



- (51) **International Patent Classification:**
C12Q 1/68 (2006.01) *G06F 19/20* (2011.01)
- (21) **International Application Number:**
PCT/US2015/028343
- (22) **International Filing Date:**
29 April 2015 (29.04.2015)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
62/029,868 28 July 2014 (28.07.2014) US
- (71) **Applicant:** THE REGENTS OF THE UNIVERSITY OF CALIFORNIA [US/US]; Office of the President, 1111 Franklin Street, 5th Floor, Oakland, CA 94607-5200 (US).
- (72) **Inventors:** IDEKER, Trey; Pharmaceutical Sciences Building, Room 4230, 9500 Gilman Drive, La Jolla, CA 92093 (US). HOFREE, Matan; Pharmaceutical Sciences Building, Room 4230, 9500 Gilman Drive, La Jolla, CA 92093 (US). SHEN, John, Paul; Pharmaceutical Sciences Building, Room 4230, 9500 Gilman Drive, La Jolla, CA 92093 (US).
- (74) **Agent:** ADRIANO, Sarah, B.; Adriano & Associates, 690 East Green Street, Suite 300, Pasadena, CA 91101 (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).

Declarations under Rule 4.17:

— of inventorship (Rule 4.17(iv))

Published:

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

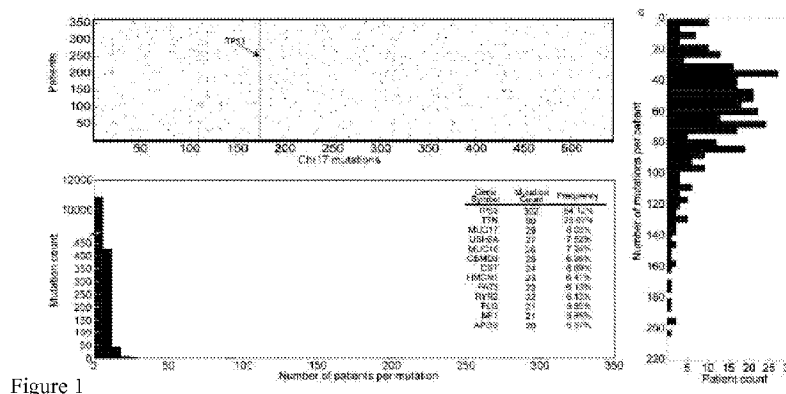
(54) **Title:** NETWORK BASED STRATIFICATION OF TUMOR MUTATIONS

Figure 1

(57) **Abstract:** The embodiments provide a method of for stratification of cancer into one or more informative subtypes of a subject in need thereof. The embodiments, further provide assigning a subject in need into one or more informative subtypes, including assigning a subject in need into an informative subtype wherein the subtype is an ovarian cancer subtype.

NETWORK BASED STRATIFICATION OF TUMOR MUTATIONS

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to U.S. Provisional Patent Application 61/865,510 filed on August 13, 2013 and entitled “NETWORK BASED STRATIFICATION OF TUMOR MUTATIONS,” the entirety of which is hereby incorporated by reference herein.

[0002] This invention was made with government support under Grant No. 5 R01 GM070743 awarded by NIH. The government has certain rights in the invention.

[0003] Throughout this application various publications are referenced. The disclosures of these publications in their entireties are hereby incorporated by reference into this application in order to more fully describe the state of the art to which this invention pertains.

FIELD OF THE INVENTION

[0004] Many forms of cancer have multiple subtypes with different causes and clinical outcomes. Somatic tumor genomes can provide a rich new source of data for uncovering these subtypes but it can be difficult to compare tumors, as two tumors rarely share the same mutations. A new method called Network Based Stratification (NBS) can be implemented that integrates somatic tumor genomes with gene networks to uncover tumor subtypes. Methods are also presented for assigning new patients to subtypes uncovered using NBS. Further, methods are presented for predicting patient survival or response to treatment.

BACKGROUND OF THE INVENTION

[0005] One of the fundamental goals of cancer informatics and bioinformatics is tumor stratification. Previous attempts to stratify tumors with molecular profiles have used mRNA expression data, which have resulted in the discovery of informative subtypes in diseases, for example, glioblastoma and breast cancer. The use of microarray techniques have been used to define several subtypes according to gene expression signatures. However, in many cancers, signatures or clinical correlations that were identified in one study could not be reproduced in many other studies as well. For example, in TCGA cohorts including Colorectal Adenocarcinoma and Small-Cell Lung Cancer, subtypes derived from expression profiles did not correlate with any clinical phenotype including patient survival and response to chemotherapy. These results can be due to limitations of expression-based analysis, such as issues with RNA sample quality, lack of reproducibility between biological replicates, and ample opportunities for

over-fitting of data. As such there has been a pressing need for improved stratification methods that can be used to correlate signatures with clinical outcomes.

SUMMARY OF THE INVENTION

[0006] The invention provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining nucleic acid sequence information from the subject, determining mutational status from the nucleic acid sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0007] The invention also provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining protein sequence information from the subject, determining mutational status from the protein sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0008] The invention further provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining epigenetic modification information for genomic DNA from the subject, determining epigenetic modification status from the epigenetic modification information so obtained, transforming the epigenetic status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0009] The invention further provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining RNA modification information for RNAs from the subject, determining RNA modification status from the RNA modification information so obtained, transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0010] The invention also provides methods for diagnosing a subject in need thereof with one or more informative subtypes of a cancer or tumor. The method comprises obtaining post-

translational modification information for proteins from the subject, determining post-translational modification status from the post-translational modification information so obtained, transforming the post-translational modification status into a transformed profile of the subject based on a reference molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile of the subject with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0011] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining nucleic acid sequence information from subjects with a cancer or tumor, determining mutational status for each subject from the nucleic acid sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0012] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining protein sequence information from subjects with a cancer or tumor, determining mutational status for each subject from the protein sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0013] The invention further provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining epigenetic modification information from subjects with a cancer or tumor, determining epigenetic modification status for each subject from the epigenetic modification information so obtained, transforming the epigenetic modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0014] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining RNA modification information from subjects with a cancer or tumor, determining RNA modification status for each subject from the RNA modification information so obtained, transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0015] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining post-translational modification information from subjects with a cancer or tumor, determining post-translational modification status for each subject from the post-translational modification information so obtained, transforming the post-translational modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles obtained into one or more clusters so as to obtain one or more subtypes.

[0016] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor; determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes; obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered and the subject of interest having cancer or tumor by using a supervised learning approach to derive a subtype classifier based on profiles from the subjects and their assignment to subtypes; and comparing the subtype classifier so derived to assign the subject of interest to a cancer or tumor subtype.

[0017] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor; determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor; obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered and the subject of interest having cancer or tumor; and applying a nearest shrunken centroid approach (Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002)), a supervised learning approach based on decision tree classifiers (Additive logistic regression: a statistical view of boosting. *Annals of Statistics* **28**(2), 2000. 337–407; "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences* **55**. 1997) or another supervised or unsupervised learning approach (Hastie, Trevor, Robert Tibshirani, Friedman, Jerome (2009). *The Elements of*

Statistical Learning: Data mining, Inference, and Prediction. New York: Springer. pp. 485–586; Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258) or combination of such approaches to assign the subject of interest to a cancer or tumor subtype.

[0018] The invention further provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from subjects with a cancer or tumor; determining mutational status for each subject from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes; characterizing the subjects grouped into one or more informative subtypes by determining status or profile of one or more measurable or quantifiable biological parameter(s) or feature(s); characterizing the subject of interest by determining status or profile of one or more measurable or quantifiable biological parameter(s); and assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on status or profile(s) of the subjects grouped into one or more informative subtypes and the status or profile of the subject of interest.

[0019] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining biological profiles of subjects grouped into one or more informative subtypes, obtaining biological profile of the subject of interest, and assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on biological profile(s) of the subjects grouped into one or more informative subtypes and the biological profile of the subject of interest.

[0020] The invention also provides methods for increasing efficiency of a bioinformatics process for network-based stratification of tumor or cancer. The method comprises obtaining a biological sample from a subject with tumor or cancer; selecting a set of genes for which nucleic acid sequence is to be determined; determining nucleic acid sequence for protein coding sequences in the set of genes selected; projecting mutations found within sequence onto a network; propagating the mutations in the network; and clustering the mutations so propagated so as to divide biological samples from subjects with tumor or cancer into subtypes, wherein, the set of genes so selected excludes whole exome or genome sequencing.

BRIEF DESCRIPTION OF THE DRAWINGS

[0021] **Figure 1** illustrates an overview of the somatic mutation landscape of a TCGA ovarian cancer cohort. As shown in panel A of **Figure 1**, somatic mutations are shown along the length of chromosome 17. In panel B of **Figure 1**, a histogram is illustrated summing the frequency of mutations per gene for the entire exome. In panel C of **Figure 1**, a histogram is illustrated that sums the frequency of genes mutated per patient in the cohort.

[0022] **Figure 2** illustrates a flowchart of the approach of network-based stratification.

[0023] **Figure 3** illustrates smoothing of patient somatic mutation profiles over a molecular interaction network.

[0024] **Figure 4** illustrates clustering mutation profiles using Non-negative Matrix Factorization (NMF) regularized by a network.

[0025] **Figure 5** illustrates the final tumor subtypes obtained from the consensus assignments of each tumor after several applications of the procedures shown in **Figures 3-4**.

[0026] **Figure 6** illustrates TCGA somatic mutations for ovarian cancer (top left) that are combined with the STRING human protein interaction network (bottom left) to generate simulated mutation datasets embedded with known network structure (center right).

[0027] **Figure 7** illustrates the accuracy with which NBS clusters recover simulated subtype assignments, evaluated with and without network smoothing and using NMF versus hierarchical clustering.

[0028] **Figure 8** illustrates the accuracy landscape of NBS across varying driver mutation frequency and module size.

[0029] **Figure 9** illustrates a standard non-network-based clustering approach (i.e., no network smoothing and substituting NMF for NetNMF) as in **Figure 8**.

[0030] **Figure 10** illustrates using a permuted network as in **Figure 8**.

[0031] **Figure 11** illustrates co-clustering matrices for uterine cancer patients, comparing NBS (STRING) to standard consensus clustering.

[0032] **Figure 12** illustrates the association of NBS subtypes with histology for uterine cancer.

[0033] **Figure 13** illustrates the composition of NBS subtypes in terms of histological type and tumor grade for uterine cancer.

[0034] **Figure 14** illustrates association of NBS subtypes with patient survival time for ovarian cancer.

[0035] **Figure 15** illustrates Kaplan-Meier survival plots for NBS subtypes for ovarian cancer.

[0036] **Figure 16** illustrates association of NBS subtypes with patient survival time for lung cancer.

[0037] **Figure 17** illustrates Kaplan-Meier survival plot for NBS subtypes and lung cancer.

[0038] **Figure 18** illustrates a comparison of data types. (a,c) A comparison of the predictive value for patient survival as estimated using a Cox proportional-hazards model, and association with histological type (e), across different data types and methods. Subtypes resulting from clustering of data from CNVs, mRNA, microRNA (miRNA), methylation and reverse phase protein arrays (RPPA) were obtained from the Broad Firehose web portal. These subtype definitions were compared to the subtypes identified by network-based stratification of somatic mutations using HumanNet with four subtypes for ovarian (b), HumanNet with six for lung (d) and STRING with three for uterine (f). In each case (b,d,f), the p-value of significance is reported from a χ^2 test of association between the assignment of patients to subtypes for each data type with NBS subtypes of a fixed number of subtypes.

[0039] **Figure 19** illustrates a network view of genes with high network smoothed mutation scores in HumanNet subtype 1 relative to other subtypes.

[0040] **Figure 20** illustrates using expression signatures derived from mutation subtypes. (a) Classification accuracy (1 – classification error) when using a supervised learning method to learn a signature based on either somatic mutation profiles or gene expression, showing training error and cross-validation error. Dashed line shows the accuracy for a random predictor. (b) Kaplan-Meier survival plots for the TCGA ovarian cancer cohort patients when predicted using a classifier trained on subtype labels derived from network-based stratification of mutation data in TCGA. (c) Applying the same classifier to serous ovarian cancer samples from Tothill *et al.*

[0041] **Figure 21** illustrates the effects of different types of mutations on stratification. (a–b) The effects of permuting a progressively larger fraction of mutation per patient for different types of somatic mutation, for the uterine (a) and ovarian (b) tumor cohorts. Lines show the median performance and colored regions represent the median absolute deviation (MAD). (c–e) Different types of filters were applied as a preprocessing step prior to running NBS on the uterine (c), ovarian (d) and lung (e) cohorts.

[0042] **Figure 22** illustrates a Kaplan-Meier plot of NBS subtypes of OV. The three subtypes are predicted in ICGC using a decision tree classifier trained on TCGA OV cohort and discovered using the NBS method.

[0043] **Figure 23** illustrates a Box-plot comparing Cisplatin sensitivity in CCLE of OV subtype 1. Using a decision tree classifier trained on TCGA we score all CCLE cell-lines for belonging to NBS OV subtype 1. Top 20 scoring subtype 1 cell-lines in CCLE are compared to bottom 80 scoring cell-lines and exhibit a significant difference in cisplatin IC50. The cell lines classified to subtype 1 show significantly less sensitivity to Cisplatin.

[0044] **Figure 24** illustrates the final tumor subtypes obtained from the consensus (majority) assignments of each tumor after 1000 applications of this procedure to samples of the original data set. A darker color coincides with higher co-clustering for pairs of patients. The overall outcome of network-based stratification is to capture informative clusters within somatic mutation data, in contrast to standard consensus clustering (Figure 5) which generally fails to produce such clusters.

[0045] **Figure 25** illustrates a network view of genes with high network smoothed mutation scores in HumanNet subtype 1 relative to other subtypes. Subtype 1 has the lowest survival and highest platinum resistance rates amongst the four recovered subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes included in the COSMIC cancer gene census.

[0046] **Figure 26** illustrates simulation across different networks. In this simulation network modules from the NCI-Nature cancer pathways network were used for the simulation and were recovered by NBS using the HumanNet network. Each subtype included between 2-6 driver modules totaling the specified size of genes and the driver gene frequency. Driver frequencies of 10%, 7.5%, 5% and driver modules comprising 100-120, 60-80, 20-40 were used in panels (a),(b) and (c) respectively. Furthermore, a subset (0-4) of the modules was assigned to overlap across multiple subtypes.

[0047] **Figure 27** illustrates uterine cancer association with histological type. (a-c) Association with histological subtype vs. the number of clusters (K). (d-f) Association with tumor grade vs. the number of clusters (K) (g) Summary of histological types for each subtype. (h) Summary of tumor grade vs each subtype.

[0048] **Figure 28** illustrates ovarian cancer association with overall survival. (a) Co-clustering matrices for ovarian cancer patients, comparing NBS (HumanNet) to standard consensus clustering. (b-c) Cox proportional hazards model logrank statistic for STRING and PathwayCommons. (d) Hazard ratio of each of the HumanNet subtypes compared to subtype 2 with confidence intervals (0.95, 0.8, 0.6 denoted in blue, yellow and orange respectively). (e) Mean and S.E survival in months for each of the subtypes. (f) A Kaplan-Meier plot of the probability of developing platinum drug resistance for HumanNet with four clusters, Logrank P=0.046. (Subtype 4 is dropped due to missing annotations for PFI for the majority of patients).

[0049] **Figure 29** illustrates lung cancer association with overall survival. (a) Co-clustering matrices for lung cancer patients, comparing NBS (HumanNet) to standard consensus clustering. (b) Lung cancer patient survival cox proportional hazard model logrank statistic for PathwayCommons. (c) A Kaplan-Meier survival plot with six subtypes.

[0050] **Figure 30** illustrates standard predictors of survival independent of ovarian subtype. (a) Percentage of patients receiving an optimal surgical resection (defined as less than 10mm of residual tumor) does not vary significantly between subtypes (χ^2 P-value = 0.77). (b) Federation of Gynaecological Oncologists (FIGO) tumor stage does not show evidence for dependence on tumor subtype (χ^2 P-value = 0.48). (c) Age at diagnosis does not show dependence on tumor subtype (One-way ANOVA P-value = 0.89).

[0051] **Figure 31** illustrates a network view of genes with high network smoothed mutation scores in ovarian, HumanNet, subtype 2 relative to other subtypes. Node size corresponds to smoothed mutation score. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.

[0052] **Figure 32** illustrates a network view of genes with high network smoothed mutation scores in ovarian, HumanNet, subtype 3 relative to other subtypes. Node size corresponds to smoothed mutation score. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.

[0053] **Figure 33** illustrates a network view of genes with high smoothed mutation scores in ovarian, HumanNet, subtype 4 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes.

[0054] **Figure 34** illustrates from mutation-derived subtypes to expression signatures. (a) A Kaplan-Meier analysis of the proportion of patients who acquire platinum resistance in the Tothill *et al.* expression cohort for subtypes defined in the TCGA dataset using somatic mutations and NBS. (b) Kaplan-Meier survival plots for the Bonome *et al.* ovarian cancer patients (c) Kaplan-Meier survival plots for a metastudy of ovarian cancer patients by Györfy *et al.*. These subtypes were recovered using a shrunken centroid model trained on the TCGA expression data with somatic mutation NBS subtypes as labels.

[0055] **Figure 35** illustrates standard consensus clustering NMF used to recover subtypes in the Tothill *et al.* expression cohort of ovarian tumors. (a) Standard consensus clustering NMF was performed for 1000 rounds with random restarts on the top 4000 most variable genes in the cohort. Average linkage hierarchical clustering was performed on the co-occurrence matrix to recover the following subtypes. Kaplan-Meier plots are shown for three (b), four (c), and five subtypes (d).

[0056] **Figure 36** illustrates the effects of progressively permuting proportions of the lung cancer dataset. Permuting a progressively larger number of mutation uniformly from the entire lung cohort. We report the median likelihood difference of a full model to a base model including just clinical covariates (age, grade, stage, mutation rate, residual tumor after surgery, as well as smoking). The colored regions represent the median absolute deviation (MAD).

[0057] **Figure 37** illustrates a network view of genes with high smoothed mutation scores in uterine cancer, STRING, subtype 1 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0058] **Figure 38** illustrates a network view of genes with high smoothed mutation scores in uterine cancer, STRING, subtype 2 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0059] **Figure 39** illustrates a network view of genes with high smoothed mutation scores in uterine cancer, STRING, subtype 3 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0060] **Figure 40** illustrates a network view of genes with high smoothed mutation scores in lung cancer, HumanNet, subtype 1 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0061] **Figure 41** illustrates a network view of genes with high smoothed mutation scores in lung cancer, HumanNet, subtype 2 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0062] **Figure 42** illustrates a network view of genes with high smoothed mutation scores in lung cancer, HumanNet, subtype 3 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0063] **Figure 43** illustrates a network view of genes with high smoothed mutation scores in lung cancer, HumanNet, subtype 5 relative to other subtypes. Node size corresponds to smoothed mutation scores. Node color corresponds to a set of functional classes of interest recovered through manual examination of the resulting network with the aid of the GeneMania Cytoscape plugin. Thickened node outlines indicate genes which are known cancer genes from the Sanger list of cancer genes. Edge thickness corresponds to relative edge confidence in the network, underlined gene names indicate the gene is mutated in this subtype.

[0064] **Figure 44** illustrates (A) Network-based stratification is a novel method that using somatic mutation data and knowledge of genetic interaction networks can stratify a heterogeneous population of cancer patients (e.g. all high grade serous ovarian cancer patients) into subtypes that are predictive of clinical outcomes (e.g. subtype 1 does not need chemotherapy at all, subtype 2 needs chemotherapy A, subtype 3 needs chemotherapy B, etc.). (B) Once subtypes are defined, a new gene expression based biomarker is developed that can classify a patient into a specific subtype. Oncologist can now make clinical decision based on past experience of other patients from that same subtype.

DETAILED DESCRIPTION OF THE INVENTION

[0065] With respect to the use of substantially any plural and/or singular terms herein, those having skill in the art can translate from the plural to the singular and/or from the singular to the plural as is appropriate to the context and/or application. The various singular/plural permutations may be expressly set forth herein for sake of clarity.

[0066] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (*e.g.*, bodies of the appended claims) are generally intended as “open” terms (*e.g.*, the term “including” should be interpreted as “including but not limited to,” the term “having” should be interpreted as “having at least,” the term “includes” should be interpreted as “includes but is not limited to,” etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present. For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases “at least one” and “one or more” to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles “a” or “an” limits any particular claim containing such introduced claim recitation to embodiments containing only one such recitation, even when the same claim includes the introductory phrases “one or more” or “at least one” and indefinite articles such as “a” or “an” (*e.g.*, “a” and/or “an” should be interpreted to mean “at least one” or “one or more”); the same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (*e.g.*, the bare recitation of “two recitations,” without other modifiers, means at least two recitations, or two or more recitations). Furthermore, in those instances where a convention analogous to “at least one of A, B, and C, etc.” is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (*e.g.*, “a system having at least one of A, B, and C” would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). In those instances where a convention analogous to “at least one of A, B, or C, etc.” is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (*e.g.*, “a system having at least one of A, B, or C” would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms, or both terms. For example, the phrase “A or B” will be understood to include the possibilities of “A” or “B” or “A and B.”

[0067] As will be understood by one skilled in the art, for any and all purposes, such as in terms of providing a written description, all ranges disclosed herein also encompass any and all possible sub-ranges and combinations of sub-ranges thereof. Any listed range can be easily recognized as sufficiently describing and enabling the same range being broken down into at least equal halves, thirds,

quarters, fifths, tenths, etc. As a non-limiting example, each range discussed herein can be readily broken down into a lower third, middle third and upper third, etc. As will also be understood by one skilled in the art all language such as “up to,” “at least,” “greater than,” “less than,” and the like include the number recited and refer to ranges which can be subsequently broken down into sub-ranges as discussed above. Finally, as will be understood by one skilled in the art, a range includes each individual member. Thus, for example, a group having 1-3 articles refers to groups having 1, 2, or 3 articles. Similarly, a group having 1-5 articles refers to groups having 1, 2, 3, 4, or 5 articles, and so forth.

[0068] The term “comprising” as used herein is synonymous with “including,” “containing,” or “characterized by,” and is inclusive or open-ended and does not exclude additional, unrecited elements or method steps.

[0069] All numbers expressing quantities of ingredients, reaction conditions, and so forth used in the specification are to be understood as being modified in all instances by the term “about.” Accordingly, unless indicated to the contrary, the numerical parameters set forth herein are approximations that may vary depending upon the desired properties sought to be obtained. At the very least, and not as an attempt to limit the application of the doctrine of equivalents to the scope of any claims in any application claiming priority to the present application, each numerical parameter should be construed in light of the number of significant digits and ordinary rounding approaches.

[0070] Cancer is a disease that can be complex. For example, cancer can be driven by a combination of genes. Cancer can also be extremely heterogeneous, in that gene combinations can vary greatly between patients. To understand the complexities of cancer, major projects such as The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) can systematically profile thousands of tumors at multiple layers of genome-scale information, including mRNA and microRNA expression, DNA copy number and methylation, and DNA sequence. There is now a strong need for informatics methods, for example bioinformatics methods, that can integrate and interpret genome-scale molecular information to provide insight into the molecular processes that can drive tumor progression. Informatics methods, such as bioinformatics methods, can also be of pressing need in the clinic, where the impact of genome-scale tumor profiling can be limited by the inability of current analysis techniques to derive clinically-relevant conclusions from the data.

[0071] “Bioinformatics” as described herein, is a study of information science that can utilize large databases of biochemical and/or pharmaceutical information. As applied to life sciences, the technology can be used for the collection and analysis of biological data. Biological data for bioinformatics can include but are not limited to data from microarrays, sequencing data, proteomic data, genomic data, and many types of biological data that are known to those skilled in the art. Bioinformatics technologies can be used for developing methods and software tools for storing, retrieving, organizing,

and analyzing multiple types of biological data. A primary goal for bioinformatics is to increase the understanding of biological processes and pathways, by the application of computational techniques. Bioinformatics can combine databases, computer science, algorithms, statistics, biostatistics, mathematics, and engineering to study, process, and analyze biological data. There are many commonly used software tools and technologies in bioinformatics that can include but are not limited to Bioconductor, Galaxy, GenePattern, GenomeSpace, Integrated Genome Browser, Cytoscape, Java, C, XML, Perl, C++, Python, R, SQL, CUDA, MATLAB, spreadsheet applications. In some embodiments, bioinformatics is used to organize and analyze biological data. In some embodiments, bioinformatics is used to analyze genomic data. In some embodiments, methods for stratification of cancer into one or more informative subtypes of a subject in need thereof are provided. In some embodiments, the method is carried out by an informatics platform. In some embodiments, the informatics platform is a bioinformatics platform comprising a computer and software.

[0072] As used herein, a “subject” means a human or animal. Usually the animal is a vertebrate such as a primate, rodent, domestic animal or game animal. In certain embodiments of the aspects described herein, the subject is a mammal, e.g., a primate, e.g. a human. The terms, “patient” and “subject” are used interchangeably. A subject can be male or female.

[0073] Preferably, the subject is a mammal. The mammal can be a human, non-human primate, mouse, rat, dog, cat, horse, or cow, but are not limited to these examples. Mammals, other than humans, can be advantageously used as subjects that represent animal models of disorders associated with, e.g., cancer. In addition, the methods and compositions described herein can be used to treat domesticated animals and/or pets.

[0074] “Tumor stratification” as described herein includes dividing a heterogeneous population of tumors into clinically-meaningful subtypes based on the similarity of molecular profiles. The identification of specific molecular markers can be used to stratify the tumor samples into meaningful subtypes and is also an important goal in cancer genomics and other types of cancer studies that are known to those skilled in the art. The subtypes may correlate with specific clinical features for example, the aggressiveness of a tumor, the response to drugs, and an overall outcome during the prognosis. The subtype can be a clinical phenotype. The clinical phenotype can be predictive of a survival rate, drug response, and/or a tumor grade. Overall, the method of tumor stratification can lead to providing new areas of cancer research or treatment or patient care such as providing new drug targets, precision cancer treatments for personalized care for patients with specific subtypes, and precision oncology. Stratification can also lead to predicting the efficiency of personalized and precision medicine and therapeutics, which can provide the safest and more effective therapeutic strategy, based on e.g., the gene and protein variations of each patient. Therefore stratification can improve diagnosis and treatment through therapy

design. In some embodiments, a method of tumor stratification is provided. In some embodiments, the method comprises obtaining sequence information from a bootstrap sample of genes from a tumor sample of the subject, projecting a mutation found within the sequence information onto a network, propagating the mutation in the network, clustering the mutation(s) so propagated so as to divide subjects with the mutation(s) into subtypes thereby stratifying cancer into informative subtypes and assigning of the subject to an informative subtype. In some embodiments, stratification is performed by a bioinformatics platform. In some embodiments, the informative subtype is a clinical phenotype. In some embodiments the clinical phenotype is predictive of a survival rate, drug response, or a tumor grade.

[0075] A source of data for stratification can be the somatic mutation profile, in which the genome or exome of a patient's tumor and that of the germline are compared to identify mutations that have become enriched in the tumor cell population. "Next-generation sequencing", Sanger sequencing or other means of obtaining genomic information known to those skilled in the art can be used to derive tumor and germline genomes or exomes in whole or in part. "Somatic mutation" as described herein refers to a genetic mutation occurring in a somatic cell, and can provide the basis for a mosaic condition. These mutations occur in the DNA after conception and can occur in any of the cells of the body except for germ-line cells. Somatic mutations in a cancer cell can encompass distinct classes of DNA sequence changes. These changes can be the substitution of one base by another, insertions or deletions of small or large segments of DNA, rearrangements in which DNA has been broken and then rejoined to another DNA segment located elsewhere within the genome, copy number increases from the two copies present in the normal diploid genome, sometimes to several hundred copies (known as gene amplification); and copy number reductions that can result in complete absence of a DNA sequence from the cancer genome. As this set of mutations is presumed to contain the causal drivers of tumor progression, similarities and differences in mutations across patients can provide invaluable information for stratification. In some embodiments, methods are provided for stratification based on a somatic mutation profile. In some embodiments, the methods are performed using a bioinformatics platform comprising a computer and software.

[0076] "Next-generation sequencing" as described herein includes high speed and high through put sequencing techniques. Instruments for next generation sequencing can include but are not limited to Illumina HiSeq2000 (Illumina), Ion Torrent (Life Technologies), MiSeq (Illumina), GS FLX+ (Roche Diagnostics Corp), and other instruments for sequencing that are known to those skilled in the art. Techniques can be used to analyze and sequence millions or billions of DNA strands in parallel to yield more through-put and minimize the need for the fragment cloning methods that are used in Sanger sequencing of genomes. Several types of next generation sequencing programs include EagleView genome viewer, Galaxy, BWA, Bowtie, MUMmerGPU, Batman, Alta-Cyclic, FindPeaks 3.1,

ALLPATHS, SHARCGS, Velvet, EDENA, SSAKE, apalma, SOAP, SOAPdenovo, SOAPsnp, CLCbiogenomicsWorkbench, NextGENE, SeqMan Genome Analyser, ELAND, GMAP, MOSAIK, MAQ, MUMmer, Novocraft, RMAP, SHRiMP, SSAHA, ZOOM, CisGenome, CloudBurst, ChiPmeta, and other programs for next-generation sequencing and data analysis that are known to those skilled in the art.

[0077] While individual mutations in well-established cancer genes have long been used to stratify patients in a straightforward manner, stratification of the entire mutation profile of a patient can be more challenging. Somatic mutations are fundamentally unlike other data types such as expression or methylation, in which nearly all genes or markers can be assigned a quantitative value in every patient. Cancer is a disease that can be driven by such somatic mutations, which accumulate in the genome during the lifetime of the individual. However, somatic mutation profiles are extremely sparse, with typically fewer than 100 mutated bases in an entire exome. Reference is made to **Figure 1**, which shows an illustration of the somatic mutation landscape of the TCGA ovarian cancer cohort. For examples, as shown in **panel a**, of **Figure 1**, somatic mutations located along the length of chromosome 17 are indicated. In **panel b** of **Figure 1**, a histogram summing the frequency of mutations per gene for the entire exome is shown. In **panel c** of **Figure 1**, a histogram summing the frequency of genes mutated per patient in the cohort is indicated. The data indicate that they are also remarkably heterogeneous, such that it is very common for clinically-identical patients to share no more than a single mutation. From the results of this example, the data shows why the clustering of mutation profiles is particularly challenging and why the previous methods of stratification using standard approaches for clustering have failed to produce meaningful stratification results.

[0078] Data scarcity and other problems associated with clustering of mutation profiles may be largely overcome by integrating somatic mutation profiles with knowledge of the molecular network architecture of human cells. It is widely known that cancer is a not just a disease of individual mutations, nor of genes, but it can manifest due to the combinations of genes acting in molecular networks corresponding to hallmark processes, for example, cell division, signaling pathways in cancer metabolism, cell proliferation, apoptosis, and other cellular processes controlled by genes that are known to those skilled in the art. For example, two tumors may not share any mutations in common, but they can share remarkable similarity in the networks impacted by these mutations. In reference to C.H Waddington, the term “canalization” in genetics refers to the measure of the ability of a population to produce the same phenotype regardless of the variability of its environment or the genotype.

[0079] Although current cancer pathway maps are incomplete, relevant information can be widely available in the current public databases of human genomes, human protein-protein interactions, and pathway interactions. For example the use of proteomics, genomics, and the study of protein-protein

interactions are important to understand human diseases and cancer on a system-wide level. The use of genomics and proteomics databases to understand human diseases and cancer can include but are not limited to Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), AllFuse, Asedb, Binding Interface Database (BID), BioGrid, Biomolecular Object Network Databank (BIND), Database of Interacting Proteins (DIP, UCLA), Genomic Knowledge Database, Human Unidentified Gene Encoded large proteins (HUGE), HumanNet, Human Protein Reference Database, Inter-Chain Beta Sheets database (ICBS), IntAct, database of Kinetic Data of Biomolecular Interactions (KDBI), Biomolecular Relations in Information Transmission and Expression (KEGG BRITE), Molecular Interactions Database (MINT), Domain peptide Interactions database (DOMINO), molmovdb.org, Mammalian Protein Protein Interaction database (MPPI), PathwayCommons, PepCyber, POINT, Protein Interactions and Molecular Information database (PRIME), Protein Interaction Database, and other programs known to those skilled in the art. HumanNet, for example, is an interaction database that can integrate many different lines of evidence, presently catalogues approximately 400,000 pairwise interactions among more than 16,000 human genes.

[0080] The increased number of approaches can be successful in integrating network databases with tumor molecular profiles to map the molecular pathways of cancer. However, as described herein, the focus is e.g., on a method of using network knowledge to stratify a cohort into meaningful subsets, for example the stratifying of the somatic mutation profiles of major cancers. Using a method of stratification, somatic mutation profiles can be clustered into robust tumor subtypes with strong association to clinical outcomes. Clinical outcomes, for example, can refer to patient survival time, aggressiveness of cancer, drug response, emergence of drug resistance, and other processes known to those skilled in the art. This method can be applied to stratify the somatic mutation profiles of major cancers, for example, three major cancers catalogued in TCGA: ovarian, uterine and lung, as described in the **Examples** below. In some embodiments, somatic mutation profiles can be subtyped.

[0081] Many cancers, for example, stomach, bladder, renal, ovarian, colorectal, lymphoma, lung, uterine, breast, ovarian, etc. can be classified into many molecular subtypes. For example stomach cancer can have 4 subtypes: tumors positive for Epstein-Barr virus (EBV), tumors with high microsatellite instability, tumors that can differ in the level of somatic copy number alterations (SCNAs), and tumors classified as chromosomally unstable, with a high level of SCNAs. The ability to stratify tumors into subtypes can advance research by giving genomic insights into many causes of a deadly form of cancer.

[0082] Cancer such as ovarian cancer, for example, can have several subtypes affected by specific genes. For example, ovarian cancer subtype 1 can have one or more or all of the mutations in the following genes: TTN (titin), NEB (nebulin), AP1G2 (adaptor-related protein complex 1, gamma 2

subunit), SYNRG (synergisin, gamma), SPTBN4 (spectrin, beta, non-erythrocytic 4), ANK1 (ankyrin 1, erythrocytic), SLC12A8 (solute carrier family 12 (potassium/chloride transporters), member 8), CACNA1A (calcium channel, voltage-dependent, P/Q type, alpha 1A subunit), MPP1 (membrane protein, palmitoylated 1, 55kDa), RIMS1 (regulating synaptic membrane exocytosis 1), SCML2 (sex comb on midleg-like 2 (Drosophila)), CIDEB (cell death-inducing DFFA-like effector b), RHAG (Rh-associated glycoprotein), GAD2 (glutamate decarboxylase 2 (pancreatic islets and brain, 65kDa)), FGFR4 (fibroblast growth factor receptor 4), ATP6V0D2 (ATPase, H⁺ transporting, lysosomal 38kDa, V0 subunit d2), MYOM2 (myomesin 2), AP1M1 (adaptor-related protein complex 1, mu 1 subunit), INA (internexin neuronal intermediate filament protein, alpha), KCNH7 (potassium voltage-gated channel, subfamily H (eag-related), member 7), SLC2A5 (solute carrier family 2 (facilitated glucose/fructose transporter), member 5), and/or PTPRN (protein tyrosine phosphatase, receptor type, N).

[0083] In another example, ovarian cancer subtype 2 can have one or more or all of the mutations in the following genes: TP53 (tumor protein p53), BRCA1 (breast cancer 1, early onset), BRCA2 (breast cancer 2, early onset), CREBBP (CREB binding protein), USP7 (ubiquitin specific peptidase 7 (herpes virus-associated)), ST18 (suppression of tumorigenicity 18 (breast carcinoma) (zinc finger protein)), NUP155 (nucleoporin 155kDa), NUP160 (nucleoporin 160kDa), SLC11A1 (solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1), PRRC2C (proline-rich coiled-coil 2C), DMBT1 (deleted in malignant brain tumors 1), NUP62 (nucleoporin 62kDa), RANBP2 (RAN binding protein 2), CRMP1 (collapsin response mediator protein 1), TPR (translocated promoter region, nuclear basket protein), TNPO3 (transportin 3), CEBPE (CCAAT/enhancer binding protein (C/EBP), epsilon), NUP133 (nucleoporin 133kDa), MAP1B (microtubule-associated protein 1B), TP53BP2 (tumor protein p53 binding protein, 2), ADAMTS4 (ADAM metalloproteinase with thrombospondin type 1 motif, 4), PTEN (phosphatase and tensin homolog), NUP188 (nucleoporin 188kDa), NUP214 (nucleoporin 214kDa), NUP153 (nucleoporin 153kDa), DPYSL5 (dihydropyrimidinase-like 5), N6AMT1 (N-6 adenine-specific DNA methyltransferase 1 (putative)), NUP98 (nucleoporin 98kDa), DPYSL4 (dihydropyrimidinase-like 4), FOSB (FBJ murine osteosarcoma viral oncogene homolog B), NUP205 (nucleoporin 205kDa), CUL9 (cullin 9), MDM4 (Mdm4 p53 binding protein homolog (mouse)), USP30 (ubiquitin specific peptidase 30), EP300 (E1A binding protein p300), CHEK2 (checkpoint kinase 2), NF2 (neurofibromin 2 (merlin)), SMURF1 (SMAD specific E3 ubiquitin protein ligase 1), SIRT5 (sirtuin 5), NUP35 (nucleoporin 35kDa), POM121 (POM121 transmembrane nucleoporin), NUP85 (nucleoporin 85kDa), ARID5B (AT rich interactive domain 5B (MRF1-like)), SIRT6 (sirtuin 6), CREB3 (cAMP responsive element binding protein 3), NUP93 (nucleoporin 93kDa), BATF3 (basic leucine zipper transcription factor, ATF-like 3), SENP2 (SUMO1/sentrin/SMT3 specific peptidase 2), EGR2 (early growth response 2), PSIP1 (PC4 and SFRS1

interacting protein 1), RAE1 (RAE1 RNA export 1 homolog (*S. pombe*)), BRIP1 (BRCA1 interacting protein C-terminal helicase 1), NUP107 (nucleoporin 107kDa), MAP1A (microtubule-associated protein 1A), FMOD (fibromodulin), BATF (basic leucine zipper transcription factor, ATF-like), IPO7 (importin 7), GABPA (GA binding protein transcription factor, alpha subunit 60kDa), ATF1 (activating transcription factor 1), SIRT1 (sirtuin 1), E4F1 (E4F transcription factor 1), THNSL2 (threonine synthase-like 2 (*S. cerevisiae*)), NPEPPS (aminopeptidase puromycin sensitive), NUP37 (nucleoporin 37kDa), DDX1 (DEAD (Asp-Glu-Ala-Asp) box helicase 1), GARS (glycyl-tRNA synthetase), KPNB1 (karyopherin (importin) beta 1), RPRD1A (regulation of nuclear pre-mRNA domain containing 1A), EGR1 (early growth response 1), EVI2A (ecotropic viral integration site 2A), TBL1XR1 (transducin (beta)-like 1 X-linked receptor 1), FOS (FBJ murine osteosarcoma viral oncogene homolog), CCNH (cyclin H), SMAD4 (SMAD family member 4), SSTR3 (somatostatin receptor 3), SDCBP2 (syndecan binding protein (syntenin) 2), MED25 (mediator complex subunit 25), ADAMTS2 (ADAM metalloproteinase with thrombospondin type 1 motif, 2), ACVRL1 (activin A receptor type II-like 1), PHAX (phosphorylated adaptor for RNA export), and/or XPO1 (exportin 1 (CRM1 homolog, yeast)).

[0084] In another example ovarian cancer subtype 3 can have one or more or all of the mutations in the following genes: AHNAK (AHNAK nucleoprotein), RPS6KL1 (ribosomal protein S6 kinase-like 1), IFNA13 (interferon, alpha 13), IRF8 (interferon regulatory factor 8), HDAC5 (histone deacetylase 5), and/or PIGR (polymeric immunoglobulin receptor).

[0085] In another example ovarian cancer subtype 4 can have one or more or all of the mutations in the following genes: MYH4 (myosin, heavy chain 4, skeletal muscle), MYH2 (myosin, heavy chain 2, skeletal muscle, adult), SWAP70 (SWAP switching B-cell complex 70kDa subunit), FGF10 (fibroblast growth factor 10), FOLR1 (folate receptor 1 (adult)), GLUD2 (glutamate dehydrogenase 2), GYG1 (glycogenin 1), GYS1 (glycogen synthase 1 (muscle)), PHKA1 (phosphorylase kinase, alpha 1 (muscle)), PRKAG1 (protein kinase, AMP-activated, gamma 1 non-catalytic subunit), and/or ROM1 (retinal outer segment membrane protein 1).

[0086] Uterine cancer can have, for example, 3 subtypes. Uterine cancer subtype 1 can have mutation(s) in one or more or all of the following genes: TAPBP (TAP binding protein (tapasin)), HIST1H1C (histone cluster 1, H1c), ARID3A (AT rich interactive domain 3A (BRIGHT-like)), ATF3 (activating transcription factor 3), HLA-A (major histocompatibility complex, class I, A), PHB (prohibitin), PADI4 (peptidyl arginine deiminase, type IV), TP53 (tumor protein p53), EPCAM (epithelial cell adhesion molecule), DYRK2 (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2), PRDM1 (PR domain containing 1, with ZNF domain), RB1CC1 (RB1-inducible coiled-coil 1), RNF20 (ring finger protein 20, E3 ubiquitin protein ligase), IRF5 (interferon regulatory factor 5), PPP1R13B (protein phosphatase 1, regulatory subunit 13B), SIK1 (salt-inducible kinase 1), CUL9 (cullin

9), PRKAB1 (protein kinase, AMP-activated, beta 1 non-catalytic subunit), RNF144B (ring finger protein 144B), CD59 (CD59 molecule, complement regulatory protein), DUSP1 (dual specificity phosphatase 1), BCL2L12 (BCL2-like 12 (proline rich)), JMY (junction mediating and regulatory protein, p53 cofactor), BAI1 (brain-specific angiogenesis inhibitor 1), CD82 (CD82 molecule), RRAD (Ras-related associated with diabetes), CAMK2D (calcium/calmodulin-dependent protein kinase II delta), PAK3 (p21 protein (Cdc42/Rac)-activated kinase 3), FBXO11 (F-box protein 11), C12orf5 (chromosome 12 open reading frame 5), ZACN (zinc activated ligand-gated ion channel), E4F1 (E4F transcription factor 1), CHEK1 (checkpoint kinase 1), UCHL1 (ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)), CSE1L (CSE1 chromosome segregation 1-like (yeast)), STEAP3 (STEAP family member 3, metalloredutase), SUMO1 (SMT3 suppressor of mif two 3 homolog 1 (*S. cerevisiae*)), CSNK1G3 (casein kinase 1, gamma 3), RAD54L (RAD54-like (*S. cerevisiae*)), COL18A1 (collagen, type XVIII, alpha 1), PIAS2 (protein inhibitor of activated STAT, 2), FAS (Fas (TNF receptor superfamily, member 6)), CTSL1 (cathepsin L1), LMLN (leishmanolysin-like (metallopeptidase M8 family)), HIC1 (hypermethylated in cancer 1), PLK3 (polo-like kinase 3), RPRM (reprimin, TP53 dependent G2 arrest mediator candidate), IFI16 (interferon, gamma-inducible protein 16), GNL3 (guanine nucleotide binding protein-like 3 (nucleolar)), NOX1 (NADPH oxidase 1), WWOX (WW domain containing oxidoreductase), ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)), HYAL2 (hyaluronoglucosaminidase 2), TNK2 (tyrosine kinase, non-receptor, 2), SERTAD4 (SERTA domain containing 4), ZCCHC8 (zinc finger, CCHC domain containing 8), CEP41 (centrosomal protein 41kDa), EXOSC5 (exosome component 5), SKIV2L2 (superkiller viralicidic activity 2-like 2 (*S. cerevisiae*)), SLMAP (sarcolemma associated protein), NEUROD6 (neuronal differentiation 6), HABP4 (hyaluronan binding protein 4), DLX2 (distal-less homeobox 2), PPP2R1A (protein phosphatase 2, regulatory subunit A, alpha), PPP2R5C (protein phosphatase 2, regulatory subunit B', gamma), PPP2R3A (protein phosphatase 2, regulatory subunit B", alpha), NDN (necdin, melanoma antigen (MAGE) family member), PRR14 (proline rich 14), POLR2J1 (polymerase (RNA) II (DNA directed) polypeptide J, 13.3kDa), PAF1 (Paf1, RNA polymerase II associated factor, homolog (*S. cerevisiae*)), CSNK1E (casein kinase 1, epsilon), TAF9B (TAF9B RNA polymerase II, TATA box binding protein (TBP)-associated factor, 31kDa), TAF3 (TAF3 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 140kDa), PRMT5 (protein arginine methyltransferase 5), ANKS1B (ankyrin repeat and sterile alpha motif domain containing 1B), MMS19 (MMS19 nucleotide excision repair homolog (*S. cerevisiae*)), INTS6 (integrator complex subunit 6), BRD7 (bromodomain containing 7), TAF5L (TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa), GTF2A1 (general transcription factor IIA, 1, 19/37kDa), GTF2E1 (general transcription factor IIE, polypeptide 1, alpha 56kDa), HNRNPA1 (heterogeneous nuclear ribonucleoprotein A1), NFKBIA (nuclear factor of kappa light polypeptide gene enhancer in B-cells

inhibitor, alpha), ERCC2 (excision repair cross-complementing rodent repair deficiency, and/or C19orf2 (unconventional prefoldin RPB5 interactor).

[0087] Uterine cancer subtype 2 can have mutation(s) in one or more or all of the following genes: PTEN (phosphatase and tensin homolog), CTNNB1 (catenin (cadherin-associated protein), beta 1, 88kDa), ARID1A (AT rich interactive domain 1A (SWI-like)), PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1 (alpha)), MUC4 (mucin 4, cell surface associated), CTCF (CCCTC-binding factor (zinc finger protein)), FGFR2 (fibroblast growth factor receptor 2), PRG4 (p53-responsive gene 4), SOX17 (SRY (sex determining region Y)-box 17), EIF3C (eukaryotic translation initiation factor 3, subunit C), IRS4 (insulin receptor substrate 4), INVS (inversin), TLE1 (transducin-like enhancer of split 1 (E(spl) homolog, *Drosophila*)), TNIK (TRAF2 and NCK interacting kinase), INPPL1 (inositol polyphosphate phosphatase-like 1), PIKFYVE (phosphoinositide kinase, FYVE finger containing), PDYN (prodynorphin), C4BPA (complement component 4 binding protein, alpha), PIK3CB (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit beta), AGAP2 (ArfGAP with GTPase domain, ankyrin repeat and PH domain 2), FGF13 (fibroblast growth factor 13), NKTR (natural killer-tumor recognition sequence), CYSLTR2 (cysteinyl leukotriene receptor 2), MCRS1 (microspherule protein 1), SOX9 (SRY (sex determining region Y)-box 9), FGFR4 (fibroblast growth factor receptor 4), FIG4 (FIG4 homolog, SAC1 lipid phosphatase domain containing (*S. cerevisiae*)), CDON (cell adhesion associated, oncogene regulated), INPP4A (inositol polyphosphate-4-phosphatase, type I, 107kDa), DMBT1 (deleted in malignant brain tumors 1), PARD3 (par-3 partitioning defective 3 homolog (*C. elegans*)), SMARCA2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2), ARID1B (AT rich interactive domain 1B (SWI-like)), IHH (indian hedgehog), RHEB (Ras homolog enriched in brain), OPRL1 (opiate receptor-like 1), CDKN2A (cyclin-dependent kinase inhibitor 2A), KITLG (KIT ligand), FPR2 (formyl peptide receptor 2), FIGF (c-fos induced growth factor (vascular endothelial growth factor D)), TACR2 (tachykinin receptor 2), IGFBP2 (insulin-like growth factor binding protein 2, 36kDa), EIF3J (eukaryotic translation initiation factor 3, subunit J), PROKR1 (prokineticin receptor 1), SMARCD2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 2), THRA (thyroid hormone receptor, alpha), ERFFI1 (ERBB receptor feedback inhibitor 1), INPP5B (inositol polyphosphate-5-phosphatase, 75kDa), ITK (IL2-inducible T-cell kinase), PMPCA (peptidase (mitochondrial processing) alpha), CSNK1A1L (casein kinase 1, alpha 1-like), INPP5J (inositol polyphosphate-5-phosphatase J), EPHB4 (EPH receptor B4), PROKR2 (prokineticin receptor 2), EPHA2 (EPH receptor A2), DMP1 (dentin matrix acidic phosphoprotein 1), VWCE (von Willebrand factor C and EGF domains), FGF12 (fibroblast growth factor 12), FRK (fyn-related kinase), MIB2 (mindbomb E3 ubiquitin protein ligase 2), LIMA1 (LIM domain and actin binding 1), MUC7 (mucin 7, secreted), PI4KB (phosphatidylinositol 4-kinase, catalytic, beta),

MTMR3 (myotubularin related protein 3), SMARCC1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1), MAST3 (microtubule associated serine/threonine kinase 3), GEN1 (Gen endonuclease homolog 1 (Drosophila)), PIK3C2B (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 beta), INPP4B (inositol polyphosphate-4-phosphatase, type II, 105kDa), ESRP1 (epithelial splicing regulatory protein 1), PTPRF (protein tyrosine phosphatase, receptor type, F), PDC (phosducin), FGF2 (fibroblast growth factor 2 (basic)), CRH (corticotropin releasing hormone), IL17A (interleukin 17A), CRK (v-crk sarcoma virus CT10 oncogene homolog (avian)), FIGLA (folliculogenesis specific basic helix-loop-helix), SLC9A3R2 (solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 2), WNT4 (wingless-type MMTV integration site family, member 4), CD83 (CD83 molecule), MED31 (mediator complex subunit 31), SUB1 (SUB1 homolog (S. cerevisiae)), SH2D2A (SH2 domain containing 2A), FHL2 (four and a half LIM domains 2), NANOG (Nanog homeobox), SLC9A3R1 (solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 1), IGF2 (insulin-like growth factor 2 (somatomedin A)), WNT1 (wingless-type MMTV integration site family, member 1), IL2RA (interleukin 2 receptor, alpha), C17orf72 (chromosome 17 open reading frame 72), NOG (noggin), PRDX1 (peroxiredoxin 1), SYT8 (synaptotagmin VIII), F2RL2 (coagulation factor II (thrombin) receptor-like 2), TWIST2 (twist basic helix-loop-helix transcription factor 2), PDPK1 (3-phosphoinositide dependent protein kinase-1), PI4K2A (phosphatidylinositol 4-kinase type 2 alpha), CACYBP (calcyclin binding protein), DVL1 (dishevelled, dsh homolog 1 (Drosophila)), CD28 (CD28 molecule), THEM4 (thioesterase superfamily member 4), CSNK2A2 (casein kinase 2, alpha prime polypeptide), XCR1 (chemokine (C motif) receptor 1), FZD8 (frizzled family receptor 8), FZD5 (frizzled family receptor 5), ICOS (inducible T-cell co-stimulator), ICOSLG (inducible T-cell co-stimulator ligand), FGF9 (fibroblast growth factor 9 (glia-activating factor)), MED16 (mediator complex subunit 16), MDFIC (MyoD family inhibitor domain containing), TBC1D10A (TBC1 domain family, member 10A), ADRA1D (adrenoceptor alpha 1D), AVPR1B (arginine vasopressin receptor 1B), MED4 (mediator complex subunit 4), ASCC1 (activating signal cointegrator 1 complex subunit 1), FZD1 (frizzled family receptor 1), RELB (v-rel reticuloendotheliosis viral oncogene homolog B), TNFRSF13B (tumor necrosis factor receptor superfamily, member 13B), TNFRSF9 (tumor necrosis factor receptor superfamily, member 9), TMEM55B (transmembrane protein 55B), STX4 (syntaxin 4), AP1M1 (adaptor-related protein complex 1, mu 1 subunit), C5orf4, TCN1 (transcobalamin I (vitamin B12 binding protein, R binder family)), HCK (hemopoietic cell kinase), MS4A1 (membrane-spanning 4-domains, subfamily A, member 1), FRS3 (fibroblast growth factor receptor substrate 3), SHB (Src homology 2 domain containing adaptor protein B), GFRA1 (GDNF family receptor alpha 1), GNA14 (guanine nucleotide binding protein (G protein), alpha 14), GP6 (glycoprotein VI (platelet)), IRAK2 (interleukin-1 receptor-

associated kinase 2), MED27 (mediator complex subunit 27), IL23R (interleukin 23 receptor), TCF7 (transcription factor 7 (T-cell specific, HMG-box)), SMARCB1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1), TMEM55A (transmembrane protein 55A), IPMK (inositol polyphosphate multikinase), CTNNAL1 (catenin (cadherin-associated protein), alpha-like 1), PRKCI (protein kinase C, iota), EPHB3 (EPH receptor B3), FLT4 (fms-related tyrosine kinase 4), TLR1 (toll-like receptor 1), DDX17 (DEAD (Asp-Glu-Ala-Asp) box helicase 17), WNT16 (wingless-type MMTV integration site family, member 16), PIP4K2A (phosphatidylinositol-5-phosphate 4-kinase, type II, alpha), CARD10 (caspase recruitment domain family, member 10), FOXO4 (forkhead box O4), IGF2BP1 (insulin-like growth factor 2 mRNA binding protein 1), PIK3R2 (phosphoinositide-3-kinase, regulatory subunit 2 (beta)), CDX4 (caudal type homeobox 4), WNT2B (wingless-type MMTV integration site family, member 2B), PIK3R3 (phosphoinositide-3-kinase, regulatory subunit 3 (gamma)), PLCD4 (phospholipase C, delta 4), PLCB2 (phospholipase C, beta 2), BMP7 (bone morphogenetic protein 7), PIK3R5 (phosphoinositide-3-kinase, regulatory subunit 5), IBSP (integrin-binding sialoprotein), PLCZ1 (phospholipase C, zeta 1), BCL9L (B-cell CLL/lymphoma 9-like), PBRM1 (polybromo 1), TLE4 (transducin-like enhancer of split 4 (E(sp1) homolog, Drosophila)), ARHGAP10 (Rho GTPase activating protein 10), AXIN2 (axin 2), PLCB3 (phospholipase C, beta 3 (phosphatidylinositol-specific)), MAGI3 (membrane associated guanylate kinase, WW and PDZ domain containing 3), SMARCC2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2), VAV3 (vav 3 guanine nucleotide exchange factor), PIK3C2G (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 gamma), ROCK1 (Rho-associated, coiled-coil containing protein kinase 1), INPP5D (inositol polyphosphate-5-phosphatase, 145kDa), MAST1 (microtubule associated serine/threonine kinase 1), PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha), INPP5F (inositol polyphosphate-5-phosphatase F), PTPRM (protein tyrosine phosphatase, receptor type, M), SYNJ2 (synaptojanin 2), and/or MED22 (mediator complex subunit 22).

[0088] Uterine cancer subtype 3 can have mutation(s) in one or more or all of the following genes: TTN (titin), NEB (nebulin), DST (dystonin), FAT3 (FAT tumor suppressor homolog 3 (Drosophila)), SYNE1 (spectrin repeat containing, nuclear envelope 1), DMD (dystrophin), RYR1 (ryanodine receptor 1 (skeletal)), MKI67 (antigen identified by monoclonal antibody Ki-67), FAT4 (FAT tumor suppressor homolog 4 (Drosophila)), TAF1 (TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa), DNAH5 (dynein, axonemal, heavy chain 5), DNAH3 (dynein, axonemal, heavy chain 3), LAMA2 (laminin, alpha 2), ASPM (asp (abnormal spindle) homolog, microcephaly associated (Drosophila)), CEP290 (centrosomal protein 290kDa), DYNC1H1 (dynein, cytoplasmic 1, heavy chain 1), DNAH2 (dynein, axonemal, heavy chain 2), TRRAP

(transformation/transcription domain-associated protein), NIPBL (Nipped-B homolog (Drosophila)), NBEA (neurobeachin), SYNE2 (spectrin repeat containing, nuclear envelope 2), DNAH7 (dynein, axonemal, heavy chain 7), DNAH1 (dynein, axonemal, heavy chain 1), NSD1 (nuclear receptor binding SET domain protein 1), DNAH8 (dynein, axonemal, heavy chain 8), JAK1 (Janus kinase 1), ITPR1 (inositol 1,4,5-trisphosphate receptor, type 1), APC (adenomatous polyposis coli), GPR98 (G protein-coupled receptor 98), FN1 (fibronectin 1), SPEN (spen homolog, transcriptional regulator (Drosophila)), SVIL (supervillin), DNAH9 (dynein, axonemal, heavy chain 9), TPR (translocated promoter region, nuclear basket protein), CREBBP (CREB binding protein), PCNT (pericentrin), PKD1L1 (polycystic kidney disease 1 like 1), HCFC1 (host cell factor C1 (VP16-accessory protein)), MYO7A (myosin VIIA), DNAH17 (dynein, axonemal, heavy chain 17), EP300 (E1A binding protein p300), ANK1 (ankyrin 1, erythrocytic), RIF1 (RAP1 interacting factor homolog (yeast)), RB1 (retinoblastoma 1), NCOR2 (nuclear receptor corepressor 2), NRXN1 (neurexin 1), DICER1 (dicer 1, ribonuclease type III), CENPF (centromere protein F, 350/400kDa), ROS1 (c-ros oncogene 1, receptor tyrosine kinase), PKD1L2 (polycystic kidney disease 1-like 2), CENPE (centromere protein E, 312kDa), SPAG17 (sperm associated antigen 17), CAD (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase), TOP2A (topoisomerase (DNA) II alpha 170kDa), NCOR1 (nuclear receptor corepressor 1), NUP98 (nucleoporin 98kDa), CASK (calcium/calmodulin-dependent serine protein kinase (MAGUK family)), HDAC6 (histone deacetylase 6), CLASP1 (cytoplasmic linker associated protein 1), KIF4A (kinesin family member 4A), ATP1A4 (ATPase, Na⁺/K⁺ transporting, alpha 4 polypeptide), EGF (epidermal growth factor), SEC24D (SEC24 family, member D (S. cerevisiae)), CKAP5 (cytoskeleton associated protein 5), DLGAP2 (discs, large (Drosophila) homolog-associated protein 2), CATSPER1 (cation channel, sperm associated 1), C9orf174, TRPM8 (transient receptor potential cation channel, subfamily M, member 8), TJP1 (tight junction protein 1), BRCA1 (breast cancer 1, early onset), TRIP11 (thyroid hormone receptor interactor 11), DCTN1 (dynactin 1), SHANK2 (SH3 and multiple ankyrin repeat domains 2), TDRD1 (tudor domain containing 1), NDST1 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1), ABI3BP (ABI family, member 3 (NESH) binding protein), SPAG16 (sperm associated antigen 16), PTCHD1 (patched domain containing 1), ASMTL (acetylserotonin O-methyltransferase-like), CATSPERG (catsper channel auxiliary subunit gamma), UBN1 (ubiquitin 1), EFHC1 (EF-hand domain (C-terminal) containing 1), ABCC1 (ATP-binding cassette, sub-family C (CFTR/MRP), member 1), PIWIL1 (piwi-like RNA-mediated gene silencing 1), SLC16A2 (solute carrier family 16, member 2 (thyroid hormone transporter)), DARS2 (aspartyl-tRNA synthetase 2, mitochondrial), ANKFY1 (ankyrin repeat and FYVE domain containing 1), CDK17 (cyclin-dependent kinase 17), SUN1 (Sad1 and UNC84 domain containing 1), SPICE1 (spindle and centriole associated protein 1), DDX53 (DEAD (Asp-Glu-Ala-Asp) box polypeptide 53), ST5 (suppression of tumorigenicity 5), LPHN1 (latrophilin 1), UBE3B

(ubiquitin protein ligase E3B), PPP6R1 (protein phosphatase 6, regulatory subunit 1), INTU (inturned planar cell polarity protein), EXT2 (exostosin glycosyltransferase 2), PIWIL2 (piwi-like RNA-mediated gene silencing 2), NDST4 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 4), GABRA1 (gamma-aminobutyric acid (GABA) A receptor, alpha 1), KIF1C (kinesin family member 1C), AKAP17A (A kinase (PRKA) anchor protein 17A), ANKH (ankylosis, progressive homolog (mouse)), AARS (alanyl-tRNA synthetase), MOV10L1 (Mov10l1, Moloney leukemia virus 10-like 1, homolog (mouse)), NDST2 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 2), SPAG6 (sperm associated antigen 6), SLC17A5 (solute carrier family 17 (anion/sugar transporter), member 5), LINS (lines homolog (Drosophila)), CLCN2 (chloride channel, voltage-sensitive 2), QARS (glutamyl-tRNA synthetase), MAB21L1 (mab-21-like 1 (C. elegans)), ZRANB2 (zinc finger, RAN-binding domain containing 2), SLC17A8 (solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 8), CEP120 (centrosomal protein 120kDa), CATSPERB (catsper channel auxiliary subunit beta), SLC01C1 (solute carrier organic anion transporter family, member 1C1), STMN4 (stathmin-like 4), MEIG1 (meiosis expressed gene 1 homolog (mouse)), ABI3 (ABI family, member 3), FJX1 (four jointed box 1 (Drosophila)), POLR2A (polymerase (RNA) II (DNA directed) polypeptide A, 220kDa), ATM (ataxia telangiectasia mutated), and/or PRKDC (protein kinase, DNA-activated, catalytic polypeptide).

[0089] Lung cancer can have, for example, 54 subtypes. Lung cancer subtype 1 can have mutation(s) in one or more or all of the following genes: TTN (titin), EGFR (epidermal growth factor receptor), NEB (nebulin), MYPN (myopalladin), ZNF423 (zinc finger protein 423), HTRA1 (HtrA serine peptidase 1), SMAD4 (SMAD family member 4), XPO1 (exportin 1 (CRM1 homolog, yeast)), PTK2B (protein tyrosine kinase 2 beta), SETD2 (SET domain containing 2), KRT1 (keratin 1), MYOM2 (myomesin 2), ANK1 (ankyrin 1, erythrocytic), PITX1 (paired-like homeodomain 1), SLC20A1 (solute carrier family 20 (phosphate transporter), member 1), CRISPLD1 (cysteine-rich secretory protein LCCL domain containing 1), EEF1B2 (eukaryotic translation elongation factor 1 beta 2), MAP3K8 (mitogen-activated protein kinase kinase kinase 8), UFD1L (ubiquitin fusion degradation 1 like (yeast)), SYP (synaptophysin), SLC11A1 (solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1), KCNAB1 (potassium voltage-gated channel, shaker-related subfamily, beta member 1), LONP1 (lon peptidase 1, mitochondrial), CCT3 (chaperonin containing TCP1, subunit 3 (gamma)), TOM1 (target of myb1 (chicken)), GAB2 (GRB2-associated binding protein 2), TUBB3 (tubulin, beta 3 class III), NAA16 (N(alpha)-acetyltransferase 16, NatA auxiliary subunit), NXF1 (nuclear RNA export factor 1), CROT (carnitine O-octanoyltransferase), BTF3 (basic transcription factor 3), RPLP2 (ribosomal protein, large, P2), EIF2S2 (eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa), MTHFD1 (methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase), ELOF1 (elongation factor 1 homolog (S. cerevisiae)),

AC007182.1, CSNK2A1 (casein kinase 2, alpha 1 polypeptide), FBXO17 (F-box protein 17), ANKRD23 (ankyrin repeat domain 23), HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1), TDG (thymine-DNA glycosylase), DNTT (deoxynucleotidyltransferase, terminal), NOS3 (nitric oxide synthase 3 (endothelial cell)), TOP2A (topoisomerase (DNA) II alpha 170kDa), TNKS2 (tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2), EBF1 (early B-cell factor 1), RHAG (Rh-associated glycoprotein), CACNA2D3 (calcium channel, voltage-dependent, alpha 2/delta subunit 3), RPS7 (ribosomal protein S7), TMBIM4 (transmembrane BAX inhibitor motif containing 4), EIF3K (eukaryotic translation initiation factor 3, subunit K), RPS26 (ribosomal protein S26), CCNH (cyclin H), PSMD7 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 7), SLC39A9 (solute carrier family 39 (zinc transporter), member 9), TUBA1C (tubulin, alpha 1c), GMCL1 (germ cell-less, spermatogenesis associated 1), RPL5 (ribosomal protein L5), PSMD2 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 2), KCNAB2 (potassium voltage-gated channel, shaker-related subfamily, beta member 2), ING4 (inhibitor of growth family, member 4), CHRNB1 (cholinergic receptor, nicotinic, beta 1 (muscle)), ATP6V1B2 (ATPase, H⁺ transporting, lysosomal 56/58kDa, V1 subunit B2), NPLOC4 (nuclear protein localization 4 homolog (S. cerevisiae)), SEL1L (sel-1 suppressor of lin-12-like (C. elegans)), AKR7A3 (aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)), UBA2 (ubiquitin-like modifier activating enzyme 2), FAM46A (family with sequence similarity 46, member A), ZAP70 (zeta-chain (TCR) associated protein kinase 70kDa), RDH8 (retinol dehydrogenase 8 (all-trans)), PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha), EIF4G2 (eukaryotic translation initiation factor 4 gamma, 2), WSCD1 (WSC domain containing 1), EIF4G1 (eukaryotic translation initiation factor 4 gamma, 1), KIF1B (kinesin family member 1B), KIF5A (kinesin family member 5A), GADD45A (growth arrest and DNA-damage-inducible, alpha), EIF3C (eukaryotic translation initiation factor 3, subunit C), EIF4E (eukaryotic translation initiation factor 4E), TUBB6 (tubulin, beta 6 class V), CEPT1 (choline/ethanolamine phosphotransferase 1), STMN1 (stathmin 1), CSH1 (chorionic somatomammotropin hormone 1 (placental lactogen)), TDP2 (tyrosyl-DNA phosphodiesterase 2), RPL14 (ribosomal protein L14), FAU (Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed), EIF3I (eukaryotic translation initiation factor 3, subunit I), CLPX (ClpX caseinolytic peptidase X homolog (E. coli)), TBCA (tubulin folding cofactor A), TCEA2 (transcription elongation factor A (SII), 2), SMAD2 (SMAD family member 2), PTPN6 (protein tyrosine phosphatase, non-receptor type 6), TREML1 (triggering receptor expressed on myeloid cells-like 1), RPL6 (ribosomal protein L6), PSMD1 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 1), CD2 (CD2 molecule), SDC3 (syndecan 3), ACAA2 (acetyl-CoA acyltransferase 2), SLAMF6 (SLAM family member 6), TCF12 (transcription factor 12), ATP5B (ATP synthase, H⁺ transporting, mitochondrial F1 complex, beta polypeptide), ERCC3 (excision repair cross-complementing rodent repair

deficiency, complementation group 3), CD5 (CD5 molecule), LRCH1 (leucine-rich repeats and calponin homology (CH) domain containing 1), FOXB1 (forkhead box B1), CTTN (cortactin), UPF3A (UPF3 regulator of nonsense transcripts homolog A (yeast)), LONP2 (lon peptidase 2, peroxisomal), SULT1A1 (sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1), UBQLN1 (ubiquilin 1), NAA15 (N(alpha)-acetyltransferase 15, NatA auxiliary subunit), RPL3L (ribosomal protein L3-like), UGT1A9 (UDP glucuronosyltransferase 1 family, polypeptide A9), SYBU (syntabulin (syntaxin-interacting)), AKD1 (adenylate kinase domain containing 1), HSD11B1 (hydroxysteroid (11-beta) dehydrogenase 1), PIPNM2 (phosphatidylinositol transfer protein, membrane-associated 2), SLC13A1 (solute carrier family 13 (sodium/sulfate symporters), member 1), USP11 (ubiquitin specific peptidase 11), DNTTIP2 (deoxynucleotidyltransferase, terminal, interacting protein 2), UBQLN2 (ubiquilin 2), EIF5B (eukaryotic translation initiation factor 5B), ZFYVE9 (zinc finger, FYVE domain containing 9), MECOM (MDS1 and EVI1 complex locus), JAK2 (Janus kinase 2), MCF2L2 (MCF.2 cell line derived transforming sequence-like 2), SV2B (synaptic vesicle glycoprotein 2B), PLD1 (phospholipase D1, phosphatidylcholine-specific), DLG2 (discs, large homolog 2 (Drosophila)), FCRL3 (Fc receptor-like 3), ARMC3 (armadillo repeat containing 3), DCC (deleted in colorectal carcinoma), PSMD9 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 9), TWF2 (twinstin, actin-binding protein, homolog 2 (Drosophila)), DTD1 (D-tyrosyl-tRNA deacylase 1), TOB1 (transducer of ERBB2, 1), PSMD13 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 13), HIST2H2AB (histone cluster 2, H2ab), NHP2 (NHP2 ribonucleoprotein), TIPIN (TIMELESS interacting protein), OTUD6B (OTU domain containing 6B), DUSP7 (dual specificity phosphatase 7), HIST1H2AA (histone cluster 1, H2aa), YY1 (YY1 transcription factor), ACO2 (aconitase 2, mitochondrial), MLST8 (MTOR associated protein, LST8 homolog (S. cerevisiae)), SKI (v-ski sarcoma viral oncogene homolog (avian)), CHEK1 (checkpoint kinase 1), HNRNPA3 (heterogeneous nuclear ribonucleoprotein A3), SHPK (sedoheptulokinase), TNPO2 (transportin 2), FLAD1 (flavin adenine dinucleotide synthetase 1), NACA2 (nascent polypeptide-associated complex alpha subunit 2), PAPSS2 (3'-phosphoadenosine 5'-phosphosulfate synthase 2), PRKD2 (protein kinase D2), ENAH (enabled homolog (Drosophila)), LCK (lymphocyte-specific protein tyrosine kinase), XRCC5 (X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)), KANK1 (KN motif and ankyrin repeat domains 1), SKIL (SKI-like oncogene), EDEM3 (ER degradation enhancer, mannosidase alpha-like 3), LEF1 (lymphoid enhancer-binding factor 1), RB1CC1 (RB1-inducible coiled-coil 1), USP13 (ubiquitin specific peptidase 13 (isopeptidase T-3)), UBE3B (ubiquitin protein ligase E3B), KCNJ4 (potassium inwardly-rectifying channel, subfamily J, member 4), NPHP1 (nephronophthisis 1 (juvenile)), GOLGA4 (golgin A4), RPTOR (regulatory associated protein of MTOR, complex 1), PML (promyelocytic leukemia),

ARHGEF6 (Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6), and/or CAD (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase).

[0090] Lung cancer subtype 2 can have mutation(s) in one or more or all of the following genes: KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog), ADAMTS2 (ADAM metalloproteinase with thrombospondin type 1 motif, 2), EIF2AK4 (eukaryotic translation initiation factor 2 alpha kinase 4), PDGFRB (platelet-derived growth factor receptor, beta polypeptide), XRN1 (5'-3' exoribonuclease 1), A2M (alpha-2-macroglobulin), ADAMTS1 (ADAM metalloproteinase with thrombospondin type 1 motif, 1), APC (adenomatous polyposis coli), CAMK2B (calcium/calmodulin-dependent protein kinase II beta), DYRK1B (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 1B), EIF4B (eukaryotic translation initiation factor 4B), GNAZ (guanine nucleotide binding protein (G protein), alpha z polypeptide), GSPT1 (G1 to S phase transition 1), HNRNP (heterogeneous nuclear ribonucleoprotein R), KALRN (kalirin, RhoGEF kinase), MAZ (MYC-associated zinc finger protein (purine-binding transcription factor)), NEDD8 (neural precursor cell expressed, developmentally down-regulated 8), PLK2 (polo-like kinase 2), PSKH2 (protein serine kinase H2), RHEB (Ras homolog enriched in brain), SSR4 (signal sequence receptor, delta), TBPL2 (TATA box binding protein like 2), TEX14 (testis expressed 14), TTK (TTK protein kinase), ADAMTS5 (ADAM metalloproteinase with thrombospondin type 1 motif, 5), ADAMTS7 (ADAM metalloproteinase with thrombospondin type 1 motif, 7), AXL (AXL receptor tyrosine kinase), BNIP2 (BCL2/adenovirus E1B 19kDa interacting protein 2), BRAF (v-raf murine sarcoma viral oncogene homolog B1), BRSK2 (BR serine/threonine kinase 2), CACNA1D (calcium channel, voltage-dependent, L type, alpha 1D subunit), CDC25A (cell division cycle 25A), CPVL (carboxypeptidase, vitellogenic-like), CRK (v-crk sarcoma virus CT10 oncogene homolog (avian)), CSNK1A1L (casein kinase 1, alpha 1-like), CTSL1 (cathepsin L1), DDX1 (DEAD (Asp-Glu-Ala-Asp) box helicase 1), DDX17 (DEAD (Asp-Glu-Ala-Asp) box helicase 17), DDX3X (DEAD (Asp-Glu-Ala-Asp) box polypeptide 3, X-linked), DLG2 (discs, large homolog 2 (Drosophila)), EDEM3 (ER degradation enhancer, mannosidase alpha-like 3), EEF1A1 (eukaryotic translation elongation factor 1 alpha 1), EEF2K (eukaryotic elongation factor-2 kinase), EIF4E2 (eukaryotic translation initiation factor 4E family member 2), FGF4 (fibroblast growth factor 4), FGFR2 (fibroblast growth factor receptor 2), GJA1 (gap junction protein, alpha 1, 43kDa), GJA4 (gap junction protein, alpha 4, 37kDa), GJA5 (gap junction protein, alpha 5, 40kDa), GLUD1 (glutamate dehydrogenase 1), GLUD2 (glutamate dehydrogenase 2), GMPR (guanosine monophosphate reductase), HAL (histidine ammonia-lyase), HEATR1 (HEAT repeat containing 1), HSD11B1 (hydroxysteroid (11-beta) dehydrogenase 1), IDH3G (isocitrate dehydrogenase 3 (NAD+) gamma), KHSRP (KH-type splicing regulatory protein), KRI1 (KRI1 homolog (S. cerevisiae)), LARP1B (La ribonucleoprotein domain family, member 1B), LATS2 (LATS, large tumor suppressor, homolog 2 (Drosophila)), LEP (leptin), LHX1 (LIM homeobox

1), LHX3 (LIM homeobox 3), LYN (v-yes-1 Yamaguchi sarcoma viral related oncogene homolog), MAGEA6 (melanoma antigen family A, 6), MAP2K4 (mitogen-activated protein kinase kinase 4), MAP3K12 (mitogen-activated protein kinase kinase kinase 12), MAP3K3 (mitogen-activated protein kinase kinase kinase 3), MAPK9 (mitogen-activated protein kinase 9), MAPKAPK3 (mitogen-activated protein kinase-activated protein kinase 3), MARK3 (MAP/microtubule affinity-regulating kinase 3), MINK1 (misshapen-like kinase 1), NAE1 (NEDD8 activating enzyme E1 subunit 1), NEK10 (NIMA-related kinase 10), NEK2 (NIMA-related kinase 2), NEK8 (NIMA-related kinase 8), NFXL1 (nuclear transcription factor, X-box binding-like 1), NUA1 (NUAK family, SNF1-like kinase, 1), PAK1 (p21 protein (Cdc42/Rac)-activated kinase 1), PAK3 (p21 protein (Cdc42/Rac)-activated kinase 3), PARD6A (par-6 partitioning defective 6 homolog alpha (C. elegans)), PDLIM5 (PDZ and LIM domain 5), PIK3CG (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit gamma), PIK3R5 (phosphoinositide-3-kinase, regulatory subunit 5), POMT1 (protein-O-mannosyltransferase 1), POMT2 (protein-O-mannosyltransferase 2), PPM1B (protein phosphatase, Mg²⁺/Mn²⁺ dependent, 1B), PPP2R2D (protein phosphatase 2, regulatory subunit B, delta), PPP4C (protein phosphatase 4, catalytic subunit), PRKAG3 (protein kinase, AMP-activated, gamma 3 non-catalytic subunit), PSMD6 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 6), PTK2 (protein tyrosine kinase 2), PTPN7 (protein tyrosine phosphatase, non-receptor type 7), RAB36 (RAB36, member RAS oncogene family), RASAL1 (RAS protein activator like 1 (GAP1 like)), RBPJ (recombination signal binding protein for immunoglobulin kappa J region), REV1 (REV1, polymerase (DNA directed)), RIT2 (Ras-like without CAAX 2), ROR2 (receptor tyrosine kinase-like orphan receptor 2), RPL36A (ribosomal protein L36a), RPL3L (ribosomal protein L3-like), RPL5 (ribosomal protein L5), RPS6KA6 (ribosomal protein S6 kinase, 90kDa, polypeptide 6), RRAGB (Ras-related GTP binding B), RRAGD (Ras-related GTP binding D), SEL1L (sel-1 suppressor of lin-12-like (C. elegans)), SLC30A1 (solute carrier family 30 (zinc transporter), member 1), SMARCB1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1), SMARCE1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1), SOAT1 (sterol O-acyltransferase 1), SOS1 (son of sevenless homolog 1 (Drosophila)), SPATA13 (spermatogenesis associated 13), SRMS (src-related kinase lacking C-terminal regulatory tyrosine and N-terminal myristylation sites), SRPK2 (SRSF protein kinase 2), TLK1 (tousled-like kinase 1), UBA3 (ubiquitin-like modifier activating enzyme 3), UGT1A9 (UDP glucuronosyltransferase 1 family, polypeptide A9), USF1 (upstream transcription factor 1), USF2 (upstream transcription factor 2, c-fos interacting), UTY (ubiquitously transcribed tetratricopeptide repeat containing, Y-linked), and/or WNT16 (wingless-type MMTV integration site family, member 16).

[0091] Lung cancer subtype 3 can have mutation(s) in one or more or all of the following genes: NAV3 (neuron navigator 3), SPTA1 (spectrin, alpha, erythrocytic 1 (elliptocytosis 2)), PTPRD

(protein tyrosine phosphatase, receptor type, D), COL11A1 (collagen, type XI, alpha 1), CTNND2 (catenin (cadherin-associated protein), delta 2), NRXN1 (neurexin 1), NEB (nebulin), MYH2 (myosin, heavy chain 2, skeletal muscle, adult), TNFR (tenascin R), SORCS1 (sortilin-related VPS10 domain containing receptor 1), BAI3 (brain-specific angiogenesis inhibitor 3), VCAN (versican), DMD (dystrophin), COL3A1 (collagen, type III, alpha 1), SORCS3 (sortilin-related VPS10 domain containing receptor 3), CDH18 (cadherin 18, type 2), EPHA5 (EPH receptor A5), CDH9 (cadherin 9, type 2 (T1-cadherin)), ADCY2 (adenylate cyclase 2 (brain)), MYH1 (myosin, heavy chain 1, skeletal muscle, adult), MYH6 (myosin, heavy chain 6, cardiac muscle, alpha), GRID2 (glutamate receptor, ionotropic, delta 2), EPHB1 (EPH receptor B1), ZFPM2 (zinc finger protein, FOG family member 2), DSCAM (Down syndrome cell adhesion molecule), CDH6 (cadherin 6, type 2, K-cadherin (fetal kidney)), COL6A3 (collagen, type VI, alpha 3), EPHA3 (EPH receptor A3), COL5A2 (collagen, type V, alpha 2), MYH7 (myosin, heavy chain 7, cardiac muscle, beta), DSCAML1 (Down syndrome cell adhesion molecule like 1), ACAN (aggrecan), COL12A1 (collagen, type XII, alpha 1), EPHA7 (EPH receptor A7), UNC5D (unc-5 homolog D (*C. elegans*)), CNGB3 (cyclic nucleotide gated channel beta 3), DTNA (dystrobrevin, alpha), CDH7 (cadherin 7, type 2), ADCY8 (adenylate cyclase 8 (brain)), GRIN2B (glutamate receptor, ionotropic, N-methyl D-aspartate 2B), DST (dystonin), CDH4 (cadherin 4, type 1, R-cadherin (retinal)), COL2A1 (collagen, type II, alpha 1), CDH2 (cadherin 2, type 1, N-cadherin (neuronal)), MYH4 (myosin, heavy chain 4, skeletal muscle), GRIK3 (glutamate receptor, ionotropic, kainate 3), ADCY5 (adenylate cyclase 5), POSTN (periostin, osteoblast specific factor), PDE1C (phosphodiesterase 1C, calmodulin-dependent 70kDa), SORL1 (sortilin-related receptor, L(DLR class) A repeats containing), GRIK2 (glutamate receptor, ionotropic, kainate 2), COL9A1 (collagen, type IX, alpha 1), GRIN3A (glutamate receptor, ionotropic, N-methyl-D-aspartate 3A), EPB41L1 (erythrocyte membrane protein band 4.1-like 1), SCN5A (sodium channel, voltage-gated, type V, alpha subunit), NCAM1 (neural cell adhesion molecule 1), CHL1 (cell adhesion molecule with homology to L1CAM (close homolog of L1)), EPHA8 (EPH receptor A8), BAI1 (brain-specific angiogenesis inhibitor 1), EPB41L3 (erythrocyte membrane protein band 4.1-like 3), GRIK1 (glutamate receptor, ionotropic, kainate 1), SRGAP3 (SLIT-ROBO Rho GTPase activating protein 3), LAMB1 (laminin, beta 1), COL9A3 (collagen, type IX, alpha 3), GRIK4 (glutamate receptor, ionotropic, kainate 4), CNGA2 (cyclic nucleotide gated channel alpha 2), GRIA2 (glutamate receptor, ionotropic, AMPA 2), NPR1 (natriuretic peptide receptor A/guanylate cyclase A (atrionatriuretic peptide receptor A)), TBX5 (T-box 5), PDZD2 (PDZ domain containing 2), C7 (complement component 7), OTOF (otoferlin), ADCY7 (adenylate cyclase 7), C8A (complement component 8, alpha polypeptide), ADARB2 (adenosine deaminase, RNA-specific, B2), MYF6 (myogenic factor 6 (herculin)), IGDCC3 (immunoglobulin superfamily, DCC subclass, member 3), EPHA4 (EPH receptor A4), PARK2 (parkinson protein 2, E3 ubiquitin protein ligase (parkin)), DRD2 (dopamine

receptor D2), NCAN (neurocan), ADCY10 (adenylate cyclase 10 (soluble)), EPHA1 (EPH receptor A1), PKP4 (plakophilin 4), C8B (complement component 8, beta polypeptide), GRIA3 (glutamate receptor, ionotropic, AMPA 3), GRIP2 (glutamate receptor interacting protein 2), GRIN2D (glutamate receptor, ionotropic, N-methyl D-aspartate 2D), EPHB3 (EPH receptor B3), ADCY6 (adenylate cyclase 6), EPHB4 (EPH receptor B4), EPB41L2 (erythrocyte membrane protein band 4.1-like 2), GRIK5 (glutamate receptor, ionotropic, kainate 5), SIGLEC1 (sialic acid binding Ig-like lectin 1, sialoadhesin), INADL (InaD-like (*Drosophila*)), MYOD1 (myogenic differentiation 1), ADAR (adenosine deaminase, RNA-specific), ST8SIA3 (ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 3), ADCY4 (adenylate cyclase 4), MYF5 (myogenic factor 5), PDE2A (phosphodiesterase 2A, cGMP-stimulated), MAGI3 (membrane associated guanylate kinase, WW and PDZ domain containing 3), FCER1A (Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide), GPR37 (G protein-coupled receptor 37 (endothelin receptor type B-like)), PTPRS (protein tyrosine phosphatase, receptor type, S), CFI (complement factor I), COL9A2 (collagen, type IX, alpha 2), ADCY3 (adenylate cyclase 3), COL6A1 (collagen, type VI, alpha 1), GATA4 (GATA binding protein 4), COL8A1 (collagen, type VIII, alpha 1), PDE4B (phosphodiesterase 4B, cAMP-specific), NIN (ninein (GSK3B interacting protein)), COMP (cartilage oligomeric matrix protein), MYOG (myogenin (myogenic factor 4)), CNGA1 (cyclic nucleotide gated channel alpha 1), CDH16 (cadherin 16, KSP-cadherin), PDE5A (phosphodiesterase 5A, cGMP-specific), NPR2 (natriuretic peptide receptor B/guanylate cyclase B (atrionatriuretic peptide receptor B)), PDE4A (phosphodiesterase 4A, cAMP-specific), CDH15 (cadherin 15, type 1, M-cadherin (myotubule)), FGF12 (fibroblast growth factor 12), SLC17A7 (solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 7), CRX (cone-rod homeobox), B4GALT4 (UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase, polypeptide 4), EFNA2 (ephrin-A2), PDE8A (phosphodiesterase 8A), VPS35 (vacuolar protein sorting 35 homolog (*S. cerevisiae*)), LGALS3BP (lectin, galactoside-binding, soluble, 3 binding protein), PPFIA3 (protein tyrosine phosphatase, receptor type, f polypeptide (PTPRF), interacting protein (liprin), alpha 3), SNTB1 (syntrophin, beta 1 (dystrophin-associated protein A1, 59kDa, basic component 1)), EPHA2 (EPH receptor A2), HAND2 (heart and neural crest derivatives expressed 2), PDE4C (phosphodiesterase 4C, cAMP-specific), GRIN1 (glutamate receptor, ionotropic, N-methyl D-aspartate 1), SYNM (synemin, intermediate filament protein), ADCY9 (adenylate cyclase 9), NEURL (neuralized homolog (*Drosophila*)), EFNA5 (ephrin-A5), EFNB3 (ephrin-B3), EFNA1 (ephrin-A1), CSRP3 (cysteine and glycine-rich protein 3 (cardiac LIM protein)), PDE7A (phosphodiesterase 7A), C14orf166 (chromosome 14 open reading frame 166), ABI1 (abl-interactor 1), EFNB1 (ephrin-B1), SLC24A1 (solute carrier family 24 (sodium/potassium/calcium exchanger), member 1), ARVCF (armadillo repeat gene deleted in velocardiofacial syndrome), C5 (complement component 5), and/or ADARB1 (adenosine deaminase, RNA-specific, B1).

[0092] Lung cancer subtype 4 can have mutation(s) in one or more or all of the following genes: NLGN4X (neuroligin 4, X-linked), PLCB1 (phospholipase C, beta 1 (phosphoinositide-specific)), KCNH7 (potassium voltage-gated channel, subfamily H (eag-related), member 7), BAI2 (brain-specific angiogenesis inhibitor 2), ROS1 (c-ros oncogene 1, receptor tyrosine kinase), UGT8 (UDP glycosyltransferase 8), SLC35A2 (solute carrier family 35 (UDP-galactose transporter), member A2), PLCL1 (phospholipase C-like 1), MRPL1 (mitochondrial ribosomal protein L1), MRPL11 (mitochondrial ribosomal protein L11), AGTR1 (angiotensin II receptor, type 1), MAS1 (MAS1 oncogene), and/or KCNH6 (potassium voltage-gated channel, subfamily H (eag-related), member 6).

[0093] Lung cancer subtype 5 can have mutation(s) in one or more or all of the following genes: POLDIP2 (polymerase (DNA-directed), delta interacting protein 2), SKIV2L2 (superkiller viralicidic activity 2-like 2 (*S. cerevisiae*)), CHEK2 (checkpoint kinase 2), TDP1 (tyrosyl-DNA phosphodiesterase 1), RAD54B (RAD54 homolog B (*S. cerevisiae*)), DIS3 (DIS3 mitotic control homolog (*S. cerevisiae*)), TTC37 (tetratricopeptide repeat domain 37), PABPC3 (poly(A) binding protein, cytoplasmic 3), EXOSC10 (exosome component 10), TSR1 (TSR1, 20S rRNA accumulation, homolog (*S. cerevisiae*)), PSME2 (proteasome (prosome, macropain) activator subunit 2 (PA28 beta)), CCNA2 (cyclin A2), RIOK2 (RIO kinase 2), PRPS1L1 (phosphoribosyl pyrophosphate synthetase 1-like 1), REL (v-rel reticuloendotheliosis viral oncogene homolog (avian)), XAB2 (XPA binding protein 2), CDT1 (chromatin licensing and DNA replication factor 1), FERMT3 (fermitin family member 3), CEBPZ (CCAAT/enhancer binding protein (C/EBP), zeta), ALX4 (ALX homeobox 4), KANK1 (KN motif and ankyrin repeat domains 1), MAT1A (methionine adenosyltransferase I, alpha), CELF4 (CUGBP, Elav-like family member 4), LSS (lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)), YTHDC1 (YTH domain containing 1), NAT10 (N-acetyltransferase 10 (GCN5-related)), CDC27 (cell division cycle 27), ZBTB20 (zinc finger and BTB domain containing 20), DCTN1 (dynactin 1), TGFBR3 (transforming growth factor, beta receptor III), CDKN2A (cyclin-dependent kinase inhibitor 2A), SLC39A6 (solute carrier family 39 (zinc transporter), member 6), CHRNA4 (cholinergic receptor, nicotinic, alpha 4 (neuronal)), UBE4B (ubiquitination factor E4B), PSME1 (proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)), BBS4 (Bardet-Biedl syndrome 4), GORASP1 (golgi reassembly stacking protein 1, 65kDa), POLR2K (polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa), RPS27A (ribosomal protein S27a), EIF4A1 (eukaryotic translation initiation factor 4A1), PSMC2 (proteasome (prosome, macropain) 26S subunit, ATPase, 2), ATF4 (activating transcription factor 4 (tax-responsive enhancer element B67)), PIAS4 (protein inhibitor of activated STAT, 4), MPST (mercaptopyruvate sulfurtransferase), SAE1 (SUMO1 activating enzyme subunit 1), GTF2E2 (general transcription factor IIE, polypeptide 2, beta 34kDa), MAGOHB (mago-nashi homolog B (*Drosophila*)), SRP68 (signal recognition particle 68kDa), SUMO1 (SMT3 suppressor of mif two 3 homolog 1 (*S.*

cerevisiae)), RFC5 (replication factor C (activator 1) 5, 36.5kDa), PSMA4 (proteasome (prosome, macropain) subunit, alpha type, 4), KPNA1 (karyopherin alpha 1 (importin alpha 5)), CCNE2 (cyclin E2), PTGES3 (prostaglandin E synthase 3 (cytosolic)), NTHL1 (nth endonuclease III-like 1 (E. coli)), DARS (aspartyl-tRNA synthetase), IMPDH2 (IMP (inosine 5'-monophosphate) dehydrogenase 2), RAD52 (RAD52 homolog (S. cerevisiae)), RMND5B (required for meiotic nuclear division 5 homolog B (S. cerevisiae)), PAN3 (PAN3 poly(A) specific ribonuclease subunit homolog (S. cerevisiae)), EDEM1 (ER degradation enhancer, mannosidase alpha-like 1), TMEM106A (transmembrane protein 106A), METAP1 (methionyl aminopeptidase 1), NR6A1 (nuclear receptor subfamily 6, group A, member 1), PSMA3 (proteasome (prosome, macropain) subunit, alpha type, 3), GSPT1 (G1 to S phase transition 1), EIF3D (eukaryotic translation initiation factor 3, subunit D), SRP19 (signal recognition particle 19kDa), MRPS9 (mitochondrial ribosomal protein S9), APEX1 (APEX nuclease (multifunctional DNA repair enzyme) 1), MCTS1 (malignant T cell amplified sequence 1), GPS1 (G protein pathway suppressor 1), TMPO (thymopoietin), METTL1 (methyltransferase like 1), POLR3H (polymerase (RNA) III (DNA directed) polypeptide H (22.9kD)), UBE2E3 (ubiquitin-conjugating enzyme E2E 3), TTL (tubulin tyrosine ligase), STOX2 (storkhead box 2), DNAJC3 (DnaJ (Hsp40) homolog, subfamily C, member 3), HNRPLL (heterogeneous nuclear ribonucleoprotein L-like), XPNPEP3 (X-prolyl aminopeptidase (aminopeptidase P) 3, putative), SETD4 (SET domain containing 4), LSM11 (LSM11, U7 small nuclear RNA associated), RPL6 (ribosomal protein L6), TYMS (thymidylate synthetase), FZR1 (fizzy/cell division cycle 20 related 1 (Drosophila)), TNPO1 (transportin 1), CCNT1 (cyclin T1), PAK1IP1 (PAK1 interacting protein 1), SYT1 (synaptotagmin I), FTSJ2 (FtsJ RNA methyltransferase homolog 2 (E. coli)), SIAH2 (siaha E3 ubiquitin protein ligase 2), COBLL1 (cordon-bleu WH2 repeat protein-like 1), APOBEC3G (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G), FOXN2 (forkhead box N2), PSMF1 (proteasome (prosome, macropain) inhibitor subunit 1 (PI31)), WDR89 (WD repeat domain 89), MSRB2 (methionine sulfoxide reductase B2), RGS13 (regulator of G-protein signaling 13), HARS (histidyl-tRNA synthetase), CHEK1 (checkpoint kinase 1), KLHDC4 (kelch domain containing 4), NFKB2 (nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)), LEO1 (Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)), POLD2 (polymerase (DNA directed), delta 2, accessory subunit), TOP1 (topoisomerase (DNA) I), NONO (non-POU domain containing, octamer-binding), COX10 (cytochrome c oxidase assembly homolog 10 (yeast)), CCNT2 (cyclin T2), MUTYH (mutY homolog (E. coli)), ZNF600 (zinc finger protein 600), UPF2 (UPF2 regulator of nonsense transcripts homolog (yeast)), RPIA (ribose 5-phosphate isomerase A), SLC13A4 (solute carrier family 13 (sodium/sulfate symporters), member 4), EIF3L (eukaryotic translation initiation factor 3, subunit L), MAF1 (MAF1 homolog (S. cerevisiae)), HNRNPF (heterogeneous nuclear ribonucleoprotein F), FAM46A (family with sequence similarity 46, member A), CWC22 (CWC22

spliceosome-associated protein homolog (*S. cerevisiae*)), CDS2 (CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2), KHDRBS3 (KH domain containing, RNA binding, signal transduction associated 3), RPL4 (ribosomal protein L4), FTSJ3 (FtsJ homolog 3 (*E. coli*)), CCNE1 (cyclin E1), GEMIN4 (gem (nuclear organelle) associated protein 4), HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1), RUSC2 (RUN and SH3 domain containing 2), CUL2 (cullin 2), KHSRP (KH-type splicing regulatory protein), EIF4B (eukaryotic translation initiation factor 4B), ZFP36 (ZFP36 ring finger protein), TBL1X (transducin (beta)-like 1X-linked), TOP3A (topoisomerase (DNA) III alpha), MFN2 (mitofusin 2), PABPC1 (poly(A) binding protein, cytoplasmic 1), STIP1 (stress-induced-phosphoprotein 1), UBQLN1 (ubiquilin 1), MAPK8IP3 (mitogen-activated protein kinase 8 interacting protein 3), PCBP3 (poly(rC) binding protein 3), CD2BP2 (CD2 (cytoplasmic tail) binding protein 2), RPA4 (replication protein A4, 30kDa), TAF1C (TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa), HSP90AB1 (heat shock protein 90kDa alpha (cytosolic), class B member 1), PLK1 (polo-like kinase 1), POLR2B (polymerase (RNA) II (DNA directed) polypeptide B, 140kDa), SUPT5H (suppressor of Ty 5 homolog (*S. cerevisiae*)), GNL3L (guanine nucleotide binding protein-like 3 (nucleolar)-like), SPAG5 (sperm associated antigen 5), SMARCAD1 (SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1), GOLGA2 (golgin A2), MCF2L (MCF.2 cell line derived transforming sequence-like), ELF1 (E74-like factor 1 (ets domain transcription factor)), DNTTIP2 (deoxynucleotidyltransferase, terminal, interacting protein 2), MECOM (MDS1 and EVI1 complex locus), CPVL (carboxypeptidase, vitellogenic-like), PC (pyruvate carboxylase), EIF4G2 (eukaryotic translation initiation factor 4 gamma, 2), CHRNB2 (cholinergic receptor, nicotinic, beta 2 (neuronal)), ABLIM3 (actin binding LIM protein family, member 3), TROAP (trophinin associated protein), RANBP6 (RAN binding protein 6), SP100 (SP100 nuclear antigen), WSCD1 (WSC domain containing 1), BRCA1 (breast cancer 1, early onset), EEF1B2 (eukaryotic translation elongation factor 1 beta 2), NUF2 (NUF2, NDC80 kinetochore complex component, homolog (*S. cerevisiae*)), ERCC6 (excision repair cross-complementing rodent repair deficiency, complementation group 6), POLR3A (polymerase (RNA) III (DNA directed) polypeptide A, 155kDa), MYO9A (myosin IXA), POLR3B (polymerase (RNA) III (DNA directed) polypeptide B), KDM5C (lysine (K)-specific demethylase 5C), and/or PCDH1 (protocadherin 1).

[0094] The term “Network-based Stratification” as described herein, includes a technique that combines genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes (**Figure 2**). Subtypes can be an informative subtype such as those correlated with a clinical phenotype. Clinical phenotype may be based on or characterized by observable and diagnosable symptoms that may be correlated to a medical treatment, practice observation or a diagnosis. Clinical phenotypes can be predicative of a survival rate, drug response, and tumor grade.

For example, reference is made to **Figure 2** which illustrates a flow chart approach for network based stratification (NBS). Network based stratification may be performed as shown in the flowchart of **Figure 2**. As shown, the first step of NBS includes a procedure to obtain a somatic mutation matrix (patient x genes mutation matrix) (200). A sample of genes from patients is then provided (210 of **Figure 2**). Genes with somatic mutations can be provided, for example, from breast, lung, prostate, ovarian, skin (melanoma, squamous cells), colorectal, pancreatic, thyroid, endometrial, uterine, bladder, and kidney, a solid tumor (leukemia, non-Hodgkin lymphoma, and tumors from a drug-resistant cancer). Genes sequences can be provided by sequencing tumor or tumor and healthy tissues which can in turn be obtained by methods known to those skilled in the art. For example, fine needle aspiration (FNA) can be performed by inserting a needled through the abdomen and directed into an organ to obtain cells from a specific tissue or a tumor, in order to obtain the genetic material. Somatic mutations can then be obtained by comparing the genetic sequences from tumor and healthy tissues.

[0095] Merely by way of example, sampling can be performed by bootstrap sampling. "Bootstrap sampling" as described herein includes a method of assigning measures of accuracy to sample estimates allowing estimation of the sampling distribution of almost any statistic using very simple methods and is known to those skilled in the art. Bootstrap sampling is a practice of estimating the properties of an estimator by measuring properties when sampling a distribution. For example, this can be performed by estimating the precision of sample statistics such as means, medians, variances and percentiles by using subsets of available data, also known as jackknifing, or drawing randomly with the replacement from a set of data points.

[0096] In some embodiments, methods are provided for network based stratification. In some embodiments, the methods provide genome scale somatic mutation profiles with a gene interaction network to assign a subject in need a subtype. In some embodiments, a method for stratification of cancer into one or more informative subtypes of a subject in need thereof is provided. In some embodiments, the method comprises obtaining sequence information from a bootstrap sample of genes from a tumor sample of the subject, projecting a mutation found within the sequence information onto a network, propagating the mutation in the network and clustering the mutation(s) so propagated so as to divide subjects with the mutation(s) into subtypes thereby stratifying cancer into informative subtypes and assigning of the subject to an informative subtype. In some embodiments, the informative subtype is a clinical phenotype. In some embodiments, the clinical phenotype is predictive of a survival rate, drug response, or a tumor grade. In some embodiments, the mutation is a somatic mutation. In some embodiments, the cancer is breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder

cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer. In some embodiments, the informative subtype is ovarian cancer subtype 1, 2, 3, or 4.

[0097] Somatic mutations for each patient can be represented as a profile of binary (1,0) states on genes, in which a '1' indicates a gene for which mutation has occurred in the tumor relative to germline (i.e. a single nucleotide base change or the insertion or deletion of bases). For each patient independently, the mutation profiles can be projected onto a human gene interaction network obtained from public databases, which are known to those skilled in the art. Human gene interaction networks that can be used for projection can include but are not limited to HumanNet, Pathway Commons, STRING, and other human gene interaction networks known to those skilled in the art. Next, the technique of network propagation can then be applied to spread the influence of each subsampled mutation profile over its network neighborhood (**220 of Figure 2**). Reference is made to **Figure 3**, in which an example illustrating smoothing of patient somatic mutation profiles over a molecular interaction network is demonstrated. As shown, is the result, a 'network-smoothed' profile (also known as a 'transformed' profile) in which the state of each gene is no longer binary but reflects its network proximity to the mutated genes in that patient, along a continuous range [0,1] (**Figure 3**).

[0098] For example, a "network-smoothed" or transformed profile may include a continuous range of values for the one or more or all of following genes for ovarian cancer subtype 1: TTN (titin), NEB (nebulin), AP1G2 (adaptor-related protein complex 1, gamma 2 subunit), SYNRG (synergyn, gamma), SPTBN4 (spectrin, beta, non-erythrocytic 4), ANK1 (ankyrin 1, erythrocytic), SLC12A8 (solute carrier family 12 (potassium/chloride transporters), member 8), CACNA1A (calcium channel, voltage-dependent, P/Q type, alpha 1A subunit), MPP1 (membrane protein, palmitoylated 1, 55kDa), RIMS1 (regulating synaptic membrane exocytosis 1), SCML2 (sex comb on midleg-like 2 (Drosophila)), CIDEA (cell death-inducing DFFA-like effector b), RHAG (Rh-associated glycoprotein), GAD2 (glutamate decarboxylase 2 (pancreatic islets and brain, 65kDa)), FGFR4 (fibroblast growth factor receptor 4), ATP6V0D2 (ATPase, H⁺ transporting, lysosomal 38kDa, V0 subunit d2), MYOM2 (myomesin 2), AP1M1 (adaptor-related protein complex 1, mu 1 subunit), INA (internexin neuronal intermediate filament protein, alpha), KCNH7 (potassium voltage-gated channel, subfamily H (eag-related), member 7), SLC2A5 (solute carrier family 2 (facilitated glucose/fructose transporter), member 5), PTPRN (protein tyrosine phosphatase, receptor type, N), FGF4 (fibroblast growth factor 4), FGF19 (fibroblast growth factor 19), FGF18 (fibroblast growth factor 18), RASL10A (RAS-like, family 10, member A), GAS2 (growth arrest-specific 2), FGF6 (fibroblast growth factor 6), FGF17 (fibroblast growth factor 17), FGF8 (fibroblast growth factor 8 (androgen-induced)), FGF23 (fibroblast growth factor 23), GYPC (glycophorin C (Gerbich blood group)), GYPE (glycophorin E (MNS blood group)), FGF5 (fibroblast growth factor 5), ANKRD23 (ankyrin repeat domain 23), MYPN (myopalladin), ZRSR2 (zinc finger

(CCCH type), RNA-binding motif and serine/arginine rich 2), FGF9 (fibroblast growth factor 9 (glia-activating factor)), ANKRD1 (ankyrin repeat domain 1 (cardiac muscle)), FRS2 (fibroblast growth factor receptor substrate 2), SLC5A1 (solute carrier family 5 (sodium/glucose cotransporter), member 1), DFFB (DNA fragmentation factor, 40kDa, beta polypeptide (caspase-activated DNase)), FGF20 (fibroblast growth factor 20), FRS3 (fibroblast growth factor receptor substrate 3), DFFA (DNA fragmentation factor, 45kDa, alpha polypeptide), NLGN4Y (neuroligin 4, Y-linked), RFPL3 (ret finger protein-like 3), NAGLU (N-acetylglucosaminidase, alpha), GYPA (glycophorin A (MNS blood group)), RS1 (retinoschisin 1), RFPL1 (ret finger protein-like 1), RFPL2 (ret finger protein-like 2), FGFRL1 (fibroblast growth factor receptor-like 1), ECE1 (endothelin converting enzyme 1), FGF1 (fibroblast growth factor 1 (acidic)), FGF2 (fibroblast growth factor 2 (basic)), ATXN10 (ataxin 10), TMEM169 (transmembrane protein 169), TMCO4 (transmembrane and coiled-coil domains 4), UNC80 (unc-80 homolog (C. elegans)), AP1B1 (adaptor-related protein complex 1, beta 1 subunit), AP1S1 (adaptor-related protein complex 1, sigma 1 subunit), GAD1 (glutamate decarboxylase 1 (brain, 67kDa)), SLC32A1 (solute carrier family 32 (GABA vesicular transporter), member 1), SGCE (sarcoglycan, epsilon), FGF13 (fibroblast growth factor 13), NLGN4X (neuroligin 4, X-linked), AES (amino-terminal enhancer of split), GAS2L1 (growth arrest-specific 2 like 1), FCER2 (Fc fragment of IgE, low affinity II, receptor for (CD23)), CD47 (CD47 molecule), MFSD6 (major facilitator superfamily domain containing 6), PLCL1 (phospholipase C-like 1), PTPRN2 (protein tyrosine phosphatase, receptor type, N polypeptide 2), PHKA2 (phosphorylase kinase, alpha 2 (liver)), GYPB (glycophorin B (MNS blood group)), SLC4A1 (solute carrier family 4, anion exchanger, member 1 (erythrocyte membrane protein band 3, Diego blood group)), ICAM4 (intercellular adhesion molecule 4 (Landsteiner-Wiener blood group)), KIF1B (kinesin family member 1B), AP1S2 (adaptor-related protein complex 1, sigma 2 subunit), CADPS (Ca⁺⁺-dependent secretion activator), STX2 (syntaxin 2), SCML1 (sex comb on midleg-like 1 (Drosophila)), PEG10 (paternally expressed 10), FGFR3 (fibroblast growth factor receptor 3), AP1M2 (adaptor-related protein complex 1, mu 2 subunit), TLE3 (transducin-like enhancer of split 3 (E(sp1) homolog, Drosophila)), WRB (tryptophan rich basic protein), KLF1 (Kruppel-like factor 1 (erythroid)), MYBPC3 (myosin binding protein C, cardiac), FAHD1 (fumarylacetoacetate hydrolase domain containing 1), ARMC3 (armadillo repeat containing 3), AP1G1 (adaptor-related protein complex 1, gamma 1 subunit), SSFA2 (sperm specific antigen 2), CD58 (CD58 molecule), STOML1 (stomatin (EPB72)-like 1), AANAT (aralkylamine N-acetyltransferase), CASP9 (caspase 9, apoptosis-related cysteine peptidase), BIRC7 (baculoviral IAP repeat containing 7), FOXA3 (forkhead box A3), NKX2-8 (NK2 homeobox 8), SLC4A7 (solute carrier family 4, sodium bicarbonate cotransporter, member 7), NET1 (neuroepithelial cell transforming 1), ATP6V0A4 (ATPase, H⁺ transporting, lysosomal V0 subunit a4), ITGA4 (integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)), KCNH6 (potassium voltage-gated

channel, subfamily H (eag-related), member 6), TLE4 (transducin-like enhancer of split 4 (E(sp1) homolog, *Drosophila*)), CABP1 (calcium binding protein 1), CACNA2D3 (calcium channel, voltage-dependent, alpha 2/delta subunit 3), TACC3 (transforming, acidic coiled-coil containing protein 3), GATA1 (GATA binding protein 1 (globin transcription factor 1)), KIF5A (kinesin family member 5A), APAF1 (apoptotic peptidase activating factor 1), TLE2 (transducin-like enhancer of split 2 (E(sp1) homolog, *Drosophila*)), TLE1 (transducin-like enhancer of split 1 (E(sp1) homolog, *Drosophila*)), FGFR1 (fibroblast growth factor receptor 1), TACC1 (transforming, acidic coiled-coil containing protein 1), CASP3 (caspase 3, apoptosis-related cysteine peptidase), STX1A (syntaxin 1A (brain)), EBF1 (early B-cell factor 1), ZNF423 (zinc finger protein 423), RIPK3 (receptor-interacting serine-threonine kinase 3), PITX2 (paired-like homeodomain 2), ZIC3 (Zic family member 3), and/or PICALM (phosphatidylinositol binding clathrin assembly protein).

[0099] For ovarian cancer subtype 2, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: TP53 (tumor protein p53), BRCA1 (breast cancer 1, early onset), BRCA2 (breast cancer 2, early onset), CREBBP (CREB binding protein), USP7 (ubiquitin specific peptidase 7 (herpes virus-associated)), ST18 (suppression of tumorigenicity 18 (breast carcinoma) (zinc finger protein)), NUP155 (nucleoporin 155kDa), NUP160 (nucleoporin 160kDa), SLC11A1 (solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1), PRRC2C (proline-rich coiled-coil 2C), DMBT1 (deleted in malignant brain tumors 1), NUP62 (nucleoporin 62kDa), RANBP2 (RAN binding protein 2), CRMP1 (collapsin response mediator protein 1), TPR (translocated promoter region, nuclear basket protein), TNPO3 (transportin 3), CEBPE (CCAAT/enhancer binding protein (C/EBP), epsilon), NUP133 (nucleoporin 133kDa), MAP1B (microtubule-associated protein 1B), TP53BP2 (tumor protein p53 binding protein, 2), ADAMTS4 (ADAM metalloproteinase with thrombospondin type 1 motif, 4), PTEN (phosphatase and tensin homolog), NUP188 (nucleoporin 188kDa), NUP214 (nucleoporin 214kDa), NUP153 (nucleoporin 153kDa), DPYSL5 (dihydropyrimidinase-like 5), N6AMT1 (N-6 adenine-specific DNA methyltransferase 1 (putative)), NUP98 (nucleoporin 98kDa), DPYSL4 (dihydropyrimidinase-like 4), FOSB (FBJ murine osteosarcoma viral oncogene homolog B), NUP205 (nucleoporin 205kDa), CUL9 (cullin 9), MDM4 (Mdm4 p53 binding protein homolog (mouse)), USP30 (ubiquitin specific peptidase 30), EP300 (E1A binding protein p300), CHEK2 (checkpoint kinase 2), NF2 (neurofibromin 2 (merlin)), SMURF1 (SMAD specific E3 ubiquitin protein ligase 1), SIRT5 (sirtuin 5), NUP35 (nucleoporin 35kDa), POM121 (POM121 transmembrane nucleoporin), NUP85 (nucleoporin 85kDa), ARID5B (AT rich interactive domain 5B (MRF1-like)), SIRT6 (sirtuin 6), CREB3 (cAMP responsive element binding protein 3), NUP93 (nucleoporin 93kDa), BATF3 (basic leucine zipper transcription factor, ATF-like 3), SENP2 (SUMO1/sentrin/SMT3 specific peptidase 2), EGR2 (early growth response 2), PSIP1 (PC4 and

SFRS1 interacting protein 1), RAE1 (RAE1 RNA export 1 homolog (*S. pombe*)), BRIP1 (BRCA1 interacting protein C-terminal helicase 1), NUP107 (nucleoporin 107kDa), MAP1A (microtubule-associated protein 1A), FMOD (fibromodulin), BATF (basic leucine zipper transcription factor, ATF-like), IPO7 (importin 7), GABPA (GA binding protein transcription factor, alpha subunit 60kDa), ATF1 (activating transcription factor 1), SIRT1 (sirtuin 1), E4F1 (E4F transcription factor 1), THNSL2 (threonine synthase-like 2 (*S. cerevisiae*)), NPEPPS (aminopeptidase puromycin sensitive), NUP37 (nucleoporin 37kDa), DDX1 (DEAD (Asp-Glu-Ala-Asp) box helicase 1), GARS (glycyl-tRNA synthetase), KPNB1 (karyopherin (importin) beta 1), RPRD1A (regulation of nuclear pre-mRNA domain containing 1A), EGR1 (early growth response 1), EVI2A (ecotropic viral integration site 2A), TBL1XR1 (transducin (beta)-like 1 X-linked receptor 1), FOS (FBJ murine osteosarcoma viral oncogene homolog), CCNH (cyclin H), SMAD4 (SMAD family member 4), SSTR3 (somatostatin receptor 3), SDCBP2 (syndecan binding protein (syntenin) 2), MED25 (mediator complex subunit 25), ADAMTS2 (ADAM metalloproteinase with thrombospondin type 1 motif, 2), ACVRL1 (activin A receptor type II-like 1), PHAX (phosphorylated adaptor for RNA export), XPO1 (exportin 1 (CRM1 homolog, yeast)), NUPL1 (nucleoporin like 1), SIRT4 (sirtuin 4), SIRT7 (sirtuin 7), NUP88 (nucleoporin 88kDa), NUPL2 (nucleoporin like 2), EGR3 (early growth response 3), EGR4 (early growth response 4), DPYSL3 (dihydropyrimidinase-like 3), CEBPG (CCAAT/enhancer binding protein (C/EBP), gamma), RANBP3L (RAN binding protein 3-like), NUP50 (nucleoporin 50kDa), SSR2 (signal sequence receptor, beta (translocon-associated protein beta)), SSR1 (signal sequence receptor, alpha), RANGRF (RAN guanine nucleotide release factor), RANBP1 (RAN binding protein 1), FAM82B, CEP68 (centrosomal protein 68kDa), NBR1 (neighbor of BRCA1 gene 1), RANBP3 (RAN binding protein 3), DPYSL2 (dihydropyrimidinase-like 2), DYNC1LI1 (dynein, cytoplasmic 1, light intermediate chain 1), NUTF2 (nuclear transport factor 2), ZC3H15 (zinc finger CCCH-type containing 15), NUP54 (nucleoporin 54kDa), CREB3L4 (cAMP responsive element binding protein 3-like 4), ATF3 (activating transcription factor 3), TFAP2A (transcription factor AP-2 alpha (activating enhancer binding protein 2 alpha)), SLC11A2 (solute carrier family 11 (proton-coupled divalent metal ion transporters), member 2), IPO8 (importin 8), HMGA1 (high mobility group AT-hook 1), DCTN4 (dynactin 4 (p62)), XPOT (exportin, tRNA), MTRF1L (mitochondrial translational release factor 1-like), TRMT112 (tRNA methyltransferase 11-2 homolog (*S. cerevisiae*)), MTRF1 (mitochondrial translational release factor 1), FOSL2 (FOS-like antigen 2), SPOP (speckle-type POZ protein), SERTAD1 (SERTA domain containing 1), UBE2CBP, TBL1Y (transducin (beta)-like 1, Y-linked), RPRD1B (regulation of nuclear pre-mRNA domain containing 1B), TGFB3 (transforming growth factor, beta 3), NAB1 (NGFI-A binding protein 1 (EGR1 binding protein 1)), NAB2 (NGFI-A binding protein 2 (EGR1 binding protein 2)), ATF5 (activating transcription factor 5), PPIF (peptidylprolyl isomerase F), BANF1 (barrier to autointegration factor 1),

CDKN2A (cyclin-dependent kinase inhibitor 2A), JUND (jun D proto-oncogene), SDSL (serine dehydratase-like), ANP32A (acidic (leucine-rich) nuclear phosphoprotein 32 family, member A), HMGA2 (high mobility group AT-hook 2), ALX4 (ALX homeobox 4), MSX2 (msh homeobox 2), MYCN (v-myc myelocytomatosis viral related oncogene, neuroblastoma derived (avian)), MDM2 (MDM2 oncogene, E3 ubiquitin protein ligase), TBL1X (transducin (beta)-like 1X-linked), SEH1L (SEH1-like (*S. cerevisiae*)), HEMK1 (HemK methyltransferase family member 1), UBE2L3 (ubiquitin-conjugating enzyme E2L 3), ATF4 (activating transcription factor 4 (tax-responsive enhancer element B67)), MIOS (missing oocyte, meiosis regulator, homolog (*Drosophila*)), AAAS (achalasia, adrenocortical insufficiency, alacrimia), CREB5 (cAMP responsive element binding protein 5), MAPRE1 (microtubule-associated protein, RP/EB family, member 1), JUNB (jun B proto-oncogene), WWP1 (WW domain containing E3 ubiquitin protein ligase 1), HARS2 (histidyl-tRNA synthetase 2, mitochondrial), BRAP (BRCA1 associated protein), PIAS4 (protein inhibitor of activated STAT, 4), WDR5 (WD repeat domain 5), SLMO2 (slowmo homolog 2 (*Drosophila*)), MXD3 (MAX dimerization protein 3), OXCT2 (3-oxoacid CoA transferase 2), CENPV (centromere protein V), UCHL3 (ubiquitin carboxyl-terminal esterase L3 (ubiquitin thiolesterase)), MAX (MYC associated factor X), IPO5 (importin 5), TEAD1 (TEA domain family member 1 (SV40 transcriptional enhancer factor)), MXD4 (MAX dimerization protein 4), RRAD (Ras-related associated with diabetes), WWP2 (WW domain containing E3 ubiquitin protein ligase 2), XRCC2 (X-ray repair complementing defective repair in Chinese hamster cells 2), RAD51 (RAD51 homolog (*S. cerevisiae*)), UBE2I (ubiquitin-conjugating enzyme E2I), BCL2L1 (BCL2-like 1), HBG2 (hemoglobin, gamma G), RAN (RAN, member RAS oncogene family), ASAP2 (ArfGAP with SH3 domain, ankyrin repeat and PH domain 2), KPNA2 (karyopherin alpha 2 (RAG cohort 1, importin alpha 1)), JUN (jun proto-oncogene), PTMA (prothymosin, alpha), ATM (ataxia telangiectasia mutated), NBR2 (neighbor of BRCA1 gene 2 (non-protein coding)), and/or UBR5 (ubiquitin protein ligase E3 component n-recognin 5).

[0100] For ovarian cancer subtype 3, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: AHNK (AHNAK nucleoprotein), RPS6KL1 (ribosomal protein S6 kinase-like 1), IFNA13 (interferon, alpha 13), IRF8 (interferon regulatory factor 8), HDAC5 (histone deacetylase 5), PIGR (polymeric immunoglobulin receptor), IFNA10 (interferon, alpha 10), DEDD2 (death effector domain containing 2), DEDD (death effector domain containing), IFNA17 (interferon, alpha 17), IFNA1 (interferon, alpha 1), TAL2 (T-cell acute lymphocytic leukemia 2), LYL1 (lymphoblastic leukemia derived sequence 1), IDO1 (indoleamine 2,3-dioxygenase 1), ADAMDEC1 (ADAM-like, decysin 1), NOD1 (nucleotide-binding oligomerization domain containing 1), CARD6 (caspase recruitment domain family, member 6), LMO1 (LIM domain only 1 (rhombotin 1)), HDHD1 (haloacid dehalogenase-like hydrolase domain containing 1), PNPLA4

(patatin-like phospholipase domain containing 4), TAL1 (T-cell acute lymphocytic leukemia 1), IRF1 (interferon regulatory factor 1), CASP10 (caspase 10, apoptosis-related cysteine peptidase), CFLAR (CASP8 and FADD-like apoptosis regulator), RIPK2 (receptor-interacting serine-threonine kinase 2), NOD2 (nucleotide-binding oligomerization domain containing 2), SDC2 (syndecan 2), GSTP1 (glutathione S-transferase pi 1), TICAM1 (toll-like receptor adaptor molecule 1), IRF3 (interferon regulatory factor 3), IL18 (interleukin 18 (interferon-gamma-inducing factor)), ANKRD11 (ankyrin repeat domain 11), IL1F5, IL18BP (interleukin 18 binding protein), ETHE1 (ethylmalonic encephalopathy 1), RNF25 (ring finger protein 25), FLAD1 (flavin adenine dinucleotide synthetase 1), RELA (v-rel reticuloendotheliosis viral oncogene homolog A (avian)), KAL1 (Kallmann syndrome 1 sequence), SCD (stearoyl-CoA desaturase (delta-9-desaturase)), GLDC (glycine dehydrogenase (decarboxylating)), FASLG (Fas ligand (TNF superfamily, member 6)), PRKCSH (protein kinase C substrate 80K-H), IL1R1 (interleukin 1 receptor, type I), NRIP1 (nuclear receptor interacting protein 1), PLG (plasminogen), MECR (mitochondrial trans-2-enoyl-CoA reductase), ZYX (zyxin), AIFM1 (apoptosis-inducing factor, mitochondrion-associated, 1), and/or TMED7 (transmembrane emp24 protein transport domain containing 7).

[0101] For ovarian cancer subtype 4, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: MYH4 (myosin, heavy chain 4, skeletal muscle), MYH2 (myosin, heavy chain 2, skeletal muscle, adult), SWAP70 (SWAP switching B-cell complex 70kDa subunit), FGF10 (fibroblast growth factor 10), FOLR1 (folate receptor 1 (adult)), GLUD2 (glutamate dehydrogenase 2), GYG1 (glycogenin 1), GYS1 (glycogen synthase 1 (muscle)), PHKA1 (phosphorylase kinase, alpha 1 (muscle)), PRKAG1 (protein kinase, AMP-activated, gamma 1 non-catalytic subunit), ROM1 (retinal outer segment membrane protein 1), AC008810.1, ADRA1B (adrenoceptor alpha 1B), APOD (apolipoprotein D), APOE (apolipoprotein E), BLK (B lymphoid tyrosine kinase), CCNL1 (cyclin L1), CD34 (CD34 molecule), CD52 (CD52 molecule), CD55 (CD55 molecule, decay accelerating factor for complement (Cromer blood group)), CHL1 (cell adhesion molecule with homology to L1CAM (close homolog of L1)), CLU (clusterin), EMR1 (egf-like module containing, mucin-like, hormone receptor-like 1), FGF3 (fibroblast growth factor 3), FTH1 (ferritin, heavy polypeptide 1), GRK1 (G protein-coupled receptor kinase 1), GRK5 (G protein-coupled receptor kinase 5), GSK3B (glycogen synthase kinase 3 beta), GYS2 (glycogen synthase 2 (liver)), HSD17B1 (hydroxysteroid (17-beta) dehydrogenase 1), KCNV2 (potassium channel, subfamily V, member 2), LIPC (lipase, hepatic), LRP8 (low density lipoprotein receptor-related protein 8, apolipoprotein e receptor), MEN1 (multiple endocrine neoplasia I), MFAP4 (microfibrillar-associated protein 4), MLX (MLX, MAX dimerization protein), MLXIPL (MLX interacting protein-like), MPP1 (membrane protein, palmitoylated 1, 55kDa), PHKB (phosphorylase kinase, beta), PHKG1 (phosphorylase kinase, gamma 1 (muscle)),

PHKG2 (phosphorylase kinase, gamma 2 (testis)), PRKAB1 (protein kinase, AMP-activated, beta 1 non-catalytic subunit), PRKAB2 (protein kinase, AMP-activated, beta 2 non-catalytic subunit), PRKAG3 (protein kinase, AMP-activated, gamma 3 non-catalytic subunit), PYGM (phosphorylase, glycogen, muscle), SAG (S-antigen; retina and pineal gland (arrestin)), SH3GL2 (SH3-domain GRB2-like 2), and/or SMARCB1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1).

[0102] For uterine cancer subtype 1, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: TAPBP (TAP binding protein (tapasin)), HIST1H1C (histone cluster 1, H1c), ARID3A (AT rich interactive domain 3A (BRIGHT-like)), ATF3 (activating transcription factor 3), HLA-A (major histocompatibility complex, class I, A), PHB (prohibitin), PADI4 (peptidyl arginine deiminase, type IV), TP53 (tumor protein p53), EPCAM (epithelial cell adhesion molecule), DYRK2 (dual-specificity tyrosine-(Y)-phosphorylation regulated kinase 2), PRDM1 (PR domain containing 1, with ZNF domain), RB1CC1 (RB1-inducible coiled-coil 1), RNF20 (ring finger protein 20, E3 ubiquitin protein ligase), IRF5 (interferon regulatory factor 5), PPP1R13B (protein phosphatase 1, regulatory subunit 13B), SIK1 (salt-inducible kinase 1), CUL9 (cullin 9), PRKAB1 (protein kinase, AMP-activated, beta 1 non-catalytic subunit), RNF144B (ring finger protein 144B), CD59 (CD59 molecule, complement regulatory protein), DUSP1 (dual specificity phosphatase 1), BCL2L12 (BCL2-like 12 (proline rich)), JMY (junction mediating and regulatory protein, p53 cofactor), BAI1 (brain-specific angiogenesis inhibitor 1), CD82 (CD82 molecule), RRAD (Ras-related associated with diabetes), CAMK2D (calcium/calmodulin-dependent protein kinase II delta), PAK3 (p21 protein (Cdc42/Rac)-activated kinase 3), FBXO11 (F-box protein 11), C12orf5 (chromosome 12 open reading frame 5), ZACN (zinc activated ligand-gated ion channel), E4F1 (E4F transcription factor 1), CHEK1 (checkpoint kinase 1), UCHL1 (ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)), CSE1L (CSE1 chromosome segregation 1-like (yeast)), STEAP3 (STEAP family member 3, metalloredoxase), SUMO1 (SMT3 suppressor of mif two 3 homolog 1 (*S. cerevisiae*)), CSNK1G3 (casein kinase 1, gamma 3), RAD54L (RAD54-like (*S. cerevisiae*)), COL18A1 (collagen, type XVIII, alpha 1), PIAS2 (protein inhibitor of activated STAT, 2), FAS (Fas (TNF receptor superfamily, member 6)), CTSL1 (cathepsin L1), LMLN (leishmanolysin-like (metallopeptidase M8 family)), HIC1 (hypermethylated in cancer 1), PLK3 (polo-like kinase 3), RPRM (reprimin, TP53 dependent G2 arrest mediator candidate), IFI16 (interferon, gamma-inducible protein 16), GNL3 (guanine nucleotide binding protein-like 3 (nucleolar)), NOX1 (NADPH oxidase 1), WWOX (WW domain containing oxidoreductase), ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)), HYAL2 (hyaluronoglucosaminidase 2), TNK2 (tyrosine kinase, non-receptor, 2), SERTAD4 (SERTA domain containing 4), ZCCHC8 (zinc finger, CCHC domain containing 8), CEP41 (centrosomal protein 41kDa),

EXOSC5 (exosome component 5), SKIV2L2 (superkiller viralicidic activity 2-like 2 (*S. cerevisiae*)), SLMAP (sarcolemma associated protein), NEUROD6 (neuronal differentiation 6), HABP4 (hyaluronan binding protein 4), DLX2 (distal-less homeobox 2), PPP2R1A (protein phosphatase 2, regulatory subunit A, alpha), PPP2R5C (protein phosphatase 2, regulatory subunit B', gamma), PPP2R3A (protein phosphatase 2, regulatory subunit B'', alpha), NDN (necdin, melanoma antigen (MAGE) family member), PRR14 (proline rich 14), POLR2J (polymerase (RNA) II (DNA directed) polypeptide J, 13.3kDa), PAF1 (Paf1, RNA polymerase II associated factor, homolog (*S. cerevisiae*)), CSNK1E (casein kinase 1, epsilon), TAF9B (TAF9B RNA polymerase II, TATA box binding protein (TBP)-associated factor, 31kDa), TAF3 (TAF3 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 140kDa), PRMT5 (protein arginine methyltransferase 5), ANKS1B (ankyrin repeat and sterile alpha motif domain containing 1B), MMS19 (MMS19 nucleotide excision repair homolog (*S. cerevisiae*)), INTS6 (integrator complex subunit 6), BRD7 (bromodomain containing 7), TAF5L (TAF5-like RNA polymerase II, p300/CBP-associated factor (PCAF)-associated factor, 65kDa), GTF2A1 (general transcription factor IIA, 1, 19/37kDa), GTF2E1 (general transcription factor IIE, polypeptide 1, alpha 56kDa), HNRNPA1 (heterogeneous nuclear ribonucleoprotein A1), NFKBIA (nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha), ERCC2 (excision repair cross-complementing rodent repair deficiency, complementation group 2), and/or C19orf2 (unconventional prefoldin RPB5 interactor).

[0103] For uterine cancer subtype 2, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: PTEN (phosphatase and tensin homolog), CTNNB1 (catenin (cadherin-associated protein), beta 1, 88kDa), ARID1A (AT rich interactive domain 1A (SWI-like)), PIK3R1 (phosphoinositide-3-kinase, regulatory subunit 1 (alpha)), MUC4 (mucin 4, cell surface associated), CTCF (CCCTC-binding factor (zinc finger protein)), FGFR2 (fibroblast growth factor receptor 2), PRG4 (p53-responsive gene 4), SOX17 (SRY (sex determining region Y)-box 17), EIF3C (eukaryotic translation initiation factor 3, subunit C), IRS4 (insulin receptor substrate 4), INVS (inversin), TLE1 (transducin-like enhancer of split 1 (E(sp1) homolog, *Drosophila*)), TNK1 (TRAF2 and NCK interacting kinase), INPPL1 (inositol polyphosphate phosphatase-like 1), PIKFYVE (phosphoinositide kinase, FYVE finger containing), PDYN (prodynorphin), C4BPA (complement component 4 binding protein, alpha), PIK3CB (phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit beta), AGAP2 (ArfGAP with GTPase domain, ankyrin repeat and PH domain 2), FGF13 (fibroblast growth factor 13), NKTR (natural killer-tumor recognition sequence), CYSLTR2 (cysteinyl leukotriene receptor 2), MCERS1 (microspherule protein 1), SOX9 (SRY (sex determining region Y)-box 9), FGFR4 (fibroblast growth factor receptor 4), FIG4 (FIG4 homolog, SAC1 lipid phosphatase domain containing (*S. cerevisiae*)), CDON (cell adhesion associated, oncogene regulated), INPP4A (inositol polyphosphate-4-phosphatase, type I, 107kDa), DMBT1 (deleted in malignant brain

tumors 1), PARD3 (par-3 partitioning defective 3 homolog (*C. elegans*)), SMARCA2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2), ARID1B (AT rich interactive domain 1B (SWI1-like)), IHH (indian hedgehog), RHEB (Ras homolog enriched in brain), OPRL1 (opiate receptor-like 1), CDKN2A (cyclin-dependent kinase inhibitor 2A), KITLG (KIT ligand), FPR2 (formyl peptide receptor 2), FIGF (c-fos induced growth factor (vascular endothelial growth factor D)), TACR2 (tachykinin receptor 2), IGFBP2 (insulin-like growth factor binding protein 2, 36kDa), EIF3J (eukaryotic translation initiation factor 3, subunit J), PROKR1 (prokineticin receptor 1), SMARCD2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 2), THRA (thyroid hormone receptor, alpha), ERFFI1 (ERBB receptor feedback inhibitor 1), INPP5B (inositol polyphosphate-5-phosphatase, 75kDa), ITK (IL2-inducible T-cell kinase), PMPCA (peptidase (mitochondrial processing) alpha), CSNK1A1L (casein kinase 1, alpha 1-like), INPP5J (inositol polyphosphate-5-phosphatase J), EPHB4 (EPH receptor B4), PROKR2 (prokineticin receptor 2), EPHA2 (EPH receptor A2), DMP1 (dentin matrix acidic phosphoprotein 1), VWCE (von Willebrand factor C and EGF domains), FGF12 (fibroblast growth factor 12), FRK (fyn-related kinase), MIB2 (mindbomb E3 ubiquitin protein ligase 2), LIMA1 (LIM domain and actin binding 1), MUC7 (mucin 7, secreted), PI4KB (phosphatidylinositol 4-kinase, catalytic, beta), MTMR3 (myotubularin related protein 3), SMARCC1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1), MAST3 (microtubule associated serine/threonine kinase 3), GEN1 (Gen endonuclease homolog 1 (*Drosophila*)), PIK3C2B (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 beta), INPP4B (inositol polyphosphate-4-phosphatase, type II, 105kDa), ESRP1 (epithelial splicing regulatory protein 1), PTPRF (protein tyrosine phosphatase, receptor type, F), PDC (phosducin), FGF2 (fibroblast growth factor 2 (basic)), CRH (corticotropin releasing hormone), IL17A (interleukin 17A), CRK (v-crk sarcoma virus CT10 oncogene homolog (avian)), FIGLA (folliculogenesis specific basic helix-loop-helix), SLC9A3R2 (solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 2), WNT4 (wingless-type MMTV integration site family, member 4), CD83 (CD83 molecule), MED31 (mediator complex subunit 31), SUB1 (SUB1 homolog (*S. cerevisiae*)), SH2D2A (SH2 domain containing 2A), FHL2 (four and a half LIM domains 2), NANOG (Nanog homeobox), SLC9A3R1 (solute carrier family 9, subfamily A (NHE3, cation proton antiporter 3), member 3 regulator 1), IGF2 (insulin-like growth factor 2 (somatomedin A)), WNT1 (wingless-type MMTV integration site family, member 1), IL2RA (interleukin 2 receptor, alpha), C17orf72 (chromosome 17 open reading frame 72), NOG (noggin), PRDX1 (peroxiredoxin 1), SYT8 (synaptotagmin VIII), F2RL2 (coagulation factor II (thrombin) receptor-like 2), TWIST2 (twist basic helix-loop-helix transcription factor 2), PDPK1 (3-phosphoinositide dependent protein kinase-1), PI4K2A (phosphatidylinositol 4-kinase type 2 alpha), CACYBP (calcyclin binding protein), DVL1 (dishevelled, dsh homolog 1 (*Drosophila*)), CD28 (CD28

molecule), THEM4 (thioesterase superfamily member 4), CSNK2A2 (casein kinase 2, alpha prime polypeptide), XCR1 (chemokine (C motif) receptor 1), FZD8 (frizzled family receptor 8), FZD5 (frizzled family receptor 5), ICOS (inducible T-cell co-stimulator), ICOSLG (inducible T-cell co-stimulator ligand), FGF9 (fibroblast growth factor 9 (glia-activating factor)), MED16 (mediator complex subunit 16), MDFIC (MyoD family inhibitor domain containing), TBC1D10A (TBC1 domain family, member 10A), ADRA1D (adrenoceptor alpha 1D), AVPR1B (arginine vasopressin receptor 1B), MED4 (mediator complex subunit 4), ASCC1 (activating signal cointegrator 1 complex subunit 1), FZD1 (frizzled family receptor 1), RELB (v-rel reticuloendotheliosis viral oncogene homolog B), TNFRSF13B (tumor necrosis factor receptor superfamily, member 13B), TNFRSF9 (tumor necrosis factor receptor superfamily, member 9), TMEM55B (transmembrane protein 55B), STX4 (syntaxin 4), AP1M1 (adaptor-related protein complex 1, mu 1 subunit), C5orf4, TCN1 (transcobalamin I (vitamin B12 binding protein, R binder family)), HCK (hemopoietic cell kinase), MS4A1 (membrane-spanning 4-domains, subfamily A, member 1), FRS3 (fibroblast growth factor receptor substrate 3), SHB (Src homology 2 domain containing adaptor protein B), GFRA1 (GDNF family receptor alpha 1), GNA14 (guanine nucleotide binding protein (G protein), alpha 14), GP6 (glycoprotein VI (platelet)), IRAK2 (interleukin-1 receptor-associated kinase 2), MED27 (mediator complex subunit 27), IL23R (interleukin 23 receptor), TCF7 (transcription factor 7 (T-cell specific, HMG-box)), SMARCB1 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1), TMEM55A (transmembrane protein 55A), IPMK (inositol polyphosphate multikinase), CTNNAL1 (catenin (cadherin-associated protein), alpha-like 1), PRKCI (protein kinase C, iota), EPHB3 (EPH receptor B3), FLT4 (fms-related tyrosine kinase 4), TLR1 (toll-like receptor 1), DDX17 (DEAD (Asp-Glu-Ala-Asp) box helicase 17), WNT16 (wingless-type MMTV integration site family, member 16), PIP4K2A (phosphatidylinositol-5-phosphate 4-kinase, type II, alpha), CARD10 (caspase recruitment domain family, member 10), FOXO4 (forkhead box O4), IGF2BP1 (insulin-like growth factor 2 mRNA binding protein 1), PIK3R2 (phosphoinositide-3-kinase, regulatory subunit 2 (beta)), CDX4 (caudal type homeobox 4), WNT2B (wingless-type MMTV integration site family, member 2B), PIK3R3 (phosphoinositide-3-kinase, regulatory subunit 3 (gamma)), PLCD4 (phospholipase C, delta 4), PLCB2 (phospholipase C, beta 2), BMP7 (bone morphogenetic protein 7), PIK3R5 (phosphoinositide-3-kinase, regulatory subunit 5), IBSP (integrin-binding sialoprotein), PLCZ1 (phospholipase C, zeta 1), BCL9L (B-cell CLL/lymphoma 9-like), PBRM1 (polybromo 1), TLE4 (transducin-like enhancer of split 4 (E(sp1) homolog, Drosophila)), ARHGAP10 (Rho GTPase activating protein 10), AXIN2 (axin 2), PLCB3 (phospholipase C, beta 3 (phosphatidylinositol-specific)), MAGI3 (membrane associated guanylate kinase, WW and PDZ domain containing 3), SMARCC2 (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2), VAV3 (vav 3 guanine nucleotide exchange factor), PIK3C2G

(phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 gamma), ROCK1 (Rho-associated, coiled-coil containing protein kinase 1), INPP5D (inositol polyphosphate-5-phosphatase, 145kDa), MAST1 (microtubule associated serine/threonine kinase 1), PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha), INPP5F (inositol polyphosphate-5-phosphatase F), PTPRM (protein tyrosine phosphatase, receptor type, M), SYNJ2 (synaptojanin 2), and/or MED22 (mediator complex subunit 22).

[0104] For uterine cancer subtype 3, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: TTN (titin), NEB (nebulin), DST (dystonin), FAT3 (FAT tumor suppressor homolog 3 (Drosophila)), SYNE1 (spectrin repeat containing, nuclear envelope 1), DMD (dystrophin), RYR1 (ryanodine receptor 1 (skeletal)), MKI67 (antigen identified by monoclonal antibody Ki-67), FAT4 (FAT tumor suppressor homolog 4 (Drosophila)), TAF1 (TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa), DNAH5 (dynein, axonemal, heavy chain 5), DNAH3 (dynein, axonemal, heavy chain 3), LAMA2 (laminin, alpha 2), ASPM (asp (abnormal spindle) homolog, microcephaly associated (Drosophila)), CEP290 (centrosomal protein 290kDa), DYNC1H1 (dynein, cytoplasmic 1, heavy chain 1), DNAH2 (dynein, axonemal, heavy chain 2), TRRAP (transformation/transcription domain-associated protein), NIPBL (Nipped-B homolog (Drosophila)), NBEA (neurobeachin), SYNE2 (spectrin repeat containing, nuclear envelope 2), DNAH7 (dynein, axonemal, heavy chain 7), DNAH1 (dynein, axonemal, heavy chain 1), NSD1 (nuclear receptor binding SET domain protein 1), DNAH8 (dynein, axonemal, heavy chain 8), JAK1 (Janus kinase 1), ITPR1 (inositol 1,4,5-trisphosphate receptor, type 1), APC (adenomatous polyposis coli), GPR98 (G protein-coupled receptor 98), FN1 (fibronectin 1), SPEN (spen homolog, transcriptional regulator (Drosophila)), SVIL (supervillin), DNAH9 (dynein, axonemal, heavy chain 9), TPR (translocated promoter region, nuclear basket protein), CREBBP (CREB binding protein), PCNT (pericentrin), PKD1L1 (polycystic kidney disease 1 like 1), HCFC1 (host cell factor C1 (VP16-accessory protein)), MYO7A (myosin VIIA), DNAH17 (dynein, axonemal, heavy chain 17), EP300 (E1A binding protein p300), ANK1 (ankyrin 1, erythrocytic), RIF1 (RAP1 interacting factor homolog (yeast)), RB1 (retinoblastoma 1), NCOR2 (nuclear receptor corepressor 2), NRXN1 (neurexin 1), DICER1 (dicer 1, ribonuclease type III), CENPF (centromere protein F, 350/400kDa), ROS1 (c-ros oncogene 1, receptor tyrosine kinase), PKD1L2 (polycystic kidney disease 1-like 2), CENPE (centromere protein E, 312kDa), SPAG17 (sperm associated antigen 17), CAD (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase), TOP2A (topoisomerase (DNA) II alpha 170kDa), NCOR1 (nuclear receptor corepressor 1), NUP98 (nucleoporin 98kDa), CASK (calcium/calmodulin-dependent serine protein kinase (MAGUK family)), HDAC6 (histone deacetylase 6), CLASP1 (cytoplasmic linker associated protein 1), KIF4A (kinesin family member 4A), ATP1A4 (ATPase, Na⁺/K⁺ transporting,

alpha 4 polypeptide), EGF (epidermal growth factor), SEC24D (SEC24 family, member D (*S. cerevisiae*)), CKAP5 (cytoskeleton associated protein 5), DLGAP2 (discs, large (*Drosophila*) homolog-associated protein 2), CATSPER1 (cation channel, sperm associated 1), C9orf174, TRPM8 (transient receptor potential cation channel, subfamily M, member 8), TJP1 (tight junction protein 1), BRCA1 (breast cancer 1, early onset), TRIP11 (thyroid hormone receptor interactor 11), DCTN1 (dynactin 1), SHANK2 (SH3 and multiple ankyrin repeat domains 2), TDRD1 (tudor domain containing 1), NDST1 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1), ABI3BP (ABI family, member 3 (NESH) binding protein), SPAG16 (sperm associated antigen 16), PTCHD1 (patched domain containing 1), ASMTL (acetylserotonin O-methyltransferase-like), CATSPERG (catsper channel auxiliary subunit gamma), UBN1 (ubiquitin 1), EFHC1 (EF-hand domain (C-terminal) containing 1), ABCC1 (ATP-binding cassette, sub-family C (CFTR/MRP), member 1), PIWIL1 (piwi-like RNA-mediated gene silencing 1), SLC16A2 (solute carrier family 16, member 2 (thyroid hormone transporter)), DARS2 (aspartyl-tRNA synthetase 2, mitochondrial), ANKFY1 (ankyrin repeat and FYVE domain containing 1), CDK17 (cyclin-dependent kinase 17), SUN1 (Sad1 and UNC84 domain containing 1), SPICE1 (spindle and centriole associated protein 1), DDX53 (DEAD (Asp-Glu-Ala-Asp) box polypeptide 53), ST5 (suppression of tumorigenicity 5), LPHN1 (latrophilin 1), UBE3B (ubiquitin protein ligase E3B), PPP6R1 (protein phosphatase 6, regulatory subunit 1), INTU (inturned planar cell polarity protein), EXT2 (exostosin glycosyltransferase 2), PIWIL2 (piwi-like RNA-mediated gene silencing 2), NDST4 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 4), GABRA1 (gamma-aminobutyric acid (GABA) A receptor, alpha 1), KIF1C (kinesin family member 1C), AKAP17A (A kinase (PRKA) anchor protein 17A), ANKH (ankylosis, progressive homolog (mouse)), AARS (alanyl-tRNA synthetase), MOV10L1 (Mov10l1, Moloney leukemia virus 10-like 1, homolog (mouse)), NDST2 (N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 2), SPAG6 (sperm associated antigen 6), SLC17A5 (solute carrier family 17 (anion/sugar transporter), member 5), LINS (lines homolog (*Drosophila*)), CLCN2 (chloride channel, voltage-sensitive 2), QARS (glutamyl-tRNA synthetase), MAB21L1 (mab-21-like 1 (*C. elegans*)), ZRANB2 (zinc finger, RAN-binding domain containing 2), SLC17A8 (solute carrier family 17 (sodium-dependent inorganic phosphate cotransporter), member 8), CEP120 (centrosomal protein 120kDa), CATSPERB (catsper channel auxiliary subunit beta), SLC01C1 (solute carrier organic anion transporter family, member 1C1), STMN4 (stathmin-like 4), MEIG1 (meiosis expressed gene 1 homolog (mouse)), ABI3 (ABI family, member 3), and/or FJX1 (four jointed box 1 (*Drosophila*)).

[0105] For lung cancer subtype 1, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: TTN (titin), EGFR (epidermal growth factor receptor), NEB (nebulin), MYPN (myopalladin), ZNF423 (zinc finger protein 423), HTRA1 (HtrA serine peptidase 1), SMAD4 (SMAD family member 4), XPO1 (exportin 1 (CRM1

homolog, yeast)), PTK2B (protein tyrosine kinase 2 beta), SETD2 (SET domain containing 2), KRT1 (keratin 1), MYOM2 (myomesin 2), ANK1 (ankyrin 1, erythrocytic), PITX1 (paired-like homeodomain 1), SLC20A1 (solute carrier family 20 (phosphate transporter), member 1), CRISPLD1 (cysteine-rich secretory protein LCCL domain containing 1), EEF1B2 (eukaryotic translation elongation factor 1 beta 2), MAP3K8 (mitogen-activated protein kinase kinase kinase 8), UFD1L (ubiquitin fusion degradation 1 like (yeast)), SYP (synaptophysin), SLC11A1 (solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1), KCNAB1 (potassium voltage-gated channel, shaker-related subfamily, beta member 1), LONP1 (lon peptidase 1, mitochondrial), CCT3 (chaperonin containing TCP1, subunit 3 (gamma)), TOM1 (target of myb1 (chicken)), GAB2 (GRB2-associated binding protein 2), TUBB3 (tubulin, beta 3 class III), NAA16 (N(alpha)-acetyltransferase 16, NatA auxiliary subunit), NXF1 (nuclear RNA export factor 1), CROT (carnitine O-octanoyltransferase), BTF3 (basic transcription factor 3), RPLP2 (ribosomal protein, large, P2), EIF2S2 (eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa), MTHFD1 (methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1, methenyltetrahydrofolate cyclohydrolase, formyltetrahydrofolate synthetase), ELOF1 (elongation factor 1 homolog (S. cerevisiae)), AC007182.1, CSNK2A1 (casein kinase 2, alpha 1 polypeptide), FBXO17 (F-box protein 17), ANKRD23 (ankyrin repeat domain 23), HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1), TDG (thymine-DNA glycosylase), DNTT (deoxynucleotidyltransferase, terminal), NOS3 (nitric oxide synthase 3 (endothelial cell)), TOP2A (topoisomerase (DNA) II alpha 170kDa), TNKS2 (tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase 2), EBF1 (early B-cell factor 1), RHAG (Rh-associated glycoprotein), CACNA2D3 (calcium channel, voltage-dependent, alpha 2/delta subunit 3), RPS7 (ribosomal protein S7), TMBIM4 (transmembrane BAX inhibitor motif containing 4), EIF3K (eukaryotic translation initiation factor 3, subunit K), RPS26 (ribosomal protein S26), CCNH (cyclin H), PSMD7 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 7), SLC39A9 (solute carrier family 39 (zinc transporter), member 9), TUBA1C (tubulin, alpha 1c), GMCL1 (germ cell-less, spermatogenesis associated 1), RPL5 (ribosomal protein L5), PSMD2 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 2), KCNAB2 (potassium voltage-gated channel, shaker-related subfamily, beta member 2), ING4 (inhibitor of growth family, member 4), CHRN1 (cholinergic receptor, nicotinic, beta 1 (muscle)), ATP6V1B2 (ATPase, H+ transporting, lysosomal 56/58kDa, V1 subunit B2), NPLOC4 (nuclear protein localization 4 homolog (S. cerevisiae)), SEL1L (sel-1 suppressor of lin-12-like (C. elegans)), AKR7A3 (aldo-keto reductase family 7, member A3 (aflatoxin aldehyde reductase)), UBA2 (ubiquitin-like modifier activating enzyme 2), FAM46A (family with sequence similarity 46, member A), ZAP70 (zeta-chain (TCR) associated protein kinase 70kDa), RDH8 (retinol dehydrogenase 8 (all-trans)), PIK3C2A (phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha), EIF4G2 (eukaryotic translation initiation factor 4 gamma, 2), WSCD1 (WSC domain containing

1), EIF4G1 (eukaryotic translation initiation factor 4 gamma, 1), KIF1B (kinesin family member 1B), KIF5A (kinesin family member 5A), GADD45A (growth arrest and DNA-damage-inducible, alpha), EIF3C (eukaryotic translation initiation factor 3, subunit C), EIF4E (eukaryotic translation initiation factor 4E), TUBB6 (tubulin, beta 6 class V), CEPT1 (choline/ethanolamine phosphotransferase 1), STMN1 (stathmin 1), CSH1 (chorionic somatomammotropin hormone 1 (placental lactogen)), TDP2 (tyrosyl-DNA phosphodiesterase 2), RPL14 (ribosomal protein L14), FAU (Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed), EIF3I (eukaryotic translation initiation factor 3, subunit I), CLPX (ClpX caseinolytic peptidase X homolog (E. coli)), TBCA (tubulin folding cofactor A), TCEA2 (transcription elongation factor A (SII), 2), SMAD2 (SMAD family member 2), PTPN6 (protein tyrosine phosphatase, non-receptor type 6), TREML1 (triggering receptor expressed on myeloid cells-like 1), RPL6 (ribosomal protein L6), PSMD1 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 1), CD2 (CD2 molecule), SDC3 (syndecan 3), ACAA2 (acetyl-CoA acyltransferase 2), SLAMF6 (SLAM family member 6), TCF12 (transcription factor 12), ATP5B (ATP synthase, H⁺ transporting, mitochondrial F1 complex, beta polypeptide), ERCC3 (excision repair cross-complementing rodent repair deficiency, complementation group 3), CD5 (CD5 molecule), LRCH1 (leucine-rich repeats and calponin homology (CH) domain containing 1), FOXB1 (forkhead box B1), CTTN (cortactin), UPF3A (UPF3 regulator of nonsense transcripts homolog A (yeast)), LONP2 (lon peptidase 2, peroxisomal), SULT1A1 (sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1), UBQLN1 (ubiquilin 1), NAA15 (N(alpha)-acetyltransferase 15, NatA auxiliary subunit), RPL3L (ribosomal protein L3-like), UGT1A9 (UDP glucuronosyltransferase 1 family, polypeptide A9), SYBU (syntabulin (syntaxin-interacting)), AKD1 (adenylate kinase domain containing 1), HSD11B1 (hydroxysteroid (11-beta) dehydrogenase 1), PITPNM2 (phosphatidylinositol transfer protein, membrane-associated 2), SLC13A1 (solute carrier family 13 (sodium/sulfate symporters), member 1), USP11 (ubiquitin specific peptidase 11), DNTTIP2 (deoxynucleotidyltransferase, terminal, interacting protein 2), UBQLN2 (ubiquilin 2), EIF5B (eukaryotic translation initiation factor 5B), ZFYVE9 (zinc finger, FYVE domain containing 9), MECOM (MDS1 and EVI1 complex locus), JAK2 (Janus kinase 2), MCF2L2 (MCF.2 cell line derived transforming sequence-like 2), SV2B (synaptic vesicle glycoprotein 2B), PLD1 (phospholipase D1, phosphatidylcholine-specific), DLG2 (discs, large homolog 2 (Drosophila)), FCRL3 (Fc receptor-like 3), ARMC3 (armadillo repeat containing 3), DCC (deleted in colorectal carcinoma), PSMD9 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 9), TWF2 (twinfilin, actin-binding protein, homolog 2 (Drosophila)), DTD1 (D-tyrosyl-tRNA deacylase 1), TOB1 (transducer of ERBB2, 1), PSMD13 (proteasome (prosome, macropain) 26S subunit, non-ATPase, 13), HIST2H2AB (histone cluster 2, H2ab), NHP2 (NHP2 ribonucleoprotein), TIPIN (TIMELESS interacting protein), OTUD6B (OTU domain containing 6B), DUSP7 (dual specificity phosphatase 7), HIST1H2AA (histone cluster 1, H2aa),

YY1 (YY1 transcription factor), ACO2 (aconitase 2, mitochondrial), MLST8 (MTOR associated protein, LST8 homolog (*S. cerevisiae*)), SKI (v-ski sarcoma viral oncogene homolog (avian)), CHEK1 (checkpoint kinase 1), HNRNPA3 (heterogeneous nuclear ribonucleoprotein A3), SHPK (sedoheptulokinase), TNPO2 (transportin 2), FLAD1 (flavin adenine dinucleotide synthetase 1), NACA2 (nascent polypeptide-associated complex alpha subunit 2), PAPSS2 (3'-phosphoadenosine 5'-phosphosulfate synthase 2), PRKD2 (protein kinase D2), ENAH (enabled homolog (*Drosophila*)), LCK (lymphocyte-specific protein tyrosine kinase), XRCC5 (X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining)), KANK1 (KN motif and ankyrin repeat domains 1), SKIL (SKI-like oncogene), EDEM3 (ER degradation enhancer, mannosidase alpha-like 3), LEF1 (lymphoid enhancer-binding factor 1), RB1CC1 (RB1-inducible coiled-coil 1), USP13 (ubiquitin specific peptidase 13 (isopeptidase T-3)), UBE3B (ubiquitin protein ligase E3B), KCNJ4 (potassium inwardly-rectifying channel, subfamily J, member 4), NPHP1 (nephronophthisis 1 (juvenile)), GOLGA4 (golgin A4), RPTOR (regulatory associated protein of MTOR, complex 1), PML (promyelocytic leukemia), ARHGEF6 (Rac/Cdc42 guanine nucleotide exchange factor (GEF) 6), and/or CAD (carbamoyl-phosphate synthetase 2, aspartate transcarbamylase, and dihydroorotase).

[0106] For lung cancer subtype 2, a “network-smoothed” or transformed profile may include a continuous range of values for the following genes: KRAS (v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog).

[0107] For lung cancer subtype 3, a “network-smoothed” or transformed profile may include a continuous range of values for the following genes: NAV3 (neuron navigator 3).

[0108] For lung cancer subtype 4, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: NLGN4X (neuroligin 4, X-linked), PLCB1 (phospholipase C, beta 1 (phosphoinositide-specific)), KCNH7 (potassium voltage-gated channel, subfamily H (eag-related), member 7), BAI2 (brain-specific angiogenesis inhibitor 2), ROS1 (c-ros oncogene 1, receptor tyrosine kinase), UGT8 (UDP glycosyltransferase 8), SLC35A2 (solute carrier family 35 (UDP-galactose transporter), member A2), PLCL1 (phospholipase C-like 1), MRPL1 (mitochondrial ribosomal protein L1), MRPL11 (mitochondrial ribosomal protein L11), AGTR1 (angiotensin II receptor, type 1), MAS1 (MAS1 oncogene), KCNH6 (potassium voltage-gated channel, subfamily H (eag-related), member 6), AGT (angiotensinogen (serpin peptidase inhibitor, clade A, member 8)), NPPB (natriuretic peptide B), MEP1A (meprin A, alpha (PABA peptide hydrolase)), MEP1B (meprin A, beta), C1R (complement component 1, r subcomponent), and/or MRPL10 (mitochondrial ribosomal protein L10).

[0109] For lung cancer subtype 5, a “network-smoothed” or transformed profile may include a continuous range of values for one or more or all of the following genes: POLDIP2 (polymerase (DNA-

directed), delta interacting protein 2), SKIV2L2 (superkiller viralicidic activity 2-like 2 (*S. cerevisiae*)), CHEK2 (checkpoint kinase 2), TDP1 (tyrosyl-DNA phosphodiesterase 1), RAD54B (RAD54 homolog B (*S. cerevisiae*)), DIS3 (DIS3 mitotic control homolog (*S. cerevisiae*)), TTC37 (tetratricopeptide repeat domain 37), PABPC3 (poly(A) binding protein, cytoplasmic 3), EXOSC10 (exosome component 10), TSR1 (TSR1, 20S rRNA accumulation, homolog (*S. cerevisiae*)), PSME2 (proteasome (prosome, macropain) activator subunit 2 (PA28 beta)), CCNA2 (cyclin A2), RIOK2 (RIO kinase 2), PRPS1L1 (phosphoribosyl pyrophosphate synthetase 1-like 1), REL (v-rel reticuloendotheliosis viral oncogene homolog (avian)), XAB2 (XPA binding protein 2), CDT1 (chromatin licensing and DNA replication factor 1), FERMT3 (fermitin family member 3), CEBPZ (CCAAT/enhancer binding protein (C/EBP), zeta), ALX4 (ALX homeobox 4), KANK1 (KN motif and ankyrin repeat domains 1), MAT1A (methionine adenosyltransferase I, alpha), CELF4 (CUGBP, Elav-like family member 4), LSS (lanosterol synthase (2,3-oxidosqualene-lanosterol cyclase)), YTHDC1 (YTH domain containing 1), NAT10 (N-acetyltransferase 10 (GCN5-related)), CDC27 (cell division cycle 27), ZBTB20 (zinc finger and BTB domain containing 20), DCTN1 (dynactin 1), TGFBR3 (transforming growth factor, beta receptor III), CDKN2A (cyclin-dependent kinase inhibitor 2A), SLC39A6 (solute carrier family 39 (zinc transporter), member 6), CHRNA4 (cholinergic receptor, nicotinic, alpha 4 (neuronal)), UBE4B (ubiquitination factor E4B), PSME1 (proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)), BBS4 (Bardet-Biedl syndrome 4), GORASP1 (golgi reassembly stacking protein 1, 65kDa), POLR2K (polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa), RPS27A (ribosomal protein S27a), EIF4A1 (eukaryotic translation initiation factor 4A1), ATF4 (activating transcription factor 4 (tax-responsive enhancer element B67)), PSMC2 (proteasome (prosome, macropain) 26S subunit, ATPase, 2), PIAS4 (protein inhibitor of activated STAT, 4), MPST (mercaptopyruvate sulfurtransferase), SAE1 (SUMO1 activating enzyme subunit 1), GTF2E2 (general transcription factor IIE, polypeptide 2, beta 34kDa), MAGOHB (mago-nashi homolog B (*Drosophila*)), SRP68 (signal recognition particle 68kDa), SUMO1 (SMT3 suppressor of mif two 3 homolog 1 (*S. cerevisiae*)), RFC5 (replication factor C (activator 1) 5, 36.5kDa), PSMA4 (proteasome (prosome, macropain) subunit, alpha type, 4), KPNA1 (karyopherin alpha 1 (importin alpha 5)), CCNE2 (cyclin E2), PTGES3 (prostaglandin E synthase 3 (cytosolic)), NTHL1 (nth endonuclease III-like 1 (*E. coli*)), DARS (aspartyl-tRNA synthetase), IMPDH2 (IMP (inosine 5'-monophosphate) dehydrogenase 2), RAD52 (RAD52 homolog (*S. cerevisiae*)), RMND5B (required for meiotic nuclear division 5 homolog B (*S. cerevisiae*)), PAN3 (PAN3 poly(A) specific ribonuclease subunit homolog (*S. cerevisiae*)), EDEM1 (ER degradation enhancer, mannosidase alpha-like 1), TMEM106A (transmembrane protein 106A), METAP1 (methionyl aminopeptidase 1), NR6A1 (nuclear receptor subfamily 6, group A, member 1), PSMA3 (proteasome (prosome, macropain) subunit, alpha type, 3), GSPT1 (G1 to S phase transition 1), EIF3D (eukaryotic translation initiation factor 3, subunit D),

SRP19 (signal recognition particle 19kDa), MRPS9 (mitochondrial ribosomal protein S9), APEX1 (APEX nuclease (multifunctional DNA repair enzyme) 1), MCTS1 (malignant T cell amplified sequence 1), GPS1 (G protein pathway suppressor 1), TMPO (thymopoietin), METTL1 (methyltransferase like 1), POLR3H (polymerase (RNA) III (DNA directed) polypeptide H (22.9kD)), UBE2E3 (ubiquitin-conjugating enzyme E2E 3), TTL (tubulin tyrosine ligase), STOX2 (storkhead box 2), DNAJC3 (DnaJ (Hsp40) homolog, subfamily C, member 3), HNRPLL (heterogeneous nuclear ribonucleoprotein L-like), XPNPEP3 (X-prolyl aminopeptidase (aminopeptidase P) 3, putative), SETD4 (SET domain containing 4), LSM11 (LSM11, U7 small nuclear RNA associated), RPL6 (ribosomal protein L6), TYMS (thymidylate synthetase), FZR1 (fizzy/cell division cycle 20 related 1 (Drosophila)), TNPO1 (transportin 1), CCNT1 (cyclin T1), PAK1IP1 (PAK1 interacting protein 1), SYT1 (synaptotagmin I), FTSJ2 (FtsJ RNA methyltransferase homolog 2 (E. coli)), SIAH2 (siah E3 ubiquitin protein ligase 2), COBLL1 (cordon-bleu WH2 repeat protein-like 1), APOBEC3G (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G), FOXN2 (forkhead box N2), PSMF1 (proteasome (prosome, macropain) inhibitor subunit 1 (PI31)), WDR89 (WD repeat domain 89), MSRB2 (methionine sulfoxide reductase B2), RGS13 (regulator of G-protein signaling 13), HARS (histidyl-tRNA synthetase), CHEK1 (checkpoint kinase 1), KLHDC4 (kelch domain containing 4), NFKB2 (nuclear factor of kappa light polypeptide gene enhancer in B-cells 2 (p49/p100)), LEO1 (Leo1, Paf1/RNA polymerase II complex component, homolog (S. cerevisiae)), POLD2 (polymerase (DNA directed), delta 2, accessory subunit), TOP1 (topoisomerase (DNA) I), NONO (non-POU domain containing, octamer-binding), COX10 (cytochrome c oxidase assembly homolog 10 (yeast)), CCNT2 (cyclin T2), MUTYH (mutY homolog (E. coli)), ZNF600 (zinc finger protein 600), UPF2 (UPF2 regulator of nonsense transcripts homolog (yeast)), RPIA (ribose 5-phosphate isomerase A), SLC13A4 (solute carrier family 13 (sodium/sulfate symporters), member 4), EIF3L (eukaryotic translation initiation factor 3, subunit L), MAF1 (MAF1 homolog (S. cerevisiae)), HNRNPF (heterogeneous nuclear ribonucleoprotein F), FAM46A (family with sequence similarity 46, member A), CWC22 (CWC22 spliceosome-associated protein homolog (S. cerevisiae)), CDS2 (CDP-diacylglycerol synthase (phosphatidate cytidylyltransferase) 2), KHDRBS3 (KH domain containing, RNA binding, signal transduction associated 3), RPL4 (ribosomal protein L4), FTSJ3 (FtsJ homolog 3 (E. coli)), CCNE1 (cyclin E1), GEMIN4 (gem (nuclear organelle) associated protein 4), HSP90AA1 (heat shock protein 90kDa alpha (cytosolic), class A member 1), RUSC2 (RUN and SH3 domain containing 2), CUL2 (cullin 2), KHSRP (KH-type splicing regulatory protein), EIF4B (eukaryotic translation initiation factor 4B), ZFP36 (ZFP36 ring finger protein), TBL1X (transducin (beta)-like 1X-linked), TOP3A (topoisomerase (DNA) III alpha), MFN2 (mitofusin 2), PABPC1 (poly(A) binding protein, cytoplasmic 1), STIP1 (stress-induced-phosphoprotein 1), UBQLN1 (ubiquilin 1), MAPK8IP3 (mitogen-activated protein kinase 8 interacting protein 3), PCBP3 (poly(rC) binding protein 3), CD2BP2 (CD2 (cytoplasmic

tail) binding protein 2), RPA4 (replication protein A4, 30kDa), TAF1C (TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa), HSP90AB1 (heat shock protein 90kDa alpha (cytosolic), class B member 1), PLK1 (polo-like kinase 1), POLR2B (polymerase (RNA) II (DNA directed) polypeptide B, 140kDa), SUPT5H (suppressor of Ty 5 homolog (*S. cerevisiae*)), GNL3L (guanine nucleotide binding protein-like 3 (nucleolar)-like), SPAG5 (sperm associated antigen 5), SMARCAD1 (SWI/SNF-related, matrix-associated actin-dependent regulator of chromatin, subfamily a, containing DEAD/H box 1), GOLGA2 (golgin A2), MCF2L (MCF.2 cell line derived transforming sequence-like), ELF1 (E74-like factor 1 (ets domain transcription factor)), DNTTIP2 (deoxynucleotidyltransferase, terminal, interacting protein 2), MECOM (MDS1 and EVI1 complex locus), CPVL (carboxypeptidase, vitellogenic-like), PC (pyruvate carboxylase), EIF4G2 (eukaryotic translation initiation factor 4 gamma, 2), CHRNB2 (cholinergic receptor, nicotinic, beta 2 (neuronal)), ABLIM3 (actin binding LIM protein family, member 3), TROAP (troponin associated protein), RANBP6 (RAN binding protein 6), SP100 (SP100 nuclear antigen), WSCD1 (WSC domain containing 1), BRCA1 (breast cancer 1, early onset), EEF1B2 (eukaryotic translation elongation factor 1 beta 2), NUF2 (NUF2, NDC80 kinetochore complex component, homolog (*S. cerevisiae*)), ERCC6 (excision repair cross-complementing rodent repair deficiency, complementation group 6), POLR3A (polymerase (RNA) III (DNA directed) polypeptide A, 155kDa), MYO9A (myosin IXA), POLR3B (polymerase (RNA) III (DNA directed) polypeptide B), KDM5C (lysine (K)-specific demethylase 5C), PCDH1 (protocadherin 1), ANAPC2 (anaphase promoting complex subunit 2), ANAPC1 (anaphase promoting complex subunit 1), HMGB3 (high mobility group box 3), and/or CHCHD2 (coiled-coil-helix-coiled-coil-helix domain containing 2).

[0110] In another embodiment, a “network-smoothed” or transformed profile for a subtype of a cancer or tumor may include a continuous range of values for one or more or all of the genes identified as being mutated and associated for respective subtype of a cancer or tumor, as provided above.

[0111] In another embodiment, the mutation may be in the nucleic acid, DNA or RNA; the mutation may be in a protein coding region, non-protein coding region (such as untranslated region, 5' UTR or 3' UTR), transcriptional regulatory region (such as promoter or enhancer), RNA processing signals (such as splicing signals, 5' splice donor, 3' splice acceptor, splicing branch site, polyadenylation signal), transcribed region of a gene, non-transcribed region of a gene, RNA structural elements and/or other genetic elements.

[0112] In another embodiment, mutation may be determined by characterizing nucleic acid, DNA or RNA, conceptual translation of a nucleic acid sequence, and/or expressed proteins. In addition to mutation at the nucleic acid level, epigenetic modification changes as well as changes in RNA modification and/or post-translational modification of proteins are anticipated as being useful biological

features for network-based stratification of subject(s) with a cancer or tumor or assigning subject of interest to a subtype of a cancer or tumor.

[0113] Following ‘network smoothing’, patient profiles may be clustered into a predefined number of subtypes (for example, $k = 2 \dots 12$) using e.g., an unsupervised learning technique of non-negative matrix factorization (NMF) (**230 of Figure 2**). “Non-negative matrix factorization” refers to a group of algorithms in a multivariate analysis and linear algebra where a matrix is factorized into two matrices, with the property that all three matrices have no negative elements. “Unsupervised learning” includes the finding of hidden structure in unlabeled data, as they are unlabeled, there is no error or reward signal to evaluate a potential solution. Approaches to unsupervised learning include clustering (for example, k-means, mixture models, hierarchical clustering), hidden Markov models, blind signal separation using feature extraction techniques for dimensionality reduction such as principal component analysis, independent component analysis, non-negative matrix factorization, and singular value decomposition, and other approaches known to those skilled in the art. “Supervised learning” as described herein is a task of inferring a function from labeled training data, in which the training data consist of a set of training examples. In supervised learning each example is a pair consisting of an input object and a desired output value. A supervised learning algorithm analyzes the training data and produces an inferred function that can be used for mapping new examples. Merely by way of example, reference is made to **Figure 4**, which illustrates the clustering of mutation profiles using non-negative matrix factorization (NMF) regulated by a network. The input data matrix (F) is decomposed into the product of two matrices, one of subtype prototypes (W) and an assignments matrix of each mutation profile to the prototypes (H). The decomposition attempts to minimize the objective function shown, which includes a network regularization term on the subtype prototypes.

[0114] In some embodiments, methods for stratification of cancer into one or more informative subtypes of a subject in need thereof are provided. In some embodiments, the method is carried out by an informatics platform. In some embodiments, the informatics platform is a bioinformatics platform comprising a computer and software. In some embodiments, the software uses supervised learning and/or unsupervised learning methods.

[0115] For network based stratification, a variant of NMF which encourages the selection of gene sets supporting each subtype based on high network connectivity (NetNMF) may be used. To promote robust cluster assignments a technique of consensus clustering can also be used, in which the above procedure is repeated for e.g., about 1000 different subsamples in which subsets of about 80% of patients and genes are drawn randomly without replacement from the entire data set (**210**, **220**, and **230** may be repeated). The results of all the e.g., about 1000 runs may be aggregated into a (patient x patient) co-occurrence matrix, which summarizes the frequency of times each pair of patients has co-segregated

into the same cluster. This co-occurrence matrix may be then clustered to recover a final stratification of the patients into clusters / subtypes (240)

[0116] Merely by way of example, reference is made to **Figure 5**, which illustrates the final tumor subtypes that can be obtained from the consensus (majority) assignments of each tumor after 1000 applications of this procedure to samples of the original data set. As shown is an aggregate consensus matrix (patient x patient) (250). A darker color coincides with higher co-clustering for pairs of patients. As shown in the **Examples**, for NBS a variant of NMF was used which encourages the selection of gene sets supporting each subtype based on high network connectivity (NetNMF). To promote robust cluster assignments the technique of consensus clustering was used, in which the above procedure is repeated for 1000 different subsamples in which subsets of 80% of patients and genes are drawn randomly without replacement from the entire data set. The results of all 1000 runs are aggregated into a (patient x patient) co-occurrence matrix, which summarizes the frequency of times each pair of patients has co-segregated into the same cluster. This co-occurrence matrix is then clustered to recover a final stratification of the patients into clusters / subtypes (**Figure 5**).

[0117] High-grade cancer mutation data for network stratification methods can be downloaded from a public data portal. Databases for cancer mutation information can include but are not limited to the Cosmic cancer database, cBioPortal for Cancer Genomics, the TCGA data portal, and other databases for cancer mutation data known to those skilled in the art. Mutational data can be generated using a computational platform. Mutational data can be generated by Illumina next generation sequencing platforms (Illumina GAIIx), Life Technology next generation sequencing platforms, and other systems known to those skilled in the art. Patient mutation profiles can be constructed as binary vectors such that a bit is set if the gene or part of a gene corresponding to that position in the vector harbors a mutation in that patient. Additional details on processing and organization of the data are available in a previous TCGA publication, "The International Cancer Genome Consortium International network of cancer genome projects" *Nature*, **4643**, (2010), and is incorporated herein.

[0118] Patient mutation profiles can be mapped onto gene interaction networks. Gene interaction networks can include but are not limited to STRING v.9, HumanNet v.1, and PathwayCommons. All network sources can comprise a combination of interaction types, including direct protein-protein interactions between a pair of gene products and indirect genetic interactions representing regulatory relationships between pairs of genes (e.g. co-expression or TF activation). For example, the PathwayCommons network can be filtered to remove any non-human genes and interactions and all remaining interactions can be used for subsequent analysis. Only the most confident 10% of interactions can be used for this work, ordered according to the quantitative interaction score provided as part of both networks. This threshold can be chosen using an independent ROC analysis with respect to a set of Gene

Ontology derived standards or other means for selecting high-confidence interactions known to those skilled in the art. The edges of all networks can be and used as unweighted, undirected networks.

[0119] After mapping a patient mutation profile onto a molecular network, network propagation can be applied to ‘smooth’ the mutation signal across the network. Network propagation can use for example, a process that can simulate a random walk on a network.

[0120] For example, network propagation can use a process that simulates a random walk on a network (with restarts) according to the function:

$$F_{t+1} = \alpha F_t A + (1 - \alpha) F_0,$$

where, F_0 is a patient-by-gene matrix, A is a degree-normalized adjacency matrix of the gene interaction network, created by multiplying the adjacency matrix by a diagonal matrix with the inverse of its row (or column) sums on the diagonal. α is a tuning parameter governing the distance that a mutation signal is allowed to diffuse through the network during propagation. The optimal value of α can be network-dependent, for example, 0.7, 0.5 and 0.7, for HumanNet, PathwayCommons and STRING respectively, but the specific value seems to have only a minor effect on the results of NBS over a sizable range (e.g. 0.5 - 0.8). The propagation function can be run iteratively with $t = [0, 1, 2, \dots]$ until F_{t+1} converges (the matrix norm of $F_{t+1} - F_t < 1 \times 10^{-6}$). Following propagation, the rows of the resultant matrix F_t can be quantile normalized to ensure that the smoothed mutation profile for each patient follows the same distribution.

[0121] Network-regularized NMF is an extension of non-negative matrix factorization (NMF) that can constrain NMF to respect the structure of an underlying gene interaction network. This can be accomplished by minimizing the following objective function. For example an iterative method with the following function can be used:

$$\min_{W, H > 0} \|F - WH\|^2 + \text{trace}(W^t K W).$$

[0122] W and H form a decomposition of the patient x gene matrix F (resulting from network smoothing as described above) such that W is a collection of basis vectors, or ‘metagenes’, and H is the basis vector loadings. The $\text{trace}(W^t K W)$ function constrains the basis vectors in W to respect local network neighborhoods. The term K is the Graph Laplacian of a nearest-neighbors influence distance matrix derived from the original network. The degree to which local network topology versus global network topology constrains W is determined by the number of nearest neighbors.

[0123] Clustering can be performed with a standard consensus clustering framework. Consensus clustering frameworks are discussed in detail by Monti *et al.*, Machine Learning 52, (2003), The Cancer Genome Atlas Research Network integrated genomic analyses of ovarian carcinoma, Nature 497, (2013), The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumors, Nature, (2012), and Verhaak *et al.*, Integrated genomic analysis identifies clinically relevant

subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1, Cancer Cell (2010), all incorporated in their entirety, herein. Network-regularized NMF (see above) can be used to derive a stratification of the input cohort. To ensure robust clustering, network-regularized NMF can be performed multiple times on subsamples of the dataset, for example, 1000 times. For a subsample, for example 80% of the patients and for example 80% of the mutated genes can be drawn at random without replacement. The set of clustering outcomes for several hundred samples, for example, 1000 samples, can then be transformed into a co-clustering matrix. This matrix can then record the frequency with which each patient pair can be observed to have membership in the same subtype over all clustering iterations in which both patients of the pair are sampled. The end result can comprise a similarity matrix of patients, which can be used to stratify the patients by applying either average linkage hierarchical clustering or a second symmetric NMF step.

[0124] Simulations can be used to determine the ability of NBS to recover subtypes from somatic mutation profiles. Simulations can be performed by computational methods that are known to those skilled in the art. To quantify the performance of NBS a cohort is needed with specified subtypes as a “ground truth” reference, which can allow control over the properties of the signal to be detected. An example of a simulated somatic mutation cohort can be provided as follows. Patient mutation profiles can be sampled with replacement from the TCGA ovarian dataset. For each patient, the mutation profile can be permuted while keeping the per-patient mutation frequency invariant, which can result in a background mutation matrix with no subtype signal. To simulate an underlying network structure for NBS to detect, a network-based signal can be added to the patient-by-mutation matrix as follows. First, a set of network communities can be established, for example, connected components enriched for edges shared within community members, in an input network (i.e., STRING, HumanNet, or PathwayCommons) using a network community detection algorithm. For example, an algorithm, such as Qcut. Next, the patient cohort can be divided randomly into a specified number of equal-sized subtypes. Each subtype can then be assigned a small number (e.g. 1-6) of network modules. These network modules can represent ‘driver’ sub-networks characterizing the subtype. For each patient, a fraction of the patient’s mutations f can be reassigned to genes covered by the driver modules for that patient’s subtype. This procedure can then result in a patient x gene mutation matrix with underlying network structure, while maintaining the per-patient mutation frequency.

[0125] After applying NBS, genes can be identified that are enriched for mutation in each of the subtypes relative to the whole cohort. In order to identify such genes, a method can be applied that assigns a score to each gene on the basis of comparing the propagated mutation score within one subtype against the remaining cohort. This method can be derived computationally and is known to those skilled in the art. For example, the Significance Analysis of Microarrays (SAM) method, previously used to

compare gene expression measurements, can be applied in order to achieve this. Significance Analysis of Microarrays (SAM) is described by Tusher *et al*, Proc Natl Acad Sci (2001), and is incorporated in its entirety herein. SAM is a non-parametric method developed for discovering differentially expressed genes in microarray experiments. Other statistical methods can also be used to compare each subtype against the remaining cohort. Statistical methods for comparison are known to those skilled in the art. For example, a rank based Wilcoxon type statistic can be used, and comparisons can be performed between each subtype against the remaining cohort.

[0126] A regression analysis can be performed to determine a relationship between an NBS-assigned subtype and the patient survival. Regression analysis is a statistical process for estimating the relationships among variables and can include many techniques for modeling and analyzing multiple variables. There are multiple statistical software packages to perform a regression analysis and are known to those skilled in the art. For example, without being limiting, survival analysis can be performed using the R ‘survival’ package. For example, a Cox-proportional hazards model can be used to determine the relationship between the NBS-assigned subtypes and patient survival. A likelihood ratio test and associated p-value can then be calculated by comparing the full model, which can include subtypes and clinical covariates, against a baseline model which includes covariates only. Clinical covariates available in TCGA and included in the model can include, for example, age, grade, stage, residual surgical resection, and mutation rate.

[0127] A method to derive an expression signature equivalent to the somatic mutation based NBS subtypes can be performed. Methods such as shrunken centroids, for example, can be used to derive an expression signature equivalent to the somatic mutation-based NBS subtypes.

[0128] Missense mutations in the genes can also be scored using methods known to those skilled in the art. There are several methods for example, CHASM, VEST and MutationAssessor. CHASM and VEST use supervised machine learning to score mutations. The CHASM training set is composed of a positive class of driver mutations from the COSMIC database and a negative class of synthetic passenger mutations simulated according to the mutation spectrum observed in the tumor type under study. The VEST training set comprises a positive class of disease mutations from the Human Gene Mutation Database and a negative class of variants detected in the ESP6500 cohort with an allele frequency > 1%. MutationAssessor can use patterns of conservation from protein alignments of large numbers of homologous sequences to assess the functional impact of missense mutations. CHASM and VEST scores can be obtained from the CRAVAT webserver (www.cravat.us). Mutation scorers were also obtained by the MutationAssessor method (Reva, Boris, Yevgeniy Antipin, and Chris Sander. "Predicting the functional impact of protein mutations: application to cancer genomics." *Nucleic acids research* (2011)).

[0129] The hyperlink “www.cravat.us” and the contents in the link are shown in CD #1 and are hereby incorporated by reference. Information regarding the contents of the CD (i.e., file name, date of creation and file size) can also be found in the “Appendix to Compact Discs” table below.

[0130] Methods to assign a new tumor sample to a subtype previously identified by NBS can be performed. Methods such as shrunken centroids, for example, can be used for sample classification by summarizing each subtype with a class ‘centroid’ and assigning new samples to the subtype with closest centroid. Such a method may be performed on the smoothed mutation profiles or on the derived mRNA expression signatures equivalent to the somatic mutation-based profiles. Smoothed mutation profiles or mRNA expression profiles can be used to learn an expression signature for each subtype defined earlier by NBS. The nearest shrunken centroid approach can be used to recover stratification predictive of survival as in **Example 5**. Alternatively, a supervised learning approach such as for example decision tree classifiers using, for example, the Logit-Boost algorithm may be used to recover NBS subtypes in the training cohort. A classifier may be trained to recover one subtype vs. the rest of the cohort or a classifier may be trained to recover multiple subtypes in a cohort. Next, such classifier may be used to assign samples from an independent cohort to subtypes as, for example, done in **Example 7**. For subtypes associated with certain clinical phenotypes such as survival rate or response to treatment such a method can predict these phenotypes for a new subject in need by assigning the subject to a subtype. A reference is made to **Examples 5 and 7**.

[0131] As used herein translating includes obtaining a network or map of physical, genetic, biochemical or molecular interactions based on knowledge of molecular biology of a cell; the network is defined by the presence of nodes and links or edges between nodes.

[0132] Nodes may be units within a network which may be connected to other units. They may be described by features such as genes, RNAs, proteins, epigenetic modifications, RNA modifications, post-translational protein modifications, genetic elements (such as promoters, enhancers, exons, introns, splice sites, splicing signals, exon/intron borders, protein coding sequences, non-protein coding sequences, untranslated regions (5' UTR or 3' UTR), polyadenylation signal, transcriptional termination signal, repetitive sequences either (e.g., simple repetitive co-polymers, Alu sequences, LINE sequence, highly repetitive sequences or middle repetitive sequences), SNPs and others known to those in the field of molecular biology or cell biology). Features are selected which will be informative for a certain human or subject's condition, such as for example cancer or tumor, due to changes to the features which permits distinguishing the human or subjects or sample afflicted with the condition from humans or subjects or samples without the condition. Furthermore, the changes to the feature may be further grouped or stratified to divide groups of humans or subjects or samples into informative sub-groups or sub-types that share common characteristics based on features and the network used.

[0133] An edge or link may connect two nodes and describes the relationship of one node to another. Such relationship may include for example information about the relatedness of one node to another or strength of an interaction. Such relatedness could be in the form of common function within a biochemical pathway or process, relatedness in the form of a genetic interaction, related in the form of a physical interaction, relatedness in the form of a hierarchical interaction, regulatory interaction or co-regulatory interaction, relatedness in the form of a developmental process, relatedness in the form of a temporal sequence or order, relatedness in the form of a spatial sequence or order, relatedness in the form of a temporal and spatial sequence or order, relatedness in the form of co-expression or co-modification, relatedness in the form of physical distance or functional distance, relatedness in the form of mutational or recombination hotspots, etc. Relatedness may include within it proximity information, either physical or functional. While one edge or link connects two nodes, each node can have multiple edges or links describing the relatedness or interaction of one node with another.

[0134] A network includes multiple nodes and edges/links that provide a fuller picture of the various relatedness or interactions between all the nodes within the network based on a single feature or combination of features. Networks may include protein-protein interaction networks, gene regulatory networks (such as DNA-protein interaction networks), RNA-protein interaction networks, gene expression network, gene co-expression network (such as transcript association networks), RNA expression network, RNA co-expression network, protein expression network, protein co-expression network, metabolic networks, signaling networks, neuronal networks, food webs, between-species interaction networks, within-species interaction networks or other networks known to or constructed by a person skilled in the art of bioinformatics, molecular biology or cell biology and based on molecular and/or cell biological features. The networks may be publically available, privately held, or commercially available.

[0135] Molecular profile is a set of features that defines the state of one or more molecular entities or molecule species in a patient, subject or sample. For example a gene mutation profile may be a set of genes and their mutation status, e.g. either “mutated” or “not mutated”. In another example, a gene expression profile may be a molecular profile in which, for a selected set of genes, a continuous value is assigned to each gene to denote the level of gene expression. Other molecular profiles may describe other states or changes of state in DNA sequence, genes, RNAs, proteins, epigenetic modifications, RNA modifications, post-translational protein modifications, genetic elements (such as promoters, enhancers, exons, introns, splice sites, splicing signals, exon/intron borders, protein coding sequences, non-protein coding sequences, untranslated regions (5' UTR or 3' UTR), polyadenylation signal, transcriptional termination signal, repetitive sequences either (e.g., simple repetitive co-polymers, Alu sequences, LINE

sequence, highly repetitive sequences or middle repetitive sequences), SNPs and others known to those in the field of molecular biology or cell biology), or a combination of the above.

[0136] Such molecular profiles may be obtained for a patient, subject or sample using methods known to those skilled in the art. Further a profile may be transformed using a network. The transformation may involve the following steps:

- (a) picking a feature(s) to be analyzed, such as somatic mutation in tumor or cancer samples;
- (b) determining the profile of somatic mutation(s) in subjects with a tumor or cancer;
- (c) mapping the profile features to nodes in a selected network, for example by marking the network nodes that correspond to the genes that have mutations as identified by a mutation profile;
- (d) propagating or spreading the influence of the marked nodes (or node features) to the neighbors within the network for example using the algorithm described by Vanunu et al.; and
- (e) obtaining a transformed profile of a subject based on the propagation in step (d).

[0137] As used herein the effect of network propagation may yield a transformed profile wherein the transformed profile may be used to assign a subject into an informative sub-type or group or, alternatively, transformed profiles obtained from a population of subjects may be used to divide the subjects into informative sub-types or groups, for example, through application of various algorithms designed to cluster values. This division of a population of subjects into informative sub-types or groups is commonly known as segregating or stratifying subjects into sub-types or groups or alternatively into informative sub-types or groups. Such informative sub-types or groups may be used to correlate with severity of cancer or tumor, clinical phenotypes, clinically measured parameters, drug response, survival, tumor grade, quality of life, etc. Such informative sub-types or groups may be used to obtain surrogate biological markers, such as gene expression profile of each sub-type or group.

[0138] Projecting a mutation is the act of placing, locating, mapping, identifying or marking a gene or protein in or onto a genetic or protein network, i.e., identifying a “node” in a gene or protein network with the mutation.

[0139] As used herein genes may include both protein coding genes as well as non-protein coding genes. Non-protein coding genes may include among others, rRNA genes, tRNA genes, snRNA genes, and microRNA (miRNA) genes. Typically protein coding genes are transcribed by RNA polymerase II and have introns except in the case of histone genes. While genes are typically in the nucleus, they may also be outside of the nucleus such as in mitochondria or chloroplasts. Within the nucleus, they could be compartmentalized such as in a nucleolus for rRNA genes. Furthermore, genes

may be genomic DNA residing on host chromosomes, or alternatively, they may be extra-chromosomal, such as a result of gene amplification or viral infection. Genes may also be host genes or foreign genes, such as genes acquired by a host cell, through uptake of a nucleic acid or viral infection.

[0140] In applying the method of the invention, mutation(s) associated with nucleic acid sequence or protein sequences may be used to assign or stratify subject(s) into informative sub-types or groups. For nucleic acids, the mutation(s) may occur within protein coding sequences or translated sequence with no change to the amino acid sequence of the resulting translated protein (synonymous or silent mutation). Alternatively, the mutation(s) may occur in the protein coding sequence and change the amino acid sequence of the resulting translated protein or produce a truncated protein (non-synonymous mutation). Mutation may also occur outside of the protein coding sequence, such as in transcriptional regulatory sequences (such as enhancers, promoters, transcriptional terminators, insulator sequences and other transcriptional elements), untranslated regions of a mRNA (such as 5' UTR or 3' UTR), introns, splicing signals (such as exon/intron junctions, splice acceptor site, splice donor site, branch site, etc.), polyadenylation signal, and other genetic elements encoded by the genome. Outside of a gene, mutations may occur in middle (such as LINEs, SINEs, LINE-1, Alu sequence, etc.) or in highly repetitive sequences (e.g., simple copolymeric sequences, direct repeats, etc.) or other extra-genic elements. Such nucleic acid mutations may be inherited or newly acquired. Newly acquired mutations are somatic mutations.

[0141] The method of the invention may also be used to assign or stratify subjects on the basis of naturally occurring genetic variation within a species or across species by using the information associated with single nucleotide polymorphism (SNP).

[0142] The method of the invention may be applied to epigenetic changes (such as changes in methylation patterns at CpG dinucleotides), changes in transcription or gene expression, changes in RNA modifications or RNA processing, changes in the sequence of the primary structure of a protein, changes in protein-DNA interaction, protein-RNA or protein-protein interaction, changes in post-translational modification of proteins or proteome, and other measurable changes in a biological sample from a subject or subjects.

[0143] The method of the invention may be used on information gather for "features" from public database or privately held database. Alternatively, the method of the invention may be used on information generated from biological samples obtained from a subject or subjects.

[0144] The invention provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining nucleic acid sequence information from the subject, determining mutational status from the nucleic acid sequence information so obtained, transforming the mutational status to a transformed profile of the subject based

on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0145] In one embodiment, the nucleic acid sequence information may be obtained from genomic DNA of a subject(s).

[0146] In one embodiment, determining the mutational status from the nucleic acid sequence information may be effected by comparing the nucleic acid sequence information to a reference information for nucleic acid sequence and determining the presence of differences between the nucleic acid sequence information and the reference information. The difference being indicative of the mutational status of the nucleic acid sequence information.

[0147] In one embodiment, transforming the mutational status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the nucleic acid sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of genes, within a network based on network proximity to mutated genes.

[0148] The invention also provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining protein sequence information from the subject, determining mutational status from the protein sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0149] In one embodiment, the protein sequence information may be obtained from conceptual translation of protein coding sequences or expressed proteins of a subject(s).

[0150] In one embodiment, determining the mutational status from the protein sequence information may be effected by comparing it to a reference information for protein sequence and determining the presence of differences from the reference information.

[0151] In one embodiment, transforming the mutational status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the protein sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of proteins within a network based on network proximity to mutated proteins.

[0152] The invention further provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining

epigenetic modification information for genomic DNA from the subject, determining epigenetic modification status from the epigenetic modification information so obtained, transforming the epigenetic status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0153] In one embodiment, the epigenetic modification information may be obtained from genomic DNA of a subject(s).

[0154] In one embodiment, determining the epigenetic modification status from the epigenetic modification information may be effected by comparing it to a reference epigenetic information and determining the presence of differences from the reference information.

[0155] In one embodiment, transforming the epigenetic modification status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the epigenetic modification information onto a network and (b) propagating the mutation(s) or change(s) in the network so as to obtain a continuous range of values for all or subset of epigenetic markers within a network based on network proximity to epigenetic modification changes.

[0156] The invention further provides methods for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor. The method comprises obtaining RNA modification information for RNAs from the subject, determining RNA modification status from the RNA modification information so obtained, transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0157] In one embodiment, the RNA modification information may be obtained from RNAs of a subject(s).

[0158] In one embodiment, determining the RNA modification status from the RNA modification information may be effected by comparing it to a reference RNA modification information and determining the presence of differences from the reference information.

[0159] In one embodiment, transforming the RNA modification status into a transformed profile of the subject may be effected by (a) projecting any difference(s) found within the RNA modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of RNAs, genes encoding RNAs or nucleic acids encoding RNAs within a network based on network proximity to RNA modification differences.

[0160] The invention also provides methods for diagnosing a subject in need thereof with one or more informative subtypes of a cancer or tumor. The method comprises obtaining post-

translational modification information for proteins from the subject, determining post-translational modification status from the post-translational modification information so obtained, transforming the post-translational modification status into a transformed profile of the subject based on a reference molecular network(s) of gene, RNA and/or protein interaction or expression, and comparing the transformed profile of the subject with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor.

[0161] In one embodiment, the post-translational modification information for proteins may be obtained from proteins of a subject(s).

[0162] In one embodiment, determining the post-translational modification status from the post-translational modification information may be effected by comparing it to a reference post-translational modification information and determining the presence of differences from the reference information.

[0163] In one embodiment, transforming the post-translational modification status into a transformed profile of the subject may be effected by (a) projecting any difference(s) found within the post-translational modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of proteins, genes encoding proteins or nucleic acids encoding proteins within a network based on network proximity to post-translational modification differences.

[0164] In one embodiment, the transformed profile comprises one or more or all of the genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0165] In one embodiment, the reference information may be obtained from subjects without cancer or tumor or healthy cells of subjects with cancer.

[0166] In an embodiment, comparing the transformed profile with reference transformed profiles may be effected by assigning the subject to a subtype of a cancer or tumor with closest reference profile. The closest reference profile may be effected by application of a nearest shrunken centroid approach (Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002)), a supervised learning approach based on decision tree classifiers (Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 28(2), 2000. 337–407; "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences* **55**. 1997) or another supervised or unsupervised learning approach (Hastie, Trevor, Robert Tibshirani, Friedman, Jerome (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. New York:

Springer. pp. 485--586; Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258) or combination of such approaches.

[0167] In one embodiment, the informative subtype may be a clinical phenotype. The clinical phenotype may be predictive of a survival rate, drug response or tumor grade.

[0168] In one embodiment, the mutation may be a somatic mutation. The somatic mutation may be in genomic DNA.

[0169] In an embodiment, the somatic mutation may be an exonic mutation or a mutation in an exon.

[0170] In one embodiment, the exonic mutation or the mutation in an exon may alter a protein coding sequence. In another embodiment, the exonic mutation or the mutation in an exon may be a synonymous mutation or a silent mutation that does not alter a protein sequence. In a further embodiment, the exonic mutation or the mutation in an exon may be a non-synonymous mutation that alters a protein sequence. In yet another embodiment, the somatic mutation may be a synonymous mutation or a non-synonymous mutation.

[0171] In one embodiment, the somatic mutation may be in a gene. The gene may be a protein coding gene or a non-protein coding gene. The protein coding gene may be transcribed by RNA polymerase II. The non-protein coding gene may encode a rRNA gene, a tRNA gene, a snRNA gene, a miRNA gene or a gene for a structural RNA.

[0172] In one embodiment, the somatic mutation in a gene may be in a promoter, enhancer, transcriptional terminator, intron, untranslated region (5' UTR or 3' UTR), exon-intron junction, splice site, splicing branch site, polyadenylation signal or other genetic elements.

[0173] In one embodiment, the somatic mutation may be in an extragenic region or a mutation outside of a gene of a subject's genome.

[0174] In one embodiment, the somatic mutation may be in a middle repetitive DNA sequence or highly repetitive DNA sequence.

[0175] In one embodiment, the somatic mutation may be in a transcribed or an untranscribed region of a subject's genome.

[0176] In one embodiment, the somatic mutation may be in nuclear or mitochondrial DNA.

[0177] In an embodiment, the cancer may be breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer.

[0178] In an embodiment, the method may be carried out by an informatics platform. The informatics platform may be a bioinformatics platform comprising a computer and software. In one embodiment, the software may use supervised learning and/or unsupervised learning methods.

[0179] In one embodiment, the method may be an automated method.

[0180] In an embodiment, the method may require selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information.

[0181] Selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprises one or more or all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with the genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0182] Additionally, selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprise selecting all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with each group of genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0183] Further, selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprise selecting two or more groups of genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with groups of genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 42, 42, and 43.

[0184] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining nucleic acid sequence information from subjects with a cancer or tumor, determining mutational status for each subject from the nucleic acid sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0185] In one embodiment, the nucleic acid sequence information may be obtained from genomic DNA of subjects.

[0186] In an embodiment, determining the mutational status from the nucleic acid sequence information may be effected by comparing it to a reference information for nucleic acid sequence and determining the presence of differences from the reference information.

[0187] In an embodiment, transforming the mutational status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the nucleic acid sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of genes, within a network based on network proximity to mutated genes.

[0188] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining protein sequence information from subjects with a cancer or tumor, determining mutational status for each subject from the protein sequence information so obtained, transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0189] In one embodiment, the protein sequence information may be obtained from conceptual translation of protein coding sequences or expressed proteins of subjects.

[0190] In an embodiment, determining the mutational status from the protein sequence information is effected by comparing it to a reference information for protein sequence and determining the presence of differences from the reference information.

[0191] In an embodiment, transforming the mutational status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the protein sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of proteins within a network based on network proximity to mutated proteins.

[0192] The invention further provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining epigenetic modification information from subjects with a cancer or tumor, determining epigenetic modification status for each subject from the epigenetic modification information so obtained, transforming the epigenetic modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0193] In one embodiment, the epigenetic modification information may be obtained from genomic DNA of subjects.

[0194] In an embodiment, determining the epigenetic modification status from the epigenetic modification information may be effected by comparing it to a reference epigenetic information and determining the presence of differences from the reference information.

[0195] In an embodiment, transforming the epigenetic modification status into a transformed profile of the subject may be effected by (a) projecting any mutation(s) found within the epigenetic modification information onto a network and (b) propagating the mutation(s) or change(s) in the network so as to obtain a continuous range of values for all or subset of epigenetic markers within a network based on network proximity to epigenetic modification changes.

[0196] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining RNA modification information from subjects with a cancer or tumor, determining RNA modification status for each subject from the RNA modification information so obtained, transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes.

[0197] In one embodiment, the RNA modification information may be obtained from RNAs of subjects.

[0198] In an embodiment, determining the RNA modification status from the RNA modification information may be effected by comparing it to a reference RNA modification information and determining the presence of differences from the reference information.

[0199] In an embodiment, transforming the RNA modification status into a transformed profile of the subject may be effected by (a) projecting any difference(s) found within the RNA modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of RNAs, genes encoding RNAs or nucleic acids encoding RNAs within a network based on network proximity to RNA modification differences.

[0200] The invention also provides methods for identifying one or more informative subtypes of a cancer or tumor. The method comprises obtaining post-translational modification information from subjects with a cancer or tumor, determining post-translational modification status for each subject from the post-translational modification information so obtained, transforming the post-translational modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression, and clustering the transformed profiles obtained into one or more clusters so as to obtain one or more subtypes.

[0201] In one embodiment, the post-translational modification information for proteins may be obtained from proteins of subjects.

[0202] In an embodiment, determining the post-translational modification status from the post-translational modification information may be effected by comparing it to a reference post-

translational modification information and determining the presence of differences from the reference information.

[0203] In an embodiment, transforming the post-translational modification status into a transformed profile of the subject may be effected by (a) projecting any difference(s) found within the post-translational modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of proteins, genes encoding proteins or nucleic acids encoding proteins within a network based on network proximity to post-translational modification differences.

[0204] In one embodiment, the transformed profile may comprise one or more or all of the genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0205] In one embodiment, clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes may be effected by grouping subjects with similar transformed profiles into one or more groups or subtypes.

[0206] In one embodiment, the reference information may be obtained from subjects without cancer or tumor.

[0207] In one embodiment, the informative subtype may be a clinical phenotype. The clinical phenotype may be predictive of a survival rate, drug response or tumor grade.

[0208] In an embodiment, the mutation may be a somatic mutation.

[0209] In one embodiment, the somatic mutation may be in genomic DNA. In another embodiment, the somatic mutation may be an exonic mutation or a mutation in an exon.

[0210] In one embodiment, the exonic mutation or the mutation in an exon may alter a protein coding sequence. In another embodiment, the exonic mutation or the mutation in an exon may be a synonymous mutation or a silent mutation that does not alter a protein sequence. In another embodiment, the exonic mutation or the mutation in an exon may be a non-synonymous mutation that alters a protein sequence. In yet another embodiment, the somatic mutation may be a synonymous mutation or a non-synonymous mutation.

[0211] In one embodiment, the somatic mutation may be in a gene. The gene may be a protein coding gene or a non-protein coding gene.

[0212] In one embodiment, the protein coding gene may be transcribed by RNA polymerase II.

[0213] In one embodiment, the non-protein coding gene may encode a rRNA gene, a tRNA gene, a snRNA gene, a miRNA gene or a gene for a structural RNA.

[0214] In one embodiment, the somatic mutation in a gene may be in a promoter, enhancer, transcriptional terminator, intron, untranslated region (5' UTR or 3' UTR), exon-intron junction, splice site, splicing branch site, polyadenylation signal or other genetic elements.

[0215] In one embodiment, the somatic mutation may be in an extragenic region or a mutation outside of a gene of a subject's genome.

[0216] In one embodiment, the somatic mutation may be in a middle repetitive DNA sequence or highly repetitive DNA sequence.

[0217] In one embodiment, the somatic mutation may be in a transcribed or an untranscribed region of a subject's genome.

[0218] In one embodiment, the somatic mutation may be in nuclear or mitochondrial DNA.

[0219] In an embodiment, the cancer may be breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer.

[0220] In an embodiment, the method may be carried out by an informatics platform. The informatics platform may be a bioinformatics platform comprising a computer and software.

[0221] In one embodiment, the software may use supervised learning and/or unsupervised learning methods.

[0222] In an embodiment, the method may be an automated method.

[0223] In one embodiment, the method may require selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information.

[0224] Selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprise one or more or all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with the genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0225] Additionally, selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprise selecting all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with each group of genes described herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0226] Further, selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information may comprise selecting two or more groups of genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with groups of genes described

herein and including the genes described in Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

[0227] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor; determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes; obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered and the subject of interest having cancer or tumor by using a supervised learning approach to derive a subtype classifier based on profiles from the subjects and their assignment to subtypes; and comparing the subtype classifier so derived to assign the subject of interest to a cancer or tumor subtype.

[0228] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor; determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor; obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered and the subject of interest having cancer or tumor; and applying a nearest shrunken centroid approach (Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567–6572 (2002)), a supervised learning approach based on decision tree classifiers (Additive logistic regression: a statistical view of boosting. *Annals of Statistics* 28(2), 2000. 337–407; "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences* **55**. 1997) or another supervised or unsupervised learning approach (Hastie, Trevor, Robert Tibshirani, Friedman, Jerome (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. New York: Springer. pp. 485–586; Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) *Foundations of Machine Learning*, The MIT Press ISBN 9780262018258) or combination of such approaches to assign the subject of interest to a cancer or tumor subtype.

[0229] The invention further provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining nucleic acid sequence information from subjects with a cancer or tumor; determining mutational status for each subject from the nucleic acid sequence information so obtained; transforming the mutational status to a transformed profile of each of the subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression; clustering the transformed profiles for subjects obtained into one or more clusters so as to obtain one or more subtypes; characterizing the subjects grouped into one or more informative subtypes by determining status or profile of one or more measurable or quantifiable biological parameter(s) or feature(s); characterizing the subject of interest by determining status or profile of one or more measurable or quantifiable biological parameter(s); and assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on status or profile(s) of the subjects grouped into one or more informative subtypes and the status or profile of the subject of interest.

[0230] The invention also provides methods of assigning a subject of interest having a cancer or tumor into one or more informative subtypes. The method comprises obtaining biological profiles of subjects grouped into one or more informative subtypes, obtaining biological profile of the subject of interest, and assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on biological profile(s) of the subjects grouped into one or more informative subtypes and the biological profile of the subject of interest.

[0231] In one embodiment, the informative subtype may be associated with a clinical phenotype. The clinical phenotype may be predictive of a survival rate, drug response or tumor grade.

[0232] In one embodiment, the cancer may be breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer. In a preferred embodiment, the cancer is ovarian cancer.

[0233] In one embodiment, the tumor is an ovarian tumor.

[0234] In one embodiment, the subtype may be ovarian cancer subtype 1, 2, 3, or 4.

[0235] In one embodiment, the subtype may be predictive of survival.

[0236] In one embodiment, the subtype may be predictive of response to treatment. The treatment may involve chemotherapy.

[0237] In one embodiment, the mutation may be a somatic mutation.

[0238] In one embodiment, the method may be carried out by an informatics platform. The informatics platform may be a bioinformatics platform comprising a computer and software.

[0239] In one embodiment, the software may use supervised learning and/or unsupervised learning methods.

[0240] In one embodiment, the method may be an automated method.

[0241] The invention also provides methods for increasing efficiency of a bioinformatics process for network-based stratification of tumor or cancer. The method comprises obtaining a biological sample from a subject with tumor or cancer; selecting a set of genes for which nucleic acid sequence is to be determined; determining nucleic acid sequence for protein coding sequences in the set of genes selected; projecting mutations found within sequence onto a network; propagating the mutations in the network; and clustering the mutations so propagated so as to divide biological samples from subjects with tumor or cancer into subtypes, wherein, the set of genes so selected excludes whole exome or genome sequencing.

[0242] While various aspects and embodiments have been disclosed herein, other aspects and embodiments will be apparent to those skilled in the art. The various aspects and embodiments disclosed herein are for purposes of illustration and are not intended to be limiting, with the true scope and spirit being indicated by the following claims.

[0243] While the present invention has been described in some detail for purposes of clarity and understanding, one skilled in the art will appreciate that various changes in form and detail can be made without departing from the true scope of the invention.

APPENDIX TO COMPACT DISCS

	CONTENT FILE NAME	DATE CREATED	FILE SIZE
CD 1: Cravat			
A. LUAD			
	1. PR_TCGA_LUAD_PAIR_Capture_All_Pairs_QCP ASS.aggregated.capture.tcga.uuid.somatic.maf.trim.g enomic.driver.chasm.Lung-Adenocarcinoma.output	8/7/14	15.9MB
	2. PR_TCGA_LUAD_PAIR_Capture_All_Pairs_QCP ASS.aggregated.capture.tcga.uuid.somatic.maf.trim.g enomic.functional.vest.output	8/7/14	16.6MB
B. OV			
	1. broad.mit.edu_OV.IlluminaGA_DNASeq.Level_2.7. somatic.maf.trim.genomic.driver.chasm.Ovary.outpu t.tsv.formatted	8/7/14	545KB
	2. broad.mit.edu_OV.IlluminaGA_DNASeq.Level_2.7. somatic.maf.trim.genomic.functional.vest.output.tsv.f ormatted	8/7/14	577KB
	3. genome.wustl.edu_OV.IlluminaGA_DNASeq.l.som atic.maf.trim.genomic.driver.chasm.Ovary.output.tsv .formatted	8/7/14	514KB
	4. genome.wustl.edu_OV.IlluminaGA_DNASeq.l.som	8/7/14	544KB

	CONTENT FILE NAME	DATE CREATED	FILE SIZE
	atic.maf.trim.genomic.functional.vest.output.tsv.formatted		
	5. genome.wustl.edu_OV.IlluminaGA_DNASeq.Level_2.2.0.0.somatic.maf.trim.genomic.driver.chasm.Ovary.output.tsv.formatted	8/7/14	873KB
	6. genome.wustl.edu_OV.IlluminaGA_DNASeq.Level_2.2.0.0.somatic.maf.trim.genomic.functional.vest.output.tsv.formatted	8/7/14	924KB
C. UCEC			
	1. An_UCEC_194.aggregated.tcga.somatic.maf.trim.genomic.driver.chasm.Uterus.output	8/7/14	4.23MB
	2. An_UCEC_194.aggregated.tcga.somatic.maf.trim.genomic.functional.vest.output	8/7/14	4.42MB
	3. genome.wustl.edu_UCEC.IlluminaGA_DNASeq.Level_2.1.somatic.maf.trim.functional.vest.output	8/7/14	17.8MB
	4. genome.wustl.edu_UCEC.IlluminaGA_DNASeq.Level_2.1.somatic.maf.trim.genomic.driver.chasm.Uterus.output	8/7/14	17.0MB
D.	README	8/10/14	681bytes
CD 2: HumanNetv0.1			
A.	HumanNet.v1.join	9/6/10	48.7MB
B.	README	8/10/14	258bytes
CD 3: PathwayCommons			
A.	homo-sapiens-9606-edge-attributes	10/26/11	141MB
B.	homo-sapiens-9606-node-attributes	10/26/11	1.65MB
C.	README	8/10/14	227bytes
CD 4: RepliSeq			
A.	README	8/10/14	70bytes
B.	refGene.mRNA.avg.hg19.GM12878.Rep1	8/7/14	773KB
C.	refGene.mRNA.avg.hg19.GM12878.sum.Rep1	8/7/14	761KB
D.	UW Repli-seq	8/8/14	499KB
E.	wgEncodeUwRepliSeqGm12878PkRep1.bed	8/7/14	199KB
F.	wgEncodeUwRepliSeqGm12878SumSignalRep1.bigWig	8/7/14	12.5MB
G.	wgEncodeUwRepliSeqGm12878ValleysRep1.bed	8/7/14	198KB
H.	wgEncodeUwRepliSeqGm12878WaveSignalRep1.bigWig	8/7/14	9.72MB
CD 5: STRINGv9.0			
A.	protein.links.detailed.v9.0.human.tsv	1/24/12	194MB
B.	README	8/10/14	269bytes
CD 6: yandell-lab			
A.	refGene_hg19.gff3	8/13/14	61.6MB

EXAMPLES

Methods

[0244] The following methods were performed with reference to the following Examples, unless stated otherwise.

Overview of network based stratification

[0245] The technique of Network-based Stratification (NBS) combined genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes. Briefly, somatic mutations for each patient are represented as a profile of binary (1,0) states on genes, in which a '1' indicates a gene for which mutation has occurred in the tumor relative to germline (i.e. a single nucleotide base change or the insertion or deletion of bases). For each patient independently the mutation profiles were projected onto a human gene interaction network obtained from public databases. Next, the technique of network propagation was applied to spread the influence of each subsampled mutation profile over its network neighborhood (**Figure 3**). The result was a 'network-smoothed' profile in which the state of each gene is no longer binary but reflects its network proximity to the mutated genes in that patient, along a continuous range [0,1]. Following this 'network smoothing', patient profiles were clustered into a predefined number of subtypes ($k = 2...12$) using the unsupervised technique of non-negative matrix factorization (**Figure 4**). For NBS a variant of NMF was used which encourages the selection of gene sets supporting each subtype based on high network connectivity (NetNMF). Finally, to promote robust cluster assignments the technique of consensus clustering was used, in which the above procedure is repeated for 1000 different subsamples in which subsets of 80% of patients and genes are drawn randomly without replacement from the entire data set. The results of all 1000 runs were aggregated into a (patient x patient) co-occurrence matrix, which summarizes the frequency of times each pair of patients has co-segregated into the same cluster. This co-occurrence matrix was then clustered to recover a final stratification of the patients into clusters/subtypes (**Figure 5 and Figure 24**).

Processing of patient profiles

[0246] High-grade serous ovarian cancer, uterine endometrial carcinoma, and lung adenocarcinoma somatic mutation data were downloaded from the TCGA data portal. Only mutation data generated using the Illumina GAIIX platform were retained for subsequent analysis, and patients with fewer than 10 mutations were discarded. Overall, there were 356 patients with mutations in 9,850 genes for the TCGA ovarian cohort, 248 patients with mutations in 17,968 genes for the TCGA uterine

endometrial cohort, and 381 patients with mutations in 15,967 genes in the TCGA lung adenocarcinoma cohort. Patient mutation profiles were constructed as binary vectors such that a bit is set if the gene corresponding to that position in the vector harbors a mutation in that patient. Additional details on processing and organization of the data are available in a previous TCGA publication, The Cancer Genome Atlas Research, N. et al., Integrated genomic characterization of endometrial carcinoma, *Nature* **497**, 67-74 (2013), and is incorporated in its entirety.

Sources of molecular network data

[0247] Patient mutation profiles were mapped onto gene interaction networks from three sources: STRING v.9, HumanNet v.1, and PathwayCommons. All network sources comprise a combination of interaction types, including direct protein-protein interactions between a pair of gene products and indirect genetic interactions representing regulatory relationships between pairs of genes (e.g. co-expression or TF activation). The PathwayCommons network was filtered to remove any non-human genes and interactions and all remaining interactions were used for subsequent analysis. Only the most confident 10% of interactions for both the STRING and HumanNet networks were used for this work, and were ordered according to the quantitative interaction score provided as part of both networks. This threshold was chosen using an independent ROC analysis with respect to a set of Gene Ontology derived gold standards. After filtering of edges all networks were used as unweighted, undirected networks.

Network smoothing

[0248] After mapping a patient mutation profile onto a molecular network, network propagation was applied to ‘smooth’ the mutation signal across the network. For network propagation a process that simulates a random walk on a network (with restarts) was used and the function applied:

$$F_{t+1} = \alpha F_t A + (1 - \alpha) F_0$$

where, F_0 is a patient-by-gene matrix, A is a degree-normalized adjacency matrix of the gene interaction network, created by multiplying the adjacency matrix by a diagonal matrix with the inverse of its row (or column) sums on the diagonal. α is a tuning parameter governing the distance that a mutation signal is allowed to diffuse through the network during propagation. The optimal value of α is network-dependent (0.7, 0.5 and 0.7, for HumanNet, PathwayCommons and STRING respectively), but the specific value seems to have only a minor effect on the results of NBS over a sizable range (e.g. 0.5 - 0.8). The propagation function was run iteratively with $t = [0, 1, 2, \dots]$ until F_{t+1} converges (the matrix norm of $F_{t+1} - F_t < 1 \times 10^{-6}$). Following propagation, the rows of the resultant matrix F_t were quantile normalized to ensure that the smoothed mutation profile for each patient follows the same distribution.

Network-regularized NMF

[0249] Network-regularized NMF, an extension of non-negative matrix factorization (NMF), constrains NMF in respect to the structure of an underlying gene interaction network. Network-regularized NMF was accomplished by minimizing the following objective function using an iterative method:

$$\min_{W, H > 0} \|F - WH\|^2 + \text{trace}(W^t KW),$$

where, W and H form a decomposition of the patient x gene matrix F (resulting from network smoothing as described above) such that W is a collection of basis vectors, or ‘metagenes’, and H is the basis vector loadings. The $\text{trace}(W^t KW)$ function constrains the basis vectors in W to respect local network neighborhoods. The term K is the Graph Laplacian of a nearest-neighbors influence distance matrix derived from the original network. The degree to which local network topology versus global network topology constrains W is determined by the number of nearest neighbors. Neighbor counts ranging from 5 to 50 were implemented to include in the nearest network and only small changes in outcome were observed. As shown in the **Examples**, the 11 most influential neighbors of each gene in the network as determined by network influence distance were used.

Consensus clustering

[0250] Clustering was performed with a standard consensus clustering framework, discussed in detail by Monti *et al.* (Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52, 91-118 (2003); incorporated herein) and used in previous TCGA publications. Network-regularized NMF was used to derive a stratification of the input cohort. In order to ensure robust clustering, network-regularized NMF was performed 1000 times on subsamples of the dataset. In each subsample, 80% of the patients and 80% of the mutated genes were sampled at random without replacement. The set of clustering outcomes for the 1000 samples was then transformed into a co-clustering matrix. This matrix records the frequency with which each patient pair was observed to have membership in the same subtype over all clustering iterations in which both patients of the pair were sampled. The result is a similarity matrix of patients, which were then used to stratify the patients by applying either average linkage hierarchical clustering or a second symmetric NMF step.

Simulation of somatic mutation cohorts

[0251] Simulations to determine the ability of NBS to recover subtypes from somatic mutation profiles were used. In order to quantify the performance of NBS, a cohort with specified

subtypes as a “ground truth” reference was needed, while allowing control over the properties of the signal to be detected. A somatic mutation cohort was simulated as follows: Patient mutation profiles were sampled with replacement from the TCGA ovarian dataset. For each patient, the mutation profile was permuted while keeping the per-patient mutation frequency invariant, resulting in a background mutation matrix with no subtype signal. To simulate an underlying network structure for NBS to detect, a network-based signal was added to the patient-by-mutation matrix as follows. First, a set of network communities was established (i.e. connected components enriched for edges shared within community members) in the input network (STRING, HumanNet, or PathwayCommons) using the network community detection algorithm Qcut. Next the patient cohort was divided randomly into four equal-sized subtypes (four was selected as reasonable due to the four expression-based subtypes that have been identified for glioblastoma, ovarian and breast cancers). Each subtype was assigned a small number (e.g. 1-6) of network modules which together had a combined size s ranging from 10 to 250 genes. These network modules represent ‘driver’ sub-networks characterizing the subtype. For each patient, a fraction of the patient’s mutations f to genes covered by the driver modules for that patient’s subtype was reassigned. This procedure resulted in a patient x gene mutation matrix with underlying network structure, while maintaining the per-patient mutation frequency.

[0252] A plausible range for the number of driver mutation in a tumor was proposed to be between 2 to 8 driver mutations. In the simulation framework, a 4% mutation rate corresponds to between 1 and 9 mutations with a median of 3, on par with the aforementioned estimates. In order to estimate the appropriate size of cancer pathways (s) the known cancer pathways in the NCI-Nature cancer interaction database were examined. Pathways in the database of varying sizes were observed that were 2 – 139 genes, with a median size of 34, and over 23% of pathways include over 50 genes.

Identifying differentially mutated sub-networks

[0253] After applying NBS, genes that were enriched for mutation in each of the subtypes relative to the whole cohort were identified. The Significance Analysis of Microarrays (SAM) method were applied on the network smoothed mutation profiles. This is a non-parametric method previously developed for discovering differentially expressed genes in microarray experiments. A rank based Wilcoxon type statistic was used and each subtype was compared against the remaining cohort. Significance was assessed using the SAM permutation scheme with 1000 permutations. The resulting set of genes for each subtype was overlaid on the network used for network smoothing.

Survival analysis

[0254] Survival analysis was performed using the R ‘survival’ package. A Cox-proportional hazards model was fit to determine the relationship between the NBS-assigned subtypes and patient survival. A likelihood ratio test and associated p-value is calculated by comparing the full model, which includes subtypes and clinical covariates, against a baseline model which includes covariates only. Clinical covariates available in TCGA and included in the model were age, grade, stage, residual surgical resection, and mutation rate, as well as cigarette smoking status for the lung cancer cohort.

Shrunken centroid prediction on expression profiles

[0255] Shrunken centroids were used to derive an expression signature equivalent to the somatic mutation-based NBS subtypes. Expression data were provided by Györfy *et al.* who aggregated several expression datasets as part of a meta-analysis of ovarian cancer. In this analysis, all data were regularized using quantile and MAS5 normalization. This analysis was performed on the Tothill *et al.* (ovarian serous samples only), Bonome *et al.*, and TCGA datasets, as well as across the full meta-analysis cohort. The ‘pamr’ R package was used, with default parameters to train a shrunken centroid model on mRNA expression levels for all genes in the TCGA ovarian dataset with subtype assignment as the class label. The trained model was next used to predict subtype labels on the held-out Tothill *et al.* and Bonome *et al.* data or the full meta-analysis expression cohorts.

Missense Mutation Scoring

[0256] Missense mutations were scored using three methods: CHASM, VEST and MutationAssessor. CHASM and VEST use supervised machine learning to score mutations. The CHASM training set is composed of a positive class of driver mutations from the COSMIC database and a negative class of synthetic passenger mutations simulated according to the mutation spectrum observed in the tumor type under study. The VEST training set comprises a positive class of disease mutations from the Human Gene Mutation Database and a negative class of variants detected in the ESP6500 cohort with an allele frequency > 1%. MutationAssessor uses patterns of conservation from protein alignments of large numbers of homologous sequences to assess the functional impact of missense mutations. CHASM and VEST scores were obtained from the CRAVAT webserver (cravat.us). Mutation scorers were also obtained by the MutationAssessor method (Reva, Boris, Yevgeniy Antipin, and Chris Sander. "Predicting the functional impact of protein mutations: application to cancer genomics." *Nucleic acids research* (2011)).

Replication Timing

[0257] RepliSeq data for GM12878 were downloaded from the ENCODE project website (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/>). Summed normalized tag densities were used as a proxy for replication time (higher counts indicating that a transcript was replicated earlier in the cell cycle). Normalized tag densities for RefSeq protein coding regions were retrieved using bigWigAverageOverBed with RefSeq gene sequence features in gff3 format downloaded from (yandell-lab.org/software/VAAST/data/hg19/Features/refGene_hg19.gff3). Tag densities were averaged for each transcript and the longest transcript was selected to represent each gene.

[0258] The hyperlinks “<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/>” and “yandell-lab.org/software/VAAST/data/hg19/Features/refGene_hg19.gff3” and the contents in the links are shown in CDs #4 and 6, respectively, and are hereby incorporated by reference. Information regarding the contents of the CDs (i.e., file name, date of creation and file size) can also be found in the “Appendix to Compact Discs” table below.

Example 1: Method of Network based stratification

[0259] The technique of Network-based Stratification (NBS) combined genome-scale somatic mutation profiles with a gene interaction network to produce a robust subdivision of patients into subtypes as indicated in the flowchart of **Figure 2**. Briefly, somatic mutations for each patient were represented as a profile of binary (1,0) states on genes, in which a ‘1’ indicates a gene for which mutation has occurred in the tumor relative to germline (i.e. a single nucleotide base change or the insertion or deletion of bases). For each patient independently the mutation profiles were projected onto a human gene interaction network obtained from public databases. Next, the technique of network propagation was applied to spread the influence of mutations over network neighborhoods (**Figure 3**). The resulting ‘network-smoothed’ patient profiles were clustered into a predefined number of subtypes ($k = 2 \dots 12$) using an unsupervised technique of non-negative matrix factorization (NMF) (**Figure 4**). Finally, to promote robust cluster assignments the technique of consensus clustering were used, aggregating the results of 1000 different subsamples from the entire data set into a single clustering result. Additional details may be found in Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods* 10(11):1108-15 (2013), all of which is incorporated by reference herein.

[0260] To evaluate the impact of different sources of network data, three interaction databases for this analysis: STRING, HumanNet or PathwayCommons were used. STRING integrates protein-protein interactions from literature curation, computationally-predicted interactions, and interactions transferred from model organisms based on orthology. HumanNet uses a naïve bayes

approach to weight different types of evidence together into a single interaction score focusing on data collected in humans, yeast, worm and fly. PathwayCommons aggregates interactions from several pathway and interaction databases, focused primarily on physical protein-protein interactions (PPIs) and functional relationships between genes in canonical regulatory, signaling, and metabolic pathways (including hallmark pathways of cancer). Table 1 summarizes the number of genes and interactions used in the analysis from each of these three networks.

	<u>Nodes</u>	<u>Edges</u>	<u>Links and Description</u>
HumanNet v.123	16,243 (7,949)	476,399 (47,641)	www.functionalnet.org/humannet A database of gene interactions, derived using a naïve bayes approach by combining multiple lines of experimental evidence. Comprised of both protein-protein interactions (PPIs) and genetic interactions
STRING v.929	16,560 (12,233)	1,638,830 (164,034)	www.string-db.org/ A database integrating a variety of evidence types including, experimental expression and literature mining approaches to derive a globally weighted network of gene interactions. Comprised of multiple types of gene interaction
PathwayCommons	14,355	507,757	www.pathwaycommons.org/pc/ An aggregated repository of gene interactions from several sources including BioGrid, HPRD, IntAct and the NCI set of cancer specific pathways. Comprised of mostly physical PPIs

Table 1: **Summary of gene interaction networks.** The table shows the networks used as part of the analysis. The HumanNet and STRING networks were filtered to include the top 10% of interactions according to the interaction weights. After filtering all edges were treated as unweighted.

[0261] The hyperlinks “www.functionalnet.org/humannet”, “www.string-db.org/” and “www.pathwaycommons.org/pc/” and the contents in the links are shown in CDs #2, 5 and 3, respectively, and are hereby incorporated by reference. Information regarding the contents of the CDs (i.e., file name, date of creation and file size) can also be found in the “Appendix to Compact Discs” table below.

Example 2: Benchmarking and performance analysis

[0262] As an initial exploration of NBS, a somatic mutation dataset was simulated based on the structure of the TCGA ovarian tumor mutation data and the STRING gene interaction network

(**Figure 6**). Mutation profiles were permuted and patients were divided randomly and uniformly into a predefined number of subtypes ($k = 4$). Next, a fraction of mutations in each patient was reassigned to fall within genes of a single ‘network module’ characteristic of that patient’s subtype (the ‘driver’ mutation frequency f , varied from 1 to 15%), and the remaining mutations were left to occur randomly. The network modules were selected randomly from the set of all network modules in STRING, defined as sets of densely-interacting genes with size range $s = 10$ to 250 as indicated in Methods. Although it is unknown whether these assumptions completely mirror the biology of cancer, they provided a reasonable model of a pathway-based genetic disease which is driven by genetic circuits embedded in a molecular network whose activity can be altered by mutations at multiple genes and is characterized by many additional mutations that are non-causal ‘passengers.’

[0263] Using this simulation framework, the ability of NBS was measured to recover the correct subtype assignments in comparison to a standard consensus clustering approach not based on network knowledge (i.e., the same NBS pipeline in (**Figure 2**)) without network smoothing and substituting NMF for NetNMF). NBS showed a striking improvement in performance, especially for large network modules as these can be associated with any of numerous different mutations across the patient population (**Figure 7**). Accuracy was calculated as the Adjusted Rand Index of overlap between the clusters and correct subtype assignments, for which a score of zero represents random overlap. Simulation was performed with a driver mutation frequency $f = 7.5\%$ with a single network module assigned to each subclass. As module size decreases, the chance of observing the same mutated gene in patients of the same subtype increases and the standard clustering algorithm performs increasingly well. As shown, the high performance of NBS depends not only on network smoothing but also NMF clustering approach, as substitution with an alternative method such as hierarchical clustering performs relatively poorly (**Figure 7**).

[0264] Next, it was investigated how NBS performance was impacted as a function of mutation frequency and network module size (**Figure 8**). Standard consensus clustering was sufficient for high mutation frequencies and small modules, for which there is substantial overlap in mutations among patients of the same subtype (**Figure 10**); however NBS was able to accurately recover the correct subtypes for a much larger range of both variables. Applying NBS on a permuted network resulted in poor performance (**Figure 9**), on par with that observed with standard consensus clustering. These results were qualitatively similar when using multiple network modules per patient (2–6) and/or a different network.

Example 3: Network-based stratification of tumor mutations

[0265] NBS was applied to stratify patients profiled by TCGA full exome sequencing, separately for three different cancers – uterine, ovarian, and lung. In all three cancers, NBS resulted in robust subtype structure, whereas standard consensus clustering was unable to stratify the patient cohort (**Figure 11**, for uterine cancer; **Figure 28a** for ovarian cancer; and **Figure 29a** for lung cancer). Similar results were obtained when using any of the three human networks STRING, HumanNet, and PathwayCommons.

[0266] The identified subtypes were then investigated whether they were predictive of observed clinical data such as histological appearance and patient survival time, in order to determine the biological importance of the identified. In uterine cancer, NBS subtypes were closely associated with the recorded subtype based on histology (**Figures 12 13 and 27**). Survival analysis was not possible due to low mortality rates for this cohort. In ovarian cancer, the identified subtypes were significant predictors of patient survival time (**Figures 14, 15, 28b and 28c**). The most aggressive ovarian tumor subtype had a mean survival of approximately 32 months while the least aggressive subtype had a mean survival of more than 80 months, a 2.5-fold difference (**Figures 28d and 28e**). Moreover, these subtypes were predictive of survival independently of clinical covariates including tumor stage, age, mutation rate and residual tumor after surgery (Likelihood ratio test, $P = 3.75 \times 10^{-5}$) (**Figure 30**). Furthermore, subtypes were predictive of time until the onset of platinum resistance (**Figure 28f**), as measured using a Kaplan-Meier analysis of platinum free survival. Finally, in lung cancer the identified subtypes were also found to be significant predictors of patient survival time (**Figures 16, 17 and 29**), median survival of 12 months versus approximately 50 months for the best surviving subtype. As for ovarian cancer, the lung cancer subtypes had predictive value beyond known clinical covariates such as tumor stage, grade, mutation frequency, age at diagnosis and smoking status (Likelihood ratio test, $P = 3.3 \times 10^{-4}$). Finally, stratification using a network in which the mapping between mutated genes and the network was permuted, disrupting the relationship between mutations and network structure, resulted in degraded predictive performance (**Figures 12, 14, and 16**).

[0267] Next, the results were compared to subtypes derived from other data types in TCGA, including CNV, methylation, mRNA expression, microRNA expression, and protein profiles. For ovarian cancer, all other data types had inferior ability to predict survival beyond what can be predicted from clinical covariates (**Figure 18a**) and led to very different subtype assignments than NBS (**Figure 18b**). In lung cancer, NBS subtypes and those based on mRNA-seq both had good predictive power (**Figure 18c**) and had some overlap in terms of patient assignments (**Figure 18d**), whereas other data types were not predictive of survival. In uterine cancer, subtypes derived from all data types were highly predictive of

histology (**Figure 18c**, CNVs had highest predictive power overall) and also had very high overlap with NBS subtype assignments (**Figure 18f**).

Example 4: Distinct network modules associate with each tumor subtype

[0268] Identification of the regions of the network responsible for discriminating the somatic mutation profiles of tumors of different subtypes was performed. Ovarian cancer was used as a proof-of-principle, and for each subtype genes were identified for which the network-smoothed mutation state differs significantly for patients of that subtype versus the others ($FDR < 0.05$, see under Methods). The set of genes were projected onto the HumanNet network and visualized using Cytoscape. The network for subtype 1 (**Figures 19 and 25**), which had the worst overall survival and shortest platinum-free interval, contained over 20 genes in the fibroblast growth pathway, which had previously been implicated as a driver of tumor progression and associated with resistance to platinum and anti-VEGF therapy. Subtype 1 has the lowest survival and highest platinum resistance rates amongst the four recovered subtypes. Node size corresponds to smoothed mutation scores. Thickened node outlines indicate genes which are known cancer genes included in the COSMIC cancer gene census. The network for subtype 2 was enriched in DNA damage response genes including *ATM*, *ATR*, *BRCA1/2*, *RAD51* and *CHEK2* (**Figure 31**). Collectively these are characteristic of a functional deficit in response to DNA damage, which has been referred to as ‘BRCAness’. Consistent with this finding, this subtype also included the vast majority of patients with *BRCA1* and *BRCA2* germline mutations (15/20 and 5/6 patients in the cohort, respectively). The network for subtype 3 was enriched for genes in the NF- κ B pathway (**Figure 32**), while subtype 4 was enriched for genes involved in cholesterol transport and fat and glycogen metabolism (**Figure 33**). A similar analysis in uterine and lung cancers produced other sub-networks with unique characteristics, including enrichments for DNA damage response, WNT signaling and histone modification to name a few.

[0269] From the results, the NBS approach was able to stratify patients into clinically informative subtypes and was also useful in identifying the molecular network regions commonly mutated in each subtype.

Example 5: Translation to predictive signatures

[0270] For network-based stratification to be applicable to new patients not in the TCGA, it was necessary to complement it with a procedure for assigning a single patient to one of the existing NBS subtypes. For this purpose, the nearest shrunken centroid approach, a standard method for sample classification which summarizes each subtype with a class ‘centroid’ and assigns new samples to the

subtype with closest centroid, was explored. The shrunken centroid approach method was able to classify the network-smoothed mutation profile of an individual patient with over 95% accuracy (**Figure 20**).

[0271] mRNA expression data are more widely available than full genome or exome sequences, such that there are numerous existing cohorts of cancer patients that have been profiled in mRNA expression but not in somatic mutations. To test the possibility of assigning a new patient to subtypes using an expression signature, mRNA expression profiles available for the TCGA ovarian tumor cohort were used to learn an expression signature for each subtype defined earlier by NBS. The nearest shrunken centroid approach was used again and expression performed as an adequate surrogate for mutation profile, albeit at a reduced accuracy (**Figure 20a**, >95% for mutations, ~60% for expression, ~30% at random). This expression signature was nonetheless able to recover stratification predictive of survival (**Figure 20b**).

[0272] The predictive value of this gene expression signature in two independent studies of serous ovarian tumors by Tothill *et al.* and Bonome *et al.*, as well as a meta-analysis including over 1000 patients (**Figures 20c and 34**), which subsumes Tothill, Bonome and TCGA samples which included expression profiles but lacked somatic mutation profiles, were examined (**Figures 20c and 34**). It was also noted that this meta-analysis included an unknown number of non-serous ovarian cancer samples. Using the expression signature, all patients were assigned to one of the four NBS subtypes. In the Tothill dataset, the subtype assignments were found to be significantly predictive of patient survival and platinum drug resistance (Logrank $P = 6.1 \times 10^{-3}$ and 1.65×10^{-6} respectively, **Figures 20c and 34**), following the same trends observed in the original TCGA cohort. In the Bonome and the meta-analysis datasets, the gene expression signature again recovered four subtypes that were significantly associated with patient survival (Logrank $P = 1.40 \times 10^{-3}$ and 1.22×10^{-4} , respectively, **Figure 34**). As a final control, clustering of the Tothill expression profiles independent of NBS subtypes was performed, resulting in a different set of subtypes that associated with survival to a more limited extent ($P = 0.01$, **Figure 35**). These results show that tumor subtypes identified by NBS are identifiable in independent data sets using gene expression as a surrogate biomarker.

Example 6: Effects of different classes of mutation on stratification

[0273] The impacts of different classes of somatic mutation on the NBS approach were investigated. Although TCGA somatic mutations are profiled only in coding regions (i.e. exomes), a fraction of these mutations do not alter the protein sequence and thus are classified as ‘synonymous’ (e.g. 23% in ovarian cancer). Therefore, the effect on NBS of disrupting synonymous mutations were tested, by reassigning them to new randomly-chosen gene locations. For uterine and lung cancer (**Figures 21a and 36**), disruption of synonymous mutations had little effect on NBS performance, in sharp contrast to

disruption of non-synonymous mutations or all mutations which greatly affected performance. Interestingly, in the ovarian cancer cohort (**Figure 21b**), disruption of both synonymous and non-synonymous mutations was detrimental to stratification performance.

[0274] The effect of removing mutations judged to be non-functional in cancer by methods such as MutationAssessor, CHASM and VEST were also investigated, which use features such as sequence conservation and protein structural information to assess the likely impact of mutations. After using each method to score all mutations across all patients, a permissive threshold was picked for retaining mutations to use for NBS (retaining the top 75% of mutations as scored by CHASM and VEST, and using MutationAssessor with the ‘low threshold’ setting). Filtering mutations with these tools resulted in decreased association of NBS subtypes with patient survival in all three cancers (**Figures 21c–21e**, with possible exception of VEST for ovarian tumors, **Figure 21d**). Finally, the effect of removing long genes or genes with late cell-cycle replication times were investigated, both of which have been postulated to accrue high numbers of mutations that may be unrelated to tumor progression. Removal of long genes substantially degraded the ability to identify ovarian and lung subtypes predictive of survival (**Figures 21d–21e**). However, removal of late-replicating genes had little effect, and in the case of the lung tumor cohort actually increased predictive power (**Figure 21e**).

Example 7: Assigning an independent cohort of patients to ovarian cancer subtypes identified using NBS

[0275] Decision tree classifiers were trained using the Logit-Boost algorithm to recover NBS subtypes in the TCGA cohort directly on mutation profiles. One classifier is trained to recover subtype 1 vs. a set of genes indicated in **Table 2**, and another one is trained to recover subtype 3 or 4 vs. the genes indicated in **Tables 3 and 4**.

Table 2. Predictor genes in a decision tree classifier of subtype 1.

Gene Symbol	Predictor weight
TTN	1.20316723
AHNAK	0.03705563
MUC17	0.01333318
COL4A6	0.00910614
FREM2	0.00840288
USH2A	0.00787923
NEB	0.00666242
RYR2	0.00655503
HUWE1	0.0062256
PKHD1L1	0.00529436
CD163L1	0.00429516
LRRK2	0.00345253

ANKRD30A	0.00114832
MUC16	0.00043608
TP53	0.00033373
LRP2	0.00014659

Table 3: Predictor genes in a decision tree classifier of subtype 3/4.

AHNAK	2.6896E-01
MYH4	2.4122E-01
MAP2	4.0769E-02
RNF213	1.6100E-02
SMARCA4	1.3451E-02
COL4A6	7.4772E-03
HUWE1	7.3140E-03
STAB2	3.2544E-03
ANK3	2.4655E-03
CUBN	1.5438E-03
DNAH17	1.4777E-03

Table 4: Predictor genes in a decision tree classifier of subtype 3/4.

AHNAK	2.6896E-01
MYH4	2.4122E-01
MAP2	4.0769E-02
RNF213	1.6100E-02
SMARCA4	1.3451E-02
COL4A6	7.4772E-03
HUWE1	7.3140E-03
STAB2	3.2544E-03
ANK3	2.4655E-03
CUBN	1.5438E-03
DNAH17	1.4777E-03
FAT4	7.5730E-04
TTN	6.1382E-04
BRCA2	5.6685E-04
WWC3	3.7552E-04
PPP1R3A	2.7560E-04
PRKDC	2.5551E-04
CD163L1	1.9744E-04
CDK12	1.4573E-04
USH2A	1.3120E-04
NEB	1.2085E-04
MYO18B	5.2858E-05
PTPRT	3.8900E-05
LRRC7	3.3668E-05
FREM2	3.2133E-05
AHNAK2	2.8186E-05
MYH1	2.4526E-05
SPTB	2.0480E-05

LRRK2	1.9006E-05
LRP2	1.3731E-05
NF1	6.9520E-06
DST	6.7974E-06
DNAH10	6.1227E-06
CCDC141	4.7854E-06
LATS1	4.4014E-06
PKHD1	2.7760E-06
MUC16	2.4884E-06
UNC79	2.1611E-06
PKHD1L1	2.0802E-06
MGAM	1.9194E-06
HERC1	1.6881E-06
ITK	1.4060E-06
FAT1	1.3651E-06
TRRAP	1.3104E-06
MUC17	1.0609E-06
DNAH5	8.5052E-07
RIMS1	8.4248E-07
ZFYVE26	8.1133E-07
FAT3	5.6177E-07
RYR2	4.2752E-07
FLG	2.6766E-07
CACNA1C	2.4265E-07
FCGBP	2.1816E-07
PCDH15	2.0032E-07
GPR98	1.8421E-07
INSRR	1.7746E-07
PEG3	1.4167E-07
ZNF236	1.0974E-07
PCDHA3	1.0025E-07
GLI2	6.8926E-08
DYNC1H1	5.5331E-08
CSMD3	4.9092E-08
TP53	2.8674E-08
ADAMTS20	2.5597E-08
TRPA1	2.1639E-08
COL22A1	7.7865E-09
L1CAM	6.7045E-09
RIMS2	6.5216E-09
ZFHX4	6.0407E-09
DSCAM	5.2414E-09
STXBP5L	4.6067E-09
HMCN1	2.2665E-09
DNAH3	4.7000E-10
COL6A3	8.9601E-11

[0276] In cross validation these classifiers achieve an area under the ROC curve of 95% and 94% respectively. Next, the classifiers are used to assign a subtype in an independent cohort of patients from the International Cancer Genome Consortium (ICGC). In case of ambiguity (i.e. a patient is assigned to both type 1 and types 3/4, we assign the patient to the latter). Survival analysis is performed after excluding stage IV patients and patients older than 75 years of age. The resulting 3 subtypes follow a survival trend similar to that observed in the TCGA cohort (**Figure 22**). We further apply the predictor for subtype 1 on sequence data collected for the cancer cell line encyclopedia using targeted sequencing and not whole exome sequencing. The top scoring subtype 1 cell-lines differ significantly from the bottom scoring cell lines (**Figure 23**).

[0277] The techniques described in this disclosure may be implemented in hardware, software, firmware, or combinations thereof. If implemented in software, the techniques may be realized at least in part by a computer-readable medium comprising instructions that, when executed, performs one or more of the methods described above. The computer-readable medium may form part of a computer program product, which may include packaging materials. The computer-readable medium may comprise random access memory (RAM) such as synchronous dynamic random access memory (SDRAM), read-only memory (ROM), non-volatile random access memory (NVRAM), electrically erasable programmable read-only memory (EEPROM), FLASH memory, magnetic or optical data storage media, and the like. The techniques additionally, or alternatively, may be realized at least in part by a computer-readable communication medium that carries or communicates code in the form of instructions or data structures and that can be accessed, read, and/or executed by a computer.

[0278] The methods described herein can be implemented on any conventional host computer system, such as those based on Intel® or AMD® microprocessors and running Microsoft Windows operating systems. Other systems, such as those using the UNIX or LINUX operating system and based on IBM®, DEC® or Motorola® microprocessors are also contemplated. The systems and methods described herein can also be implemented to run on client-server systems and wide-area networks, such as the Internet.

[0279] Software to implement a method or model of the invention can be written in any well-known computer language, such as Java, C, C++, Visual Basic, FORTRAN or COBOL and compiled using any well-known compatible compiler. The software of the invention normally runs from instructions stored in a memory on a host computer system. A memory or computer readable medium can be a hard disk, floppy disc, compact disc, DVD, magneto-optical disc, Random Access Memory, Read Only Memory or Flash Memory. The memory or computer readable medium used in the invention can be contained within a single computer or distributed in a network. A network can be any of a number of conventional network systems known in the art such as a local area network (LAN) or a wide area

network (WAN). Client-server environments, database servers and networks that can be used in the invention are well known in the art. For example, the database server can run on an operating system such as UNIX, running a relational database management system, a World Wide Web application and a World Wide Web server. Other types of memories and computer readable media are also contemplated to function within the scope of the invention.

[0280] The data matrices constructed by the methods described in this invention can be represented without limitation in a flat text file, in an SQL or noSQL database, or in a markup language format including, for example, Standard Generalized Markup Language (SGML), Hypertext markup language (HTML) or Extensible Markup language (XML). Markup languages can be used to tag the information stored in a database or data structure of the invention, thereby providing convenient annotation and transfer of data between databases and data structures. In particular, an XML format can be useful for structuring the data representation of reactions, reactants and their annotations; for exchanging database contents, for example, over a network or internet; for updating individual elements using the document object model; or for providing differential access to multiple users for different information content of a data base or data structure of the invention. XML programming methods and editors for writing XML code are known in the art as described, for example, in Ray, *Learning XML* O'Reilly and Associates, Sebastopol, CA (2001).

[0281] A computer system of the invention can further include a user interface capable of receiving a representation of one or more reactions. A user interface of the invention can also be capable of sending at least one command for modifying the data structure, the constraint set or the commands for applying the constraint set to the data representation, or a combination thereof. The interface can be a graphic user interface having graphical means for making selections such as menus or dialog boxes. The interface can be arranged with layered screens accessible by making selections from a main screen. The user interface can provide access to other databases useful in the invention such as other gene or protein networks, gene mutation data, a metabolic reaction database or links to other databases having information relevant to the reactions or reactants in the reaction network data structure or to human physiology. Also, the user interface can display a graphical representation of a gene or protein network or another biological network or the results of the stratification, clinical phenotypes or subtypes or subtype assignment derived using the invention.

[0282] The above description discloses several methods and materials of the present invention. This invention is susceptible to modifications in the methods and materials, as well as alterations in the fabrication methods and equipment. Such modifications will become apparent to those skilled in the art from a consideration of this disclosure or practice of the invention disclosed herein. Consequently, it is not intended that this invention be limited to the specific embodiments disclosed

herein, but that it cover all modifications and alternatives coming within the true scope and spirit of the invention.

[0283] The foregoing description and Examples detail certain embodiments. It will be appreciated, however, that no matter how detailed the foregoing may appear in text, the invention may be practiced in many ways and the invention should be construed in accordance with the appended claims and any equivalents thereof.

REFERENCES

1. The International Cancer Genome Consortium International network of cancer genome projects. *Nature* **464**, 993-996 (2010).
2. The Cancer Genome Atlas Research Network Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
3. Cancer Genome Atlas Research, N. et al. Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73 (2013).
4. Brunham, L.R. & Hayden, M.R. Whole-genome sequencing: the new standard of care? *Science* **336**, 1112-1113 (2012).
5. Chin, L., Andersen, J.N. & Futreal, P.A. Cancer genomics: from discovery science to personalized medicine. *Nature medicine* **17**, 297-303 (2011).
6. Konstantinopoulos, P.A., Spentzos, D. & Cannistra, S.A. Gene-expression profiling in epithelial ovarian cancer. *Nature clinical practice. Oncology* **5**, 577-587 (2008).
7. Konstantinopoulos, P.A. et al. Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 3555-3561 (2010).
8. Reis-Filho, J.S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* **378**, 1812-1823 (2011).
9. Esteva, F.J. et al. Prognostic role of a multigene reverse transcriptase-PCR assay in patients with node-negative breast cancer not receiving adjuvant systemic therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **11**, 3315-3319 (2005).
10. The Cancer Genome Atlas Network Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012).
11. Raspe, E., Decraene, C. & Berx, G. Gene expression profiling to dissect the complexity of cancer biology: pitfalls and promise. *Seminars in cancer biology* **22**, 250-260 (2012).

12. Mardis, E.R. Genome sequencing and cancer. *Current opinion in genetics & development* **22**, 245-250 (2012).
13. Carter, H. et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research* **69**, 6660-6667 (2009).
14. Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).
15. Wang, K. et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nature genetics* **43**, 1219-1223 (2011).
16. Dulak, A.M. et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature genetics* **45**, 478-486 (2013).
17. Allegra, C.J. et al. American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 2091-2096 (2009).
18. The Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature advance online publication* (2012).
19. Muzny, D.M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337 (2012).
20. Kreeger, P.K. & Lauffenburger, D.A. Cancer systems biology: a network modeling perspective. *Carcinogenesis* **31**, 2-8 (2010).
21. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).
22. CH, W. Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563-565 (1942).
23. Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. & Marcotte, E.M. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research* **21**, 1109-1121 (2011).
24. Vandin, F., Upfal, E. & Raphael, B.J. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* **18**, 507-522 (2011).
25. Vaske, C.J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237-245 (2010).
26. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research* **22**, 398-406 (2012).

27. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Molecular systems biology* **3**, 140 (2007).
28. Dutkowski, J. & Ideker, T. Protein networks as logic functions in development and cancer. *PLoS computational biology* **7**, e1002180 (2011).
29. Szklarczyk, D. et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561-568 (2011).
30. Cerami, E.G. et al. Pathway Commons, a web resource for biological pathway data. *Nucleic acids research* **39**, D685-690 (2011).
31. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology* **6**, e1000641 (2010).
32. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791 (1999).
33. Cai, D., He, X., Wu, X. & Han, J. Non-negative matrix factorization on manifold. *IEEE Xplore digital library*, 63-72 (2008).
34. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**, 91-118 (2003).
35. Yang, D. et al. Association of BRCA1 and BRCA2 mutations with survival, chemotherapy sensitivity, and gene mutator phenotype in patients with ovarian cancer. *JAMA : the journal of the American Medical Association* **306**, 1557-1565 (2011).
36. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L. & Ideker, T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27**, 431-432 (2011).
37. Cole, C. et al. Inhibition of FGFR2 and FGFR1 increases cisplatin sensitivity in ovarian cancer. *Cancer biology & therapy* **10**, 495-504 (2010).
38. Wysham, W.Z. et al. BRCAness profile of sporadic ovarian cancer predicts disease recurrence. *PloS one* **7**, e30042 (2012).
39. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunk centroids of gene expression. *Proceedings of the National Academy of Sciences* **99**, 6567-6572 (2002).
40. Le Page, C. et al. Gene expression profiling of primary cultures of ovarian epithelial cells identifies novel molecular classifiers of ovarian cancer. *British journal of cancer* **94**, 436-445 (2006).

41. Tothill, R.W. et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical cancer research : an official journal of the American Association for Cancer Research* **14**, 5198-5208 (2008).
42. Györfy, B., Lánckzy, A. & Szállási, Z. Implementing an online tool for genome-wide validation of survival-associated biomarkers in ovarian-cancer using microarray data from 1287 patients. *Endocrine-related cancer* **19**, 197-208 (2012).
43. Bonome, T. et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer research* **68**, 5478-5486 (2008).
44. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118 (2011).
45. Douville, C. et al. CRAVAT: cancer-related analysis of variants toolkit. *Bioinformatics* **29**, 647-648 (2013).
46. Stamatoyannopoulos, J.A. et al. Human mutation rate associated with DNA replication timing. *Nature genetics* **41**, 393-395 (2009).
47. Rabiner, L.R. & Gold, B. Theory and application of digital signal processing. (Prentice-Hall, Englewood Cliffs, N.J.; 1975).
48. Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. *Nature Reviews Cancer* **10**, 116-129 (2010).
49. Birrer, M.J. et al. Whole genome oligonucleotide-based array comparative genomic hybridization analysis identified fibroblast growth factor 1 as a prognostic marker for advanced-stage serous ovarian adenocarcinomas. *Journal of Clinical Oncology* **25**, 2281-2287 (2007).
50. Futreal, P.A. et al. A census of human cancer genes. *Nature reviews. Cancer* **4**, 177-183 (2004).
51. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546-1558 (2013).
52. Simon, D.N. & Wilson, K.L. The nucleoskeleton as a genome-associated dynamic 'network of networks'. *Nature reviews. Molecular cell biology* **12**, 695-708 (2011).
53. Liu, Y. et al. Integrated analysis of gene expression and tumor nuclear image profiles associated with chemotherapy response in serous ovarian carcinoma. *PloS one* **7**, e36383 (2012).
54. Strauss, B.S. Role in tumorigenesis of silent mutations in the TP53 gene. *Mutation research* **457**, 93-104 (2000).
55. Kimchi-Sarfaty, C. et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525-528 (2007).
56. Sauna, Z.E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nature reviews. Genetics* **12**, 683-691 (2011).

57. Salzman, D.W. & Weidhaas, J.B. miRNAs in the spotlight: Making 'silent' mutations speak up. *Nature medicine* **17**, 934-935 (2011).
58. Rand, W.M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66**, 846-850 (1971).
59. The Broad Institute Firehose portal analyses run from May 25th 2012, http://gdac.broadinstitute.org/runs/analyses_2012_05_25/reports/cancer/OV/. (2012).
60. Brunet, J.P., Tamayo, P., Golub, T.R. & Mesirov, J.P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **101**, 4164-4169 (2004).
61. Verhaak, R.G. et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98-110 (2010).
62. Ruan, J. & Zhang, W. Identifying network communities with a high resolution. *Physical Review E* **77**, 016104 (2008).
63. Verhaak, R.G. et al. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *The Journal of clinical investigation* **123**, 517-525 (2013).
64. Schaefer, C.F. et al. PID: the Pathway Interaction Database. *Nucleic acids research* **37**, D674-679 (2009).
65. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* **98**, 5116-5121 (2001).
66. Andersen, P.K. & Gill, R.D. Cox's regression model for counting processes: a large sample study. *The annals of statistics* **10**, 1100-1120 (1982).
67. Stenson, P.D. et al. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.] Chapter 1*, Unit1 13 (2012).
68. Hansen, R.S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 139-144 (2010).
69. Kent, W.J. et al. The human genome browser at UCSC. *Genome research* **12**, 996-1006 (2002).

WHAT IS CLAIMED IS:

1. A method for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor comprising:
 - (a) Obtaining nucleic acid sequence information from the subject;
 - (b) Determining mutational status from the nucleic acid sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor, thereby diagnosing the subject as having one or more informative subtypes of a cancer or tumor.
2. A method for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor comprising:
 - (a) Obtaining protein sequence information from the subject;
 - (b) Determining mutational status from the protein sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor, thereby diagnosing the subject as having one or more informative subtypes of a cancer or tumor.
3. A method for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor comprising:
 - (a) Obtaining epigenetic modification information for genomic DNA from the subject;
 - (b) Determining epigenetic modification status from the epigenetic modification information so obtained;
 - (c) Transforming the epigenetic status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or

tumor, thereby diagnosing the subject as having one or more informative subtypes of a cancer or tumor.

4. A method for diagnosing a subject in need thereof as having one or more informative subtypes of a cancer or tumor comprising:
 - (a) Obtaining RNA modification information for RNAs from the subject;
 - (b) Determining RNA modification status from the RNA modification information so obtained;
 - (c) Transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Comparing the transformed profile of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor, thereby diagnosing the subject as having one or more informative subtypes of a cancer or tumor.
5. A method for diagnosing a subject in need thereof with one or more informative subtypes of a cancer or tumor comprising:
 - (a) Obtaining post-translational modification information for proteins from the subject;
 - (b) Determining post-translational modification status from the post-translational modification information so obtained;
 - (c) Transforming the post-translational modification status of step (b) into a transformed profile of the subject based on a reference molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Comparing the transformed profile of the subject of step (c) with reference transformed profiles corresponding to subjects grouped into one or more informative subtypes of a cancer or tumor, thereby diagnosing the subject with one or more informative subtypes of a cancer or tumor.
6. The method of claim 1, wherein the nucleic acid sequence information is obtained from genomic DNA of a subject(s).
7. The method of claim 2, wherein the protein sequence information is obtained from conceptual translation of protein coding sequences or expressed proteins of a subject(s).

8. The method of claim 3, wherein the epigenetic modification information is obtained from genomic DNA of a subject(s).
9. The method of claim 4, wherein the RNA modification information is obtained from RNAs of a subject(s).
10. The method of claim 5, wherein the post-translational modification information for proteins is obtained from proteins of a subject(s).
11. The method of claim 1, 3 or 4, wherein the transformed profile in step (c) comprises one or more or all of the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
12. The method of claim 2 or 5, wherein the transformed profile in step (c) comprises one or more or all of the proteins encoded by the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
13. The method of claim 1, wherein in step (b) determining the mutational status from the nucleic acid sequence information is effected by comparing the nucleic acid sequence information to a reference information for nucleic acid sequence and determining the presence of differences between the nucleic acid sequence information and the reference information, the difference being indicative of the mutational status of the nucleic acid sequence information.
14. The method of claim 2, wherein in step (b) determining the mutational status from the protein sequence information is effected by comparing it to a reference information for protein sequence and determining the presence of differences from the reference information.
15. The method of claim 3, wherein in step (b) determining the epigenetic modification status from the epigenetic modification information is effected by comparing it to a reference epigenetic information and determining the presence of differences from the reference information.
16. The method of claim 4, wherein in step (b) determining the RNA modification status from the RNA modification information is effected by comparing it to a reference RNA modification information and determining the presence of differences from the reference information.

17. The method of claim 5, wherein in step (b) determining the post-translational modification status from the post-translational modification information is effected by comparing it to a reference post-translational modification information and determining the presence of differences from the reference information.
18. The method of claim 13, 14, 15, 16 or 17, wherein the reference information is obtained from subjects without cancer or tumor.
19. The method of claim 1, wherein in step (c) transforming the mutational status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the nucleic acid sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of genes, within a network based on network proximity to mutated genes, thereby producing a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression.
20. The method of claim 2, wherein in step (c) transforming the mutational status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the protein sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of proteins within a network based on network proximity to mutated proteins thereby producing a transformed profile of the subject based on molecular network(s) of protein interaction or expression.
21. The method of claim 3, wherein in step (c) transforming the epigenetic modification status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the epigenetic modification information onto a network and (b) propagating the mutation(s) or change(s) in the network so as to obtain a continuous range of values for all or subset of epigenetic markers within a network based on network proximity to epigenetic modification changes, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.
22. The method of claim 4, wherein in step (c) transforming the RNA modification status into a transformed profile of the subject is effected by (a) projecting any difference(s) found within the RNA modification information onto a network and (b) propagating the difference(s) in the

network so as to obtain a continuous range of values for all or subset of RNAs, genes encoding RNAs or nucleic acids encoding RNAs within a network based on network proximity to RNA modification differences, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.

23. The method of claim 5, wherein in step (c) transforming the post-translational modification status into a transformed profile of the subject is effected by (a) projecting any difference(s) found within the post-translational modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of proteins, genes encoding proteins or nucleic acids encoding proteins within a network based on network proximity to post-translational modification differences, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.
24. The method of claim 1, 2, 3, 4 or 5, wherein in step (d) comparing the transformed profile of step (c) with reference transformed profiles is effected by assigning the subject to a subtype of a cancer or tumor with closest reference profile, thereby diagnosing which one or more informative subtypes of the cancer or tumor the subject belongs.
25. The method of claim 24, wherein the closest reference profile is effected by application of a nearest shrunken centroid approach, a supervised learning approach based on decision tree classifiers or another supervised or unsupervised learning approach or combination of such approaches.
26. The method of claim 1, 2, 3, 4 or 5, wherein the informative subtype is a clinical phenotype.
27. The method of claim 26, wherein the clinical phenotype is predictive of a survival rate, drug response or tumor grade.
28. The method of claim 1, wherein the mutation is a somatic mutation.
29. The method of claim 28, wherein the somatic mutation is in genomic DNA.

30. The method of claim 28, wherein the somatic mutation is an exonic mutation or a mutation in an exon.
31. The method of claim 30, wherein the exonic mutation or the mutation in an exon alters a protein coding sequence.
32. The method of claim 31, wherein the exonic mutation or the mutation in an exon is a synonymous mutation or a silent mutation that does not alter a protein sequence.
33. The method of claim 31, wherein the exonic mutation or the mutation in an exon is a non-synonymous mutation that alters a protein sequence.
34. The method of claim 31, wherein the somatic mutation is a synonymous mutation or a non-synonymous mutation.
35. The method of claim 28, wherein the somatic mutation is in a gene.
36. The method of claim 35, wherein the gene is a protein coding gene.
37. The method of claim 36, wherein the protein coding gene is transcribed by RNA polymerase II.
38. The method of claim 35, wherein the somatic mutation in a gene is in a promoter, enhancer, transcriptional terminator, intron, untranslated region (5' UTR or 3' UTR), exon-intron junction, splice site, splicing branch site, polyadenylation signal or other genetic elements.
39. The method of claim 35, wherein the gene is a non-protein coding gene.
40. The method of claim 39, wherein the non-protein coding gene encodes a rRNA gene, a tRNA gene, a snRNA gene, a miRNA gene or a gene for a structural RNA.
41. The method of claim 28, wherein the somatic mutation is in an extragenic region or a mutation outside of a gene of a subject's genome.

42. The method of claim 28, wherein the somatic mutation is in a middle repetitive DNA sequence or highly repetitive DNA sequence.
43. The method of claim 28, wherein the somatic mutation is in a transcribed region of a subject's genome.
44. The method of claim 28, wherein the somatic mutation is in an untranscribed region of a subject's genome.
45. The method of claim 28, wherein the somatic mutation is in nuclear or mitochondrial DNA.
46. The method of claim 1, 2, 3, 4 or 5, wherein the cancer is breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer.
47. The method of claim 1, 2, 3, 4, or 5, wherein the method is carried out by an informatics platform.
48. The method of claim 1, 2, 3, 4, or 5, wherein the informatics platform is a bioinformatics platform comprising a computer and software.
49. The method of claim 48, wherein the software uses supervised learning and/or unsupervised learning methods.
50. The method of claim 1, 2, 3, 4, or 5, wherein the method is an automated method.
51. The method of claim 1, 2, 3, 4, or 5, wherein step (a) of the method requires selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5.
52. The method of claim 51, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5 comprises one or more or all of the genes, proteins, RNAs, nucleic acid

sequences or protein sequences associated with the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.

53. The method of claim 52, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5 comprises selecting all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with each group of genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
54. The method of claim 52, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5 comprises selecting two or more groups of genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with groups of genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 42, 42, and 43.
55. A method for identifying one or more informative subtypes of a cancer or tumor comprising:
- (a) Obtaining nucleic acid sequence information from subjects with a cancer or tumor;
 - (b) Determining mutational status for each subject from the nucleic acid sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.
56. A method for identifying one or more informative subtypes of a cancer or tumor comprising:
- (a) Obtaining protein sequence information from subjects with a cancer or tumor;
 - (b) Determining mutational status for each subject from the protein sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.

57. A method for identifying one or more informative subtypes of a cancer or tumor comprising:
- (a) Obtaining epigenetic modification information from subjects with a cancer or tumor;
 - (b) Determining epigenetic modification status for each subject from the epigenetic modification information so obtained;
 - (c) Transforming the epigenetic modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.
58. A method for identifying one or more informative subtypes of a cancer or tumor comprising:
- (a) Obtaining RNA modification information from subjects with a cancer or tumor;
 - (b) Determining RNA modification status for each subject from the RNA modification information so obtained;
 - (c) Transforming the RNA modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.
59. A method for identifying one or more informative subtypes of a cancer or tumor comprising:
- (a) Obtaining post-translational modification information from subjects with a cancer or tumor;
 - (b) Determining post-translational modification status for each subject from the post-translational modification information so obtained;
 - (c) Transforming the post-translational modification status to a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.

60. The method of claim 55, wherein the nucleic acid sequence information is obtained from genomic DNA of subjects.
61. The method of claim 56, wherein the protein sequence information is obtained from conceptual translation of protein coding sequences or expressed proteins of subjects.
62. The method of claim 57, wherein the epigenetic modification information is obtained from genomic DNA of subjects.
63. The method of claim 58, wherein the RNA modification information is obtained from RNAs of subjects.
64. The method of claim 59, wherein the post-translational modification information for proteins is obtained from proteins of subjects.
65. The method of claim 55, 57 or 58, wherein the transformed profile in step (c) comprises one or more or all of the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
66. The method of claim 56 or 59, wherein the transformed profile in step (c) comprises one or more or all of the proteins encoded by the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
67. The method of claim 55, wherein in step (b) determining the mutational status from the nucleic acid sequence information is effected by comparing it to a reference information for nucleic acid sequence and determining the presence of differences from the reference information.
68. The method of claim 56, wherein in step (b) determining the mutational status from the protein sequence information is effected by comparing it to a reference information for protein sequence and determining the presence of differences from the reference information.
69. The method of claim 57, wherein in step (b) determining the epigenetic modification status from the epigenetic modification information is effected by comparing it to a reference epigenetic information and determining the presence of differences from the reference information.

70. The method of claim 58, wherein in step (b) determining the RNA modification status from the RNA modification information is effected by comparing it to a reference RNA modification information and determining the presence of differences from the reference information.
71. The method of claim 59, wherein in step (b) determining the post-translational modification status from the post-translational modification information is effected by comparing it to a reference post-translational modification information and determining the presence of differences from the reference information.
72. The method of claim 67, 68, 69, 70 or 71, wherein the reference information is obtained from subjects without cancer or tumor.
73. The method of claim 55, wherein in step (c) transforming the mutational status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the nucleic acid sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of genes, within a network based on network proximity to mutated genes, thereby producing a transformed profile of the subject based on molecular network(s) of gene, RNA and/or protein interaction or expression.
74. The method of claim 56, wherein in step (c) transforming the mutational status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the protein sequence information onto a network and (b) propagating the mutation(s) in the network so as to obtain a continuous range of values for all or subset of proteins within a network based on network proximity to mutated proteins thereby producing a transformed profile of the subject based on molecular network(s) of protein interaction or expression.
75. The method of claim 57, wherein in step (c) transforming the epigenetic modification status into a transformed profile of the subject is effected by (a) projecting any mutation(s) found within the epigenetic modification information onto a network and (b) propagating the mutation(s) or change(s) in the network so as to obtain a continuous range of values for all or subset of epigenetic markers within a network based on network proximity to epigenetic modification changes, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.

76. The method of claim 58, wherein in step (c) transforming the RNA modification status into a transformed profile of the subject is effected by (a) projecting any difference(s) found within the RNA modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of RNAs, genes encoding RNAs or nucleic acids encoding RNAs within a network based on network proximity to RNA modification differences, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.
77. The method of claim 59, wherein in step (c) transforming the post-translational modification status into a transformed profile of the subject is effected by (a) projecting any difference(s) found within the post-translational modification information onto a network and (b) propagating the difference(s) in the network so as to obtain a continuous range of values for all or subset of proteins, genes encoding proteins or nucleic acids encoding proteins within a network based on network proximity to post-translational modification differences, thereby producing a transformed profile of the subject based on molecular network(s) of the gene, RNA and/or protein interaction or expression.
78. The method of claim 55, 56, 57, 58 or 59, wherein in step (d) clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes is effected by grouping subjects with similar transformed profiles into one or more groups or subtypes, thereby identifying one or more informative subtypes of a cancer or tumor.
79. The method of claim 55, 56, 57, 58 or 59, wherein the informative subtype is a clinical phenotype.
80. The method of claim 79, wherein the clinical phenotype is predictive of a survival rate, drug response or tumor grade.
81. The method of claim 55, wherein the mutation is a somatic mutation.
82. The method of claim 81, wherein the somatic mutation is in genomic DNA.

83. The method of claim 81, wherein the somatic mutation is an exonic mutation or a mutation in an exon.
84. The method of claim 83, wherein the exonic mutation or the mutation in an exon alters a protein coding sequence.
85. The method of claim 84, wherein the exonic mutation or the mutation in an exon is a synonymous mutation or a silent mutation that does not alter a protein sequence.
86. The method of claim 84, wherein the exonic mutation or the mutation in an exon is a non-synonymous mutation that alters a protein sequence.
87. The method of claim 84, wherein the somatic mutation is a synonymous mutation or a non-synonymous mutation.
88. The method of claim 81, wherein the somatic mutation is in a gene.
89. The method of claim 88, wherein the gene is a protein coding gene.
90. The method of claim 89, wherein the protein coding gene is transcribed by RNA polymerase II.
91. The method of claim 88, wherein the somatic mutation in a gene is in a promoter, enhancer, transcriptional terminator, intron, untranslated region (5' UTR or 3' UTR), exon-intron junction, splice site, splicing branch site, polyadenylation signal or other genetic elements.
92. The method of claim 88, wherein the gene is a non-protein coding gene.
93. The method of claim 92, wherein the non-protein coding gene encodes a rRNA gene, a tRNA gene, a snRNA gene, a miRNA gene or a gene for a structural RNA.
94. The method of claim 81, wherein the somatic mutation is in an extragenic region or a mutation outside of a gene of a subject's genome.

95. The method of claim 81, wherein the somatic mutation is in a middle repetitive DNA sequence or highly repetitive DNA sequence.
96. The method of claim 81, wherein the somatic mutation is in a transcribed region of a subject's genome.
97. The method of claim 81, wherein the somatic mutation is in an untranscribed region of a subject's genome.
98. The method of claim 81, wherein the somatic mutation is in nuclear or mitochondrial DNA.
99. The method of claim 55, 56, 57, 58 or 59, wherein the cancer is breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer.
100. The method of claim 55, 56, 57, 58 or 59, wherein the method is carried out by an informatics platform.
101. The method of claim 55, 56, 57, 58 or 59, wherein the informatics platform is a bioinformatics platform comprising a computer and software.
102. The method of claim 101, wherein the software uses supervised learning and/or unsupervised learning methods.
103. The method of claim 55, 56, 57, 58 or 59, wherein the method is an automated method.
104. The method of claim 55, 56, 57, 58 or 59, wherein step (a) of the method requires selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5.
105. The method of claim 104, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of

- claim 1, 2, 3, 4, or 5 comprises one or more or all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with the genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
106. The method of claim 105, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5 comprises selecting all of the genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with each group of genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
107. The method of claim 105, wherein selecting a subset of genes, proteins, RNAs, nucleic acid sequences or protein sequences from which to obtain information at step (a) of the method of claim 1, 2, 3, 4, or 5 comprises selecting two or more groups of genes, proteins, RNAs, nucleic acid sequences or protein sequences associated with groups of genes shown in any of Tables 2, 3 and 4 and Figures 19, 25, 31, 32, 33, 37, 38, 39, 40, 41, 42, and 43.
108. A method of assigning a subject of interest having a cancer or tumor into one or more informative subtypes comprising:
- (a) Obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor;
 - (b) Determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor;
 - (e) obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered in (d) and the subject of interest having cancer or tumor by using a supervised learning approach to derive a subtype classifier based on profiles from the subjects in (a) and their assignment to subtypes; and

- (f) comparing the subtype classifier so derived to assign the subject of interest to a cancer or tumor subtype, thereby assigning a subject of interest having a cancer or tumor into one or more informative subtypes.
109. A method of assigning a subject of interest having a cancer or tumor into one or more informative subtypes comprising:
- (a) Obtaining nucleic acid sequence information from a plurality of subjects with a cancer or tumor;
 - (b) Determining mutational status for each of the plurality of subjects from the nucleic acid sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of each of the plurality of subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes, thereby identifying one or more informative subtypes of a cancer or tumor;
 - (e) obtaining mutation profiles, transformed profiles using a network, biological profiles or gene expression profiles for subjects clustered in (d) and the subject of interest having cancer or tumor; and
 - (f) applying a nearest shrunken centroid approach, a supervised learning approach based on decision tree classifiers or another supervised or unsupervised learning approach or combination of such approaches to assign the subject of interest to a cancer or tumor subtype, thereby assigning a subject of interest having a cancer or tumor into one or more informative subtypes.
110. A method of assigning a subject of interest having a cancer or tumor into one or more informative subtypes comprising:
- (a) Obtaining nucleic acid sequence information from subjects with a cancer or tumor;
 - (b) Determining mutational status for each subject from the nucleic acid sequence information so obtained;
 - (c) Transforming the mutational status to a transformed profile of each of the subjects based on molecular network(s) of gene, RNA and/or protein interaction or expression;
 - (d) Clustering the transformed profiles for subjects obtained in (c) into one or more clusters so as to obtain one or more subtypes;

- (e) Characterizing the subjects grouped into one or more informative subtypes in (d) by determining status or profile of one or more measurable or quantifiable biological parameter(s) or feature(s);
 - (f) Characterizing the subject of interest by determining status or profile of one or more measurable or quantifiable biological parameter(s); and
 - (g) Assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on status or profile(s) of the subjects grouped into one or more informative subtypes in (e) and the status or profile of the subject of interest in (f), thereby identifying one or more informative subtypes of a cancer or tumor to which a subject of interest is to be assigned.
111. A method of assigning a subject of interest having a cancer or tumor into one or more informative subtypes comprising:
- (a) Obtaining biological profiles of subjects grouped into one or more informative subtypes;
 - (b) Obtaining biological profile of the subject of interest; and
 - (c) Assigning a subject of interest having a cancer or tumor into one or more informative subtypes, based on biological profile(s) of the subjects grouped into one or more informative subtypes in (a) and the biological profile of the subject of interest in (b), thereby assigning a subject of interest having a cancer or tumor into one or more informative subtypes.
112. The method of claim 108, 109, 110 or 111, wherein the informative subtype is associated with a clinical phenotype.
113. The method of claim 112, wherein the clinical phenotype is predictive of a survival rate, drug response or tumor grade.
114. The method of claim 108, 109, 110 or 111, wherein the cancer is breast cancer, lung cancer, prostate cancer, ovarian cancer, melanoma, squamous cell carcinoma, colorectal cancer, pancreatic cancer, thyroid cancer, endometrial cancer, uterine sarcoma, uterine cancer, bladder cancer, kidney cancer, a solid tumor, leukemia, non-Hodgkin lymphoma, or a drug-resistant cancer.
115. The method of claim 108, 109, 110 or 111, wherein the tumor is an ovarian tumor.

116. The method of claim 108, 109, 110 or 111, wherein the cancer is ovarian cancer.
117. The method of claim 108, 109, 110 or 111, wherein the subtype is ovarian cancer subtype 1, 2, 3, or 4.
118. The method of claim 108, 109, 110 or 111, wherein the subtype is predictive of survival.
119. The method of claim 108, 109, 110 or 111, wherein the subtype is predictive of response to treatment.
120. The method of claim 119, wherein the treatment involves chemotherapy.
121. The method of claim 108, 109 or 110, wherein the mutation is a somatic mutation.
122. The method of claim 108, 109, 110 or 111, wherein the method is carried out by an informatics platform.
123. The method of claim 122, wherein the informatics platform is a bioinformatics platform comprising a computer and software.
124. The method of claim 123, wherein the software uses supervised learning and/or unsupervised learning methods.
125. The method of claim 108, 109, 110 or 111, wherein the method is an automated method.
126. A method for increasing efficiency of a bioinformatics process for network-based stratification of tumor or cancer comprising:
 - (a) Obtaining a biological sample from a subject with tumor or cancer;
 - (b) Selecting a set of genes for which nucleic acid sequence is to be determined;
 - (c) Determining nucleic acid sequence for protein coding sequences in the set of genes selected in step (b);
 - (d) Projecting mutations found within sequence of step (c) onto a network;
 - (e) Propagating the mutations in the network; and

- (f) Clustering the mutations so propagated so as to divide biological samples from subjects with tumor or cancer into subtypes, wherein in step (b), the set of genes so selected excludes whole exome or genome sequencing, thereby increasing the efficiency of a bioinformatics process for network-based stratification of tumor or cancer.

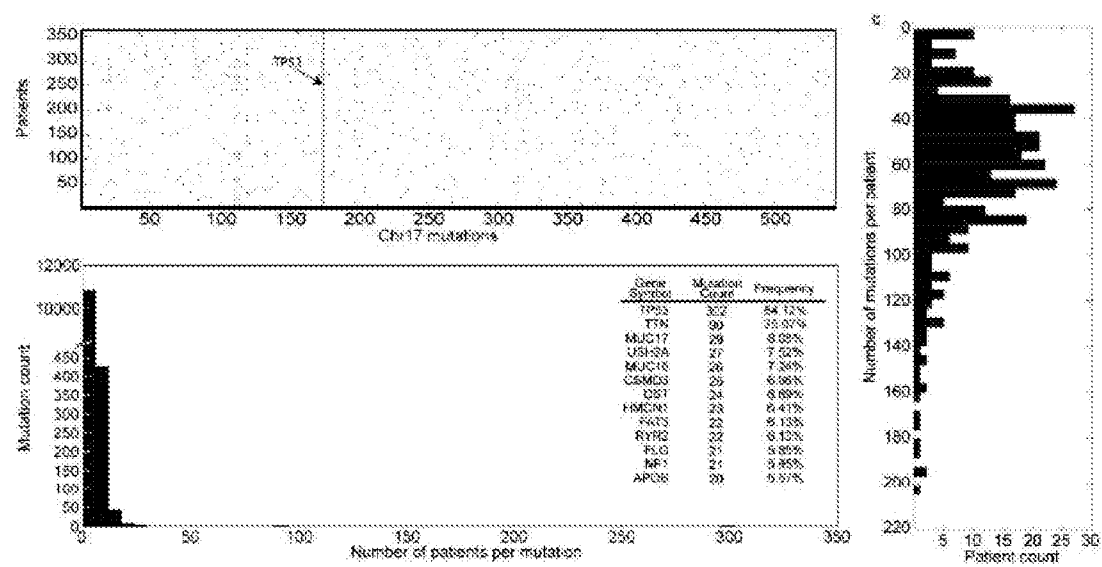


Figure 1

2/44

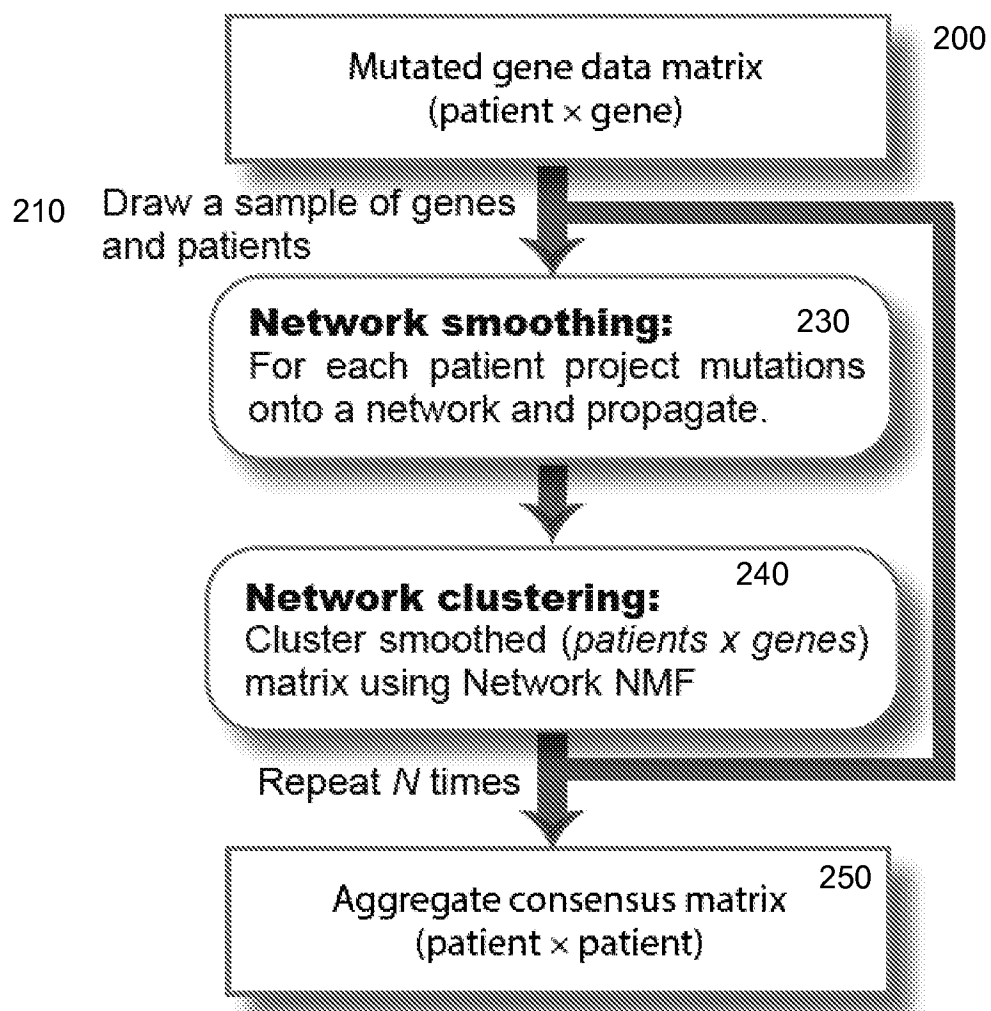


Figure 2

3/44

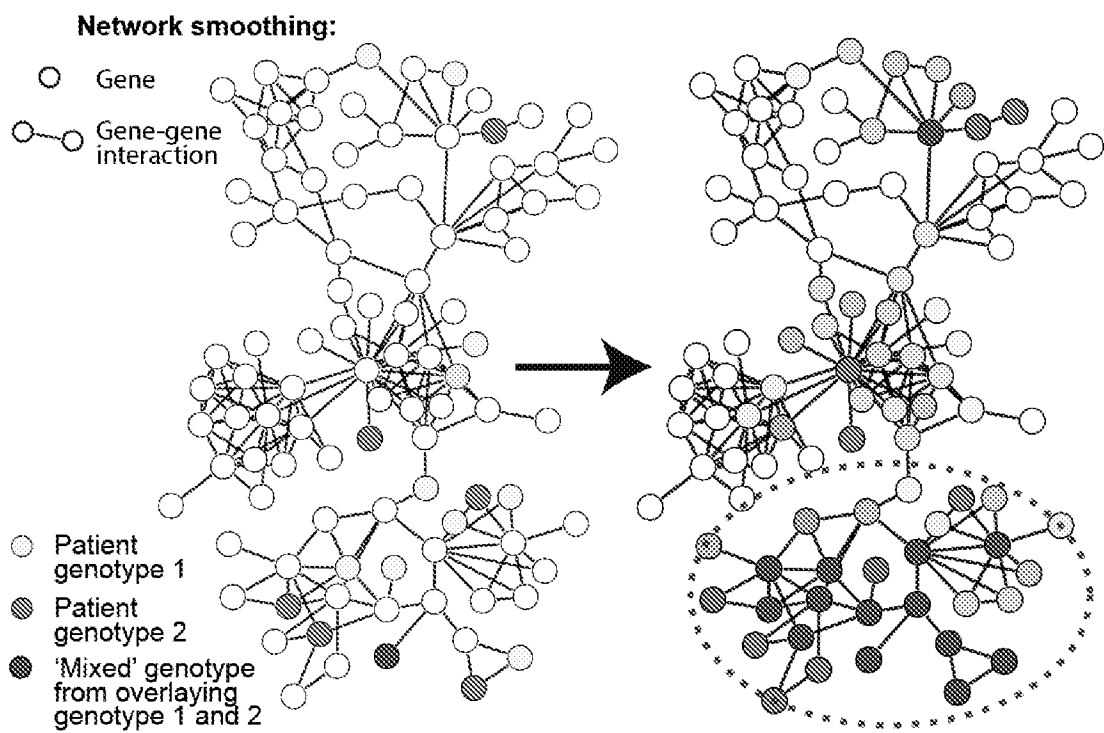


Figure 3

4 / 44

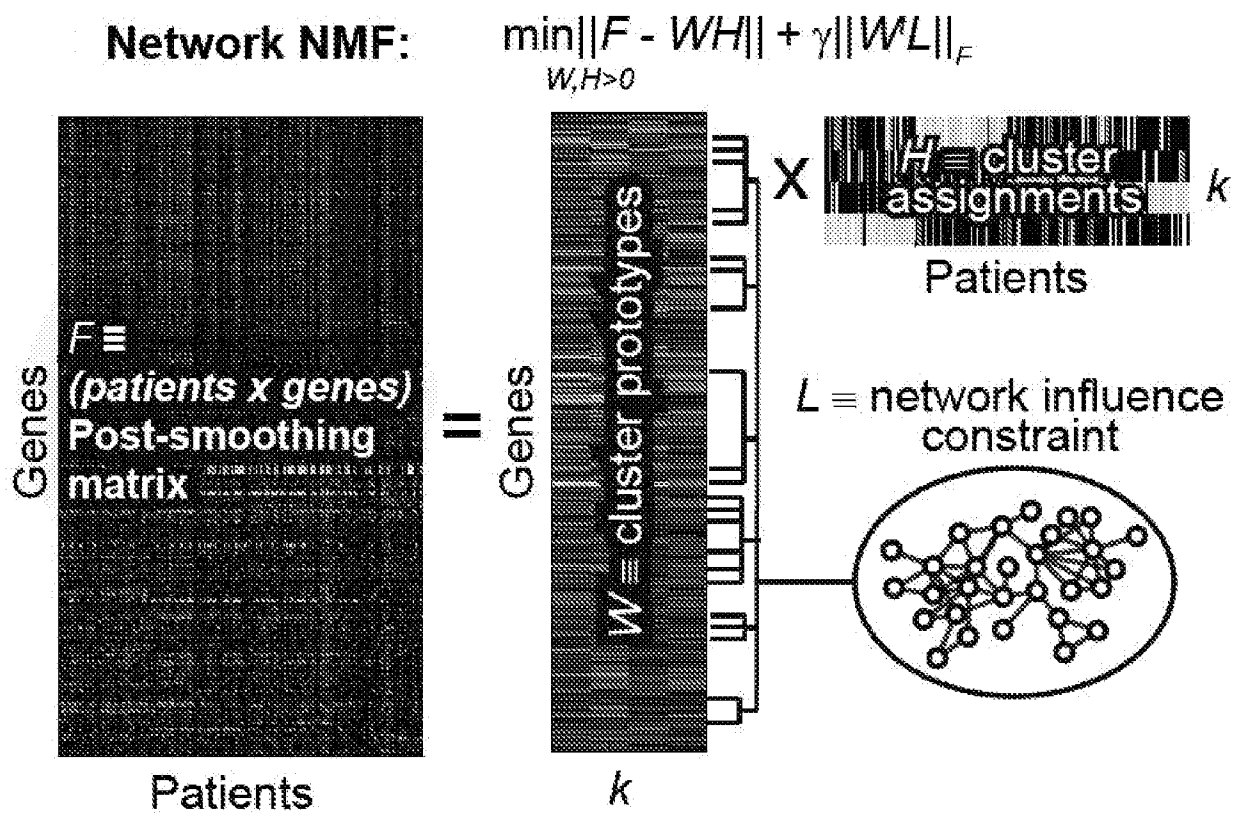


Figure 4

5/44

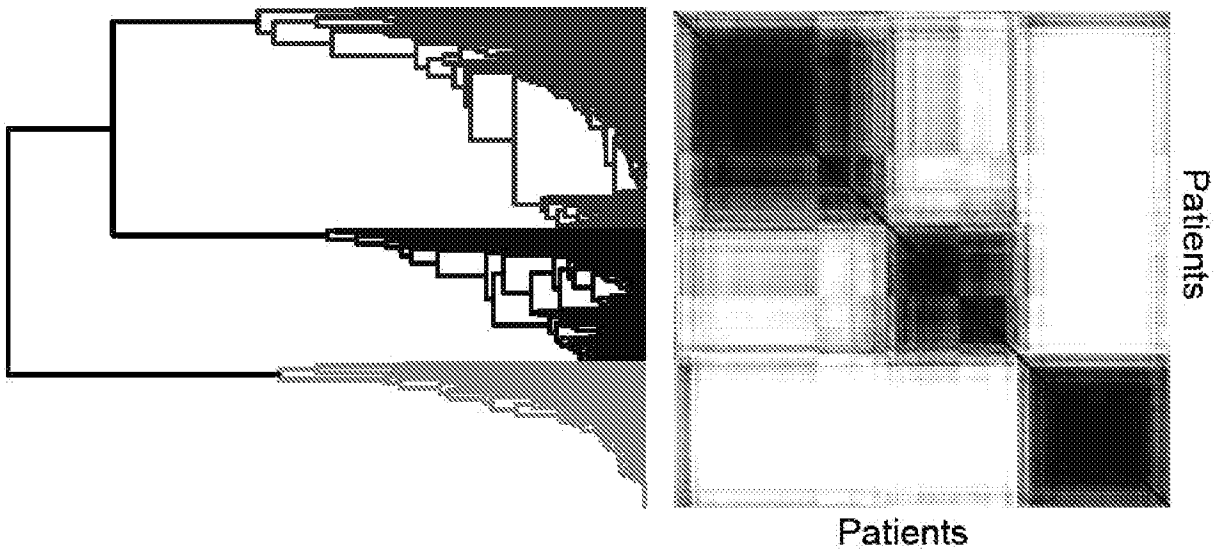
Network-based stratification

Figure 5

6/44

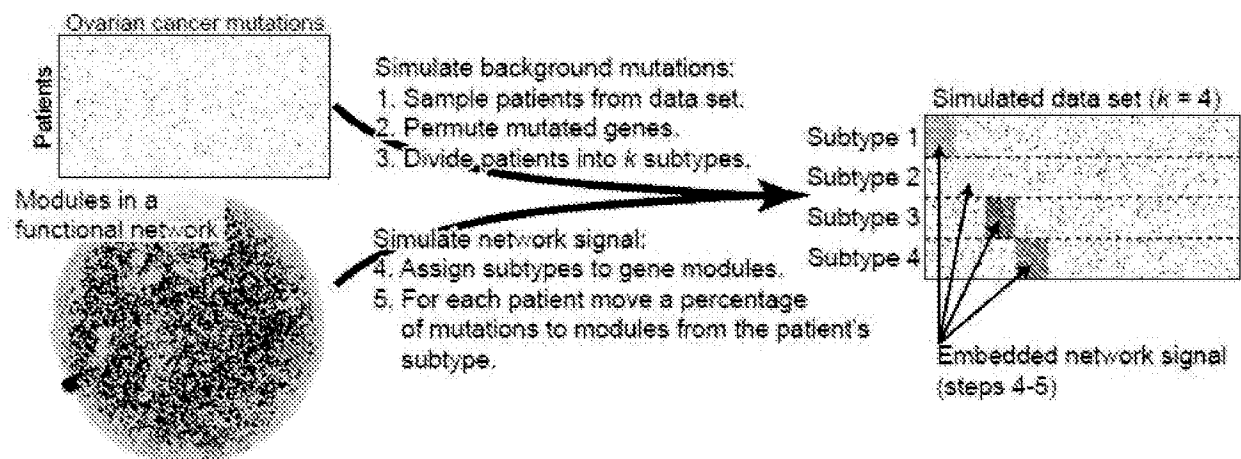


Figure 6

7/44

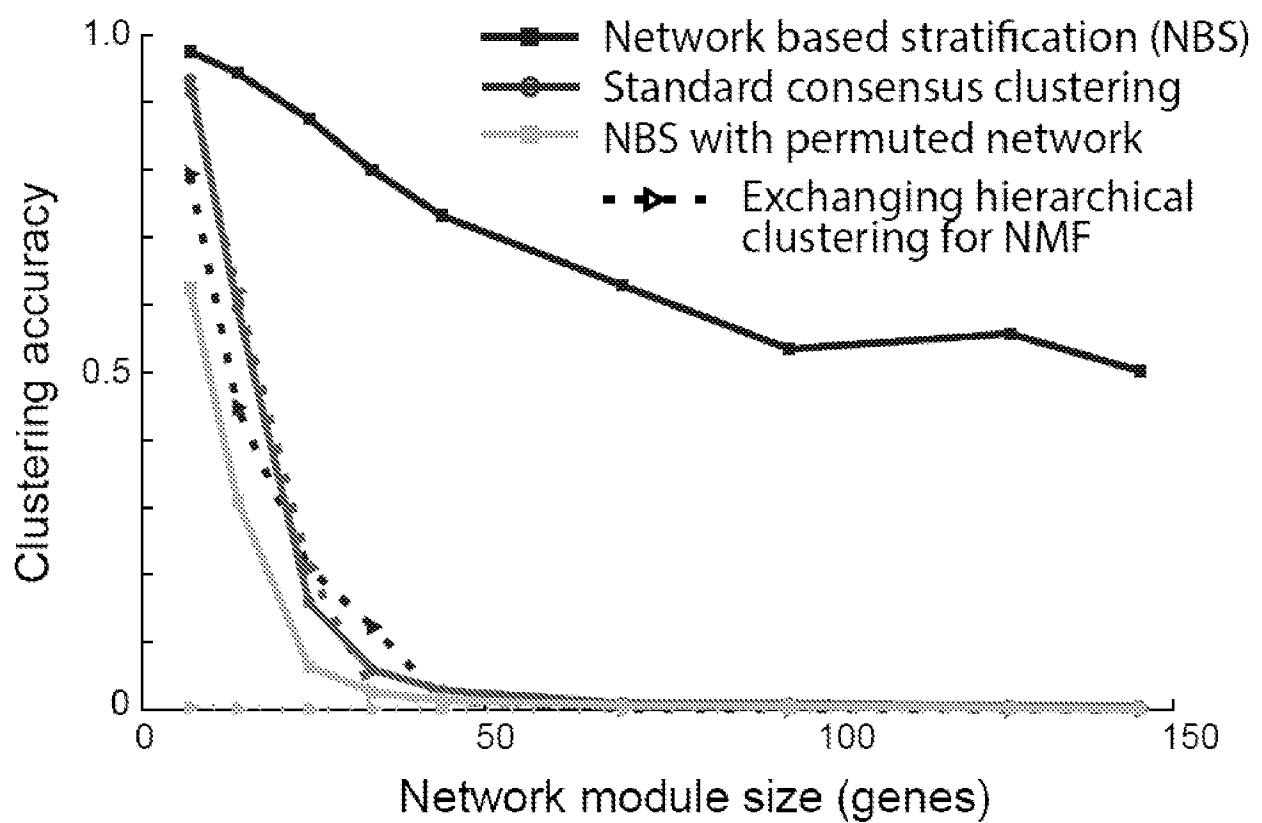


Figure 7

8/44

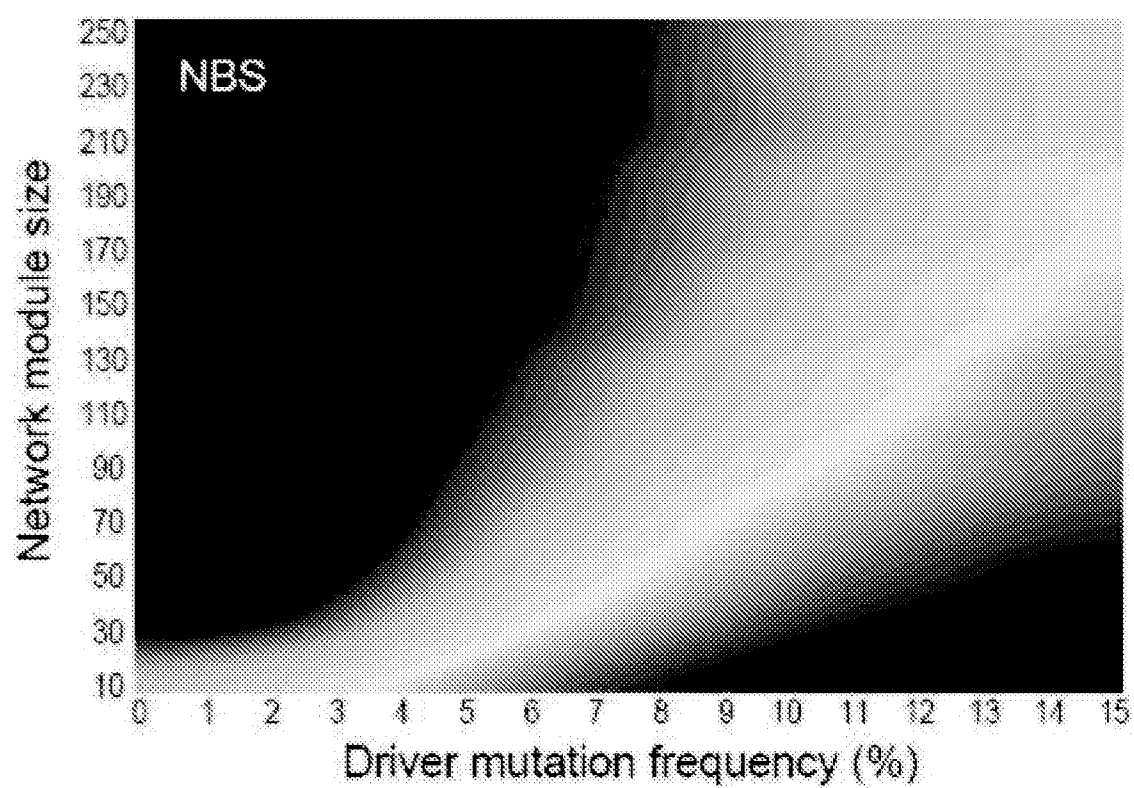


Figure 8

9/44

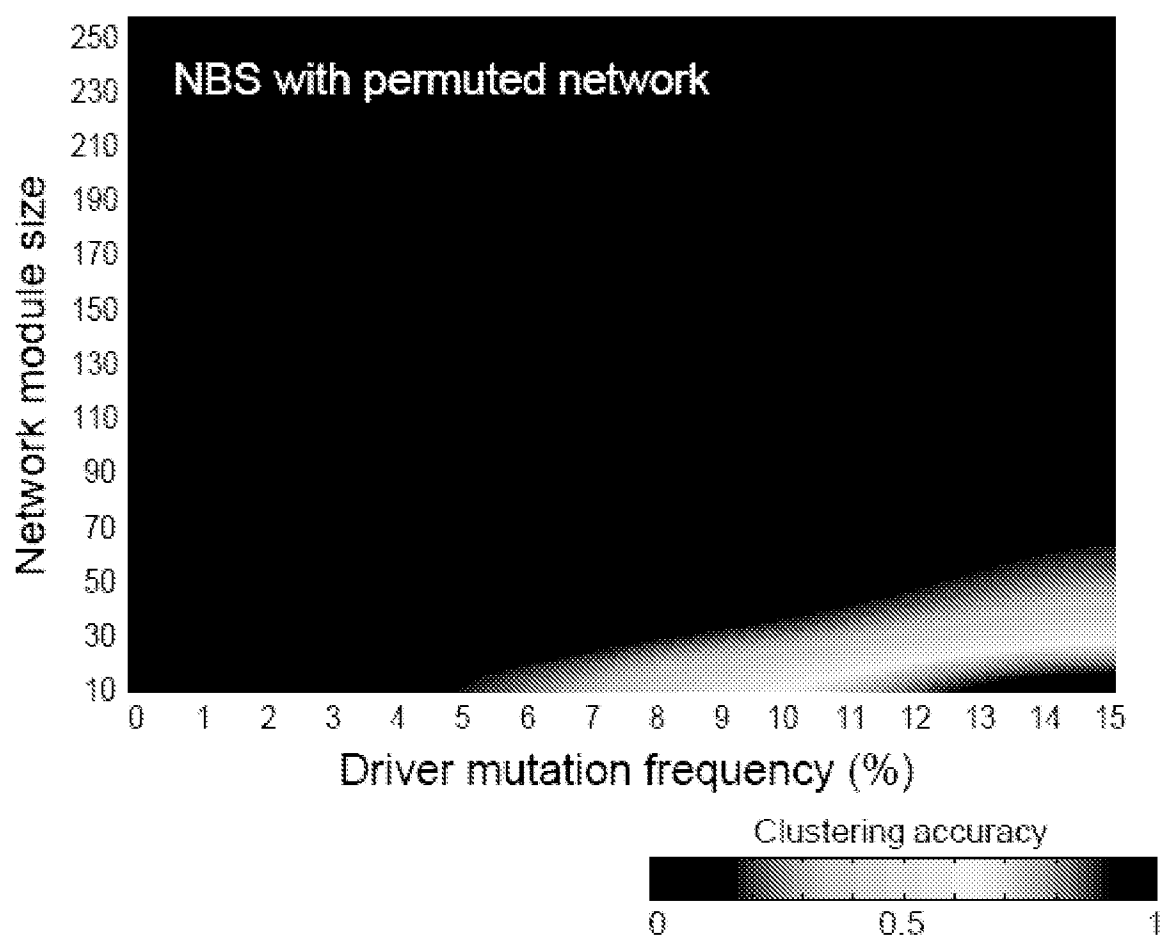


Figure 9

10/44

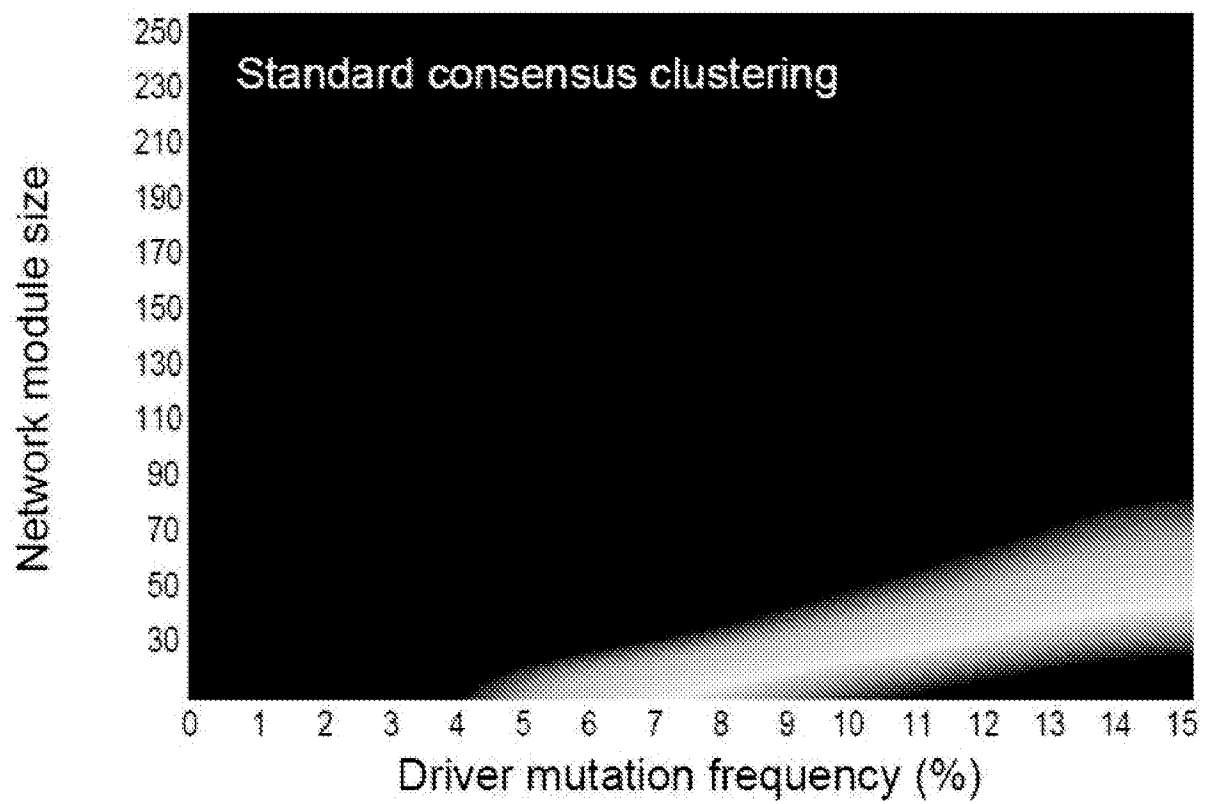


Figure 10

11/44

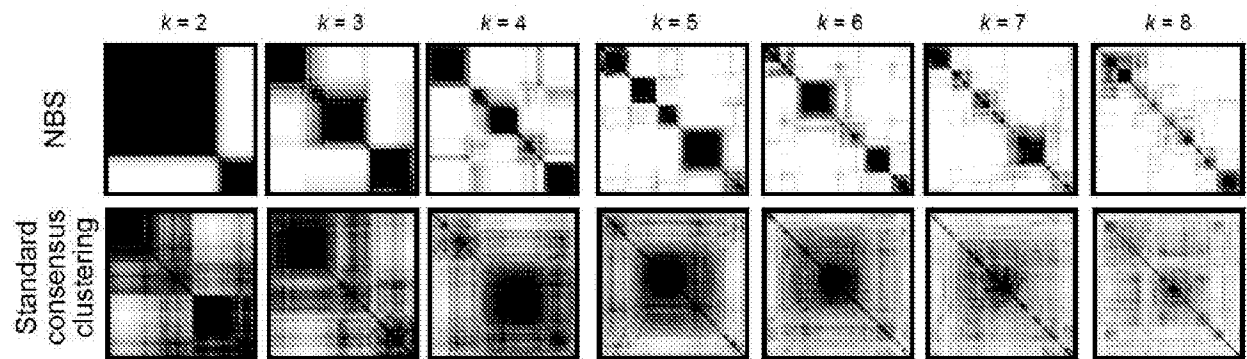


Figure 11

12/44

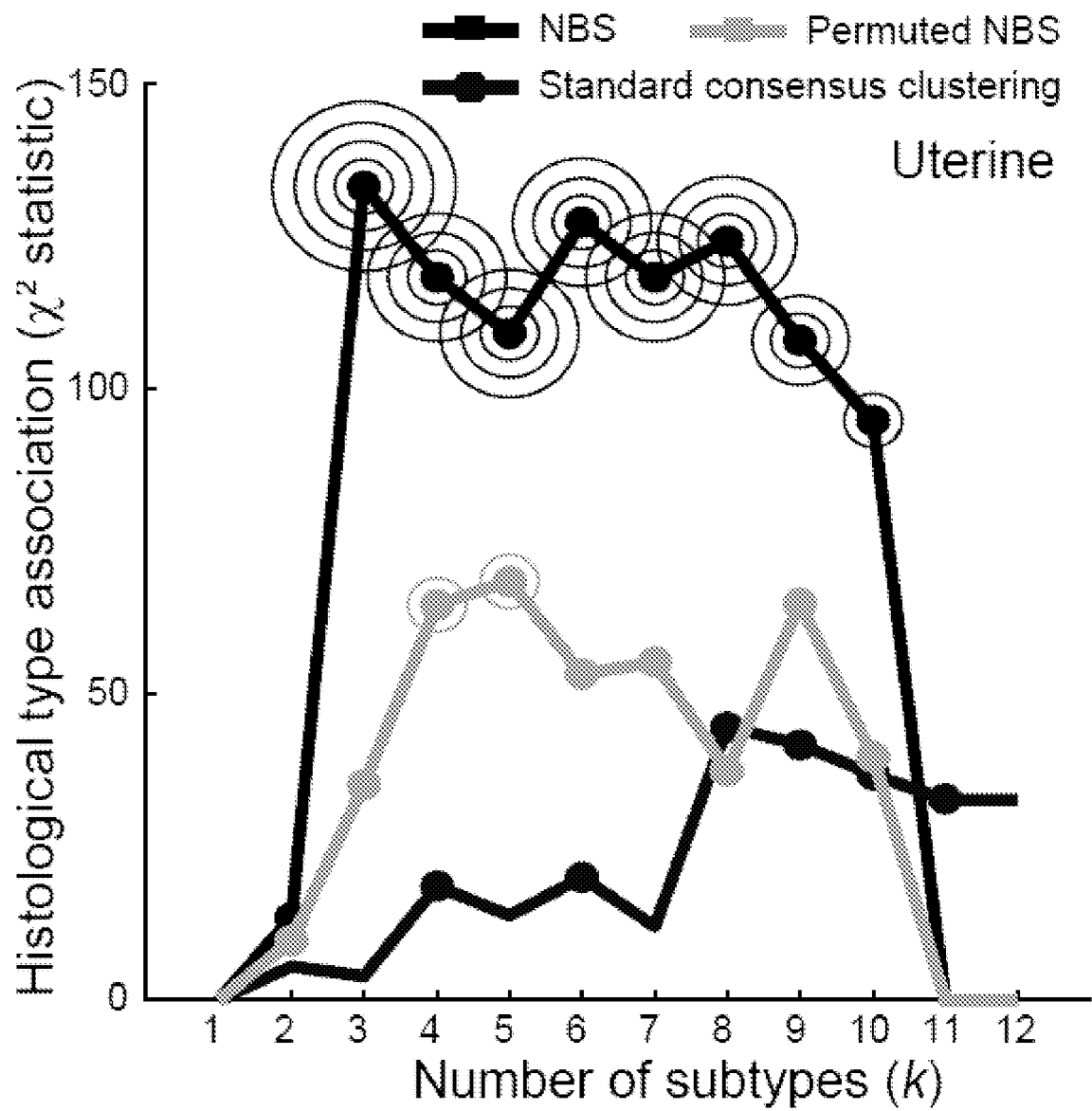


Figure 12

13/44

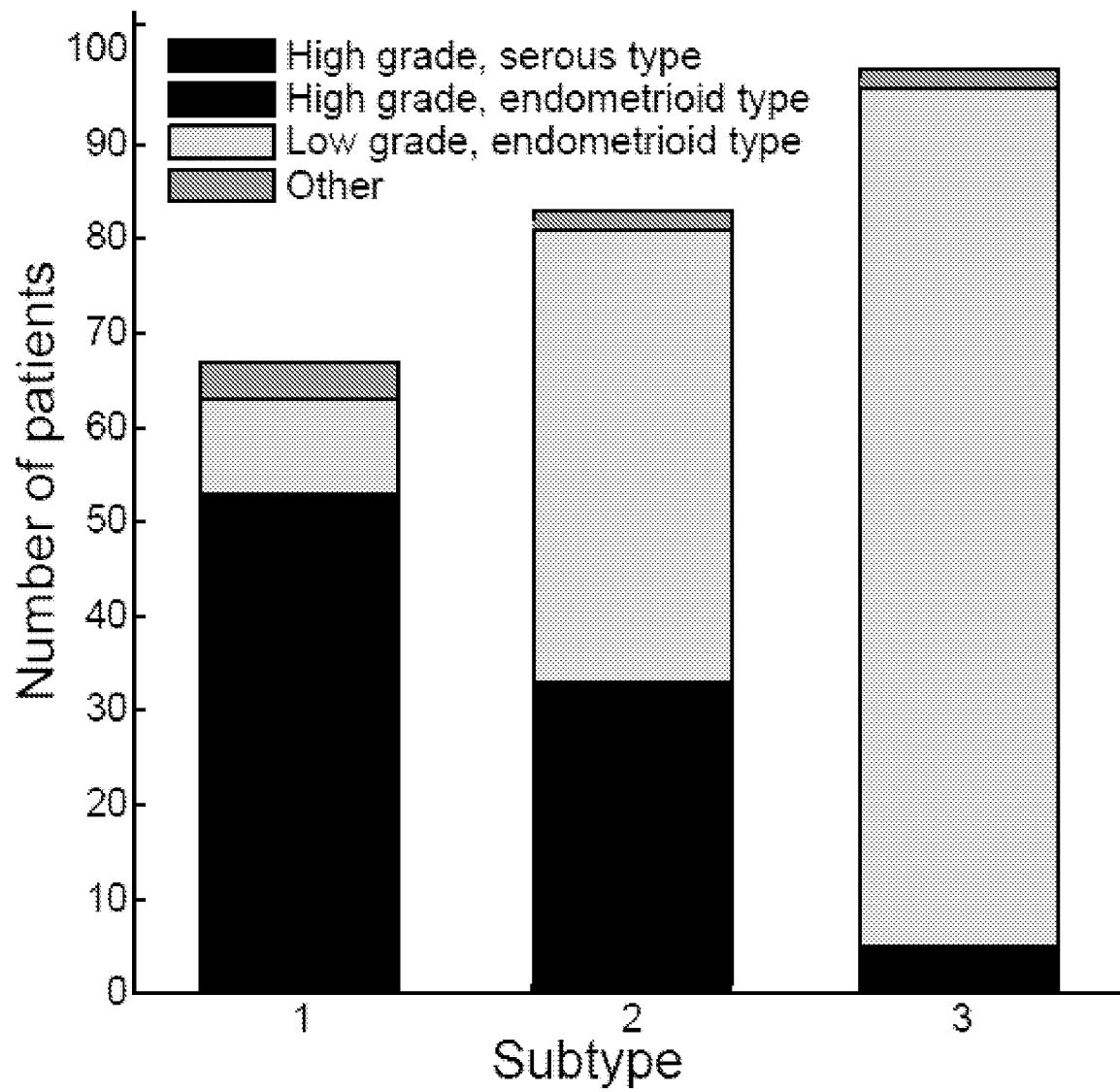


Figure 13

14/44

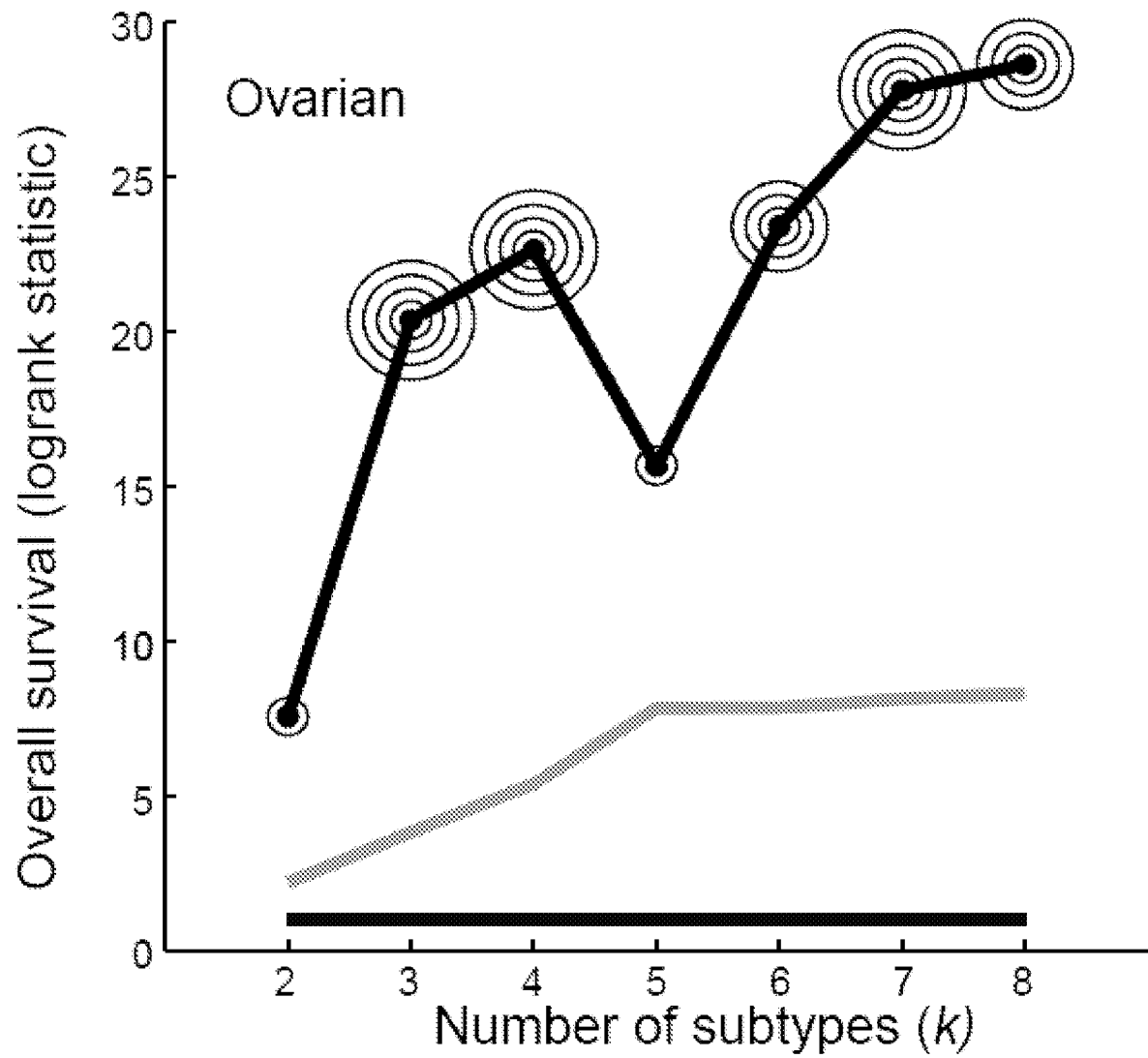


Figure 14

15/44

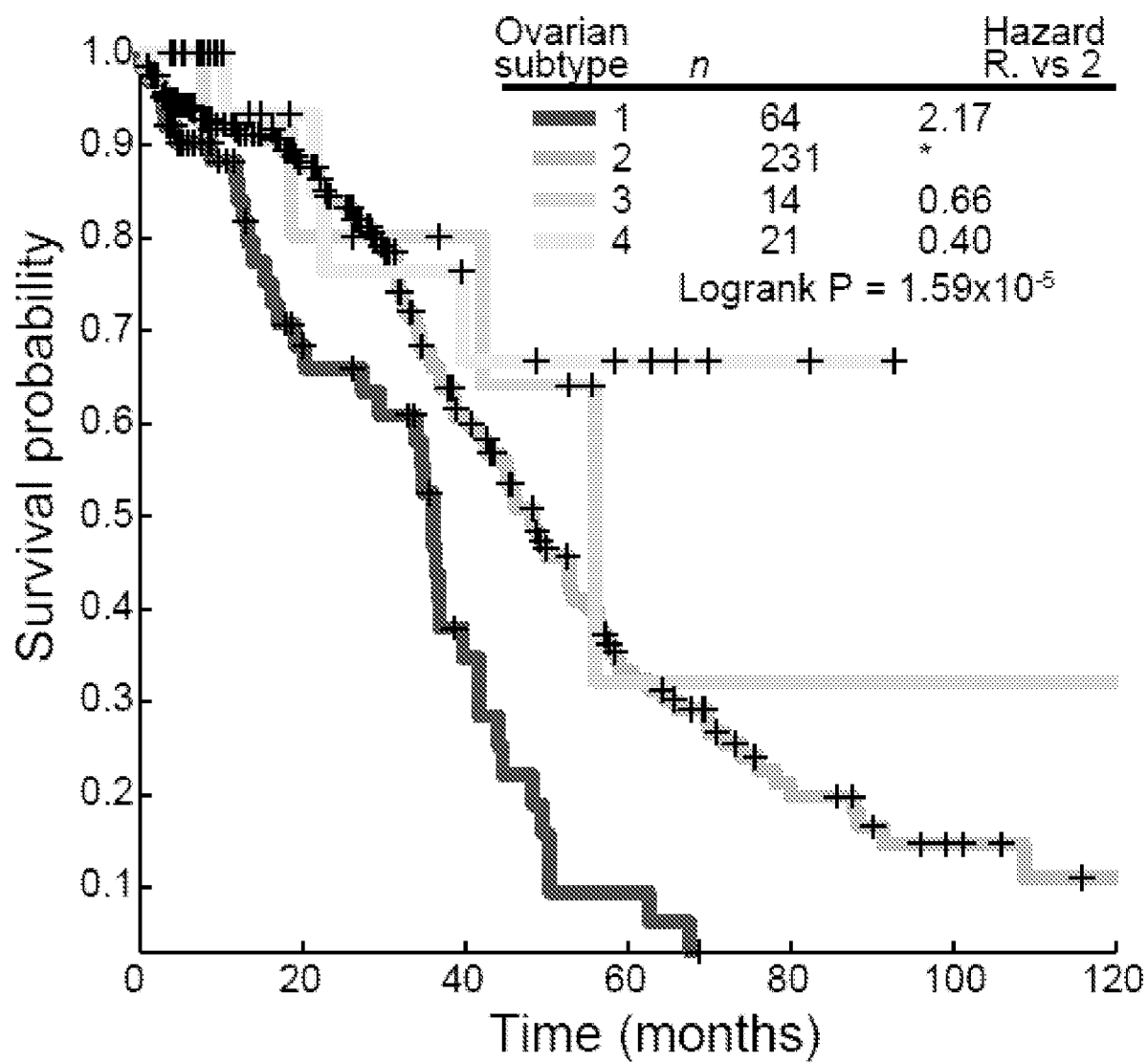


Figure 15

16/44

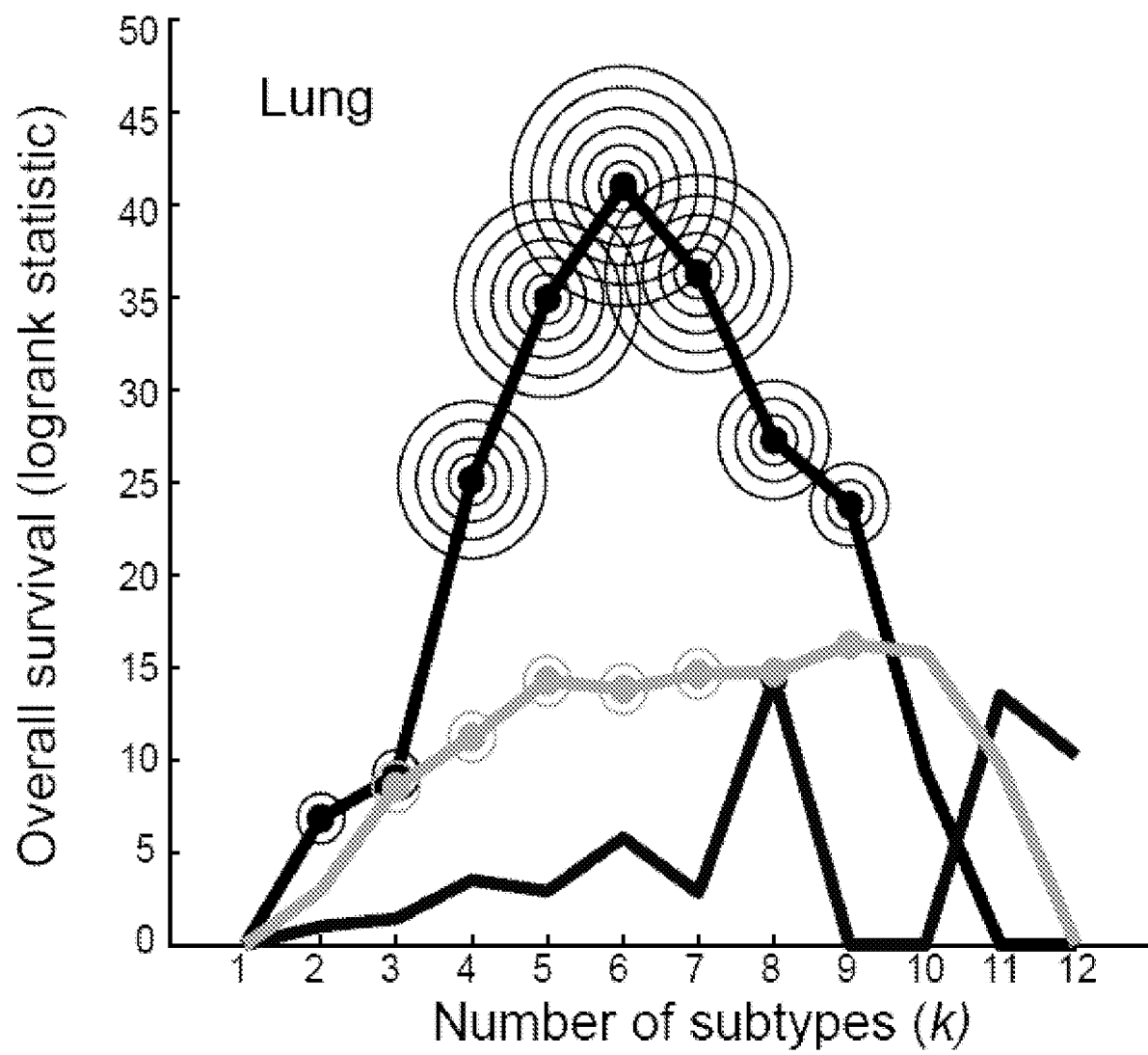


Figure 16

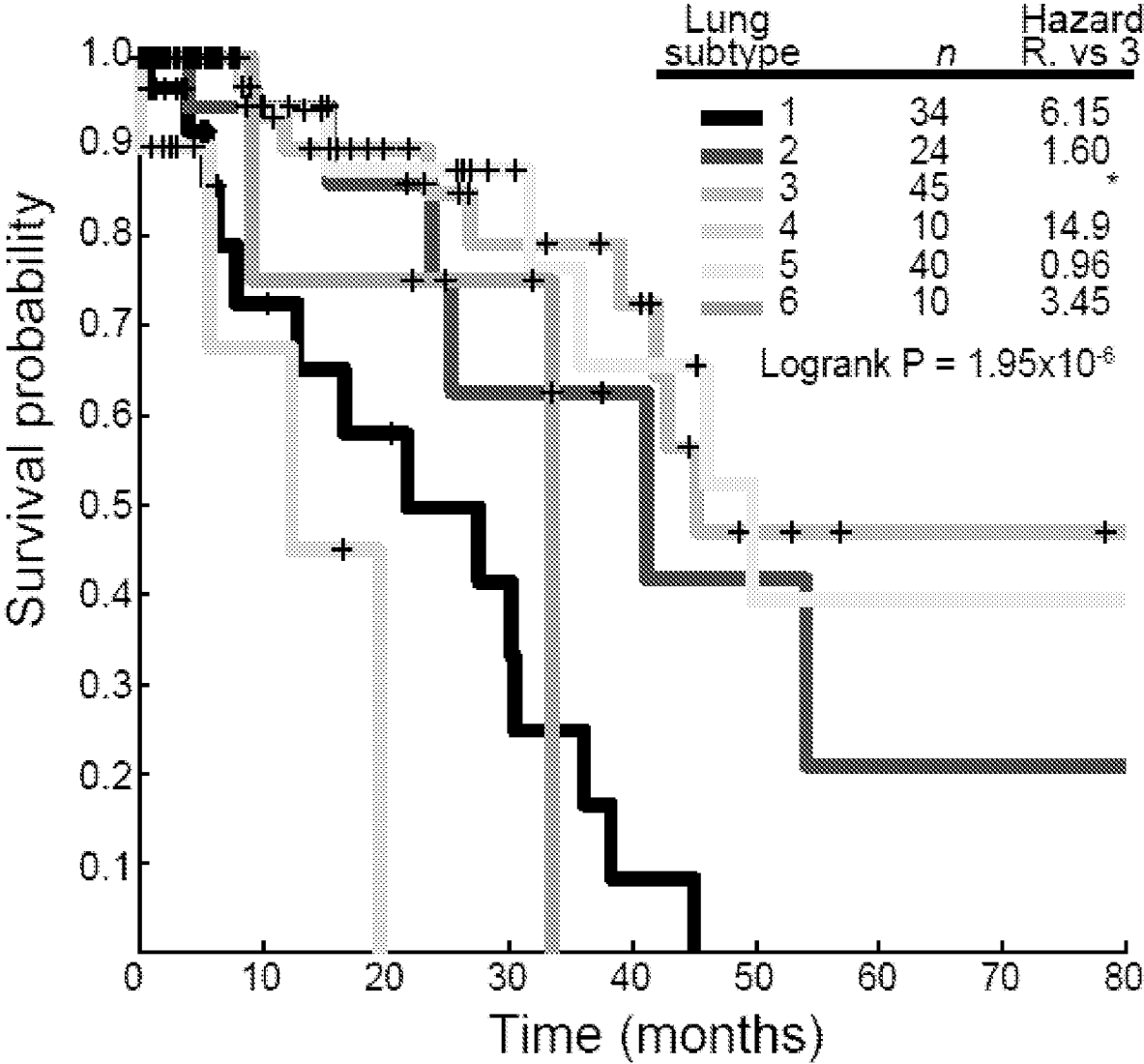


Figure 17

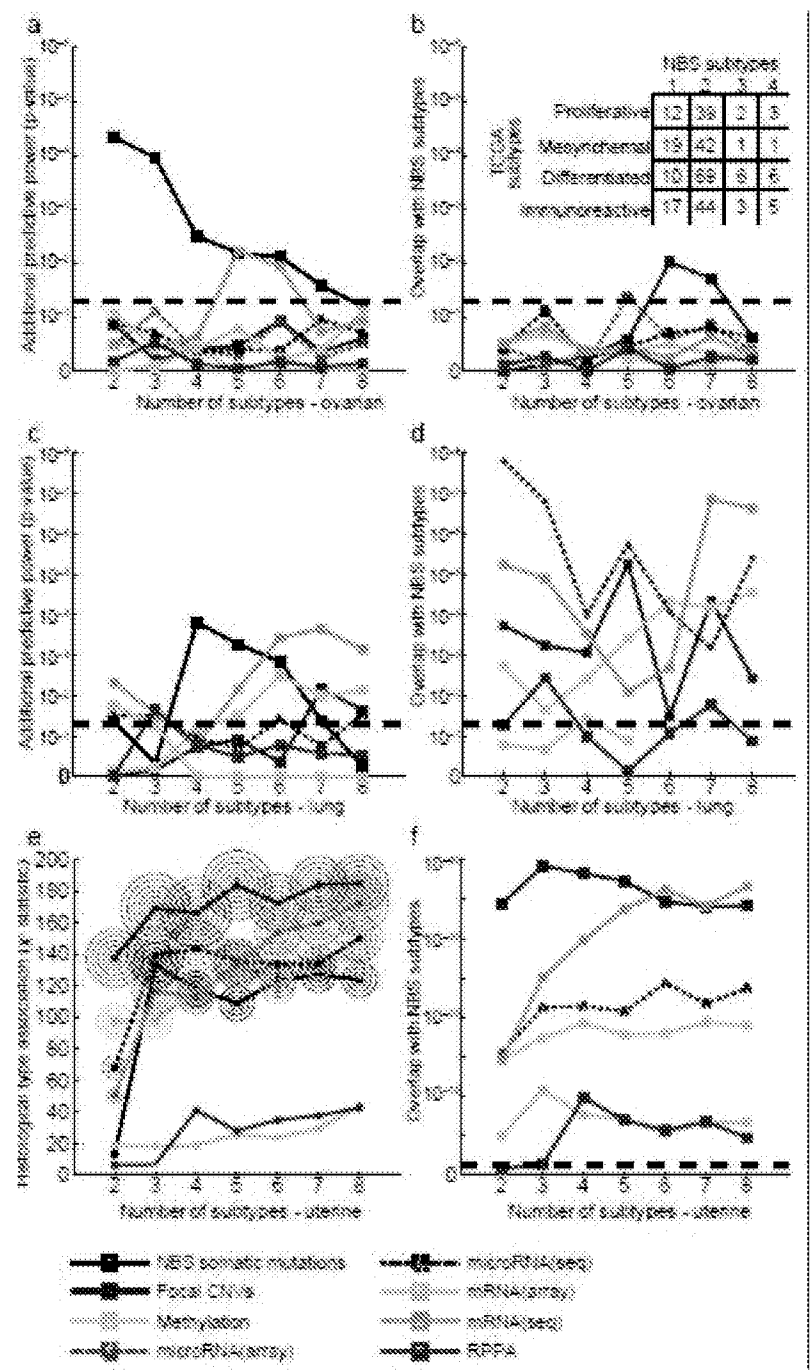


Figure 18

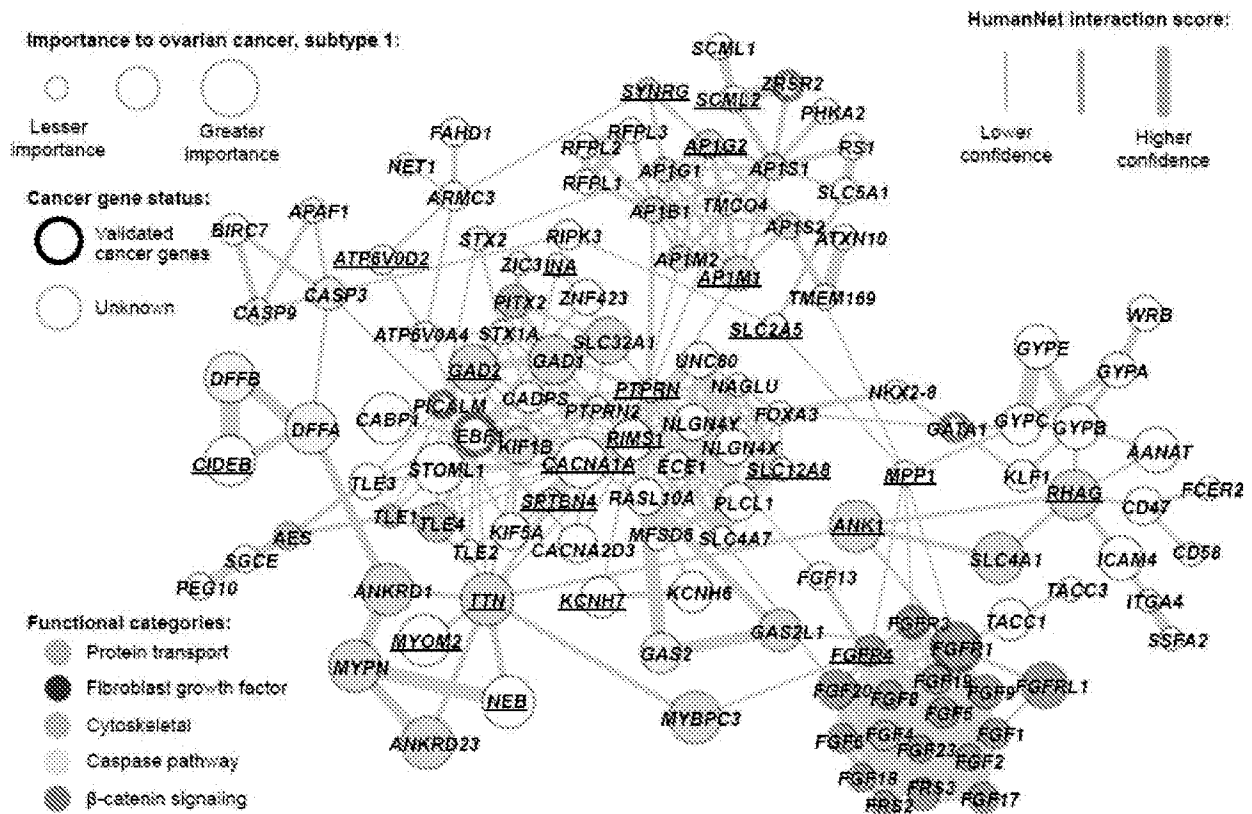


Figure 19

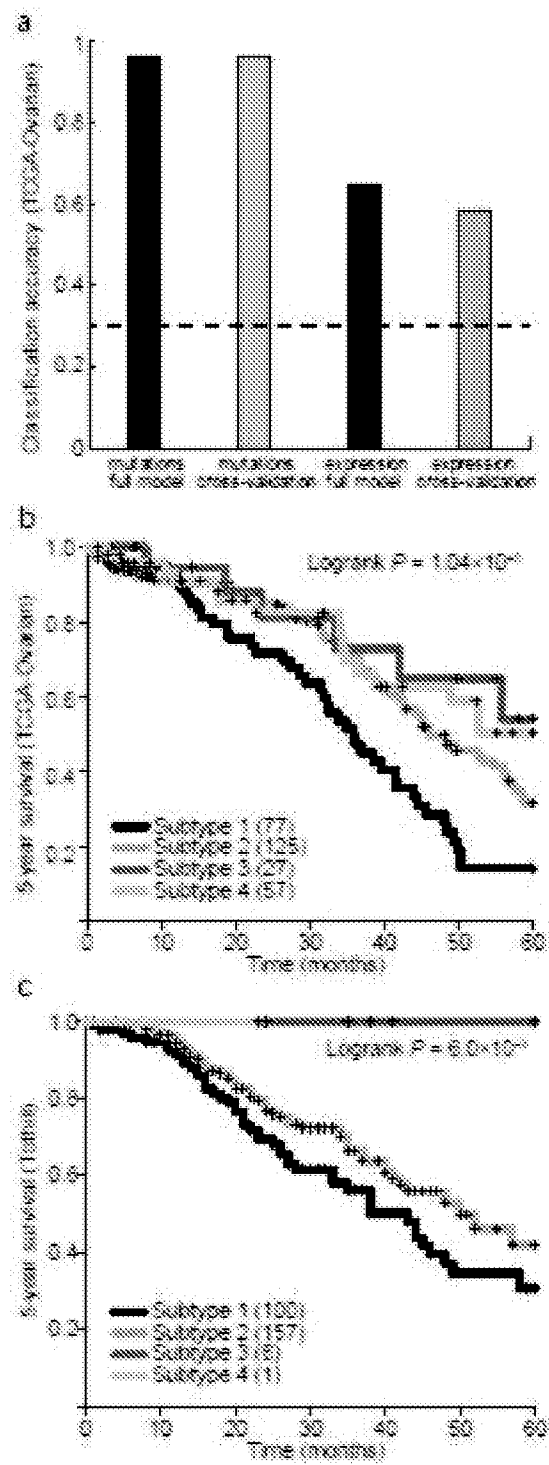


Figure 20

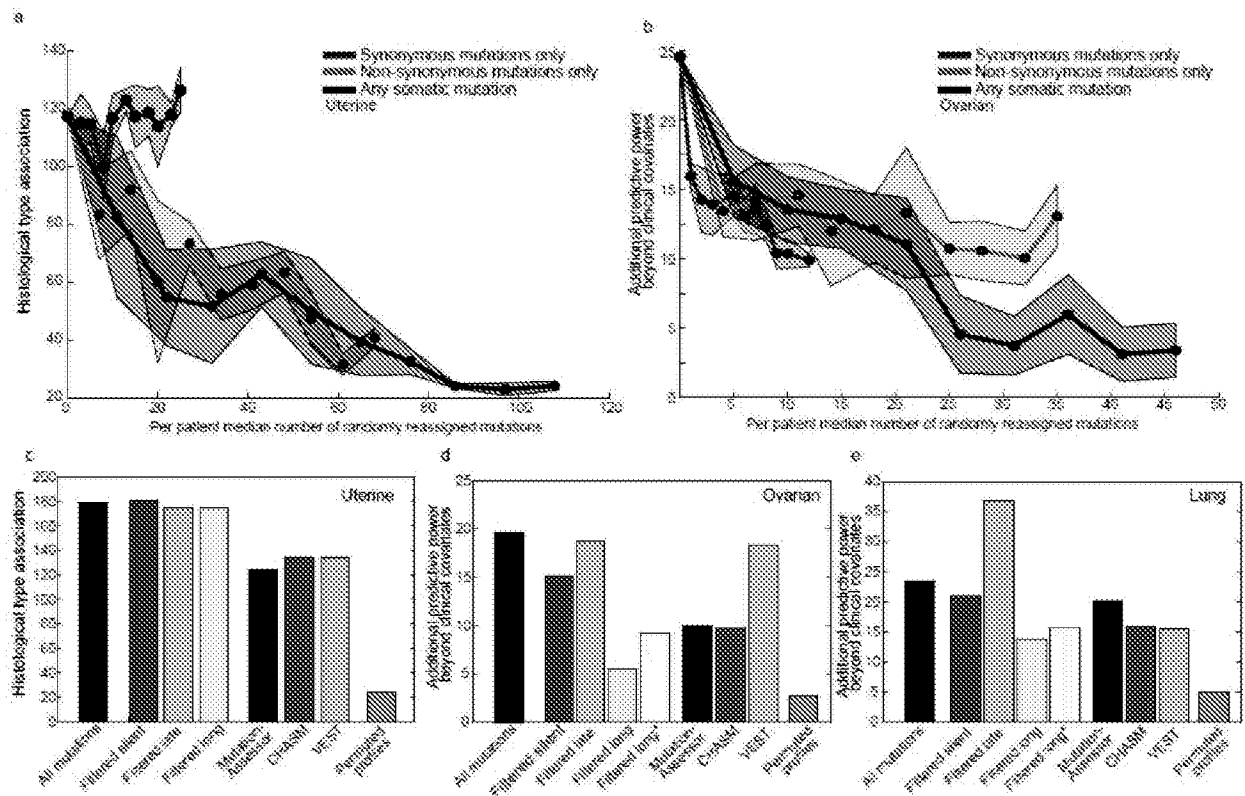


Figure 21

22/44

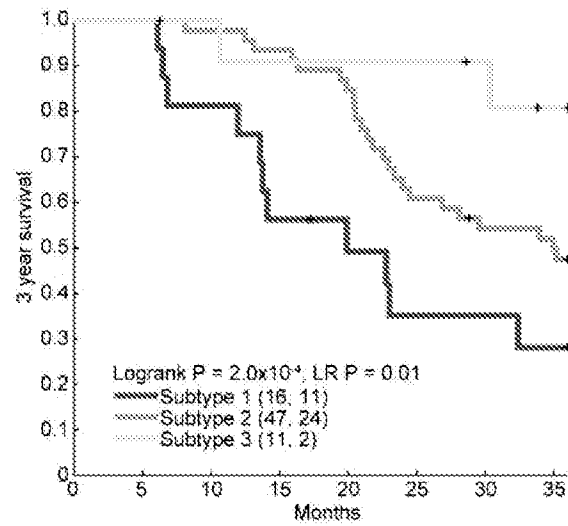


Figure 22

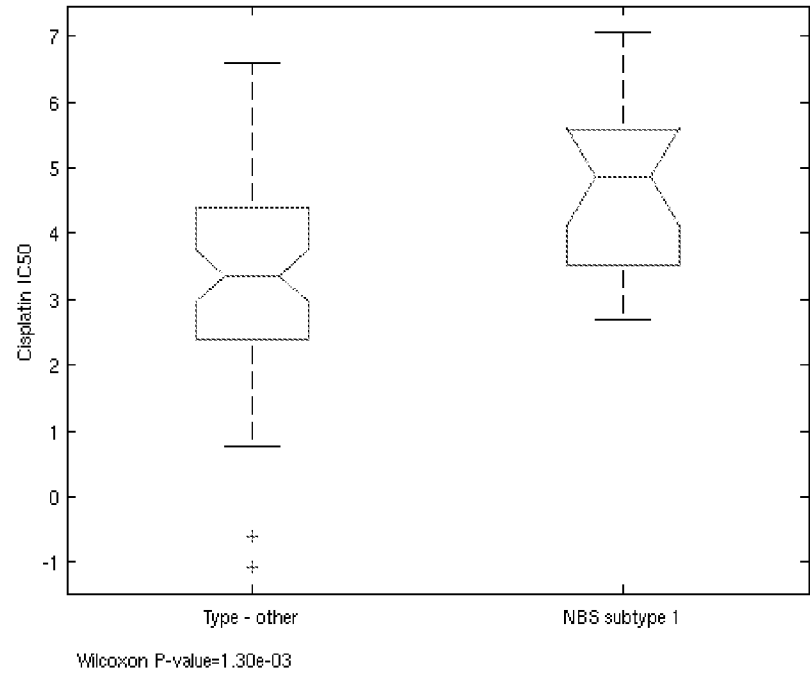


Figure 23

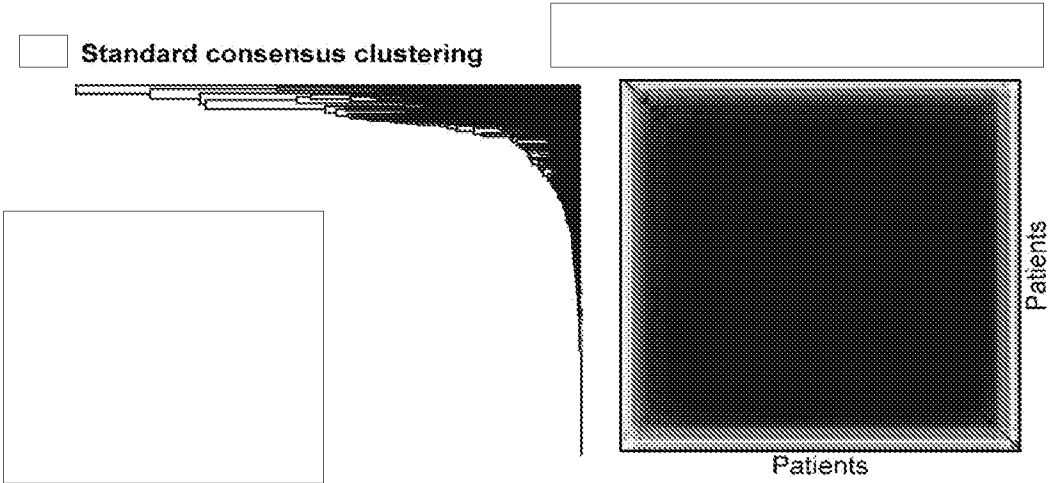


FIGURE 24

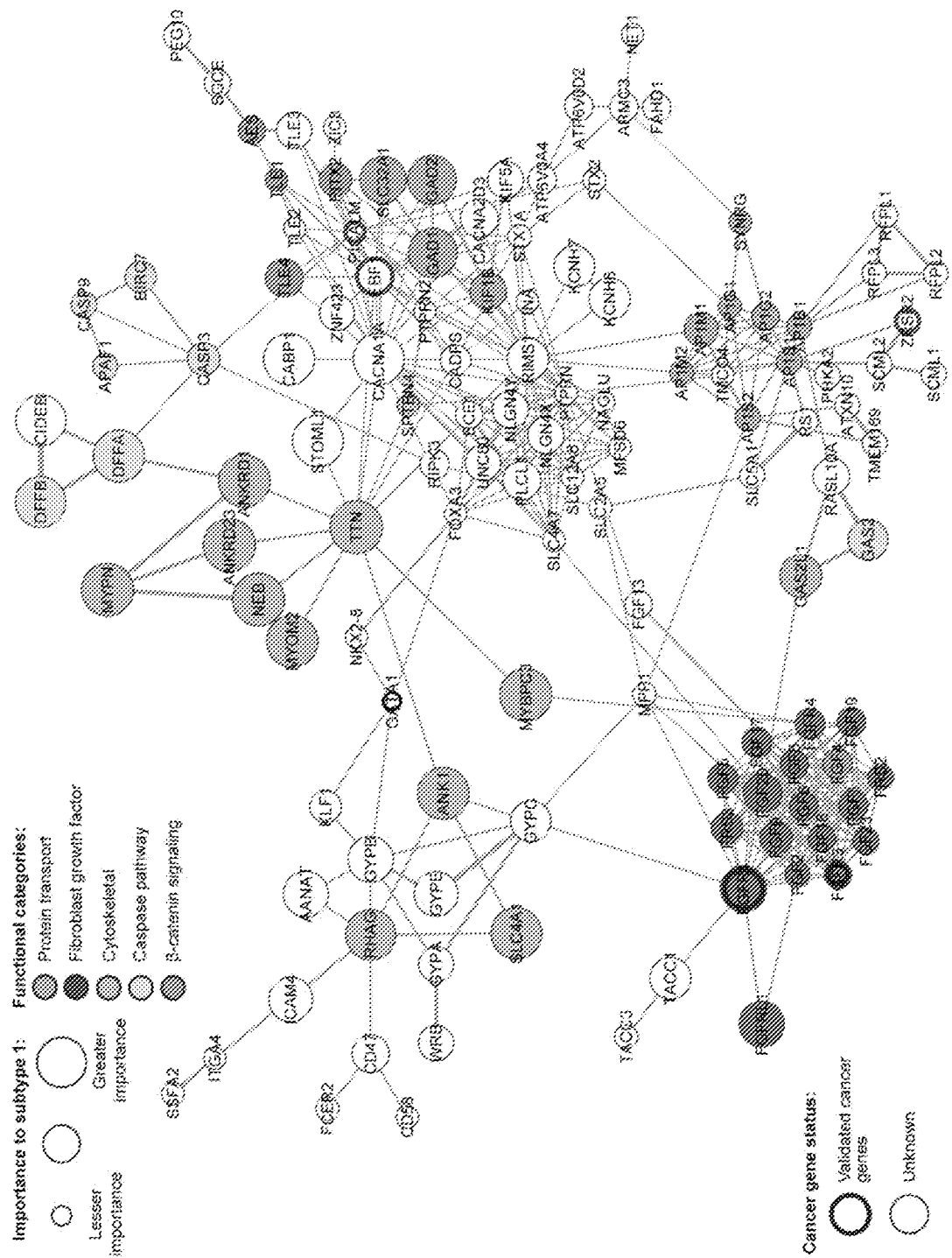


FIGURE 25

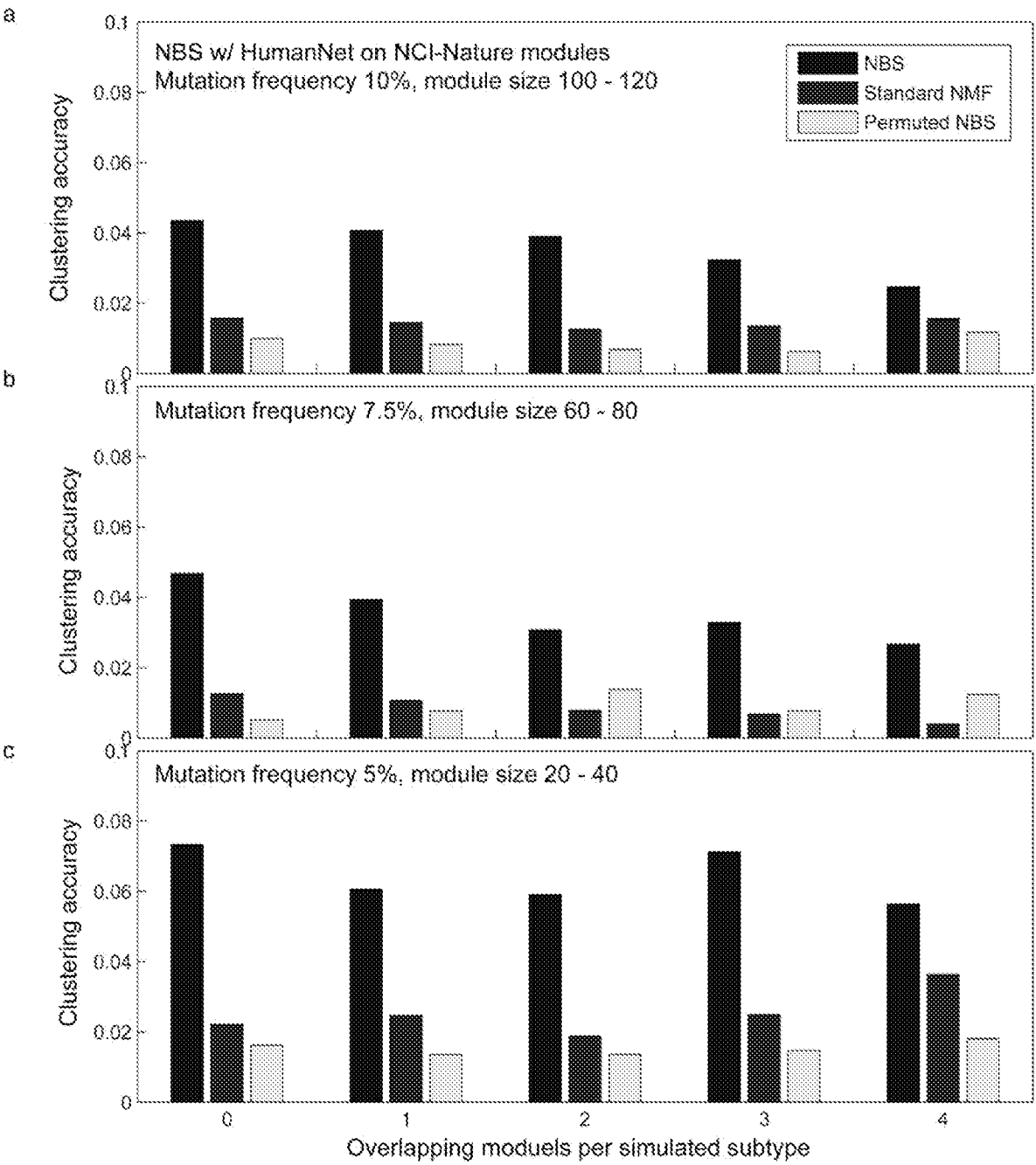


FIGURE 26

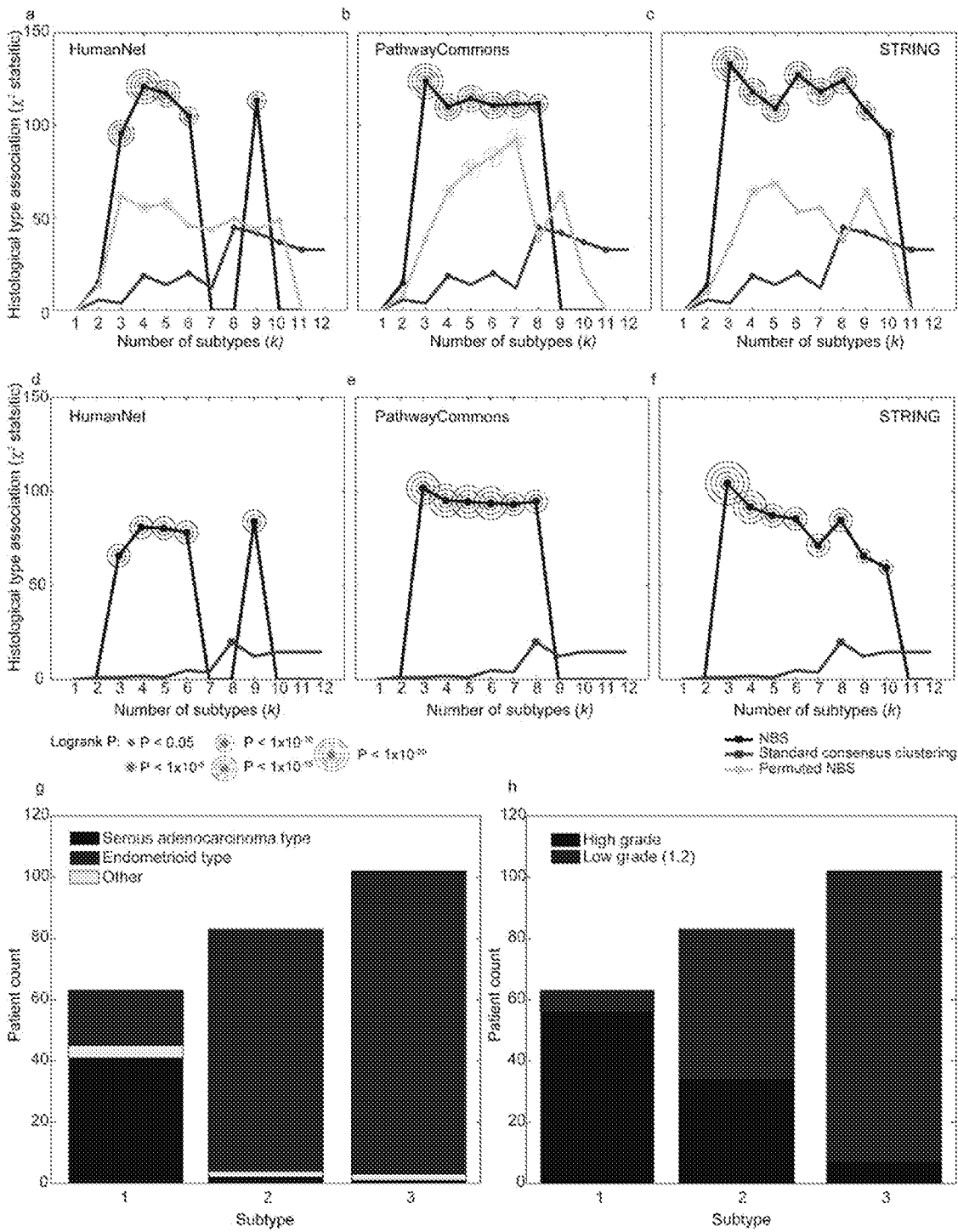


FIGURE 27

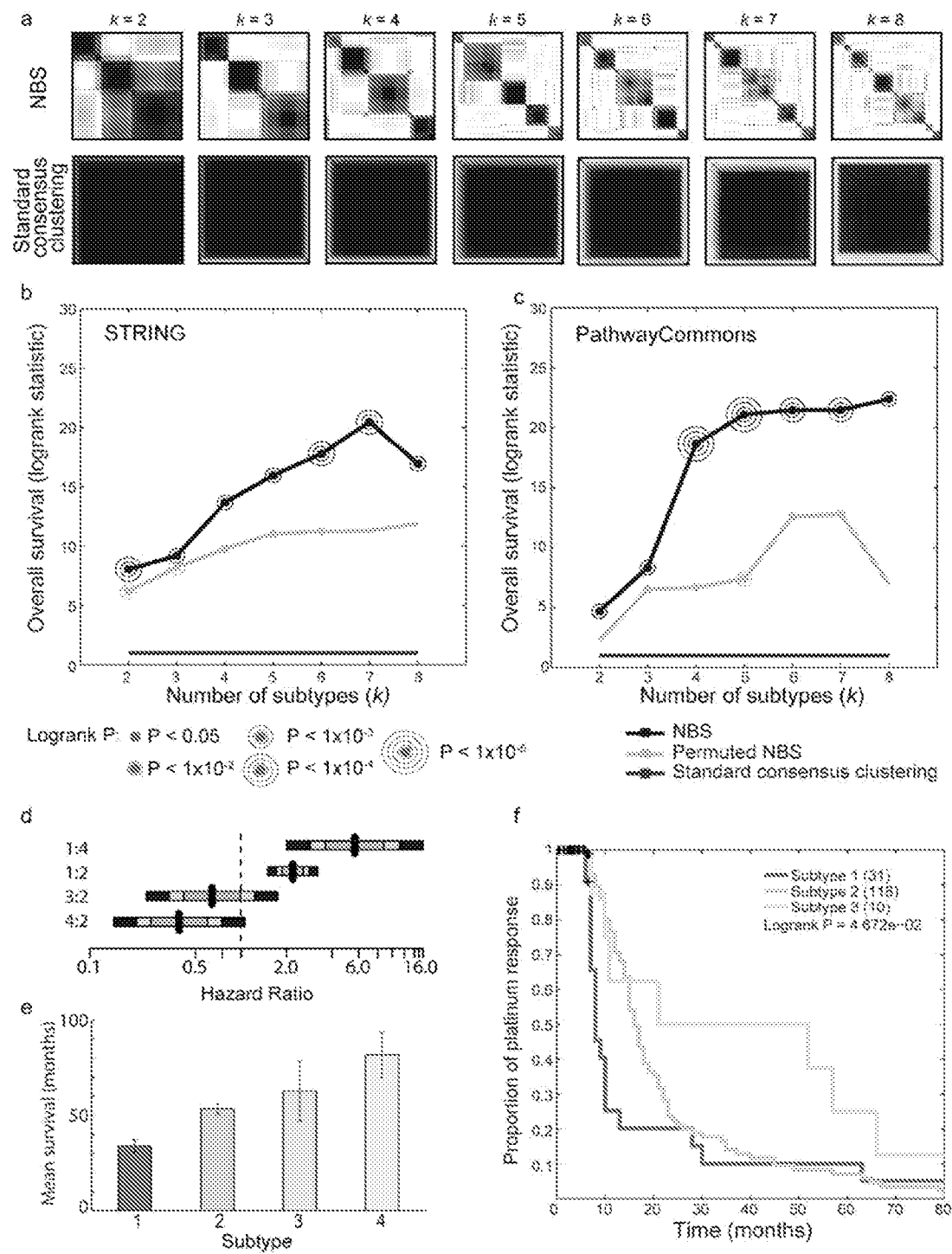


FIGURE 28

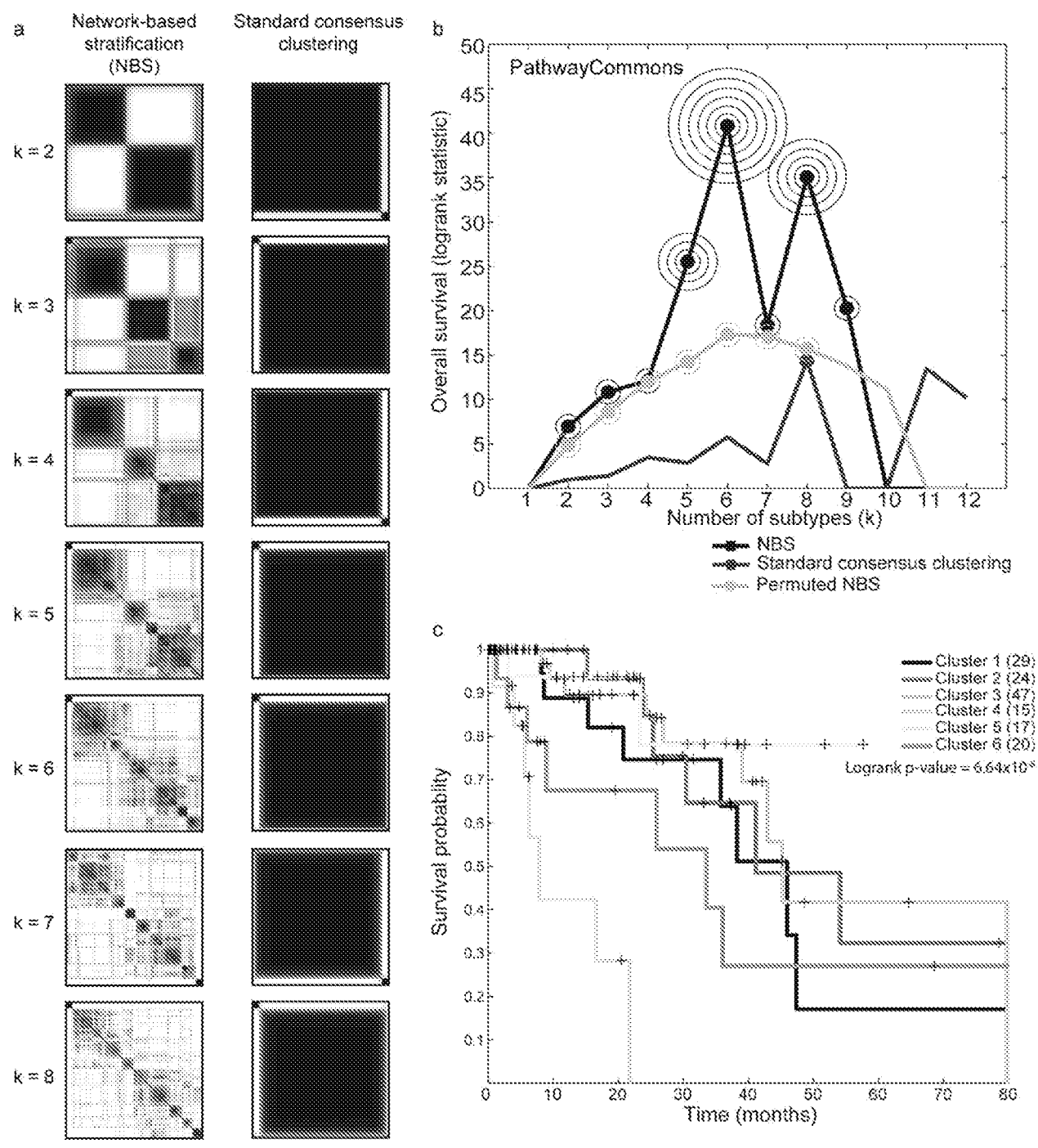


FIGURE 29

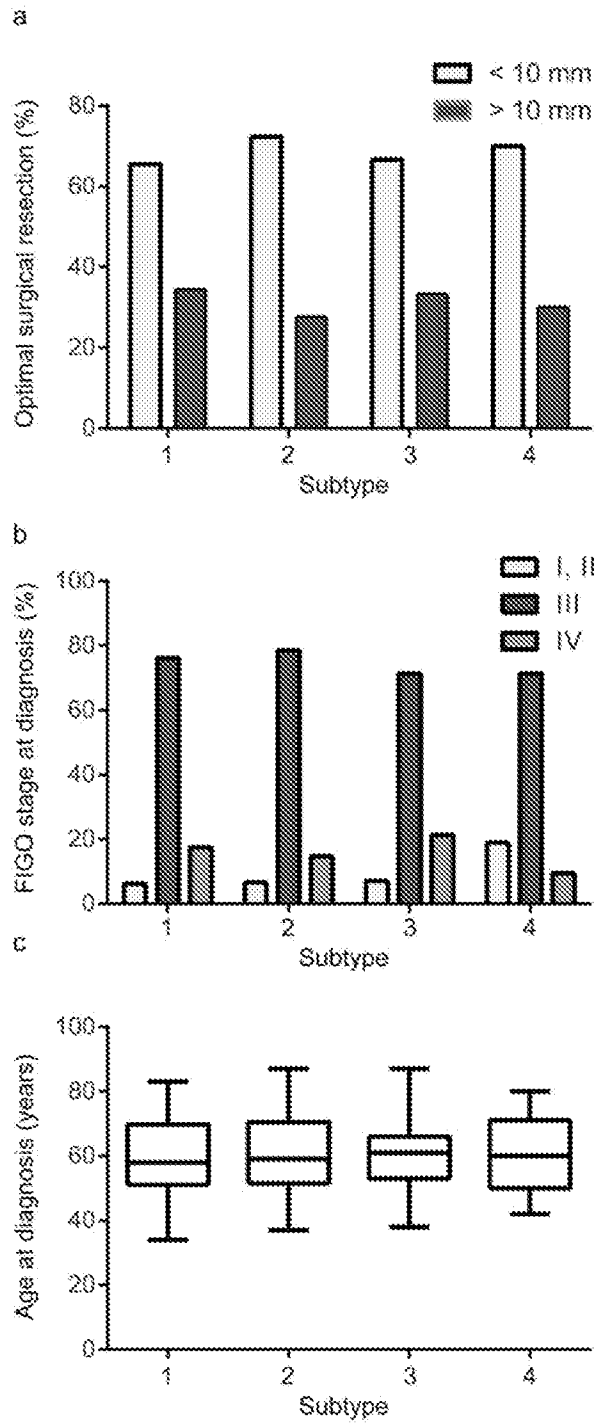


FIGURE 30

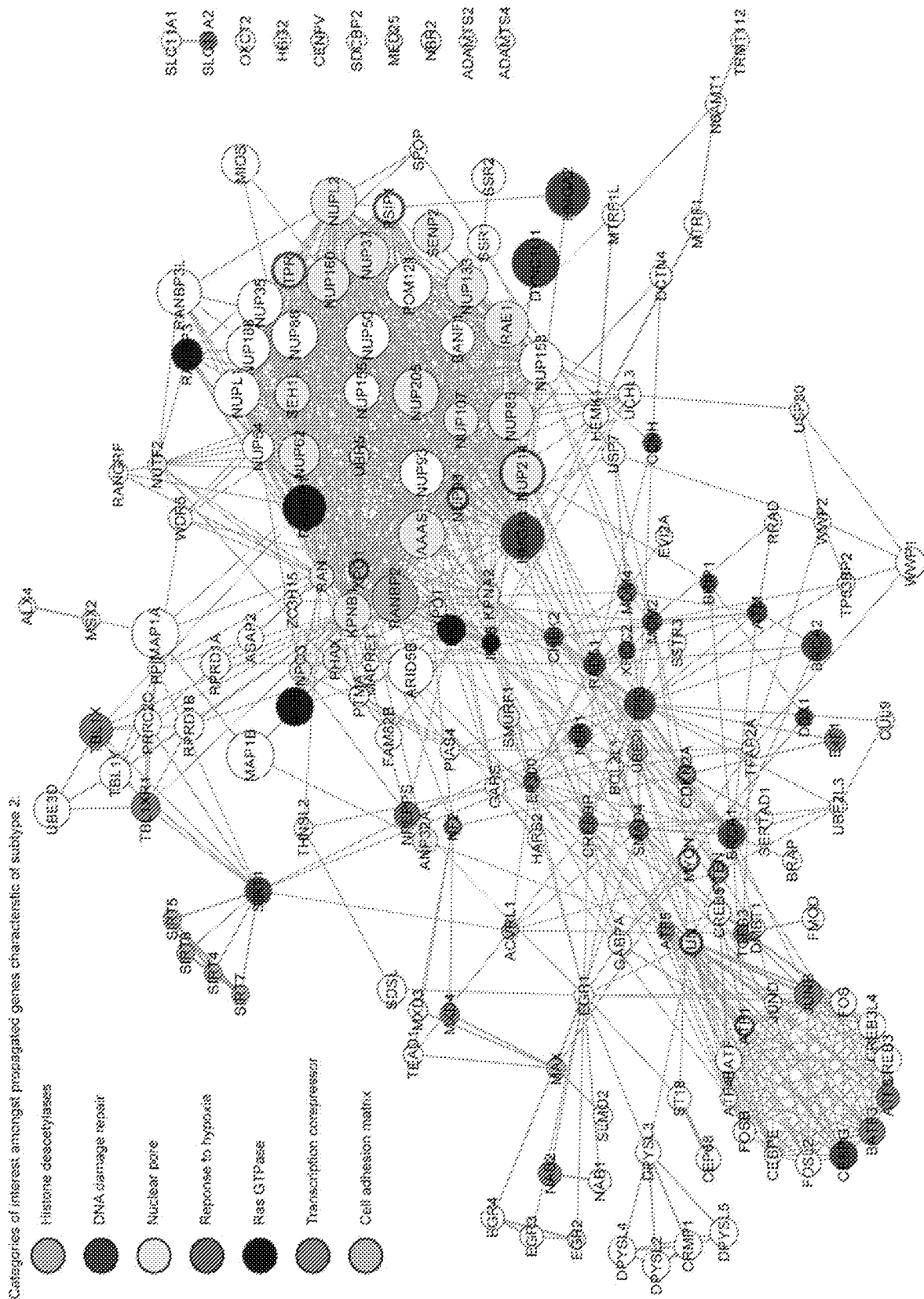


FIGURE 31

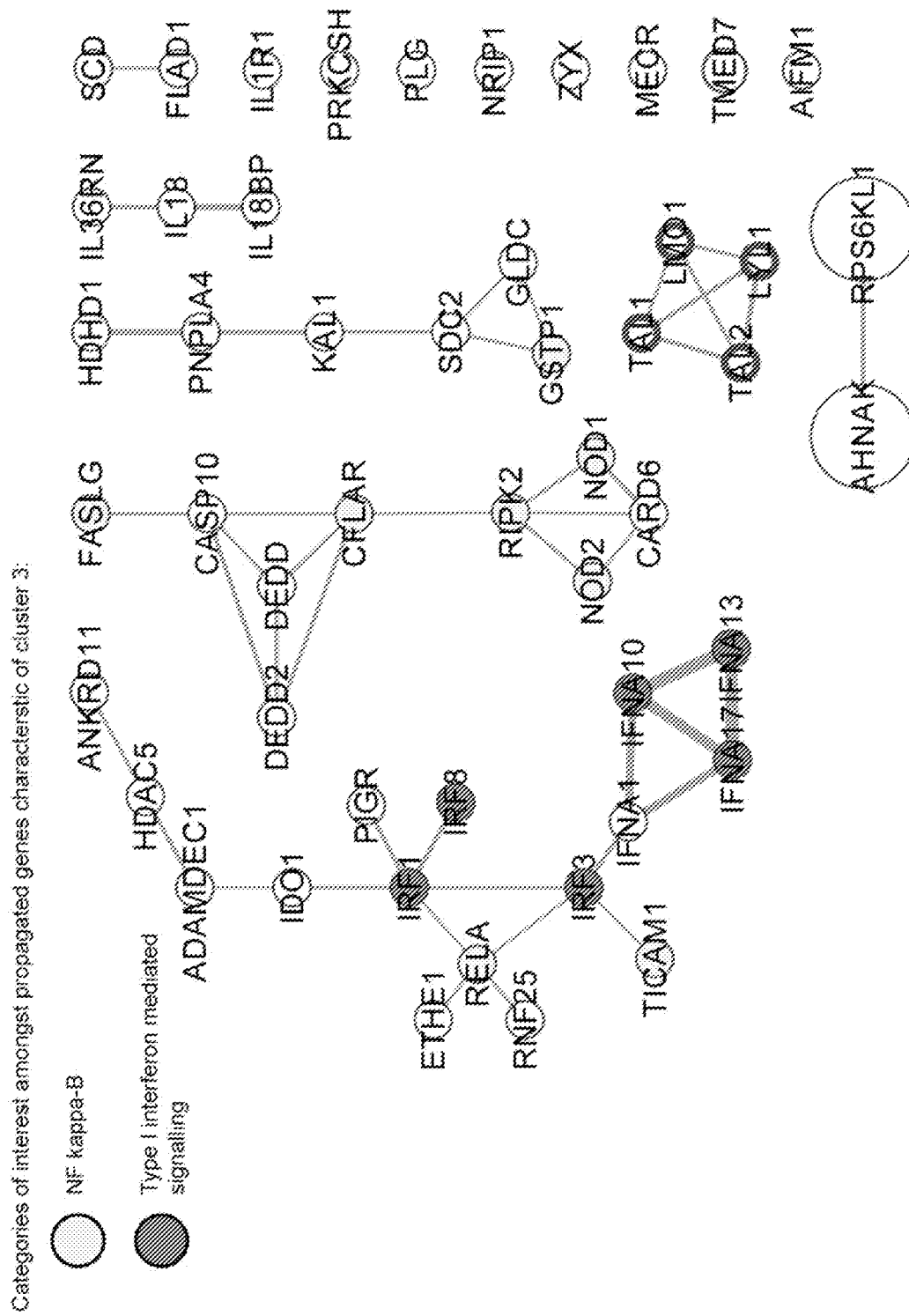


FIGURE 32

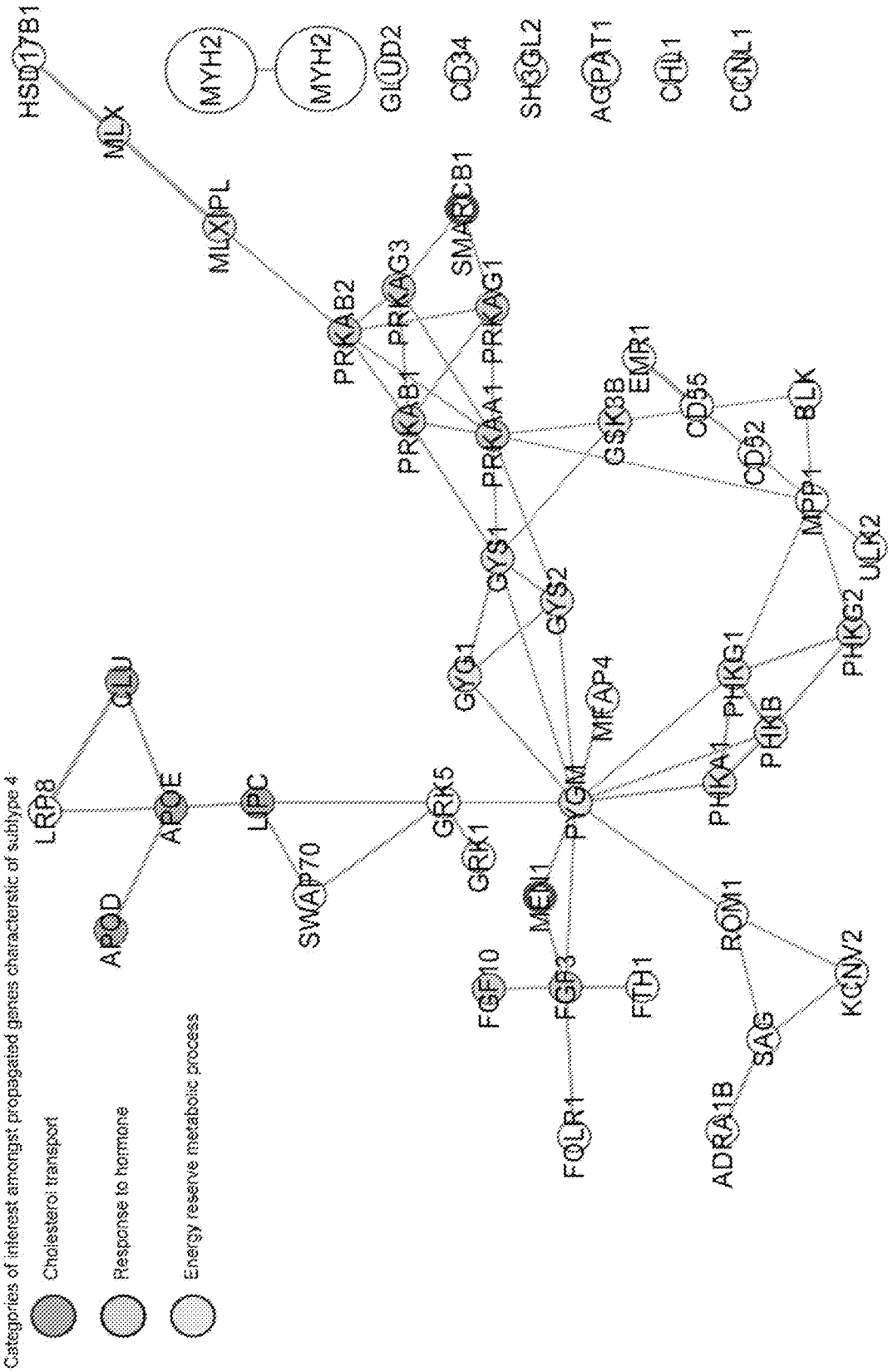


FIGURE 33

34/44

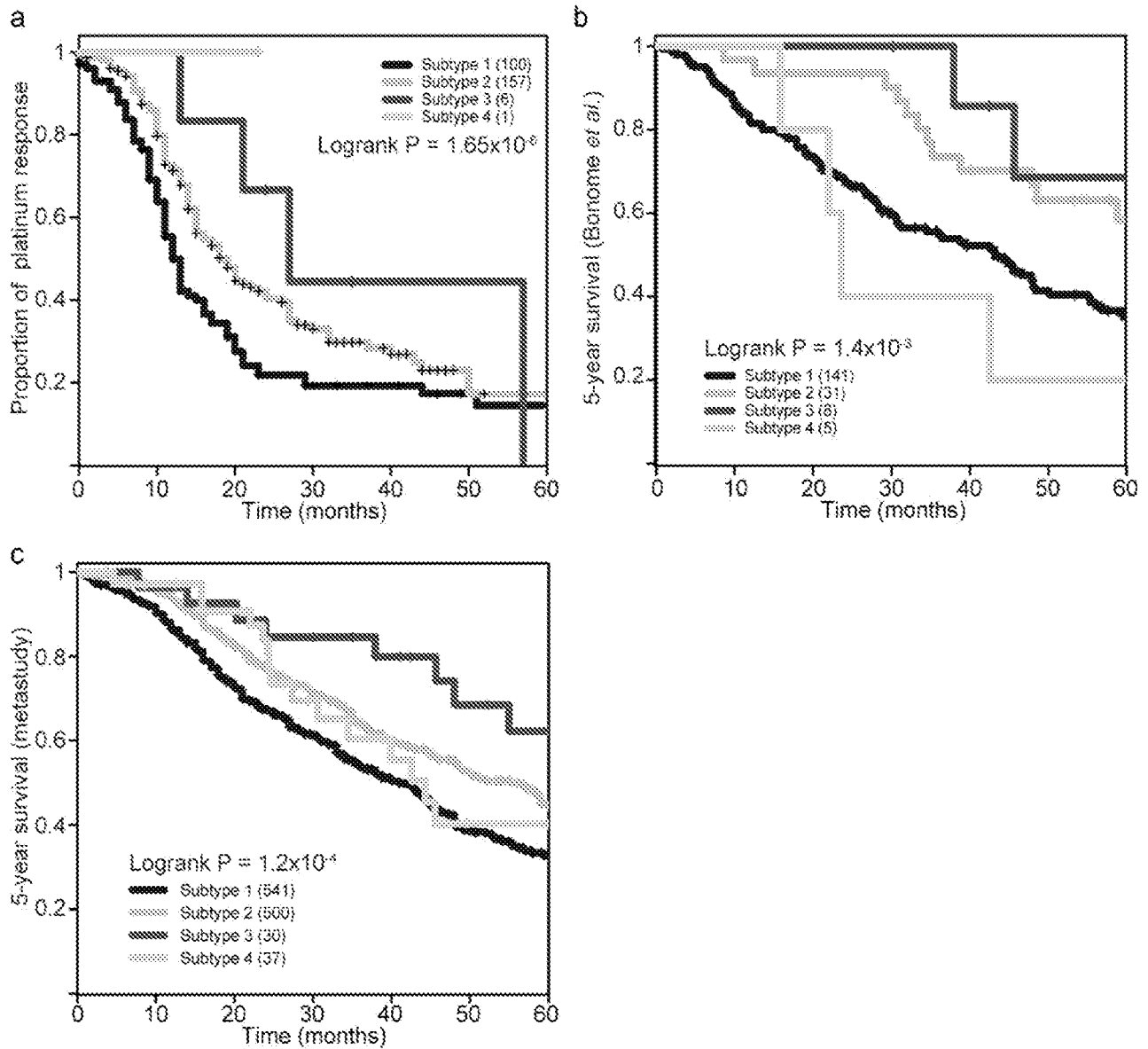


FIGURE 34

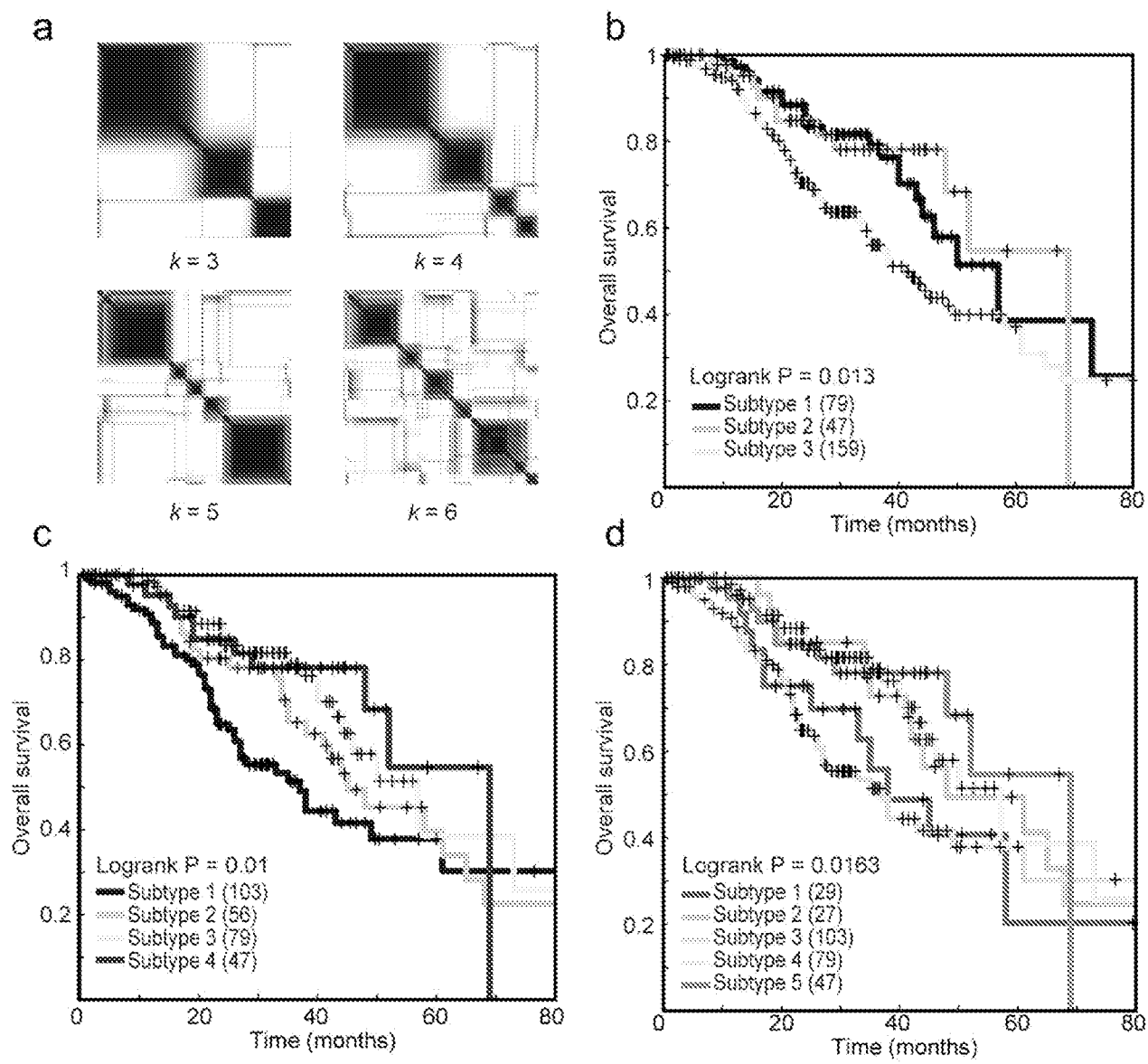


FIGURE 35

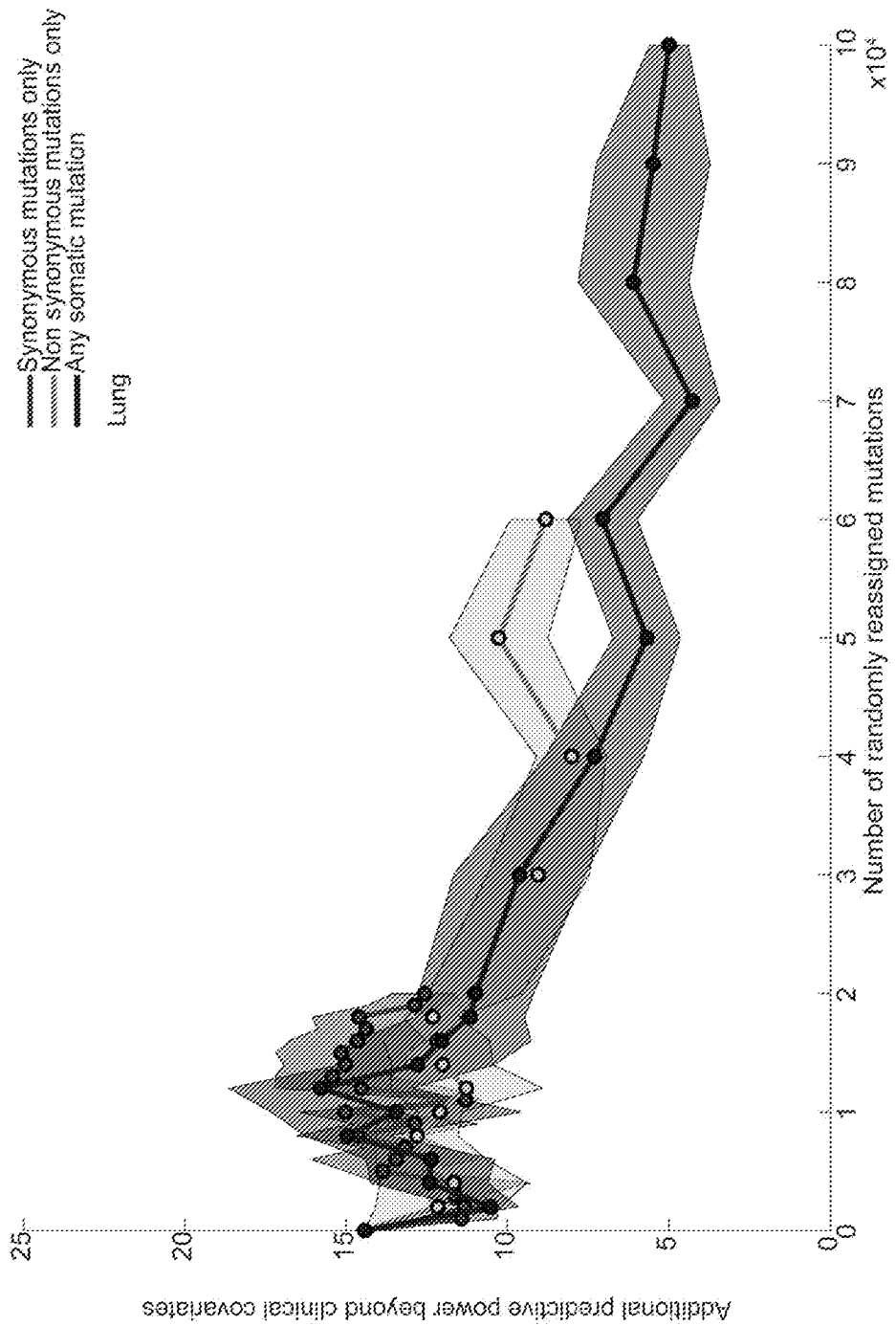


FIGURE 36

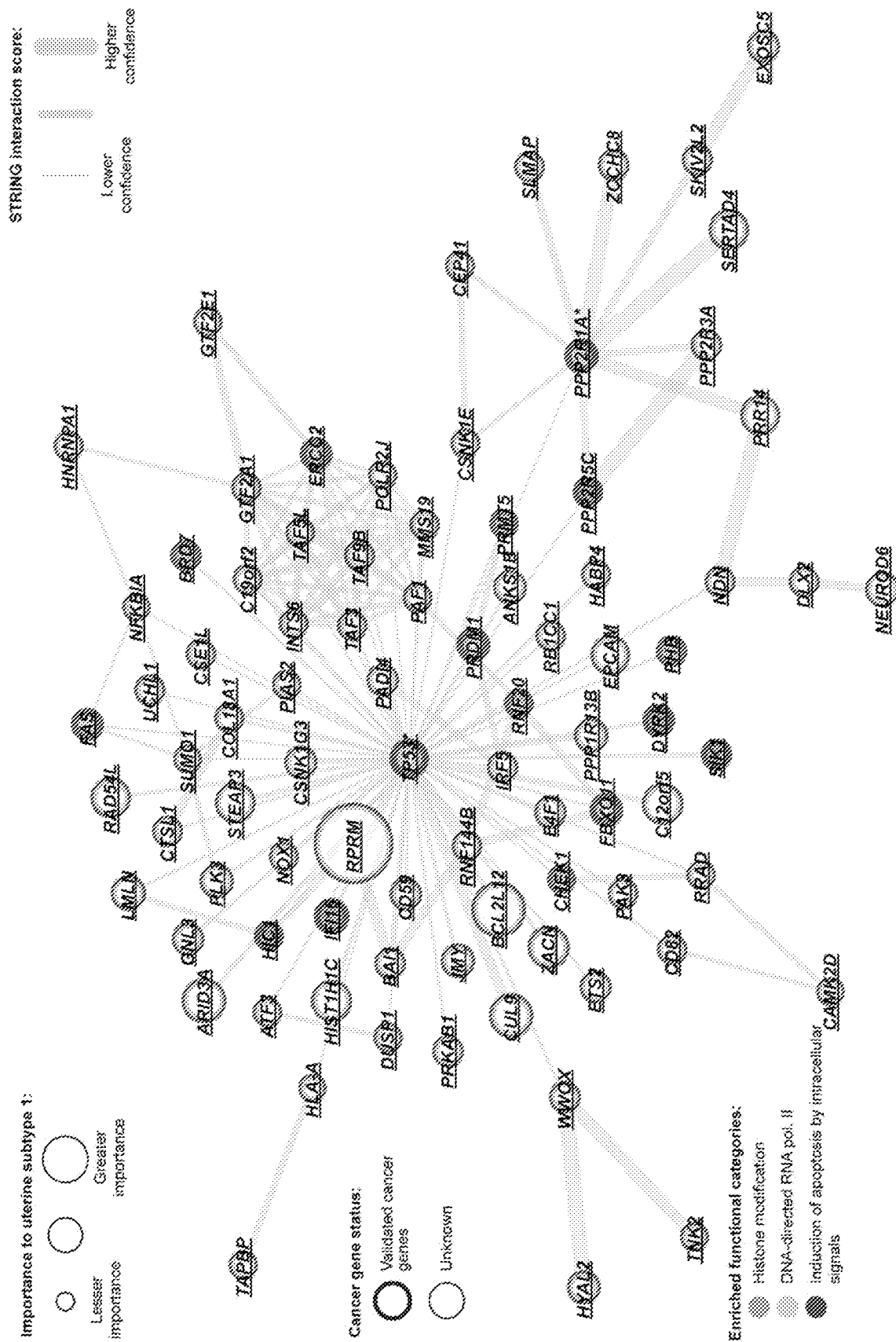


FIGURE 37

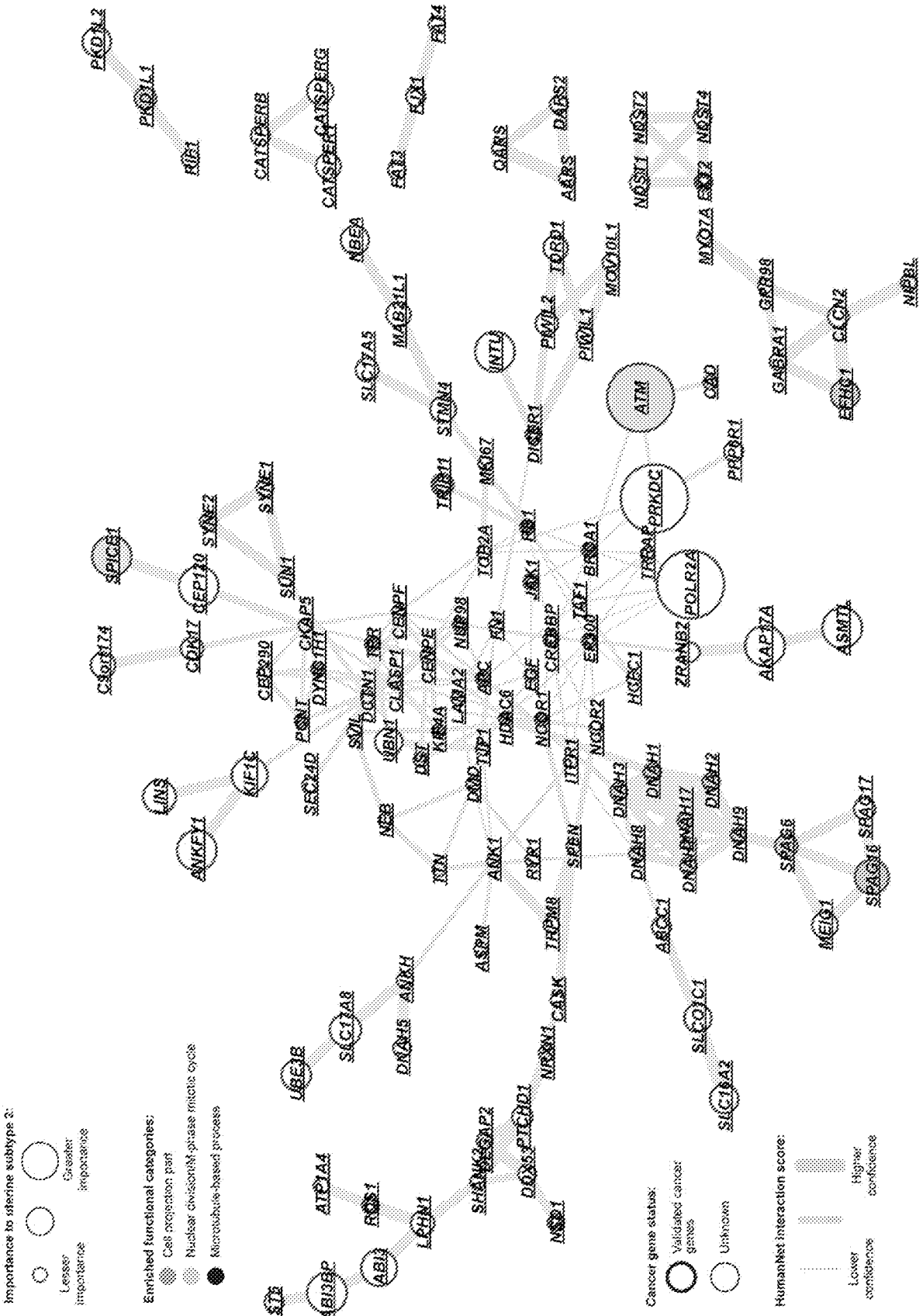


FIGURE 38

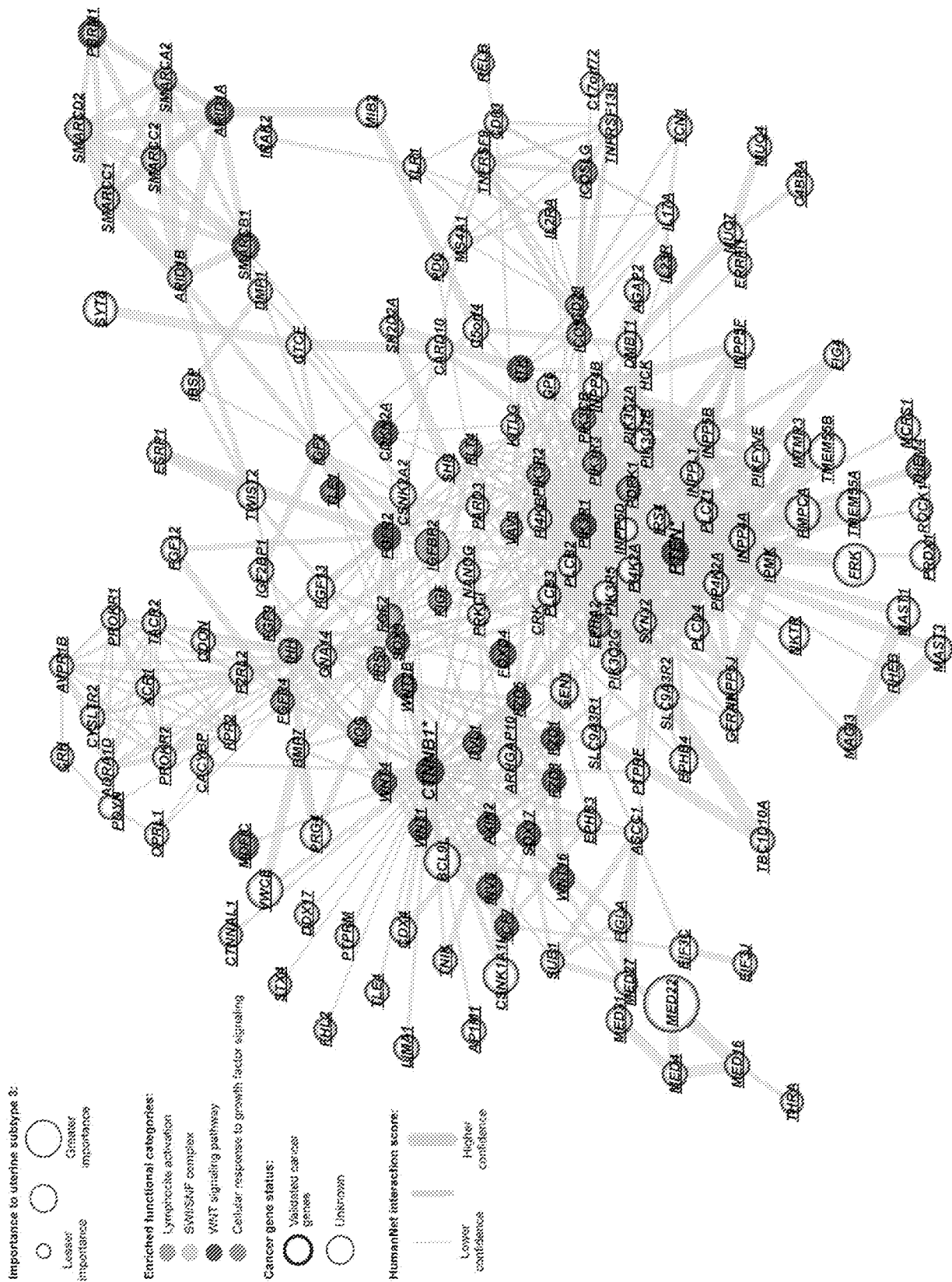


FIGURE 39

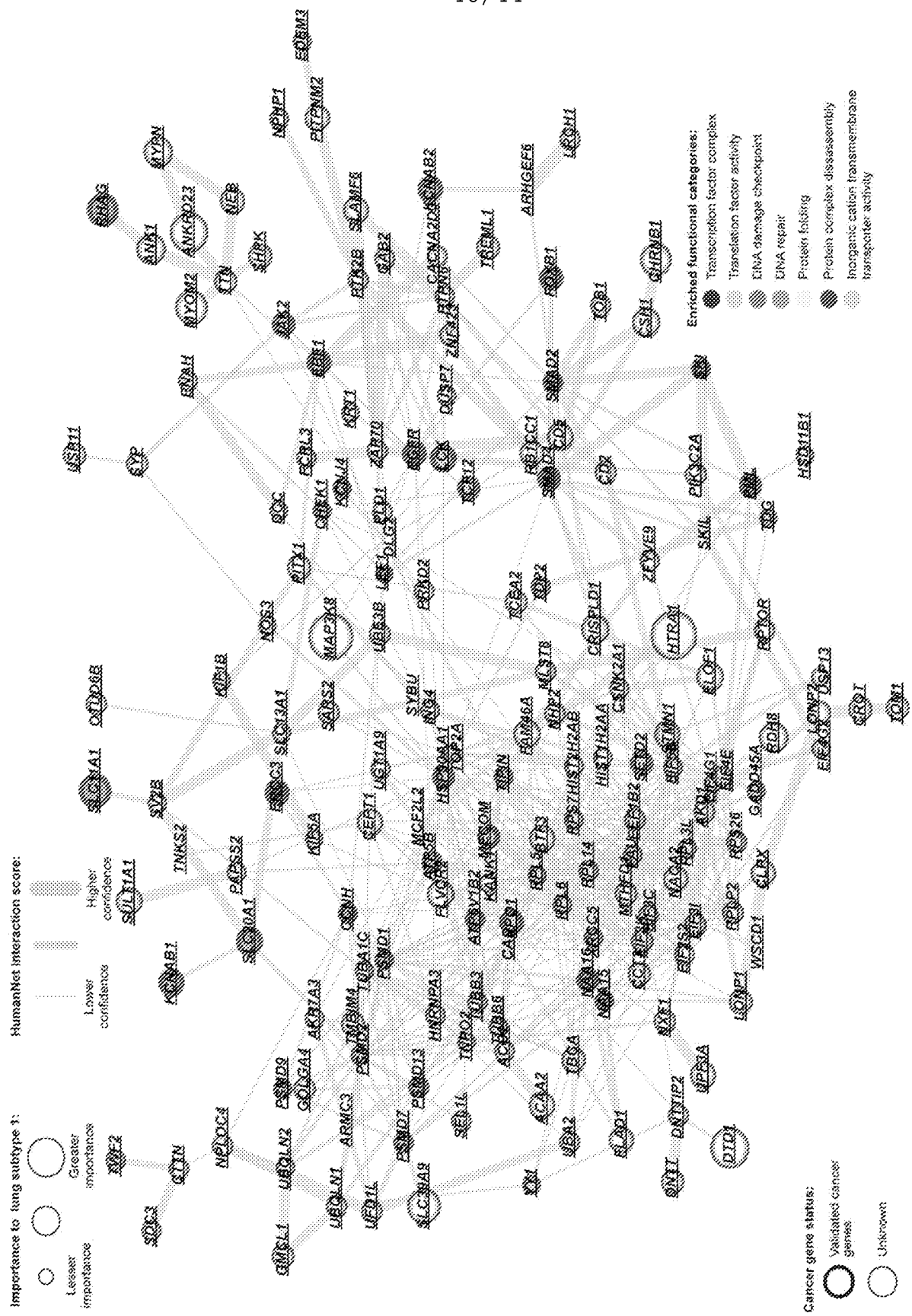


FIGURE 40

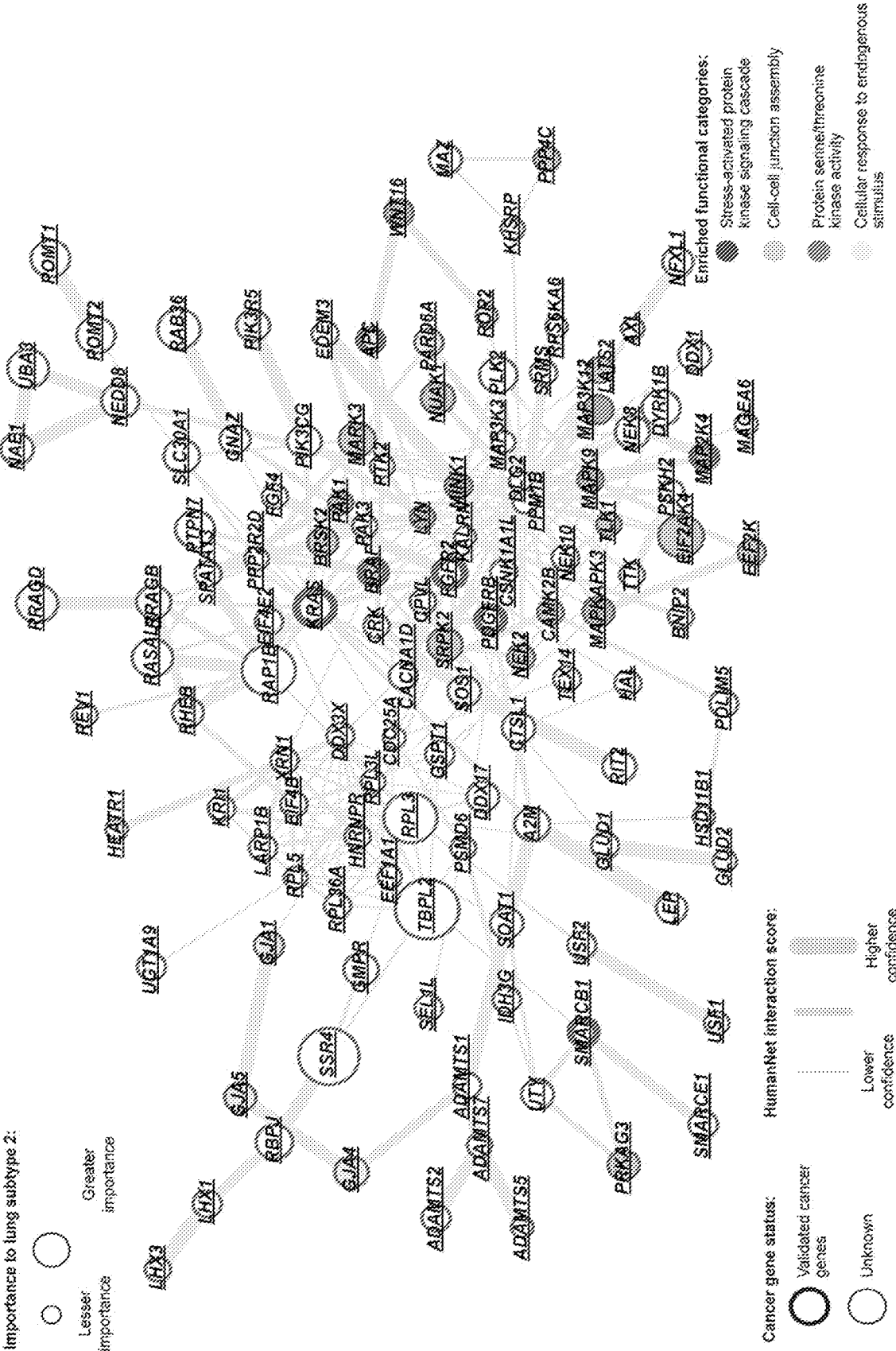


FIGURE 41

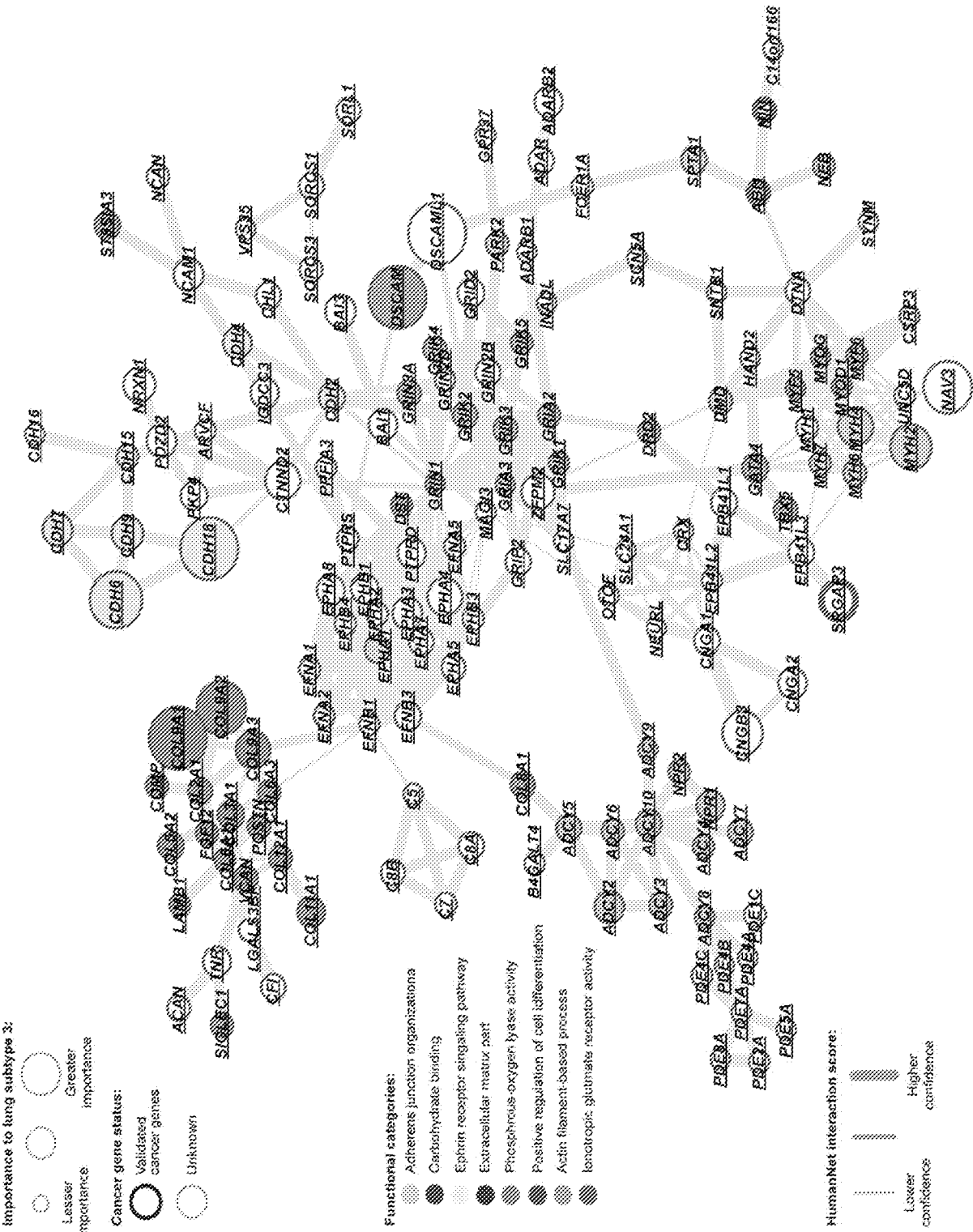
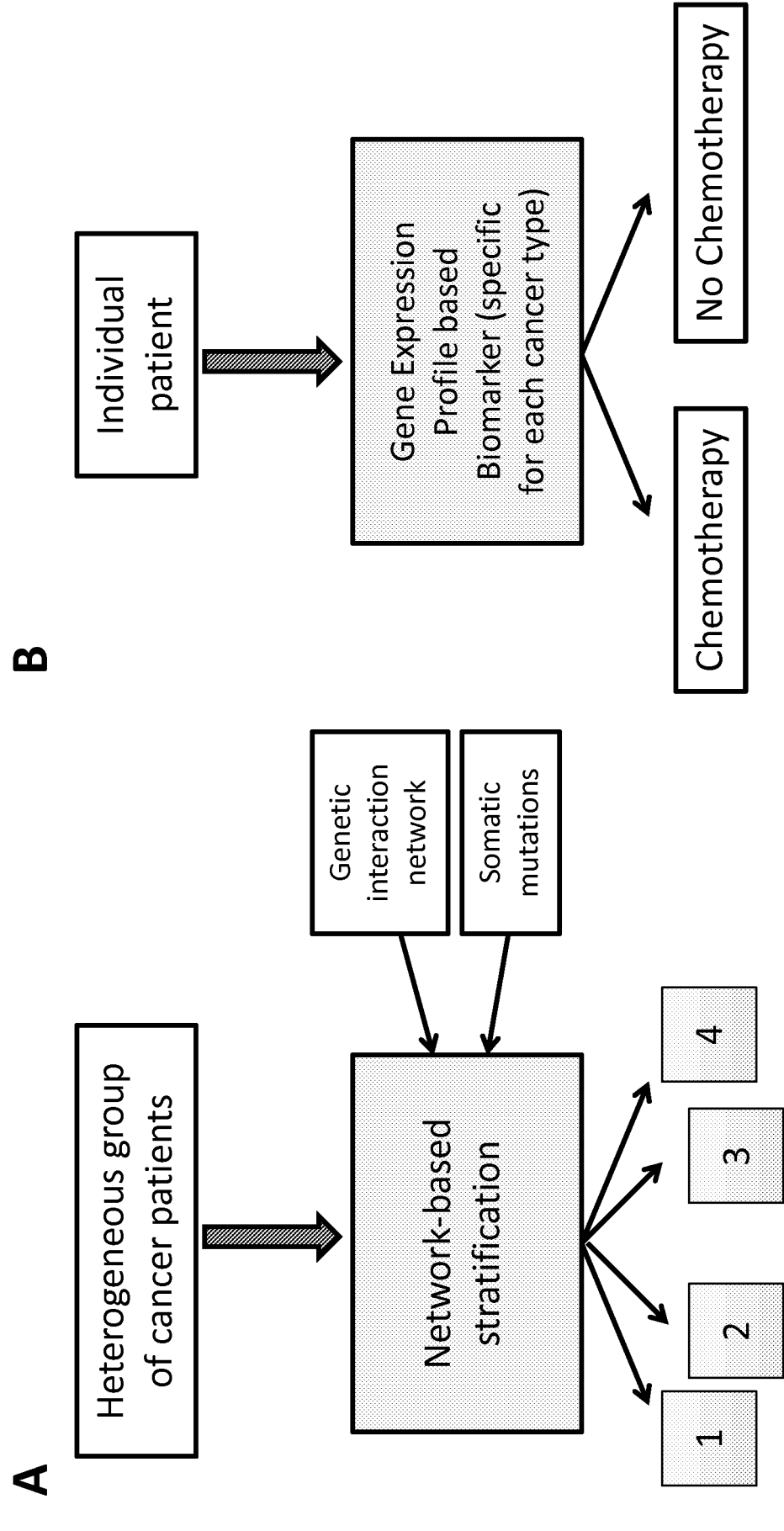


FIGURE 42



Homogeneous subtypes, predictive of clinical outcomes

FIGURE 44