LONGITUDINAL SAMPLE SETS AND METHODS OF MAKING AND USING THE SAME

CROSS-REFERENCE TO RELATED APPLICATION

[0001] Pursuant to 35 U.S.C. § 119(e), this application claims priority to the filing date of U.S. Provisional Application Serial No. 63/567,746 filed on March 20, 2024; the disclosure of which application is herein incorporated by reference.

FIELD OF THE INVENTION

[0002] This disclosure relates generally to methods, systems and libraries for carrying out analyte analysis, e.g. genomic or proteomic analysis.

BACKGROUND OF THE INVENTION

[0003] Blood plasma, the liquid component of blood, is a valuable source for analyte detection in mammals. Plasma can be used in clinical laboratories and research to identify pathophysiological conditions, monitor drug concentrations, and identify biomarkers associated with disease presence and/or progression. Plasma contains a wide array of proteins, metabolites, and nucleic acids, making it a rich source of circulating biomarkers for detecting diseases like inflammatory conditions, autoimmune disorders, neurodegenerative disorders, fetal abnormalities and cancer in a clinical setting. Plasma can be used in research, *e.g.,* to identify diagnostic and therapeutic targets for various diseases and disorders.

[0004] Plasma has certain advantages over the use of whole blood or serum for detection of analytes. Plasma is versatile, and suitable as a detection matrix for a wide range of tests, including DNA and RNA analysis, proteomics, metabolomics, and coagulation studies. For example, using plasma instead of whole blood minimizes interference of cells in detection methods, improving the accuracy and sensitivity of analyte detection. Plasma also generally displays greater stability than other cell-free blood products, *e.g.,* serum, making it a preferred choice for some analyses.

[0005] Plasma can be obtained using various methods, including centrifugation and use of microfluidic devices, which enable miniaturized and automated plasma separation, enhancing the speed and efficiency of analysis. Certain of these devices utilize capillary-

driven blood plasma separation, especially for point-of-care diagnostics. Plasmapheresis is another method for obtaining plasma from a subject, and although it has certain advantages, its utility for obtaining stable samples for long-term, longitudinal analyses remained in question.

[0006] The present disclosure addresses the utility of plasmapheresis-obtained plasma samples and provides methods, systems and libraries of plasmapheresis samples for analytical analyses.

SUMMARY OF THE INVENTION

[0007] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify all key or essential features of the claimed subject matter, nor is it intended to be used to limit the scope of the claimed subject matter.

[0008] The present disclosure provides libraries of plasmapheresis samples and databases comprising data obtained therefrom. Exemplary libraries of the disclosure include libraries comprising a plurality of longitudinal sample sets which include multiple samples obtained from the same donor over a period of time. *e.g.,* years. Also provided are datasets, data collections and systems obtained using these longitudinal sample sets.

[0009] Accordingly, the disclosure provides libraries of plasmapheresis samples, comprising a plurality of longitudinal sample sets, wherein a longitudinal sample set comprises multiple samples obtained from the same donor over a period of time, and wherein the samples are obtained by plasmapheresis.

[0010] In some aspects, the disclosure provides a method for identifying analytes that are associated with a phenotype of interest using a computer system that includes one or more processors and system memory, the method comprising selecting two or more longitudinal data sets in a database derived from two or more longitudinal sample sets, wherein the longitudinal sample sets comprise samples from a plurality of subjects determining one or more analyte scores from the plurality of analytes detected in the longitudinal data sets using the one or more processors, wherein the analyte scores are correlated with the phenotype of interest, and identifying one or more analytes associated with the phenotype of interest based on the analyte scores for each longitudinal data set.

[0011] In specific aspects, the analyte scores are generated using the one or more processors.

[0012] In specific aspects, selecting two or more longitudinal data sets in the database derived from two or more longitudinal sample sets is performed using the one or more processors. In specific aspects, the longitudinal sample sets to be queried are limited to samples collected over a designated timeframe, *e.g.,* six months, one year, five years, ten years, etc.

[0013] In some aspects, the data derived from the samples of the longitudinal sample sets are stored in the database in conjunction with additional information that can be associated with the sample donors. In specific aspects, the additional information in the database regarding the samples of the longitudinal sample sets is de-identified.

[0014] In some aspects, the disclosure provides a computer system, comprising one or more processors, a system memory; and one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed using the one or more processors, cause the computer system to implement a method for identifying genes that are associated with a disease-related phenotype of interest.

[0015] In some aspects, the disclosure provides a computer system, comprising one or more processors, a system memory; and one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed using the one or more processors, cause the computer system to implement a method for identifying longitudinal data sets that are associated with a phenotype of interest, the method including: (a) selecting, using the one or more processors, two or more longitudinal data sets in a database derived from two or more longitudinal sample sets, wherein the longitudinal sample sets comprise samples from a plurality of subjects collected over a designated time frame; (b) determining one or more analyte scores from the plurality of analytes detected in the longitudinal data sets using the one or more processors, wherein the analyte scores are correlated with the phenotype of interest; and (c) identifying one or more analytes associated with the phenotype of interest based on the analyte scores for each longitudinal data set using the one or more processors.

[0016] In some aspects, the method for identifying longitudinal data sets that are associated with a phenotype of interest in the systems further utilizes additional information in the database regarding the samples of the longitudinal sample sets. In specific aspects, the

method for identifying longitudinal data sets that are associated with a phenotype of interest in the systems further comprises additional information that can be associated with the sample donors. In specific aspects, the additional information in the database regarding the samples of the longitudinal sample sets is de-identified.

[0017] Other features, details, utilities, and advantages of the claimed subject matter will be apparent from the following written Detailed Description including those aspects illustrated in the accompanying drawings and defined in the appended claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0018] It should be understood that the drawings are not necessarily to scale, and that like reference numbers refer to like features.

[0019] Figure 1 depicts an example workflow used in the development of a model that characterizes the molecular profile of a given disease.

[0020] Figure 2 depicts a plot of the available plasmapheresis donations for each donor that developed a given disease. The dotted line represents the first date of diagnosis for the disease.

[0021] Figure 3 is a diagram depicting exemplary additional information that can be associated with the plasma donor's samples.

[0022] Figure 4 illustrates a process for de-identification that allows selected donor information to be associated with their respective samples.

[0023] Figure 5 depicts an example of correlating the health status information for a given donor to the molecular information that is contained in their plasmapheresis donation samples.

[0024] Figure 6 illustrates exemplary diseases that donors developed within the sample collection timeframe. The number of samples depicted as being above 100,000 are not drawn to scale *e.g.,* certain disorders shown have samples of greater than 300,000 samples in the collection.

[0025] Figure 7 depicts a tornado plot for Chronic Kidney Disease in the plasmapheresis biobank.

[0026] Figure 8 depicts a tornado plot for Type 1 Diabetes represented in the plasmapheresis biobank.

[0027] Figure 9 depicts a tornado plot for Type 2 Diabetes represented in the plasmapheresis biobank.

[0028] Figure 10 is a diagrammatic representation of a computer system that can be used with the methods and systems described herein.

[0029] Figures 11A and 11B are a set of components plots illustrating detectable age-related changes in 247 cross-sectional samples from the plasmapheresis biobank.

[0030] Figures 12A and 12B show illustrative examples of the detection of longitudinal molecular changes reflecting life events.

[0031] Figure 13 is a visualization of plasma samples in the plasmapheresis biobank for 1,497 donors who developed Parkinson's disease ("PD").

[0032] Figure 14 shows a funnel graph with the initial criteria to define the PD cohort from the plasmapheresis biobank for further study.

[0033] Figure 15 is a visualization of the plasma samples for 348 donors selected from the plasmapheresis biobank using the criteria of Figure 14.

[0034] Figure 16 summarizes the characteristics of the PD and control samples used in the longitudinal study of Examples 4 and 5.

[0035] Figures 17A and 17B show the samples collected with respect to clinical PD progression for PD patients and their respective control samples.

[0036] Figure 18 is a PLS-DA components plot for PD subjects and their matched control subjects following SomaScan™ analysis of plasmapheresis samples over a multi-year timeframe.

DETAILED DESCRIPTION OF EMBODIMENTS

[0036] The following examples merely illustrate particularly preferred methods and selected aspects in connection therewith. The following examples are not to be construed as limiting the scope of the claims.

[0037] In the following description, numerous specific details are set forth to provide a more thorough understanding of the present disclosure. However, it will be apparent to one of skill in the art that the present disclosure may be practiced without one or more of these specific details. In other instances, features and procedures well known to those skilled in the art have not been described in order to avoid obscuring the invention. The terms used herein are intended to have the plain and ordinary meaning as understood by those of ordinary skill in the art.

[0038] All of the functionalities described in connection with one aspect of the systems and/or methods described herein are intended to be applicable to the additional embodiments of the systems and/or methods except where expressly stated or where the feature or function is incompatible with the additional embodiments. For example, where a given feature or function is expressly described in connection with one aspect but not expressly mentioned in connection with an alternative embodiment, it should be understood that the feature or function may be deployed, utilized, or implemented in connection with the alternative embodiment unless the feature or function is incompatible with the alternative embodiment.

[0039] Note that as used herein and in the appended claims, the singular forms include plural referents unless the context clearly dictates otherwise. Thus, for example, reference to "a subject" may refer to one or more subjects (depending on context), and reference to "a system" includes reference to equivalent steps, methods and devices known to those skilled in the art, and so forth.

[0040] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this disclosure belongs. All publications mentioned herein are incorporated by reference for all purposes, including describing and disclosing devices, formulations and methodologies that may be used in connection with the presently described disclosure. Conventional methods are used for the procedures described herein, such as those provided in the art and demonstrated in the Examples and various general references. Unless otherwise stated, nucleic acid sequences described herein are given, when read from left to right, in the 5' to 3' direction. Nucleic acid sequences may be provided as DNA, as RNA, or a combination of DNA and RNA (*e.g.,* a chimeric nucleic acid) and may contain non-natural nucleotides and/or bases. Unless otherwise stated, protein sequences described herein are given, when read from left to right, in the N-terminal to C-terminal direction. Protein sequences may contain non-natural amino acids and may contain chemical modifications at specified locations within the sequence and with various degrees of modification.

[0041] The term "and/or" where used herein is to be taken as specific disclosure of each of the multiple specified features or components with or without another. Thus, the term "and/or" as used in a phrase such as "A and/or B" herein is intended to include "A and B," "A or B," "A" (alone), and "B" (alone). Likewise, the term "and/or" as used in a phrase such as "A, B, and/or

C" is intended to encompass each of the following embodiments: A, B, and C; A, B, or C; A or C; A or B; B or C; A and C; A and B; B and C; A (alone); B (alone); and C (alone).

[0042] As used herein, the term "about," as applied to one or more values of interest, refers to a value that falls within 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1%, or less in either direction (greater than or less than) of a stated reference value, unless otherwise stated or otherwise evident from the context (except where such number would exceed 100% of a possible value).

[0043] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (*e.g.,* bodies of the appended claims) are generally intended as "open" terms (*e.g.,* the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present. For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases "at least one" and "one or more" to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles "a" or "an" limits any particular claim containing such introduced claim recitation to embodiments containing only one such recitation, even when the same claim includes the introductory phrases "one or more" or "at least one" and indefinite articles such as "a" or "an" (*e.g.,* "a" and/or "an" should be interpreted to mean "at least one" or "one or more"); the same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (*e.g.,* the bare recitation of "two recitations," without other modifiers, means at least two recitations, or two or more recitations).

[0044] In addition, where features or aspects of the disclosure are described in terms of Markush groups, those skilled in the art will recognize that the disclosure is also thereby described in terms of any individual member or subgroup of members of the Markush group.

[0045] As will be understood by one skilled in the art, for any and all purposes, such as in terms of providing a written description, all ranges disclosed herein also encompass any and all possible sub-ranges and combinations of sub-ranges thereof. Any listed range can be easily

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

[0046] While the apparatus and method has or will be described for the sake of grammatical fluidity with functional explanations, it is to be expressly understood that the claims, unless expressly formulated under 35 U.S.C. §112, are not to be construed as necessarily limited in any way by the construction of "means" or "steps" limitations, but are to be accorded the full scope of the meaning and equivalents of the definition provided by the claims under the judicial doctrine of equivalents, and in the case where the claims are expressly formulated under 35 U.S.C. §112 are to be accorded full statutory equivalents under 35 U.S.C. §112.

The Invention in General

[0047] Plasmapheresis is an extracorporeal technique that involves the removal, treatment, and return or exchange of blood plasma or components thereof from and to the blood circulation. During a plasmapheresis session, the patient's blood is drawn from the subject and introduced at a peripheral or central access site into an apheresis system. Separation of the plasma from other blood components can be performed using two different major techniques: centrifugal separation and membrane separation (Ward, D.M. Conventional apheresis therapies: A review. *J. Clin. Apher.* 2011, *26*, 230–238).

[0048] In the centrifugal separation method, the blood is spun in an apheresis device and the different components are separated via specific gravity. In the membrane separation method, the blood crosses non-selective microporous membranes, which allow for the passage of molecules less than a certain weight and also allows for the retention of blood cells. After the plasma is separated from the blood cells, the blood cells are mixed with a liquid to replace the plasma and are returned to the subject from which the blood was drawn.

[0049] Prior to the experimental data presented in the present disclosure, it was unclear, however, whether plasmapheresis samples would be viable samples for various analyses, including genomic and proteomic analyses, due to the nature of the process and the resulting attributes and limitations of the plasma sample.

[0050] For example, plasmapheresis is used therapeutically to remove extra antibodies, abnormal proteins, or other harmful substances from the blood, and is routinely used in this manner to treat certain types of blood disorders, autoimmune disorders, nervous system disorders, kidney ailments or other conditions. Since such "therapeutic plasmapheresis" nonspecifically removes plasma molecules (including pathogenic molecules), it is unclear whether obtaining plasma samples using plasmapheresis would abnormally skew analyses using plasmapheresis sample due to inadvertent removal of target analytes.

[0051] The potential utility and reproducibility of using plasmapheresis samples for analyte detection was also unclear due to certain additives to the plasma needed for successful sample collection. During plasmapheresis, the foreign surface of the tubing of the apheresis device can activate platelets and the coagulation cascade, leading to clot formation. Anticoagulants like citrate and heparin are used to prevent clotting in the extracorporeal circuit, ensuring the blood remains fluid while being processed. Depending on the anticoagulant and device used, it was believed the addition of these anticoagulants could interfere with the detection and/or quantification of specific analytes in the plasmapheresis sample.

[0052] In addition, the volume of liquid obtained during the plasmapheresis process is quite large as compared to a venous whole blood draw, and the effect of the collection and storage in this larger volume was also a concern for use of plasmapheresis samples, as it potentially impacts the concentration of analytes in the sample.

[0053] The present disclosure provides a demonstration that plasmapheresis samples can indeed be effectively used for analyte analyses, *e.g.,* proteomic and genomic analysis, and that such analysis can reliably track analyte changes in subjects over time.

[0054] The present disclosure demonstrates that plasmapheresis samples obtained from donors in which the integrity of the plasma is maintained allow successful use of various analytical tools, and thus are essentially equivalent to plasma obtained using other methods, *e.g.,* venous withdrawal of whole blood followed by centrifugation. Specifically, the present disclosure

provides a library of plasmapheresis samples that can be successfully used for disease analysis and analyte detection.

Library and Plasmapheresis Biobank Samples

[0055] The libraries and plasmapheresis biobank of the disclosure include a plurality of distinct samples from multiple individuals over a designated timeframe. While the number of distinct samples in a given sample library of the invention may vary, in some instances the number of distinct samples in the library is 1,000,000 or more, such as 2,000,0000, 5,000,000 or more such as 10,000,0000 or more, including 100,000,000 or more. Two samples are "distinct" if they are different from each other, *e.g.,* in that they were obtained from a different donor or obtained from the same donor at a different time.

[0056] The distinct samples can be aggregated to form a large sample library that constitutes a "plasmapheresis biobank" of plasma samples. A "plasmapheresis biobank" is thus a collection of plasma samples that originates from at least 100,000, preferably 500,000, 1,000,000, 1,500,000, 2,000,000, 2,500,000, or 3M+ donors and offers a unique opportunity to study the preclinical and clinical phases of diseases over a multi-year period. Presently, Applicants have generated a plasmapheresis biobank that contains more than 80 million samples dating back to 2010. This establishes the Applicants' plasmapheresis biobank as an extensive collection of distinct samples from individuals providing a valuable tool for numerous medical uses including diagnostic, prognostic, therapeutic and drug development applications.

[0057] One advantage of plasmapheresis donation as compared to whole blood donation or plasma donation is that it can take place with a shorter time period between donations. For example, whole blood can generally be donated every 56 days, or about once every two months, and plasma donors can generally be donated every 28 days, samples obtained through apheresis may donate as often as twice a week, every 7 days, or up to 24 times in a 12-month period. This allows the ability to obtain more samples from a donor over a designated time frame, and provides for a more robust set of longitudinal samples from individual donors.

[0058] The plasmapheresis biobank includes stored data from a plurality of longitudinal sample sets obtained from individual donors. A "longitudinal sample set" includes multiple samples obtained from the same donor over a period of time. Generally, a given longitudinal

sample set is made up of a plurality of distinct samples that were obtained from the same donor over a date range. While the number of distinct samples that make up a given longitudinal sample set may vary, in some instances the number of distinct samples making up a given longitudinal sample set is 2 samples or more, 5 samples or more, 10 samples or more, 25 samples or more, or 50 samples or more. The period of time (i.e., date range) over which a given longitudinal sample set is collected may vary. In some instances, the period of time is 1 year or longer, such as 5 years or longer, including 10 years or longer, *e.g.,* 15 years or longer, such as 20 years or longer.

[0059] The data sets that are obtained from the longitudinal sample sets may include any data derived from the sets using various analytical methods, *e.g.,* various "omics" methods, that provide information on the presence and/or abundance of one or more analytes in the longitudinal sample sets. These longitudinal data sets can be used to determine one or more analyte scores (*e.g.,* the presence and/or quantitative data on the analyte from individual samples) from the plurality of analytes detected in the longitudinal data sets. These analyte scores can be generated, *e.g.,* using the computer systems as described in more detail herein, and can be correlated with various phenotypes based on the interrogation of the longitudinal data sets.

[0060] It is a distinct advantage of plasmapheresis donation that samples may be obtained in shorter timeframes than that usually required between whole blood sample donations. Accordingly, in some instances, a longitudinal sample set includes samples obtained from a donor with a minimal collection interval between any two samples of the set. In other words, the distinct samples making up a given longitudinal sample set are samples that were obtained from a donor with a shortened time duration, *e.g.,* a minimal interval, between when the samples were obtained from the donor. While the minimal interval may vary, in some aspects, the minimal interval is 2 days or longer, *e.g.,* 5 days or longer. Samples making up a given longitudinal sample set may be obtained from a donor according to a regular or irregular schedule.

[0061] Any two given longitudinal sample sets are considered "distinct" if they are made up of samples obtained from different donors, i.e., distinct individuals. While the number of distinct longitudinal sample sets present in a given sample library of the invention may vary, in some instances the number of distinct longitudinal sample sets present in a sample library is

100 or more, 500 or more, such as 1,000 or more, such as 10,000 or more, such as 100,000 or more, such as 500,000 or more, including 1,000,000 or more. In such instances, a sample library is made up of longitudinal sample sets obtained from 100 or more different donors, such as 500 or more different donors, such as 1,000 or more different donors, such as 10,000 or more different donors, such as 100,000 or more different donors, such as 500,000 or more different donors, including 1,000,000 or more different donors.

[0062] In some aspects, the plurality of longitudinal sample sets includes longitudinal sample sets obtained from a geographically diverse group of donors. While a geographically diverse group of donors may vary, geographically diverse groups of donors that may be represented in a sample library may be from different countries or different geographic regions of the same country.

[0063] Specific longitudinal sample sets of the disclosure are characterized in that at least the initial samples of a given set are obtained from a donor when the donor had no reported disease. By "at least initial" is meant at least the first sample (i.e., the starting sample obtained at the beginning of the time period) is obtained prior to a clinical diagnosis of the relevant disease or disorder, where in some instances "at least initial" refers to the first 5% or more, such as 10% or more, *e.g.,* 15% or more of the samples obtained from the donor.

[0064] In some instances, all or a majority of the samples of a given sample set are obtained from a donor when the donor had not reported clinical diagnosis of the specific disease state being analyzed. As such, in certain aspects 50% or more, 60% or more, 70% or more, 80% or more, 90% or more, 95% or more, or 100% of the samples making up a given longitudinal sample set may be obtained from a donor with no reported disease. Where samples are obtained from a donor with no clinical diagnosis of the relevant disease, such samples are obtained from a donor that, at the time the sample is collected, reported that they did suffer from any disease.

[0065] The plasmapheresis samples often include an added anti-coagulant, such as sodium citrate, EDTA, heparin and the like. When present, the added anti-coagulant may be present in an amount ranging from 2% to 4%, such as 68 mmol/L to 136 mmol/L citrate concentration. Sample donors may vary, wherein some instances the samples are obtained from mammalian donors, i.e., they are mammalian samples. As used herein, the terms "mammals" and "mammalian" are used broadly to describe organisms which are within the class mammalia, including the orders carnivore (*e.g.,* dogs and cats), rodentia (*e.g.,* mice,

guinea pigs, and rats), and primates (*e.g.,* humans, cynomolgus, chimpanzees, and monkeys). In specific aspects, the samples are human samples.

[0066] The volume of the samples contained within the sample library may vary, and may also differ from the total volume of the plasma obtained from the subject at collection. In some instances, the volume in the sample is 1 ml or greater, such as 2 ml or greater, where in some instances the volume ranges from 1 mL to 10 mL, such as 1 to 3 mL. These volumes can be a fraction of the total volume obtained from a subject through plasmapheresis (*e.g.,* an aliquot from a larger collected sample of approximately 750ml). The samples may be present in any suitable container, *e.g.,* vials, bottles and the like. In a specific aspect, the container is tube that can hold up to 5mL of a liquid such as a liquid sample.

[0067] In some aspects, the longitudinal sample sets are stable under long-term storage stable. By storage stable is meant that during the storage period, *e.g.,* 1 year or longer, 5 years or longer, 10 years or longer, 15 years or longer, or 20 years or longer under standard conditions the samples are both chemically and physically stable. In some instances, the samples may be frozen, *e.g.,* where the samples are maintained at a temperature of -20 °C or lower, such as -30 °C or lower, including -80 °C or lower. In another aspect, the samples may be frozen, *e.g.,* where the samples are maintained at a temperature of 0 °C or lower, such as -10 °C or lower, including -15 °C or lower.

[0068] The distinct samples of a library or biobank may be stored in a single location or separate locations, *e.g.,* samples do not need to be co-located to constitute a longitudinal sample set or library. In some aspects, the samples of the library or biobank are distributed among two or more different locations, i.e., some of the samples are present at first location while other samples are present at one or more locations different from the first.

Exemplary Work Flows

[0069] An exemplary work flow for use of the plasmapheresis samples in the libraries is illustrated in Figure 1. First, plasma is collected from donors via plasmapheresis and sample tubes are stored, preferably in long-term storage, *e.g.,* sub-zero temperatures (101). The individual plasmapheresis plasma samples from the individual donors are then matched (102) with related information through "tokens," as described in more detail herein, to assign samples to potential disease or health states in a de-identified manner. Analytes are extracted from the

individual samples and analyzed using the applicable techniques for identification and/or quantification of the analyte (103). For example, genomics or proteomics data can be generated from a specific sample to identify and/or quantify the analyte(s) of interest in the sample. A comparison between protein patterns before and after the onset of disease or between healthy cohorts of plasma donors and those with disease are made to identify analytes with clinical significance (104, 105). For example, actionable biomarker data can be generated to predict disease onset and inform targeted treatment strategies. This information is subsequently utilized in clinical setting, *e.g.,* disease diagnosis and/or therapeutic intervention (106).

[0070] The libraries of the present disclosure include available plasmapheresis donations for individual donors over a period of time which spans the identification, development and/or progression of a specific disease of interest. Figure 2 illustrates a "tornado plot" where the x-axis represents the time difference, in years, from prior to clinical diagnosis of a disease through the first clinical diagnosis and progression. The y-axis represents each donor per row that developed a disease of interest, ordered by time difference from last donation in the system to a first diagnosis code associated with the specific disease.

Information Linkage for Individual Samples

[0071] In certain aspects, at least some, if not all, of the longitudinal sample sets of a given library are linked to real world data of the individuals from which the samples were obtained. As such, for each longitudinal sample set in the library, the longitudinal sample set is linked with, i.e., associated with, data for the donor from which that longitudinal sample set was obtained.

[0072] Information of interest that can be linked to distinct samples includes, but is not limited to: clinical trial data, real world data (i.e., the data relating to donor health status and/or the delivery of health care routinely collected from a variety of sources, such as electronic health records, claims and billing activity, product and disease registries, social determinants of health, imaging data, medical notes data, and data gathered from other sources such as mobile devices, wearables (such as pedometers, heart rate monitors, smart watches, etc.), lifestyle data, consumer behavior data, internet usage data, geospatial data, mortality data, demographics data, and combinations thereof. In some instances, the donor data associated with a given

longitudinal sample set is de-identified, such that the identity of the donor is not associated with the longitudinal data set. De-identification may be achieved using any convenient protocol. Figure 3 illustrates examples of additional information that can be associated with the plasma donor's samples. Each plasma donation that comes from a donor can be linked through that donor to other information sources that were not collected at the time and location of donation.

[0073] In specific aspects, the donor data is tokenized. The process of "tokenization", when applied to data security, is the process of substituting a sensitive data element, *e.g.,* donor identity or other protected health information, with a non-sensitive equivalent, referred to as a token, that has no intrinsic or exploitable meaning or value. The token is a reference that maps back to the sensitive data through a tokenization system. Figure 4 shows a de-identification process that leads to tokens which allow donor information to be associated with their samples without needing to know any personally identifiable information (PII). The process shows two major components: a rules engine that removes or modifies data and a mathematical function that transforms the PII in a one-way manner (i.e., so that it cannot be reverse engineered) into a code that can be used as a token for data linking.

[0074] Figure 5 depicts a process for correlating the health status information for a given donor to the molecular information that is contained in their plasma donation samples. This particular example shows a donor that begins donating plasmapheresis samples while "healthy" but who eventually develops Parkinson's disease (PD) during the donation timeframe. The clinical management of the donor progresses toward an eventual PD diagnosis (represented by observations made on the "Diagnoses" track along with events on the "Procedures" track and the "Medication" track) where changes in the levels of proteins (represented by additional dots in the "Protein X level" track) may be associated with the progression of PD. The correlation of this information provides an opportunity to determine factors associated with the earliest signs of disease.

[0075] Figure 6 depicts a selection of diseases that donors in the longitudinal sample sets eventually developed during the timeframe of their sample collection. The y-axis indicates the number of donors that have a given disease diagnosis and the x-axis indicates broad indication classes that the diseases of interest may be categorized into. Many more diseases are present in the sample collection with greater than 25 thousand ICD-10 codes

(International Classification of Diseases or "ICD") that have 50 or more donors linked with each code. Figures 7-9 depict Tornado plots for several other diseases represented in the sample collection (Chronic Kidney Disease, Type 1 Diabetes, and Type 2 Diabetes, respectively). These plots show different levels of coverage and the identification and progression of the particular disease over the timeframe in which individual samples were collected.

Multi-omics analysis

[0076] The samples present in the libraries and plasmapheresis biobank can be subjected to various analyses that determine the presence and/or abundance of analytes in the distinct samples of the longitudinal samples sets. Exemplary analyte detection techniques depend on the nature of the analyte to be detected, and can include various "omic" assays that are tailored to the identification of one or more class of analyte, *e.g.,* DNA, RNA, proteins, and the like.

[0077] The selection of which "omic" assays to use will depend upon the nature of the analyte sought as well as the clinical or research goal of the system developed. In various examples, biological assays are performed on different portions of a sample to provide an analyte data set corresponding to the biological assay for various analytes. Various assays are known to those of skill in the art and are useful to interrogate a sample.

[0078] As used herein the term "assay" includes known biological assays and may also include computational biology approaches for transforming biological information into useful inputs for systems analysis and modeling. Various pre-processing computational tools may be included with the assays described herein and the term "assay" is not intended to be limiting to solely physical aspects of the molecular techniques.

[0079] Examples of such assays include but are not limited to: whole-genome sequencing (WGS), whole-genome bisulfite sequencing (WGSB), small-RNA sequencing, quantitative immunoassay, enzyme-linked immunosorbent assay (ELISA), proximity extension assay (PEA), protein microarray, mass spectrometry, low-coverage Whole-Genome Sequencing (lcWGS); selective tagging 5mC sequencing (WO2019/051484), CNV calling; tumor fraction (TF) estimation; Whole Genome Bisulfite Sequencing; LINE-1 CpG methylation; 56 genes CpG methylation; cf-Protein Immuno-Quant ELISAs, SIMOA; and cf-miRNA

sequencing, and cell type or cell phenotype mixture proportions derived from any of the above assays. This ability to analyze multiple analytes (such as but not limited to DNA, RNA, proteins, autoantibodies, metabolites, or combinations thereof) simultaneously from the same sample, or fractions thereof can increase the sensitivity and specificity of such bodily fluid tests by exploiting independent information between signals.

[0080] In one example, cell-free DNA (cfDNA) content is assessed by lcWGS or targeted sequencing, or whole-genome bisulfite sequencing (WGBS) or whole-genome enzymatic methyl sequencing, cell-free microRNA (cf-miRNA) is assessed by small-RNA sequencing or PCR (digital droplet or quantitative), and levels of circulating proteins are measured by quantitative immunoassay. In one example, cell-free DNA (cfDNA) content is assessed by whole-genome bisulfite sequencing (WGBS), proteins are measured by quantitative immunoassay (including ELISA or proximity extension assay), and autoantibodies are measured by protein microarrays.

[0081] Extracellular nucleic acid molecules include cell-free DNA (cfDNA) and cell-free RNA (cfRNA). These molecules are typically fragmented but resist full degradation in plasma because of protection from extracellular vesicles or binding proteins (*e.g.,* nucleosomes for cfDNA and ribonucleoproteins for cfRNA).

[0082] Many cfDNA features (*e.g.,* methylation pattern, mutation, copy number, fragment pattern, and nucleosome footprint) have been utilized for noninvasive assessments of disease diagnosis and prognosis. *See, e.g.,* van der Pol Y. and Mouliere F. *Cancer Cell.* 2019; 36:350–368; Jamshidi A. et al., *Cancer Cell.* 2022;40:1537–1549.e12. Shen S.Y. et al., *Nature.* 2018;563:579–583. Biological information may also include information regarding transcription start sites, transcription factor binding sites, nucleosomal positioning or occupancy, transposase-accessible chromatin using sequencing (ATAC-seq) data, histone marker data, DNAse hypersensitivity sites (DHSs), and the like. The plasma concentration of cfDNA may be assayed as an analyte that in various examples indicates the presence of a pathology.

[0083] Many cfRNA features can also serve as analytes; such features include the abundance of microRNAs (Zhou J. et al., *J. Clin. Oncol.* 2011;29:4781–4788) and circular RNAs (Wang S. et al., *Mol. Cancer.* 2021;20:13), fragment copies, and alternative splicing patterns of

mRNAs and long noncoding RNAs. Zhu Y. et al., *Theranostics*. 2021;11:181–193; Larson M.H. et al., *Nat. Commun.* 2021;12:2357.

[0084] In various examples, assays that profile the characteristics of cfDNA and/or cfRNA are used to generate signatures useful in the computational applications. In one example, characteristics of cf-DNA are used in machine learning models and to generate classifiers to stratify individuals or detect disease as described herein. Exemplary signatures include but are not limited to those that provide biological information regarding gene expression, 3D chromatin, chromatin states, copy number variants, tissue of origin and cell composition in cfDNA samples. Metrics of cfDNA concentration that may be used as input signatures for machine learning methods and models may be obtained by methods that include but are not limited to methods that quantitate dsDNA within specified size ranges (*e.g.,* Agilent TapeStation™, Bioanalyzer, Fragment Analyzer), methods that quantitate all dsDNA using dsDNA-binding dyes (*e.g.,* QuantiFluor™, PicoGreen, SYBR Green), and methods quantify DNA fragments (either dsDNA or ssDNA) at or below specific sizes (*e.g.,* short fragment qPCR, long fragment qPCR, and long/short qPCR ratio).

[0085] In some aspects, changes in gene expression are identified by measuring plasma cfDNA or cfRNA concentration levels and methods such as microarray analysis are used to assess changes in gene expression levels in a sample. Metrics of cfDNA or cfRNA concentration that may be used as input signatures for machine learning methods and models include but are not limited to Tape Station, short qPCR, long qPCR, and long/short qPCR ratio.

[0086] In some aspects, lcWGS can be used to sequence the cf-DNA in a sample and then interrogated for somatic mutations associated with a particular neuropathological condition. Using somatic mutations from lcWGS, deep WGS, or targeted sequencing (by NGS or other techniques) may generate data which may be used as an input to analyte signatures or analyzed in conjunction with analyte signatures to identify the onset, diagnosis, prognosis, progression, etc. of a neuropathological condition in a subject.

[0087] Somatic mutation analysis allows multiplexed detection of specific loci information in a single test. Such panel-based analysis can range in gene number from several to several hundred in a single assay. Other types of gene panels include whole-exon or whole-gene sequencing, or genomic analysis may be directed to a selected class of genes (*e.g.,* kinome or

transcription factor profiling) and offer the advantage of identifying novel mutations in a known gene set.

[0088] In other examples, assays are used to infer the three-dimensional structure of a genome using cell-free DNA (cfDNA). In particular, the present disclosure provides methods and systems for detecting chromatin abnormalities associated with diseases or conditions. For example, the abundance of a cfDNA fragment in a sample is believed to be predictive of the chromatin state of the gene from which the cfDNA fragment originated, and these states can change in certain pathological disorders. Identifying changes in the chromatin state of genes can thus serve as a method to identify the presence or progression of a particular disease in a subject. The chromatin state of genes can be predicted from the abundance and position of cfDNA fragments in samples using computer-aided techniques. The chromatin state may also be useful in inferring gene expression in a sample. Pliner H. A. et al., *Molecular Cell.* 71 (5) (2018) 858–871. The expression of a gene can be controlled by controlling access of the cellular machinery to the transcription start site. Access to the transcription start site can be determined by the state of the chromatin on which the transcription start site is located. Chromatin state can be controlled through chromatin remodeling, which can condense (close) or loosen (open) transcription start site. A closed transcription start site results in decreased gene expression while an open transcription start site results in increased gene expression. Also, the length of cfDNA fragments may depend on chromatin state. Chromatin remodeling can occur through the modification of histone and other related proteins. Non-limiting examples of histone modifications that can control the state of chromatin and transcription start sites include, for example, methylation, acetylation, phosphorylation, and ubiquitination.

[0089] Expression of genes is also controlled by more distal elements such as enhancers, which interact with transcriptional machinery in the 3D space of the physical genome. ATAC-seq and DNAse-seq provide measurements of open chromatin, which correlate with the binding of these more distal elements which may not be obviously associated with a particular gene. For example, ATAC-seq data can be obtained for a multitude of cell types and states and be used to identify regions of the genome with open chromatin for a variety of underlying regions such as active transcription start sites or bound enhancers or repressors.

[0090] The half-life of cfDNA once released from cells can depend on chromatin remodeling states. Thus, the abundance of a cfDNA fragment in a sample can be indicative of the chromatin

state of the gene from which the cfDNA fragment originated (referred to herein as a cfDNA's "position"). Chromatin states of genes can change in diseases. Identifying changes in the chromatin state of genes can serve as a method to identify the presence of a disease in a subject. When comparing expressed and unexpressed genes, there is a quantitative shift in both the number and positional distribution of cell-free DNA (cfDNA) fragments.

[0091] The position of cfDNA sequence reads within the genome can be determined by "mapping" the sequence to a reference genome. Mapping can be performed with the aid of computer algorithms including, for example, the Needleman-Wunsch algorithm, the BLAST algorithm, the Smith-Waterman algorithm, a Burrows-wheeler alignment, a suffix tree, or a custom-developed algorithm.

[0092] DNA methylation, which refers to the addition of the methyl group to DNA, is one of the most extensively characterized epigenetic modification with important functional consequences. Typically, DNA methylation occurs at cytosine bases of nucleic acid sequences. Enzymatic methyl sequencing is especially useful since it uses a three-step conversion requiring lower volume of sample for analysis. Various methods of determining changes in methylation state of nucleic acids also can be used to identify analytes in the samples from the plasmapheresis biobank. Enzymatic methyl sequencing ("EMseq") is capable of characterizing DNA methylation of nearly every nucleotide in the genome. In certain methods, the cfDNA is subjected to conditions sufficient to convert cytosine nucleobases into uracil nucleobases by performing bisulfite conversion, *e.g.,* oxidizing the cfDNA. In some aspects, the bisulfite conversion comprises reduced representation bisulfite sequencing.

[0093] In other aspects, the assay that is used for methylation analysis is selected from mass spectrometry, methylation-Specific PCR (MSP), reduced representation bisulfite sequencing, (RRBS), HELP assay, GLAD-PCR assay, ChIP-on-chip assays, restriction landmark genomic scanning, methylated DNA immunoprecipitation (MeDIP), pyrosequencing of bisulfite treated DNA, molecular break light assay, methyl Sensitive Southern Blotting, High Resolution Melt Analysis (HRM or HRMA, ancient DNA methylation reconstruction, whole genome bisulfite sequencing (WGBS), or Methylation Sensitive Single Nucleotide Primer Extension Assay (msSNuPE).

[0094] In some aspects, the analytes are identified using cfRNA analysis. Methods for performing this include, but are not limited to, RNA sequencing, whole transcriptome shotgun

sequencing, northern blot, in situ hybridization, hybridization array, serial analysis of gene expression (SAGE), reverse transcription PCR, real-time PCR, real-time reverse transcription PCR, quantitative PCR, digital droplet PCR, microarray, NanoString mRNA analysis, FISH assays or a combination thereof.

[0095] Methods of "quantitative" amplification are a variety of suitable methods. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be used to calibrate the PCR reaction. Detailed protocols for quantitative PCR are provided in Innis, et al. (1990) PCR Protocols, A Guide to Methods and Applications (Academic Press, Inc. N.Y.). Measurement of DNA copy number at microsatellite loci using quantitative PCR analysis is described in Ginzinger DG, et al. (2000) *Cancer Research.* 60:5405-5409. The known nucleic acid sequence for the genes is sufficient to enable one to routinely select primers to amplify any portion of the gene. Fluorogenic quantitative PCR may also be used in the methods of the disclosure. In fluorogenic quantitative PCR, quantitation is based on amount of fluorescence signals, *e.g.,* TaqMan and SYBR green. Other suitable amplification methods include, but are not limited to, ligase chain reaction (LCR) (see Wu and Wallace (1989) *Genomics.* 4: 560, Landegren, et al. (1988) *Science.* 241:1077, and Barringer et al. (1990) *Gen.e* 89: 117), transcription amplification (Kwoh, et al. (1989) *Proc. Natl. Acad. Sci. USA* 86: 1173), self-sustained sequence replication (Guatelli, et al. (1990) *Proc. Nat. Acad. Sci. USA* 87: 1874), dot PCR, and linker adapter PCR, etc.

[0096] In various examples, proteins are assayed using immunoassay or mass spectrometry. For example, proteins may be measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

[0097] In various examples, proteins are measured by affinity reagents or immunoassays such as protein arrays, SIMOA (Quanterix, Billerica, MA, USA), ELISA (Abcam, Cambridge, UK), the Olink™ (Explore HT, Upsalla, Sweden) or SomaScan (Somalogic, Boulder, CO, USA), Luminex and Meso Scale Discovery.

[0098] In other examples, proteins in the samples are assayed based on characteristics such as post-translational modifications, phosphorylation status, or other protein modifications that may impact their activity, localization, lifespan and the like.

[0099] In some aspects, the protein data is normalized by a standard curve. In various examples, each protein is treated as an essentially unique immunoassay, each with a standard curve that can be calculated in various ways. The concentration relationship is typically non-linear. Then the sample may be run. and calculated based on the expected fluorescence concentration in the primary sample.

[00100] Following data collection and analysis, one or more longitudinal sample sets of the plasmapheresis biobank may be linked to analysis data (*e.g.,* "omics" data) for samples of the longitudinal sample set. As such, a longitudinal sample set of the library may be associated with omics data for one or more samples of the longitudinal sample set. In such instances, one or more of the samples of the set may be linked to omics data for that sample, where the number of samples of a given set linked to their omics data may vary. In some instances, 1% or more, such as 5% or more, including 10% or more, *e.g.,* 25% or more of the samples of a given set may be linked their respective omics data. In some aspects, the majority of the samples of a given set may be linked their omics data, *e.g.,* more than 50%, such as 60% or more, *e.g.,* 70% or more, such as 80% or more, including 90% or more, and in some instances all of the samples of a given set may be linked to their respective omics data. Omics data to which a given sample may be linked include, but are not limited to, proteomics data (including subsets thereof, *e.g.,* glycoproteomics, post-translation modifications, oxidation state, acetylation state, methylation state, ubiquitylation state, etc.), genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data. In some instances, the omics data to which a given sample is linked includes one or more of proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data, such as two or more of proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data, such as three or more of proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data, such as four or more of proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data, such as five or more of proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data, where in some instances the omics data includes proteomics data, genomics data, epigenomics data, metabolomics data, transcriptomics data and lipidomics data.

[00101] Where desired, longitudinal sample sets of a given library may be categorized (i.e., grouped or classified) into one or more cohorts. Cohorts of interest may vary, where in some instances the cohorts include disease cohorts, age cohorts, ethnicity cohorts, gender cohorts, lifestyle cohorts, genotypic cohorts, molecular cohorts, control cohorts and the like. In some aspects, longitudinal sample sets of a given library are categorized into one or more disease cohorts. Disease cohorts of interest include, but are not limited to: oncological diseases, central nervous system diseases, immunology and inflammatory diseases, cardiovascular diseases, infectious diseases, respiratory diseases and metabolic diseases.

Predictive Modeling

[00102] In some aspects, disease prognosis, status, subtype, progression and the like in a subject is determined by measuring the levels of two or more analytes in a given disease subject's samples over time and applying an interpretation function to transform the biomarker levels into a predictive value, which provides a quantitative measurement of disease presence, progression, severity, etc. in the subject . In specific aspects this qualitative measurement is correlated with traditional clinical assessments for clinical diagnosis, monitoring and/or intervention.

[00103] In some aspects, the interpretation function is based on a predictive model. Established statistical algorithms and methods well-known in the art, useful as models or useful in designing predictive models, can include but are not limited to: analysis of variants (ANOVA); Bayesian networks; boosting and Ada-boosting; bootstrap aggregating algorithms; decision trees classification techniques, such as Classification and Regression Trees (CART), boosted CART, Random Forest (RF), Recursive Partitioning Trees (RPART), and others; Curds and Whey (CW); Curds and Whey-Lasso; dimension reduction methods, such as principal component analysis (PCA) and factor rotation or factor analysis; discriminant analysis, including Linear Discriminant Analysis (LDA), Eigengene Linear Discriminant Analysis (ELDA), and quadratic discriminant analysis; Discriminant Function Analysis (DFA); factor rotation or factor analysis; genetic algorithms; Hidden Markov Models; kernel based machine algorithms such as kernel density estimation, kernel partial least squares algorithms, kernel matching pursuit algorithms, kernel Fisher's discriminate analysis algorithms, and kernel principal components analysis algorithms; linear regression and

generalized linear models, including or utilizing Forward Linear Stepwise Regression, Lasso (or LASSO) shrinkage and selection method, and Elastic Net regularization and selection method; glmnet (Lasso and Elastic Net-regularized generalized linear model); Logistic Regression (LogReg); meta-learner algorithms; nearest neighbor methods for classification or regression, e.g. Kth-nearest neighbor (KNN); non-linear regression or classification algorithms; neural networks; partial least square; rules based classifiers; shrunken centroids (SC); sliced inverse regression; Standard for the Exchange of Product model data, Application Interpreted Constructs (StepAIC); super principal component (SPC) regression; and, Support Vector Machines (SVM) and Recursive Support Vector Machines (RSVM), among others. Additionally, clustering algorithms are well known in the art and can be used in determining subject sub-groups.

[00104] In some aspects, the interpretation function is based on data-driven disease progression modeling combined with multimodel integration methods. Such modeling strategy could help to reconstruct disease progression timeline and identify a pathophysiological cascade of biomarkers from multiple sources of omics data from one or more cross-sectional and longitudinal studies, including large cohorts of patients, healthy controls and the individuals at prodromal/at-risk stages. Data-driven disease progression models could be derived from one or more of following methods: event-based model and discriminative event-based model from mixture modeling, subtype and stage inference by combining clustering and disease progression model, latent-time joint mixed model and differential equation models. Multimodel data integration methods combine heterogeneous information from physiological, clinical and biological datasets through multivariate statistical analysis and latent variable modeling, including partial least square (PLS), canonical correlation analysis (CCA), and its extension to Deep CCA.

[00105] Logistic Regression is a well-known and often preferred traditional predictive modeling method for dichotomous response variables such as in comparisons between two treatments (*e.g.,* treatment 1 versus treatment 2). It can be used to model both linear and non-linear aspects of the data variables and provides easily interpretable odds ratios.

[00106] Discriminant Function Analysis (DFA) uses a set of analytes as variables (roots) to discriminate between two or more naturally occurring groups. DFA can be used in aspects of the invention that test analytes which are significantly different between groups. A

forward step-wise DFA can be used to select a set of analytes that maximally discriminate among the groups studied. Specifically, at each step all variables can be reviewed to determine which will maximally discriminate among groups. This information is then included in a discriminative function, denoted a root, which is an equation consisting of linear combinations of analyte concentrations for the prediction of group membership. The discriminatory potential of the final equation can be observed as a line plot of the root values obtained for each group. This approach identifies groups of analytes whose changes in concentration levels can be used to delineate profiles, diagnose disease, and assess therapeutic efficacy. The DFA model can also create an arbitrary score by which new subjects can be classified as either "healthy" or "diseased." To facilitate the use of this score for the medical community the score can be rescaled so a value of 0 indicates a healthy individual and scores greater than 0 indicate increasing disease activity.

[00107] Classification and regression trees (CART) can be used in aspects of the invention to perform logical splits (if/then) of data to create a decision tree. All observations that fall in a given node are classified according to the most common outcome in that node. CART results are easily interpretable since one can follow a series of if/then tree branches until a classification results.

[00108] Support vector machines (SVM) can be used in aspects of the invention to classify objects into two or more classes. Examples of such classes include sets of treatment alternatives, sets of diagnostic alternatives, or sets of prognostic alternatives. Each object is assigned to a class based on its similarity to (or distance from) objects in the training data set in which the correct class assignment of each object is known. The measure of similarity of a new object to the known objects is determined using support vectors, which define a region in a potentially high dimensional space.

[00109] The process of bootstrap aggregating is computationally simple and can be used in aspects of the invention. In the first step, a given dataset is randomly resampled a specified number of times (*e.g.*, hundreds, thousands, or more), effectively providing that number of new datasets, which are referred to as "bootstrapped resamples" of data. Each of these bootstrapped resamples of data can then be used to build a model. Then, in the example of classification models, the class of every new observation is predicted by the number of classification models created in the first step. The final class decision is based upon a "majority

vote" of the classification models; i.e., a final classification call is determined by counting the number of times a new observation is classified into a given group, and taking the majority classification (33%+ for a three-class system). In the example of logistical regression models, if a logistical regression is bagged 1000 times, there will be 1000 logistical models, and each will provide the probability of a sample belonging to class 1 or 2.

[00110] Curds and Whey (CW) using ordinary least squares (OLS) is another predictive modeling method that can be used in aspects of the invention. *See* L. Breiman and JH Friedman, J. Royal. Stat. Soc. B 1997, 59(l):3-54, herein incorporated by reference herein in its entirety. This method takes advantage of the correlations between response variables to improve predictive accuracy, compared with the usual procedure of performing an individual regression of each response variable on the common set of predictor variables X. In CW, Y = XB * S, where Y = (v¾) with k for the k* patient and j for j* response (j =1 for TJC, j = 2 for SJC, etc.), B is obtained using OLS, and S is the shrinkage matrix computed from the canonical coordinate system. Another method that can be used in aspects of the invention is Curds and Whey and Lasso in combination (CW-Lasso). Instead of using OLS to obtain B, as in CW, here Lasso is used, and parameters are adjusted accordingly for the Lasso approach.

[00111] The embodied techniques above can be useful either combined with: a biomarker selection technique (*e.g.,* forward selection, backwards selection, or stepwise selection); with complete enumeration of all potential panels of a given size; with genetic algorithms; neural networks; gradient boosting; large language models or they can themselves include biomarker selection methodologies in their own techniques. These techniques can be coupled with information criteria, such as Akaike's Information Criterion (AIC), Bayes Information Criterion (BIC), or cross-validation, to quantify the tradeoff between the inclusion of additional biomarkers and model improvement, and to minimize overfit. The resulting predictive models can be validated in other studies, or cross-validated in the study they were originally trained in, using such techniques as, for example, Leave-One-Out (LOO) and 10-Fold cross-validation (10-Fold CV).

[00112] In some aspects of the present disclosure, candidate drug targets can be identified through the use of one or more of the following methods Mendelian Randomization, network-based approaches, graph-based approaches and propagation. Identification of

candidate drug targets are, however, not limited to the aforementioned methods and said methods are listed by way of example and not limitation.

[00113] Mendelian randomization is a method that uses genetic variants as instrumental variables to assess causal relationships between modifiable exposures and disease outcomes. This aids in determining potential therapeutic targets. Network-Based approaches have different applications including the analysis of protein-protein interaction (PPI) networks to identify key proteins or hubs involved in disease pathways. The modeling of gene regulatory networks (GRNs) is a network-based approach to uncover transcription factors or genes with regulatory roles in disease processes. The investigation of metabolic networks is a network-based approach to identify metabolites or enzymes involved in disease metabolism. And the construction of drug-target interaction networks is another network-based approach that can be used to identify potential drug targets based on drug-protein interactions.

[00114] Graph-based approaches involve analyzing the interconnected relationships between entities represented as nodes and edges in a network, aiding in the identification of key targets within biological systems for therapeutic intervention. Propagation diffusion involves algorithms that spread information or influence through networks, helping identify relevant nodes or entities in complex systems such as biological networks for potential therapeutic targeting.

[00115] In some aspects of the present disclosure, it is not required that the disease score be compared to any pre-determined "reference," "normal," "control," "standard," "healthy," "pre-disease" or other like index, in order for the score to provide a quantitative measure of disease activity in the subject.

[00116] In other aspects of the present disclosure, the amount of the biomarker(s) can be measured in a sample and used to derive a disease score, which is then compared to a "normal" or "control" level or value, utilizing techniques such as, reference or discrimination limits or risk defining thresholds, in order to define cut-off points and/or abnormal values for a disease. The normal level then is the level of one or more biomarkers or combined biomarker indices typically found in a subject who is not suffering from the disease under evaluation. Other terms for "normal" or "control" are, *e.g.,* "reference," "index," "baseline," "standard," "healthy," "pre-disease," and the like. Such normal levels can vary, based on whether a biomarker is used alone or in a formula combined with other biomarkers to output a score.

Alternatively, the normal level can be a database of biomarker patterns from previously tested subjects who did not develop the disease under evaluation over a pre-clinical or clinically relevant time period. Reference (normal, control) values can also be derived from, a control subject or population whose disease activity level or state is known. In some aspects of the present disclosure, the reference value can be derived from one or more subjects who have been exposed to treatment for disease, or from one or more subjects who are at low risk of developing disease, or from subjects who have shown improvements in disease activity factors (*e.g.,* clinical parameters as defined herein) as a result of exposure to treatment. In some aspects the reference value can be derived from one or more subjects who have not been exposed to treatment. An example comprises samples that can be collected from (a) subjects who have received initial treatment for disease, and (b) subjects who have received subsequent treatment for disease, to monitor the progress of the treatment. A reference value can also be derived from disease activity algorithms or computed indices from population studies.

Databases

[00117] Aspects of the present disclosure further include databases of data and associated information (*e.g.,* storage in the form of non-transitory computer readable storage media) that comprise data derived from one or more longitudinal sample sets. As such, aspects of the invention include databases of longitudinal data sets from one or a plurality of longitudinal sample sets, wherein a longitudinal sample set includes multiple samples obtained from the same donor over a period of time, *e.g.,* where the data includes donor data linked the longitudinal sample set as described in more detail herein. Computer readable storage media may be employed on one or more computers for complete automation or partial automation of a system for practicing methods and models as described herein.

[00118] In certain aspects, instructions in accordance with the method described herein can be coded onto a computer-readable medium in the form of "programming," where the term "computer readable medium" as used herein refers to any non-transitory storage medium that participates in providing instructions and data to a computer for execution and processing. Examples of suitable non-transitory storage media include a floppy disk, hard disk, optical disk, magneto-optical disk, CD-ROM, CD-R, magnetic tape, non-volatile memory card, ROM, DVD-ROM, Blue-ray disk, solid state disk, and network attached storage (NAS), whether or

not such devices are internal or external to the computer. A file containing information can be "stored" on a computer readable medium, where "storing" means recording information such that it is accessible and retrievable at a later date by a computer. The computer-implemented method described herein can be executed using programming that can be written in one or more of any number of computer programming languages. Such languages include, for example, Java (Sun Microsystems, Inc., Santa Clara, CA), Visual Basic (Microsoft Corp., Redmond, WA), C++ (AT&T Corp., Bedminster, NJ), Python, as well as any many others.

[00119] The format of the database and the manner in which the data records, *e.g.,* sample records, linked donor data, sample omics data, etc., are indicated as being associated / linked with one another (because they relate to the same sample) may follow conventional database techniques. The database control unit comprises a processor which is suitably configured/programmed to provide the desired functionality described herein using conventional programming/configuration techniques for controlling database generation, maintenance and access operations. The functionality of a suitable control unit may be provided in various different ways, for example using a suitably programmed general purpose computer, or suitably configured application-specific integrated circuit(s)/circuitry or using a plurality of discrete circuitry/processing elements for providing different elements of the desired functionality.

[00120] Database storage units of aspects of the invention may include a memory for the various records making up the database. The database storage unit may be based on any conventional memory technology, for example a solid-state memory or a disk-based memory, such as a magnetic disk or optical disc-based memory. Furthermore, although the memory comprising the storage unit may in principle comprise a single memory unit, in general the memory comprising the storage unit may be made up of a distributed array of memory units providing redundancy for failure recovery in accordance with conventional techniques.

Computer Systems

[00121] As summarized above, aspects of the present disclosure include systems for practicing the subject methods, *e.g.,* querying a database. Systems according to certain aspects comprise a processor comprising memory operably coupled to the processor, wherein the memory comprises instructions stored thereon, which, when executed by the processor, cause

the processor to: receive input from an input device, and outputting, to an output device, a result, wherein the processor and the memory are operably connected to each of the input device and the output device.

[00122] The databases containing the longitudinal data sets can be configured in useful data sets stored on computer readable media. Certain systems also provide functionality (*e.g.,* code and processes) for storing any of the results (*e.g.,* query results) or data sets generated as described herein. Such results or data sets are typically stored, at least temporarily, on a computer readable medium such as those presented in the following discussion. The results or data sets may also be output in any of various manners such as displaying, printing, and the like.

[00123] Examples of displays suitable for interfacing with a user in accordance with the invention include but are not limited to cathode ray tube displays, liquid crystal displays, plasma displays, touch screen displays, video projection displays, light-emitting diode and organic light-emitting diode displays, surface-conduction electron-emitter displays and the like. Examples of printers include toner-based printers, liquid inkjet printers, solid ink printers, dye-sublimation printers as well as inkless printers such as thermal printers. Printing may be to a tangible medium such as paper or transparencies.

[00124] Examples of tangible computer-readable media suitable for use computer program products and computational apparatus of this invention include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media; semiconductor memory devices (*e.g.,* flash memory), and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM) and sometimes application-specific integrated circuits (ASICs), programmable logic devices (PLDs) and signal transmission media for delivering computer-readable instructions, such as local area networks, wide area networks, and the Internet. The data and program instructions provided herein may also be embodied on a carrier wave or other transport medium (including electronic or optically conductive pathways). The data and program instructions of this invention may also be embodied on a carrier wave or other transport medium (*e.g.,* optical lines, electrical lines, and/or airwaves).

[00125] Examples of program instructions include low-level code, such as that produced by a compiler, as well as higher-level code that may be executed by the computer using an interpreter. Further, the program instructions may be machine code, source code and/or any other code that directly or indirectly controls operation of a computing machine. The code may specify input, output, calculations, conditionals, branches, iterative loops, etc.

[00126] Figure 10 illustrates, in simple block format, a typical computer system that, when appropriately configured or designed, can serve as a computational apparatus according to certain aspects. The computer system 1000 includes any number of processors 1002 (also referred to as central processing units, or CPUs) that are coupled to storage devices including primary storage 1006 (typically a random access memory, or RAM), primary storage 1004 (typically a read only memory, or ROM). CPU 1002 may be of various types including microcontrollers and microprocessors such as programmable devices (*e.g.,* CPLDs and FPGAs) and non-programmable devices such as gate array ASICs or general-purpose microprocessors. In the depicted aspect, primary storage 1004 acts to transfer data and instructions uni-directionally to the CPU and primary storage 1006 is used typically to transfer data and instructions in a bi-directional manner. Both of these primary storage devices may include any suitable computer-readable media such as those described above. A mass storage device 1008 is also coupled bi-directionally to primary storage 1006 and provides additional data storage capacity and may include any of the computer-readable media described above. Mass storage device 1008 may be used to store programs, data and the like and is typically a secondary storage medium such as a hard disk. Frequently, such programs, data and the like are temporarily copied to primary memory 1006 for execution on CPU 1002. It will be appreciated that the information retained within the mass storage device 1008, may, in appropriate cases, be incorporated in standard fashion as part of primary storage 1004. A specific mass storage device such as a CD-ROM 1014 may also pass data uni-directionally to the CPU or primary storage.

[00127] CPU 1002 is also coupled to an interface 1010 that connects to one or more input/output devices such as such as video monitors, track balls, mice, keyboards, microphones, touch-sensitive displays, transducer card readers, magnetic or paper tape readers, tablets, styluses, voice or handwriting recognition peripherals, USB ports, or other well-known input devices such as, of course, other computers. Finally, CPU 1002 optionally may be

coupled to an external device such as a database or a computer or telecommunications network using an external connection as shown generally at 1012. With such a connection, it is contemplated that the CPU might receive information from the network, or might output information to the network in the course of performing the method steps described herein.

[00128] In one aspect, a system such as computer system 1000 is used as a data import, data correlation, and querying system capable of performing some or all of the tasks described herein. System 1000 may also serve as various other tools associated with Knowledge Bases and querying such as a data capture tool. Information and programs, including data files can be provided via a network connection 1012 for access or downloading by a researcher. Alternatively, such information, programs and files can be provided to the researcher on a storage device.

[00129] In a specific aspect, the computer system 1000 is directly coupled to a data acquisition system such as a microarray or high-throughput screening system that captures data from samples. Data from such systems are provided via interface 1010 for analysis by system 1000. Alternatively, the data processed by system 1000 are provided from a data storage source such as a database or other repository of relevant data. Once in apparatus 900, a memory device such as primary storage 1006 or mass storage 1008 buffers or stores, at least temporarily, relevant data. The memory may also store various routines and/or programs for importing, analyzing and presenting the data, including importing Feature Sets, correlating Feature Sets with one another and with Feature Groups, generating and running queries, etc.

[00130] In certain aspects user terminals may include any type of computer (*e.g.,* desktop, laptop, tablet, etc.), media computing platforms (*e.g.,* cable, satellite set top boxes, digital video recorders, etc.), handheld computing devices (*e.g.,* PDAs, e-mail clients, etc.), cell phones or any other type of computing or communication platforms. A server system in communication with a user terminal may include a server device or decentralized server devices, and may include mainframe computers, mini computers, super computers, personal computers, or combinations thereof. A plurality of server systems may also be used without departing from the scope of the present invention. User terminals and a server system may communicate with each other through a network. The network may comprise, *e.g.,* wired networks such as LANs (local area networks), WANs (wide area networks), MANs (metropolitan area networks), ISDNs (Integrated Service Digital Networks), etc. as well as

wireless networks such as wireless LANs, CDMA, Bluetooth, and satellite communication networks, etc. without limiting the scope of the present invention.

[00131] In at least some of the previously described aspects, one or more elements used in an embodiment can interchangeably be used in another embodiment unless such a replacement is not technically feasible. It will be appreciated by those skilled in the art that various other omissions, additions and modifications may be made to the methods and structures described above without departing from the scope of the claimed subject matter. All such modifications and changes are intended to fall within the scope of the subject matter, as defined by the appended claims.

EXAMPLES

[00132] The following examples are put forth so as to provide those of ordinary skill in the art with a complete disclosure and description of how to make and use the present disclosure and are not intended to limit the scope of what the inventors regard as their invention, nor are they intended to represent or imply that the experiments below are all of or the only experiments performed. It will be appreciated by persons skilled in the art that numerous variations and/or modifications may be made to the invention as shown in the specific aspects without departing from the spirit or scope of the disclosure as broadly described. The present aspects are, therefore, to be considered in all respects as illustrative and not restrictive.

Example 1: Generation of a Plasmapheresis Biobank

[00133] An exemplary plasmapheresis biobank of the invention was generated using a large collection of longitudinal sample sets obtained by plasmapheresis over a multi-year period. Through tokenization, the samples of this biobank were connected to various data regarding the samples and their donors in de-identified fashion in a database. This data included, but was not limited to, real world data, such as diagnostic codes (*e.g.,* ICD-10 codes, *infra*). Examples of data used included, but were not limited to, counts of donors diagnosed with various ICD-10 codes (see Table 1) and temporal relationship between plasma donations and disease diagnosis (see, *e.g.,* Figures 7-9). Table 1 depicts a portion of the data derived from the biobank including a selection of disease counts represented in a

sample collection as of January 2024. The ICD-10 code refers to the International Classification of Diseases (ICD, 10th Revision) that is a medical classification list by the World Health Organization (WHO), and is a diagnostic tool used to classify and monitor causes of injury and death and that maintains information for health analyses, such as the study of mortality and morbidity trends.

TABLE 1: ICD Codes and Donor Count in Plasmapheresis Biobank

[ICD-10 Code]	[ICD-10 Description]	[Donor Count]
C259	Malignant neoplasm of pancreas, unspecified	597
C569	Malignant neoplasm of unspecified ovary	691
C719	Malignant neoplasm of brain, unspecified	869
C3490	Malignant neoplasm of unspecified part of unspecified bronchus	2,238
H409	Unspecified glaucoma	3,575
F329	Major depressive disorder, single episode, unspecified	279,138
G35	Multiple sclerosis	4,362
G20	Parkinson's disease	1,150
G300	Alzheimer's disease with early onset	153
G309	Alzheimer's disease, unspecified	391
I67850	Cerebral autosomal dominant arteriopathy with subcortical infarcts and	20
F319	Bipolar disorder, unspecified	114,631
F209	Schizophrenia, unspecified	36,555
G1221	Amyotrophic lateral sclerosis	152
M0579	Rheumatoid arthritis with	1,658

	rheumatoid factor of multiple sites	
L409	Psoriasis, unspecified	10,917
K5090	Crohn's disease, unspecified, without complications	3,668
M329	Systemic lupus erythematosus, unspecified	4,542
I10	Essential (primary) hypertension	374,089
I509	Heart failure, unspecified	20,132
I739	Peripheral vascular disease, unspecified	16,485
I4891	Unspecified atrial fibrillation	13,330
U071	COVID-19	217,487
J1100	Influenza due to unidentified influenza virus with unspecified	810
Z227	Latent tuberculosis	848
B54	Unspecified malaria	53
B182	Chronic viral hepatitis C	10,181
J45909	Unspecified asthma, uncomplicated	175,113
J449	Chronic obstructive pulmonary disease, unspecified	53,323
J84112	Idiopathic pulmonary fibrosis	165
Z141	Cystic fibrosis carrier	1,011
E109	Type 1 diabetes mellitus without complications	5,318
E119	Type 2 diabetes mellitus without complications	105,035
K700	Alcoholic fatty liver	1,512
K7581	Nonalcoholic steatohepatitis (NASH)	3,261

K760	Fatty (change of) liver, not elsewhere classified	70,924

Example 2: Samples from the Plasmapheresis Biobank for Identification of Biomarkers in Parkinson's Disease

[00134] Earlier stage diagnosis can be especially useful for chronic, progressive neurodegenerative diseases in which the symptoms follow the earlier pathology of the disease. For example, around 60-80% of dopaminergic neurons in the substantia nigra of a subject must die before noticeable Parkinson's symptoms appear, meaning a significant loss of these neurons is needed before motor symptoms become evident. *See* Lang AE et al., *N Engl J Med.* 1998;339:1044–1053; Dauer W et al. *Neuron.* 2003;39:889–909. Identification of such progressive neuropathological conditions through non-invasive assessment of analytes and temporal signatures allows for earlier treatment and perhaps even intervention in the disease state before it progresses to the symptomatic stage.

[00135] Ongoing initiatives such as The Parkinson Progression Marker Initiative (PPMI) (*Prog Neurobiol.* 2011 Dec;95(4):629-35) have played a crucial role in identifying potential biomarkers associated with Parkinson's disease and to understand the development of the disease and identify progression markers. PPMI and research cohorts rely on a detailed longitudinal follow up of hundreds to several thousands of individuals with early symptoms or established PD diagnosis. In contrast, population-based investigations like those involving the UK Biobank (Sudlow C et al., *PLoS Med.* 2015 Mar 31;12(3):e1001779) follow hundreds of thousands of individuals during the entire development of diseases but only a small fraction of them will develop rare diseases like PD. While tailored research cohorts like PPMI and population-based studies like those using the UK Biobank are complementary by nature, they both suffer from a major limitation: the number of biospecimens collected during the pre-clinical phases of diseases such as PD is by nature extremely limited. Indeed, research cohorts are mostly focused on the clinical phase of diseases while the collection of biospecimens in population-based cohorts primarily occurs at the baseline. This prevents a longitudinal understanding of the early phases of disease progression from a molecular perspective.

[00136] These limitations were overcome in the present studies by the use of a unique set of samples that provide the ability to perform retrospective analysis on a large cohort of

affected individuals spanning years from prior to onset of a disease through diagnosis and progression of the disease. The identification of analytes for use in the systems and methods of the disclosure was empowered by a unique asset of plasmapheresis samples taken from a cohort of over 3,000,000 individuals over two decades, with as many as thirty samples taken from a single individual during this timeframe, allowing a unique opportunity for retrospective analysis of this population. Importantly, as the subjects donating these samples were phenotypically designated as "normal" at the time of donation, the retrospective analysis of these samples was able to provide molecular identification of disease onset and progression in an agnostic fashion, and potentially years prior to clinical identification of the disease state.

Example 3: Confirmation of Robust Biological Signals in Samples from the Plasmapheresis Biobank

[00137] Prior to using samples from the plasmapheresis biobank described in Example 1 for the identification of such predictive analytes, first the samples were confirmed to be suitable for such analysis. Specifically, the feasibility of using the samples from the plasmapheresis biobank for the analysis of neuropathological conditions was investigated using longitudinal profiling of all phases of Parkinson's disease ("PD") and by demonstrating the experimental feasibility of measuring longitudinal molecular changes in ~700 samples from human plasma donors.

[00138] More than 58,000 samples from donors who developed PD were stored within the plasmapheresis biobank, establishing the plasmapheresis biobank as the most extensive collection of biospecimens from individuals undergoing PD development. The samples used for analysis of sample integrity and subject aging process were selected using linear regression adjusted for gender, race, storage weeks and number of donations within 30 days. Correction for multiple comparisons was performed using Benjamini-Hochberg (BH) adjustment, a statistical method that controls the false discovery rate (FDR) in multiple comparisons. Benjamini, Y., & Hochberg, Y. (1995) *Journal of the Royal Statistical Society: Series B (Methodological).* 57(1), 289-300.

[00139] Briefly, the potential impact of age and storage conditions on the plasma proteome in the stored plasmapheresis samples was measured using two proteomics detection platforms: the Olink Explore HT (5k) (Olink, Uppsala, Sweden), a proteomics antibody-based

detection platform that can measure over 5,400 proteins in a plasma sample and the SomaScan™ assay v4.1 (Somalogic, Boulder, CO, USA), a proteomics aptamer-based detection platform capable of measuring 7288 human proteins in a plasma sample. Results are shown in Figures 11A and 11B.

[00140] Age-related proteomics changes were compared to those from the UK Biobank cohort (Sun, B. B. *et al. Nature.* (2023) 622, 329–338) and a meta-analysis of four independent cohorts (Coenen, L. et al., *Front. Aging.* 4, 112109 (2023)). Proteomic changes were connected to biological events associated with aging, such as menopause or cardiovascular events (respectively shown in Figures 12A and 12B). Plasma proteomics measurements were further performed using the Somalogic SomaScan™ platform (Somalogic, Boulder, CO, USA) in 504 samples from 40 donors. Lines in Figures 12A and 12B connect samples from the same donors. Observed temporal molecular changes can be associated with menopause (Follicle stimulating hormone, Figure 12A) or cardiovascular event as identified by medical records (Natriuretic peptides B, Figure 12B).

[00141] These data demonstrate that the samples from the stored plasmapheresis biobank asset were robust and suitable for various multi-omic assays for identification of analytes for systems and method for improved drug development. In particular, the samples were uniquely positioned to allow identification of targets, pathways, and various other biological indicia that are altered over time in pathological conditions, and in assisting in the development of drug development systems, tools, and reagents.

Example 4: Longitudinal Profiling of Biological Processes in for Parkinson's Disease in Samples from the Plasmapheresis Biobank

[00142] The feasibility of using the samples from the plasmapheresis biobank for multi-omic analysis of neuropathological conditions was investigated using longitudinal profiling of all phases of Parkinson's disease ("PD") and by demonstrating the experimental feasibility of measuring longitudinal molecular changes in samples from ~700 human plasma donors.

[00143] It was determined that more than 58,000 samples from donors who developed PD were stored within the plasmapheresis biobank, making it one of the most extensive collections of biospecimens from individuals undergoing PD development. This unique asset allowed the retrospective analysis and deep longitudinal molecular profiling of plasma samples

from these individuals who developed PD and a unique opportunity to study all phases of PD development, from onset to progression, at a molecular level. Longitudinal samples were selected from a subset of plasma donors who were identified as being affected with PD and carefully matched control donors, who were then each assayed across multiple proteomics technologies to discover new pre-clinical analytes of PD and new candidate therapeutics targets and pathways for further drug development.

[00144] As illustrated in Figure 13, the plasmapheresis biobank was annotated with real-world data (RWD), and the individual members of the sample library (each designated by a black dot) was shown to cover all phases of PD development from approximately twelve years before clinical diagnosis to eight years after clinical diagnosis. Only donors with complete and consistent demographics from the plasmapheresis biobank were included in this visualization. PD diagnosis was established based on the presence of at least one G20 ICD10 code in their medical records (claims or electronic health record ("EHR") data). Plasma samples were represented relative to the time of their first diagnosis of PD (G20 ICD10 code) and donors were ranked based on the time difference between their last donation and their first PD diagnosis.

[00145] The confirmed data were used to provide an initial definition of the PD cohort and to identify potential plasma donors with a high likelihood of having developed PD (Figure 14). The selection process involved the definition of specific filtering criteria, which may include but are not limited to demographics, medical history (G20 code, PD medication), and sample availability. This initial PD cohort comprised around 700 donors who had developed PD based on claims data. In specific cases, EHR data could potentially be used to provide additional information on donors with specific medical or proteomics patterns when needed.

[00146] Three main cohorts were defined based primarily on the claims data: a cohort of established PD; an expert reviewed PD cohort; and a non-PD controls cohort.

[00147] The data-driven cohort of established PD can be identified, *e.g.,* using the following set of inclusion criteria: two appearance of inclusion ICD-10-CM codes (Table 2), two or more prescriptions for PD medications (Table 3), and at least three or more diagnosis

claims. 125 donors with at least one plasma donation were included in the established PD cohort based on such exemplary claims data.

[00148] In addition, exemplary criteria such as the following can be used as exclusion criteria: exclusion ICD-10-CM codes (Table 4) and exclusion of medications that can induce symptoms similar to PD (Table 5).

[00149] Individuals considered for the expert reviewed cohort were selected using different approaches relying on presence/absence of ICD10 codes related to diagnosis and medication, and machine learning algorithms. The health journey of each of these individuals was examined by CNS scientists trained to identify PD based on medical health records. 223 donors were included in the expert-reviewed cohort.

[00150] The samples from these donors who developed PD covered all phases of PD development (Figure 15). Plasma samples were represented relative to the time of their first diagnosis of PD (G20 ICD10 code) and donors were ranked based on the time difference between their last donation and their first PD diagnosis.

[00151] Finally, the control cohort exclusion criteria are a lack of sufficient information and any ICD-10-CM codes from Tables 2 and 4 and all medications from Tables 3 and 5.

Table 2: Inclusion ICD-10-CM codes used for the selection of PD cohorts and for the exclusion of donors from the control cohort:

ICD-10-CM code	Description
G20	Parkinson's disease
G20.A	Parkinson's disease without dyskinesia
G20.A1	Parkinson's disease without dyskinesia without mention of fluctuations
G20.A2	Parkinson's disease without dyskinesia with fluctuations
G20.B	Parkinson's disease with dyskinesia
G20.B1	Parkinson's disease with dyskinesia without mention of fluctuations
G20.B2	Parkinson's disease with dyskinesia with fluctuations
G20.C	Parkinsonism, unspecified

Table 3: Parkinson's disease medications used for the selection of PD cohorts and for the exclusion of donors from the control cohort:

Proprietary medication name	Nonproprietary medication name
Amantadine	Amantadine
Gocovri	Amantadine
osmolex er	Amantadine
amantadine hcl	amantadine hcl
Apokyn	apomorphine hydrochloride
apomorphine hydrochloride	apomorphine hydrochloride
apomorphinum muriaticum	apomorphine hydrochloride
Kynmobi	apomorphine hydrochloride
Benztropine	Benztropine
benztropine mesylate	benztropine mesylate

Cogentin	benztropine mesylate
bromocriptine mesylate	bromocriptine mesylate
Cycloset	bromocriptine mesylate
Parlodel	bromocriptine mesylate
Cabergoline	Cabergoline
Carbidopa	Carbidopa
Lodosyn	Carbidopa
carbidopa and levodopa	carbidopa and levodopa
Duopa	carbidopa and levodopa
Parcopa	carbidopa and levodopa
Rytary	carbidopa and levodopa
Sinemet	carbidopa and levodopa
sinemet – cr	carbidopa and levodopa
Dhivy	carbidopa levodopa
carbidopa, levodopa, and entacapone	carbidopa, levodopa, and entacapone
Stalevo	carbidopa, levodopa, and entacapone
Comtan	Entacapone
Entacapone	Entacapone
Nourianz	Istradefylline
Inbrija	Levodopa
Ongentys	Opicapone
Pramipexole	Pramipexole
pramipexole dihydrochloride	Pramipexole
pramipexole dihydrochloride - extended-release	Pramipexole
Mirapex	pramipexole dihydrochloride
mirapex er	pramipexole dihydrochloride
pramipexole dihydrochloride	pramipexole dihydrochloride
Rasagiline	Rasagiline
Azilect	rasagiline mesylate

rasagiline mesylate	rasagiline mesylate
Ropinirole	Ropinirole
ropinirole hydrochloride	Ropinirole
Requip	ropinirole hydrochloride
requip – xl	ropinirole hydrochloride
ropinirole hydrochloride	ropinirole hydrochloride
Neupro	Rotigotine
Xadago	safinamide mesylate
Emsam	Selegiline
Eldepryl	selegiline hydrochloride
selegiline hydrochloride	selegiline hydrochloride
Zelapar	selegiline hydrochloride
Tasmar	Tolcapone
Tolcapone	Tolcapone
trihexyphenidyl hydrochloride	trihexyphenidyl hydrochloride

Table 4: Exclusion ICD-10-CM codes for the data-driven PD cohort and control cohort (list also referred as diseases like PD):

ICD-10-CM code	Description
G10	Huntington's disease
G12	Spinal muscular atrophy and related syndromes
G12.0	Infantile spinal muscular atrophy, type I [Werdnig-Hoffman]
G12.1	Other inherited spinal muscular atrophy
G12.2	Motor neuron disease
G12.20	Motor neuron disease, unspecified
G12.21	Amyotrophic lateral sclerosis
G12.22	Progressive bulbar palsy
G12.23	Primary lateral sclerosis
G12.24	Familial motor neuron disease
G12.25	Progressive spinal muscle atrophy
G12.29	Other motor neuron disease
G12.8	Other spinal muscular atrophies and related syndromes
G12.9	Spinal muscular atrophy, unspecified
G21	Secondary parkinsonism
G21.0	Malignant neuroleptic syndrome
G21.1	Other drug-induced secondary parkinsonism
G21.11	Neuroleptic induced parkinsonism
G21.19	Other drug induced secondary parkinsonism
G21.2	Secondary parkinsonism due to other external agents
G21.3	Postencephalitic parkinsonism
G21.4	Vascular parkinsonism

G21.8	Other secondary parkinsonism
G21.9	Secondary parkinsonism, unspecified
G23.1	Progressive supranuclear ophthalmoplegia [Steele-Richardson-Olszewski]
G23.2	Striatonigral degeneration
G31.0	Frontotemporal dementia
G31.83	Neurocognitive disorder with Lewy bodies
G31.85	Corticobasal degeneration
G35	Multiple sclerosis
G90.3	Multi-system degeneration of the autonomic nervous system
G91.2	(Idiopathic) normal pressure hydrocephalus

Table 5: Parkinson's disease medications used for the exclusion of donors from the data-driven PD cohort and control cohort (list also referred as antipsychotics):

Proprietary medication name	Nonproprietary medication name
Abilify	Abilify
Abilify	Aripiprazole
abilify asimtufii	Aripiprazole
abilify maintena	Aripiprazole
abilify mycite	Aripiprazole
Aripiprazole	Aripiprazole
Aristada	aripiprazole lauroxil
aristada initio	aripiprazole lauroxil
Asenapine	Asenapine
Secuado	Asenapine
Asenapine	asenapine maleate
asenapine maleate	asenapine maleate
Saphris	asenapine maleate

Brexpiprazole	Brexpiprazole
Rexulti	Brexpiprazole
Vraylar	Cariprazine
Chlorpromazine	Chlorpromazine
chlorpromazine hydrochloride	Chlorpromazine
chlorpromazinum	Chlorpromazine
chlorpromazine hci	chlorpromazine hci
Clozapine	Clozapine
Clozaril	Clozapine
Fazaclo	Clozapine
Versacloz	Clozapine
Haldol	Haloperidol
Haloperidol	Haloperidol
haloperidol lactate	Haloperidol
haldol decanoate	haloperidol decanoate
haloperidol decanoate	haloperidol decanoate
Haldol	haloperidol lactate
Haloperidol	haloperidol lactate
Latuda	lurasidone hydrochloride
lurasidone hydrochloride	lurasidone hydrochloride
Olanzapine	Olanzapine
Zyprexa	Olanzapine
zyprexa zydis	Olanzapine
olanzapine and fluoxetine	olanzapine and fluoxetine
Symbyax	olanzapine and fluoxetine hydrochloride
olanzapine and fluoxetine	olanzapine and fluoxetine
Lybalvi	olanzapine and samidorphan l-malate
Zyprexa	olanzapine pamoate
Invega	Paliperidone

paliperidone	Paliperidone
invega hafyera	paliperidone palmitate
invega sustenna	paliperidone palmitate
invega trinza	paliperidone palmitate
Orap	Pimozide
Pimozide	Pimozide
Quetiapine	Quetiapine
quetiapine - extended release	Quetiapine
quetiapine fumarate	Quetiapine
Seroquel	Quetiapine
Quetiapine	quetiapine extended-release
quetiapine fumarate	quetiapine fumarate
quetiapine fumarate - extended release	quetiapine fumarate
Seroquel	quetiapine fumarate
seroquel - xr	quetiapine fumarate
seroquel xr	quetiapine fumarate
quetiapine fumarate er	quetiapine fumarate er
Perseris	Risperidone
Risperdal	Risperidone
risperdal consta	Risperidone
risperdal m-tab	Risperidone
Risperidone	Risperidone
risperidone m-tab	Risperidone
Rykindo	Risperidone
Uzedy	Risperidone
Geodon	Ziprasidone
Ziprasidone	Ziprasidone
ziprasidone hydrochloride	ziprasidone hcl
Geodon	ziprasidone hydrochloride
Ziprasidone	ziprasidone hydrochloride

Geodon	ziprasidone mesylate
ziprasidone mesylate	ziprasidone mesylate

[00152] Donor cohorts were assessed using *in silico* and manual approaches to confirm the relevance of individuals selected for the study. Due to the large amount of medical information and the number of individuals considered for this study, an *in silico* framework was used to identify outliers in the donors-PD and controls cohorts for an in-depth evaluation of their medical information. The *in silico* approach leveraged network analysis of donor similarities of medical history based RWD. This analysis encompasses the evaluation of comorbidities and medication patterns, facilitating the identification and exclusion of mislabeled donors and instances of atypical PD. Furthermore, the health journey of each donor was reviewed manually by PD experts, leading to the final selection of donors who have developed PD.

[00153] Adhering to the concept of digital twins, a similar number of paired control samples were identified based on demographics and sample availability. Co-diagnoses for major diseases such as diabetes and hypertension were matched and patients with all major central nervous system disorders (Table 3) were excluded from the control group. In total, ~2600 samples from ~700 individuals were selected for deep molecular profiling, ensuring a comprehensive and representative dataset for advancing the understanding of PD.

Example 5: In-depth characterization of PD development in subjects from the plasmapheresis biobank using real world data

[00154] To uncover novel analytes of PD and identify potential therapeutic targets and pathways, it was imperative to gain a precise understanding of PD development and its associated co-morbidities. The limitations of smaller data sets that had been available in conventional resources were overcome by analyzing RWD from an extensive cohort of samples from the plasmapheresis biobank in conjunction with claims data and EHR of approximately 1.3 million individuals that have developed PD, providing a unique opportunity to decipher the multifaceted nature of PD onset and progression.

[00155] Deep molecular profiling of 2609 samples from PD and control subjects was performed by deep proteomics profiling of individuals developing PD and their respective

paired control samples (See Figure 16). The proteomics used for analysis of these samples included the previously used Olink Explore HT (5k) (Olink, Uppsala, Sweden), and SomaScan assay v5 (Somalogic, Boulder, CO, USA) platforms as well as Biognosys™ TrueDiscovery™, a platform that uses mass spectrometry to provide proteomics solutions for drug development (Biognosys, Newton, MA, USA) and the NULISAseq CNS Disease Panel from Alamar Biosciences, a CNS disease-targeted panel using multiplexed quantification (Alamar Biosciences, Fremont, CA, USA). These different proteomics platforms have been shown to provide complementary signals, and their combination allows the investigation of around 15k proteins. These molecular changes were contextualized with the medical records from RWD to integrate the molecular and proteomic analysis with the clinical and phenotypic data available for the subjects.

[00156] Samples were identified at all phases of PD progression, including prior to identified clinical onset (Figures 17A and 17B). 348 PD donors with a total of 1323 samples were identified. 245 individuals (70%) associated with a total of 913 samples were identified with at least one sample donated prior to the first doctor's visit associated with a G20 code. 135 individuals (39%) associated with a total of 410 samples were identified with at least one sample donated after the first doctor visit associated with a G20 code. 32 individuals (95) associated with a total of 207 samples were identified with at least one sample donated before and at least one sample donated after the first doctor visit were associated with a G20 code.

[00157] Partial least squares discriminant analysis (PLS-DA) was performed on both the PD and control samples. The PLS-DA showed proteomic variance in all samples over time, but also showed particular variance of certain proteins in the PD samples as compared to their matched control samples (Figure 18).

[00158] Although the foregoing invention has been described in some detail by way of illustration and example for purposes of clarity of understanding, it is readily apparent to those of ordinary skill in the art in light of the teachings of this invention that certain changes and modifications may be made thereto without departing from the spirit or scope of the appended claims.

[00159] Accordingly, the preceding merely illustrates the principles of the invention. It will be appreciated that those skilled in the art will be able to devise various arrangements which, although not explicitly described or shown herein, embody the principles of the

invention and are included within its spirit and scope. Furthermore, all examples and conditional language recited herein are principally intended to aid the reader in understanding the principles of the invention and the concepts contributed by the inventors to furthering the art, and are to be construed as being without limitation to such specifically recited examples and conditions. Moreover, all statements herein reciting principles, aspects, and embodiments of the invention as well as specific examples thereof, are intended to encompass both structural and functional equivalents thereof. Additionally, it is intended that such equivalents include both currently known equivalents and equivalents developed in the future, i.e., any elements developed that perform the same function, regardless of structure. Moreover, nothing disclosed herein is intended to be dedicated to the public regardless of whether such disclosure is explicitly recited in the claims.

[00160] The scope of the present disclosure, therefore, is not intended to be limited to the exemplary embodiments shown and described herein. Rather, the scope and spirit of present disclosure is embodied by the appended claims. In the claims, 35 U.S.C. §112(f) or 35 U.S.C. §112(6) is expressly defined as being invoked for a limitation in the claim only when the exact phrase "means for" or the exact phrase "step for" is recited at the beginning of such limitation in the claim; if such exact phrase is not used in a limitation in the claim, then 35 U.S.C. § 112 (f) or 35 U.S.C. §112(6) is not invoked.

What is claimed is:

1. A method for identifying analytes that are associated with a phenotype of interest using a computer system that includes one or more processors and system memory, the method comprising:

selecting, using the one or more processors, two or more longitudinal data sets in a database derived from one or more longitudinal sample sets, wherein the longitudinal sample sets comprise samples from individual subjects collected over a designated time frame;

determining one or more analyte scores from the plurality of analytes detected in the longitudinal data sets using the one or more processors, wherein the analyte scores are correlated with the phenotype of interest; and

identifying one or more analytes associated with the phenotype of interest based on the analyte scores for each longitudinal data set using the one or more processors.

2. The method of claim 1, wherein the data derived from the samples of the longitudinal sample sets are stored in the database with additional information that can be associated with the sample donors.

3. The method of claim 2, wherein the additional information in the database regarding the samples of the longitudinal sample sets is de-identified.

4. The method of claim 1, further comprising obtaining longitudinal data sets from two or more longitudinal sample sets, wherein the longitudinal data sets are stored in a database.

5. A computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed using the one or more processors, cause the computer system to implement a method for identifying genes that are associated with a disease-related phenotype of interest, the method being any one of the methods of claims 1-4.

6. A computer system, comprising:

one or more processors;

system memory; and

one or more computer-readable storage media having stored thereon computer-executable instructions that, when executed using the one or more processors, cause the computer system to implement a method for identifying longitudinal data sets that are associated with a phenotype of interest, the method including:

(a) selecting, using the one or more processors, two or more longitudinal data sets in a database derived from one or more longitudinal sample sets, wherein the longitudinal sample sets comprise samples from a plurality of subjects collected over a designated time frame;

(b) determining one or more analyte scores from the plurality of analytes detected in the longitudinal data sets using the one or more processors, wherein the analyte scores are correlated with the phenotype of interest;

(c) identifying one or more analytes associated with the phenotype of interest based on the analyte scores for each longitudinal data set using the one or more processors.

7. The system of claim 6, wherein the method for identifying longitudinal data sets that are associated with a phenotype of interest further utilizes additional information in the database regarding the samples of the longitudinal sample sets.

8. The system of claims 6 and 7, wherein the method for identifying longitudinal data sets that are associated with a phenotype of interest further comprises additional information that can be associated with the sample donors.

9. The system of claims 7 and 8, wherein the additional information in the database regarding the samples of the longitudinal sample sets is de-identified.

10. A library of plasmapheresis samples, comprising:

a plurality of longitudinal sample sets, wherein a longitudinal sample set comprises multiple samples obtained from the same donor over a period of time, and wherein the samples are obtained by plasmapheresis.