(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
28 July 2016 (28.07.2016)

WIPO | PCT

(10) International Publication Number
WO 2016/118527 A1

(51) International Patent Classification:
G06F 19/24 (2011.01) *G06F 19/18* (2011.01)
G06F 19/22 (2011.01)

(21) International Application Number:
PCT/US2016/013959

(22) International Filing Date:
19 January 2016 (19.01.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/105,697 20 January 2015 (20.01.2015) US
62/127,546 3 March 2015 (03.03.2015) US

(71) Applicant: **NANTOMICS, LLC** [US/US]; 9920 Jefferson Boulevard, Culver City, California 90232 (US).

(72) Inventor: **SZETO, Christopher**; 4530 W. Walnut Street #1, Soquel, California 95073 (US).

(74) Agents: **FESSENMAIER, Martin** et al.; Fish & Tsang, LLP, 2603 Main Street, Ste 1000, Irvine, California 92614 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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))*

— *with amended claims (Art. 19(1))*

(54) Title: SYSTEMS AND METHODS FOR RESPONSE PREDICTION TO CHEMOTHERAPY IN HIGH GRADE BLADDER CANCER

(57) Abstract: Contemplated systems and methods allow for prediction of chemotherapy outcome for patients diagnosed with high-grade bladder cancer. In particularly preferred aspects, the prediction is performed using a model based on machine learning wherein the model has a minimum predetermined accuracy gain and wherein a thusly identified model provides the identity and weight factors for omics data used in the outcome prediction.

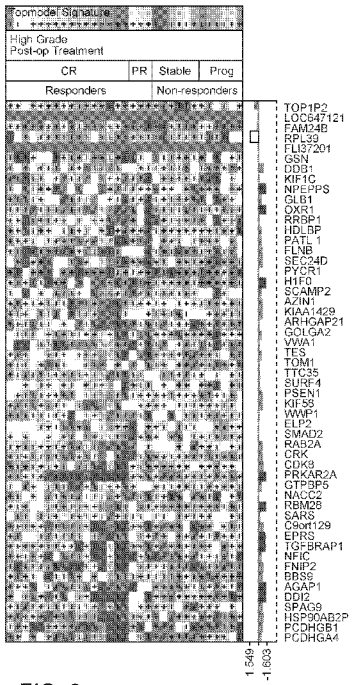


FIG. 2

WO 2016/118527 A1

SYSTEMS AND METHODS FOR RESPONSE PREDICTION TO CHEMOTHERAPY IN HIGH GRADE BLADDER CANCER

[0001] This application claims priority to US provisional application with the serial number 62/105697, which was filed 20-Jan-15, and US provisional application with the serial number 62/127546, which was filed 03-Mar-15, both of which are incorporated by reference herein.

Field of the Invention

[0002] The field of the invention is *in silico* systems and methods for prediction of treatment outcome for chemotherapy in bladder cancer.

Background of the Invention

[0003] The background description includes information that may be useful in understanding the present invention. It is not an admission that any of the information provided herein is prior art or relevant to the presently claimed invention, or that any publication specifically or implicitly referenced is prior art.

[0004] All publications herein are incorporated by reference to the same extent as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference. Where a definition or use of a term in an incorporated reference is inconsistent or contrary to the definition of that term provided herein, the definition of that term provided herein applies and the definition of that term in the reference does not apply.

[0005] Selection of pharmaceutical treatment options for cancer has historically been limited to empirical data and histological findings to so match a drug to a particular cancer type. More recently, advances in molecular medicine have allowed a more personalized approach in the choice of chemotherapy, taking into account presence or absence of specific receptors on a cell, mutational status of signaling molecules, etc. While such improvements have translated at least in some cases to increased survival time, response to a chemotherapeutic drug is in all or almost all cases not entirely predictable. Moreover, once a patient is committed to a specific treatment regimen, changes in treatment protocol are often not advised and/or poorly tolerated by the patient.

[0006] To help predict likely treatment outcome for pharmaceutical interventions, various computational systems and methods have been developed. Most notably, WO 2014/193982

describes systems and methods in which pathway elements (corresponding to cellular *in vivo* features) of a pathway model are modified *in silico* to simulate treatment of a cell with a drug. The modified model can then be used to help predict the effect of the drug on one or more pathways, and indirectly predict the effect of the drug on a diseased tissue. While such system has provided remarkable predictive power in certain circumstances, such system was based on cell culture data and as such did not fully reflect *in vivo* environments. Moreover, simulation of the treatment was performed using a single model that was rooted in measured and assumed attributes and therefore relied on specific assumptions genuine to the model. The described approach fails to provide insight into mitigating risks associated with the specific assumptions of model.

[0007] To accommodate large quantities of data from complex *in vivo* systems, computer-based machine learning technologies have been developed that can ingest large data sets that exceed the capacity of human beings to assimilate. In general, machine learning algorithms are often configured to identify patterns in training data sets so that the algorithms "learn" or become "trained" how to predict possible outcomes when presented with new input data. Notably, there are numerous types of machine learning algorithms, each having their own specific underlying mode of analysis (*e.g.*, support vector machines, Bayesian statistics, Random Forests, etc.), and with that inherent bias. An example for such analysis is presented in US2004/0193019 to Wei in which discriminant analysis-based pattern recognition is used to generate a prediction model that correlates biological profile information with treatment outcome information. The so formed prediction model is then used to rank possible responses to treatment. Wei simply builds prediction outcome models to make an assessment of likely outcome based patient-specific profile information. Unfortunately, not all algorithms will be suitable for predictive analysis of drug treatment as each algorithm has built in assumptions that might not be valid for the specific disease and/or drug treatment. Furthermore, models that are maximized for a particular prediction will not necessarily provide the best accuracy as compared to a random event and/or other model.

[0008] To address such difficulties, US 2014/0199273 to Cesano et al. discusses selection of specific models/statistical methods that are suitable for prediction or prognosis in a healthcare setting. While Cesano discusses selection of suitable models, these models, once selected still suffer from the same difficulties of inherent bias.

[0009] Thus, even though various system and methods of treatment prediction are known in the art, all or almost all of them suffer from various disadvantages. Therefore, there is still a need for systems and methods that help to more accurately predict drug treatment response of a cancer patient to an intended chemotherapy before commencing treatment.

Summary of The Invention

[0010] The inventor has discovered that a predictive model for treatment outcome for high-grade bladder cancer can be derived from a collection of models that were prepared using various machine learning algorithms trained on previously known high-grade bladder cancer omics information that was associated with treatment outcome. Most preferably, prediction accuracy is improved by identification of a model with high accuracy gain and selection of omics parameters and associated weighting from the identified model.

[0011] In one aspect of the inventive subject matter, the inventor contemplates a method of predicting treatment outcome for a patient having high-grade bladder cancer. In preferred aspects contemplated methods include a step of obtaining a plurality of omics data from the patient, and a further step of (a) using an accuracy gain metric to select at least a single model for prediction of the treatment outcome of high grade bladder cancer treatment or (b) selecting at least a single model on the basis of a previously determined accuracy gain metric for prediction of the treatment outcome of high grade bladder cancer treatment. Models may be selected from among a large number, for example, from among at least 10 trained models or from among at least 100 trained models or even more. In yet another step, an analysis engine then calculates a prediction outcome (*e.g.*, complete response to treatment, partial response to treatment, stable non-response to treatment, and progressive non-response to treatment) using the single model and the plurality of omics data from the patient.

[0012] Most typically, the omics data include whole genome differential objects, exome differential objects, SNP data, copy number data, RNA transcription data, protein expression data, and/or protein activity data, and it is further preferred that the accuracy gain metric may be an accuracy gain, an accuracy gain distribution, an area under curve metric, an R^2 metric, a p-value metric, a silhouette coefficient, and/or a confusion matrix. While not limiting the inventive subject matter, it is also contemplated that the accuracy gain metric of the single model is within the upper quartile of all models, or within the top 5% of all models, or wherein the accuracy gain metric of the single model exceeds all other models.

[0013] In further contemplated aspects, the single model may be generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

[0014] Most preferably, the step of calculating comprises a step of selecting features of the single model having minimum absolute predetermined weights (*e.g.*, within the top quartile of all weights in the single model). While numerous features may be suitable, it is contemplated that the step of calculating uses at least 10 distinct selected features in the single model. In particularly preferred methods for high-grade bladder cancer, the features of the single model are RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN. Moreover, it is contemplated that the RNA transcription values for the genes are calculated with respective factors, that the respective factors are weighted, and that (using absolute values), the weights are in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

[0015] Viewed from a different perspective, the inventors therefore also contemplate a method of predicting treatment outcome for a patient having high-grade bladder cancer. Such methods will preferably include a step of obtaining plurality of RNA transcription data of the patient, and a further step of calculating, by an analysis engine and using the plurality of

RNA transcription data of the patient, a treatment outcome score using a model. Most typically, the model uses RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

[0016] Most preferably, the plurality of RNA transcription data are obtained from polyA RNA, and/or the treatment outcome score is indicative of a complete response to treatment, a partial response to treatment, a stable non-response to treatment, or a progressive non-response to treatment. As already noted above it is contemplated that the model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear), and/or that the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

[0017] Consequently, the inventors also contemplate a method of predicting treatment outcome for a patient having high-grade bladder cancer. Especially preferred such methods include a step of obtaining plurality of RNA transcription data of the patient, wherein the RNA transcription values are values for at least two genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2,

LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN; and a further step of using the RNA transcription values in a model generated by a machine learning algorithm to so predict treatment outcome for the patient.

[0018] While not limiting to the inventive subject matter, it is typically preferred that the machine learning algorithm uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear). Moreover, it is contemplated that the RNA transcription values for the genes are calculated with respective factors, and that the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

[0019] Thus, the inventors also contemplate use of RNA transcription values for prediction of the treatment outcome of high grade bladder cancer treatment, wherein the prediction uses a single model obtained from a machine learning algorithm, and wherein the RNA transcription values are for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN. Typically, but not necessarily, the RNA transcription

values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

[0020] Various objects, features, aspects and advantages of the inventive subject matter will become more apparent from the following detailed description of preferred embodiments, along with the accompanying drawing figures in which like numerals represent like components.

Brief Description of The Drawing

[0021] Figure 1 is an exemplary table of features and feature weights derived from a model with high accuracy gain using TCGA high-grade bladder cancer data.

[0022] Figure 2 is an exemplary heat map of RNA transcription values from TCGA high-grade bladder cancer data for responders to drug treatment and non-responders.

Detailed Description

[0023] The inventive subject matter is directed to various computer systems and methods in which genomic information for a relatively large class of patients suffering from a particular neoplastic disease (*e.g.*, bladder cancer) is subjected to a relatively large number of machine learning algorithms to so identify a corresponding large number of predictive models. The predictive models are then analyzed for accuracy gain, and the model(s) with the highest accuracy gain will then be used to identify relevant omics factors for the prediction.

[0024] Thus, it should be especially appreciated that contemplated systems and methods are neither based on prediction optimization of a singular model nor based on identification of best correlations of selected omics parameters with a treatment prediction. Instead, it should be recognized that contemplated systems and methods rely on the identification of omics parameters and associated weighting factors that are derived from one or more

implementations of machine learning algorithms that result in trained models having a predetermined or minimum accuracy gain. Notably, the so identified omics parameters will typically have no statistically predictive power by themselves and as such would not be used in any omics based test system. However, where such identified omics parameters are used in the context of a trained model that has high accuracy gain, multiple omics parameters will provide a system with high predictive power, particularly when applied in the system using weighting factors associated with the trained model. Of course, it should also be appreciated that such model and omics parameters and weightings are unique to the particular training sets and/or type of outcome prediction, and that other diseases (*e.g.*, lung cancer) and/or outcome predictions (*e.g.*, survival time past 5 years) may lead to entirely different models, omics parameters, and weightings. Thus, the inventor is considered to have discovered weightings and/or trained models that have high predictive power associated with high-grade bladder cancer. In addition, treatment prediction can be validated from the *a priori* identified pathway(s) and/or pathway element(s), or identified pathways and/or pathway elements by *in silico* modulation using known pathway modeling system and methods to so help confirm treatment strategy predicted by the system.

[0025] It is therefore contemplated that the inventive subject matter is directed to various systems and methods in which genomic information and associated meta data for a relatively large class of patients suffering from a high-grade bladder cancer is subjected to multiple and distinct machine learning algorithms. In one preferred aspect of the inventive subject matter, RNA transcription values and associated meta data (*e.g.*, treatment outcome) are subject to training and validation splitting in a preprocessing step that then provides the data to different machine-learning packages for analysis.

[0026] It should be noted that the focus of the disclosed inventive subject matter is to enable construction or configuration of a computing device(s) to operate on vast quantities of digital data, beyond the capabilities of a human. Although the digital data can represent machine-trained computer models of omics data and treatment outcomes, it should be appreciated that the digital data is a representation of one or more digital models of such real-world items, not the actual items. Rather, by properly configuring or programming the devices as disclosed herein, through the instantiation of such digital models in the memory of the computing devices, the computing devices are able to manage the digital data or models in a manner that would be beyond the capability of a human. Furthermore, the computing devices lack *a*

priori capabilities without such configuration. In addition, it should be appreciated that the present inventive subject matter significantly improves/alleviates problems inherent to computational analysis of complex omics calculations.

[0027] Viewed from a different perspective, it should be appreciated that the present systems and methods in computer technology is being used to solve a problem inherent in computing models for omics data. Thus, without computers, the problem, and thus the present inventive subject matter, would not exist. More specifically, the disclosed approach results in one or more optimized trained models having greater accuracy gain than other trained models that are less capable, which results in less latency in generating predictive results based on patient data.

[0028] It should be noted that any language directed to a computer should be read to include any suitable combination of computing devices, including servers, interfaces, systems, databases, agents, peers, engines, controllers, modules, or other types of computing devices operating individually or collectively. One should appreciate the computing devices comprise a processor configured to execute software instructions stored on a tangible, non-transitory computer readable storage medium (*e.g.*, hard drive, FPGA, PLA, solid state drive, RAM, flash, ROM, etc.). The software instructions configure or otherwise program the computing device to provide the roles, responsibilities, or other functionality as discussed below with respect to the disclosed apparatus. Further, the disclosed technologies can be embodied as a computer program product that includes a non-transitory computer readable medium storing the software instructions that causes a processor to execute the disclosed steps associated with implementations of computer-based algorithms, processes, methods, or other instructions. In some embodiments, the various servers, systems, databases, or interfaces exchange data using standardized protocols or algorithms, possibly based on HTTP, HTTPS, AES, public-private key exchanges, web service APIs, known financial transaction protocols, or other electronic information exchanging methods. Data exchanges among devices can be conducted over a packet-switched network, the Internet, LAN, WAN, VPN, or other type of packet switched network, circuit switched network, and/or cell switched network.

[0029] As used in the description herein and throughout the claims that follow, when a system, engine, server, device, module, or other computing element is described as configured to perform or execute functions on data in a memory, the meaning of "configured

to" or "programmed to" is defined as one or more processors or cores of the computing element being programmed by a set of software instructions stored in the memory of the computing element to execute the set of functions or operate on target data or data objects stored in the memory.

[0030] For example, in the analysis of high-grade bladder cancer, a large number of genomic data with respective meta data from patients diagnosed with high-grade bladder cancer were processed to create training data sets that were then fed into a collection of model templates (i.e., software implementations of machine learning algorithms). Using the data sets and machine learning systems, corresponding trained models were created that were subsequently analyzed (and ranked) for accuracy gain as further described below. From the model with the highest accuracy gain, omics parameters and weighting factors for each of the parameters were extracted and used as the predictive model.

[0031] More specifically, and using the above approach, the inventor investigated by analysis of publicly available data (here: TCGA BLCA data) which of the high-grade bladder cancer patients in the data would respond to chemotherapy, which could at least potentially eliminate surgery. In this dataset, 116 drug treatment courses were tracked in 50 patients. Of these 116 treatments, 111 were chemotherapy agents, including Adriamycin, Avastin, Carboplatin, Cisplatin, Docetaxel, Doxorubicin, Etopside, Gemcitabine, Ifosfamide, Methotrexate, Paclitaxel and Vinblastine (or equivalent brand names for these drugs). Of these 111 chemotherapy treatments 78 had 'treatment best response' recorded. If a patient had a chemotherapy agent with Complete or Partial Response recorded, they were considered a "chemotherapy responder". If they had Clinical Progressive or Stable disease, they were considered a "chemotherapy non-responder". A total of 33 patients had a chemotherapy response recorded (15 non-responders and 18 responders). All 33 patients were confirmed to be high-grade bladder cancer patients using further TCGA clinical information.

[0032] These data were used to generate 72 candidate predictive models of which patients with high grade tumors could respond to chemotherapy. Each model was trained using k-fold cross-validation by splitting the data set into training sets and validation sets. Twenty-four predictive models were calculated for each of the available data sets using prediction model templates available via scikit-learn (scikit-learn developers, online scikit-learn.org), using various classifiers, including linear classifiers, NMF-based classifiers, graphical-based classifiers, tree-based classifiers, Bayesian-based classifiers, and net-based classifiers,

yielding 360 evaluation models. All of the so constructed evaluation models were then subjected to accuracy gain analysis to identify the model building process with the highest accuracy gain. In this example, accuracy gain was calculated by comparison of the correct prediction percentage using the validation data set against the percentage (frequency) of occurrence of the majority classifier (here: treatment is responsive). For example, where responsive treatment frequency is 60% in the known data set and where the model correctly predicts 85% of the treatment outcome as responsive, the accuracy gain is 25%. Notably, over all models constructed, the best model building process was 88% accurate in cross-validation testing folds (which was 33% better than majority) and used an elastic net classifier. The final fully-trained model that used the most accurate build process was selected from the 72 candidate models.

[0033] It should be appreciated that using such approach will rapidly generate a relatively large number of trained models. For example, where n algorithms are used with m types of input data sets using p fold cross validation, the overall number of trained models is n x m x p. All of the so constructed models were then subjected to accuracy gain analysis to identify the model with the highest accuracy gain. In this example, accuracy gain was calculated by comparison of the correct prediction percentage using the validation data set against the percentage (frequency) of occurrence of the majority classifier (here: treatment is responsive). For example, where responsive treatment frequency is 60% in the known data set and where the model correctly predicts 85% of the treatment outcome as responsive, the accuracy gain is 25%. Notably, over all models constructed, the best model was 88% accurate in cross-validation testing folds (which was 33% better than majority) and used an elastic net classifier.

[0034] In this context it must be appreciated that each type of model includes inherent biases or assumptions, which may influence how a resulting trained model would operate relative to other types of trained models, even when trained on identical data. Accordingly, different models will produce different predictions/accuracy gains when using the same training data set. Heretofore, in an attempt to improve prediction outcome, single machine learning algorithms were optimized to increase correct prediction on the same data set. However, due to inherent bias of the algorithms, such optimization will not necessarily increase accuracy (*i.e.*, accurate prediction capability against 'coin flip') in predictability. Such bias can be overcome by training numerous diverse models with different underlying principles and

classifiers on disease-specific data sets with associated metadata and by selecting from the so trained models those with desirable accuracy gain or robustness.

[0035] Once a desired model with high accuracy gain is selected, omic parameters with high relevance can then be selected from the model to produce a predictive model with improved accuracy of prediction. **Figure 1** exemplarily depicts a collection of genes encoding an RNA where the omics data from a patient are RNA transcription data (transcription strength). Here, the predictive model was built as described above from the *a priori* known TGCA data using RNA transcription levels from the gene expression panel. The best predictive model had 88% accuracy in cross-validation testing folds and the top 53 genes with highest weighting factor are shown. For example, the PCDHGA4 gene (Protocadherin Gamma Subfamily A, 4) had a weighting factor of -121543.6202 with respect to the RNA expression, with further genes and weighting factors listed below the PCDHGA4 gene. It should be appreciated that multiple, different types of data beyond RNA transcription data were also used to create trained models. The inventor discovered that using the RNA transcription data as training data resulting in the best models (i.e., models having the highest accuracy gain) relative to other trained models that were trained on other types of omic data (e.g., WGS, SNP copy number, proteomics, etc.).

[0036] **Figure 2** exemplarily illustrates a heat map for the actual patient data where each row in the map corresponds to a single patient, and each column to a specific gene (here, the genes listed in the graph of Figure 1. As can also be seen from the heat map, the patient data are grouped into responders (categorized in CR: complete response and PR: partial response) and non-responders (categorized in Prog: with disease progression and Stable: without disease progression). Color depth/grayscale value corresponds to measured transcription level and is expressed as color/gray scale value between -1.8 and 1.8. Taken with the weighting factors of Figure 1, the final predictive score for each patient is the sum of the expression value of Figure 2 for each gene multiplied by the weighting factor. Any final predictive score above zero (red/grey with + symbol) is indicative of likely treatment response, while a final predictive score below zero (blue/grey with - symbol) is indicative of a likely lack of treatment response. As can be taken from the 'topmodel signature' (final predictive score), only one false positive result was present in the 'Responders' category (top row in Responders category) while the Non-Responders had two false negative results (bottom row in Prog category, bottom row in Stable category).

[0037] Moreover, with further reference to the heat map of Figure 2, it should be appreciated that the statistical significance of each of the RNA transcription data would by itself not be sufficient for an accurate prediction as shown in the bar graph at the bottom portion of the map. Here the bars represent signed t-test values between the results of a responder group and the non-responder group that were corrected for multiple hypothesis testing using Bonferroni correction. As is readily apparent, only a limited set of data exhibited statistically significant differences between responders and non-responders as is shown in the black bars (e.g., DDI2, AGAP1, etc.) and white bar (RPL39). However, when at least some of the individual results are taken together (particularly in combination with the calculated weighting), the predictive power of the model will outperform most, if not all competing other models.

[0038] Moreover, it should also be appreciated that using a pathway modeling algorithm (see *e.g.*, WO 2011/139345, WO 2013/062505, WO 2014/059036, and WO 2014/193982) patient data can be used to validate and/or simulate treatment before the patient undergoes actual treatment, and such validation can then be reassessed using the best models for high-grade bladder cancer. For example, highly weighted RNA transcription can be clamped off *in silico* in the pathway modeling system, and activities are re-inferred, which in effect simulates *in silico* the anticipated effect of a drug intervention *in vivo*. The prediction model can then be used to reassess the newly inferred post-intervention data.

[0039] In further contemplated aspects of the inventive subject matter it should be recognized that while the example above used RNA transcription data, one or more other (or additional) omics data are also suitable for use in conjunction with the teachings herein. For example, suitable alternative or additional omics data include whole genome differential object data, exome differential object data, SNP data, copy number data, protein expression data, and/or protein activity data. Likewise, meta data associated with the omics data need not be limited to treatment outcome, but may include a large number of alternative patient or care-relevant metrics. For example, contemplated metadata may include treatment cost, likelihood of resistance, likelihood of metastatic disease, 5-year survival, suitability for immunotherapy, patient demographic information, etc.

[0040] Similarly, it should be noted that the number of models created is not limiting to the inventive subject matter and that (in general) higher numbers of models are preferred. Such models are preferably based on multiple and distinct machine learning algorithms, and it should be appreciated that all known machine learning algorithms are deemed suitable for use

herein. For example, contemplated classifiers include one or more of a linear classifier, an NMF-based classifier, a graphical-based classifier, a tree-based classifier, a Bayesian-based classifier, a rules-based classifier, a net-based classifier, and a kNN classifier. However, especially preferred algorithms will include those that use a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear). Beyond the above classifiers, additional suitable algorithms include various forms of neural networks (e.g., artificial neural networks, convolution neural networks, etc.), binary decision trees, or other types of learning. Sources for such algorithms are readily available via TensorFlow (see URL www.tensorflow.com), OpenAI (see URL www.openai.com), and Baidu (see URL research.baidu.com/warp-ctc). Thus, the inventor contemplates that at least 5, at least 10, at least 20, at least 50, at least 100, at least 500, at least 1,000, at least 5,000, or at least 10,000 trained models are created. Depending on the number of possible training data sets, the number of validations, and the number of types of algorithms, the number of resulting trained models could even exceed 1,000,000 trained models.

[0041] Once the models are created, model quality is assessed and most preferably models are retained that have a prediction power that exceeds random selection. Viewed from a different perspective, models will be assessed on their gain in accuracy. There are numerous manners of assessing accuracy, and the particular choice may depend at least in part on the algorithm used. For example, suitable metrics include an accuracy value, an accuracy gain, a performance metric, or other measure of the corresponding model. Additional example metrics include an area under curve metric, an R^2, a p-value metric, a silhouette coefficient, a confusion matrix, or other metric that relates to the nature of the model or its corresponding model template.

[0042] For example, accuracy of a model can be derived through use of known data sets and corresponding known clinical outcome data. Thus, for a specific model template a number of evaluation models can be built that are both trained and validated against the input known data sets (*e.g.*, k-fold cross validation). For example, a trained model can be trained based on

80% of the input data. Once the evaluation model has been trained, the remaining 20% of the genomic data can be run through the evaluation model to see if it generates prediction data similar to or closet to the remaining 20% of the known clinical outcome data. The accuracy of the trained evaluation model is then considered to be the ratio of the number of correct predictions to the total number of outcomes.

[0043] For example, a RNA transcription data set/clinical outcome data set represents a cohort of 500 patients. The data sets can then be partitioned into one or more groups of evaluation training sets, *e.g.*, containing 400 patient samples. Models are then created based on the 400 patient samples, and the so trained models are validated by executing the model on the remaining 100 patients' transcription data set to generate 100 prediction outcomes. The 100 prediction outcomes are then compared to the actual 100 outcomes from the patient data in the clinical outcome data set. The accuracy of the trained model is the number of correct prediction outcomes relative to the total number of outcomes. If, out of the 100 prediction outcomes, the trained evaluation model generates 85 correct outcomes that match the actual or known clinical outcomes from the patient data, then the accuracy of the trained evaluation model is considered 85%. Alternatively, where the observed outcome (*e.g.*, drug responder) has a frequency of 60% in the meta data of the RNA transcription data set, and where the model generates 85 correct outcomes out of the 100 prediction outcomes, the accuracy gain would be 25% (*i.e.*, 25% above randomly observed results; predicted event occurs at 60%, correct prediction at 85%, accuracy gain is 25%)

[0044] Depending on the number of models/ accuracy distribution, it should be appreciated that the model used for prediction may be selected as the top model (having highest accuracy gain, or highest accuracy score, etc.), or as being in the top n-tile (tertile, quartile, quintile, etc.), or as being in the top n% of all models (top 5%, top 10%, etc.). Thus suitable models have may have an accuracy gain metric that exceeds all other models.

[0045] With respect to the single model, it should be appreciated that the prediction based on the top (or other selected single) model may be based on all of the omics data that were part of the input data (*i.e.*, uses all RNA expression levels used for training the models) or only a fraction of the omics data. For example, where only fractions of the omics data are used for final prediction, the omics data with the highest or minimum absolute predetermined weight (*e.g.*, top quartile of all weights in the single model) in the model will be generally preferred as is shown in the selected features (genes) of Figure 1. Thus, suitable models will employ at

least 5, or at least 10, or at least 20, or at least 50, or at least 100 features in the prediction. Moreover, it should also be appreciated that where features are identified that have substantial statistical significance between the treatment outcomes, these features may be used, preferably in combination, in an gene expression array rather than in a predictive algorithm (*e.g.*, significant features in Figure 2).

[0046] It should be apparent to those skilled in the art that many more modifications besides those already described are possible without departing from the inventive concepts herein. The inventive subject matter, therefore, is not to be restricted except in the scope of the appended claims. Moreover, in interpreting both the specification and the claims, all terms should be interpreted in the broadest possible manner consistent with the context. In particular, the terms "comprises" and "comprising" should be interpreted as referring to elements, components, or steps in a non-exclusive manner, indicating that the referenced elements, components, or steps may be present, or utilized, or combined with other elements, components, or steps that are not expressly referenced. Where the specification claims refers to at least one of something selected from the group consisting of A, B, C .... and N, the text should be interpreted as requiring only one element from the group, not A plus N, or B plus N, etc. Furthermore, and as used in the description herein and throughout the claims that follow, the meaning of "a," "an," and "the" includes plural reference unless the context clearly dictates otherwise. Also, as used in the description herein, the meaning of "in" includes "in" and "on" unless the context clearly dictates otherwise.

CLAIMS

What is claimed is:

1. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:
   obtaining a plurality of omics data from the patient;
   using an accuracy gain metric to select a single model for prediction of the treatment outcome of high grade bladder cancer treatment, or selecting a single model on the basis of a previously determined accuracy gain metric for prediction of the treatment outcome of high grade bladder cancer treatment;
   calculating, by an analysis engine, a prediction outcome using the single model and the plurality of omics data from the patient.

2. The method of claim 1 wherein the omics data are selected from the group consisting of whole genome differential objects, exome differential objects, SNP data, copy number data, RNA transcription data, protein expression data, and protein activity data.

3. The method of any one of the preceding claims wherein the accuracy gain metric is selected form the group consisting of accuracy gain, accuracy gain distribution, an area under curve metric, an R^2, a p-value metric, a silhouette coefficient, and a confusion matrix.

4. The method of any one of the preceding claims wherein the single model is selected from among at least 100 models.

5. The method of any one of the preceding claims wherein the accuracy gain metric of the single model is within the upper quartile of all models.

6. The method of any one of the preceding claims wherein the accuracy gain metric of the single model is within the top 5% of all models.

7. The method of any one of the preceding claims wherein the accuracy gain metric of the single model exceeds all other models.

8. The method of any one of the preceding claims wherein the prediction outcome is selected from the group consisting of complete response to treatment, partial response to treatment, stable non-response to treatment, and progressive non-response to treatment.

9. The method of any one of the preceding claims wherein the single model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

10. The method of any one of the preceding claims wherein the step of calculating comprises a step of selecting features of the single model having minimum absolute predetermined weights.

11. The method of claim 10 wherein the minimum absolute predetermined weights are within the top quartile of all weights in the single model.

12. The method of any one of the preceding claims wherein the step of calculating uses at least 10 distinct selected features in the single model.

13. The method of claim 10 wherein the features are RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

14. The method of claim 13 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1,

SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

15. The method of claim 1 wherein the accuracy gain metric is selected form the group consisting of accuracy gain, accuracy gain distribution, an area under curve metric, an R^2, a p-value metric, a silhouette coefficient, and a confusion matrix.

16. The method of claim 1 wherein the single model is selected from among at least 100 models.

17. The method of claim 1 wherein the accuracy gain metric of the single model is within the upper quartile of all models.

18. The method of claim 1 wherein the accuracy gain metric of the single model is within the top 5% of all models.

19. The method of claim 1 wherein the accuracy gain metric of the single model exceeds all other models.

20. The method of claim 1 wherein the prediction outcome is selected from the group consisting of complete response to treatment, partial response to treatment, stable non-response to treatment, and progressive non-response to treatment.

21. The method of claim 1 wherein the single model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

22. The method of claim 1 wherein the step of calculating comprises a step of selecting features of the single model having minimum absolute predetermined weights.

23. The method of claim 22 wherein the minimum absolute predetermined weights are within the top quartile of all weights in the single model.

24. The method of claim 1 wherein the step of calculating uses at least 10 distinct selected features in the single model.

25. The method of claim 22 wherein the features are RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

26. The method of claim 25 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

27. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:

obtaining plurality of RNA transcription data of the patient; and

calculating, by an analysis engine and using the plurality of RNA transcription data of the patient, a treatment outcome score using a model;

wherein the model uses RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

28. The method of claim 27 wherein the plurality of RNA transcription data are obtained from polyA RNA.

29. The method of claim 27 or 28 wherein the treatment outcome score is indicative of a complete response to treatment, a partial response to treatment, a stable non-response to treatment, or a progressive non-response to treatment.

30. The method of any one of claims 27 to 29 wherein the model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

31. The method of any one of claims 27 to 30 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

32. The method of claim 27 wherein the treatment outcome score is indicative of a complete response to treatment, a partial response to treatment, a stable non-response to treatment, or a progressive non-response to treatment.

33. The method of claim 27 wherein the model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA

JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

34. The method of claim 27 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

35. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:

obtaining plurality of RNA transcription data of the patient;

wherein the RNA transcription values are values for at least two genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN; and

using the RNA transcription values in a model generated by a machine learning algorithm to so predict treatment outcome for the patient.

36. The method of claim 35 wherein the machine learning algorithm uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

37. The method of claim 36 wherein the machine learning algorithm uses a glmnet elastic nets (sparse linear) classifier.

38. The method of claim 35 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

39. Use of RNA transcription values for prediction of the treatment outcome of high grade bladder cancer treatment, wherein the prediction uses a single model obtained from a machine learning algorithm, and wherein the RNA transcription values are for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

40. The use of claim 39 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

41. The use of claim 39 wherein the machine learning algorithm uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

42. The use of claim 41 wherein the machine learning algorithm uses a glmnet elastic nets (sparse linear).

AMENDED CLAIMS
received by the International Bureau on 01 July 2016 (01.07.2016)

What is claimed is:

1. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:

obtaining a plurality of omics data from the patient;

generating a plurality of models using a plurality of machine learning algorithms and *a priori* omics data;

using an accuracy gain metric to select a single model from the plurality of models for prediction of the treatment outcome of high grade bladder cancer treatment, or selecting a single model from the plurality of models on the basis of a previously determined accuracy gain metric for prediction of the treatment outcome of high grade bladder cancer treatment; and

calculating, by an analysis engine, a prediction outcome using the single model and the plurality of omics data from the patient.

2. The method of claim 1 wherein the omics data are selected from the group consisting of whole genome differential objects, exome differential objects, SNP data, copy number data, RNA transcription data, protein expression data, and protein activity data.

3. The method of any one of the preceding claims wherein the accuracy gain metric is selected form the group consisting of accuracy gain, accuracy gain distribution, an area under curve metric, an R^2, a p-value metric, a silhouette coefficient, and a confusion matrix.

4. The method of any one of the preceding claims wherein the single model is selected from among at least 100 models.

5. The method of any one of the preceding claims wherein the accuracy gain metric of the single model is within the upper quartile of all models.

6. The method of any one of the preceding claims wherein the accuracy gain metric of the single model is within the top 5% of all models.

7. The method of any one of the preceding claims wherein the accuracy gain metric of the single model exceeds all other models.

8. The method of any one of the preceding claims wherein the prediction outcome is selected from the group consisting of complete response to treatment, partial response to treatment, stable non-response to treatment, and progressive non-response to treatment.

9. The method of any one of the preceding claims wherein the single model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

10. The method of any one of the preceding claims wherein the step of calculating comprises a step of selecting features of the single model having minimum absolute predetermined weights.

11. The method of claim 10 wherein the minimum absolute predetermined weights are within the top quartile of all weights in the single model.

12. The method of any one of the preceding claims wherein the step of calculating uses at least 10 distinct selected features in the single model.

13. The method of claim 10 wherein the features are RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

14. The method of claim 13 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129,

SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

15. The method of claim 1 wherein the accuracy gain metric is selected form the group consisting of accuracy gain, accuracy gain distribution, an area under curve metric, an R^2, a p-value metric, a silhouette coefficient, and a confusion matrix.

16. The method of claim 1 wherein the single model is selected from among at least 100 models.

17. The method of claim 1 wherein the accuracy gain metric of the single model is within an upper quartile of all models.

18. The method of claim 1 wherein the accuracy gain metric of the single model is within a top 5% of all models.

19. The method of claim 1 wherein the accuracy gain metric of the single model exceeds all other models.

20. The method of claim 1 wherein the prediction outcome is selected from the group consisting of complete response to treatment, partial response to treatment, stable non-response to treatment, and progressive non-response to treatment.

21. The method of claim 1 wherein the single model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

22. The method of claim 1 wherein the step of calculating comprises a step of selecting features of the single model having minimum absolute predetermined weights.

23. The method of claim 22 wherein the minimum absolute predetermined weights are within a top quartile of all weights in the single model.

24. The method of claim 1 wherein the step of calculating uses at least 10 distinct selected features in the single model.

25. The method of claim 22 wherein the features are RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

26. The method of claim 25 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

27. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:

obtaining plurality of RNA transcription data of the patient; and

calculating, by an analysis engine and using the plurality of RNA transcription data of the patient, a treatment outcome score using a model;

wherein the model uses RNA transcription values for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35,

TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

28. The method of claim 27 wherein the plurality of RNA transcription data are obtained from polyA RNA.

29. The method of claim 27 or 28 wherein the treatment outcome score is indicative of a complete response to treatment, a partial response to treatment, a stable non-response to treatment, or a progressive non-response to treatment.

30. The method of any one of claims 27 to 29 wherein the model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

31. The method of any one of claims 27 to 30 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

32. The method of claim 27 wherein the treatment outcome score is indicative of a complete response to treatment, a partial response to treatment, a stable non-response to treatment, or a progressive non-response to treatment.

33. The method of claim 27 wherein the model was generated using a machine learning algorithm that uses a classifier selected form the group consisting of NMFpredictor

(linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

34. The method of claim 27 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

35. A method of predicting treatment outcome for a patient having high-grade bladder cancer, comprising:

obtaining a plurality of RNA transcription values of the patient;

wherein the RNA transcription values are values for at least two genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN; and

using the RNA transcription values in a model generated by a machine learning algorithm to so predict treatment outcome for the patient.

36. The method of claim 35 wherein the machine learning algorithm uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based),

WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

37. The method of claim 36 wherein the machine learning algorithm uses a glmnet elastic nets (sparse linear) classifier.

38. The method of claim 35 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

39. Use of a plurality of RNA transcription values for prediction of a treatment outcome of high grade bladder cancer treatment, wherein the prediction uses a single model obtained from a plurality of machine learning algorithms, and wherein the RNA transcription values are for genes selected from the group consisting of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1, SEC24D, FLNB, PATL1, HDLBP, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

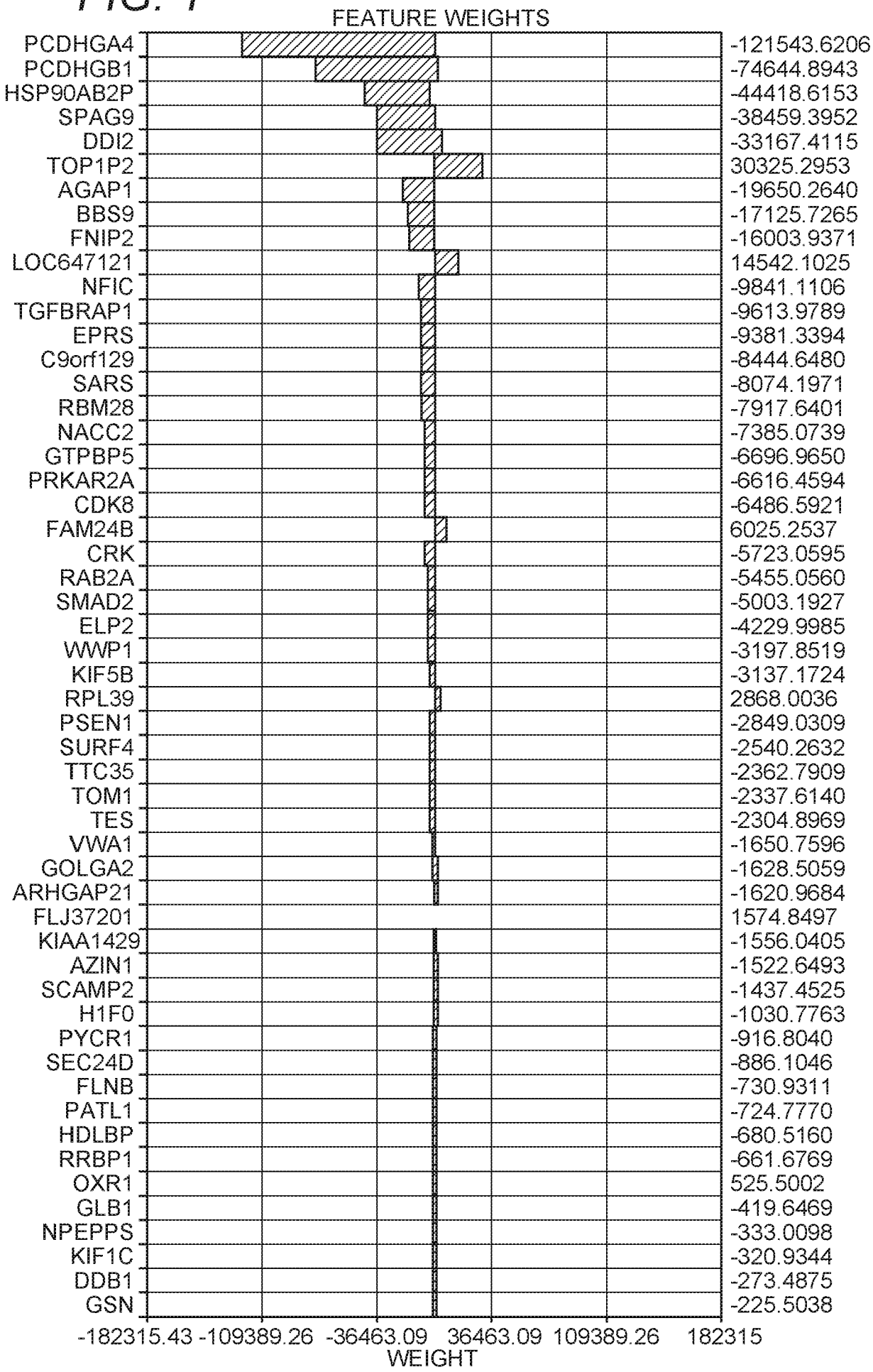
40. The use of claim 39 wherein the RNA transcription values for the genes are calculated with respective factors, and wherein the respective factors are weighted, using absolute values, in the order of PCDHGA4, PCDHGB1, HSP90AB2P, SPAG9, DDI2, TOP1P2, AGAP1, BBS9, FNIP2, LOC647121, NFIC, TGFBRAP1, EPRS, C9orf129, SARS, RBM28, NACC2, GTPBP5, PRKAR2A, CDK8, FAM24B, CRK, RAB2A, SMAD2, ELP2, WWP1, KIF5B, RPL39, PSEN1, SURF4, TTC35, TOM1, TES, VWA1, GOLGA2, ARHGAP21, FLJ37201, KIAA1429, AZIN1, SCAMP2, H1F0, PYCR1,

SEC24D, FLNB, PATL1, HDLBP, RRBP1, OXR1, GLB1, NPEPPS, KIF1C, DDB1, and GSN.

41. The use of claim 39 wherein the machine learning algorithm uses a classifier selected form the group consisting of NMFpredictor (linear), SVMlight (linear), SVMlight first order polynomial kernel (degree-d polynomial), SVMlight second order polynomial kernel (degree-d polynomial), WEKA SMO (linear), WEKA j48 trees (trees-based), WEKA hyper pipes (distribution-based), WEKA random forests (trees-based), WEKA naive Bayes (probabilistic/bayes), WEKA JRip (rules-based), glmnet lasso (sparse linear), glmnet ridge regression (sparse linear), and glmnet elastic nets (sparse linear).

42. The use of claim 41 wherein the machine learning algorithm uses a glmnet elastic nets (sparse linear).

FIG. 1


FEATURE WEIGHTS
PCDHGA4 -121543.6206
PCDHGB1 -74644.8943
HSP90AB2P -44418.6153
SPAG9 -38459.3952
DDI2 -33167.4115
TOP1P2 30325.2953
AGAP1 -19650.2640
BBS9 -17125.7265
FNIP2 -16003.9371
LOC647121 14542.1025
NFIC -9841.1106
TGFBRAP1 -9613.9789
EPRS -9381.3394
C9orf129 -8444.6480
SARS -8074.1971
RBM28 -7917.6401
NACC2 -7385.0739
GTPBP5 -6696.9650
PRKAR2A -6616.4594
CDK8 -6486.5921
FAM24B 6025.2537
CRK -5723.0595
RAB2A -5455.0560
SMAD2 -5003.1927
ELP2 -4229.9985
WWP1 -3197.8519
KIF5B -3137.1724
RPL39 2868.0036
PSEN1 -2849.0309
SURF4 -2540.2632
TTC35 -2362.7909
TOM1 -2337.6140
TES -2304.8969
VWA1 -1650.7596
GOLGA2 -1628.5059
ARHGAP21 -1620.9684
FLJ37201 1574.8497
KIAA1429 -1556.0405
AZIN1 -1522.6493
SCAMP2 -1437.4525
H1F0 -1030.7763
PYCR1 -916.8040
SEC24D -886.1046
FLNB -730.9311
PATL1 -724.7770
HDLBP -680.5160
RRBP1 -661.6769
OXR1 525.5002
GLB1 -419.6469
NPEPPS -333.0098
KIF1C -320.9344
DDB1 -273.4875
GSN -225.5038
-182315.43 -109389.26 -36463.09 36463.09 109389.26 182315
WEIGHT

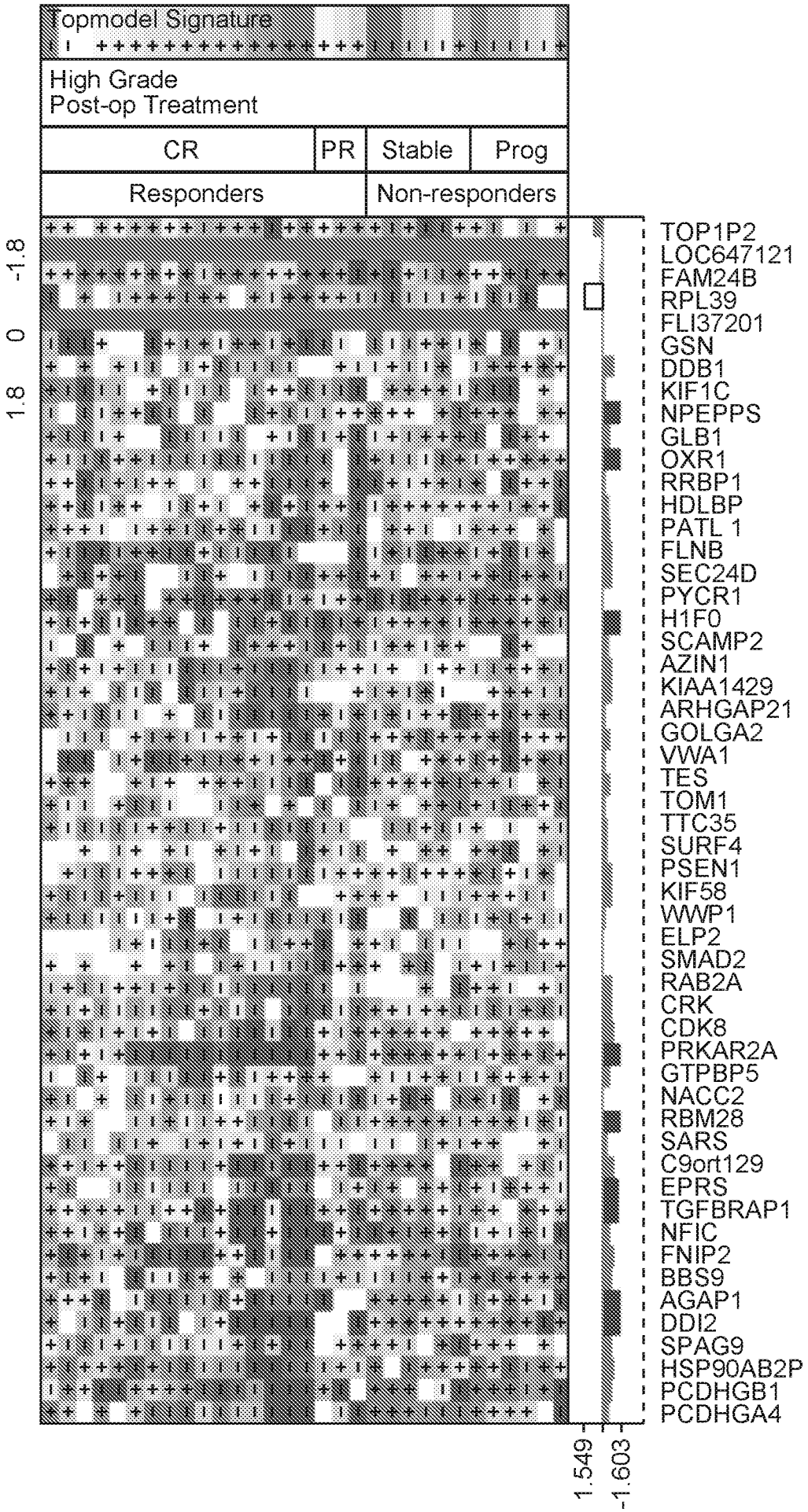

Topmodel Signature
High Grade
Post-op Treatment
CR
PR
Stable
Prog
Responders
Non-responders
1.8
0
-1.8
TOP1P2
LOC647121
FAM24B
RPL39
FLI37201
GSN
DDB1
KIF1C
NPEPPS
GLB1
OXR1
RRBP1
HDLBP
PATL 1
FLNB
SEC24D
PYCR1
H1F0
SCAMP2
AZIN1
KIAA1429
ARHGAP21
GOLGA2
VWA1
TES
TOM1
TTC35
SURF4
PSEN1
KIF58
WWP1
ELP2
SMAD2
RAB2A
CRK
CDK8
PRKAR2A
GTPBP5
NACC2
RBM28
SARS
C9ort129
EPRS
TGFBRAP1
NFIC
FNIP2
BBS9
AGAP1
DDI2
SPAG9
HSP90AB2P
PCDHGB1
PCDHGA4
1.549
-1.603

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/013959

A. CLASSIFICATION OF SUBJECT MATTER

G06F 19/24(2011.01)i, G06F 19/22(2011.01)i, G06F 19/18(2011.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F 19/24; G06F 19/00; C40B 40/06; C07H 21/00; C12Q 1/68; C12M 1/34; G06F 19/22; G06F 19/18

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & keywords: high grade bladder cancer, treatment outcome, omics data, accuracy gain metric, single model, machine learning

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2014-043803 A1 (GENOMEDX BIOSCIENCES, INC.) 27 March 2014 See paragraph [0019] and claims 1, 4, 8-9.	1-3,15-20,22-24
Y		21,25-26
A		27-29,32-42
Y	WO 2013-090620 A1 (GENOMEDX BIOSCIENCES, INC.) 20 June 2013 See paragraphs [00309], [00313], [00531]-[00556] and claim 115.	21,33,35-42
X	WO 2012-009382 A2 (THE REGENTS OF THE UNIVERSITY OF COLORADO) 19 January 2012 See claims 1, 11.	27-29,32,34
Y		25-26,33,35-42
A	WO 2005-008213 A2 (GENOMIC HEALTH, INC.) 27 January 2005 See claims 17-18, 24.	1-3,15-29,32-42
A	US 2007-0128636 A1 (BAKER et al.) 07 June 2007 See claims 1-25.	1-3,15-29,32-42

☐ 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	Date of mailing of the international search report
04 May 2016 (04.05.2016)	**04 May 2016 (04.05.2016)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea
Facsimile No. +82-42-481-8578

Authorized officer
KIM, Seung Beom
Telephone No. +82-42-481-3371

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/013959

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 11,13-14
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
Claims 11,13-14 are referring to the multiple dependent claims which do not comply with PCT Rule 6.4(a).

3. ☒ Claims Nos.: 4-10,12,30-31
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required addtional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of any additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2016/013959

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014-043803 A1	27/03/2014	AU 2013-317645 A1	09/04/2015
		CA 2885202 A1	27/03/2014
		EP 2898125 A1	29/07/2015
		US 2014-0080731 A1	20/03/2014
WO 2013-090620 A1	20/06/2013	AU 2012-352153 A1	26/06/2014
		CA 2858581 A1	20/06/2013
		EP 2791359 A1	22/10/2014
		EP 2791359 A4	07/10/2015
		US 2015-0011401 A1	08/01/2015
WO 2012-009382 A2	19/01/2012	WO 2012-009382 A3	19/04/2012
WO 2005-008213 A2	27/01/2005	AU 2004-258085 A1	27/01/2005
		AU 2004-258085 B2	27/05/2010
		CA 2531967 A1	27/01/2005
		CA 2531967 C	16/07/2013
		EP 1644858 A2	12/04/2006
		EP 1644858 A4	13/05/2009
		JP 2007-527220 A	27/09/2007
		JP 4906505 B2	28/03/2012
		US 2005-0048542 A1	03/03/2005
		US 2009-0280490 A1	12/11/2009
		US 7526387 B2	28/04/2009
		US 7939261 B2	10/05/2011
		WO 2005-008213 A2	24/03/2005
US 2007-0128636 A1	07/06/2007	WO 2007-067500 A2	14/06/2007
		WO 2007-067500 A3	20/03/2008