



(51) International Patent Classification:

A61K 31/277 (2006.01) A61K 31/44 (2006.01)

A61K 31/337 (2006.01) A61K 31/473 (2006.01)

A61K 31/437 (2006.01) A61K 31/475 (2006.01)

(21) International Application Number:

PCT/US2017/044712

(22) International Filing Date:

31 July 2017 (31.07.2017)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/368,997 29 July 2016 (29.07.2016) US

(71) Applicant: UNIVERSITY OF MIAMI [US/US]; 1951 NW 7th Avenue, Miami, FL 33136 (US).

(72) Inventors: RAO, Jonnagadda, Sunil; 1951 NW 7th Avenue, Miami, FL 33136 (US). LIU, Hongmei; 1951 NW 7th Avenue, Miami, FL 33136 (US).

(74) Agent: STEPHENS, Paul, B.; Marshall, Gerstein & Borun LLP, 233 S. Wacker Drive, 6300 Willis Tower, Chicago, IL 60606-6357 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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, SA, 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, ST, 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, KM, ML, MR, NE, SN, TD, TG).

Published:

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

(54) Title: PRECISION THERAPEUTIC BIOMARKER SCREENING FOR CANCER

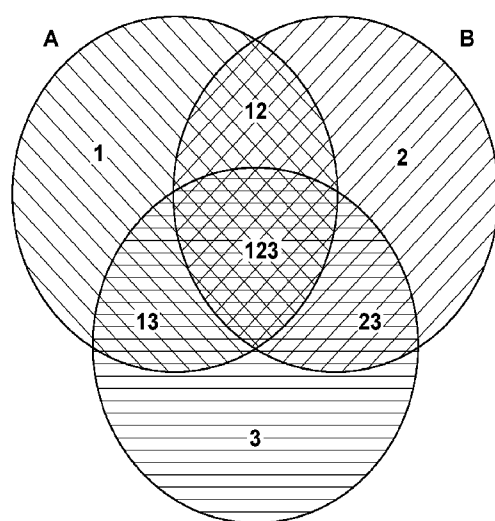


FIG. 1

(57) Abstract: Computer-implemented processes for precision therapeutic biomarker screening for cancer include using a model-based overlapping clustering framework to assess large numbers of possible drugs and drug combinations against patient data, including cell line responsiveness. A multivariate regression model has been developed, along with a latent overlapping cluster indicator variable. The techniques employ a new finite mixture of multivariate regression (FMMR) model and expectation-maximization (EM) algorithm for modeling. The techniques can analyze large amounts of drug data and identify complex overlapping drug clusters, as well as cluster-wise drivers that facilitate identification of new drugs for treating pathologies, such as cancer, in patients.



PRECISION THERAPEUTIC BIOMARKER SCREENING FOR CANCER

CROSS REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to U.S. Provisional Application No. 62/368,997, entitled Precision Therapeutic Biomarker Screening for Cancer, and filed July 29, 2016, which is incorporated herein by reference in its entirety.

FIELD OF THE DISCLOSURE

[0002] The present disclosure relates generally to precision treatment of cancer tissue and therapeutic biomarker screening, and, more particularly, to techniques screening biomarkers based on an overlapping cluster grouping technique.

BACKGROUND

[0003] The background description provided herein is for the purpose of generally presenting the context of the disclosure. Work of the presently named inventor, to the extent it is described in this background section, as well as aspects of the description that may not otherwise qualify as prior art at the time of filing, are neither expressly nor impliedly admitted as prior art against the present disclosure.

[0004] The use of drugs to selectively target specific genetic alterations in defined patient subpopulations has seen significant successes. One example can be found in the treatment of chronic myeloid leukemia (CML) where the first consistent chromosomal abnormality associated with a human cancer was identified back in the 1960s. Fast forward to the 1980s where the consequence of this abnormality was discovered to be the production of an abnormal gene called BCR-ABL. Intense drug discovery programs were initiated to shut down the activity of BCR-ABL, and in 1992, imatinib (Gleevec) was developed. In 1998, the drug was tested in CML patients who had exhausted standard treatment options and whose life expectancy was limited, with remarkable results in their blood counts returning to normal. In 2001, the FDA approved imatinib. The result, a once commonly fatal cancer now has a five-year survival rate of 95% (Druker, Guilhot, O'Brien, Gathmann, Kantarjian, Gattermann, Deininger, Silver, Goldman, Stone et al. 2006).

[0005] Achievements like this largely inspire today's high throughput studies of linking cancer drugs (known or in development) to specific genomic changes which could be used as therapeutic biomarkers. The hope is that such analyses will shed light on biological mechanisms underlying drug sensitivity, tumor resistance and potential drug combination synergies. Such discoveries not only motivate development of precision cancer therapies, improvements of current therapies, and explorations of new therapeutic avenues, they also guide early developments of new cancer drugs (Garnett, Edelman, Heidorn, Greenman, Dastur, Lau, Greninger, Thompson, Luo, Soares et al. 2012; Barretina, Caponigro, Stransky, Venkatesan, Margolin, Kim, Wilson, Lehár, Kryukov, Sonkin et al. 2012).

[0006] Cancer cell lines have frequently been used as a convenient way of conducting such studies. For a systematic search of therapeutic biomarkers to a variety of cancer drugs, Garnett et al. (2012) screened 639 human tumor cell lines which represent much of the tissue-type and genetic diversity of human cancers with 130 drugs. These drugs, including approved drugs, drugs in development as well as experimental tool compounds, cover a wide range of targets and processes involved in cancer biology. In the Garnett et al. work, a range of 275-507 cell lines were screened for each drug. The effect of a 72h drug treatment on cell viability was used to derive measures of drug sensitivity, including the half- maximal inhibitory concentration (IC_{50}). In addition, the cell lines underwent sequencing of 64 known cancer genes, genome-wide analysis of copy number gains and losses, and expression profiling of 14,500 genes.

[0007] In analysis of this data, Garnett et al. used multivariate analysis of variance (MANOVA) to examine pairwise drug-(genomic) feature associations. This process largely limits knowledge discovery in that: 1) only categorical features can be used – as a result over 95% of the features were not examined; 2) the drug-feature associations can change dramatically by considering other features' impacts, hence the marginal associations rarely reflect true relationships. It is more likely that sensitivity of cancer cells to drugs depends on a multiplicity of genomic and epigenomic features with potential interactions. Therefore linear regression with an elastic-net penalty (Zou and Hastie 2005) was further applied to each drug to identify significantly related features while adjusting for other effects.

[0008] However given the degree of complexity of the data, this later elastic-net regression applied for each drug was far from sufficient. Firstly, the 639 tumor cell lines came from a variety of cancer tissue types, hence there is likely additional heterogeneity manifested as subpopulations in data. Secondly, the 130 drugs are hardly independent with each other. It has long been recognized that one can improve the prediction performance by modeling with multiple response variables (Breiman and Friedman 1997).

SUMMARY OF THE INVENTION

[0009] As a result of the foregoing shortcomings in the state of the art, the present application addresses some direct questions of interest. These include, i) can cancer-specific therapeutic biomarkers be detected, ii) can drug resistance patterns be identified along with predictive strategies to circumvent resistance using alternate drugs, and iii) can biomarkers of combination therapies be identified to help predict synergies in drug activities?

[0010] To address these questions a number of statistical challenges exist: i) therapeutic biomarkers of drug sensitivity cluster among the cell lines; ii) clusters can overlap (e.g. a cell line may belong to multiple clusters); and iii) can a multivariate framework be developed so that drugs can be modeled jointly.

[0011] The present application describes new statistical modeling techniques that address these issues using a finite mixture of multivariate regression (FMMR) model generalized to enable overlapping clustering and the numerical solutions of drug data, genomic data, biomarker data, demographic data, pathology data (e.g., cell line data, etc.).

[0012] In accordance with an example, a method of identifying alternative drug therapies, the method comprises: receiving, at a processing unit, drug responsiveness data for a treatment drug applied to a plurality of different cell lines, each cell line differing in genomic profile from one another such that the genomic profiles for the plurality of different cell lines are mutually exclusive, the drug responsiveness data being determined from live cell testing using the treatment drug; identifying at least two cell lines from the plurality of different cell lines, and determining, at the processing unit, an overlap cell line cluster where one of the at least two cell lines has a high drug responsiveness value and another of the at least two cell lines has a low drug responsiveness value, and wherein the overlap cell line cluster has a

drug responsiveness value for the treatment drug that is between the high drug responsiveness value and the low drug responsiveness value, wherein determining the overlap cell line cluster using a generalized finite mixture of multivariate regression (FMMR) model; performing, at the processing unit, an optimality algorithm process on data for a cluster of drugs to identify one or more alternative drugs exhibiting a drug responsiveness value between the high drug responsiveness value and the low drug responsiveness value; and administering to cells of one of the at least two cell lines the one or more alternative drugs.

[0013] In accordance with another example, a method of grouping drugs for treating a pathology, the method comprises: receiving, at a processing unit, drug responsiveness data for a set of drugs and for a set of cell lines; grouping, using the processing unit, each of the drugs into a different drug grouping, using (i) a generalized finite mixture of multivariate regression (FMMR) model as follows

$$f(\mathbf{y}_i | \mathbf{x}_i, \Theta) = \sum_{(l_1, \dots, l_s) \in T} \pi_{(l_1, \dots, l_s)} f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \Sigma).$$

wherein $f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \Sigma)$ is a multivariate normal density function generalized for both non-overlapping clusters and overlapping clusters, y_i are drug responsiveness values, $\mathbf{x}_i$ are genomic biomarkers, $\mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}$ are coefficient matrix values for a cluster, $l_1 \dots l_s$ are clusters, π is an observation value indicating whether a drug belongs in a grouping ($\pi = 1$) or not ($\pi = 0$), and Σ is an error term for the coefficient matrix values, i represents a drug, wherein the generalized FMMR model allows for modeling drugs for both non-overlapping clusters and overlapping clusters, and (ii) a clustering algorithm configured to group drugs based on the coefficient matrix values, $\mathbf{B}$, for different drugs and based on cell line assignments; and administering one or more drugs from one of the drug groupings to a cell line expressing the pathology.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The figures described below depict various aspects of the system and methods disclosed herein. It should be understood that each figure depicts an embodiment of a particular aspect of the disclosed system and methods, and that each of the figures is intended to accord with a possible embodiment thereof. Further, wherever possible, the following description refers to the reference numerals included in the following figures, in which features depicted in multiple figures are designated with consistent reference numerals.

[0015] Fig. 1 illustrates a plurality ($K = 3$) object clusters, showing that clusters can overlap resulting in $2^K - 1$ overlapping patterns.

[0016] Figs. 2A and 2B illustrate specificity measurement results for an example first analyzed scenario where there are no overlapping clusters, showing that techniques herein retain capture the un-overlapping clusters. Fig. 2C illustrates F1 measurements for this first analyzed scenario. Fig. 2D illustrates a coefficient error for this first analyzed scenario.

[0017] Figs. 3A and 3B illustrate specificity measurement results for an example second analyzed scenario where there are (30%) overlapping clusters. Fig. 3C illustrates F1 measurements for this second analyzed scenario. Fig. 3D illustrates a coefficient error for this second analyzed scenario.

[0018] Figs. 4A-4C illustrate drug groupings that result from application of an affinity-propagation clustering (APC) algorithm, showing different nodes, each node corresponding to a different drug as a second-level cluster.

[0019] Fig. 5A illustrates a mapping of estimation results of cell lines with a fitting of second-level drug cluster using a generalized FMMR model, in accordance with an example.

[0020] Fig. 5B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with an example.

[0021] Fig. 6A illustrates a mapping of estimation results of cell lines with a fitting of second-level drug cluster using a generalized FMMR model, in accordance with another example.

[0022] Fig. 6B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with another example.

[0023] Fig. 7A illustrates a mapping of estimation results of cell lines with a fitting of second-level drug cluster using a generalized FMMR model, in accordance with yet another example.

[0024] Fig. 7B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with yet another example.

[0025] Fig. 8A illustrates a mapping of estimation results of cell lines with a fitting of second-level drug cluster using a generalized FMMR model, in accordance with yet another example.

[0026] Fig. 8B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with an example.

[0027] Fig. 9A illustrates drug responsive data for a plurality of drugs for a soft-tissue cancer.

[0028] Fig. 9B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with an example.

[0029] Fig. 10A illustrates drug responsive data for a plurality of drugs for a soft-tissue cancer.

[0030] Fig. 10B illustrates a plot of estimate coefficient matrix showing drugs for different genes, in accordance with an example.

[0031] Figs. 11A-11E illustrate the drug responsiveness data for a plurality of drugs out of a larger examined listing of drugs.

[0032] Figs. 12A and 12B illustrate predicted drug responsive data for different combination combinations, in accordance with an example.

[0033] Fig. 13 illustrates a process for clustering (or grouping) drugs and identifying alternative drugs based on drug responsive data such as that of Figs. 12A and 12B, in accordance with an example.

[0034] Fig. 14 illustrates a process for clustering (or grouping) drugs and generating biomarker mapping of drug groupings corresponding to the cell line clusters of Fig. 13, in accordance with an example.

[0035] Fig. 15 illustrates a drug analysis system for implementing the techniques herein, in accordance with an example.

[0036] Fig. 16 illustrates Table 1 showing sensitive versus drug resistant drugs for soft tissue cancer, identified by the drug analysis system of Fig. 15, in accordance with an example.

[0037] Fig. 17 illustrates a plaid model Algorithm 1, in accordance with an example.

[0038] Fig. 18 illustrates a all-plaid model Algorithm 2, in accordance with an example.

DETAILED DESCRIPTION

[0039] The present techniques provide computer-implemented processes for precision therapeutic biomarker screening for cancer. The techniques use a model-based overlapping clustering framework to assess large numbers of possible drugs and drug combinations against patient data, including cell line responsiveness, genetic data, phenotype data, demographic data, etc. A multivariate regression model has been developed, along with a latent overlapping cluster indicator variable. The techniques employ a new finite mixture of multivariate regression (FMMR) model and EM algorithm for modeling. The techniques have been used to analyze large amounts of drug data and have identified complex overlapping drug clusters, as well as cluster-wise drivers that facilitate identification of new drugs for treating pathologies, such as cancer, in patients.

[0040] Clustering is a well-established statistical technique which involves grouping data elements according to a measure of similarity. Traditional clustering techniques generate partitions so that each data element belongs to one and only one cluster. It has long been recognized that such an ideal clustering seldom exists in real data. It is more likely that clusters overlap in some parts.

[0041] Some statistical methods have been developed to handle the overlapping clustering problem. Lazzeroni and Owen (2002), for example, put forward the so-called plaid model for two-sided overlapping clustering of gene expression data. However, the numerical solutions proposed in the paper can produce unsatisfactory cluster retrieval results (see Turner, Bailey and Krzanowski (2005) and Turner, Bailey, Krzanowski and Hemingway (2005) for improved and extended (to structured data) plaid models, and see, Zhang (2010) for the Bayesian plaid model formulation). Segal, Battle and Koller (2002) introduced a probabilistic model for one-sided overlapping clustering. Banerjee, Krumpelman, Ghosh, Basu and Mooney (2005) later generalized this probabilistic model to any regular exponential family distribution. However both of the methods suffer from computational complexity in estimating a binary membership matrix $\mathbf{M} \in \{0, 1\}^{n \times K}$, which is an integer optimization problem.

[0042] An alternative “naive” overlapping clustering method is the finite mixture (FM) model with a hard threshold α on cluster membership probabilities. This approach assigns a data point to a cluster k if $P(z_{ik} = 1 \mid \mathbf{X}_i, \Theta) > \alpha$ and hence enables an item to belong to multiple clusters.

[0043] As pointed out by Banerjee et al. (2005), this conventional FM method is problematic because: 1) it is hard to choose the value of α ; and 2) it is not a natural generative model for overlapping clustering since one underlying assumption of the mixture model is that each object comes from only one mixture component. Later Fu and Banerjee (2008) introduced an overlapping clustering approach based on the multiplicative mixture model (Heller and Ghahramani 2007). However, this model ended up with the same computational issue as in Banerjee et al. (2005).

[0044] Some have focused on investigating the relationship between response variables $\mathbf{Y} \in \mathbb{R}^q$ and covariates $\mathbf{X} \in \mathbb{R}^p$ by fitting a regression model rather than to explore the X or Y on its own. In the context of regression analysis for $q = 1$, finite mixture of regression (FMR) models are commonly used to capture unobserved cross-sectional heterogeneity in the data (see, e.g., Jedidi, Ramaswamy, DeSarbo and Wedel 1996). The FMR model postulates

that a sample of observations come from a finite mixture of latent sub-populations with each sub-population represented by a specific regression model.

[0045] In 1972, Quandt proposed a two component mixture of univariate Gaussian linear regression model with the idea of using maximum likelihood function to estimate model parameters. Subsequently, the well-known expectation-maximization (EM) algorithm was used for maximum likelihood estimation (MLE) with incomplete data, revolutionizing studies of the FMR model, with various researchers contributing the advancing efforts.

[0046] However, looking at each of these conventional techniques, the results insufficient for true therapeutic applications, especially as data sets grow in size and complexity. Notably, none of these conventional techniques offer a truly generalized model, one that provides for overlapping (i.e., clustering) data sets for optimized variable selection. Data distributions depart from expected normality. In terms of statistical accuracy, breakdowns occur from forced approximations, such as setting mixture components as a constant. But most importantly, none of the above listed methods allow for simultaneous variable selection. In order to do variable selection, these data models were constrained to the point where response variables were treated as independent from one another. Yet, clinical testing shows, especially for pathologies like various cancers, that selection variables can be interdependent. Therefore, these proposed techniques are not suitable for cancer drug screening.

[0047] Overcoming these and other limitations of conventional techniques, with the advanced techniques described herein, we are able to show that biomarkers indicative of a first pathology, say soft tissue cancer, may well have demonstrated drug treatments that alone or combined with other drug treatments may be used to treat a second pathology, such as cancer in the bone or kidney.

[0048] The present techniques provide, among other things, a computer-implemented process using a model-based overlapping clustering framework. As detailed in examples below, a multivariate regression model is provided and a latent overlapping cluster indicator variable is introduced. The present techniques employ a new finite mixture of multivariate regression (FMMR) model and expectation-maximization (EM) algorithm to fit the new model. Penalized likelihood with a lasso, Elastic-Net, or other penalty function may be used

to estimate model parameters meanwhile allowing variable selection. Simulation studies show excellent finite-sample performance that outperforms other methods. Analysis of drug data, for example, has found complex overlapping drug clusters as well as cluster-wise drivers that facilitate identification of new drugs for treating pathologies such as cancer.

[0049] As we discuss, the present techniques are able to assess multiple variables across cancer cell line targets to record drugs and drug combinations for treatment.

[0050] For a sample (of cell lines) of n observations, denote $\mathbf{y}_i = (y_{i1}, \dots, y_{iq})^T \in \mathbb{R}^q$ a vector of responses (IC_{50} values), $\mathbf{x}_i = (x_{i1}, \dots, x_{ip_n})^T \in \mathbb{R}^{p_n}$ a vector of predictors (genomic markers) and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iq})^T$ a vector of random errors for the i th observation. Assume $\boldsymbol{\varepsilon}_i \stackrel{iid}{\sim} N_q(\mathbf{0}, \boldsymbol{\Sigma})$ for $i = 1, \dots, n$. The following allows overlapping clustering of cell lines within the context of a multivariate regression model,

$$\mathbf{y}_i = \sum_{k=1}^K \mathbf{B}_k^T \mathbf{x}_i P_{ik} + \boldsymbol{\varepsilon}_i, \quad i = 1, \dots, n, \quad (2.1)$$

where K is the total number of clusters, $\mathbf{B}_k$ is a $p_n \times q$ coefficient matrix for the k th cluster, and $P_{ik} \in \{0, 1\}$ is 1 if observation i belongs to the k th cluster, otherwise 0. In traditional clustering problem, it assumes that each observation belongs to exactly one cluster, namely $\sum_{k=1}^K P_{ik} = 1$ for all i . Therefore we allow $\sum_{k=1}^K P_{ik} \geq 1$ so that each observation can belong to multiple clusters.

[0051] We provide some interpretation of the clusters in model (2.1). A cluster k in the model contains a subset of observations for whom $P_{ik} = 1$. We will assume that not all genomic feature are relevant in describing the cluster of observations and thus will assume a sparse coefficient matrix $\mathbf{B}_k$. Since the sparse pattern can vary with k , each cluster can be represented by a unique set biomarkers. For samples belonging to multiple clusters, their response variables are explained by multiple sets of biomarkers, indicative of involving several biological processes simultaneously.

[0052] We examined two approaches for fitting (2.1): a generalized finite mixture of multivariate regressions model and a generalization of the plaid model previously discussed.

ESTIMATION PROCEDURE

Finite Mixture of Multivariate Linear Regression

[0053] When $\sum_{k=1}^K P_{ik} = 1$, model (2.1) can be characterized by a hierarchical structure which ends up with an FMMR model. Consider a latent cluster membership random variable z_i for observation i . Given $\sum_{k=1}^K P_{ik} = 1$, the range of z_i equals to $\{1, \dots, K\}$. Assume that $P(z_i = k) = \pi_k$ for each k . By properties of probability, it satisfies that $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$. Moreover for model (2.1), it has

$$(\mathbf{y}_i | z_i = k, \mathbf{x}_i, \mathbf{B}_k, \boldsymbol{\Sigma}) \sim N_q(\mathbf{B}_k^T \mathbf{x}_i, \boldsymbol{\Sigma}). \quad (3.1)$$

Denote $\boldsymbol{\Theta} = (\mathbf{B}_1, \dots, \mathbf{B}_K, \boldsymbol{\Sigma}, \boldsymbol{\pi})$ with $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{K-1})^T$. We can derive the joint density of $\mathbf{y}_i$ and z_i as

$$f(\mathbf{y}_i, z_i | \mathbf{x}_i, \boldsymbol{\Theta}) = \prod_{k=1}^K \{\pi_k f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_k, \boldsymbol{\Sigma})\}^{I(z_i=k)}, \quad (3.2)$$

where $f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_k, \boldsymbol{\Sigma})$ is a multivariate normal density function defined in (3.1).

Summarizing (3.2) over $z_i = 1, \dots, K$ yields the FMMR model

$$f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{k=1}^K \pi_k f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_k, \boldsymbol{\Sigma}). \quad (3.3)$$

[0054] When using the EM algorithm to estimate model (3.3), $(\mathbf{x}_i; \mathbf{y}_i)$ is regarded as an incomplete sample observation for missing cluster membership variable z_i , one instead deals with the joint density of (3.2) with complete data. The complete (conditional) log-likelihood of $\boldsymbol{\Theta}$ given a sample of n independent observations from model (2.1) becomes

$$l_n(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log \pi_k + \sum_{i=1}^n \sum_{k=1}^K z_{ik} \log f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_k, \boldsymbol{\Sigma}), \quad (3.4)$$

where $z_{ik} = I(z_i = k)$.

[0055] Variable selection, to obtain a parsimonious model, is essential especially to high dimensional statistical modeling for enhanced scientific discovery and improved prediction accuracy (Fan and Li 2006). Khalili and Chen (2012) introduced a penalized likelihood approach for variable selection in FMR models, which was shown to be variable selection consistent and computationally much more efficient than all-subset selections. Similarly, we derived a penalized log-likelihood of Θ as

$$\tilde{l}_n(\Theta) = l_n(\Theta) - \rho_n(\Theta), \quad (3.5)$$

where

$$\rho_n(\Theta) = \sum_{k=1}^K \pi_k \rho_{nk}(\mathbf{B}_k).$$

[0056] The following two penalty functions were used to meet different demands in variable selection:

1) the L_1 -penalty in LASSO (Tibshirani 1996) for simultaneous estimation and variable selection

$$\rho_{nk}(\mathbf{B}_k) = \lambda_k \|\mathbf{B}_k\|_1$$

2) a linear combination of the L_1 - and L_2 -penalty in Elastic-Net (Zou and Hastie, 2005) for simultaneous estimation and selection of grouped features

$$\rho_{nk}(\mathbf{B}_k) = \lambda_k^1 \|\mathbf{B}_k\|_1 + \lambda_k^2 \|\mathbf{B}_k\|_2^2,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are respectively the L_1 - and L_2 -norms and the tuning parameters $\lambda_k, \lambda_k^1, \lambda_k^2 \geq 0$.

[0057] In next section, revised FMMR model and EM algorithm are devised for overlapping clustering multivariate regression data.

GENERALIZED FMMR MODEL AND EM ALGORITHM FOR OVERLAPPING CLUSTERING

[0058] The FMMR model solves (2.1) given $\sum_{k=1}^K P_{ik} = 1$ for each i ,

for each i , wherein the clusters are partitional. When the clusters are overlapping, namely $\sum_{k=1}^K P_{ik} \geq 1$, we propose a generalized FMMR model to address the issue, meanwhile this generalized model preserves the ability in fitting partitions.

[0059] Suppose we have K objective clusters indexed by 1 to K . Note that these clusters can overlap with each other, resulting in $2^K - 1$ types of overlapping patterns. Let K equal to 3 for an example, the overall overlapping patterns are composed of $S = \{1; 2; 3; 12; 13; 23; 123\}$. Here element 12 in S represents exclusively the overlapping part of objective clusters 1 and 2, whereas element 1 represents a subset of objective cluster 1 which does not have overlappings with all other objective clusters. Fig. 1 illustrates the particular part each overlapping pattern in S represents. By definition, overlapping patterns in S are mutually exclusive. And each observation i from (2.1), given $\sum_{k=1}^K P_{ik} \geq 1$, belongs to one and only one overlapping pattern of S .

[0060] We define $2^K - 1$ hypothetical clusters. Each hypothetical cluster (called “cluster” herein) indicates an overlapping pattern defined above and is indexed by an element in T wherein

$$T = \bigcup_{s=1}^K \{(l_1 \dots l_s) : \{l_1, \dots, l_s\} \subseteq \{1, \dots, K\}\}.$$

Cluster $(l_1 \dots l_s)$ implies that its members belong to objective clusters $l_1, \dots, l_s$.

[0061] We introduce a latent cluster membership random variable z_i for observation i from (2.1), and characterize model (2.1) by a hierarchical structure. Given $\sum_{k=1}^K P_{ik} \geq 1$, the range of z_i becomes the set T . We further postulate that

$$P(z_i = (l_1 \dots l_s)) = \pi_{(l_1 \dots l_s)}. \quad (3.6)$$

By properties of probability, vector $\boldsymbol{\pi} = (\pi_{(l_1 \dots l_s)})_{(l_1 \dots l_s) \in T}^T$ satisfies that

$$\sum_{(l_1 \dots l_s) \in T} \pi_{(l_1 \dots l_s)} = 1, \quad \pi_{(l_1 \dots l_s)} \geq 0.$$

Moreover for (2.1), it amounts to $P_{d_1}, \dots, P_{d_s} = 1$ given $z_i = (l_1 \dots l_s)$. Hence it has

$$(y_i | z_i = (l_1 \dots l_s), \mathbf{x}_i, \{\mathbf{B}_k\}_{k=l_1}^{l_s}, \boldsymbol{\Sigma}) \sim N_q \left(\sum_{k=l_1}^{l_s} \mathbf{B}_k^T \mathbf{x}_i, \boldsymbol{\Sigma} \right). \quad (3.7)$$

Let $\boldsymbol{\Theta} = (\mathbf{B}_1, \dots, \mathbf{B}_K, \boldsymbol{\Sigma}, \boldsymbol{\pi})$. By (3.6) and (3.7), the joint density of $\mathbf{y}_i$ and z_i equals to

$$f(\mathbf{y}_i, z_i | \mathbf{x}_i, \boldsymbol{\Theta}) = \prod_{(l_1 \dots l_s) \in T} \{ \pi_{(l_1 \dots l_s)} f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \boldsymbol{\Sigma}) \}^{I(z_i = (l_1 \dots l_s))}, \quad (3.8)$$

where $f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \boldsymbol{\Sigma})$ is a multivariate normal density function defined in (3.7).

[0062] By summarizing (3.8) over z_i , we obtain a generalized FMMR model

$$f(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\Theta}) = \sum_{(l_1 \dots l_s) \in T} \pi_{(l_1 \dots l_s)} f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \boldsymbol{\Sigma}). \quad (3.9)$$

[0063] Note that if $\pi_{(l_1 \dots l_s)} = 0$ for $s > 1$, (3.9) reduces to traditional FMMR models in (3.3).

[0064] For a sample of n independent observations from model (2.1), the complete (conditional) log-likelihood of $\boldsymbol{\Theta}$ becomes

$$l_n(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{(l_1 \dots l_s) \in T} z_{i,(l_1 \dots l_s)} \log \pi_{(l_1 \dots l_s)} + \sum_{i=1}^n \sum_{(l_1 \dots l_s) \in T} z_{i,(l_1 \dots l_s)} \log f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \boldsymbol{\Sigma}),$$

where $z_{i,(l_1 \dots l_s)} = I(z_i = (l_1 \dots l_s))$.

[0065] We derived a penalized log-likelihood of $\boldsymbol{\Theta}$ which enforces sparsity in genomic features.

$$\tilde{l}_n(\Theta) = l_n(\Theta) - \sum_{k=1}^K \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} \pi_{(l_1 \dots l_s)} \rho_{nk}(\mathbf{B}_k), \quad (3.10)$$

where $\rho_{nk}(\mathbf{B}_k)$ utilizes one of two penalty functions, the LASSO penalty ($\rho_{nk}(\|\mathbf{B}_k\|) = \lambda_k \|\mathbf{B}_k\|_1$) or the Elastic-Net penalty ($\rho_{nk}(\|\mathbf{B}_k\|) = \lambda_k^1 \|\mathbf{B}_k\|_1 + \lambda_k^2 \|\mathbf{B}_k\|_{2,\infty}^2$) for

grouped features as defined herein. The notation $\sum_{k \in \{l_1, \dots, l_s\}} \pi_{(l_1 \dots l_s)}$ implies that it summarizes over $(l_1, \dots, l_s) \in T$ for which $k \in \{l_1, \dots, l_s\}$. In (3.10), $\mathbf{B}_k$ is a coefficient matrix corresponding to the k th objective cluster. The penalty imposed on $\mathbf{B}_k$ is proportional to

$\sum_{k \in \{l_1, \dots, l_s\}} \pi_{(l_1 \dots l_s)}$ which depicts the proportion of observations involving in the k th objective cluster. This is a strategy of relating the penalty to sample sizes similar to Khalili and Chen (2012) for enhanced power of the method.

Numerical Solution to the Generalized FMMR Model

[0066] A new expectation maximization (EM) algorithm to maximize the penalized log-likelihood in ($\tilde{l}_n(\Theta)$) (3.10) with a LASSO penalty function. Due to the additive property for overlapping clusters, each $\mathbf{B}_k$ involves in multiple clusters. Therefore coefficient matrix $\mathbf{B}_k$ for $k = 1; \dots; K$ cannot be optimized independently within each cluster like in usual EM algorithm for FMR models. Instead we sequentially update each $\mathbf{B}_k$ given the rest is known. Specifically, the revised EM algorithm iteratively maximizes the $\tilde{l}_n(\Theta)$ in following two steps:

[0067] E-step: Given Θ^m , estimate $z_{i,(l_1 \dots l_s)}^m$ with its posterior probability by applying the Bayes' rule to (3.8) and (3.9),

$$\hat{P}(z_{i,(l_1 \dots l_s)}^m = 1 \mid \mathbf{y}_i, \mathbf{x}_i, \Theta^m) = \frac{\pi_{(l_1 \dots l_s)}^m f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{B}_{l_1}^m, \dots, \mathbf{B}_{l_s}^m, \Sigma^m)}{\sum_{(l'_1 \dots l'_s) \in T} \pi_{(l'_1 \dots l'_s)}^m f(\mathbf{y}_i \mid \mathbf{x}_i, \mathbf{B}_{l'_1}^m, \dots, \mathbf{B}_{l'_s}^m, \Sigma^m)}$$

for $i = 1, \dots, n$ and $(l_1 \dots l_s) \in T$.

M-step: Given $\mathbf{Z}^m = (z_{i,(l_1 \dots l_s)}^m)$, update the mixing proportions π^{m+1} by

$$\pi_{(l_1 \dots l_s)}^{m+1} = \frac{1}{n} \sum_{i=1}^n z_{i,(l_1 \dots l_s)}^m$$

for $(l_1 \dots l_s) \in T$. Note that this is obtained by maximizing the leading term of (3.10) with respect to π for simplicity. This simplified updating scheme however works well in simulation studies.

Given $\mathbf{Z}^m$, π^{m+1} and Σ^m , sequentially update $\mathbf{B}_k^{m+1} \mid \{\mathbf{B}_{s,s \neq k}^m\}$ by

$$\begin{aligned} \mathbf{B}_k^{m+1} = \max_{\mathbf{B}_k} \sum_{i=1}^n \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} z_{i,(l_1 \dots l_s)}^m \log f(\mathbf{y}_i \mid \mathbf{x}_i, \{\mathbf{B}_l^m\}_{l=(l_1 \dots l_s) \neq k}, \Sigma^m) \\ - \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} \pi_{(l_1 \dots l_s)}^{m+1} \lambda_k \|\mathbf{B}_k\|_1, \end{aligned} \quad (3.11)$$

for $k = 1, \dots, K$.

[0068] According to (3.7), $\mathbf{y}_i$ for $i = 1, \dots, n$ are i.i.d. in multivariate normal. Hence (3.11) becomes

$$\begin{aligned} \mathbf{B}_k^{m+1} = \min_{\mathbf{B}_k} \sum_{i=1}^n \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} z_{i,(l_1 \dots l_s)}^m (\mathbf{y}_i^* - \mathbf{B}_k^T \mathbf{x}_i)^T (\Sigma^m)^{-1} (\mathbf{y}_i^* - \mathbf{B}_k^T \mathbf{x}_i) \\ + 2 \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} \pi_{(l_1 \dots l_s)}^{m+1} \lambda_k \|\mathbf{B}_k\|_1, \\ = \min_{\mathbf{B}_k} \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} \text{tr}((\mathbf{Y}^* - \mathbf{X} \mathbf{B}_k)^T \mathbf{z}_{(l_1 \dots l_s)}^m (\mathbf{Y}^* - \mathbf{X} \mathbf{B}_k) (\Sigma^m)^{-1}) \\ + 2 \sum_{\substack{(l_1 \dots l_s) \\ k \in \{l_1, \dots, l_s\}}} \pi_{(l_1 \dots l_s)}^{m+1} \lambda_k \|\mathbf{B}_k\|_1, \end{aligned} \quad (3.12)$$

where

$$\begin{aligned} \mathbf{y}_i^* &= \mathbf{y}_i - \sum_{l \in (l_1, \dots, l_s) \setminus k} (\mathbf{B}_l^m)^T \mathbf{x}_i, \\ \mathbf{Y}^* &= (\mathbf{y}_1^*, \dots, \mathbf{y}_n^*)^T \quad \text{and} \quad \mathbf{z}_{(l_1 \dots l_s)}^m = \text{diag} \left(z_{1,(l_1 \dots l_s)}^m, \dots, z_{n,(l_1 \dots l_s)}^m \right). \end{aligned}$$

[0069] Optimization of (3.12) is a multivariate regression problem with a LASSO penalty for estimation, which can be solved by the MRCE algorithm (Rothman, Levina and Zhu 2010). If we ignore the covariance structure of Σ^m in (3.12), the optimization reduces to q independent LASSO regression problems. In this case, more complex penalty functions say the Elastic-Net or the fused LASSO penalty can be easily applied. Therefore we also investigated the performance of this simplified strategy in simulation studies, and found the present techniques work surprisingly well comparable, in terms of clustering accuracy, to the original method utilizing the MRCE algorithm.

Given $\mathbf{Z}^m$ and $\{\mathbf{B}_k^{m+1}\}_{k=1}^K$, update Σ^{m+1} by

$$\begin{aligned}\Sigma^{m+1} &= \max_{\Sigma} \sum_{i=1}^n \sum_{(l_1 \dots l_s) \in T} z_{i(l_1 \dots l_s)}^m \log f(y_i | \mathbf{x}_i, \{\mathbf{B}_l^{m+1}\}_{l=l_1}^{l_s}, \Sigma) \\ &= \min_{\Sigma} \sum_{i=1}^n \sum_{(l_1 \dots l_s) \in T} -z_{i(l_1 \dots l_s)}^m \log |\Sigma^{-1}| \\ &\quad + \sum_{(l_1 \dots l_s) \in T} \text{tr}((\mathbf{Y} - \mathbf{X} \sum_{l=l_1}^{l_s} \mathbf{B}_l^{m+1})^T \mathbf{z}_{i(l_1 \dots l_s)}^m (\mathbf{Y} - \mathbf{X} \sum_{l=l_1}^{l_s} \mathbf{B}_l^{m+1}) \Sigma^{-1}).\end{aligned}\quad (3.13)$$

By taking the derivative of (3.13) according to Σ^{-1} , we get

$$\Sigma^{m+1} = \frac{\sum_{(l_1 \dots l_s) \in T} (\mathbf{Y} - \mathbf{X} \sum_{l=l_1}^{l_s} \mathbf{B}_l^{m+1})^T \mathbf{z}_{i(l_1 \dots l_s)}^m (\mathbf{Y} - \mathbf{X} \sum_{l=l_1}^{l_s} \mathbf{B}_l^{m+1})}{\sum_{(l_1 \dots l_s) \in T} \sum_{i=1}^n z_{i(l_1 \dots l_s)}^m},$$

where $\sum_{(l_1 \dots l_s) \in T} \sum_{i=1}^n z_{i(l_1 \dots l_s)}^m = n$.

[0070] Although we can also penalize the inverse covariance matrix Σ^{-1} in (3.13) like in Rothman et al. (2010), our simulations have shown that penalizing the Σ^{-1} results in a lower degree of clustering accuracy than not penalizing the Σ^{-1} . This is because by penalizing the Σ^{-1} one introduces bias in its estimation, and a biased estimate of Σ^{-1} in return deteriorates the estimate of $z_i(l_1 \dots l_s)$ in the E-step. Thus we choose to not penalize the Σ^{-1} .

[0071] Commencing with an initial value Θ^0 , the present models may iterate between the E- and M-steps until the relative change in log-likelihood, $|(l_n^{m+1}(\hat{\Theta}) - l_n^m(\hat{\Theta})) / l_n^m(\hat{\Theta})|$, is smaller than some threshold value, taken as 10^{-5} in simulation studies and 10^{-3} in real data

analysis. Additionally, a cluster, whose mixing proportion is smaller than some threshold value taken as 0.01 in the paper, was removed during iterations to avoid over estimations.

Selection of the Tuning Parameters and K

[0072] In preceding penalized likelihood approach, the techniques can choose the sizes of component-wise tuning parameters λ_k for $k = 1, \dots, K$, which controls the complexity of an estimated model (Fan and Lv 2010). The data-driven method cross-validation (CV) (Stone 1977) is frequently adopted in literature (Tibshirani 1996; Zou and Hastie 2005). In an example, we used a component-wise 10-fold CV method for optimal tuning parameter selection in the expression (3.12).

[0073] Selection of the number of components K is performed in finite mixture models (including the FM, FMR and FMMR). In applications, the choice can be based on prior knowledge of data analysts. With respect to formatted methodologies, information criteria (IC) remains by far the most popular strategy for selection of K . See Fraley and Raftery (1998), Anderson and Burnham (2002) and Claeskens, Hjort et al. (2008) for general treatments on this topic. Here we choose the K that minimizes the IC,

$$IC_n(K) = -2l_n(\Theta) + N_K a_n. \quad (3.14)$$

In (3.14), N_K is the effective number of parameters in the model with

$$N_K = |\{\mathbf{B}_k, k = 1, \dots, K\}| + |\boldsymbol{\pi}| - 1 + |\boldsymbol{\Sigma}|,$$

where $|A|$ calculates the number of nonzero elements in A . Also in (3.14), a_n is a positive sequence depending on n . The well-known AIC (Akaike 1974) and BIC (Schwarz et al. 1978) correspond to $a_n = 2$ and $a_n = \log(n)$ respectively. It has been shown (in Keribin (2000)) that under general regularity conditions, BIC can correctly identify the order of an FM model as n increases to infinity. The feasibility of their results to the FMR or FMMR models is however unknown. However, we have examined the performance of BIC for selection of K in an generalized FMMR model via simulations in supplementary material (Exhibit A, attached herewith). It obtains a high degree of accuracy in identifying the true K in our new model.

GENERALIZATION OF THE PLAID MODEL

[0074] To make comparisons with the generalized FMMR model introduced herein, we generalize the plaid model by introducing covariates into the model such that it can as well cluster overlapping multivariate regression data. To date, none of the conventional plaid model techniques have been conducted on overlapping clustering for (univariate/multivariate) regression data to best of our knowledge. The plaid model and probabilistic model are the two most popular methods for overlapping clustering ordinary data. The reason to select the plaid model rather than the probabilistic model for comparison is because the latter one revolves around Bayesian model like ours. The plaid model however brought in the concept of layers (subsets of rows and columns of a data matrix), wherein a data matrix is decomposed as the sum of the layers. Model (2.1) is the generalized plaid model. It seeks to find the parameter estimators that minimize the Q

$$Q = \frac{1}{2} \sum_{i=1}^n \|\mathbf{y}_i - \sum_{k=1}^K \mathbf{x}_i^T \mathbf{B}_k P_{ik}\|_2^2 + \sum_{k=1}^K \rho_{nk}(\mathbf{B}_k),$$

where $p_{nk}(\mathbf{B}_k)$ is the penalty function defined herein.

[0075] Detailed procedures for optimizing Q are shown in Algorithm 1 in Fig. 17. At each time, the algorithm searches a layer which explains as much of the rest data as possible; once a layer has been found, it is subtracted from the data and remains unchanged; the algorithm stops when no further layers can be found. However this method introduces bias to parameter estimators due to the discrete updating scheme, where previous recruited layers cannot be refined as new layers are recruited into the model. To address this issue, we further revised Algorithm 1 to enable joint optimization of all layers. Detailed procedures are presented in Algorithm 2 in Fig. 18.

ASYMPTOTIC PROPERTIES AND SOME SIMULATION STUDIES

Asymptotic Properties

[0076] Khalili and Lin (2013) studied the penalized likelihood approach for variable selection and estimation in finite mixture of regression models with a diverging p_n . It was shown that this approach leads to consistent estimation as well as variable selection under

certain regularity conditions. Let $f(w; \Theta)$ be the joint density function of $w = (x; y)$ with parameter space $\Theta \in \Omega$. The conditional density function of y given x follows (3.9). Since the regularity conditions given for asymptotic establishments in Khalili and Lin (2013) are made on $p_{nk}(B_k)$ and $f(w; \Theta)$ directly, corresponding theoretical properties can be extended to current problem.

[0077] Denote by B_k^j the j th column of B_k for $j = 1, \dots, q$. Let $B_k^j = (B_{k1}^j, B_{k2}^j)$ to divide the coefficient vector into non-zero and zero subsets. Denote by Θ^0 the true value of Θ , $\Theta^0 = (\Theta_1^0, \Theta_2^0)$ is the corresponding decomposition such that Θ_2^0 contains all zero coefficients, i.e. B_{k2}^j , and $\hat{\Theta}_n = (\hat{\Theta}_{n1}, \hat{\Theta}_{n2})$ is the estimate. Since the dimension of $\hat{\Theta}_{n1}$ increases with n as p_n is diverging, we investigate the asymptotic distribution of a finite linear transformation, $D_n \hat{\Theta}_{n1}$, where D_n is a constant $l \times d_{n1}$ matrix with a finite l , and d_{n1} is the dimension of Θ_1^0 . Moreover $D_n D_n^T \rightarrow D$ and D is a positive definite symmetric matrix. In the theory, we assume that the order K is independent of the sample size n and also is known beforehand. Strategies for its selection in applications are discussed herein and performance of the selection is studied in simulations.

[0078] Theorem 1. Suppose the penalty function $p_{nk}(B_k)$ satisfies conditions $P_0 - P_2$, and the joint density function $f(w; \Theta)$ of a random sample $w = (x; y)$ satisfies conditions $R_1 - R_5$.

I. If $\frac{p_n^A}{n} \rightarrow 0$, then there exists a local maximizer $\hat{\Theta}_n$ for the penalized log-likelihood $\tilde{l}_n(\Theta)$ with

$$\|\hat{\Theta}_n - \Theta^0\| = O_p \left\{ \sqrt{\frac{p_n}{n}} (1 + q_{2n}) \right\},$$

where $q_{2n} = \max_{k,at} \left\{ \frac{|\rho'_{n,k}(b_{k,at}^0)|}{\sqrt{n}}; b_{k,at}^0 \neq 0 \right\}$.

II. If $\rho_{nk}(\mathbf{B}_k)$ also satisfies condition $\mathcal{P}_3$ and $\frac{p_n^S}{n} \rightarrow 0$, for any $\sqrt{n/p_n}$ -consistent maximum penalized log-likelihood estimator $\hat{\Theta}_n$, as $n \rightarrow \infty$ it has:

1. Variable selection consistency:

$$P(\hat{\mathbf{B}}_{k2}^j = 0) \rightarrow 1, \quad k = 1, \dots, K \text{ and } j = 1, \dots, q.$$

2. Asymptotic normality:

$$\sqrt{n} \mathbf{D}_n \mathcal{I}_1^{-1/2}(\Theta_1^0) \left\{ \left[\mathcal{I}_1(\Theta_1^0) - \frac{\rho'_n(\Theta_1^0)}{n} \right] (\hat{\Theta}_{n1} - \Theta_1^0) + \frac{\rho'_n(\Theta_1^0)}{n} \right\} \rightarrow^d N(0, \mathbf{D}), \quad (5.1)$$

where $\mathcal{I}_1(\Theta_1^0)$ is the Fisher information matrix under the true subset model.

Proof. Write $\Theta = (\theta_1, \theta_2, \dots, \theta_{t_n})^T$, where t_n is the total number of parameters in the model. Regularity conditions $\mathcal{P}_0 - \mathcal{P}_3$ and $\mathcal{R}_1 - \mathcal{R}_5$ has the same form as in Khalili and Lin (2013). By following the proofs in Khalili and Lin (2013), conclusions below can be established.

1) Let $\sigma_n = \sqrt{\frac{p_n}{n}} (1 + q_{2n})$. For any given $\varepsilon > 0$, there exists a constant C_ε such that

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{\|\mathbf{u}\| = C_\varepsilon} \tilde{l}_n(\Theta^0 + \sigma_n \mathbf{u}) < \tilde{l}_n(\Theta^0) \right\} \geq 1 - \varepsilon.$$

2) For any $\Theta = (\Theta_1, \Theta_2)$ in the neighborhood $\|\Theta - \Theta^0\| = O(\sqrt{p_n/n})$, it has that

$$P \left\{ \tilde{l}_n(\Theta_1, \Theta_2) < \tilde{l}_n(\Theta_1, 0) \right\} \rightarrow 1 \quad \text{as } n \rightarrow \infty.$$

3)

$$\left\{ -\frac{1}{n} E\{l_n''(\Theta_1^0)\} + \frac{\rho_n''(\Theta_1^0)}{n} \right\} (\Theta_{n1} - \Theta_1^0) + \frac{\rho_n'(\Theta_1^0)}{n} = \frac{l_n'(\Theta_1^0)}{n} + o(1/\sqrt{n}). \quad (5.2)$$

By (5.2), the left side of (5.1), denoted as A , becomes

$$A = \sum_{i=1}^n \mathbf{W}_i + o(1),$$

where $W_i = n^{-\frac{1}{2}} D_n l_n^{-\frac{1}{2}}(\Theta_1^0) [\partial \log f(w_i; \Theta_1^0) / \partial \Theta_1]$, $i = 1, \dots, n$, are independent and identically distributed random vectors. They satisfy the condition of the Lindeberg-Feller central limit theorem and also

$$\sum_{i=1}^n \text{Cov}(\mathbf{W}_i) = \mathbf{D}_n \mathbf{D}_n^T \rightarrow \mathbf{D} \quad \text{as } n \rightarrow \infty.$$

[0079] Theorem 1 can be easily derived from above conclusions. We refer readers to Khalili and Lin (2013) for detailed proofs.

[0080] By Theorem 1.I, $\hat{\Theta}_n$ with respect to LASSO or Elastic Net penalty has convergence rate $\sqrt{p_n/n}$ under appropriate choice of tuning parameters. Note that consistent estimator of Θ^0 does not necessarily guarantee the variable selection consistency. Theorem 1.II looks into conditions under which the estimator is consistent in variable selection. However the LASSO or Elastic Net penalty cannot attain both properties simultaneously because the bias term q_{2n} is proportional to $\lambda k / \sqrt{n}$, while λk must be large enough to achieve sparsity. This problem can be efficiently solved by using the adaptive LASSO or adaptive Elastic Net penalty instead, which fulfills the conditions required for both variable selection and estimation consistency.

Design of the Simulations

[0081] Two scenarios are designed for the unoverlapping and overlapping situations respectively. In each scenario, data are generated from model (2.1) with $K = 3$, $p_n = 15$, $q = 3$

and $n = 150; 450$ respectively. Predictors x_i are IID from $N_{pn}(0; \Sigma_1)$ and random errors ε_i are IID from $N_q(0, \Sigma)$ with $\Sigma_1(i, j) = 0.5^{|i-j|}$ and $\Sigma(i, j) = 0.75^{|i-j|}$.

[0082] The sparse coefficient matrixes B_k , $k = 1, 2, 3$ are generated as follows,

$$B_k = W \otimes S \otimes T, \quad (5.3)$$

where $\otimes$ indicates the element-wise product. Each entry of W is drawn independently from $N(0, 1)$, each entry of S independently from the Bernoulli distribution $B(1, p_1)$, and rows of T (either all 1 or all 0) are determined by p_n independent Bernoulli draws from $B(1, p_2)$. Thus for each response variable, we expect $p_1 p_2 p_n$ relevant predictors, and there will be $(1 - p_2) p_n$ predictors in expectation to be irrelevant to all q responses. We have for all scenarios $p_1 = 0.5$ and $p_2 = 0.9$. Simulations are repeated for 50 times. For each repetition, a new sequence of B_k , $k = 1; 2; 3$ are generated. To show that our proposed method has the ability to identify overlapping clusters, we assume K is known in following simulations.

[0083] Scenario 1: The 3 clusters are unoverlapping, each cluster contains $n/3$ observations.

[0084] Scenario 2: 70% of the observations involve in single clusters, 22% of the observations involve in two clusters and 8% of the observations involve in three clusters.

Cluster Recovery Quality Measures

[0085] Quality measures in Baeza-Yates, Ribeiro-Neto et al. (1999) are used to evaluate the clustering performance of the proposed method. Suppose we want to compare a target cluster A and a retrieved cluster $\hat{A}$. N_S denotes the number of observations in set S . The quality measures are defined as

$$\begin{aligned} \text{"specificity"} &= \frac{N_{A \cap \hat{A}}}{N_{\hat{A}}}, \\ \text{"sensitivity"} &= \frac{N_{A \cap \hat{A}}}{N_A}, \\ \text{"F}_1 \text{ measure"} &= \frac{2N_{A \cap \hat{A}}}{N_A + N_{\hat{A}}}. \end{aligned}$$

[0086] The F_1 measure, taken as the harmonic mean of specificity and sensitivity, gives an overall measure of the clustering performance.

[0087] In order to compare a sequence of target clusters, we use a one-to-one correspondence match approach (Turner, Bailey and Krzanowski 2005). We first make the number of retrieved clusters to be the same as the number of target clusters by adding null clusters to retrieved clusters or dropping additional poorly retrieved clusters. Retrieved clusters are then matched to target clusters via a one-to-one correspondence. Finally calculate the mean pair-wise quality measures. The optimal one-to-one match is the one producing the highest mean pair-wise F_1 measure.

Cluster Model Comparisons

[0088] A sequence of methods have been implemented in simulations including the plaid model (*plaid*) and its revised counterparts (*aplaid*); the FMR model with 'flexmix' R package (*EM*) in Leisch (2004) and Grun and Leisch (2008) where the multivariate responses are treated as independent and overlapping problems are not considered; our proposed generalized FMMR model with the multivariate responses treated as independent (*gEMseplasso0*); our proposed generalized FMMR model with separate lasso estimation for B_k in (3.12) (*gEMseplasso*); and finally our proposed original method (*gEMmrce*).

[0089] The clustering performances were evaluated via quality measures as well as the sum of squared estimation errors (SSE) for model coefficients. Results for the example simulations are summarized in Figs. 2A-2D and Figs. 3A-3D. In scenario 1 when there is no overlappings, *EM* and *gEMseplasso0* do equally well showing that the generalized FMMR model retains the ability in recovering unoverlapping clusters. By taking into account the covariance structure among multivariate responses in generalized FMMR model, *gEMseplasso* and *gEMmrce* produce a higher degree of clustering accuracy than *EM* and *gEMseplasso0*. There is little difference between *gEMseplasso* and *gEMmrce* however. When there are 30% overlappings in scenario 2, specificity of the *EM* drops down dramatically resulting in a poor F_1 measure. The *gEMseplasso0* does much better than the *EM* by estimating overlapping clusters, but on the other hand it is worse than the *gEMseplasso* and *gEMmrce*. There is again little difference between *gEMseplasso* and

gEMmrce. Moreover, the *gEMseplasso* and *gEMmrce* converge fast with n , achieving a median 95% accuracy in terms of F_1 measure when $n = 450$. The *plaid* and *aplaid* however have a poor performance in all scenarios.

THERAPEUTIC BIOMARKER SCREENING FOR THE SANGER DATASET

[0090] We applied the innovative techniques herein to analyzing the SANGER high throughput drug sensitivity dataset for cancer. An updated version of this data set (<http://www.cancerrxgene.org/downloads/>) contains 140 drugs as response variables, 13831 genomic features as covariates (including tissue type, rearrangements, mutation status of 71 cancer genes, continuous copy number data of 426 genes causally implicated in cancer, as well as genome-wide transcriptional profiles), and 707 human tumor cell lines as samples. The responses, made up with IC_{50} values from pairwise drug-cell-line screening, has some missing values. Therefore the data was first filtered by removing cell lines for which less than 50% of the drugs were tested, resulting in 591 cell lines remaining. In the remaining data, about 37.4% of the cell lines are with missing values. These missing values are imputed via the random forest imputation algorithm (Ishwaran, Kogalur, Blackstone and Lauer 2008; Stekhoven and Buhlmann 2012).

[0091] In order to identify patterns of cancer-specific therapeutic biomarkers, we employed method *gEMseplasso* in the analysis. Note that by “cancer-specific” we do not assume separate patterns for each cancer type but rather clusters that may be *driven* by one or a small number of cancer types. Throughout the analysis, we fixed K to a value of 3, although as previously shown, a BIC model selection approach could also be used. However, fixing the value of K does not limit the interesting findings that we can still find and saves on computational time.

[0092] Due to the scale and complexity of the data, the analysis, in this example, was conducted in three steps. Although simulation studies show that modeling with multivariate responses yields much higher clustering accuracy than modeling with single response, it is unreasonable to simply fit all 140 drugs (responses) in one generalized FMMR model. Because by doing so, one assumes that the cell line assignments to components are the same for all 140 drugs, which can hardly be true. Thus in our analysis, we first divide the 140

drugs into groups and then fit each drug group with a generalized FMMR model, strategies of which lead to following three steps in detail.

[0093] In the first step, for example as implemented in a drug analysis computing system, the processing takes received drug data and fits each drug c by a generalized FMMR model. As a result, we obtain drug-specific cell line assignments, $\hat{Z}_c$, as well as component-wise coefficient estimates, $\hat{B}_k^c$ for $k = 1, \dots, 3$.

[0094] In the second step, the process uses an affinity-propagation clustering (APC) algorithm (Frey and Dueck 2007) to group the 140 drugs based on results from the first step. The grouping is conducted in a two-level nested manner. In the first level, the APC algorithm is applied to all 140 drugs. The pair-wise similarity matrix required by the algorithm as input data is calculated from the Euclidean distance between $\hat{Z}_a$ and $\hat{Z}_b$. In the second level, the APC algorithm is utilized again within every first-level drug group. The pair-wise similarity matrix is calculated from the Euclidean distance between $\hat{B}_k^a$ and $\hat{B}_k^b$ for $k = 1, \dots, 3$. Consequently, drugs having similar cell line assignments and component-wise coefficient estimates are grouped together. Resulting drug groups are shown in Figs. 4A-4C (which are quadrants of an overall drug grouping (or cluster) display, where the coloring and thickness of connecting lines indicate degree of closeness).

[0095] Finally, in the third step of the process, each second-level drug group from step two is fitted by a generalized FMMR model. We get specific cell line assignments, $\hat{Z}_C$, and component-wise coefficient estimates $\hat{B}_k^C$, $k = 1, \dots, 3$, for every second-level drug group C .

[0096] As an illustration, the membership and coefficient estimates of cell line component 1 and 12 for every second-level drug group are shown in Figs. 5A and 5B (parts 1-3, where part 3 is the legend showing drugs and drug groupings by color corresponding to the corresponding biomarker responsiveness data for each drug on the w-axis of parts 1 and 2) and Figs. 6A and 6B (with the same sub-parts indications), respectively. The memberships are depicted via frequencies of cancer types that cell lines belong to. These are represented by the size of the colored bubbles in the top panels. Estimates of other clusters are not presented in the paper due to limited space. It's very clear that clusters being driven by different cancer types have very different therapeutic genomic profiles and yet no clusters are

homogeneous in a particular cancer type. This speaks to the great heterogeneity seen in cancer overall where in fact, cancer is not thought of as a single disease but rather many diseases characterized by different underlying biological changes.

[0097] Next, we show how the present techniques can be used to identify predictive strategies for circumventing drug resistance based on mutation data.

[0098] Figs. 7A-7B and Figs. 8A-8B are similar to Figs. 5A, 5B, 6A, and 6B, except that they only show a subset of $\hat{B}_1^C$ and $\hat{B}_{12}^C$ with regard to the mutations of 71 cancer genes, where $\hat{B}_{12}^C = \hat{B}_1^C + \hat{B}_2^C$ are coefficient estimates of component 12 for drug group C. The present techniques can be used for drug repurposing (Martins, Zhou, Corella, Horiuchi, Yau, Rakshandehroo, Gordan, Levin, Johnson, Jascur et al. 2015) for resistant tumors. Figs. 7A-7B and Figs. 8A-8B are regression based drug-genetic association maps (bottom panel) with red indicating drug-sensitivity biomarkers and blue indicating drug-resistance biomarkers, conditioning on a component of cell lines (upper panel) identified through the generalized FMMR model. We focus the process on soft-tissue cancers as an example of this drug repurposing. The process extracted those cell line components which contain a high percentage of soft-tissue cancer (large purple bubbles), along with the subset component-wise coefficient estimates. The extracted information is presented in Figs. 9A and 9B. Fig. 9A showing the IC_{50} values of soft-tissue cell lines in extracted cell line components. Extracted drugs in Figs. 9A and 9B are further filtered, by the processes herein, by only keeping those that have a low IC_{50} value without resistance biomarkers and that have a high IC_{50} value with resistance biomarkers, leading to Figs. 10A and 10B. As shown in these figures, the process herein is able to determine that the former set of drugs can be used as alternatives for the later set of drugs, which progress resistance to soft tissue cancer. Table 1, Fig. 16, provides a summary of all pairwise drug combinations found in the dataset that are predicted in this way for soft tissue cancers. As a positive control (represented by a * in Table 1), of this strategy, we find that drug Gemcitabine was found to be effective in soft tissue cancer resistant to standard chemotherapy (doxyrubicin) in a small sample size phase II study (Merimsky, Meller, Flusser, Kollender, Issakov, Weil-Ben-Arush, Fenig, Neuman, Sapir, Ariad et al. 2000).

[0099] Overlapping cell lines can be used to guide drug combinations and identify potential drug synergies. Based on estimates of cell line assignments $\hat{Z}_C$, the process can cluster the IC_{50} values with respect to cell line components for each drug.

[0100] Fig. 11A-11E are boxplots of the resulting component-wise IC_{50} values for 15 out of 140 drugs. Due to limited space, the rest are not shown. As indicated in the illustrated example, the present techniques are able to automatically identify drugs for which the magnitude of IC_{50} s for overlapping component, say 12, is between those for component 1 and 2. Drugs MS.275 and GW843682X show such interesting pattern. Next, the process can determine the component with highest IC_{50} s and try to decrease its IC_{50} s towards overlapping component. Take GW843682X for an example, component 2 has the highest IC_{50} s among components 1, 2, and 12. The process can do initial predictions that the IC_{50} s for cell lines in component 2 using $\hat{B}_1^{gw843682x}$, $\hat{B}_2^{gw843682x}$ and $\hat{B}_{12}^{gw843682x}$ respectively. Predicted results as well as the observations are shown in top panel (middle) of Fig. 12A. By doing so, the process can determine if coefficients $\hat{B}_1^{gw843682x}$ and $\hat{B}_{12}^{gw843682x}$ applied to cell lines in component 2 can generate lower IC_{50} s similar to that in component 1 and 12. This type of determination can be performed by the processes herein without actually physically testing the drug data on a cell line, but rather the determination is made from drug data, biomarker response, cell line data, etc. Since $\hat{B}_1^{gw843682x}$ corresponds to the coefficient estimates of drug GW843682X regressing on its cell line component 1, we want to find another drug which produces similar coefficient estimates as $\hat{B}_1^{gw843682x}$. The process automatically identifies coefficient estimates $\hat{B}_3^{epothilone.b}$ are found highly correlated with $\hat{B}_1^{gw843682x}$. And the predicted IC_{50} s for cell lines in component 2 using $\hat{B}_3^{epothilone.b}$, $\hat{B}_2^{gw843682x}$ and $\hat{B}_3^{epothilone.b} + \hat{B}_2^{gw843682x}$ are presented in bottom panel (middle) of Fig. 12B. It is shown that the drug combination of GW843682X and Epothilone.B decreases the IC_{50} s dramatically.

[0101] Similar studies were conducted for drug MS.275 with respect to component 1, 2, and 12 and for drug GW843682X with respect to component 2, 3 and 23. Results are shown in left and right columns of Figs. 12A and 12B, respectively. The combination of GW843682X and NVP.BEZ235 is effective in decreasing the IC_{50} s, while the combination of

MS.275 and Bicalutamide does not show the effectiveness. As a positive control, Wildey, Chen, Lent, Stetson, Pink, Barnholtz-Sloan and Dowlati (2014) pointed out that a group of drugs (including GW-843682X and Epothilone B) may provide a practical starting point to investigate combinatorial drug therapies for synergistic effect in small-cell lung cancer (SCLC). Among the cluster of drugs, they found synergism between GW-843682X and CGP-60474 via a preliminary study.

[0102] In Figs. 12A and 12B, the indications Prc1 and Prc1h indicate a cell line of a first genomic profile and Prc2 and represent a second genomic profile for that cell line. For example, the plot for Prc1(h) may represent IC_{50} drug responsiveness data for the plotted drug for a soft tissue cancer cell line having a first genomic profile. Prc2 may then represent a mutation of that soft tissue cancer cell line. The present techniques, as described, can then analyze drugs by their drug grouping to identify drugs producing an IC_{50} value in a desired range, between the max and mix IC_{50} values of the two cell lines, and specifically a drug that corresponds to an overlapping cluster drug cell line (e.g., Prc4h) that includes at least some of the genomic profile of the first cell line (Prc1(h)) and at least some of the genomic profile of the second cell line (Prc2). As shown, for example, because GW843682X+Epothilone.B produces an IC_{50} responsive to an overlapping cluster cell line Prc4H, that drug GW843682X+Epothilone.B is automatically identified by the processes as a potential for treating the cancer of a patient having either the Prc1 or Prc2 cancer cell lines.

[0103] Any number of cell lines can be examined against drugs and drug combinations, with the present techniques. Examples include cells that are cancer cells, including but not limited to thyroid cancer cells, soft tissue cancer cells, skin cancer cells, pancreatic cancer cells, nervous system cancer cells, lung cancer cells, kidney cancer cells, digestive system cancer cells, breast cancer cells, bone cancer cells, blood cancer cells, digestive tract cancer cells, and urogenital cancers cells.

[0104] In some of the examples, at least some examined cell lines correspond to a primary cancer, and wherein an overlapping cell line cluster corresponds to the primary cancer. In yet other examples, at least some of cell lines correspond to a primary cancer, and an overlapping cell line cluster includes a secondary cancer different than the primary cancer.

[0105] The present techniques can automatically identify, from a database of drugs, plurality of alternative drugs exhibiting a drug responsiveness value between a high drug responsiveness value and a low drug responsiveness value, such that the alternative drugs have the highest drug responsiveness for the overlapping cell line cluster. This can be used to maximize the likelihood of success of the alternative drugs. In automated drug delivery systems, the alternative drug may then be administered to the overlapping cell line cluster for testing.

[0106] We note that an alternative drug may comprise at least one drug combination comprising a treatment drug and an additive drug. In other examples, the alternative drug comprises at least one drug combination that does not contain the treatment drug.

[0107] Figs. 13 and 14 illustrate example processes for identifying alternative treatment drugs from drug cluster groupings and processes for generating biomarker mappings of drug cluster groupings for different cell lines, respectively. For process 1300 (Fig. 13), at a block 1302, drug responsiveness data is received at a drug analysis computing system. The drug responsiveness data may corresponding to drug treatment data for a drug applied to a one or more different cell lines. Such drug treatment data may include, by way of example, drug test data on drugs applied to patient cell lines. In response to the treatment data, the drug analysis computing system (at a block 1304) identifies two or more cell lines from the plurality of different cell lines. For example, the drug analysis computing system may be configured to use any number of criteria to identify initial cell lines. Cell lines could be selected randomly in some examples. In some examples patient data may be used to determine initial cell lines. In some examples, cell lines are chosen to introduce a sufficient amount of diversity in patients represented for drug analysis. At a block 1306, the drug analysis computing system implements one or more of the clustering grouping analysis techniques herein to determine an overlapping cell line cluster, from the among the identified cell lines. For example, the drug analysis computing system may apply a model-based overlapping clustering framework, as described herein, e.g., a multivariate regression model analysis, such as a finite mixture of multivariate regression model and EM algorithm.

[0108] At a block 1308, an optimality algorithm process is performed on data for a cluster of drugs 1308, as the implementation of Algorithm 1. From the optimization, one or more

alternative drugs are identified (block 1310), drugs exhibiting a desired drug responsiveness, measured as a IC_{50} value. With the drugs being identified, the one or more alternative drugs may then be applied to cell lines for treatment (1312) and for collecting new drug responsiveness data.

[0109] Fig. 14 illustrates a process 1400, which may be an example implementation of the process 1300, in accordance with examples herein. Initially, drug responsiveness data is received to a drug analysis computing system at a block 1402. At a block 1404, the drug analysis computing system fits each drug by a generalized finite multivariate mixture regression (FMMR) model and identifies individual cell line assignments to clusters and cluster-wise coefficient estimates. An affinity propagation clustering (APC) algorithm is applied to the drug data, based on estimates, at a block 1406. In some examples, the APC may be applied in a nested manner. From there, APC generated drug groupings are generated and stored (block 1408). These groupings may group drugs based on correlations of responsiveness, commonality of drug targets or biological pathways, or other factors, for example. A generalized FMMR model is then applied to each drug group to produce a subset of estimation results of drug responsiveness, at a block 1410. At block 1412, the drug analysis computing system generates a biomarker mapping of drug groupings and corresponding cell line clusters for treatment by these drug groupings. The mappings may be based on cell lines as described, while in other examples, the mappings may be based on gene expression, as shown in the example generated biomarker mapping in Figs. 7, 8, 9, and 10.

[0110] Fig. 15 illustrates an example drug analysis system 100 that illustrates various components used in implementing an example embodiment of the present techniques. A signal-processing device 102 (or “signal processor” or “diagnostic device”) is configured to collect cancer cell, gene, genomic, demographic, drug, and other data from example databases 116 and 115 connected to the single processing device through a wireless (or wired) network 117. The signal-processing device 102 may have a generalized FMMR model generator 104 operatively connected to an internal database 114 (that may alternatively drug protein and/or cancer cell and other data) via a link 122 connected to an input/output (I/O) circuit 112. It should be noted that, while not shown, additional databases may be linked to the generalized FMMR model generator 104 in a known manner. The generator processor

104 includes a program memory 106, one or more processors 108 (may be called microcontrollers or a microprocessors), a random-access memory (RAM) 110, and the input/output (I/O) circuit 112, all of which are interconnected via an address/data bus 120. It should be appreciated that although only one processor 108 is shown, the generator 104 may include multiple microprocessors 108. Similarly, the memory of the generator 104 may include multiple RAMs 110 and multiple program memories 106. Although the I/O circuit 112 is shown as a single block, it should be appreciated that the I/O circuit 112 may include a number of different types of I/O circuits. The RAM(s) 110 and the program memories 106 may be implemented as semiconductor memories, magnetically readable memories, and/or optically readable memories, for example. A link 124, which may include one or more wired and/or wireless (Bluetooth, WLAN, etc.) connections, may operatively connect the matrix generator 104 to the databases 115 and 116 through the I/O circuit 112. In other examples, the databases 115 and 116 may be part of the signal-processing device 102.

[0111] The program memory 106 and/or the RAM 110 may store various applications (i.e., machine readable instructions) for execution by the processor 108. For example, an operating system 130 may generally control the operation of the signal-processing device 102 and provide a user interface to the signal-processing device 102 to implement data processing operations. The program memory 106 and/or the RAM 110 may also store a variety of subroutines 132 for accessing specific functions of the signal-processing device 102. By way of example, and without limitation, the subroutines 132 may include, among other things: a subroutine to perform the process steps of Figs. 13 and/or 14 or any of the other processes described herein.

[0112] The subroutines 132 may also include other subroutines, for example, implementing software keyboard functionality, interfacing with other hardware in the signal processing device 102, etc. The program memory 106 and/or the RAM 110 may further store data related to the configuration and/or operation of the signal-processing device 102, and/or related to the operation of the one or more subroutines 132. For example, the data may be data gathered from the databases 115 and 116, data determined and/or calculated by the processor 108, etc. In addition to the matrix generator 104, the signal-processing device 102 may include other hardware resources. The signal-processing device 102 may also include

various types of input/output hardware such as a visual display 126 and input device(s) 128 (e.g., keypad, keyboard, etc.). In an embodiment, the display 126 is touch-sensitive, and may cooperate with a software keyboard routine as one of the software routines 132 to accept user input.

[0113] It may be advantageous for the signal-processing device 102 to communicate with a medical treatment device, medical drug testing databases, biomarker/genomic testing databases, medical data records storage device, through the network 117 or through any of a number of known networking devices and techniques (e.g., through a commuter network such as a hospital or clinic intranet, the Internet, etc.). For example, the signal-processing device may be connected to a medical records database, hospital management processing system, healthcare professional terminals (e.g., doctor stations, nurse stations), high throughput screening framework, drug/biomarker databases, or other system.

[0114] The system 100 may be implemented as computer-readable instructions stored on a single dedicated machine, for example, one with one or more computer processing units. In some examples, the dedicated machine performs only the functions described in the processes of Figs. 1-14, and any other functions needed to perform those processes. The dedicated machine may be a standalone machine or embedded within another computing machine.

[0115] In some examples, one or more of the functions of the system 100 may be performed remotely, including, for example, on a server connected to a remote computing device, through a wired or wireless interface at 112 and the network 117. Such distributed processing may include having all or a portion of the processing of system 100 performed on a remote server. In some embodiments, the techniques herein may be implemented as software-as-a-service (SaaS) with the computer-readable instructions to perform the method steps being stored on one or more of the computer processing devices and communicating with one or more user devices, including but not limited to personal computers, handheld devices, etc.

[0116] Provided are computer-implemented processes built upon new models for identifying therapeutic biomarkers for cancer and other pathologies, where these techniques

can answer specific questions regarding sensitivity, resistance and synergy across large sets of possible treatment drugs. In an example, a penalized likelihood approach for an FMMR model was used and which enforced sparsity in the genomic features. To enable overlapping clustering, the FMMR model was then generalized and a new EM algorithm derived for estimation of model parameters implemented. While the noteworthy plaid model can also be generalized for overlapping clustering multivariate regression data, the generalized FMMR model markedly outperforms that method.

[0117] Throughout this specification, plural instances may implement components, operations, or structures described as a single instance. Although individual operations of one or more methods are illustrated and described as separate operations, one or more of the individual operations may be performed concurrently, and nothing requires that the operations be performed in the order illustrated. Structures and functionality presented as separate components in example configurations may be implemented as a combined structure or component. Similarly, structures and functionality presented as a single component may be implemented as separate components. These and other variations, modifications, additions, and improvements fall within the scope of the subject matter herein.

[0118] Additionally, certain embodiments are described herein as including logic or a number of routines, subroutines, applications, or instructions. These may constitute either software (e.g., code embodied on a machine-readable medium or in a transmission signal) or hardware. In hardware, the routines, etc., are tangible units capable of performing certain operations and may be configured or arranged in a certain manner. In example embodiments, one or more computer systems (e.g., a standalone, client or server computer system) or one or more hardware modules of a computer system (e.g., a processor or a group of processors) may be configured by software (e.g., an application or application portion) as a hardware module that operates to perform certain operations as described herein.

[0119] In various embodiments, a hardware module may be implemented mechanically or electronically. For example, a hardware module may comprise dedicated circuitry or logic that is permanently configured (e.g., as a special-purpose processor, such as a field programmable gate array (FPGA) or an application-specific integrated circuit (ASIC)) to perform certain operations. A hardware module may also comprise programmable logic or

circuitry (e.g., as encompassed within a general-purpose processor or other programmable processor) that is temporarily configured by software to perform certain operations. It will be appreciated that the decision to implement a hardware module mechanically, in dedicated and permanently configured circuitry, or in temporarily configured circuitry (e.g., configured by software) may be driven by cost and time considerations.

[0120] Accordingly, the term "hardware module" should be understood to encompass a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired), or temporarily configured (e.g., programmed) to operate in a certain manner or to perform certain operations described herein. Considering embodiments in which hardware modules are temporarily configured (e.g., programmed), each of the hardware modules need not be configured or instantiated at any one instance in time. For example, where the hardware modules comprise a general-purpose processor configured using software, the general-purpose processor may be configured as respective different hardware modules at different times. Software may accordingly configure a processor, for example, to constitute a particular hardware module at one instance of time and to constitute a different hardware module at a different instance of time.

[0121] Hardware modules can provide information to, and receive information from, other hardware modules. Accordingly, the described hardware modules may be regarded as being communicatively coupled. Where multiple of such hardware modules exist contemporaneously, communications may be achieved through signal transmission (e.g., over appropriate circuits and buses) that connects the hardware modules. In embodiments in which multiple hardware modules are configured or instantiated at different times, communications between such hardware modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple hardware modules have access. For example, one hardware module may perform an operation and store the output of that operation in a memory device to which it is communicatively coupled. A further hardware module may then, at a later time, access the memory device to retrieve and process the stored output. Hardware modules may also initiate communications with input or output devices, and can operate on a resource (e.g., a collection of information).

[0122] The various operations of the example methods described herein may be performed, at least partially, by one or more processors that are temporarily configured (e.g., by software) or permanently configured to perform the relevant operations. Whether temporarily or permanently configured, such processors may constitute processor-implemented modules that operate to perform one or more operations or functions. The modules referred to herein may, in some example embodiments, comprise processor-implemented modules.

[0123] Similarly, the methods or routines described herein may be at least partially processor-implemented. For example, at least some of the operations of a method may be performed by one or more processors or processor-implemented hardware modules. The performance of certain of the operations may be distributed among the one or more processors, not only residing within a single machine, but also deployed across a number of machines. In some example embodiments, the processor or processors may be located in a single location (e.g., within a home environment, an office environment or as a server farm), while in other embodiments the processors may be distributed across a number of locations.

[0124] The one or more processors may also operate to support performance of the relevant operations in a “cloud computing” environment or as a “software as a service” (SaaS). For example, at least some of the operations may be performed by a group of computers (as examples of machines including processors), these operations being accessible via a network (e.g., the Internet) and via one or more appropriate interfaces (e.g., application program interfaces (APIs).)

[0125] The performance of certain of the operations may be distributed among the one or more processors, not only residing within a single machine, but also deployed across a number of machines. In some example embodiments, the one or more processors or processor-implemented modules may be located in a single geographic location (e.g., within a home environment, an office environment, or a server farm). In other example embodiments, the one or more processors or processor-implemented modules may be distributed across a number of geographic locations.

[0126] Unless specifically stated otherwise, discussions herein using words such as "processing," "computing," "calculating," "determining," "presenting," "displaying," or the like may refer to actions or processes of a machine (e.g., a computer) that manipulates or transforms data represented as physical (e.g., electronic, magnetic, or optical) quantities within one or more memories (e.g., volatile memory, non-volatile memory, or a combination thereof), registers, or other machine components that receive, store, transmit, or display information.

[0127] As used herein any reference to "one embodiment" or "an embodiment" means that a particular element, feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

[0128] Some embodiments may be described using the expression "coupled" and "connected" along with their derivatives. For example, some embodiments may be described using the term "coupled" to indicate that two or more elements are in direct physical or electrical contact. The term "coupled," however, may also mean that two or more elements are not in direct contact with each other, but yet still co-operate or interact with each other. The embodiments are not limited in this context.

[0129] As used herein, the terms "comprises," "comprising," "includes," "including," "has," "having" or any other variation thereof, are intended to cover a non-exclusive inclusion. For example, a process, method, article, or apparatus that comprises a list of elements is not necessarily limited to only those elements but may include other elements not expressly listed or inherent to such process, method, article, or apparatus. Further, unless expressly stated to the contrary, "or" refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).

[0130] In addition, use of the "a" or "an" are employed to describe elements and components of the embodiments herein. This is done merely for convenience and to give a

general sense of the description. This description, and the claims that follow, should be read to include one or at least one and the singular also includes the plural unless it is obvious that it is meant otherwise.

[0131] While the present invention has been described with reference to specific examples, which are intended to be illustrative only and not to be limiting of the invention, it will be apparent to those of ordinary skill in the art that changes, additions and/or deletions may be made to the disclosed embodiments without departing from the spirit and scope of the invention.

[0132] The foregoing description is given for clearness of understanding; and no unnecessary limitations should be understood therefrom, as modifications within the scope of the invention may be apparent to those having ordinary skill in the art.

REFERENCES

- Akaike, H. (1974), "A new look at the statistical model identification," IEEE transactions on automatic control, 19(6), 716-723.
- Anderson, D. R., and Burnham, K. P. (2002), "Avoiding pitfalls when using information-theoretic methods," The Journal of Wildlife Management, pp. 912-918.
- Azzalini, A. (2005), "The skew-normal distribution and related multivariate families," Scandinavian Journal of Statistics, 32(2), 159-188.
- Baeza-Yates, R., Ribeiro-Neto, B. et al. (1999), Modern information retrieval, Vol. 463 ACM press New York.
- Bailey, A. M., Mao, Y., Zeng, J., Holla, V., Johnson, A., Brusco, L., Chen, K., Mendelsohn, J., Routbort, M. J., Mills, G. B. et al. (2014), "Implementation of biomarker-driven cancer therapy: existing tools and remaining gaps," Discovery medicine, 17(92), 101.
- Banerjee, A., Krumpelman, C., Ghosh, J., Basu, S., and Mooney, R. J. (2005), Model-based overlapping clustering., in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, ACM, pp. 532-537.
- Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Leh ar, J., Kryukov, G. V., Sonkin, D. et al. (2012), "The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity," Nature, 483(7391), 603-607.
- Branco, M. D., and Dey, D. K. (2001), "A general class of multivariate skew-elliptical distributions," Journal of Multivariate Analysis, 79(1), 99-113.
- Breiman, L., and Friedman, J. H. (1997), "Predicting multivariate responses in multiple linear regression," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59(1), 3-54.
- Chen, L., Pourahmadi, M., and Maadooliat, M. (2014), "Regularized multivariate regression models with skew-t error distributions," Journal of Statistical Planning and Inference, 149, 125-139.

Claeskens, G., Hjort, N. L. et al. (2008), Model selection and model averaging, Vol. 330 Cambridge University Press Cambridge.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," Journal of the royal statistical society. Series B (methodological), pp. 1-38.

DeSarbo, W. S., and Cron, W. L. (1988), "A maximum likelihood methodology for clusterwise linear regression," Journal of classification, 5(2), 249-282.

Druker, B. J., Guilhot, F., O'Brien, S. G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M. W., Silver, R. T., Goldman, J. M., Stone, R. M. et al. (2006), "Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia," New England Journal of Medicine, 355(23), 2408-2417.

Estivill-Castro, V. (2002), "Why so many clustering algorithms: a position paper," ACM SIGKDD explorations newsletter, 4(1), 65-75.

Fan, J., and Li, R. (2006), "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," arXiv preprint math/0602133.

Fan, J., and Lv, J. (2010), "A selective overview of variable selection in high dimensional feature space," Statistica Sinica, 20(1), 101.

Fraley, C., and Raftery, A. E. (1998), "How many clusters? Which clustering method? Answers via model-based cluster analysis," The computer journal, 41(8), 578-588.

Frey, B. J., and Dueck, D. (2007), "Clustering by passing messages between data points," science, 315(5814), 972-976.

Fu, Q., and Banerjee, A. (2008), Multiplicative mixture models for overlapping clustering, in 2008 Eighth IEEE International Conference on Data Mining, IEEE, pp. 791-796.

Galimberti, G., and Soffritti, G. (2014), "A multivariate linear regression analysis using finite mixtures of t distributions," Computational Statistics & Data Analysis, 71, 138-150.

- Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J. et al. (2012), "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, 483(7391), 570-575.
- Grun, B., and Leisch, F. (2008), "FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters,".
- Heller, K. A., and Ghahramani, Z. (2007), A Nonparametric Bayesian Approach to Modeling Overlapping Clusters, in *AISTATS*, pp. 187-194.
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random survival forests," *The annals of applied statistics*, pp. 841-860.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999), "Data clustering: a review," *ACM computing surveys (CSUR)*, 31(3), 264-323.
- Jedidi, K., Ramaswamy, V., DeSarbo, W. S., and Wedel, M. (1996), "On estimating finite mixtures of multivariate regression and simultaneous equation models," *Structural Equation Modeling: A Multidisciplinary Journal*, 3(3), 266-289.
- Jiang, J., Nguyen, T., and Rao, J. S. (2015), "The E-MS algorithm: model selection with incomplete data," *Journal of the American Statistical Association*, 110(511), 1136-1147.
- Jones, P., and McLachlan, G. (1992), "Fitting finite mixture models in a regression context," *Australian Journal of Statistics*, 34(2), 233-240.
- Keribin, C. (2000), "Consistent estimation of the order of mixture models," *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 49-66.
- Khalili, A., and Chen, J. (2012), "Variable selection in finite mixture of regression models," *Journal of the American Statistical Association*.
- Khalili, A., and Lin, S. (2013), "Regularization in finite mixture of regression models with diverging number of parameters," *Biometrics*, 69(2), 436-446.
- Lazzeroni, L., and Owen, A. (2002), "Plaid models for gene expression data," *Statistica sinica*, pp. 61-86.

Leisch, F. (2004), "Flexmix: A general framework for finite mixture models and latent glass regression in R".

Martins, M. M., Zhou, A. Y., Corella, A., Horiuchi, D., Yau, C., Rakshandehroo, T., Gordan, J. D., Levin, R. S., Johnson, J., Jascur, J. et al. (2015), "Linking Tumor Mutations to Drug Responses via a Quantitative Chemical-Genetic Interaction Map," *Cancer discovery*, 5(2), 154-167.

Merimsky, O., Meller, I., Flusser, G., Kollender, Y., Issakov, J., Weil-Ben-Arush, M., Fenig, E., Neuman, G., Sapir, D., Ariad, S. et al. (2000), "Gemcitabine in soft tissue or bone sarcoma resistant to standard chemotherapy: a phase II study," *Cancer chemotherapy and pharmacology*, 45(2), 177-181.

Needham, R. (1965), "Computer methods for classification and grouping," *The Use of Computers in Anthropology*, I. Hymes, ed, pp. 345-356.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010), "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *The annals of applied statistics*, 4(1), 53.

Quandt, R. E. (1972), "A new approach to estimating switching regressions," *Journal of the American statistical association*, 67(338), 306-310.

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*, 19(4), 947-962.

Schwarz, G. et al. (1978), "Estimating the dimension of a model," *The annals of statistics*, 6(2), 461-464.

Segal, E., Battle, A., and Koller, D. (2002), Decomposing gene expression into cellular processes, in *Proceedings of the Pacific Symposium on Biocomputing*, Vol. 8, pp. 89-100.

Similä, T., and Tikka, J. (2007), "Input selection and shrinkage in multiresponse linear regression," *Computational Statistics & Data Analysis*, 52(1), 406-422.

Soffritti, G., and Galimberti, G. (2011), "Multivariate linear regression with non-normal errors: a solution based on mixture models," *Statistics and Computing*, 21(4), 523-536.

- Stekhoven, D. J., and Bühlmann, P. (2012), “MissForestnon-parametric missing value imputation for mixed-type data,” *Bioinformatics*, 28(1), 112-118.
- Stone, M. (1977), “An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 44-47.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267-288.
- Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), “Simultaneous variable selection,” *Technometrics*, 47(3), 349-363.
- Turner, H., Bailey, T., and Krzanowski, W. (2005), “Improved biclustering of microarray data demonstrated through systematic performance tests,” *Computational statistics & data analysis*, 48(2), 235-254.
- Turner, H. L., Bailey, T. C., Krzanowski, W. J., and Hemingway, C. A. (2005), “Biclustering models for structured microarray data,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2(4), 316-329.
- Wildey, G., Chen, Y., Lent, I., Stetson, L., Pink, J., Barnholtz-Sloan, J. S., and Dowlati, A. (2014), “Pharmacogenomic approach to identify drug sensitivity in small-cell lung cancer,” *PloS one*, 9(9), e106784.
- Zhang, J. (2010), “A Bayesian model for biclustering with applications,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(4), 635-656.
- Zou, H., and Hastie, T. (2005), “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

WHAT IS CLAIMED:

1. A method of identifying alternative drug therapies, the method comprising:

receiving, at a processing unit, drug responsiveness data for a treatment drug applied to a plurality of different cell lines, each cell line differing in genomic profile from one another such that the genomic profiles for the plurality of different cell lines are mutually exclusive, the drug responsiveness data being determined from live cell testing using the treatment drug;

identifying at least two cell lines from the plurality of different cell lines, and determining, at the processing unit, an overlap cell line cluster where one of the at least two cell lines has a high drug responsiveness value and another of the at least two cell lines has a low drug responsiveness value, and wherein the overlap cell line cluster has a drug responsiveness value for the treatment drug that is between the high drug responsiveness value and the low drug responsiveness value, wherein determining the overlap cell line cluster using a generalized finite mixture of multivariate regression (FMMR) model;

performing, at the processing unit, an optimality algorithm process on data for a cluster of drugs to identify one or more alternative drugs exhibiting a drug responsiveness value between the high drug responsiveness value and the low drug responsiveness value; and

administering to cells of one of the at least two cell lines the one or more alternative drugs.
2. The method of claim 1, wherein the cells are cancer cells, including but not limited to thyroid cancer cells, soft tissue cancer cells, skin cancer cells, pancreatic cancer cells, nervous system cancer cells, lung cancer cells, kidney cancer cells, digestive system cancer cells, breast cancer cells, bone cancer cells, blood cancer cells, digestive tract cancer cells, and urogenital cancers cells.
3. The method of claim 1, wherein drug responsiveness value is an IC₅₀ value.
4. The method of claim 3, wherein high drug responsiveness value is a lowest IC₅₀ value for any of the plurality of cell lines.
5. The method of claim 3, wherein low drug responsiveness value is a highest IC₅₀ value for any of the plurality of cell lines.
6. The method of claim 1, wherein the at least two cell lines each correspond to a primary cancer, and wherein the overlap cell line cluster corresponds to the primary cancer.
7. The method of claim 1, wherein the at least two cell lines each correspond to a primary cancer, and wherein the overlap cell line cluster includes a secondary cancer different than the primary cancer.

8. The method of claim 1, wherein if a plurality of alternative drugs exhibiting a drug responsiveness value between the high drug responsiveness value and the low drug responsiveness value are identified, identifying the alternative drug having the highest drug responsiveness for the overlapping cell line cluster; and

administering to cells of one of the at least two cell lines the alternative drug having the highest drug responsiveness for the overlapping cell line cluster.

9. The method of claim 1, wherein the one or more alternative drugs comprises at least one drug combination comprising the treatment drug and an additive drug.

10. The method of claim 1, wherein the one or more alternative drugs comprises at least one drug combination that does not contain the treatment drug.

11. A method of grouping drugs for treating a pathology, the method comprising:
receiving, at a processing unit, drug responsiveness data for a set of drugs and for a set of cell lines;

grouping, using the processing unit, each of the drugs into a different drug grouping, using

(i) a generalized finite mixture of multivariate regression (FMMR) model as follows

$$f(\mathbf{y}_i | \mathbf{x}_i, \Theta) = \sum_{(l_1, \dots, l_s) \in T} \pi_{(l_1, \dots, l_s)} f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \Sigma).$$

wherein $f(\mathbf{y}_i | \mathbf{x}_i, \mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}, \Sigma)$ is a multivariate normal density function generalized for both non-overlapping clusters and overlapping clusters, $\mathbf{y}_i$ are drug responsiveness values, $\mathbf{x}_i$ are genomic biomarkers, $\mathbf{B}_{l_1}, \dots, \mathbf{B}_{l_s}$ are coefficient matrix values for a cluster, $l_1, \dots, l_s$, are clusters, π is an observation value indicating whether a drug belongs in a grouping ($\pi = 1$) or not ($\pi = 0$), and Σ is an error term for the coefficient matrix values, i represents a drug,

wherein the generalized FMMR model allows for modeling drugs for both non-overlapping clusters and overlapping clusters, and

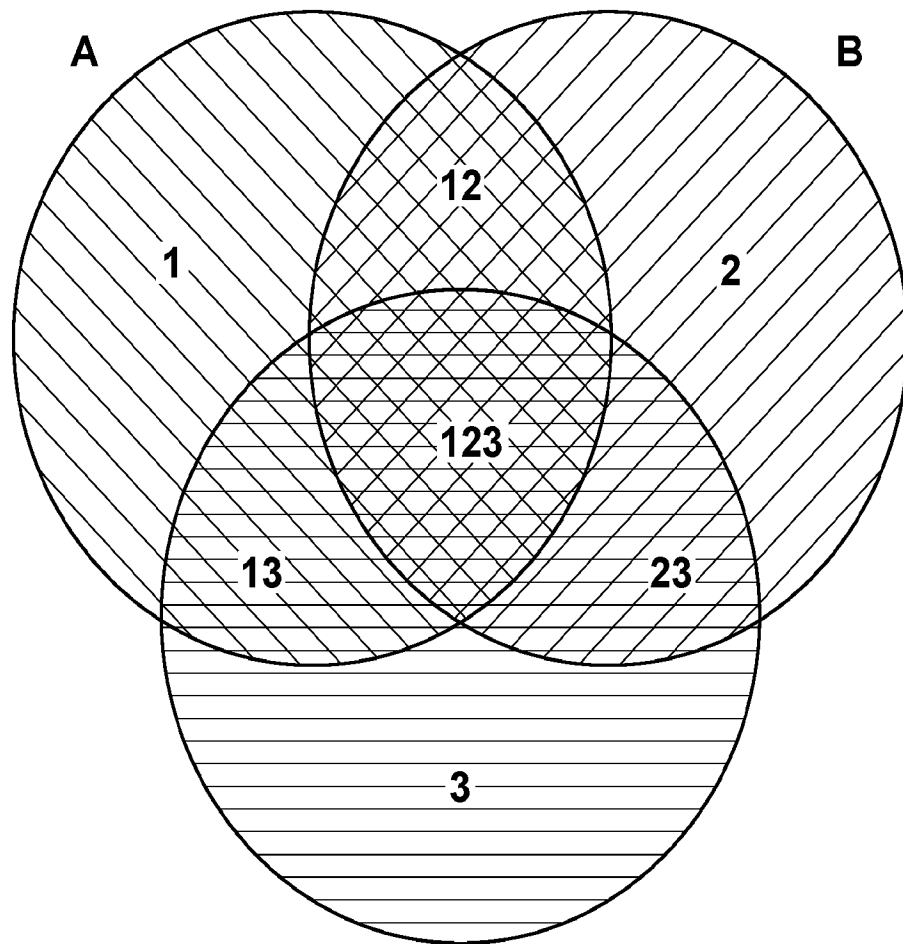
(ii) a clustering algorithm configured to group drugs based on the coefficient matrix values, $\mathbf{B}$, for different drugs and based on cell line assignments; and

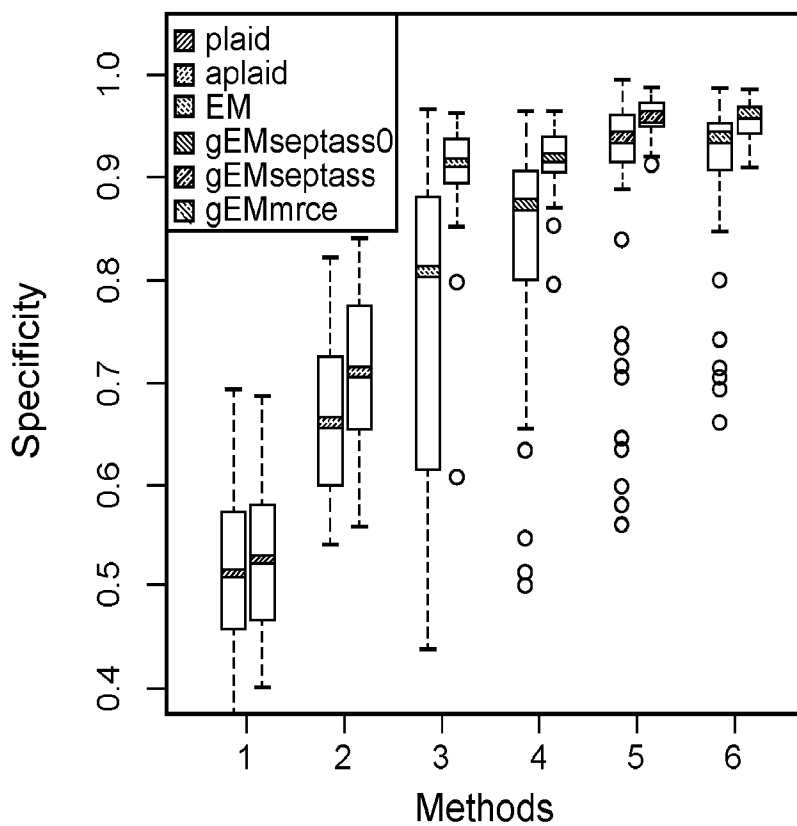
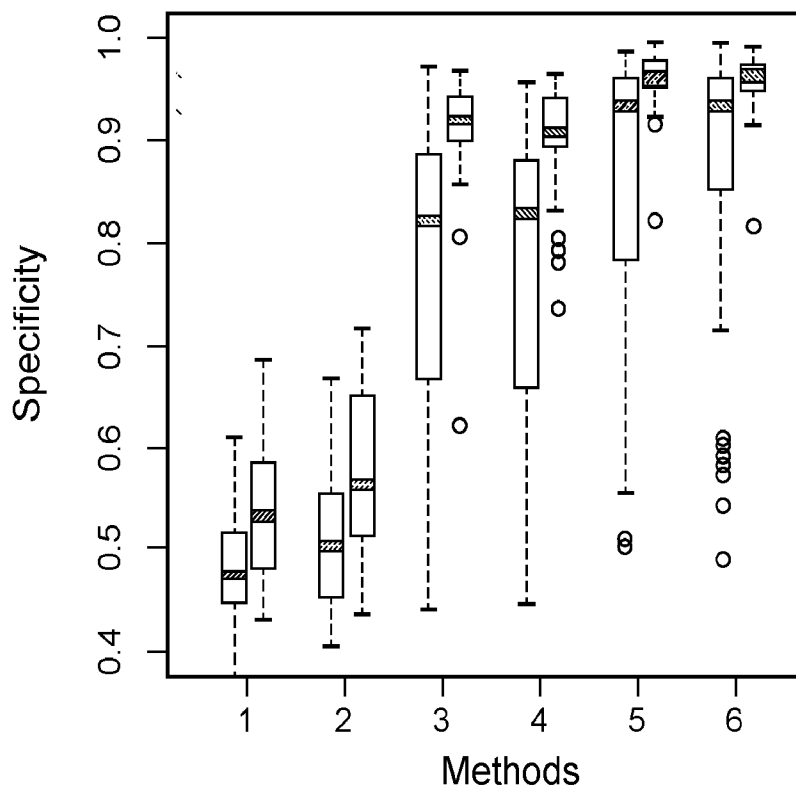
administering one or more drugs from one of the drug groupings to a cell line expressing the pathology.

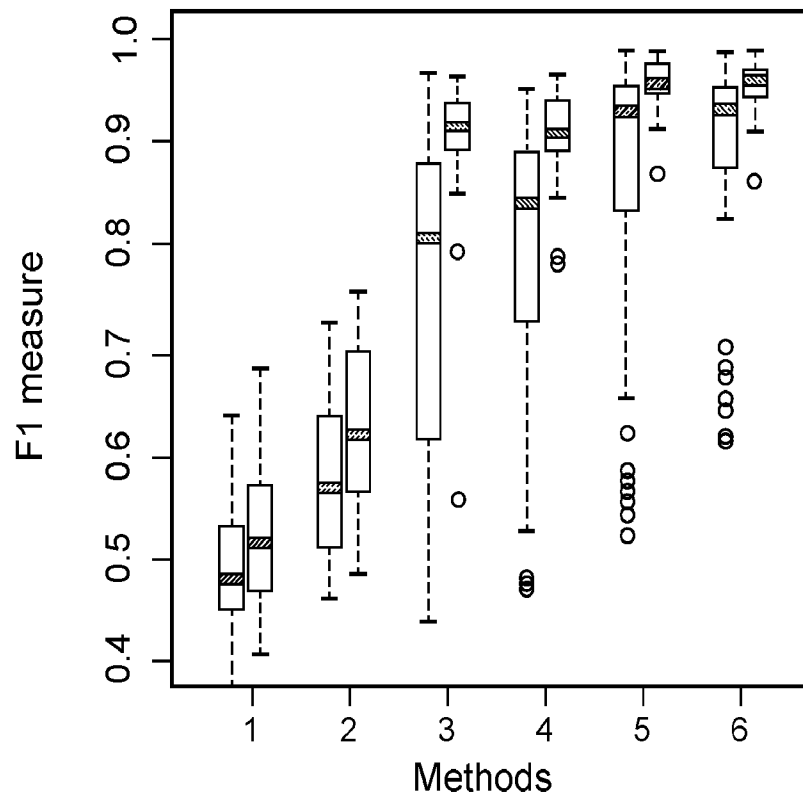
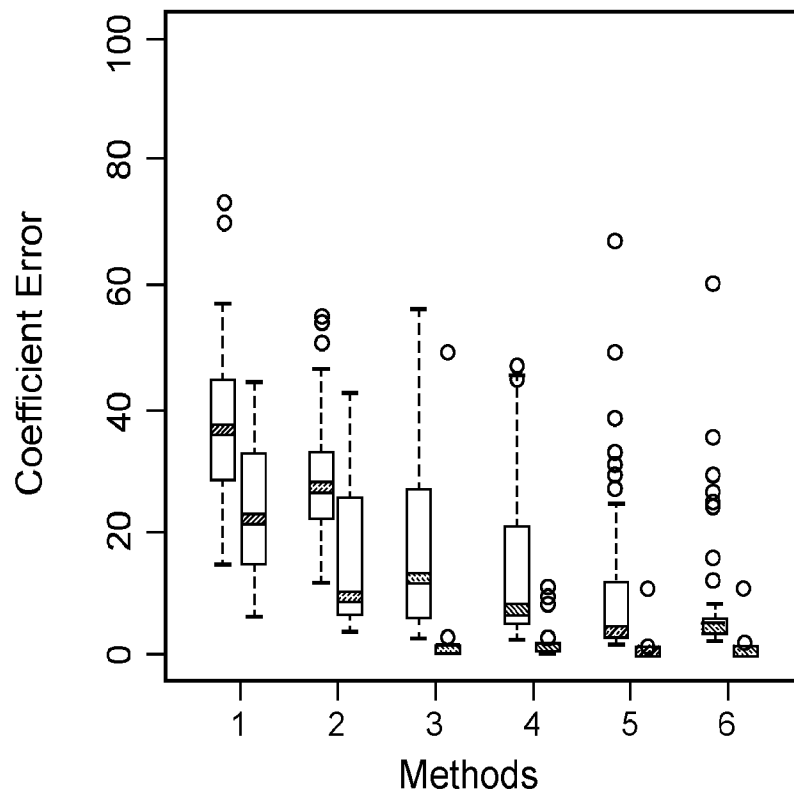
12. The method of claim 11, wherein the clustering algorithm is a multi-level nested clustering algorithm.

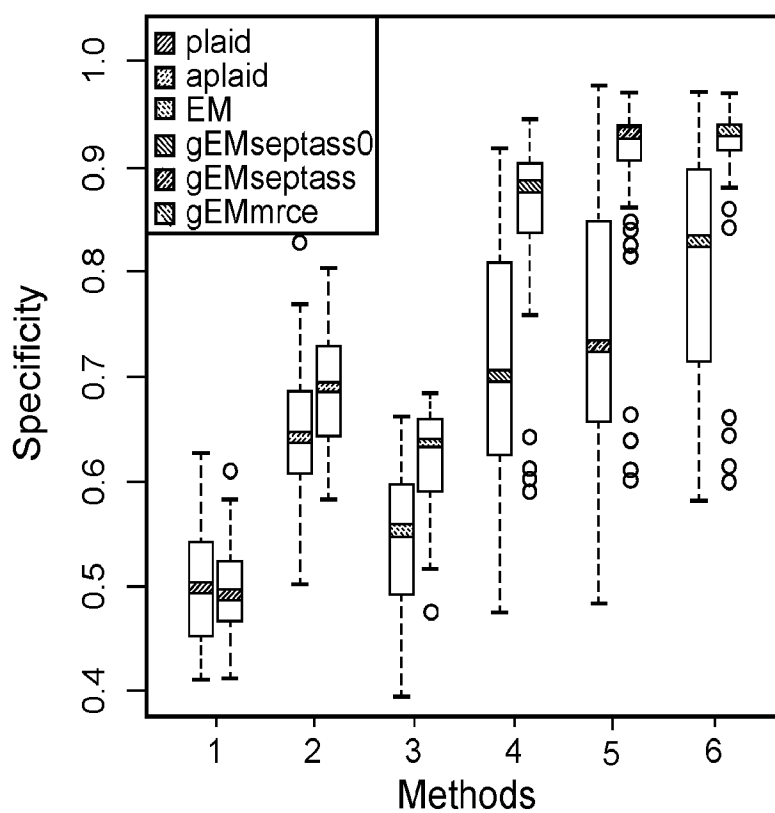
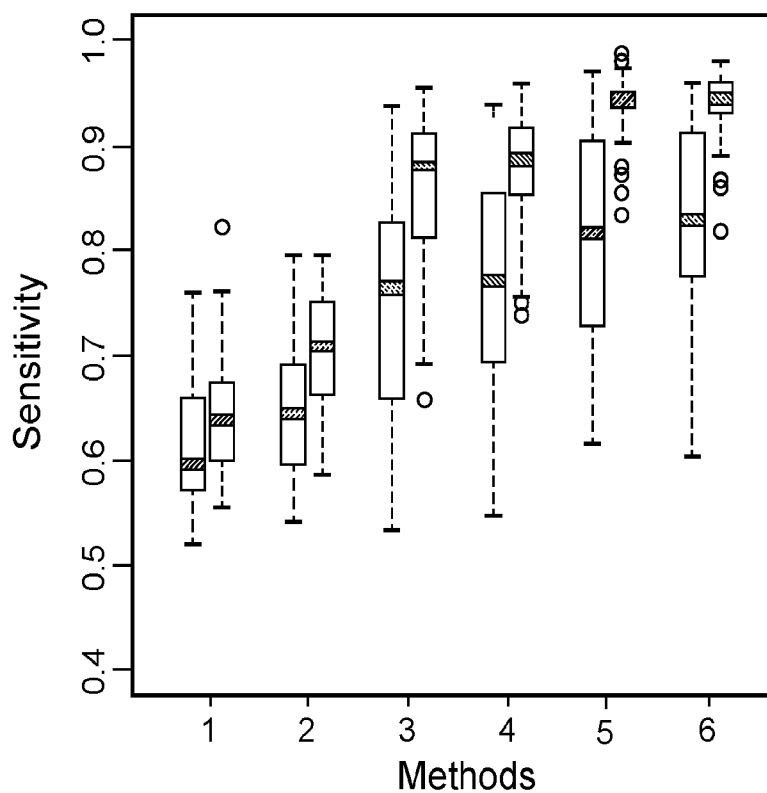
13. The method of claim 12, wherein clustering algorithm is an affinity-propagation clustering algorithm.

14. The method of claim 13, further comprising:
applying the affinity-propagation clustering algorithm at a first level to produce a first level of groupings; and
applying the affinity-propagation clustering algorithm to each of the first level of drug groupings, to produce a subsequent level of drug groupings.

**FIG. 1**

**FIG. 2A****FIG. 2B**

**FIG. 2C****FIG. 2D**

**FIG. 3A****FIG. 3B**

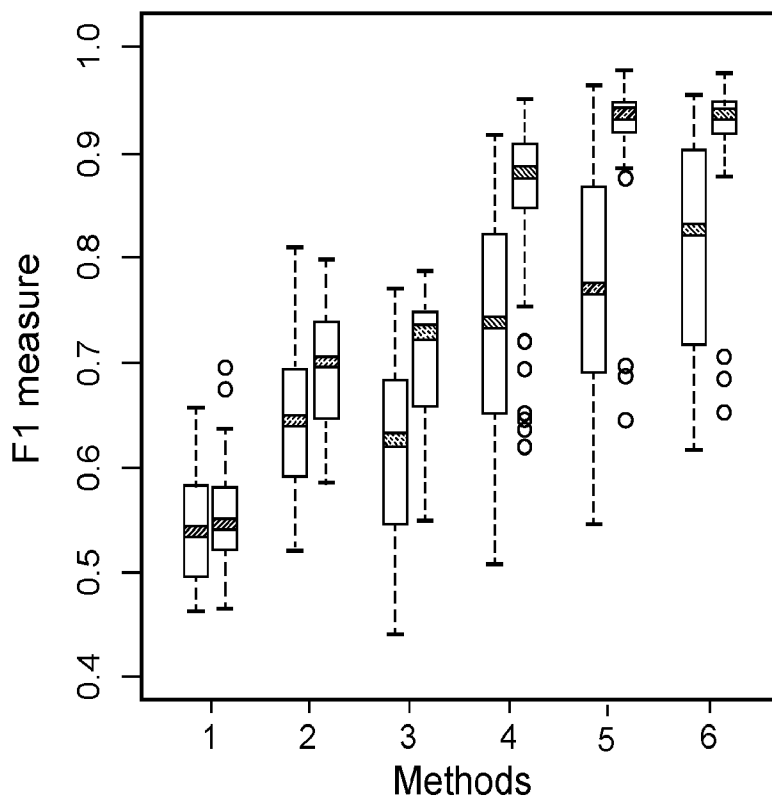


FIG. 3C

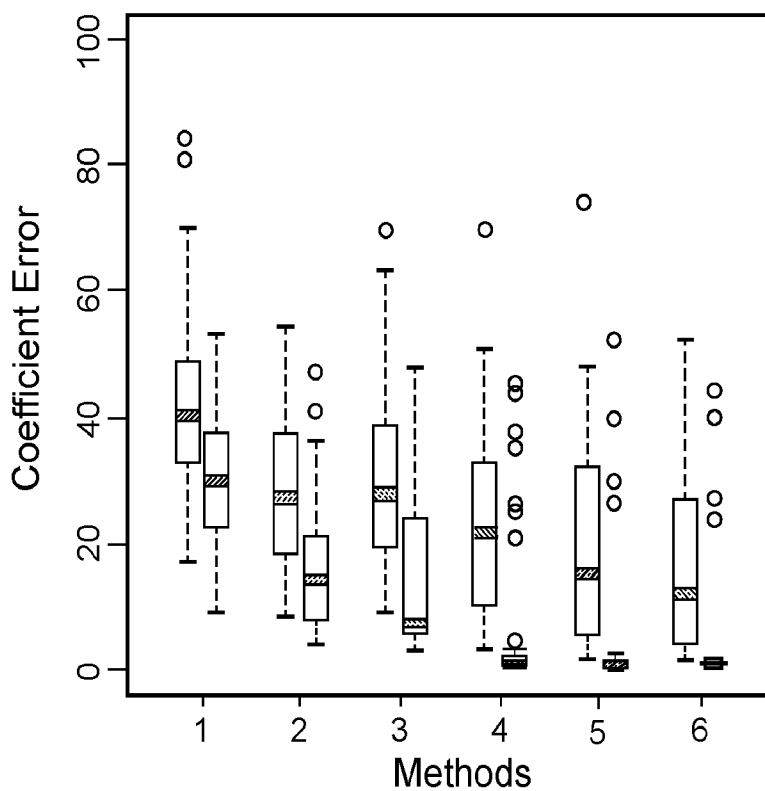


FIG. 3D

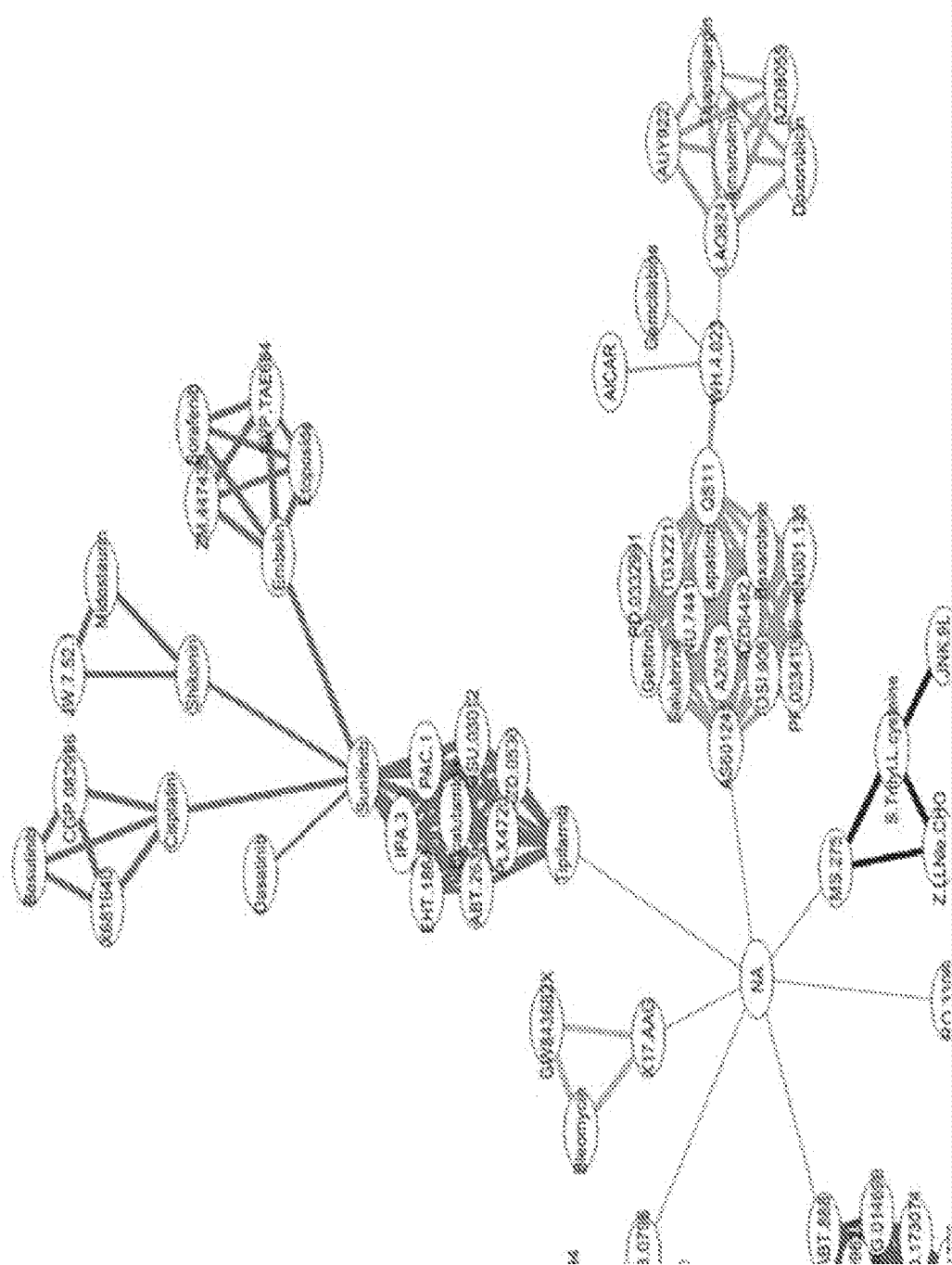


FIG. 4A

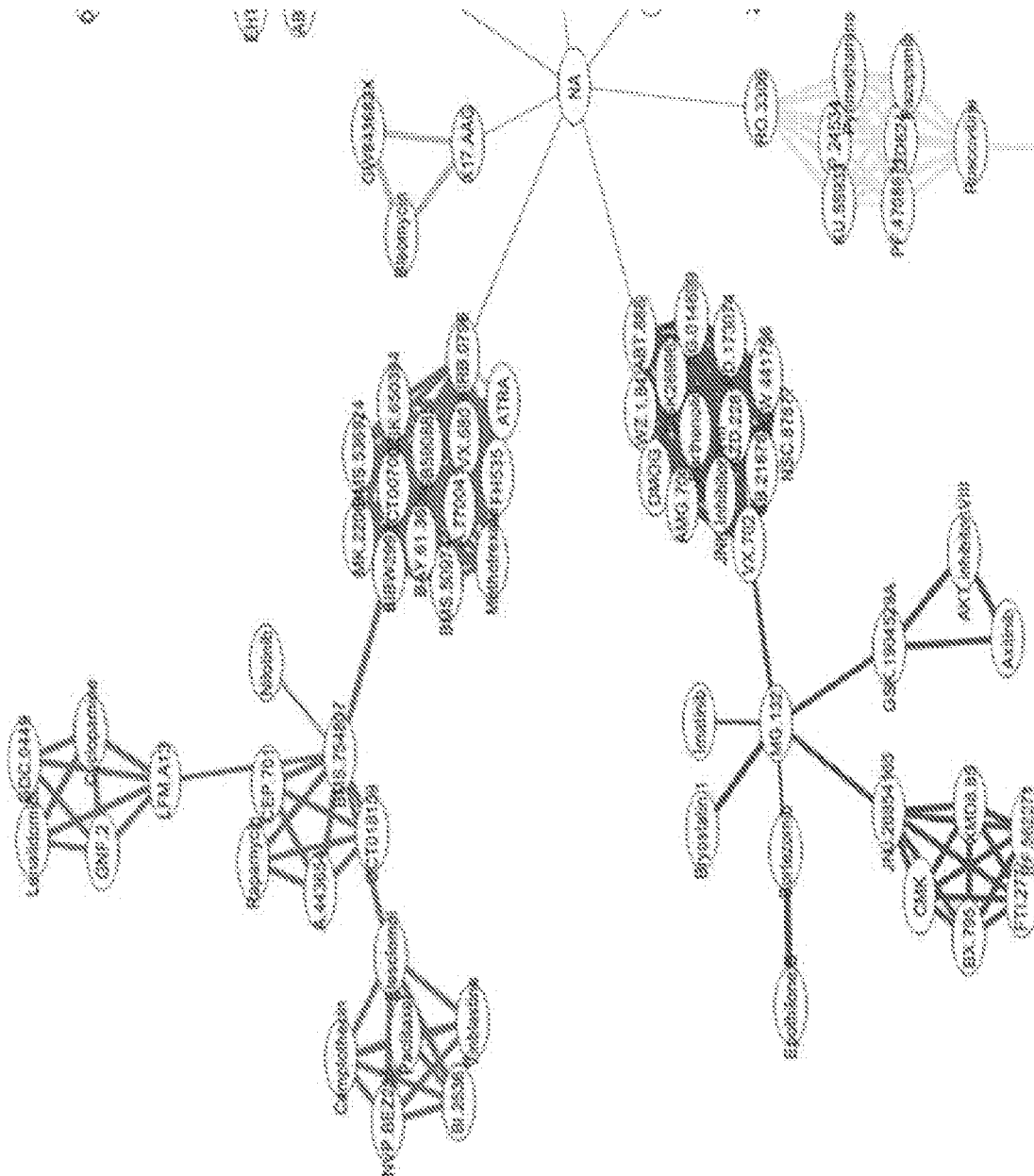


FIG. 4B

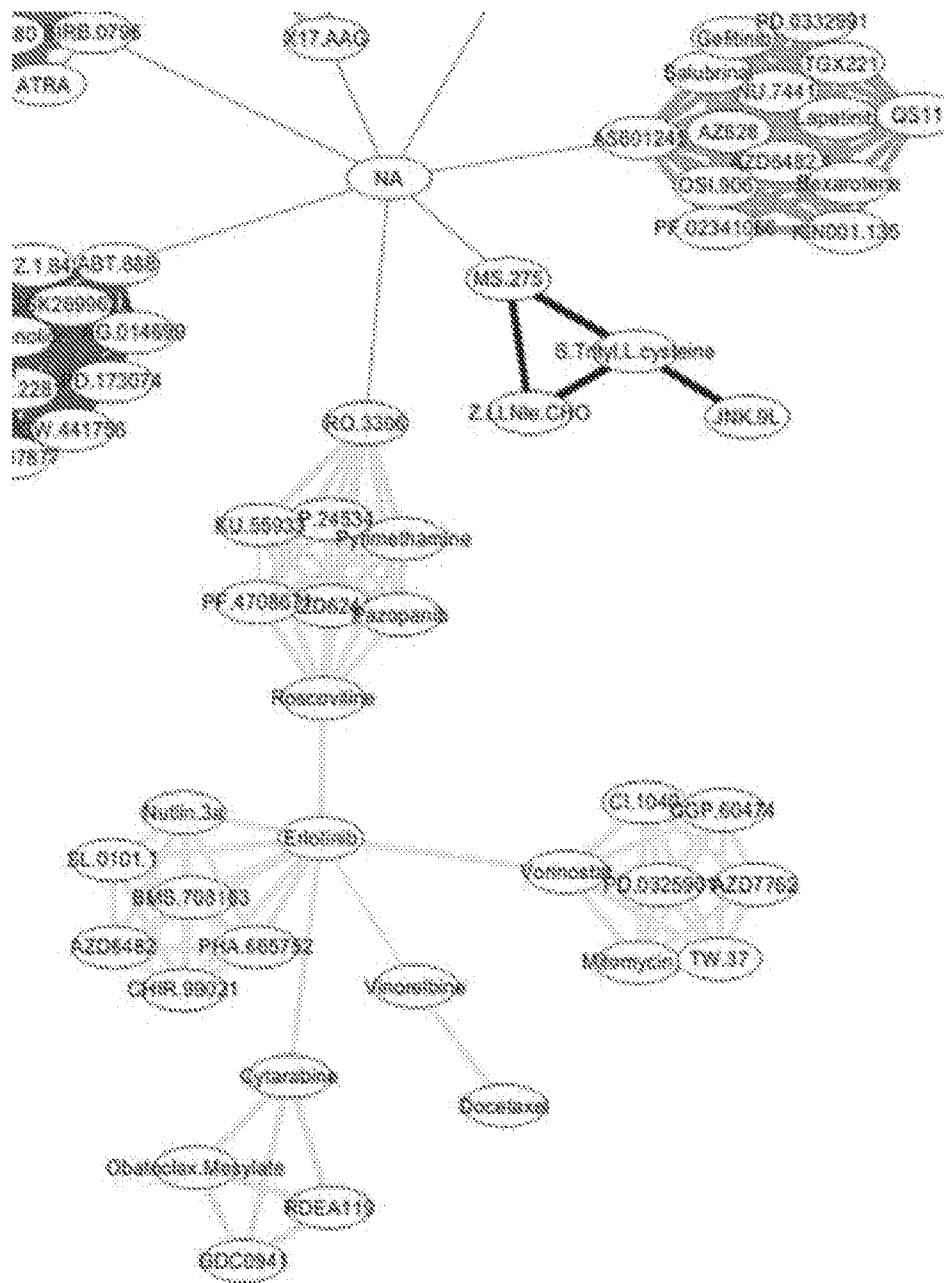


FIG. 4C

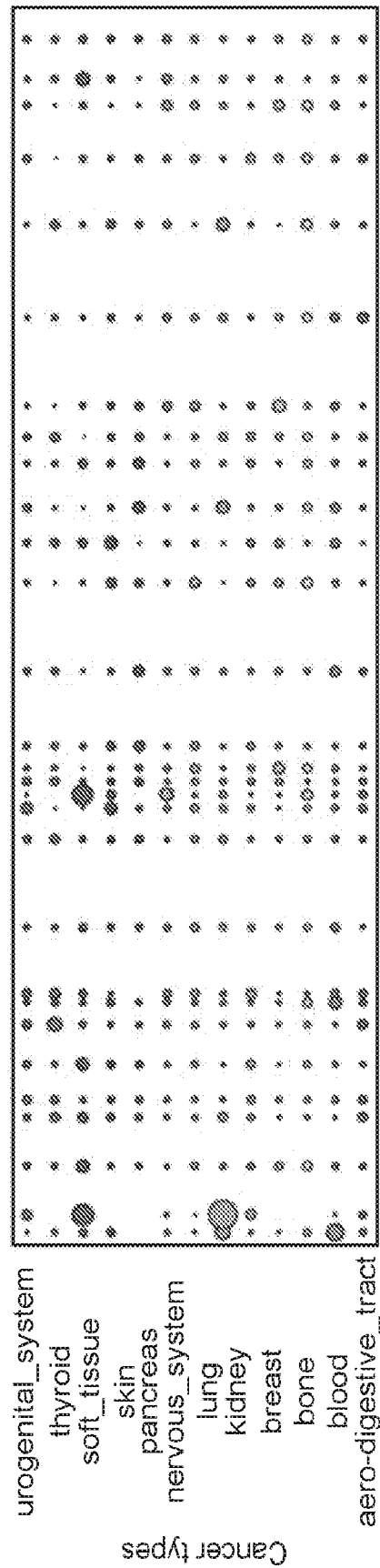


FIG. 5A

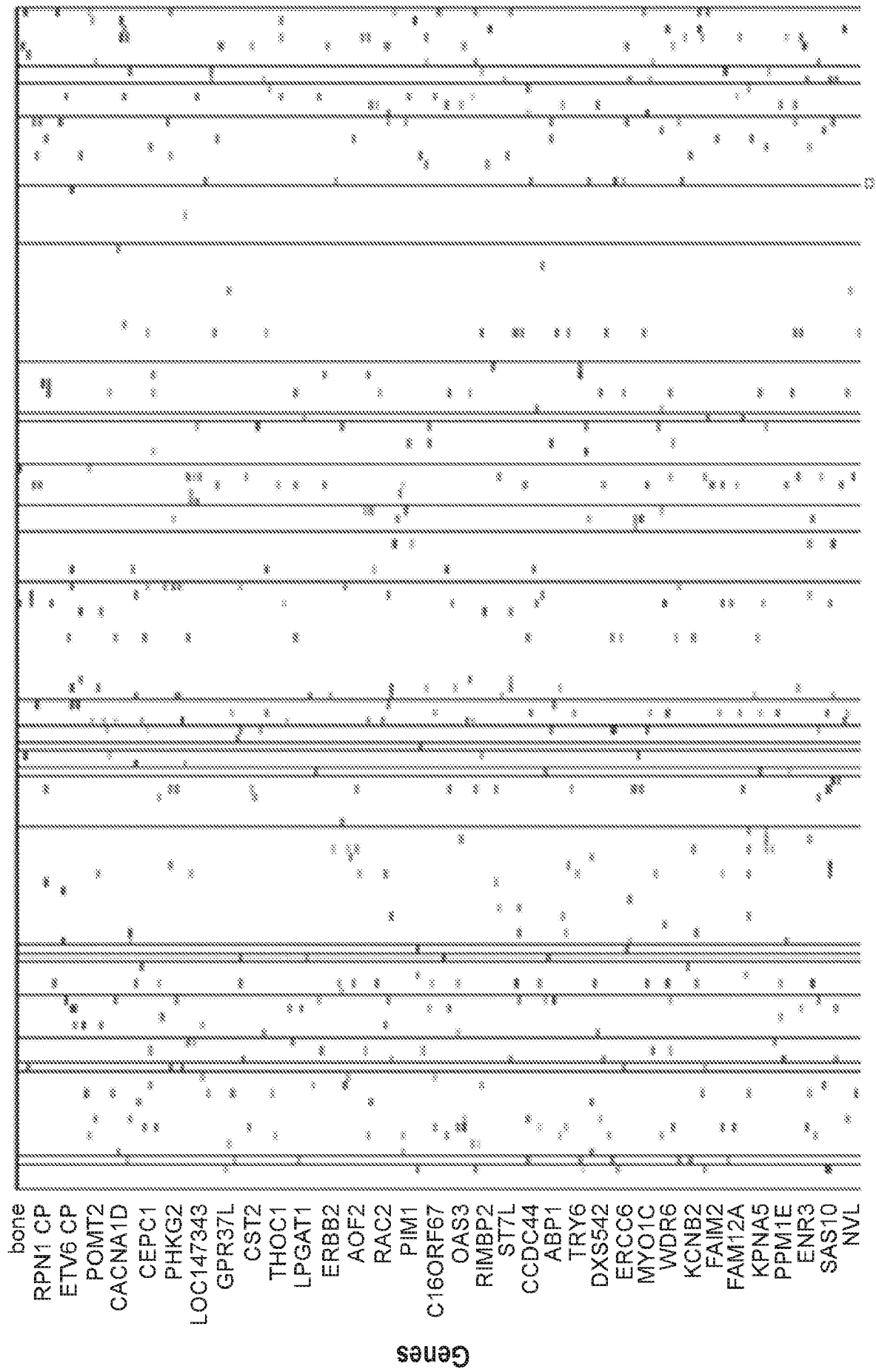


FIG. 5B (PART 1)

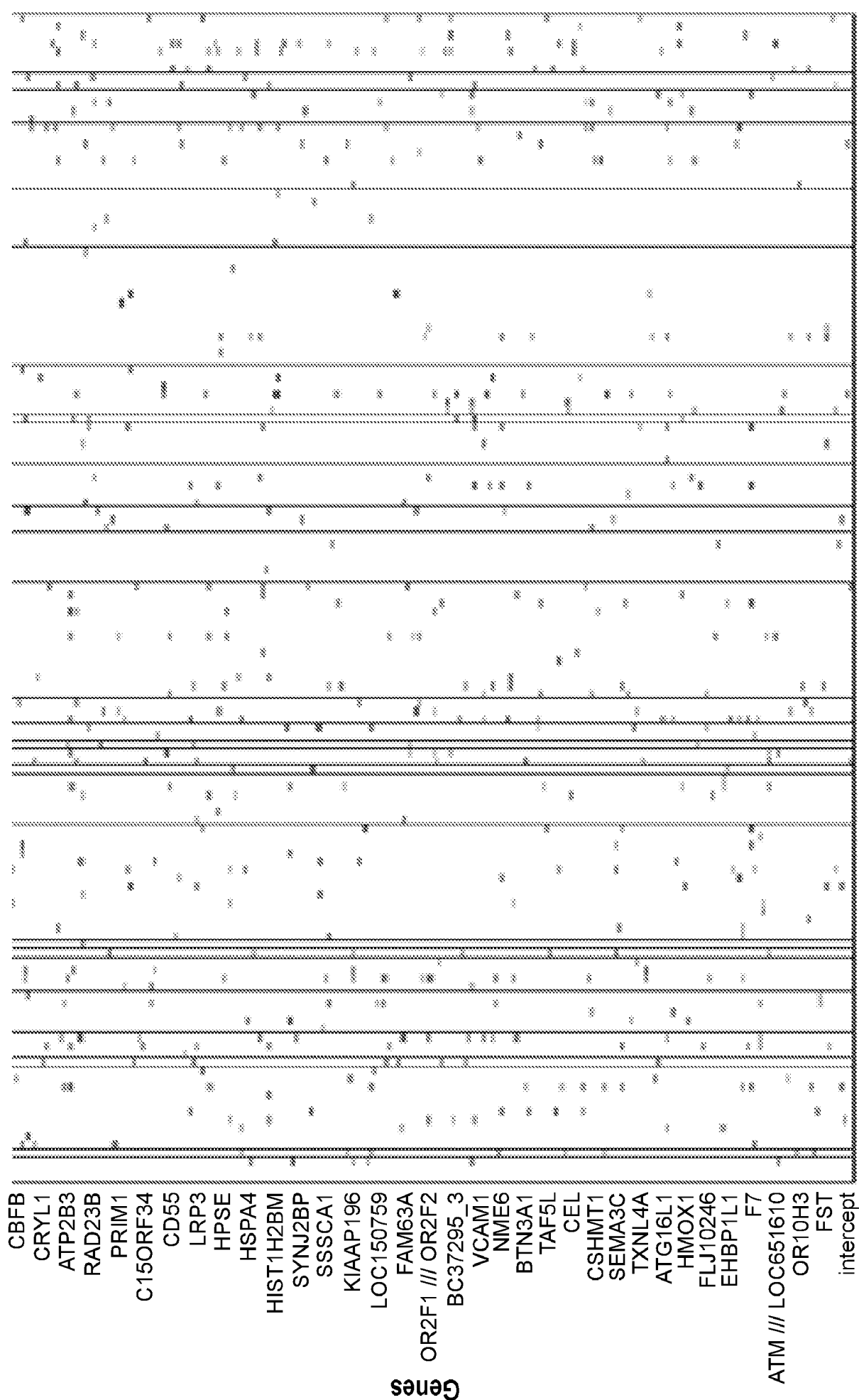


FIG. 5B (PART 2)

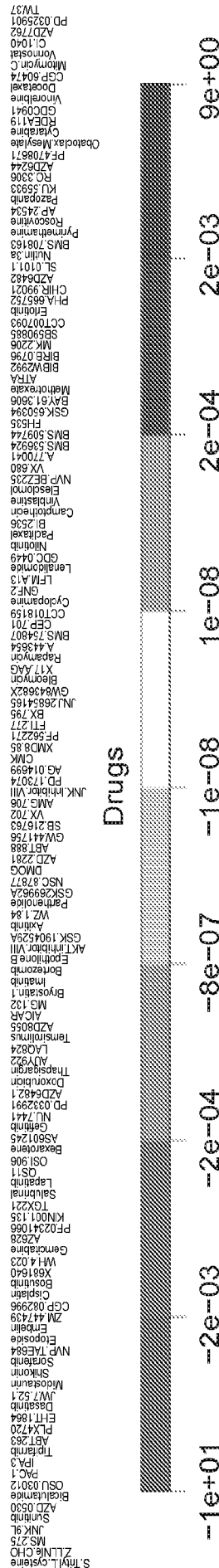


FIG. 5B (PART 3)

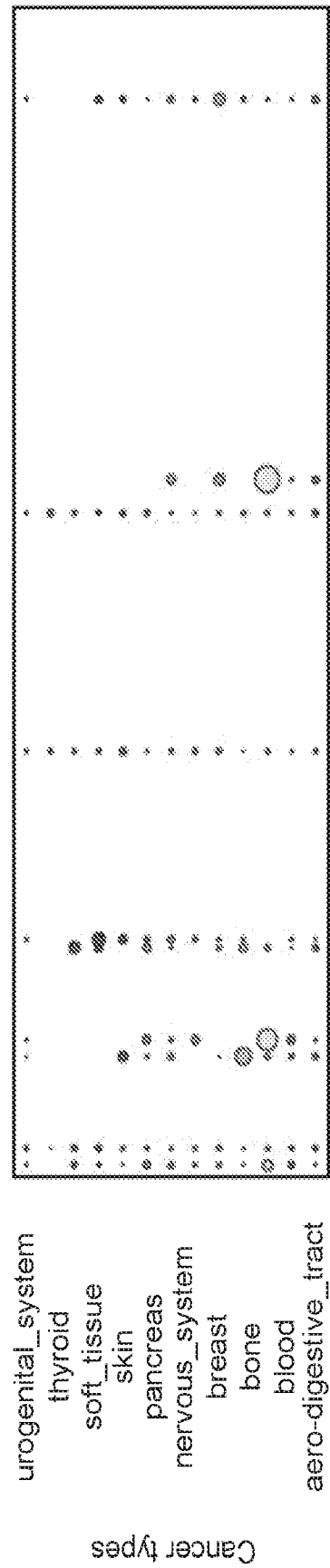


FIG. 6A

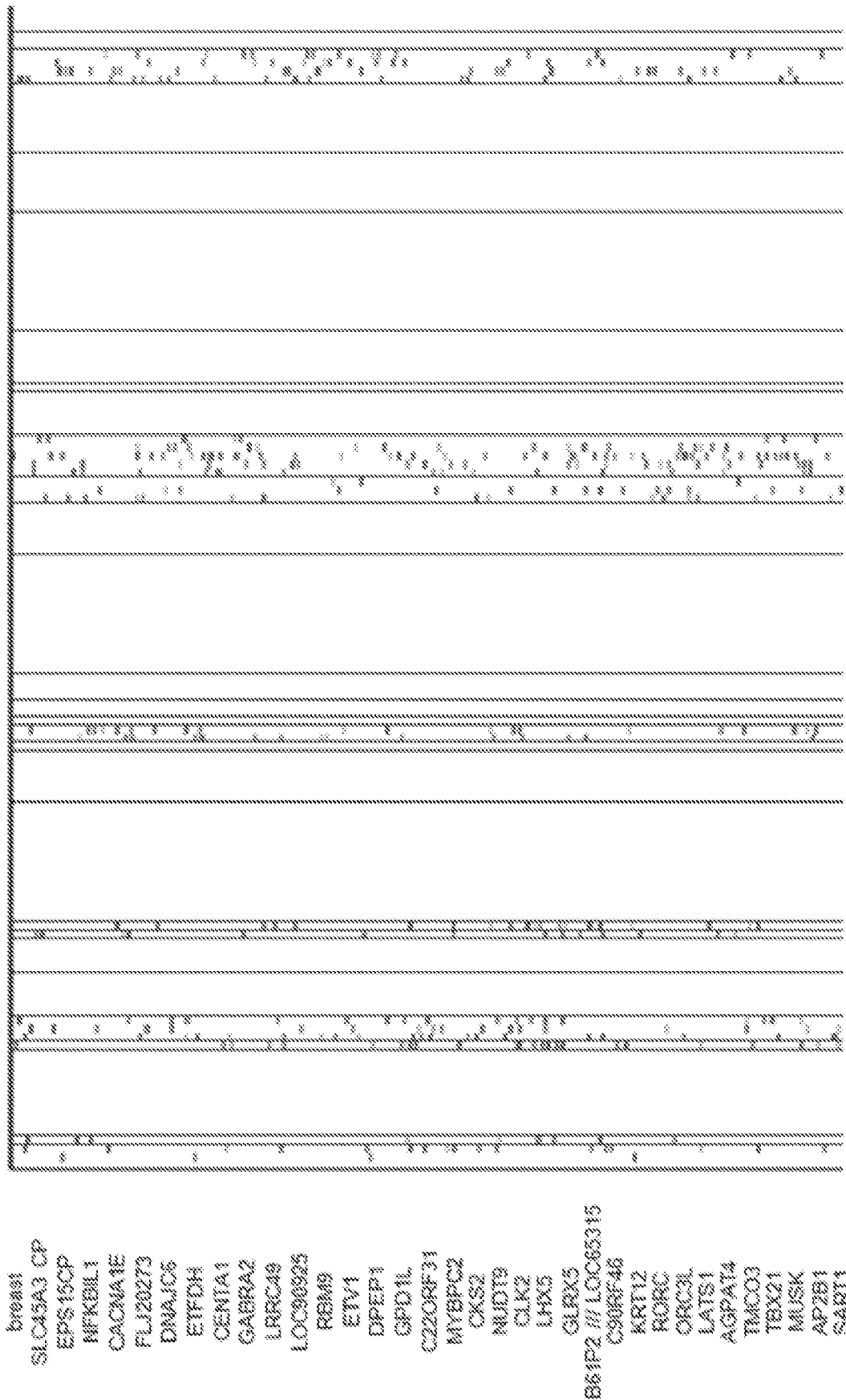


FIG. 6B (PART 1)

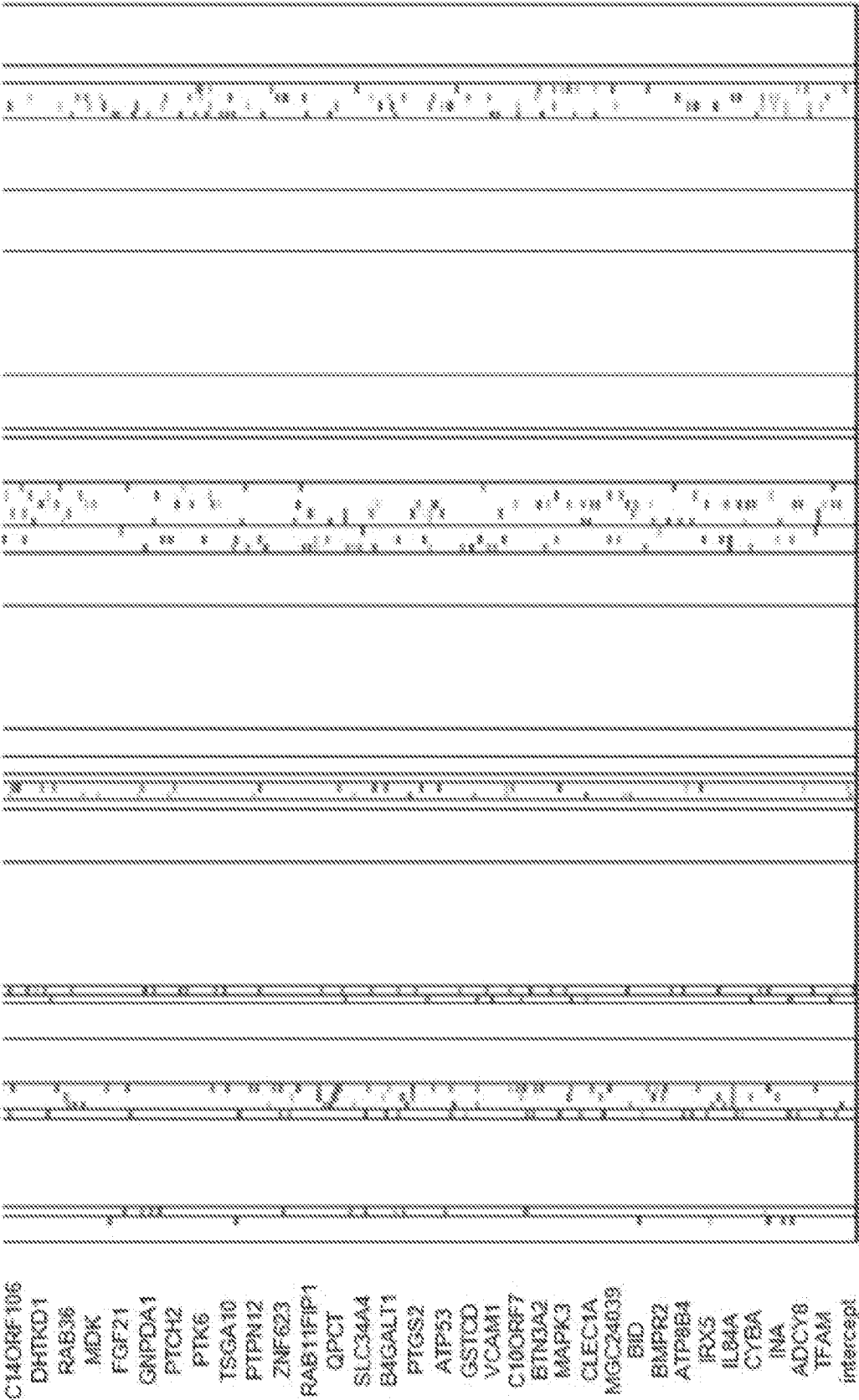


FIG. 6B (PART 2)

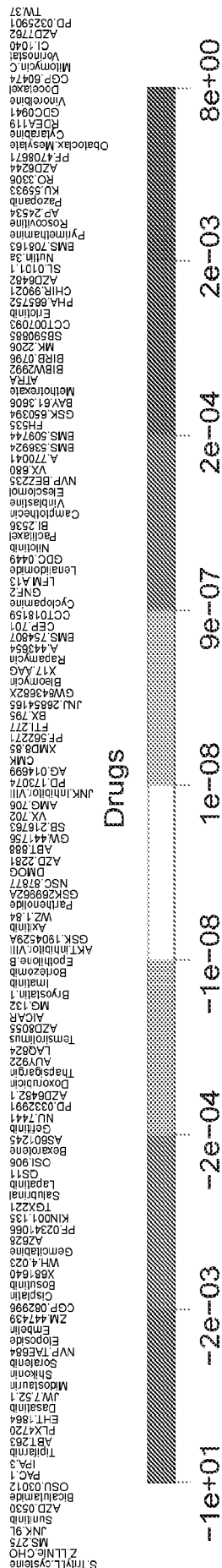


FIG. 6B (PART 3)

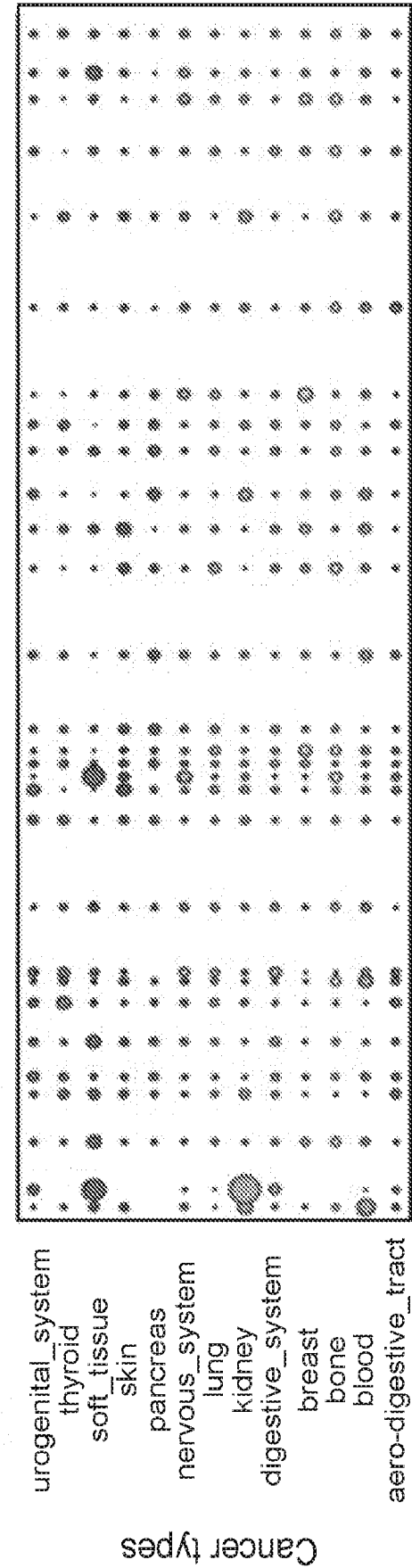


FIG. 7A

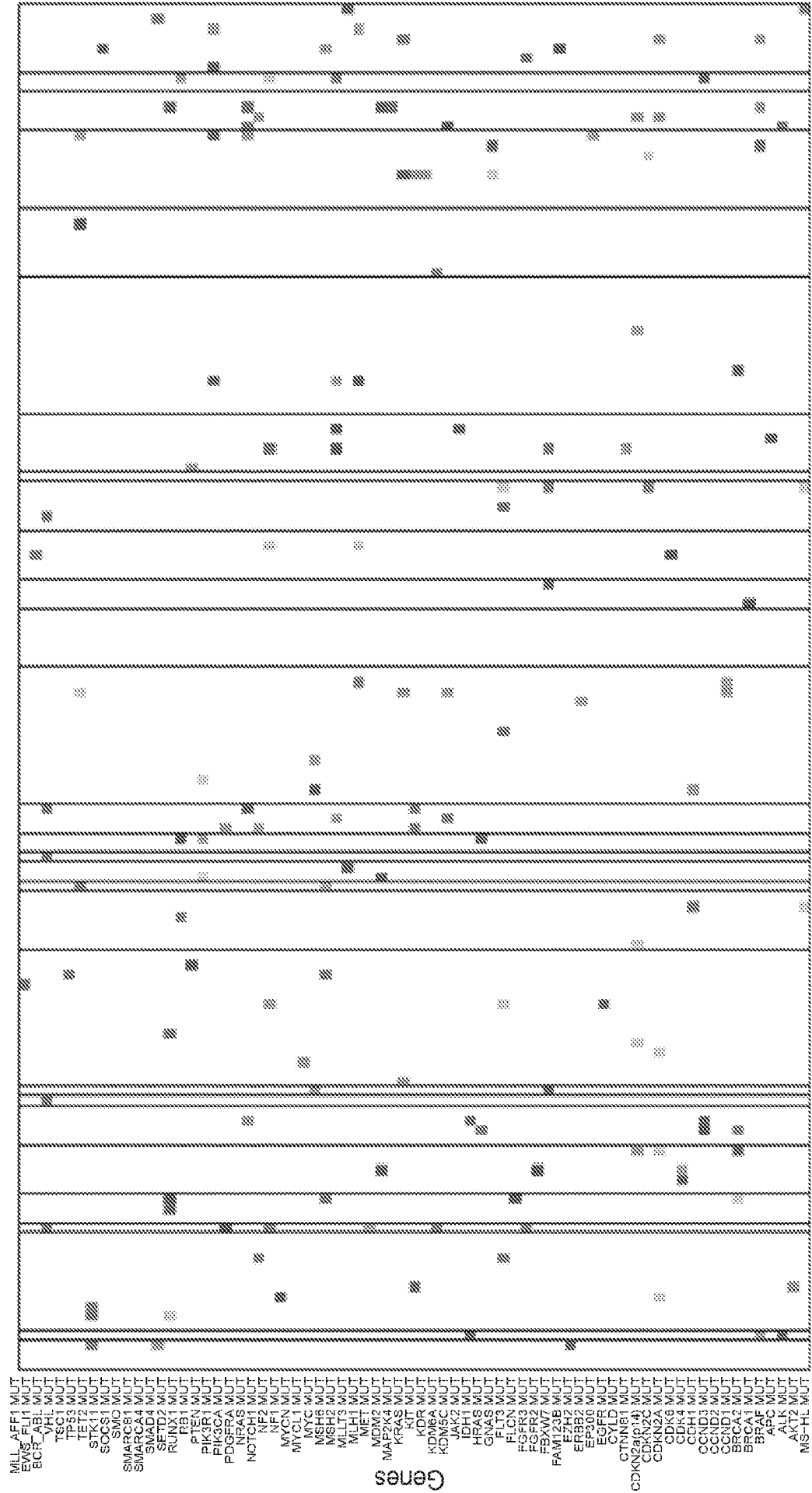


FIG. 7B

19/39

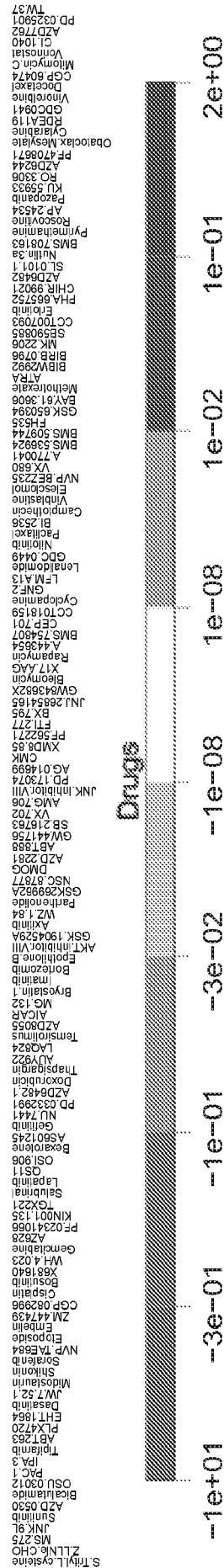


FIG. 7C

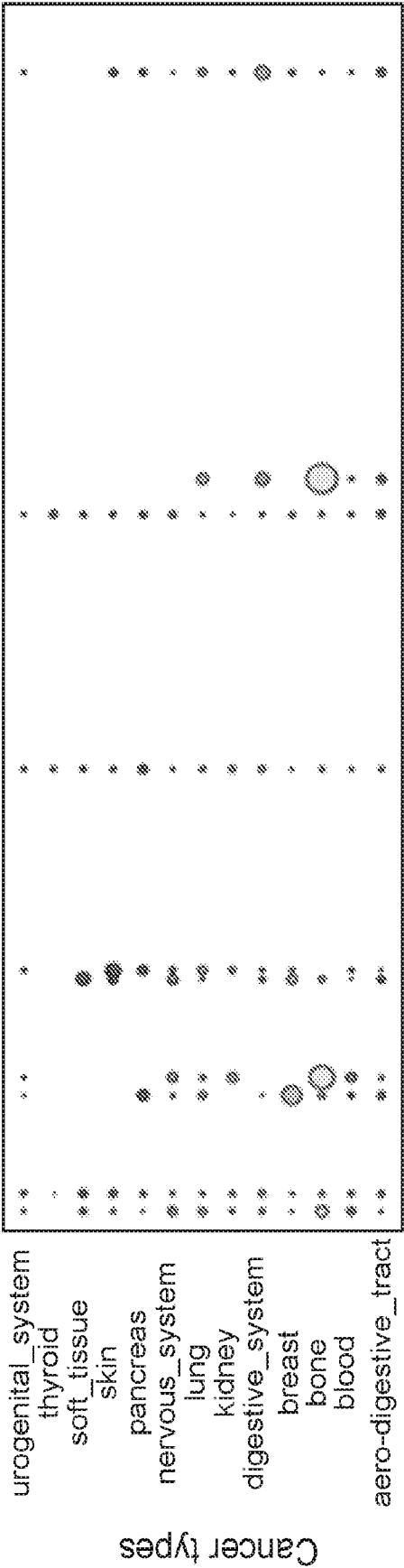


FIG. 8A

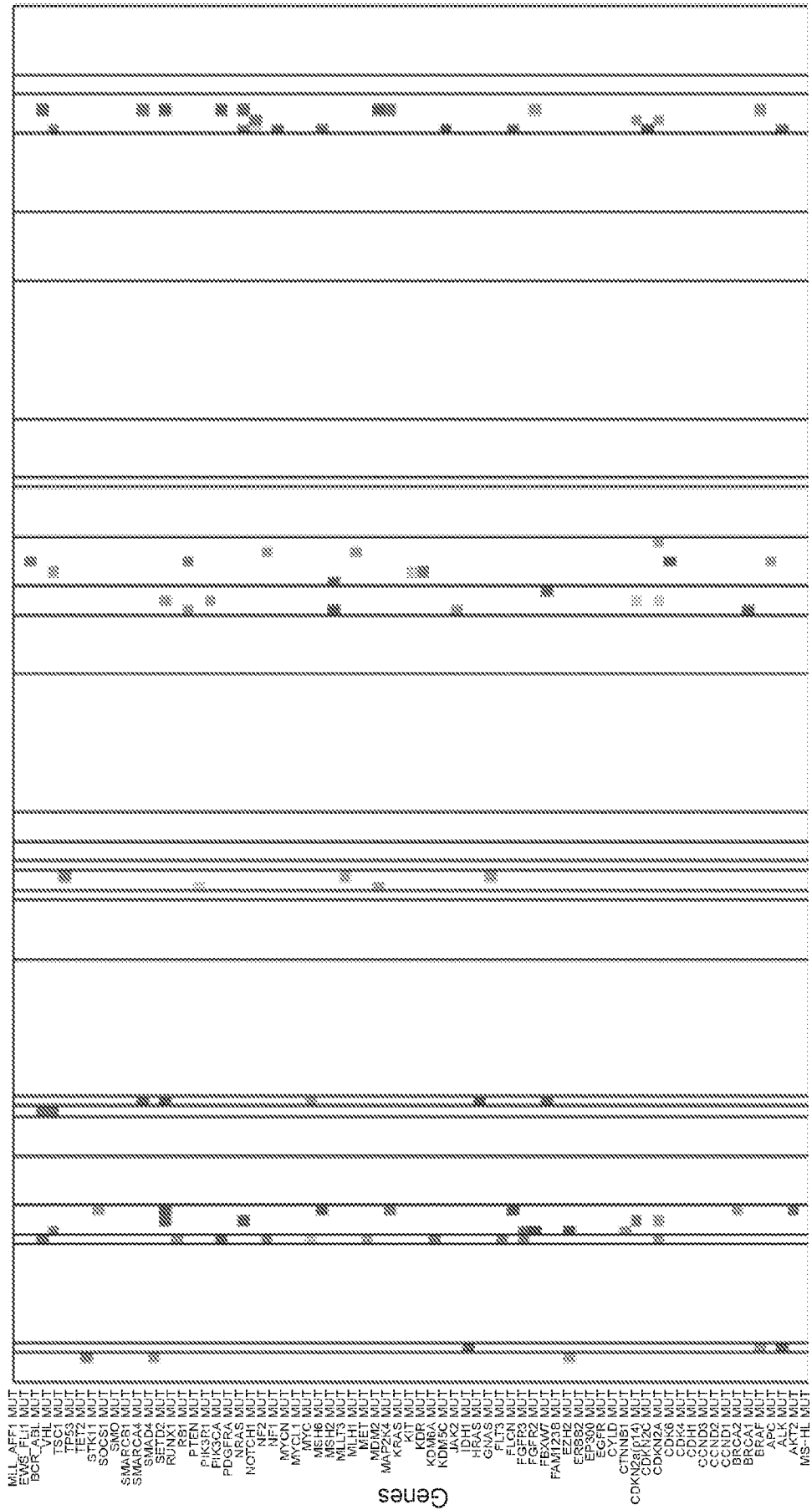


FIG. 8B

22/39

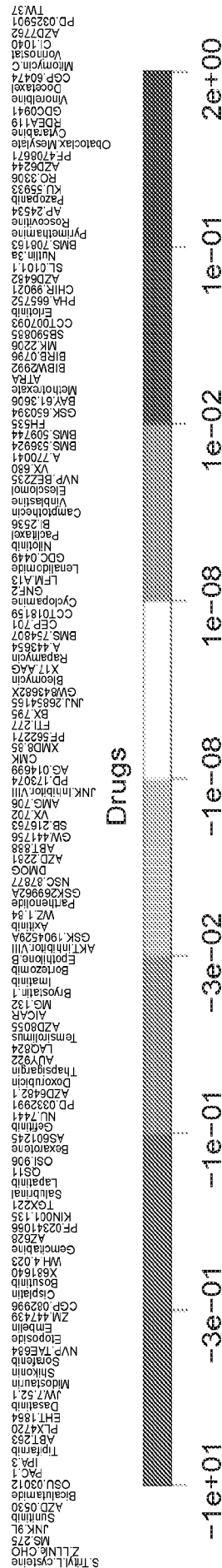


FIG. 8C

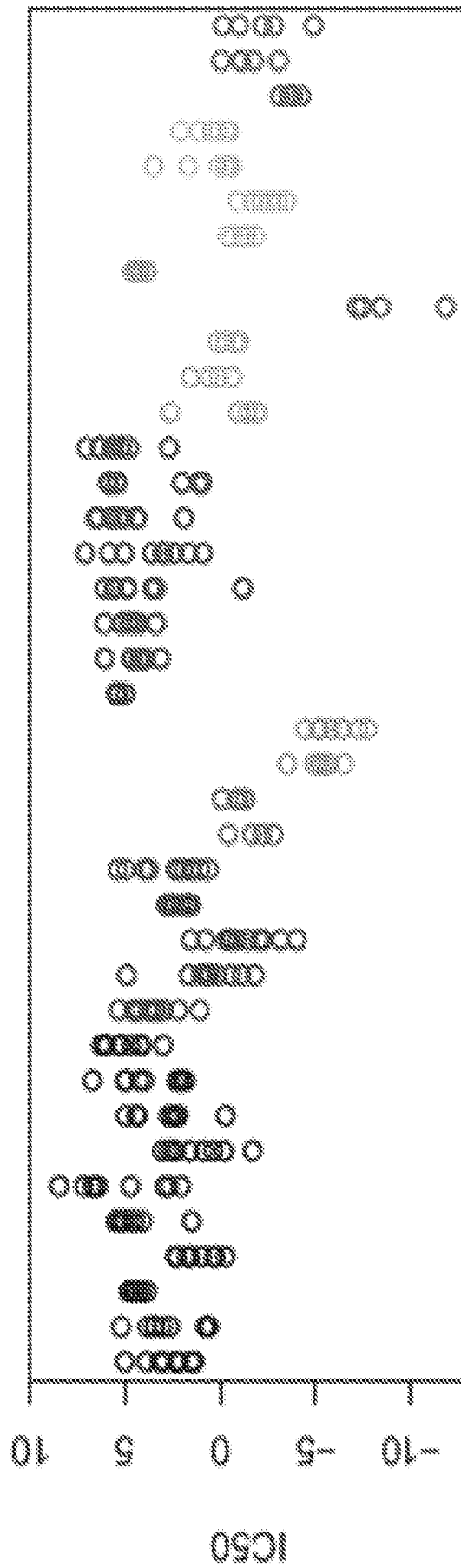
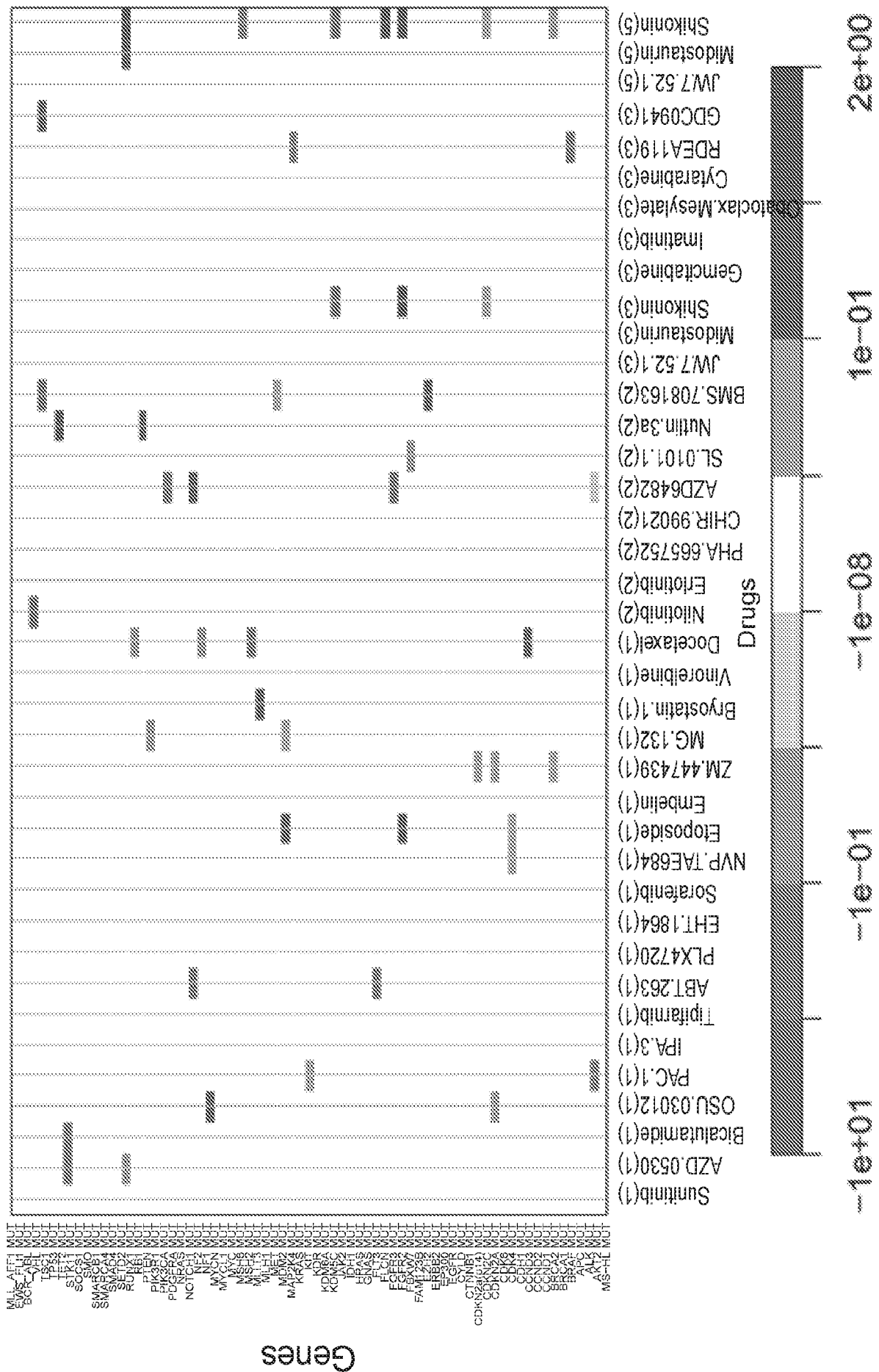


FIG. 9A



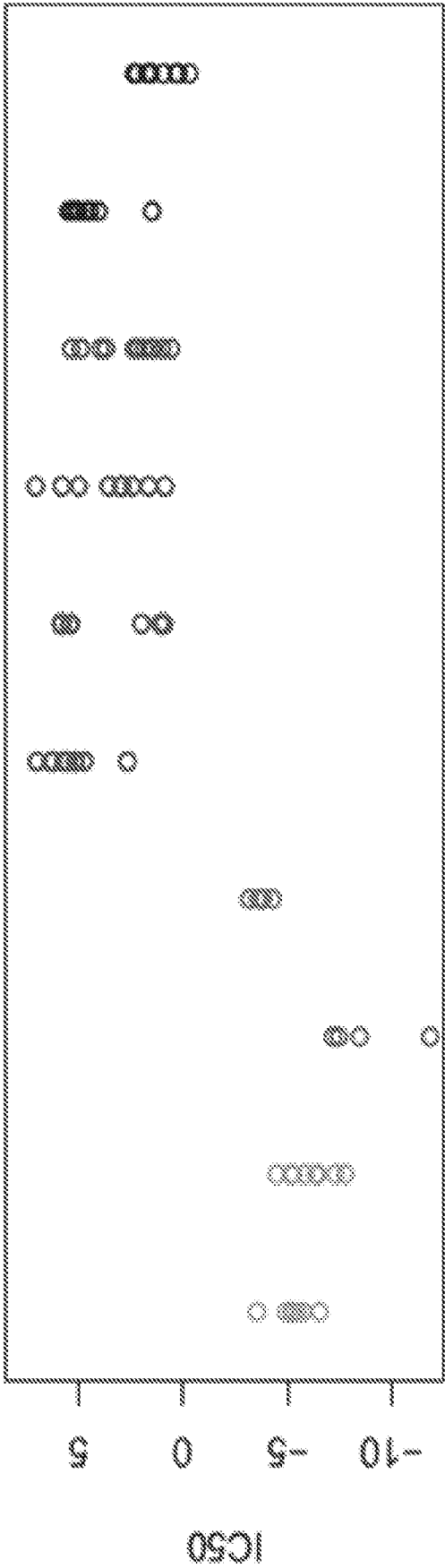


FIG. 10A

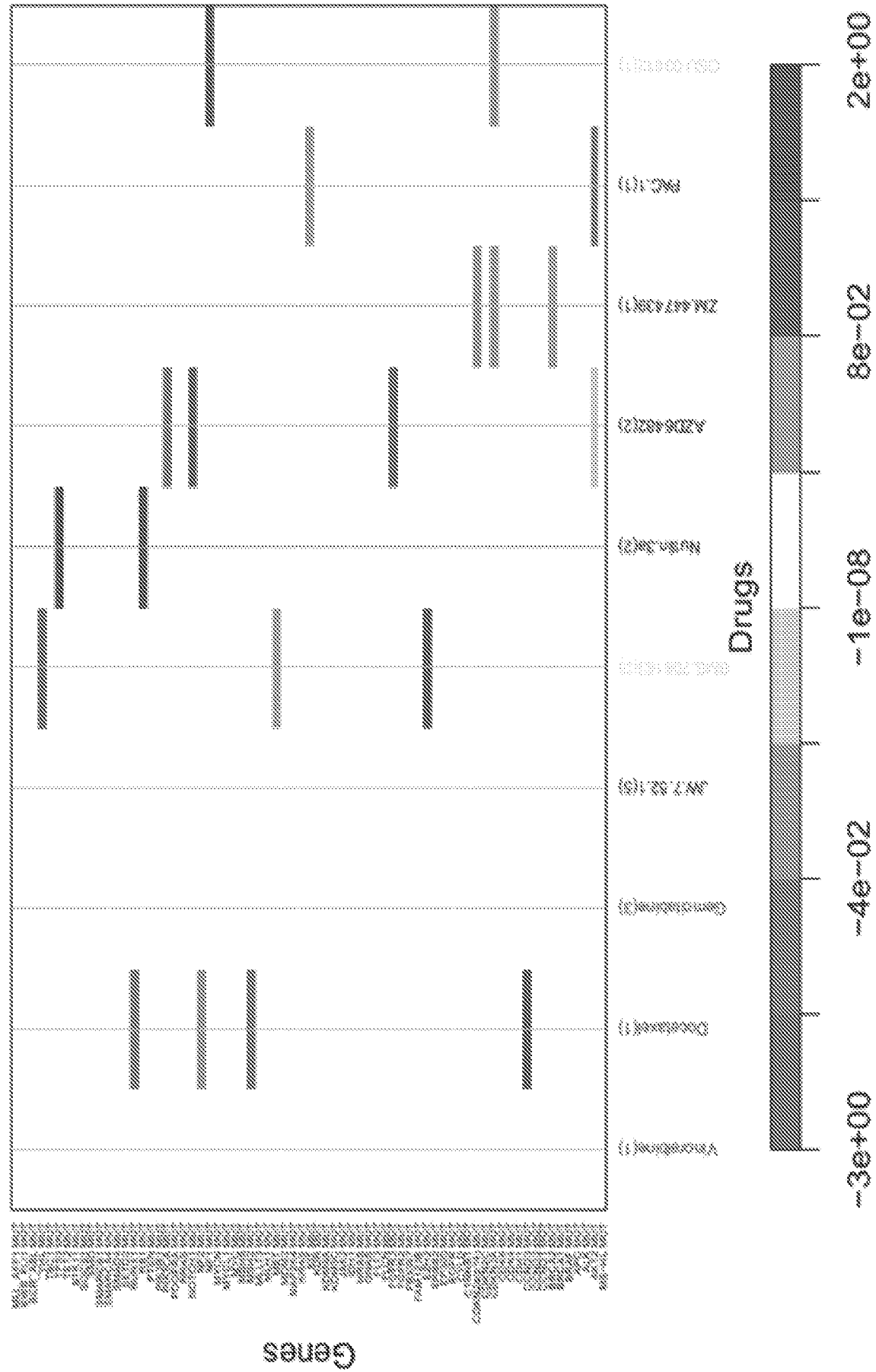


FIG. 10B

27/39

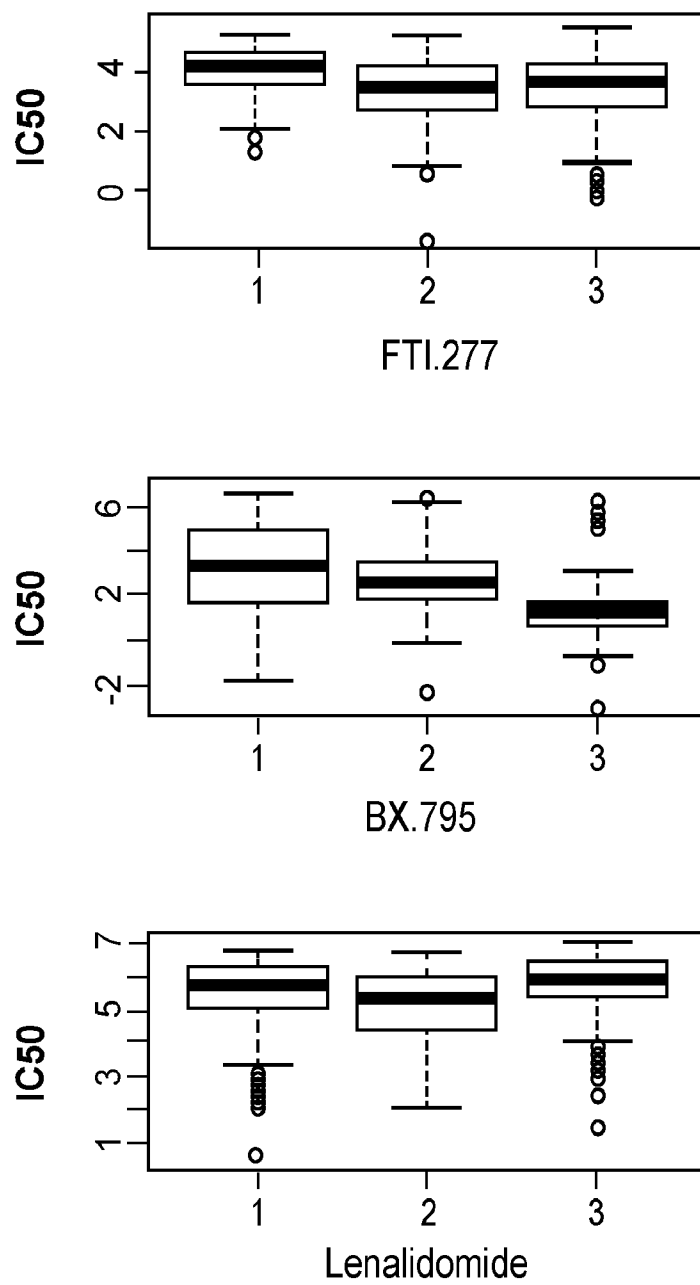


FIG. 11A

28/39

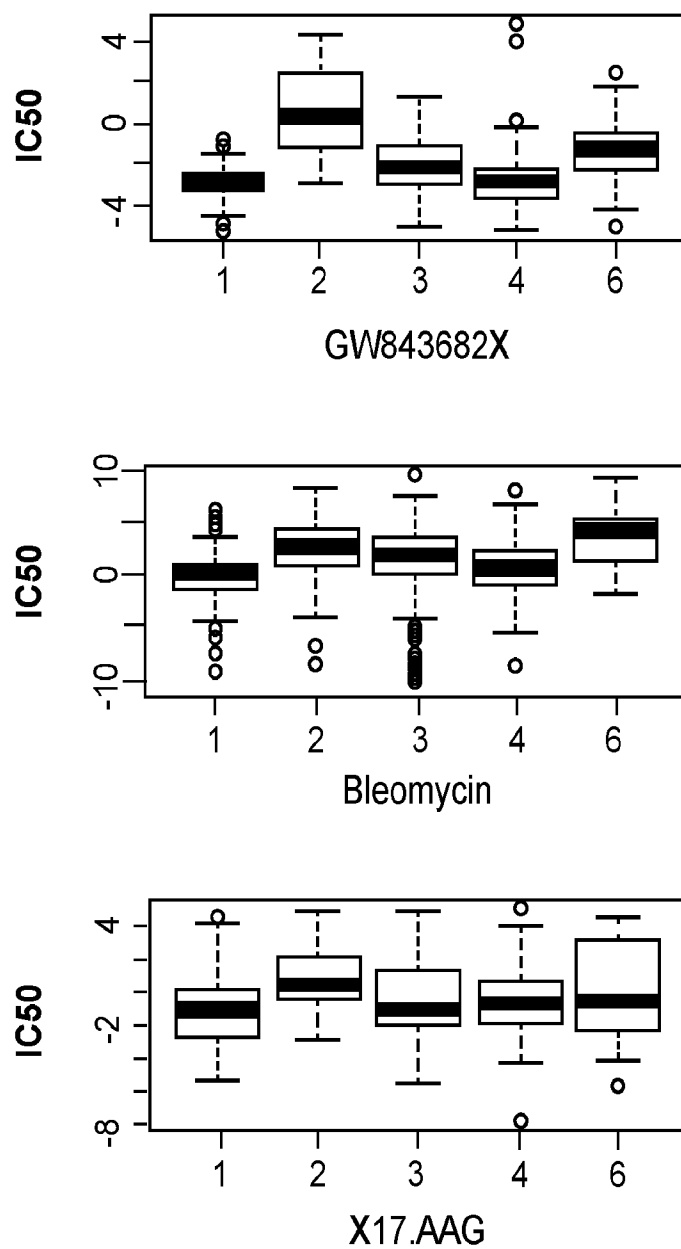


FIG. 11B

29/39

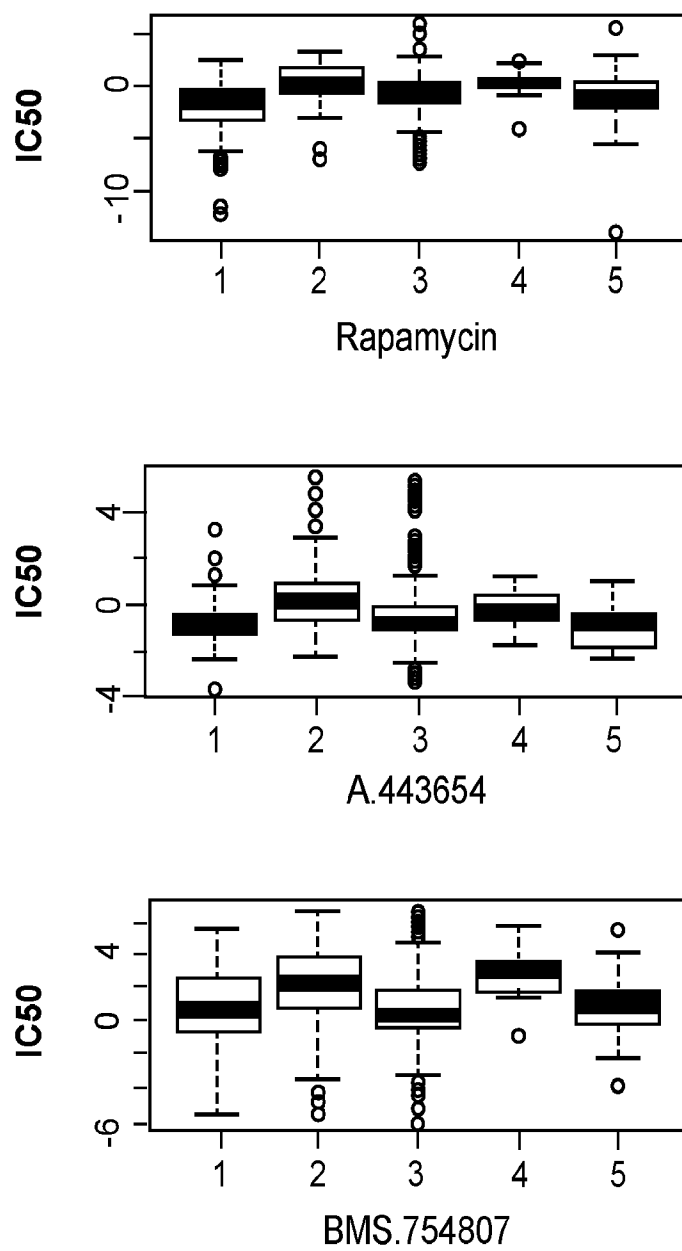
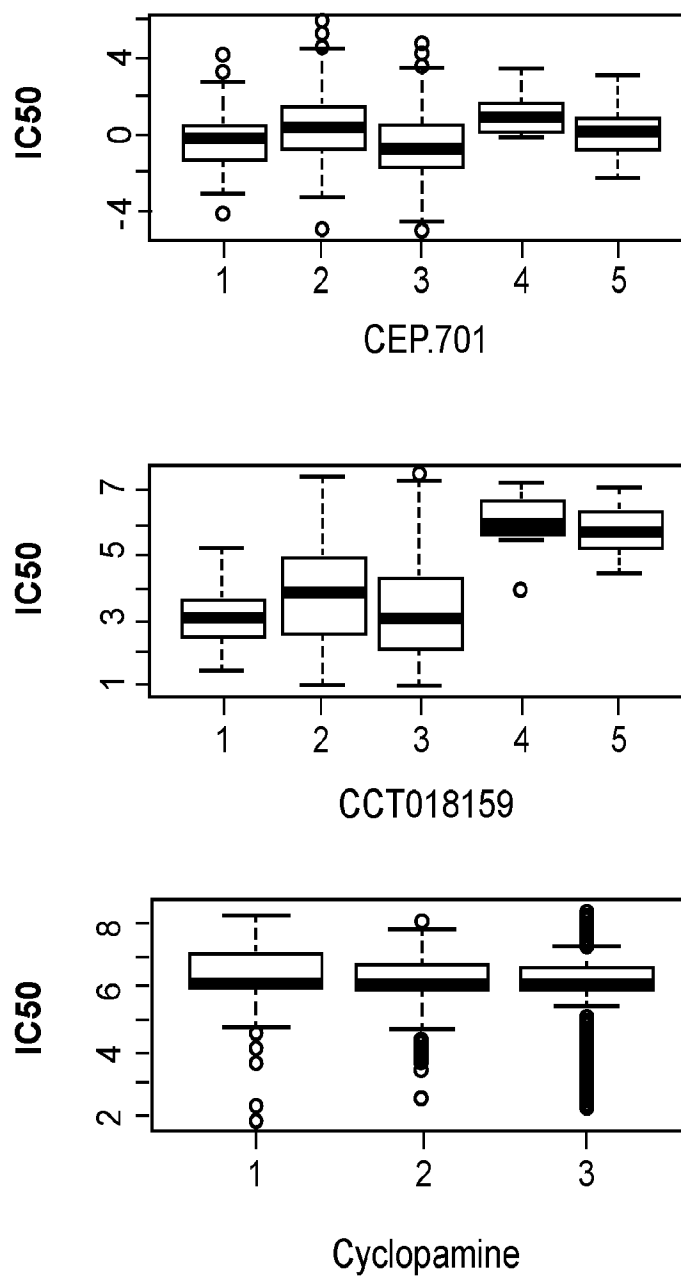


FIG. 11C

30/39

**FIG. 11D**

31/39

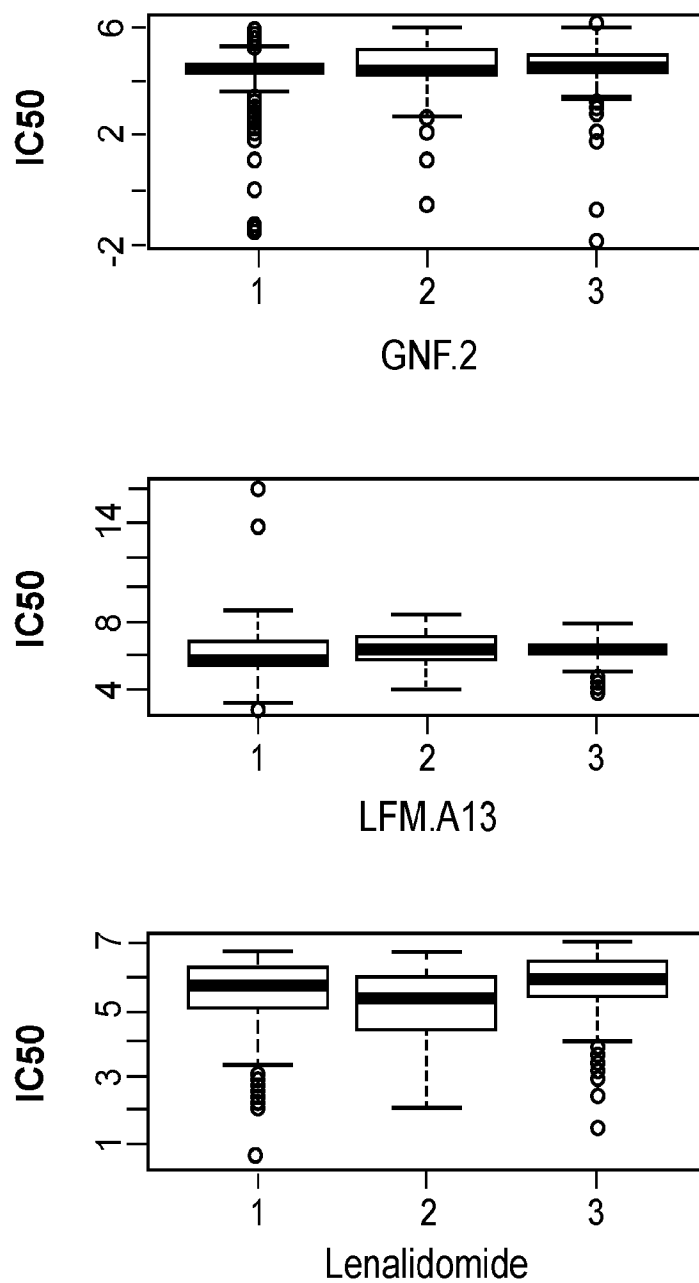


FIG. 11E

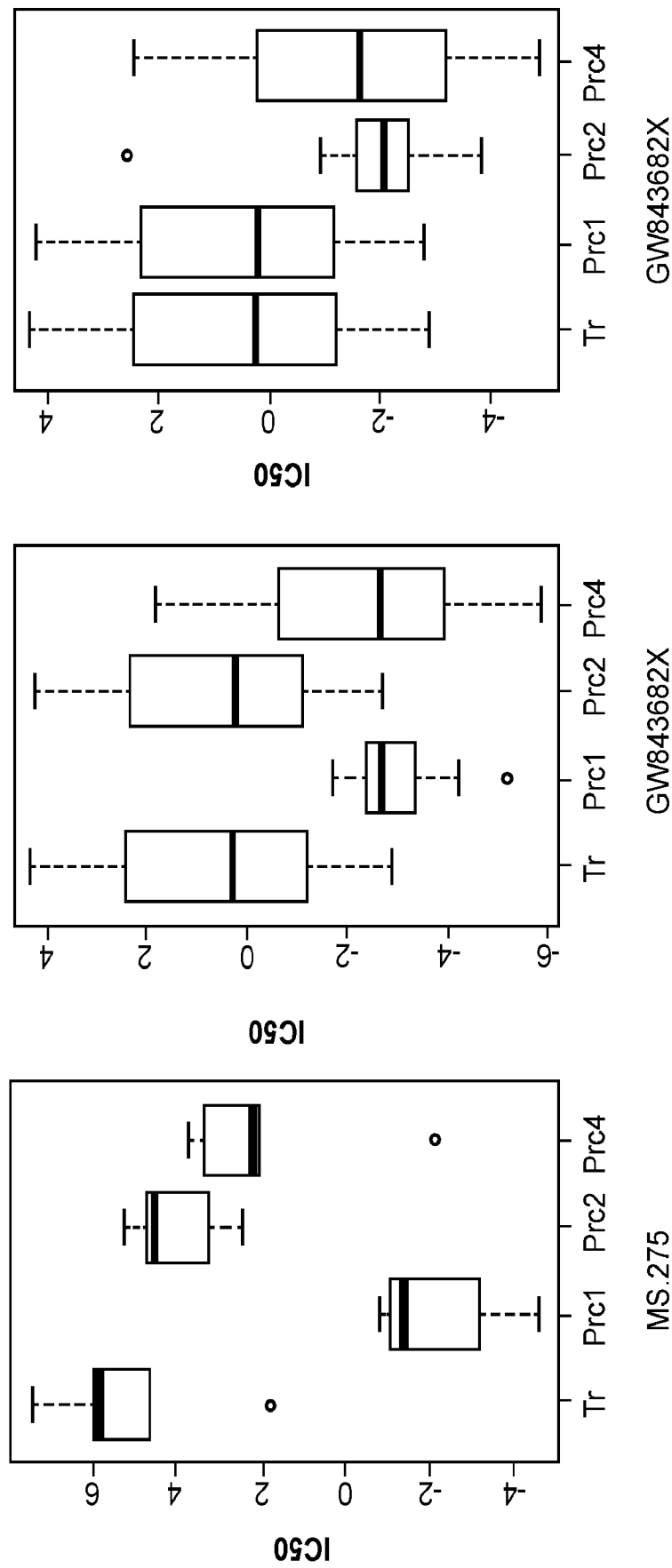


FIG. 12A

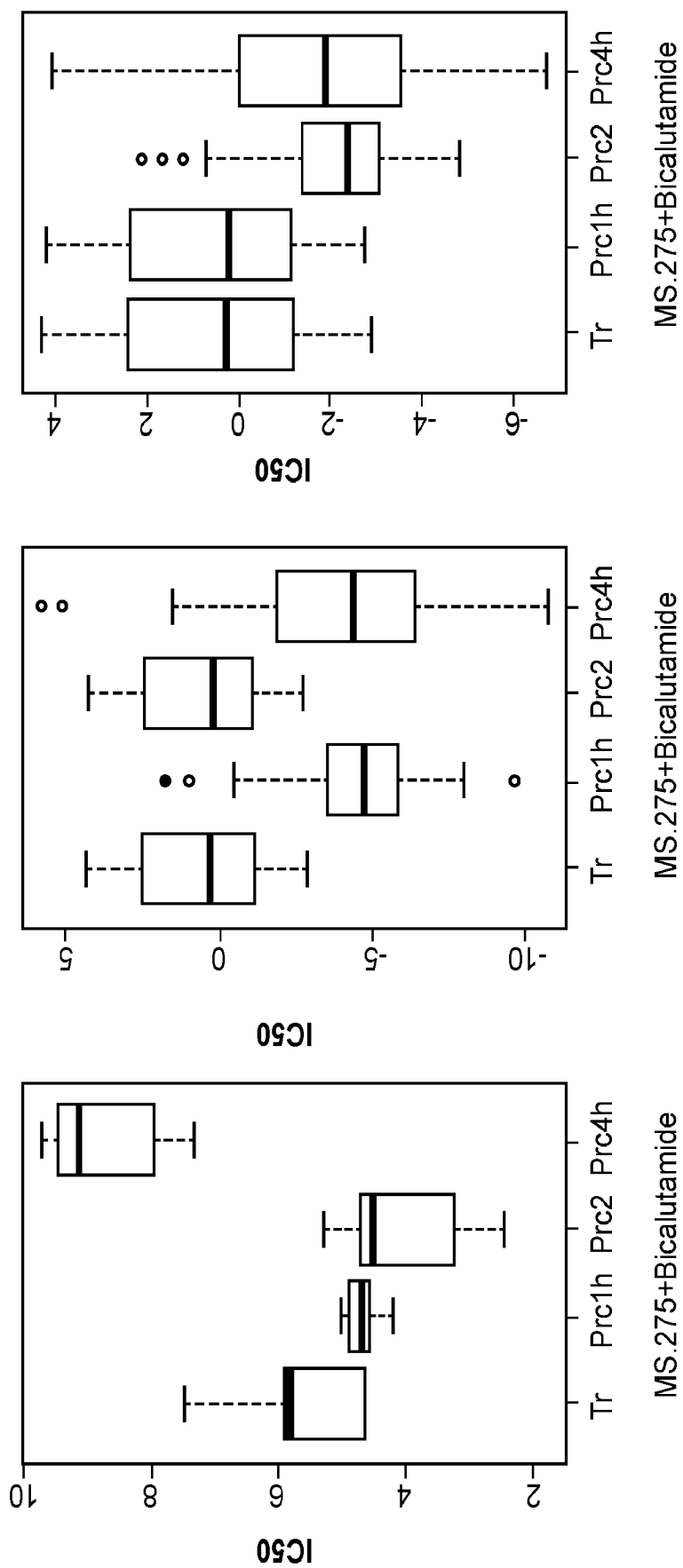


FIG. 12B

34/39

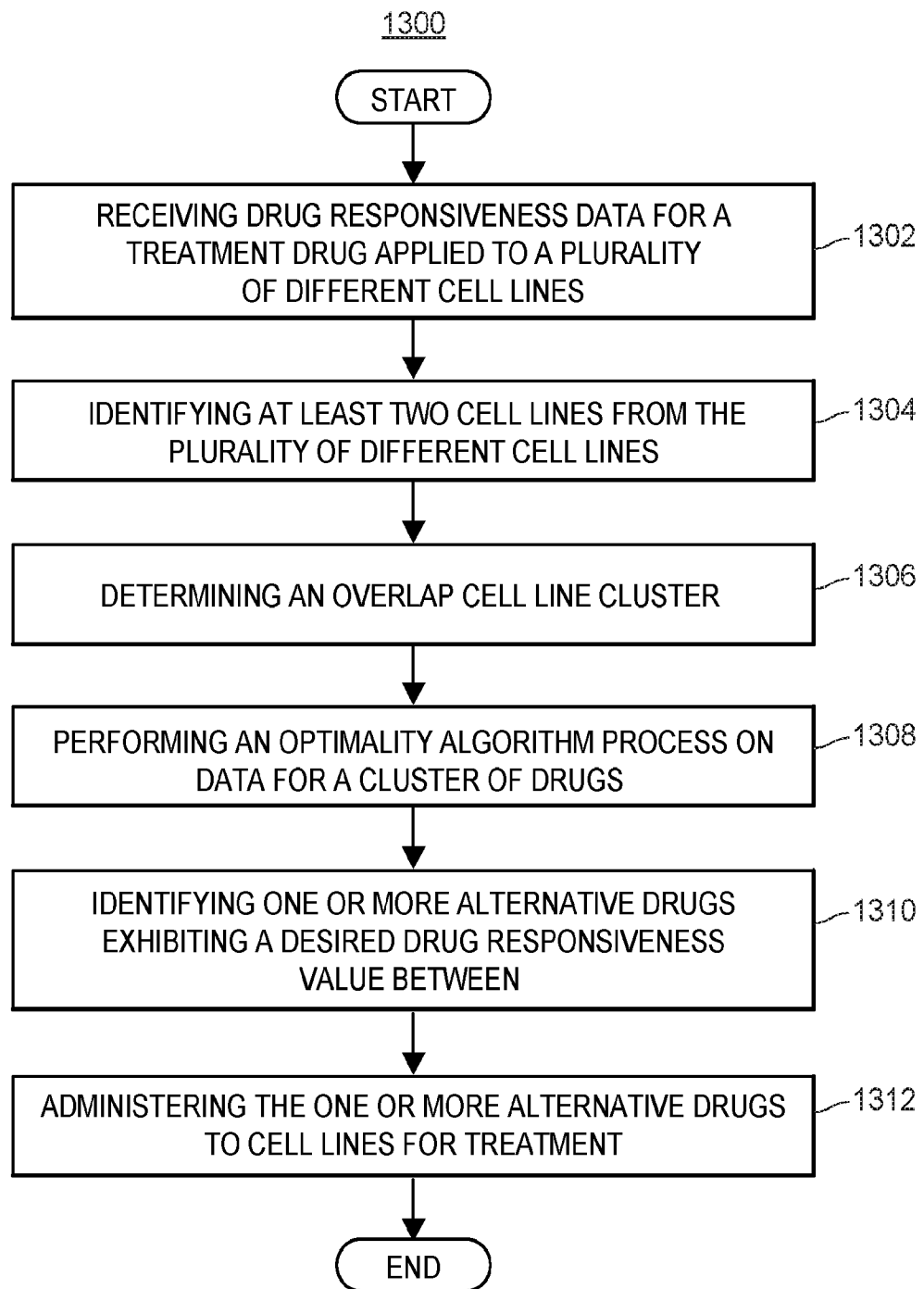


FIG. 13

35/39

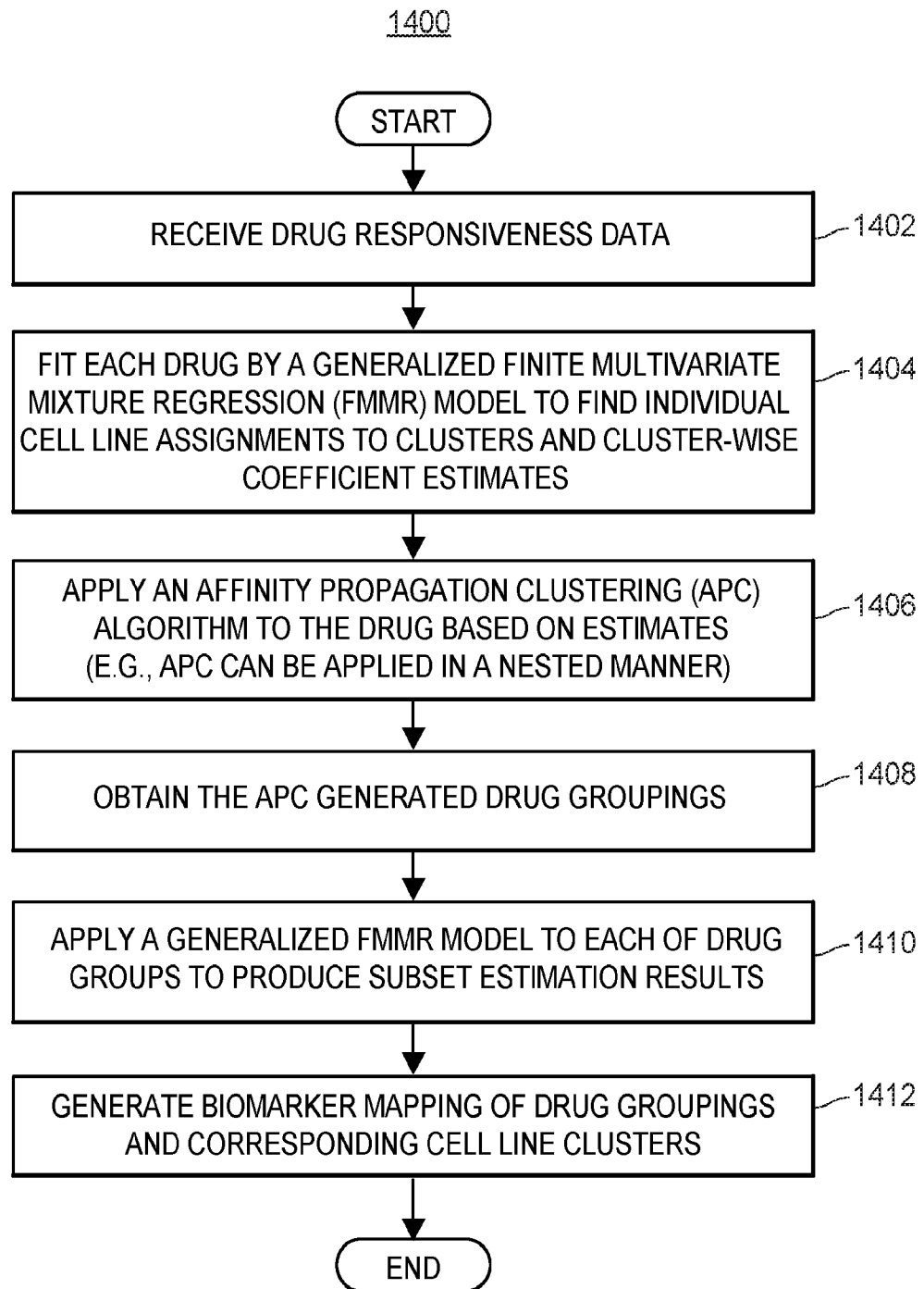


FIG. 14

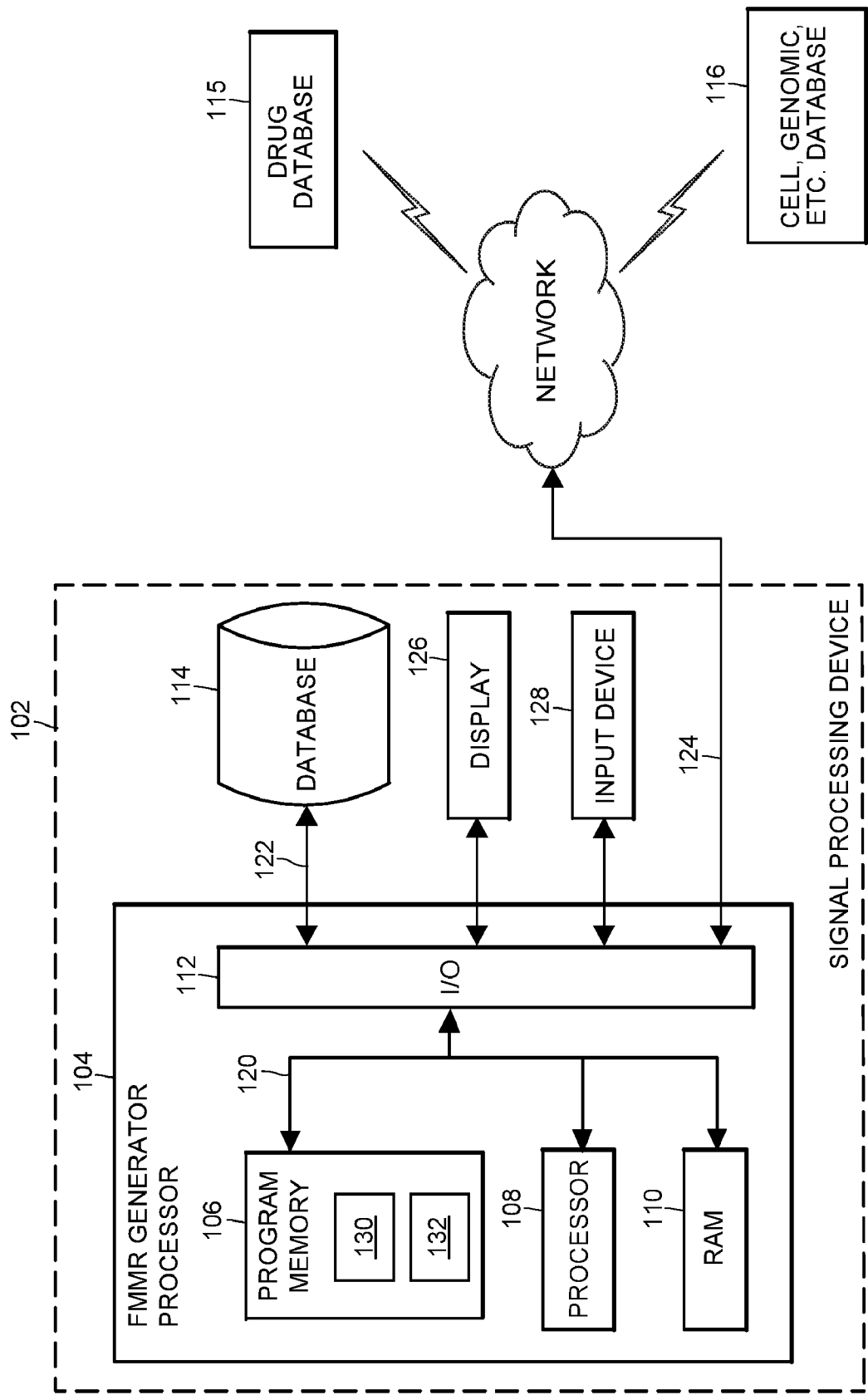


FIG. 15

Sensitive drug names	Resistant drug names
Vinorelbine	BMS.708163
Docetaxel	Nutlin.3a
Gemcitabine*	AZD6482
JW.7.52.1	ZM.447439
	PAC.1
	OSU.03012

Table 1: Sensitive versus resistant drugs for soft tissue cancer.

FIG. 16

for $k = 1 : K$ do

At this point, we have already found $k - 1$ layers and are searching for the k th layer.

Calculate the residuals from previous $k - 1$ layers by

$$z_i = y_i - \sum_{l=1}^{k-1} x_i^T B_l P_{il}, \quad i = 1, \dots, n.$$

Initialize P_k^0 .

for $s = 1 : S$ do

Given P_k^{s-1} , compute

$$B_k^s = \arg \min_{B_k} \frac{1}{2} \sum_{i=1}^n \|z_i - x_i^T B_k P_{ik}^{s-1}\|_2^2 + \sum_{j=1}^g \lambda_{jk} \|b_{jk}\|_1,$$

where $B_k = (b_{1k}, \dots, b_{gk})$.

Given B_k^s , compute

$$P_{ik}^s = \begin{cases} 0.5 + \min\{0.5, \frac{S}{2(S-T)}\}, & \|z_i - x_i^T B_k^s\|_2^2 \leq \|z_i\|_2^2, \\ 0.5 - \min\{0.5, \frac{S}{2(S-T)}\}, & \text{otherwise,} \end{cases}$$

where T was taken as $0.2S$ for $i = 1, \dots, n$.

end

Given P_k^S , update B_k^{S+1} .

Given B_k^{S+1} , prune layer memberships by

$$P_{ik}^{S+1} = \begin{cases} 1, & \|z_i - x_i^T B_k^{S+1}\|_2^2 \leq \tau \|z_i\|_2^2, \\ 0, & \text{otherwise,} \end{cases}$$

where $\tau \in (0, 1)$ for $i = 1, \dots, n$.

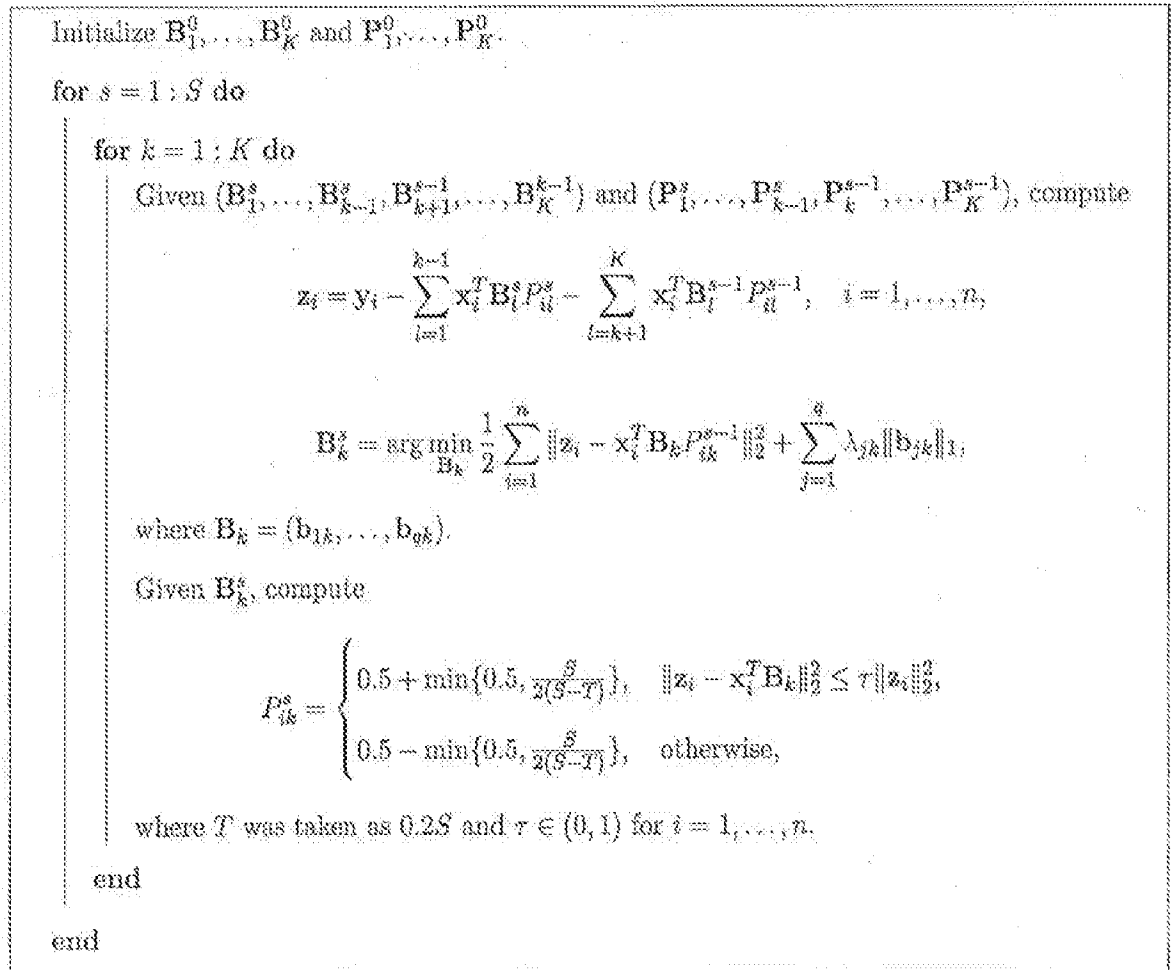
Given P_k^{S+1} , update B_k^{S+2} .

Given $P_1^{S+1}, \dots, P_k^{S+1}$, back fit the layers R times. We set $R = 2$.

end

Algorithm 1: Plaid

FIG. 17



Algorithm 2: All-plaid

FIG. 18

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US2017/044712

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - A61K 31/277; A61K 31/337; A61K 31/437; A61K 31/44; A61K 31/473; A61K 31/475 (2017.01)

CPC - A61K 31/277; A61K 31/337; A61K 31/437; A61K 31/44; A61K 31/473; A61K 31/475; A61K 31/513; A61K 31/704; A61K 33/24; C12Q 1/6886; C12Q 2600/106; C12Q 2600/136; C12Q 2600/158; G01N 33/5011; G01N 33/57449; G01N 2800/52 (2017.08)

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)

See Search History document

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

USPC - 424/649; 435/6.11; 435/6.12; 435/7.1; 435/7.23; 435/7.92; 435/29; 506/2; 506/9; 506/16 (keyword delimited)

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

See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GARNETT et al. "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells," Nature, 29 March 2012 (29.03.2012), Vol. 483, Pgs. 570-575 and Methods. entire document	11-14
Y		1-10
Y	US 2010/0144554 A1 (SEMIZAROV et al) 10 June 2010 (10.06.2010) entire document	1-10
Y	US 2008/0118576 A1 (THEODORESCU et al) 22 May 2008 (22.05.2008) entire document	9, 10
P, X	LIU et al. "Precision Therapeutic Biomarker Identification with Application to the Cancer Genome Project" arXiv preprint arXiv:1702.02264. 2017 Feb 8. (08.02.2017) entire document	1-14
A	US 2010/0198900 A1 (GIFFORD) 05 August 2010 (05.08.2010) entire document	1-14
A	BASU et al. "An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules," Cell, 29 August 2013 (29.08.2013), Vol. 154, Pgs. 1151-1161. entire document	1-14
A	US 2016/0102365 A1 (UNIVERSITY OF MIAMI) 14 April 2016 (14.04.2016) entire document	1-14

☐ 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 another 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

"&" document member of the same patent family

Date of the actual completion of the international search

18 September 2017

Date of mailing of the international search report

06 OCT 2017

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, VA 22313-1450
Facsimile No. 571-273-8300

Authorized officer

Blaine R. Copenheaver

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774