

SYSTEM, METHOD AND DEVICE FOR PREDICTION AND/OR IDENTIFICATION OF GENETIC AND PHENOTYPIC TRAITS

TECHNICAL FIELD

[0001] The present disclosure relates to a system, method and device for prediction and/or identification of genetic and phenotypic traits.

BACKGROUND

[0002] The following discussion of the background is intended to facilitate an understanding of the present disclosure only. It should be appreciated that the discussion is not an acknowledgement or admission that any of the material referred to was published, known or is part of the common general knowledge of the person skilled in the art in any jurisdiction as of the priority date of the invention.

[0003] Assays have been used as analysis tools for the provision of quicker leads in criminal investigations. Such assays typically have the ability to detect multiple phenotypic traits, including but not limited to race, gender, ABO blood grouping, Rhesus D, HLA-DM, Hair colour, Eye colour, Colour blindness, Baldness, and sample origin (whether human being or otherwise) etc.

[0004] An all-in-one assay has been contemplated to detect up to ten phenotypic traits. A preliminary genetics and bioinformatics study however indicated that such a phenotypic coverage may be too extensive and may not be feasible within one assay due to the total number of probes required, the complexity of multiplexing, and the potential reduction in accuracy as such.

[0005] There exists a need for an improved solution to reduce the complexity of multiplexing without comprising on the accuracy of identification/ detection.

SUMMARY

[0006] Other aspects and features of the present invention will become apparent to those of ordinary skill in the art upon review of the following description of specific embodiments of the invention in conjunction with the accompanying

figures.

[0007] The disclosure seeks to provide a system and method to identify and/or predict one or more genetic and/or phenotypic traits. Compared to prior art, the disclosure seeks to reduce the total number of probes and/or primers required for an assay system, reduce the complexity of multiplexing and improving speed/accuracy of detection. The technical solution adopts the use of artificial intelligence and machine learning, for example support vector machine (SVM), for the prediction a plurality of phenotypical traits, preferably four phenotypic traits. The four phenotypic traits may include a race, a sample origin, a gender and an ABO blood type. The sample origin refers to a determination on whether the sample is obtained from a human or a non-human species.

[0008] An aspect of the disclosure is the development of two assays to support forensic investigations by enabling sample screening to provide rapid and valuable leads. The developed assays have been rigorously tested and shown to be effective over numerous sample types (e.g. blood, saliva, buccal, touch) on numerous matrices (e.g. cloth, denim, cigarette butts, bottles).

[0009] In an aspect of the disclosure, there is a method for developing an assay for identification of a human race from a sample. The method comprises the steps of: (a.) shortlisting a plurality of genetic variations representative of a plurality of human race types, preferably at least four human races types; (b.) configuring, via the use of a microarray, probes and primers associated with each of the plurality of SNPs, the configuration comprises a step of introducing mismatches to the microarray in order to strengthen or reduce binding to specific DNA strands corresponding to the shortlisted SNPs; (c.) detecting whether a plurality of probe signals are present and if present, sending the detected probe signals to a first ML/AI engine; and (d.) distinguishing between probe signals by the first ML/AI engine to provide at least one SNP call; wherein the first ML/AI engine generates at least one SNP combination having a predetermined number of characters.

[0010] The plurality of genetic variations may include single-nucleotide polymorphisms (SNPs).

[0011] In some embodiments, each of the at least one SNP combination comprises ten (10) characters.

[0012] In some embodiments, the method further comprises the step of training a second ML/AI engine for a prediction of human race based on known race group(s) obtained from at least one correspondence database. The training of the second ML/AI engine may include the use of one or more known sample SNP data for calculating one or more model parameters associated with the second ML/AI engine.

[0013] In some embodiments, the calculation of model parameters associated with the second ML/AI engine may be based on a Mahalanobis distance calculation method. The model parameters may include scaler and classifier assigned to respective nodes in the second ML/AI engine to generate a calibration graph.

[0014] After the model parameters have been determined after training, the second ML/AI engine may be used to predict/determine the human race when any new unknown sample (in the form of at least one SNP combination) is received, based on a probabilistic outcome. Preferably, the Mahalanobis distance calculation may comprise the sub-steps of (i.) translating the unknown sample SNP into a first numerical vector; (ii.) translating a plurality of known sample SNPs into a set of second numerical vectors, each of the second numerical vectors associated with a known human race type; (iii.) calculating or deriving a first distance between the first numerical vector and each of the second numerical vectors; (iii.) calculating or deriving a second distance between the first numerical vector and a subset of the set of second numerical vectors; (iv.) deriving a probability value based on the first distance and the second distance, wherein the probability value is indicative of the unknown sample belonging to one of the known human race types.

[0015] In some embodiments, the step of shortlisting includes shortlisting ten SNPs via an analysis of SNP allele frequencies and corresponding ethnicities via the at least one correspondence database. Preferably the SNP allele frequencies associated with each human race were first arranged according to a highest frequency to a lowest frequency based on predetermined threshold. The

predetermined threshold may include a higher allele frequency of 0.9 and a lower allele frequency of 0.1, with SNPs with the highest and lowest frequencies were selected and compared across the different races. Preferably, where two or more human races could not be sufficiently discriminated with certain SNPs based on the thresholds, more SNPs may be shortlisted by relaxing the threshold further to help discriminate the races based on the allele frequency. This way, 10 SNP loci as follows are chosen for the assay. An example of the 10 SNP loci is shown in Table 1.

Table 1.

Gene	Loci
HERC2	rs12913832
SLC24A5	rs1426654
SLC45A2	rs28777
EDAR	rs3827760
IgChr6	rs1015869
OCA2	rs1800414
IgChr2	rs3935223
AKAP7	rs6569733
SLC6A9	rs6429644
MPG	rs3176383

[0016] In accordance with another aspect of the disclosure there is a LOC device or platform for use in an assay. The LOC device comprises a plurality of microarray probes arranged to receive samples and to distinguish between SNPs. The microarray probes are designed such that mismatches are introduced to strengthen binding or reduce binding to the specific DNA strands corresponding to different SNPs. In-silico calculations may be based on the melting temperatures of the primer and probe design. The design is implemented via printing the different probes on the microarray and actual hybridization experiments may be conducted to verify the in-silico calculations.

[0017] In another aspect of the disclosure there is a method for developing an assay for identification of a plurality of phenotypic traits from a sample. The method comprises the steps of analysing if the sample belongs to a human being; if determined to be a human being; (a.) analysing if the human being is a male or a female; (b.) determining the blood type of the human being; wherein the step of determining the blood type of the human being includes performing at least one SNP allele discrimination for ABO gene.

[0018] Preferably, the SNP allele discrimination for ABO gene includes sending the at least one SNP allele to a ML/AI engine for prediction of blood group, and calculating at least one model parameters associated with the SNP allele. The model parameters may be determined through generating microarray signal data using known samples and using a support vector machine algorithm of the ML/AI engine to calculate the model parameters.

[0019] Scaler and classifier are examples of the model's parameters as defined in the support vector machine supervised machine learning model.

[0020] Once the model parameters are derived, the parameters are then embedded in an algorithm to predict/classify the right SNP allele judgement when presented with the microarray signals from an unknown sample and subsequently, a blood type judgement of the sample.

[0021] In some embodiments, the development of primer and probe sequence(s) for identification of ABO blood type is designed based on an introduction of artificial base-substitutions. In some embodiments, the introduction of artificial base-substitutions is followed by an in-silico (performed by a computer processor/chip) calculation and verification.

[0022] Although the invention is applied to forensics, other applications making use of the predicted phenotypic traits can be contemplated.

BRIEF DESCRIPTION OF DRAWINGS

[0023] In the figures, which illustrate, by way of example only, embodiments of the present invention,

[0024] Figure 1 is a flowchart of a method for developing an assay for identification of a human race from a sample.

[0025] Figure 2 is a flowchart of a method for developing an assay for identification of a plurality of phenotypic traits from a sample.

[0026] Figure 3 is a flowchart of a method for the identification of a blood group and a race type of one or more unknown samples, based on the developing of assays in Figure 1 and 2.

[0027] Figure 4 is high-level diagram of a system for the development of an assay for identification of a plurality of phenotypic traits from a sample.

[0028] Figures 5a to 5e illustrate the implementation of a Mahalanobis distance method on a ML/AI engine for predication of a human race.

DESCRIPTION OF EMBODIMENTS

[0029] As used herein, the term “comprise” or variations such as “comprises” or “comprising”, will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers.

[0030] As used herein, the term “include” or variations such as “includes” or “including”, will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers.

[0031] As used herein, the term “may” or “can” will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers.

[0032] As used herein, the term ‘lab-on-chip’ (LOC) device refers broadly to devices that integrates a plurality of laboratory functions on a single integrated

circuit (IC) to achieve automation and diagnostics. An LOC device can combine molecular biology, microfluidics and microelectronics to achieve diagnostics and surveillance. It is contemplated that some LOC device may be used to implement Polymerase chain reaction (PCR) and Hybridization.

[0033] As used herein, the term 'system' can include one or more LOC devices and different components such as, but not limited to, temperature control system, image capturing devices/ processors (e.g. camera, optical reader), touch monitor/laptop, and software to drive the PCR reaction, provide imaging optics and image processing and result interpretation. The term system also refers to hardware, software, and/or combinations of hardware and software to achieve the desired intended functions of the system.

[0034] As used herein, the term 'primers' and 'probes' are broadly understood to be used for initialization or starting a PCR chain reaction. Design or implementation of 'primers' and 'probes' may include one or more microarrays used to detect expression of thousands of genes at the same time. Microarray probes may be designed by an algorithm of introducing mismatches in order to strengthen binding or reduce binding to the specific DNA strands corresponding to different SNPs.

[0035] As used herein, the term 'sample' broadly refers to input fed to a LOC device. Samples can include blood, saliva, buccal, fingerprints, skin found on different matrices such as cloth, denim, cigarette butts, bottles, etc.

[0036] As used herein, the term 'artificial intelligence (AI) engine' refers to a combination of hardware and software to implement one or more AI algorithms. The AI engine can include one or more processors, one or more servers, input modules, intermediate modules, and output modules to implement the AI algorithm. The AI engine may also include machine learning algorithms and can be referred to as an ML/AI engine.

[0037] As used herein, the term 'network' can be any means of providing communication between one or more devices and/or content stored elsewhere. As used herein, network can be a personal area network, local area network, a storage area network, a system area network, a wide area network, a virtual private

network, and an enterprise private network. The network can include one or more gateways or no gateways. The network communication can be conducted via published standard protocols or proprietary protocols.

[0038] As used herein, communication of data through any network can be: (i) encoded or unencoded; (ii) encrypted or unencrypted; (iii) delivered via a wired network, a wireless network, or a combination of wired and wireless networks. Wireless communication can be accomplished in any practical manner including a Wi-Fi 802.11 network, a Bluetooth™ network, or mobile phone network (such as 3G, 4G, LTE, and 5G). The terms “connect”, “connected”, and “connecting” as used herein refer to a communication link between at least two devices and can be accomplished as discussed in this paragraph.

[0039] As used herein, the term “associate”, “associated”, and “associating” indicate a defined relationship (or cross-reference) between at least two items. Such defined relationship may be expressed in one or more mathematical equations and/or formulas.

[0040] As used herein, the term “computing device” may be a single stand-alone computer such as a desktop computer or a laptop computer, a thin client, a tablet computer, an embedded computer or a mobile phone. The computing device may run a local operating system and store computer files on a local storage drive. The computing device may access files and application through a gateway to one or more content repositories, the content repository can host files and/or run virtual applications and generate a virtual desktop for the computing device.

[0041] As used herein, the term ‘server’ or ‘processor’ may include a single stand-alone computer, a single dedicated server, multiple dedicated servers, and/or a virtual server running on a larger network of servers and/or cloud-based service. A server may also include virtual machines, and can include one or more computing services hosted on cloud. Cloud could be either private or public commercial cloud. A server of processor may execute instructions, codes,

computer programs, scripts which it accesses from at least one of a hard disk, a floppy disk, an optical disk, a flash drive, a ROM, a RAM.

[0042] According with an aspect of the disclosure and with reference to Figure. 1 there is a method 100 for identification of a human race from a sample comprising the steps of: (a.) shortlisting a plurality of genetic variations (e.g. single-nucleotide polymorphisms- SNPs) representative of a plurality of races types (step S102); (b.) configuring, via the use of a microarray, probes and primers associated with each of the plurality of SNPs, the configuration comprises a step of introducing mismatches to the microarray in order to strengthen or reduce binding to specific DNA strands corresponding to the shortlisted SNPs (step S104); (c.) detecting whether a plurality of probe signals are present and if present, sending the detected probe signals to a first ML/AI engine (step S106); and (d.) distinguishing between probe signals by the first ML/AI engine to provide at least one SNP call; wherein the first ML/AI engine generates at least one SNP combination having a predetermined number of characters (step S108).

[0043] The step S102 of shortlisting the plurality of genetic variations may be made with reference to a database, such as, but not limited to, a genome database. An example of such genome database is a DNA database correlated to different phenotypic traits. An example of the DNA database is the 1000 genome database derived from a global or country project.

[0044] In step S104, the step of introducing one or more mismatches may be determined through an in-silico process. This process may comprise one or more experiments and calculation steps.

[0045] In some embodiments, the calculation step comprises calculating a binding affinity parameter of a probe sequence associated with the test sample sequences. The calculation step may be preceded by a step of making one or more random DNA sequence modifications in-silico to locate an optimal probe DNA sequence that shows the potential to differentiate the SNPs present in test samples. The optimal probe DNA is then synthesized and printed on the microarray. The mismatch between the probe versus the test sample will affect the

signal intensity that leads to SNP determination through the data analysis algorithm.

[0046] It is appreciable that the binding affinity may be defined by the melting temperature of the two DNA strands (probe and test sample) calculated via a commercial package.

[0047] In some embodiments, the first ML/AI engine may be implemented as a support vector machine (SVM). The support vector machine may then be used to derive a set of model parameters comprising a scalar and/or a classifier. Once the set of model parameters is derived, a calibration graph based on the model parameters is generated.

[0048] In some embodiments, the at least one SNP combination comprises ten (10) characters.

[0049] In some embodiments, the method 100 further comprises the step of training a second ML/AI engine for a prediction of human race based on known race group(s) obtained from at least one correspondence database (step S110). The training of the second ML/AI engine may include the use of one or more known sample SNP data. The prediction of human race may include a step of calculating model parameters. The calculation of model parameters may be based on a Mahalanobis distance calculation method. The model parameters may include scaler and classifier assigned to respective nodes in the second ML/AI engine as part of a training process.

[0050] After the model parameters have been determined after training, the second ML/AI engine may be used to predict the human race when any new samples are received. Preferably, the Mahalanobis distance calculation may be based on the sub-steps of (i.) translating the known sample SNP data into a numerical vector; (ii.) calculating or deriving a first distance between the new sample and known race groups (e.g. 29 race groups); (iii.) calculating or deriving a second distance between the new sample and a subset of the known race groups (e.g. 25 race groups); (iv.) deriving a probability value based on the first distance and the second distance.

[0051] In some embodiments, the step of shortlisting a plurality of genetic variations includes shortlisting ten (10) SNPs via analyzing SNP allele frequencies and corresponding ethnicities via the correspondence database. The SNP allele frequencies for each race were first arranged according to a highest frequency to a lowest frequency based on predetermined threshold. The predetermined threshold may include a higher allele frequency of 0.9 and a lower allele frequency of 0.1. The SNPs with the highest and lowest frequencies may be selected and compared across the different races.

[0052] It is contemplated that other genetic variations other than SNPs may be used. Non-limiting examples include mRNA expression levels, its splicing patterns, base insertions and/or deletions in larger scale than SNPs, and epigenetic alterations such as methylation status of DNA base in the genome.

[0053] According with another aspect there is a method 200 for developing an assay for identification of a plurality of phenotypic traits from a sample comprising the steps of analysing if the sample belongs to a human being (step S202); and if determined to be a human being;

(a.) analysing if the human being is a male or a female (step S204); and

(b.) determining the blood type of the human being; wherein the step of determining the blood type of the human being includes performing at least one SNP allele discrimination for an ABO gene (step S206).

[0054] To determine the blood type, a trained ML/AI engine may be used. The ML/AI engine may be trained via a step of calculating at least one model parameter, which include a step of generating microarray signal data using known samples and using a support vector machine algorithm of the ML/AI engine to calculate the model parameters (step S208).

[0055] The SNP allele discrimination for the ABO gene may include the step of sending the at least one SNP allele to a ML/AI engine for prediction of blood group (step S210).

[0056] The model parameters may also be derived or trained via a reference database of known blood groups. The model parameters may comprise one or more scalers and classifiers assigned to respective nodes in the support vector machine as part of a training process.

[0057] After the model parameters are derived, the model parameters may be input to an algorithm to predict/classify the right SNP allele judgement when presented with the microarray signals from an unknown sample and subsequently, a blood type judgement/determination of the sample.

[0058] In some embodiments, the primer and probe sequence(s) for identification of ABO blood type is developed based on an introduction of one or more artificial base-substitutions. The introduction of artificial base-substitutions may be followed by an in-silico (performed by a computer processor/chip) calculation and verification.

[0059] According with another aspect of the disclosure and with reference to Figure. 3 there is a computer-implemented method 300 of analyzing a biological sample to identify/predict or determine a plurality of phenotypic traits, and in particular a race type and a blood group of the sample. It is appreciable that the method 300 may include combinations of steps from method 100 and method 200. The computer-implemented method 300 comprises the steps of: reading a probe intensity of the sample (step S302); inputting the probe intensity readings into an input module (step S304); determining, using a first ML/AI engine, at least one SNP call, wherein the first ML/AI engine generates at least one SNP combination having a predetermined number of characters (step S306). The blood group is then determined by the SNP combination (step S308) and output (step S310).

[0060] The output from step S306, i.e. the at least one SNP combination is fed into a trained second ML/AI engine, which is in turn used to determine a race group based on a Mahalanobis distance model (step S312). The output of the second ML/AI engine may be an output in the form of a probability value, for example, 0.8 or 80% probability that it is a certain race group/type (step S314).

[0061] The first ML/AI engine may be trained using probe intensity signals from chips containing samples of known blood/race types. The known correlation between probe intensity signals and blood group/race types may be expressed as data entries in a reference database (step S316) as known to a skilled person. The first ML/AI engine will utilize the data entries in the reference database for training. In a model where the first ML/AI engine is a support vector machine (SVM) implemented in the form of a supervised training model, the outcome of the training will be the allocation of model parameters to generate a calibration graph (also referred to as a model). (step S318).

[0062] Similarly, the second ML/AI engine may be trained with reference to a database, such as a 1000 genome database (step S320). The second ML/AI engine will utilize the data entries 1000 genome database for training. The outcome of the training will be a derived set of model parameters which is in turned used to generate a calibration graph. (step S324). Training of the second ML/AI engine is based on the Mahalanobis distance model.

[0063] According with an aspect of the disclosure and with reference to Figure 4 there is a system 400 for identification of at least one phenotypic trait comprising a LOC device 402 and a first AI engine 404 arranged to receive data from the LOC device 402.

[0064] The system 400 may be a system suitably configured and arranged to facilitate investigative work. The investigative work may be in the form of forensics investigative work so as to analyze one or more biological samples to identify one or more phenotypic traits of a suspect from a crime scene.

[0065] The LOC device 402 may combine Micro-Electro-Mechanical-Systems (MEMS) with micro-fluidics to automate and integrate multiplexed DNA amplification with microarray detection. The LOC device 402 may be a known device such as the VereplexTM Biosystem. The LOC device 402 may include a microarray 424 and a PCR chamber 426. The device 402 combines Micro-Electro-Mechanical-Systems (MEMS) with micro-fluidics to automate and integrate multiplexed DNA amplification with microarray detection. The LOC device 402

typically does not include a separate microreactor. In other words, the LOC device 402 may be integrated with a microreactor.

[0066] The AI engine 404 may include an input module 442 arranged to receive data from the LOC device 402. The AI engine 404 may be installed on any processor. In some embodiments, the AI engine 404 may form part of a specialized system for identification and diagnosis. In some embodiments, the AI engine 404 may be installed on a specialized LOC system for performing two or more functions, such as PCR and hybridization.

[0067] The ML/AI engine 404 may be installed with an AI algorithm, such as, but not limited to, a support vector machine (SVM), a neural network algorithm, a genetic algorithm, evolutionary algorithm, expert rule base, and/or any supervised learning algorithm. In some embodiments the ML/AI engine 404 is an SVM. The SVM is arranged to perform machine learning for identification and/or prediction of at least one phenotypic trait. The SVM is trained based on a supervised learning model, i.e. a predetermined number of data points from known correlation between each of the phenotypic trait and SNP combinations is used to train the SVM. In some embodiments, the at least one phenotypic trait includes the race of a suspect from the crime scene.

[0068] In some embodiments, the AI engine 404 is used to distinguish between the different microarray probe signals so as to correlate the single-nucleotide polymorphism (SNP) with the associated phenotypic trait(s).

[0069] For some phenotypic traits, for example race type, the aforementioned correlation may not be established based on a single machine learning/AI algorithm. In such a case the AI engine 404 may be regarded as a first AI engine 404, which may be used to distinguish between probe signals to provide at least one SNP call; wherein the first AI engine 404 generates at least one SNP combination having a predetermined number of characters, and then sending the generated at least one SNP combination to a second ML/AI engine 406 for a prediction of race.

[0070] It is contemplated that the prediction of race by the second ML/AI engine

406 may include a step of calculating model parameters. The calculation of model parameters may be based on a statistical and/or probabilistic calculation model, such as, but not limited to, the Mahalanobis distance calculation method. It is contemplated that other suitable ML/AI algorithms may be used for the second ML/AI engine 406.

[0071] The system 400 may be used for the development of the assay for human race type identification, ten (10) SNPs may be shortlisted through an analysis of SNP allele frequencies and corresponding ethnicities (race type) which can be found in two databases. These databases are the Singapore Genome Variation Project (SGVP) and the 1000 Genomes Project which are publicly available.

[0072] The SNP allele frequencies for each race were first arranged according to the highest frequency to the lowest frequency. Using a threshold of 0.9 and 0.1 allele frequency, SNPs with the highest and lowest frequencies were selected and compared across the different races. If two races could not be sufficiently discriminated with certain SNPs based on the threshold, more SNPs would be shortlisted by relaxing the threshold further to help discriminate the races based on the allele frequency. This way, 10 SNP loci were chosen for the assay.

[0073] As part of the development of an assay for race type identification, primers and probes will have to be configured/designed. The primer design for the assay followed standard molecular biology design principles. However, the microarray probes to distinguish between SNPs are derived based on an algorithm of introducing mismatches in order to strengthen binding or reduce binding to the specific DNA strands corresponding to different SNPs. In this way, different probes were then printed on the microarray and actual hybridization experiments were conducted to further verify the in-silico calculations and the process is iterated till the probes are able to distinguish SNPs.

[0074] It is observed that the DNA base mismatches for differentiating between SNPs did not completely eliminate microarray probe signals in non-specific binding. In other words, the actual probe signals may not be clear cut. For example, the

probe that is designed to bind the 'A' allele, may still give a signal when the 'G' allele is present in the sample. The first ML/AI engine 404 is used to distinguish between probe signals to provide the correct SNP call. For each SNP used in the assay, different samples homozygous for 'A' or 'G' alleles, and heterozygous alleles ('R') were run and the corresponding probe signals were used to train the machine learning algorithm. The machine learning algorithm, support vector machine, uses available training data to calculate a model containing parameters which will be used to then predict the SNP outcomes when new data is presented to it. The microarray analysis algorithm will give a ten-character SNP combination that is then presented to the race call interpreter.

[0075] The ten-character SNP combinations of samples contained in the two databases, SGVP and 1000 Genomes Project were tabulated with the corresponding race, and used to generate a model for predicting the race based on a given ten-character SNP combination. This required another layer of machine learning where model parameters were calculated using Mahalanobis distance calculation. The model parameters form the basis of the second layer of software algorithm which translates the ten-character SNP combination into a race prediction.

[0076] The second layer of software algorithm may be implemented as part of the second ML/AI engine 406.

[0077] The system 400 may also be used for the development of the assay for identification of various phenotypic traits. In some embodiments, the system 400 may be used to implement a method for developing an assay for identification of a plurality of phenotypic traits from a sample comprising the steps of analysing if the sample belongs to a human being; if determined to be a human being; (a.) analysing if the human being is a male or a female; (b.) determining the blood type of the human being; wherein the step of determining the blood type of the human being includes performing at least one SNP allele discrimination for ABO gene.

Non-limiting example of a Mahalanobis distance calculation

[0078] Reference to the Mahalanobis distance calculation method as mentioned

in the description is elaborated as follows.

[0079] As an introductory note, a distinction is made between the Mahalanobis distance and the Euclidean distance, which is defined as a length of line between two points in a space. The Mahalanobis distance provides a spatial distance between a point and a group of points. Based on Mahalanobis distance calculation, a distance between, (1) a SNP data of a sample obtained from a LOC device that is yet to be analyzed and, (2) a group of SNP base data from different race groups obtained from a reference database, such as a public 1000 genome database and Singapore Genome Variation Project, SGVP, can be calculated. The shorter the distance is, the higher the likelihood of the sample belongs to the race group.

[0080] An example of the steps of deriving a probability value associated with a phenotypic trait based on the Mahalanobis distance calculation method is as follows.

[0081] Translating pre-analysed SNP data (which may be obtained from a LOC) into a numerical vector. This may be based on a 10-character/alphabet SNP data information, which include bases reading, where the bases are concatenated into one string, "AGMGGCARTA".

[0082] Each alphabet is translated into a 3-dimensional numerical vector in accordance with a pre-defined translation table. For example, A is translated to (1, 0, 0), M is translated to (0, 1, 0) etc. The 10-character string may therefore be translated into 10 of 3-dimensional ($3 \times 10 = 30$, binary [0 or 1], in this case) vector. For example, sample_input_vector = [1, 0, 0, 0, 0, 1, 0, 1, 0, ..., 1, 0, 0].

[0083] Applying the same translation rule, SNP data from the 1000 genome database and SGVP are translated into numerical vectors and stored in a database. In total, data from the reference database, which includes 2771 individuals over 29 race groups (across 5 super populations) were processed.

[0084] The Mahalanobis distance between the translated sample and each of 29 race groups are calculated and stored into a first distance vector. The resulting vector is a 29-dimensional vector. Four (4) of the 29 race groups are against a

particular population groups, which may not be in the current assay scope, but may be included to derive better resolution between 2 or more other race groups). An illustration is shown in Figure 5a and 5b.

[0085] Using the first distance vector data *d*, a second distance vector *d2* is derived for a subset of the 29 race groups, for example, 25 race groups. The subset excludes the aforementioned 4 of the 29 race groups. The second distance vector results in 25-dimensional vector. An illustration is shown in Figure 5c and 5d.

[0086] Using the second distance data in the *d2_sample*, a final probability value *PCHB* for each of 25 groups is calculated with the formula as shown in Figure 5e. The obtained probability data are then sorted in descending order and reported with super population group name.

[0087] It is appreciable to a skilled person that any resultant assay comes with chemical reagents which follow simple and common laboratory techniques and are combined in a reaction mix for loading onto a LOC device.

[0088] The analysis process may be based on the following steps: (i.) pre-amplification; (ii.) PCR preparation and chip loading; (iii.) PCR analysis; (iv.) hybridization; (v.) wash; and (vi.) image capture & results. It is contemplated that the total time taken to complete the six steps above will be less than 3 hours. This is an improvement over prior known techniques.

[0089] In some embodiments, the output of the system 400 will be reflected as a report for operators to follow up on the protocol they have. The resulting data analysis may require an in-house developed software algorithm to provide a straightforward call/result.

[0090] It is appreciable that through the use of a LOC device, potential areas for contamination are minimized as the operations are unified.

[0091] It is appreciable that as with any DNA-based technique, enough starting DNA template is necessary.

[0092] It is appreciable that parameters of the assay, as with any assay, are to be well defined, in order to obtain optimum results. This includes parameters such as: input quantity, temperature, duration, and number of PCR cycles, temperature and duration of microarray analysis.

[0093] It should be further appreciated by the person skilled in the art that variations and combinations of features described above, not being alternatives or substitutes, may be combined to form yet further embodiments falling within the intended scope of the invention.

CLAIMS

1. A method for developing an assay for identification of a human race from a sample comprising the steps of:

(a.) shortlisting a plurality of genetic variations representative of at least four races types;

(b.) configuring, via the use of a microarray, probes and primers associated with each of the plurality of SNPs, the configuration comprises a step of introducing mismatches to the microarray in order to strengthen or reduce binding to specific DNA strands corresponding to the shortlisted SNPs;

(c.) detecting whether a plurality of probe signals are present and if present, sending the detected probe signals to a first ML/AI engine; and

(d.) distinguishing between probe signals by the first ML/AI engine to provide at least one SNP call; wherein the first ML/AI engine generates at least one SNP combination having a predetermined number of characters.

2. The method of claim 1, wherein the plurality of genetic variations includes single-nucleotide polymorphisms (SNP), and wherein each of the at least one SNP combination comprises ten (10) characters.

3. The method of claim 1 or 2, the method further comprises the step of sending the at least one SNP combination to a second ML/AI engine for a prediction of human race.

4. The method of claim 3, wherein the prediction of human race may include a step of training the second ML/AI engine, the step of training the second ML/AI engine includes a step of calculating model parameters.

5. The method of claim 4, wherein the calculation of model parameters is based on a Mahalanobis distance calculation method.

6. The method of any one of the preceding claims, wherein the step of shortlisting

a plurality of genetic variations includes shortlisting ten 10 SNPs via analyzing SNP allele frequencies and corresponding ethnicities via at least one correspondence database.

7. The method of claim 6, wherein the SNP allele frequencies for each human race were first arranged according to a highest frequency to a lowest frequency based on predetermined threshold.

8. The method of claim 7, wherein the predetermined threshold may include a higher allele frequency of 0.9 and a lower allele frequency of 0.1.

9. The method of claim 8, wherein the SNPs with the highest and lowest frequencies were selected and compared across the different human races.

10. A method for developing an assay for identification of a plurality of phenotypic traits from a sample comprising the steps of analysing if the sample belongs to a human being; if determined to be a human being;

(a.) analysing if the human being is a male or a female;

(b.) determining the blood type of the human being; wherein the step of determining the blood type of the human being includes performing at least one SNP allele discrimination for ABO gene.

11. The method of claim 10, wherein the SNP allele discrimination for ABO gene includes sending the at least one SNP allele to a ML/AI engine for prediction of blood group.

12. The method of claim 10, further comprises a step of training the ML/AI engine, wherein the step of training includes a step of generating microarray signal data using known samples and using a support vector machine algorithm of the ML/AI engine to calculate the model parameters.

13. The method of claim 12, wherein after the model parameters are derived, the model parameters are input to an algorithm to predict/classify the right SNP allele judgement when presented with the microarray signals from an unknown sample

and subsequently, a blood type judgement of the sample.

14. The method of claim 13, wherein the development of primer and probe sequence(s) for identification of ABO blood type is designed based on an introduction of artificial base-substitutions.

15. The method of claim 14, wherein the introduction of artificial base-substitutions is followed by an in-silico (performed by a computer processor/chip) calculation and verification.

16. A system for developing an assay for identification of a human race from a sample comprising:

a LOC device configured to receive the sample, the LOC device comprises at least one microarray;

a processor configured to: (a.) shortlist a plurality of genetic variations representative of a plurality of races types; (b.) configure, via the use of the microarray, probes and primers associated with each of the plurality of genetic variations, the configuration comprises a step of introducing mismatches to the microarray in order to strengthen or reduce binding to specific DNA strands corresponding to the shortlisted SNPs; (c.) detect whether a plurality of probe signals are present; and

a first ML/AI engine to receive the plurality of probe signals and if present, send the detected probe signals to a first ML/AI engine to distinguish between probe signals to provide at least one SNP call; wherein the first ML/AI engine generates at least one SNP combination having a predetermined number of characters.

17. The system of claim 16, further comprises a second ML/AI engine to receive the at least one SNP combination to determine or predict a human race type associated with the at least one SNP combination, the second ML/AI engine implements a Mahalanobis distance calculation method.

18. A system for developing an assay for identification of a plurality of phenotypic

traits from a sample comprising:

a LOC device configured to receive the sample, the LOC device comprises at least one microarray;

a processor configured to analysing if the sample belongs to a human being, and if determined to be a human being, the processor is configured to (a.) analyse if the human being is a male or a female; (b.) determine the blood type of the human being; wherein the step of determining the blood type of the human being includes performing at least one SNP allele discrimination for ABO gene.

19. A non-transitory computer readable medium having stored thereon software instructions that, when executed by a processor, cause the processor to execute the steps of claims 1 to 9, and/or claims 10 to 15.