



(51) International Patent Classification:

G16H 50/20 (2018.01) G16B 20/20 (2019.01)
G16B 40/20 (2019.01)

(21) International Application Number:

PCT/IL2022/051279

(22) International Filing Date:

01 December 2022 (01.12.2022)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/284,817 01 December 2021 (01.12.2021) US

(71) Applicants: **HADASIT MEDICAL RESEARCH SERVICES AND DEVELOPMENT LTD.** [IL/IL]; Jerusalem BioPark, Hadassah Ein Kerem Medical Center, P.O.B. 12000, 9112001 Jerusalem (IL). **INSERM (INSTITUT NATIONAL DE LA SANTÉ ET DE LA RECHERCHE MÉDICALE)** [FR/FR]; 101 Rue de Tol-

biac, 75013 Paris (FR). **SORBONNE UNIVERSITÉ** [FR/FR]; 21 rue de l'École de Médecine, 75006 Paris (FR).

(72) Inventors: **ROSENBERG, Shai**; 22 Ha'gdud Ha'Ivry St., 9234517 Jerusalem (IL). **SOUSSI, Thierry**; 101 Rue de Tolbiac, 75013 Paris (FR).

(74) Agent: **MINTOS, Moshe** et al.; GASSNER MINTOS, POBox 3393, 6103301 Tel Aviv (IL).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, 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, ST, SV, SY, TH,

(54) Title: MACHINE LEARNING PREDICTION OF GENETIC MUTATIONS IMPACT

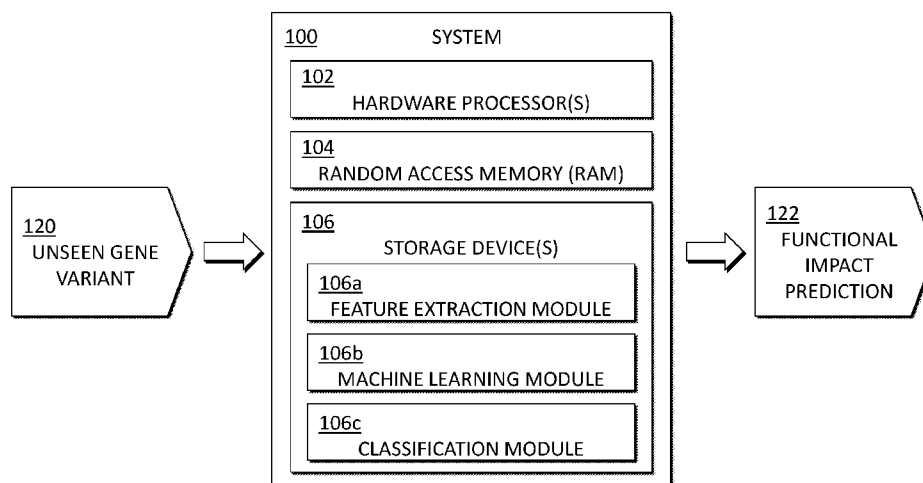


FIG. 1

(57) Abstract: A computer-implemented method comprising: receiving genetic information with respect to a plurality of known variants of a gene of interest; calculating, based on the received genetic information, for each of the variants, a set of features representing a functional estimation or a loss of function (LOF) associated with the variant; at a training stage, training a machine learning model on a training dataset comprising: (i) all of the sets of features, and (ii) labels indicating a pathogenicity associated with each of the variants; and at an inference stage, applying the trained machine learning model to an unseen target variant of the gene of interest, to predict a pathogenicity of the unseen target variant of the gene of interest.

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS,
ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, CV, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

MACHINE LEARNING PREDICTION OF GENETIC MUTATIONS IMPACT**CROSS-REFERENCE TO RELATED APPLICATIONS**

[0001] This application claims priority from U.S. Application Ser. No. 63/284,817, filed December 1, 2021, entitled “MACHINE LEARNING PREDICTION OF GENETIC MUTATIONS IMPACT,” the contents of which are hereby incorporated herein in their entirety by reference.

FIELD OF THE INVENTION

[0002] The present invention relates to the field of machine learning.

BACKGROUND

[0003] This invention relates to the field of machine learning.

[0004] Cancer is caused by a sequence of acquired somatic genomic aberrations. Large scale sequencing studies have shown that individual patients have unique mutations profiles, some of which may be susceptible to different particular drug therapies. This personalized medicine approach already led to major clinical achievements, such as in targeting BRAF V600E mutations in melanoma, and in targeting EGFR mutations in lung cancer. It has also been shown in a meta-analysis of phase II clinical trials that a personalized approach is more beneficial in clinical trials as compared to non-personalized approach.

[0005] However, application of personalized genomic medicine in cancer is still limited, and only a small minority of cancer patients are assigned to this approach. One of the major obstacles is that tumors usually have many mutations, and it is difficult to define the major drivers in a given pathology, and to prioritize drug selections accordingly. Thus, correctly identifying the true driver mutations in a patient's tumor is a major challenge in precision oncology.

[0006] Two key approaches were proposed and are being used to address this challenge. The first is attempting to predict the consequence of mutations based on biological reasoning, such as the appearance of specific mutations in active areas of cancer genes or the occurrence of similar mutations in other cancer patients. The second

is curating literature and knowledge databases for clinical response of tumors with similar mutations to candidate drugs. Nevertheless, these approaches offer only limited relief in the case of frequent mutations, and almost none when it comes to medium- and low-frequency mutations.

[0007] The foregoing examples of the related art and limitations related therewith are intended to be illustrative and not exclusive. Other limitations of the related art will become apparent to those of skill in the art upon a reading of the specification and a study of the figures.

SUMMARY OF THE INVENTION

[0008] The following embodiments and aspects thereof are described and illustrated in conjunction with systems, tools and methods which are meant to be exemplary and illustrative, not limiting in scope.

[0009] There is provided, in an embodiment, a system comprising at least one hardware processor; and a non-transitory computer-readable storage medium having stored thereon program instructions, the program instructions executable by the at least one hardware processor to: receive genetic information with respect to a plurality of known variants of a gene of interest, calculate, based on the received genetic information, for each of the variants, a set of features representing a functional estimation or a loss of function (LOF) associated with the variant, at a training stage, train a machine learning model on a training dataset comprising: (i) all of the sets of features with respect to the variants, and (ii) labels indicating a pathogenicity associated with each of the variants, and at an inference stage, apply the trained machine learning model to genetic information from an unseen target variant of the gene of interest, to predict a pathogenicity of the unseen target variant of the gene of interest.

[0010] There is also provided, in an embodiment, a computer-implemented method comprising: receiving genetic information with respect to a plurality of known variants of a gene of interest; calculating, based on the received genetic information, for each of the variants, a set of features representing a functional estimation or a loss of function (LOF) associated with the variant; at a training stage, training a machine learning model on a training dataset comprising: (i) all of the sets of features with respect to the variants, and (ii) labels indicating a pathogenicity associated with each of the variants; and at an inference stage, applying the trained machine learning model to genetic information from

an unseen target variant of the gene of interest, to predict a pathogenicity of the unseen target variant of the gene of interest.

[0011] There is further provided, in an embodiment, a computer program product comprising a non-transitory computer-readable storage medium having program instructions embodied therewith, the program instructions executable by at least one hardware processor to: receive genetic information with respect to a plurality of known variants of a gene of interest; calculate, based on the received genetic information, for each of the variants, a set of features representing a functional estimation or a loss of function (LOF) associated with the variant; at a training stage, train a machine learning model on a training dataset comprising: (i) all of the sets of features with respect to the variants, and (ii) labels indicating a pathogenicity associated with each of the variants; and at an inference stage, apply the trained machine learning model to genetic information from an unseen target variant of the gene of interest, to predict a pathogenicity of the unseen target variant of the gene of interest.

[0012] In some embodiments, the gene of interest is tumor protein 53 (tp53).

[0013] In some embodiments, at least some of the features are obtained using functional assays performed with respect to the plurality of known variants.

[0014] In some embodiments, the training dataset comprises only the features obtained using functional assays performed with respect to the plurality of known variants.

[0015] In some embodiments, at least some of the features are calculated using computational methods.

[0016] In some embodiments, the plurality of known variants comprises at least a first portion comprising variants having a number of occurrences in human cancer that is greater than one.

[0017] In some embodiments, the plurality of known variants comprises at least a second portion comprising variants having a number of occurrences in human cancer equal to one or zero.

[0018] In some embodiments, the labels indicating a pathogenicity are binary labels selected from the group consisting of: pathogenic and non-pathogenic.

[0019] In some embodiments, the unseen target variant of the gene of interest is obtained from a biological sample collected from a subject of interest.

[0020] There is further provided, in an embodiment, a method comprising: receiving genetic information from an unseen target variant of a gene of interest taken from a biological sample collected from a subject of interest; and applying, to the genetic information from the unseen target variant, a machine learning model trained to predict a pathogenicity of variants of the gene of interest, to predict a pathogenicity of the unseen target variant of the gene of interest, wherein a prediction of the unseen target variant of the gene of interest as pathogenic indicates a negative prognosis for the biological sample, and (ii) a prediction of the unseen target variant of the gene of interest as non-pathogenic indicates a positive prognosis for the biological sample.

[0021] There is further provided, in an embodiment, a method of classifying a sample from a subject, comprising the steps of: (i) determining the sequence of a gene of interest; (ii) identifying an unseen target variant of the gene of interest; and (iii) applying the computer-implemented method of claim 10 to determine the pathogenicity of the unseen target variant, wherein the presence of a (a) pathogenic unseen target variant of the gene of interest indicates a negative prognosis for the sample, and (b) non-pathogenic unseen target variant of the gene of interest indicates a positive prognosis for the sample.

[0022] In addition to the exemplary aspects and embodiments described above, further aspects and embodiments will become apparent by reference to the figures and by study of the following detailed description.

BRIEF DESCRIPTION OF THE FIGURES

[0023] The present invention will be understood and appreciated more comprehensively from the following detailed description taken in conjunction with the appended drawings in which:

[0024] Fig. 1 is a block diagram of an exemplary system which provides for training a machine learning model to predict a functional impact of an unseen variant of a gene of interest, according to some embodiments of the present disclosure;

[0025] Fig. 2 illustrates the functional steps in a method for training a machine learning model to predict a functional impact of an unseen variant of a gene of interest, in accordance with some embodiments of the present invention;

[0026] Fig. 3 shows a workflow diagram illustrating the training pipeline used for the development of the present machine learning model, in accordance with some embodiments of the present invention;

[0027] Figs. 4A-4B illustrates the p53 mutation landscape, in accordance with some embodiments of the present invention;

[0028] Figs. 5A-5C show test results for the model trained using only the functional features (pie chart A), the model trained using only computational features (pie chart B), and for the model trained using all the features (pie chart C) , in accordance with some embodiments of the present invention;

[0029] Figs. 6A-6C show multidimensional scaling (MDS) calculations based on LOF features of p53 variants, according to some embodiments of the present disclosure;

[0030] Figs. 7A-7B show functional analysis of p53 variants, according to some embodiments of the present disclosure;

[0031] Fig. 8 shows the frequency of 196 variants in gnomAD and UMD databases, according to some embodiments of the present disclosure;

[0032] Figs. 9A-9B show ClinVar comparison with the present machine learning model's predictions, according to some embodiments of the present disclosure; and

[0033] Fig. 10 shows the survival curve of tumors from TCGA database, according to some embodiments of the present disclosure.

DETAILED DESCRIPTION

[0034] Disclosed herein is a technique, embodied in a system, computer-implemented method, and computer program product, for training a machine learning model to predict a pathogenicity of an unseen variant of a gene of interest.

[0035] The present disclosure will use the terms “functionally normal” or “functionally abnormal” to denote the functional impact of a variant as measured in a given assay.

[0036] In some embodiments, the unseen target variant of a gene of interest may be obtained from a biological sample collected from a tumor site in a subject. The disclosed trained machine learning model may then be inferenced on genetic information extracted from the unseen target variant, to predict a pathogenicity of the unseen target variant of the gene of interest.

[0037] The following disclosure will discuss extensively embodiments of the present invention specifically with respect to tumor protein 53 (p53). However, the principles underlying the present disclosure may be equally and effectively applied to other similar genes for which loss of function (LOF) data as may be associated with variants of the gene are available.

[0038] As noted above, predicting the pathogenicity of somatic or germline genomic variations represents an unmet need in genetic consultation and precision genomic medicine for cancer treatment. In the context of the present disclosure, there is an unmet need to have accurate information regarding p53 variants, because mutations in p53 gene occur in a large number of tumors, with something like half of all tumors harboring such mutations.

[0039] By way of background, p53 is any isoform of a protein encoded by homologous genes in humans. This homolog serves a crucial function of preventing cancer formation and conserving stability by preventing genome mutation. Further, many studies have suggested that p53TPs53 mutations have prognostic importance and sometimes are a significant factor in determining the response of tumors to therapy.

[0040] Currently, somatic p53 status is used in routine clinical practice in several types of cancer, such as chronic lymphocytic leukemia (CLL), acute myeloid leukemia (AML), and myelodysplastic syndrome, in order to identify patients likely to benefit from specific treatments. For example, a germline mutation in p53 causes the Li-Fraumeni Syndrome (LFS) with severe genetic predisposition to cancer. Furthermore, it has been clearly established that germline p53 variants are frequent in familial cancer syndromes, such as LFS, or in families with hereditary breast and ovarian cancer, and surveillance of individuals with an identified germline p53 mutation is highly beneficial to improve the likelihood of early tumor detection and subsequently improved outcomes.

[0041] However, classification of p53 variants from human cancer based on their pathogenicity is highly challenging. Although the coding sequence of the p53 gene is small (1800 nucleotides for a 393 amino acid protein), distinguishing true driver variants from sequencing artefacts, passenger mutations and benign polymorphisms is particularly difficult, because missense variants have been found at nearly every p53 codons (albeit at various frequencies) with a high concentration in the 200 residues of the DNA binding domain of the protein.

[0042] This challenge is complicated by the landscape of p53 variants which is composed predominantly of multiple missense mutations spread-out on the entire gene. Among the 2,314 possible missense variants in the coding region, 1,621 (70%) have been described in at least one tumor and among them only 190 have interpretation in ClinVar, the leading genomic variant database. The greatest advantage for the analysis of missense mutations in p53 is that the read-out of p53 functions can be easily monitored. Currently, the functional activity of more than 10,000 p53 variants from 12 different readouts are available.

[0043] Accordingly, in some embodiments, the present disclosure provides for a machine learning model configured to predict the functional consequences or impact (e.g., pathogenicity) of unseen missense mutation in p53 variants. The present inventors have validated the model using multiple independent datasets of normal and cancer patients, and it has been shown to provide a significant predictive value for survival analysis.

[0044] In some embodiments, the present disclosure provides for training a machine learning model to obtain a trained machine learning model configured to predict a loss of function of an unseen missense variant in a gene of interest.

[0045] In some embodiments, a trained machine learning model of the present disclosure is trained on a training dataset comprising genetic information with respect to a plurality of known variants of a gene of interest (e.g., p53). In some embodiments, the genetic information comprises, with respect to each of the known variants of the gene, one or more scores representing a loss of function (LOF) associated with that variant. In some embodiments, at least some of the one or more scores representing LOF are obtained using functional assays and/or computational methods.

[0046] In some embodiments, in the case of p53, the present disclosure takes advantage of publicly-available data associated with p53, e.g., p53 variants reported in multiple datasets, which allows for a robust construction of a training dataset of the present disclosure.

[0047] In some embodiments, each known variant of the gene of interest (e.g., p53) comprised in the training dataset of the present disclosure is labeled with a label indicating a pathogenicity of the variant, e.g., pathogenic or non-pathogenic (which may be also expressed as 'deleterious' or 'non-deleterious'). In some embodiments, a

pathogenic or deleterious label may be assigned to a specific variant of the plurality of known variants when such specific variant is associated with a diagnosis of cancer in one or more subjects.

[0048] The trained machine learning model of the present disclosure was validated using various independent datasets enriched in non-pathogenic variants or pathogenic variants. Validation analysis of the present trained machine learning model over the ClinVAR database (a public database of variant interpretations) reflected an accuracy level of 96.3% and a sensitivity of 95.7%, wherein benign and likely-benign p53 variants were detected with a sensitivity of 100%.

[0049] The trained machine learning model of the present disclosure was further validated using the survival data of the Cancer Genome Atlas (TCGA) Program. Survival of patients with missense variants predicted as non-functional by the present trained machine learning model is comparable to the survival of patients with truncating variants in p53. By contrast, patients with variants predicted as functional by the present machine learning model had longer overall survival which was comparable to patients with no p53 mutations. These analyses portray a picture of a highly robust model, with the capability of correctly identifying p53 mutations whether somatic or germline, benign or pathogenic, in healthy and in sick patients.

[0050] A potential advantage of the present invention is, therefore, in that it may accurately predict a functional impact of an unseen mutational variant of a gene of interest. Thus, the present invention may have significant implications in situations where somatic and germline mutations gene diagnostics are of clinical relevance. A machine learning model of the present disclosure may be used in conjunction with a patient's family history, to provide a probability score of pathogenicity for the individual patient. In some embodiments, a machine learning model of the present disclosure may also be applied for the purpose of classification of somatic mutations in the p53 gene for therapeutic decision making such as for CLL patients.

[0051] Fig. 1 is a block diagram of an exemplary system **100** which provides for training a machine learning model to predict a functional impact of an unseen variant of a gene of interest, in accordance with some embodiments of the present invention.

[0052] In some embodiments, system **100** may comprise a hardware processor **102**, and a random-access memory (RAM) **104**, and/or one or more non-transitory computer-

readable storage device **106**. In some embodiments, system **100** may store in storage device **106** software instructions or components configured to operate a processing unit (also ‘hardware processor,’ ‘CPU,’ ‘quantum computer processor,’ or simply ‘processor’), such as hardware processor **102**. In some embodiments, the software components may include an operating system, including various software components and/or drivers for controlling and managing general system tasks (e.g., memory management, storage device control, power management, etc.) and facilitating communication between various hardware and software components. Components of system **100** may be co-located or distributed, or the system may be configured to run as one or more cloud computing ‘instances,’ ‘containers,’ ‘virtual machines,’ or other types of encapsulated software applications, as known in the art.

[0053] Storage device(s) **106** may have stored thereon program instructions and/or components configured to operate hardware processor(s) **102**. The program instructions may include one or more software modules, such as a feature extraction module **106a**, a machine learning module **106b**, and a classification module **106c**.

[0054] Feature extraction module **106a** is configured to extract feature data from input genetic information with respect to a plurality of mutational variants of a gene of interest. In some embodiments, feature extraction module **106a** may be configured to receive genetic information with respect to a plurality of known variants of a gene of interest (e.g., p53), and to calculate one or more scores representing a loss of function (LOF) associated with each of the variants. In some embodiments, at least some of the one or more scores representing LOF are obtained using functional assays and/or computational methods. As used herein, the term ‘LOF’ refers broadly to a type of mutation in which the altered gene product lacks the molecular function associated with that gene.

[0055] Machine learning module **106b** may comprise any one or more suitable neural networks architectures (i.e., which include one or more neural network layers), and can be implemented using any suitable optimization algorithm. In some embodiments, machine learning module **106b** may be configured to construct a training dataset comprising one or more sets of features extracted and/or calculated by feature extraction module **106a**. Machine learning module may then be configured to train a machine learning model of the present disclosure (which may be implemented by classification module **106c**), on the constructed dataset.

[0056] In some embodiments, classification module **106c** may comprise one or more machine learning algorithms which may be trained on a training dataset construed by machine learning module **106b**. The trained model may be inferenced on input genetic and related information **120** from an unseen variant of a gene of interest, and to output a prediction **122** with respect to a functional impact of the unseen variant of the gene of interest.

[0057] System **100** as described herein is only an exemplary embodiment of the present invention, and in practice may be implemented in hardware only, software only, or a combination of both hardware and software. System **100** may have more or fewer components and modules than shown, may combine two or more of the components, or may have a different configuration or arrangement of the components. System **100** may include any additional component enabling it to function as an operable computer system, such as a motherboard, data busses, power supply, a network interface card, a display, an input device (e.g., keyboard, pointing device, touch-sensitive display), etc. (not shown). Moreover, components of system **100** may be co-located or distributed, or the system may be configured to run as one or more cloud computing ‘instances,’ ‘containers,’ ‘virtual machines,’ or other types of encapsulated software applications, as known in the art.

[0058] The instructions of system **100** will now be discussed with reference to the flowchart of Fig. 2, which illustrates the functional steps in a method **200** for training a machine learning model to predict a functional impact of an unseen variant of a gene of interest, in accordance with some embodiments of the present invention. The various steps of method **200** will be discussed with continued reference to system **100** shown in Fig. 1, and to Fig. 3 which shows a workflow diagram illustrating the training pipeline used for the development of the present machine learning model.

[0059] The various steps of method **200** will be described with continuous reference to exemplary system **100** shown in Fig. 1. The various steps of method **200** may either be performed in the order they are presented or in a different order (or even in parallel), as long as the order allows for a necessary input to a certain step to be obtained from an output of an earlier step. In addition, the steps of method **200** may be performed automatically (e.g., by system **100** of Fig. 1), unless specifically stated otherwise. In addition, the steps of method **200** are set forth for exemplary purposes, and it is expected that modifications to the flowchart may be implemented as necessary or desirable.

[0060] Method **200** begins at step **202**, wherein system **100** receives, as input, genetic information with respect to a plurality of mutation variants of a gene of interest. In some embodiments, the received genetic information may be acquired from available source databases. For example, in the exemplary case of p53, the information may be received from one or more of the following available sources:

- The UMD_p53 database: The UMD_p53 includes the p53 status of more than 80,400 tumors, individuals with germline mutations, and cell lines, analyzed both by conventional Sanger sequencing as well as NGS. (See, Leroy B. Fournier et al., The p53 website: an integrative resource centre for the p53 mutation database and p53 mutant analysis. Nucleic acids research.).
- Mutant Loss of Activity Database (MULTLOAD): This database was first released in 2012 and includes comprehensive details on the properties of p53 variants based on 600 publications. MULTLOAD includes multiple activity fields, such as change of transactivation on various promoters, apoptosis or growth arrest performed in multiple experimental conditions). For several hot spot mutants, multiple gain of function activities are also included. MULTLOAD includes more than 150,000 entries with multiple entries for most variants.
- Cancer mutation databases: Including data from the Cancer Genome Atlas (TCGA) Program, Memorial Sloan Kettering Cancer Center, and International Cancer Genome Consortium (ICGC). In some embodiments, genomic data, such as genomic coordinates and genetic events, were extracted from each dataset to define the correct annotation according to Human Genome Variation Society (HGVS) recommendations. In some embodiments, variant annotation were validated by using the Name Checker tool developed by Mutalyzer (see, <https://mutalyzer.nl/>).
- TCGA survival data: Includes tumor samples with their p53 mutational status and survival information (n= 10,322). Samples with more than one mutation were excluded from the analysis to prevent conflicting conclusions. Patients with more than one sample were also removed to prevent conflict. The four categories of p53 mutational status includes:
 - (i) no mutation in the p53 gene – 6,987 samples,
 - (ii) missense mutation in p53 classified as non-deleterious – 66 samples,

- (iii) missense mutation in p53 classified as deleterious – 2,118 samples, and
 - (iv) Tumors with a truncating non-missense mutation in p53 – 1,151 samples. Frameshift deletion or insertion, nonsense, mutations in splice region or in splice site were considered truncating. Tumors with other non-missense mutations in p53 were excluded from the analysis.
- Population database: The Genome Aggregation Database (gnomAD) is a resource developed for aggregating and harmonizing exome and genome sequencing from normal population. It is the largest source of SNP available and includes data from 141,456 individuals. p53 variants were extracted from version 2.1 (version no cancer) and validated by using the Name Checker tool developed by Mutalyzer (<https://mutalyzer.nl/>).
 - Predictive data: dbNSFP is a database that compiles prediction scores from multiple algorithms, along with conservation scores and other related information, for every potential non-synonymous variant in the human genome. Data for p53 was extracted from version 3.5 and manually curated to be specific to the full p53 protein and 21 dbNSFP scores were retained for the analysis. Scores originating from seven other in-silico predictive software were also included in the present study leading to a total of 28 different scores used for the training analysis.
 - Functional data: The UMD_p53 database includes three sets of functional data for p53 variants, comprising a total of 14 different readouts for p53 function. The first set includes p53 transcriptional activity which is essential for its tumor suppressive function, as tested on eight different promoter sequences in a yeast assay. The average and median value of the eight activities were also included as readouts as they can improve the training. The second set of functional data integrated in the UMD_p53 database includes assessment of cell cycle arrest activity of all variants localized in the DNA binding domain of p53, in H1299 cells. The third set includes dominant negative activity, loss of function and response to etoposide as analyzed in mammalian cells for 8,258 p53 variants.

[0061] In the exemplary case of p53, the present disclosure applies the concept of cancer shared dataset (CSD) to the information received in step 202. Accordingly, the present disclosure provides for defining a common p53 dataset, which combines selected

information from all data sources presented above. In the present case, the common p53 dataset includes only p53 variants which are found at least once in each of the source databases. Because the source databases are all derived from independent studies using different patients and different methodologies, it is likely that variants shared amongst all databases represent true recurrent cancer-associated variants.

[0062] Accordingly, in the exemplary case of p53, the common p53 dataset comprises 290 missense variants were found to be shared by all source databases.

[0063] In some embodiments, the common p53 dataset also includes a ‘negative’ portion, comprising p53 variants that were never found (693 variants) or found only once (323 variants) in human cancer. The ‘negative’ portion of the common p53 dataset includes only missense variants, because they are the most common alterations detected for p53, and the most difficult to classify.

[0064] In some embodiments, the instructions of system **100** may comprise applying one or more data selection process with respect to the common p53 dataset assembled in step **202**. In some embodiments, the data selection process may comprise selecting a subset of gene variants for use in training the machine learning model of the present disclosure.

[0065] In the exemplary case of p53, the data selection process results in a ‘positive,’ cancer-related portion of the common p53 dataset assembled in step **202** comprising 290 p53 variants, and a ‘negative,’ non cancer-related portion comprising 1016 variants (1011 after removing variants kept for experimental validation).

[0066] Figs. 4A-4B show the p53 mutation landscape. As can be seen in Fig. 4A, the open reading frame of the major transcript of p53 (NM_000546.6) can sustain 3546 single nucleotide substitutions leading to 2569 different cDNA variants (c-variants) and the synthesis of 2314 potential protein variants (p-variants). About 30% of these variants have never been identified in any publicly-available database. Variants selected for the negative and positive sets are shown in the lower part of the figure.

[0067] Fig. 4B shows the occurrence and frequency of variants in the UMD_p53 database (2019 release). Variants have been split in four classes according to their occurrence in the database (1-10, 11-100, 101-1,000, and more than 1,000). Left axis: occurrence of each variant in patients included in UMD (log scale). Nine variants (p.Arg175His, p.Arg248Trp, p.Arg273Cys, p.Arg273His, p.Arg282Trp, p.Arg248Gln,

p.Tyr220Cys, p.Gly245Ser and p.Arg249Ser) have been reported in more than 1,000 patients.

[0068] Among the 1750 c-variants (1624 p-variants) that have been described in the 2019 release of UMD_p53 database, 103 c-variants (102 p-variants) were described in more than 100 cases corresponding to 70% of missense variants detected in human tumors with eight protein variants described more than 1,000 times and corresponding to 29% of the patients. On the other hand, 1,158 c-variants (1077 p-variants) were described at low frequency (1–9 times) and correspond to 6% of patients in the database. Four hundred eighty nine c-variants (470 p-variants) are found at intermediate frequency (10–99 times) and correspond to 24% of the patients. Although for a few hot spot variants oncogenic activity was profusely validated in multiple cellular or mouse models, information related to less frequent p53 variants is scarce.

[0069] Table 1 below summarizes the spectrum of p53 variants in different tumor types.

Table 1:

Tumor	Missense	Nonsense	Frameshift	Inframe	Splice	Syn.	other
Lung	69.1%	12.7%	9.2%	0.6%	5.5%	1.2%	0.6%
Colon	72.8%	12.6	7.6	0.4	3.1	1.3	2.2
Breast	63.6%	12.8	13.6	0.5	4.7	2.2	2.6
Ovary	65.6%	11.1	13.0	0.5	6.0	1.7	2.2
Pancreas	66.1%	12.3	13.1	0.5	4.7	0.7	2.5
Brain	83.2%	6.1	4.7	0.4	2.9	0.9	1.8
Oral	66.6%	13.7	9.7	0.6	4.2	2.9	2.3
Gastric	69.6%	12.1	9.7	0.4	3.7	2.5	2.1

[0070] In some embodiments, in step **204**, the instructions of feature extraction module **106a** may cause system **100** to extract and/or calculate one or more sets of features from the information received in step **202**.

[0071] For example, in the exemplary case of p53, the instructions of feature extraction module **106a** may cause system **100** to extract and/or calculate, from the input information received in step **202**, one or more features representing a loss of function (LOF) in each gene variant included in the received information. Accordingly, in some embodiments, with respect to each gene variant, the one or more features represent a lack of the molecular function associated with a gene of interest.

[0072] In some embodiments, one or more of the following exemplary LOF and similar features may be extracted and/or calculated with respect to at least some of the p53 variants included in the received information:

- Exemplary functional assay-based features, including, but not limited to:
 - Residual transcriptional activity of mutant p53 on the WAF1 promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the MDM2 promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the BAX promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the 14-3-3-s promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the AIP promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the GADD45 promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the NOXA promoter (% compared to wild-type).
 - Residual transcriptional activity of mutant p53 on the p52R2 promoter (% compared to wild-type).
 - Median value of the 8 promoter activity
 - Average value of the 8 promoter activity
 - variants dominant negative activity tested in A549 cells treated with nutlin.

- variants loss of function tested in p53_ko A 549 cells treated with nutlin.
- variants response to etoposide tested in p53_koA 549 cells treated with nutlin.
- variant loss of function tested in H1299; only variants in the DBD are available
- Exemplary computational features, including, but not limited to:
 - Impact assessment of amino acid substitutions based on evolutionary conservation in protein homologs.
 - Likelihood Ratio Test based on comparative genomics data set of 32 vertebrate species to identify deleterious mutations.
 - DNA sequence variants evaluated for disease causing potential based on in silico tests of amino acid substitutions.
 - Logistic regression based on 10 component scores (SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy, PhyloP) and the maximum frequency observed in the 1000 genomes populations.
 - Conservation score based on 46way alignment primate set.
 - Conservation score based on 46way alignment placental set.
 - Evolutionary constraint score using maximum likelihood evolutionary rate estimation, based on "Rejected Substitutions" (RS) score.
 - Conservation score using rigorous statistical tests to detect bases under selection based on 29way alignment.
 - Conservation score based on 100way alignment vertebrate set.
 - Conservation score based on 46way alignment primate set.
 - Conservation score based on 46way alignment placental set.
 - Conservation score based on 100way alignment vertebrate set.
 - Score predicting the pathogenicity of missense variants based on individual tools (mutpred, fathmm, vest, polyphen, sift, provean,

mutassessor, mutationTaster, LRT, GERP, SiPhy, phyloP and PhastCons).

- EA scores for each mutation were calculated based on a simple model of the phenotype–genotype relationship which hypothesizes that protein evolution is a continuous and differentiable process.
- RF method for predicting protein-ligand affinity, using Random forests. The features for each protein are the number of occurrences of a particular protein-ligand atom type pair interacting within a certain distance range.
- Score predicting the effect of amino acid substitution on protein function based on sequence homology and physical properties.
- Score predicting impact of amino acid substitution based on multiple sequence alignment and physical considerations and using machine learning (uses alleles encoding human proteins and their closely related mammalian homologs as True Negative observations).
- Score predicting impact of amino acid substitution based on multiple sequence alignment and physical considerations and using machine learning (applies common human nsSNVs as true negative observations).
- Pairwise sequence alignment scores used to generate predictions for the effect of amino acid substitutions.

[0073] In some embodiments, in step **204**, the instructions of feature extraction module **106a** may further cause system **100** to calculate additional feature, including, but not limited to:

- Germline-to-somatic (GVS) ratio score: The UMD_p53 database includes p53 variants identified in various types of tumors. However, because in most cases normal DNA is not available, it is possible that rare non-pathogenic SNPs are misclassified as somatic variants. Nevertheless, the large number of variants included in p53_UMD allows some specific analysis to identify potential constitutional SNPs. This allows the calculation of an additional feature—the GVS ratio. Because the UMD_p53 database includes both germline and somatic mutations, and since the distribution of variants is similar in both, it is possible to

define for each variant the GVS ratio. This will define if a p53 variant is found at higher frequency as germline variants.

- Frequency of multiple mutations (MMF) score: MMF score reflects the frequency at which each p53 variant is found, associated with one or more other p53 variants in the same tumor. This score will detect variants that are frequently co-selected because they are either benign passenger variants or low frequency SNPs. For all variants included in UMD_p53, the number of p53 variants per tumor has been fully recorded. Although the majority of tumors (91%) express a single p53 variant, 7% and 2% express either two or more than two p53 variants, respectively.

[0074] In some embodiments, different, additional, and/or other LOF-related features may be extracted and/or calculated in this step, with respect to at least one of the variants of the gene of interest.

[0075] In some embodiments, in the exemplary case of p53, step **204** includes a validation step with respect to Li Fraumeni Syndrome (LFS) p53 variants. For this purpose, two independent datasets were analyzed. The first one is issued from the IARC database and includes 144 families with certified LFS. The second LFS dataset was collected from 4 centers: the MD Anderson, The National Cancer Institute (NCI) LFS dataset, the Dana Farer Cancer Institute (DFCI) LFS dataset, and the Children's Hospital Of Philadelphia (CHOP) cancer predisposition program. The dataset contains 324 LFS families. The p53 founder variant LRG321t1:c. 1010G>A, (p.R337H) found predominantly in Brazil and included at high frequency in both datasets has been excluded.

[0076] In some embodiments, in the exemplary case of p53, step **204** includes a validation step based on data included in the ClinVar database with respect to 748 missense p53 variants. The p53 variants in ClinVar are graded on a pathology scale using a five-tier classification system recommended by the American College of Medical Genetics and Genomics (ACMG):

- Pathogenic (P)
- Likely pathogenic (LP)
- Uncertain significance (VUS)

- Likely benign (LB)
- Benign (B).

[0077] For the validation step, variants classified as P and LP by ClinVar were grouped and defined as D (deleterious), whereas variants classified as LB and B by ClinVar were defined as ND (Not deleterious).

[0078] With reference back to Fig. 2, in step **206**, the instructions of machine learning module **106b** may cause system **100** to construct one or more training datasets for training a machine learning model, based on the information received in step **202** and the features extracted and/or calculated in step **204**.

[0079] In the exemplary case of p53, a training dataset of the present disclosure may comprise the following data:

- 1294 variants of p53, comprising:
 - o 290 cancer-related (i.e., 'positive') variants.
 - o 1011 non cancer-related (i.e., 'negative') variants.

[0080] In the exemplary case of p53, the training dataset comprises, for each of the p53 variants, one or more of the features extracted and/or calculated in step **204**.

[0081] In the exemplary case of p53, the training dataset comprises, each of the p53 variants is annotated with a binary label denoting a pathogenicity of the variants, e.g., D/ND (indicating deleterious/non-deleterious), yes/no, 1/0, etc.

[0082] In some embodiments, the training dataset may be divided into a training portion, a validation portion, and a test portion. The training portion may be used to train one or more machine learning algorithms, to obtain a trained machine learning model of the present disclosure. The validation portion may be used to validate the prediction results various machine learning algorithms, to select the optimal algorithm for the task. The test portion may be ultimately used to test the resulting trained machine learning model.

[0083] Accordingly, in some embodiments, a validation dataset comprises multiple missense p53 variants with a minor allele frequency ranging from 10^{-6} to 0.8, identified from an analysis of 14 independent genomic databases from the human population. Of these variants, 41 are suspected to be either potential non-deleterious SNP or pathogenic

variants from asymptomatic individuals. In addition, variants from the ‘positive’ and the ‘negative’ training datasets are included in the validation dataset, based on a selection process which identified variants detected in three or more population datasets at frequency above $5 * 10^{-6}$ in at least one database.

[0084] With reference back to Fig. 2, in step **208**, the instructions of machine learning module **106b** may cause system **100** to train one or more machine learning models on the training dataset constructed in step **206**, to obtain a trained machine learning classifier **106c**.

[0085] In the exemplary case of p53, the instructions of machine learning module **106b** may cause system **100** to train machine learning classifier **106c** using various combinations of the features comprised in the training dataset constructed in step **206**, such as, but not limited to:

- All features included in the training dataset.
- Computational-based features only.
- Functional assay-based features only.

[0086] The instructions of machine learning module **106b** may cause system **100** to employ any suitable machine learning algorithm to create the present machine learning model. Examples of such machine learning algorithms may be Random Forest (RF) and Gradient Boosting Machine (GBM) algorithms.

[0087] In the exemplary case of p53, the present inventors trained several test machine learning models using different machine learning algorithms, such as RF and GBM, and training dataset compositions. The accuracy of these various trained models was tested on a validation dataset comprising 20% of the training dataset constructed in step **206**. The analyses performed using GBM and RF showed similar performances. GBM performed with 99.66% AUC when trained using all features included in the training dataset and using only the functional-based features, and with 97.64% AUC when trained using only the computational-based features. RF performed with 99.76% AUC when trained using only the functional-based features, 99.56% AUC when trained using all features included in the training dataset, and with 97.62% when trained using only the computational-based features.

[0088] To select the optimal machine learning algorithm, the present inventors performed 10 tuned runs for each of the RF and GBM algorithms, using variations that were alternately trained on all features included in the training dataset, and only the functional-based features. The results indicate that GBM-based models performed better than the RF models, by 1.2% for the model trained using all features included in the training dataset, and by 0.02% for the model trained using only the functional-based features. GBM also performed better than RF in accuracy, by 1.17% for the model trained using all features included in the training dataset, and by 3.1% for the model trained using only the functional-based features.

[0089] Accordingly, the present inventors selected to train the machine learning classifier **106c** using the GBM algorithm, with only the functional-based features included in the training dataset.

[0090] With reference back to Fig. 2, in step **210**, machine learning classifier **106c** trained in step **208** may be inferenced on genetic sequencing and related information from an unseen target variant of the gene of interest **120** (shown in Fig. 1), to output a prediction **122** with respect to a functional impact of the unseen target mutation variant of the gene of interest.

Experimental Results

[0091] Table 2 below summarizes the performance results of the GBM-based machine learning models trained according to method **200**, as detailed herein above. The results shown in Table 2 compare the prediction outputted by each model to the ‘ground truth’ pathogenicity of the tested sample, using ‘D’ to denote deleterious (or pathogenic), and ‘ND’ to denote non-deleterious (or non-pathogenic).

Table 2:

	All Features		Functional Features only		Computational Features only	
	Ground Truth					
Prediction	D	ND	D	ND	D	ND
D	55	6	54	27	53	2
ND	1	196	2	175	3	200

	All Features	Functional Features only	Computational Features only
AUC	99.66%	97.64%	99.66%
Accuracy	97.29%	88.76%	98.06%

Fig. 5A shows results on the test portion of the training dataset (see step **206** of method **200**), for the model trained using only the functional features (pie chart A), the model trained using only computational features (pie chart B), and for the model trained using all the features (pie chart C). In Fig. 5B, panel D is a Venn diagram showing discrepancies in the negative set of the three models and their intersections. Panel E is a Venn diagram similar to D, for the positive set. Fig. 5C shows ROC curves for the model trained using only the functional features (panel A), the model trained using only computational features (panel B), and for the model trained using all the features (panel C).

[0092] As shown in Fig. 5A, the functional features model had an AUC of 96.8% and an accuracy of 96.5% (pie chart A). The sensitivity was 92.8% and specificity was 97.5%. The performance of the other two models is presented in pie charts B and C in Fig. 5A. The discrepancy variants issued from the model's predictions on the test set (Fig. 5B, panels D-E. Machine learning classifier **106c** (trained using only the functional features) predicted 4 variants to be non-deleterious (i.e., functionally active), although they are included in the pathogenic training portion of the training dataset. A close examination of these four variants shows that they are indeed functionally impaired cancer-associated variants. Two of these variants (p.E224D and p.Q331H) are localized at the vicinity of exon/intron junction sequences and are well known to be associated with dysfunctional splicing and nonsense-mediated mRNA decay. As all functional scores are issued from experimental data based on forced expression of protein variants, functional data as well as in-silico protein function predictors used for these null variants will not be accurate, thus causing false negative scores. The third variant is localized in codon 181 (p.R181C) that has been shown to be essential with codon 180 for dimer stability. Variants at codon 181 such as p.R181C or p.R181H do not fully abolish p53 function and have differential loss of function depending on the p53 target genes. Germline variants of the form p.R181C were found to be a founder mutation associated with an increased risk of breast cancer in Arab families and is the only p53 variant that can be identified in a situation of

homozygosity, suggesting a low penetrance associated with a partial loss of the tumor suppressive function. The fourth variant, p.G334V, is located in the tetramerization domain and is well known to disrupt p53 oligomerization. Variants in this domain are difficult to interrogate in cellular assays as artificial overexpression alleviate this defect with forced oligomerization and a potent activity. Taken together, these four false negative variants are included in the CSD and are indeed pathogenic but cannot be accurately assessed via the present machine learning model. On the other hand, five variants labeled as functional (never detected in human cancer) were predicted to be deleterious (i.e., functionally impaired). Although multiple explanations can be considered, such as no sufficient loss of activity to impair the tumor suppressive effect of p53, it is also possible that these variants are counter selected in normal cells due to a toxic effect.

[0093] The present machine learning model was trained on selected variants rather than on randomly chosen ones, since it is possible to label only a subset of variants as pathogenic or non-pathogenic. To minimize biases, general and simple rules for the inclusion of variants in the training data sets were defined, as detailed above. Using a training set that was taken from the 1,294 labeled variants, rather than chosen randomly from all the 2,314 variants may result in a biased model that cannot predict accurately the remaining unlabeled 1,020 variants from the test portion of the training dataset. Hence, it is important to verify that the training set is not distinct from the test portion of the training dataset in its properties. Accordingly, a dimensionality reduction algorithm was applied to examine whether the training variants are dispersed evenly in the features space with all the rest of the variants and thus represent the whole mutational landscape. Multidimensional Scaling (MDS) reduced the 42 features into a two-dimensional space and was performed on variants used for the training as well as on the rest of variants from p53_UMD database.

[0094] Figs. 6A-6B show the results of the Multidimensional Scaling (MDS). Euclidean distance between each pairwise variants of the UMD database was calculated using all variant features. Fig. 6A shows MDS discerning the variants in the training set (orange) and the rest of the variants predicted by the present machine learning model (green). The two groups disperse evenly in the variants space. Fig. 6B shows MDS discerning deleterious (blue) vs. non-deleterious variants (yellow).

[0095] Fig. 6C shows variant frequency in UMD of the variants predicted by the algorithm. Panel C shows variants used for training in red and the variants from the test portion of the training dataset in turquoise. X-axis: p53 variants ranked according to their frequency in UMD from left to right. Y-axis: frequency of each variant in UMD (Log2 scale). Panel D in Fig. 6C is similar to panel C, with variants from the test portion of the training dataset only. The deleterious (D) and non-deleterious (ND) variants in this set present with a mixed frequency in UMD. Panel E in Fig. 6C is Similar to panel D, with variants from the training set only.

[0096] Variants used for training and the rest of the variants were presented with even distributions on the two MDS axes, supporting the assumption that these variants represent well the mutational landscape of the p53 gene (Fig. 6A). On the other hand, as expected, functionally inactive (deleterious) variants are more centered compared to active (non-deleterious) variants (Fig. 6B). Moreover, these findings allow grouping the labeling of the training set and the machine learning models prediction of the test portion of the training dataset into a unified approach to predict pathogenicity to all 2,314 missense variants of p53 gene. Fig. 6B, panels C-E show the machine learning model's predictions as related to variants frequency in the UMD_p53 database. As expected, pathogenic variants tend to have higher frequency compared to non-pathogenic variants. Nevertheless, the two groups cannot be separated accurately using frequency information alone— thus highlighting the importance of the predictive algorithm.

SNP Prediction and Experimental Validation

[0097] A validation of machine learning classifier **106c** was performed using the validation portion of the training dataset comprising 41 p53 variants ('set 41') constructed in step **206** of method **200**, as detailed hereinabove. This set comprises p53 variants found in population databases and suspected to be either potential benign SNPs, or pathogenic variants from asymptomatic individuals. Among the validation portion of the training dataset, 15 variants that were validated as SNPs, including the two most common polymorphisms p.P72R and p.P47S, were classified as non-deleterious (i.e., functional) by machine learning classifier **106c**. None of these variants has been described as deleterious or non-functional either in the three large scale functional datasets or in the MUTLOAD database. Furthermore, their Frequency of Multiple Mutations (MMF) and Germinal to Somatic (GVS) scores show that they were outliers

found at high frequency as germline variants or associated with other p53 variants in human tumors. Among the remaining 20 variants also classified as non-deleterious (i.e., functional), nine variants are likely SNPs found at low frequency in the human population, ten cannot be classified precisely and the remaining one is a well-known passenger variant (p.R175C). None of these variants display any loss of activity in the three large scale functional datasets and have high GVS and/or MMF scores. Finally, the six variants that were classified as deleterious (i.e., non-functional) by machine learning classifier **106c** include five bona fide cancer-associated variants found in multiple cancer patients.

[0098] Figs. 7A-7B show functional analysis of p53 variants. Fig. 7A shows representative photographs of colony formation assays. Plates were stained 2 weeks after transfection. Wild type p53 as well as p53 variants can inhibit colony formation whereas the cancer associated variant, p.R175H, used as a positive control, does not inhibit colony formation. Inclusion of p53 variants in the training or in the predictive set is indicated under the name of the variant as 'D' (deleterious), 'ND' (non-deleterious), or '-' (not included). For variants in the test set (20% of the training set) both labeling and prediction are given. Fig. 7B shows quantitative summary of the colony formation assay in Fig. 7A. Inclusion of p53 variants in the training or in the predictive set is shown under the name of the variant.

[0099] Growth arrest activity of p53 variants confirmed the prediction of machine learning classifier **106c** (Figs. 7A-7B). As expected, the variant at position p.R181C have an intermediated phenotype. Variant, p.G334R, located in the tetramerization domain shows a partial defect in the functional assay but displays an intact activity in all published multiple large-scale studies, stressing the difficulty to analyze variants in this particular domain. Nevertheless, this variant was recently shown to be a pathogenic, Ashkenazi Jewish–predominant mutation associated with a familial multiple cancer syndrome. Taken together, analysis of p53 variants from this set of 41 variants, as well as the experimental data, are in good agreement with the predictions of machine learning classifier **106c**.

[00100] The set of 41 variants was also used to compare the present model with other available scores. This set is enriched with non-deleterious variants. The predictions by machine learning classifier **106c** were compared with six scores that provided formal classification cutoff in the p53_UMD database (Polyphen2 HumVar, Polyphen2

HumDiv, Sift, Condel, Provean and Mutassessor. Machine learning classifier **106c** outperformed the other scores, with accuracy ranging between 97.5-100% (97.1% specificity, as this set is enriched for non-deleterious variants), where the second best score was Polyphen2 HumVar with accuracy ranging between 80.9-86.8% (80-83.9% specificity).

[00101] For further comparison, the six scores were also compared to the performance of machine learning classifier **106c** on the test set (n = 258, also enriched for non-deleterious variants) and on ClinVar's non-VUS variants set (n = 198, enriched for deleterious variants). On the test set, machine learning classifier **106c** maintained its advantage with 96.5% accuracy (97.5% specificity), where Condel performed second best with 81.8% (76.9% specificity).

[00102] On the ClinVar set, which is highly enriched with deleterious variants, machine learning classifier **106c** maintained best performance of 94.8% accuracy, with SIFT presenting a similar score. The rest of the scores also performed relatively well on ClinVar (accuracy ranging between 85.7-93.3%). This boldens emphasize the fact that most models perform well on classifying deleterious variants (sensitivity) but perform poorly for the classification of non-deleterious variants (specificity). Machine learning classifier **106c** by contrast performs well for both deleterious and non-deleterious variants.

Testing the Model on Population Data

[00103] Machine learning classifier **106c** was applied on gnomAD, the largest set of genetic variations found in the human population. Although gnomAD includes mostly benign variants, recent studies indicate that it also contains pathogenic variants in tumor suppressor genes such as BRCA1 or p53. Indeed, 39 out of the 196 missense p53 variants included in gnomAD have been classified as deleterious by the algorithm, including 22 CSD variants.

[00104] Fig. 8 shows the frequency of 196 variants in the gnomAD and UMD databases. Bars showing frequency in gnomAD are colored in red. Bars showing frequency in UMD are colored by the model's prediction—blue for deleterious (D) and yellow for non-deleterious (ND) variants.

[00105] Frequency is represented in log2 scale. They are identified both in the complete database as well as the no-cancer versions of gnomAD, indicating that they are associated with asymptomatic individuals carrying pathogenic p53 variants. This high frequency is in accordance with the elevated (15-30%) number of p53 *de novo* mutations in the early onset cancer patients not associated with familial history. On the other hand, 157/196 variants were defined as non-deleterious, including the 15 p53 SNPs described above that were recently validated as bona fide SNPs. Although these include the three pathogenic variants described above (splice variants p.E224D and p.T125M and the Brazilian variant p.R337H, known to be associated with adrenocortical carcinoma and whose loss of activity have been difficult to appraise), the remaining variants are found at very low frequency both in gnomAD and UMD and are likely very infrequent private SNPs or sequencing errors.

Testing on Data of Li Fraumeni Syndrome Patients

[00106] The present inventors performed another validation using variants taken from datasets of Li Fraumeni Syndrome (LFS) patients that should include only pathogenic deleterious (i.e., non-functional) variants. The first cohort included 147 families (60 different p53 variants) included in the LFS dataset from the IARC database. Machine learning classifier **106c** classified 54 variants as deleterious or non-functional, including 24 variants that were not included in the training portion of the training dataset and are from the test portion of the training dataset. The 4 remaining variants predicted as non-deleterious or functional were low frequency variants without any obvious loss of activity. Overall, this led to a machine learning classifier **106c** prediction accuracy of 93.1% (54/58) after removing the two miss-identified SNPs.

[00107] The second cohort includes 77 p53 variants issued from 324 LFS families. Machine learning classifier **106c** classified 71 variants as D (63 from the training dataset and eight from the test portion of the training dataset). Among the six remaining variants predicted to be non-deleterious or functional, three were miss-identified SNP (p.I254V, p.R283C and p.N235S), two are pathogenic variants localized in the tetramerization domain of p53 and the remaining one is the pathogenic variant p.181C discussed in the previous section (model prediction accuracy is 95.9% after removing the three SNPs).

Testing on The ClinVar Database

[00108] The ClinVar database which is extensively used in genetic testing programs includes 748 p53 missense variants that were classified using the 5-tier classification system (pathogenic, likely pathogenic, uncertain significance, likely benign, or benign). First, matching ClinVar with the training dataset constructed in step **206** of method **200** detailed hereinabove, shows that the negative portion of the dataset includes only variants classified as benign (B), likely benign (LB) or VUS, whereas the positive portion of the dataset includes only variants defined as pathogenic(P), likely pathogenic (LP) or VUS, thus supporting our labeling selection procedure.

[00109] Machine learning classifier **106c** predicted that the 26 B or LB variants included in ClinVar are non-deleterious or functional (true negative 100%, no false positive). On the other hand, among the 164 P or LP variants, 157 are predicted to be deleterious or non-functional. Hence, the non-deleterious predictions are 95.73% sensitive, 100% specific, the accuracy is 96.32%, and the area under the curve (AUC) is 98.8%. Among the seven false negative variants, three were localized in the tetramerization domain of the protein.

[00110] Five hundred fifty-one ClinVar variants are annotated as “variants of unknown significance” (VUS), 223 (40%) of which were classified as deleterious or non-functional by machine learning classifier **106c**. Although the VUS status suggests an unresolved issue for these variants, the observation that 68 (30%) of them have been described as somatic variants in more than 100 independent studies is highly suggestive of a pathogenic class.

[00111] Figs. 9A-9B show ClinVar comparison with predictions by machine learning classifier **106c**. Fig. 9A, pie chart A, shows variants annotated by ClinVar as Benign (B), Likely benign (LB), Pathogenic (P) or Likely pathogenic (LP), compared with predictions output by machine learning classifier **106c** as deleterious (D) and non-deleterious (ND). Twenty-six variants were predicted as B or LB by ClinVar and as ND by machine learning classifier **106c** and are considered True Negative (TN), given in light green. 157 variants were predicted as P or LP by ClinVar and as D by machine learning classifier **106c** and are considered True Positive (TP), given in green. 7 variants were predicted as P or LP by ClinVar and as ND by machine learning classifier **106c** and are considered False Negative (FN), given in purple. All B or LB variants were predicted

ND by machine learning classifier **106c**, with no false positives. Fig. 9A, pie chart B shows similar results to pie chart A, with the addition of 551 variants annotated by ClinVar as Variants of Unknown Significance (VUS). The VUS variants make up 74.3% of the variants in the analysis. Fig. 9A, pie chart C, shows similar results to pie chart B, with VUS variants also separated by their respective prediction. 211 VUS variants were predicted as D, given in blue. 340 VUS variants were predicted as ND, given in yellow. Fig. 9B shows a Venn diagram showing discrepancies in machine learning classifier **106c** - ClinVar comparison. The discrepancies were similar for the functional and all-features models. The computational model showed better performance in this analysis, with four of the seven false negatives correctly classified, and a remaining discrepancy of three variants.

Testing the Model on Survival Information

[00112] The present inventors further tested machine learning classifier **106c** predictions against survival data, using pan-cancer tumor samples from TCGA. The samples were divided into four categories of p53 mutational status: (i) no mutation in the p53 gene, (ii) missense mutation in p53 predicted by machine learning classifier **106c** to be non-deleterious functional, (iii) missense mutation in p53 predicted by machine learning classifier **106c** to be deleterious or non-functional, (iv) Tumors with a truncating non-missense mutation in p53.

[00113] Patients with p53 mutations predicted to be non-deleterious or functional had longer survival times compared to patients with p53 mutations predicted to be deleterious or non-functional ($p=0.00188$) and compared to patients with truncating mutations ($p=0.00125$). Patients with p53 mutations predicted to be deleterious or non-functional by machine learning classifier **106c** had significantly shorter survival time as compared to patients with no p53 mutations ($p<2e-16$). There was no significant survival difference between patients with functional mutations and patients with no p53 mutations ($p=0.109$), nor between patients with non-functional mutations and patients with known truncating mutations ($p=0.759$). Interestingly – minority of mutations in the somatic database of TCGA were predicted to be non-deleterious or functional (66/3,335, 1.98%) and this seems to reflect that such mutations are not positively selected in cancer. These findings provide an independent validation that the algorithm's prediction has accurate

clinical implication. Similar analysis performed for the other models (based on computation features and based on all features) provided comparable results.

[00114] Fig. 10 shows the survival curve of tumors from TCGA database. TCGA tumor samples are presented in four groups, by their p53 mutational status: (i) samples with no p53 mutation (No p53) (green); (ii) missense p53 mutations predicted by machine learning classifier **106c** to be non-deleterious (ND) (yellow); (iii) missense p53 mutations predicted by machine learning classifier **106c** to be deleterious (D) (blue); and (iv) samples with a truncating p53 mutation (Truncating) (purple). P-values for the comparison between these groups are also shown. Survival predictions are distinct when comparing No p53 and D ($p < 2e-16$), comparing ND and D ($p = 0.00188$) and comparing ND and Truncating ($p = 0.00125$). Survival predictions are indistinctive for comparisons between D and Truncating ($p = 0.759$) and between ND and No p53 ($p = 0.1$).

[00115] In some embodiments, the present disclosure provides for a predictive machine learning model configured to predict the loss of function of every possible missense variant in p53 gene.

[00116] The present invention may be a system, a method, and/or a computer program product. The computer program product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.

[00117] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as

radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire. Rather, the computer readable storage medium is a non-transient (i.e., not-volatile) medium.

[00118] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[00119] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-setting data, or either source code or object code written in any combination of one or more programming languages, including an object-oriented programming language such as Java, Smalltalk, C++ or the like, and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, a field-programmable gate array (FPGA), or a programmable logic array (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention. In some embodiments, electronic

circuitry including, for example, an application-specific integrated circuit (ASIC), may be incorporate the computer readable program instructions already at time of fabrication, such that the ASIC is configured to execute these instructions without programming.

[00120] Aspects of the present invention are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[00121] These computer readable program instructions may be provided to a processor of a general-purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[00122] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer-implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[00123] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module,

segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[00124] In the description and claims, each of the terms “substantially,” “essentially,” and forms thereof, when describing a numerical value, means up to a 20% deviation (namely, $\pm 20\%$) from that value. Similarly, when such a term describes a numerical range, it means up to a 20% broader range – 10% over that explicit range and 10% below it).

[00125] In the description, any given numerical range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range, such that each such subrange and individual numerical value constitutes an embodiment of the invention. This applies regardless of the breadth of the range. For example, description of a range of integers from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6, etc., as well as individual numbers within that range, for example, 1, 4, and 6. Similarly, description of a range of fractions, for example from 0.6 to 1.1, should be considered to have specifically disclosed subranges such as from 0.6 to 0.9, from 0.7 to 1.1, from 0.9 to 1, from 0.8 to 0.9, from 0.6 to 1.1, from 1 to 1.1 etc., as well as individual numbers within that range, for example 0.7, 1, and 1.1.

[00126] The descriptions of the various embodiments of the present invention have been presented for purposes of illustration, but are not intended to be exhaustive or limited to the explicit descriptions. Many modifications and variations will be apparent to those of ordinary skill in the art without departing from the scope and spirit of the described embodiments. The terminology used herein was chosen to best explain the principles of the embodiments, the practical application or technical improvement over technologies found in the marketplace, or to enable others of ordinary skill in the art to understand the embodiments disclosed herein.

[00127] In the description and claims of the application, each of the words “comprise,” “include,” and “have,” as well as forms thereof, are not necessarily limited to members in a list with which the words may be associated.

[00128] Where there are inconsistencies between the description and any document incorporated by reference or otherwise relied upon, it is intended that the present description controls.

CLAIMS

What is claimed is:

1. A system comprising:
 - at least one hardware processor; and
 - a non-transitory computer-readable storage medium having stored thereon program instructions, the program instructions executable by the at least one hardware processor to:
 - receive genetic information with respect to a plurality of known variants of a gene of interest,
 - calculate, based on said received genetic information, for each of said variants, a set of features representing a functional estimation or a loss of function (LOF) associated with said variant,
 - at a training stage, train a machine learning model on a training dataset comprising:
 - (i) all of said sets of features with respect to said variants, and
 - (ii) labels indicating a pathogenicity associated with each of said variants,
 - and
 - at an inference stage, apply said trained machine learning model to genetic information from an unseen target variant of said gene of interest, to predict a pathogenicity of said unseen target variant of said gene of interest.
2. The system of claim 1, wherein said gene of interest is tumor protein 53 (tp53).
3. The system of any one of claims 1 or 2, wherein at least some of said features are obtained using functional assays performed with respect to said plurality of known variants.
4. The system of any one of claims 1-3, wherein said training dataset comprises only said features obtained using functional assays performed with respect to said plurality of known variants.
5. The system of any one of claims 1-3, wherein at least some of said features are calculated using computational methods.

6. The system of any one of claims 1-5, wherein said plurality of known variants comprises at least a first portion comprising variants having a number of occurrences in human cancer that is greater than one.

7. The system of any one of claims 1-6, wherein said plurality of known variants comprises at least a second portion comprising variants having a number of occurrences in human cancer equal to one or zero.

8. The system of any one of claims 1-7, wherein said labels indicating a pathogenicity are binary labels selected from the group consisting of: pathogenic and non-pathogenic.

9. The system of any one of claims 1-8, wherein said unseen target variant of said gene of interest is obtained from a biological sample collected from a subject of interest.

10. A computer-implemented method comprising:

receiving genetic information with respect to a plurality of known variants of a gene of interest;

calculating, based on said received genetic information, for each of said variants, a set of features representing a functional estimation or a loss of function (LOF) associated with said variant;

at a training stage, training a machine learning model on a training dataset comprising:

(i) all of said sets of features with respect to said variants, and

(ii) labels indicating a pathogenicity associated with each of said variants; and

at an inference stage, applying said trained machine learning model to genetic information from an unseen target variant of said gene of interest, to predict a pathogenicity of said unseen target variant of said gene of interest.

11. The computer-implemented method of claim 10, wherein said gene of interest is tumor protein 53 (tp53).

12. The computer-implemented method of any one of claims 10 or 11, wherein at least some of said features are obtained using functional assays performed with respect to said plurality of known variants.

13. The computer-implemented method of any one of claims 10-12, wherein said training dataset comprises only said features obtained using functional assays performed with respect to said plurality of known variants.

14. The computer-implemented method of any one of claims 10-12, wherein at least some of said features are calculated using computational methods.

15. The computer-implemented method of any one of claims 10-14, wherein said plurality of known variants comprises at least a first portion comprising variants having a number of occurrences in human cancer that is greater than one.

16. The computer-implemented method of any one of claims 10-15, wherein said plurality of known variants comprises at least a second portion comprising variants having a number of occurrences in human cancer equal to one or zero.

17. The computer-implemented method of any one of claims 10-16, wherein said labels indicating a pathogenicity are binary labels selected from the group consisting of: pathogenic and non-pathogenic.

18. The computer-implemented method of any one of claims 10-17, wherein said unseen target variant of said gene of interest is obtained from a biological sample collected from a subject of interest.

19. A computer program product comprising a non-transitory computer-readable storage medium having program instructions embodied therewith, the program instructions executable by at least one hardware processor to:

receive genetic information with respect to a plurality of known variants of a gene of interest;

calculate, based on said received genetic information, for each of said variants, a set of features representing a functional estimation or a loss of function (LOF) associated with said variant;

at a training stage, train a machine learning model on a training dataset comprising:

- (i) all of said sets of features with respect to said variants, and
- (ii) labels indicating a pathogenicity associated with each of said variants; and

at an inference stage, apply said trained machine learning model to genetic information from an unseen target variant of said gene of interest, to predict a pathogenicity of said unseen target variant of said gene of interest.

20. The computer program product of claim 19, wherein said gene of interest is tumor protein 53 (tp53).

21. The computer program product of any one of claims 19 or 20, wherein at least some of said features are obtained using functional assays performed with respect to said plurality of known variants.

22. The computer program product of any one of claims 19-21, wherein said training dataset comprises only said features obtained using functional assays performed with respect to said plurality of known variants.

23. The computer program product of any one of claims 19-21, wherein at least some of said features are calculated using computational methods.

24. The computer program product of any one of claims 19-23, wherein said plurality of known variants comprises at least a first portion comprising variants having a number of occurrences in human cancer that is greater than one.

25. The computer program product of any one of claims 19-24, wherein said plurality of known variants comprises at least a second portion comprising variants having a number of occurrences in human cancer equal to one or zero.

26. The computer program product of any one of claims 19-25, wherein said labels indicating a pathogenicity are binary labels selected from the group consisting of: pathogenic and non-pathogenic.

27. The computer program product of any one of claims 19-26, wherein said unseen target variant of said gene of interest is obtained from a biological sample collected from a subject of interest.

28. A method comprising:

receiving genetic information from an unseen target variant of a gene of interest taken from a biological sample collected from a subject of interest; and

applying, to said genetic information from said unseen target variant, a machine learning model trained to predict a pathogenicity of variants of said gene of interest, to predict a pathogenicity of said unseen target variant of said gene of interest,

wherein a prediction of said unseen target variant of said gene of interest as pathogenic indicates a negative prognosis for said biological sample, and (ii) a prediction of said unseen target variant of said gene of interest as non-pathogenic indicates a positive prognosis for said biological sample.

29. A method of classifying a sample from a subject, comprising the steps of: (i) determining the sequence of a gene of interest; (ii) identifying an unseen target variant of said gene of interest; and (iii) applying the computer-implemented method of claim 10 to determine the pathogenicity of said unseen target variant, wherein the presence of a (a) pathogenic unseen target variant of said gene of interest indicates a negative prognosis for said sample, and (b) non-pathogenic unseen target variant of said gene of interest indicates a positive prognosis for said sample.

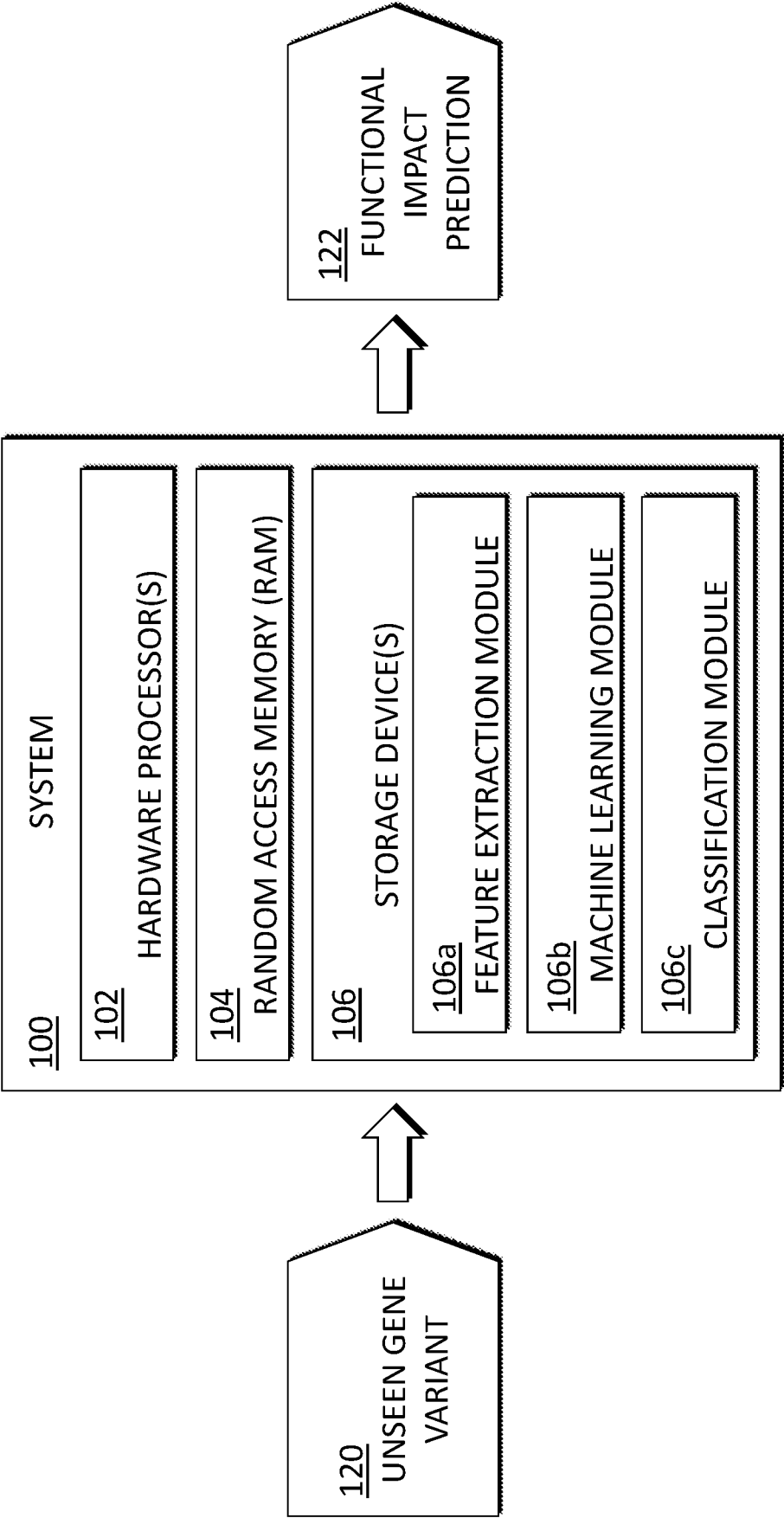


FIG. 1

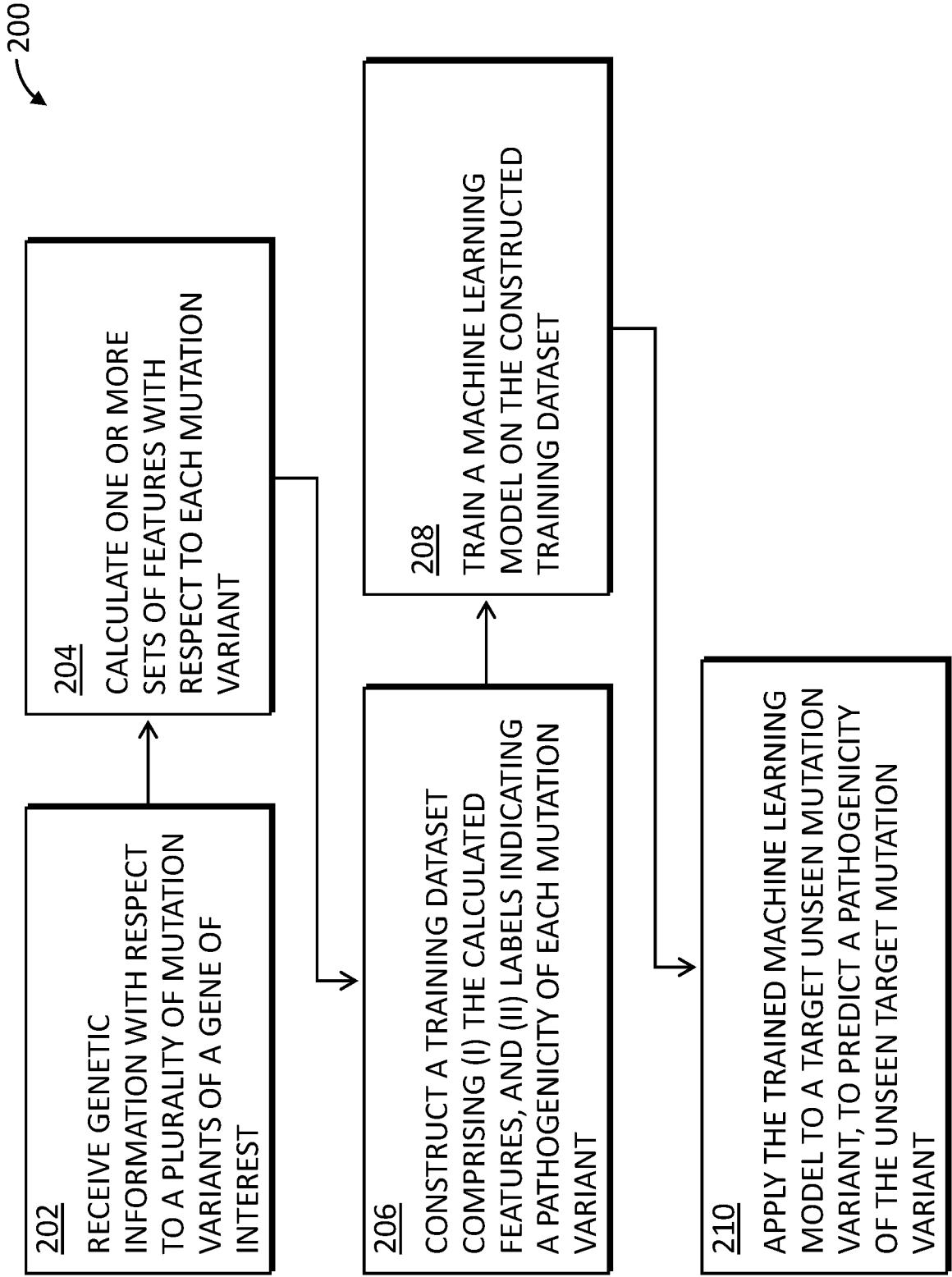


FIG. 2

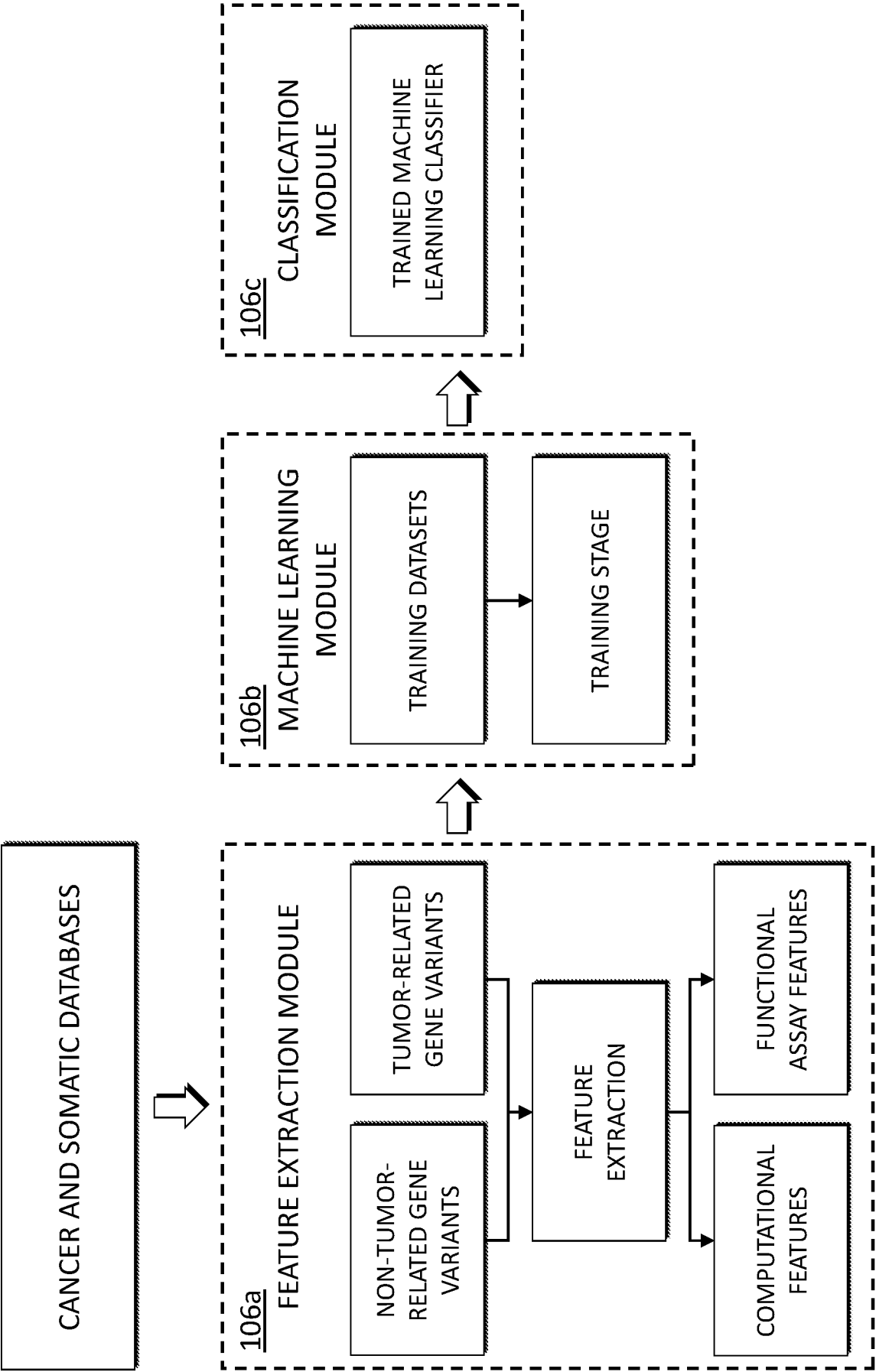


FIG. 3

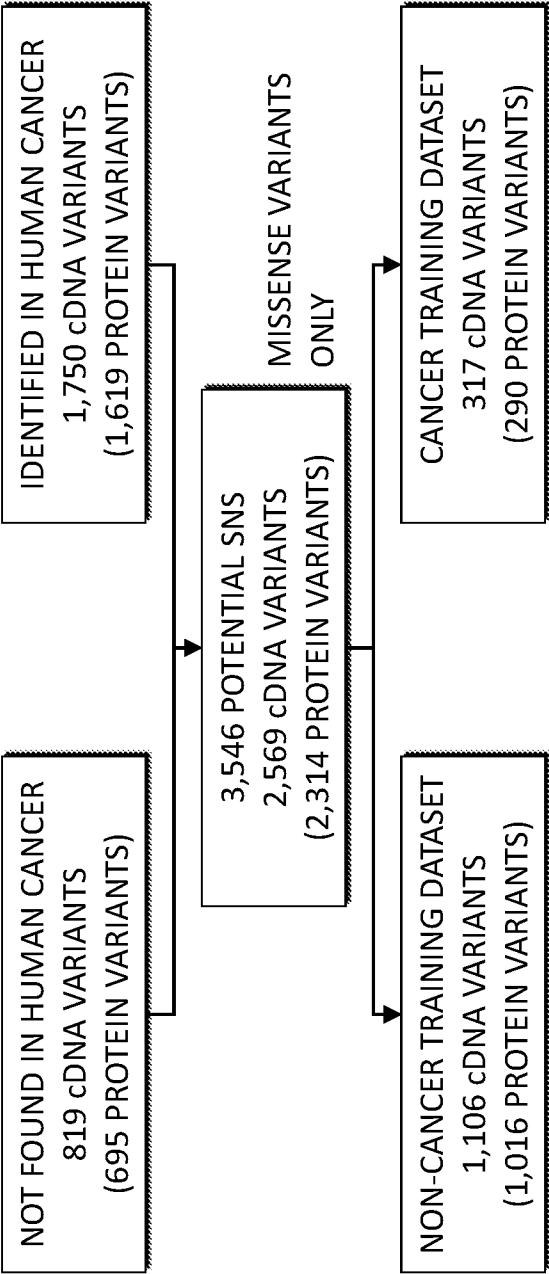


FIG. 4A

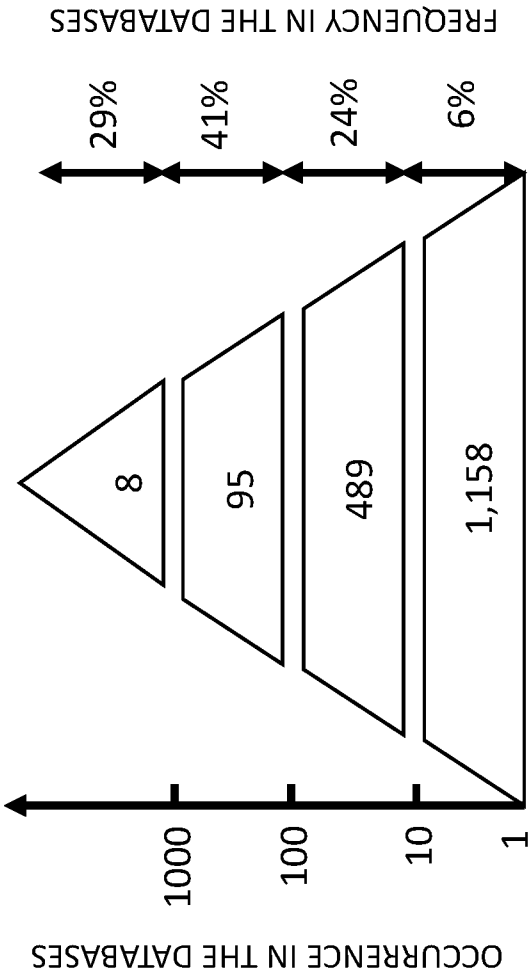


FIG. 4B

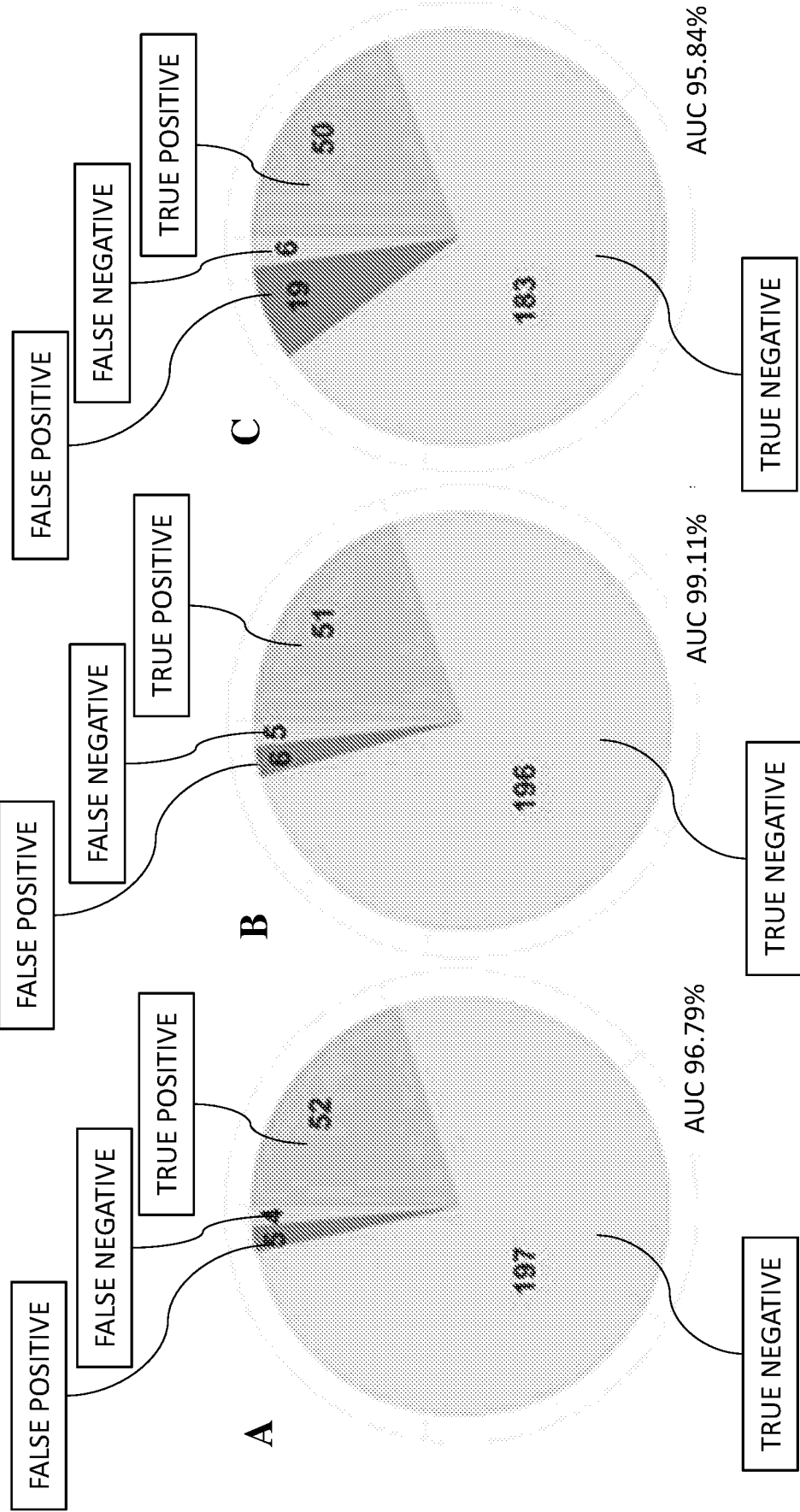


FIG. 5A

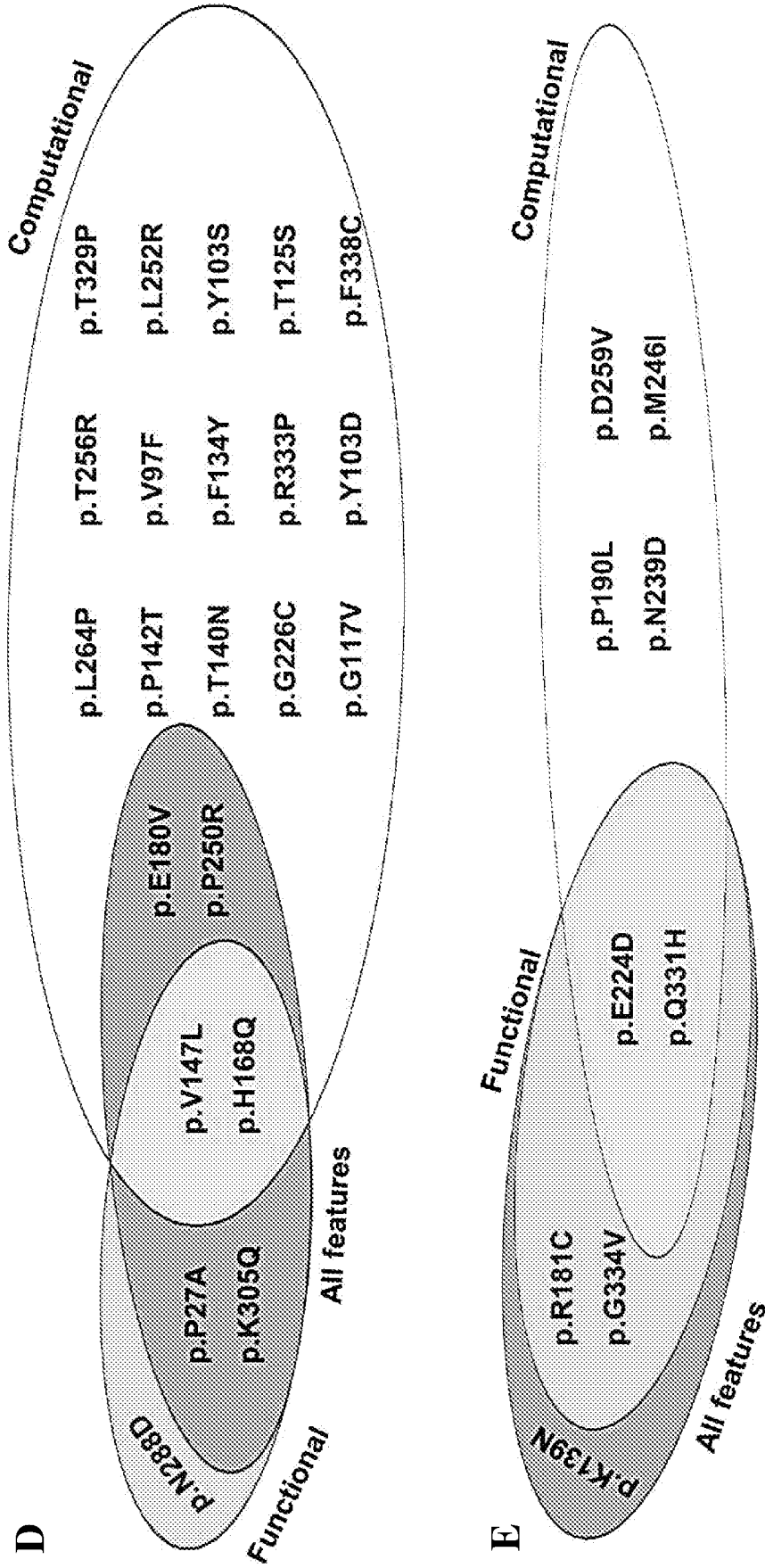


FIG. 5B

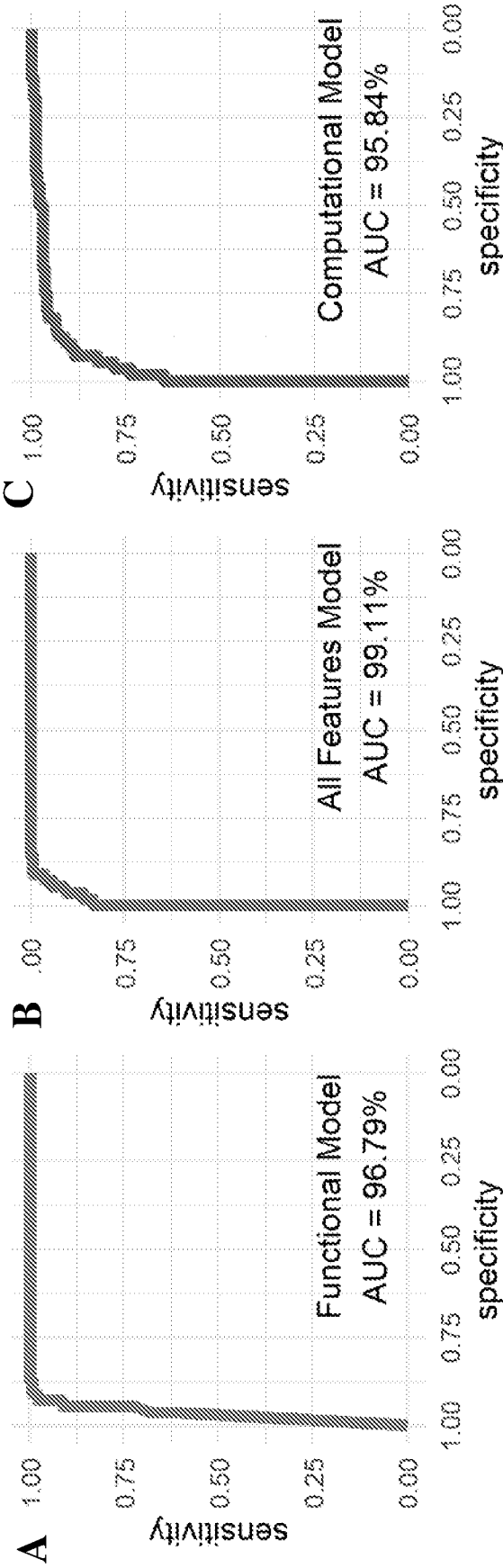


FIG. 5C

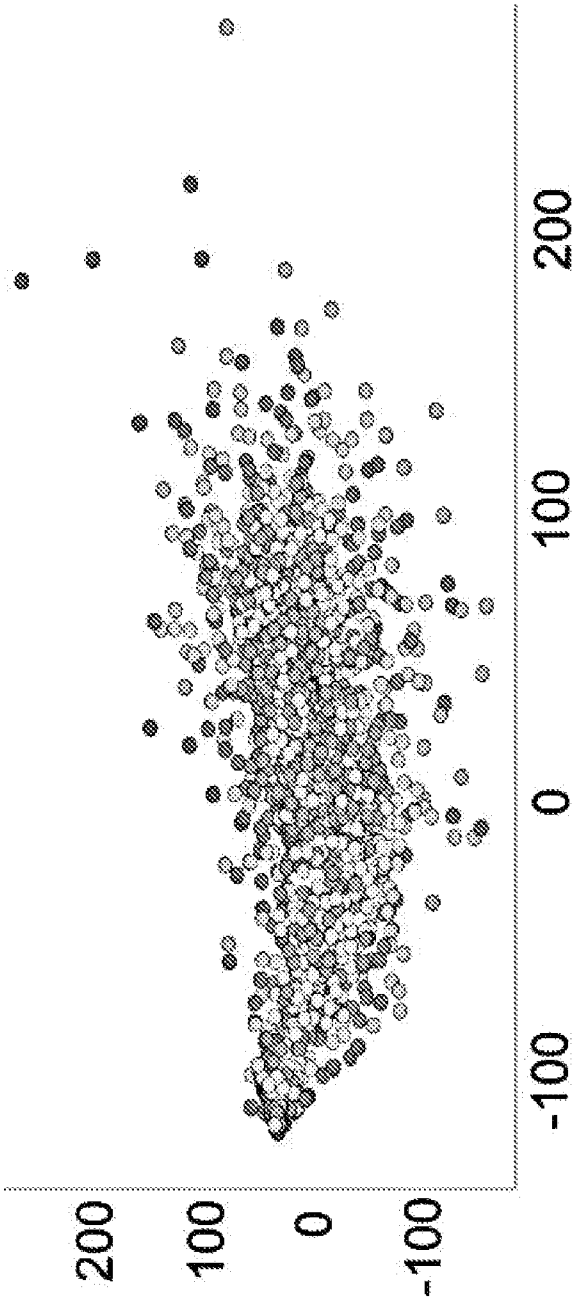


FIG. 6A

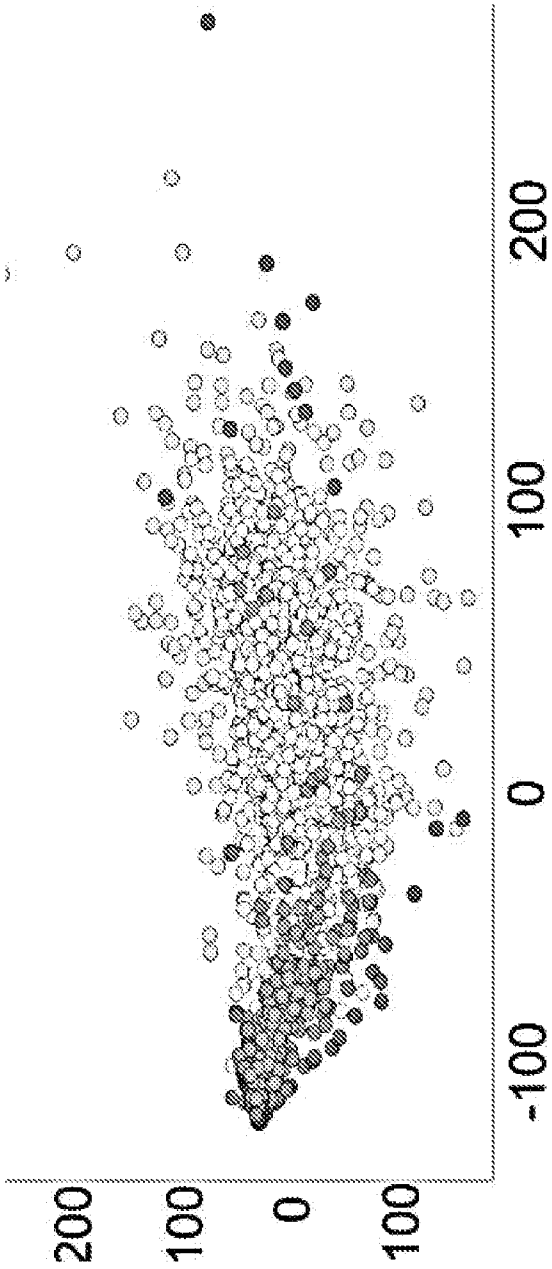
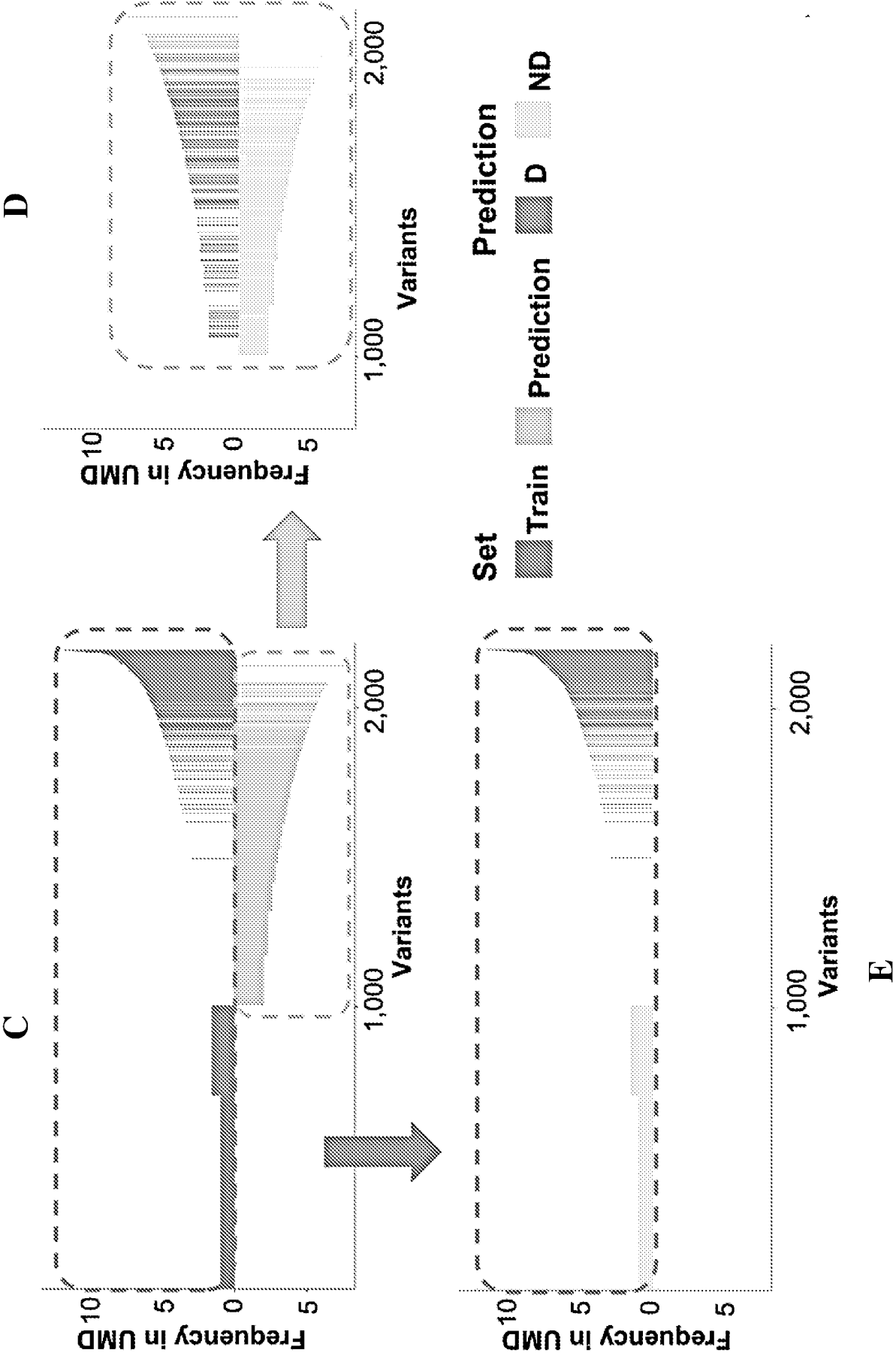


FIG. 6B



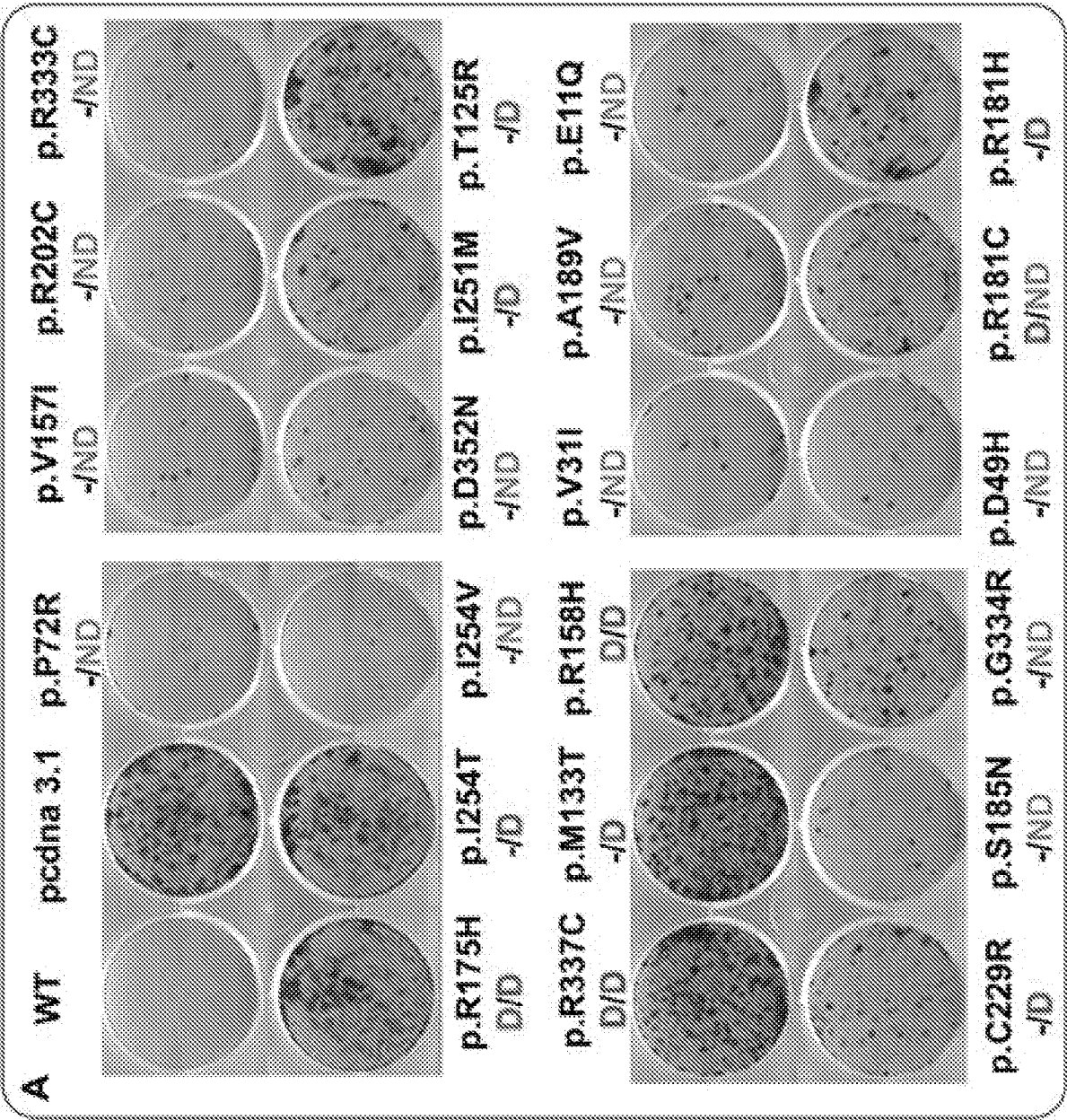


FIG. 7A

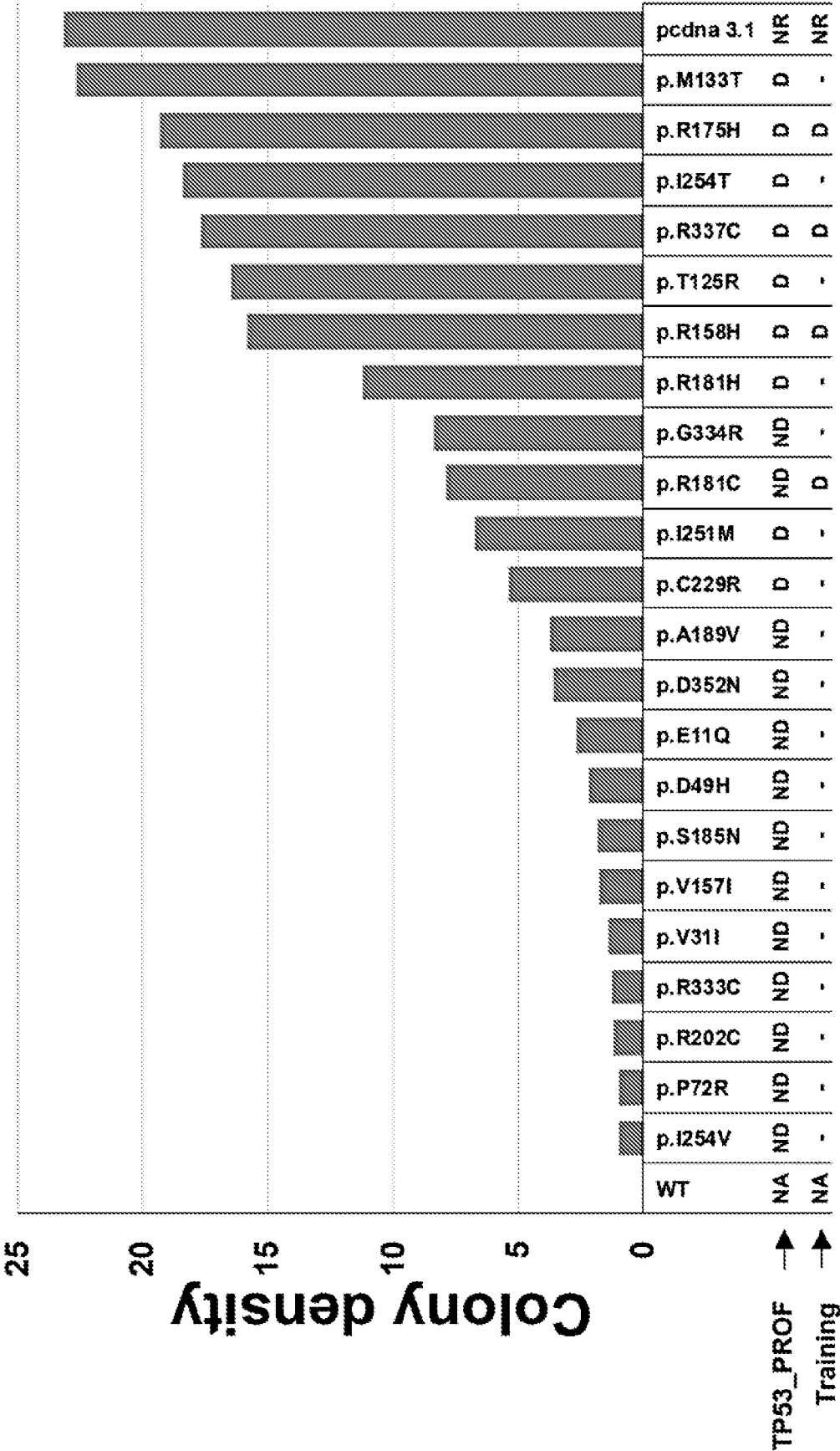


FIG. 7B

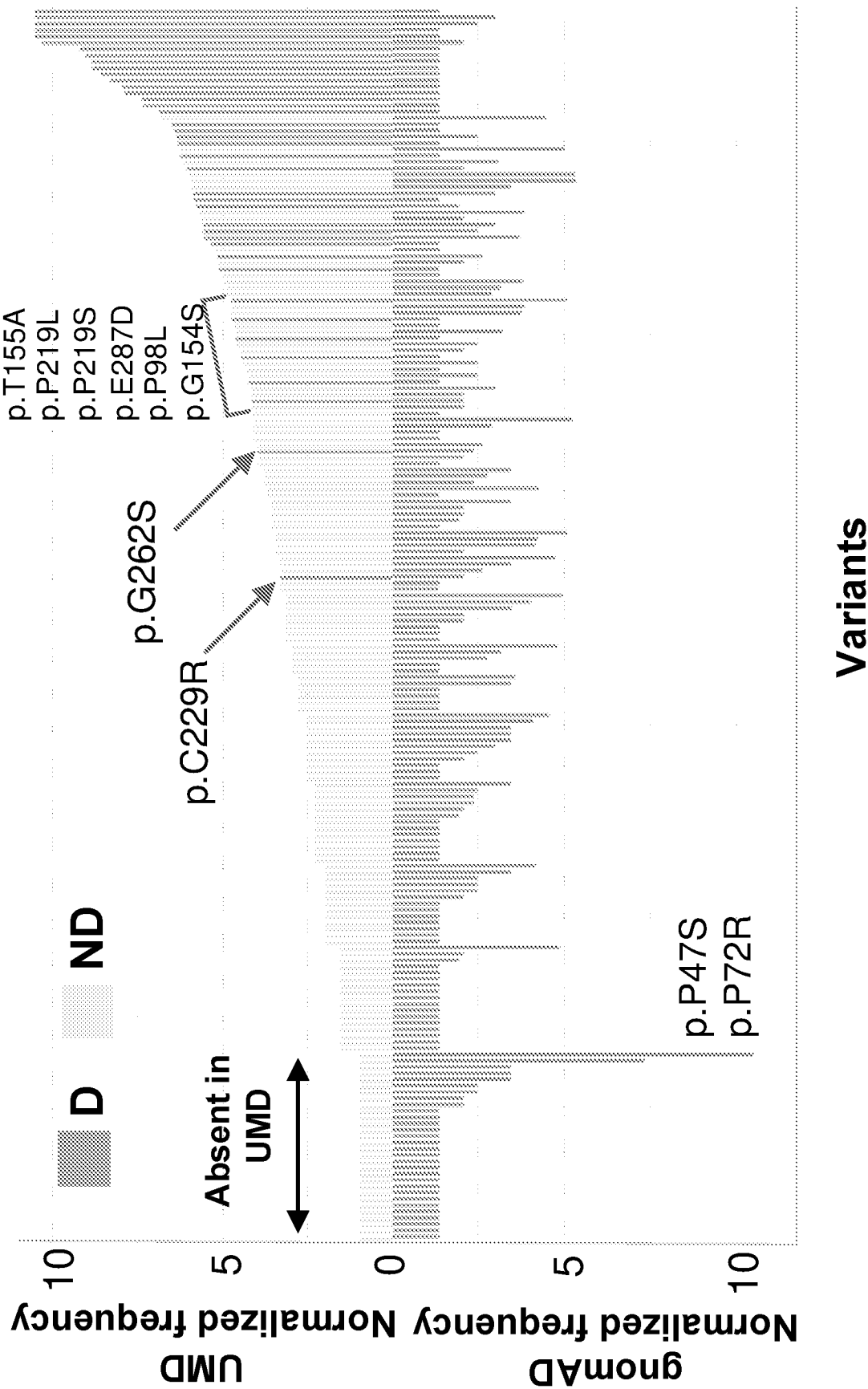


FIG. 8

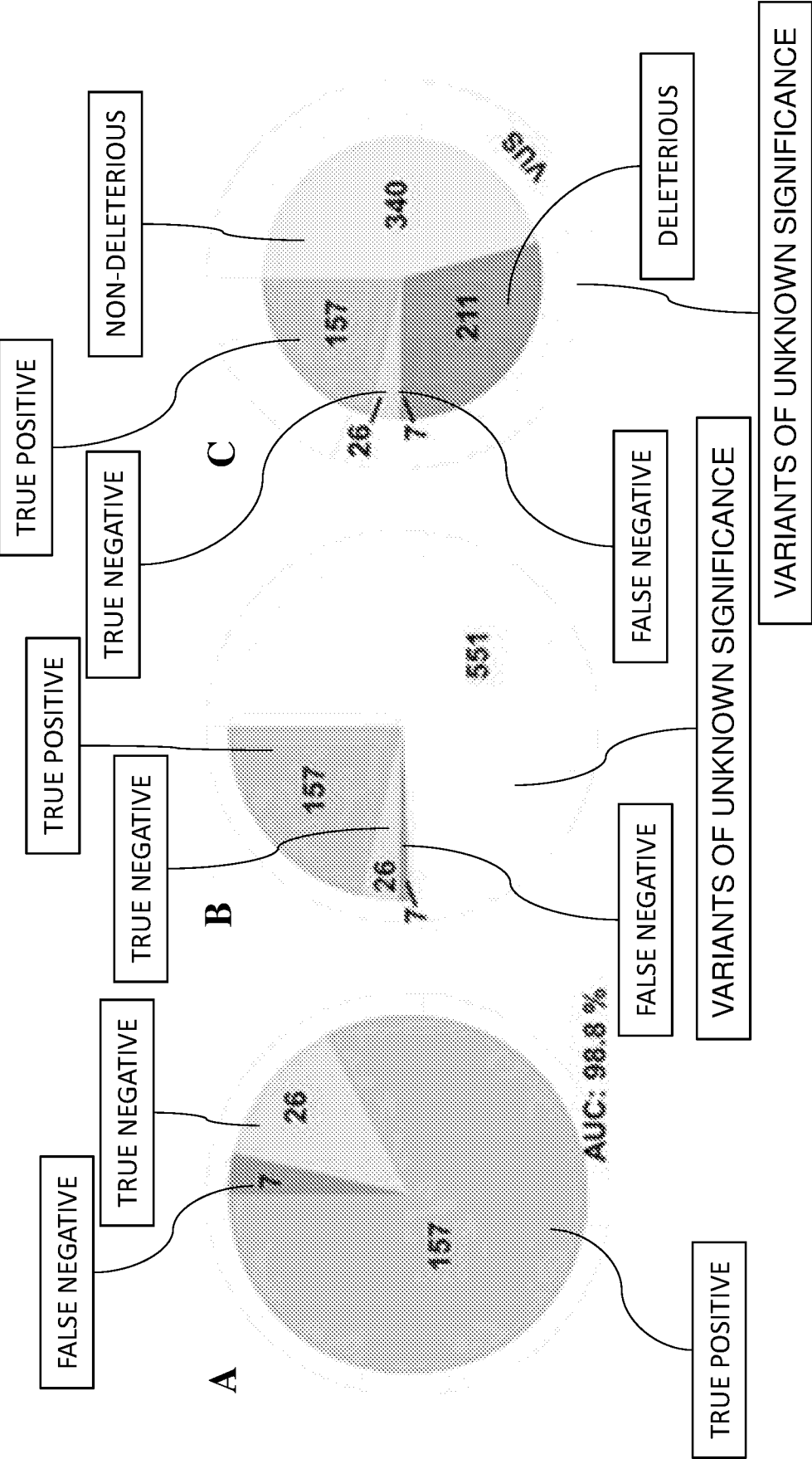


FIG. 9A

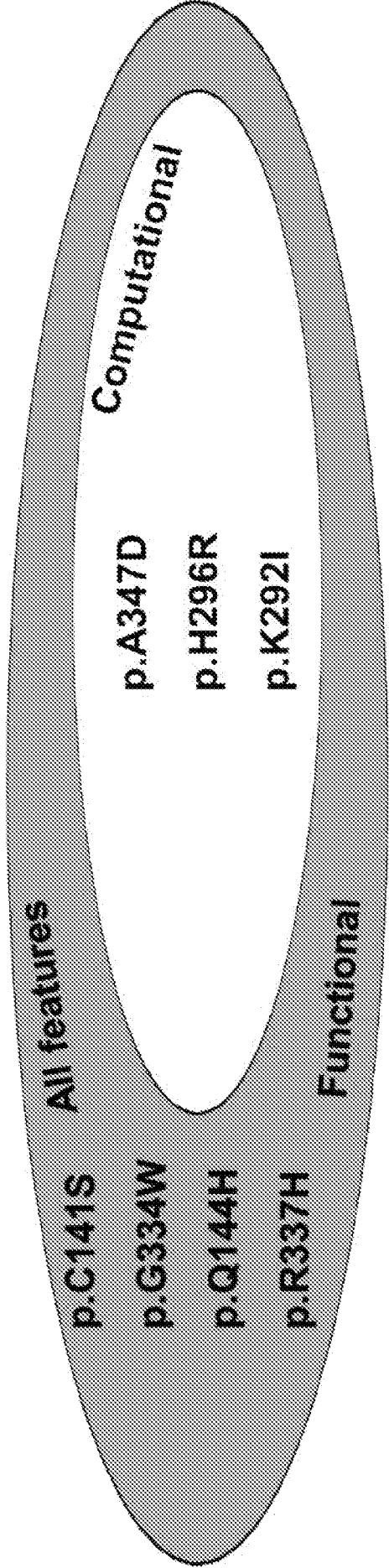


FIG. 9B

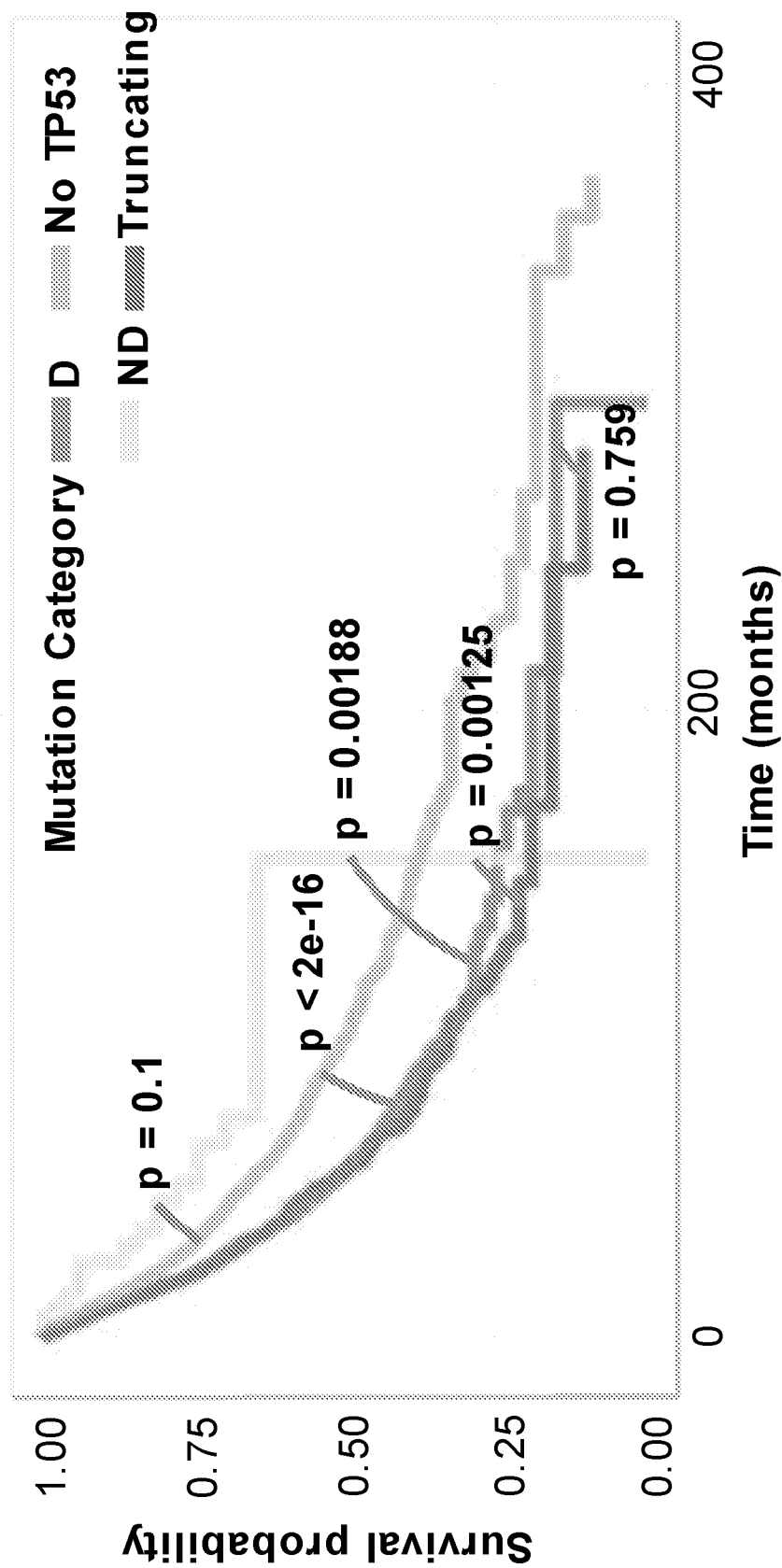


FIG. 10

INTERNATIONAL SEARCH REPORT

International application No
PCT/IL2022/051279

A. CLASSIFICATION OF SUBJECT MATTER INV. G16H50/20 G16B40/20 G16B20/20 ADD.		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED Minimum documentation searched (classification system followed by classification symbols) G16B G16H		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used) EPO-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	Kamada Mayumi ET AL: "Network-based pathogenicity prediction for variants of uncertain significance", bioRxiv, 16 July 2021 (2021-07-16), XP055872406, DOI: 10.1101/2021.07.15.452566 Retrieved from the Internet: URL:https://www.biorxiv.org/content/10.1101/2021.07.15.452566v1.full.pdf [retrieved on 2021-12-13] the whole document in particular page 4 – page 7; figures 1-4; tables 1-2 page 12 – page 18 ----- -/--	1-29
☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
13 February 2023		22/02/2023
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Rákossy, Z

INTERNATIONAL SEARCH REPORT

International application No
PCT/IL2022/051279

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2021/057042 A1 (BEAUBIER NIKE T [US] ET AL) 25 February 2021 (2021-02-25) the whole document in particular paragraph [0357] - paragraph [0383]; figures 2a, 5 -----	1-29
X	US 2020/342955 A1 (GUO XIN [US] ET AL) 29 October 2020 (2020-10-29) the whole document in particular paragraph [0087] - paragraph [0111]; figure 2B -----	1-29
X,P	BEN-COHEN GIL ET AL: "TP53_PROF: a machine learning model to predict impact of missense mutations in TP53", BRIEFINGS IN BIOINFORMATICS, [Online] vol. 23, no. 2, 18 January 2022 (2022-01-18), pages 1-19, XP93022971, GB ISSN: 1467-5463, DOI: 10.1093/bib/bbab524 Retrieved from the Internet: URL:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8921628/pdf/bbab524.pdf> [retrieved on 2023-02-10] the whole document -----	1-29

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/IL2022/051279

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2021057042 A1	25-02-2021	AU 2020334901 A1	24-03-2022
		CA 3148023 A1	25-02-2021
		EP 4013866 A1	22-06-2022
		JP 2022544604 A	19-10-2022
		US 2021057042 A1	25-02-2021
		US 2022351805 A1	03-11-2022
		WO 2021034712 A1	25-02-2021

US 2020342955 A1	29-10-2020	EP 3704640 A1	09-09-2020
		US 2020342955 A1	29-10-2020
		WO 2019084559 A1	02-05-2019
