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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 January 2020 (09.01.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/008192 A2

(51) International Patent Classification:
G16B 20/00 (2019.01)

(21) International Application Number:
PCT/GB2019/051885

(22) International Filing Date:
03 July 2019 (03.07.2019)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1810897.7 03 July 2018 (03.07.2018) GB

(71) Applicant: **CHRONOMICS LIMITED** [GB/GB]; 1 St James Court, Norwich Norfolk NR3 1RU (GB).

(72) Inventors: **MARTIN-HERRANZ, Daniel**; 1 St James Court, Norwich Norfolk NR3 1RU (GB). **DIANES-SANTOS, Jose**; 1 St James Court, Norwich Norfolk NR3 1RU (GB). **STUBBS, Thomas**; 1 St James Court, Norwich Norfolk NR3 1RU (GB).

(74) Agent: **BARKER BRETTELL LLP**; 100 Hagley Road, Edgbaston, Birmingham West Midlands B16 8QQ (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: PHENOTYPE PREDICTION

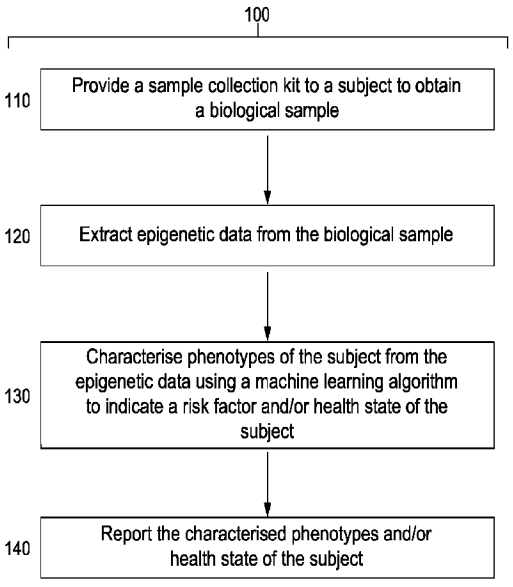

FIGURE 3

(57) Abstract: A method of characterising phenotypes of a subject is provided. The method comprises characterising one or more phenotypes of a subject from epigenetic data of the subject using a machine learning algorithm (130). The one or more phenotypes are influenced by environmental factors and are risk factors for one or more diseases.

WO 2020/008192 A2

[Continued on next page]

Published:

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

PHENOTYPE PREDICTION

FIELD OF THE INVENTION

The invention relates to prediction of phenotypes of a subject, and more particularly to prediction of phenotypes of a subject based on epigenetic data in order to determine a health state of the subject.

BACKGROUND OF THE INVENTION

Epigenetics is normally defined as the study of mitotically and/or meiotically stable heritable changes in gene function that cannot be explained by changes in DNA sequence. Epigenetic mechanisms include DNA methylation, histone modifications and non-coding RNAs (see Figure 1). Together those mechanisms regulate gene expression and allow the creation of different cell types and functions starting from the same set of genes, which is needed to build complex multi-cellular organisms from a single blueprint.

Unlike the human genome (i.e., DNA), which is fixed from birth, epigenetic marks change during the lifetime of a person upon exposure to different cues, including intrinsic signals (e.g. hormone responses triggered by a certain mental state), extrinsic environmental factors (e.g. exposure to air pollution or chemical compounds) or life choices (e.g. type of diet, smoking, etc.). Therefore, epigenetic information constitutes a biological layer that holistically captures the complex state of a biological system, combining multiple effects, including genetics, genetic-environmental interactions or the cellular composition of the tissue. Given this complexity, it is difficult to assign causality or mechanisms to epigenetic measurements, which has been the focus of many research studies in the academic community.

SUMMARY OF THE INVENTION

According to an aspect of the invention, there is provided a method of characterising phenotypes of a subject from epigenetic data of the subject using a machine learning algorithm. The one or more phenotypes may be influenced by environmental factors. The one or more phenotypes may be risk factors for one or more diseases. The method may provide a person with a risk of developing a condition before the condition is developed or begins to develop. This may enable pre-clinical, preventative healthcare

and/or health management to be implemented to reduce a risk of a condition developing.

The method may further comprise obtaining epigenetic data of the subject. Obtaining epigenetic data of the subject may comprise extracting epigenetic data from a biological sample taken from the subject. Alternatively, obtaining epigenetic data may comprise receiving epigenetic data of the subject (for example, previously extracted from a biological sample taken from the subject). The epigenetic data of the subject may be or comprise one or more data files.

The one or more phenotypes may be or comprise one or more of biological age, smoke exposure, and metabolic state. In particular, higher biological ages are associated with a higher risk of developing age-related diseases (such as different types of cancer, type II diabetes, cardiovascular disease or neurodegenerative diseases) and all-cause mortality. Therefore, biological age is a particularly illuminative risk factor or environmental phenotype. It is also possible to slow down the processes that lead to increases in biological age, making biological age a useful metric of the effectiveness of lifestyle or environmental interventions.

The machine learning algorithm may be or comprise a random forest algorithm. The machine learning algorithm may be trained using epigenetic data (for example, DNA methylation data). The machine learning algorithm may be trained on a subset of epigenetic data selected from a set of epigenetic data. This may reduce the overall number of features used to train the machine learning algorithm. The subset of epigenetic data may be selected using a machine learning algorithm (for example, a different random forest algorithm). An output of a first machine learning algorithm may therefore provide or be used as an input for a second machine learning algorithm, wherein the second machine learning algorithm is trained and/or configured to characterise phenotypes of the subject from epigenetic data of the subject. The first machine learning algorithm may be the same type of machine learning algorithm as the second machine learning algorithm. For example, the first machine learning algorithm and the second machine learning algorithm may both be a random forest algorithm. The subset of epigenetic data may be selected based on feature importance. The subset of epigenetic data may be selected using a threshold. The threshold may be a pre-determined threshold, for example a set value (for example, absolute feature

importance). Alternatively, the threshold may be based on statistics of the set of epigenetic data (for example, relative feature importance).

The epigenetic data may be derived from a biological sample of the subject. The biological sample may be at least one of urine, blood or semen. Alternatively, the biological sample may be saliva. A biological sample of saliva is easy to obtain via a non-invasive sampling process. Epigenetic profiles (for example, a DNA methylation profile) of saliva also correlate well with epigenetic profiles found in blood and brain tissues.

According to a second aspect of the invention, there is provided a method of determining and providing information regarding a health state of a subject, the method comprising: providing a sample collection kit to the subject to obtain a biological sample; extracting epigenetic data from the biological sample; characterising phenotypes of the subject from the epigenetic data using a machine learning algorithm to indicate the health state of the subject; and reporting the characterised phenotypes and/or health state of the subject.

Such a method of determining and providing information regarding a health state of a subject may enable key information relating to the health of a subject to be communicated without requiring a face-to-face interaction between the user and a medical professional. The method may provide a person with a risk of developing a condition before the condition is developed or begins to develop. This may enable pre-clinical, preventative healthcare and/or health management to be implemented to reduce a risk of a condition developing. This may be particularly useful for health-conscious individuals who are interested in understanding their past, current or future health states without requiring a medical professional to diagnose the individual with a particular medical condition. Such a method may also be useful for caregivers providing care to individuals who are primarily or indefinitely based at a location from which travelling to a medical professional requires significant effort and/or time. In cases where professional medical care is not necessary, it may be quicker and/or simpler for the caregiver to use the method to obtain insights into the health state of the individual (i.e., the subject), rather than requiring an opinion from a medical professional. In this way, both the cost of, and time associated with the care may be reduced.

The method may further comprise proposing one or more medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or health state of the subject.

The proposed medical and/or lifestyle interventions may be personalized with regard to the specific characterised phenotypes and/or the health state of the subject.

The method may further comprise: extracting genetic data from the biological sample of the subject; and characterising phenotypes of the subject from at least the genetic data and the epigenetic data using the machine learning algorithm to indicate the health state of the subject.

Using genetic data in conjunction with the epigenetic data to characterise phenotypes using the machine learning algorithm may enable the genetic predisposition of the subject to certain phenotypes to be compared with the phenotypes actually exhibited by the subject. As such, the genetic profile of the subject may be compared with the epigenetic profile of the subject which governs the genetic expression of the subject.

The method may further comprise: obtaining at least one of lifestyle data, health data, well-being data and environmental data from the subject; and characterising phenotypes of the subject from the at least one of lifestyle data, health data, well-being data and environmental data and the epigenetic data using the machine learning algorithm to indicate the health state of the subject.

The accuracy and personalisation of the characterised phenotypes and/or the health state of the subject as determined by the machine learning algorithm may be improved by utilising additional data specific to the subject, for example electronic health records.

The lifestyle data, health data, well-being data and environment data may comprise at least one of microbiome data, metabolomic data, proteomic data, imaging data, medical or clinical records from the subject and close relatives (including information about past and current diseases or conditions), age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue

levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth), information from social networking platforms and/or social media platforms, and chronotype.

Both strands of DNA from the biological sample of the subject may be used to extract genetic information from the biological sample. Using both strands of DNA from the biological sample of the subject may enable epigenetic data to be extracted from both sides of the DNA helix of the subject, whilst simultaneously extracting genetic data from the biological sample.

The biological sample may comprise urine, blood or semen or another biological material. The biological sample may comprise saliva obtained from the subject. Saliva may provide a good epigenetic profile for analysis, whilst being an easy and non-invasive tissue for obtaining a biological sample from the subject. Saliva may also be easily collected and stored in the sample collection kit provided to the subject.

The characterised phenotypes comprise environmental phenotypes. By characterising environmental phenotypes of the subject and not only pathological phenotypes of the subject, a greater insight into environmental factor correlation and possible causation for particular medical conditions or diseases may be obtained. The characterised phenotypes may comprise at least one of microbiome data, metabolomic data, proteomic data, information about past and current diseases or conditions, age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and

fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

The lifestyle data, health data, well-being data and environmental data may comprise at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking sites and social media sites, and information uploaded from third-party providers.

By obtaining additional data from a wide range of information sources, the accuracy and relevance of the characterised phenotypes and/or the health state of the subject as determined by the machine learning algorithm may be improved. Furthermore, utilising information from sources which are widely available and frequently used by a large number of people allows subjects to provide additional data without changing personal routines or habits to any great degree.

The method may further comprise repeating the method at a plurality of points in time to indicate a stability of the characterised phenotypes and/or health state of the subject over time. By obtaining data at a plurality of points in time, a feedback loop can be generated by tracking changes in the phenotypes of the subject with time, and comparing those phenotypes to updated characterised phenotypes determined by the machine learning algorithm based on more recent epigenetic data from the subject. Improvements in the characterised phenotypes and/or the health state of the subject may therefore be quantitatively or qualitatively tracked over a period of time using longitudinal epigenetic data and additional data.

The method may further comprise refining the machine learning algorithm using at least one of epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data obtained from the subject. In this way, the predictive power of the machine learning algorithm to characterise phenotypes of the subject from the epigenetic data of the subject may be improved, without requiring an extensive

additional training process requiring a new set of training data. Periodic updates of the machine learning algorithm may be enabled by periodically providing at least one of epigenetic data, lifestyle data, health data, well-being data and environmental data obtained from the subject.

The method may further comprise repeating the method a pre-determined time period after proposing the one or more medical and/or lifestyle interventions. In this way, subjects may receive quantitative and qualitative assessment of the effectiveness of medical and/or lifestyle interventions without requiring subjection to any invasive, time consuming or stressful medical procedures.

Characterising phenotypes of the subject may comprise at least one of determining current phenotypes of the subject, and predicting past and/or future phenotypes of the subject. Characterising phenotypes of the subject may comprise at least one of calculating a value of a continuous variable within a phenotypic class, and calculating a probability of the subject belonging to a phenotypic class. Subjects may be able to determine improvements in characterised phenotypes and/or health state, and may also be able to identify phenotypic variable or phenotypic risk targets to aim towards in the future. Subjects may also be able to see differences in phenotypic variables or phenotypic risks with and without proposed medical and/or lifestyle interventions.

The extracted epigenetic data may comprise epigenetic data extracted from at least 1 million CpG sites. The epigenetic data may be captured from the DNA of the subject using a pull-down approach. The pull-down approach may be based on oligonucleotide probes.

Reporting the characterised phenotypes and/or health status of the subject may comprise reporting the characterised phenotypes and/or health status of the subject to at least one of the subject, a medical professional, a significant other, family member or next of kin of the subject, and a third party such as an employer or an insurer.

According to a third aspect of the invention, there is provided a method of developing a machine learning algorithm configured to characterise phenotypes of a subject, the method comprising: providing a sample collection kit to each of a population of individuals to obtain biological samples from at least a subset of the population; extracting epigenetic data from at least a subset of the biological samples; obtaining at least one of lifestyle data, health data, well-being data and environmental data from at

least a subset of the population; collating at least a part of the epigenetic data and at least a part of the at least one of lifestyle data, health data, well-being data and environmental data in a training data set; training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set.

The lifestyle data, health data, well-being data and environmental data may comprise at least one of microbiome data, metabolomic data, proteomic data, imaging data, medical or clinical records from the subject and close relatives (including information about past and current diseases or conditions), age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

The lifestyle data, health data, well-being data and environmental data may comprise at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking platforms and social media platforms, and information uploaded from third-party providers.

The method may further comprise: extracting genetic data from at least a subset of the biological samples; collating the genetic data in the training data set; and training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set. In this way, the machine learning algorithm may be trained to

identify how characterised phenotypes relate to genetic predispositions of the subject to those characterised (or additional uncharacterised) phenotypes.

The method may further comprise: repeating the steps of providing a sample collection kit to a population of individuals to obtain biological samples from at least a subset of the population, extracting epigenetic data from at least a subset of the biological samples, and obtaining at least one of lifestyle data, health data, well-being data and environmental data from at least a subset of the population at a plurality of different points in time, to obtain longitudinal epigenetic data and at least one of longitudinal lifestyle data, longitudinal health data, well-being data and longitudinal environmental data; collating the longitudinal epigenetic data and the at least one of longitudinal lifestyle data, longitudinal health data, well-being data and longitudinal environmental data in the training data set; and training the machine learning algorithm to characterise predicted past and/or future phenotypes from epigenetic data using the training data set. By training the machine learning algorithm to predict past or future phenotypes, this data can be used to augment the existing training data set obtained from the population of individuals, thereby potentially improving the accuracy and predictive power of the machine learning algorithm. Additionally, subjects may be informed as to a future phenotypic risk depending on whether or not medical and/or lifestyle interventions are adhered to. Furthermore, changes or improvements in characterised phenotypes compared to a previous point in time may be predicted from the epigenetic data from the subject.

The method may further comprise collating at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the first aspect of the invention in the training data set; and further training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set.

In this way, the predictive power of the machine learning algorithm to characterise phenotypes of the subject from the epigenetic data of the subject may be improved, without requiring an extensive additional training process requiring a new set of training data. Periodic updates of the machine learning algorithm may be enabled by periodically providing at least one of epigenetic data, lifestyle data, health data, well-being data and environmental data obtained from the subject.

The phenotypes the machine learning algorithm is configured to characterise may comprise at least one of microbiome data, metabolomic data, proteomic data, information about past and current diseases or conditions, age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) chronotype or other phenotypic data.

According to a fourth aspect of the invention, there is provided a digital platform for determining a health state of a subject, the platform comprising: a data storage module, the data storage module configured to store epigenetic data of the subject; a data analysis module in communication with the data storage module, the data analysis module configured to use a machine learning algorithm to characterise phenotypes of the subject from the epigenetic data of the subject to indicate a health state of the subject; and a user module in communication with the data analysis module, the user module configured to display the characterised phenotypes and/or the health state of the subject on the user device and controllable by a user via the user device.

A digital platform configured to store and analyse data of a subject, and report pertinent characterised phenotypes and a health state of the subject to a user of the digital platform, may enable key information relating to the health and well-being of a subject to be communicated to a user of the digital platform without requiring a face-to-face interaction between the user and a medical professional. This may be particularly useful for health-conscious individuals who are interested in understanding their past, current or future health states without requiring a medical professional to diagnose the individual with a particular medical condition. Such a

digital platform may also be useful for caregivers providing care to individuals who are primarily or indefinitely based at a location from which travelling to a medical professional requires significant effort and/or time. In cases where professional medical care is not necessary, it may be quicker and/or simpler for the caregiver to use the digital platform to obtain insights into the health state of the individual (i.e., the subject), rather than requiring an opinion from a medical professional. In this way, both the cost of, and time associated with the care may be reduced whilst simultaneously providing personalized health insights.

The user module may be further configured to display on the user device one or more proposed medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or health state of the subject.

The user or subject may take action based on the proposed medical and/or lifestyle interventions in order to improve the health state of the subject, and/or may be able to reduce a phenotypic variable or phenotypic risk of the subject.

The data storage module may be further configured to store at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject.

The lifestyle data, health data, well-being data and environmental data of the subject may comprise at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking platforms and social media platforms, and information uploaded from third-party providers.

The digital platform may further comprise a training module in communication with the data storage module, the data analysis module and the user module. The training module may be configured to: selectively update the machine learning algorithm used by the data analysis module using epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject obtained from the data storage module; and provide an updated machine

learning algorithm to characterise phenotypes of the subject from epigenetic data to indicate a health state of the subject to the data analysis module.

Data provided by the user to the data storage module 710 can be used to expand a population from which a training data set for training the machine learning algorithm may be selected. Periodically (e.g., once a month, once every three months, once every six months, annually) retraining the machine learning algorithm using additional data provided by the user enables the predictive power and accuracy of the machine learning algorithm to characterise phenotypes from epigenetic data to be continuously improved. In this way, a plurality of users may benefit from the improved predictive and characterising power of the machine learning algorithm updated using the subject data of one or more others users stored in the data storage module 710.

The digital platform may further comprise a first security module. The first security module may be configured to: determine whether or not the training module has been granted permission by the user, via the user module, to obtain epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module; and allow the training module to obtain epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module only if the first security module has determined that the training module has been granted permission to do so.

The user module may be in communication with the data storage module. The user module may be further configured to selectively provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module.

The user module may be further configured to selectively: automatically access and retrieve at least one of lifestyle data, health data, well-being data and environmental data; and automatically provide the at least one of lifestyle data, health data, well-being data and environmental data to the data storage module.

The user module being configured to automatically access and retrieve data from sources of lifestyle data, health data, well-being data and environmental data, may reduce the burden on the user to provide the information from each source of data to

the data storage module separately. As such, longitudinal data obtained from one or more sources of lifestyle data, health data, well-being data and environmental data may be periodically or constantly updated and provided to the data storage module automatically. Furthermore, the user module may retrieve data that the user would not consider reporting to the data storage module, or data that the user may be unaware is being collated in sources of lifestyle data, health data, well-being data and environmental data (although not in breach of the personal privacy of the user).

The digital platform may further comprise a second security module. The second security module may be configured to: determine whether or not the user module has been granted permission by the user to provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module; and allow the user module to provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module only if the second security module has determined that the user module has been granted permission to do so.

The first security module may be further configured to anonymise the at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject after determining that the training module has been granted permission to obtain the at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module.

The second security module may be further configured to anonymise the at least one of lifestyle data, health data, well-being data and environmental data of the subject after determining that the user module has been granted permission to provide the at least one of lifestyle data, health data, well-being data and environmental data to the data storage module.

Users may be more willing to share subject data knowing that a component of the digital platform provides security against the unwarranted sharing of subject data with third parties outside of the service provided by the digital platform.

The user module may be further configured to display one or more of a plurality of apps, each app configured to display one or more related characterised phenotypes on the user device.

The user module may be further configured to enable a user to select one or more of the plurality of apps to be displayed simultaneously on the user device. In this way, the user may be able to view, compare and contrast different aspects of the health state of the subject by simultaneously viewing one or more different characterised phenotypes or groups of characterise phenotypes on the user device. The user may be able to see which phenotypic variable or phenotypic risk is the most important or relevant from a health perspective according to the information displayed by one or more of the plurality of apps. The user may then determine which proposed medical and/or lifestyle intervention may have the greatest effect on reducing and/or improving the phenotypic variable or phenotypic risk of the subject, in order to improve the health state of the subject.

The digital platform may be further configured to motivate the user to share subject data by offering incentives and/or rewards in return for the user granting permission, via the user module, for at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject stored in the data storage module to be shared with third parties for research purposes.

The digital platform may be further configured to motivate the user to share subject data by offering incentives and/or rewards in return for the user granting permission for the training module to obtain at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject stored in the data storage module.

Obtaining and collating data relating to a health state of a subject can be difficult without incentive. In particular, building databases which comprise large amounts of data taken from a large cross-section of individuals in a population can be difficult without incentive. It can be difficult to stress the benefits of sharing such data to individuals in a population to convince such individuals to share their data. By offering incentives and/or rewards for users to share subject data via the digital platform, users may be encouraged or motivated to share subject data. As a result,

third party research projects may benefit greatly from a larger pool of data (e.g., results obtained from such studies may be much more reliable or accurate), and may not need to resort to data augmentation. Furthermore, the machine learning algorithm used by the data analysis module may be updated and/or refined by training the machine learning algorithm further using additional data shared by users. This may enable the machine learning algorithm to improve its characterising and/or predictive power to identify phenotypes from epigenetic data.

The incentives and/or rewards offered by the digital platform may comprise additional services and/or functionality provided by the digital platform. The incentives and/or rewards may comprise a currency. The currency may be configured to be exchanged for additional services and/or functionality provided by the digital platform, and/or withdrawn from the digital platform by the user.

The user may be able to obtain a greater insight into the characterised phenotypes and/or the health state of the subject by being granted access to, or exchanging currency for, additional services and/or functionalities that would otherwise not be provided by the digital platform. As such, the user may be able to access different (e.g., more specific and personalized) medical and/or lifestyle interventions that may act to improve the health state of the subject than would otherwise be available through the digital platform.

The currency may be an internal currency that is only recognised by the digital platform.

Two or more of the modules of the digital platform may be in wireless communication with one another. For example, the data storage module and/or the data analysis module may be located remotely (e.g., accessible via cloud computing) from the user module, which may be located locally on a user device. The data storage module and/or the data analysis module may therefore be in wireless communication with both one another and the user module.

The optional features from any aspect may be combined with the features of any other aspect, in any combination. For example, the method of the first aspect may comprise any one or more of the features described with reference to the method of the second

aspect and/or third aspect, and vice versa. Furthermore, the digital platform of the fourth aspect may comprise any of the optional features described with reference to the methods of any of the first, second and third aspects, and vice versa. Features may be interchangeable between different aspects and embodiments and may be removed from different aspects and embodiments and may be added to different aspects and embodiments.

Features which are described in the context of separate aspects and embodiments of the invention may be used together and/or be interchangeable wherever possible. Similarly, where features are, for brevity, described in the context of a single embodiment, those features may also be provided separately or in any suitable sub-combination. Features described in connection with the methods of the first, second and third aspects may have corresponding features definable with respect to the digital platform of the fourth aspect, and these embodiments are specifically envisaged.

BRIEF DESCRIPTION OF THE FIGURES

The invention will now be described by way of example with reference to the accompanying drawings in which:

FIG. 1 shows a schematic of addition of a methyl group to cytosine letters in human DNA by a DNA methyltransferase or DNMT (Jeremy J Day and J David Sweatt. DNA methylation and memory formation. *Nature Neuroscience*, 13(11):1319, 2010;

FIG. 2 shows a schematic of changes in human DNA methylation through human lifespan (Martin Widschwendter, Allison Jones, Iona Evans, Daniel Reisel, Joakim Dillner, Karin Sundström, Ewout W. Steyerberg, Yvonne Vergouwe, Odette Wegwarth, Felix G. Rebitschek, Uwe Siebert, Gaby Sroczynski, Inez D. de Beaufort, Ineke Bolt, David Cibula, Michal Zikan, Line Bjørge, Nicoletta Colombo, Nadia Harbeck, Frank Dud-bridge, Anne-Marie Tasse, Bartha M. Knoppers, Yann Joly, Andrew E. Teschendorff, and Nora Pashayan. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nature Reviews Clinical Oncology*, 15:292–309, 2018);

FIG. 3 shows a schematic of a process for determining and providing information regarding a health state of a subject using epigenetic data;

FIG. 4 shows a schematic of a process for extracting at least epigenetic data from a biological sample of a subject;

FIG. 5 shows charts depicting phenotypic variables and/or phenotypic risks based on characterised phenotypes of a subject;

FIG. 6 shows a schematic of a process for developing a machine learning algorithm configured to characterise phenotypes of a subject using epigenetic data;

FIG. 7 shows a schematic of a method for training a machine learning algorithm to characterise phenotypes from epigenetic data;

FIG. 8 shows a schematic of a further method for training a machine learning algorithm to characterise phenotypes from epigenetic data;

FIG. 9 shows the predictions for the biological age environmental phenotype using DNA methylation data for a population of individuals. 9A: test set in the first random forest (used for feature selection); 9B: test set in the second random forest (trained on the selected features).

FIG. 10 shows the predictions for the smoke exposure environmental phenotype using DNA methylation data for a population of individuals. The distributions of the probabilities of belonging to a specific class of self-reported smoking status are displayed in different colours.

FIG. 11 shows the predictions for the metabolic state environmental phenotype using DNA methylation data for a population of individuals. The distributions of the probabilities of belonging to a specific class of self-reported BMI range are displayed in different colours.

FIG. 12 shows how different environmental phenotypes (which behave as risk factors or pre-disease health states and can be estimated using epigenetic data) can contribute in different proportions to the development of different diseases (or pathological phenotypes), which in turn affect mortality rate.

FIG. 13 shows a schematic of a digital platform for determining a health state of a subject;

FIG. 14 shows an example of an "app" of a digital platform for determining a health state of a subject displayed on a user device through which the digital platform is accessed; and

FIG. 15 shows an example of a "home" page of a digital platform for determining a health state of a subject displayed on a user device through which the digital platform is accessed.

Features which are described in the context of separate aspects and embodiments of the invention may be used together and/or be interchangeable wherever possible. Similarly, where features are, for brevity, described in the context of a single embodiment, these may also be provided separately or in any suitable sub-combination. Features described in connection with the method of the first aspect may have corresponding features definable with respect to the method of the second aspect and the digital platform, and these embodiments are specifically envisaged.

DETAILED DESCRIPTION OF THE INVENTION

In this specification, the term "machine learning algorithm" refers to an algorithm developed using a machine learning process. A "machine learning algorithm" developed using one or more sets of training data may be further updated or refined using further training data. Similarly, "machine learning" refers to a process in which an algorithm is automatically adapted or updated (e.g., by a computer) in response to an error in an output of the algorithm.

There are several ways in which epigenetic information can be encoded in cells. The archetypal among them is DNA methylation, the addition of a methyl group to the cytosine letters in human DNA (see Figure 1). In humans, this predominantly occurs at cytosine bases followed by guanine bases (commonly referred to as in CpG context), of which there are more than 28 million instances in the human genome. A methyl group (CH_3) is added to cytosine by a DNA methyltransferase (DNMT) to form 5-methylcytosine, the most commonly modified base in the mammalian genome.

In a given DNA strand of a given cell, the methylation readout of each one of these CpG sites is a binary variable: the site is either methylated or not. However, when a sample is taken from a tissue many cells are analysed at the same time. That is why DNA methylation in a CpG site is normally expressed as a percentage, which

approximately reflects how many cells in that tissue are methylated at that genomic location.

DNA methylation levels have been shown to be influenced by both genetics and the environment, representing a very attractive data source to monitor and predict complex phenotypes. This is shown in Figure 2, wherein DNA methylation is obtained from surrogate tissues (such as blood, buccal cells, and cervical cells). DNA methylation is influenced by many factors during a lifetime, of which various stages are shown in Figure 2. In particular, Figure 2 illustrates how DNA methylation (where methylated cytosines are shown by black circles 10, and unmethylated cytosines are shown by white circles 20) is typically affected by the following factors (although other factors may also affect DNA methylation): i) lifestyle of previous generations (such as father's diet), ii) intrauterine environment during early development (such as exposure to diethylstilbestrol), iii) genetic background (such as meQTLs – methylation Quantitative Trait Loci), iv) environment (such as smoking), v) reproductive factors (such as pregnancy), and iv) lifestyle (such as obesity). Figure 2 shows how DNA methylation changes with age through early development, childhood, puberty, adult life (including pregnancy) and old age.

Currently, risk models only include epidemiological factors, and those models do not enable differentiation of individuals with good and poor prognosis. By integrating epigenetic and genetic information, the interaction of an individual with environmental factors (for example stress, nutrition, smoking and/or absence of physical exercise), which are currently key components of multivariable risk algorithms, can be measured at the molecular level.

DNA methylation is also a biologically stable and reproducible matter, and is easily obtained from a number of surrogate organs and tissues (for example, saliva and blood). DNA methylation also enables monitoring of the effectiveness of risk-reducing interventions.

Different cells have different epigenomes. As such, different cells, starting with the same genetic code, can express different subsets of genes and specialize in their function, such as transmitting nerve impulses in the case of neurons or secreting antibodies in the case of B cells. Therefore, since different tissues are composed of different cell types, they show different profiles when assessed for their DNA methylation patterns.

Common tissues used to obtain genetic and epigenetic information include blood, urine, semen and saliva. Saliva in particular represents an attractive choice for obtaining genetic and/or epigenetic information because it contains a mix of buccal epithelial cells and leukocytes, from which DNA can be successfully extracted. Saliva is also easy to obtain via a non-invasive sampling process (e.g., an oral swab or spitting into a container), and is easy to use during testing.

DNA methylation profiling in saliva samples is becoming increasingly common and has proven useful in studies looking at many different diseases and conditions, such as Parkinson's, respiratory allergy, attention-deficit/hyperactivity disorder, head and neck cancer and obesity. DNA methylation profiles of saliva are also correlated with those in blood ($R^2 = 0.97$) and brain ($R^2 = 0.86$) tissues, which makes saliva a useful surrogate tissue to capture effect associated with many complex traits.

Buccal cells also seem to exhibit more stable DNA methylation profiles in longitudinal studies when compared with other tissues.

The most widely used single base resolution methods for measuring DNA methylation currently rely on a chemical reaction step called bisulfite conversion. Treatment of DNA with sodium bisulfite enables cytosines to be differentiated based upon whether they are methylated or unmethylated. Methylated cytosines remain unchanged during bisulfite conversion, whereas unmethylated cytosines are converted into the base uracil (U) and then, after PCR amplification (with A,C,G,T deoxynucleotides), into thymine (T). The amplified sequencing library can then be sequenced or hybridized, and the data processed to map methylation patterns.

There are two main types of technologies that are used to read out DNA methylation genome-wide at single-base pair resolution: next-generation bisulfite sequencing and Infinium methylation arrays. A comparison of the main features of the two technologies is shown in Table 1.

Feature	Next-generation bisulfite sequencing	Illumina Infinium methylation arrays
Core technology	Next-generation DNA sequencing	Hybridisation
Number of CpG sites assayed	Customisable	27000, 450000 (450K), or 850000
Cost	Customisable	Fixed
Amount of input DNA	10 ng to 3 µg[1]	500 ng to 5 µg[2]
Reproducibility[3]	0.97	0.98
Captures genetic information?	Yes	No
Allele-specific methylation?	Yes	No

[1] Depending on the number of CpG sites assayed

[2] This precludes the use of Infinium methylation arrays for scarce samples, such as micro-dissected cancer biopsies

[3] Pearson correlation coefficient among technical replicates

Next-generation sequencing technology enables specific targeting of CpG sites to be sequenced, resulting in a minimized sequencing cost. This can be achieved via a pull-down approach, in which oligonucleotide probes are used to enrich for specific regions in the genome.

Next-generation sequencing technology allows capturing and sequencing both strands of DNA, which enables accurate recovery of genetic information together with epigenetic information. This allows individuals to be informed of both epigenetic data, and genetic data (such as ancestry estimates and predisposition to certain diseases).

Most of the publicly available epigenetic data has been generated using the Illumina Infinium methylation array 450K. Next-generation sequencing technology can overlap

its coverage with the CpG sites present in the 450K array (e.g. with a specific pull-down approach over 90% overlap with 450K array). In addition to these 450,000 sites, therefore, next-generation sequencing technology allows for the interrogation of up to a further 4.5 million CpG sites for the same price per sample as the Illumina Infinium methylation array 450K. This enables many of the discoveries already made in the literature using the Illumina Infinium methylation array 450K to be translated to next-generation sequencing technology in order to evaluate regions of the genome for which there is currently no literature information.

There is a minor reduction in reproducibility for next-generation sequencing technology when compared to the Illumina Infiunium methylation array 450K (0.97 Pearson correlation coefficient among technical replicates for next-generation sequencing technology, compared to 0.98 for Illumina Infinium methylation array 450K). However, the advancements in DNA methylation imputation algorithms mean that such a discrepancy is not nearly large enough to outweigh the benefits of next-generation sequencing technology when compared to using an Illumina Infinium methylation array 450K.

Figure 3 shows a flowchart for a method 100 of determining and providing information regarding a health state of a subject, according to an embodiment of the invention. The method comprises providing a sample collection kit to the subject to obtain a biological sample (e.g., urine, blood, semen, saliva) at step 110. At step 120, epigenetic data is extracted from the biological sample. Phenotypes of the subject are characterised from the epigenetic data using a machine learning algorithm to indicate the health state of the subject at step 130. Finally, at step 140, the characterised phenotypes and/or the health state of the subject are reported.

The biological sample of the subject may be obtained from peripheral sources (such as blood, saliva, urine or sperm) in a non-invasive manner at step 110, without the need for the intervention of a medical professional. The subject receives a sample collection kit that includes all items and instructions (e.g. Oragene (OG) 500 DNA Collection Kit from DNA Genotek etc.) necessary for the extraction, registration, storage and transport of the biological sample to a laboratory. However, in other embodiments, other more invasive sources may instead or additionally be used (e.g., biopsy of solid tissue). In such embodiments, a medical professional should carry out the sample collection procedure.

Once the biological sample is collected, a unique anonymized barcode is associated to the biological sample so that the biological sample can be tracked through a processing pipeline (used to extract epigenetic data from the biological sample) and assigned to the (data of the) subject. In alternative embodiments, if and/or when wet-lab processing steps are simple enough, a subject may also extract and sequence DNA from a biological sample using a portable DNA sequencer (e.g., MinION), and then upload the raw sequencing data to a secure server. In such embodiments, epigenetic data may be extracted from the raw sequencing data uploaded to the secure server.

Epigenetic data extracted from the biological sample includes information about 5mC DNA methylation (i.e., 5-methylcytosine modifications). In alternative embodiments, epigenetic data extracted from the biological sample may include information about other DNA modifications (such as 5-hydroxymethylcytosine, 5-formylcytosine, 5-carboxylcytosine or 6-methyladenine), histone modifications (such as acetylation, methylation, phosphorylation, ubiquitylation, sumoylation or biotinylation), chromatin accessibility, nucleosome positioning, the binding of transcription factors or other nuclear proteins to the DNA, non-coding RNAs, RNA modifications (such as N6-methyladenosine, N1-methyladenosine, pseudouridine, 5-methylcytosine, 2'-O-methylations or any other modifications registered in a database like RMBase), aggregation of prion-like proteins and any other molecular manifestations of cellular memory and gene expression regulation.

Figure 4 shows a flowchart of an example method 200 by which the step 120 (extracting epigenetic data from the biological sample) of method 100 may be performed. At step 205, DNA is extracted from the sample. Optionally, the DNA may be sonicated to produce fragmented DNA.

At step 210, DNA methylation (or other DNA modifications and/or epigenetic marks) is differentiated (e.g. through a chemical conversion step, an enrichment step, an enzymatic conversion, combinations of them, ...). Following step 210, once the DNA methylation (or other DNA modifications and/or epigenetic marks) has been differentiated, a sequencing library is generated at step 215. A sequencing library can also be generated from the DNA extracted from the biological sample without differentiating the DNA methylation, in order to simultaneously obtain genetic data in addition to epigenetic data. In this case, the process moves from step 205 directly to step 215. Alternatively, computational methods can be used further downstream (e.g.

see 245) to obtain genetic data from a sequencing library where the epigenetic marks have been differentiated. Both strands of DNA of the subject may be used to extract genetic data, allowing epigenetic information from both sides of the DNA helix to be extracted in addition to genetic data. Optionally, DNA and differentiated DNA methylation may be amplified using well-known DNA amplification techniques. At step 220, DNA and differentiated DNA methylation is sequenced. The DNA and differentiated DNA methylation may be sequenced using an Illumina next-generation sequencing machine.

Steps 205 to 220 represent a "wet-lab pipeline" of the method 200 for extracting epigenetic data from a biological sample.

Regarding step 210, differentiation of DNA methylation (or other DNA modifications or epigenetic marks) may be performed by using a biosensor to detect a biomarker. As used herein, the term "biosensor" means any system capable of detecting the presence of a biomarker. As used herein, the term "biomarker" means any epigenetic mark (e.g., DNA methylation or other DNA modification, or any other epigenetic mark) that can be detected in a biological sample.

Examples of biosensors may comprise a ligand or ligands capable of specific binding to the biomarker. Such biosensors are useful in detecting and/or measuring a biomarker. The DNA methylation may be differentiated via antibody-based enrichment methods, such as pulling down of methyl-cytosine in a specific context; Biotin-streptavidin or click chemistry-based enrichment methods; and other DNA or protein-based chemical or other enrichment methods.

The biomarker may be detected and/or measured using enrichment methods, mass readout or base calling methods. An enrichment method may be selected from: PCR-based enrichment of the sites of interest or nearby regions; restriction enzymes; pull-down approaches with oligonucleotide probes; and selection based upon other properties of the DNA sequence in the surrounding region, such as weight-based enrichment methods, or charge-based enrichment methods.

A base calling method may be selected from Illumina sequencing-based approaches for detection of methylation levels using DNA libraries that have undergone some form of DNA modification, such as bisulfite treatment (converting cytosine to uracil –

read as thymine, while leaving methylated cytosine unaltered). This can be achieved in a number of different methods: Amplicon sequencing – specific primers are designed such that the methylation state at a given site can be determined, such as methylation sensitive PCR (PCR). Alternatively, enzymatic conversion approaches can be used, for example instead of bisulfite conversion.

Alternatively, a mass readout method may be selected from mass spectrometry-based approaches for detection of methylation levels for specific CG sites by their mass differences. One instance of this is the AGENA epiTYPER. In whole genome bisulfite sequencing, the genome is sonicated, tagmented or random primed to define the library. In reduced-representation bisulfite sequencing, the library is defined through the use of restriction enzymes, commonly MspI. Additional base calling approaches could be used for the detection of cytosine modifications, such as pore methods where the base modification is detected by voltage differences across a channel occupied by a given base combination (Oxford Nanopore) such pores are not limited to biological pores (e.g. protein pores vs grapheme pores).

Approaches based on stalling of polymerases during polymerisation could be used instead, such as those utilised by Pacific Bioscience for the detection of methyl-adenine. The biomarker may be directly detected, e.g. by SELDI or MALDI-TOF.

Alternatively, the biomarker may be detected directly or indirectly via interaction with a ligand or ligands such as an antibody or a biomarker-binding fragment thereof, or other peptide, or ligand, e.g. aptamer, or oligonucleotide, capable of specifically binding the biomarker. The ligand may possess a detectable label, such as a luminescent, fluorescent or radioactive label, and/or an affinity tag. For example, detecting and/or measuring can be performed by one or more method(s) selected from the group consisting of: SELDI (-TOF), MALDI (-TOF), a 1-D gel-based analysis, a 2-D gel-based analysis, mass spectroscopy (MS) such as selected reaction monitoring (SRM), reverse phase (RP) LC, size permeation (gel filtration), ion exchange, affinity, HPLC, UPLC and other LC or LC MS-based techniques. Appropriate LC MS techniques include ICAT® (Applied Biosystems, CA, USA), or iTRAQ® (Applied Biosystems, CA, USA).Liquid chromatography (e.g. high pressure liquid chromatography (HPLC) or low pressure liquid chromatography (LPLC)), thin-layer chromatography, NMR (nuclear magnetic resonance) spectroscopy could also be used.

Detecting and/or measuring biomarkers may alternatively be performed using mass spectroscopy (MS). Detecting and/or measuring may be performed using selected reaction monitoring (SRM). SRM is a method used in tandem mass spectrometry in which an ion of a particular mass is selected in the first stage of a tandem mass spectrometer and an ion product of a fragmentation reaction of the precursor ion is selected in the second mass spectrometer stage for detection. Specific analyte panels can be developed for SRM matching the analytes on the biomarker panel. The analyte panels can quantitatively measure the protein analytes with high precision. This methodology has the advantage of allowing raw blood to be used instead of blood serum which minimizes the number of intermediate processing steps. For example, detecting and/or measuring can be performed by one or more method(s) selected from the group consisting of: SELDI (-TOF), MALDI (-TOF), a 1-D gel-based analysis, a 2-D gel-based analysis, mass spectroscopy (MS) such as selected reaction monitoring (SRM), reverse phase (RP) LC, size permeation (gel filtration), ion exchange, affinity, HPLC, UPLC and other LC or LC MS-based techniques. Appropriate LC MS techniques include ICAT® (Applied Biosystems, CA, USA), or iTRAQ® (Applied Biosystems, CA, USA). Liquid chromatography (e.g. high pressure liquid chromatography (HPLC) or low pressure liquid chromatography (LPLC)), thin-layer chromatography, NMR (nuclear magnetic resonance) spectroscopy could also be used.

Detecting and/or measuring the biomarkers may alternatively be performed using an immunological method, involving an antibody, or a fragment thereof capable of specific binding to the biomarker. Suitable immunological methods include sandwich immunoassays, such as sandwich ELISA, in which the detection of the biomarkers is performed using two antibodies which recognize different epitopes on a biomarker; radioimmunoassays (RIA), direct, indirect or competitive enzyme linked immunosorbent assays (ELISA), enzyme immunoassays (EIA), Fluorescence immunoassays (FIA), western blotting, immunoprecipitation and any particle- based immunoassay (e.g. using gold, silver, or latex particles, magnetic particles, or Q-dots).

Immunological methods may be performed, for example, in microtitre plate or strip format. Immunological methods in accordance with the invention may be based, for example, on any of the following methods.

In immunonephelometry, the interaction of an antibody and target antigen on the biomarker results in the formation of immune complexes that are too small to precipitate. However, these complexes will scatter incident light and this can be measured using a nephelometer. The antigen, i.e. biomarker, concentration can be determined within minutes of the reaction.

Radioimmunoassay (RIA) methods employ radioactive isotopes such as I125 to label either the antigen or antibody. The isotope used emits gamma rays, which are usually measured following removal of unbound (free) radiolabel. The major advantages of RIA, compared with other immunoassays, are higher sensitivity, easy signal detection, and well-established, rapid assays. The major disadvantages are the health and safety risks posed by the use of radiation and the time and expense associated with maintaining a licensed radiation safety and disposal program. For this reason, RIA has been largely replaced in routine clinical laboratory practice by enzyme immunoassays.

Enzyme (EIA) immunoassays were developed as an alternative to radioimmunoassays (RIA). These methods use an enzyme to label either the antibody or target antigen. The sensitivity of EIA approaches that of RIA, without the danger posed by radioactive isotopes. One of the most widely used EIA methods for detection is the enzyme-linked immunosorbent assay. Enzyme-linked immunosorbent assay (ELISA) methods may use two antibodies one of which is specific for the target antigen and the other of which is coupled to an enzyme, addition of the substrate for the enzyme results in production of a chemiluminescent or fluorescent signal.

Fluorescent immunoassay (FIA) refers to immunoassays which utilize a fluorescent label or an enzyme label which acts on the substrate to form a fluorescent product. Fluorescent measurements are inherently more sensitive than colorimetric (spectrophotometric) measurements. Therefore, FIA methods have greater analytical sensitivity than EIA methods, which employ absorbance (optical density) measurement. Chemiluminescent immunoassays utilize a chemiluminescent label, which produces light when excited by chemical energy; the emissions are measured using a light detector.

Immunological methods can thus be performed using well-known methods to detect and/ or measure biomarkers to differentiate epigenetic marks in the biological sample.

Any direct (e.g., using a sensor chip) or indirect procedure may be used in the detection of the biomarker. The Biotin-Avidin or Biotin-Streptavidin systems are generic labelling systems that can be adapted for use in immunological methods. One binding partner (hapten, antigen, ligand, aptamer, antibody, enzyme etc) is labelled with biotin and the other partner (surface, e.g. well, bead, sensor etc) is labelled with avidin or streptavidin. This is conventional technology for immunoassays, gene probe assays and (bio)sensors, but is an indirect immobilisation route rather than a direct one. For example a biotinylated ligand (e.g. antibody or aptamer) specific for a biomarker may be immobilised on an avidin or streptavidin surface, the immobilised ligand may then be exposed to a sample containing or suspected of containing the biomarker in order to detect and/or quantify a biomarker. Detection and/or quantification of the immobilised antigen may then be performed by an immunological method as described herein.

The term “antibody” as used herein includes, but is not limited to: polyclonal, monoclonal, bispecific, humanised or chimeric antibodies, single chain antibodies, Fab fragments and F(ab')2 fragments, fragments produced by a Fab expression library, anti-idiotypic (anti-Id) antibodies and epitope-binding fragments of any of the above.

The term "antibody" as used herein also refers to immunoglobulin molecules and immunologically-active portions of immunoglobulin molecules, i.e., molecules that contain an antigen binding site that specifically binds an antigen. The immunoglobulin molecules can be of any class (e.g., IgG, IgE, IgM, IgD and IgA) or subclass of immunoglobulin molecule.

The identification of key biomarkers specific to a disease is central to integration of diagnostic procedures and therapeutic regimes. Using predictive biomarkers, appropriate diagnostic tools such as biosensors can be developed, accordingly.

Detecting and quantifying can be performed using a biosensor, microanalytical system, microengineered system, microseparation system, immunochromatography system or other suitable analytical devices. The biosensor may incorporate an immunological method for detection of the biomarker, electrical, thermal, magnetic, optical (e.g. hologram) or acoustic technologies. Using such biosensors, it is possible

to detect the target biomarker at the anticipated concentrations found in biological samples.

The biomarker can be detected using a biosensor incorporating technologies based on "smart" holograms, or high frequency acoustic systems, such systems are particularly amenable to "barcode" or array configurations. In smart hologram sensors (Smart Holograms Ltd, Cambridge, UK), a holographic image is stored in a thin polymer film that is sensitised to react specifically with the biomarker. On exposure, the biomarker reacts with the polymer leading to an alteration in the image displayed by the hologram. The test result read-out can be a change in the optical brightness, image, colour and/or position of the image. For qualitative and semi-quantitative applications, a sensor hologram can be read by eye, thus removing the need for detection equipment. A simple colour sensor can be used to read the signal when quantitative measurements are required. Opacity or colour of the sample does not interfere with operation of the sensor. The format of the sensor allows multiplexing for simultaneous detection of several substances. Reversible and irreversible sensors can be designed to meet different requirements, and continuous monitoring of a particular biomarker of interest is feasible. Suitably, biosensors for detection of the biomarker combine biomolecular recognition with appropriate means to convert detection of the presence, or quantitation, of the biomarker in the sample into a signal.

Biosensors can be adapted for "alternate site" diagnostic testing, e.g. in the ward, outpatients' department, surgery, home, field and workplace. Biosensors to detect the biomarker include acoustic, plasmon resonance, holographic and microengineered sensors. Imprinted recognition elements, thin film transistor technology, magnetic acoustic resonator devices and other acousto-electrical systems may be employed in biosensors for detection of a biomarker.

Following step 220, raw sequencing data is generated and collated at step 225. The raw sequencing data may be stored in FASTQ format. The raw sequencing data is then inputted in a quality control pipeline at step 230 that can include generating quality control reports, removing low-quality data, demultiplexing, removing adapter sequences and removing PCR duplicates. Step 230 may be performed using in-house software or publicly available tools (such as FastQC, Trim Galore, Cutadapt or Trimmomatic).

At step 235, the sequencing data that has been subjected to quality control procedures is aligned or mapped to a reference genome. The reference genome may be obtained from an arbitrary subject, the subject under study or a graph genome that takes into account genetic variation in a population of subjects. Unaligned sequences can also be aligned to genomes from other species different to the subject (e.g., bacterial or viral species), rendering information about contamination of the sample or insights into a microbiome of the subject. Information relating to the microbiome of the subject may be used as additional data, discussed later. In-house software or publically available tools may be used for these purposes (e.g., BWA, Bismark, BSMAP, graph aligners), depending on the type of wet-lab protocol and epigenetic marks involved.

An alignment file (e.g., BAM, SAM) is generated at step 235 that can be further processed (e.g., sorting, removing secondary alignments, removing PCR duplicates, indexing) using in-house software or publically available tools (such as SAMTools, BAMTools, Picard, bamUtil). The alignment file contains information detailing alignment of sequences to the reference genome.

In some embodiments, *de novo* assembly and/or direct calling of epigenetic marks or genetic variants can be directly performed in the sequencing data after quality control at step 230, without an intermediate alignment step (i.e., step 235). In such embodiments, the method may proceed from step 230 directly to step 240.

Finally, calls are generated from the mapped sequences. At step 240, DNA methylation (or other DNA modifications and/or epigenetic marks) are called, resulting in epigenetic data (e.g., DNA modifications, histone marks peaks, chromatin accessibility peaks, DNA-binding proteins peaks, non-coding RNA counts, RNA modifications) being obtained at step 250. Similarly, at step 245, genetic variants may be called to obtain genetic data (e.g., single nucleotide polymorphisms, multiple nucleotide polymorphisms, insertions, deletions, structural rearrangements, copy number changes) at step 255. Algorithms employed for these purposes vary depending on the type of epigenetic data and genetic data obtained after calling.

Steps 225 to 255 of the method 200 represent a “bioinformatic pipeline” for extracting epigenetic data from a biological sample of a subject.

In some cases, the epigenetic data may need to be normalized within a given sample or subject, in which case several statistical and algorithmic implementations can be used

(e.g., quantile normalization, empirical Bayes methods). Furthermore, in some cases the epigenetic data needs to be corrected for cell composition, which can be accomplished using reference-free (e.g., SVA, ISVA) and reference-based (e.g., EpiDISH) methods. Additionally, the epigenetic data can be expanded (e.g., when the technology (e.g. sequencing technology) has missed one or more sites in the genome of the subject) using imputation algorithms, which can be constructed in-house or obtained from publically available software (e.g., ChromImpute).

With regard to step 130 of method 100, the epigenetic data extracted from the biological sample of the subject is then used to characterise phenotypes of the subject using a machine learning algorithm to indicate a health status of the subject. The primary goal is to map epigenetic data to one or more phenotypic variables (e.g., a phenotypic continuous variable or phenotypic classification) using the machine learning algorithm. Given the high dimensionality of the epigenetic data extracted from the biological sample, feature selection and/or extraction from the epigenetic data may be required in some cases. Different strategies can be employed for feature selection and/or extraction from the epigenetic data.

For example, selected and/or extracted features can include (but are not limited to) epigenetic data from: genomic regions or positions that present different epigenetic profiles between subjects with different phenotypic characteristics or within a subject in a longitudinal dataset (differences in mean or median); genomic regions or positions that present epigenetic profiles that are variable between subjects with the same or different phenotypic characteristics or within a subject in a longitudinal dataset (differences in variance); genomic regions or positions with specific genomic features (such as genes, exons, introns, promoters, bivalent promoters, transcription start sites, intragenic regions, enhancers, nucleosome-depleted regions, repressed chromatin, transcribed chromatin, hypomethylated footprints, DNA methylation valleys, transcription factor binding sites, binding of architectural proteins, binding of Polycomb and Trithorax complexes, repetitive elements, lamina-associated domains, topologically associated domain boundaries and other features derived from the 3D nuclear organisation, recombination hotspots, mutational hotspots, epimutational hotspots, regions that change their DNA methylation patterns with age, replication origins, early and late replicating regions, DNase I hypersensitive sites and other chromatin accessible regions, CpG content and other specific k-mer content, GWAS hits, eQTLS, meQTLS, conserved regions, certain histone modifications that lead to

specific chromatin states, and any other genomic features derived from segmentation algorithms); and combinations of the above.

In other cases, features can be selected and/or extracted in a more unsupervised manner using dimensionality reduction techniques (e.g., principal component analysis, linear discriminant analysis, t-distributed stochastic neighbor embedding, non-negative matrix factorisation, autoencoders). Furthermore, features can be selected and/or extracted, and prioritized, based on significance measures (e.g., p-value, q-value, effect size) derived from some form of statistical testing (e.g., t-test, Wilcoxon test, Bartlett test) or modelling (e.g., linear models) and/or based on their contributions to the machine learning algorithm (e.g., conditional feature contribution in a random forest, recursive feature elimination).

The selected and/or extracted features from the epigenetic data can be used to gain functional biological insights for a given phenotype with the help of several statistical and algorithmic methods, including information about the genes or biological pathways that may be involved (e.g., gene ontology, pathway analysis), genomic features that seem to be associated with it (e.g., enrichment analysis), clustering of subjects or samples (e.g., hierarchical clustering, k-means clustering, PCA-based clustering, tSNE-based clustering), clustering of features based on their co-regulation (e.g., weighted correlation network analysis) or inferring causality (e.g., exposure-outcome mediation, causal inference test, Mendelian randomisation).

The computational steps for selecting and/or extracting features from the epigenetic data can be carried out in one or more of the following platforms: computers, servers, cloud-computing services, mobile phones, tablet devices and any other suitable systems.

The machine learning algorithm is configured to generate one or more phenotypic variables (e.g., continuous values and probabilities of belonging to one or more phenotypic classes) from the epigenetic data for the subject. In characterizing phenotypes of the subject, the machine learning algorithm can indicate a health state of the subject by quantifying effects relating to lifestyle (e.g., smoking exposure, sleep deprivation, dietary components) and environment (e.g., air pollution exposure, heavy metals exposure) factors at a molecular level.

In addition to the epigenetic data, other additional data can be collected from the subject. This additional dataset contains information about variables related to the health, well-being, lifestyle and environmental exposure of the subject, such as: genetic data, microbiome data, metabolomic data, proteomic data, imaging data, medical or clinical records from the subject and close relatives (including information about past and current diseases or conditions), age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth), and any other sources of personal and biological information that may be useful for the purposes of this method.

The additional data can be collected as static variables or it can include dynamic variables that change longitudinally. The data can be obtained from different sources, such as trackers and wearables, medical devices, biosensor devices, mobile phone, physical devices connected to the Internet of Things, any other sensors, self-reported questionnaires or surveys or uploaded data from third-party providers. All the additional data is stored and organised in a database that links the epigenetic data extracted from the subject with all the additional data for the subject, allowing for the easy access and computational use of the information.

The machine learning algorithm may be selected from linear models, logistic regression, nearest neighbour algorithms, support vector machines, decision trees, random forests, gradient boosting, Gaussian processes, shallow artificial neural networks or deep learning approaches. In particular, convolutional neural networks can be used to efficiently combine epigenetic and genetic data into the prediction and

sequential models, such as tensorial recurrent neural networks, can be used to explicitly integrate the longitudinal dimension of the input data.

Alternatively, longitudinal input data can be implemented in computational models that, for example, predict the risk of developing a certain phenotype or the time until the phenotype is developed, or predict a past phenotype of the subject. For such purposes, algorithms and statistical approaches derived from time series analysis (e.g., autocorrelation, cross-correlation, ARMA models, ARIMA models, ARFIMA models, hidden Markov models, Gaussian processes and other stochastic modelling approached, recurrent neural networks, tensorial recurrent neural networks) or survival or time-to-event analysis (e.g., Kaplan-Meier statistics, log-rank testing, Cox models). Additionally, such approaches can be used to model future or past instances of epigenetic data or additional data (e.g., predicted epigenetic profile for the subject in, for example, three years' time), which can serve for the purposes of data augmentation or to minimise the number of biological samples required from the subject. This in turn reduces cost associated with performing the above method and reduces undue hassle and interference with the subject.

Characterising phenotypes of the subject from at least the genetic data and the epigenetic data using the machine learning algorithm to indicate a health state of the subject enables a comparison of the subject's genetic predisposition to certain phenotypes with the phenotypes actually exhibited by the subject. In this way, the impact of epigenetic factors on genetic expression can be assessed in detail.

Additional data (for example, lifestyle, health and well-being and environmental data) obtained from the subject may also be used as input data for the machine learning algorithm, such that phenotypes of the subject can be characterised from at least the lifestyle, health and well-being and environmental data and the epigenetic data to indicate a health state of the subject.

The characterised phenotypes of the subject comprise both environmental and pathological phenotypes, including but not limited to microbiome data, metabolomic data, proteomic data, information about past and current diseases or conditions, age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies,

alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth), and any other phenotypes that may be characterised for the purposes of this method.

The ability to characterise a broad range of environmental phenotypes represents a distinct improvement over the majority of previous work relating to epigenetics, which has primarily focused on epigenetic factors relating to specific pathologies (i.e., medical diseases, medical conditions). By characterising environmental phenotypes of the subject and not only pathological phenotypes of the subject, a greater insight into environmental factor correlation and possible causation for particular medical conditions or diseases may be obtained (see Figure 12).

The epigenetic data, genetic data and additional data (e.g., lifestyle, health and well-being, and environmental data) may be used to update and refine the machine learning algorithm, thereby improving the characterizing and/or predictive power of the machine learning algorithm.

At step 140 of the method 100, the characterised phenotypes and/or the health state of the subject are reported. This information is made accessible in an easy-to-understand and self-contained way, so the subject is able to interpret the results by them self. Information can be directly accessed through a digital platform that is executed in at least one of a computer, mobile phone, tablet, wearable or any other similar device. The digital platform may also contain educational material of different levels of complexity to facilitate the understanding of the content and the results can be reported through the use of different types of graphs, plots, animation, videos and text. Alternatively, the information can be provided through some intermediary entity that has a connection with the subject (e.g., employer, insurer) or through an intermediary person associated with the subject (e.g., medical professional, caretaker, parent).

Alternative ways of accessing the data (e.g., physical reports) may also be used. The information obtained from the computational models can be provided in a descriptive way (see Figure 5, discussed below), so the subject can assess the current state of a given phenotype.

The results from the machine learning algorithm quantify the effects of different lifestyle or environmental factors in a subject using epigenetic data and can be interpreted as updated risk scores that change longitudinally (both in the case of phenotypic continuous variables and the probability of belonging to a certain phenotypic class). A given phenotypic variable derived from the models can also be associated with the risk of developing other health or well-being related conditions (e.g., an exposure to air pollutants can be associated with a certain risk of developing lung complications). Measurements of the confidence of the result or error associated with it can also be provided.

The results can be compared with the results reported by wearables, trackers or self-reported surveys of the subject, with the values from the subject from different time points, or with at least a subset of a population (e.g., by age, by geographical location, by gender).

Additionally, personalised recommendations for interventions can be presented to the subject. This can include changes in the subject's lifestyle (e.g., increased or decreased physical exercise, dietary change, weight loss, downloading a certain mobile app to quit smoking) or environment (e.g., expending more time in less polluted areas, buying an air or water filter, avoiding commuting at certain hours) that are predicted to reduce phenotypic risk, or prevent an increase in phenotypic risk.

The recommendations may be based on longitudinal information from other subjects that undertook the given recommendation and were shown to improve the aforementioned phenotype at the epigenetic level, therefore providing a source for evidence-based lifestyle recommendations. The subject can be followed before, during and after the intervention(s) and more data (both epigenetic and additional) can be collected to improve the resolution and the accuracy of the interventions.

For example, Figure 5A shows an example chart 300 in which a phenotypic variable or phenotypic risk of a subject at a point in time, as calculated solely from information obtained from health trackers and wearable sensors configured to monitor a

physiological parameter, is illustrated in the left-hand column 305. The same phenotypic variable or phenotypic risk is shown in the right-hand column 310, with the value instead calculated using the machine learning algorithm of the method 100. A recommended threshold 315 for the phenotypic variable or phenotypic risk is also provided on the chart 300. The phenotypic variable or risk as calculated from information obtained from health trackers and wearable sensors is shown below the recommended threshold 315, while the phenotypic variable or risk as calculated using the machine learning algorithm of the method 100 is shown to be above the recommended threshold 315. From the chart 300, the subject may gain a more accurate understanding of their phenotypic characteristics than may otherwise be available if the subject was to rely only on health trackers and wearable sensors. For example, the subject can easily determine whether or not the phenotypic variable or risk is above or below a recommended threshold. If necessary, based on the determination of whether the phenotypic variable or risk is above or below the recommended threshold, the subject can take measures (such as medical and/or lifestyle interventions) to address lifestyle, health and well-being and/or environmental factors which may aid in reducing their phenotypic variable or risk.

Similarly, Figure 5B shows an example chart 320 in which a phenotypic variable or phenotypic risk of a subject at a point in time, illustrated in the left-hand column 325 is compared with both a typical phenotypic variable or phenotypic risk of two other populations, and a recommended threshold 340 for the phenotypic variable or phenotypic risk. In the example shown, the middle column 330 illustrates the typical phenotypic variable or phenotypic risk for a population of people aged 25 to 35 years old, shown to be below the recommended threshold 340 in chart 320. In the example shown, the right-hand column 335 illustrates the typical phenotypic variable or phenotypic risk for people from the same country as the subject, shown to be above the recommended threshold 340 in chart 320. The populations with which the subject may select to compare their phenotypic variable or phenotypic risk may be selected from any population for which a typical phenotypic variable or phenotypic risk has been determined. In this way, the subject can easily compare their own phenotypic variable or phenotypic risk with one or more different populations to aid their understanding of their own phenotypic variable or phenotypic risk.

Figure 5C shows a chart 360 illustrating how two different phenotypic variables or phenotypic risks from the same subject (curve 365 and curve 370) change over time

(i.e. longitudinally). The curves can be constructed using data points obtained after phenotype/epigenetic characterisation of the subject at different time points and/or using data points predicted using the machine learning algorithm. This prediction may take into account changes in the lifestyle, health, well-being and environmental factors affecting the epigenetic profile of the subject. As can be seen by the shape of curve 370, this phenotypic variable or phenotypic risk rises steadily, followed by a sharp decrease after time t (375), which is the time at which a certain medical and/or lifestyle intervention takes place. Therefore, the subject can visualise how this intervention reduces the phenotypic risk portrayed by curve 370. In contrast, the phenotypic risk in curve 365 is only partially stabilised after time t, which allows the subject to compare, in a simple, intuitive and easy to understand manner, the effectiveness and/or impact of a given intervention on different phenotypic variables or phenotypic risks. Alternatively, curve 370 could represent a phenotypic variable or phenotypic risk for a subject and curve 365 the same phenotypic variable or phenotypic risk for a defined population of subjects (e.g. selected by age, geographical region, ...). In this case, chart 360 would allow to track the evolution of a phenotypic risk at the population level and compare the effectiveness of the intervention at the population and the individual subject level, opening the door for the personalised recommendation of lifestyle interventions in an evidence-based manner.

As shown in the charts 300, 320 and 360 of Figures 5A, 5B and 5C, the characterised phenotypes and/or health status of the subject as determined by the machine learning algorithm can be reported and/or displayed to the subject under analysis in an intuitive manner on the digital platform.

Figure 6 shows a method 400 for developing a machine learning algorithm configured to characterise phenotypes of a subject according to a second aspect of the invention. The method 400 comprises providing a sample collection to each of a population of individuals to obtain biological samples from at least a subset of the population at step 410. At step 420, epigenetic data is extracted from at least a subset of the biological samples. At least one of lifestyle data, health data, well-being data and environmental data is obtained from at least a subset of the population at step 430. At step 440, the epigenetic data and the at least one of lifestyle data, health data, well-being data and environmental data is collated in a training data set. Finally, the machine learning

algorithm is trained to characterise phenotypes from epigenetic data using the training data set at step 450.

Epigenetic data may be extracted from at least a subset of the biological samples at step 420 using one or more of the methods described above with respect to the method 200. Similarly, the at least one of the lifestyle data, health data, well-being data and environmental data may be obtained from at least a subset of the population at step 430 in substantially the same manner as described above with respect to the method 100.

The population of individuals may be a population of humans, although the method 400 can be used to train a machine learning algorithm to characterise phenotypes from epigenetic data for non-human species too (e.g., pets, protected species, zoo animals). The population of individuals may comprise individuals with a wide variety of phenotypic characteristics, or may comprise a more unique and/or stratified cohort of individuals (especially in the case that the method is to be applied to a specific research project). The biological samples obtained from at least a subset of the population can serve as the basis for a cross-sectional study. Additionally, at least a subset (or all) of the same population of individuals can be sampled again at future points in time, allowing for the machine learning algorithm to be trained for longitudinal characterisation and prediction of phenotypes from epigenetic data.

The method 400 may further comprise extracting genetic data from at least a subset of the biological samples, as described above with respect to the method 200. The extracted genetic data may then be collated in the training data set, alongside the extracted epigenetic data and the at least one of lifestyle data, health data, well-being data and environmental data. The machine learning algorithm can then be trained to characterise phenotypes from epigenetic data using the training data set (wherein genetic data extracted from at least a subset of the biological samples). Training the machine learning algorithm to characterise phenotypes by using genetic data as an input into the machine learning algorithm during the training process allows a comparison of genetic predispositions to certain phenotypes with phenotypes actually exhibited by a subject (once the machine learning algorithm is trained using the training data set).

The machine learning algorithm is trained, at step 450, using the training data set from the population of individuals to solve both regression and classification problems. In

the regression case, the machine learning algorithm is trained to map the input data (from the training data set) to a phenotypic continuous variable. The phenotypic continuous variable is generally obtained from the at least one of lifestyle data, health data, well-being data and environmental data. The phenotypic continuous variable may comprise age, body mass index, number of cigarettes or packs smoked in a time interval, alcohol consumption, coffee or tea consumption, exposure to certain pollutants or chemicals, mental health questionnaires scores (e.g., stressful events questionnaire), blood biomarkers concentration (e.g., C-reactive protein, folate and other vitamins), cell type counts, hours and types of physical exercise, hours and quality of sleep and any other continuous variables derived from the at least one of lifestyle data, health data, well-being data and environmental data.

In the classification case, the machine learning algorithm is trained to map the input data (from the training data set) to a probability of belonging to one or more phenotypic classes. The one or more phenotypic classes are generally obtained from the at least one of lifestyle data, health data, well-being data and environmental data. The one or more phenotypic classes may comprise smoking status, obesity, substance abuse, excessive exposure to pollutants or chemicals, insulin resistance, anxiety, childhood trauma or other types of trauma, chronic fatigue, chronic inflammation, depression, pregnancy-related traits (e.g. fetal alcohol disorder, post-partum depression, pregnancy anxiety), diet deficiencies (e.g. folate), food and respiratory allergies, infection status (e.g. *Helicobacter pylori* infection, HIV infection), sleep problems (e.g. obstructive sleep apnea, sleep deprivation), chronotype, physical exercise status, stress status and any other categorical variables derived from the at least one of lifestyle data, health data, well-being data and environmental data.

Features may be selected and/or extracted from the epigenetic data in the training data set in substantially the same manner as described with respect to the method 100.

The training data set generated from the population of individuals may be partitioned into training, validation and test data sets. If the training data set is not large or complete enough, data augmentation algorithms (e.g., transformations, generative adversarial networks) can be used to expand the training data set in order to avoid training the machine learning algorithm to overfit input data.

Different machine learning algorithms with different architectures can be trained in a supervised manner according to step 450, including linear models, logistic regression,

nearest neighbour algorithms, support vector machines, decision trees, random forests, gradient boosting, gaussian processes, shallow artificial neural networks or deep learning approaches. In particular, convolutional neural networks can be used to efficiently combine epigenetic and genetic data into the prediction and sequential models, such as tensorial recurrent neural networks, can be used to explicitly integrate the longitudinal dimension of the input data. Alternatively, machine learning algorithms to characterise phenotypes from epigenetic data can be trained in an unsupervised manner, the machine learning algorithm learning to infer one or more relationships between data variables without the data in the training data set being labelled.

Alternatively, longitudinal input data can be used to build machine learning algorithms that predict the risk of developing a certain phenotype or the time until this happens. For these purposes, machine learning algorithms derived from time series analysis (e.g,. autocorrelation, cross-correlation, ARMA models, ARIMA models, ARFIMA models, hidden Markov models, Gaussian processes and other stochastic modelling approaches, recurrent neural networks, tensorial recurrent neural networks), or survival or time-to-event analysis (e.g., Kaplan-Meier statistics, log-rank testing, Cox models) can be employed. Additionally, these approaches can be used to train the machine learning algorithm to model future or past instances of epigenetic data or additional data (e.g. predicted epigenetic profile for a given subject in 3 years' time), which can serve for the purposes of data augmentation or to minimise the sampling of the subject, reducing the cost of the technology and avoiding the hassle to the subject.

Figure 7 shows a schematic of a method 500 for training a machine learning algorithm to characterise phenotypes from epigenetic data. A training data set 510 includes epigenetic data and lifestyle data, health data, well-being data and environmental data (e.g., obtained from health trackers and wearable sensors, questionnaires, blood-based measurements etc.). The training data set 510 also includes additional metadata (e.g. variables that describe the method used to generate the epigenetic data, including: batch numbers for reagents, details of sequencing run, plate position etc.). This metadata may be important for: reducing the impact of any process-induced confounders, reducing model overfitting, among other use cases.

A machine learning algorithm 520 uses the information collated in the training data set 510 as input data, and is trained to output characterise phenotypes 530 from epigenetic data using the information in the training data set 510.

The exact method of training is specific to the type of machine learning algorithm being trained to characterise phenotypes from epigenetic data. Typically, training a machine learning algorithm comprises iteratively updating parameters of the machine learning algorithm (the parameters defining how the machine learning algorithm behaves) in response to an error (e.g., magnitude of an error) in one or more outputs predicted from input data (i.e., a training data set). The parameters defining the machine learning algorithm are typically set to zeros or an arbitrary distribution of random values initially (i.e., before training begins). The number of training iterations of updating parameters of the machine learning algorithm needed is dependent on the type of machine learning algorithm, and the type and/or complexity of information in the training data set. Once a sufficient number of iterations has been performed for the machine learning algorithm to successfully characterise phenotypes from epigenetic data to within a satisfactory degree of error, the machine learning algorithm is considered to be trained.

Once the machine learning algorithm is trained to characterise phenotypes from epigenetic data using a training data set comprising information from a population of individuals, the machine learning algorithm can be used to infer (i.e., characterise) phenotypic variables of subject(s) that were not part of the population of individuals comprising the training data set.

Figure 8 shows a schematic for a method 600 for training a machine learning algorithm to characterise phenotypes from epigenetic data. The method 600 is substantially similar to the method 500 described above, with the distinction that the epigenetic data, at least one of lifestyle data, health data, well-being data and environmental data (which may be obtained from health trackers and wearable sensors, other sensors such as those in mobile phones of devices connected to the Internet of Things, questionnaires etc.) and the additional metadata forming the training data set 610 is collected at a plurality of points in time. The plurality of points in time are depicted as times t1, t2 and t3 in Figure 8, although any number of distinct time points at which to obtain data from at least a subset of the population may be used. The epigenetic data, and the at least one of lifestyle data, health data,

well-being data and environmental data therefore comprise longitudinal epigenetic data, and at least one of longitudinal lifestyle data, longitudinal health and well-being data and longitudinal environmental data. Optionally at step 620, features may be selected and/or extracted from the data in the training data set 610, before the data is input into a machine learning algorithm 630 in order to train the machine learning algorithm 630 to characterise phenotypes from epigenetic data.

Collecting longitudinal data from at least a subset of the population at a plurality of points in time allows for the machine learning algorithm 630 to be trained to characterise past, present and future phenotypes from epigenetic data, as discussed previously.

The following passages are illustrative examples of obtaining high-quality DNA from a human saliva sample, suitable for generating high-throughput epigenetic data.

In some embodiments, a collection device with unique identifiers is distributed to an individual that is to provide the saliva sample. The collection device contains buffering reagents to ensure that the biological material is lysed in order to be transported safely. In addition, the collection device contains stabilising reagents to ensure that the DNA remains intact for the duration of the return and storage.

Upon receipt of the collection device, it must be clear to the individual providing the sample how to proceed. This includes ensuring that: a) they have not eaten or drunk anything in the preceding 30 minutes; b) they understand how to record the date, time and unique identifier for the collection device within an online schema to enable the order and collection device to be associated with said individual; and c) they proceed safely with providing the sample and ensuring that the sample is stabilised and returned in conjunction with the legal requirements of the countries involved.

Upon receipt of the sample at a laboratory, the unique identifier must be cross-checked against a database of unique identifiers to determine the correct assignment of the sample (and to check that the sample is expected to be received). Once correct assignment has been clarified, the sample is stored (in conditions compliant with the collection device), or is processed downstream.

The DNA-containing lysate can be processed using standard workflows to extract a pure DNA sample. In some embodiments, this is achieved by taking substantially 1 mL of solution from the collection device (for example, using the QIAsymphony SP with the QIAsymphony DSP DNA Midi Kit). Specific protocols may be required depending on the collection device used. For example, specific protocols have been developed for extractions from Oragene saliva samples, including a heated elution (at substantially 37°C) into substantially 60 µL of elution buffer.

DNA yields and purity can be calculated downstream of the extraction (for example, using Quant-iT® PicoGreen ® reagent from Life Technologies). In some embodiments, to evaluate DNA purity absorbance (260, 280 and 320 nm), measurements can be made using a microplate reader to determine an A260:A280 ratio. DNA integrity can subsequently be evaluated with a substantially 0.8% agarose gel at substantially 90 V for substantially 50 minutes (for example, using SYBR-Safe from Life Technologies). Once a high-purity DNA sample has been achieved and confirmed through testing, it can then be processed to assess epigenetic marks.

Saliva is the tissue of choice in this embodiment. This is because saliva contains a mix of buccal epithelial cells and leukocytes, with several associated benefits including: ease of use and a non-invasive sampling process; high correlation between the epigenetic profiles of saliva and those found in the blood and the brain; and a high longitudinal stability of buccal cells when compared with other tissues.

The next step is to generate raw DNA methylation data. Firstly, library preparation is required. In some embodiments, library preparation comprises mixing substantially 1 µg of sample human DNA with a diluted bisulfite conversion control (for example, DNA from lambda phage). Using a sonication device, the DNA sample is then fragmented into double-stranded DNA fragments. In some embodiments, between 5 and 15 sonication cycles, for example 10 sonication cycles, are used. Each cycle of sonication may alternate between periods of activity and periods of inactivity (for example, between 15 and 45 seconds or substantially 30 seconds on or sonicating the DNA sample, followed by between 15 and 45 seconds or substantially 30 seconds off or not sonicating the DNA sample). This generates double-stranded DNA fragments comprised of 3' and 5' overhangs (for example, having a median insert size of between substantially 300 and 400 base pairs). In some embodiments, end repair of

the fragments is then performed. Subsequently, adaptors can be ligated to the DNA fragments. Adaptors add the DNA sequence to bind the flow cell in the sequencing instrument, provide binding sites for sequencing primers and include sample-specific indexes that allow multiplexing of several sample libraries (for example, unique dual index). Furthermore, adaptors may be optimised for libraries that undergo bisulfite conversion (e.g., methylation adaptors). In some embodiments, the NEBNext Ultra II Library Prep Kit can be employed together with a ligation enhancer and IDT Methylation Index Adaptors (thermocycler parameters may be incubate for between 10 and 20 minutes, for example substantially 15 minutes at a temperature of substantially 20°C). Size selection is then performed to capture the DNA fragments. In some embodiments, bead-based size selection is used (for example, using AMPure XP beads). Quality control may be performed in the libraries using the BioAnalyzer (for a low number of samples) or the Fragment Analyzer (for a high number of samples). The above steps may be performed in a 96-well plate.

Following library preparation, bisulfite conversion and pre-capture amplification takes place (although it will be appreciated that other techniques such as enzymatic conversion approaches may be used in place of bisulfite conversion). Treatment of DNA with sodium bisulfite converts cytosine residues into uracil but leaves 5-methylcytosine residues unaffected, which allows differentiating distinct DNA methylation states at base-pair resolution. In some embodiments, the EX DNA Methylation-Lightning Kit (from Zymo Research) is used to perform bisulfite conversion. In some embodiments, bisulfite conversion comprises denaturing DNA and incubating with a conversion reagent. The thermocycler parameters may be as follows: substantially 98°C for substantially 8 minutes, followed by substantially 54°C for substantially 60 minutes, and then hold at substantially 4°C for a maximum of 20 hours). Subsequently, the bisulfite conversion may comprise binding the DNA to magnetic beads, before incubating with desulfonation reagent (for example, at room temperature for between substantially 15 and 20 minutes). The magnetic beads may then be washed and dried (for example, at substantially 55°C for between substantially 20 and 30 minutes), before eluting the bisulfite-converted DNA.

Next, PCR amplification is performed to obtained double stranded DNA. The original cytosine residues result in thymine residues and the original 5-methylcytosine residues result in cytosine residues. In some embodiments, the KAPA HiFi HotStart Uracil+

ReadyMix PCR kit together with appropriate IDT NGS primers (for example, primers that allow for ligation mediated PCR) can be employed. Thermocycler parameters may be as follows: substantially 95°C for substantially 2 minutes; substantially 10 cycles of substantially 30 seconds at substantially 98°C, followed by substantially 60°C for substantially 30 seconds, followed by substantially 72°C for substantially 4 minutes; substantially 72°C for substantially 10 minutes; hold at substantially 4°C. The hold at 4°C is effectively an open-ended wait until the next step in the process is taken (for example, either freezing the amplified bisulfite-converted libraries or moving to purification). In some embodiments, PCR amplified libraries are left overnight at 4°C before being purified the following day. Finally, the amplified bisulfite-converted libraries can be purified (for example, using AMPure XP beads) and quality controlled (for example, using the BioAnalyzer or Fragment Analyzer). Final DNA fragments should be between substantially 150 and 500 base pairs. In some embodiments, other fragment lengths are present. Additionally, in some embodiments, bisulfite conversion percentage can be determined using droplet digital PCR (ddPCR). In some embodiments, 4 sample libraries can be pooled together (for example, with approximately 250 ng of DNA from each sample) to make the method more cost effective).

In some embodiments, enrichment of genomic regions takes place. This may be done, for example, using oligonucleotide probes. It is possible to select bisulfite-converted DNA fragments that map to specific genomic regions, which can reduce the cost of the method and increase the accuracy of the DNA methylation data for, for example, CpG sites that are more relevant for the training of machine learning models. In some embodiments, this is achieved by hybridizing the denatured DNA fragments to oligonucleotide probes that are complementary to the sequences to be targeted. Given that bisulfite-converted DNA fragments can exist in different states (depending on the methylation status), oligonucleotide probes for all the possible sequence combinations may be used. Bisulfite-converted DNA fragments that have hybridized to the probes can be purified from the rest through different means. In some embodiments, the SeqCap Epi CpGiant probes kit (Roche) is used. In this case, more than 5 million CpG sites are retrieved from genomic regions that are functionally relevant. Both DNA strands are targeted, which allows obtaining accurate genetic information from the raw DNA sequencing data on top of the epigenetic (for example, DNA methylation) data (for example, it allows to differentiate C>T mutations, which commonly occur during

mammalian aging, from unmethylated cytosines). In some embodiments, substantially 1 μg of bisulfite-converted DNA is mixed with bisulfite capture enhancer and xGen Universal Blockers. This mixture is then dried (for example using a DNA vacuum concentration, at substantially 60°C). Next, hybridization buffers are added and the DNA is denatured. The thermocycler parameters for this step may be substantially 95°C for substantially 10 minutes. Substantially 4.5 μL of the SeqCap Epi probe pool reagent is added per bisulfite-converted sample library to the previous mixture, and incubated to allow hybridization of the oligonucleotide probes with the bisulfite-converted DNA. The thermocycler parameters may be substantially 47°C for between substantially 64 and 72 hours. A heated lid held at substantially 57°C may be used. Capture beads are then used to bind the oligonucleotide probes (which themselves have hybridized with the sample DNA). Thermocycler parameters may be substantially 47°C for substantially 45 minutes. A heated lid held at substantially 57°C may be used. After several washes, the bead-bound captured DNA can be further PCR-amplified. In some embodiments, the KAPA HiFi HotStart ReadyMix PCR kit together with appropriate IDT NGS primers can be employed. Thermocycler parameters may be as follows: substantially 98°C for substantially 45 seconds; between 10 and 15 cycles, for example 11 cycles of substantially 98°C for substantially 15 seconds, followed by substantially 60°C for substantially 30 seconds, followed by substantially 72°C for substantially 30 seconds; substantially 72°C for substantially 1 minute; hold at 4°C (as described above). The post-capture amplified libraries can be purified (for example, using AMPure XP beads) and quality controlled (for example, using the BioAnalyzer or the Fragment Analyzer). The fragment DNA sizes should be between substantially 150 and 500 base pairs.

The prepared libraries can then be sequenced. In some embodiments, paired end sequencing (for example, with a read length of 2 x 100 base pairs) is performed to generate raw DNA sequencing data. In some embodiments, the sequencing technology used is Illumina technology (for example, using the NovaSeq 6000 system with an S4 flowcell).

The following passages are illustrative examples of quantifying one or more environmental phenotypes (e.g., a phenotype which may be highly influenced by the environment and lifestyle of the individual) from epigenetic data.

Bioinformatics Processing

In some embodiments, if several samples (libraries) have been pulled together in the same sequencing lane, the resulting raw reads may be separated into different files depending on the DNA sample that they came from (a process known as demultiplexing). In the case of paired-end sequencing, this generates one or more pairs of files (for example, FASTQ files) that represent the input data to be processed. The first file of the pair contains the forward reads, and the second file contains the reverse reads.

FASTQ files are the standard format used to capture raw DNA sequencing data. Each read entry (DNA sequence fragment) is composed of 4 lines: a sequence identifier, raw sequence letters, a separator line (for example, this is normally the '+' character, and sometimes the sequence identifier again), and quality values (for the different sequence letters). Quality control is applied to the FASTQ files in order to identify technical artifacts and remove low quality reads. Residual sequence coming from the adaptors must also be removed (a process known as adaptor trimming). In some embodiments, Trim Galore can be used as the software for this purpose. Trim Galore also allows quantification of the original read duplication levels (that is, the number of reads with an identical sequence, therefore likely to have originated from the same DNA fragment).

Next, filtered reads are aligned to the human reference genome (for example, using a mapper than can handle bisulfite-converted DNA sequences). In some embodiments, Bismark is used as the software for this purpose. This requires first indexing a bisulfite-converted version of the genome previous to the mapping step. Default parameters together with the '--pbat' option (which selects the DNA strand types of the reference) allows accurate mapping of the reads (for example, generated with the previously described SeqCap Epi CpGiant probes kit). The output file from the mapper is a BAM file, a binary version of a SAM file (a tab-delimited text file that contains sequence alignment data). If more than one pair of FASTQ files was originally generated for the sample under consideration, the resulting BAM files can be merged (for example, using *samtools*). Next, duplicated reads (for example, those reads that likely originated from the same original DNA fragment and that were PCR-amplified) must be removed to ensure accurate DNA methylation estimates. In some embodiments, the deduplication tool provided by Bismark can be used with default

parameters. Finally, the number of reads supporting a methylated state and the number of reads supporting an unmodified or unmethylated state for all of the cytosines can be extracted. In some embodiments, the methylation extractor tool provided as part of Bismark (for example, using the parameters '-p --comprehensive --scaffolds --merge_non_CpG --bedGraph') can be used to generated the final COV file with the DNA methylation data.

From the various file types and processing steps with quality outputs, inclusive of the BAM files, it is possible to determine metrics such as median depth of coverage for a given sample (for example, the total number of reads supporting the methylation measurement, including the number of reads supporting a methylated cytosine and the number of reads supporting an unmethylated cytosine), duplication rate, percentage of uniquely mapping reads and other quality control parameters linked to the sample. This metadata may serve several purposes. Firstly, the metadata enables assessment of whether the requisite median depth of coverage of uniquely mapping reads has been achieved. In some embodiments, a required median depth of coverage could be 30 reads. In some embodiments, if such a depth is not obtained from the BAM file, additional sequencing is required. In other embodiments, other quality parameters are used to determine whether additional samples or sequencing is required. For example, duplication rate (and optionally in combination with other quality metrics) informs whether a given DNA library needs to be sequenced deeper, or whether a new library needs to be created from the existing sample, or in turn whether a new sample needs to be obtained from the individual. Such decisions can be modelled using software tools such as PreSeq (using functions such as: presep c_curve and preseq lc_extrap). Secondly, the quality control data can be used in conjunction with other metadata from the individual (for example, a sex of the person) to validate its identity.

In other embodiments, the reads generated from the bisulfite-conversion control are used to estimate the bisulfite conversion rate. This is an important source of technical variation among different sample batches. In the case that lambda phage DNA is used as such a control, this DNA sequence needs to be included as another 'chromosome' in the reference genome.

In other embodiments, a different DNA sequencing technology (not Illumina technology) is used. For example, sequencing may be performed using BGI

technology, Oxford Nanopore technology, or PacBio technology. The sequencing technology used may affect one or more aspects of the bioinformatic processing described above. For example, different data types may require different alignment parameters, different adapter trimming protocols, different quality filters etc. It will be appreciated that the skilled person would employ or adapt the requirements as necessary based upon the sequencing technology used.

In some embodiments, the bioinformatic processing procedure is run in a cloud-computing environment (for example AWS). The processing procedure can be automatically triggered (for example, using lambda functions in AWS) and parallelized.

In some embodiments, a workflow language (for example, Workflow Description Language, interpreted by software such as Cromwell) is used to organize the different steps of the processing procedure and the respective inputs and outputs. For example, operating-system-level virtualization (for example, Docker containers) may be used to deploy the appropriate software in the different processing steps.

DNA Methylation Data Wrangling

For each individual sample, DNA methylation data generated with next-generation sequencing technology is normally stored in a tab-delimited file (COV format). In this file, each row represents a different cytosine in the genome and the following information is provided in the columns: chromosome, start position (1-based), end position (1-based), methylation percentage, number of reads supporting a methylated cytosine, number of reads supporting an unmethylated cytosine. It is possible to merge the information of all the individual samples. For example, two matrices (where rows represent cytosines and columns the different individual samples) may be generated:

- Matrix 1: containing the methylation percentages.
- Matrix 2: containing the total number of reads supporting the methylation measurement i.e. depth of coverage (number of reads supporting a methylated cytosine + number of reads supporting an unmethylated cytosine).

The cytosines in matrix 1 are then filtered based on a depth of coverage threshold (provided in matrix 2). Normally, only those cytosines which have a minimum depth of coverage of 5 reads are kept (although other thresholds such as 10 reads may also be used). Imputation methods that estimate the methylation percentages of the missing cytosines can be further applied (including k-nearest neighbours, mean or median methylation values across a population of samples, mean or median values for a population of samples with similar phenotypes, imputation methods that take into account the methylation state of cytosines that are close in the linear DNA sequence or in the 3D genomic space or other imputation methods that make use of the biological information provided by genomic context). In other embodiments, cytosines from opposite strands and belonging to the same CpG site can be merged into a single cytosine to increase the depth of coverage or accuracy of the measurement.

Training a machine learning model to predict biological age from DNA methylation data

The ageing process is highly influenced by lifestyle and environmental factors, such as exercise, diet, sleep or exposure to toxins (such as those present in tobacco or air pollutants) and pathogens. Recent heritability estimates suggest that only 10-15% of variability in ageing rates can be explained by genetic variants. Thus, age can be considered an environmental phenotype.

Epigenetic data, (for example DNA methylation data), can be used to quantify age at the molecular level, which is normally referred to as biological age. Differences between biological age and the chronological age of an individual are the best molecular proxy to estimate age-associated risk. As such, higher biological ages are associated with a higher risk of developing age-related diseases (such as different types of cancer, diabetes type II, cardiovascular disease or neurodegenerative diseases) and all-cause mortality. It is possible to slow down the processes that lead to increases in biological age, for example using lifestyle interventions, which makes biological age a fantastic metric of the effectiveness of different lifestyle interventions.

The method comprises estimating the biological age of an individual given the DNA methylation patterns of the cells in his or her saliva, although it will be appreciated that the method could equally be applied using other forms of epigenetic data such as

alternative DNA modifications (such as 5-hydroxymethylcytosine, 5-formylcytosine, 5-carboxylcytosine or 6-methyladenine), histone modifications (such as acetylation, methylation, phosphorylation, ubiquitylation, sumoylation or biotinylation), chromatin accessibility, nucleosome position, the binding of transcription factors or other nuclear proteins to the DNA, non-coding RNAs, RNA modifications (such as N6-methyladenosine, N1-methyladenosine, pseudouridine, 5-methylcytosine, 2'-O-methylations or any other modifications registered in a database like RMBase), aggregation of prion-like proteins and any other molecular manifestations of cellular memory and gene expression regulation. In this embodiment, methylation values for > 5,500,000 cytosines in CpG context are used as the independent variables (also known as covariates) and the dependent variable (the one to be predicted) is the age of the sample. No other methods are available that use such a high-dimensional DNA methylation dataset generated using next-generation sequencing technologies for this purpose.

Given the high dimensionality and redundancy of the data, the number of features (i.e. covariates) may need to be reduced before the final training with the machine learning model. In an embodiment of this method, a concatenated random forest regression modelling strategy (e.g. implemented using the *scikit-learn* Python library) can be applied to yield accurate results:

- Random forest one (RF1). The purpose of this step is to find those cytosines whose methylation status can better predict age during human lifespan. Those cytosines will have a high value for their feature importance (for example, a threshold of 10^{-5} is set). This reduces the number of cytosines (in this example, to 1211). The characteristics and error of the predictions in the test set can be seen in Figure 9A. The Pearson correlation coefficient between chronological age and biological age is determined to be 0.9826, with a median absolute error of 2.3537 years.

- Random forest two (RF2). Once the best features have been found, another random forest regressor is fit to obtain the final model, as shown in Figure 9B. This ensures a reduction in the median absolute error of the model (from 2.3537 years in RF1 to 1.8932 years in RF2) and further decreases the

likelihood of overfitting. The Pearson correlation coefficient between chronological age and biological age is determined to be 0.9846.

Model hyperparameters for the above-mentioned random forest algorithms RF1 and RF2 are shown in Table 2A and Table 2B respectively. It will be appreciated that the model hyperparameters shown in Tables 2A and 2B are exemplary, and that other hyperparameters could also be used. For example, the hyperparameter 'max_features' in relation to RF1 could be between 50 and 50000. A 70% training:30% test split in the data set was used in this case, although other splits may be used depending on the data set (for example, a split of between 50% training:50% test split and 90% training:10% test split, for example, 80% training:20% test, or 60% training:40% test). In some embodiments, the hyperparameters are further tuned using cross-validation with the *GridSearchCV* function in Python.

In some embodiments, the dependent variable is a transformed version of age (that accounts for different ageing rates during different lifespan periods). Alternatively, the relationship between biological age and chronological age can be modelled *ad hoc* after prediction.

In some embodiments, the methylation covariates can be derived from further transformations of the data (e.g. selecting cytosines with high levels of correlation in their methylation values with those found as a result of RF1, or constructing differentially methylated regions that summarise several cytosines, or performing dimensionality reduction with PCA and using the PCs as covariates, etc.) Alternatively, it will be appreciated that other epigenetic data could be used in place of DNA methylation data to train a machine learning model to characterise phenotypes of an individual, wherein the phenotypes are environmentally influenced and are risk factors for one or more diseases.

Additional covariates can be added during the training process, such as sex, principal components that capture technical variation between batches, cell composition counts (that can be inferred from bulk DNA methylation data using different types of software), the genotype of specific variants, etc.

Training machine learning models to predict other environmental phenotypes from DNA methylation data

Other environmental phenotypes that capture the effects of lifestyle factors can also be predicted from epigenetic data (for example, DNA methylation data) using machine learning algorithms. These include, but are not limited to, smoke exposure and metabolic state.

Quantifying exposure to different chemicals from tobacco smoke and potentially air pollutants can be achieved using machine learning algorithms analyzing epigenetic data. The DNA methylation signature follows a dose-response curve and disappears to a great extent (and over the period of several years) after exposure has stopped. Furthermore, it allows quantification of indirect sources of smoke exposure (such as second-hand smoke). This environmental phenotype (e.g., a phenotype which is influenced by environmental factors) is associated with a risk of developing smoking-related diseases, such as lung cancer.

In some embodiments, in order to train a machine learning model to predict or characterise the phenotype of smoke exposure, self-reported smoking status is used as the dependent variable. In other embodiments, other variables that capture smoke exposure such as the number of cigarettes smoked per unit time, or the level of urinary nicotine equivalents, can be used during the training procedure.

In one embodiment, the same training strategy as described above with respect to determining or characterising biological age was used, but using a classification random forest framework with the following categories: ‘never smoker’ (currently 0 cigarettes per day and no past smoking history) and ‘current smoker’ (currently >0 cigarettes per day). Figure 10 shows the results for all samples after running the prediction or characterisation in RF2. It is worth noting that former smokers (currently 0 cigarettes per day but with past smoking history) follow a probability distribution with a median value located between the median values of the probability distributions for ‘never smoker’ and ‘current smoker’, as expected (indicated by the central spot 1001 in the boxplot 1000). Therefore, the probability of belonging to the ‘current smoker’ category can be used as a continuous variable proportional to the exposure (and its dynamics over time). The shaded region 1002 surrounding the box plot 1000 indicates the probability distribution within the ‘former smoker’ classification. The

box plots 1003, 1006, the central spots 1004, 1007 and shaded regions 1005, 1008 depict similar information for the 'never smoker' and 'current smoker' classifications respectively. The number of coefficients (i.e., cytosines in CpG context) found by RF1 was 2265 when characterising smoke exposure.

Turning to Figure 11, quantifying metabolic state can be achieved in a similar manner to that described above. Doing so comprises quantifying the degree of similarity of a person's epigenetic data with the epigenome of an obese individual (for example, characterised by a gene expression profile of high adiposity, insulin resistance and inflammation). As a consequence, metabolic state can be considered a risk factor for several diseases such as metabolic syndrome or type II diabetes. Using self-reported body mass index (BMI), different metabolic categories can be defined, such as underweight (for example, BMI < 18.5), normal weight ($18.5 \leq$ BMI < 25), overweight ($25 \leq$ BMI < 30) and obese (BMI ≥ 30). In one embodiment, the same training strategy as described above with respect to determining or characterising biological age was used, but using a classification random forest framework with the following categories: low weight (underweight) and high weight (overweight and obese). Afterwards, the model was tested in individuals of normal weight. Figure 11 shows the results for all of the samples after running the prediction or characterisation in RF2. It is worth nothing that individuals of normal weight follow a probability distribution with a median value located between the median values of the probability distributions for low weight and high weight categories, as expected. In other words, the probability of belonging to the high weight category can be used as a 'molecular BMI' measurement that captures the risk of developing metabolic-associated disease. The box plots, median values and probability distributions are depicted in Figure 11 using a substantially identical approach to those depicted in Figure 10. In other embodiments, other variables that capture the metabolic state of an individual, such as waist-to-hip ratio, can be used during the training procedure. The number of coefficients (e.g., DNA methylation sites or CpG sites) found by RF1 was 5409 when characterising metabolic state.

These examples highlight the ability of epigenetic data (for example, DNA methylation data to characterise environmental phenotypes (e.g., phenotypes that are influenced by environmental factors) and provide molecular assessments for the up-to-date risk of developing one or more diseases. It will be appreciated that a machine

learning algorithm could be trained on a combination of two or more environmental phenotypes or risk factors.

Figure 12 shows a schematic illustration of how environmental phenotypes can be used once characterised from epigenetic data using a machine learning algorithm. Each environmental phenotype A, B ... and so on may be a risk factor for one or more diseases D1, D2 ... and so on. The magnitude of the risk factor in respect of each disease is the environmental phenotype (for example, a continuous variable such as a probability of belonging to a 'current smoker' category) multiplied by a weighting coefficient, illustrated in Figure 12 by the weighting coefficients W_{XN}, wherein X is the environmental phenotype (selected from A, B ... and so on) and N is the disease for which the environmental phenotype is a risk factor. For example, the weighting coefficient in respect of disease D1 for environmental phenotype A is given as W_{A1}. A larger weighting coefficient may indicate that an environmental phenotype is a significant risk factor for a certain disease. For example, the weighting coefficient may be between 0 and 1, wherein 0 indicates that the environmental phenotype is an insignificant or non-existent risk factor for a certain disease, and wherein 1 indicates that the environmental phenotype is a significant or certain risk factor for a certain disease. Each environmental phenotype may have a different value of weighting coefficient in respect of different diseases. For example, W_{A1} may have a different value to W_{A2}. Once the environmental phenotypes have been multiplied by the relevant weighting coefficients to give a magnitude of the risk factor that impacts a given disease, a mortality coefficient is determined for each disease. The mortality coefficient may indicate a probability of mortality based on the development of the given disease (which in turn depends on the magnitude of the risk factors). Each of a plurality of mortality coefficients may be used to provide an overall mortality rate. The mortality rate may alternatively be substituted by a likely time to death based on the magnitude of the risk factors.

Figure 13 shows a schematic of a digital platform 700 for determining a health state of a subject according to a third aspect of the invention. The digital platform 700 comprises a data storage module 710, a data analysis module 720 and a user module 730.

The data storage module 710 is configured to store epigenetic data of the subject. The data storage module 710 may also be further configured to store at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject. The lifestyle data, health data, well-being data and environmental data of the subject may comprise information from blood-based measurements and/or other biological and/or medical sampling procedures, information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking platforms and social media platforms, and information uploaded from third-party providers.

The data analysis module 720 is in communication with the data storage module 710. The data analysis module 720 is configured to use a machine learning algorithm to characterise phenotypes of the subject from the epigenetic data of the subject stored in the data storage module 710, to indicate a health state of the subject.

The user module 730 is in communication with the data analysis module 720. The user module 730 is configured to display the characterised phenotypes and/or the health state of the subject on the user device 740. The user module 730 is also configured to be controllable via a user interface of the user device 740. The user module 730 may be located on, or remotely from, the user device 740. The user device 740 may be a computer (e.g., a personal computer such as a desktop PC or a laptop PC), a smartphone, a tablet, or any other suitable electronic device.

The digital platform 700 may be used by a plurality of users. The plurality of users may use the digital platform 700 simultaneously.

The user module 730 is also further configured to display, on the user device 740, one or more proposed medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or the health state of the subject.

In the embodiment shown, the digital platform 700 also comprises a training module 755. The training module 755 is in communication with the data storage module 710, the data analysis module 720 and the user module 730. The training module 755 is configured to selectively update the machine learning algorithm used by the data

analysis module 720 using epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject. The training module 755 is configured to selectively update the machine learning algorithm using data obtained from the data storage module 710. The training module 755 is further configured to provide an updated machine learning algorithm (to characterise phenotypes of the subject from epigenetic data to indicate a health state of the subject) to the data analysis module 720. In some embodiments, the training module 755 may be configured to replace the machine learning algorithm with the updated machine learning algorithm on the data analysis module 720 (i.e., by deleting the machine learning algorithm from the data analysis module 720 and providing the updated machine learning algorithm instead). In alternative embodiments, the digital platform 700 may not comprise a training module.

The training module 755 may be configured to periodically obtain data from the data storage module 710 to update the machine learning algorithm used by the data analysis module 720. The training module 755 may be configured to periodically check whether any additional data has been provided to the data storage module 710 by one or more users. Data provided by the user to the data storage module 710 can be used to expand a population from which a training data set for training the machine learning algorithm may be selected. Periodically (e.g., once a month, once every three months, once every six months, annually) retraining the machine learning algorithm using additional data provided by the user enables the predictive power and accuracy of the machine learning algorithm to characterise phenotypes from epigenetic data to be continuously improved. In this way, a plurality of users may benefit from the improved predictive and characterising power of the machine learning algorithm updated using the subject data of one or more other users stored in the data storage module 710.

In the embodiment shown, the digital platform 700 further comprises a first security module 750a. The first security module 750a is configured to determine whether or not the training module 755 has been granted permission to obtain epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module 710. The user may grant permission for the training module 755 to obtain subject data associated with the user from the data storage module 710 through a user command input via the user interface of the user device 740.

Once the first security module 750a has determined that the training module 755 has been granted permission to obtain subject data associated with the user from the data storage module 710, the first security module 750a is configured to allow the training module 755 to obtain subject data associated with the user from the data storage module 710. The training module 755 can then update the machine learning algorithm used by the data analysis module 720 using the obtained subject data associated with the user. If instead the first security module 750a determines that the training module 755 has not been granted permission to obtain subject data associated with the user from the data storage module 710, the first security module 750a is configured to deny the training module 755 access to subject data associated with the user stored in the data storage module 710. In alternative embodiments, the digital platform 700 may not comprise a first security module.

In the embodiment shown, the user module 730 is in communication with the data storage module 710. The user module 730 is configured to selectively provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module 710. The user module 730 may be configured to selectively provide the additional data to the data storage module responsive to a user command input via the user interface of the user device 740. In alternative embodiments, the user module 730 may not be in communication with the data storage module 710.

In the embodiment shown, the user module 730 is configured to automatically access and retrieve at least one of lifestyle data, health data, well-being data and environmental data of the subject. The user module 730 may be in wired or wireless communication (via the user device 740) with one or more of health trackers associated with the subject, wearable sensors configured to monitor physiological parameters of the subject, biosensor devices associated with the subject, a mobile telephone of the subject, and physical devices connected to the Internet of Things and associated with the subject. From such communication with external devices, the user module 730 may access and retrieve information obtained by those devices (or at least a part of the information obtained by those devices). The user module 730 may also, from such communication, access and retrieve information from online surveys completed by the subject, information from social networking platforms and social media platforms used by or visited by the subject, and information uploaded from third-party providers relating to the subject (or at least a part of such information). The user may selectively allow the user module 730 to access one or more external

devices individually, or may selectively grant blanket permission for the user module 730 to access all external devices of the user or subject which are in communication with the user module 730.

Once the user module 730 has accessed and retrieved the additional data from the above described sources of information, the user module 730 is configured to automatically provide the additional data to the data storage module 710. In alternative embodiments, the user module 730 may require permission from the user to automatically access and retrieve additional data (or at least a part of the additional data) via wireless communication with one or more of the above described sources of information. The user module 730 may be configured to selectively automatically access and retrieve additional data (or at least a part of the additional data), and provide the additional data (or at least a part of the additional data) to the data storage module 710, responsive to a user command input via the user interface of the user device 740. For example, the user may only wish to provide additional data relating to some aspects of his or her lifestyle, health and well-being and environment (such as smoking, activity levels etc.), whilst not sharing or providing other aspects of that data (such as location via GPS data etc.).

The digital platform 700 further comprises a second security module 750b. In embodiments where the user module 730 is configured to be in communication with the data storage module 710, the second security module 750b is configured to determine whether or not the user module 730 has been granted permission by the user to provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module 730. The second security module 750 is further configured to allow the user module 730 to provide the additional data to the data storage module 710 on determining that the user module 730 has been granted permission by the user to do so (e.g., by a user command input via the user interface of the user device 740).

To increase the security of the additional data associated with the subject, the second security module 750b may be further configured to temporarily store a copy of the at least one of lifestyle data, health data, well-being data and environmental data. The additional data may be stored in a memory of the second security module 750b. The second security module 750b may then determine whether or not the user module 730 has been granted permission to provide the additional data to the data storage module

710. If the second security module 750b determines that permission has been granted by the user, the second security module 750b is configured to provide the copy of the additional data to the data storage module 710 and delete the copy of the additional data stored in the memory of the second security module 750b. If the second security module 750b determines that permission has not been granted by the user, the second security module 750 deletes the copy of the additional data and does not provide the copy of the additional data to the data storage module 710.

Similar security provisions may be provided for the first security module 750a.

The user module 730 and the second security module 750b may comprise one or more encrypters to encrypt the at least one lifestyle data, health data, well-being data and environmental data. The data storage module 710 may be further configured to only receive the data from the user module 730 or the second security module 750b if the data is successfully unencrypted by one or more corresponding decrypters located at the data storage module 710. Instead, or additionally, the data may be associated with a security key by the user module 730 and/or the second security module 750b. The data storage module 710 may comprise a memory, the memory of the data storage module 710 storing a security key. The data storage module may be configured to only receive the data from the user module 730 or the second security module 750b if the security key associated with the data (provided by the user module 730 and/or the second security module 750b) matches the security key stored in the memory of the data storage module 710.

Similar security provisions may be provided for the training module 755, data storage module 710 and first security module 750a in respect of the training module obtaining subject data associated with a user from the data storage module 710.

In addition to the security provisions described above, the second security module 750b may be further configured to anonymise the at least one of lifestyle data, health data, well-being data and environmental data. The second security module 750b may be configured to anonymise the additional data after determining that the user module 730 has been granted permission to provide the additional data to the data storage module 710.

Similar security provisions may be provided for the first security module 750a.

In alternative embodiments, the digital platform 700 may not comprise a second security module.

Each of the data storage module 710, the data analysis module 720, the user module 730, the first security module 750a, the second security module 750b, and the training module 755 may be implemented on the user device 740, or may be implemented remotely from the user device 740 (i.e., on a computing device located remotely from the user device 740, for example a remote server). Each of the modules may be implemented on a separate computing device, for example an individual server. One or more of the modules of the digital platform 700 may be in wired or wireless communication with one another.

The user module 730 comprises a plurality of apps 760 (an example of an app 760, relating to a characterised air pollutants phenotype, is shown in Figure 14). Each of the apps 760 is a software application executable on the user device 740 Each of the plurality of apps 760 is configured to display a single characterised phenotype, or one or more related characterised phenotypes on the user device 740. The apps 760 each contain information about one area or category of health and well-being (for example lifestyle, mental wellbeing, immune system, metabolism, family/pregnancy, toxin exposure). Related characterised phenotypes may be, for example, sleep quality and/or quality, and chronotype. Another example of related characterised phenotypes may be dietary components, and obesity. In general, related characterised phenotypes are characterised phenotypes of which one of the phenotypes may affect one or more of the related phenotypes. The user module 730 is configured to enable the user to select one or more of the plurality of apps 760 to be displayed simultaneously on the user device 740. Each of the plurality of apps 760 is an interactive report, wherein the user may interact with the report via the user interface of the user device 740. Each app 760 may contain one or more widgets 770 (e.g., graphs, interactive trackers).

Furthermore, the user module 730 is configured to display a “home” page on the user device 740 when the user selects to display the home page on the user device 740. The home page displays options which may be selected by the user via the user interface of the user device 740 in order to access specific functionalities of the digital platform. For example, the user may be able to access an option to upload at least one of lifestyle data, health data, well-being data and environmental data to the data storage module 710 from the home page. The user may also be able to access one or more

apps 760 from the home page in order to display the apps 760 that are of most interest to the user by using the user interface of the user device 740. Individual widgets 770 from one or more apps 760 may also be exported to the home page such that the widgets 770 are temporarily or permanently displayed on the home page. Individual widgets 770 from one or more apps 760 may also be exported to electronic locations (e.g., a computing device or internet location) separately from the one or more apps 760. In this way, key data contained in and/or displayed by the widgets 770 may be communicated separately from the one or more apps 760.

The apps 760 allow the user to interact in a continuous manner with many of the widgets and/or interactive trackers (e.g., by inputting sleep hours of the subject to the widget or trackers), and allow for integration of other sources of additional data (such as wearable sensors and health trackers etc.). An advantage of this feature is that users may feel more personally invested in the platform, and therefore will remain engaged with the digital platform.

The functionality of each app 760 may be augmented by up-to-date scientific information relating to the characterised phenotypes or categories which each app 760 relates to, with curated scientific references for users wishing to expand their knowledge.

Once within the digital platform 700, users will be invited to share subject data in an anonymised manner for use in third-party research initiatives. Users will be rewarded for agreeing to share their data through an altruistic scoring system. Through subsequent engagement in additional studies, users will again be invited to contribute and will be rewarded for their contribution. Importantly, this contribution will be variable depending on the nature, impact and size of the study being conducted, resulting in a variable reward. Subsequent to agreeing to the sharing of said data, the user will be required to provide (in various manners depending on the study) metadata attributes that will be required for any given study (investment). These contributions have the added benefit that they allow the user to more trivially engage in additional future studies since certain metadata may already be present. This behavior is encouraged through making these sharing characteristics visible within the digital platform 700, for example on the home page. By exploiting this altruistic side of human psychology, a greater number of users may be retained, greatly reducing user turnover rate in the long-term and benefiting society as a whole.

Users may also be engaged with and retained in the digital platform 700 via the provision of personalized medical and/or lifestyle recommendations or interventions. The recommendations or interventions link with the tracking information shown in the widgets 770. The recommendations may include courses, programmes, products or information (e.g., videos or text). The recommendations may be contained within the digital platform 700, enabling customers to simply and easily make proactive changes to improve the characterised phenotypes and/or the health state of the subject. The initial trigger for the user to engage in actions defined in such recommendations may come upon obtaining access to a specific app 760 showing a certain phenotypic variable or risk. The user may be prompted to ensure that the phenotypic variable or phenotypic risk improves in a predetermined time period (for example, a week, a month or a year). Users may receive additional triggers as derived from the tracking information in the widgets 770 found within each app 760 that will encourage the user to continue enacting the proactive change.

Weekly, daily and monthly tracking information is also provided in the widgets 770 to encourage the users to remain engaged with the digital platform 700. The time period over which tracking information is provided in the widgets 770 will vary between apps 760, depending upon the nature of the characterised phenotype. Users may be encouraged to share certain data with the digital platform by providing additional data to the data storage module 710 (e.g., GPS location for air pollution data, health tracker and/or wearable sensor data for exercise). Subsequently, users are rewarded through the visualization and analysis of this data in the widgets 770. Users are encouraged to improve characterised phenotypes (i.e., phenotypic variables or phenotypic risks) by making medical and/or lifestyle interventions in their daily lives. Subsequent triggers may be defined at fixed analysis time points (e.g., once a week reporting, or once a month reporting).

The digital platform 700 is configured to motivate users to provide and share data (i.e., epigenetic data, genetic data, lifestyle data, health and well-being data, environmental data, metadata), for example, to third parties for research projects. The digital platform is configured to offer incentives and rewards in return for the user sharing data. Providing and sharing of data by the user may comprise providing at least lifestyle data, health data, well-being data and environmental data to the data storage module 710. For example, the user may be required to provide or generate additional new data (e.g., via a blood sample to be analysed, or a self-assessment

questionnaire) for a given study that the user wishes to take part in. Alternatively or additionally, providing and sharing of data by the user may also comprise granting permission, via the user module 730, for at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data to be shared with third parties for research purposes, or to be provided to the training module 755 to update the machine learning algorithm used by the data analysis module 720 for the benefit of both the user and a plurality of other users (due to improved accuracy and predictive/characterising power of the machine learning algorithm after updating/retraining via subject data). The user may be able to see (e.g., via the user device 740) to what extent the user is contributing by sharing data via contribution analytics (e.g., performed by the data analysis module 720) identifying, for example, the number and/or type of studies the user has taken part in. Being informed of this contribution analysis data may help to incentivise the user to share further data in future.

The incentives and/or rewards offered by the digital platform 700 may comprise a currency configured to be exchanged for additional services and/or functionality provided by the digital platform 700. For example, the user may exchange currency obtained as a result of providing and/or sharing data for access to apps 760 which relate to characterised chronotypes that are otherwise not available on the digital platform (referred to hereinafter as “internal currency”). The internal currency may be a currency that is recognised only by the digital platform 700 (i.e., the currency is unique to the digital platform 700).

Alternatively, the internal currency may be a currency that is recognised worldwide, for example GBP, USD, EUR JPY, AUD, CHF etc. In such cases, the user module 730 is configured to enable the user to transfer the value of the internal currency obtained to a bank account of the user via the user interface of the user device 740.

Alternatively, additional services and/or functionality of the digital platform 700 (i.e., access to additional apps 760) may be exchanged for currency not obtained via the user providing and sharing data (known hereinafter as “external currency”). External currency may be a currency that is recognised internationally, such as GBP, USD, EUR, JPY, AUD, CHF etc. External currency may also be exchanged within the digital platform 700 for a currency recognised only by the digital platform 700. Once exchanged, the currency recognised only by the digital platform may be exchanged for

additional services and/or functionality of the digital platform 700. This is shown in Figure 15, which depicts a display of a user device displaying a "store page" 780 of the user module 730, from which additional apps 760 may be purchased.

As shown in Figure 15, the apps 760 may be grouped in related categories on the store page 780, such as lifestyle, mental health, immune system etc. Apps made available to the digital platform 700 most recently may be located in a "featured" category containing the newest apps 760. Apps 760 may be purchased individually or a plurality of apps 760 may be purchased simultaneously.

The digital platform 700 could be used in a business or corporate setting (e.g., a business-to-business setting) if one or more users or subjects give their permission for their data to be used in such a capacity. In this way, for example, the characterised phenotypes of one or more subjects (e.g., employees) could be used as an indicator of workplace wellness for the one or more subjects or employees (e.g, individual employees, groups of subjects forming a part or a whole of company departments, groups of subjects forming a part of a whole of a company). The user module 730 could display or suggest one or more proposed medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or the health state of the subject or subjects. The user module 730 and/or the data analysis module 720 may be configured to average or aggregate the characterised phenotypes of the subjects before providing displaying or suggested one or more proposed medical and/or lifestyle interventions, or may display or suggest one or more proposed medical and/or lifestyle interventions for one or more specific subjects.

Alternatively or additionally, the characterised phenotypes of the subject or subjects in a business or corporate setting could be used to aid insurance providers in determining accurate and appropriate workplace insurance premiums, both for individual specific subjects and for a general workforce (e.g., using averaged or aggregated characterised phenotypes of a plurality of subjects).

From reading the present disclosure, other variations and modifications will be apparent to the skilled person. Such variations and modifications may involve equivalent and other features which are already known in the art of epigenetic analysis, and which may be used instead of, or in addition to, features already described herein.

Although the appended claims are directed to particular combinations of features, it should be understood that the scope of the disclosure of the present invention also includes any novel feature or any novel combination of features disclosed herein either explicitly or implicitly or any generalisation thereof, whether or not it relates to the same invention as presently claimed in any claim and whether or not it mitigates any or all of the same technical problems as does the present invention.

Features which are described in the context of separate embodiments may also be provided in combination in a single embodiment. Conversely, various features which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable sub-combination. The applicant hereby gives notice that new claims may be formulated to such features and/or combinations of such features during the prosecution of the present application or of any further application derived therefrom.

For the sake of completeness, it is also stated that the term "comprising" does not exclude other elements or steps, the term "a" or "an" does not exclude a plurality, a single processor or other unit may fulfill the functions of several means recited in the claims and any reference signs in the claims shall not be construed as limiting the scope of the claims.

RandomForestRegressor(bootstrap=True	criterion='mse'	max_depth=None	
max_features=20000	max_leaf_nodes=None		
min_impurity_decrease=0.0	min_impurity_split=None		
min_samples_leaf=1	min_samples_split=2		
min_weight_fraction_leaf=0.0	n_estimators=5000	n_jobs=-1	
oob_score=False	random_state=0	verbose=1	warm_start=False)

Table 2A

RandomForestRegressor(bootstrap=True	criterion='mse'	max_depth=None	
max_features='auto'	max_leaf_nodes=None		
min_impurity_decrease=0.0	min_impurity_split=None		
min_samples_leaf=1	min_samples_split=2		
min_weight_fraction_leaf=0.0	n_estimators=5000	n_jobs=-1	
oob_score=False	random_state=0	verbose=1	warm_start=False)

Table 2B

LIST OF REFERENCES

Vincenzo EA Russo, Robert A Martienssen, and Arthur D Riggs. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, 1996.

Robert Feil and Mario F Fraga. Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, 13(2):97, 2012.

Keith D Robertson. DNA methylation and human disease. *Nature Reviews Genetics*, 6 (8):597, 2005.

Jeremy J Day and J David Sweatt. DNA methylation and memory formation. *Nature neuroscience*, 13(11):1319, 2010.

Peter W Laird. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191, 2010.

Andrew E Teschendorff and Shijie C Zheng. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. *Epigenomics*, 9(5):757–768, 2017.

Christof Angermueller, Heather J Lee, Wolf Reik, and Oliver Stegle. Deepcpg: accurate prediction of single-cell DNA methylation states using deep learning. *Genome biology*, 18(1):67, 2017.

Jenny Van Dongen, Michel G Nivard, Gonneke Willemsen, Jouke-Jan Hottenga, Quinta Helmer, Conor V Dolan, Erik A Ehli, Gareth E Davies, Maarten Van Iterson, Charles E Breeze, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature communications*, 7:11115, 2016.

Michael J Ziller, Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T-Y Tsai, Oliver Kohlbacher, Philip L De Jager, Evan D Rosen, David A Bennett, Bradley E Bernstein, et al. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500(7463):477, 2013.

Alicia K Smith, Varun Kilaru, Torsten Klengel, Kristina B Mercer, Bekh Bradley, Karen N Conneely, Kerry J Ressler, and Elisabeth B Binder. DNA extracted from saliva for methylation studies of psychiatric traits: evidence tissue specificity and

relatedness to brain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(1):36–44, 2015.

Yu-Hsuan Chuang, Kimberly C Paul, Jeff M Bronstein, Yvette Bordelon, Steve Horvath, and Beate Ritz. Parkinson's disease is associated with DNA methylation levels in human blood and saliva. *Genome medicine*, 9(1):76, 2017.

Sabine AS Langie, Katarzyna Szarc vel Szic, Ken Declerck, Sophie Traen, Gudrun Koppen, Guy Van Camp, Greet Schoeters, Wim Vanden Berghe, and Patrick De Boever. Whole-genome saliva and blood DNA methylation profiling in individuals with a respiratory allergy. *PloS one*, 11(3):e0151109, 2016.

Beth Wilmot, Rebecca Fry, Lisa Smeester, Erica D Musser, Jonathan Mill, and Joel T Nigg. Methylomic analysis of salivary DNA in childhood ADHD identifies altered DNA methylation in vipr2. *Journal of Child Psychology and Psychiatry*, 57(2):152–160, 2016.

Yenkai Lim, Yunxia Wan, Dimitrios Vagenas, Dmitry A Ovchinnikov, Chris FL Perry, Melissa J Davis, and Chamindie Punyadeera. Salivary DNA methylation panel to diagnose HPV-positive and HPV-negative head and neck cancers. *BMC cancer*, 16(1):749, 2016.

Kathryn Tully Oelsner, Yan Guo, Sophie Bao-Chieu To, Amy L Non, and Shari L Barkin. Maternal BMI as a predictor of methylation of obesity-related genes in saliva samples from preschool-age hispanic children at-risk for obesity. *BMC genomics*, 18 (1):57, 2017.

Trine B Rounge, Christian M Page, Maija Lepistö, Pekka Ellonen, Bettina K Andreassen, and Elisabete Weiderpass. Genome-wide DNA methylation in saliva and body size of adolescent girls. *Epigenomics*, 8(11):1495–1505, 2016.

Patricia Braun, Marie Hafner, Yasunori Nagahama, Benjamin Hing, Melissa McKane, Andrew Grossbach, Matthew Howard, Hiroto Kawasaki, James Potash, and Gen Shinozaki. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *European Neuropsychopharmacology*, 27:S506, 2017.

Nicklas Heine Staunstrup, Anna Starnawska, Mette Nyegaard, Anders Lade Nielsen, Anders Børglum, and Ole Mors. Saliva as a blood alternative for genome-wide DNA

methylation profiling by methylated DNA immunoprecipitation (MeDIP) sequencing. *Epigenomes*, 1(3):14, 2017.

Robert Lowe, Carolina Gemma, Huriya Beyan, Mohammed I Hawa, Alexandra Bazeos, R David Leslie, Alexandre Montpetit, Vardhman K Rakyan, and Sreeram V Ram-agopalan. Buccals are likely to be a more informative surrogate tissue than blood for epigenome-wide association studies. *Epigenetics*, 8(4):445–454, 2013.

Marie Forest, Kieran J O'Donnell, Greg Voisin, Helene Gaudreau, Julia L MacIsaac, Lisa M McEwen, Patricia P Silveira, Meir Steiner, Michael S Kobor, Michael J Meaney, et al. Agreement in DNA methylation levels from the Illumina 450K array across batches, tissues, and time. *Epigenetics*, 13(1):19–32, 2018.

Sascha Tierling, Beate Schmitt, and Jörn Walter. Comprehensive evaluation of commercial bisulfite-based DNA methylation kits and development of an alternative protocol with improved conversion performance. *Genetics & epigenetics*, 10:1179237X18766097, 2018.

Juan J Carmona, William P Accomando, Alexandra M Binder, John N Hutchinson, Lorena Pantano, Benedetta Izzi, Allan C Just, Xihong Lin, Joel Schwartz, Pantel S Vokonas, et al. Empirical comparison of reduced representation bisulfite sequencing and Infinium Beadchip reproducibility and coverage of DNA methylation in humans. *NPJ genomic medicine*, 2(1):13, 2017.

Fiona Allum, Xiaojian Shao, Frédéric Guénard, Marie-Michelle Simon, Stephan Busche, Maxime Caron, John Lambourne, Julie Lessard, Karolina Tandre, Åsa K Hedman, et al. Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nature communications*, 6: 7211, 2015.

Milos Pavlovic, Pradipta Ray, Kristina Pavlovic, Aaron Kotamarti, Min Chen, and Michael Q Zhang. Direction: a machine learning framework for predicting and characterizing DNA methylation and hydroxymethylation in mammalian genomes. *Bioinformatics*, 33(19):2986–2994, 2017.

Martin Widschwendter, Allison Jones, Iona Evans, Daniel Reisel, Joakim Dillner, Karin Sundström, Ewout W. Steyerberg, Yvonne Vergouwe, Odette Wegwarth, Felix G. Rebitschek, Uwe Siebert, Gaby Sroczynski, Inez D. de Beaufort, Ineke Bolt, David

Cibula, Michal Zikan, Line Bjørge, Nicoletta Colombo, Nadia Harbeck, Frank Dudbridge, Anne-Marie Tasse, Bartha M. Knoppers, Yann Joly, Andrew E. Teschendorff, and Nora Pashayan. Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nature Reviews Clinical Oncology*, 15:292–309, 2018.

Claims

1. A method of characterising phenotypes of a subject comprising:

characterising one or more phenotypes of a subject from epigenetic data of the subject using a machine learning algorithm;

wherein the one or more phenotypes are influenced by environmental factors and are risk factors for one or more diseases.

2. The method of claim 1, further comprising obtaining epigenetic data of the subject.

3. A method of determining and providing information regarding a health state of a subject, the method comprising:

a) providing a sample collection kit to the subject to obtain a biological sample;

b) extracting epigenetic data from the biological sample;

c) characterising one or more phenotypes of the subject from the epigenetic data using a machine learning algorithm to indicate the health state of the subject, wherein the phenotypes are influenced by environmental factors and are risk factors for one or more diseases; and

d) reporting the characterised phenotypes and/or health state of the subject.

4. The method of any preceding claim, further comprising proposing one or more medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or health state of the subject.

5. The method of any preceding claim, further comprising:

extracting genetic data from the biological sample of the subject; and

characterising phenotypes of the subject from at least the genetic data and the epigenetic data using the machine learning algorithm to indicate the health state of the subject.

6. The method of any preceding claim, further comprising:

obtaining at least one of lifestyle data, health data, well-being data and environmental data from the subject; and

characterising phenotypes of the subject from the at least one of lifestyle data, health data, well-being data and environmental data and the epigenetic data using the machine learning algorithm to indicate the health state of the subject.

7. The method of claim 6, wherein the lifestyle data, health data, well-being data and environment data comprises at least one of microbiome data, metabolomic data, proteomic data, imaging data, medical or clinical records from the subject and close relatives (including information about past and current diseases or conditions), age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

8. The method of claim 5, wherein both strands of DNA from the biological sample of the subject are used to extract genetic information from the biological sample.

9. The method of any preceding claim, wherein the biological sample comprises at least one of saliva, urine, blood or semen obtained from the subject.

10. The method of any preceding claim, wherein the characterised phenotypes comprise at least one of biological age, metabolic state and smoke exposure.

11. The method of any preceding claim, wherein the characterised phenotypes comprise at least one of microbiome data, metabolomic data, proteomic data, information about past and current diseases or conditions, age, gender, date of birth,

ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

12. The method of claim 6 or any claim dependent directly or indirectly from claim 6, wherein the lifestyle data, health data, well-being data and environmental data comprises at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking sites and social media sites, and information uploaded from third-party providers.

13. The method of any preceding claim, further comprising repeating the method at a plurality of points in time to indicate a stability of the characterised phenotypes and/or health state of the subject over time.

14. The method of any preceding claim, further comprising refining the machine learning algorithm using at least one of epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data obtained from the subject.

15. The method of claim 4 or any claim dependent directly or indirectly from claim 4, further comprising repeating the method a pre-determined time period after proposing the one or more medical and/or lifestyle interventions.

16. The method of any preceding claim, wherein characterising phenotypes of the subject comprises at least one of:

i) determining current phenotypes of the subject; and

ii) predicting past and/or future phenotypes of the subject;

17. The method of any preceding claim, wherein characterising phenotypes of the subject comprises at least one of:

i) calculating a value of a continuous variable within a phenotypic class; and

ii) calculating a probability of the subject belonging to a phenotypic class.

18. The method of any preceding claim, wherein the extracted epigenetic data comprises epigenetic data extracted from at least 1 million CpG sites.

19. The method of any preceding claim, wherein reporting the characterised phenotypes and/or health status of the subject comprises reporting the characterised phenotypes and/or health status of the subject to at least one of:

i) the subject;

ii) a medical professional;

iii) a significant other, family member or next of kin of the subject; and

iv) a third party such as an insurer or employer.

20. A method of developing a machine learning algorithm configured to characterise phenotypes of a subject, the method comprising:

a) providing a sample collection kit to each of a population of individuals to obtain biological samples from at least a subset of the population;

b) extracting epigenetic data from at least a subset of the biological samples;

c) obtaining at least one of lifestyle data, health data, well-being data and environmental data from at least a subset of the population;

d) collating the epigenetic data and the at least one of lifestyle data, health data, well-being data and environmental data in a training data set;

e) training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set.

21. The method of claim 20, wherein the lifestyle data, health data, well-being data and environmental data comprises at least one of microbiome data, metabolomic data, proteomic data, imaging data, medical or clinical records from the subject and close relatives (including information about past and current diseases or conditions), age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

22. The method of claim 20 or claim 21, wherein the lifestyle data, health data, well-being data and environmental data comprises at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking sites and social media sites, and information uploaded from third-party providers.

23. The method of any of claims 20 to 22, further comprising:

extracting genetic data from at least a subset of the biological samples;

collating the genetic data in the training data set; and

training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set.

24. The method of any of claims 20 to 23, further comprising:

performing steps a), b) and c) of claim 20 at a plurality of different points in time to obtain longitudinal epigenetic data and at least one of longitudinal lifestyle data, longitudinal health and well-being data and longitudinal environmental data;

collating the longitudinal epigenetic data and the at least one of longitudinal lifestyle data, longitudinal health and well-being data and longitudinal environmental data in the training data set; and

training the machine learning algorithm to characterise predicted past and/or future phenotypes from epigenetic data using the training data set.

25. The method of any of claims 20 to 24, further comprising:

collating at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the method of any of claims 1 to 19 in the training data set; and

further training the machine learning algorithm to characterise phenotypes from epigenetic data using the training data set.

26. The method of any of claims 20 to 25, wherein the phenotypes the machine learning algorithm is configured to characterise comprise at least one of microbiome data, metabolomic data, proteomic data, information about past and current diseases or conditions, age, gender, date of birth, ancestry, racial background, ethnicity, educational history, professional occupation, geographical and location history, smoking history, height, weight, body mass index, blood glucose level and other blood-based markers, blood pressure, heart rate, diet composition, folate levels and other nutrients deficiencies, alcohol consumption, coffee and tea consumption, medication history, mental status, anxiety levels, fatigue levels, stress levels, allergic reactions and predispositions, inflammation-related markers, infection history, childhood abuse history or other sources of traumatic experiences, sleeping patterns, travelling patterns, exercise and fitness-related variables, lung function, air pollutants exposure, pesticides exposure, heavy metals exposure (such as cadmium or lead), diesel exhaust exposure, persistent organic pollutants exposure, and exposure to any chemicals that leave a specific epigenetic signature (such as bisphenol A or diethylstilbestrol), pregnancy status, pregnancy-related conditions (pre, during and post birth) and chronotype.

27. A digital platform for determining a health state of a subject, the platform comprising:

a data storage module, the data storage module configured to store epigenetic data of the subject;

a data analysis module in communication with the data storage module, the data analysis module configured to use a machine learning algorithm to characterise phenotypes of the subject from the epigenetic data of the subject to indicate a health state of the subject;

a user module in communication with the data analysis module, the user module configured to display the characterised phenotypes and/or the health state of the subject on a user device and controllable by a user via the user device.

28. The digital platform of claim 27, wherein the user module is further configured to display on the user device one or more proposed medical and/or lifestyle interventions based upon the reported characterised phenotypes and/or health state of the subject.

29. The digital platform of claim 27 or claim 28, wherein the data storage module is further configured to store at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject.

30. The digital platform of claim 29, wherein the lifestyle data, health data, well-being data and environmental data of the subject comprises at least one of information from health trackers, information from wearable sensors configured to monitor physiological parameters of the subject, information from biosensor devices, information from a mobile telephone, information from physical devices connected to the Internet of Things, information from self-reported questionnaires or written surveys, information from online surveys, information from social networking platforms and social media platforms, and information uploaded from third-party providers.

31. The digital platform of claim 29 or claim 30, further comprising a training module in communication with the data storage module, the data analysis module and the user module, the training module configured to:

selectively update the machine learning algorithm used by the data analysis module using epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject obtained from the data storage module; and

provide an updated machine learning algorithm to characterise phenotypes of the subject from epigenetic data to indicate a health state of the subject to the data analysis module.

32. The digital platform of claim 31, further comprising a first security module, the first security module configured to:

determine whether or not the training module has been granted permission by the user, via the user module, to obtain epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module; and

allow the training module to obtain epigenetic data of the subject and at least one of genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module only if the first security module has determined that the training module has been granted permission to do so.

33. The digital platform of any of claims 29 to 32, wherein:

the user module is in communication with the data storage module; and

the user module is further configured to selectively provide at least one of lifestyle data, health data, well-being data and environmental data of the subject to the data storage module.

34. The digital platform of claim 33, wherein the user module is further configured to selectively:

automatically access and retrieve at least one of lifestyle data, health data, well-being data and environmental data of the subject; and

automatically provide at least part of the retrieved at least one of lifestyle data, health data, well-being data and environmental data of the subject to the data storage module.

35. The digital platform of claim 33 or claim 34, wherein the digital platform further comprises:

a second security module, the second security module configured to:

determine whether or not the user module has been granted permission by the user to provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module; and

allow the user module to provide at least one of lifestyle data, health data, well-being data and environmental data to the data storage module only if the second security module has determined that the user module has been granted permission to do so.

36. The digital platform of any of claims 32 to 35, wherein the first security module is further configured to anonymise the at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject after determining that the training module has been granted permission to obtain the at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject from the data storage module.

37. The digital platform of claim 35 or claim 36, wherein the second security module is further configured to anonymise the at least one of lifestyle data, health data, well-being data and environmental data of the subject after determining that the user module has been granted permission to provide the at least one of lifestyle data, health data, well-being data and environmental data to the data storage module.

38. The digital platform of any of claims 27 to 35, wherein the user module is further configured to display one or more of a plurality of apps on the user device, each app configured to display one or more related characterised phenotypes on the user device.

39. The digital platform of claim 29 or any claim dependent directly or indirectly from claim 29, wherein the digital platform is configured to motivate the user to share subject data by offering incentives and/or rewards in return for the user granting permission, via the user module, for at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject stored in the data storage module to be shared with third parties for research purposes.

40. The digital platform of claim 31 of any claim dependent directly or indirectly dependent from claim 31, wherein the digital platform is configured to motivate the user to share subject data by offering incentives and/or rewards in return for the user granting permission for the training module to obtain at least one of the epigenetic data, genetic data, lifestyle data, health data, well-being data and environmental data of the subject stored in the data storage module.

41. The digital platform of claim 39 or claim 40, wherein the incentives and/or rewards offered by the digital platform comprise:

additional services and/or functionality provided by the digital platform; and/or

a currency; wherein

the currency is configured to be:

i) exchanged for additional services and/or functionality provided by the digital platform; and/or

ii) withdrawn from digital platform by the user.

42. The digital platform of claim 41, wherein the currency is a currency that is only recognised by the digital platform.

43. The digital platform of any of claims 27 to 42, wherein two or more of the modules of the digital platform are in wireless communication with one another.

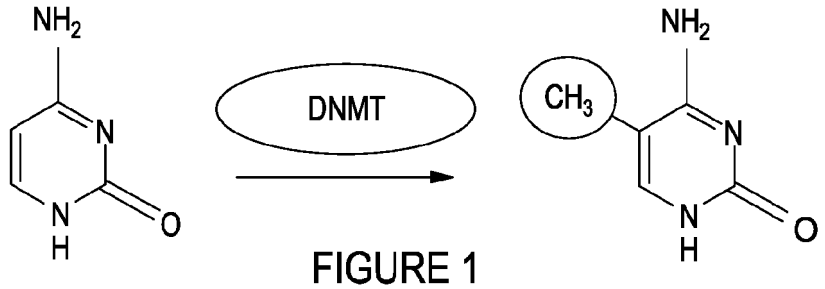

FIGURE 1

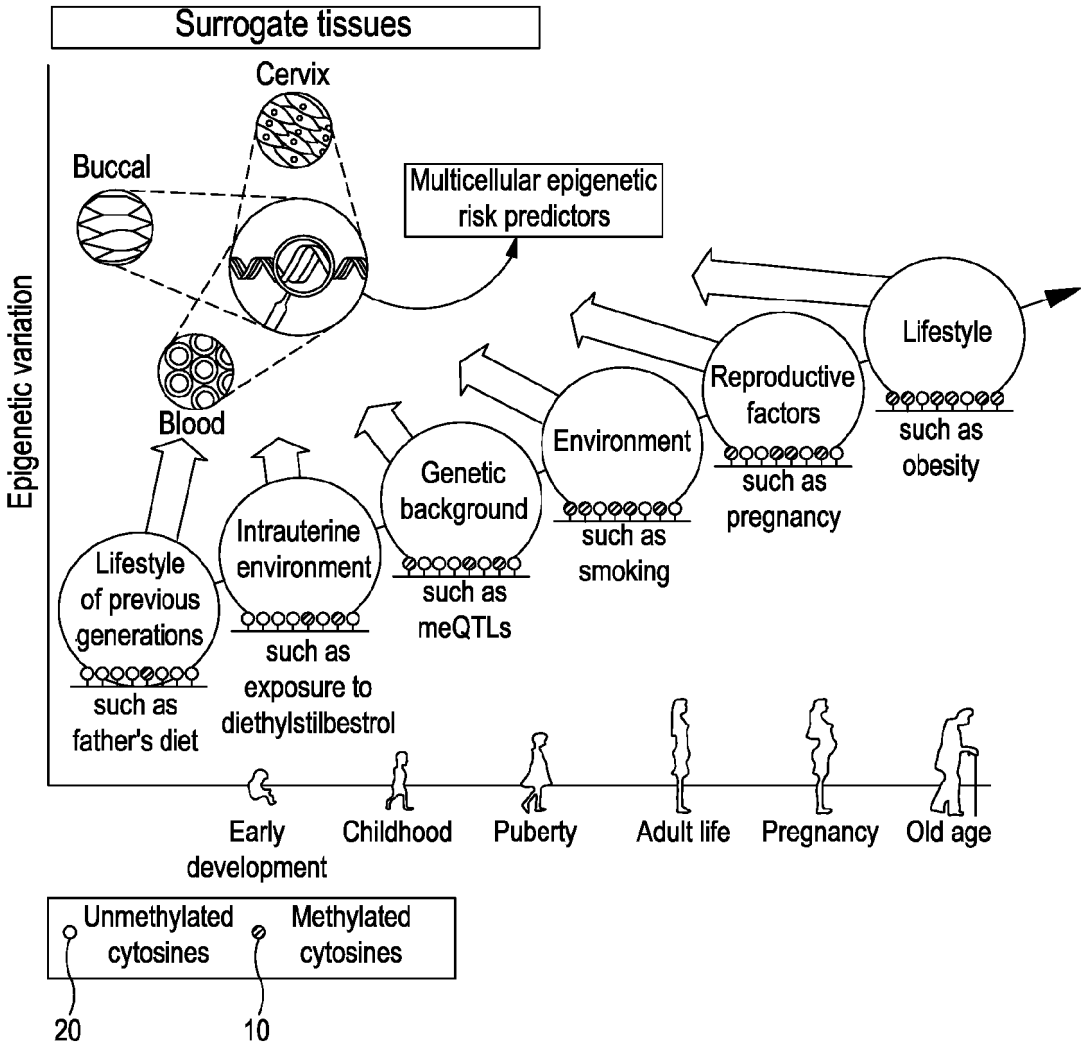

FIGURE 2

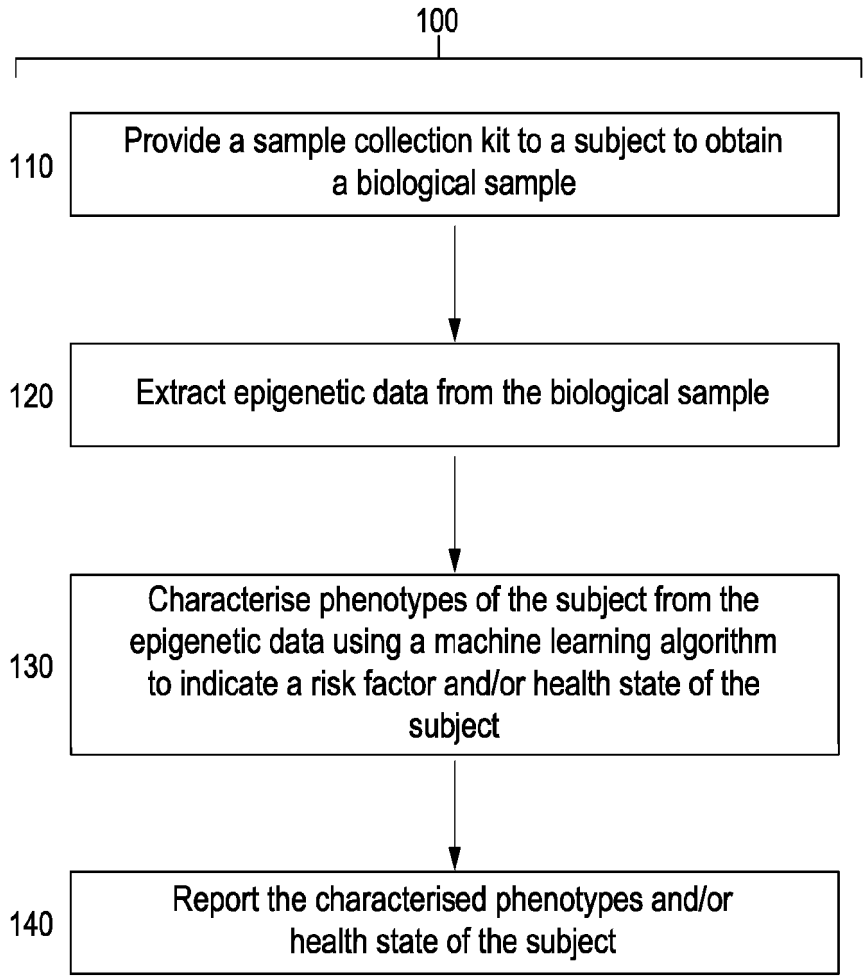
100
110
Provide a sample collection kit to a subject to obtain a biological sample
120
Extract epigenetic data from the biological sample
130
Characterise phenotypes of the subject from the epigenetic data using a machine learning algorithm to indicate a risk factor and/or health state of the subject
140
Report the characterised phenotypes and/or health state of the subject

FIGURE 3

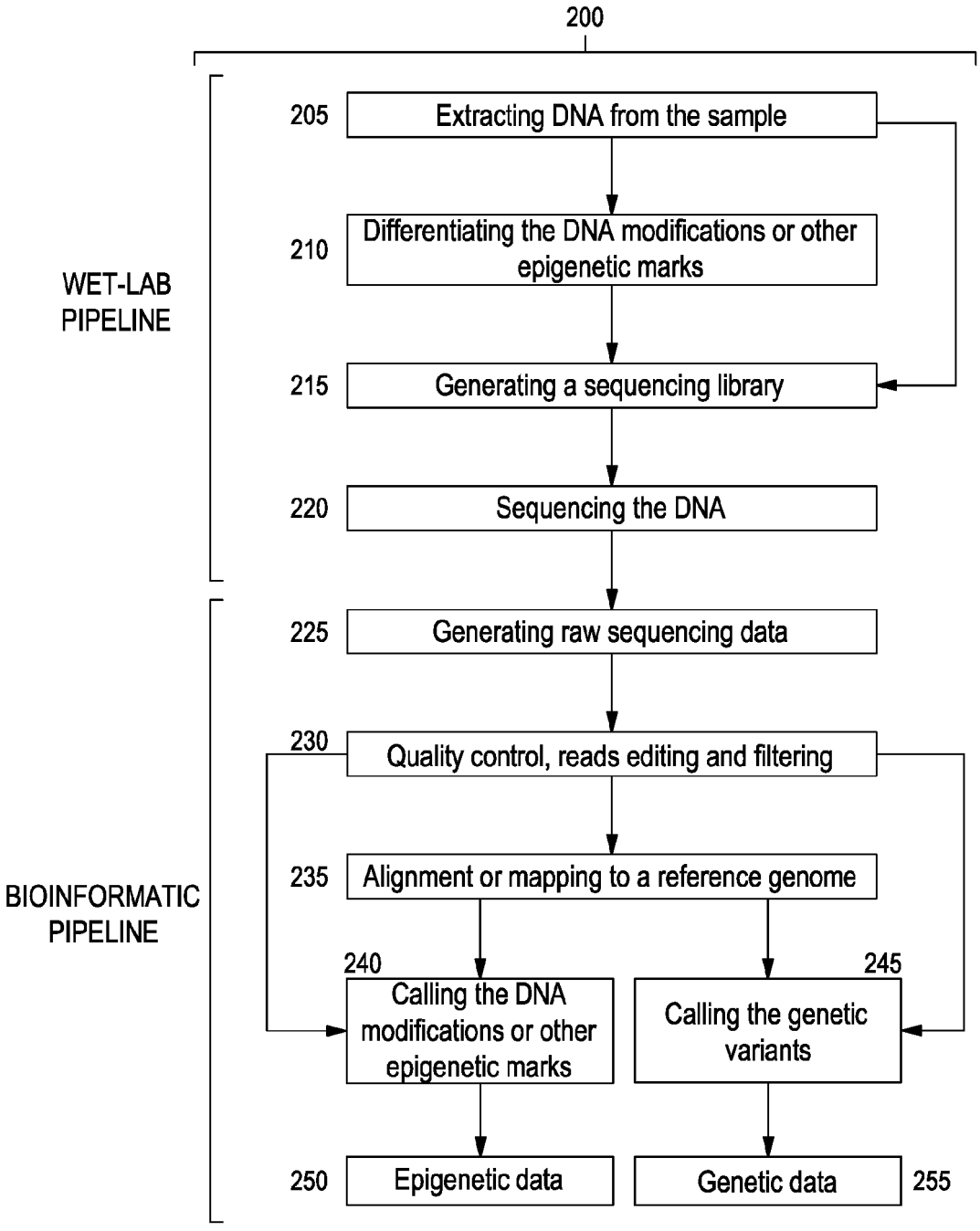
200
WET-LAB PIPELINE
205
Extracting DNA from the sample
210
Differentiating the DNA modifications or other epigenetic marks
215
Generating a sequencing library
220
Sequencing the DNA
BIOINFORMATIC PIPELINE
225
Generating raw sequencing data
230
Quality control, reads editing and filtering
235
Alignment or mapping to a reference genome
240
Calling the DNA modifications or other epigenetic marks
245
Calling the genetic variants
250
Epigenetic data
Genetic data
255

FIGURE 4

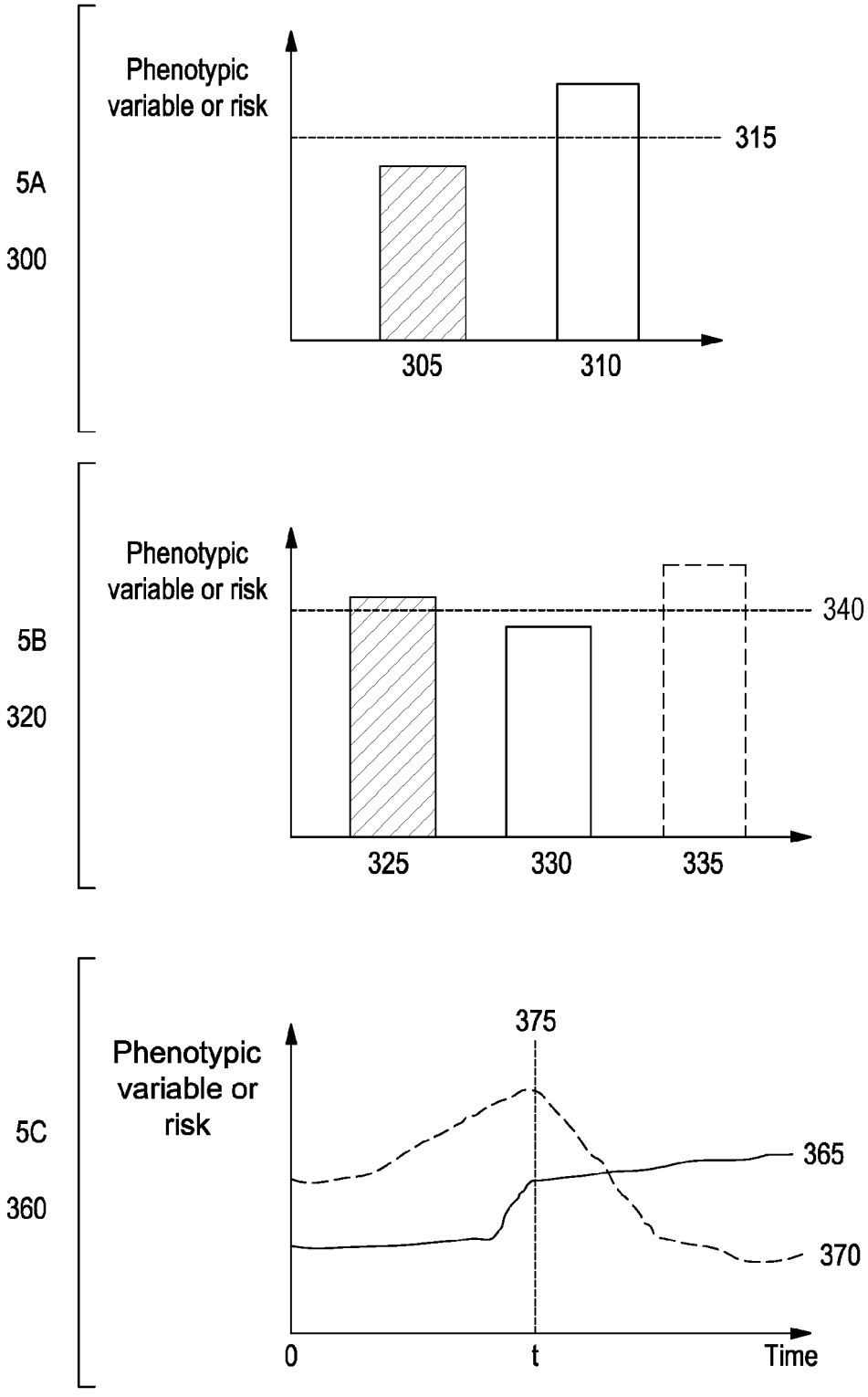
5A
300
Phenotypic variable or risk
315
305
310
5B
320
Phenotypic variable or risk
340
325
330
335
5C
360
Phenotypic variable or risk
375
365
370
0
t
Time

FIGURE 5

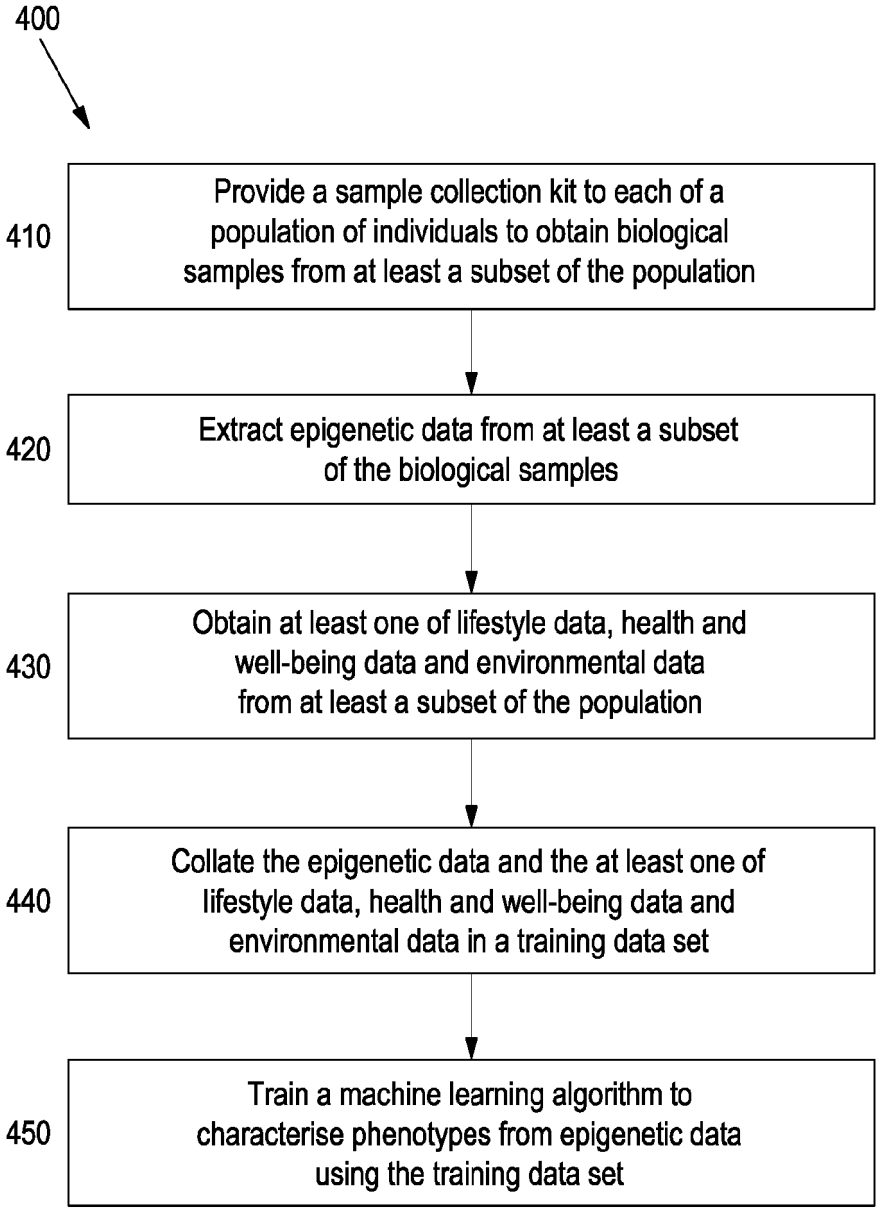
400
410
Provide a sample collection kit to each of a population of individuals to obtain biological samples from at least a subset of the population
420
Extract epigenetic data from at least a subset of the biological samples
430
Obtain at least one of lifestyle data, health and well-being data and environmental data from at least a subset of the population
440
Collate the epigenetic data and the at least one of lifestyle data, health and well-being data and environmental data in a training data set
450
Train a machine learning algorithm to characterise phenotypes from epigenetic data using the training data set

FIGURE 6

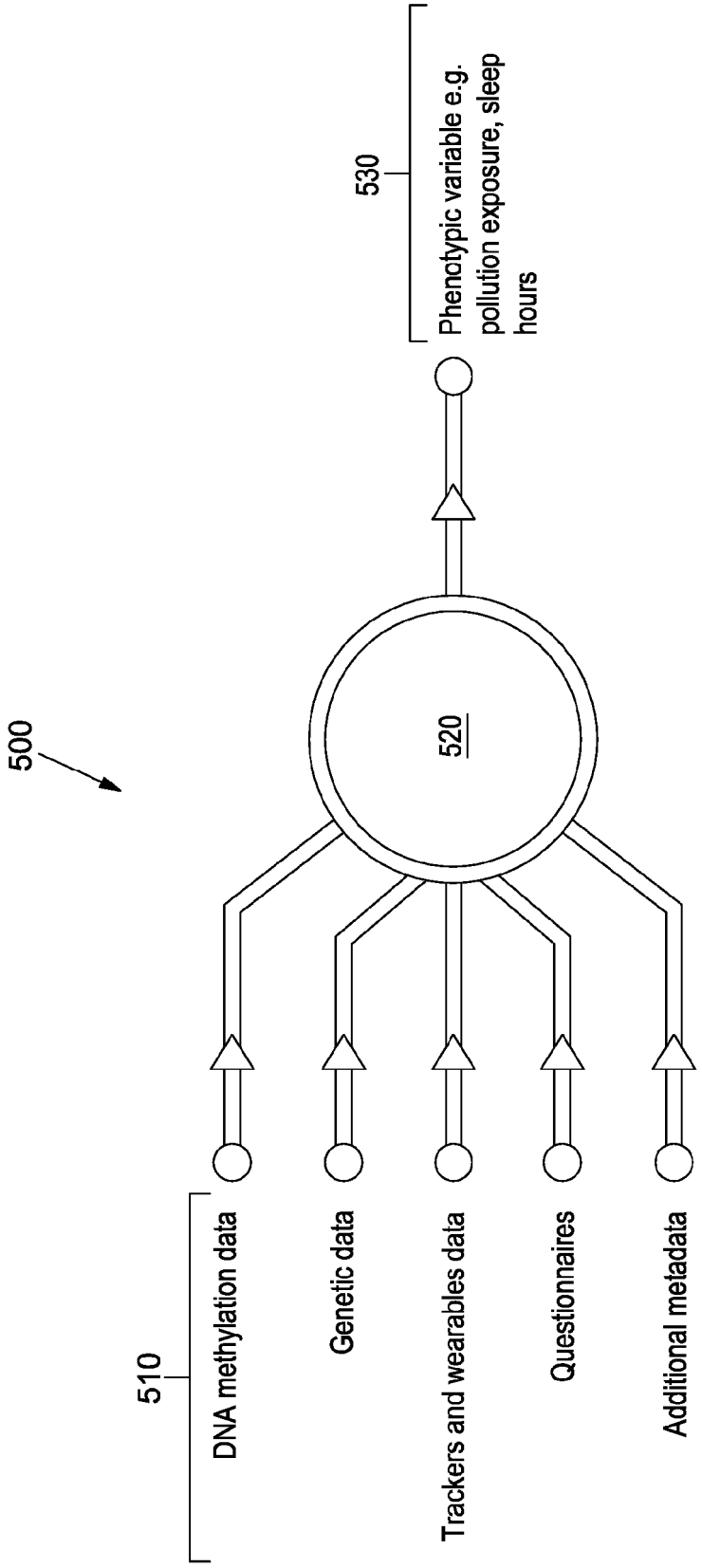
500
520
530
Phenotypic variable e.g. pollution exposure, sleep hours
510
DNA methylation data
Genetic data
Trackers and wearables data
Questionnaires
Additional metadata

FIGURE 7

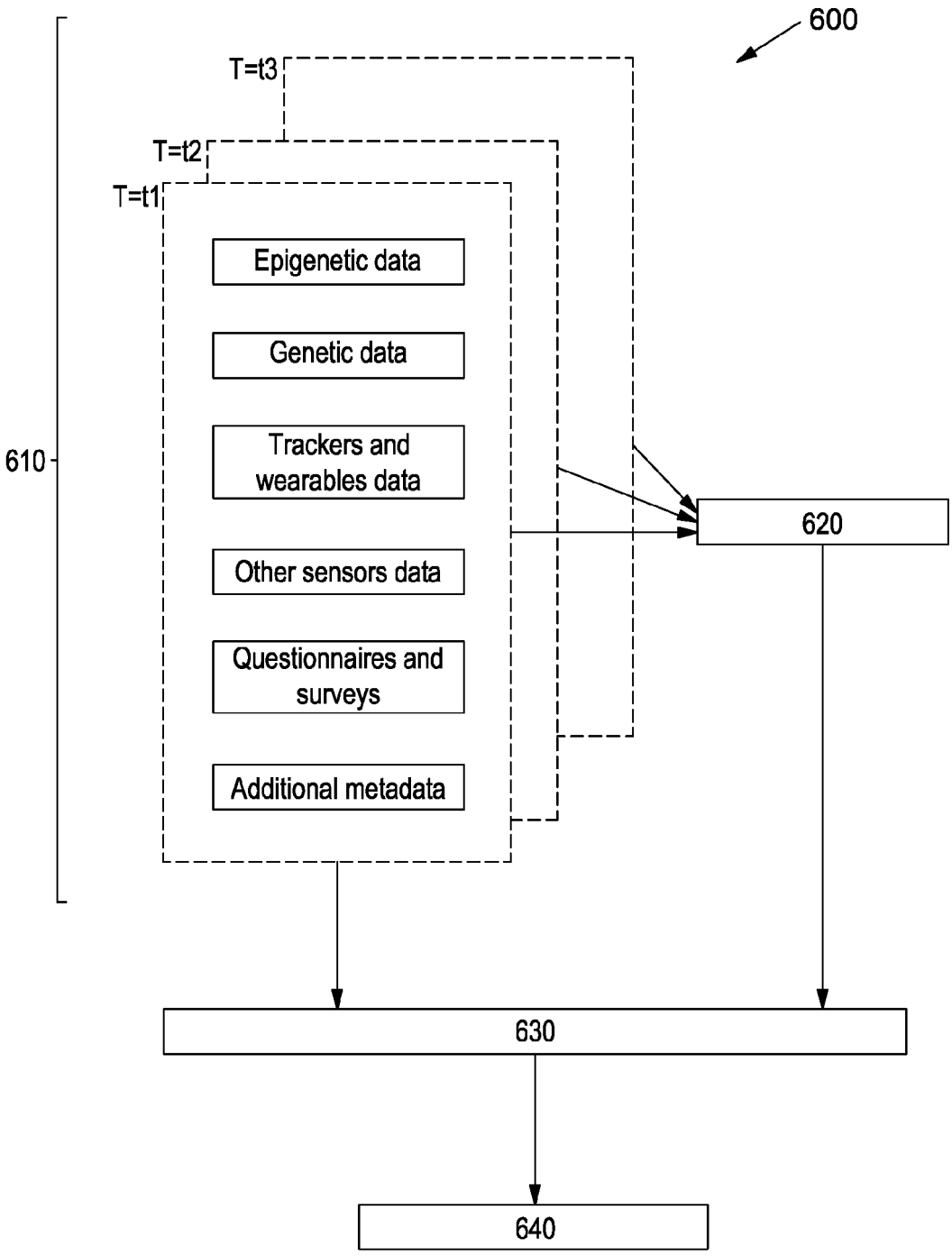
600
T=t3
T=t2
T=t1
Epigenetic data
Genetic data
Trackers and wearables data
Other sensors data
Questionnaires and surveys
Additional metadata
610
620
630
640

FIGURE 8

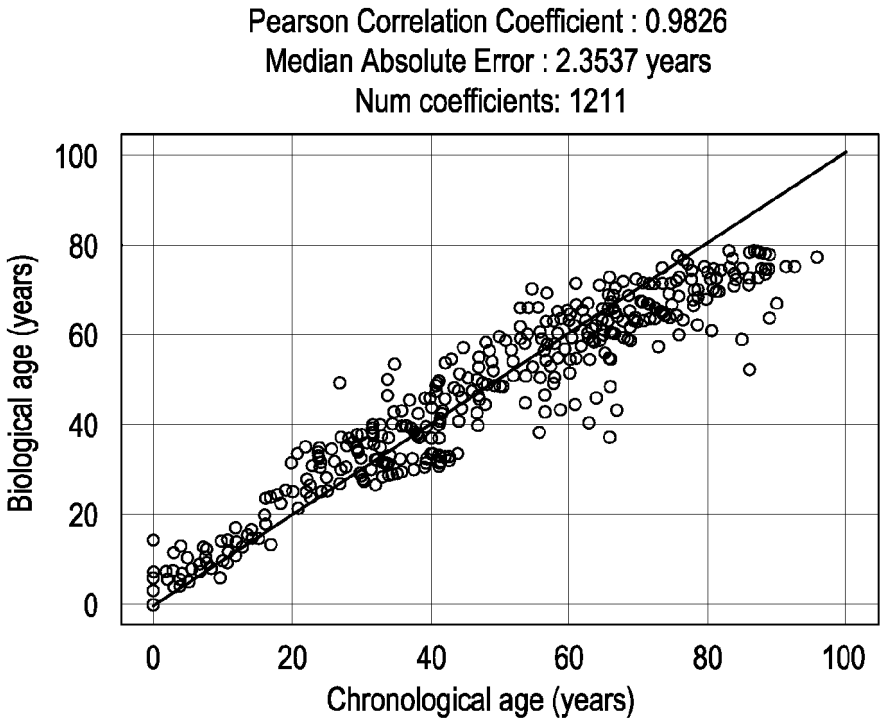

FIGURE 9A

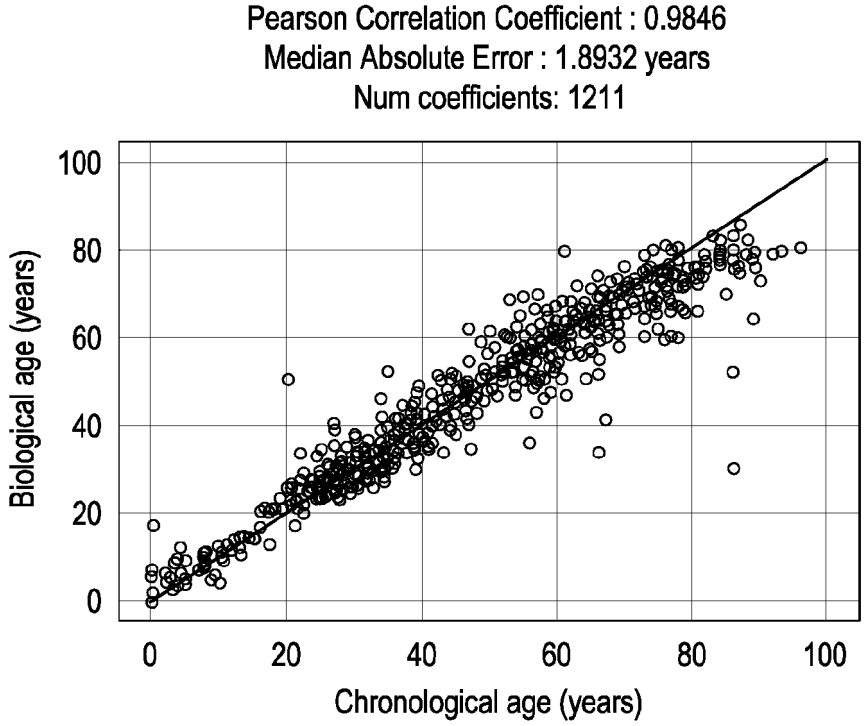

FIGURE 9B

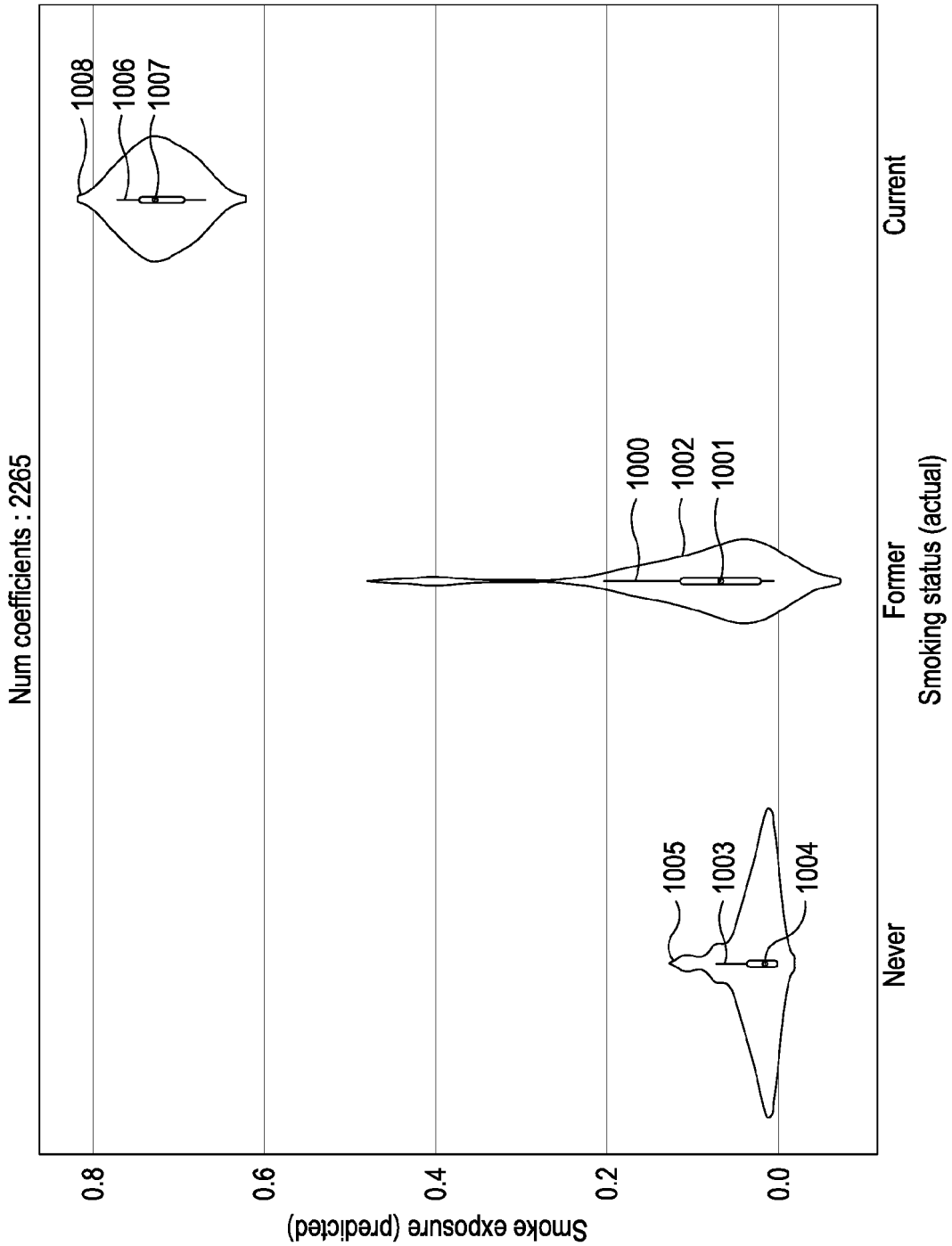
Num coefficients : 2265
Smoke exposure (predicted)
0.0
0.2
0.4
0.6
0.8
Never
Former
Current
Smoking status (actual)
1008
1006
1007
1000
1002
1001
1005
1003
1004

FIGURE 10

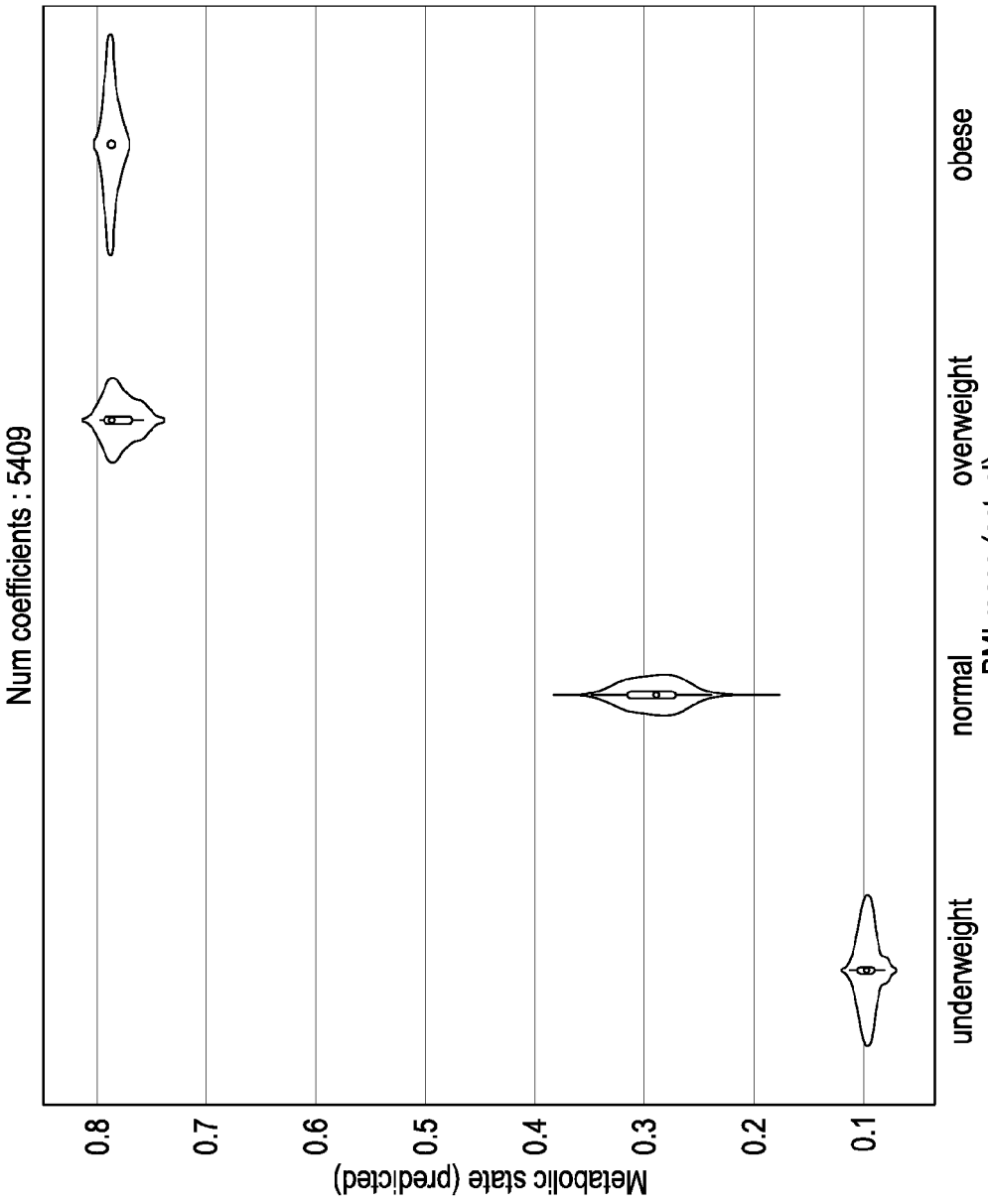
Num coefficients : 5409
Metabolic state (predicted)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
underweight
normal
overweight
obese
BMI range (actual)

FIGURE 11

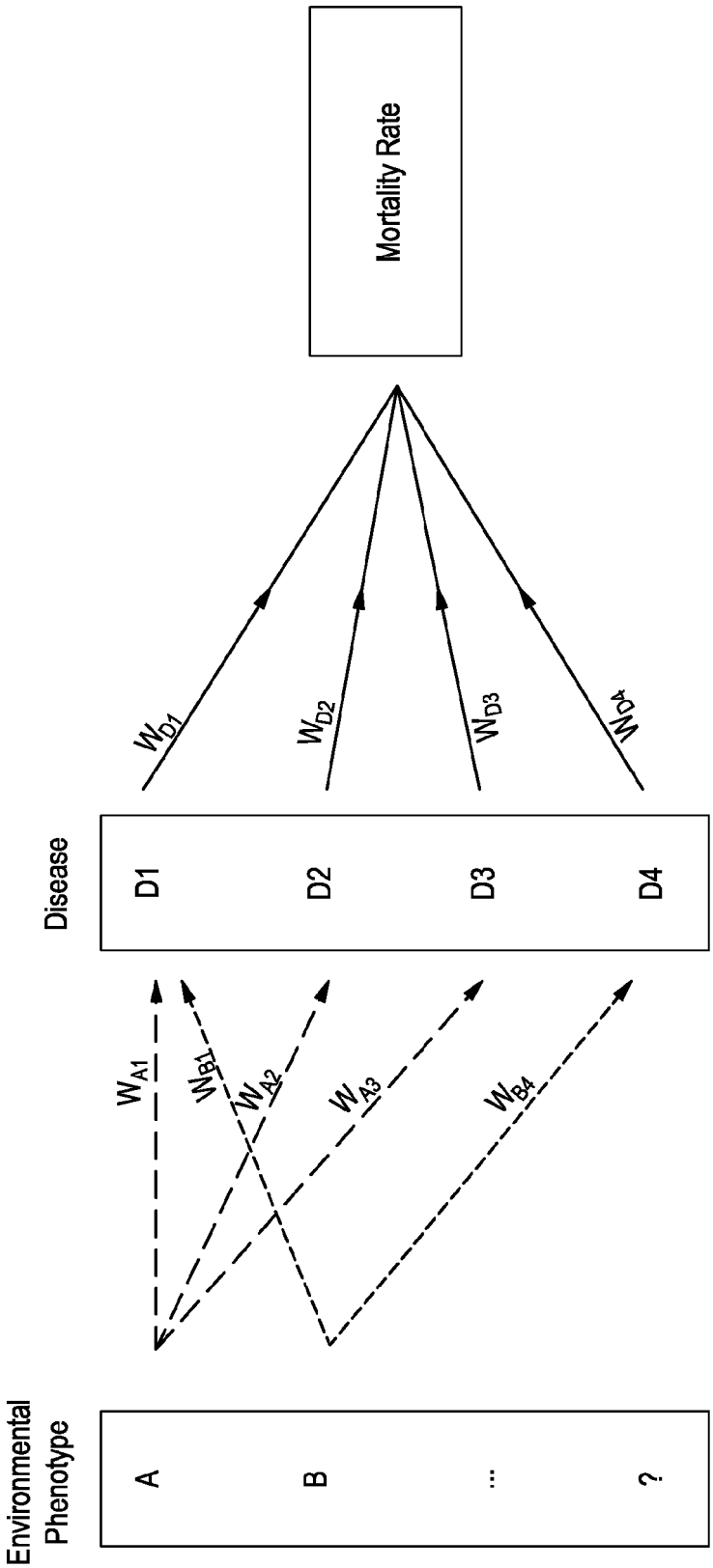
Mortality Rate
W_{D1}
W_{D2}
W_{D3}
W_{D4}
Disease
D1
D2
D3
D4
W_{A1}
W_{B1}
W_{A2}
W_{A3}
W_{B4}
Environmental Phenotype
A
B
...
?

FIGURE 12

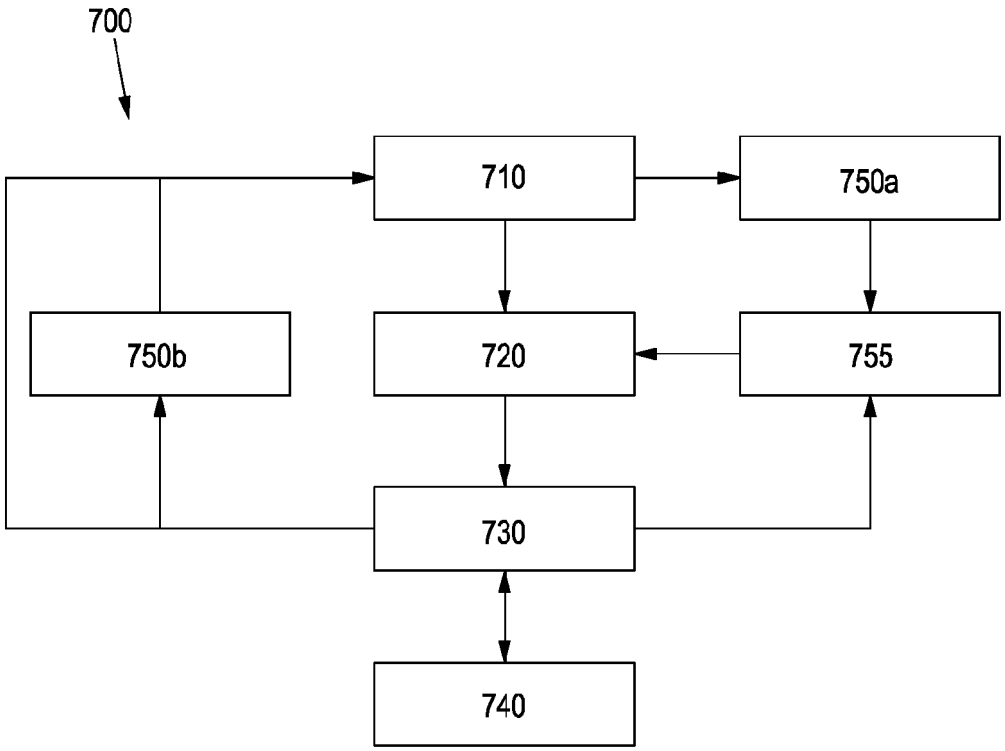

FIGURE 13

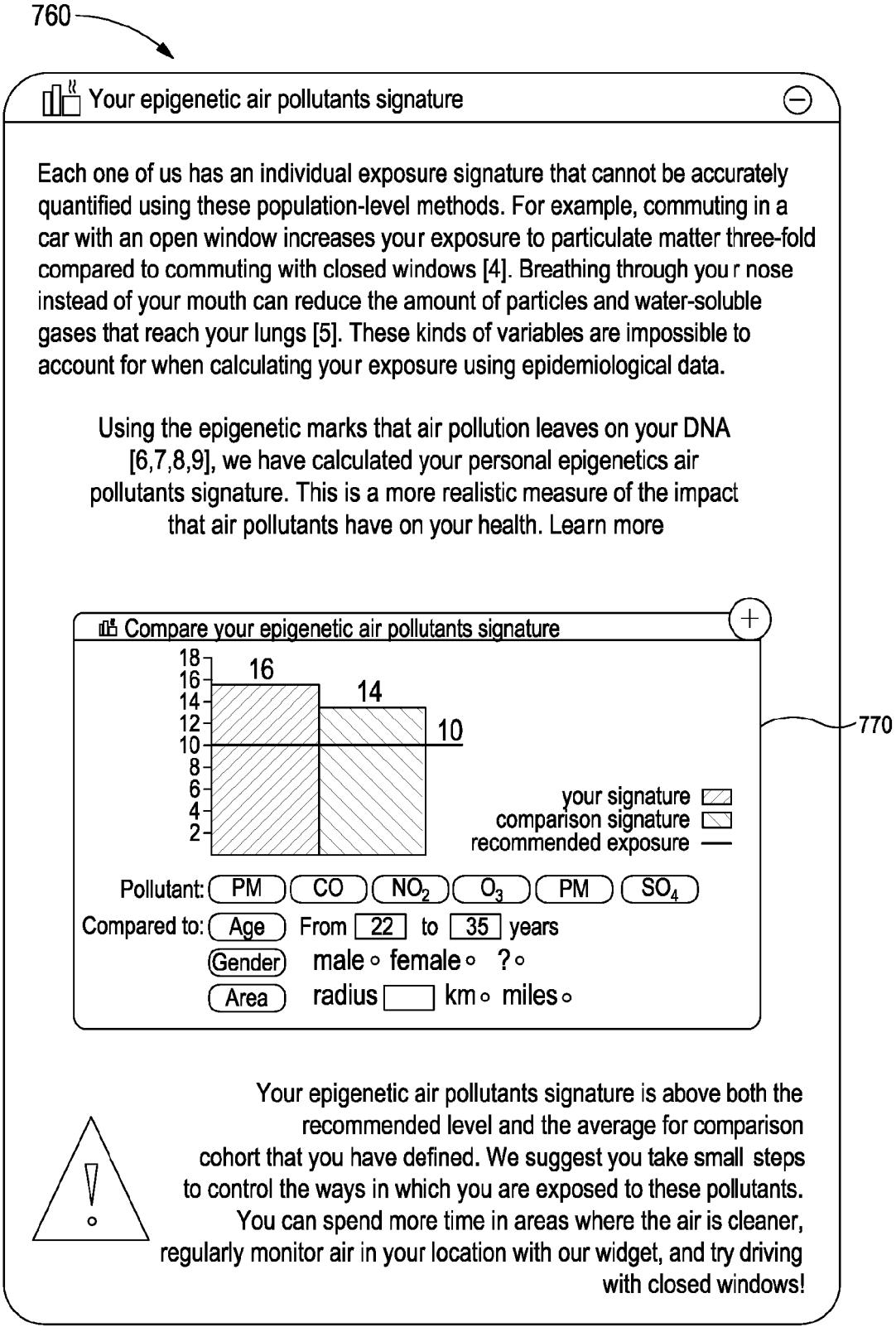
760
Your epigenetic air pollutants signature
Each one of us has an individual exposure signature that cannot be accurately quantified using these population-level methods. For example, commuting in a car with an open window increases your exposure to particulate matter three-fold compared to commuting with closed windows [4]. Breathing through your nose instead of your mouth can reduce the amount of particles and water-soluble gases that reach your lungs [5]. These kinds of variables are impossible to account for when calculating your exposure using epidemiological data.
Using the epigenetic marks that air pollution leaves on your DNA [6,7,8,9], we have calculated your personal epigenetics air pollutants signature. This is a more realistic measure of the impact that air pollutants have on your health. Learn more
Compare your epigenetic air pollutants signature
18
16
14
12
10
8
6
4
2
16
14
10
your signature
comparison signature
recommended exposure
770
Pollutant: PM CO NO_2 O_3 PM SO_4
Compared to: Age From 22 to 35 years
Gender male female ?
Area radius km miles
Your epigenetic air pollutants signature is above both the recommended level and the average for comparison cohort that you have defined. We suggest you take small steps to control the ways in which you are exposed to these pollutants. You can spend more time in areas where the air is cleaner, regularly monitor air in your location with our widget, and try driving with closed windows!

FIGURE 14

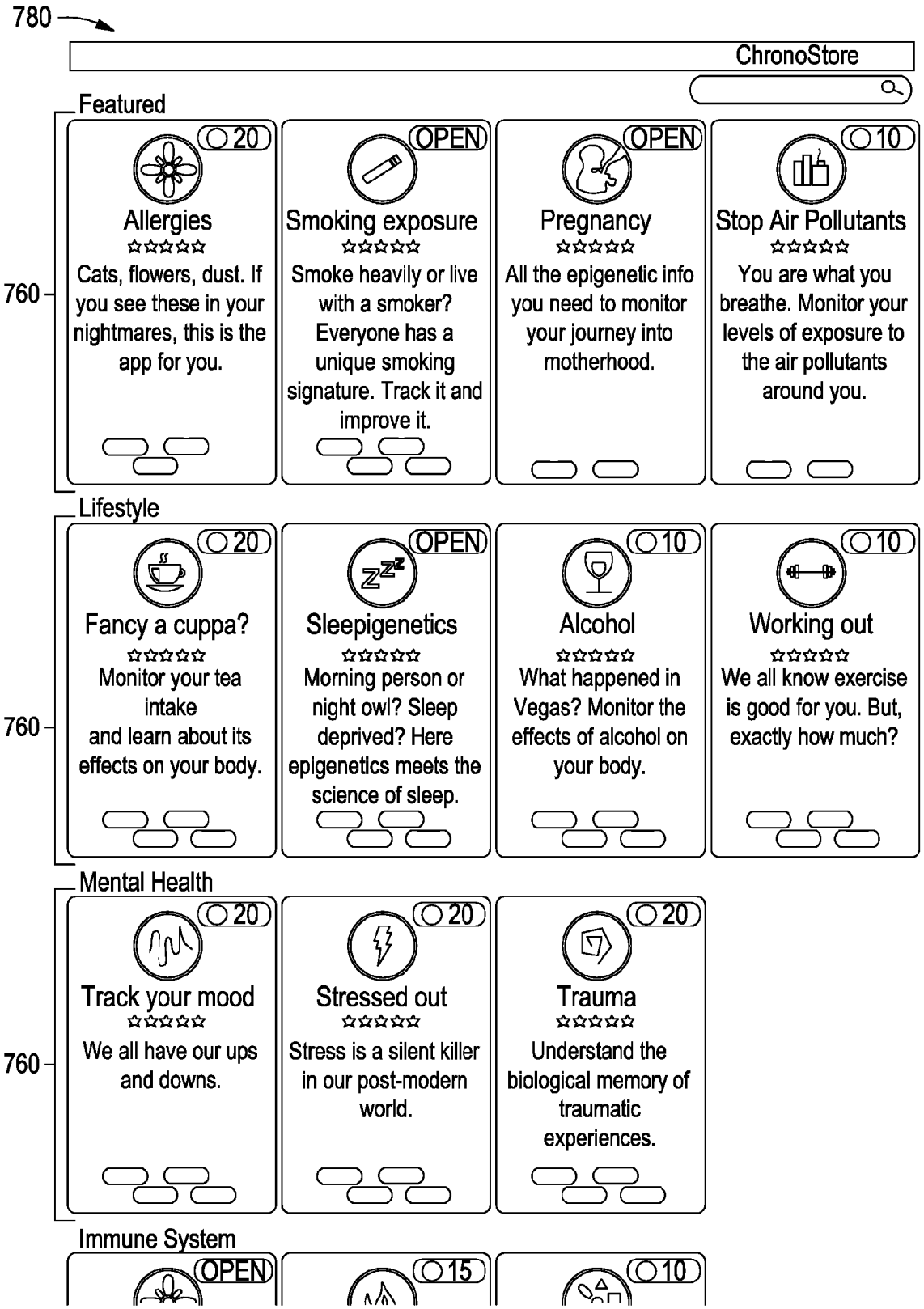
780
ChronoStore
Featured
760
20
Allergies
Cats, flowers, dust. If you see these in your nightmares, this is the app for you.
OPEN
Smoking exposure
Smoke heavily or live with a smoker? Everyone has a unique smoking signature. Track it and improve it.
OPEN
Pregnancy
All the epigenetic info you need to monitor your journey into motherhood.
10
Stop Air Pollutants
You are what you breathe. Monitor your levels of exposure to the air pollutants around you.
Lifestyle
760
20
Fancy a cuppa?
Monitor your tea intake and learn about its effects on your body.
OPEN
Sleepigenetics
Morning person or night owl? Sleep deprived? Here epigenetics meets the science of sleep.
10
Alcohol
What happened in Vegas? Monitor the effects of alcohol on your body.
10
Working out
We all know exercise is good for you. But, exactly how much?
Mental Health
760
20
Track your mood
We all have our ups and downs.
20
Stressed out
Stress is a silent killer in our post-modern world.
20
Trauma
Understand the biological memory of traumatic experiences.
Immune System
OPEN
15
10

FIGURE 15