of N-genes by N-genes. Where the network K may represent a number of studies on the genes. The tensor decomposition (e.g., HOEVD) in this case may allow, for example, uncovering important relations among the genes (e.g., pheromone-response-dependent relation or orthogonal cell-cycle-dependent relation). An example of a tensor represented by a three-dimensional array is discussed below with respect to Figure 2.

**[0060]** Figure 2 is a high-level diagram illustrating a linear transformation of a number of two dimensional (2-D) arrays forming a three-dimensional (3-D) array 200, according to some embodiments. The 3-D array 200 may be stored in memory 300 (see Figure 3). The 3-D array 200 may include a number N of biological datasets that correspond to genetic sequences. In some embodiments, the number N can be greater than two. Each biological dataset may correspond to a tissue type and can include a number M of biological data arrays. Each biological data array may be associated with a patient or, more generally, an organism). Each biological data array may include a plurality of data units (e.g., chromosomes). A linear transformation, such as a tensor decomposition algorithm may be applied to the 3-D array 200 to generate a plurality of eigen 2-D arrays 220, 230 and 240. The generated eigen 2-D arrays 220, 230 and 240 can be analyzed to determine one or more characteristics related to a disease (e.g., changes in glioblastoma multiforme (GBM) tumor with respect to normal tissue). The 3-D array 200 may comprise a number N of 2-D data arrays (D1, D2, D3, ... DN) (for clarity only D1-D3 are shown in Figure 2). Each of the 2-D data arrays (D1, D2, D3, ... DN) can store one set of the biological datasets and includes M columns. Each column can store one of the M biological data arrays corresponding to a subject such as a patient.

**[0061]** As used herein, "health status" may refer to the presence, absence, quality, rank, or severity of any disease or health condition, history and physical examination finding, laboratory value, and the like. As used herein, a "health parameter" can include a differential

diagnosis, meaning a diagnosis that is potential, confirmed, unconfirmed, based on a likelihood, ranked, or the like.

**[0062]** In some embodiments, each biological data array may comprise biological data measurable by a DNA microarray (e.g., genomic DNA copy numbers, genome-wide mRNA expressions, binding of proteins to DNA and binding of proteins to RNA), a sequencing technology (e.g., using a different technology that covers the same ground as microarrays), a protein microarray or mass spectrometry, where protein abundance levels are measured on a large proteomic scale and a traditional measurement (e.g., immunohistochemical staining). The biological data may include chromatin or histone modification, a DNA copy number, an mRNA expression, a micro-RNA expression, a DNA methylation, binding of proteins to DNA, binding of proteins to RNA or protein abundance levels.

**[0063]** In some embodiments, the biological data may be derived from a patient-specific sample including a normal tissue, a disease-related tissue or a culture of a patient's cell. The biological datasets may also be associated with genes and the one or more subjects comprises at least one of time points or conditions. The tensor decomposition of the Nth-order tensor may allow for identifying abnormal patterns to identify genes or proteins which enable including or excluding a diagnosis. Further, the tensor decomposition may allow classifying a patient into a subgroup of patients based on patient-specific genomic data, resulting in an improved diagnosis by identifying the patient's disease subtype. The tensor decomposition may also be advantageous in patients therapy planning, for example, by allowing patient-specific therapy to be designed based criteria, such as, a correlation between an outcome of a therapeutic method and a global genomic predictor.

**[0064]** In patients disease prognosis, the tensor decomposition may facilitate designing at least one of predicting a patient's survival or a patient's response to a therapeutic method such as chemotherapy. The Nth-order tensor may include a patient's routine examination data, in which case decomposition of the tensor may allow designing of a personalized preventive regimen for a patient based on analyses of the patient's routine examinations data. In some embodiments, the biological datasets may be associated with imaging data including

magnetic resonance imaging (MRI) data, electro cardiogram (ECG) data, electromyography (EMG) data or electroencephalogram (EEG) data. The biological datasets may associated with vital statistics or phenotypic data.

**[0065]** In some embodiments, the tensor decomposition of the Nth-order tensor may allow removing normal pattern copy number variations (CNVs) and an experimental variation from a genomic sequence. The tensor decomposition of the Nth-order tensor may permit an improved prognostic prediction of the disease by revealing disease-associated changes in chromosome copy numbers, focal copy number variations (CNVs) nonfocal CNVs and the like. The tensor decomposition of the Nth-order tensor may also allow integrating global mRNA expressions measured in multiple time courses, removal of experimental artifacts and identification of significant combinations of patterns of expression variation across the genes, the time points and the conditions.

**[0066]** In embodiments, applying the tensor decomposition algorithm may comprise applying at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigen-value decomposition (HOEVD) or parallel factor analysis (PARAFAC) to the Nth-order tensor. Some of the present embodiments apply HOSVD to decompose the 3-D array 200, as described in more detail herein. The PARAFAC method is known in the art and will not be described with respect to the present embodiments.

**[0067]** The HOSVD generated eigen 2-D arrays may comprise a set of N left-basis 2-D arrays 220. Each of the left-basis arrays 220 (e.g.,  $U_1, U_2, U_3, \dots U_N$ ) (for clarity only  $U_1-U_3$  are shown in Figure 2) may correspond to a tissue type and can include a number M of columns, each of which stores a left-basis vector 222 associated with a patient. The eigen 2-D arrays 230 comprise a set of N diagonal arrays ( $\Sigma_1, \Sigma_2, \Sigma_3, \dots \Sigma_N$ ) (for clarity only  $\Sigma_1-\Sigma_3$  are shown in Figure 2). Each diagonal array (e.g.,  $\Sigma_1, \Sigma_2, \Sigma_3, \dots$  or  $\Sigma_N$ ) may correspond to a tissue type and can include a number N of diagonal elements 232. The 2-D array 240 comprises a right-basis array, which can include a number of right-basis vectors 242.

[0068] In some embodiments, decomposition of the Nth-order tensor may be employed for disease related characterization such as diagnosing, tracking a clinical course or estimating a prognosis, associated with the disease.

[0069] Figure 3 is a block diagram illustrating a biological data characterization system 300 coupled to a database 350, according to some embodiments. The system 300 includes a processors 310, memory 320, an analysis module 330 and a display module 340. Processor 310 may include one or more processors and may be coupled to memory 320. Memory 320 may comprise volatile memory such as random access memory (RAM) or nonvolatile memory (e.g., read only memory (ROM), flash memory, etc.). Memory 320 may also include machine-readable medium, such as magnetic or optical disks. Memory 320 may retrieve information related to the Nth-order tensors 100 of Figure 1 or the 3-D array 200 of Figure 2 from a database 350 coupled to the system 300 and store tensors 100 or the 3-D array 200 along with 2-D eigen-arrays 220, 230 and 240 of Figure 2. Database 350 may be coupled to system 300 via a network (e.g., Internet, wide area network (WNA), local area network (LNA), etc.). In some embodiments, system 300 may encompass database 350.

[0070] Processor 310 can apply a tensor decomposition algorithm, such as HOSVD, HO GSVD or HOEVD to the tensors 100 or 3-D array 200 and generate eigen 2-D arrays 220, 230 and 240. In some embodiments, processor 310 may apply the HOSVD or HO GSVD algorithms to array comparative genomic hybridization (aCGH) data from patient-matched normal and glioblastoma multiforme (GBM) blood samples. Application of HOSVD algorithm may remove one or more normal pattern copy number variations (CNVs) or experimental variations from the aCGH data. The HOSVD algorithm can also reveal GBM-associated changes in at least one of chromosome copy numbers, focal CNVs and unreported CNVs existing in the aCGH data. In some embodiments, processor 310 may apply a decomposition algorithm to an Nth-order tensor representing data ( $N \geq 2$ ) to generate, from two or more submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ . The data may comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters. Analysis module 330 can perform disease related characterizations as discussed above. For example, analysis module 330 can facilitate various analyses of eigen 2-D

arrays 230 of Figure 2, for example, by assigning each diagonal element 232 of Figure 2 to an indicator of a significance of a respective element of a right-basis vector 222 of Figure 2, as described herein in more detail. In some embodiments, Analysis module 330 can determine an indicator of a health parameter of a subject, based on the eigenvectors and on values, associated with the subject, of the two or more index parameters. The display module 240 can display 2-D arrays 220, 230 and 240 and any other graphical or tabulated data resulting from analyses performed by analysis module 330. Display module 330 can display the indicator of the health parameter of the subject in various ways including digital readout, graphical display, or the like. In embodiments, the indicator of the health parameter may be communicated, to a user or a printer device, over a phone line, a computer network, or the like. Display module 330 may comprise software and/or firmware and may use one or more display units such as cathode ray tubes (CRTs) or flat panel displays.

**[0071]** Figure 4 is a flowchart of a method 400 for genomic prognostic prediction, according to some embodiments. Method 400 includes storing the nth-tensors 100 of Figure 1 or 3-D array 200 of Figure 2 in memory 320 of Figure 3 (410). A tensor decomposition algorithm such as HOSVD, HO GSVD or HOEVD may be applied, by processor 310 of Figure 3, to the datasets stored in tensors 100 or 3-D array 200 to generate eigen 2-D arrays 220, 230 and 240 of Figure 2 (420). The generated eigen 2-D arrays 220, 230 and 240 may be analyzed by analysis module 330 to determine one or more disease-related characteristics (430). The HOSVD algorithm is mathematically described herein with respect to  $N > 2$  matrices (i.e., arrays  $D_1$ - $D_N$ ) of 3-D array 200. Each matrix can be a real  $m_i \times n$  matrix. Each matrix is exactly factored as  $D_i = U_i \Sigma_i V^T$ , where  $V$ , identical in all factorizations, is obtained from the balanced eigensystem  $SV = VA$  of the arithmetic mean  $S$  of all pairwise quotients  $A_i A_j^{-1}$  of the matrices  $A_i = D_i^T D_i$ , where  $i$  is not equal to  $j$ , independent of the order of the matrices  $D_i$ . It can be proved that this decomposition extends to higher orders all of the mathematical properties of the GSVD except for column-wise orthogonality of the matrices  $U_i$  (e.g., 2-D arrays 220 of Figure 2).

**[0072]** It can be proved that matrix  $S$  is nondefective, i.e.,  $S$  has  $n$  independent eigenvectors and that  $V$  is real and that the eigenvalues of  $S$  (i.e.,  $\lambda_1, \lambda_2, \dots, \lambda_N$ ) satisfy  $\lambda_k \geq 1$ . In the described HO GSVD comparison of two matrices, the  $k_{th}$  diagonal element of  $\Sigma_i = \text{diag}(\sigma_{i,k})$

(e.g., the  $k_{th}$  element 232 of Figure 2) is interpreted in the factorization of the  $i_{th}$  matrix  $D_i$  as indicating the significance of the  $k_{th}$  right basis vector  $v_k$  in  $D_i$  in terms of the overall information that  $v_k$  captures in  $D_i$ . The ratio  $\sigma_{i,k}/\sigma_{j,k}$  indicates the significance of  $v_k$  in  $D_i$  relative to its significance in  $D_j$ . It can also be proved that an eigenvalue  $\lambda_k = 1$  corresponds to a right basis vector  $v_k$  of equal significance in all matrices  $D_i$  and  $D_j$  for all  $i$  and  $j$ , when the corresponding left basis vector  $u_{i,k}$  is orthonormal to all other left basis vectors in  $U_i$  for all  $i$ . Detailed description of various analysis results corresponding to application of the HOSVD to a number of datasets related to patients and other subjects will be discussed with respect to Figures 5-43 below. For clarity, more detail treatment of mathematical aspects of HOSVD is skipped here and is provided in documents attached as Appendices A, B, and C. Disclosures in Appendix A have also been published as Lee et al., (2012) *GSVD Comparison of Patient-Matched Normal and Tumor aCGH Profiles Reveals Global Copy-Number Alterations Predicting Glioblastoma Multiforme Survival*, in PLoS ONE 7(1): e30098. doi:10.1371/journal.pone.0030098. Disclosures in Appendices B and C have been published as Ponnappalli et al., (2011) *A Higher-Order Generalized Singular Value Decomposition for Comparison of Global mRNA Expression from Multiple Organisms* in PLoS ONE 6(12): e28072. doi:10.1371/journal.pone.0028072.

**[0073]** The HOEVD tensor decomposition method can be used for decomposition of higher order tensors. Herein, as an example, the HOEVD tensor decomposition method is described in relation with a the third-order tensor of size K-networks x N-genes x N-genes as follows:

[0074]

**Higher-Order EVD (HOEVD).** Let the third-order tensor  $\{\hat{a}_k\}$  of size  $K$ -networks  $\times$   $N$ -genes  $\times$   $N$ -genes tabulate a series of  $K$  genome-scale networks computed from a series of  $K$  genome-scale signals  $\{\hat{e}_k\}$ , of size  $N$ -genes  $\times$   $M_k$ -arrays each, such that  $\hat{a}_k = \hat{e}_k \hat{e}_k^T$  for all  $k = 1, 2, \dots, K$ . We define and compute a HOEVD of the tensor of networks  $\{\hat{a}_k\}$ ,

$$\hat{a} \equiv \sum_{k=1}^K \hat{a}_k = \hat{u} \left( \sum_{k=1}^K \hat{e}_k^2 \right) \hat{u}^T = \hat{u} \hat{\varepsilon}^2 \hat{u}^T, \quad [5]$$

using the SVD of the appended signals  $\hat{e} \equiv (\hat{e}_1, \hat{e}_2, \dots, \hat{e}_K) = \hat{u} \hat{\varepsilon} \hat{v}^T$ , where the  $m$ th column of  $\hat{u}$ ,  $|\alpha_m\rangle \equiv \hat{u}|m\rangle$ , lists the genome-scale expression of the  $m$ th eigenarray of  $\hat{e}$ . Whereas the matrix EVD is equivalent to the matrix SVD for a symmetric nonnegative matrix, this tensor HOEVD is different from the tensor higher-order SVD (14–16) for the series of symmetric nonnegative matrices  $\{\hat{a}_k\}$ , where the higher-order SVD is computed from the SVD of the appended networks  $(\hat{a}_1, \hat{a}_2, \dots, \hat{a}_K)$  rather than the appended signals. This HOEVD formulates the overall network computed from the appended signals  $\hat{a} = \hat{e} \hat{e}^T$  as a linear superposition of a series of  $M \equiv \sum_{k=1}^K M_k$  rank-1

symmetric “subnetworks” that are decorrelated of each other,  $\hat{a} = \sum_{m=1}^M \varepsilon_m^2 |\alpha_m\rangle\langle\alpha_m|$ . Each subnetwork is also decoupled of all other subnetworks in the overall network  $\hat{a}$ , since  $\hat{\varepsilon}$  is diagonal.

This HOEVD formulates each individual network in the tensor  $\{\hat{a}_k\}$  as a linear superposition of this series of  $M$  rank-1 symmetric decorrelated subnetworks and the series of  $M(M-1)/2$  rank-2 symmetric couplings among these subnetworks (Figure 39), such that

$$\begin{aligned} \hat{a}_k = & \sum_{m=1}^M \varepsilon_{k,m}^2 |\alpha_m\rangle\langle\alpha_m| \\ & + \sum_{m=1}^M \sum_{l=m+1}^M \varepsilon_{k,lm}^2 (|\alpha_l\rangle\langle\alpha_m| + |\alpha_m\rangle\langle\alpha_l|), \end{aligned} \quad [6]$$

for all  $k = 1, 2, \dots, K$ . The subnetworks are not decoupled in any one of the networks  $\{\hat{a}_k\}$ , since, in general,  $\{\hat{\varepsilon}_k^2\}$  are symmetric but not diagonal, such that  $\varepsilon_{k,lm}^2 \equiv \langle l | \hat{\varepsilon}_k^2 | m \rangle = \langle m | \hat{\varepsilon}_k^2 | l \rangle \neq 0$ . The significance of the  $m$ th subnetwork in the  $k$ th network is indicated by the  $m$ th fraction of eigenexpression of the  $k$ th network  $p_{k,m} = \varepsilon_{k,m}^2 / (\sum_{k=1}^K \sum_{m=1}^M \varepsilon_{k,m}^2) \geq 0$ , i.e., the expression correlation captured by the  $m$ th subnetwork in the  $k$ th network relative to that captured by all subnetworks (and all couplings among them, where  $\sum_{k=1}^K \varepsilon_{k,lm}^2 = 0$  for all  $l \neq m$ ) in all networks. Similarly, the amplitude of the fraction  $p_{k,lm} = \varepsilon_{k,lm}^2 / (\sum_{k=1}^K \sum_{m=1}^M \varepsilon_{k,m}^2)$  indicates the significance of the coupling between the  $l$ th and  $m$ th subnetworks in the  $k$ th network. The sign of this fraction indicates the direction of the coupling, such that  $p_{k,lm} > 0$  corresponds to a transition from the  $l$ th to the  $m$ th subnetwork and  $p_{k,lm} < 0$  corresponds to the transition from the  $m$ th to the

/th. For real signals  $\{\hat{e}_k\}$ , the subnetworks are unique, and the couplings among them are unique up to phase factors of  $\pm 1$ , except in degenerate subspaces of  $\hat{e}$ .

[0075]

**Interpretation of the Subnetworks and Their Couplings.** We parallel- and antiparallel-associate each subnetwork or coupling with most likely expression correlations, or none thereof, according to the annotations of the two groups of  $x$  pairs of genes each, with largest and smallest levels of correlations in this subnetwork or coupling among all  $X = N(N - 1)/2$  pairs of genes, respectively. The  $P$  value of a given association by annotation is calculated by using combinatorics and assuming hypergeometric probability distribution of the  $Y$  pairs of annotations among the  $X$  pairs of genes, and of the subset of  $y \subseteq Y$  pairs of annotations among the subset of  $x \subseteq X$  pairs of genes,  $P(x; y, Y, X) = \binom{X}{x}^{-1} \sum_{z=y}^x \binom{Y}{z} \binom{X-Y}{x-z}$ , where  $\binom{X}{x} = X!x!^{-1}(X - x)^{-1}$  is the binomial coefficient (17). The most likely association of a subnetwork with a pathway or of a coupling between two subnetworks with a transition between two pathways is that which corresponds to the smallest  $P$  value. Independently, we also parallel- and antiparallel-associate each eigenarray with most likely cellular states, or none thereof, assuming hypergeo-

metric distribution of the annotations among the  $N$ -genes and the subsets of  $n \subseteq N$  genes with largest and smallest levels of expression in this eigenarray. The corresponding eigengene might be inferred to represent the corresponding biological process from its pattern of expression.

For visualization, we set the  $x$  correlations among the  $X$  pairs of genes largest in amplitude in each subnetwork and coupling equal to  $\pm 1$ , i.e., correlated or anticorrelated, respectively, according to their signs. The remaining correlations are set equal to 0, i.e., decorrelated. We compare the discretized subnetworks and couplings using Boolean functions (6).

[0076]

Fig. 39 is a higher-order EVD (HOEVD) of the third-order series of the three networks  $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ . The network  $\hat{a}_3$  is the pseudoinverse projection of the network  $\hat{a}_1$  onto a genome-scale proteins' DNA-binding basis signal of 2,476-genes  $\times$  12-samples of development transcription factors (Mathematica Notebook 3 and Data Set 4), computed for the 1,827 genes at the intersection of  $\hat{a}_1$  and the basis signal. The HOEVD is computed for the 868 genes at the intersection of  $\hat{a}_1, \hat{a}_2$  and  $\hat{a}_3$ . Raster display of  $\hat{a}_k \approx \sum_{m=1}^3 \epsilon_{k,m}^2 |\alpha_m\rangle \langle \alpha_m| + \sum_{m=1}^3 \sum_{l=m+1}^3 \epsilon_{k,lm}^2 (|\alpha_l\rangle \langle \alpha_m| + |\alpha_m\rangle \langle \alpha_l|)$ , for all  $k = 1, 2, 3$ , visualizing each of the three networks as an approximate superposition of only the three most significant HOEVD subnetworks and the three couplings among them, in the subset of 26 genes which constitute the 100 correlations in each subnetwork and coupling that are largest in amplitude among the 435 correlations of 30 traditionally-classified cell cycle-regulated genes. This tensor HOEVD is different from the tensor higher-order SVD for the series of symmetric nonnegative matrices  $\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$ . The subnetworks correlate with the genomic pathways that are manifest in the series of networks. The most significant subnetwork correlates with the response to the pheromone. This subnetwork does not contribute to the expression correlations of the cell cycle-projected network  $\hat{a}_2$ , where  $\epsilon_{2,1}^2 \approx 0$ . The second and third subnetworks correlate with the two pathways of antipodal cell cycle expression oscillations, at the cell cycle stage  $G_1$  vs. those at  $G_2$ , and at S vs. M, respectively. These subnetworks do not contribute to the expression correlations of the development-projected network  $\hat{a}_3$ , where  $\epsilon_{3,2}^2 \approx \epsilon_{3,3}^2 \approx 0$ . The couplings correlate with the transitions among these independent pathways that are manifest in the individual networks only. The coupling between the first and second subnetworks is associated with the transition between the two pathways of response to pheromone and cell cycle expression oscillations at  $G_1$  vs. those at  $G_2$ , i.e., the exit from pheromone-induced arrest and entry into cell cycle progression. The coupling between the first and third subnetworks is associated with the transition between the response to pheromone and cell cycle expression oscillations at S vs. those at M, i.e., cell cycle expression oscillations at  $G_1/S$  vs. those at M. The coupling between the second and third subnetworks is associated with the transition between the orthogonal cell cycle expression oscillations at  $G_1$  vs. those at  $G_2$  and at S vs. M, i.e., cell cycle expression oscillations at the two antipodal cell cycle checkpoints of  $G_1/S$  vs.  $G_2/M$ . All these couplings add to the expression correlation of the cell cycle-projected  $\hat{a}_2$ , where  $\epsilon_{2,12}^2, \epsilon_{2,13}^2, \epsilon_{2,23}^2 > 0$ ; their contributions to the expression correlations of  $\hat{a}_1$  and the development-projected  $\hat{a}_3$  are negligible.

[0077] Figures 5A-5C show Kaplan-Meier survival analyses of an initial set of 251 patients classified by GBM-associated chromosome number changes. Figure 5A shows KM survival analysis for 247 patients with TCGA annotations in the initial set of 251 patients, classified by number changes in chromosome 10. This figure shows almost overlapping KM curves with a KM median survival time difference of  $\sim 2$  months, and a corresponding log-rank test  $P$ -value  $\sim 10^{-1}$ , meaning that chromosome 10 loss, frequently observed in GBM, is a poor predictor of GBM survival. Figure 5B shows KM survival analysis for 247 patients classified by number changes in chromosome 7. This figure shows almost overlapping KM curves with a KM median survival time difference of  $< 1$  month and a corresponding log-rank test  $P$ -value  $> 5 \times 10^{-1}$ , meaning that chromosome 7 gain is a poor predictor of GBM survival. Figure 5C is a KM survival analysis for 247 patients classified by number changes in chromosome 9p. This figure shows a KM median survival time difference of  $\sim 3$  months, and a log-rank test  $P$ -value  $> 10^{-1}$ , meaning that chromosome 9p loss is a poor predictor of GBM survival.

**[0078]** Previously unreported CNAs identified by GSVD include TLK2, METTL2A, METTL2B, KDM5A, SLC6A12, SLC6A13, IQSEC3, CCNE1, POP4, PLEKHF1, C19orf12, and C19orf2. For example, TLK2/METTL2A (17q23.2) is amplified in ~22% of the patients; METTL2B (7q32.1) is amplified in ~8% of the patients; and KDM5A (12p13.33) is amplified in ~4% of the patients. Moreover, these identified genes primarily reside in 4 genetic segments: chr17:57,851,812- chr17:57,973,757 encompassing TLK2 and METTL2A (Fig. 6); chr7:127,892,509- chr7:127,947,649 encompassing METTL2B (Fig. 7); chr12:33,854- chr12:264,310 encompassing KDM5A, SLC6A12, SLC6A13, and IQSEC3 (Fig. 8); and chr19:33,329,393- chr19:35,322,055 encompassing CCNE1, POP4, PLEKHF1, C19orf12, and C19orf2 (Fig. 9).

**[0079]** Figure 6 is a diagram illustrating genes that are found in chromosomal segment 17:57,851,812-17:57,973,757 of the human genome, according to some embodiments.

**[0080]** Figure 7 shows a diagram of a genetic map illustrating the coordinates of TLK2 and METTL2A on segment chr17:57,851,812-chr17:57,973,757 on NCBI36/hg18 assembly of the human genome. The amplification of this segment is correlated with GBM prognosis. Copy-number amplification of TLK2 has been correlated with overexpression in several other cancers. Previous studies have shown that the human gene TLK2, with homologs in the plant *Arabidopsis thaliana* but not in the yeast *Saccharomyces cerevisiae*, encodes for a multicellular organisms-specific serine/threonine protein kinase, a biochemically putative drug target, whose activity directly depends on ongoing DNA replication.

**[0081]** Figure 8 shows a diagram of a genetic map illustrating the coordinates of METTL2B on segment chr7:127,892,509- chr7:127,947,649 on NCBI36/hg assembly of the human genome. Previous studies have shown that overexpression of METTL2A/B has been linked to metastatic samples relative to primary prostate tumor samples; cAMP response element-binding (CREB) regulation in myeloid leukemia, and response to chemotherapy in breast cancer patients.

**[0082]** Figure 9 shows a diagram of a genetic map illustrating the coordinates of CCNE1, POP4, PLEKHF1, C19orf12, and C19orf2 on segment chr19:33,329,393-

chr19:35,322,055 on NCBI36/hg assembly of the human genome. Previous studies have shown that CCNE1 regulates entry into the DNA synthesis phase of the cell division cycle and copy number amplification of CCNE1 has been linked with several cancers but not GBM. Recent studies suggest that there is a link between amplicon-dependent expression of CCNE1 together with the flanking genes POP4, PLEKHF1, C19orf12, and C19orf2 on the segment. and primary treatment of ovarian cancer may be due to rapid repopulation of the tumor after chemotherapy.

**[0083]** Figure 10 is a diagram illustrating survival analyses of a set of patients classified by copy number changes in selected segments, according to some embodiments. Survival analyses of the patients from the three sets classified by chemotherapy alone or GSVD and chemotherapy both. (a) KM and Cox survival analyses of the 236 patients with TCGA chemotherapy annotations in the initial set of 251 patients, classified by chemotherapy, show that lack of chemotherapy, with a KM median survival time difference of about 10 months and a univariate hazard ratio of 2.6 (Figure 40), confers more than twice the hazard of chemotherapy. (b) Survival analyses of the 236 patients classified by both GSVD and chemotherapy, show similar multivariate Cox hazard ratios, of 3 and 3.1, respectively. This means that GSVD and chemotherapy are independent prognostic predictors. With a KM median survival time difference of about 30 months, GSVD and chemotherapy combined make a better predictor than chemotherapy alone. (c) Survival analyses of the 317 patients with TCGA chemotherapy annotations in the inclusive confirmation set of 344 patients, classified by chemotherapy, show a KM median survival time difference of about 11 months and a univariate hazard ratio of 2.7, and confirm the survival analyses of the initial set of 251 patients. (d) Survival analyses of the 317 patients classified by both GSVD and chemotherapy show similar multivariate Cox hazard ratios, of 3.1 and 3.2, and a KM median survival time difference of about 30 months, with the corresponding log-rank test P-value  $<10^{-17}$ . This confirms that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone. (e) Survival analyses for the 154 patients with TCGA chemotherapy annotations in the independent validation set of 184 patients, classified by chemotherapy, show a KM median survival time difference of about 11 months and a univariate hazard ratio of 2.2, and validate the survival analyses of the initial set of 251 patients. (f)

Survival analyses for the 154 patients classified by both GSVD and chemotherapy, show similar multivariate Cox hazard ratios, of 3.3 and 2.7, and a KM median survival time difference of about 43 months. This validates that the prognostic contribution of GSVD is independent of chemotherapy, and that combined with chemotherapy, GSVD makes a better predictor than chemotherapy alone, also for patients with measured GBM aCGH profiles in the absence of matched normal blood profiles.

**[0084]** Figure 11 is a diagram illustrating a HO GSVD of biological data, according to some embodiments. In this raster display, the *S. pombe*, *S. cerevisiae* and human global mRNA expression datasets are tabulated as organism-specific genes x 17-arrays matrices  $D_1$ ,  $D_2$  and  $D_3$ . Overexpression, no change in expression, and underexpression have been centered at gene- and array-invariant expression. The underlying assumption is that there exists a one-to-one mapping among the 17 columns of the three matrices but not necessarily among their rows. These matrices are transformed to the reduced diagonalized matrices  $\Sigma_1$ ,  $\Sigma_2$  and  $\Sigma_3$ , each of 17-“arraylets,” i.e., left basis vectors x 17-“genelets,” i.e., right basis vectors, by using the organism-specific genes x 17-arraylets transformation matrices  $U_1$ ,  $U_2$  and  $U_3$  and the shared 17-genelets x 17-arrays transformation matrix  $V^T$ . For this particular  $V$ , this decomposition extends to higher orders all of the mathematical properties of the GSVD except for orthogonality of the arraylets, i.e., left basis vectors that form the matrices  $U_1$ ,  $U_2$  and  $U_3$ . Thus, the genelets, i.e., right basis vectors  $v_k$  are defined to be of equal significance in all the datasets when the corresponding arraylets  $u_{1,k}$ ,  $u_{2,k}$  and  $u_{3,k}$  are orthonormal to all other arraylets in  $U_1$ ,  $U_2$  and  $U_3$ , and when the corresponding higher-order generalized singular values are equal:  $\sigma_{1,k} = \sigma_{2,k} = \sigma_{3,k}$ . Like the GSVD for two organisms, the HO GSVD provides a sequence-independent comparative mathematical framework for datasets from more than two organisms, where the mathematical variables and operations represent biological reality: genelets of common significance in the multiple datasets, and the corresponding arraylets, represent cell-cycle checkpoints or transitions from one phase to the next, common to *S. pombe*, *S. cerevisiae* and human. Simultaneous reconstruction and classification of the three datasets in the common subspace that these patterns span outlines the biological similarity in the regulation of their cell-cycle programs. Notably,

genes of significantly different cell-cycle peak times but highly conserved sequences are correctly classified.

**[0085]** Figure 12 is a diagram illustrating a right basis array 1210 and patterns of expression variation across time, according to some embodiments. The right basis array 1210 and bar chart 1220 and graphs 1230 and 1240 relate to application of HO GSVD algorithm for decomposition of global mRNA expression for multiple organisms. (a) Right basis array 1210 displays the expression of 17 genelets across 17 time points, with overexpression, no change in expression, and underexpression around the array-invariant, i.e., time-invariant expression. (b) The bar chart 1220 depicts the corresponding inverse eigenvalues  $\lambda_k^{-1}$  showing that the 13th through the 17th genelets may be approximately equally significant in the three data sets with  $\lambda_k$  having a value approximately between 1 and 2, where the five corresponding arraylets in each data set are  $\varepsilon = 0.33$ -orthonormal to all other arraylets (see Figure 22). (c) The line-joined graph 1230 of the 13th (1), 14th (3) and 15th (2) genelets in the two-dimensional subspace that approximates the five-dimensional HO GSVD subspace, normalized to zero average and unit variance. (d) The line-joined graphs 1240 show the projected 16th (4) and 17th (5) genelets in the two-dimensional subspace. The five genelets describe expression oscillations of two periods in the three time courses.

**[0086]** Figure 13 is a diagram illustrating an HO GSVD reconstruction and classification of a number of mRNA expressions, according to some embodiments. Specifically, charts (a) to (i) shown in Figure 13, relate to the simultaneous HO GSVD reconstruction and classification of *S. pombe*, *S. cerevisiae* and human global mRNA expression in the approximately common HO GSVD subspace. In charts (a-c) *S. pombe*, *S. cerevisiae* and human array expression are projected from the five-dimensional common HO GSVD subspace onto the two-dimensional subspace that approximates the common subspace. The arrays are color-coded according to their previous cell-cycle classification. The arrows describe the projections of the  $k = 13, \dots, 17$  arraylets of each data set. The dashed unit and half-unit circles outline 100% and 50% of added-up (rather than canceled-out) contributions of these five arraylets to the overall projected expression. In charts (d-f) Expression of 380, 641 and 787 cell cycle-regulated genes of *S. pombe*, *S. cerevisiae* and human, respectively, are color-coded according to previous

classifications. Charts (g-i) show the HO GSVD pictures of the *S. pombe*, *S. cerevisiae* and human cell-cycle programs. The arrows describe the projections of the  $k = 13, \dots, 17$  shared genelets and organism-specific arraylets that span the common HO GSVD subspace and represent cell-cycle checkpoints or transitions from one phase to the next.

**[0087]** Figure 14 is a diagram illustrating simultaneous HO GSVD sequence-independent classification of a number of genes, according to some embodiments. The genes under consideration in Figure 14 are genes of significantly different cell-cycle peak times but highly conserved sequences. Chart (a) shows the *S. pombe* gene BFR1 and chart (b) shows its closest *S. cerevisiae* homologs. In Chart (c), the *S. pombe* and in chart (d), *S. cerevisiae* closest homologs of the *S. cerevisiae* gene PLB1 are shown. Chart (e) shows the *S. pombe* cyclin-encoding gene CIG2 and its closest *S. pombe*. Shown in chart (f) and (g) are the *S. cerevisiae* and human homologs, respectively.

**[0088]** Figure 15 is a diagram illustrating simultaneous correlations among  $n=17$  arraylets in one organism, according to some embodiments. Raster displays of  $U_i^T U_i$ , with correlations  $\geq \varepsilon = 0.33$ ,  $\leq -\varepsilon$ , and  $\in (-\varepsilon, \varepsilon)$ , show that for  $k = 13, \dots, 17$  the arraylets  $u_{i,k}$  with  $k = 13, \dots, 17$ , that correspond to  $1 \leq \lambda_k \leq 2$ , are  $\varepsilon = 0.33$ -orthonormal to all other arraylets in each data set. The corresponding five genelets,  $v_k$  are approximately equally significant in the three data sets with  $\sigma_{1,k} : \sigma_{2,k} : \sigma_{3,k} \sim 1 : 1 : 1$  in the *S. pombe*, *S. cerevisiae* and human datasets, respectively (Figure 12). Following Theorem 3, therefore, these genelets span the, these arraylets and genelets may span the approximately "common HO GSVD subspace" for the three data sets.

**[0089]** Figure 16 is a diagram illustrating three dimensional least squares approximation of the five-dimensional approximately common HO GSVD subspace, according to some embodiments. Line-joined graphs of the first (1), second (2) and third (3) most significant orthonormal vectors in the least squares approximation of the genelets  $v_k$  with  $k = 13, \dots, 17$  are shown. These orthonormal vectors span the common HO GSVD subspace. This five-dimensional subspace may be approximated with the two orthonormal vectors  $x$  and  $y$ , which fit normalized cosine functions of two periods, and 0- and  $-\pi/2$ - initial phases, i.e., normalized zero-phase cosine and sine functions of two periods, respectively.

[0090] Figure 17 is a diagram illustrating an example of an mRNA expression (*S. pombe* global mRNA expression) reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments. The example mRNA expression may include *S. pombe* global mRNA expression reconstructed in the five-dimensional common HO GSVD subspace with genes sorted according to their phases in the two-dimensional subspace that approximates it. Chart (a) is an expression of the sorted 3167 genes in the 17 arrays, centered at their gene- and array-invariant levels, showing a traveling wave of expression. Chart (b) shows an expression of the sorted genes in the 17 arraylets, centered at their arraylet-invariant levels. Arraylets  $k = 13, \dots, 17$  display the sorting. Chart (c) depicts line-joined graphs of the 13th (1), 14th (2), 15th (3), 16th (4) and 17th (5) arraylets fit one-period cosines with initial phases similar to those of the corresponding genelets (similar to probelets in Figure 11).

[0091] Figure 18 is a diagram illustrating another example of an mRNA expression (*S. cerevisiae* global mRNA expression) reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments. The example mRNA expression includes *S. cerevisiae* global mRNA expression reconstructed in the five-dimensional common HO GSVD subspace with genes sorted according to their phases in the two-dimensional subspace that approximates it. Chart (a) is an expression of the sorted 4772 genes in the 17 arrays, centered at their gene- and array-invariant levels, showing a traveling wave of expression. Chart (b) shows an expression of the sorted genes in the 17 arraylets, centered at their arraylet-invariant levels, where arraylets  $k = 13, \dots, 17$  display the sorting. Chart (c) depicts line-joined graphs of the 13th (1), 14th (2), 15th (3), 16th (4) and 17th (5) arraylets fit one-period cosines with initial phases similar to those of the corresponding genelets.

[0092] Figure 19 is a diagram illustrating a human global mRNA expression reconstructed in the five-dimensional approximately common HO GSVD subspace, according to some embodiments. The genes are sorted according to their phases in the two-dimensional subspace that approximates them. Chart (a) is an expression of the sorted 13,068 genes in the 17 arrays, centered at their gene- and array-invariant levels, showing a traveling wave of expression. Chart (b) shows an expression of the sorted genes in the 17 arraylets, centered at their arraylet-invariant levels, where arraylets  $k = 13, \dots, 17$  display the sorting. Chart (c) shows line-joined

graphs of the 13th (1), 14th (2), 15th (3), 16th (4) and 17th (5) arraylets fit one-period cosines with initial phases that may be similar to those of the corresponding genelets.

**[0093]** Figure 20 is a diagram illustrating significant probelets and corresponding tumor and normal arraylets uncovered by GSVD of the patient-matched GBM and normal blood aCGH profiles, according to some embodiments. (a) A chart 2010 is a plot of the second tumor arraylet and describes a global pattern of tumor-exclusive co-occurring CNAs across the tumor probes. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. Segments (black lines) identified by circular binary segmentation (CBS) include most known GBM-associated focal CNAs, e.g., Epidermal growth factor receptor (EGFR) amplification. CNAs previously unrecognized in GBM may include an amplification of a segment containing the biochemically putative drug target-encoding. (b) A chart 2015 shows a plot of a probelet that may be identified as the second most tumor-exclusive probelet, which may also be identified as the most significant probelet in the tumor data set, describes the corresponding variation across the patients. The patients are ordered and classified according to each patient's relative copy number in this probelet. There are 227 patients with high ( $>0.02$ ) and 23 patients with low, approximately zero, numbers in the second probelet. One patient remains unclassified with a large negative ( $<-0.02$ ) number. This classification may significantly correlate with GBM survival times. (c) A chart 2020 is a raster display of the tumor data set, with relative gain, no change, and loss of DNA copy numbers, which may show the correspondence between the GBM profiles and the second probelet and tumor arraylet. Chromosome 7 gain and losses of chromosomes 9p and 10, which may be dominant in the second tumor arraylet (see 2220 in Figure 22), may be negligible in the patients with low copy numbers in the second probelet, but may be distinct in the remaining patients (see 2240 in Figure 22). This may illustrate that the copy numbers listed in the second probelet correspond to the weights of the second tumor arraylet in the GBM profiles of the patients. (d) A chart 2030 is a plot of the 246th normal arraylet, which describes an X chromosome-exclusive amplification across the normal probes. (e) A chart 2035 shows a plot of the 246th probelet, which may be approximately common to both the normal and tumor data sets, and second most significant in the normal data set (see 2240 in Figure 22), may describe the corresponding copy-number

amplification in the female relative to the male patients. Classification of the patients by the 246th probelet may agree with the copy-number gender assignments (see table in Figure 34), also for three patients with missing TCGA gender annotations and three additional patients with conflicting TCGA annotations and copy-number gender assignments. (f) Chart 2040 is a raster display of the normal data set, which may show the correspondence between the normal blood profiles and the 246th probelet and normal arraylet. X chromosome amplification, which may be dominant in the 246th normal arraylet (Chart 2040), may be distinct in the female but nonexistent in the male patients (Figure Chart 2035). Note also that although the tumor samples exhibit female-specific X chromosome amplification (Chart 2020), the second tumor arraylet (Chart 2010) exhibits an unsegmented X chromosome copy-number distribution, that is approximately centered at zero with a relatively small width.

**[0094]** Figure 21 is a diagram illustrating survival analyses of three sets of patients classified by GSVD, age at diagnosis or both, according to some embodiments. (a) A graph 2110 shows Kaplan-Meier curves for the 247 patients with TCGA annotations in the initial set of 251 patients, classified by copy numbers in the second probelet, which is computed by GSVD for 251 patients, which may indicate a KM median survival time difference of nearly 16 months, with the corresponding log-rank test  $P$ -value  $<10^{-3}$ . The univariate Cox proportional hazard ratio is 2.3, with a  $P$ -value  $<10^{-2}$  (see table in Figure 34), which may suggest that high relative copy numbers in the second probelet confer more than twice the hazard of low numbers. (b) A graph 2120 shows KM and Cox survival analyses for the 247 patients classified by age, i.e.,  $>50$  or  $<50$  years old at diagnosis, which may indicate that the prognostic contribution of age, with a KM median survival time difference of nearly 11 months and a univariate hazard ratio of 2, is comparable to that of GSVD. (c) A graph 2130 shows Survival analyses for the 247 patients classified by both GSVD and age, which may indicate similar multivariate Cox hazard ratios, of 1.8 and 1.7, that do not differ significantly from the corresponding univariate hazard ratios, of 2.3 and 2, respectively. This may signify that GSVD and age may be independent prognostic predictors. With a KM median survival time difference of approximately 22 months, GSVD and age combined make a better predictor than age alone. (d) A graph 2140 shows Survival analyses for the 334 patients with TCGA annotations and a GSVD classification in the inclusive

confirmation set of 344 patients, classified by copy numbers in the second probelet, which is computed by GSVD for the 344 patients, which may indicate a KM median survival time difference of nearly 16 months and a univariate hazard ratio of 2.4, and confirm the survival analyses of the initial set of 251 patients. (e) A graph 2150 shows Survival analyses for the 334 patients classified by age confirm that the prognostic contribution of age, with a KM median survival time difference of approximately 10 months and a univariate hazard ratio of 2, is comparable to that of GSVD. (f) A graph 2160 shows Survival analyses for the 334 patients classified by both GSVD and age, which may indicate similar multivariate Cox hazard ratios, of 1.9 and 1.8, that may not differ significantly from the corresponding univariate hazard ratios, and a KM median survival time difference of nearly 22 months, with the corresponding log-rank test  $P$ -value  $<10^{-5}$ . This result may confirm that the prognostic contribution of GSVD is independent of age, and that combined with age, GSVD makes a better predictor than age alone. (g) A graph 2170 shows survival analyses for the 183 patients with a GSVD classification in the independent validation set of 184 patients, classified by correlations of each patient's GBM profile with the second tumor arraylet, which can be computed by GSVD for the 251 patients, which may indicate a KM median survival time difference of nearly 12 months and a univariate hazard ratio of 2.9, and may validate the survival analyses of the initial set of 251 patients. (h) A graph 2180 shows survival analyses for the 183 patients classified by age, which may validate that the prognostic contribution of age is comparable to that of GSVD. (i) A graph 2190 shows survival analyses for the 183 patients classified by both GSVD and age, which may indicate similar multivariate Cox hazard ratios, of 2 and 2.2, and a KM median survival time difference of nearly 41 months, with the corresponding log-rank test  $P$ -value  $<10^{-5}$ . This result may validate that the prognostic contribution of GSVD is independent of age, and that combined with age, GSVD may make a better predictor than age alone, also for patients with measured GBM aCGH profiles in the absence of matched normal blood profiles.

**[0095]** Figure 22 is a diagram illustrating most significant probelets in tumor and normal data sets, age at diagnosis or both, according to some embodiments. (a) Bar charts 2220 and 2240 show the ten significant probelets in the tumor data set and the generalized fraction that

each probelet captures in this data set. The generalized fraction are given as  $P_{1,n}$  and  $P_{2,n}$  below in terms of the normalized values for  $\sigma_{1,n}^2$  and  $\sigma_{2,n}^2$ :

$$P_{1,n} = \sigma_{1,n}^2 / \sum_{n=1}^N \sigma_{1,n}^2,$$

$$P_{2,n} = \sigma_{2,n}^2 / \sum_{n=1}^N \sigma_{2,n}^2.$$

The results shown in bar charts 2220 and 2240 may indicate that the two most tumor-exclusive probelets, i.e., the first probelet (see Figure 26) and the second probelet (see Figure 20, 2010-2020), with angular distances  $>2\pi/9$ , may also be the two most significant probelets in the tumor data set, with  $\sim 11\%$  and  $22\%$  of the information in this data set, respectively. The “generalized normalized Shannon entropy” (Equation 3 in Appendix A) of the tumor dataset is  $d_1=0.73$ . (b) Bar chart 2240 shows ten significant probelets in the normal data set and the generalized fraction that each probelet captures in this data set, which may indicate that the five most normal-exclusive probelets, the 247th to 251st probelets (see Figures 27-31), with angular distances approximately  $\approx -\pi/6$ , may be among the seven most significant probelets in the normal data set, capturing together  $\sim 56\%$  of the information in this data set. The 246th probelet (see Figure 20, 2030-2040), which is relatively common to the normal and tumor data sets with an angular distance  $>-\pi/6$ , may be the second most significant probelet in the normal data set with  $\sim 8\%$  of the information. The generalized entropy of the normal dataset,  $d_2=0.59$ , is smaller than that of the tumor dataset. This means that the normal dataset is more redundant and less complex than the tumor dataset.

**[0096]** Figure 23 is a diagram illustrating a survival analysis of an initial set of a number of patients classified by GBM-associated chromosome number changes, according to some embodiments. Graphs 2320- 2360 shown in Figure 23 are Kaplan-Meier survival analyses of the initial set of 251 patients classified by GBM-associated chromosome number changes. (a) The graph 2320 shows a Kaplan-Meier survival analysis for the 247 patients with TCGA annotations in the initial set of 251 patients, classified by number changes in chromosome 10, which may indicate almost overlapping KM curves with a KM median survival time difference

of  $\sim 2$  months, and a corresponding log-rank test  $P$ -value  $\sim 10^{-1}$ . This result may mean that chromosome 10 loss, frequently observed in GBM, is a poor predictor of GBM survival. (b) The graph 2340 depicts a KM survival analysis for the 247 patients classified by number changes in chromosome 7, which may indicate almost overlapping KM curves with a KM median survival time difference of  $< \text{one month}$ , and a corresponding log-rank test  $P$ -value  $> 5 \times 10^{-1}$ . This result may suggest that chromosome 7 gain is a poor predictor of GBM survival. (c) The graph 2360 shows a KM survival analysis for the 247 patients classified by number changes in chromosome 9p, which may indicate a KM median survival time difference of  $\sim 3$  months, and a log-rank test  $P$ -value  $> 10^{-1}$ . This result may signify that chromosome 9p loss is a poor predictor of GBM survival.

[0097] Figure 24 is a diagram illustrating a survival analysis of an initial set of a number of patients classified by copy number changes in selected segments, according to some embodiments. Graphs 2405-2460 show KM survival analyses of the initial set of 251 patients classified by copy number changes in selected segments containing GBM-associated genes or genes previously unrecognized in GBM. In the KM survival analyses for the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern, i.e., the second tumor-exclusive arraylet (Dataset S3), log-rank test  $P$ -values  $< 5 \times 10^{-2}$  are calculated for only 12 of the classifications. Of these, only six may correspond to a KM median survival time difference that is  $\approx > 5$  months, approximately a third of the  $\sim 16$  months difference observed for the GSVD classification. One of these segments may contain the genes TLK2 and METTL2A, previously unrecognized in GBM. The KM median survival time can be calculated for the 56 patients with TLK2 amplification, which is  $\sim 5$  months longer than that for the remaining patients. This may suggest that drug-targeting the kinase and/or the methyltransferase-like protein that TLK2 and METTL2A encode, respectively, may affect not only the pathogenesis but also the prognosis of GBM.

[0098] Figure 25 is a diagram illustrating a survival analysis of an initial set of a number of patients, according to some embodiments. Graph 2500 shows a result of a KM survival analysis of an initial set of 251 patients classified by a mutation in the gene *IDH1*.

**[0099]** Figure 26 is a diagram illustrating a significant probelet and corresponding tumor arraylet, according to some embodiments. This probelet may be the first most tumor-exclusive probelet, which is shown with corresponding tumor arraylet uncovered by GSVD of the patient-matched GBM and normal blood aCGH profiles. (a) A plot 2620 of the first tumor arraylet describes unsegmented chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (b) A graph 2630 of the first most tumor-exclusive probelet, which is also the second most significant probelet in the tumor data set (see 2220 in Figure 22), describes the corresponding variation across the patients. The patients are ordered according to each patient's relative copy number in this probelet. These copy numbers may significantly correlate with the genomic center where the GBM samples were hybridized at, HMS, MSKCC, or multiple locations, with the  $P$ -values  $<10^{-5}$  (see Table in Figure 35 and Figure 32). (c) A raster display 2640 of the tumor data set, with relative gain, no change, and loss of DNA copy numbers, may indicate the correspondence between the GBM profiles and the first probelet and tumor arraylet.

**[00100]** Figure 27 is a diagram illustrating a normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments. The normal-exclusive probelet is 247th, normal-exclusive probelet and corresponding normal arraylet is uncovered by GSVD. (a) A plot 2720 of the 247th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. The normal probes are ordered, and their copy numbers are colored, according to each probe's chromosomal location. (b) A plot 2730 of the 247th probelet may describe the corresponding variation across the patients. Copy numbers in this probelet may correlate with the date of hybridization of the normal blood samples, 7.22.2009, 10.8.2009, or other, with the  $P$ -values  $<10^{-3}$  (see the Table in Figure 35 and Figure 32). (c) A raster display 2740 of the normal data set shows the correspondence between the normal blood profiles and the 247th probelet and normal arraylet.

**[00101]** Figure 28 is a diagram illustrating a normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments. The

normal-exclusive probelet is 248th, normal-exclusive probelet and the corresponding normal arraylet is uncovered by GSVD. (a) A Plot 2820 of the 248th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) A Plot 2830 of the 248th probelet describes the corresponding variation across the patients. Copy numbers in this probelet may significantly correlate with the tissue batch/hybridization scanner of the normal blood samples, HMS 8/2331 and other, with the  $P$ -values  $<10^{-12}$  (see the Table in Figure 35 and Figure 32). (c) A raster display 2840 of the normal data set may show the correspondence between the normal blood profiles and the 248th probelet and normal arraylet.

**[00102]** Figure 29 is a diagram illustrating another normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments. The normal-exclusive probelet is 249th, normal-exclusive probelet and the corresponding normal arraylet is uncovered by GSVD. (a) A Plot 2920 of the 249th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) A Plot 2930 of the 249th probelet describes the corresponding variation across the patients. Copy numbers in this probelet may significantly correlate with the tissue batch/hybridization scanner of the normal blood samples, HMS 8/2331 and other, with the  $P$ -values  $<10^{-12}$  (see the Table in Figure 35 and Figure 32). (c) A raster display 2940 of the normal data set may show the correspondence between the normal blood profiles and the 249th probelet and normal arraylet.

**[00103]** Figure 30 is a diagram illustrating yet another normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments. The normal-exclusive probelet is 250th, normal-exclusive probelet and the corresponding normal arraylet is uncovered by GSVD. (a) A Plot 3020 of the 250th normal arraylet describes copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) A Plot 3030 of the 248th probelet may describe the corresponding variation across the patients. Copy numbers in this probelet may correlate with the date of hybridization of the normal blood samples, 4.18.2007, 7.22.2009, or other, with the  $P$ -values  $<10^{-3}$  (see the Table in Figure 35 and Figure 32). (c) A raster display 3040 of the normal

data set may show the correspondence between the normal blood profiles and the 250th probelet and normal arraylet.

**[00104]** Figure 31 is a diagram illustrating a first most normal-exclusive probelet and corresponding normal arraylet uncovered by GSVD, according to some embodiments. The normal-exclusive probelet is 251st, normal-exclusive probelet and the corresponding normal arraylet is uncovered by GSVD. (a) A Plot 3120 of the 251st normal arraylet describes unsegmented chromosomes (black lines), each with copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths. (b) A Plot 3130 Plot of the first most normal-exclusive probelet, which may also be the most significant probelet in the normal data set (see Figure 22, 2240), describes the corresponding variation across the patients. Copy numbers in this probelet may significantly correlate with the genomic center where the normal blood samples were hybridized at, HMS, MSKCC, or multiple locations, with the  $P$ -values  $<10^{-13}$  (see the Table in Figure 35 and Figure 32). (c) A raster display 3140 of the normal data set may show the correspondence between the normal blood profiles and the 251st probelet and normal arraylet.

**[00105]** Figure 32 is a diagram illustrating differences in copy numbers among the TCGA annotations associated with the significant probelets, according to some embodiments. Boxplot visualization of the distribution of copy numbers are shown of the (a) first, possibly the most tumor-exclusive probelet among the associated genomic centers where the GBM samples were hybridized at (Table in Figure 35); (b) 247th, normal-exclusive probelet among the dates of hybridization of the normal blood samples; (c) 248th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (d) 249th, normal-exclusive probelet between the associated tissue batches/hybridization scanners of the normal samples; (e) 250th, normal-exclusive probelet among the dates of hybridization of the normal blood samples; (f) 251st, possibly the most normal-exclusive probelet among the associated genomic centers where the normal blood samples were hybridized at. The Mann-Whitney-Wilcoxon  $P$ -values correspond to the two annotations that may be associated with largest or smallest relative copy numbers in each probelet.

**[00106]** Figure 33 is a diagram illustrating copy-number distributions of one of the probelet and the corresponding normal arraylet and tumor arraylet, according to some embodiments. Copy-number distributions relates to the 246th probelet and the corresponding 246th normal arraylet and 246th tumor arraylet. Boxplot visualization and Mann-Whitney-Wilcoxon  $P$ -values of the distribution of copy numbers are shown of the (a) 246th probelet, which may be approximately common to both the normal and tumor data sets, and may be the second most significant in the normal data set (see Figure 22, 2240), between the gender annotations (Table in Figure 35); (b) 246th normal arraylet between the autosomal and X chromosome normal probes; (c) 246th tumor arraylet between the autosomal and X chromosome tumor probes.

**[00107]** Figure 34 is a table illustrating proportional hazard models of three sets of patients classified by GSVD, according to some embodiments. The Cox proportional hazard models of the three sets of patients are classified by GSVD, age at diagnosis or both. In each set of patients, the multivariate Cox proportional hazard ratios for GSVD and age may be similar and may not differ significantly from the corresponding univariate hazard ratios. This may indicate that GSVD and age are independent prognostic predictors.

**[00108]** Figure 35 is a table illustrating enrichment of significant probelets in TCGA annotations, according to some embodiments. Probabilistic significance of the enrichment of the  $n$  patients, that may correspond to the largest or smallest relative copy numbers in each significant probelet, in the respective TCGA annotations are shown. The  $P$ -value of each enrichment can be calculated assuming hypergeometric probability distribution of the  $K$  annotations among  $N=251$  patients of the initial set, and of the subset of  $k \subseteq K$  annotations among the subset of  $n$  patients, as described by:

$$P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^n \binom{K}{i} \binom{N-K}{n-i}.$$

**[00109]** Figure 36 is a diagram illustrating HO GSVD of biological data related to patient and normal samples, according to some embodiments. It shows the generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles.

The structure of the patient-matched but probe-independent tumor and normal datasets  $D_1$  and  $D_2$ , of the initial set of  $N=251$  patients, i.e.,  $N$ -arrays  $\times M_1=212,696$ -tumor probes and  $M_2=211,227$ -normal probes, is of an order higher than that of a single matrix. The patients, the tumor and normal probes as well as the tissue types, each represent a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost. The GSVD simultaneously separated the paired data sets into paired weighted sums of  $N$  outer products of two patterns each: one pattern of copy-number variation across the patients, i.e., a “probelet”  $v_n^T$  (e.g., a row of right basis array), which is identical for both the tumor and normal data sets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the “tumor arraylet”  $u_{1,n}$  (e.g., vectors of array  $U_1$  of left basis arrays) or the corresponding normal-specific pattern across the normal probes, i.e., the “normal arraylet”  $u_{2,n}$  (e.g., vectors of array  $U_2$  of left basis arrays). This can be depicted in a raster display, with relative copy-number gain, no change, and loss, explicitly showing the first through the 10th and the 242nd through the 251st probelets and corresponding tumor and normal arraylets, which may capture approximately 52% and 71% of the information in the tumor and normal data set, respectively.

[00110] The significance of the probelet  $v_n^T$  (e.g., rows of right basis array) in the tumor data set (e.g.,  $D_1$  of the 3-D array) relative to its significance in the normal data set (e.g.,  $D_2$  of the 3-D array) is defined in terms of an “angular distance” that is proportional to the ratio of these weights, as shown in the following expression:

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

[00111] This significance is depicted in a bar chart display, showing that the first and second probelets are almost exclusive to the tumor data set with angular distances  $>2\pi/9$ , the 247th to 251st probelets are approximately exclusive to the normal data set with angular distances  $<\approx \pi/6$ , and the 246th probelet is relatively common to the normal and tumor data sets with an angular distance  $>-\pi/6$ . It may be found and confirmed that the second most tumor-exclusive probelet, the most significant probelet in the tumor data set, significantly correlates with GBM prognosis. The corresponding tumor arraylet describes a global pattern of tumor-

exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs, including the biochemically putative drug target-encoding *TLK2*. It can also be found and validated that a negligible weight of the global pattern in a patient's GBM aCGH profile is indicative of a significantly longer GBM survival time. It was shown that the GSVD provides a mathematical framework for comparative modeling of DNA microarray data from two organisms. Recent experimental results verify a computationally predicted genomewide mode of regulation, and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms. The GSVD comparative modeling of aCGH data from patient-matched tumor and normal samples draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

[00112] The foregoing description, for purpose of explanation, has been described with reference to specific embodiments. However, the illustrative discussions above are not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The embodiments were chosen and described in order to best explain the principles of the invention and its practical applications, to thereby enable others skilled in the art to best utilize the invention and various embodiments with various modifications as are suited to the particular use contemplated.

[00113] All publications and patents, and NCBI gene ID sequences cited in this disclosure are incorporated by reference in their entirety. To the extent the material incorporated by reference contradicts or is inconsistent with this specification, the specification will supersede any such material. The citation of any references herein is not an admission that such references are prior art to the present invention.

[00114] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following embodiments.

# Appendix

## A

# GSVD Comparison of Patient-Matched Normal and Tumor aCGH Profiles Reveals Global Copy-Number Alterations Predicting Glioblastoma Multiforme Survival

Cheng H. Lee<sup>1#</sup>, Benjamin O. Alpert<sup>1,2#</sup>, Preethi Sankaranarayanan<sup>1,2</sup>, and Orly Alter<sup>1,2,3\*</sup>

**1** Scientific Computing and Imaging (SCI) Institute,

**2** Department of Bioengineering and

**3** Department of Human Genetics, University of Utah, Salt Lake City, UT 84112, United States of America

\* E-mail: orly@sci.utah.edu

# These authors contributed equally to this work.

## Abstract

Despite recent large-scale profiling efforts, the best prognostic predictor of glioblastoma multiforme (GBM) remains the patient's age at diagnosis. We describe a global pattern of tumor-exclusive co-occurring copy-number alterations (CNAs) that is correlated, possibly coordinated with GBM patients' survival and response to chemotherapy. The pattern is revealed by GSVD comparison of patient-matched but probe-independent GBM and normal aCGH datasets from The Cancer Genome Atlas (TCGA). We find that, first, the GSVD, formulated as a framework for comparatively modeling two composite datasets, removes from the pattern copy-number variations (CNVs) that occur in the normal human genome (e.g., female-specific X chromosome amplification) and experimental variations (e.g., in tissue batch, genomic center, hybridization date and scanner), without a-priori knowledge of these variations. Second, 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. These include the biochemically putative drug target, cell cycle-regulated serine/threonine kinase-encoding *TLK2*, the cyclin E1-encoding *CCNE1*, and the Rb-binding histone demethylase-encoding *KDM5A*. Third, the pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone. GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set, confirms the global pattern. GSVD classification of the GBM profiles of an independent set of patients validates the prognostic contribution of the pattern.

## Introduction

Glioblastoma multiforme (GBM), the most common brain tumor in adults, is characterized by poor prognosis [1]. GBM tumors exhibit a range of copy-number alterations (CNAs), many of which play roles in the cancer's pathogenesis [2–4]. Recent large-scale gene expression [5–7] and DNA methylation [8] profiling efforts identified GBM molecular subtypes, distinguished by small numbers of biomarkers. However, despite these efforts, GBM's best prognostic predictor remains the patient's age at diagnosis [9, 10].

To identify CNAs that might predict GBM patients' survival, we comparatively model patient-matched GBM and normal array CGH (aCGH) profiles from The Cancer Genome Atlas (TCGA) by using the generalized singular value decomposition (GSVD) [11]. Previously, we formulated the GSVD as a framework for comparatively modeling two composite datasets [12] (see also [13]), and illustrated its application in sequence-independent comparison of DNA microarray data from two organisms, where, as we showed, the mathematical variables and operations of the GSVD represent experimental or biological reality. The variables, subspaces of significant patterns that are uncovered in the simultaneous decomposition of the two datasets and are mathematically significant in either both (i.e., common to

both) datasets or only one (i.e., exclusive to one) of the datasets, correlate with cellular programs that are either conserved in both or unique to only one of the organisms, respectively. The operation of reconstruction in the subspaces that are mathematically common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism.

We now find that also in probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared tumor and normal patterns of copy-number variation across the patients and the corresponding tumor- and normal-specific patterns of copy-number variation across the tumor and normal probes, represent experimental or biological reality. Patterns that are mathematically significant in both datasets represent copy-number variations (CNVs) in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification). Patterns that are mathematically significant in the normal but not the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some patterns that are mathematically significant in the tumor but not the normal dataset represent experimental variations that exclusively affect the tumor dataset.

One pattern, that is mathematically significant in the tumor but not the normal dataset, represents tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs in >3% of the patients [14]. This pattern is correlated, possibly coordinated with GBM patients' survival and response to therapy. We find that the pattern provides a prognostic predictor that is better than the chromosome numbers or any one focal CNA that it identifies, suggesting that the GBM survival phenotype is an outcome of its global genotype. The pattern is independent of age, and combined with age, makes a better predictor than age alone.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [15].

## Methods

To compare TCGA patient-matched GBM and normal (mostly blood) aCGH profiles (Dataset S1 and Mathematica Notebooks S1 and S2), Agilent Human aCGH 244A-measured 365 tumor and 360 normal profiles were selected, corresponding to the same  $N=251$  patients. Each profile lists  $\log_2$  of the TCGA level 1 background-subtracted intensity in the sample relative to the Promega DNA reference, with signal to background >2.5 for both the sample and reference in more than 90% of the 223,603 autosomal probes on the microarray. The profiles are organized in one tumor and one normal dataset, of  $M_1=212,696$  and  $M_2=211,227$  autosomal and X chromosome probes, each probe with valid data in at least 99% of either the tumor or normal arrays, respectively. Each profile is centered at its autosomal median copy number. The <0.2% missing data entries in the tumor and normal datasets are estimated by using singular value decomposition (SVD) as described [12, 16]. Within each set, the medians of profiles of samples from the same patient are taken.

The structure of the patient-matched but probe-independent tumor and normal datasets  $D_1$  and  $D_2$ , of  $N$  patients, i.e.,  $N$ -arrays  $\times M_1$ -tumor and  $M_2$ -normal probes, is of an order higher than that of a single matrix. The patients, the tumor and normal probes as well as the tissue types, each represent a degree of freedom. Unfolded into a single matrix, some of the degrees of freedom are lost and much of the information in the datasets might also be lost.

To compare the tumor and normal datasets, therefore, we use the GSVD, formulated to simultaneously separate the paired datasets into paired weighted sums of  $N$  outer products of two patterns each: One

pattern of copy-number variation across the patients, i.e., a “probelet”  $v_n^T$ , which is identical for both the tumor and normal datasets, combined with either the corresponding tumor-specific pattern of copy-number variation across the tumor probes, i.e., the “tumor arraylet”  $u_{1,n}$ , or the corresponding normal-specific pattern across the normal probes, i.e., the “normal arraylet”  $u_{2,n}$  (Figure 36),

$$\begin{aligned} D_1 &= U_1 \Sigma_1 V^T = \sum_{n=1}^N \sigma_{1,n} u_{1,n} \otimes v_n^T, \\ D_2 &= U_2 \Sigma_2 V^T = \sum_{n=1}^N \sigma_{2,n} u_{2,n} \otimes v_n^T. \end{aligned} \quad (1)$$

The probelets are, in general, non-orthonormal, but are normalized, such that  $v_n^T v_n = 1$ . The tumor and normal arraylets are orthonormal, such that  $U_1^T U_1 = U_2^T U_2 = I$ .

The significance of the probelet  $v_n^T$  in either the tumor or normal dataset, in terms of the overall information that it captures in this dataset, is proportional to either of the weights  $\sigma_{1,n}$  or  $\sigma_{2,n}$ , respectively (Figure 22),

$$\begin{aligned} p_{1,n} &= \sigma_{1,n}^2 / \sum_{n=1}^N \sigma_{1,n}^2, \\ p_{2,n} &= \sigma_{2,n}^2 / \sum_{n=1}^N \sigma_{2,n}^2. \end{aligned} \quad (2)$$

The “generalized normalized Shannon entropy” of each dataset,

$$\begin{aligned} 0 \leq d_1 &= (\log N)^{-1} \sum_{n=1}^N p_{1,n} \log p_{1,n} \leq 1, \\ 0 \leq d_2 &= (\log N)^{-1} \sum_{n=1}^N p_{2,n} \log p_{2,n} \leq 1, \end{aligned} \quad (3)$$

measures the complexity of the data from the distribution of the overall information among the different probelets and corresponding arraylets. An entropy of zero corresponds to an ordered and redundant dataset in which all the information is captured by a single probelet and its corresponding arraylet. An entropy of one corresponds to a disordered and random dataset in which all probelets and arraylets are of equal significance. The significance of the probelet  $v_n^T$  in the tumor dataset relative to its significance in the normal dataset is defined in terms of an “angular distance”  $\theta_n$  that is proportional to the ratio of these weights,

$$-\pi/4 \leq \theta_n = \arctan(\sigma_{1,n}/\sigma_{2,n}) - \pi/4 \leq \pi/4. \quad (4)$$

An angular distance of  $\pm\pi/4$  indicates a probelet that is exclusive to either the tumor or normal dataset, respectively, whereas an angular distance of zero indicates a probelet that is common to both the tumor and normal datasets. The probelets are arranged in decreasing order of their angular distances, i.e., their significance in the tumor dataset relative to the normal dataset.

We find that the two most tumor-exclusive mathematical patterns of copy-number variation across the patients, i.e., the first probelet (Figure 26) and the second probelet (Figure 20 a-c), with angular distances  $> 2\pi/9$ , are also the two most significant probelets in the tumor dataset, with  $\sim 11\%$  and  $22\%$  of the information in this dataset, respectively. Similarly, the five most normal-exclusive probelets, the 247th to 251st probelets (Figures 27-31), with angular distances  $\lesssim -\pi/6$ , are among the seven most significant probelets in the normal dataset, capturing together  $\sim 56\%$  of the information in this dataset.

The 246th probelet (Figure 20 *d-f*), which is the second most significant probelet in the normal dataset with  $\sim 8\%$  of the information, is relatively common to the normal and tumor datasets with an angular distance  $> -\pi/6$ .

To biologically or experimentally interpret these significant probelets, we correlate or anticorrelate each probelet with relative copy-number gain or loss across a group of patients according to the TCGA annotations of the group of  $n$  patients with largest or smallest relative copy numbers in this probelet among all  $N$  patients, respectively. The  $P$ -value of a given association is calculated assuming hypergeometric probability distribution of the  $K$  annotations among the  $N$  patients, and of the subset of  $k \subseteq K$  annotations among the subset of  $n$  patients, as described [17],  $P(k; n, N, K) = \binom{N}{n}^{-1} \sum_{i=k}^n \binom{K}{i} \binom{N-K}{n-i}$  (Table 1). We visualize the copy-number distribution between the annotations that are associated with largest or smallest relative copy numbers in each probelet by using boxplots, and by calculating the corresponding Mann-Whitney-Wilcoxon  $P$ -value (Figures 32-33). To interpret the corresponding tumor and normal arraylets, we map the tumor and normal probes onto the National Center for Biotechnology Information (NCBI) human genome sequence build 36, by using the Agilent Technologies probe annotations posted at the University of California at Santa Cruz (UCSC) human genome browser [18, 19]. We segment each arraylet and assign each segment a  $P$ -value by using the circular binary segmentation (CBS) algorithm as described (Dataset S2) [20, 21]. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to variations in the preprocessing of the data, e.g., in the data selection cutoffs.

## Results

We find that, first, the GSVD identifies significant experimental variations that exclusively affect either the tumor or the normal dataset, as well as CNVs that occur in the normal human genome and are common to both datasets, without a-priori knowledge of these variations. The mathematically most tumor-exclusive probelet, i.e., the first probelet (Figure 26), correlates with tumor-exclusive experimental variation in the genomic center where the GBM samples were hybridized at, with the  $P$ -values  $< 10^{-5}$  (Figures 35 and 32). Similarly, the five most normal-exclusive probelets, i.e., the 247th to 251st probelets (Figures 27-31), correlate with experimental variations among the normal samples in genomic center, DNA microarray hybridization or scan date as well as the tissue batch and hybridization scanner, with  $P$ -values  $< 10^{-3}$ . Consistently, the corresponding arraylets, i.e., the first tumor arraylet and the 247th to 251st normal arraylets, describe copy-number distributions which are approximately centered at zero with relatively large, chromosome-invariant widths.

The 246th probelet (Figure 20*e*), which is mathematically approximately common to both the normal and tumor datasets, describes copy-number amplification in the female relative to the male patients that is biologically common to both the normal and tumor datasets. Consistently, both the 246th normal arraylet (Figure 20*d*) and 246th tumor arraylet describe an X chromosome-exclusive amplification. The  $P$ -values are  $< 10^{-38}$  (Table 1 and Figure 33). To assign the patients gender, we calculate for each patient the standard deviation of the mean X chromosome number from the autosomal genomic mean in the patient's normal profile (Figure 20*f*). Patients with X chromosome amplification greater than twice the standard deviation are assigned the female gender. For three of the patients, this copy-number gender assignment conflicts with the TCGA gender annotation. For three additional patients, the TCGA gender annotation is missing. In all these cases, the classification of the patients by the 246th probelet agrees with the copy-number assignment.

Second, we find that the GSVD identifies a global pattern of tumor-exclusive co-occurring CNAs that includes most known GBM-associated changes in chromosome numbers and focal CNAs. This global pattern is described by the second tumor arraylet (Figure 20*a* and Dataset S3). The second most tumor-exclusive probelet (Figure 20*b*), which describes the corresponding copy-number variation across the patients, is the most significant probelet in the tumor dataset. Dominant in the global pattern,

and frequently observed in GBM samples [2], is a co-occurrence of a gain of chromosome 7 and losses of chromosome 10 and the short arm of chromosome 9 (9p). To assign a chromosome gain or loss, we calculate for each tumor profile the standard deviation of the mean chromosome number from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10. The gain of chromosome 7 and the losses of chromosomes 10 and 9p are greater than twice the standard deviation in the global pattern as well as the tumor profiles of ~20%, 41% and 12% of the patients, respectively.

Focal CNAs that are known to play roles in the origination and development of GBM and are described by the global pattern include amplifications of segments containing the genes *MDM4* (1q32.1), *AKT3* (1q44), *EGFR* (7p11.2), *MET* (7q31.2), *CDK4* (12q14.1) and *MDM2* (12q15), and deletions of segments containing the genes *CDKN2A/B* (9p21.3) and *PTEN* (10q23.31), that occur in >3% of the patients. To assign a CNA in a segment, we calculate for each tumor profile the mean segment copy number. Profiles with segment amplification or deletion greater than twice the standard deviation from the autosomal genomic mean, excluding the outlying chromosomes 7, 9p and 10, or greater than one standard deviation from the chromosomal mean, when this deviation is consistent with the deviation from the genomic mean, are assigned a segment gain or loss, respectively. The frequencies of amplification or deletion we observe for these segments are similar to the reported frequencies of the corresponding focal CNAs [4].

Novel CNAs, previously unrecognized in GBM, are also revealed by the global pattern [14]. These include an amplification of a segment that contains *TLK2* (17q23.2) in ~22% of the patients, with the corresponding CBS  $P$ -value <  $10^{-140}$ . Copy-number amplification of *TLK2* has been correlated with overexpression in several other cancers [22, 23]. The human gene *TLK2*, with homologs in the plant *Arabidopsis thaliana* but not in the yeast *Saccharomyces cerevisiae*, encodes for a multicellular organisms-specific serine/threonine protein kinase, a biochemically putative drug target [24], which activity is directly dependent on ongoing DNA replication [25]. On the same segment with *TLK2*, we also find the gene *METTL2A*. Another amplified segment (CBS  $P$ -value <  $10^{-13}$ ) contains the homologous gene *METTL2B* (7q32.1). Overexpression of *METTL2A/B* was linked with prostate cancer metastasis [26], cAMP response element-binding (CREB) regulation in myeloid leukemia [27], and breast cancer patients' response to chemotherapy [28].

An amplification of a segment (CBS  $P$ -value <  $10^{-145}$ ) encompassing the cyclin E1-encoding *CCNE1* (19q12) is revealed in ~4% of the patients. Cyclin E1 regulates entry into the DNA synthesis phase of the cell division cycle. Copy number increases of *CCNE1* have been linked with multiple cancers [29, 30], but not GBM. Amplicon-dependent expression of *CCNE1*, together with the genes *POP4*, *PLEKHF1*, *C19orf12* and *C19orf2* that flank *CCNE1* on this segment, was linked with primary treatment failure in ovarian cancer, possibly due to rapid repopulation of the tumor after chemotherapy [31].

Another rare amplification in ~4% of the patients, of a segment (CBS  $P$ -value <  $10^{-28}$ ) that overlaps with the 5' end of *KDM5A* (12p13.33), is also revealed. The protein encoded by *KDM5A*, a retinoblastoma tumor suppressor (Rb)-binding lysine-specific histone demethylase [32], has been recently implicated in cancer drug tolerance [33]. The same amplified segment includes the solute carrier (SLC) sodium-neurotransmitter symporters *SLC6A12/13*, biochemically putative carriers of drugs that might overcome the blood-brain barrier [34]. On the same segment we also find *IQSEC3*, a mature neuron-specific guanine nucleotide exchange factor (GEF) for the ADP-ribosylation factor (ARF) *ARF1*, a key regulator of intracellular membrane traffic [35].

Note that although the tumor samples exhibit female-specific X chromosome amplification (Figure 20c), the second tumor arraylet exhibits an unsegmented X chromosome copy-number distribution, that is approximately centered at zero with a relatively small width. This illustrates the mathematical separation of the global pattern of tumor-exclusive co-occurring CNAs, that is described by the second tumor arraylet, from all other biological and experimental variations that compose either the tumor or the normal dataset, such as the gender variation that is common to both datasets, and is described by the 246th probelet and the corresponding 246th tumor and 246th normal arraylets.

Third, we find that the GSVD classifies the patients into two groups of significantly different prognoses.

The classification is according to the copy numbers listed in the second probelet, which correspond to the weights of the second tumor arraylet in the GBM aCGH profiles of the patients. A group of 227 patients, 224 of which with TCGA annotations, displays high ( $>0.02$ ) relative copy numbers in the second probelet, and a Kaplan-Meier (KM) [36] median survival time of  $\sim 13$  months (Figure 21a). A group of 23 patients, i.e.,  $\sim 10\%$  of the patients, displays low, approximately zero, relative copy numbers in the second probelet, and a KM median survival time of  $\sim 29$  months, which is more than twice longer than that of the previous group. The corresponding log-rank test  $P$ -value is  $< 10^{-3}$ . The univariate Cox [37] proportional hazard ratio is 2.3, with a  $P$ -value  $< 10^{-2}$  (Table S1 in Appendix S1), meaning that high relative copy numbers in the second probelet confer more than twice the hazard of low numbers. Note that the cutoff of  $\pm 0.02$  was selected to enable classification of as many of the patients as possible. Only one of the 251 patients has a negative copy number in the second probelet  $< -0.02$ , and remains unclassified. This patient is also missing the TCGA annotations. Survival analysis of only the chemotherapy patients classified by GSVD gives similar results (Figures 40 and 41a). The  $P$ -values are calculated without adjusting for multiple comparisons [38]. We observe, therefore, that a negligible weight of the global pattern in a patient's GBM aCGH profile is indicative of a significantly longer survival time, as well as an improved response to treatment among chemotherapy patients.

A mutation in the gene *IDH1* was recently linked with improved GBM prognosis [1,6] and associated with a CpG island methylator phenotype [8]. We find, however, only seven patients (six chemotherapy patients), i.e.,  $< 3\%$ , with *IDH1* mutation. This is less than a third of the 23 patients in the long-term survival group defined by the global pattern. The corresponding survival analyses are, therefore, statistically insignificant (Figures 25 and 42).

Chromosome 10 loss, chromosome 7 gain and even loss of 9p, which are dominant in the global pattern, have been suggested as indicators of poorer GBM prognoses for over two decades [2,3]. However, the KM survival curves for the groups of patients with either one of these chromosome number changes almost overlap the curves for the patients with no changes (Figure 23). The log-rank test  $P$ -values for all three classifications are  $\gtrsim 10^{-1}$ , with the median survival time differences  $\lesssim 3$  months. Similarly, in the KM survival analyses of the groups of patients with either a CNA or no CNA in either one of the 130 segments identified by the global pattern (Figure 24), log-rank test  $P$ -values  $< 5 \times 10^{-2}$  are calculated for only 12 of the classifications. Of these, only six correspond to a KM median survival time difference that is  $\gtrsim 5$  months, approximately a third of the  $\sim 16$  months difference observed for the GSVD classification.

One of these segments contains the genes *TLK2* and *METTL2A* and another segment contains the homologous gene *METTL2B*, previously unrecognized in GBM. The KM median survival times we calculate for the 56 patients with *TLK2/METTL2A* amplification and, separately, for the 19 patients with *METTL2B* amplifications are  $\sim 5$  and 8 months longer than that for the remaining patients in each case. Similarly, the KM median survival times we calculate for the 43 chemotherapy patients with *TLK2/METTL2A* amplification and, separately, for the 15 chemotherapy patients with *METTL2B* amplification, are both  $\sim 7$  months longer than that for the remaining chemotherapy patients in each case (Figure 43). This suggests that drug-targeting the kinase that *TLK2* encodes and/or the methyltransferase-like proteins that *METTL2A/B* encode may affect not only the pathogenesis but also the prognosis of GBM as well as the patient's response to chemotherapy.

Taken together, we find that the global pattern provides a better prognostic predictor than the chromosome numbers or any one focal CNA that it identifies. This suggests that the GBM survival phenotype is an outcome of its global genotype.

Despite the recent genome-scale molecular profiling efforts, age at diagnosis remains the best prognostic predictor for GBM in clinical use. The KM median survival time difference between the patients  $> 50$  or  $< 50$  years old at diagnosis is  $\sim 11$  months, approximately two thirds of the  $\sim 16$  months difference observed for the global pattern, with the log-rank test  $P$ -value  $< 10^{-4}$  (Figure 21b). The univariate Cox proportional hazard ratio we calculate for age is 2, i.e., similar to that for the global pattern. Taken together, the prognostic contribution of the global pattern is comparable to that of age. Similarly we

find that the prognostic contribution of the global pattern is comparable to that of chemotherapy (Figure 10a).

To examine whether the weight of the global pattern in a patient's GBM aCGH profile is correlated with the patient's age at diagnosis, we classify the patients into four groups, with prognosis of longer-term survival according to both, only one or neither of the classifications (Figure 21c). The KM curves for these four groups are significantly different, with the log-rank test  $P$ -value  $< 10^{-4}$ . Within each age group, the subgroup of patients with low relative copy numbers in the second probelet consistently exhibits longer survival than the remaining patients. The median survival time of the 16 patients  $< 50$  years old at diagnosis with low copy numbers in the second probelet is  $\sim 34$  months, almost three times longer than the  $\sim 12$  months median survival time of the patients  $> 50$  years old at diagnosis with high numbers in the second probelet. The multivariate Cox proportional hazard ratios for the global pattern and age are 1.8 and 1.7, respectively, with both corresponding  $P$ -values  $< 3 \times 10^{-2}$ . These ratios are similar, meaning that both a high weight of the global pattern in a patient's GBM aCGH profile and an age  $> 50$  years old at diagnosis confer similar relative hazard. These ratios also do not differ significantly from the univariate ratios of 2.3 and 2 for the global pattern and age, respectively. Taken together, the prognostic contribution of the global pattern is not only comparable to that of age, but is also independent of age. Combined with age, the global pattern makes a better predictor than age alone. Similarly, we find that the global pattern is independent of chemotherapy (Figure 10b).

To confirm the global pattern, we use GSVD to compare matched profiles of a larger, more recent, set of 344 TCGA patients, that is inclusive of the initial set of 251 patients [15]. Agilent Human aCGH 244A-measured 458 tumor and 459 normal profiles were selected, corresponding to the inclusive confirmation set of  $N=344$  patients (Dataset S4). The profiles, centered at their autosomal median copy numbers, are organized in one tumor and one normal dataset, of  $M_1=200,139$  and  $M_2=198,342$  probes, respectively. Within each set, the medians of profiles of samples from the same patient are taken after estimating missing data by using SVD. We find that the significant probelets and corresponding tumor and normal arraylets, as well as their interpretations, are robust to the increase from 251 patients in the initial set to 344 patients in the inclusive confirmation set, and the accompanying decreases in tumor and normal probes, respectively.

The second tumor arraylet computed by GSVD for the 344 patients of the inclusive confirmation set correlates with that of the initial set, with the correlation  $\sim 0.99$ . To classify the patients according to the copy numbers listed in the corresponding second probelet of the inclusive confirmation set, the classification cutoff  $\pm 0.02$  of the initial set of 251 patients is scaled by the norm of the copy numbers listed for these patients, resulting in the cutoff  $\pm 0.017$ . Only four of the 251 patients in the initial set, i.e.,  $\sim 1.5\%$ , with copy numbers that are near the classification cutoffs of both sets, change classification. Of the 344 patients, we find that 315 patients, 309 with TCGA annotations, display high ( $> 0.017$ ) and 27, i.e.,  $\sim 8\%$ , display low, approximately zero, relative copy numbers in the second probelet. Only two patients, one missing TCGA annotations, remain unclassified with large negative ( $< -0.017$ ) copy numbers in the second probelet. Survival analyses of the inclusive confirmation set of 344 patients give qualitatively the same results as these of the initial set of 251 patients. These analyses confirm that a negligible weight of the global pattern, which is described by the second tumor arraylet, i.e., a low copy number in the second probelet, is indicative of a significantly longer survival time (Figure 21d). Survival analysis of only the chemotherapy patients in the inclusive confirmation set classified by GSVD gives similar results (Figure 41b). These analyses confirm that the prognostic contribution of the global pattern is comparable to that of age (Figure 21e) and is independent of age (Figure 21f). Similarly, we confirm that the global pattern is independent of chemotherapy (Figures 10 c and d in Appendix S1).

To validate the prognostic contribution of the global pattern, we classify GBM profiles of an independent set of 184 TCGA patients, that is mutually exclusive of the initial set of 251 patients. Agilent Human aCGH 244A-measured 280 tumor profiles were selected, corresponding to the independent validation set of 184 patients with available TCGA status annotations (Dataset S5). Each profile lists relative copy

numbers in more than 97.5% of the 206,820 autosomal probes among the  $M_1=212,696$  probes that define the second tumor arraylet computed by GSVD for the 251 patients of the initial set. Medians of profiles of samples from the same patient are taken. To classify the 184 patients according to the correlations of their GBM profiles with the second tumor arraylet of the initial set, the classification cutoff of the initial set of 251 patients is scaled by the norm of the correlations calculated for these patients, resulting in the cutoff  $\pm 0.15$ . For the profiles of 162 patients we calculate high ( $>0.15$ ) and for 21, i.e.,  $\sim 11\%$ , low, approximately zero, correlation with the second tumor arraylet. One patient remains unclassified with a large negative ( $<-0.15$ ) correlation.

We find that survival analyses of the independent validation set of 184 patients give qualitatively the same results as these of the initial set of 251 patients and the inclusive confirmation set of 344 patients (Figures 21 *g-i* and Figures 41 *c*, and 10 *e* and *f* in Appendix S1). These analyses validate the prognostic contribution of the global pattern, which is computed by GSVD of patient-matched tumor and normal aCGH profiles, also for patients with measured GBM aCGH profiles in the absence of matched normal profiles.

## Discussion

Previously, we showed that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent experimental or biological reality [12, 39]. The variables, subspaces of significant patterns that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [40] verify a computationally predicted genome-wide mode of regulation [41, 42], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

Recently, we mathematically defined a higher-order GSVD (HO GSVD) for more than two large-scale matrices with different row dimensions and the same column dimension [13]. We proved that this novel HO GSVD extends to higher orders almost all of the mathematical properties of the GSVD. We showed, comparing global mRNA expression from the three disparate organisms *S. pombe*, *S. cerevisiae* and human, that the HO GSVD provides a sequence-independent comparative framework for more than two genomic datasets, where the variables and operations represent experimental or biological reality. The approximately common HO GSVD subspace represents biological similarity among the organisms. Simultaneous reconstruction in the common subspace removes the experimental artifacts, which are dissimilar, from the datasets.

We now show that also in probe-independent comparison of aCGH data from patient-matched tumor and normal samples, the mathematical variables of the GSVD, i.e., shared probelets and the corresponding tumor- and normal-specific arraylets, represent experimental or biological reality. Probelets that are mathematically significant in both datasets, correspond to normal arraylets representing copy-number variations (CNVs) in the normal human genome that are conserved in the tumor genome (e.g., female-specific X chromosome amplification) and are represented by the corresponding tumor arraylets. Probelets that are mathematically significant in the normal but not the tumor dataset represent experimental variations that exclusively affect the normal dataset. Similarly, some probelets that are mathematically significant in the tumor but not the normal dataset represent experimental variations that exclusively affect the tumor dataset.

We find that the mathematically second most tumor-exclusive probelet, which is also the mathematically most significant probelet in the tumor dataset, is statistically correlated, possibly biologically coor-

dinated with GBM patients' survival and response to chemotherapy. The corresponding tumor arraylet describes a global pattern of tumor-exclusive co-occurring CNAs, including most known GBM-associated changes in chromosome numbers and focal CNAs, as well as several previously unreported CNAs, including the biochemically putative drug target-encoding *TLK2* [14]. We find that a negligible weight of the second tumor arraylet in a patient's GBM aCGH profile, mathematically defined by either the corresponding copy number in the second probelet, or by the correlation of the GBM profile with the second arraylet, is indicative of a significantly longer GBM survival time. This GSVD comparative modeling of aCGH data from patient-matched tumor and normal samples, therefore, draws a mathematical analogy between the prediction of cellular modes of regulation and the prognosis of cancers.

We confirm our results with GSVD comparison of matched profiles of a larger set of TCGA patients, inclusive of the initial set. We validate the prognostic contribution of the pattern with GSVD classification of the GBM profiles of a set of patients that is independent of both the initial set and the inclusive confirmation set [15].

Additional possible applications of the GSVD (and also the HO GSVD) in personalized medicine include comparisons of multiple patient-matched datasets, each corresponding to either (i) a set of large-scale molecular biological profiles (such as DNA copy numbers) acquired by a high-throughput technology (such as DNA microarrays) from the same tissue type (such as tumor or normal); or (ii) a set of biomedical images or signals; or (iii) a set of anatomical or clinical pathology test results or phenotypical observations (such as age). GSVD comparisons can uncover the relations and possibly even causal coordinations between these different recorded aspects of the same medical phenomenon. GSVD comparisons can be used to determine a single patient's medical status in relation to all the other patients in the set, and inform the patient's diagnosis, prognosis and treatment.

## Acknowledgements

We thank CR Johnson, MA Saunders and CF Van Loan for insightful discussions of matrix and tensor computations, and H Colman, RL Jensen and JD Schiffman for helpful discussions of GBM cancer genomics. We also thank AM Gross for technical assistance.

## References

1. Purow B, Schiff D (2009) Advances in the genetics of glioblastoma: are we reaching critical mass? *Nat Rev Neurol* 5: 419-426.
2. Wiltshire RN, Rasheed BK, Friedman HS, Friedman AH, Bigner SH (2000) Comparative genetic patterns of glioblastoma multiforme: potential diagnostic tool for tumor classification. *Neuro Oncol* 2: 164-173.
3. Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, et al (2005) Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res* 65: 1678-1686.
4. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061-1068.
5. Mischel PS, Shai R, Shi T, Horvath S, Lu KV, et al (2003) Identification of molecular subtypes of glioblastoma by gene expression profiling. *Oncogene* 22: 2361-2673.
6. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, et al (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in *PDGFRA*, *IDH1*, *EGFR*, and *NF1*. *Cancer Cell* 17: 98-110.

7. Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, et al (2010) A multigene predictor of outcome in glioblastoma. *Neuro Oncol* 12: 49-57.
8. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, et al (2010) Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* 17: 510-522.
9. Curran WJ Jr, Scott CB, Horton J, Nelson JS, Weinstein AS, et al (1993) Recursive partitioning analysis of prognostic factors in three Radiation Therapy Oncology Group malignant glioma trials. *J Natl Cancer Inst* 85: 704-710.
10. Gorlia T, van den Bent MJ, Hegi ME, Mirimanoff RO, Weller M, et al (2008) Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *Lancet Oncol* 9: 29-38.
11. Golub GH, Van Loan CF (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press, third edition, 694 p.
12. Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci USA* 100: 3351-3356.
13. Ponnappalli SP, Saunders MA, Van Loan CF, Alter O (2011) A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. *PLoS ONE* 6: e28072.
14. Lee CH, Alter O (2010) Known and novel copy number alterations in GBM and their patterns of co-occurrence are revealed by GSVD comparison of array CGH data from patient-matched normal and tumor TCGA samples. In: 60th Annual American Society of Human Genetics (ASHG) Meeting (November 2-6, 2010, Washington, DC).
15. Alpert BO, Sankaranarayanan P, Lee CH, Alter O (2011) Glioblastoma multiforme prognosis by using a patient's array CGH tumor profile and a generalized SVD-computed global pattern of copy-number alterations. In: 2nd DNA and Genome World Day (April 25-29, 2011, Dalian, China).
16. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, et al (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet* 359: 1301-1307.
17. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22: 281-285.
18. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al (2002) The human genome browser at UCSC. *Genome Res* 12: 996-1006.
19. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* 39: D876-D882.
20. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557-572.
21. Venkatraman ES, Olshen AB (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* 23: 657-663.

22. Heidenblad M, Lindgren D, Veltman JA, Jonson T, Mahlamäki EH, et al (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene* 24: 1794-1801.
23. Wang Q, Diskin S, Rappaport E, Attiyeh E, Mosse Y, et al (2006) Integrative genomics identifies distinct molecular classes of neuroblastoma and shows that multiple genes are targeted by regional alterations in DNA copy number. *Cancer Res* 66: 6050-6062.
24. Hopkins AL, Groom CR (2002) The druggable genome. *Nat Rev Drug Discov* 1: 727-730.
25. Silljé HH, Takahashi K, Tanaka K, Van Houwe G, Nigg EA (1999) Mammalian homologues of the plant Tousled gene code for cell-cycle-regulated kinases with maximal activities linked to ongoing DNA replication. *EMBO J* 18: 5691-5702.
26. Chandran UR, Ma C, Dhir R, Bisceglia M, Lyons-Weiler M, et al (2007) Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. *BMC Cancer* 7: 64.
27. Pellegrini M, Cheng JC, Voutila J, Judelson D, Taylor J, et al (2008) Expression profile of CREB knockdown in myeloid leukemia cells. *BMC Cancer* 8: 264.
28. Millour M, Charbonnel C, Magrangeas F, Minvielle S, Campion L, et al (2006) Gene expression profiles discriminate between pathological complete response and resistance to neoadjuvant FEC100 in breast cancer. *Cancer Genomics Proteomics* 3: 89-95.
29. Snijders AM, Nowee ME, Fridlyand J, Piek JM, Dorsman JC, et al (2003) Genome-wide-array-based comparative genomic hybridization reveals genetic homogeneity and frequent copy number increases encompassing CCNE1 in fallopian tube carcinoma. *Oncogene* 22: 4281-4286.
30. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, et al (2010) The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* 467: 1109-1113.
31. Etemadmoghadam D, George J, Cowin PA, Cullinane C, Kansara M, et al (2010) Amplicon-dependent CCNE1 expression is critical for clonogenic survival after cisplatin treatment and is correlated with 20q11 gain in ovarian cancer. *PLoS ONE* 5: e15498.
32. Defeo-Jones D, Huang PS, Jones RE, Haskell KM, Vuocolo GA, et al (1991) Cloning of cDNAs for cellular proteins that bind to the retinoblastoma gene product. *Nature* 352: 251-254.
33. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, et al (2010) A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* 141: 69-80.
34. Pardridge WM (2005) The blood-brain barrier: bottleneck in brain drug development. *NeuroRx* 2: 3-14.
35. Hattori Y, Ohta S, Hamada K, Yamada-Okabe H, Kanemura Y, et al (2007) Identification of a neuron-specific human gene, KIAA1110, that is a guanine nucleotide exchange factor for ARF1. *Biochem Biophys Res Commun* 364: 737-742.
36. Kaplan EL, Meier P (1958) Nonparametric estimation from incomplete observations. *J Amer Statist Assn* 53: 457-481.
37. Cox DR (1972) Regression models and life-tables. *J Roy Statist Soc B* 34: 187-220.

38. Rothman KJ (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1: 43-46.
39. Alter O (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci USA* 103: 16063-16064.
40. Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffley JF, et al (2009) Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol Syst Biol* 5: 312.
41. Alter O, Golub GH (2004) Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc Natl Acad Sci USA* 101: 16577-16582.
42. Omberg L, Golub GH, Alter O (2007) A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc Natl Acad Sci USA* 104: 18371-18376.

## Supporting Information

**Mathematica Notebook S1. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles.** A Mathematica 8.0.1 code file, executable by Mathematica 8.0.1 and readable by Mathematica Player, freely available at <http://www.wolfram.com/products/player/>.  
(NB)

**Mathematica Notebook S2. Generalized singular value decomposition (GSVD) of the TCGA patient-matched tumor and normal aCGH profiles.** A PDF format file, readable by Adobe Acrobat Reader.  
(PDF)

**Dataset S1. Initial set of 251 patients.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing The Cancer Genome Atlas (TCGA) [4] annotations of the initial set of 251 patients and the corresponding normal and tumor samples. The tumor and normal profiles of the initial set of 251 patients, in tab-delimited text format files, tabulating  $\log_2$  relative copy number variation across 212,696 and 211,227 tumor and normal probes, respectively, are available at <http://www.alterlab.org/GBM.prognosis/>.  
(TXT)

**Dataset S2. Segments of the significant tumor and normal arraylets, computed by GSVD for the initial set of 251 patients.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating segments identified by circular binary segmentation (CBS) [20, 21].  
(TXT)

**Dataset S3. Segments of the second tumor arraylet, computed by GSVD for the initial set of 251 patients.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, tabulating, for each of the 130 CBS segments of the second tumor arraylet, the segment's coordinates, the CBS  $P$ -value, and the log-rank test  $P$ -value corresponding to the Kaplan-Meier (KM) survival analysis of the initial set of 251 patients classified by either a gain or a loss of this segment.  
(TXT)

**Dataset S4. Inclusive confirmation set of 344 patients.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the inclusive confirmation set of 344 patients. The tumor and normal profiles of the inclusive confirmation set of 344 patients, in tab-delimited text format files, tabulating  $\log_2$  relative copy number variation across 200,139 and 198,342 tumor and normal probes, respectively, are available at <http://www.alterlab.org/GBM.prognosis/>.  
(TXT)

**Dataset S5. Independent validation set of 184 patients.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the TCGA annotations of the independent validation set of 184 patients. Tumor profiles of the independent validation set of 184 patients, in a tab-delimited text format file, tabulating  $\log_2$  relative copy number variation across 212,696 autosomal and X chromosome probes, is available at <http://www.alterlab.org/GBM.prognosis/>.  
(TXT)

# Appendix

## B

# A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms

Sri Priya Ponnappalli<sup>1</sup>, Michael A. Saunders<sup>2</sup>, Charles F. Van Loan<sup>3</sup>, and Orly Alter<sup>4,\*</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, University of Texas at Austin, TX 78712, USA

<sup>2</sup> Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

<sup>3</sup> Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

<sup>4</sup> Scientific Computing and Imaging (SCI) Institute and Departments of Bioengineering and Human Genetics, University of Utah, Salt Lake City, UT 84112, USA

\* E-mail: orly@sci.utah.edu

## Abstract

The number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing in many areas of science, 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 generalized singular value decomposition (GSVD), is limited to two matrices. We mathematically define a higher-order GSVD (HO GSVD) for  $N \geq 2$  matrices  $D_i \in m_i \times n$ , each with full column rank. Each matrix is exactly factored as  $D_i = U_i \Sigma_i V^T$ , where  $V$ , identical in all factorizations, is obtained from the eigensystem  $SV = V\Lambda$  of the arithmetic mean  $S$  of all pairwise quotients  $A_i A_j^{-1}$  of the matrices  $A_i = D_i^T D_i$ ,  $i \neq j$ . We prove that this decomposition extends to higher orders almost all of the mathematical properties of the GSVD. The matrix  $S$  is nondefective with  $V$  and  $\Lambda$  real. Its eigenvalues satisfy  $\lambda_k \geq 1$ . Equality holds if and only if the corresponding eigenvector  $v_k$  is a right basis vector of equal significance in all matrices  $D_i$  and  $D_j$ , that is  $\sigma_{i,k}/\sigma_{j,k} = 1$  for all  $i$  and  $j$ , and the corresponding left basis vector  $u_{i,k}$  is orthogonal to all other vectors in  $U_i$  for all  $i$ . The eigenvalues  $\lambda_k = 1$ , therefore, define the “common HO GSVD subspace.” We illustrate the HO GSVD with a comparison of genome-scale cell-cycle mRNA expression from *S. pombe*, *S. cerevisiae* and human. Unlike existing algorithms, a mapping among the genes of these disparate organisms is not required. We find that 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. In the simultaneous sequence-independent classification of the genes of the three organisms in this common subspace, genes of highly conserved sequences but significantly different cell-cycle peak times are correctly classified.

## Introduction

In many areas of science, especially in biotechnology, the number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing. This is accompanied by a fundamental need for mathematical frameworks that can compare multiple large-scale matrices with different row dimensions. For example, comparative analyses of global mRNA expression from multiple model organisms promise to enhance fundamental understanding of the universality and specialization of molecular biological mechanisms, and may prove useful in medical diagnosis, treatment and drug design [1]. Existing algorithms limit analyses to subsets of homologous genes among the different organisms, effectively introducing into the analysis the assumption that sequence and functional similarities are equivalent (e.g., [2]). However, it is well known that this assumption does not always hold, for example, in cases of nonorthologous gene displacement, when nonorthologous proteins in different organisms fulfill the same function [3]. For sequence-independent comparisons, mathematical frameworks are required that can distinguish and

separate the similar from the dissimilar among multiple large-scale datasets tabulated as matrices with different row dimensions, corresponding to the different sets of genes of the different organisms. The only such framework to date, the generalized singular value decomposition (GSVD) [4–7], is limited to two matrices.

It was shown that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent biological reality [7,8]. The variables, significant subspaces that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [9] verify a computationally predicted genome-wide mode of regulation that correlates DNA replication origin activity with mRNA expression [10,11], demonstrating that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

We now define a higher-order GSVD (HO GSVD) for the comparison of  $N \geq 2$  datasets. The datasets are tabulated as  $N$  real matrices  $D_i \in m_i \times n$ , each with full column rank, with different row dimensions and the same column dimension, where there exists a one-to-one mapping among the columns of the matrices. Like the GSVD, the HO GSVD is an exact decomposition, i.e., each matrix is exactly factored as  $D_i = U_i \Sigma_i V^T$ , where the columns of  $U_i$  and  $V$  have unit length and are the left and right basis vectors respectively, and each  $\Sigma_i$  is diagonal and positive definite. Like the GSVD, the matrix  $V$  is identical in all factorizations. In our HO GSVD, the matrix  $V$  is obtained from the eigensystem  $SV = V\Lambda$  of the arithmetic mean  $S$  of all pairwise quotients  $A_i A_j^{-1}$  of the matrices  $A_i = D_i^T D_i$ , or equivalently of all  $S_{ij} = \frac{1}{2}(A_i A_j^{-1} + A_j A_i^{-1})$ ,  $i \neq j$ .

To clarify our choice of  $S$ , we note that in the GSVD, defined by Van Loan [5], the matrix  $V$  can be formed from the eigenvectors of the unbalanced quotient  $A_1 A_2^{-1}$  (Section 1 in Supplementary Appendix C). We observe that this  $V$  can also be formed from the eigenvectors of the balanced arithmetic mean  $S_{12} = \frac{1}{2}(A_1 A_2^{-1} + A_2 A_1^{-1})$ . We prove that in the case of  $N = 2$ , our definition of  $V$  by using the eigensystem of  $S \equiv S_{12} = \frac{1}{2}(A_1 A_2^{-1} + A_2 A_1^{-1})$  leads algebraically to the GSVD (Supplementary Theorems S1–S5 in Appendix C), and therefore, as Paige and Saunders showed [6], can be computed in a stable way. We also note that in the GSVD, the matrix  $V$  does not depend upon the ordering of the matrices  $D_1$  and  $D_2$ . Therefore, we define our HO GSVD for  $N \geq 2$  matrices by using the balanced arithmetic mean  $S$  of all pairwise arithmetic means  $S_{ij}$ , each of which defines the GSVD of the corresponding pair of matrices  $D_i$  and  $D_j$ , noting that  $S$  does not depend upon the ordering of the matrices  $D_i$  and  $D_j$ .

We prove that  $S$  is nondefective (it has  $n$  independent eigenvectors), and that its eigensystem is real (Theorem 1). We prove that the eigenvalues of  $S$  satisfy  $\lambda_k \geq 1$  (Theorem 2). As in our GSVD comparison of two matrices [7], we interpret the  $k$ th diagonal of  $\Sigma_i = \text{diag}(\sigma_{i,k})$  in the factorization of the  $i$ th matrix  $D_i$  as indicating the significance of the  $k$ th right basis vector  $v_k$  in  $D_i$  in terms of the overall information that  $v_k$  captures in  $D_i$ . The ratio  $\sigma_{i,k}/\sigma_{j,k}$  indicates the significance of  $v_k$  in  $D_i$  relative to its significance in  $D_j$ . We prove that an eigenvalue of  $S$  satisfies  $\lambda_k = 1$  if and only if the corresponding eigenvector  $v_k$  is a right basis vector of equal significance in all  $D_i$  and  $D_j$ , that is,  $\sigma_{i,k}/\sigma_{j,k} = 1$  for all  $i$  and  $j$ , and the corresponding left basis vector  $u_{i,k}$  is orthonormal to all other vectors in  $U_i$  for all  $i$ . We therefore mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the  $N \geq 2$  matrices to be the subspace spanned by the right basis vectors  $v_k$  that correspond to the  $\lambda_k = 1$  eigenvalues of  $S$  (Theorem 3). We also show that each of the right basis vectors  $\{v_k\}$  that span the common HO GSVD subspace 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  $i$  and  $j$  (Corollary 1).

Recent research showed that several higher-order generalizations are possible for a given matrix decomposition, each preserving some but not all of the properties of the matrix decomposition [12–14] (see also Supplementary Theorem S6 and Conjecture S1 in Appendix C). Our new HO GSVD extends to higher orders all of the mathematical properties of the GSVD except for complete column-wise orthogonality of the left basis vectors that form the matrix  $U_i$  for all  $i$ , i.e., in each factorization.

We illustrate the HO GSVD with a comparison of cell-cycle mRNA expression from *S. pombe* [15,16], *S. cerevisiae* [17] and human [18]. Unlike existing algorithms, a mapping among the genes of these disparate organisms is not required (Supplementary Section 2 in Appendix C). We find that the common HO GSVD subspace represents the cell-cycle mRNA expression oscillations, which are similar among the datasets. Simultaneous reconstruction in this common subspace, therefore, removes the experimental artifacts, which are dissimilar, from the datasets. Simultaneous sequence-independent classification of the genes of the three organisms in the common subspace is in agreement with previous classifications into cell-cycle phases [19]. Notably, genes of highly conserved sequences across the three organisms [20,21] but significantly different cell-cycle peak times, such as genes from the ABC transporter superfamily [22–28], phospholipase B-encoding genes [29,30] and even the B cyclin-encoding genes [31,32], are correctly classified.

## Methods

### HO GSVD Construction

Suppose we have a set of  $N$  real matrices  $D_i \in m_i \times n$  each with full column rank. We define a HO GSVD of these  $N$  matrices as

$$\begin{aligned} D_1 &= U_1 \Sigma_1 V^T, \\ D_2 &= U_2 \Sigma_2 V^T, \\ &\vdots \\ D_N &= U_N \Sigma_N V^T, \end{aligned} \quad (1)$$

where each  $U_i \in m_i \times n$  is composed of normalized left basis vectors, each  $\Sigma_i = \text{diag}(\sigma_{i,k}) \in n \times n$  is diagonal with  $\sigma_{i,k} > 0$ , and  $V$ , identical in all matrix factorizations, is composed of normalized right basis vectors. As in the GSVD comparison of global mRNA expression from two organisms [7], in the HO GSVD comparison of global mRNA expression from  $N \geq 2$  organisms, the shared right basis vectors  $v_k$  of Equation (1) are the “genelets” and the  $N$  sets of left basis vectors  $u_{i,k}$  are the  $N$  sets of “arraylets” (Figure 11 and Supplementary Section 2 in Appendix C). We obtain  $V$  from the eigensystem of  $S$ , the arithmetic mean of all pairwise quotients  $A_i A_j^{-1}$  of the matrices  $A_i = D_i^T D_i$ , or equivalently of all  $S_{ij} = \frac{1}{2}(A_i A_j^{-1} + A_j A_i^{-1})$ ,  $i \neq j$ :

$$\begin{aligned} S &\equiv \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j>i}^N (A_i A_j^{-1} + A_j A_i^{-1}) \\ &= \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j>i}^N S_{ij}, \\ SV &= V\Lambda, \\ V &\equiv (v_1 \ \dots \ v_n), \quad \Lambda = \text{diag}(\lambda_k), \end{aligned} \quad (2)$$

with  $\|v_k\| = 1$ . We prove that  $S$  is nondefective, i.e.,  $S$  has  $n$  independent eigenvectors, and that its eigenvectors  $V$  and eigenvalues  $\Lambda$  are real (Theorem 1). We prove that the eigenvalues of  $S$  satisfy  $\lambda_k \geq 1$  (Theorem 2).

Given  $V$ , we compute matrices  $B_i$  by solving  $N$  linear systems:

$$\begin{aligned} VB_i^T &= D_i^T, \\ B_i &\equiv (b_{i,1} \ \dots \ b_{i,n}), \quad i = 1, \dots, N, \end{aligned} \quad (3)$$

and we construct  $\Sigma_i$  and  $U_i = (u_{i,1} \ \dots \ u_{i,n})$  by normalizing the columns of  $B_i$ :

$$\begin{aligned} \sigma_{i,k} &= \|b_{i,k}\|, \\ \Sigma_i &= \text{diag}(\sigma_{i,k}), \\ B_i &= U_i \Sigma_i. \end{aligned} \quad (4)$$

### HO GSVD Interpretation

In this construction, the rows of each of the  $N$  matrices  $D_i$  are superpositions of the same right basis vectors, the columns of  $V$  (Figures 37 and 38 and Section 1 in Appendix C). As in our GSVD comparison of two matrices, we interpret the  $k$ th diagonals of  $\Sigma_i$ , the “higher-order generalized singular value set”  $\{\sigma_{i,k}\}$ , as indicating the significance of the  $k$ th right basis vector  $v_k$  in the matrices  $D_i$ , and reflecting the overall information that  $v_k$  captures in each  $D_i$  respectively. The ratio  $\sigma_{i,k}/\sigma_{j,k}$  indicates the significance of  $v_k$  in  $D_i$  relative to its significance in  $D_j$ . A ratio of  $\sigma_{i,k}/\sigma_{j,k} = 1$  for all  $i$  and  $j$  corresponds to a right basis vector  $v_k$  of equal significance in all  $N$  matrices  $D_i$ . GSVD comparisons of two matrices showed that right basis vectors of approximately equal significance in the two matrices reflect themes that are common to both matrices under comparison [7]. A ratio of  $\sigma_{i,k}/\sigma_{j,k} \ll 1$  indicates a basis vector  $v_k$  of almost negligible significance in  $D_i$  relative to its significance in  $D_j$ . GSVD comparisons of two matrices showed that right basis vectors of negligible significance in one matrix reflect themes that are exclusive to the other matrix.

We prove that an eigenvalue of  $S$  satisfies  $\lambda_k = 1$  if and only if the corresponding eigenvector  $v_k$  is a right basis vector of equal significance in all  $D_i$  and  $D_j$ , that is,  $\sigma_{i,k}/\sigma_{j,k} = 1$  for all  $i$  and  $j$ , and the corresponding left basis vector  $u_{i,k}$  is orthonormal to all other vectors in  $U_i$  for all  $i$ . We therefore mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the  $N \geq 2$  matrices to be the subspace spanned by the right basis vectors  $\{v_k\}$  corresponding to the eigenvalues of  $S$  that satisfy  $\lambda_k = 1$  (Theorem 3).

It follows that each of the right basis vectors  $\{v_k\}$  that span the common HO GSVD subspace 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  $i$  and  $j$  (Corollary 1). Since the GSVD can be computed in a stable way [6], we note that the common HO GSVD subspace can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices  $D_i$  and  $D_j$ . This also suggests that it may be possible to formulate the HO GSVD as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33]. Such a formulation may lead to a stable numerical algorithm for computing the HO GSVD, and possibly also to a higher-order general Gauss-Markov linear statistical model [34–36].

We show, in a comparison of  $N = 3$  matrices, that the approximately common HO GSVD subspace of these three matrices reflects a theme that is common to the three matrices under comparison (Section 2).

### HO GSVD Mathematical Properties

**Theorem 1.**  *$S$  is nondefective (it has  $n$  independent eigenvectors) and its eigensystem is real.*

*Proof.* From Equation (2) it follows that

$$\begin{aligned} S &= \frac{1}{N(N-1)}(H - NI), \\ H &= \left( \sum_{i=1}^N A_i \right) \left( \sum_{j=1}^N A_j^{-1} \right), \end{aligned} \quad (5)$$

and the eigenvectors of  $S$  equal the eigenvectors of  $H$ .

Let the SVD of the matrices  $D_i$  appended along the  $n$ -columns axis be

$$\begin{aligned} \begin{bmatrix} D_1 \\ \vdots \\ D_N \end{bmatrix} &= \begin{bmatrix} \hat{U}_1 \\ \vdots \\ \hat{U}_N \end{bmatrix} \hat{\Sigma} \hat{V}^T, \\ \sum_{i=1}^N \hat{U}_i^T \hat{U}_i &= I. \end{aligned} \quad (6)$$

Since the matrices  $D_i$  are real and with full column rank, it follows from the SVD of  $\hat{U}_i$  that the symmetric matrices  $\hat{U}_i^T \hat{U}_i$  are real and positive definite, and their inverses exist. It then follows from Equations (5) and (6) that  $H$  is similar to  $\hat{H}$ ,

$$\begin{aligned} H &= \hat{V} \hat{\Sigma} \hat{H} \hat{\Sigma}^{-1} \hat{V}^T, \\ \hat{H} &= \sum_{j=1}^N (\hat{U}_j^T \hat{U}_j)^{-1}, \end{aligned} \quad (7)$$

and the eigenvalues of  $H$  equal the eigenvalues of  $\hat{H}$ .

A sum of real, symmetric and positive definite matrices,  $\hat{H}$  is also real, symmetric and positive definite; therefore, its eigensystem

$$\hat{Y}^T \hat{H} \hat{Y} = \text{diag}(\mu_k) \quad (8)$$

is real with  $\hat{Y}$  orthogonal and  $\mu_k > 0$ . Without loss of generality let  $\hat{Y}$  be orthonormal, such that  $\|\hat{y}_k\| = 1$ . It follows from the similarity of  $H$  with  $\hat{H}$  that the eigensystem of  $H$  can be written as  $V^{-1} H V = \text{diag}(\mu_k)$ , with the real and nonsingular  $V = (\hat{V} \hat{\Sigma} \hat{Y}) \hat{W}^{-1}$ , where  $\hat{W} = \text{diag}(\hat{w}_k)$  and  $\hat{w}_k = \|\hat{V} \hat{\Sigma} \hat{y}_k\|$  such that  $\|v_k\| = 1$  for all  $k$ .

Thus, from Equation (5),  $S$  is nondefective with real eigenvectors  $V$ . Also, the eigenvalues of  $S$  satisfy

$$\lambda_k = \frac{1}{N(N-1)}(\mu_k - N), \quad (9)$$

where  $\mu_k > 0$  are the eigenvalues of  $H$  and  $\hat{H}$ . Thus, the eigenvalues of  $S$  are real.  $\square$

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

*Proof.* Following Equation (9), asserting that the eigenvalues of  $S$  satisfy  $\lambda_k \geq 1$  is equivalent to asserting that the eigenvalues of  $\hat{H}$  satisfy  $\mu_k \geq N^2$ .

From Equations (6) and (7), the eigenvalues of  $\hat{H}$  satisfy

$$\mu_k \geq \min_x \sum_{j=1}^N [x^T (\hat{U}_j^T \hat{U}_j) x]^{-1}, \quad (10)$$

under the constraint that

$$\sum_{j=1}^N x^T (\hat{U}_j^T \hat{U}_j) x = 1, \quad (11)$$

where  $x$  is a real unit vector, and where it follows from the Cauchy-Schwarz inequality [37] (see also [4, 34, 38]) for the real nonzero vectors  $(\hat{U}_j^T \hat{U}_j)x$  and  $(\hat{U}_j^T \hat{U}_j)^{-1}x$  that for all  $j$

$$x^T (\hat{U}_j^T \hat{U}_j)^{-1} x \geq [x^T (\hat{U}_j^T \hat{U}_j) x]^{-1}. \quad (12)$$

With the constraint of Equation (11), which requires the sum of the  $N$  positive numbers  $x^T (\hat{U}_j^T \hat{U}_j)^{-1} x$  to equal one, the lower bound on the eigenvalues of  $\hat{H}$  in Equation (10) is at its minimum when the sum of the inverses of these numbers is at its minimum, that is, when the numbers equal

$$x^T (\hat{U}_i^T \hat{U}_i) x = x^T (\hat{U}_j^T \hat{U}_j) x = N^{-1} \quad (13)$$

for all  $i$  and  $j$ . Thus, the eigenvalues of  $\hat{H}$  satisfy  $\mu_k \geq N^2$ .  $\square$

**Theorem 3. The common HO GSVD subspace.** An eigenvalue of  $S$  satisfies  $\lambda_k = 1$  if and only if the corresponding eigenvector  $v_k$  is a right basis vector of equal significance in all  $D_i$  and  $D_j$ , that is,  $\sigma_{i,k}/\sigma_{j,k} = 1$  for all  $i$  and  $j$ , and the corresponding left basis vector  $u_{i,k}$  is orthonormal to all other vectors in  $U_i$  for all  $i$ . The “common HO GSVD subspace” of the  $N \geq 2$  matrices is, therefore, the subspace spanned by the right basis vectors  $\{v_k\}$  corresponding to the eigenvalues of  $S$  that satisfy  $\lambda_k = 1$ .

*Proof.* Without loss of generality, let  $k = n$ . From Equation (12) and the Cauchy-Schwarz inequality, an eigenvalue of  $\hat{H}$  equals its minimum lower bound  $\mu_n = N^2$  if and only if the corresponding eigenvector  $\hat{y}_n$  is also an eigenvector of  $\hat{U}_i^T \hat{U}_i$  for all  $i$  [37], where, from Equation (13), the corresponding eigenvalue equals  $N^{-1}$ ,

$$(\hat{U}_i^T \hat{U}_i) \hat{Y} = [(\hat{U}_i^T \hat{U}_i) \hat{y}_1 \quad (\hat{U}_i^T \hat{U}_i) \hat{y}_2 \quad \dots \quad N^{-1} \hat{y}_n]. \quad (14)$$

Given the eigenvectors  $V = (\hat{V} \hat{\Sigma} \hat{Y}) \hat{W}^{-1}$  of  $S$ , we solve Equation (3) for each  $D_i = \hat{U}_i \hat{\Sigma} \hat{V}^T$  of Equation (6), and obtain

$$\begin{aligned} U_i \Sigma_i = B_i &= D_i V^{-T} \\ &= \hat{U}_i \hat{Y} \hat{W}. \end{aligned} \quad (15)$$

Following Equations (14) and (15), where  $v_n = \hat{w}_n^{-1} \hat{V} \hat{\Sigma} \hat{y}_n$  corresponds to a minimum eigenvalue  $\lambda_n = 1$ , and since  $\hat{Y}$  is orthonormal, we obtain

$$\begin{aligned} \hat{W}^{-1} \Sigma_i (U_i^T U_i) \Sigma_i \hat{W}^{-1} &= \hat{Y}^T (\hat{U}_i^T \hat{U}_i) \hat{Y} \\ &= \begin{bmatrix} \hat{y}_1^T (\hat{U}_i^T \hat{U}_i) \hat{y}_1 & \hat{y}_2^T (\hat{U}_i^T \hat{U}_i) \hat{y}_1 & \dots & 0 \\ \hat{y}_1^T (\hat{U}_i^T \hat{U}_i) \hat{y}_2 & \hat{y}_2^T (\hat{U}_i^T \hat{U}_i) \hat{y}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & N^{-1} \end{bmatrix}, \end{aligned} \quad (16)$$

with zeroes in the  $n$ th row and the  $n$ th column of the matrix above everywhere except for the diagonal element. Thus, an eigenvalue of  $S$  satisfies  $\lambda_n = 1$  if and only if the corresponding left basis vectors  $u_{i,n}$  are orthonormal to all other vectors in  $U_i$ .

The corresponding higher-order generalized singular values are  $\sigma_{i,n} = N^{-1/2} \hat{w}_n > 0$ . Thus  $\sigma_{i,n}/\sigma_{j,n} = 1$  for all  $i$  and  $j$ , and the corresponding right basis vector  $v_n$  is of equal significance in all matrices  $D_i$  and  $D_j$ .  $\square$

**Corollary 1.** *An eigenvalue of  $S$  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  $i$  and  $j$ .*

*Proof.* From Equations (12) and (13), and since the pairwise quotients  $A_i A_j^{-1}$  are similar to  $(U_i^T U_i)(U_j^T U_j)^{-1}$  with the similarity transformation of  $\hat{V} \hat{\Sigma}$  for all  $i$  and  $j$ , it follows that an eigenvalue of  $S$  satisfies  $\lambda_k = 1$  if and only if the corresponding right basis vector  $v_k = \hat{w}_k^{-1} \hat{V} \hat{\Sigma} \hat{y}_k$  is also an eigenvector of each of the pairwise quotients  $A_i A_j^{-1}$  of the matrices  $A_i = D_i^T D_i$  with equal corresponding eigenvalues, or equivalently of all  $S_{ij}$  with all eigenvalues at their minimum of one,

$$S_{ij} v_k = \frac{1}{2}(A_i A_j^{-1} + A_i A_j^{-1}) v_k = v_k. \quad (17)$$

We prove (Supplementary Theorems S1–S5 in Appendix C) that in the case of  $N = 2$  matrices our definition of  $V$  by using the eigensystem of  $S_{ij}$  leads algebraically to the GSVD, where an eigenvalue of  $S_{ij}$  equals its minimum of one if and only if the two corresponding generalized singular values are equal, such that the corresponding generalized singular vector  $v_k$  is of equal significance in both matrices  $D_i$  and  $D_j$ . Thus, it follows that each of the right basis vectors  $\{v_k\}$  that span the common HO GSVD subspace 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  $i$  and  $j$ .  $\square$

Note that since the GSVD can be computed in a stable way [6], the common HO GSVD subspace we define (Theorem 3) can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices  $D_i$  and  $D_j$  (Corollary 1). It may also be possible to formulate the HO GSVD as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33]. Such a formulation may lead to a stable numerical algorithm for computing the HO GSVD, and possibly also to a higher-order general Gauss-Markov linear statistical model [34–36].

## Results

### HO GSVD Comparison of Global mRNA Expression from Three Organisms

Consider now the HO GSVD comparative analysis of global mRNA expression datasets from the  $N = 3$  organisms *S. pombe*, *S. cerevisiae* and human (Supplementary Section 2.1 in Appendix C, Mathematica Notebooks S1 and S2, and Datasets S1–S3). The datasets are tabulated as matrices of  $n = 17$  columns each, corresponding to DNA microarray-measured mRNA expression from each organism at 17 time points equally spaced during approximately two cell-cycle periods. The underlying assumption is that there exists a one-to-one mapping among the 17 columns of the three matrices but not necessarily among their rows, which correspond to either  $m_1 = 3167$ -*S. pombe* genes,  $m_2 = 4772$ -*S. cerevisiae* genes or  $m_3 = 13,068$ -human genes. The HO GSVD of Equation (1) transforms the datasets from the organism-specific genes  $\times$  17-arrays spaces to the reduced spaces of the 17-“arraylets,” i.e., left basis vectors  $\times$  17-“genelets,” i.e., right basis vectors, where the datasets  $D_i$  are represented by the diagonal nonnegative matrices  $\Sigma_i$ , by using the organism-specific genes  $\times$  17-arraylets transformation matrices  $U_i$  and the one shared 17-genelets  $\times$  17-arrays transformation matrix  $V^T$  (Figure 11).

Following Theorem 3, the approximately common HO GSVD subspace of the three datasets is spanned by the five genelets  $k = 13, \dots, 17$  that correspond to  $1 \lesssim \lambda_k \lesssim 2$ . We find that these five genelets are approximately equally significant with  $\sigma_{1,k} : \sigma_{2,k} : \sigma_{3,k} \sim 1 : 1 : 1$  in the *S. pombe*, *S. cerevisiae* and human datasets, respectively (Figure 12 a and b). The five corresponding arraylets in each dataset are  $\epsilon = 0.33$ -orthonormal to all other arraylets (Figure 15).

## Common HO GSVD Subspace Represents Similar Cell-Cycle Oscillations

The expression variations across time of the five genelets that span the approximately common HO GSVD subspace fit normalized cosine functions of two periods, superimposed on time-invariant expression (Figure 12 *c* and *d*). Consistently, the corresponding organism-specific arraylets are enriched [39] in overexpressed or underexpressed organism-specific cell cycle-regulated genes, with 24 of the 30 *P*-values  $< 10^{-8}$  (Table 1 and Supplementary Section 2.2 in Appendix C). For example, the three 17th arraylets, which correspond to the 0-phase 17th genelet, are enriched in overexpressed G2 *S. pombe* genes, G2/M and M/G1 *S. cerevisiae* genes and S and G2 human genes, respectively, representing the cell-cycle checkpoints in which the three cultures are initially synchronized.

Simultaneous sequence-independent reconstruction and classification of the three datasets in the common subspace outline cell-cycle progression in time and across the genes in the three organisms (Supplementary Sections 2.3 and 2.4 in Appendix C). Projecting the expression of the 17 arrays of either organism from the corresponding five-dimensional arraylets subspace onto the two-dimensional subspace that approximates it (Figure 16),  $\geq 50\%$  of the contributions of the arraylets add up, rather than cancel out (Figure 13 *a-c*). In these two-dimensional subspaces, the angular order of the arrays of either organism describes cell-cycle progression in time through approximately two cell-cycle periods, from the initial cell-cycle phase and back to that initial phase twice. Projecting the expression of the genes,  $\geq 50\%$  of the contributions of the five genelets add up in the overall expression of 343 of the 380 *S. pombe* genes classified as cell cycle-regulated, 554 of the 641 *S. cerevisiae* cell-cycle genes, and 632 of the 787 human cell-cycle genes (Figure 13 *d-f*). Simultaneous classification of the genes of either organism into cell-cycle phases according to their angular order in these two-dimensional subspaces is consistent with the classification of the arrays, and is in good agreement with the previous classifications of the genes (Figure 13 *g-i*). With all 3167 *S. pombe*, 4772 *S. cerevisiae* and 13,068 human genes sorted, the expression variations of the five arraylets from each organism approximately fit one-period cosines, with the initial phase of each arraylet (Figures 17-19) similar to that of its corresponding genelet (Figure 12). The global mRNA expression of each organism, reconstructed in the common HO GSVD subspace, approximately fits a traveling wave, oscillating across time and across the genes.

Note also that simultaneous reconstruction in the common HO GSVD subspace removes the experimental artifacts and batch effects, which are dissimilar, from the three datasets. Consider, for example, the second genelet. With  $\sigma_{1,2} : \sigma_{2,2} : \sigma_{3,2} \sim 1 : 8 : 3$  in the *S. pombe*, *S. cerevisiae* and human datasets, respectively, this genelet is almost exclusive to the *S. cerevisiae* dataset. This genelet is anticorrelated with a time decaying pattern of expression (Figure 12*a*). Consistently, the corresponding *S. cerevisiae*-specific arraylet is enriched in underexpressed *S. cerevisiae* genes that were classified as up-regulated by the *S. cerevisiae* synchronizing agent, the  $\alpha$ -factor pheromone, with the *P*-value  $< 10^{-46}$ . Reconstruction in the common subspace effectively removes this *S. cerevisiae*-approximately exclusive pattern of expression variation from the three datasets.

## Simultaneous HO GSVD Classification of Homologous Genes of Different Cell-Cycle Peak Times

Notably, in the simultaneous sequence-independent classification of the genes of the three organisms in the common subspace, genes of significantly different cell-cycle peak times [19] but highly conserved sequences [20, 21] are correctly classified (Supplementary Section 2.5 in Appendix C).

For example, consider the G2 *S. pombe* gene *BFR1* (Figure 14*a*), which belongs to the evolutionarily highly conserved ATP-binding cassette (ABC) transporter superfamily [22]. The closest homologs of *BFR1* in our *S. pombe*, *S. cerevisiae* and human datasets are the *S. cerevisiae* genes *SNQ2*, *PDR5*, *PDR15* and *PDR10* (Supplemental Table S1*a* in Appendix C). The expression of *SNQ2* and *PDR5* is known to peak at the S/G2 and G2/M cell-cycle phases, respectively [17]. However, sequence similarity does not imply similar cell-cycle peak times, and *PDR15* and *PDR10*, the closest homologs of *PDR5*,

are induced during stationary phase [23], which has been hypothesized to occur in G1, before the Cdc28-defined cell-cycle arrest [24]. Consistently, we find *PDR15* and *PDR10* at the M/G1 to G1 transition, antipodal to (i.e., half a cell-cycle period apart from) *SNQ2* and *PDR5*, which are projected onto S/G2 and G2/M, respectively (Figure 14b). We also find the transcription factor *PDR1* at S/G2, its known cell-cycle peak time, adjacent to *SNQ2* and *PDR5*, which it positively regulates and might be regulated by, and antipodal to *PDR15*, which it negatively regulates [25–28].

Another example is the *S. cerevisiae* phospholipase B-encoding gene *PLB1* [29], which peaks at the cell-cycle phase M/G1 [30]. Its closest homolog in our *S. cerevisiae* dataset, *PLB3*, also peaks at M/G1 [17] (Figure 14d). However, among the closest *S. pombe* and human homologs of *PLB1* (Supplementary Table S1b in Appendix C), we find the *S. pombe* genes *SPAC977.09c* and *SPAC1786.02*, which expressions peak at the almost antipodal *S. pombe* cell-cycle phases S and G2, respectively [19] (Figure 14c).

As a third example, consider the *S. pombe* G1 B-type cyclin-encoding gene *CIG2* [31,32] (Supplementary Table S1c in Appendix C). Its closest *S. pombe* homolog, *CDC13*, peaks at M [19] (Figure 14e). The closest human homologs of *CIG2*, the cyclins *CCNA2* and *CCNB2*, peak at G2 and G2/M, respectively (Figure 14g). However, while periodicity in mRNA abundance levels through the cell cycle is highly conserved among members of the cyclin family, the cell-cycle peak times are not necessarily conserved [1]: The closest homologs of *CIG2* in our *S. cerevisiae* dataset, are the G2/M promoter-encoding genes *CLB1,2* and *CLB3,4*, which expressions peak at G2/M and S respectively, and *CLB5*, which encodes a DNA synthesis promoter, and peaks at G1 (Figure 14f).

## Discussion

We mathematically defined a higher-order GSVD (HO GSVD) for two or more large-scale matrices with different row dimensions and the same column dimension. We proved that our new HO GSVD extends to higher orders almost all of the mathematical properties of the GSVD: The eigenvalues of  $S$  are always greater than or equal to one, and an eigenvalue of one corresponds to a right basis vector of equal significance in all matrices, and to a left basis vector in each matrix factorization that is orthogonal to all other left basis vectors in that factorization. We therefore mathematically defined, in analogy with the GSVD, the common HO GSVD subspace of the  $N \geq 2$  matrices to be the subspace spanned by the right basis vectors that correspond to the eigenvalues of  $S$  that equal one.

The only property that does not extend to higher orders in general is the complete column-wise orthogonality of the normalized left basis vectors in each factorization. Recent research showed that several higher-order generalizations are possible for a given matrix decomposition, each preserving some but not all of the properties of the matrix decomposition [12–14]. The HO GSVD has the interesting property of preserving the exactness and diagonality of the matrix GSVD and, in special cases, also partial or even complete column-wise orthogonality. That is, all  $N$  matrix factorizations in Equation (1) are exact, all  $N$  matrices  $\Sigma_i$  are diagonal, and when one or more of the eigenvalues of  $S$  equal one, the corresponding left basis vectors in each factorization are orthogonal to all other left basis vectors in that factorization.

The complete column-wise orthogonality of the matrix GSVD [5] enables its stable computation [6]. We showed that each of the right basis vectors that span the common HO GSVD subspace 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  $i$  and  $j$ . Since the GSVD can be computed in a stable way, the common HO GSVD subspace can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices  $D_i$  and  $D_j$ . That is, the common HO GSVD subspace exists also for  $N$  matrices  $D_i$  that are not all of full column rank. This also means that the common HO GSVD subspace can be formulated as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33].

It would be ideal if our procedure reduced to the stable computation of the matrix GSVD when  $N = 2$ . To achieve this ideal, we would need to find a procedure that allows a computation of the HO

GSVD, not just the common HO GSVD subspace, for  $N$  matrices  $D_i$  that are not all of full column rank. A formulation of the HO GSVD, not just the common HO GSVD subspace, as a solution to an optimization problem may lead to a stable numerical algorithm for computing the HO GSVD. Such a formulation may also lead to a higher-order general Gauss-Markov linear statistical model [34–36].

It was shown that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent experimental or biological reality [7, 8]. The variables, subspaces of significant patterns that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [9] verify a computationally predicted genome-wide mode of regulation [10, 11], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

Here we showed, comparing global cell-cycle mRNA expression from the three disparate organisms *S. pombe*, *S. cerevisiae* and human, that the HO GSVD provides a sequence-independent comparative framework for two or more genomic datasets, where the variables and operations represent biological reality. 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 removes the experimental artifacts, which are dissimilar, from the datasets. In the simultaneous sequence-independent classification of the genes of the three organisms in this common subspace, genes of highly conserved sequences but significantly different cell-cycle peak times are correctly classified.

Additional possible applications of our HO GSVD in biotechnology include comparison of multiple genomic datasets, each corresponding to: (i) the same experiment repeated multiple times using different experimental protocols, to separate the biological signal that is similar in all datasets from the dissimilar experimental artifacts; (ii) one of multiple types of genomic information, such as DNA copy number, DNA methylation and mRNA expression, collected from the same set of samples, e.g., tumor samples, to elucidate the molecular composition of the overall biological signal in these samples; (iii) one of multiple chromosomes of the same organism, to illustrate the relation, if any, between these chromosomes in terms of their, e.g., mRNA expression in a given set of samples; and (iv) one of multiple interacting organisms, e.g., in an ecosystem, to illuminate the exchange of biological information in these interactions.

## Supplementary Information

**Appendix S1.** A PDF format file, readable by Adobe Acrobat Reader.

**Mathematica Notebook S1. Higher-order generalized singular value decomposition (HO GSVD) of global mRNA expression datasets from three different organisms.** A Mathematica 5.2 code file, executable by Mathematica 5.2 and readable by Mathematica Player, freely available at <http://www.wolfram.com/products/player/>.

**Mathematica Notebook S2. HO GSVD of global mRNA expression datasets from three different organisms.** A PDF format file, readable by Adobe Acrobat Reader.

**Dataset S1. *S. pombe* global mRNA expression.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of  $m_1=3167$  *S. pombe* gene clones at  $n=17$  time points during about two cell-cycle periods from Rustici *et al.* [15] with the cell-cycle classifications of Rustici *et al.* or Oliva *et al.* [16].

**Dataset S2. *S. cerevisiae* global mRNA expression.** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of  $m_2=4772$  *S. cerevisiae* open reading frames (ORFs), or genes, at  $n=17$  time points during about two cell-cycle periods, including cell-cycle classifications, from Spellman *et al.* [17].

**Dataset S3. Human global mRNA expression** A tab-delimited text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of  $m_3=13,068$  human genes at  $n=17$  time points during about two cell-cycle periods, including cell-cycle classifications, from Whitfield *et al.* [18].

## Acknowledgements

We thank GH Golub for introducing us to matrix and tensor computations, and the American Institute of Mathematics in Palo Alto and Stanford University for hosting the 2004 Workshop on Tensor Decompositions and the 2006 Workshop on Algorithms for Modern Massive Data Sets, respectively, where some of this work was done. We also thank CH Lee for technical assistance, RA Horn for helpful discussions of matrix analysis and careful reading of the manuscript, and L De Lathauwer and A Goffeau for helpful comments.

## References

1. Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature* 443: 594-597.
2. Lu Y, Huggins P, Bar-Joseph Z (2009) Cross species analysis of microarray expression data. *Bioinformatics* 25: 1476-1483.
3. Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc Natl Acad Sci USA* 93: 10268-10273.
4. Golub GH, Van Loan CF (1996) *Matrix Computations*. Baltimore: Johns Hopkins University Press, third edition, 694 p.
5. Van Loan CF (1976) Generalizing the singular value decomposition. *SIAM J Numer Anal* 13: 76-83.
6. Paige CC, Saunders MA (1981) Towards a generalized singular value decomposition. *SIAM J Numer Anal* 18: 398-405.
7. Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci USA* 100: 3351-3356.
8. Alter O (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci USA* 103: 16063-16064.
9. Omberg L, Meyerson JR, Kobayashi K, Drury LS, Diffley JF, et al (2009) Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol Syst Biol* 5: 312.

10. Alter O, Golub GH (2004) Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proc Natl Acad Sci USA* 101: 16577-16582.
11. Omberg L, Golub GH, Alter O (2007) A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc Natl Acad Sci USA* 104: 18371-18376.
12. De Lathauwer L, De Moor B, Vandewalle J (2000) A multilinear singular value decomposition. *SIAM J Matrix Anal Appl* 21: 1253-1278.
13. Vandewalle J, De Lathauwer L, Comon P (2003) The generalized higher order singular value decomposition and the oriented signal-to-signal ratios of pairs of signal tensors and their use in signal processing. In: *Proc ECCTD'03 - European Conf on Circuit Theory and Design*. pp. I-389-I-392.
14. Alter O, Golub GH (2005) Reconstructing the pathways of a cellular system from genome-scale signals using matrix and tensor computations. *Proc Natl Acad Sci USA* 102: 17559-17564.
15. Rustici G, Mata J, Kivinen K, Lió P, Penkett CJ, et al (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat Genet* 36: 809-817.
16. Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, et al (2005) The cell cycle-regulated genes of *Schizosaccharomyces pombe*. *PLoS Biol* 3: e225.
17. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, et al (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 9: 3273-3297.
18. Whitfield ML, Sherlock G, Saldanha A, Murray JI, Ball CA, et al (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* 13: 1977-2000.
19. Gauthier NP, Larsen ME, Wernersson R, Brunak S, Jensen TS (2010) Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. *Nucleic Acids Res* 38: D699-D702.
20. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403-410.
21. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35: D61-D65.
22. Decottignies A, Goffeau A (1997) Complete inventory of the yeast ABC proteins. *Nat Genet* 15: 137-145.
23. Mamnun YM, Schüller C, Kuchler K (2004) Expression regulation of the yeast PDR5 ATP-binding cassette (ABC) transporter suggests a role in cellular detoxification during the exponential growth phase. *FEBS Lett* 559: 111-117.
24. Werner-Washburne M, Braun E, Johnston GC, Singer RA (1993) Stationary phase in the yeast *Saccharomyces cerevisiae*. *Microbiol Rev* 57: 383-401.
25. Meyers S, Schauer W, Balzi E, Wagner M, Goffeau A, et al (1992) Interaction of the yeast pleiotropic drug resistance genes PDR1 and PDR5. *Curr Genet* 21: 431-436.

26. Mahé Y, Parle-McDermott A, Nourani A, Delahodde A, Lamprecht A, et al (1996) The ATP-binding cassette multidrug transporter Snq2 of *Saccharomyces cerevisiae*: a novel target for the transcription factors Pdr1 and Pdr3. *Mol Microbiol* 20: 109-117.
27. Wolfger H, Mahé Y, Parle-McDermott A, Delahodde A, Kuchler K (1997) The yeast ATP binding cassette (ABC) protein genes PDR10 and PDR15 are novel targets for the Pdr1 and Pdr3 transcriptional regulators. *FEBS Lett* 418: 269-274.
28. Hlaváček O, Kučerová H, Harant K, Palková Z, Váchová L (2009) Putative role for ABC multidrug exporters in yeast quorum sensing. *FEBS Lett* 583: 1107-1113.
29. Lee KS, Patton JL, Fido M, Hines LK, Kohlwein SD, et al (1994) The *Saccharomyces cerevisiae* PLB1 gene encodes a protein required for lysophospholipase and phospholipase B activity. *J Biol Chem* 269: 19725-19730.
30. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2: 65-73.
31. Martin-Castellanos C, Labib K, Moreno S (1996) B-type cyclins regulate G1 progression in fission yeast in opposition to the p25rum1 cdk inhibitor. *EMBO J* 15: 839-849.
32. Fisher DL, Nurse P (1996) A single fission yeast mitotic cyclin B p34<sup>cdc2</sup> kinase promotes both S-phase and mitosis in the absence of G1 cyclins. *EMBO J* 15: 850-860.
33. Chu MT, Funderlic RE, Golub GH (1997) On a variational formulation of the generalized singular value decomposition. *SIAM J Matrix Anal Appl* 18: 1082-1092.
34. Rao CR (1973) *Linear Statistical Inference and Its Applications*. New York, NY: John Wiley & Sons, second edition, 656 p.
35. Rao CR (1984) Optimization of functions of matrices with applications to statistical problems. In: Rao PSRS, Sedransk J, editors, *W.G. Cochran's Impact on Statistics*, New York, NY: John Wiley & Sons. pp. 191-202.
36. Paige CC (1985) The general linear model and the generalized singular value decomposition. *Linear Algebra Appl* 70: 269-284.
37. Marshall AW, Olkin L (1990) Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae* 40: 89-93.
38. Horn RA, Johnson CR (1985) *Matrix Analysis*. Cambridge, UK: Cambridge University Press, 575 p.
39. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* 22: 281-285.

Table 1. Arraylets or left basis vectors.

| Dataset              | Arraylet | Overexpression |                       | Underexpression |                       |
|----------------------|----------|----------------|-----------------------|-----------------|-----------------------|
|                      |          | Annotation     | <i>P</i> -value       | Annotation      | <i>P</i> -value       |
| <i>S. pombe</i>      | 13       | G2             | $2.4 \times 10^{-10}$ | G1              | $1.0 \times 10^{-15}$ |
|                      | 14       | M              | $2.2 \times 10^{-21}$ | G2              | $1.3 \times 10^{-9}$  |
|                      | 15       | M              | $4.1 \times 10^{-13}$ | S               | $1.6 \times 10^{-17}$ |
|                      | 16       | G2             | $5.2 \times 10^{-18}$ | G1              | $1.2 \times 10^{-26}$ |
|                      | 17       | G2             | $2.4 \times 10^{-10}$ | S               | $5.3 \times 10^{-35}$ |
| <i>S. cerevisiae</i> | 13       | S/G2           | $4.3 \times 10^{-15}$ | M/G1            | $1.4 \times 10^{-9}$  |
|                      | 14       | M/G1           | $4.9 \times 10^{-26}$ | G2/M            | $2.2 \times 10^{-12}$ |
|                      | 15       | G1             | $7.7 \times 10^{-17}$ | S               | $1.3 \times 10^{-8}$  |
|                      | 16       | G2/M           | $2.3 \times 10^{-38}$ | G1              | $2.0 \times 10^{-32}$ |
|                      | 17       | G2/M           | $2.3 \times 10^{-41}$ | G1              | $2.6 \times 10^{-40}$ |
| Human                | 13       | G1/S           | $1.1 \times 10^{-33}$ | G2              | $2.4 \times 10^{-44}$ |
|                      | 14       | M/G1           | $5.7 \times 10^{-3}$  | G2              | $4.7 \times 10^{-2}$  |
|                      | 15       | G2             | $9.8 \times 10^{-24}$ | None            | $1.4 \times 10^{-1}$  |
|                      | 16       | G1/S           | $9.8 \times 10^{-13}$ | G2              | $4.1 \times 10^{-4}$  |
|                      | 17       | G2             | $9.3 \times 10^{-33}$ | M/G1            | $2.7 \times 10^{-2}$  |

Probabilistic significance of the enrichment of the arraylets, i.e., HO GSVD patterns of expression variation across the *S. pombe*, *S. cerevisiae* and human genes, that span the common HO GSVD subspace in each dataset, in over- or underexpressed cell cycle-regulated genes. The *P*-value of each enrichment is calculated as described [39] (Supplementary Section 2.2) assuming hypergeometric distribution of the annotations (Supplementary Datasets S1–S3) among the genes, including the  $m=100$  genes most over- or underexpressed in each arraylet.

# Appendix C

## 1. Mathematical Framework: HO GSVD

Comparative analyses of large-scale datasets promise to enhance our fundamental understanding of the data by distinguishing the similar from the dissimilar among these data. For example, comparative analyses of global mRNA expression from multiple model organisms promise to enhance fundamental understanding of the universality and specialization of molecular biological mechanisms, and may prove useful in medical diagnosis, treatment and drug design [1]. Existing algorithms limit analyses to subsets of homologous genes among the different organisms, effectively introducing into the analysis the assumption that sequence and functional similarities are equivalent [2]. For sequence-independent comparisons [3], mathematical frameworks are required that can distinguish the similar from the dissimilar among multiple large-scale datasets tabulated as matrices with the same column dimension and different row dimensions, corresponding to the different sets of genes of the different organisms. The only such framework to date, that of the generalized singular value decomposition (GSVD) [4–6] is limited to two matrices.

Recently we showed that the GSVD provides a comparative mathematical framework for global mRNA expression datasets from two different organisms, tabulated as two matrices with the same column dimension and different row dimensions, where the mathematical variables and operations represent biological reality [7]. In this application, one matrix tabulates DNA microarray-measured genome-scale mRNA expression from the yeast *S. cerevisiae*, sampled at  $n$  time points at equal time intervals during the cell-cycle program. This matrix is of size  $m_1$ -*S. cerevisiae* genes  $\times$   $n$ -DNA microarrays. The second matrix tabulates data from the HeLa human cell line, sampled at the same number of time points, also at equal time intervals, and is of size  $m_2$ -human genes  $\times$   $n$ -arrays. The underlying assumption of the GSVD as a comparative mathematical framework for the two matrices is that there exists a one-to-one mapping among the columns of the matrices, but not necessarily among their rows. The GSVD factors each matrix into a product of an organism-specific matrix of size  $m_1$ -*S. cerevisiae* genes or  $m_2$ -human genes  $\times$   $n$ -“arraylets,” i.e., left basis vectors, an organism-specific diagonal matrix of size  $n$ -arraylets  $\times$   $n$ -“genelets,” i.e., right basis vectors, and a shared matrix of size  $n$ -genelets  $\times$   $n$ -arrays.

We showed that the mathematical variables of the GSVD, i.e., the patterns of the genelets and the two sets of arraylets, represent either the similar or the dissimilar among the biological programs that compose the *S. cerevisiae* and human datasets. Genelets of common significance in both datasets, and the corresponding arraylets, represent cell-cycle checkpoints that are common to *S. cerevisiae* and human. Simultaneous reconstruction and classification of both the *S. cerevisiae* and human data in the common subspace that these patterns span outlines the biological similarity in the regulation of their

cell-cycle programs. Patterns almost exclusive to either dataset correlate with either the *S. cerevisiae* or the human exclusive synchronization responses. Reconstruction of either dataset in the subspaces of the common vs. exclusive patterns represents differential gene expression in the *S. cerevisiae* and human conserved cell-cycle programs vs. their unique synchronization-response programs, respectively. Notably, relations such as these between the expression profiles of the *S. cerevisiae* genes *KAR4* and *CIK1*, which are known to be correlated in response to the synchronizing agent, the  $\alpha$ -factor pheromone, yet anticorrelated during cell division, are correctly depicted.

We now define a higher-order GSVD (HO GSVD) of  $N \geq 2$  datasets, tabulated as  $N$  real matrices  $D_i$  with the same column dimension and, in general, different row dimensions. In our form of the HO GSVD, each data matrix  $D_i$  is assumed to have full column rank and is factored as the product  $D_i = U_i \Sigma_i V^T$ , where  $U_i$  is the same shape as  $D_i$  (rectangular),  $\Sigma_i$  is diagonal and positive definite, and  $V$  is square and nonsingular. The columns of  $U_i$  have unit length; we call them the “left basis vectors” for  $D_i$  (a different set for each  $i$ ). The columns of  $V$  also have unit length; we call them “right basis vectors,” and they are the same in all factorizations. In the application of the HO GSVD to a comparison of global mRNA expression from  $N \geq 2$  organisms, the right basis vectors are the genelets and the  $N$  sets of left basis vectors are the  $N$  sets of arraylets. We use the notation:

$$U_i \equiv (u_{i,1} \dots u_{i,n}), \quad \|u_{i,k}\| = 1, \quad (1)$$

$$\Sigma_i \equiv \text{diag}(\sigma_{i,k}), \quad \sigma_{i,k} > 0, \quad (2)$$

$$V \equiv (v_1 \dots v_n), \quad \|v_k\| = 1. \quad (3)$$

We call  $\{\sigma_{i,k}\}$  the “higher-order generalized singular value set.” They are weights in the following sums of rank-one matrices of unit norm:

$$D_i = U_i \Sigma_i V^T = \sum_k \sigma_{i,k} u_{i,k} v_k^T, \quad \|u_{i,k} v_k^T\| = 1. \quad (4)$$

Hence we regard the  $k$ th values  $\sigma_{i,k}$  as indicating the significance of the  $k$ th right basis vector  $v_k$  in the matrices  $D_i$  (reflecting the overall information that  $v_k$  captures in each  $D_i$  in turn). To obtain the factorizations, we work with the matrices  $A_i = D_i^T D_i$  and the matrix sum  $S$ , defined as the arithmetic mean of all pairwise quotients  $A_i A_j^{-1}$ , or equivalently of all  $S_{ij} = \frac{1}{2}(A_i A_j^{-1} + A_j A_i^{-1})$ ,  $i \neq j$ . The eigensystem  $SV = V\Lambda$  is used to define  $V$ , and the factors  $U_i$  and  $\Sigma_i$  are computed from  $D_i$  and  $V$ .

To clarify our choice of  $S$ , we note that in the GSVD, defined by Van Loan [4], the matrix  $V$  can be formed from the eigenvectors of the unbalanced quotient  $A_1 A_2^{-1}$  (Supplementary Section 1.1). We observe that this  $V$  can also be formed from the eigenvectors of the balanced arithmetic mean  $S_{12} = \frac{1}{2}(A_1 A_2^{-1} + A_2 A_1^{-1})$ . We prove that in the case of  $N = 2$ , our definition of  $V$  by using the eigensystem of  $S \equiv S_{12} = \frac{1}{2}(A_1 A_2^{-1} + A_2 A_1^{-1})$  leads algebraically to the GSVD and therefore, as Paige and

dx.doi.org/10.1371/journal.pone.0028072 | A-2

Ponnappalli, Saunders, Van Loan & Alter (2011)

Saunders showed [5], can be computed in a stable way. We also note that in the GSVD, the matrix  $V$  is invariant under the exchange of the two matrices  $D_1$  and  $D_2$ .

Therefore, we define our HO GSVD for  $N \geq 2$  matrices by using the balanced arithmetic mean  $S$  of all pairwise arithmetic means  $S_{ij}$ , each of which defines the GSVD of the corresponding pair of matrices  $D_i$  and  $D_j$ , noting that  $S$  is invariant under the exchange of any two matrices  $D_i$  and  $D_j$ .

We also show that the existing SVD and GSVD decompositions are in some sense special cases of our HO GSVD (Supplementary Section 1.2). Finally, we conjecture a role for our exact HO GSVD in iterative approximation algorithms (Supplementary Section 1.3).

### 1.1. The Matrix GSVD

**1.1.1. Construction of the matrix GSVD.** Suppose we have two real matrices  $D_1 \in m_1 \times n$  and  $D_2 \in m_2 \times n$  each with full column rank. Van Loan [4] defined the GSVD of  $D_1$  and  $D_2$  as

$$\begin{aligned} U_1^T D_1 X &\equiv \Sigma_1, \\ U_2^T D_2 X &\equiv \Sigma_2, \end{aligned} \quad (5)$$

where each  $U_i \in m_i \times n$  has orthonormal columns,  $X \in n \times n$  is nonsingular, and the  $\Sigma_i = \text{diag}(\sigma_{i,k}) \in n \times n$  are diagonal with  $\sigma_{i,k} > 0$  ( $i = 1, 2$ ). Paige and Saunders [5] showed that the GSVD can be computed in a stable way by orthogonal transformations. In the full column-rank case it takes the form

$$\begin{aligned} D_1 &\equiv U_1 \Sigma_1 V^T, \\ D_2 &\equiv U_2 \Sigma_2 V^T, \end{aligned} \quad (6)$$

where  $U_1$ ,  $U_2$  and  $\begin{bmatrix} \Sigma_1 \\ \Sigma_2 \end{bmatrix}$  have orthonormal columns [6], and  $V$  is square and nonsingular.

We work with the form of Supplementary Equation (6), but we find it useful and more similar to the standard SVD [6] if we assume that the columns of  $V$  are scaled to have unit length, with the columns of  $\Sigma_i$  scaled accordingly. Note that the ratios  $\sigma_{1,k}/\sigma_{2,k}$  are not altered by the scaling.

In place of the methods of Van Loan [4] and Paige and Saunders [5], we construct the GSVD of Supplementary Equation (6) as follows. We obtain  $V$  from the eigensystem of  $S$ , the arithmetic mean of the quotients  $A_1 A_2^{-1}$  and  $A_2 A_1^{-1}$  of the matrices  $A_1 = D_1^T D_1$  and  $A_2 = D_2^T D_2$ :

$$\begin{aligned} S &\equiv S_{12} = \frac{1}{2}(A_1 A_2^{-1} + A_2 A_1^{-1}), \\ SV &= V\Lambda, \\ V &\equiv (v_1 \dots v_n), \quad \Lambda = \text{diag}(\lambda_k), \end{aligned} \quad (7)$$

with  $\|v_k\| = 1$ . Given  $V$ , we compute matrices  $B_i$  by solving two linear systems

$$\begin{aligned} V B_i^T &= D_i^T, \\ B_i &\equiv (b_{i,1} \dots b_{i,n}), \quad i = 1, 2, \end{aligned} \quad (8)$$

and we construct  $\Sigma_i$  and  $U_i = (u_{i,1} \dots u_{i,n})$  by normalizing the columns of  $B_i$ :

$$\begin{aligned} \sigma_{i,k} &= \|b_{i,k}\|, \\ \Sigma_i &= \text{diag}(\sigma_{i,k}), \\ B_i &= U_i \Sigma_i. \end{aligned} \quad (9)$$

We prove below that  $S$  is nondefective (it has  $n$  independent eigenvectors) and its eigensystem is real.

From Supplementary Equations (8) and (9) we have  $D_i = B_i V^T = U_i \Sigma_i V^T$  as in Supplementary Equation (6). We see that the rows of both  $D_1$  and  $D_2$  are superpositions of the same right basis vectors, the columns of  $V$  (Figure 37). This is the construction that we generalize in Equations (1)–(4) to compute our HO GSVD.

### 1.1.2. Interpretation of the GSVD construction.

In our GSVD comparison of two matrices, we interpreted the  $k$ th diagonals of  $\Sigma_1$  and  $\Sigma_2$ , i.e., the “generalized singular value pair”  $(\sigma_{1,k}, \sigma_{2,k})$ , as indicating the significance of the  $k$ th right basis vector  $v_k$  in the matrices  $D_1$  and  $D_2$ , and reflecting the overall information that  $v_k$  captures in  $D_1$  and  $D_2$  respectively [7]. The ratio  $\sigma_{1,k}/\sigma_{2,k}$  indicates the significance of  $v_k$  in  $D_1$  relative to its significance in  $D_2$ .

A ratio of  $\sigma_{1,k}/\sigma_{2,k} = 1$  corresponds to a basis vector  $v_k$  of equal significance in  $D_1$  and  $D_2$ . GSVD comparisons of two matrices showed that right basis vectors of approximately equal significance in both matrices reflect themes that are common to the two matrices under comparison [7].

A ratio of  $\sigma_{1,k}/\sigma_{2,k} \gg 1$  corresponds to a basis vector  $v_k$  of almost negligible significance in  $D_2$  relative to its significance in  $D_1$ . Likewise, a ratio of  $\sigma_{1,k}/\sigma_{2,k} \ll 1$  indicates a basis vector  $v_k$  of almost negligible significance in  $D_1$  relative to its significance in  $D_2$ . GSVD comparisons of two matrices showed that right basis vectors of negligible significance in one matrix reflect themes that are exclusive to the other matrix.

### 1.1.3. Mathematical properties of $\Lambda$ and $V$ in the GSVD construction.

Note that our GSVD construction in Supplementary Equations (7)–(9) is well defined for any square nonsingular  $V$ . We now show that our particular  $V$  leads algebraically to the GSVD of Supplementary Equation (6), ignoring the rescaled columns of  $\Sigma_i$  and  $V$ . Recall that  $A_i = D_i^T D_i$  and  $D_i = U_i \Sigma_i V^T$  with  $\Sigma_i$  diagonal.

In practice we would prefer not to form  $A_i$  or  $S$  directly. Instead we may work with the QR factorizations  $D_i = Q_i R_i$ , where  $Q_i^T Q_i = I$  and  $R_i$  is upper triangular and nonsingular. Define another triangular matrix  $R$  and its SVD to be

$$R \equiv R_1 R_2^{-1} = \tilde{U} \tilde{\Sigma} \tilde{V}^T, \quad (10)$$

where  $\tilde{U}$  and  $\tilde{V}$  are square and orthogonal, and  $\tilde{\Sigma} =$

Ponnappalli, Saunders, Van Loan & Alter (2011)

dx.doi.org/10.1371/journal.pone.0028072 | A-3

$\text{diag}(\tilde{\sigma}_k)$ ,  $\tilde{\sigma}_k > 0$ . We then have

$$\begin{aligned} S &= \frac{1}{2}[(R_1^T R_1)(R_2^T R_2)^{-1} + (R_2^T R_2)(R_1^T R_1)^{-1}], \\ R_1^{-T} S R_1^T &= \frac{1}{2}[R_1(R_2^T R_2)^{-1} R_1^T + R_1^{-T}(R_2^T R_2) R_1^{-1}] \\ &= \frac{1}{2}[R R^T + (R R^T)^{-1}] \\ &= \tilde{U} \Lambda \tilde{U}^T, \end{aligned} \quad (11)$$

where  $\Lambda \equiv \frac{1}{2}(\tilde{\Sigma}^2 + \tilde{\Sigma}^{-2}) \equiv \text{diag}(\lambda_k)$ . Thus

$$S(R_1^T \tilde{U}) = (R_1^T \tilde{U})\Lambda, \quad (12)$$

$$S V = V \Lambda, \quad V \equiv R_1^T \tilde{U} D, \quad (13)$$

where the diagonal matrix  $D$  normalizes  $R_1^T \tilde{U}$  so that  $V$  has columns of unit length.

**Supplementary Theorem S1.** The matrices  $U_1$  and  $U_2$  constructed in Supplementary Equation (9) have orthonormal columns ( $U_1^T U_1 = U_2^T U_2 = I$ ).

*Proof.* From Supplementary Equation (8) we have  $V B_i^T B_i V^T = D_i^T D_i$ , so that

$$\begin{aligned} (R_1^T \tilde{U} D) B_i^T B_i (D \tilde{U}^T R_1) &= R_i^T R_i \\ \Rightarrow D B_1^T B_1 D &= I, \\ D B_2^T B_2 D &= \tilde{U}^T R_1^{-T} (R_2^T R_2) R_1^{-1} \tilde{U} = \tilde{\Sigma}^{-2}, \end{aligned}$$

with help from Supplementary Equations (10), (12) and (13). We see that both  $B_i^T B_i$  are diagonal, and the quantities in Supplementary Equation (9) must be

$$U_1 = B_1 D, \quad \Sigma_1 = D^{-1}, \quad (14)$$

$$U_2 = B_2 D \tilde{\Sigma}, \quad \Sigma_2 = \tilde{\Sigma}^{-1} D^{-1}, \quad (15)$$

with  $U_i$  column-wise orthonormal.  $\square$

**Supplementary Theorem S2.** The eigenvalues of  $S$  satisfy  $\lambda_k \geq 1$ , and  $S$  has a full set of  $n$  eigenvectors (it is nondefective). Also, the eigenvectors are real.

*Proof.* The equivalence transformation of Supplementary Equation (11) shows that  $S$  has the same eigenvalues as  $\tilde{U} \Lambda \tilde{U}^T$ , namely  $\Lambda$ . From the definition of  $\Lambda$  and  $\tilde{\Sigma}$ , the eigenvalues are  $\lambda_k = \frac{1}{2}(\tilde{\sigma}_k^2 + 1/\tilde{\sigma}_k^2) \geq 1$ . Also, in the eigensystem of  $S$  in Supplementary Equation (13),  $V = R_1^T \tilde{U} D$  is a product of real nonsingular matrices and hence is real and nonsingular.  $\square$

**Supplementary Theorem S3.** An eigenvalue  $\lambda_k = 1$  corresponds to a right basis vector  $v_k$  of equal significance in both matrices  $D_1$  and  $D_2$ . That is,  $\sigma_{1,k}/\sigma_{2,k} = 1$ .

*Proof.* From the orthonormality of  $U_1$  and  $U_2$  of Equations (14) and (15), and our GSVD construction of Equations (7)–(9), we have  $\lambda_k = (\sigma_{1,k}^2/\sigma_{2,k}^2 + \sigma_{2,k}^2/\sigma_{1,k}^2)/2$ . Therefore,  $\lambda_k = 1$  can occur only if  $\sigma_{1,k}/\sigma_{2,k} = 1$ . In other words,  $\lambda_k = 1$  if the  $k$ th right basis vector  $v_k$  is equally significant in  $D_1$  and  $D_2$ .  $\square$

Note that the GSVD is a generalization of the SVD in that if one of the matrices is the identity matrix, the GSVD reduces to the SVD of the other matrix.

In Equations (1)–(4), we now define a HO GSVD and in Theorems 1–3 and Corollary 1 we show that this new decomposition extends to higher orders all of the mathematical properties of the GSVD except for complete column-wise orthogonality of the left basis vectors that form the matrices  $U_i$  for all  $i$ . We proceed in the same way as in Supplementary Equations (7)–(9).

**1.2. The Matrix GSVD and SVD as Special Cases**  
Let us now show that the GSVD and the standard SVD are special cases of our HO GSVD.

**Supplementary Theorem S4.** Suppose matrices  $D$  and  $D_j$  are real and have full column rank. The HO GSVD of  $N$  matrices satisfying  $D_i = D$  for all  $i \neq j$  reduces to the GSVD of the two matrices  $D$  and  $D_j$ .

*Proof.* Substituting  $D_i = D$  for all  $i \neq j$  in Equation (2), we obtain the matrix sum  $S = \frac{1}{N}[A A_j^{-1} + A_j A^{-1} + (N-2)I]$ , with  $A = D^T D$ . The eigenvectors of  $S$  are the same as the eigenvectors  $V$  of  $\frac{1}{2}(A A_j^{-1} + A_j A^{-1})$ , and Supplementary Theorem S2 shows that  $V$  exists. Solving two linear systems for  $B$  and  $B_j$  and normalizing the solutions,

$$\begin{aligned} V B^T &= D_i^T = D^T, & B &= U \Sigma, \\ V B_j^T &= D_j^T, & B_j &= U_j \Sigma_j, \end{aligned}$$

reduces the HO GSVD of Equation (1) to the GSVD of  $D$  and  $D_j$  of Supplementary Equation (6):

$$\begin{aligned} D_i &= D = U \Sigma V^T, & i &\neq j, \\ D_j &= U_j \Sigma_j V^T. \end{aligned}$$

Supplementary Theorem S1 shows that the columns of  $U$  and  $U_j$  are orthonormal.  $\square$

**Supplementary Theorem S5.** Suppose the matrix  $D_j$  is real and has full column rank. The HO GSVD of  $N$  matrices satisfying  $D_i = I$  for all  $i \neq j$  reduces to the SVD of  $D_j$ .

*Proof.* Substituting  $D_i = I$  for all  $i \neq j$  in Supplementary Equation (2), we obtain the matrix sum  $S = \frac{1}{N}[A_j + A_j^{-1} + (N-2)I]$ . The symmetry of  $S$  implies that its eigenvectors  $V$  exist and are orthonormal. Computing the matrix  $B_j$  from

$$V B_j^T = D_j^T, \quad B_j = U_j \Sigma_j,$$

gives  $D_j = U_j \Sigma_j V^T$ , and Supplementary Theorem S1 shows that the columns of  $U_j$  are orthonormal. Hence the factorization must be the SVD of  $D_j$  [6].  $\square$

### 1.3. Role in Approximation Algorithms

Recent research showed that several higher-order generalizations are possible for a given matrix decomposition,

each preserving only some but not all of the properties of the matrix decomposition [12, 13].

Our HO GSVD preserves the exactness as well as the diagonality of the matrix GSVD, i.e., all  $N$  matrix factorizations in Equation (1) are exact and all  $N$  matrices  $\Sigma_i$  are diagonal. In general, our HO GSVD does not preserve the orthogonality of the matrix GSVD, i.e., the matrices  $U_i$  in Equation (1) are not necessarily column-wise orthonormal. For some applications, however, one might want to preserve the orthogonality instead of the exactness of the matrix GSVD. An iterative approximation algorithm might be used to compute for a set of  $N > 2$  real matrices  $D_i \in m_i \times n$ , each with full column rank, an approximate decomposition

$$\begin{aligned} D_1 &\approx U_1 \Sigma_1 V^T, \\ D_2 &\approx U_2 \Sigma_2 V^T, \\ &\vdots \\ D_N &\approx U_N \Sigma_N V^T, \end{aligned} \quad (16)$$

where each  $U_i \in m_i \times n$  is composed of orthonormal columns, each  $\Sigma_i = \text{diag}(\sigma_{i,k}) \in n \times n$  is diagonal with  $\sigma_{i,k} > 0$ , and  $V$  is identical in all  $N$  matrix factorizations.

If there exist an exact decomposition of Equation (1) where the matrices  $U_i$  are column-wise orthonormal, then it is reasonable to expect that the iterative approximation algorithm will converge to that exact decomposition. More than that, when the iterative approximation algorithm is initialized with the exact decomposition, it is reasonable to expect convergence in just one iteration. We show below that if there exist an exact decomposition of Equation (1) in which the matrices  $U_i$  are column-wise orthonormal, our HO GSVD leads algebraically to that exact decomposition.

**Supplementary Theorem S6.** *If there exist an exact HO GSVD of Equation (1) where the matrices  $U_i$  are column-wise orthonormal, then our particular  $V$  leads algebraically to the exact decomposition.*

*Proof.* If there exist an exact decomposition with column-wise orthonormal  $U_i$  for all  $i$ , then  $A_i = D_i^T D_i = V \Sigma_i^2 V^T$  and the right basis vectors, i.e., the columns of  $V$ , simultaneously diagonalize all pairwise quotients  $A_i A_j^{-1} V = V \Sigma_i^2 \Sigma_j^{-2}$  as well as their arithmetic mean  $SV = V\Lambda$ .

Therefore, the  $V$  of the eigensystem of  $S$  in Equation (2) is equivalent to the  $V$  of the HO GSVD of Equation (1) with column-wise orthonormal  $U_i$  for all  $i$ .  $\square$

We conjecture, therefore, the following role for our exact HO GSVD in iterative approximation algorithms.

**Supplementary Conjecture S1.** *An iterative approximation algorithm will converge to the optimal approximate decomposition of Supplementary Equation (16) in a significantly reduced number of iterations when initialized with our exact HO GSVD, rather than with random  $U_i$ ,  $\Sigma_i$  and  $V$ .*

## 2. Biological Application: Comparison of Global mRNA Expression Datasets from Three Different Organisms

### 2.1. *S. pombe*, *S. cerevisiae* and Human Datasets

Rustici *et al.* [15] monitored mRNA levels in the yeast *Schizosaccharomyces pombe* over about two cell-cycle periods, in a culture synchronized initially by the *cdc25-22* block-release late in the cell-cycle phase  $G_2$ , relative to reference mRNA from an asynchronous culture, at 15min intervals for 240min. The *S. pombe* dataset we analyze (Supplementary Dataset S1) tabulates the ratios of gene expression levels for the  $m_1=3167$  gene clones with no missing data in at least 14 of the  $n=17$  arrays. Of these, the mRNA expression of 380 gene clones was classified as cell cycle-regulated by Rustici *et al.* or Oliva *et al.* [16].

Spellman *et al.* [17] monitored mRNA expression in the yeast *Saccharomyces cerevisiae* over about two cell-cycle periods, in a culture synchronized initially by the  $\alpha$ -factor pheromone in the cell-cycle phase  $M/G_1$ , relative to reference mRNA from an asynchronous culture, at 7min intervals for 112min. The *S. cerevisiae* dataset we analyze (Supplementary Dataset S2) tabulates the ratios of gene expression levels for the  $m_2=4772$  open reading frames (ORFs), or genes, with no missing data in at least 14 of the  $n=17$  arrays. Of these, the mRNA expression of 641 ORFs was traditionally or microarray-classified as cell cycle-regulated.

Whitfield *et al.* [18] monitored mRNA levels in the human HeLa cell line over about two cell-cycle periods, in a culture synchronized initially by a double thymidine-block in S-phase, relative to reference mRNA from an asynchronous culture, at 2hr intervals for 34hr. The human dataset we analyze (Supplementary Dataset S3) tabulates the ratios of gene expression levels for the  $m_3=13,068$  gene clones with no missing data in at least 14 of the  $n=17$  arrays. Of these, the mRNA expression of 787 gene clones was classified as cell cycle-regulated.

Of the 53,839, 81,124 and 222,156 elements in the *S. pombe*, *S. cerevisiae* and human data matrices, 2420, 2936 and 14,680 elements, respectively, i.e.,  $\sim 5\%$ , are missing valid data. SVD [6] is used to estimate the missing data as described [7]. In each of the data matrices, SVD of the expression patterns of the genes with no missing data uncovered 17 orthogonal patterns of gene expression, i.e., "eigengenes." The five most significant of these patterns, in terms of the fraction of the mRNA expression that they capture, are used to estimate the missing data in the remaining genes. For each of the three data matrices, the five most significant eigengenes and their corresponding fractions are almost identical to those computed after the missing data are estimated, with the corresponding correlations  $>0.95$  (Supplementary Mathematica Notebooks S1 and S2). This suggests that the five most significant eigengenes, as computed for the genes with no missing data, are valid patterns for estimation of missing data. This also indicates that this SVD estimation of missing data is robust to variations in the data selection cutoffs.

Ponnappalli, Saunders, Van Loan &amp; Alter (2011)

A-5 | alterlab.org/HO\_GSVD/

We compute the HO GSVD of the three data matrices after missing data estimation by using Equations (1)–(4). Following Theorem 3, we find that the approximately common HO GSVD subspace of the three datasets is spanned by the five genelets  $v_k$   $k = 13, \dots, 17$  (Figures 12 and 15).

## 2.2. Common Subspace Interpretation

In analogy with the GSVD [7], a common HO GSVD genelet and the  $N = 3$  corresponding arraylets, which are approximately orthonormal to all other arraylets in the corresponding datasets (Theorem 3), are inferred to represent a biological process common to the three organisms and the corresponding cellular states when a consistent biological or experimental theme is reflected in the interpretations of the patterns of the genelet and the arraylets.

A genelet  $v_k$  is associated with a biological process or an experimental artifact when its pattern of expression variation across the arrays, i.e., across time, is interpretable. An arraylet  $u_{i,k}$  is parallel- and antiparallel-associated with the most likely parallel and antiparallel cellular states according to the annotations of the two groups of  $m$  genes each, with largest and smallest levels of expression in this arraylet among all  $m_i$  genes, respectively. The  $P$ -value of a given association, i.e.,  $P(y; m, m_i, z)$ , is calculated assuming hypergeometric probability distribution of the  $z$  annotations among the  $m_i$  genes, and of the subset of  $y \subseteq z$  annotations among the subset of  $m$  genes, as described [39],

$$P(y; m, m_i, z) = \binom{m_i}{m}^{-1} \sum_{x=y}^m \binom{z}{x} \binom{m_i - z}{m - x}. \quad (17)$$

We find that the approximately common HO GSVD subspace represents cell-cycle mRNA expression in the three disparate organisms (Figure 12 and Table 1).

## 2.3. HO GSVD Data Reconstruction

The decoupling of the HO GSVD genelets and  $N$  sets of arraylets, i.e., the diagonality of the matrices  $\Sigma_i$  in Equation (1), allows simultaneous reconstruction of the  $N = 3$  datasets in the common HO GSVD subspace without eliminating genes or arrays.

In analogy with the GSVD [7], given a common HO GSVD subspace that is spanned by the  $K$  genelets  $\{v_k\}$  where  $k = n - K + 1, \dots, n$ , we reconstruct each dataset in terms of only these genelets and the corresponding  $\{u_{i,k}\}$  arraylets,

$$\begin{aligned} D_i &= U_i \Sigma_i V^T = \sum_{k=1}^n \sigma_{i,k} u_{i,k} v_k^T \\ &\rightarrow \sum_{k=n-K+1}^n \sigma_{i,k} u_{i,k} v_k^T. \end{aligned} \quad (18)$$

Note that this reconstruction is mathematically equivalent to setting to zero the higher-order generalized

singular values  $\{\sigma_{i,k}\}$  in  $\Sigma_i$  for all  $k \neq n - K + 1, \dots, n$ , and then multiplying the matrices  $U_i \Sigma_i V^T$  to obtain the reconstructed  $D_i$ .

## 2.4. Simultaneous HO GSVD Classification

Identifying the subset of genelets and the corresponding arraylets that span the approximately common HO GSVD subspace allows simultaneous classification of the genes and arrays of the three datasets by similarity in their expression of these genelets or corresponding arraylets, respectively, rather than by their overall expression, as described [7].

We least squares-approximate the  $K=5$ -dimensional subspace spanned by the five genelets  $v_k$  with  $k = 13, \dots, 17$ , with the two-dimensional space spanned by two of the three orthonormal vectors  $x$ ,  $y$  and  $z \in \mathbb{R}^n$  that maximize the norm  $\sum_{k=n-K+1}^n (\|v_k^T x\|^2 + \|v_k^T y\|^2 + \|v_k^T z\|^2)$  (Supplementary Figure 16). Since the common HO GSVD subspace represents cell-cycle mRNA expression, the two vectors that we select to approximate the common subspace,  $x$  and  $y$ , describe expression oscillations of two periods in the three time courses.

We plot the projection of each gene of the dataset  $D_i$  from the  $K$ -genelets subspace onto  $y$ , i.e.,  $e_m^T \sum_{k=n-K+1}^n (\sigma_{i,k} u_{i,k} v_k^T) y / a_m$  where  $e_m$  is a unit  $m_i$ -vector, along the  $y$ -axis vs. that onto  $x$  along the  $x$ -axis, normalized by its ideal amplitude  $a_m$ , where the contribution of each genelet to the overall projected expression of the gene adds up rather than cancels out,  $a_m^2 = \sum_k \sum_j |\sigma_{i,k} \sigma_{i,j} (e_m^T u_{i,k} u_{i,j}^T e_m) v_k^T (xx^T + yy^T) v_j|$ . In this plot, the distance of each gene from the origin is the amplitude of its normalized projection. A unit amplitude indicates that the genelets add up; a zero amplitude indicates that they cancel out. The angular distance of each gene from the  $x$ -axis is its phase in the progression of expression across the genes from  $x$  to  $y$  and back to  $x$ , going through the projection of each genelet  $v_k$  in this subspace, i.e.,  $(xx^T + yy^T) v_k$ . Sorting the genes according to these angular distances gives the angular order, or classification, of the genes.

Similarly, we plot the projection of each array from the  $K$ -arraylets subspace onto  $\sum_{k=n-K+1}^n (u_{i,k} v_k^T) y$ , i.e.,  $y^T \sum_{k=n-K+1}^n (\sigma_{i,k} v_k v_k^T) e_n$  where  $e_n$  is a unit  $n$ -vector, along the  $y$ -axis vs. that onto  $\sum_{k=n-K+1}^n (u_{i,k} v_k^T) x$  along the  $x$ -axis, normalized by its ideal amplitude  $a_n$ , where the contribution of each arraylet to the overall projected expression of the array adds up rather than cancels out,  $a_n^2 = \sum_k \sum_j |\sigma_{i,k} \sigma_{i,j} (e_n^T v_k v_j^T e_n) v_k^T (xx^T + yy^T) v_j|$ . We sort the arrays according to their angular distances from the  $x$ -axis.

For classification, we set to zero the arithmetic mean of each genelet across the arrays, i.e., time, and that of each arraylet across the genes, such that the expression of each gene and array is centered at its time- or gene-invariant level, respectively.

With all 3167 *S. pombe*, 4772 *S. cerevisiae* and 13,068 human genes sorted, the expression variations of the five  $k = 13, \dots, 17$  arraylets from each organ-

ism approximately fit one period cosines, with the initial phase of each arraylet similar to that of its corresponding genelet. The global mRNA expression of each organism, reconstructed in the common HO GSVD subspace, approximately fits a traveling wave, oscillating across time and across the genes (Figures 17-19).

## 2.5. Sequence-Independence of the Classification

Our new HO GSVD provides a comparative mathematical framework for  $N \geq 2$  large-scale DNA microarray datasets from  $N$  organisms tabulated as  $N$  matrices that does not require a one-to-one mapping between the genes of the different organisms. The HO GSVD, therefore, can be used to identify genes of common function across different organisms independently of the sequence similarity among these genes, and to study, e.g., nonorthologous gene displacement [3]. The HO GSVD can also be used

to identify homologous genes, of similar DNA or protein sequences in one organism or across multiple organisms, that have different cellular functions.

We examine, for example, the HO GSVD classifications of genes of significantly different cell-cycle peak times [19] but highly conserved sequences [20, 21]. We consider three subsets of genes, the closest *S. pombe*, *S. cerevisiae* and human homologs of (i) the *S. pombe* gene *BFR1*, which belongs to the evolutionarily highly conserved ATP-binding cassette (ABC) transporter superfamily [22–28], (ii) the *S. cerevisiae* phospholipase B-encoding gene *PLB1* [29, 30], and (iii) the *S. pombe* strongly regulated S-phase cyclin-encoding gene *CIG2* [31, 32] (Supplementary Table S1). We find, notably, that these genes are correctly classified (Figure 14).

|     | Query Gene                                  | Gene              | Organism             | RefSeq ID      | Bit Score | E-value              |
|-----|---------------------------------------------|-------------------|----------------------|----------------|-----------|----------------------|
| (a) | Bfr1<br><i>S. pombe</i><br>NP_587932.3      | Snq2              | <i>S. cerevisiae</i> | NP_010294.1    | 1149      | 0                    |
|     |                                             | Pdr5              | <i>S. cerevisiae</i> | NP_014796.1    | 1103      | 0                    |
|     |                                             | Pdr18             | <i>S. cerevisiae</i> | NP_014468.1    | 1097      | 0                    |
|     |                                             | Pdr15             | <i>S. cerevisiae</i> | NP_010694.1    | 1093      | 0                    |
|     |                                             | Pdr12             | <i>S. cerevisiae</i> | NP_015267.1    | 1070      | 0                    |
|     |                                             | Pdr10             | <i>S. cerevisiae</i> | NP_014973.1    | 1029      | 0                    |
| (b) | Plb1<br><i>S. cerevisiae</i><br>NP_013721.1 | Plb2              | <i>S. cerevisiae</i> | NP_013719.1    | 825       | 0                    |
|     |                                             | Plb3              | <i>S. cerevisiae</i> | NP_014632.1    | 813       | 0                    |
|     |                                             | SPAC977.09c       | <i>S. pombe</i>      | NP_592772.1    | 385       | $7 \times 10^{-107}$ |
|     |                                             | SPAC1A6.03c       | <i>S. pombe</i>      | NP_593194.1    | 372       | $7 \times 10^{-103}$ |
|     |                                             | SPCC1450.09c      | <i>S. pombe</i>      | NP_588308.1    | 369       | $8 \times 10^{-102}$ |
|     |                                             | SPAC1786.02       | <i>S. pombe</i>      | NP_594024.1    | 355       | $1 \times 10^{-97}$  |
| (c) | Cig2<br><i>S. pombe</i><br>NP_593889.1      | Cdc13             | <i>S. pombe</i>      | NP_595171.1    | 346       | $3 \times 10^{-95}$  |
|     |                                             | Clb2              | <i>S. cerevisiae</i> | NP_015444.1    | 248       | $1 \times 10^{-65}$  |
|     |                                             | Cig1              | <i>S. pombe</i>      | NP_588110.2    | 241       | $1 \times 10^{-63}$  |
|     |                                             | Clb1              | <i>S. cerevisiae</i> | NP_011622.1    | 234       | $2 \times 10^{-61}$  |
|     |                                             | Clb4              | <i>S. cerevisiae</i> | NP_013311.1    | 222       | $5 \times 10^{-58}$  |
|     |                                             | Clb3              | <i>S. cerevisiae</i> | NP_010126.1    | 221       | $2 \times 10^{-57}$  |
|     |                                             | Ccnb2             | Human                | NP_004692.1    | 202       | $5 \times 10^{-52}$  |
|     |                                             | Clb6              | <i>S. cerevisiae</i> | NP_011623.1    | 183       | $4 \times 10^{-46}$  |
|     |                                             | Clb5              | <i>S. cerevisiae</i> | NP_015445.1    | 179       | $5 \times 10^{-45}$  |
|     |                                             | Ccnb1             | Human                | NP_114172.1    | 174       | $1 \times 10^{-43}$  |
|     |                                             | Rem1              | <i>S. pombe</i>      | NP_595798.1    | 160       | $2 \times 10^{-39}$  |
|     |                                             | Ccna1 (isoform c) | Human                | NP_001104516.1 | 159       | $7 \times 10^{-39}$  |
|     |                                             | Ccna1 (isoform a) | Human                | NP_003905.1    | 158       | $1 \times 10^{-38}$  |
|     |                                             | Ccna1 (isoform b) | Human                | NP_001104515.1 | 157       | $1 \times 10^{-38}$  |
|     |                                             | Ccna2             | Human                | NP_001228.1    | 145       | $6 \times 10^{-35}$  |

**Supplementary Table S1.** The closest *S. pombe*, *S. cerevisiae* and human homologs of (a) the *S. pombe* gene *BFR1*, (b) the *S. cerevisiae* gene *PLB1*, and (c) the *S. pombe* gene *CIG2*, as determined by an NCBI BLAST [20] of the

Ponnapalli, Saunders, Van Loan & Alter (2011)

[dx.doi.org/10.1371/journal.pone.0028072](https://doi.org/10.1371/journal.pone.0028072) | A-7

protein sequence that corresponds to each gene against the NCBI RefSeq database [21].

## WHAT IS CLAIMED IS:

1. A method, for medical characterization of a subject based on biological data, comprising:

applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

determining an indicator of a health parameter of a subject, the determining based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

2. The method of claim 1, wherein the index parameters comprise at least two of: patient identifications, tissue type identifications, a health status of one or more patients, a bioactive agent exposure status, an environmental exposure status, a nucleotide sequence copy numbers, DNA sequences, mRNA sequences, mRNA levels, a micro-RNA expression level, a DNA methylation level, a level of binding of proteins to DNA, a level of binding of proteins to RNA, gene product levels, gene product activity levels, a cell cycle status, a biochemical status, imaging data, a treatment status, biomarker levels, or time periods.

3. The method of claim 1, wherein the applying further comprises generating a diagonal matrix of singular values of each of A and B, and wherein the determining is further based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^TA$ .

4. The method of claim 1, wherein the eigenvectors of  $A^TA$  are the same as the eigenvectors of  $B^TB$ .

5. The method of claim 1, wherein the determining occurs at a first time, and further comprising repeating the determining at a second time to track a course of a health condition of the subject.

6. The method of claim 1, wherein at least one of the index parameters is measurable by at least one of a DNA microarray, DNA sequencing, a protein microarray, or mass spectrometry.

7. The method of claim 6, wherein the data comprises chromatin or histone modification, and wherein the data are derived from a patient-specific sample including at least one of a normal tissue, a disease-related tissue, or a culture of a patient's cell.

8. The method of claim 1, wherein the data comprise at least one of magnetic resonance imaging (MRI) data, electrocardiogram (ECG) data, electromyography (EMG) data, or electroencephalogram (EEG) data.

9. The method of claim 1, wherein the applying substantially removes from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

10. The method of claim 1, wherein the algorithm decomposes the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HOGSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

11. The method of claim 1, wherein the applying classifies the subject into a subgroup of patients based on at least patient-specific genomic data.

12. The method of claim 1, wherein the applying correlates an outcome of a therapeutic method and a genomic predictor in the data.

13. A system, for medical characterization of a subject based on biological data, comprising:

a processor configured to apply a decomposition algorithm to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^TA$ ,  $BB^T$ , and  $B^TB$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

an analysis module configured to determine an indicator of a health parameter of a subject, based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

14. The system of claim 13, wherein the processor is further configured to generate a diagonal matrix of singular values of each of submatrices A and B, and wherein the analysis module is further configured to determine the indicator of the health parameter based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^T A$ .

15. The system of claim 13, wherein the analysis module is further configured to determine the indicator of the health parameter at a first time, and to repeat the determination at a second time to track a course of a health condition of the subject.

16. The system of claim 13, wherein the processor is further configured to substantially remove from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

17. The system of claim 13, wherein the processor is further configured to apply the decomposition algorithm to decompose the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

18. The system of claim 13, wherein the processor is further configured to apply the decomposition algorithm to classify the subject into a subgroup of patients based on at least patient-specific genomic data.

19. The system of claim 13, wherein the processor is further configured to apply the decomposition algorithm to correlate an outcome of a therapeutic method and a genomic predictor in the data.

20. A non-transitory machine-readable medium comprising instructions that, when executed by one or more processors, perform the following acts:

applying a decomposition algorithm, by a processor, to an Nth-order tensor representing data, wherein  $N \geq 2$ , to generate, from at least two submatrices A and B of the tensor, eigenvectors of each of  $AA^T$ ,  $A^T A$ ,  $BB^T$ , and  $B^T B$ ;

wherein the data comprise indicators, represented in respective rows and columns of the tensor, of values of at least two index parameters; and

determining an indicator of a health parameter of a subject, the determining based on the eigenvectors and on values, associated with the subject, of the at least two index parameters;

wherein the health parameter comprises at least one of a differential diagnosis, a first health status of the subject, a disease subtype, at least one of an estimated probability or an estimated risk of a second health status of the subject, an indicator of a prognosis of the subject, or a predicted response to a treatment of the subject.

21. The machine-readable medium of claim 20, wherein the applying further comprises generating a diagonal matrix of singular values of each of A and B, and wherein the determining is further based on at least one of the diagonal matrices, wherein the singular values of A are the square roots of the eigenvalues of  $A^T A$ .

22. The machine-readable medium of claim 20, wherein the applying substantially removes from the data at least one of normal pattern copy number variations (CNVs) and an experimental variation.

23. The machine-readable medium of claim 20, wherein the algorithm decomposes the tensor according to at least one of a higher-order singular value decomposition (HOSVD), a higher-order generalized singular value decomposition (HO GSVD), a higher-order eigenvalue decomposition (HOEVD), or a parallel factor analysis (PARAFAC).

24. The machine-readable medium of claim 20, wherein the applying classifies the subject into a subgroup of patients based on at least patient-specific genomic data.

25. The machine-readable medium of claim 20, wherein the applying correlates an outcome of a therapeutic method and a genomic predictor in the data.

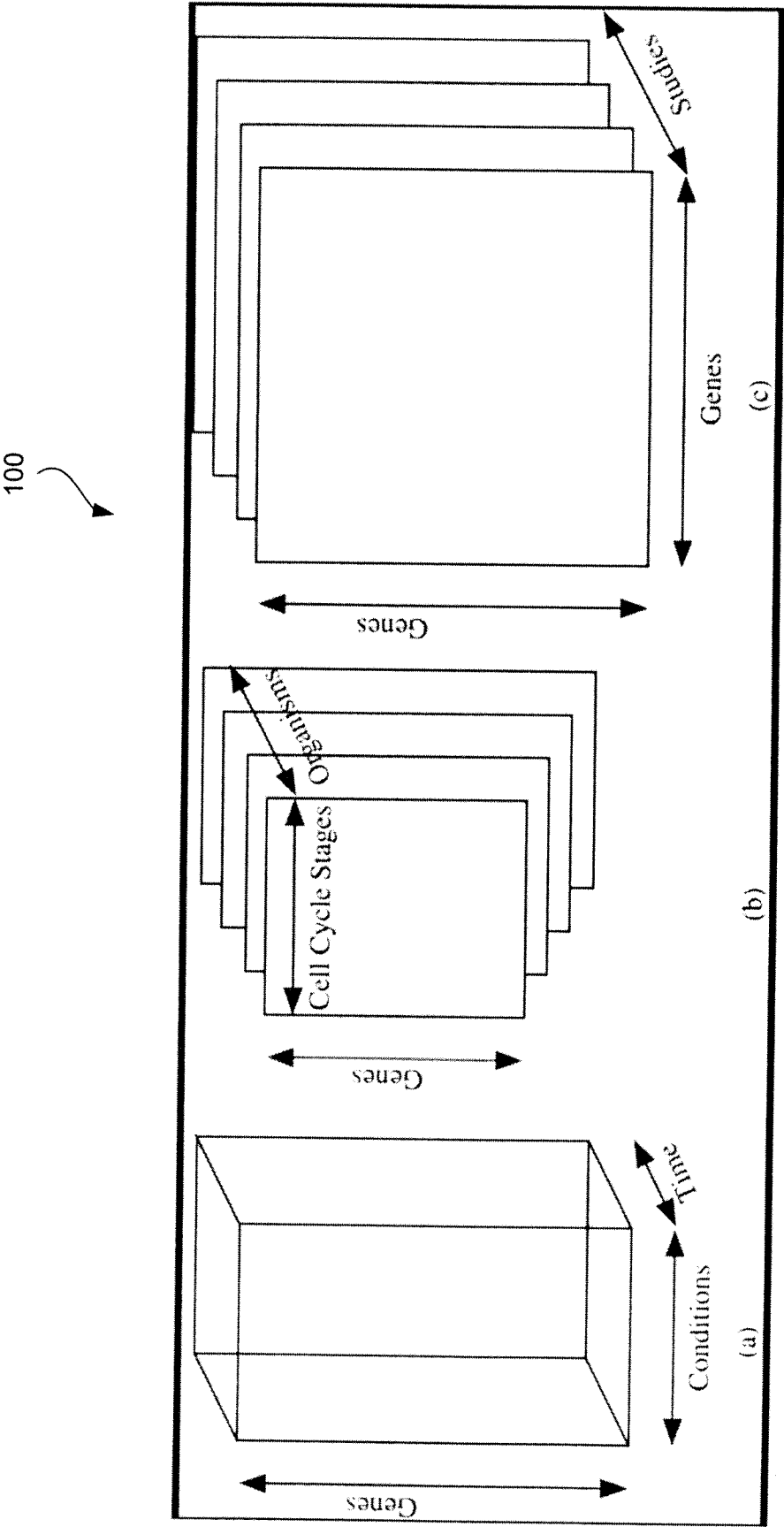


Figure 1

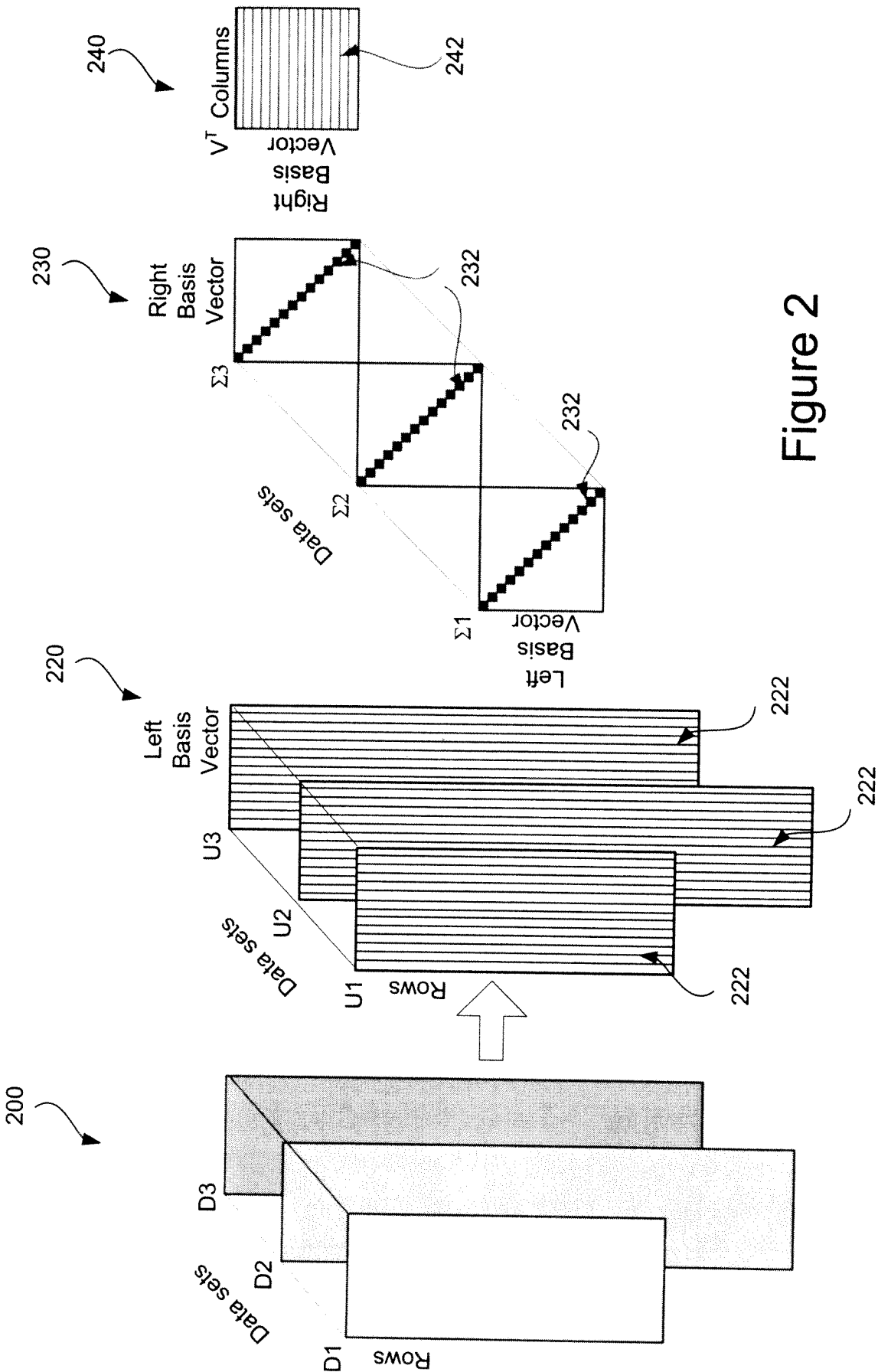


Figure 2

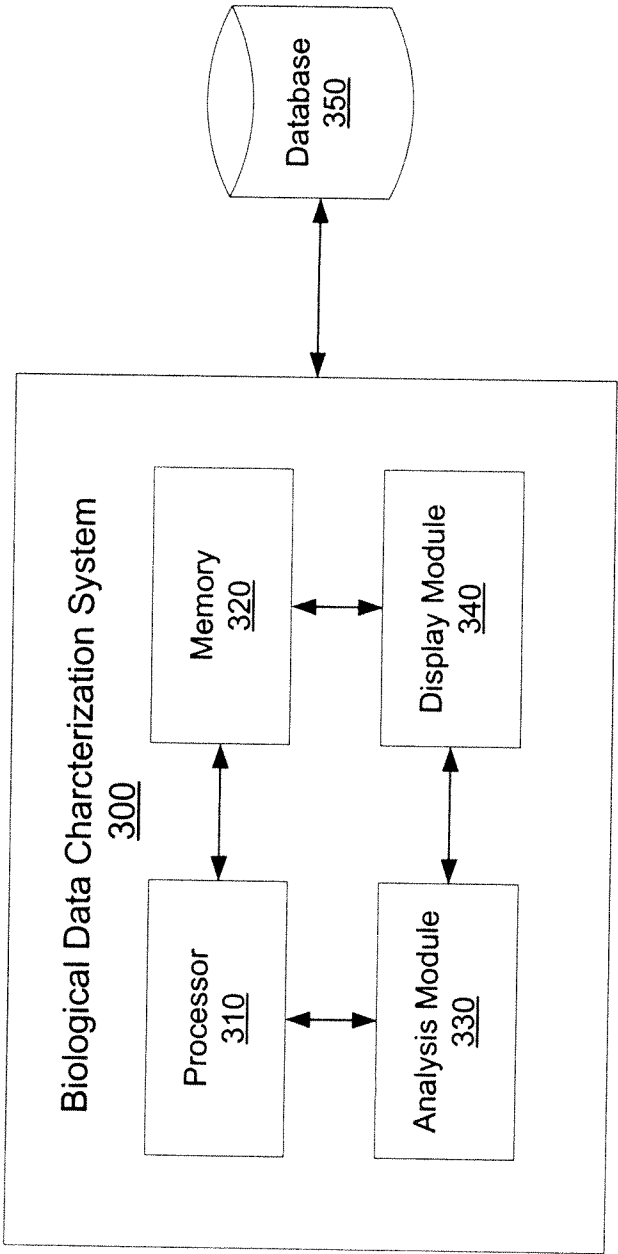


Figure 3

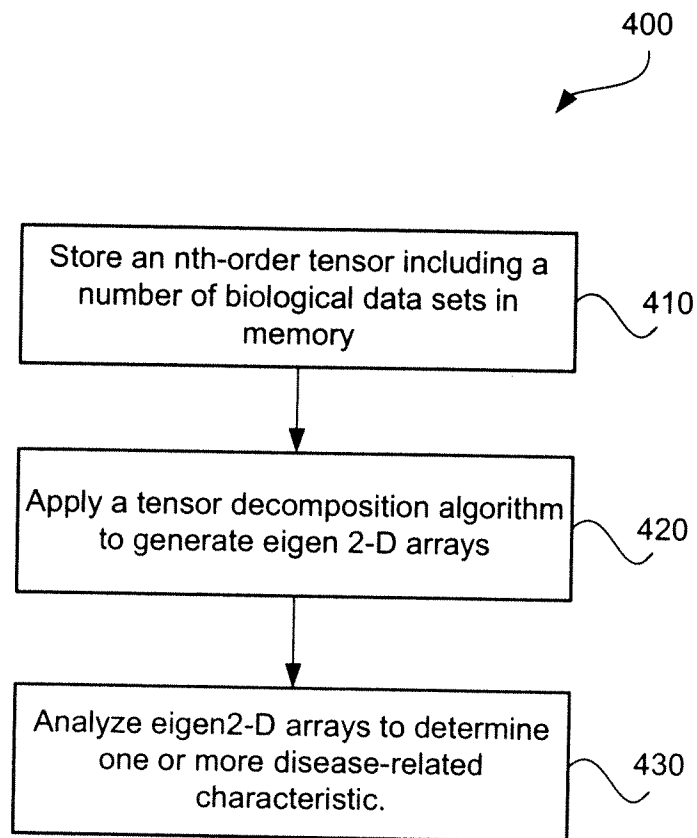


Figure 4

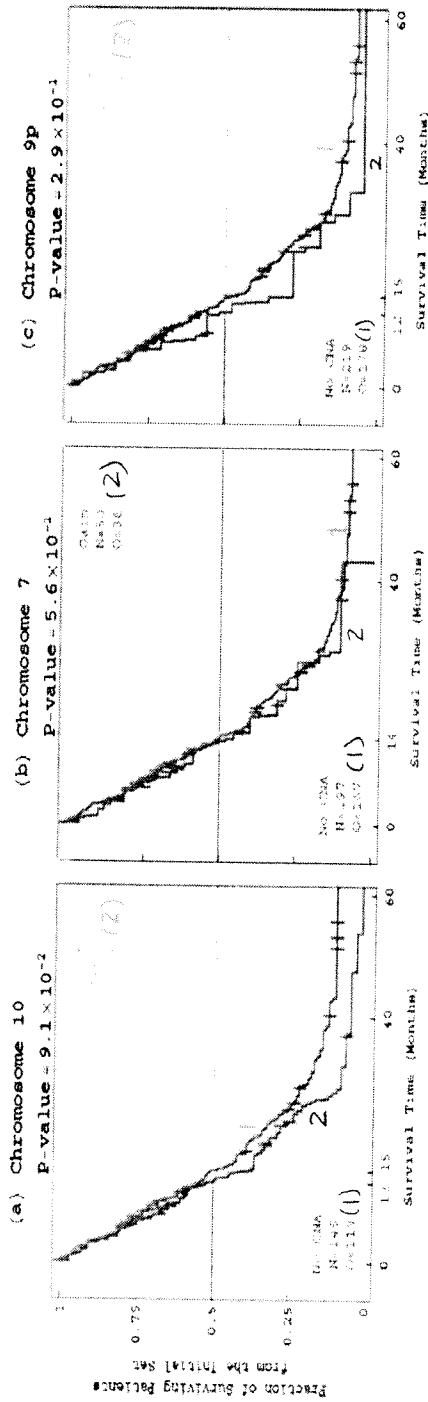


Figure 5A

Figure 5B

Figure 5C

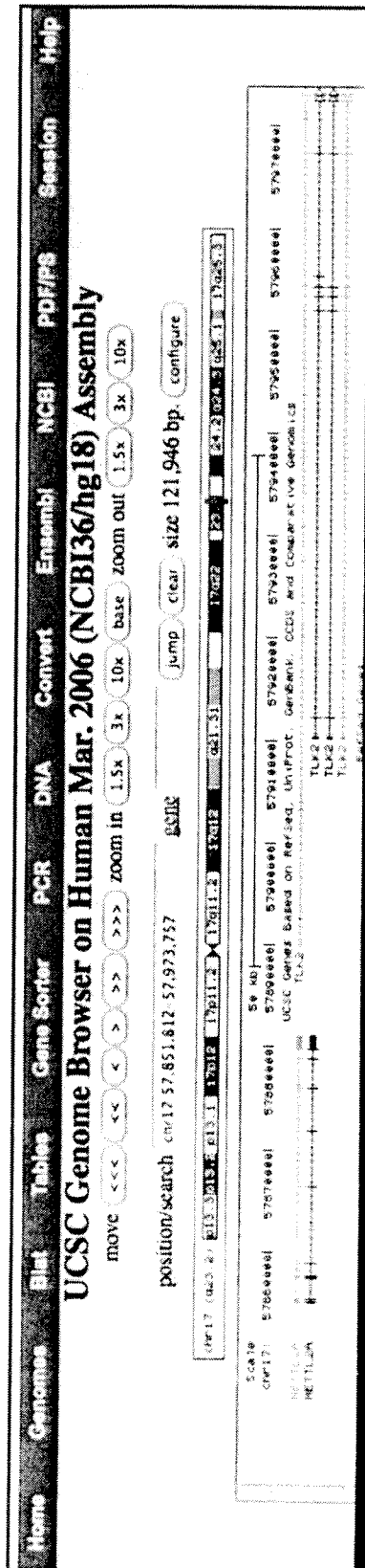


Figure 6

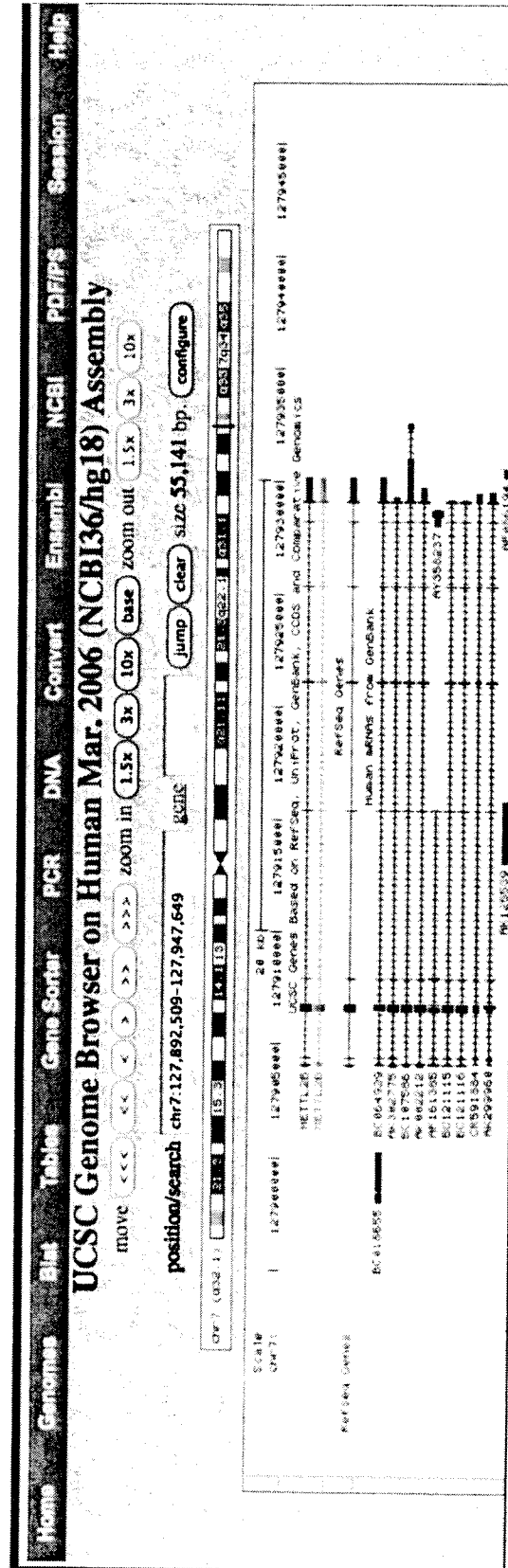
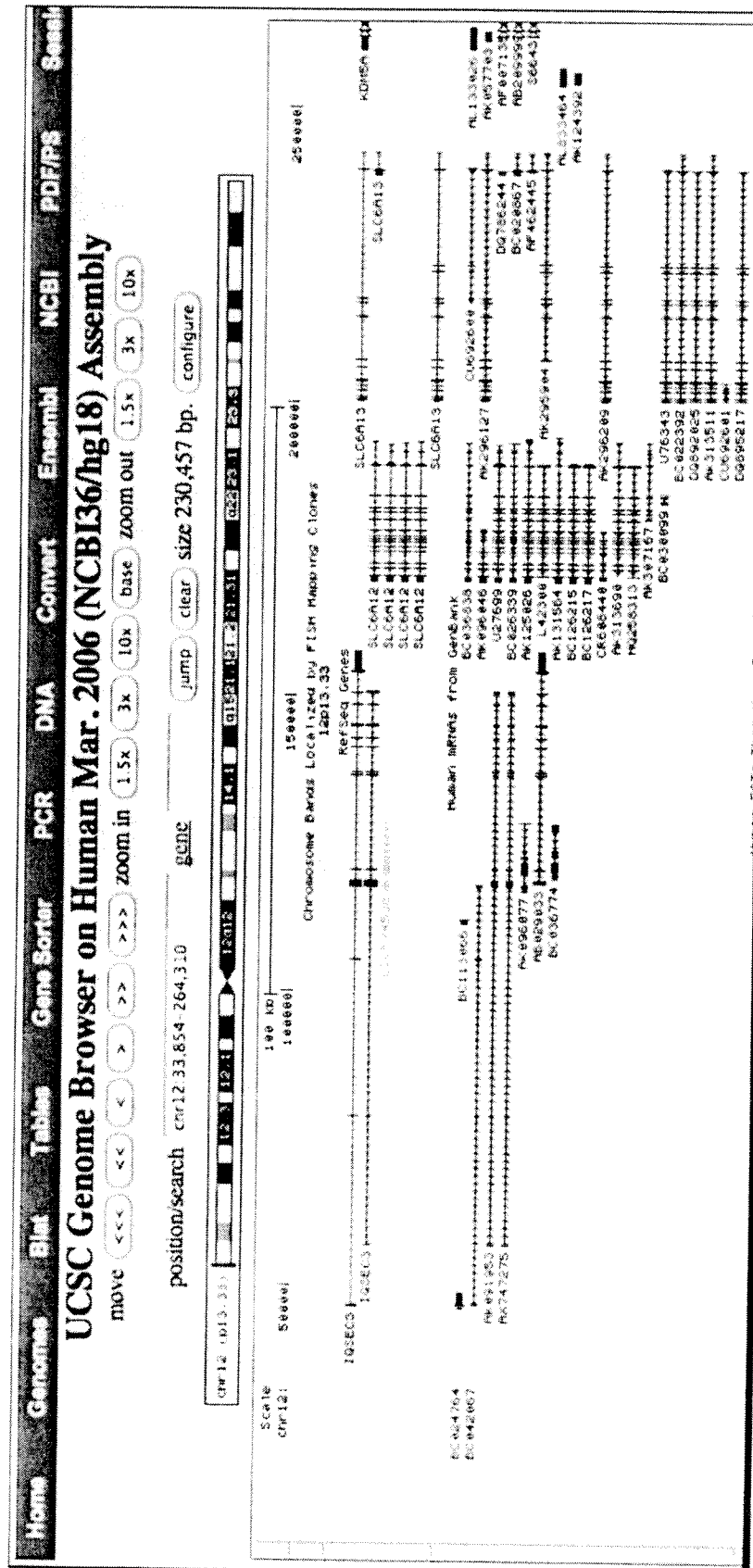


Figure 7



## Figure 8

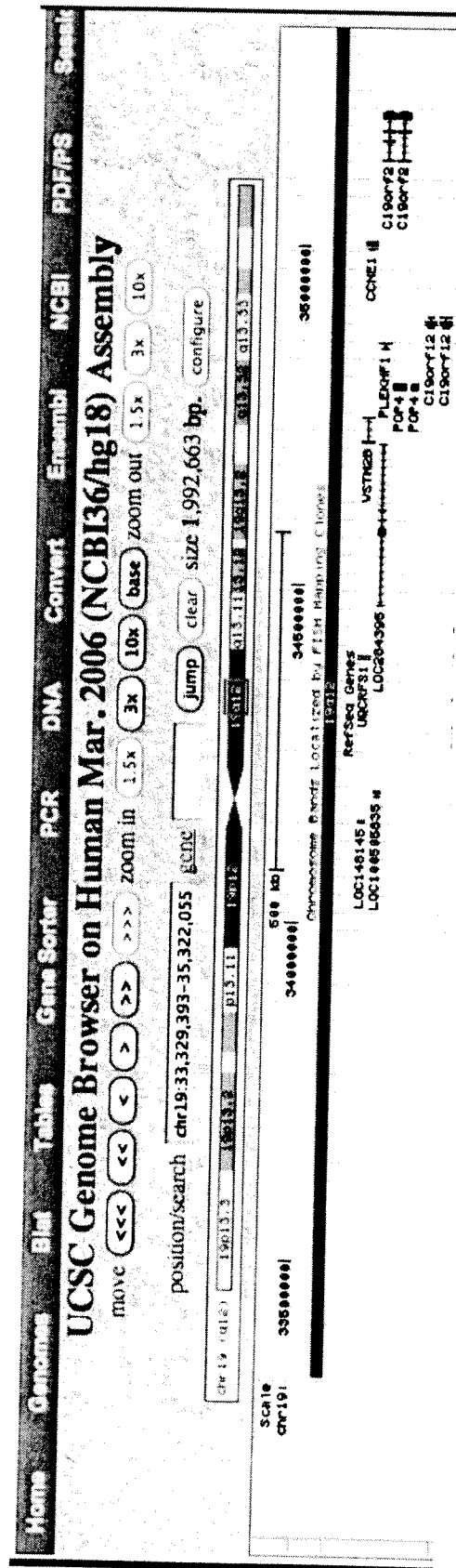


Figure 9

9/41

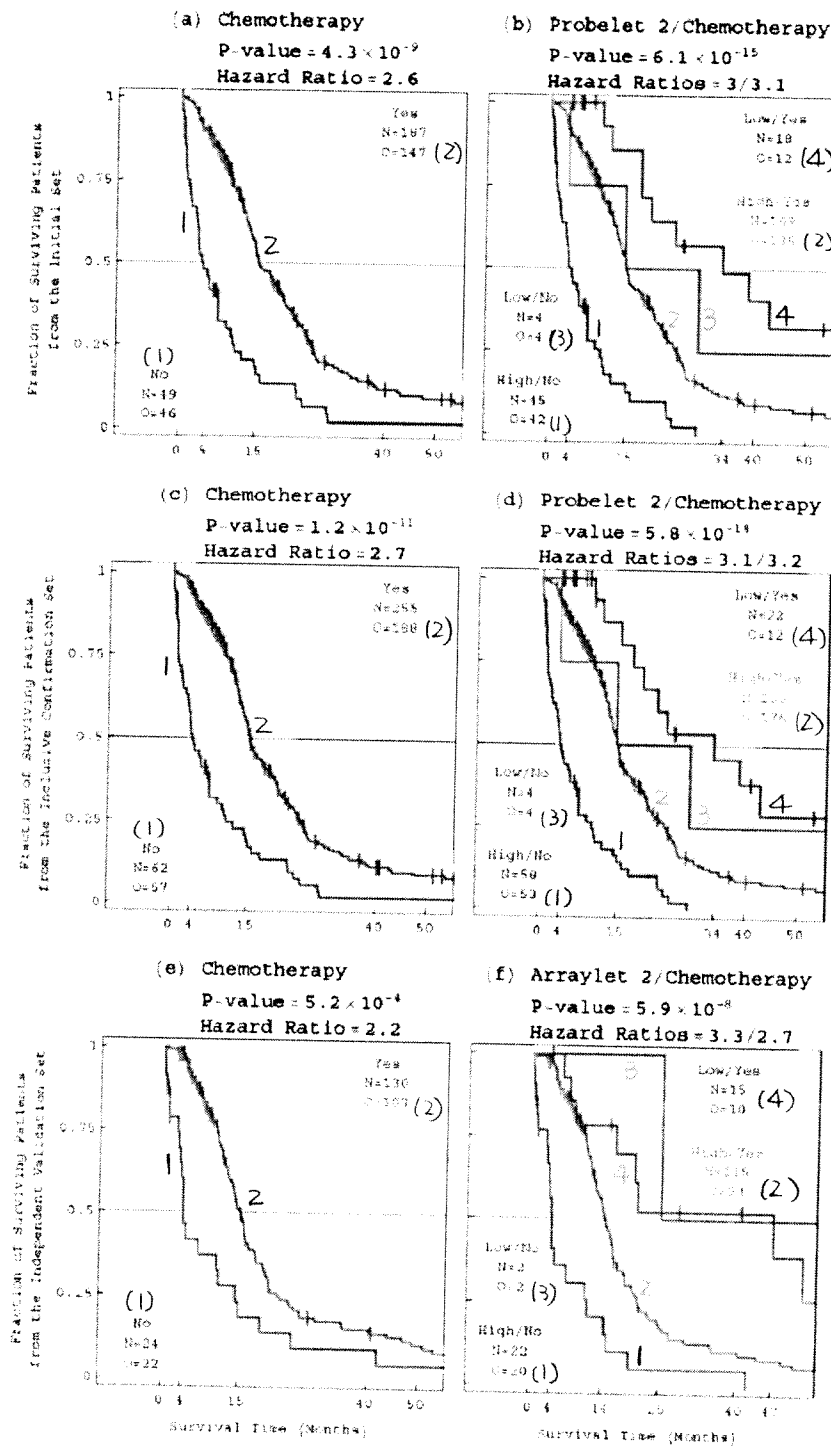


Figure 10

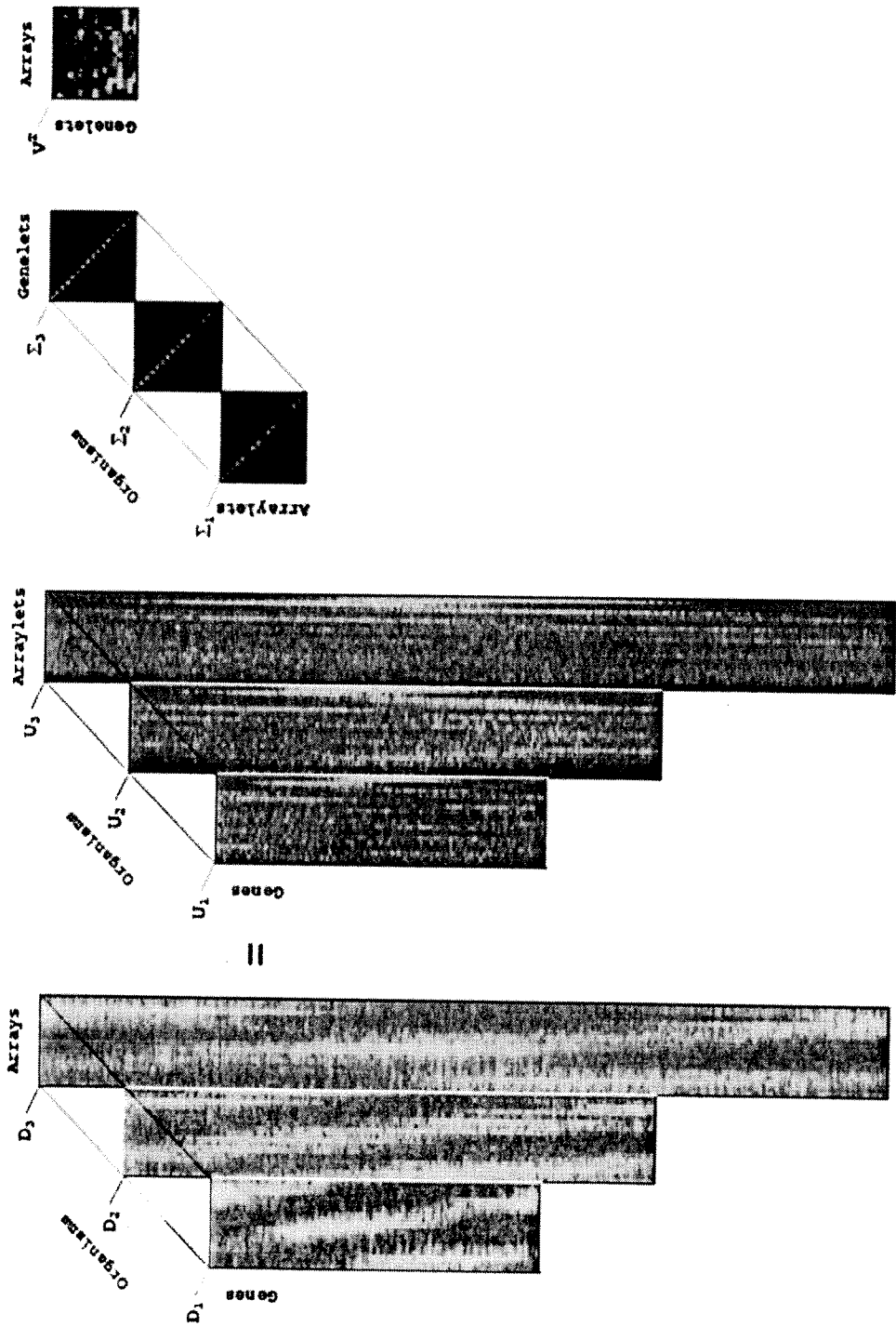


Figure 11

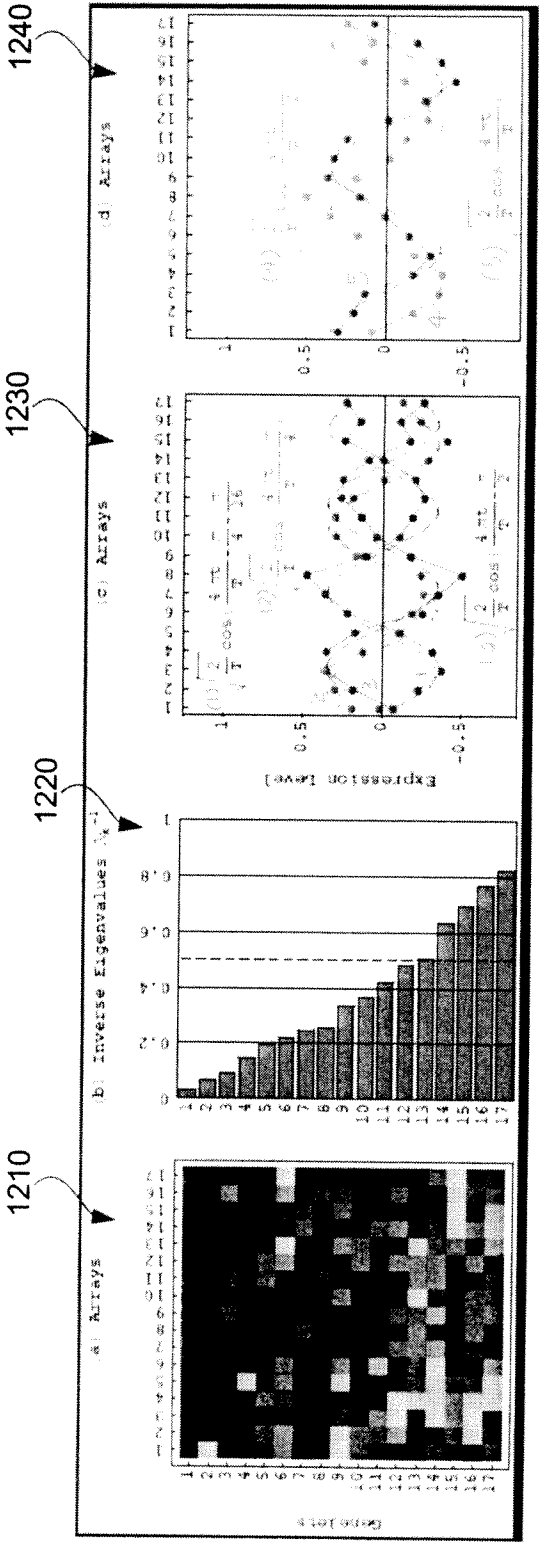


Figure 12

12/41

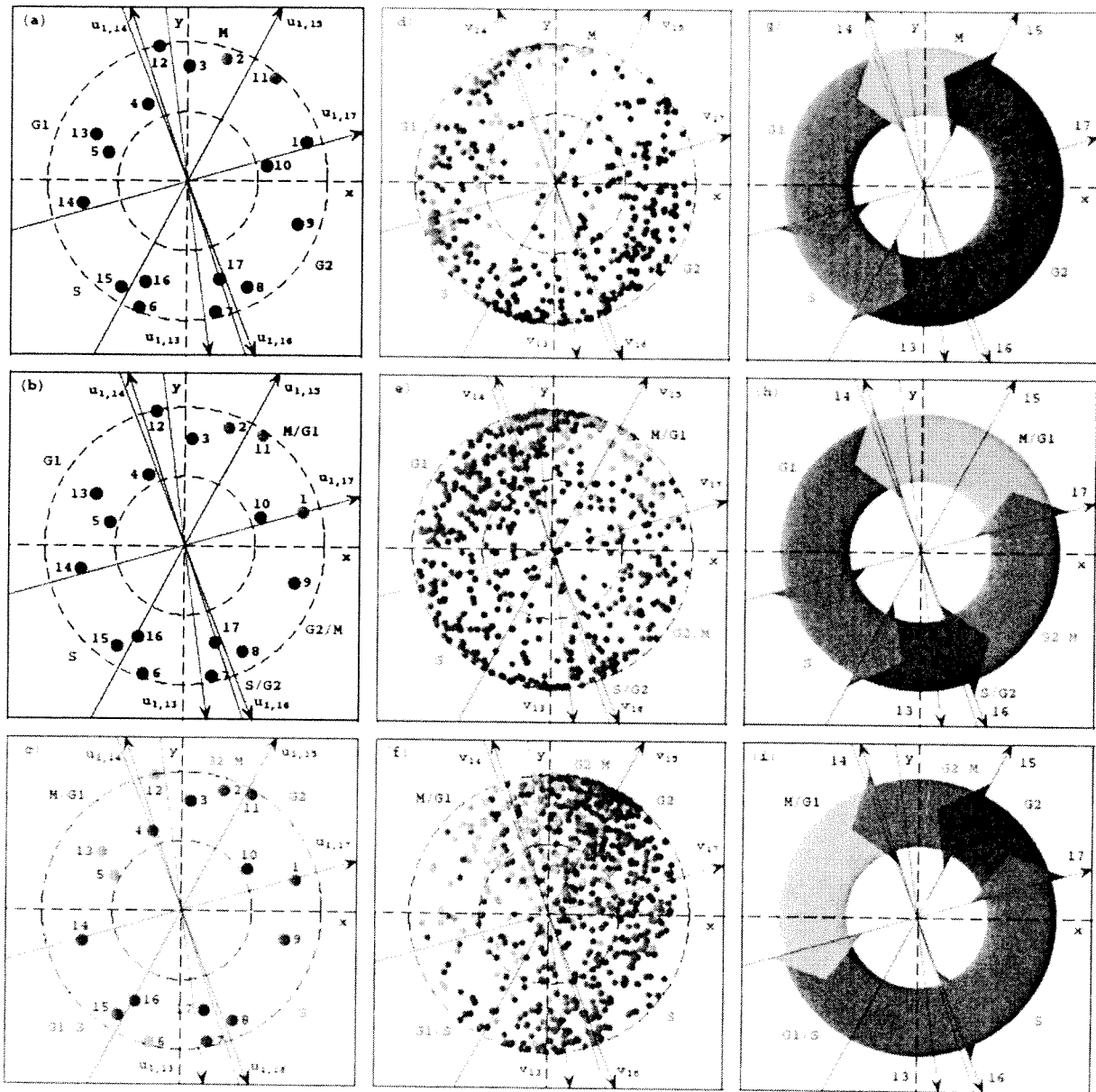


Figure 13

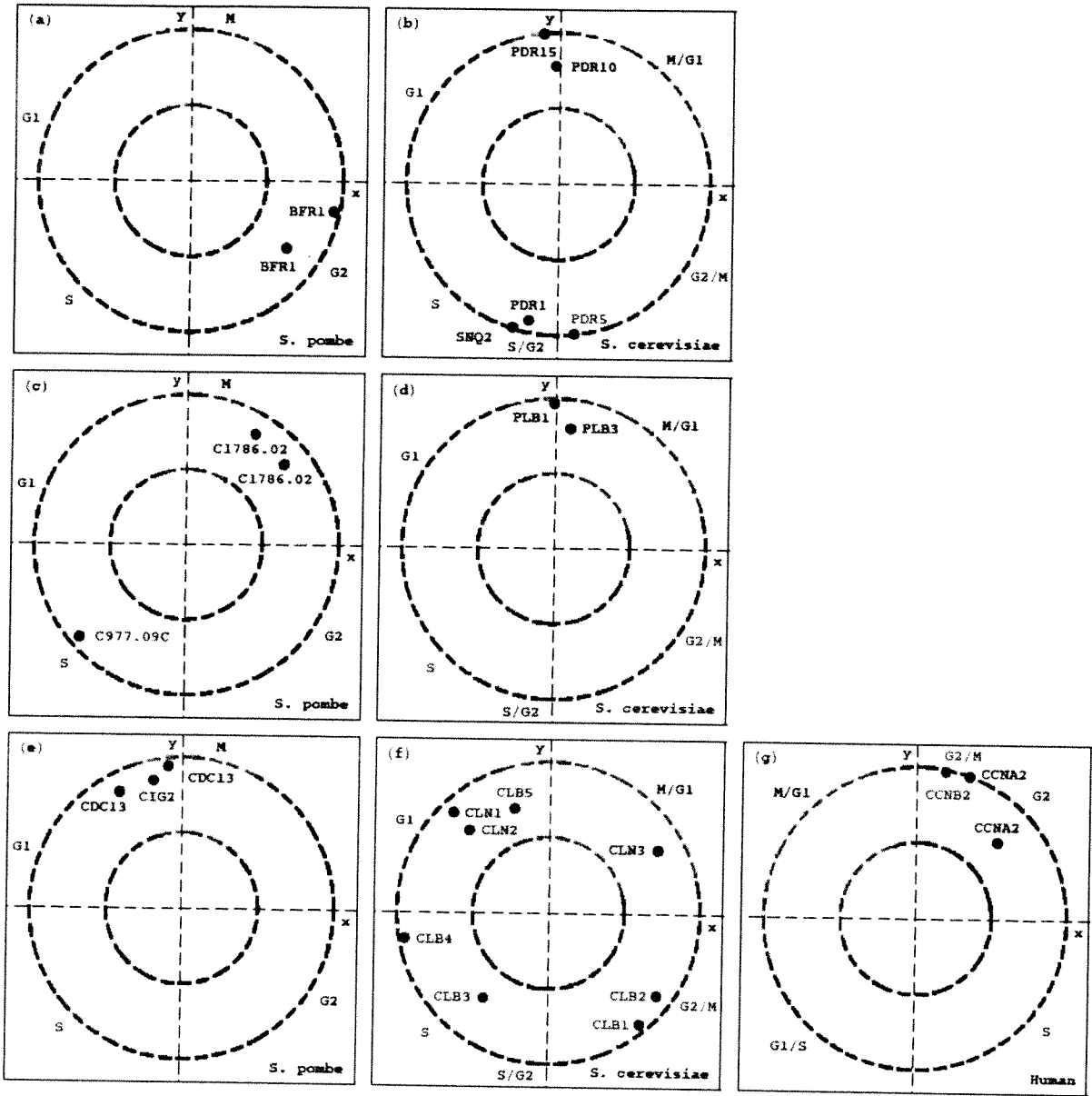


Figure 14

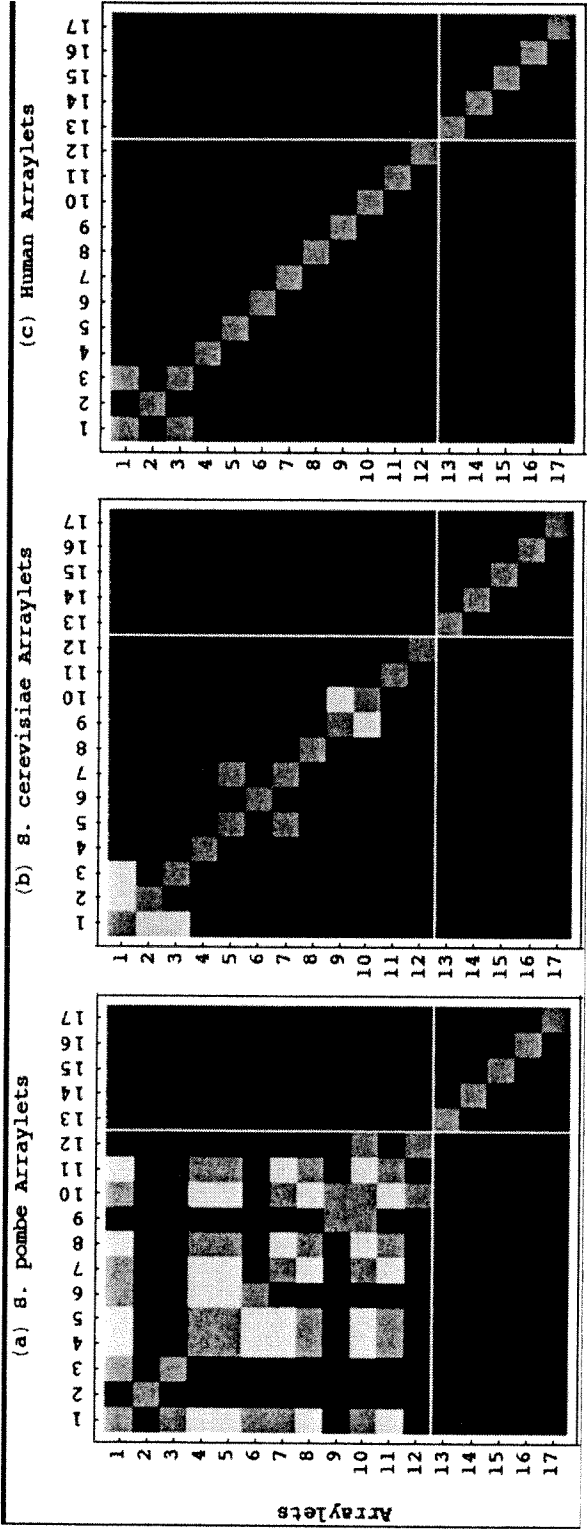


Figure 15

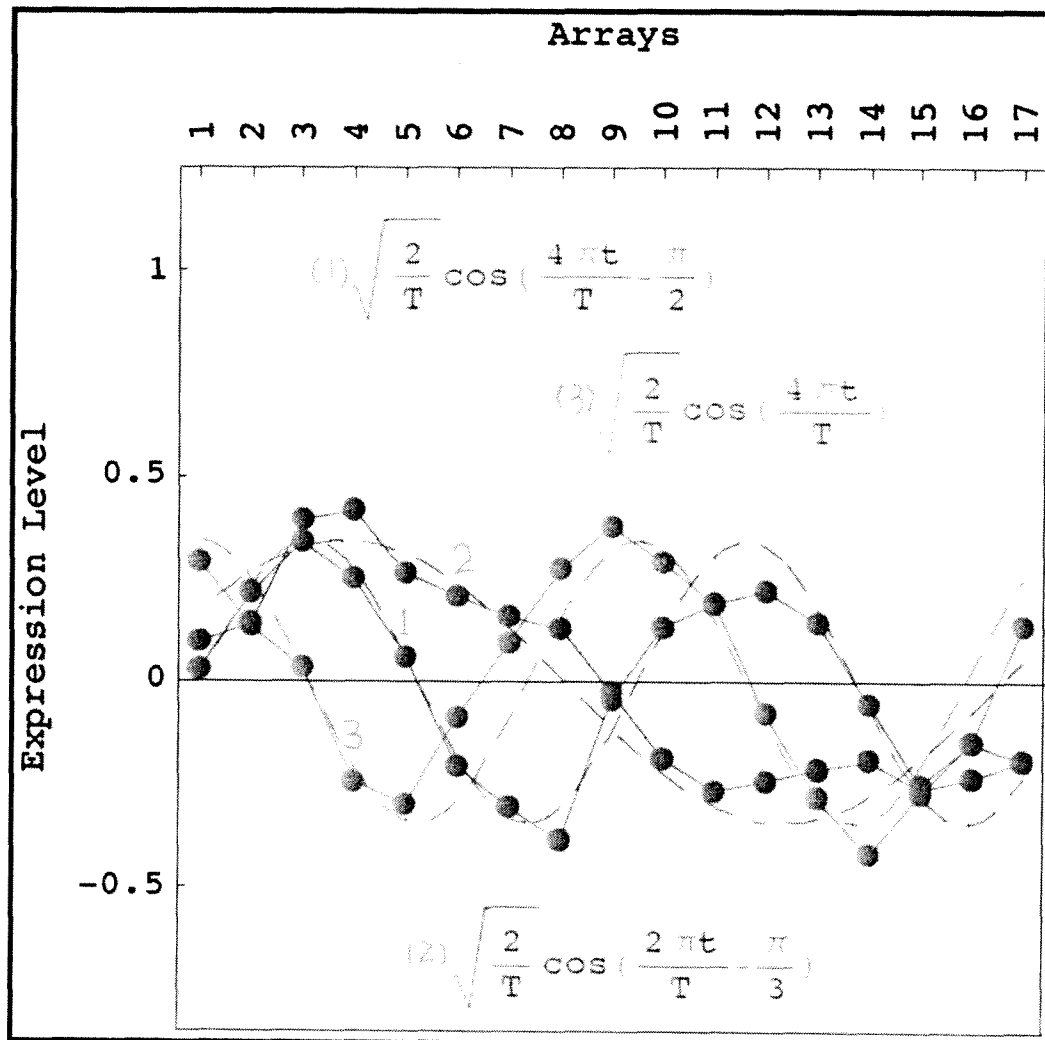


Figure 16

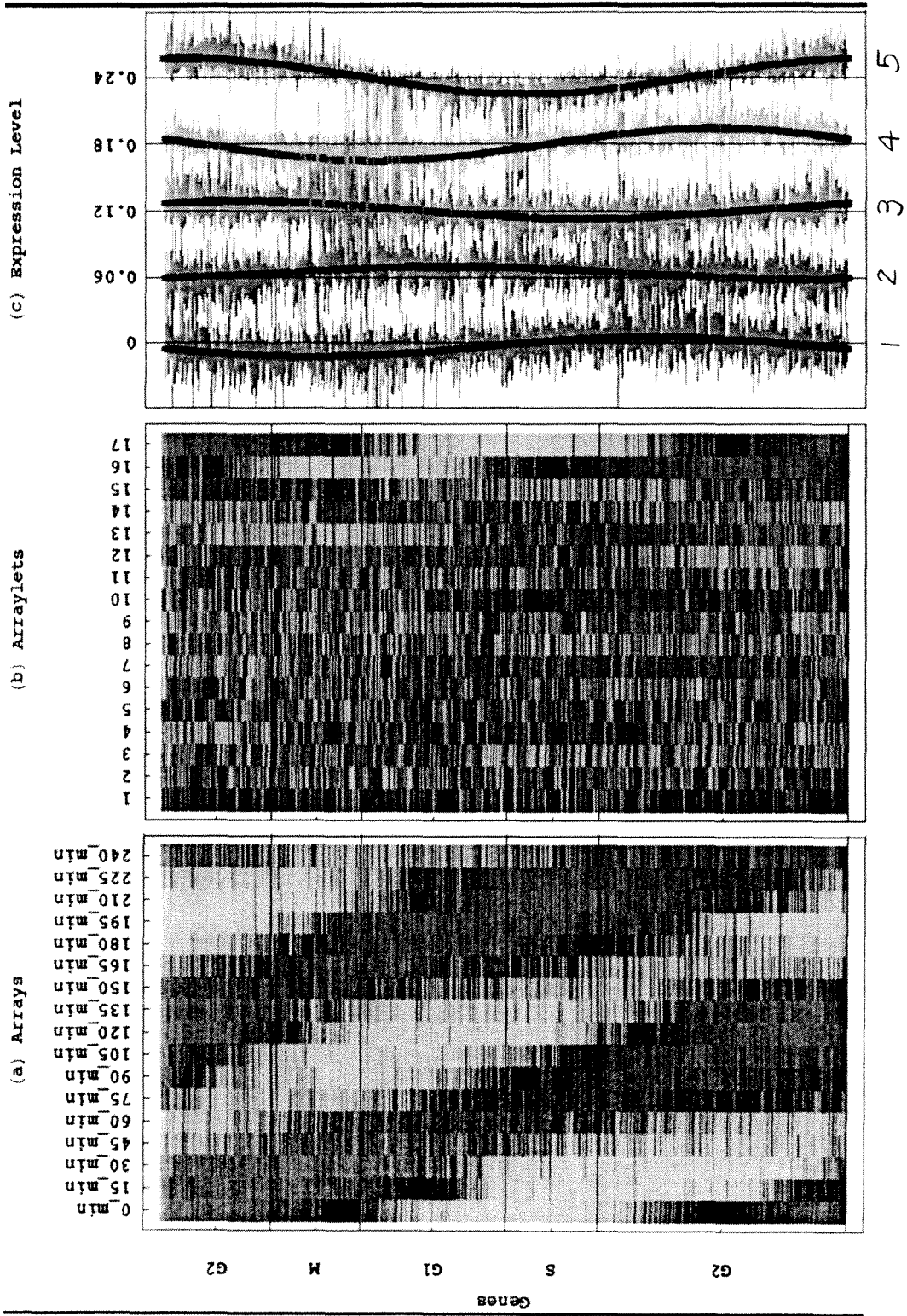


Figure 17

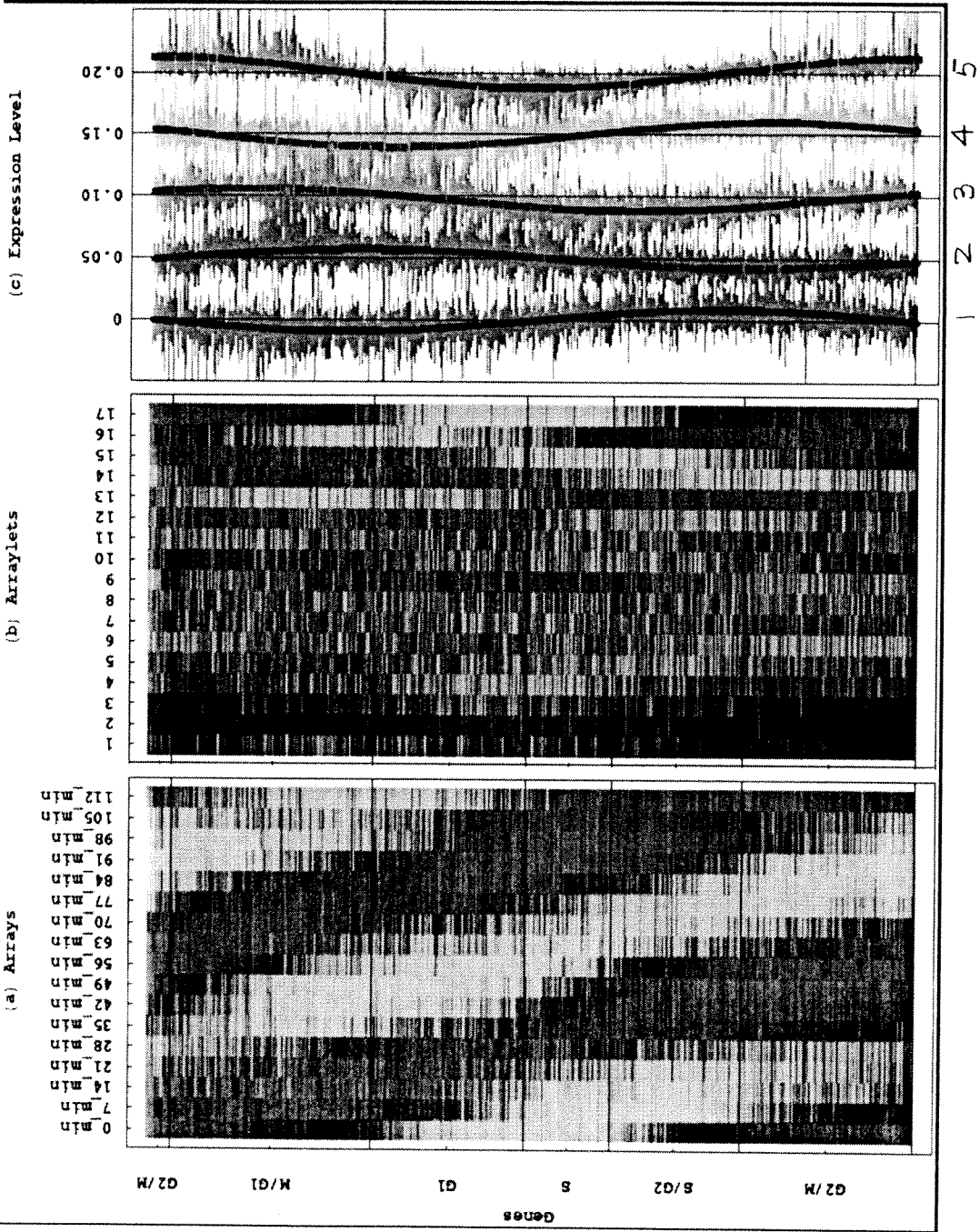


Figure 18

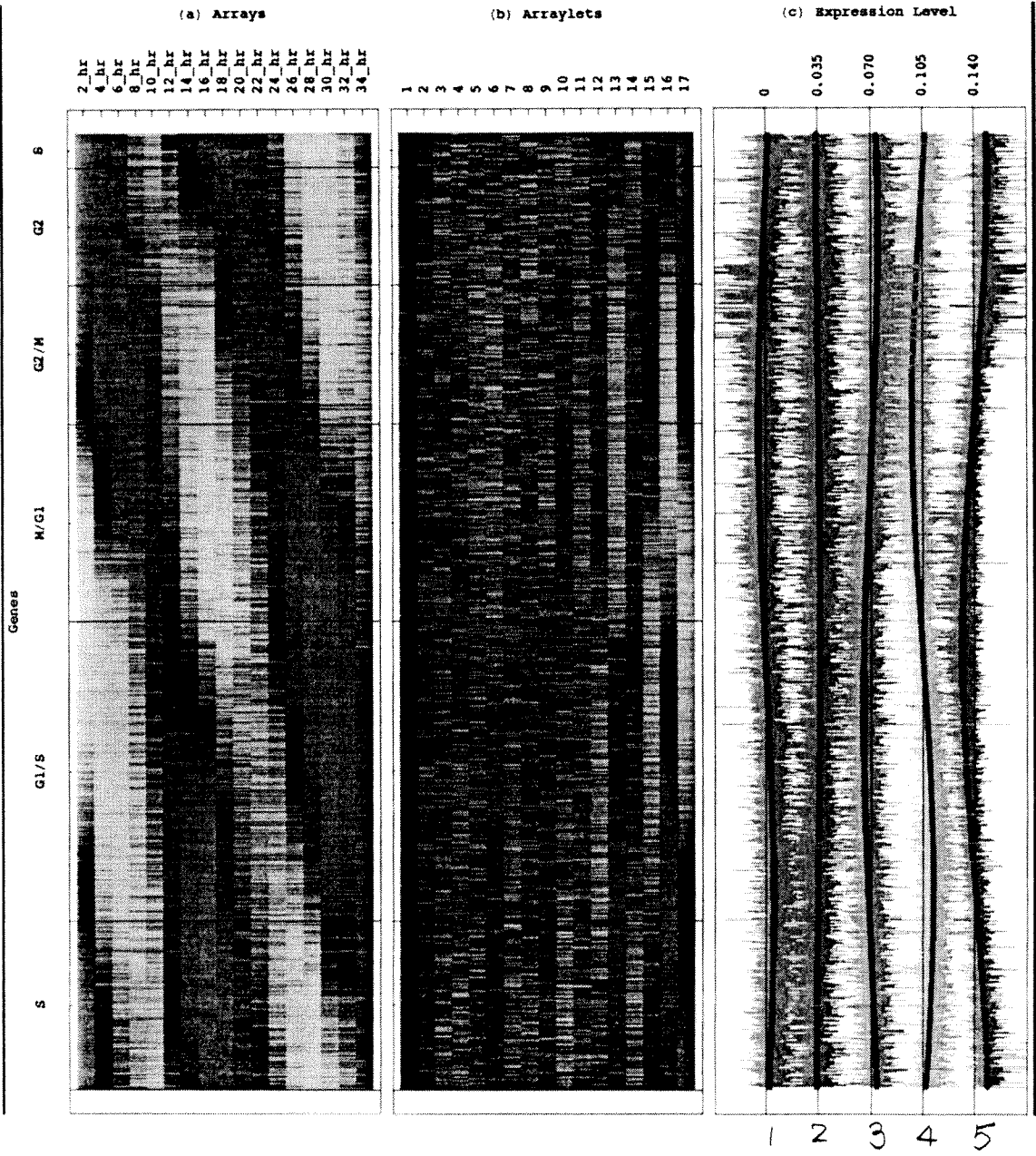


Figure 19

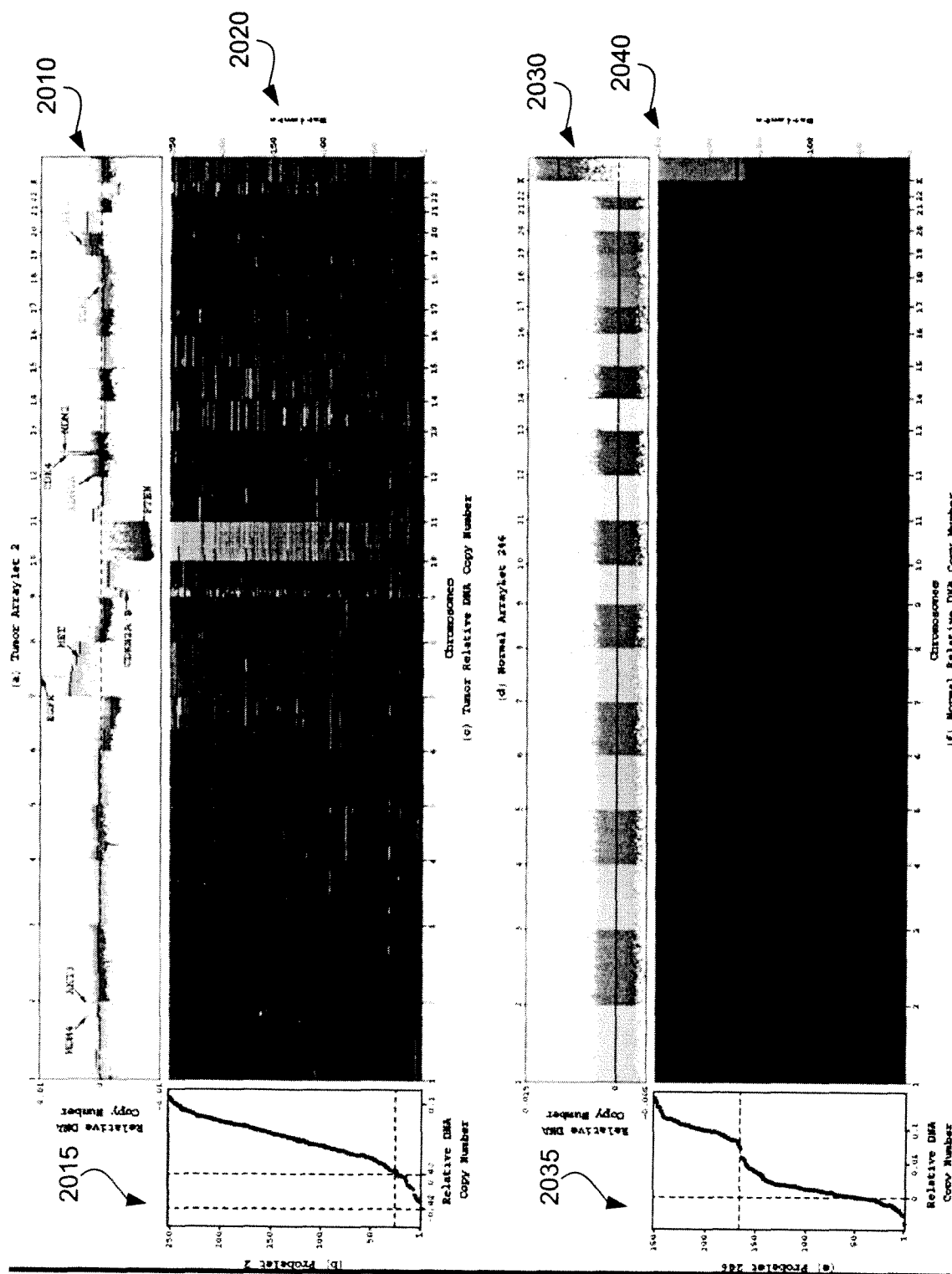


Figure 20

20/41

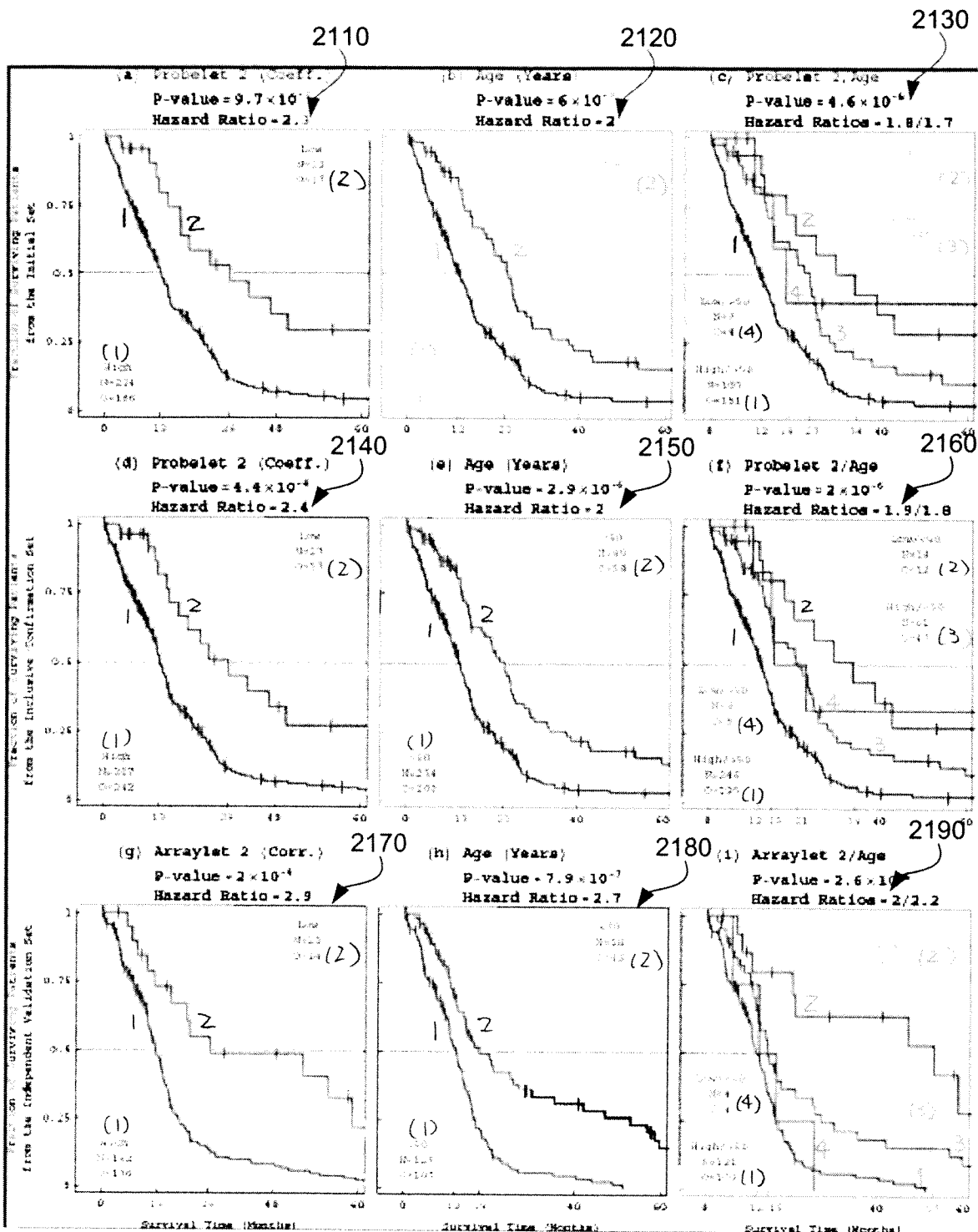
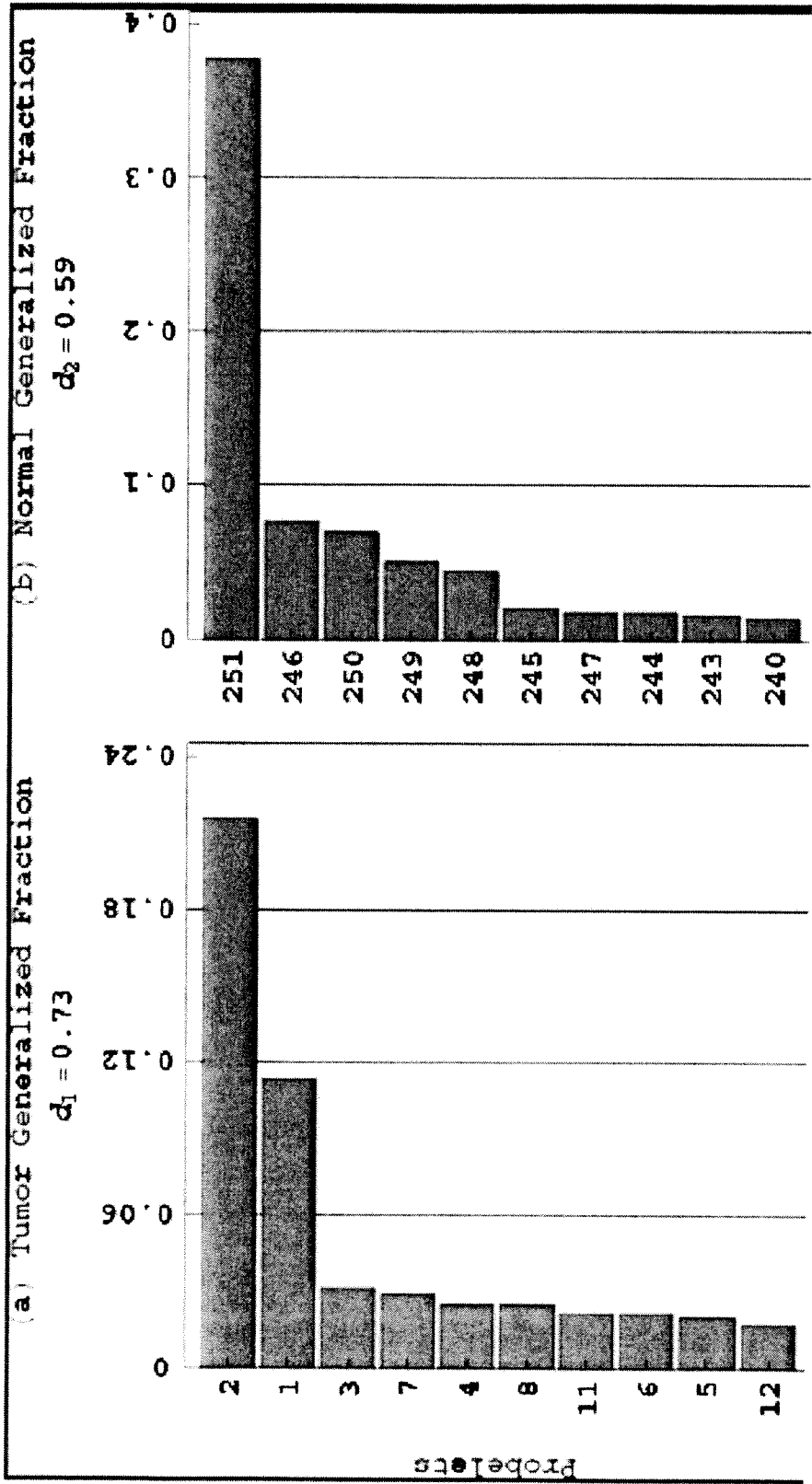


Figure 21



2240

2220

Figure 22

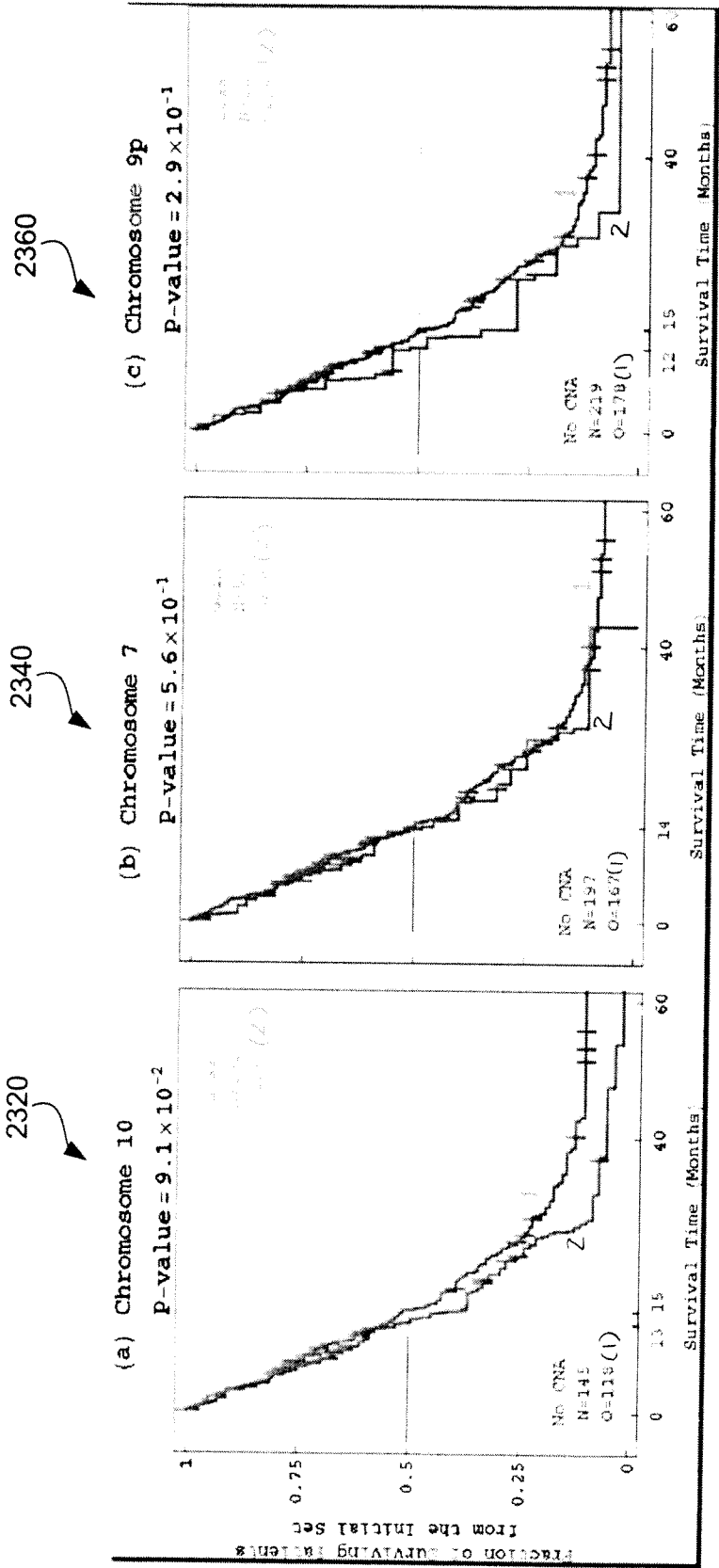


Figure 23

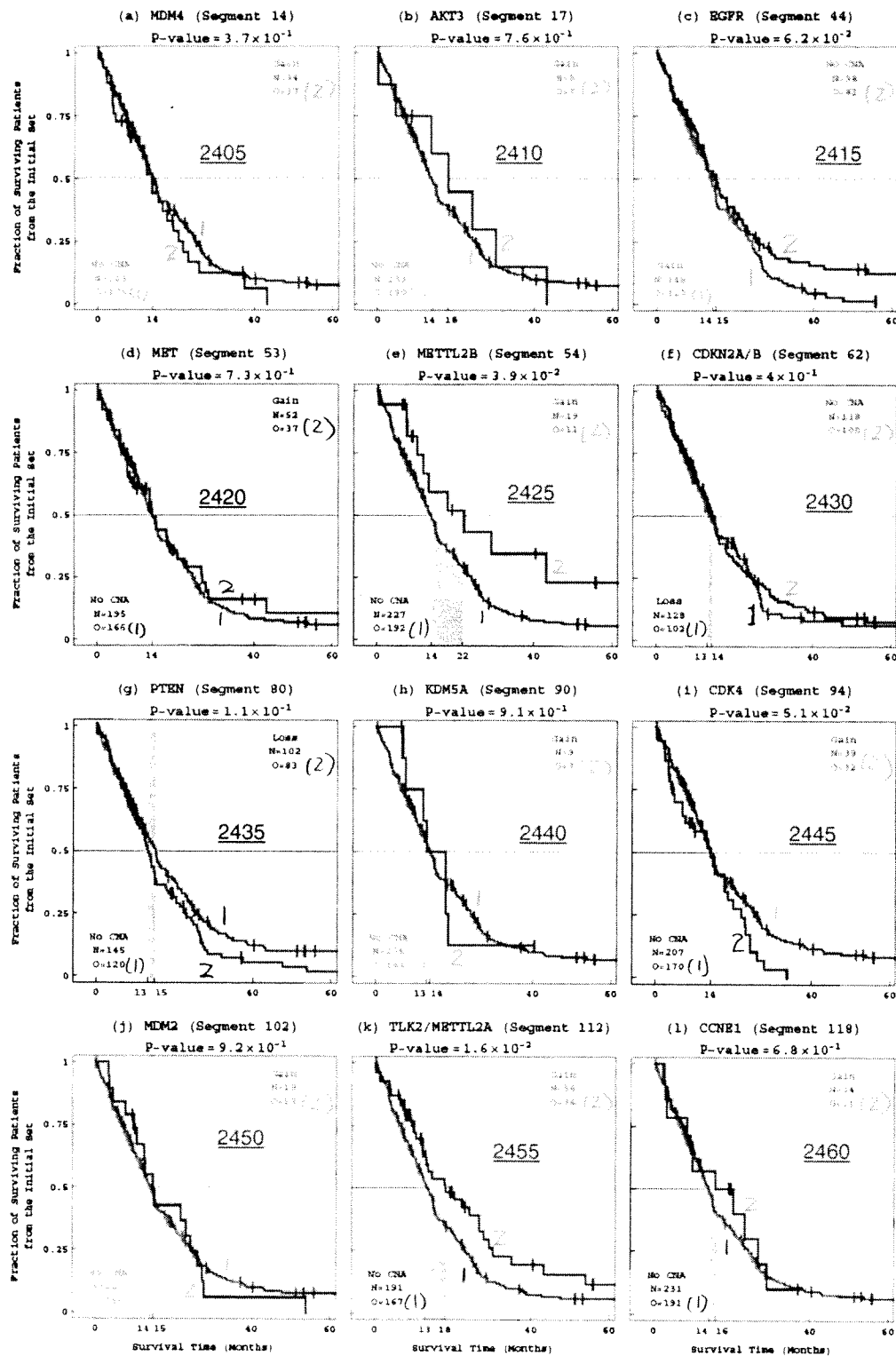


Figure 24

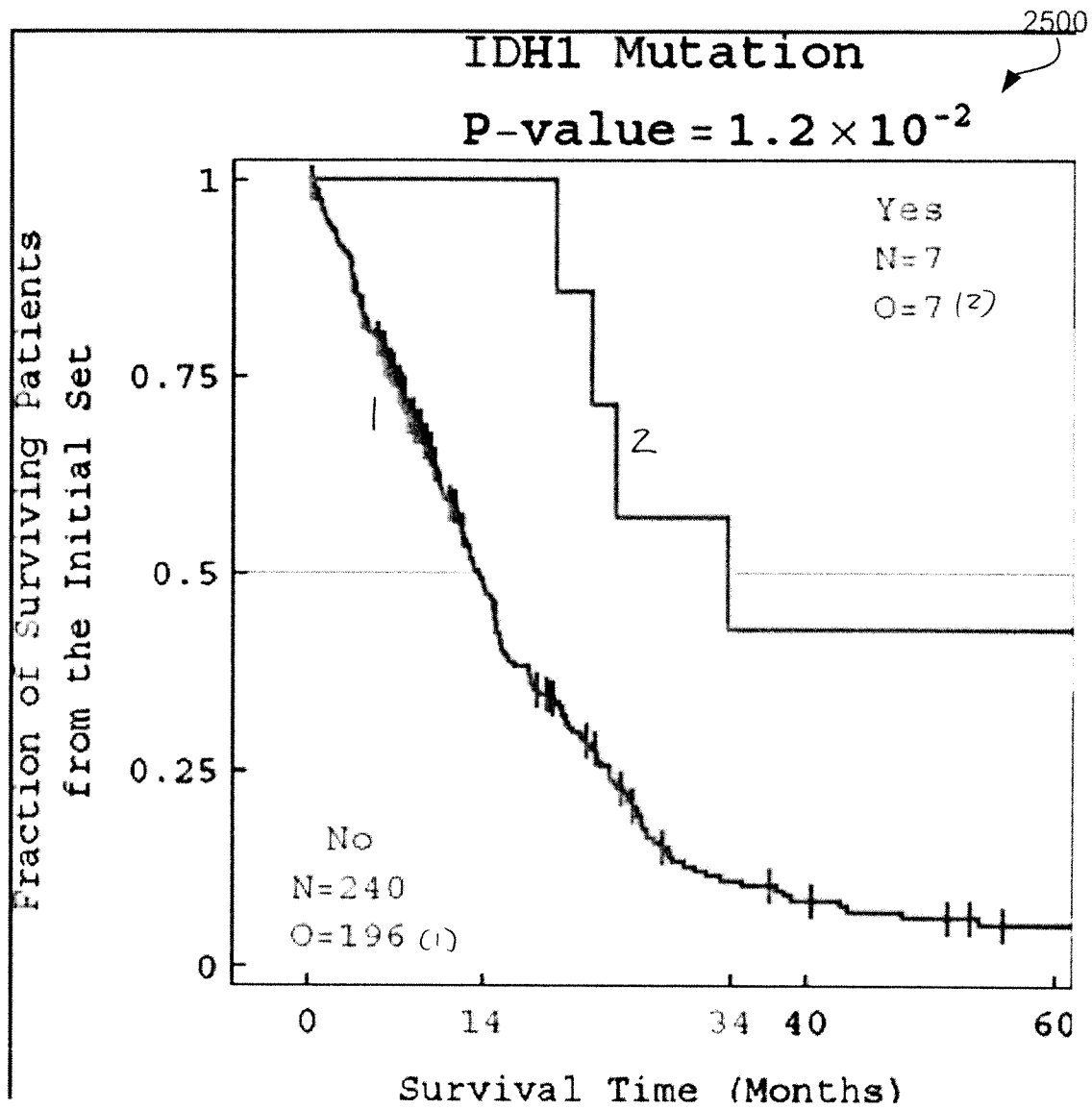


Figure 25

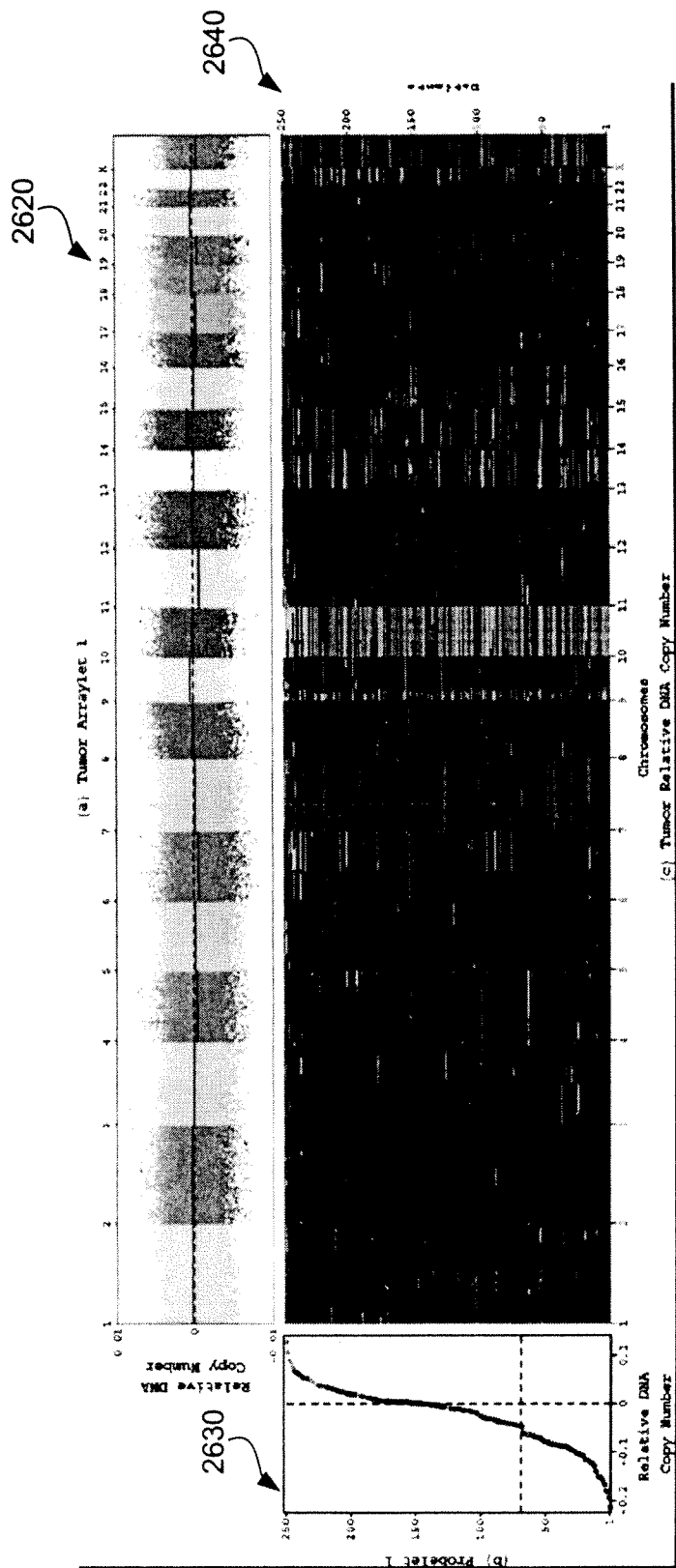


Figure 26

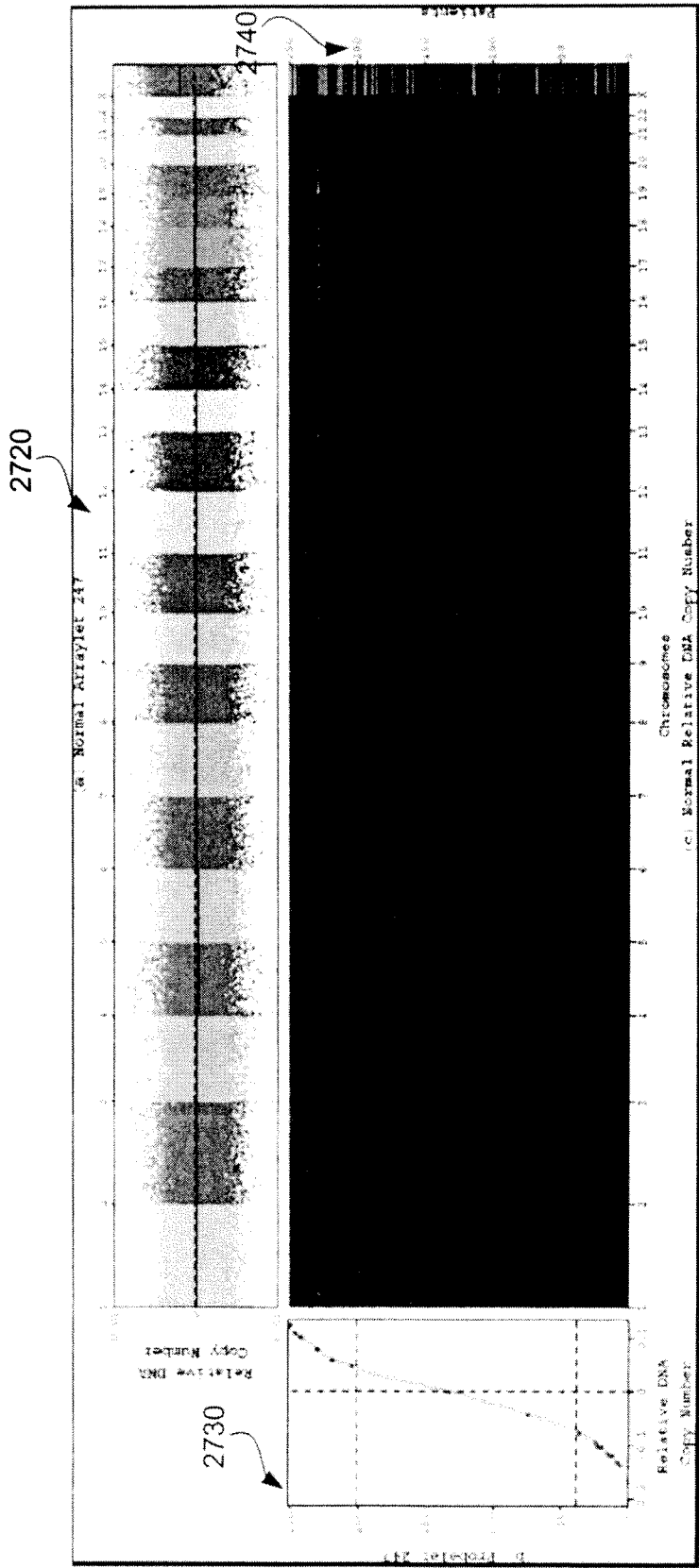


Figure 27

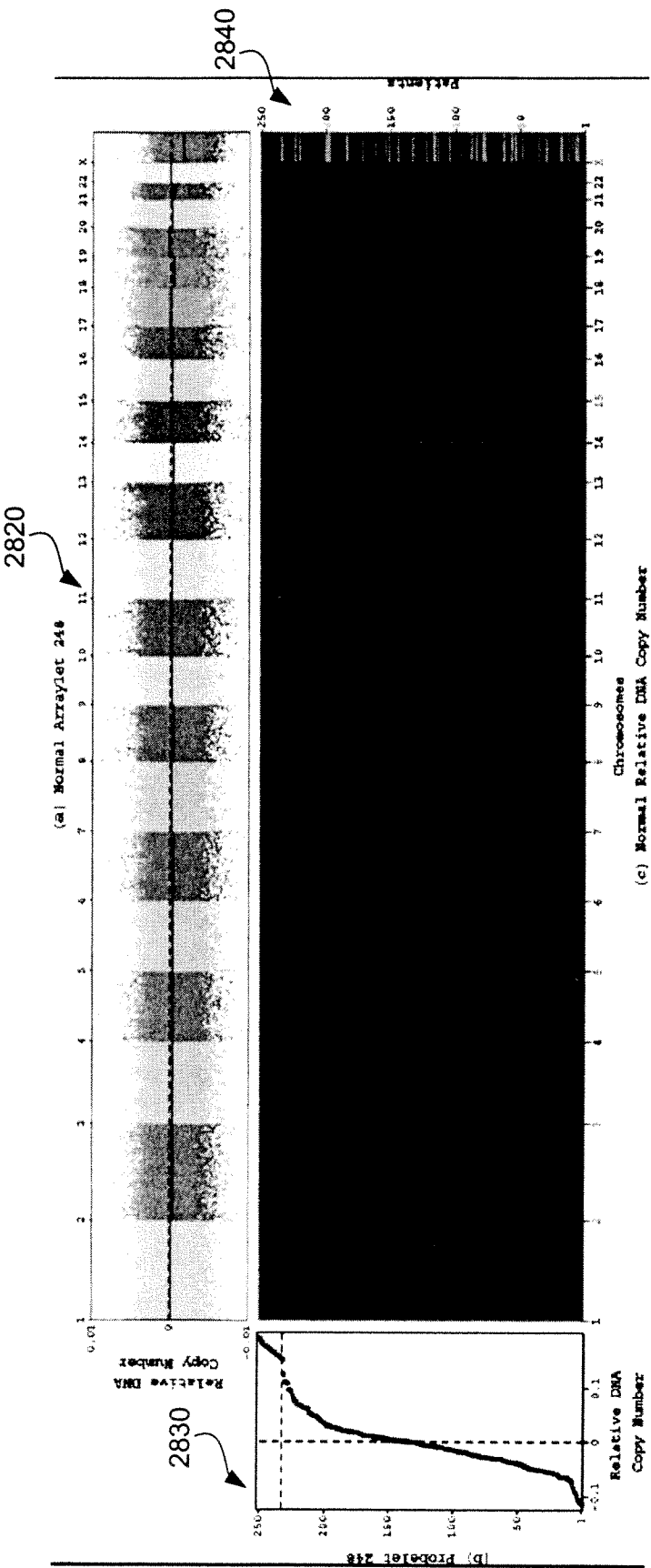


Figure 28

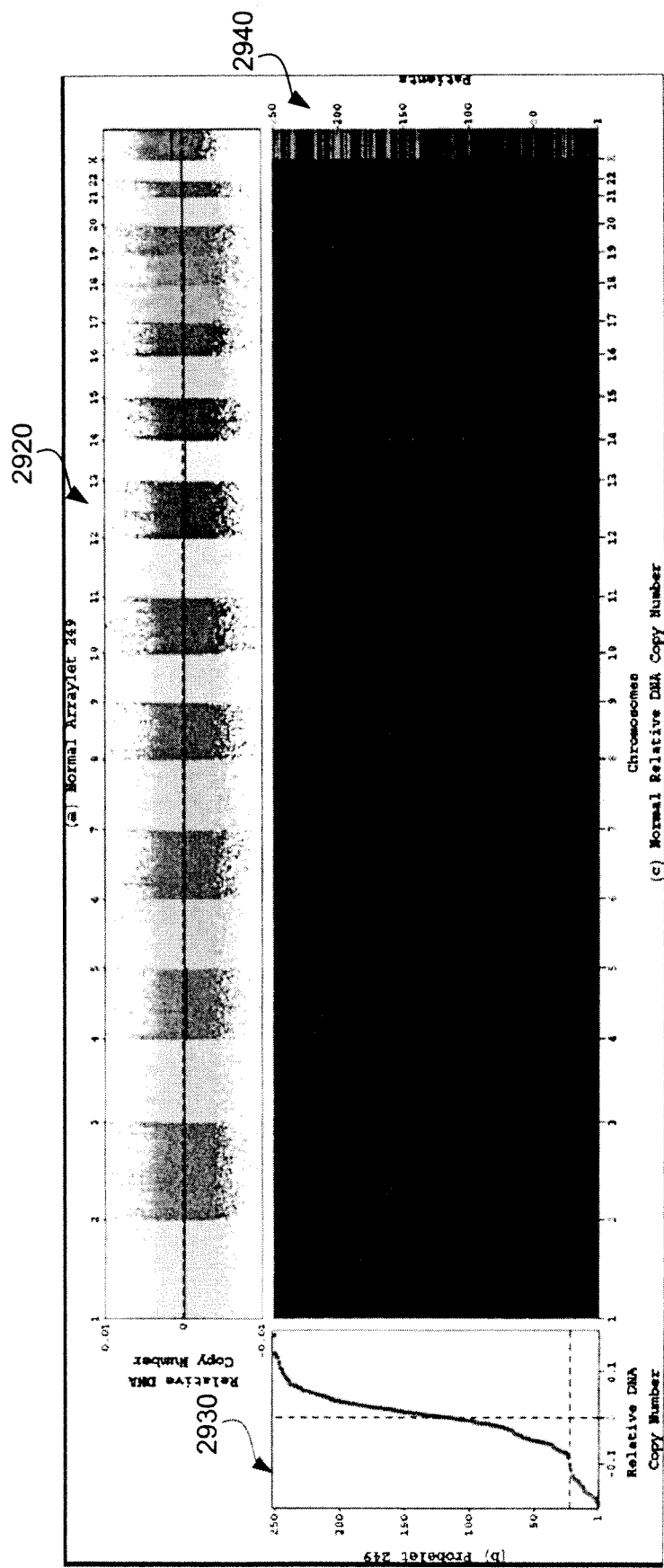


Figure 29

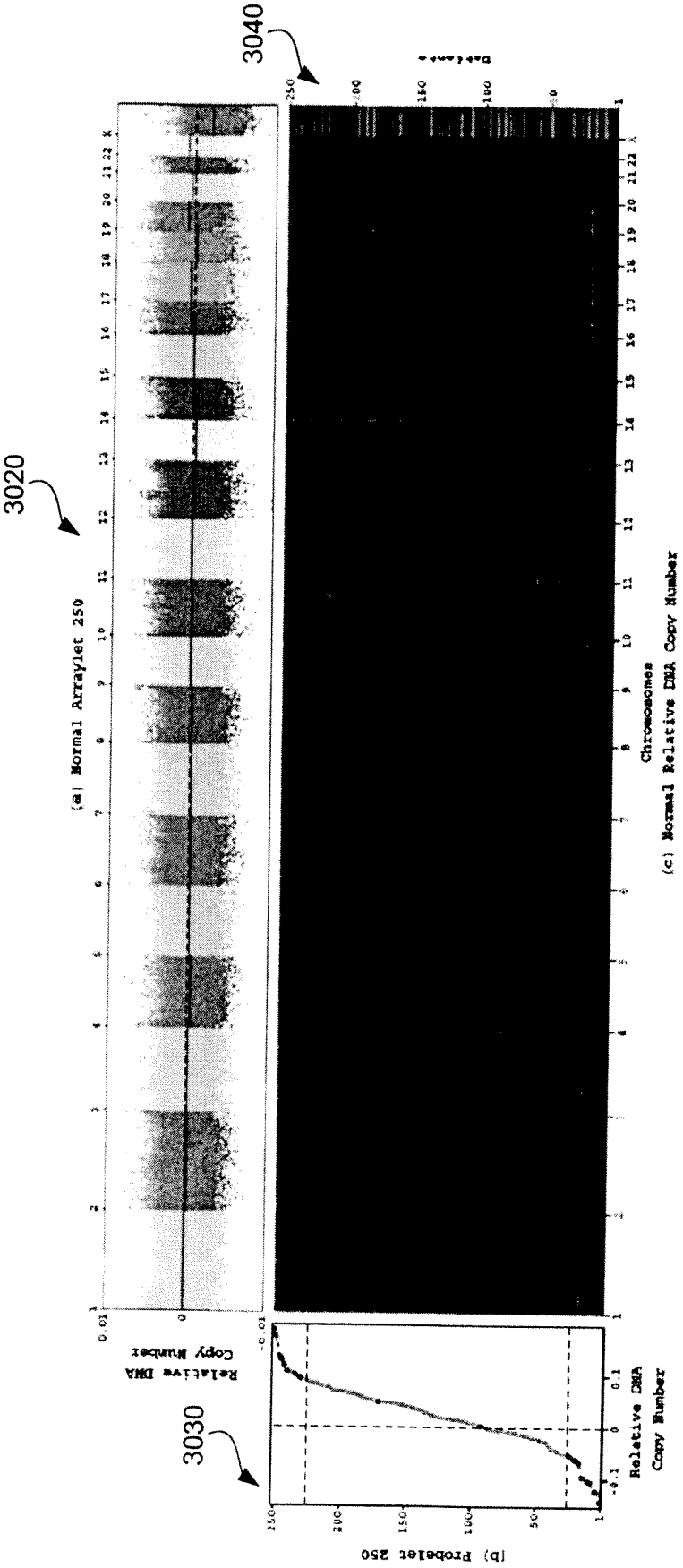


Figure 30

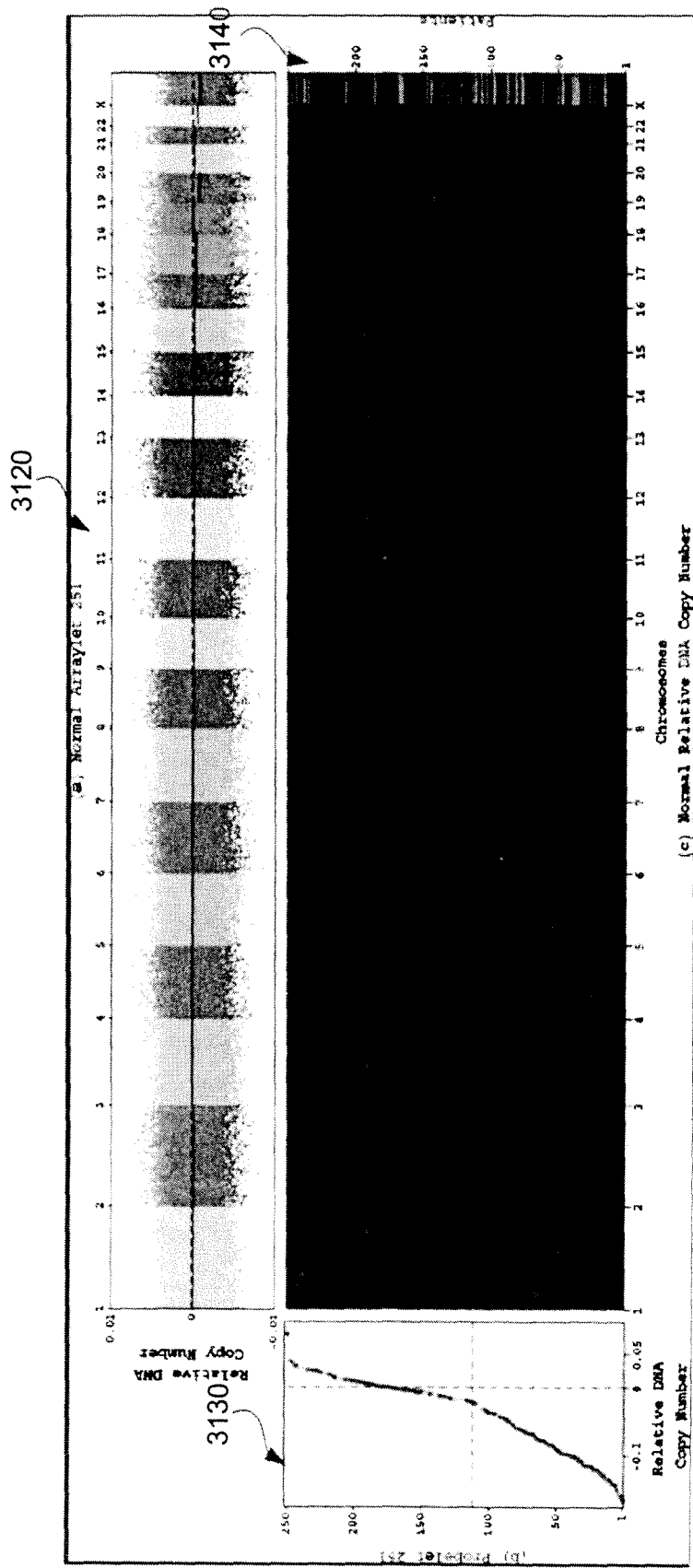


Figure 31

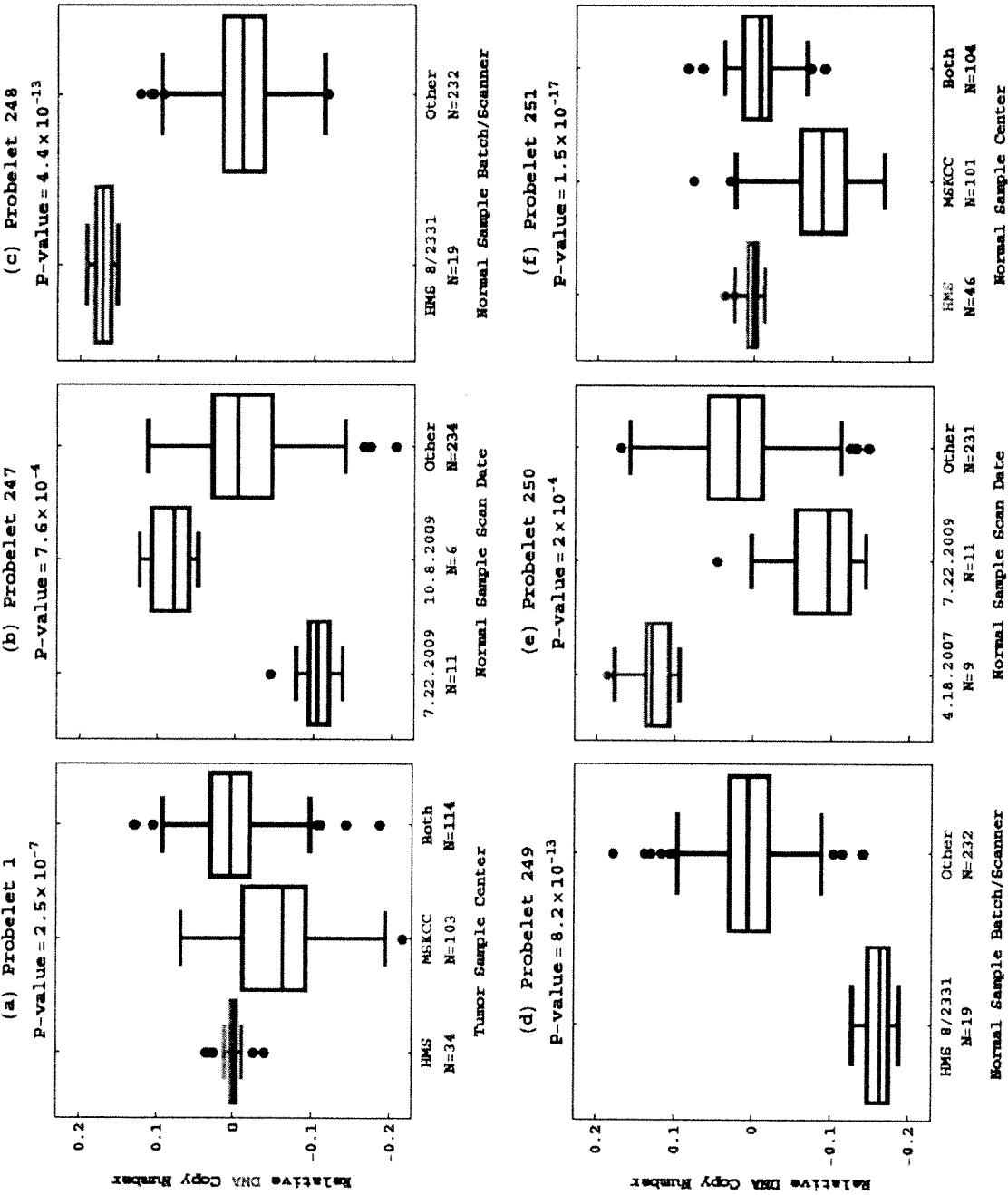


Figure 32

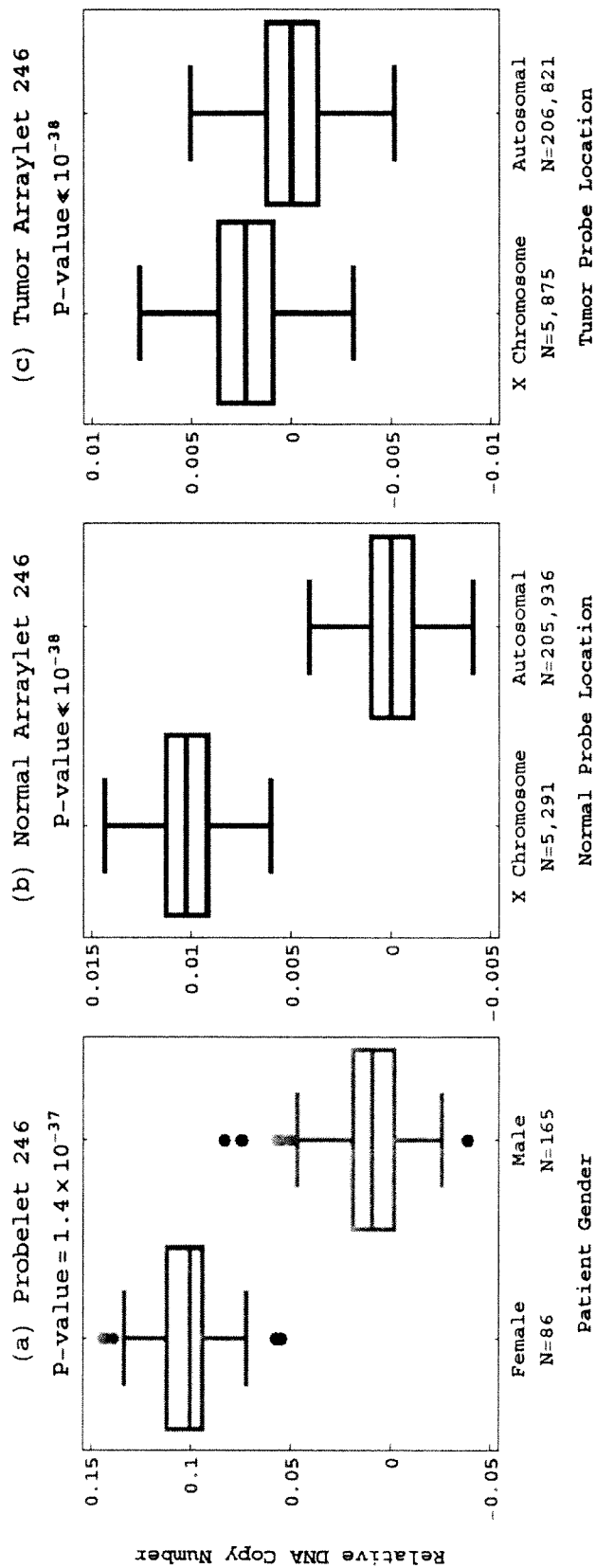


Figure 33

| Cox Proportional Hazard Model | Predictor | Initial Set  |                      | Inclusive Confirmation Set |                      | Independent Validation Set |                      |
|-------------------------------|-----------|--------------|----------------------|----------------------------|----------------------|----------------------------|----------------------|
|                               |           | Hazard Ratio | P-value              | Hazard Ratio               | P-value              | Hazard Ratio               | P-value              |
| Univariate                    | GSVD      | 2.3          | $1.3 \times 10^{-3}$ | 2.4                        | $6.5 \times 10^{-4}$ | 2.9                        | $3.6 \times 10^{-4}$ |
|                               | Age       | 2.0          | $7.9 \times 10^{-5}$ | 2.0                        | $4.3 \times 10^{-6}$ | 2.7                        | $1.7 \times 10^{-6}$ |
| Multivariate                  | GSVD      | 1.8          | $2.2 \times 10^{-2}$ | 1.9                        | $1.2 \times 10^{-2}$ | 2.0                        | $2.2 \times 10^{-2}$ |
|                               | Age       | 1.7          | $2.0 \times 10^{-3}$ | 1.8                        | $1.0 \times 10^{-4}$ | 2.2                        | $2.0 \times 10^{-4}$ |

Figure 34

| Probelet | Phenotype                   | Relative DNA Copy Number Gain |     |    |    | Relative DNA Copy Number Loss |            |     |     |     |                       |
|----------|-----------------------------|-------------------------------|-----|----|----|-------------------------------|------------|-----|-----|-----|-----------------------|
|          |                             | Annotation                    | n   | K  | k  | P-value                       | Annotation | n   | K   | k   | P-value               |
| 1        | Tumor Sample Center         | HMS                           | 183 | 34 | 34 | $8.5 \times 10^{-6}$          | MSKCC      | 68  | 103 | 55  | $3.9 \times 10^{-17}$ |
| 246      | Patient Gender              | Female                        | 86  | 86 | 84 | $8.0 \times 10^{-62}$         | Male       | 165 | 165 | 163 | $8.0 \times 10^{-62}$ |
| 247      | Normal Sample Scan Date     | 10.8.2009                     | 51  | 6  | 6  | $5.5 \times 10^{-5}$          | 7.22.2009  | 38  | 11  | 10  | $2.0 \times 10^{-8}$  |
| 248      | Normal Sample Batch/Scanner | HMS 8/2331                    | 19  | 19 | 19 | $6.2 \times 10^{-29}$         |            |     |     |     |                       |
| 249      | Normal Sample Batch/Scanner |                               |     |    |    |                               | HMS 8/2331 | 22  | 19  | 19  | $9.6 \times 10^{-26}$ |
| 250      | Normal Sample Scan Date     | 4.18.2007                     | 26  | 9  | 9  | $3.3 \times 10^{-10}$         | 7.22.2009  | 25  | 11  | 9   | $1.1 \times 10^{-5}$  |
| 251      | Normal Sample Center        | HMS                           | 139 | 46 | 46 | $2.8 \times 10^{-14}$         | MSKCC      | 112 | 101 | 89  | $2.1 \times 10^{-42}$ |

Figure 35

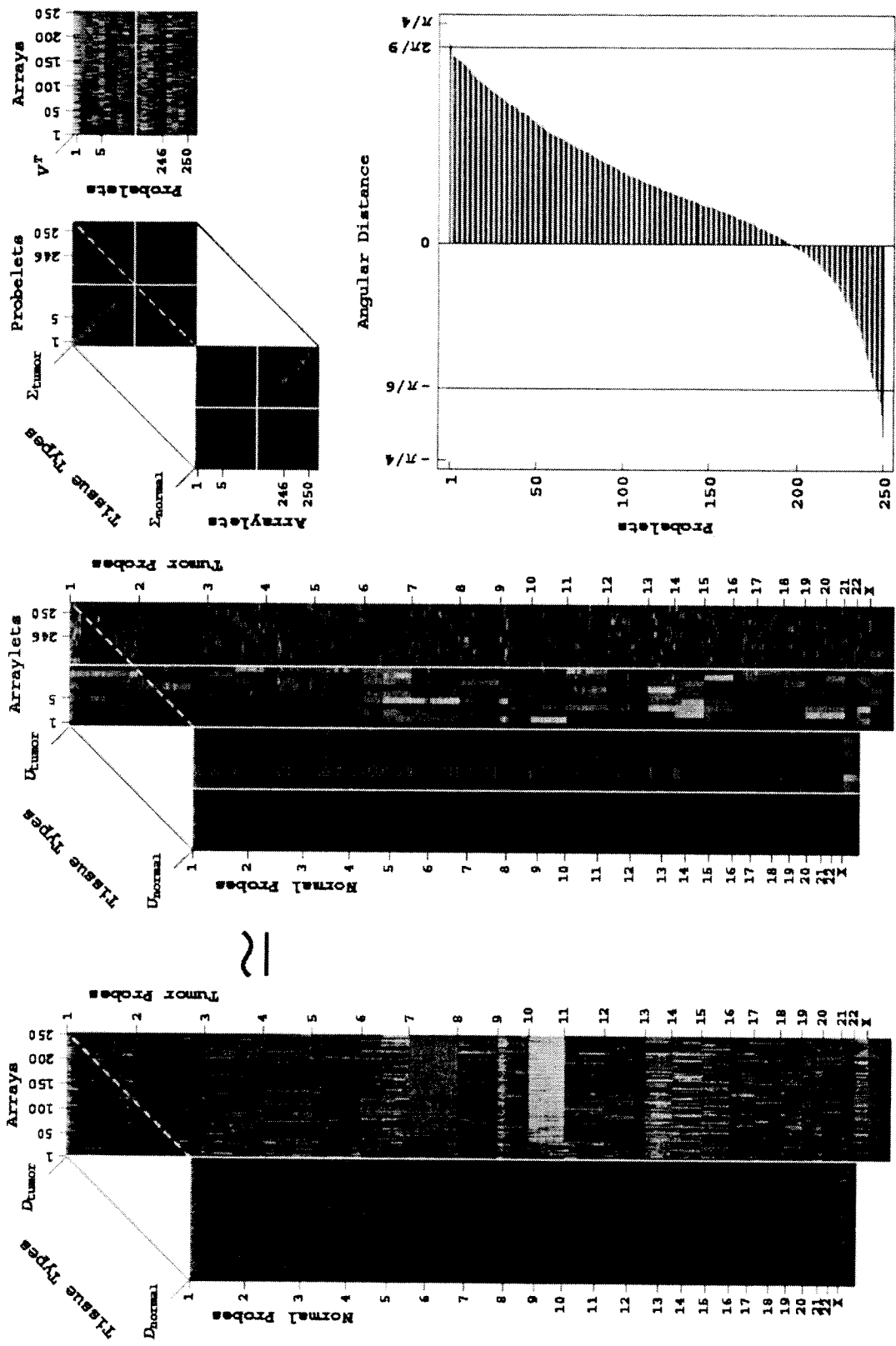


Figure 36

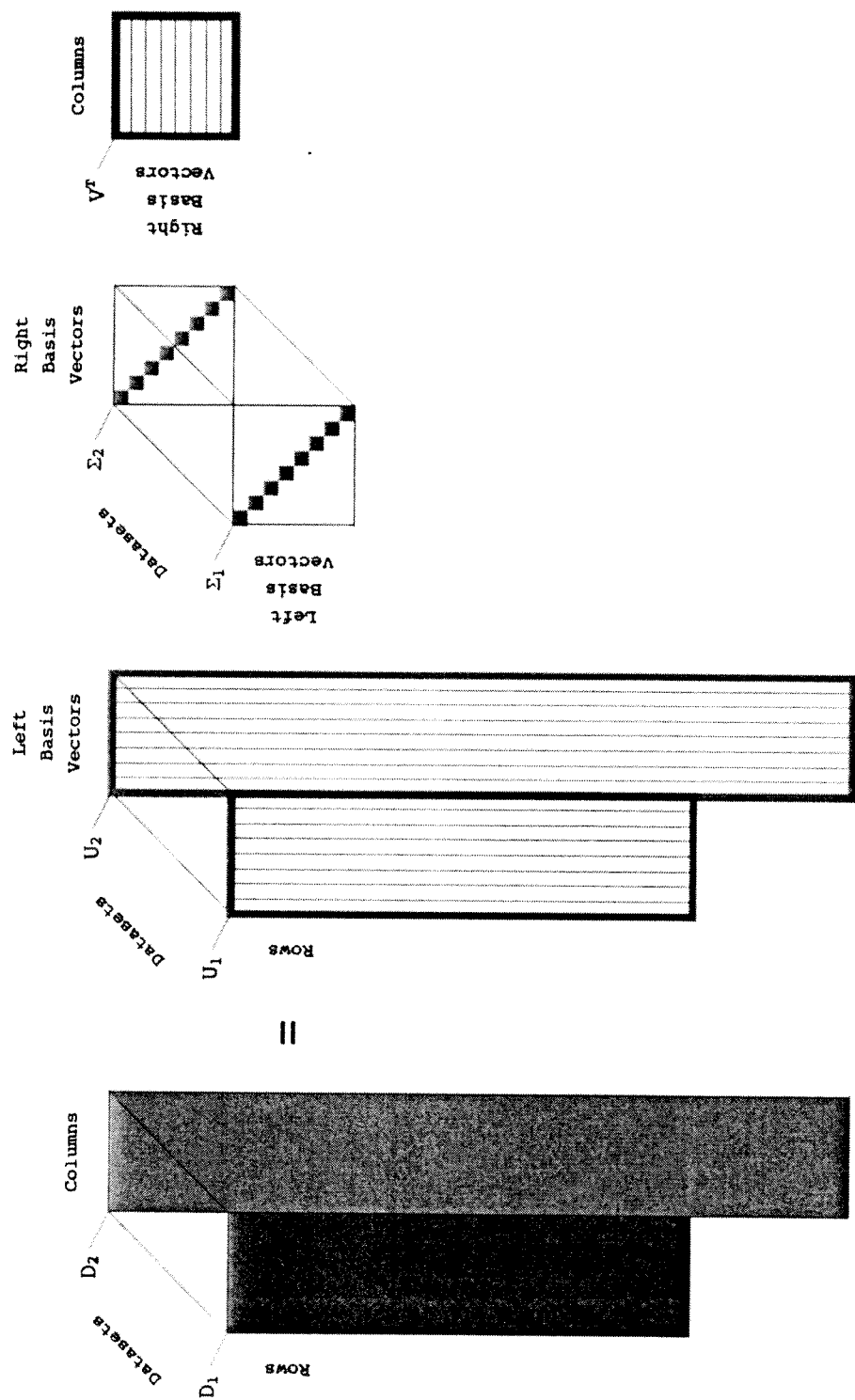


Figure 37

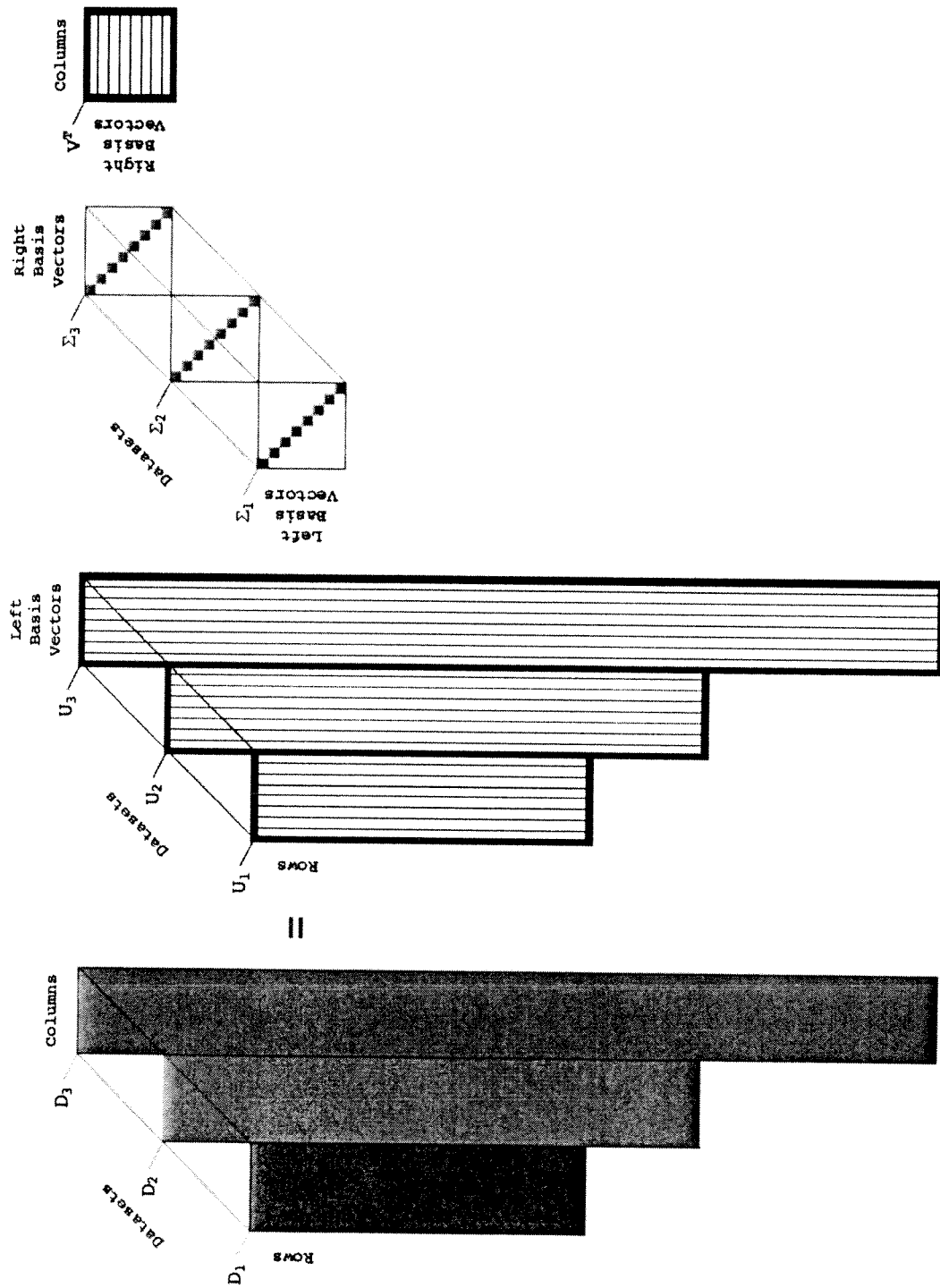


Figure 38

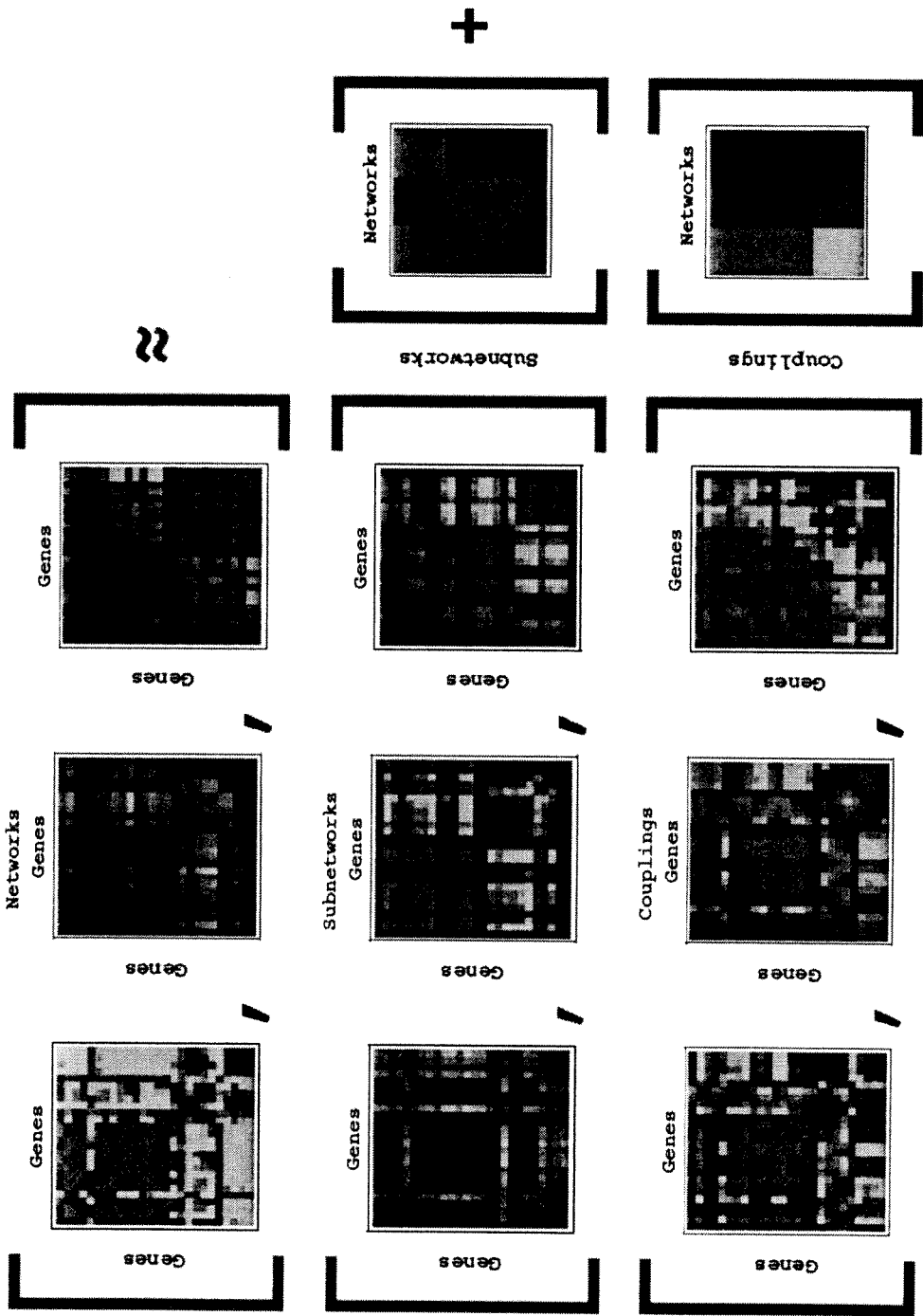


Figure 39

| Cox Proportional Hazard Model | Predictor    | Initial Set  |                       | Inclusive Confirmation Set |                       | Independent Validation Set |                      |
|-------------------------------|--------------|--------------|-----------------------|----------------------------|-----------------------|----------------------------|----------------------|
|                               |              | Hazard Ratio | P-value               | Hazard Ratio               | P-value               | Hazard Ratio               | P-value              |
| Univariate                    | GSVD         | 2.4          | $1.2 \times 10^{-3}$  | 2.4                        | $6.4 \times 10^{-4}$  | 2.8                        | $1.3 \times 10^{-3}$ |
|                               | Chemotherapy | 2.6          | $1.5 \times 10^{-8}$  | 2.7                        | $6.3 \times 10^{-11}$ | 2.2                        | $7.3 \times 10^{-4}$ |
| Multivariate                  | GSVD         | 3.0          | $5.2 \times 10^{-5}$  | 3.1                        | $2.5 \times 10^{-5}$  | 3.3                        | $2.3 \times 10^{-4}$ |
|                               | Chemotherapy | 3.1          | $7.9 \times 10^{-11}$ | 3.2                        | $1.9 \times 10^{-13}$ | 2.7                        | $3.0 \times 10^{-5}$ |

Figure 40

39/41

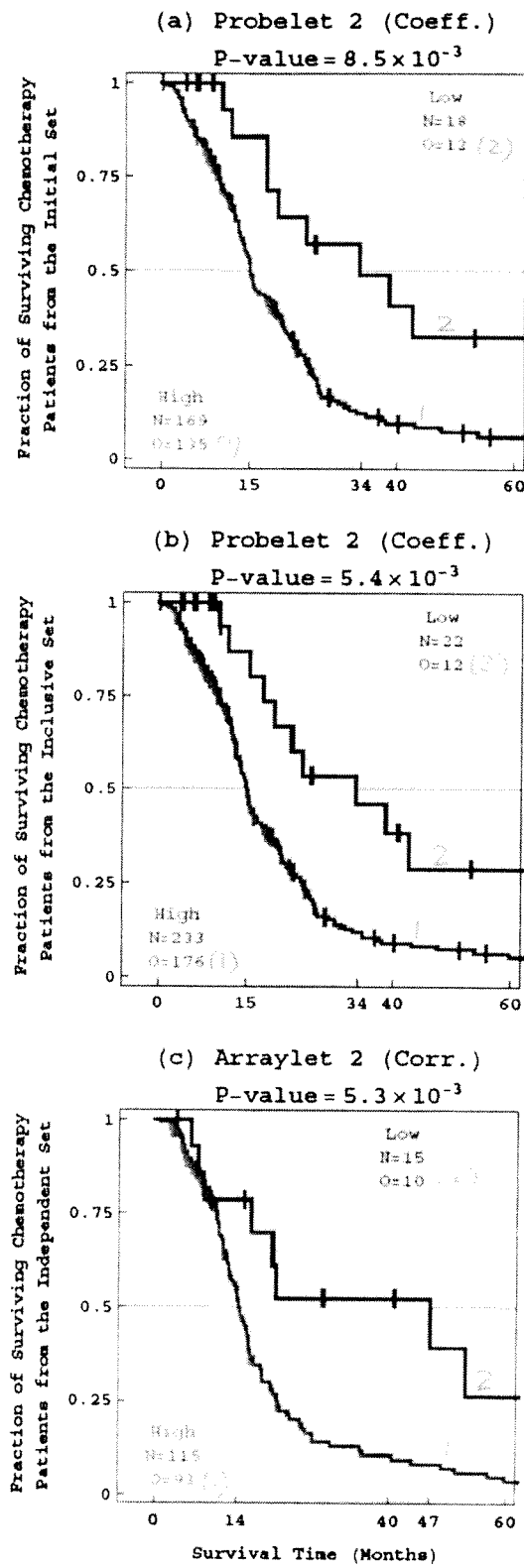


Figure 41

40/41

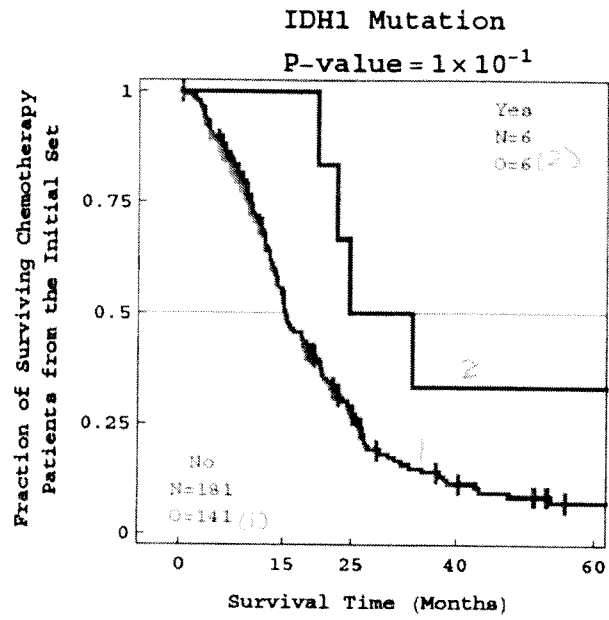


Figure 42

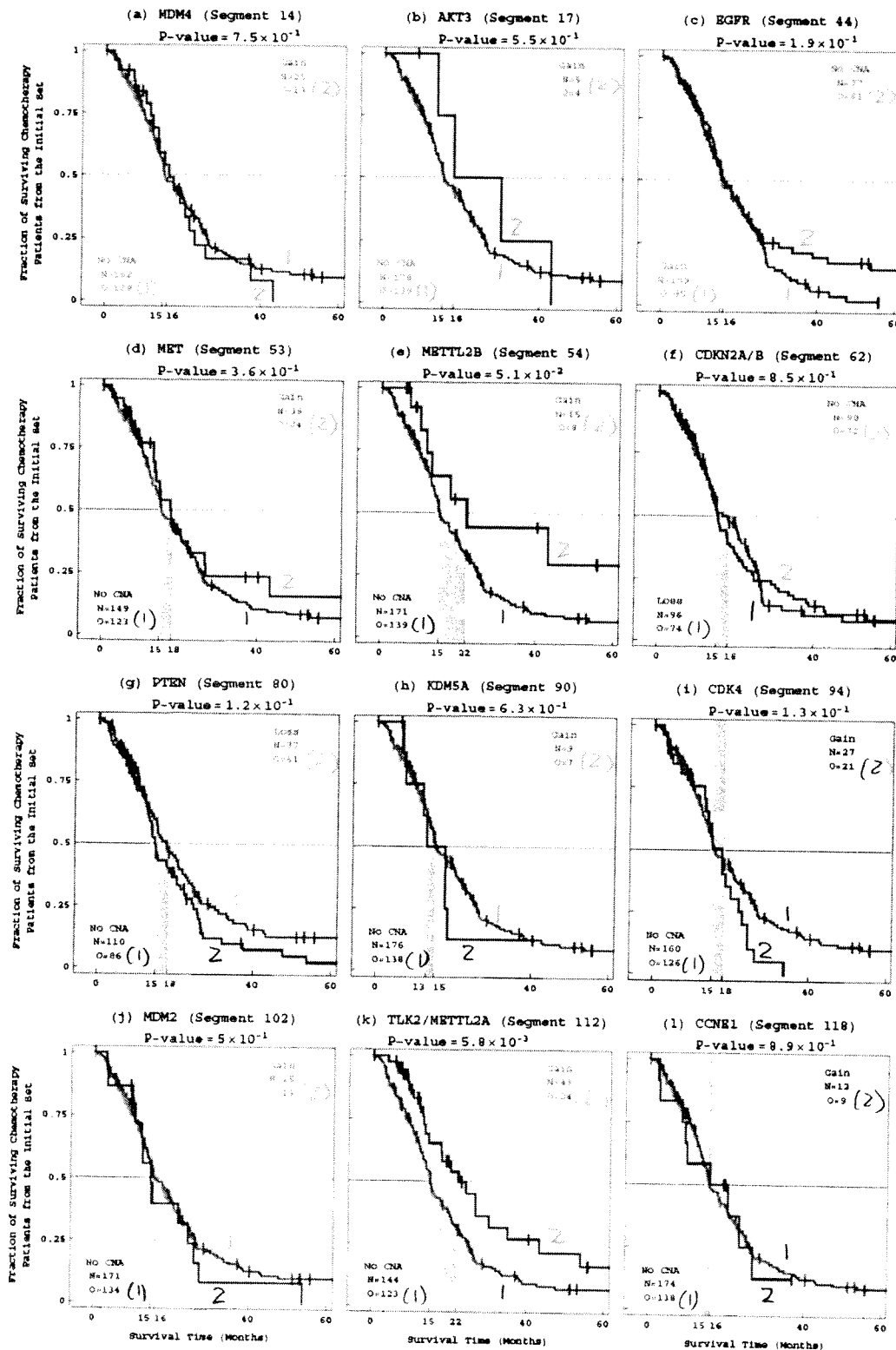


Figure 43

## INTERNATIONAL SEARCH REPORT

International application No.  
**PCT/US2012/054315****A. CLASSIFICATION OF SUBJECT MATTER****G06F 19/10(2011.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**

Minimum documentation searched (classification system followed by classification symbols)

G06F 19/10; A61N 1/372; G06F 19/00; G06F 17/00; C12Q 1/70; G01N 33/50

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) &amp; Keywords: medical, biological, gene, health

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                               | Relevant to claim No. |
|-----------|------------------------------------------------------------------------------------------------------------------|-----------------------|
| A         | WO 2009-134479 A1 (MEDTRONIC, INC. et al.) 05 November 2009<br>See abstract; claim 1; and figures 4-5, 9, 15-16. | 1-25                  |
| A         | US 2005-0119534 A1 (DONALD TROST et al.) 02 June 2005<br>See abstract; claims 1, 40, 49; and figure 1.           | 1-25                  |
| A         | US 2003-0044775 A1 (SANDEEP GULATI) 06 March 2003<br>See abstract; claim 1; and figures 1-3.                     | 1-25                  |

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

\* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&amp;" document member of the same patent family

Date of the actual completion of the international search

31 OCTOBER 2012 (31.10.2012)

Date of mailing of the international search report

**31 OCTOBER 2012 (31.10.2012)**

Name and mailing address of the ISA/KR

Korean Intellectual Property Office  
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan  
City, 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

LEE, Seok Hyung

Telephone No. 82-42-481-5983



**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2012/054315**

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s)                                                                                                                                                                        | Publication<br>date                                                                                                                                    |
|-------------------------------------------|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| WO 2009-134479 A1                         | 05.11.2009          | EP 2307093 A1<br>EP 2307093 B1<br>US 2011-0040352 A1                                                                                                                                              | 13.04.2011<br>25.07.2012<br>17.02.2011                                                                                                                 |
| US 2005-0119534 A1                        | 02.06.2005          | BR PI0415845A<br>CA 2542460 A1<br>EP 1681986 A2<br>JP 2008-502371 A<br>JP 2008-502371 T<br>MX PA06004538A<br>WO 2005-039388 A2<br>WO 2005-039388 A3                                               | 27.03.2007<br>06.05.2005<br>26.07.2006<br>31.01.2008<br>31.01.2008<br>23.06.2006<br>06.05.2005<br>04.12.2008                                           |
| US 2003-0044775 A1                        | 06.03.2003          | AU 2000-34940 A1<br>EP 1145177 A2<br>EP 1145177 B1<br>EP 1406199 A2<br>EP 1406199 A3<br>US 2004-0072232 A1<br>US 6245511 B1<br>US 6704662 B2<br>US 7047136 B2<br>WO 00-51055 A2<br>WO 00-51055 A3 | 14.09.2000<br>17.10.2001<br>20.10.2004<br>07.04.2004<br>11.10.2006<br>15.04.2004<br>12.06.2001<br>09.03.2004<br>16.05.2006<br>31.08.2000<br>31.08.2000 |