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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 November 2015 (19.11.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/173435 A1

(51) International Patent Classification:
G06F 19/18 (2011.01)

(21) International Application Number:
PCT/EP2015/060921

(22) International Filing Date:
18 May 2015 (18.05.2015)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1408687.0 16 May 2014 (16.05.2014) GB

(71) Applicant: **KATHOLIEKE UNIVERSITEIT LEUVEN, KU LEUVEN R&D** [BE/BE]; Waaistraat 6 box 5105, B-3000 Leuven (BE).

(72) Inventor: **CLAES, Peter**; Moederhoefstraat 42 bus 2.2, B-2547 Lint (BE).

(74) Agent: **LC PATENTS**; Crutzenstraat 26, 3500 Hasselt (BE).

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

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

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

[Continued on next page]

(54) Title: METHOD FOR PREDICTING A PHENOTYPE FROM A GENOTYPE

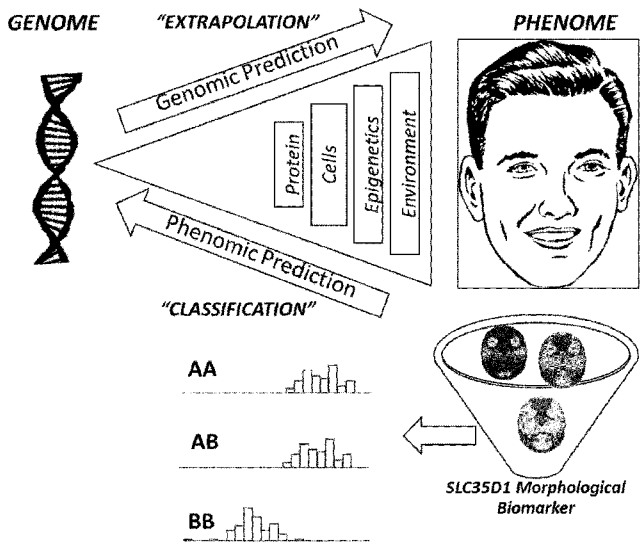


Fig. 1

(57) Abstract: The present invention provides methods for predicting an unknown phenotype from a given genotype and finding genotype-phenotype associations. In particular, the method comprises integrating a plurality of phenomic prediction models that are adapted to predict a feature of a genotype based on a known phenotype into an optimization algorithm.

WO 2015/173435 A1

— *of inventorship (Rule 4.17(iv))*

Published:

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

METHOD FOR PREDICTING A PHENOTYPE FROM A GENOTYPE

Field of the invention

The present invention provides methods for predicting an unknown phenotype from a given genotype and finding genotype-phenotype associations.

Background of the invention

The phenotype includes an organism's observable characteristics, such as health, disease, morphology, appearance and fitness. A key goal in biology is to understand phenotypic variations produced through a complex web of interactions between genetic and environmental variations. Predicting organismal phenotypes from genotype data (e.g. genomic prediction as illustrated in Figure1) or more generally genetic features such as single nucleotide polymorphisms (SNP), haplotype, copy number variation (CNV), genomic... variations, or any other type of variation that can be measured from DNA has important applications in medicine, forensics, and agriculture. In contrast to Mendelian traits, the modelling from genotype to phenotype for complex traits is difficult in part due to genetic and epigenetic interactions at different molecular and environmental levels. Prior to the genomic era, classical quantitative genetics was focused on the prediction of phenotypic response to artificial selection at the level of populations and across generations, which is useful for the purposes of animal and plant breeding. However, simplifying assumptions that are routinely made in classical quantitative genetic theory are incorrect for complex traits. More recently, the analysis of quantitative trait loci (QTL) and the incorporation of complex gene-environment networks have become important methods in the study of quantitative genetics. Today we are well aware of the growing importance of gene product networks for the study of biological systems.

Since the advent of high-throughput genotyping, bioinformatics research uses much larger datasets, stored in databases, than previously possible and in combination with advances in computer science and computational capacity, has been able to take new types of approaches. A versatile bioinformatics toolbox for genotype-to-phenotype predictions is emerging and is helping to pave the path to personalized medicine and DNA-based forensic predictions e.g. for relatively 'simple' non-Mendelian traits, like eye-color, predictions have become possible to a practical extent using only six SNPs in a straightforward multinomial logistic regression model. These SNPs were identified in association studies such as genome-wide association studies (GWAS) and linkage analysis studies, both of which are extensively used in genetic epidemiology. In such an association study, variability in SNP variation is tested against variability in phenotype variation and if the association is significant based on a chosen test statistic and p-value cut-off, the SNP is said to be associated. One of the main challenges here is the separation between true and false positive associations and in order to tackle this problem the same

test of association is typically repeated in an independent so-called replication panel of samples. The first panel used is then referred to as the discovery panel. In contrast to the very accurate estimates possible for sex, eye and hair color, much less accurate predictions have been reported for adult body height. For most complex traits, the typical approach today is to type vast numbers of SNPs (i.e. high density SNP panels) located throughout the genome. In other words, by not selecting SNPs based on a statistical significant p-value cut-off. Subsequently, adapted statistical and machine learning models such as genomic best linear unbiased prediction, Bayesian regression as well as fully non-parametric approaches are used to perform so called genomic-predictions (as illustrated in Figure 1). The paradigm in a genome-based prediction is simple and different from that of a genome-wide association studies (GWAS) as the former include the entire genome variability captured by the available marker set and do not rely on selection of single loci based on significance tests. As such SNPs in strong linkage disequilibrium (LD) with important causal variants are most probably included in the prediction. The challenge however, is to deal with the large numbers of SNPs, many of which have no or only a very small effect on the phenotype, such that sparse constraints are often incorporated into the model.

The accuracy of genomic-predictions is known to be subject to the size of the genome, the density of the markers, the levels of LD, the heritability, and genetic architecture of the trait under study. Some models can therefore try to incorporate higher order SNP interactions, which lead to the requirement of cloud computing capabilities and very large sample sizes with increasing cost. Some methods known in the art also rank SNPs by association effect and it has been shown that the accuracy of the predictions can be improved by taking aspects of the underlying genetic architecture, into account. Therefore, despite differences in approach, most if not all disciplines dealing with genotype-to-phenotype mapping of complex traits converge to the same conclusion: Progress is only anticipated through the incorporation of epigenetic and gene regulation networks. However, there is much still to learn about these methods through both applied studies and intensive computer simulations.

Besides the (epi) genetic complexity, success in predictions of phenotype from genotype also depends on the physical convolution of the trait itself. Indeed, in contrast to most traits tested for prediction so far (eye color, height, weight, starvation stress, milk fat percentage, overall type…) multipartite traits, such as facial morphology, imply multidimensional and multivariate quantitative traits. Generally, such physical complexity is simplified extensively by a priori breaking the trait down into smaller parts from which aspects are extracted as univariate scalar measures. Common aspects of facial morphology that are measured for example, are distances (i.e. mouth width), angles (i.e. nose angle) and ratios (short/long faces) between inter-landmark distances. Two examples include body height and nose width for body and facial morphology, respectively, where the prediction accuracy of both is far from promising. Furthermore, even when several of these measurements are used individually, they still oversimplify the spatial configuration too much and when they are used together, they are difficult to

interpret, as in they fail to produce an image-like recognizable view of the result. Alternatively, a dimensionality reduction technique, like principal component analysis (PCA) on landmark configurations or Fourier analysis on outline data, is applied and the resulting PC axes or harmonics are treated as separate phenotypic traits and investigated separately. A simple synthetic simulation for prediction from genotype using harmonics was recently introduced by Bo *et al* in "Shape mapping: genetic mapping meets geometric morphometrics" (Briefings in bioinformatics, 2013, 15 (4): p. 571-581). However, the simulation model was built under the assumption of one QTL for shape that is associated with one marker, therefore ignoring any possible genetic complexity. Both strategies of physical simplification, with inter-landmark distances and PCA, have been used in recent GWAS studies on facial morphology. Together these distances and PCA based phenotypic traits were used to identify five different genes influencing normal-range facial morphology. Only finding five functional loci is a very low rate of success meaning that either many more participants are needed or better analytical methods are needed, or the trait is determined by rare variants and will be more difficult to map. Considering that many genes can result in craniofacial abnormalities and the differences among populations and sexes, which had to have evolved very quickly. In combination with the poor prediction results it is safe to say that the shape effects of genes do not necessarily coincide with the shape features typically selected a priori. This major shortcoming was already noted in by Klingenberg *et al* in "Quantitative genetics of geometric shape in the mouse mandible." (Evolution, 2001. 55(11): p. 2342-2352), who instead employed the multivariate breeder's equation on shape in the mouse mandible to combine different physical measurements of the phenotype into a single association analysis. For the investigation of the mandible's genetic architecture he employed a multivariate generalization of linkage analysis based on canonical correlation.

The phenotypic complement to genomics is phenomics which aims to obtain high-throughput and multidimensional phenotyping. The paradigm shift is simple and similar to the one made in the human genome project, instead of 'phenotyping as usual' or measuring a limited set of simplified features that seem relevant, why not measure it all? In contrast to genomic technologies, which successfully measure and characterize complete genomes, the development in phenomics lags behind. However, with the advent of ever more consumer-worn sensors, the technological hardware exists for extensively (wide variety of measurements from different sensors) and intensively (in great detail and high resolution) collecting quantitative phenotypic data. For example, 2D, 3D and 4D image, surface and/or medical scanners, provide the optimal technical means to capture information of static, dynamic as well as functional phenotypes such as morphology, appearance and physical as well as chemical function as dynamic processes of motion and metabolism to the level of phenomics. An example phenomic approach to describe eye-color is the use of eye iris images instead of a single categorical color variable as illustrated in Figure 2.

Based on 3D surface scans, Claes *et al* in "Modeling 3D facial shape from DNA." (Plos Genetics, March 2014) extended the use of partial least square regression, which is a multivariate association technique similar to canonical correlation, by incorporating it into a novel bootstrapped response-based imputation modelling (BRIM) framework. In BRIM, facial morphology, as a response to predictors like genomic ancestry, self-reported sex and individual SNP genotype, was used to bootstrap the relationship between these variables with the result of both filtering out noise in the initial predictor variables clarifying and revealing their effects on the response. In this way, they were able to significantly associate and reveal the effects of 20 out of 50 candidate genes (selected from the literature on dysmorphology) on normal-range facial morphology in an admixed population. An example for a SNP in SLC35D1 is seen in Figure 1. In follow-up work, Claes *et al* in "Toward DNA-based facial composites: Preliminary results and validation." (Forensic Science International: Genetics, November 2014, 13: p.208-16), combined these pictorial effects to produce a prediction of facial morphology in a two-staged procedure. In the first stage, genomic ancestry and sex were used to create a 'base-face', which is simply the average sex- and ancestry-matched face. Subsequently, the genotype-based effects of the 24 significant SNPs are overlaid onto the base-face using a heuristically weighted photomontage approach forming the 'predicted-face'. The aspects that define the individual from the SNP information can be analyzed by comparing the predicted-face against the base-face. Two example predictions are presented in Figure 3.

Despite tremendous advances in genotyping and sequencing methods and dramatic increases in computing power, for many traits, multipartite or not, current genomic-predictions generally remain proof-of-concept demonstrations or simulation exercises. There are many unknowns that require investigation and the latest discoveries in genetics can and should be fully utilized improving these methods. Joint ventures, like genotype-phenotype mapping in a post-GWAS world, will become important tools facilitating these discoveries. However, until these discoveries are made, the difficulty for genomic-predictions is not just in dealing with vast amounts of SNP data, but also in the extrapolation of DNA information and (epi) genetic and environmental factors towards the phenotype. The same difficulty of extrapolation applies to the results shown in Figure 3, where the pictorial effects of all SNPs could have been combined differently, leading to different and maybe better results.

Therefore there is a need for new improved methods for predicting a phenotype from a genotype.

Summary of the invention

It is an object of the present invention to provide a good, improved method for associating and predicting an unknown phenotype from a given genotype. Improvements over the prior art are observed in the quality/correctness of the predicted unknown phenotype and/or in the computing power required for the prediction. Additional improvements over the prior art are observed in finding

associated genetic variations to phenotypic variations beyond simple statistical discovery and replication confirmation, leading to an increased number and stronger confirmed genotype-phenotype associations.

Embodiments of the present invention are preferably based on the novelty of predicting individual features of a genotype from a given (or synthetically generated) phenotype using phenomic prediction models. A phenomic-prediction model innovatively reframes a genomic-prediction as its inverse (Figure 1).

It is an advantage of embodiments of the present invention that the prediction of a feature of a genotype from a given (or synthetically generated) phenotype is framed within the simpler challenge of classification and/or regression in contrast to the challenging extrapolation of unknown factors in genomic-predictions.

It is an advantage that embodiments of the present invention provide a good, e.g. improved, method, which deal with the partial truth or incomplete knowledge in genetics that hamper current genotype-phenotype prediction models. Or in other words, methods that deal with the challenge of extrapolation from genotype to phenotype in a more sophisticated and improved way.

In a first aspect the invention provides methods predicting an unknown phenotype from a given genotype, the method comprising:

- providing a plurality of phenomic prediction models, said phenomic prediction models adapted to predict a feature of a genotype based on a known phenotype;

- integrating the plurality of phenomic prediction models in at least one optimization algorithm; and

- applying the at least one optimization algorithm to predict an unknown phenotype from a given genotype.

In preferred embodiments the method further comprises generating an error or matching score by the at least one optimization algorithm, whereby said error score is adapted to evaluate said known phenotype for a single or plurality of features of a given genotype. It is an advantage of the error score to evaluate a currently given (or synthetically generated) phenotype without genotype data as a solution in the optimization algorithm in the context of a single or plurality of features of a genotype to update and improve the currently given or synthetically generated phenotype as a solution to the prediction through the lowering of its error score accordingly.

In preferred embodiments the method further comprises generating multiple error scores, whereby said multiple error scores are adapted to evaluate and rank multiple known phenotypes for a single or

plurality of given features of a genotype. It is an advantage of generating multiple error scores to evaluate and rank multiple given or synthetically generated phenotypes to select the best known phenotypes as solutions in the optimization algorithm and to rank a given database of known phenotypes in order of goodness of fit or match with the given single or plurality of features of a given genotype. It is an advantage of ranking a given database of known (existing or synthetically derived) phenotypes in order of match as initial solutions to or solutions to update in the optimization algorithm or as an output of the invention on itself.

In a preferred embodiment, the phenotypes as described herein are represented by images. In particular, the methods of the invention employ an image database wherein multiple known phenotypes have been evaluated and are ranked for the given genotype.

In preferred embodiments the method further comprises the evaluation of a particular phenomic prediction model adapted to select associated and relevant single features of genotype data. In a particular embodiment, the method of the invention further comprises the evaluation of a particular phenomic prediction model to identify if an association is relevant between a) the feature of a genotype predicted by the particular phenomic prediction model, and b) the known phenotype upon which the prediction is based. It is an advantage of evaluating a particular phenomic prediction model adapted to select associated and relevant single features of genotype data to reduce the false positive detection rate in genotype-phenotype association efforts. It is an advantage to reduce the false positive detection rate in genotype-phenotype association efforts to select the best set of a plurality of phenomic prediction models to integrate into the optimization algorithm. The methods of the invention may further use the evaluation of a particular phenomic prediction model to find a relevant genetic variation in a genome-wide association study.

In preferred embodiments the phenotype is captured using high-throughput and multidimensional phenotyping and is embedded in a multidimensional or multivariate phenotype-space. In further preferred embodiments the multidimensional or multivariate phenotype-space is ordinated following genetically and/or environmentally meaningful dimensions (herein also referred to as genetically and/or environmentally meaningful directions). It is an advantage to embed a plurality of phenotypic features measured from the phenotype in a multidimensional phenotype-space to situate and compare a given phenotype to other phenotypes. It is an advantage to situate and comparing a given phenotype to other phenotypes to obtain phenomic-predictions of the given phenotype. It is an advantage to embed a plurality of phenotypic features in a multidimensional phenotype-space to generate new, synthetic phenotypes during the at least one optimization algorithm. It is an advantage to ordinate the phenotype-space following genetically and/or environmentally meaningful directions (dimensions) to

better construct and to increase the performance of phenomic-prediction models and the optimization algorithm.

In preferred embodiments, known and unknown phenotypes comprise multiple 2D, 3D or 4D images and said feature of a genotype comprises the information at an individual or a small group of SNPs, haplotypes, or genomic measurements (such as sex and genetic background also known as genomic ancestry e.g.) It is an advantage to use image-based phenotypes to obtain phenomic descriptions of static, dynamic and functional phenotypes such as morphology, appearance and physical as well as chemical function. It is an advantage to focus on individual features of genotype data at an individual or a small group of SNPs, haplotypes or genomic measures such that phenomic-prediction models remain simplified classification or regression models. Preferably said phenotypes comprise a facial phenotype or image, in particular a facial phenotype consisting of one or more 2D, 3D or 4D images. In a further embodiment, the known and unknown phenotypes further comprise non-image based phenotypic measurements (such as body mass index, age, etcetera).

In further preferred embodiments the method further comprises the step of reducing the multidimensional or multivariate phenotype-space to a one dimensional axis or curve in function of the feature of a genotype. It is an advantage to reduce the multidimensional phenotype-space to a one dimensional axis or curve that is highly correlated with the feature of a genotype to aid in the classification. Preferably said reducing is performed using imaging biomarkers, filters, or features; in particular using imaging biomarkers. In a preferred embodiment, the method further comprises tracing the genetic architecture or basis of said imaging biomarkers, filters, or features to improve classification from said imaging biomarkers, filters, or features into their associated feature of a genotype. It is an advantage but not necessary to incorporate any knowledge on genetic interactions.

In alternative preferred embodiments the method further may comprise classification from the multidimensional or multivariate phenotype-space into associated feature of a genotype. The reduction to a one dimensional axis or curve in function of the feature of a genotype is an advantage but not necessary for the invention to work.

In preferred embodiments the optimization algorithm is a global optimization algorithm, optionally followed by a local optimization algorithm. It is an advantage to use a global optimization algorithm as it is anticipated that the error function constructed from the integration of a plurality of phenomic prediction models into the optimization algorithm will contain multiple local optima due to the interaction of different features of genotype data (gene-gene interactions e.g.). Therefore to find the global optimum, a global optimization is preferred. The output of the global optimization algorithm can serve as input for a subsequent local search optimization to further refine the phenotype prediction from genotype. In further preferred embodiments the global optimization algorithm is an artificial

intelligence based algorithm such as an evolutionary algorithm. The advantage of artificial intelligence based optimization algorithms is their self-learning ability to handle very complex error functions with many local minima. Therefore the advantage is that our lack of knowledge on possible interactions between different features of genotype data is compensated for using self-learning artificial intelligence. The method allows to incorporate all the knowledge we know about the genetic architecture of a phenotype and artificial intelligence is used to deal with the partial truth or incomplete knowledge. It is an advantage of using an evolutionary optimization algorithm due to its efficient scanning of phenotype variations within a phenotype-space and efficiently uses the ability of the phenotype-space to generate synthetic phenotypes through the recombination of other phenotypes embedded in the same phenotype-space.

The developmental path from genotype-to-phenotype for complex and multipartite traits, like facial morphology, is difficult involving (epi) genetic and environmental interactions. Scientists all over the world are trying to model the involved complexity, with relatively slow progress. Therefore, genomic-prediction of the phenotype in these traits remains challenging. The central, novel and inventive theme in embodiments of the present invention is to reframe the problem as its inverse: phenomic-prediction of genotype w.r.t. a particular genetic feature such as but not limited to SNP, haplotype, CNV or genomic measures. The task is no longer to extrapolate unknown interactions, but to properly describe and filter the phenotype which is the result of all interactions. The prediction of a feature of a genotype from a given phenotype is framed within simple classification or regression challenges in contrast to the challenging extrapolation of unknown factors in current genomic-predictions. Phenomic-predictions lead to complementary applications of genomic-predictions. Instead of straightforwardly trying to predict an unknown phenotype, one can measure how well a known (existing or new, previously unseen or synthetically generated) phenotype fits a given target genotype referred to as “phenotype fit” (Figure 4). This is done by building multiple phenomic prediction models for different associated features of a genotype and evaluating and combining the predicted feature values of a genotype from the given phenotype against the feature values of a given genotype into an error or matching score or goodness of fit. This allows for high-throughput screening of phenotypes against a given genotype. This has clear value in a variety of forensic as well as medical applications. For example, thousands of reference images are available in institutional databases, e.g., departments of motor vehicles, registered residents, and arrestees and can be tested against evidentiary DNA found on a crime scene (Figure 5). Finally and most importantly, testing phenotypes against a given genotype also opens up the possibility of performing an actual phenotype prediction from genotype using optimization techniques that simply evaluate/rank and/or improve (adapt) phenotypes by minimizing error scores. In a particular embodiment, an artificial intelligence optimization technique is used to evaluate and recombine a population of known phenotypes as solutions in an evolutionary algorithm (Figure 6). In other words,

besides novel implications and applications, the current invention also provides an alternative approach to phenotype prediction from a given genotype. Evolutionary algorithms are known to cope well with complex and non-linear optimization problems and are tolerant to data imprecision and partial truth. In other words, the lack of knowledge in gene interactions going from genotype to phenotype today is compensated for by these advanced AI-based optimization techniques. In current genomic-prediction methods typically a single prediction model (explicitly incorporating genetic variations and gene interactions) is trained and tested. However, in an evolutionary algorithm as in the preferred embodiments of the present invention, multiple (one for each selected feature of a genotype) 'simpler' phenomic prediction models to predict a feature of a genotype from a known phenotype are trained and used, without explicitly having to model gene interactions. The evolutionary algorithm generates solutions based on the genotype-phenotype correlation captured in these phenomic-prediction models. Improving these phenomic prediction models or adding new ones, through new discoveries in genetics for example, will help narrow down the range of possible solutions to more specific outcomes. Small updates and discoveries for the simpler underlying phenomic prediction models may lead to large improvements in the optimization algorithm, while large updates and discoveries are needed for small improvements in current genomic prediction models found in the state-of-the-art.

Note that evolutionary algorithms have been used in the context of genomic-predictions. However, in contrast to the present invention, they have been used to optimize a genomic-prediction model based on a training dataset through the recombination and mutation of prediction models, e.g. in the form of random decision forests, which constitute the populations of solutions. Subsequently, once trained or optimized the same model is used for multiple predictions. In the current invention, the solutions to recombine are phenotypes in contrast to prediction models and for each phenotype to predict from a given genotype the algorithm starts over and performs new self-learning cycles. In that sense, each prediction involves a new optimization, finding the known or synthetically generated phenotype that best fits a given genotype, based on partial knowledge captured in one or a plurality of phenomic prediction models.

Preferred embodiments use phenomic prediction models to enable genomic prediction, whereby said phenomic prediction models go from a known phenotype to genotype (Figure 7). This inversion might seem trivial at first, but it represents a fundamental reformulation, contrary to the approaches in the prior art, which affects many of the complexities involved. As the phenome is the overt, integrated and measureable endpoint of these interactions, the task is no longer to extrapolate unknown interactions, but to properly represent and strip down the phenotype as the result of interactions onto the level of individual features of a genotype such as but not limited to SNP, haplotype, CNV or genomic variations. Instead of seeing the physical complexity of the phenotype as a burden, it is preferred to be typed intensively following the principles of phenomics. In embodiment, where static, dynamic and/or

functional morphology, appearance, and physical as well as chemical processes is applied, this involves the use of 2D, 3D or 4D imaging techniques. Subsequently, the known and unknown phenotypes are preferably embedded in a multidimensional phenotype-space (Figure 8) (shape-space in the case of morphology) that is ordinated following genetically and environmentally meaningful dimensions (directions). In contrast, the genetic complexity is preferably simplified to the level of individual or small groups of SNP, haplotype, CNV or genomic measures like sex, genetic relationship and background also known as genomic ancestry. Genotype estimation is performed through the selection of associated and simplified genotype features and the extraction of phenotype biomarkers, features or characteristics within phenotype-spaces that can be used as indicators of genotype state or condition. In one embodiment, where facial morphology is applied, as depicted in Fig. 1, this implies the selection of associated genotype features with facial morphology and the extraction of morphological biomarkers or features from facial image data, which are particular quantifiable shape characteristics. These are essentially a posteriori physical simplifications of morphology, targeting a particular configuration of changes within a phenotype-space that is found to be in association with a particular genetic or environmental factor. In the simple example of sex; facial masculinity, as a suite of facial changes as revealed from male/female differences in ancestry and age controlled population samples, can provide accurate categorical predictions of sex from facial morphology. By treating these morphological biomarkers or features, once extracted, as individual and simplified phenotypic traits, they can be investigated in any number of ways to gain insight in their genetic architecture.

In preferred embodiments, the performance of a phenomic prediction model is adapted to be used as a test-statistic to facilitate the selection of associated genotype features as an alternative to traditional association studies relying on simple test-statistics and a separated discovery and replication stage. The performance of a phenomic prediction model is defined as its ability to predict a feature of a genotype from a given phenotype and is linked with the strength of the association between the feature of a genotype and the phenotype. The stronger the association the better the performance. It is an advantage to use the performance as test-statistic to discover genotype-phenotype associations because it is testing the ability of the association in terms of prediction significance besides merely statistical significance. In further preferred embodiments, the performance of a phenomic prediction model is evaluated in a cross-validation framework. It is an advantage to test the performance of a phenomic prediction model using K-fold cross-validation (Figure 9) to test the prediction significance on partitions of the data not used for model training to avoid model over fitting and evaluate the generalizability of the phenomic prediction model (its ability to be applied onto unseen and newly created synthetic data).

In a particular embodiment, the present invention provides a method comprising:

- providing a database of known phenotypes;

- providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype;
- integrating the plurality of phenomic prediction models in at least one optimization algorithm;
- applying the at least one optimization algorithm comprising:
  - calculating an error score by comparing features of the genotype predicted by the phenomic prediction models based on known phenotypes from the database with the given genotype;
  - ranking known phenotypes in the database according to the error score;
  - selecting a subset of known phenotypes based on said ranking;
  - optionally creating synthetic phenotypes based on the known phenotypes in the subset and adding the synthetic phenotypes to the subset; and
  - iterating the steps of the optimization algorithm at least one time on the subset of known and optionally synthetic phenotypes;
- predicting the unknown phenotype for the given genotype to be one of the highest ranked known or optionally synthetic phenotypes, or to be a synthetic phenotype derived therefrom.

In an alternative embodiment, the present invention provides a method comprising:
- providing a database comprising at least one known phenotype;
- providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype;
- integrating the plurality of phenomic prediction models in at least one optimization algorithm;
- applying the at least one optimization algorithm comprising:
  - calculating a first error score by comparing a first feature of the genotype predicted by at least a first phenomic prediction model based on at least one known phenotype from the database with the given genotype;
  - calculating a second error score by comparing a second feature of the genotype predicted by at least a second phenomic prediction model based on at least one known phenotype from the database with the given genotype;
  - adapting at least one known phenotype from the database based on the first and second error score to obtain a database comprising at least one adapted phenotype;
  - optionally iterating the steps of the optimization algorithm at least one time on the database comprising the at least one adapted phenotype;
- predicting the unknown phenotype for the given genotype to be the adapted phenotype.

In a second aspect, the present invention provides a computer program product for, if implemented on a control unit, performing a method according to the first aspect of the present invention.

In a third aspect, the present invention provides a data carrier storing a computer program product according to the second aspect of the present invention. The term "data carrier" is equal to the terms "carrier medium" or "computer readable medium", and refers to any medium that participates in providing instructions to a processor for execution. Such a medium may take many forms, including but not limited to, non-volatile media, volatile media, and transmission media. Non-volatile media include, for example, optical or magnetic disks, such as a storage device which is part of mass storage. Volatile media include dynamic memory such as RAM. Common forms of computer readable media include, for example, a floppy disk, a flexible disk, a hard disk, magnetic tape, or any other magnetic medium, a CD-ROM, any other optical medium, punch cards, paper tapes, any other physical medium with patterns of holes, a RAM, a PROM, an EPROM, a FLASH-EPROM, any other memory chip or cartridge, a carrier wave as described hereafter, or any other medium from which a computer can read. Various forms of computer readable media may be involved in carrying one or more sequences of one or more instructions to a processor for execution. For example, the instructions may initially be carried on a magnetic disk of a remote computer. The remote computer can load the instructions into its dynamic memory and send the instructions over a telephone line using a modem. A modem local to the computer system can receive the data on the telephone line and use an infrared transmitter to convert the data to an infrared signal. An infrared detector coupled to a bus can receive the data carried in the infra-red signal and place the data on the bus. The bus carries data to main memory, from which a processor retrieves and executes the instructions. The instructions received by main memory may optionally be stored on a storage device either before or after execution by a processor. The instructions can also be transmitted via a carrier wave in a network, such as a LAN, a WAN or the internet. Transmission media can take the form of acoustic or light waves, such as those generated during radio wave and infrared data communications. Transmission media include coaxial cables, copper wire and fibre optics, including the wires that form a bus within a computer.

In a fourth aspect, the present invention provides in transmission of a computer program product according to the second aspect of the present invention over a network. In another embodiment, the methods of the present invention, or parts thereof, may be executed on a remote server. For example, a user may transmit a given genotype to a remote server, upon which the methods of the inventions are executed on the remote server, and the predicted unknown phenotype is transmitted from the remote server to the user.

In a fifth aspect, the present invention provides devices for predicting a phenotype from a given genotype, said device comprising a control unit whereby the control unit is adapted to perform a method according to embodiments of the present invention.

Furthermore, the present invention provides the use of a method as described herein in a genome-wide association study.

In addition, the present invention provides a database of ranked phenotypes. In particular, a database comprising multiple phenotypes, wherein the phenotypes have been ranked in relation to a given genotype in accordance with the method described herein.

Particular and preferred aspects of the invention are set out in the accompanying independent and dependent claims. Features from the dependent claims may be combined with features of the independent claims and with features of other dependent claims as appropriate and not merely as explicitly set out in the claims.

These and other aspects of the invention will be apparent from and elucidated with reference to the embodiment(s) described hereinafter.

Brief description of the drawings

In all drawings, Gx whereby x is an element of {1, 2, 3... N} refers to different features of a genotype and Py whereby y is an element of {1, 2, 3... N} refers to different features of a phenotype unless otherwise stated in the brief description of the figures below. In the drawings, several images of humans and animals are represented using sketched drawings only. It is to be understood that these represent real (e.g. photographic) images with gray and/or color intensities in pixels, voxels and/or 3D points that may be used in the methods of the invention.

Fig. 1 illustrates the fundamental difference between genomic-prediction and phenomic-prediction. In the example illustrated in Fig. 1 a single SNP in the gene *SLC35D1* is used. After the discovery of a morphological biomarker for a particular SNP (rs1074265) in this gene, the idea is to score the given facial image with the morphological biomarker and to infer its genotype for this SNP from this score. Embodiments of the present invention advantageously provide phenomic-prediction models for all or a subset of SNPs associated with facial morphology and to apply these in an evolutionary algorithm to make improved genomic-predictions for persons not in the database.

Fig. 2 illustrates an image which can be used for image-based phenomic description of eye-color used in embodiments of the present invention.

Fig. 3 illustrates facial predictions obtained using BRIM, a method known in the art. The actual face, predicted face and their facial signature of the predicted-face against the base-face (gender and ancestry matched average) is given for a female (A-E) and male (F-J) subject. The actual face of the female subject is shown in front view (A) and side view (C), the predicted face is shown in front view (B) and side view (D) as well. E represents the facial signature of the predicted face against the base-face.

Similarly, for the male subject, the actual face is shown in front view (F) and side view (H) together with the predicted face (G and I, respectively). The facial signature of the predicted face against the base-face for the male subject is shown as well (J). This is the closest attempt made today by prior art methods to generate a face from DNA. Further improvements are hampered by a lack of our knowledge in how to combine for instance pictorial SNP effects in gene-networks. The results shown are just one way of combining SNP effects, where some facial aspects are well approximated, but many others are not, such that these results remain a simple proof-of-concept.

Fig. 4 illustrates establishing a phenotypic fit, illustrated as starting from a CT scan image of the human head against given DNA, which can be done by applying embodiments of a method of the present invention. In embodiments this can be done for instance by building multiple phenomic prediction models for different associated features of a genotype and evaluating and combining the predicted feature values of a genotype from the given phenotype against the feature values of a given genotype into an error or matching score or goodness of fit.

Fig. 5 illustrates high throughput screening of multiple known phenotypes in a database against a given genotype. Thousands of reference images are available in institutional databases, e.g., departments of motor vehicles, registered residents, and arrestees and can be tested against evidentiary DNA found on a crime scene. The result is a ranking of the multiple phenotypes in a database from best to worst match with the given genotype.

Fig. 6 illustrates the flow chart of a general evolutionary optimization algorithm, a method known in the art, which may be used in the methods of the present invention.

Fig. 7 illustrates the inversion from genomic prediction to phenomic prediction illustrated on the eye as phenotype of interest according to embodiments of the present invention. This inversion might seem trivial at first, but it represents a fundamental reformulation which affects many of the complexities involved. Px represents phenotype features, whereas Gx represents genotype features.

Fig. 8 illustrates the embedding of phenotypes into a phenotype space, which is a multidimensional space constructed from multiple phenotype features (Px) from different phenotypes (i,j or l). A single phenotype (i, j or l) is represented as a single point in the phenotype space. Once constructed, new synthetic phenotypes (s), can be generated based on the phenotype-space at different multi-dimensional point locations within the phenotype-space.

Fig. 9 schematic illustration of K-fold cross-validation, a method known in the art, that may be applied in the methods of the present invention.

Fig. 10 illustrates different features of a genotype using very local SNP information (G1), multiple SNPs or Haplotypes (G3), the full genome (GN) or a combination thereof (G2).

Fig. 11 illustrates possible modes of inheritance in SNP variation.

Fig. 12 illustrates global and continuous features of a genotype. Based on dense SNP sets across the genome, a genetic distance (viz., the allele sharing distance) between any two individuals in a DNA database can be computed. All pairwise distances between individuals constitute the so-called genomic relationship (also known as identity by state) matrix, from which axes of overall genetic variation can be established using principal coordinate analysis. These constitute another type of genetic features of interest and in contrast to individual SNP variations these are not categorical but continuous in nature. The concept is illustrated based on 4 populations living across Europe. Individuals within each of the 4 populations are genetically closer in distance compared to individuals between the 4 populations.

Fig. 13 illustrates a database which can be used in embodiments of the present invention. A database to learn phenomic predictions preferably comprises a subset of individuals from which both genetic as well as phenotypic information may be sampled. The former is preferably collected in the form of DNA, from which different features of a genotype can be extracted. The latter is preferably given in the form of images of the phenotype that are embedded in a multidimensional phenotype space, illustrated for human facial appearance. The phenotype-space is capturing overall modes of variation across phenotypes as well as directions of phenotype variation in function of factors of interest, such as, but not limited to ageing, BMI and sex. Within this space, the phenomic description of for instance human facial appearance is represented as a single point. Similarities between facial phenotypes can be expressed as distances and angles in this face-space. Additionally a facial typicality can be obtained by taking the distance of a particular facial phenotype to the consensus or average face, which is located at the centroid of the face-space.

Fig. 14 illustrates exemplary embodiments of the present invention focusing on facial appearance. (A) Facial morphology is described using 7150 3D points and (B) 683 by 683 2D pixel images are used to describe facial texture. When combined they describe facial appearance (C).

Fig. 15 illustrates the construction of symmetrized facial phenotypes of morphology as explained in Claes *et al.* in "Modeling 3D facial shape from DNA." (Plos Genetics, 2014), whereby A) illustrates an original surface, B) the reflected surface of (A) and C) The symmetrized surface of (A) and (B) by taking the average of both surfaces in (A) and (B).

Fig. 16 illustrates a phenomic prediction model according to embodiments of the present invention comprising of an image-based classification and/or regression on a feature of a genotype based on a

plurality of image features or biomarkers. The example shows a classification, but the same applies for a regression.

Fig. 17 illustrates a phenomic prediction model comprising of a single strong classifier (Top) or an ensemble of multiple strong and/or weak classifiers (Bottom) used in embodiments of the present invention. In the bottom ensemble of classifiers, multiple classifiers are combined using an ensemble strategy on the classification outcomes. The same applies for regressions. Illustration is based on an image-based description of a domestic cows' appearance.

Fig. 18 illustrates an alternative embodiment of the present invention, whereby combining an ensemble of classifiers is provided by combining them into a single classifier producing a single outcome, instead of combining multiple outcomes as illustrated in Figure 17. Illustration is based on an image-based description of domestic cows' appearance.

Fig. 19 schematically illustrates multiple bootstrap sampling runs of training data embedded within a k-fold partitioning of the complete dataset. The result is an ensemble of classifiers learned from randomized training datasets. The same applies for regressions.

Fig. 20 schematically illustrates combining multiple classifiers from distinct bootstrap sampling runs into a single combined classifier for one fold on a k-fold partitioning of the complete dataset illustrated in Figure 19.

Fig. 21 schematically illustrates combining multiple classifiers from different folds from a k-fold partitioning of the complete dataset illustrated in Figure 19.

Fig. 22 illustrates how to use the output of a phenomic prediction model according to embodiments of the present invention to generate a discreet or continuous matching score that expresses how well the phenotype fits to the feature value of a given genotype for which the phenomic prediction model was trained illustrated on an image-based description of the phenotype appearance of a given domestic cow. In the first case (top section), the domestic cow phenotype poorly fits to a SNPx feature in the given genotype, because the estimated SNPx feature label (AA) using the related phenomic prediction model and the given SNPx feature label are different (BB). This results in a low binary or continuous matching score and a high error score. In the second case, using another phenomic prediction model related to the SNPy feature of a genotype, the domestic cow phenotype fits well to the SNPy feature of the given genotype, because the estimated SNPy feature label (AB) and the given SNPy feature label (AB) are the same. This results in a high binary or continuous matching score and a low error score. As a result a ranking may be provided in best matching score (i.e. low error score) and vice versa.

Fig. 23 schematically illustrates the evaluation of the performance of a phenomic prediction model according to embodiments of the present invention to select relevant features of a genotype that are associated with variations in phenotype. A feature of a genotype is selected to be relevant or associated if a phenomic-prediction model trained on one partition of the dataset performs adequately enough on a separate and independent test partition of the dataset (e.g. G1, G3, GN). In the case the phenomic-prediction model does not perform adequately enough on an independent test partition of the dataset, the feature of a genotype (e.g. G2) is said not to be associated or relevant and is preferably not used in subsequent stages of embodiments of the present invention.

Fig. 24 illustrates dealing with imbalanced dataset designs by up sampling the minority class to a desirable balance level with the majority class, a method known in the art which may be used in methods of the present invention. Synthetic minority phenotypes are preferably created from given minority phenotypes. Embodiments of the present invention use e.g. a randomized linear interpolation in the phenotype space between randomly selected minority class phenotypes. Alternative embodiments of the invention can straightforwardly use other synthetic phenotype generators as well as other strategies to deal with imbalanced dataset designs such as, but not limited to, under sampling the majority class, another method known in the art.

Fig. 25 schematically illustrates response-based imputation in BRIM, a method known in the art. Using partitions of the dataset into training and test datasets, an underlying relationship model in BRIM is learned from the training dataset and used to impute and correct the values of the predictors in the test dataset.

Fig. 26 schematically illustrates bootstrapping (in the sense of improving itself in contrast to statistical bootstrap sampling) response-based imputations in BRIM, a method known in the art.

Fig. 27 illustrates how a discovered image biomarker is able to generate new RIP values for unseen or previously unavailable known phenotypes as well as new and synthetically generated known phenotypes. The RIP value is obtained by projecting the known phenotype under investigation onto the direction associated to the image-biomarker. Subsequently the RIP value is computed as the signed distance between a reference known phenotype, the consensus phenotype e.g., and the projected known phenotype under investigation.

Fig. 28 illustrates how RIP distributions of known phenotypes in a training dataset are used to evaluate the RIP value of the newly given known phenotype to see whether it matches with a feature value of a given genotype. The illustration is based on a morphological biomarker for a SNP in the gene *SLC35D1*. The training dataset RIP distributions are used to obtain a probability that the test known phenotype matches the feature value of a given genotype. The final output is a matching and/or error score

expressing the phenotypic fit of the test known facial phenotype with the feature value of the given genotype in the SNP of *SLC35D1*.

Fig. 29 illustrates how the forced imputation aspect from the BRIM procedure, a method known in the art, is nested as an inner k-fold partitioning within the outer k-fold partitioning to learn and or evaluate phenomic prediction models (Top section). Multiple bootstrapping (in the sense of improving itself in contrast to statistical bootstrapped sampling) iterations in BRIM are performed using different inner fold partitions of the data set within a single outer fold in each iteration (Bottom section).

Fig. 30 schematically illustrates detecting an appropriate mode of inheritance of a SNP feature of a genotype (Gx) by evaluating the performance of phenomic prediction models under for instance seven different mode of inheritance assumptions and associated test setups. The learning and evaluation of the phenomic prediction models is performed on training and test partitions of the complete data set using the appropriate k-fold cross-validation with k equal to four as illustrated in Figure 9. Each test-setup generates a binary outcome, and the combination of all seven binary outcomes allows to infer the correct mode of inheritance.

Fig. 31 illustrates dealing with imbalanced dataset designs by up-sampling the minority group when two groups are pulled together in test setup 1-3 in Figure 30. The up-sampling is done as illustrated in Figure 24. Top scenario, pulling group 1 and 2 together without upsampling, when both are compared together to group 3. Bottom scenario, pulling group 1 and 2 together with upsampling of group 1, when both are compared together with group 3.

Fig. 32 schematically illustrates integrating a plurality of phenomic prediction models for a plurality of relevant features of a genotype into an overall error score, that expresses the phenotype fit of a given known phenotype in the context of the plurality of relevant features of a given genotype.

Fig. 33 schematically illustrates integrating a plurality of phenomic prediction models according to embodiments of the present invention for a plurality of relevant features of a genotype to rank a dataset of known phenotype solutions in an optimization algorithm in function of feature values of a given genotype.

Fig. 34 schematically illustrates integrating a plurality of phenomic prediction models according to embodiments of the present invention for a plurality of relevant features of a genotype to rank a database of existing known phenotypes in function of feature values of a given genotype.

Fig. 35 schematically illustrates a sequential optimization according to embodiments of the invention to for instance predict an unknown leaf appearance phenotype from genotype, using 2D leaf images, using a plurality of phenomic prediction models for a plurality of relevant features of a genotype in sequential

order to update a known leaf phenotype in function of feature values of a given genotype in order to generate the unknown leaf phenotype.

Fig. 36 schematically illustrates using an evolutionary optimization algorithm according to embodiments of the present invention to predict an unknown phenotype using a plurality of phenomic prediction models for a plurality of relevant features values of a given genotype. Populations of synthetic known phenotype solutions are preferably updated over subsequent cycles of the algorithm until a stopping criterion has been achieved. After convergence a final prediction result for the unknown phenotype is extracted from the final updated population of known phenotype solutions.

Fig. 37 illustrates a screenshot of software used for predicting facial morphology as phenotype from a given genotype using preferred embodiments of the present invention after population initialization. Relevant feature values of a genotype were extracted from DNA. A population of 2000 known phenotype solutions was initialized through a random sampling of 2000 synthetic known phenotypes from the phenotype space using a normal random number generator for each of the 78 dimensions of the phenotype space. The window on top right (C) provides 2000 current known phenotype solutions as single points on the first two dimensions of the phenotype space. The window on bottom right (D) provides the current prediction from the population of known phenotype solutions as a weighted average of all known phenotype solutions. The window left, more specifically the top panel of axes (A), provides monitoring of the mean overall error score of the current population of solutions, also printed in value on top of the panel of axes. The window left, more specifically the bottom panel of axes (B), provides monitoring of the population diversity as the mean Mahalanobis distance between all 2000 known phenotype solutions and the known consensus phenotype solution of the current population in phenotype-space.

Fig. 38 illustrates a screenshot of the software initialized in Figure 37 for predicting facial morphology as phenotype from a given genotype using preferred embodiments of the present invention after 10 algorithmic cycles as illustrated in Figure 36. It is observed that the population of solutions is moving through the phenotype space, whilst reducing the mean overall error score. The current facial prediction is updated accordingly.

Fig. 39 illustrates a screenshot of software initialized in Figure 37 for predicting facial morphology as phenotype from a given genotype using preferred embodiments of the present invention after 50 algorithmic cycles as illustrated in Figure 36. It is observed that the population of solutions is moving further through the phenotype space whilst the diversity in the population is decreasing in comparison to Figure 37 and 38. The mean overall error score is also further decreased. The current facial prediction is updated accordingly.

Fig 40. Illustrates a screenshot of software initialized in Figure 37 for predicting facial morphology as phenotype from a given genotype using preferred embodiments of the present invention after 100 algorithmic cycles and reaching the maximum of 100 iterations as stopping criterion. It is observed that the population of solutions has moved to a tight subspace in the phenotype space with a low remaining diversity in the population. The final facial prediction of the unknown phenotype is updated accordingly and returned as the final prediction output of the evolutionary optimization algorithm from the relevant features of a given genotype.

Fig. 41 provides a comparison of different methods for facial prediction from DNA. **A:** (comparative example): outcome of facial prediction using 81 SNPs, 5 Genetic Background axis, Sex, Age, (body) height and body (weight) according to the method described in Claes *et al* "Toward DNA-based facial composites: Preliminary results and validation." (Forensic Science International: Genetics, November 2014, 13: p.208-16). **B:** outcome of facial prediction based on the same 81 SNPs, 5 Genetic Background axis, Sex, Age, (body) height and body (weight) using the sequential optimization algorithm of the invention. **C:** facial prediction from the same 81 SNPs, 5 Genetic Background axis, Sex, Age, (body) height and body (weight) based on the methods of the invention employing an evolutionary optimization algorithm. **D:** the true face of the person whose DNA was used is illustrated, which was used to run the algorithm for predicting phenotype from genotype in the comparative example and according to embodiments of the present invention. In comparison to the true face, the predicted face of B is markedly better than A, while C is even better than B.

The drawings are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes. Any reference signs in the claims shall not be construed as limiting the scope. In the different drawings, the same reference signs refer to the same or analogous elements.

Detailed description of preferred embodiments

The present invention will be described with respect to particular embodiments and with reference to certain drawings but the invention is not limited thereto but only by the claims. The drawings described are only schematic and are non-limiting. In the drawings, the size of some of the elements may be exaggerated and not drawn on scale for illustrative purposes. Where the term "comprising" is used in the present description and claims, it does not exclude other elements or steps. Where an indefinite or definite article is used when referring to a singular noun e.g. "a" or "an", "the", this includes a plural of that noun unless something else is specifically stated. The term "comprising", used in the claims, should not be interpreted as being restricted to the means listed thereafter; it does not exclude other elements or steps. Thus, the scope of the expression "a device comprising means A and B" should not be limited to devices consisting only of components A and B. It means that with respect to the present invention,

the only relevant components of the device are A and B. Furthermore, the terms first, second, third and the like in the description and in the claims, are used for distinguishing between similar elements and not necessarily for describing a sequential or chronological order. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other sequences than described or illustrated herein. Moreover, the terms top, bottom, over, under and the like in the description and the claims are used for descriptive purposes and not necessarily for describing relative positions. It is to be understood that the terms so used are interchangeable under appropriate circumstances and that the embodiments of the invention described herein are capable of operation in other orientations than described or illustrated herein.

In the drawings, like reference numerals indicate like features; and, a reference numeral appearing in more than one figure refers to the same element. The drawings and the following detailed descriptions show specific embodiments of a novel method or device for predicting a phenotype from a genotype.

Where in embodiments of the present invention reference is made to “genotype”, reference is made to features of genotype data as described below. Moreover reference is made to local and/or global genetic variation. For instance genetic variations used can be defined either locally on the level of individual SNPs or globally on the level of genetic distances between individuals, also known as genetic background or ancestry. Individual SNPs are the typical locations within our DNA that vary from one individual to another with three possible genotypes each (AA, AB and BB). Typically, the variations within a single SNP are coded under an additive model assumption using a three-class categorical variable (AA=-1, AB=0, BB=1). SNPs can be extended to haplotypes, haplotype blocks and CNV, (copy number variation), epigenetic variation (molecular interaction with DNA), and combinations thereof.

A “given genotype” as used herein, refers to a genotype that is used to predict a phenotype from it. A “given genotype” may also be referred to as an “input genotype”. The “given genotype” is preferably determined from a tissue sample, such as a bodily fluid sample, e.g. a plasma or blood sample. The “given genotype” may as well be a synthetic genotype, e.g. a hypothetical genotype of a subject estimated based on the genotypes of family members of said subject.

Where in embodiments of the present invention reference is made to “genomic-prediction”, reference is made to extrapolating a phenotype, like for instance facial morphology, from a genotype, like for instance DNA.

Where in embodiments of the present invention reference is made to “phenomic-prediction”, reference is made to predicting a feature of a genotype from a known phenotype, like for instance facial morphology.

As used herein, a "known phenotype", also termed "given phenotype", refers to known, measured phenotypes as well as to synthetically generated phenotypes. In contrast, an "unknown phenotype" refers to a phenotype that is not known from the outset of the method, but that is predicted from a given genotype. As used herein, a "phenotype" without further indication refers to a "known phenotype" unless its context dictates otherwise.

As used herein, an "error score" may be a negative or positive indication value for the correlation between two parameters, unless the context herein dictates otherwise. In particular, an "error score" is calculated by comparing features of the genotype predicted by a phenomic prediction model based on a known phenotype from the database with the given genotype. Thus, the error score may be a negative indication value for the comparison between the predicted genotype and the given genotype, i.e. a higher error score indicates a worse fit between the two genotypes, or the error score may be a positive indication value for said comparison, i.e. a higher error score indicates a better fit between the genotypes. It is readily apparent to the skilled person that both types of error scores are equivalent and the one type can easily be converted to the other type. As used herein, a "matching score" refers to a type of error score, namely a score that is a positive indication value for the comparison of two features.

In a particular embodiment, the optimization algorithm will aim to improve the error score, i.e. lower a negative indication value or increase a positive indication value. In a further embodiment, the methods of the invention will be performed until the error score reaches a threshold value. The threshold may be a predetermined value or a value calculated during the methods of the invention, preferably the threshold is a predetermined value. For example, in embodiments applying iteration steps, the iteration may be performed until the threshold is reached. In an alternative embodiment, the iteration may be applied for a predetermined number of iterations, e.g. at least 5 iterations, in particular at least 10 iterations, more in particular at least 20 iterations. In a preferred embodiment, at least 25, more in particular at least 50 iterations are applied.

4D imaging is also known as dynamic 3D imaging. Here, 'dynamic' is defined in a broad sense as obtaining multiple 3D images of a phenotype of the same subject that changes in time (the 4th dimension), where the time-scale can be short (real-time) or long (follow-up). In the case of the latter, expectations on phenotypic growth, development and ageing as well as disease prognosis or survival following the framework the presented invention will become possible. In the case of the former, predictions from the analysis of the dynamics of real-time phenotypic changes like facial expression can lead to the anticipation instead of just recognition of human emotion. In a particular embodiment, a 4D image refers to a set of images wherein each image is a 3D image of the same subject represented at a different time.

Embodiments of the present invention advantageously provide a novel approach to the problem of genotype to phenotype correlations. The central theme of embodiments of the present invention is the reasoning from richness in phenotype to underlying genotype, such that a genomic-prediction is innovatively reframed as its inverse or phenomic-prediction as illustrated Figure 7: given the phenome what can we infer about the genotype at the level of individual SNPs, haplotypes, CNVs or genomic measures? To do so, a phenomics conceptual framework is adopted and adapted to move beyond 'phenotype analysis as usual' such that features of a genotype can be predicted. Methodologically, intellectual traditions both pre- and post-genomic era are preferably combined and further integrated with other techniques like for instance image computing to deal with and exploit the physical complexity in multipartite traits.

Features of genotype data: Embodiments of the present invention advantageously focus on features of genotype data either defined locally on the level of an individual or multiple SNPs or globally on taking the whole genome into account such as genetic distances between individuals (also known as genetic relationship, kinship, background or ancestry) (Figure 10). Individual SNPs are the typical locations within our DNA that vary from one individual to another with three possible genotypes each (AA, AB and BB). Typically, the variations within a single SNP are coded under an additive model assumption using a three-class categorical variable (AA=-1, AB=0, BB=1). In this case the estimation of SNP genotype from phenotype thus comes down to a three-class classification problem. Depending on a different mode of inheritance in an individual SNP, the classification can change as illustrated in Figure 11). If instead of an additive mode of inheritance, a recessive mode of inheritance (AA,AB = -1, BB = 1), dominant mode of inheritance (AA=-1, AB,BB = 1) or over dominant mode of inheritance (AA,BB = -1, AB = 1) is present the estimation of SNP genotype from phenotype comes down to a two-class classification problem. In case of the more complicated non-additive mode of inheritance, where the three classes are clearly distinct it is preferred to construct multiple phenomic prediction models for different combinations of the classes. For example, but not limited to, exemplary embodiments of the present invention use three two class phenomic prediction models defined as: (AA = -1, AB,BB=1), (AA,AB = -1, BB = 1) and (AA,BB = -1, AB = 1), which are in fact the recessive, dominant and over dominant models. In preferred embodiments of the invention, the correct mode of inheritance is detected by any means possible and taken into account to construct the phenomic prediction model.

On the other hand, taking all SNPs across the genome into account simultaneously, a genetic distance (viz., the allele sharing distance) between two individuals can be computed. All pairwise distances between individuals constitute the so-called genomic relationship (a.k.a, identity by state) matrix, from which axes of overall genetic variation can be established using principal coordinate analysis (Figure 12). These constitute another type of genetic features of interest and in contrast to individual SNP variations these are not categorical but continuous in nature. However, it should be noted that the estimation of

continuous variables can also be performed using classifications. It is straightforward to expand features on the level of individual SNPs to the level of small groups of SNPs combined as well as diplotypes, haplotypes or other measurements from DNA such as, but not limited to, Copy Number Variations (CNVs) as well as constructing different genetic distances between subjects in a dataset and establishing an ordination between them using other techniques than principal coordinate analysis onto axis of overall genetic variation.

Image-based phenotypes: Embodiments of the present invention advantageously may focus on the use of 2D, 3D and 4D image acquisition devices to obtain phenomic descriptions of phenotypes that can be captured using such 2D, 3D and 4D image acquisition devices, such as, but not limited to, phenotypes of static, dynamic and/or functional morphology and/or appearance and/or physical and chemical dynamic processes. In a particularly preferred embodiment, the methods of the invention are applied on 2D or 3D facial images.

Embodiments of the present invention advantageously may focus on the complex and multipartite phenotypic trait of facial morphology and appearance. Humans are subjective creatures constantly deciphering codes about one another based on visual cues. In this context, the face has evolved to be the single most telling part of the human body. It is a biological billboard advertising our sex, ancestry, general health, kinship, genotype, and environmental exposures. Craniofacial structures present important biomarkers for certain diseases and genetic mutations, providing many clues regarding growth and development. In dysmorphology, for example, the facial phenotype is extensively used in the diagnosis of patients. Genetically, craniofacial morphology is complex. Preferably by assuming that by capturing 3D facial morphology in all its complexity, we capture a substantial sampling of the externally evident human phenome, which will be able to predict the genotype for a range of underlying genetic features.

The idea of morphology or form, defined as size and shape independent of position and orientation, is a fundamental concept underpinning the life sciences ranging from molecular to organismal levels. The 3D shape of proteins for example, affects their interactions with other molecular structures like DNA, small molecules, and other proteins. The human visual-cognition system is capable of discriminating between different forms to identify and classify organisms into phylogenetic relationships. Conventional morphometrics is based on individual distances, distance ratios, and angles between morphological landmarks. Alternatively, geometric morphometrics uses the coordinates of homologous landmark configurations, defined as “a point of correspondence on an object that matches between and within populations”. A popular strategy to deal with the confounders of position and orientation is Procrustes superimposition which places the landmark data into a common frame or shape-space. In the shape-

space, the Euclidean distance between two configurations is known as the Procrustes distance and serves as a measure of shape difference or dissimilarity.

Technological advances over the past decade have improved the practical relevance of morphometrics tremendously. For example, with the advent of three-dimensional (3D) tomographic imaging and surface scanning, the quantification of form can now be performed in three dimensions. Furthermore, increases in computing power and advancements made in image computing methodology allowed for a substantial increase in the spatial resolution at which associated forms can be represented. For example, spatially-dense morphometric descriptions of facial shape can be obtained, which are homologous spatially-dense (>1000) quasi-landmark configurations. Note that, by homologous, we mean that each quasi-landmark occupies the same position on the face relative to all other quasi-landmarks for all individuals. Hence, it provides detail on the more overt as well as the more subtle differences and is independent of a potential 'facial perception' bias (i.e., concentration of landmarks on perceptually salient features of the face). This perfectly fits the phenomics view of intensive phenotyping of facial morphology. The main challenge here lies in formulating statistical models with appropriate distributional assumptions and with realistic variance-covariance structures that account for the morphometric characteristics of the data.

Embodiments of the present invention advantageously may focus on the complex and multipartite trait of facial morphology and appearance acquired using 3D surface scanning technology and described using spatially-dense 3D points or quasi-landmark configurations that are homologous across the facial images from different individuals in a database. It is straightforward to perform the same starting from 2D or 4D data and either using 3D points or individual pixel and/or voxel locations and/or intensities to describe morphology and appearance.

Data acquisition and intensive phenotyping according to embodiments of the invention are preferably performed by capturing, processing (establishing spatially-dense morphometric descriptions), and organization of 3D facial images from normal-range unrelated persons, twins, parent-offspring, as well as persons presenting with clinically significant facial dysmorphology. Although, ideally DNA would be available on all participants in the data collection, this is not required. In preferred embodiments the focus is largely on the collection of DNA on the normal-range. The normal-range dataset will therefore serve as the main instrument to construct the phenotype-space that can be used in association studies and to perform the prediction of phenotype from genotype (Figure 13).

In embodiments of the present invention, both a genotype and phenotype database can be expanded with genotypes and phenotypes from which the corresponding phenotype and genotype respectively is not given as long that there is a subset in the complete dataset for which both genotype and phenotype data is given.

Confounding factors and/or features: It is a well-known fact that additional genetic, epigenetic and or environmental factors can influence the phenotype under study. In the case of facial morphology confounding factors include, but are not limited to, sex, height, weight, age... Sometimes these factors can be extracted and estimated from DNA and sometimes knowledge about them can be gained by other sources of information. In these cases they can be taken into account in the prediction of phenotype from genotype, through the construction of phenomic prediction models of these factors as additional features to estimate from the phenotype. In all these and other cases most of these factors if acquired for the phenotype data in a database to train phenomic prediction models are preferably taken into account as conditional variables during the construction of phenomic prediction models of other features of a genotype. In the example of sex and facial morphology, sex can be estimated from DNA and as such a phenomic prediction model for a two-class categorical variable estimating sex from facial morphology can be built and taken into account during facial prediction from genotype. Age as another example can be estimated from DNA through the analysis of epi-genetics and as such age can be taken into account during facial prediction from genotype. The same accounts for weight and height or the combined measurement of both in body mass index if they are estimated from DNA or extracted by other sources of information (e.g. size and length of clothing). If no knowledge about features of these factors is known, one could aim to use the average value of the related features to the factors in the database of training samples as input for the prediction of phenotype from genotype or alternatively, create multiple phenotype predictions from genotype using a range of values for each of these unknown factors. Nevertheless all of these factors if of interest can be incorporated through the constructing of associated phenomic prediction models that predict features of these factors. For example, but not limited to, a phenomic prediction model can be constructed that estimates, sex, age, weight, height and/or BMI from facial morphology.

During the association of SNP features of a genotype with phenotypes, it is a well-known fact that population stratification is to be taken into account. The same applies for constructing phenomic prediction models for SNP features of a genotype. Any technique that deals with the problem of population stratification during genotype-phenotype association efforts can be straightforwardly used. Embodiments of the present invention advantageously focus on using the features of genetic background based on a principal coordinate analysis of genetic distances (e.g. allele sharing frequency) between individuals (Figure 12), as conditional variables in combination with sex, height, weight and age for the construction of phenomic prediction models of features of a genotype that are based on single or multiple SNP variations.

Learning Databases: In preferred embodiments of the invention databases used to train or learn the phenomic prediction models and to perform phenotype from genotype predictions comprises 2D, 3D or 4D image data of phenotypes combined with information and values on confounding variables of

interest and as many features of genotype data as available. Exemplary embodiments of the present invention focus on an existing dataset of 2006 3D facial images of subjects sampled across Europe, with additional knowledge on sex, age, weight, height and ~12000 SNPs. From the set of SNPs, 100 principal coordinate axes of variation in genetic distance between individuals in the dataset were constructed. The number of subjects and SNPs preferably is expanded if possible to improve the prediction of facial phenotype from genotype. Exemplary embodiments of the present invention focusing on facial morphology and appearance use 7150 3D points to describe morphology and 683 by 683 2D pixel images to describe facial texture and when combined describe facial appearance (Figure 14).

Phenotype-Spaces: In preferred embodiments of the invention a phenotype is embedded in a multidimensional or alternatively multivariate phenotype-space as illustrated in Figure 8 and 13. In a phenotype-space a single phenotype is represented as a single point in a multidimensional or alternatively multivariate space. The advantage of embedding a phenotype in a phenotype-space is the ability to analyse the phenotype in the context of other phenotypes by measuring similarities between them using a variety of multivariate distances and angles defined within the phenotype-space. The additional advantage of resulting phenotype-spaces is the ability to generate new and synthetic phenotypes based on the combination of other phenotypes that are embedded into the same phenotype-space or through the application of different directions within the phenotype-space that are linked to specific variations between phenotypes. Related to this advantage is the ability to generate a phenotype plausibility or typicality score or measurement (how common or distinct is a given phenotype?). A phenotype-space is preferably obtained by describing multiple phenotypes consistently using corresponding phenotype features between different phenotypes. Each feature that is corresponding across multiple phenotypes can add a single axis or dimension in the phenotype-space as illustrated in Figure 8. All or a subset of features combined define the complete multidimensional phenotype-space. For example, but not limited to, for image-based phenotypic descriptions all or a subset of 3D coordinates, pixel or voxel intensities after appropriate image-registration and/or point cloud registration each define a single axis of a multidimensional image-based phenotype-space. Preferred embodiments of the present invention advantageously focusing on facial morphology consider the coordinates of consistently mapped 3D point configurations across all 3D facial images using 3D point cloud registration techniques to establish a 3D face shape-space or face-space for short. The same can be straightforwardly obtained from 2D facial images using 2D image registration techniques available within the fields of computer vision and image computing.

In embodiments of the invention a dimensionality reduction can be performed on the phenotype-space. Most often information of 3D points, pixels or voxels in images is highly correlated, such that a lower dimensional image-based phenotype-space can be obtained that is still able to describe the complete phenotype or original image within an acceptable limit of accuracy and approximation. The first

advantage of a dimension reduction is an advantage in lowering computational and memory burden. A second advantage of a dimension reduction is the exposure of possible underlying patterns of variation between phenotypes. As an example, but not limited to, principal component analysis can be applied to reduce the phenotype-space onto a set of uncorrelated lower dimensional axes, that maximally capture the covariance patterns between multiple phenotype features. Principal component analysis is just one example of a linear dimensionality reduction technique and many others such as, but not limited to, independent component analysis, linear discriminant analysis... can be applied. Besides linear dimensionality reduction techniques, non-linear dimensionality reduction techniques and manifold learning can be applied as well.

Exemplary embodiments of the present invention advantageously focusing on facial morphology apply a principal component analysis to lower the dimension of the face shape-space to a final dimensionality of 78 that explains 98.5% of the total facial morphology variation observed within the dataset of 2006 3D facial images as illustrated in Figure 13. In further embodiments of the present invention advantageously focusing on facial morphology a Mahalanobis based distance or angle between two facial phenotypes within the shape-space is used to measure the similarity between both. In further embodiments of the present invention advantageously focusing on facial morphology the Mahalanobis distance of a phenotype to the centroid of the face shape-space is used to express the plausibility or typicality of the facial phenotype. The advantage of for instance using a Mahalanobis measure instead of a Euclidean measure, is that variations along different principal components are normalized or equalized such that the first set of principal components do not dominate in the measures. In further embodiments of the present invention advantageously focusing on facial morphology the face shape-space is used to create and evaluate synthetic generated facial phenotypes. In further embodiments of the present invention advantageously focusing on faces a similar phenotype-space for facial appearance instead of morphology is constructed based on the RGB intensities of 2D facial images that are connected to their respective 3D facial morphology using a 3D to 2D texture mapping and that are consistently constructed with corresponding pixels across different subjects in the database. In alternative embodiments of the present invention advantageously focusing on facial morphology and appearance any other phenotype-space construction, reduction and measurements of similarity and typicality can be used. It is straight forward to apply the same for other types of phenotypic morphology as well as appearance besides facial morphology and appearance.

In preferred embodiments of the present invention these phenotype-spaces are ordinated according to genetically and environmentally relevant variations (also referred to as the dimensions or directions of the phenotype-space). It is an advantage, but not a necessary requirement for the invention to work, to ordinate phenotypes in a phenotype-space according to genetically and environmentally relevant variations to increase the power in finding associated and predicting features of a genotype using

phenomic prediction models. In the example of facial morphology, the resistance against facial asymmetry during facial development is known to be genetically controlled. However, the patterns of fluctuating asymmetry are generally considered to be developmental noise due to influences of the environment during development. Through the construction of symmetrized phenotypes as done for faces in Claes *et al.* in "Modeling 3D facial shape from DNA." (Plos Genetics, 2014), resulting phenotype-spaces are not influenced by environmental noise captured in facial asymmetries as illustrated in Figure 15. In alternative embodiments of the invention results of heritability studies on the phenotype features can be used to focus the phenotype-space onto phenotype features that are under stronger genetic control. In the example of facial morphology, facial features such as the cheeks are more influenced by environmental diet and ageing for example in comparison of other facial features like the nose. As such, but not limited to, a weighted PCA giving higher weights to nose features and lower weights to cheek features can be applied to ordinate the resulting principal components emphasizing variations between features under stronger genetic control. Heritability studies of phenotype features can be performed based on proper datasets of for example, but not limited to, parents and offspring and identical in combination with fraternal twins.

In alternative embodiments of the invention 2D,3D or 4D image data can be used completely, partially or modular and on multiple levels of resolution to obtain image-based phenotype-spaces.

Phenomic prediction models: In preferred embodiments of the invention a phenomic prediction model comprises of an image-based classification and or regression on a feature of a genotype and other possibly confounding factors of interest based on a plurality of image features or biomarkers whether or not expanded with non-image based features or biomarkers as illustrated in Figure 16. Classification and regression are a core topic in machine learning, where a variety of classifiers and regressions, such as, but not limited to, random decision forests, artificial neural networks, statistical parametric and non-parametric methods, support vector machines… can be used in embodiments of the present invention to work. A full list of all possible techniques is not given, but expert skilled person in the field of machine learning or any form of computational science is perfectly aware of all possible techniques. All techniques typically start from a set or all possible image-based phenotype features that can be extracted from the image data and additional non-image based features present in the training database based upon which a classifier or regression for a feature of a genotype is trained and tested. In preferred embodiments of the invention all possible features are extracted from phenotype-spaces whether reduced in dimension or not to ensure consistency of features used for classification and/or regression on different phenotypes.

In embodiments of the invention a phenomic prediction model comprises of a single strong classifier and/or regression or an ensemble of multiple strong and/or weak classifiers and/or regressions as

illustrated in Figure 17. An example, but not limited to, of an ensembles of weak classifiers and/or regressions is generated in techniques similar to for instance adaptive boosting (ADABOOST) where subsequent classifiers are created based on previously misclassified training examples. The advantage of creating an ensemble of classifiers and/or regressions instead of a single classifier and/or regression is the ability to combine multiple weak and simple classifiers and/or regressions into a single strong classifier and/or regression. Different classifiers and/or regressions within such an ensemble can use different phenotype features or different combinations of phenotype features to perform the classification and/or regression. Another advantage of creating an ensemble of classifiers and/or regressions from a dataset is the ability to incorporate and deal with the variability in classifiers and/or regressions in function of the variability of underlying training data and to avoid over fitting of the training data as well as to increase the generalizability of the phenomic prediction model. An ensemble of classifiers and/or regressions can be combined into a single classifier and/or regression as illustrated in Figure 18 or used separately each predicting the feature of a genotype from phenotype after which the individual predictions can be combined in an ensemble strategy as illustrated in Figure 17. Exemplary embodiments of the present invention advantageously focusing on facial morphology learn a phenomic-prediction model based on multiple and different training datasets through a combination of k-fold data partitioning as illustrated in Figure 9 and multiple bootstrap sampling with replacement of the training data within each fold as illustrated in Figure 19. Resulting classifiers and/or regressions from multiple bootstrapped samples of training data within a single fold are combined into a single classifier and/or regression that is robust against variations in underlying training dataset variability within a single fold as illustrated in Figure 20. Resulting robust classifiers and/or regressions from different training folds are again combined in a similar fashion into a single classifier and/or regression that is robust against underlying training data variability across different folds and that does not over fit any of the training data samples in different folds as illustrated in Figure 21. In the implementation of the exemplary embodiment of the invention focusing on facial morphology, a 10-fold data partitioning combined with 50 bootstrapped samples with replacement within each fold is performed. However other k-fold and number of bootstrapped samples also without replacement can be used as well. Note again that resulting classifiers and/or regressions do not need to be combined and can be used separately but they are combined in the exemplary implementation of the embodiments of the present invention.

In embodiments of the invention the output of a phenomic prediction model comprises a discrete value or label for categorical genetic features and a continuous value for non-categorical genetic features whether or not combined with a continuous value expressing the confidence of the outcome estimation. In further embodiments of the invention the output of a phenomic prediction model is used to generate a discrete or continuous matching score that expresses how well the phenotype fits to the feature value

of a given genotype for which the phenomic prediction model was trained as illustrated in Figure 22. The continuous matching score can be converted to an error score, for example, but not limited to, by taking the negative logarithm of the matching score. Exemplary embodiments of the present invention advantageously focusing on facial morphology generate a continuous Bayesian based a-posteriori probability that a given phenotype matches to the feature value of a given genotype. The error score is preferably obtained by taking the negative logarithm of the matching Bayesian based a-posteriori probability.

In preferred embodiments of the invention the performance of a phenomic prediction model is evaluated using well-established classifier and/or regression evaluation techniques available in machine learning. In further preferred embodiments of the invention the performance is evaluated using k-fold cross-validation as illustrated in Figure 9. The advantage of a k-fold cross-validation is testing the performance of a phenomic prediction model when trained or learned and tested on separate partitioning's of the data available in the dataset to ensure a good generalization of the model onto unseen and/or previously unavailable phenotypes or new synthetic generated phenotypes within phenotype-spaces. In further preferred embodiments of the invention a performance measure is augmented with a p-value to determine if the performance is statistically significant or not by using the performance measure as a test-statistic under for example, but not limited to, permutation testing that is adapted to the classifier and/or regression used.

In further preferred embodiments of the invention the statistical evaluation of the performance of a phenomic prediction model is used to select relevant features of a genotype that are associated with variations in phenotype as illustrated in Figure 23. A feature of a genotype is preferably selected to be relevant or associated if a phenomic-prediction model trained on one partition of the dataset performs adequately enough to a certain, but not limited to, level of statistical significance on a separate and independent test partition of the dataset. In the case the phenomic-prediction model does not perform adequately enough to a certain, but not limited to, level of statistical significance on an independent test partition of the dataset, the feature of a genotype is said not to be associated or relevant and is preferably not used in subsequent stages of the invention. In a particular embodiment, an association is deemed relevant if it is statistically significant. Exemplary embodiments of the present invention advantageously focusing on facial morphology perform a 4-fold cross-validation of the 2006 samples available in the dataset. The result is that the performance of a phenomic prediction model for the feature of a genotype under investigation is tested 4 times and the p-values of the 4 test results are combined into a single p-value using the Fisher's method based upon which statistical significance and thus relevance of the feature of a genotype under investigation is determined. Other k-fold cross-validation setups than the 4-fold cross-validation can be used. In the example of using a 2-fold cross-validation, the testing for association resembles discovery and replication testing traditionally

performed within existing association studies with the additional advantage that a possible association is actually tested across both panels or folds instead of a merely within panel or fold statistical significant confirmation. The advantage is an improved approach to avoid the detection of false positive genotype-phenotype associations. It is advised to use a low number of k-folds such as k= 2, 3 or 4 when using the evaluation of phenomic prediction models to discover associated features of a genotype. The reason is mainly because the size of the test partitions of the dataset should not become too small to ensure enough statistical power in the evaluation of the phenomic-prediction model learned from the training partitions of the dataset. Exemplary embodiments of the present invention selected 81 associated SNPs from the list of 12000 SNPs, sex as estimated from DNA, 5 axis of genetic background out of a list of 100 axes of genetic background constructed by using the full list of SNPs to compute allele sharing distances between all 2006 subjects followed by a principal coordinate analysis, and age, weight and height as additional factors. All these features of a genotype and additional factors have been selected with the procedure of evaluating associated phenomic prediction models.

In preferred embodiments of the invention the training and performance evaluation of a phenomic prediction model is done using techniques in machine learning and statistical analysis that can adjust for imbalanced dataset designs. The advantage of adjusting for imbalanced datasets is to ensure a better learning and evaluation of phenomic prediction models especially for phenomic prediction models that estimate a feature of a genotype based on SNP and haplotype variations. Indeed, frequencies in individual SNP or haplotype variations are highly variable and often highly imbalanced between the minority and majority allele in the general population. Explicitly trying to balance the data, by focusing the sampling of subjects on the minority allele is tedious and not practical. However, it is a well-known fact that the learning of classifiers and/or regressions can be hampered due to imbalanced dataset designs as evidenced by He and Garcia in “Learning from imbalanced data (2009, IEEE transactions on knowledge and data engineering). Furthermore, many classifier performance measures are influenced by the imbalance in the dataset. For example, considering a two class classification problem with 10% of the dataset belonging to the minority class and the remaining 90% belonging to the majority class. An ingenuous classifier that simply returns membership into the majority class as output independent of the given phenotype, achieves an artificially high prediction accuracy of 90% when tested on the dataset but is useless when used as a phenomic prediction model. In preferred embodiments of the present invention performance measures such as an F-statistics in an Analysis of variance (ANOVA) taking the misbalancing into account or a classification performance measure that lowers the performance measure when misclassifying a minority class instance, such as, but not limited to a G-mean ROC measure, are preferably used. Finally, performing cross-validations on an imbalanced dataset can induce a dataset shift as illustrated by Moreno-Torres et al in “Study on the impact of partition-induced dataset shift on k-fold cross-validation” (2012, IEEE Transactions on neural networks and learning systems),

which results in incomparable training and test datasets. The result is that the performance of a classifier is wrongfully under estimated. Different approaches exist within machine learning to deal with the problem of unbalanced dataset designs, such as, but not limited to, up-sampling the minority class or under-sampling the majority class. An overview of these two and other, but not limited to, techniques dealing with imbalanced data is given by He and Garcia in "Learning from imbalanced data (2009, IEEE transactions on knowledge and data engineering). A possible, but not limited to, technique to deal with the danger of dataset shifts during cross-validation in imbalanced dataset designs may be the partitioning of the data taking the class labels into account, to ensure that the same frequencies of minority and majority class instances is present in each partition or fold of the complete dataset.

Exemplary embodiments of the present invention use a technique to up-sample the minority class in learning classifiers for features of a genotype at the level of individual SNPs as illustrated in Figure 24. Additional synthetic phenotypes of minority class samples are created through the combination of given phenotypes within the minority class within the phenotype-space until a desired minimum balance (number of minority class instances divided by the number of majority class instances) between the minority and a single or multiple majority classes is obtained. Exemplary embodiments of the present invention embed the up-sampling of the minority class within each of the 50 bootstrapped sampling runs of training data samples within a single cross-validation fold and this for each fold in the cross-validation as illustrated in Figure 19. The desired minimum balance is preferably set to 0.4 in the current implementation of the embodiments of the invention, but any other minimum balance can be used as long as this is not too low. It is advised not to set the minimum desired balance below 0.3. The advantage of embedding the up-sampling of the minority class into multiple runs of training datasets for a classifier and/or regression is to obtain robustness of the final phenomic prediction model against artificial variations induced by the up-sampling of the minority class. Further exemplary embodiments of the present invention take class labels into account to ensure comparability between training and test datasets created during k-fold data partitioning and cross-validation.

Although most of the research in machine learning is focusing on the algorithmic improvement of the classifiers and regressions themselves, little attention is being paid to the kind of features extracted. Image-based features typically comprises photometric (e.g. intensities, color values, gray values…), geometric (spatial relationships, curvature, area…) as well as contextual features (image changes, biomarkers…), where the latter are features extracted from the image in the context of additional non-image based information or from multiple sequences of images. The nature of features can range from local to global and low to high level features. Examples, but not limited to, of local and low level features from images may include pixel and/or voxel intensities and/or 3D point locations. Examples, but not limited to, of local and higher level features are features within individual pixels or voxels and landmarks that capture information from other spatially separated pixels or voxels and 3D points such as, but not

limited to, Scale-invariant feature transform (SIFT)- and SPIN-images, distances, directions (e.g. the surface normal in a 3D point on a surface image), angles, ratios, areas, local curvatures... Examples of global features include, but are not limited to, main variations or distribution summary statistics such as, but not limited to, principal components and average values respectively of features captured in a group of pixels, voxels and/or 3D points. However, it is well understood that better, in the sense of capturing more relevant and discriminative information for the classification and/or regression problem ad hand, features result in improved classification and/or regression performance. In preferred embodiments of the invention better features are extracted and used in phenomic prediction models. Better features can be extracted by either incorporating biological background knowledge into the model or using feature selection learning techniques. An example, but not limited to, of a more advanced classification/regression technique that includes feature learning, is deep learning, which aims to self-learn the optimal set of features to use in a specific machine learning task using advanced searching techniques. Examples, but not limited to, of incorporating biological background knowledge is to use results of scientific studies relevant to the phenotype to predict. In the example of classifying sex from facial morphology, it is of interest to focus the extraction of features that are known to differ between males and females. A range of existing publications on sexual dimorphism in facial morphology provide such information to define better features from. As an additional example in the classification of facial morphology into an individual SNP genotype, it is of interest to incorporate phenotypic facial characteristics or features seen and used to diagnose genetic disorders to which the gene in which the SNP is located is known to be associated. A final example is the discovery of image biomarkers, again by first studying the relationship between the phenotype and the feature of a genotype to predict at the level of the image. An example, but not limited to, is the use of BRIM to extract image biomarkers.

Exemplary embodiments of the present invention advantageously focus on classifiers and regressions based on image-biomarkers as strong features extracted using the methodology of BRIM as given in Claes et al in “Modeling 3D facial shape from DNA”. Generally, association studies on multipartite phenotypes such as but not limited to facial morphology e.g., have simplified the physical complexity a-priori, finding only a limited set of associated SNPs. More importantly these approaches lack the ability to transform their finding back into a biomarker or strong features for the particular SNPs discovered. For example, a SNP might be associated with nose width, but how nose width indicates the state of a genotype at the particular SNP is a different question. This embodiment of the present invention aims to expand the discovery of genetic associations with complex phenotypes without simplifying the physical complexity. BRIM is preferably used in this context such that accompanying biomarkers for significant associations can be revealed, modelled and used in phenomic prediction models. Individual SNP genotypes as predictor are preferably coded under their respective mode of inheritance and associated against phenotype-spaces as response. Exemplary embodiments of the present invention use a partial

least square regression (PLSR) embedded in a BRIM framework, which combines the multiple dimensions of a phenotype-space into a single direction that is customized to the predictor variable being modelled. Other relationship models or types of regression both linear as well as non-linear besides PLSR can be straightforwardly used as well. Within BRIM, the phenotype-spaces are used to refine and, in some cases transform the initial predictor variables. In other words and in contrast to other techniques, BRIM uses the response variables in a leave-N-out forced imputation setup to update the initial predictor variable values, creating a new type of variable: the response-based imputed predictor (RIP) variable (see figure 25). BRIM is to be used within a phenotype-space that enables the comparison of a phenotype in the context of other phenotypes by defining an appropriate measure of similarity. Exemplary embodiments of the present invention use the Mahalanobis distance between phenotypes in phenotype-space, but other measures are straightforwardly applicable as well. The process is bootstrapped (in the sense of improving itself in contrast to statistical bootstrap sampling) and RIP estimations over successive iterations can be monitored (Figure 26). The bootstrapping functions to correct observation error, misspecification of predictor values, and other sources of statistical confounding, such as an incorrect model of inheritance assumption and erroneous DNA sequencing e.g. Furthermore, the relationship between the RIP variable and the phenotype-space, allows to reveal the image biomarker (for an example of a morphological image-base biomarker see Figure 1) as a linear direction or a non-linear curved path through phenotype-space that accumulates an ensemble of image-based features that are optimally correlated with the underlying feature of a genotype. Finally, the discovered image biomarker is able to generate new RIP values for unseen or previously unavailable phenotypes as well as new and synthetically generated phenotypes in function of the associated feature of a genotype as illustrated in figure 27. Exemplary embodiments of the present invention use an underlying partial least squares regression model resulting in a linear direction through phenotype-space extracting the associated image biomarker.

In further embodiments of the present invention the image biomarker found with BRIM is used as a genetic feature-specific phenotypic quantitative trait or strong phenotype feature to infer agreement between a feature of a genotype and a new, unseen or synthetically generated phenotype under investigation using a phenomic prediction. The RIP value obtained from the test phenotype using the image biomarker is a univariate continuous variable and forms a bridge from the multivariate phenotype to the feature of a genotype. In further embodiments of the present invention RIP distributions of the training data samples are used to evaluate the RIP value of the phenotype under testing to see whether it matches with a given feature value of a genotype. The training RIP distributions are used to obtain a probability that the test phenotype matches the feature value of a given genotype (Figure 28). In the case of a categorical or group (sex, SNPs, Haplotypes…) feature value of a genotype the probability is a Bayesian a-posterior probability. In the case of a continuous (CNV, genetic background, age…) feature

value of a genotype the RIP distribution of the training samples is a weighted distribution based on the agreement of the feature values of a genotype of the training samples with the feature value of the given genotype. The result is that the distribution is shifted and centered on the feature value of the given genotype. The shifted RIP distribution generates a probability for the RIP value of a phenotype to be tested and classifies based on a probability whether the phenotype matches to the feature value of the given genotype.

In further embodiments of the invention the RIP distributions of phenotypes in test data partitions of the complete dataset are used to evaluate the performance of a phenomic prediction model. It is important to evaluate the performance of a phenomic prediction model on a test partition of the complete dataset that is independent from the training partition of the complete dataset that was used to learn the phenomic prediction model to correctly test the generalizability of the phenomic prediction model. It is important for the test partition of the complete data set to comprise of phenotypes for which the feature values of a genotype are known. In other words, phenotypes without genotype data or synthetically generated phenotypes cannot be used to evaluate the performance of a phenomic prediction model. In the case of a categorical feature of a genotype a univariate ANOVA that takes possible misbalancing into account is used to test if the RIP distributions of different categories of phenotypes in the test partition, defined by the category membership value in the related feature of a genotype, are significantly different and/or separated. Additionally in the case of a two-class (e.g. sex, recessive and/or dominant and/or over dominant individual SNP variation) categorical feature of a genotype, a two-class ROC analysis is performed on the RIP values of the phenotypes in the test partition and the G-mean accuracy measure (which properly deals with imbalanced dataset designs in contrast to the traditional accuracy score used in ROC analyses as described by He and Garcia in "Learning from imbalanced data (2009, IEEE transactions on knowledge and data engineering)) may be used as a binary classification performance measure. In exemplary embodiments of the present invention the G-mean measure is used in conjunction with the ANOVA analysis to evaluate the performance of a phenomic prediction model for a two-class feature of a genotype. The advantage is that the G-mean score is giving additional information on top of the ANOVA analysis in the ability of the phenomic prediction model to classify phenotype instances into their correct category. Therefore a more informative evaluation of a phenomic prediction model van be performed. In alternative embodiments of the present invention the G-mean performance measure from a two-class ROC analyses can be adapted to a G-mean performance measure for multi-class ROC analyses to obtain a G-mean performance measure in conjunction with an ANOVA measure for multi-class (e.g. additive/non additive SNP variation, Haplotypes…) features of a genotype. In the case of a continuous (e.g. genetic background, age, weight, height…) feature of a genotype or confounding factor a univariate correlation between RIP values and original feature values of a genotype for the phenotypes in the test partition is

used to evaluate the performance. The ANOVA F-statistic, G-mean measure as well as the correlation are tested as test-statistics for statistical significance under permutation. In the current implementation of the present invention 10000 permutation runs are performed. The advantage of permutation compared to traditional parametric statistics is its implicit ability to cope with imbalanced dataset designs and the reduced need for enforcing assumptions about the underlying null distribution.

In further embodiments of the present invention embed the forced imputation aspect from the BRIM procedure as a nested or inner k-fold partitioning within the outer k-fold partitioning to learn and or evaluate phenomic prediction models as illustrated in Figure 29. Both in the situation of learning a phenomic prediction model (outer k-fold was set with k = 10 to increase robustness against training data variability) as well as evaluating a phenomic prediction model (outer k-fold was set with k = 4 to create test partitions of the data for evaluation purposes), the current implementation uses an inner 10-fold cross-validation to obtain the leave-N-out forced imputations in BRIM, but other values for k besides 10 can be considered mainly as trade-off between computational burden, imputation precision and avoiding over fitting. The higher the number of inner folds the longer the computation time and the higher the imputation precision, but also the higher the danger of over fitting the imputation. The lower the number of inner folds the shorter the computation time but also the lower the imputation precision and the danger of over fitting. Using a 10-fold inner cross-validation obtains a good trade-off between computation time and imputation precision whilst avoiding over fitting and is therefore advised. BRIM is an iterative and bootstrapped (in the sense of improving itself in contrast to statistical bootstrapped sampling) procedure such that the inner k-fold partitioning and imputation is repeated over different iterations preferably using a different 10-fold partitioning of the data set within a single outer fold in each iteration as illustrated in Figure 29. The advantage of using different 10-fold partitioning's of the data in each iteration is to further increase the robustness against training and imputation data variability as well as over fitting. The bootstrapping iterations are stopped if subsequent imputed RIP values over different iterations converge, which is measured by computing the correlation between subsequent imputed RIP values over different iterations or if a maximum of iterations is achieved. In the current implementation the bootstrapping is stopped when the correlation between imputed values over different iterations is higher or equal to 0.98. The maximum number of iterations is user defined and is set to 5 in a preferred embodiment of the present invention. The maximum of 5 iterations was set based on the observation that convergence is most often achieved after already 3 iterations in the current implementation of embodiments of the present invention focusing on facial morphology. In further embodiments of the present invention the multiple bootstrap sampling of training datasets with or without replacement in combination with the up-sampling of minority class instances to deal with imbalanced dataset designs is performed each time within each inner fold and over subsequent BRIM bootstrapping iterations. This ensures that the RIP value imputation and any resulting image biomarker

and thus phenomic prediction model based on BRIM established image biomarkers increasingly gains robustness against the variability of underlying training data samples, avoids over fitting and can be learned from imbalanced dataset designs.

In reference to Figures 20 and 21, combining an ensemble of classifiers and/or regressions into a single classifier and/or regression, further embodiments of the present invention focusing on BRIM based directions through phenotype-space as image biomarkers, used in phenomic prediction models, combine multiple directions using PLSR through the phenotype-space from different bootstrapped training data samples with up-sampling of minority classes within each inner-fold into a single direction through the phenotype-space. This is done by taking the median value of the loadings of these directions along each of the dimensions in the phenotype-space. Additionally, the median values of the loadings are tested to be significantly different from zero using a Wilcoxon rank test and are set equal to zero in the case no statistical difference from zero is reported. This is done based on the observation that due to training data variability certain loadings fluctuate around zero and these fluctuations are noise. The result is a single and robustly estimated direction through the phenotype-space that constitutes the BRIM based image biomarker learned from the inner 10-fold cross validation. In reference to Figure 21, further embodiments of the present invention focusing on BRIM based directions through phenotype-space as image biomarkers in phenomic-predictions, the resulting robustly estimated directions through phenotype space from the inner 10-folds within each of the K outer folds are again combined into a single robustly estimated direction through phenotype-space in the exact same manner. The result is a final robustly defined image-biomarker for the phenomic-prediction model to be used in estimating a feature of a genotype for new previously unavailable and/or synthetically generated phenotypes.

Pattern of inheritance in individual SNP variations: In further preferred embodiments of the present invention if a single SNP variation is used as a feature of a genotype or as part of a feature of a genotype the appropriate mode of inheritance for the particular SNP variation is used as illustrated in Figure 11. In further preferred embodiments of the present invention the appropriate mode of inheritance is detected by evaluating the performance of phenomic prediction models under different mode of inheritance assumptions as illustrated in Figure 30. For each SNP variation used, seven test setups are created as such and phenomic prediction models are subsequently evaluated to determine the correct test setup. The first test is testing the recessive mode of inheritance, the second test is testing the dominant mode of inheritance, the third test is testing the over-dominant mode of inheritance, the fourth test is testing the additive mode of inheritance, the fifth test is testing the difference between the two homozygotes (AA = -1, BB = 1) without using the heterozygotes (AB), the sixth test is testing the difference between the first homozygotes and the heterozygotes (AA = -1, AB = 1) without using the second homozygotes (BB), the seventh test is testing the difference between the heterozygotes and the

second homozygotes (AB = -1, BB = 1) without using the first homozygotes (AA). The last three tests setups (5-7) are performed to give additional confidence and information in detecting the correct mode of inheritance. Each test generates a binary outcome, 0 if the test was not significant or successful, 1 if the test was significant or successful. The combination of all seven binary outcomes generates a 7 bit code and depending on the 7 bit code the mode of inheritance can be detected. [1 0 0 0 1 0 1] = recessive mode of inheritance, [0 1 0 0 1 1 0] = dominant mode of inheritance, [0 0 1 0 0 1 1] = over dominant mode of inheritance, [1 1 0 1 1 1 1] = additive mode of inheritance, [1 1 1 0/1 1 1 1] = non additive mode of inheritance. These constitute the set of trivial 7 bit codes that lead to a clear decision on the mode of inheritance. In other cases of 7 bit code combinations, the mode of inheritance is not trivially detected, but through the analysis of the 7 bit code in combination with the p-values of each test setup more educated decisions can be made. Additionally, depending on the analysis of the 7 bit code one can opt to deploy and learn a phenomic prediction model using either one of the 5-7 test setups instead of one of the modes of inheritance (tests setups 1-4), whilst ignoring one out of the three SNP genotype groups. It is for example possible, but not limited to, that certain SNP genotype groups are underrepresented in the complete dataset that is available, such that there is a lack of statistical power to detect their true differences at the level of the phenotype in any type of comparison against the other SNP genotype groups. In these cases it is preferred to ignore this particular SNP genotype group.

In further preferred embodiments of the present invention the testing of a specific mode of inheritance that groups two out of the three SNP genotype groups (AA, AB or BB) into a single group (test setups 1-3) deals with imbalanced dataset designs when pulling the two groups together as illustrated in Figure 31. Example modes of inheritance that pull two groups together are the recessive mode of inheritance (AA pulled together with AB, test setup 1), the dominant mode of inheritance (AB pulled together with BB, test setup 2), and the over-dominant mode of inheritance (AA pulled together with BB, test setup 3). If two groups are pulled together when they are imbalanced (as an extreme example: number of instances in group 1 = 10 and number of instances in group 2 = 1000) testing against the third group is dominated by the difference of the majority group (group 2 in the example) in the pulled group with the third group and leads to false interpretations of the test setup. Basically the first group is ignored and the test setup pulling the two groups together reduces to one of the 5-7 test setups, where only two groups are compared and no two groups are pulled together. Exemplary embodiments of the present invention perform an up sampling of the minority group within the pulled group to correct for the imbalance. Additional synthetic phenotypes of minority class samples are created through the combination of given phenotypes within the minority class within the phenotype-space until a desired minimum balance (number of minority class instances divided by the number of majority class instances) between the minority and a single or multiple majority classes is obtained. Exemplary

embodiments of the present invention embed the up sampling of the minority class within each of the 50 bootstrapped samples of training data samples within each inner fold of the BRIM imputation. This is exactly the same embedding of the up sampling of the minority group as described earlier and again comes down to an up sampling each time a single PLSR is computed. The desired minimum balance is set to 0.4 in the current implementation of the embodiments of the invention, but any other minimum balance can be used as long as this is not too low. It is advised not to set the minimum desired balance below 0.3. In further preferred embodiments of the present invention take the three group class labels into account to ensure comparability between training and test datasets created during k-fold data partitioning and cross-validation. This ensures that in despite that two groups are pulled together the frequencies of all three groups separately are consistent across training and test partitions of the dataset.

In further preferred embodiments of the present invention the detection of the mode of inheritance is performed in combination with the selection of relevant and associated features of a genotype in case the latter use individual SNP variation as feature of a genotype or as part of a feature of a genotype. In other words the selection of the kind of features of a genotype as associated or not is performed by running all 7 test setups using the appropriate outer k-fold cross-validation settings to evaluate phenomic-prediction models through an integration of Figure 30 into Figure 23. Exemplary embodiments of the present invention as such use a 4-fold outer fold cross-validation as described earlier in each of the 7 test setups. In further preferred embodiments of the present invention features of a genotype using individual SNP variation are not considered to be relevant or associated if the resulting 7 bit code equals to [0 0 0 0 0 0 0] or if in the analysis of non-trivial 7 bit codes more educated decisions were not possible. In all other cases the feature of a genotype using individual SNP variation is selected to be relevant and associated.

In further preferred embodiments of the present invention features of a genotype using individual SNP variation that have been selected to be associated use the appropriate mode of inheritance or alternatively one of the grouping designs as defined in test setups 5-7, based on the evaluation of the 7 bit code, when learning a phenomic prediction model of these features of a genotype to be used in the prediction of a phenotype from genotype. In other words, during the selection of associated features of a genotype based on individual SNP variations, phenomic prediction models are evaluated within each of the 7 test setups using an outer 4-fold cross-validation. In contrast the phenomic prediction model of features of genotype data based on individual SNP variations that is used in the matching of new previously unavailable or synthetic phenotypes against the feature of a genotype is learned in a subsequent stage using a single grouping design of the three SNP genotype groups from test setup 1-7 and learned with an outer 10-fold cross-validation as previously described.

Integrating a plurality of phenomic prediction models: In preferred embodiments of the present invention a plurality of phenomic prediction models are integrated to obtain a phenotype fit to a plurality of feature values of a given genotype. In further preferred embodiments of the present invention a phenotype fit is expressed using an overall matching and/or error score through the combination of a plurality of phenomic prediction estimates for a plurality of different features of a genotype from a given new previously unavailable or synthetically generated phenotype in a phenotype-space as illustrated in Figure 32. Each individual phenomic prediction model within the plurality of phenomic prediction models generates an estimate for a single feature of a genotype from the known phenotype under investigation that can be used to determine how well the phenotype matches a feature value of a given genotype, either through a categorical classification result, or a continuous confidence or probability as described earlier as illustrated in Figure 22. Exemplary embodiments of the present invention use a probability as a matching score. The higher the matching score, the better the phenotype fits the associated feature value of a given genotype, the lower the matching score the worst the phenotype fits the associated feature value of a given genotype. In further exemplary embodiments of the present invention the matching score is converted into an error score by taking the negative logarithm of the probability. The higher the error score, the lower the matching score. The lower the error score, the higher the matching score. In further exemplary embodiments of the present invention the individual error scores for a plurality of different features of a genotype are combined into a single overall average error score. Alternative embodiments of the present invention can combine the error scores in other ways. For example, but not limited to, the individual error scores can be summed instead of averaged. The advantage of averaging instead of summing is that overall error-score ranges are standardized and independent of the amount of features of a genotype used to establish the overall error score.

Alternative embodiments of the present invention can also combine the individual error scores in a weighted manner. For example, but not limited to, an overall error score can be obtained through a weighted average of the individual error scores by assigning different weights to the individual error scores. The advantage of giving different weights to individual error scores is that the importance of the associated features of a genotype can be controlled and adapted if needed. For example, but not limited to, a higher or lower weight can be assigned to an individual error score based on the performance evaluation of the phenomic prediction model that produces the individual error score. As such better performing phenomic prediction models can gain importance in the overall error score. As another example, but not limited to, a higher or lower weight can be given to an individual error score in function of the associated feature value of a given genotype. As such a higher or lower importance can be given to a feature of a genotype if the corresponding given feature value of a given genotype is rare (low in frequency within a population) or common (high in frequency within a population). Alternative

embodiments of the present invention can also learn the optimal weights using machine learning techniques such as, but not limited to, neural networks. This can provide an alternative means to further detect associated false positives that still remain after carefully selecting relevant features of a genotype. If the outcomes of these weight learning techniques, result in setting the optimal weights for a single or a subset of error scores equal to zero, this probably implies that the associated phenomic prediction model was modelling irrelevant noise and can thus be dropped from the plurality of phenomic prediction models. Further alternative embodiments of the present invention can adapt or change the weighting to evaluate known phenotypes multiple times with a different emphasis on different features of a genotype and over multiple iterations in an optimization algorithm.

Further alternative embodiments of the present invention can also combine the single overall error score of a phenotype in function of a genotype with additional error scores computed from the phenotype but not in function of the genotype. For example, but not limited to, the typicality of a phenotype as measured and defined within the phenotype space can be added to the error score matching the phenotype to a given genotype. The advantage of doing this is the ability to favour phenotypes that are more typical compared to other phenotypes in the phenotype-space. The current implementation of the embodiments of the present invention allows to incorporate such a typicality error for phenotypes using the Mahalanobis distance to the origin of the phenotype-space. This can be useful to avoid predicting very atypical phenotypes from a given genotype in the optimization algorithm and serves as a regularization during optimization. However, when using the preferred and exemplary embodiments of the present invention to establish errors scores matching known phenotypes to genotypes from probabilities determined using RIP distributions, no such regularization during optimization is required as observed from experimental results focusing on the prediction of facial morphology from genotype.

Predicting a phenotype from genotype: In preferred embodiments of the present invention the overall error score is preferably used to evaluate a known phenotype against a given genotype as a solution in an optimization algorithm or as a possible match in biometric identification and/or verification applications as illustrated in Figure 32. The overall error score can be defined from a single, all or a subset of features of a genotype. Exemplary embodiments of the present invention use all found associated features of a genotype. In further preferred embodiments of the present invention the overall error score is used to evaluate and rank a plurality of known phenotypes against a given genotype from lowest to highest overall error score, either as multiple solutions in an optimization algorithm as illustrated in Figure 33 or as available phenotypes in a database to be ranked according the degree of matching to the given genotype as illustrated in Figure 34.

In further preferred embodiments of the present invention an optimization algorithm that lowers and/or minimizes the overall error score or increases and/or maximizes the overall matching score of a single or a plurality of phenotypes through the creation of new and/or altered phenotypes along the dimensions of the phenotype-space is used to predict the phenotype from genotype. An example, but not limited to, sequential optimization technique is implemented in an exemplary embodiment of the present invention as illustrated in Figure 35. Starting with a selected, known phenotype from the phenotype-space as an initial and current solution, the solution is evaluated against a single feature value of a given genotype and updated to improve its match to the given feature value of the given genotype. The updated solution then becomes the current solution. Exemplary embodiments of the example within the present invention update the solution along the direction associated with the extracted image biomarker through phenotype-space. Subsequently, the current solution is evaluated against another feature value of the given genotype and is updated again. The updating of the current solution is performed sequentially for a plurality of different features of a genotype. Alternative embodiments of this example can run the sequential updating multiple times using the plurality of features of a genotype with the same or different sequence of updating.

In preferred embodiments of the present invention the optimization algorithm is a global optimization algorithm that can cope with complex and non-linear optimization problems. In further and alternative embodiments of the present invention the outcome of the global optimization serves as an initial input for a subsequent local search optimization. An example, but not limited to, local search optimization based on the overall error scores from the evaluation of phenotypes as solutions to the optimization problem are a simplex and/or a Powell-optimization. It is an advantage to first perform a global optimization before a local search optimization, because the latter are easily stuck in local optima of the overall error score to minimize or the overall matching score to maximize.

In further preferred embodiments of the present invention the global optimization algorithm is an artificial intelligence optimization algorithm as illustrated in Figure 36). The advantage of artificial intelligence algorithms is their ability for self-learning during optimization and as such to cope with complex and non-linear optimization algorithms. In further preferred embodiments of the present invention the artificial optimization algorithm is an evolutionary optimization algorithm. The advantage of an evolutionary optimization algorithm is its ability to optimally use the phenotype-space to generate new synthetic phenotypes based on the recombination of phenotypes in a population as possible solutions to the optimization problem. As such an evolutionary algorithm is able to efficiently scan all possible phenotypic variations within the multidimensional phenotype space, which otherwise can become challenging for other optimization techniques especially when the dimensions of the phenotype-space are very high. An evolutionary algorithm uses nature-inspired heuristics, like fitness, recombination, and mutation, to evolve a population of candidate solutions to the next level. Exemplary

embodiments of the present invention use the given genotype for which the unknown phenotype needs to be predicted, to screen a population of phenotypes as possible prediction solutions based on an overall error score from a plurality of phenomic prediction models. This creates a DNA-based fitness landscape and the evolutionary algorithm allows the population of phenotypes to evolve and move through this landscape, in subsequent cycles, in search of the best fitting phenotypes as prediction results. Thus an additional advantage of using an evolutionary optimization algorithm is the generation of a whole population of predicted phenotypes that match the given genotype to similar levels of agreement. From this population of resulting phenotypes a single random and/or best, but not limited to, phenotype can be selected as the final prediction result or a combination of the resulting phenotypes can be created as the final prediction outcome. Examples of a combination of resulting phenotypes in the final population include, but are not limited to, an average phenotype or a weighted average phenotype where the weights are defined in function of the fitness of each of the resulting phenotypes in the population.

Important to evolutionary algorithms is the definition of the fitness score and the way current candidate phenotypes are recombined and mutated. In preferred embodiments of the present invention the overall error/matching score is used as fitness score to minimize/maximize in the evolutionary algorithm. It is important to realize that recombination and mutation are not performed on the level of the genotype, but on the level of the phenotype, which again as before endorses the reasoning from phenotype to genotype in the present invention. An important distinction between the embodiments of the present invention and current phenotype predictions from genotype, including genomic-predictions, is that for each new phenotype from genotype prediction to be made the evolutionary algorithm starts over and performs new self-learning cycles. In further contrast, an evolutionary algorithm, which belongs to soft computing, is tolerant of imprecision and partial truth, which is certainly welcome in this complex puzzle of genotype to phenotype correlations. In other words, the lack of knowledge in associated genes and gene/environment interactions going from genotype to phenotype today is compensated for by this advanced AI-based optimization technique when combined with the preferred or alternative embodiments of the present invention using of a plurality of phenomic prediction models. In the case that only partial and not the complete knowledge of the genetic-architecture of a complex phenotype is known this partial knowledge can still be efficiently captured within a plurality of phenomic prediction models and prediction results can be generated within the precision boundaries based on the partial knowledge. Furthermore, the precision boundaries of the prediction result can be made explicit by measuring the variation left in the final population of phenotypes generated by the evolutionary algorithm after convergence. Small increases in knowledge of the genetic architecture of the complex phenotype, for example, but not limited to, the discovery and incorporation of additionally associated SNPs with variations in the phenotype, can straightforwardly be incorporated as additional

phenomic prediction models and will increase the phenotype from genotype prediction accuracies accordingly. In contrast, traditional genomic-prediction models are hampered by not knowing the complete genetic architecture and large increases in knowledge of the genetic architecture are anticipated to be needed to increase phenotype from genotype prediction accuracies in such techniques.

Exemplary embodiments of the present invention focusing on facial morphology use a population size of 2000 known phenotypes within the evolutionary algorithm as illustrated in Figures 37 - 40. The population is evolved until convergence is reached or a maximum number of iterations is achieved. Convergence is detected by monitoring the average error score and/or the best error score of the entire population. In the current implementation of the embodiments of the present invention the maximum number of iterations is set equal to 100, which seems to be more than enough to achieve convergence. The population is initialized using 2000 normally distributed random perturbations of the consensus phenotype in the phenotype-space based on the standard deviations along each of the dimensions in the phenotype-space. Alternative embodiments of the invention can straightforwardly use other population initialization methods. Examples include, but are not limited to, uniformly distributed random permutations of the consensus phenotype in the phenotype-space along each of the dimensions of the phenotype or randomly selected existing phenotypes in a database embedded within the phenotype-space. Error scores for the population of phenotypes are used to rank the population from best to worst matching phenotypes and the best set of phenotypes is selected to produce the next generation of the population, much like the survival of the fittest. An expert in the field of evolutionary algorithms is well aware of all the possibilities to select the best candidates and to generate new candidates from these for the next iterations. Exemplary embodiments of the present invention have implemented a distribution based next population generation approach, but for the invention to work any other approach can be straightforwardly applied. In the implementation of the exemplary embodiments of the invention a fraction of the existing population is selected according to a reduction parameter that is set by the user. Currently the reduction parameter is set to 0.5, which means that in each iteration the 1000 best out the 2000 known phenotypes in the population are selected to create the next generation. The 1000 best phenotypes are first used to estimate the bandwidth of the kernel smoothing window needed for the establishment of probability density estimates, based on a normal kernel function, along each of the dimensions of the phenotype-space. This is essentially capturing the variation along each of the dimensions of the phenotype-space observed within the best 1000 known phenotypes. In other words, it captures the subspace in the phenotype-space that is spanned by the best 1000 known phenotypes. Then 1000 new, synthetic phenotypes, to augment the population size back to a total number of 2000, are created by random perturbations in each of the directions of the phenotype space in function of the estimated bandwidths for each dimension separately on 1000

randomly selected phenotypes with replacement from the 1000 best phenotypes. The result is 1000 new known phenotypes, different from the best 1000 known phenotypes, but part of the subspace spanned by the best 1000 known phenotypes. As noted before this is just one way of generating subsequent generations within an evolutionary algorithm, but from experiments with the methods of the present invention, we see that this approach is computationally interesting as in fast and easy to compute, with a good and fast convergence behaviour, without easily getting stuck in local optima and therefore generating consistent prediction results when the optimization algorithm is ran multiple times for the same given genotype. After optimization convergence exemplary embodiments of the present invention generate a single prediction phenotype result from genotype as a weighted average of the final surviving population of 2000 phenotypes. The weights are defined in function of the error scores of each of the phenotypes in the final population, giving a higher weight to phenotypes with a lower error score.

In a particular embodiment, the methods of the invention involve the ranking of multiple known phenotypes (i.e. known from the outset or synthetically generated) and predicting the unknown phenotype for the given genotype to be one of the highest ranked known or optionally synthetically generated phenotypes, or a synthetic phenotype derived therefrom. The invention also provides a databases of phenotypes, wherein the phenotypes have been ranked for a given genotype according to the methods described herein.

As a non-limiting example, the methods of the invention may be applied on a database containing 2D images, optionally containing additional information such as sex and age. For example, the methods of the invention may be applied on government databases (e.g. containing images from driver's licences, identity cards…) or commercial databases (e.g. social media databases). The methods of the invention may be applied to identify the most likely or most similar 2D image for a given genotype (e.g. obtained from a blood sample). In said instance, the methods of the invention may be used to rank the images in the database and predict the unknown phenotype for the given genotype to be the highest ranking image or a set of highest ranking images. The methods of the invention may as well comprise generating a synthetic phenotype from the highest ranking images and predict the unknown phenotype to the said synthetic phenotype.

Therefore, in a preferred embodiment, the unknown phenotype is predicted to be a synthetic phenotype derived from the highest ranking phenotypes. Particularly preferred is a synthetic phenotype that is a weighted average of the highest ranking phenotypes.

Alternatively, in the sequential adaptation methods, the methods of the invention may involve predicting the unknown phenotype for the given genotype to be the adapted phenotype.

Example

***Integrated illustration of the present invention on an example facial phenotype prediction from genotype*:** Embodiments of the present invention are illustrated focusing on the prediction of human facial morphology, captured using 3D images as illustrated in Figure 14, from a given genotype. A known genotype is given based on the extraction and sequencing of DNA from a voluntarily given saliva sample of a European Male. Additional factor values for age (39 years), weight (82 kg) and height (1.82 m) are provided as well. Sex (male) is extracted from the given DNA, as well as 81 individual SNP genotypes and 5 values along 5 different axis of genetic background across Europe. For the evaluation of the prediction outcomes of the unknown facial phenotype, a 3D image of the true facial phenotype belonging to the given genotype is illustrated in Figure 41 (D).

In a first instance the unknown facial phenotype from the given genotype is predicted using embodiments of the present invention involving:

- *providing a database of known phenotypes*: 2000 synthetic known facial phenotypes are randomly created from the consensus face in a phenotype-space for 3D facial morphology (face-space), spanning 78 dimensions, using a normal distributed random generator that takes the standard deviations along each of the 78 dimensions in the face-space into account. Alternatively, 2000 existing (non-synthetic) known facial phenotypes in a database that are embedded into and scattered across the face-space can be used. The result is having 2000 synthetic known facial phenotypes that are distributed across the face-space and that are represented as 78 dimensional points from which a complete 3D facial image can be reconstructed and displayed, as illustrated in Figure 8 and 13, and this for each of the 2000 synthetic known facial phenotypes.

- *providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype*: 81 phenomic prediction models each predicting a different individual SNP feature of a genotype from the 2000 known facial phenotypes; 5 phenomic prediction models each predict a value for a different axis of genetic background feature of a genotype from the 2000 known facial phenotypes; one phenomic prediction model predicts sex as a feature of a genotype from the 2000 known facial phenotypes; one phenomic prediction model predicts age from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm; one phenomic prediction model predicts (body) weight from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm; and one phenomic prediction model predicts (body) height from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm. Thus in total 90 phenomic prediction models are provided.

- *integrating the plurality of phenomic prediction models in at least one optimization algorithm*: The 2000 known facial phenotypes are used as a population of 2000 possible solutions in an evolutionary algorithm, as illustrated in Figure 6 and 36, where the 90 phenomic prediction models are used to generate a fitness value for each of the 2000 solutions in the population that expresses how well each solution matches all the features of a given genotype. This creates a fitness landscape as defined in evolutionary algorithms, through which the population of 2000 solutions can evolve providing better solutions in the sense of solutions that better match the 90 features of a given genotype.

- *applying the at least one optimization algorithm comprising:*

  - *calculating an error score by comparing features of the genotype predicted by the phenomic prediction models based on known phenotypes from the database with the given genotype:* Each phenomic prediction model estimates a feature of a genotype from each known phenotype in the population as illustrated in Figure 16. The estimated feature of a genotype from a known phenotype is compared against the corresponding feature value of the given genotype using a probability that expresses the similarity between the estimated and given feature of a genotype. The higher the probability the higher the similarity. This probability in turn is transformed into an error score for each feature of a genotype by taking the negative logarithm of the probability as illustrated in Figure 22. Subsequently, the error scores over all 90 features of a genotype are combined into a single error score by taking the average of the error scores for each feature of a genotype, as illustrated in Figure 32, and this is done for each of the 2000 known phenotypes in the population. The result is a single error score for each of the 2000 known phenotypes in the population that expresses its match to the given genotype. The lower the error score the better the known phenotype matches the given genotype and the higher its fitness value in the evolutionary algorithm.

  - *ranking known phenotypes in the database according to the error score*: The 2000 known phenotypes in the population are sorted in order of lowest to highest error score, as illustrated in Figure 33.

  - *selecting a subset of known phenotypes based on said ranking*; The first 1000 known phenotypes in the sorted population are selected as the best matching known phenotypes to the given genotype. The last 1000 known phenotypes in the sorted populations are removed from the population. The result is a population of size 1000 known phenotypes, containing only the 1000 best matching known phenotypes.

  - *optionally creating synthetic phenotypes based on the known phenotypes in the subset and adding the synthetic phenotypes to the subset*: The distribution of the 1000 selected best

matching known phenotypes along the 78 dimensions of the face-space is used to create 1000 new synthetic phenotypes by randomly sampling 78 dimensional points from this distribution. Subsequently, the 1000 new and known synthetic phenotypes are added to the 1000 previously selected best matching known phenotypes such that the population size is again 2000. The distribution of the 1000 selected best matching known phenotypes captures the subspace that is spanned by the selected best matching known phenotypes within the complete phenotype-space. The newly created known phenotypes belong to the same subspace but are different to the 1000 selected best matching known phenotypes.

- *iterating the steps of the optimization algorithm at least one time on the subset of known and optionally synthetic phenotypes*: The newly created population of 2000 known phenotypes is an updated and improved population of solutions in function of improved matching to the given genotype. Running multiple iterations will keep on updating and improving the population of 2000 solutions, to an extent that all 2000 known phenotypes better match the given genotype. The results is also that the diversity of the 2000 known phenotypes within the population reduces as illustrated over different iterations in Figures 37-40 and this until a maximum of 100 iterations has been reached.

- *predicting the unknown phenotype for the given genotype to be one of the highest ranked known or optionally synthetic phenotypes, or to be a synthetic phenotype derived therefrom:* A single prediction of the unknown phenotype for the given genotype is generated as a weighted average of the final surviving population of 2000 known phenotypes after 100 iterations. The weights are defined in function of the error scores of each of the known phenotypes in the final population, giving a higher weight to known phenotypes with a lower error score. The final facial prediction outcome for the test given genotype is illustrated in Figure 41 (C).

In a second instance the unknown facial phenotype from the given genotype is predicted using embodiments of the present invention involving:

- *providing a database comprising at least one known phenotype*: The consensus face of the phenotype-space for 3D facial morphology (face-space), spanning 78 dimensions is provided as a single known facial phenotype.

- *providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype:* 81 phenomic prediction models each predicting a different individual SNP feature of a genotype from the 2000 known facial phenotypes; 5 phenomic prediction models each predict a value for a different axis of genetic background feature of a genotype from the 2000 known facial phenotypes; one phenomic prediction model predicts sex as a

feature of a genotype from the 2000 known facial phenotypes; one phenomic prediction model predicts age from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm; one phenomic prediction model predicts (body) weight from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm; and one phenomic prediction model predicts (body) height from the 2000 known facial phenotype as an additional factor that can be controlled in the algorithm. Thus in total 90 phenomic prediction models are provided.

- *integrating the plurality of phenomic prediction models in at least one optimization algorithm*: Each of the 90 phenomic prediction models is used one by one to sequentially update the single known facial phenotype in function of all 90 features of the given genotype as illustrated in Figure 35.

- *applying the at least one optimization algorithm comprising:*

- *calculating a first error score by comparing a first feature of the genotype predicted by at least a first phenomic prediction model based on at least one known phenotype from the database with the given genotype:* A single first selected phenomic prediction model estimates a single feature of a genotype from the known phenotype as illustrated in Figure 16. The estimated feature of a genotype from the known phenotype is compared against the corresponding feature value of the given genotype using a probability that expresses the similarity between the estimated and given feature of a genotype. The higher the probability the higher the similarity. This probability in turn is transformed into an error score for each feature of a genotype by taking the negative logarithm of the probability as illustrated in Figure 22.

- *Calculating a second error score by comparing a second feature of the genotype predicted by at least a second phenomic prediction model based on at least one known phenotype from the database with the given genotype;* Another single selected phenomic prediction model, different to the first selected phenomic prediction model estimates another single feature of a genotype from the known phenotype as illustrated in Figure 16. The estimated feature of a genotype from the known phenotype is compared against the corresponding feature value of the given genotype using a probability that expresses the similarity between the estimated and given feature of a genotype. The higher the probability the higher the similarity. This probability in turn is transformed into an error score for each feature of a genotype by taking the negative logarithm of the probability as illustrated in Figure 22. This is repeated until all 90 phenomic prediction models have been selected once.

- *adapting at least one known phenotype from the database based on the first and second error score to obtain a database comprising at least one adapted phenotype;* After calculating the first error score the known phenotype is adapted to lower this error score, or alternative until the

highest probability is achieved using a simplex optimization algorithm. The updated known phenotype is then used to calculate the second error score, and the known phenotype is updated again minimizing the second error score this time. This is repeated until all 90 error scores have been minimized in sequential order.

- *Optionally iterating the steps of the optimization algorithm at least one time on the database comprising the at least one adapted phenotype;* The whole process is repeated 10 times over all 90 phenomic prediction models, where in each repeat the order of phenomic prediction models is randomized.

- *predicting the unknown phenotype for the given genotype to be the adapted phenotype.* After 10 repeats over all 90 phenomic prediction models, the initial known phenotype given as the consensus face of the face-space, has been updated 900 times. The last updated known phenotype is given as the prediction for the unknown phenotype. The final facial prediction outcome for the test given genotype is illustrated in Figure 41 (B).

In a third and last instance, the unknown facial phenotype is predicted as a comparative example using the technique described in Claes *et al* in "Toward DNA-based facial composites: Preliminary results and validation." (Forensic Science International: Genetics, November 2014, 13: p.208-16). The implementation and deployment of this technique started from the exact same databases available and used for the embodiments of the present invention, enabling an objective and fair comparison of prediction outcomes. In other words, the difference is that the technique as described in the prior art does not use the technical embodiments in the present invention, but starts from the exact same datasets, features of a genotype and additional factors (thus 90 in total). At no time, during the execution of the technique a single or multiple features of a genotype are estimated from a single or multiple known facial phenotype. The technique first constructs a base-face, using the information of sex, age, weight, height and the 5 values of the 5 different axis of genetic background. Subsequently, the effects of the 81 SNPs are revealed using the technique of BRIM on the database of 2006 known phenotypes and are then overlaid onto the base-face using a heuristically weighted photomontage approach forming the 'predicted-face'. The prediction outcome for the test given genotype from this comparative example is illustrated in Figure 41 (A).

In order of goodness, it is observed that the first prediction instance as illustrated in Figure 41 (C) using embodiments of the present invention in an evolutionary optimization algorithm captures relevant and characteristic facial features in correspondence to the true facial phenotype as illustrated in Figure 41 (D). The second prediction instance as illustrated in Figure 41 (B) using embodiments of the present invention in a sequential optimization algorithm is still able to capture relevant facial characteristics but not as good as the first prediction instance. The third and last prediction instance as illustrated in Figure

41 (A) using the technique as comparative example clearly misses most of the specific facial characteristics and is generating an average looking European male prediction outcome.

Further embodiments of the present invention that are used in the first two prediction instances as illustrated in Figure 41 (B) and (C) involve:

Known and unknown phenotypes are embedded in a high-dimensional phenotype-space: The phenotype-space for facial morphology is constructed using each coordinate of 7150 3D points, which are consistently sampled from 2006 3D facial images in an existing database of 3D facial images, as a single dimension. This results in a 21450 (= 3 x 7150) dimensional face-space. In such a face-space, a facial phenotype is embedded or represented as a single 21450 dimensional point. In other words, from this embedding the original 7150 3D quasi-landmarks and therefore the original 3D facial image can be reconstructed and displayed. Subsequently, principal component analysis (PCA) is employed to reduce the number of dimensions from 21450 to 78 by selecting the first 78 principal components, which together capture 98.5% of the total variance observed in all 7150 3D points over all 2006 facial images. Such a dimension reduction is not required but results in a lower computational burden in exemplary embodiments of the present invention, and is still able to reconstruct and display 7150 3D points and therefore 3D facial images from 78 dimensional single points in the face-space. Although these reconstructions are not exact compared to the face-space prior to PCA reduction, they approximate 3D facial images with sub-millimeter (<1mm) accuracy for each of the 7150 3D points over all 2006 3D facial images in the database.

The high-dimensional phenotype-space is ordinated following genetically and/or environmentally meaningful dimensions: In the example of facial morphology, the resistance against facial asymmetry during facial development is known to be genetically controlled. However, the patterns of fluctuating asymmetry are generally considered to be developmental noise due to influences of the environment during development. Through the construction of symmetrized phenotypes as done for faces in Claes *et al*. in "Modeling 3D facial shape from DNA." (Plos Genetics, 2014), resulting phenotype-spaces are not influenced by environmental noise captured in facial asymmetries as illustrated in Figure 15.

The evaluation of a particular phenomic prediction model to identify if the association is relevant between a) the feature of a genotype predicted by the particular phenomic prediction model, and b) the known phenotype upon which the prediction is based: The 81 individual SNP features of a genotype, were selected to be relevant and associated to the facial phenotype from 12000 different SNP features available in our database of 2006 facial phenotypes and genotypes, through the evaluation of the corresponding phenomic prediction models under 7 different test-setups to determine their respective mode of inheritance. The 5 features of genetic background of a genotype, were selected to be relevant and associated to the facial phenotype from 100 different genetic background features of a genotype

available in our database of 2006 facial phenotypes and genotypes, through the evaluation of the corresponding phenomic prediction models. Sex as a feature of a genotype was selected to be relevant and associated to the facial phenotype through the evaluation of the corresponding phenomic prediction model. Age, (body) weight and (body) height, were selected to be relevant and associated to the facial phenotype through the evaluation of the corresponding phenomic prediction models.

Classification from multi-dimensional phenotype-space into associated feature of a genotype through the reduction of the phenotype-space using imaging biomarkers: a phenomic prediction model comprises of an image-based classification and or regression on a feature of a genotype as illustrated in Figure 16. Any classification and/or regression starts from features extracted from the images and the stronger the features the better the performance of the classification and/or regression. In the first two prediction instances of facial morphology, the phenomic predictions are based on image-biomarkers as strong features extracted using the methodology of BRIM as given in Claes et al in "Modeling 3D facial shape from DNA". BRIM reveals the image-biomarker as a reduced linear direction or a non-linear curved path through phenotype-space that accumulates an ensemble of image-based features that are optimally correlated with the underlying feature of a genotype. The image-biomarker is able to generate RIP (response-based imputed predictor) values for known, unseen or previously unavailable phenotypes as well as new and synthetically generated phenotypes in function of the associated feature of a genotype as illustrated in figure 27. These are used to generate the probabilities expressing the match of a known phenotype with the feature of a given genotype as illustrated in Figure 28. These probabilities are in turn changed in error scores by taking the negative logarithm of the probabilities. Note that the aspect of BRIM to generate a RIP value from a known phenotype is not used in the technique as described in the comparative example used to generate the third and last prediction instance as illustrated in figure 41 (A). In other words, an image-biomarker resulting from BRIM, is not used in this technique for classification and/or regression purposes in general as well as classification and/or regression of a feature of a genotype from a known phenotype in particular.

It is to be understood that this invention is not limited to the particular features of the means and/or the process steps of the methods described as such means and methods may vary. It is also to be understood that the terminology used herein is for purposes of describing particular embodiments only, and is not intended to be limiting. It is also to be understood that plural forms include singular and/or plural referents unless the context clearly dictates otherwise. It is moreover to be understood that, in case parameter ranges are given which are delimited by numeric values, the ranges are deemed to include these limitation values.

The above description is for the purpose of teaching the person of ordinary skill in the art how to practice the present invention, and it is not intended to detail all those obvious modifications and

variations of it which will become apparent to the skilled worker upon reading the description. It is intended, however, that all such obvious modifications and variations be included within the scope of the present invention, which is defined by the following claims. The claims are intended to cover the claimed components and steps in any sequence which is effective to meet the objectives there intended, unless the context specifically indicates the contrary.

Claims

1. A method for predicting an unknown phenotype from a given genotype, the method comprising:

- providing a plurality of phenomic prediction models, said phenomic prediction models adapted to predict a feature of a genotype based on a known phenotype;

- integrating the plurality of phenomic prediction models in at least one optimization algorithm; and

- applying the at least one optimization algorithm to predict an unknown phenotype from a given genotype.

2. The method according to claim 1, further comprising generating an error score used by the at least one optimization algorithm, whereby said error score is adapted to evaluate said known phenotype for the given genotype.

3. The method according to claim 1, further comprising generating error scores, whereby said error scores are adapted to evaluate and rank multiple known phenotypes for the given genotype.

4. The method according to claim 3, wherein multiple known phenotypes are represented in an image database wherein the multiple known phenotypes are evaluated and ranked for the given genotype.

5. The method according to claim 1, further comprising the evaluation of a particular phenomic prediction model to identify if the association is relevant between a) the feature of a genotype predicted by the particular phenomic prediction model, and b) the known phenotype upon which the prediction is based.

6. The method according to claim 5, whereby the evaluation of the particular phenomic prediction model is used to find a relevant genetic variation in a genome-wide association study.

7. The method according to any of claims 1 to 6, whereby said known and unknown phenotypes are embedded in a multidimensional phenotype-space.

8. The method according to claim 7, whereby the multidimensional phenotype-space is ordinated following genetically and/or environmentally meaningful dimensions.

9. The method according to any of claims 1 to 8, whereby said known and unknown phenotypes comprise a plurality of 2D, 3D or 4D image-based characteristics or features; whereby said feature of a genotype comprises the value or variation at an individual or a small group of single nucleotide polymorphisms, haplotypes, copy number variations, or genomic measures such as, sex, genetic relationship and background.

10. The method according to claim 9, wherein the known and unknown phenotypes further comprise non-image based phenotypic features.

11. The method according to any of claims 9 or 10, wherein said known and unknown phenotypes comprise facial phenotypes consisting of multiple 2D, 3D or 4D facial images.

12. The method according to any of claims 7 to 11, further comprising the step of reducing the multidimensional phenotype-space to a one dimensional axis or curve in function of the feature of a genotype.

13. The method according to claim 12, whereby said reducing is performed using imaging biomarkers, filters or features.

14. The method according to claim 13, further comprising tracing the genetic architecture or basis of said imaging biomarkers, filters or features to improve classification from said imaging biomarkers, filters or features into their associated feature of a genotype.

15. The method according to any of claims 7 to 13, further comprising classification from said multidimensional phenotype-space into associated feature of a genotype.

16. The method according to any of claims 1 to 15, whereby said optimization algorithm is a global optimization algorithm, optionally followed by a local search optimization algorithm.

17. The method according to 16, whereby said global optimization algorithm is an evolutionary algorithm.

18. The method of claim 1, comprising:

- providing a database of known phenotypes;

- providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype;

- integrating the plurality of phenomic prediction models in at least one optimization algorithm;

- applying the at least one optimization algorithm comprising:

  - calculating an error score by comparing features of the genotype predicted by the phenomic prediction models based on known phenotypes from the database with the given genotype;

  - ranking known phenotypes in the database according to the error score;

  - selecting a subset of known phenotypes based on said ranking;

- optionally creating synthetic phenotypes based on the known phenotypes in the subset and adding the synthetic phenotypes to the subset; and
- iterating the steps of the optimization algorithm at least one time on the subset of known and optionally synthetic phenotypes;

- predicting the unknown phenotype for the given genotype to be one of the highest ranked known or optionally synthetic phenotypes, or to be a synthetic phenotype derived therefrom.

19. The method of claim 1, comprising:

- providing a database comprising at least one known phenotype;

- providing a plurality of phenomic prediction models, each phenomic prediction models adapted to predict at least one feature of a genotype based on a known phenotype;

- integrating the plurality of phenomic prediction models in at least one optimization algorithm;

- applying the at least one optimization algorithm comprising:

- calculating a first error score by comparing a first feature of the genotype predicted by at least a first phenomic prediction model based on at least one known phenotype from the database with the given genotype;
- calculating a second error score by comparing a second feature of the genotype predicted by at least a second phenomic prediction model based on at least one known phenotype from the database with the given genotype;
- adapting at least one known phenotype from the database based on the first and second error score to obtain a database comprising at least one adapted phenotype;
- optionally iterating the steps of the optimization algorithm at least one time on the database comprising the at least one adapted phenotype;

- predicting the unknown phenotype for the given genotype to be the adapted phenotype.

20. A computer program product for, if implemented on a control unit, performing a method according to any of claims 1 to 19.

21. A data carrier storing a computer program product according to claim 20.

22. Transmission of a computer program product according to claim 20 over a network.

23. A device for comprising a control unit performing a method according to any of claims 1 to 19.

24. Use of a method according to any one of claims 1 to 19 in a genome-wide association study.

25. A database comprising multiple phenotypes, wherein the phenotypes have been ranked in relation to a given genotype in accordance with the method of any one of claims 1 to 18.

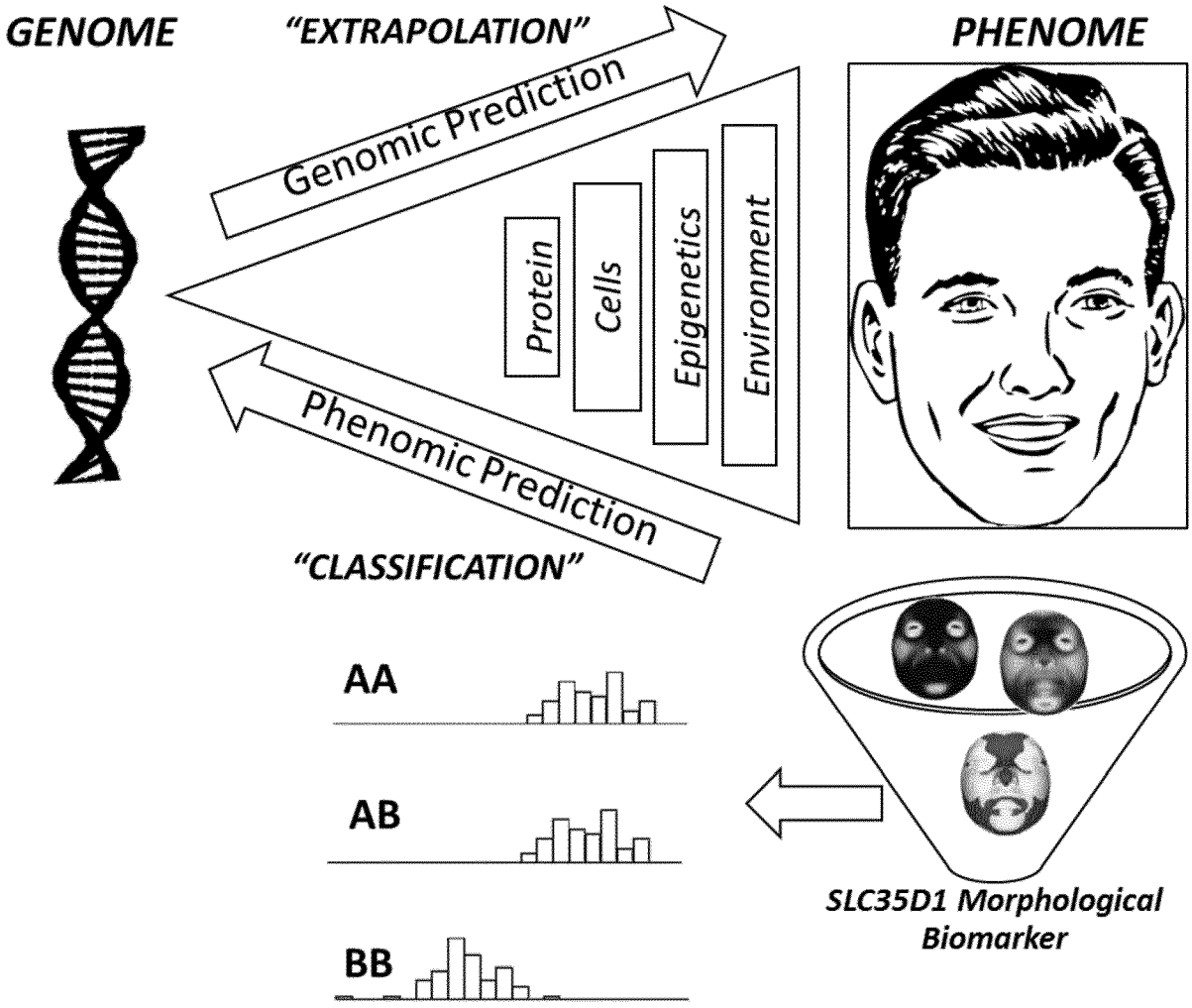

GENOME
"EXTRAPOLATION"
PHENOME
Genomic Prediction
Protein
Cells
Epigenetics
Environment
Phenomic Prediction
"CLASSIFICATION"
AA
AB
BB
SLC35D1 Morphological Biomarker

Fig. 1

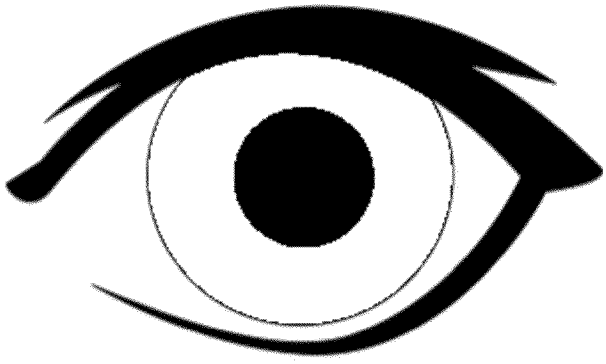


Fig. 2

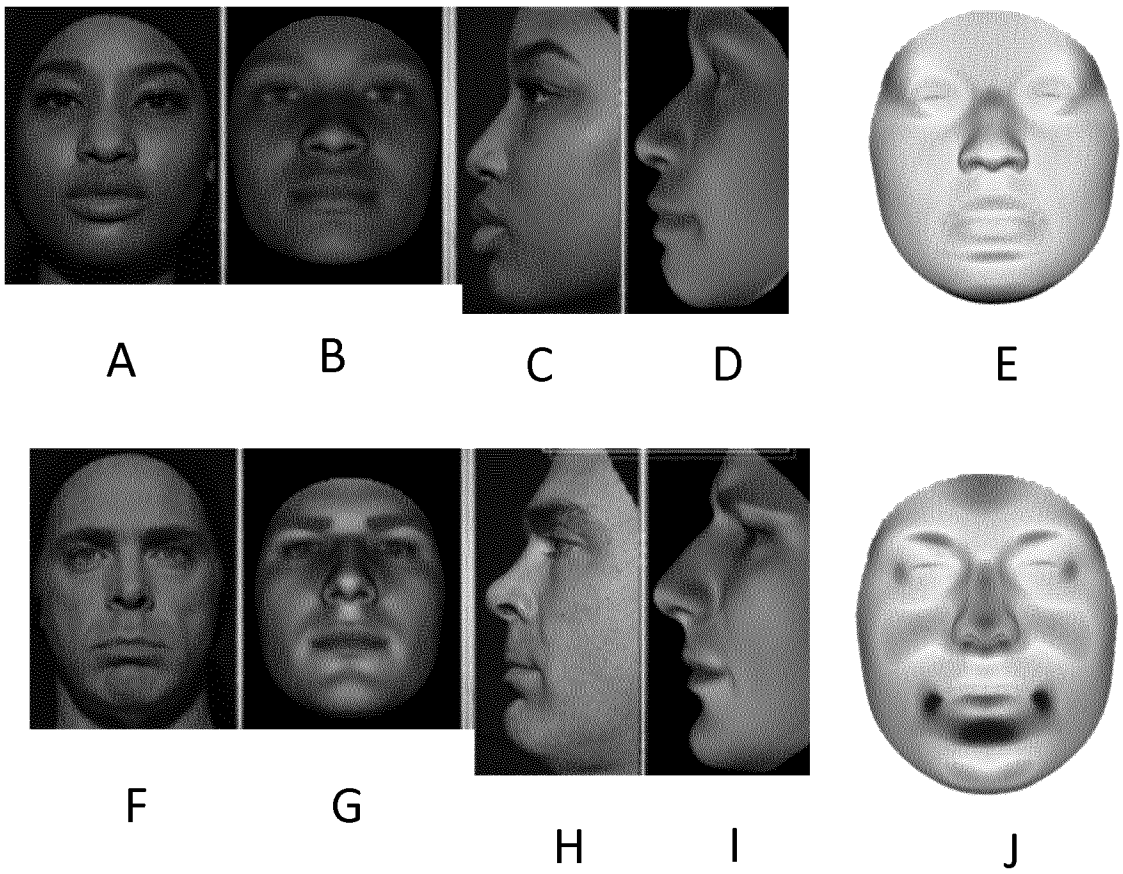


Fig. 3

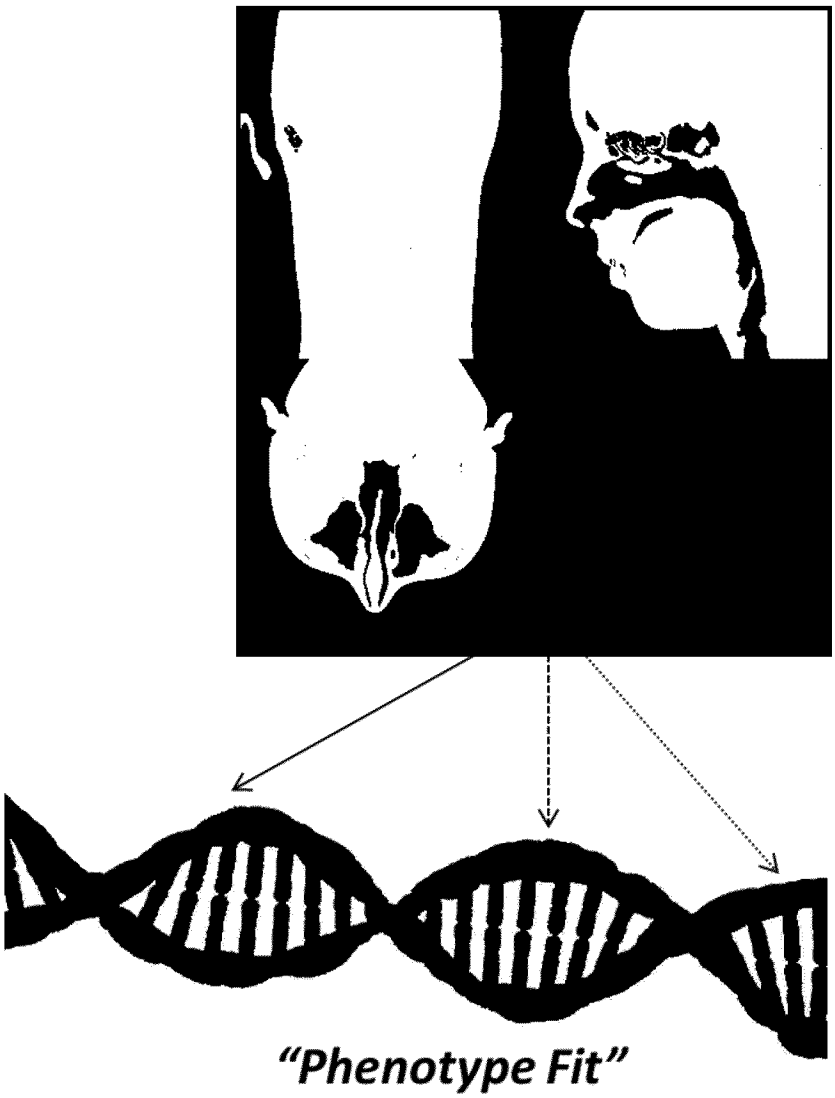


Fig. 4

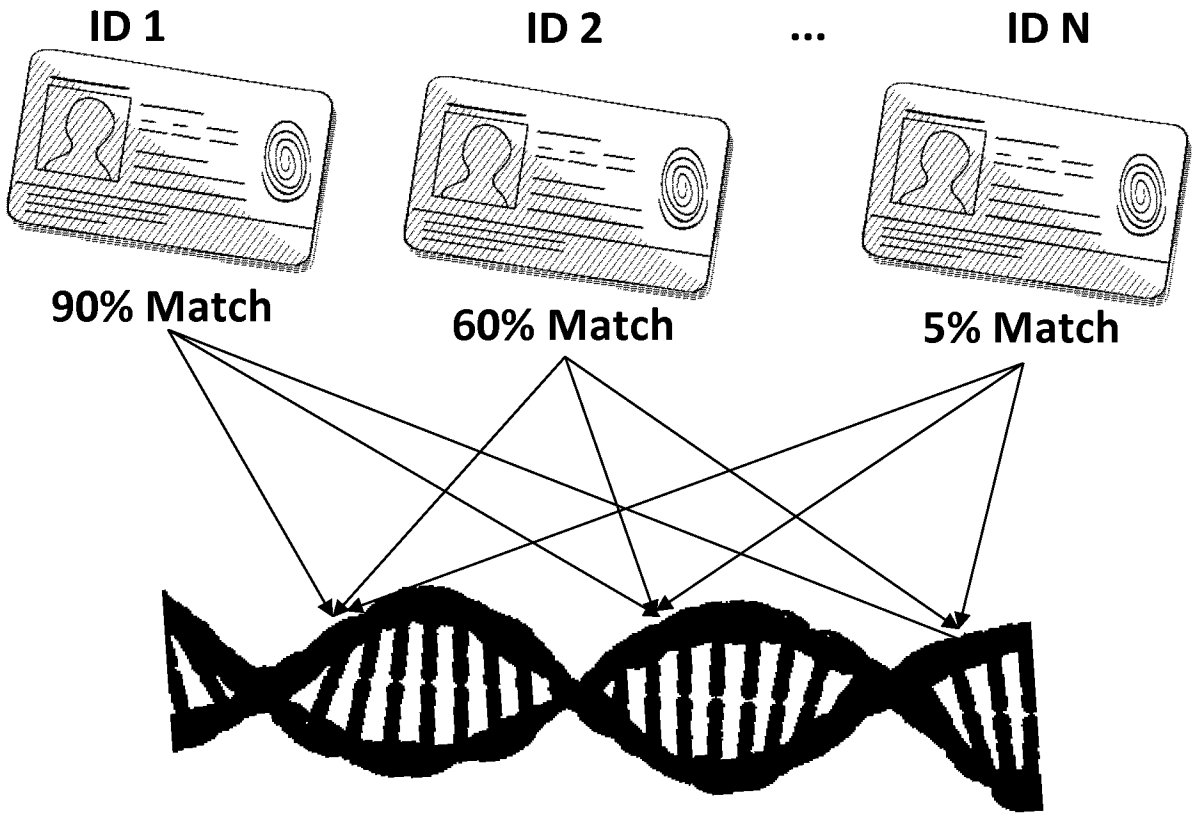


Fig. 5

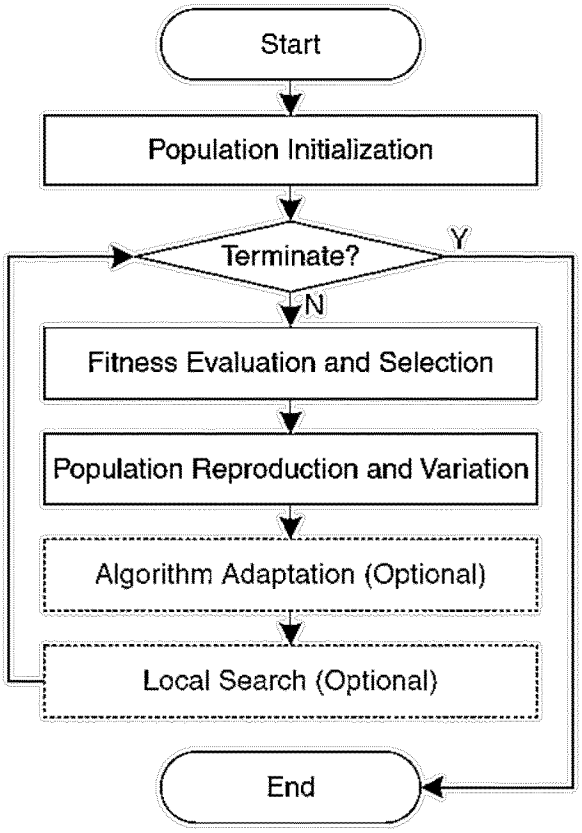

Start
Population Initialization
Terminate?
Y
N
Fitness Evaluation and Selection
Population Reproduction and Variation
Algorithm Adaptation (Optional)
Local Search (Optional)
End

Fig. 6

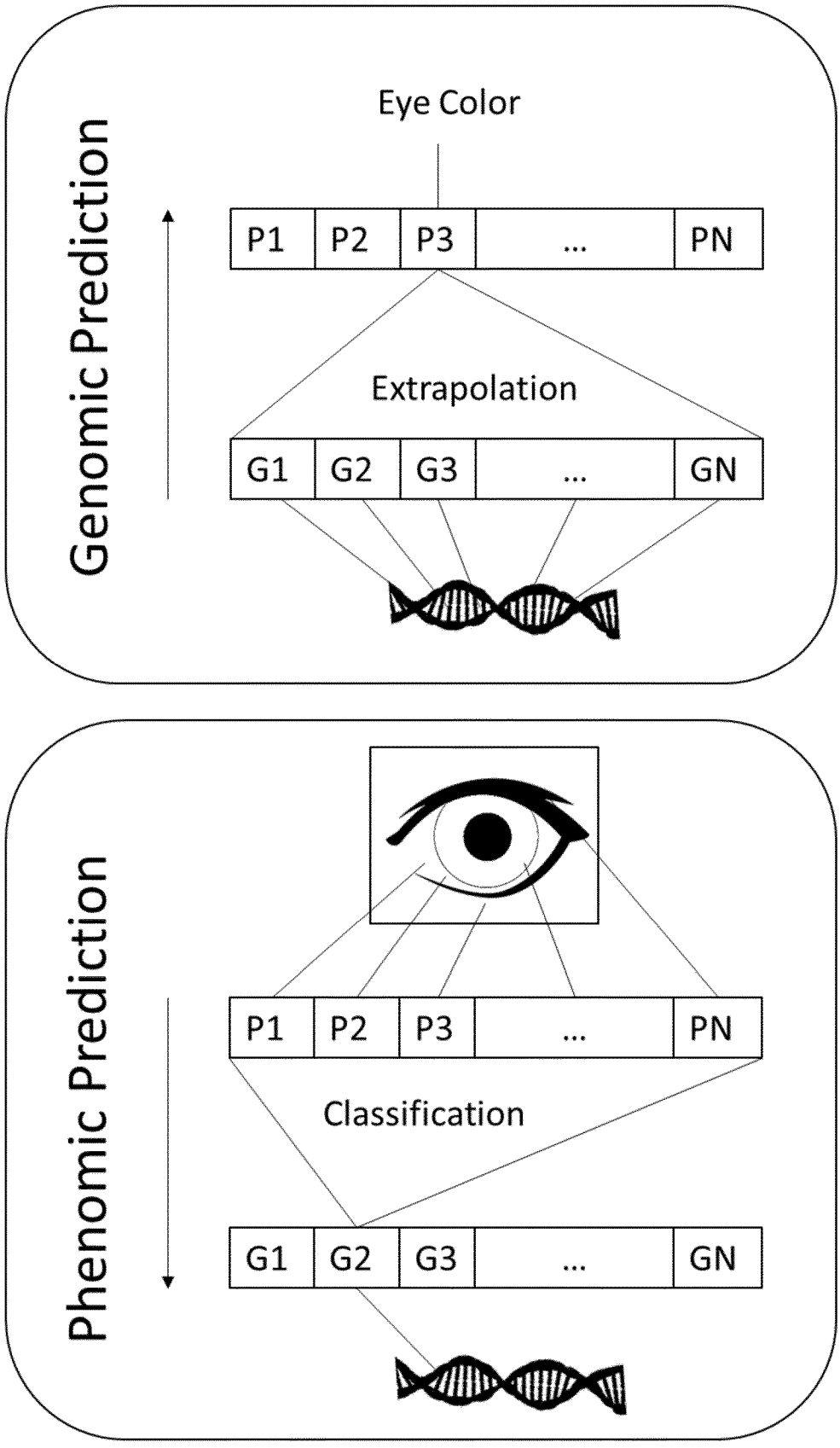

Genomic Prediction
Eye Color
P1
P2
P3
...
PN
Extrapolation
G1
G2
G3
...
GN
Phenomic Prediction
P1
P2
P3
...
PN
Classification
G1
G2
G3
...
GN

Fig. 7

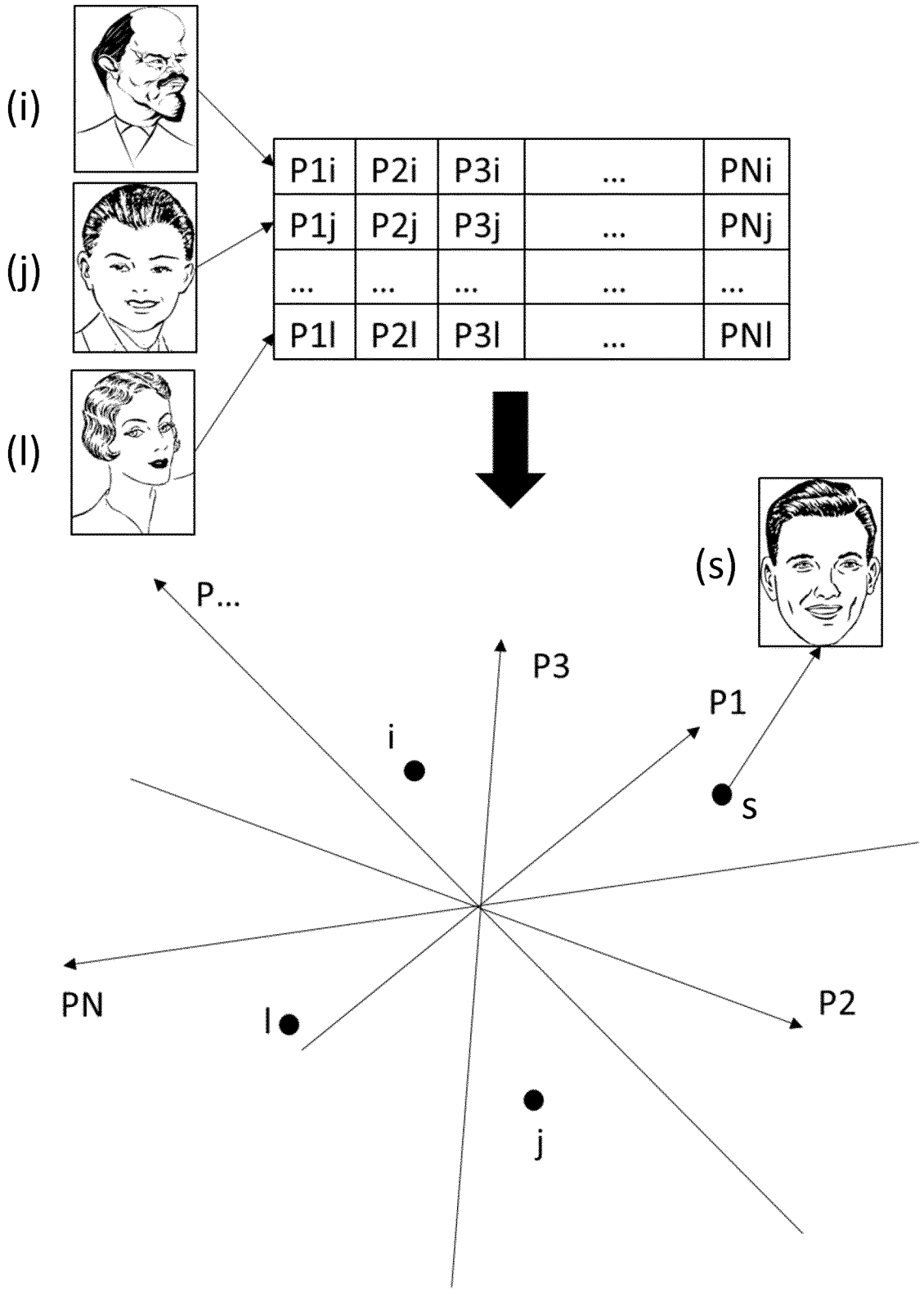

(i)
(j)
(l)
P1i P2i P3i ... PNi
P1j P2j P3j ... PNj
... ...
P1l P2l P3l ... PNl
(s)
P...
P3
P1
i
s
PN
l
P2
j

Fig. 8

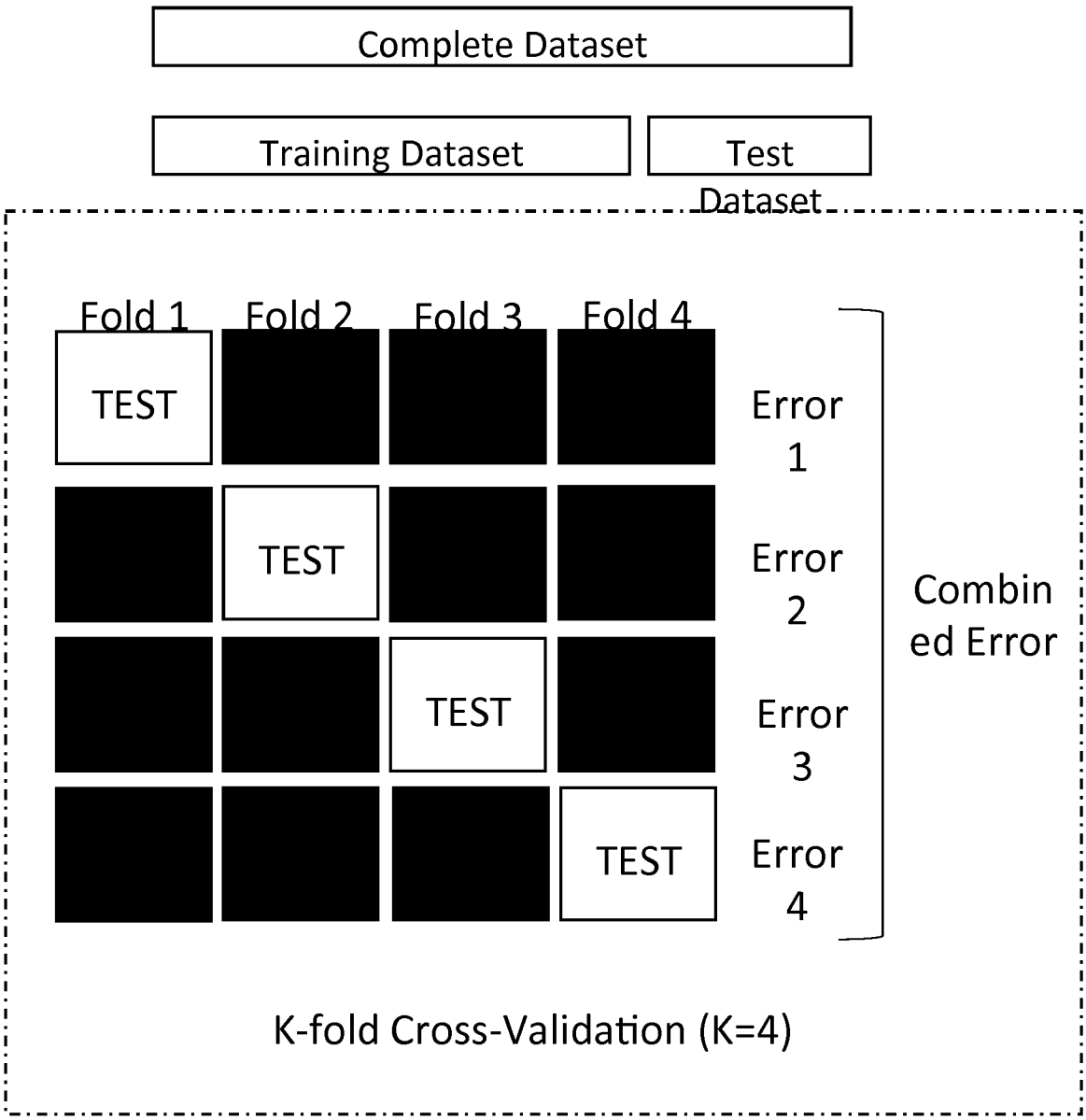

Complete Dataset
Training Dataset
Test Dataset
Fold 1
Fold 2
Fold 3
Fold 4
TEST
TEST
TEST
TEST
Error 1
Error 2
Error 3
Error 4
Combined Error
K-fold Cross-Validation (K=4)

Fig. 9

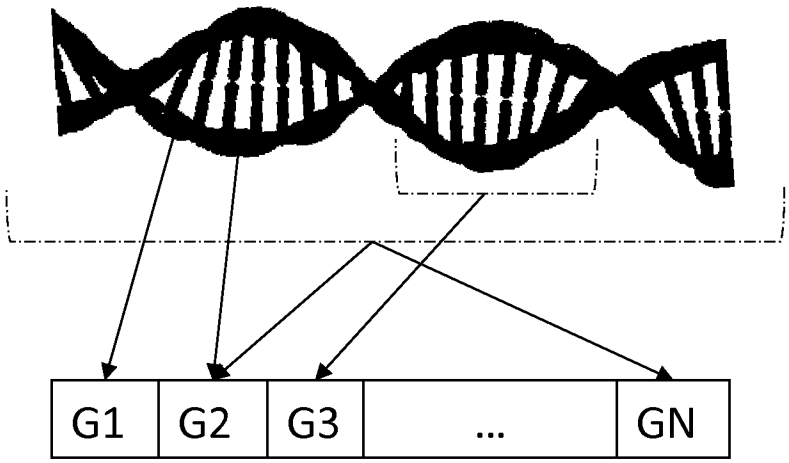
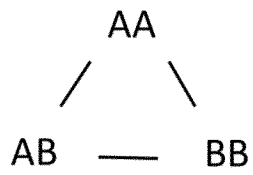


Fig. 10

Modes of Inheritance

AA – AB – BB	*Additive*
AAAB – BB	*Recessive*
AA – ABBB	*Dominant*
AABB – AB	*Over Dominant*



Non Additive

Fig. 11

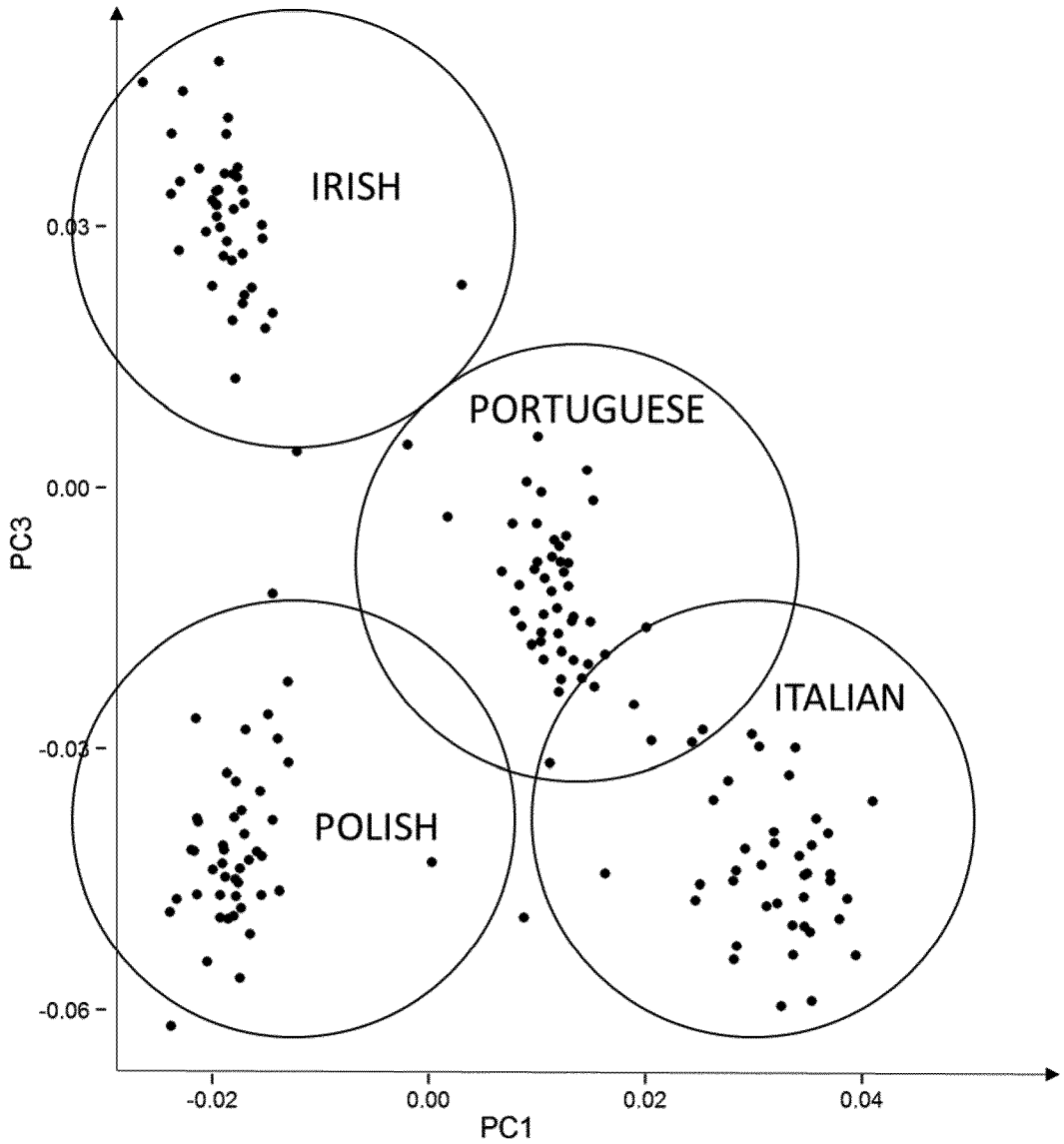

IRISH
PORTUGUESE
ITALIAN
POLISH
0.03
0.00
-0.03
-0.06
PC3
-0.02
0.00
0.02
0.04
PC1

Fig. 12

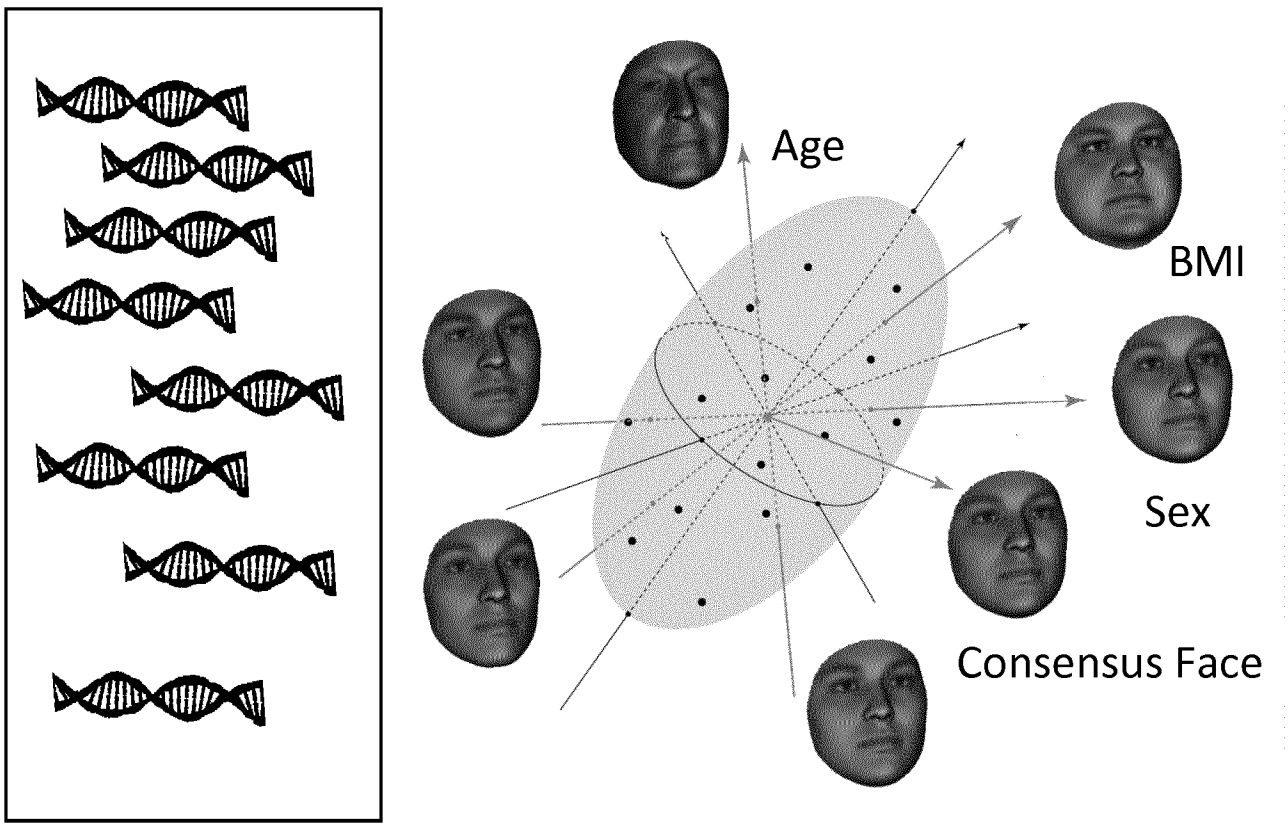
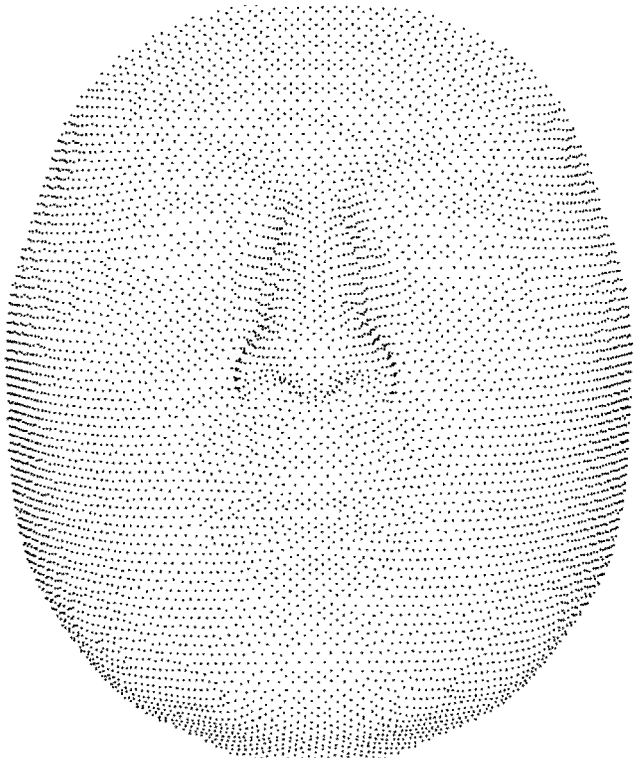
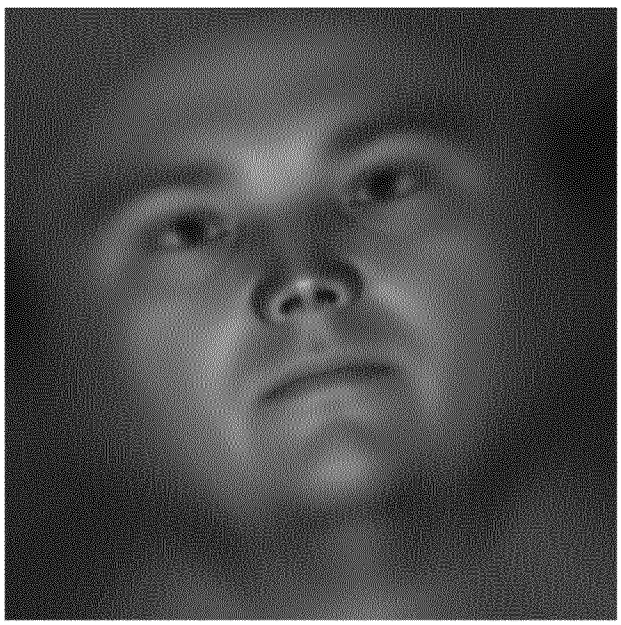


Fig. 13



A) Facial Morphology (7150 3D points)



B) Facial Texture (683x683 pixels)



C) Facial Appearance (A+B)

Fig. 14

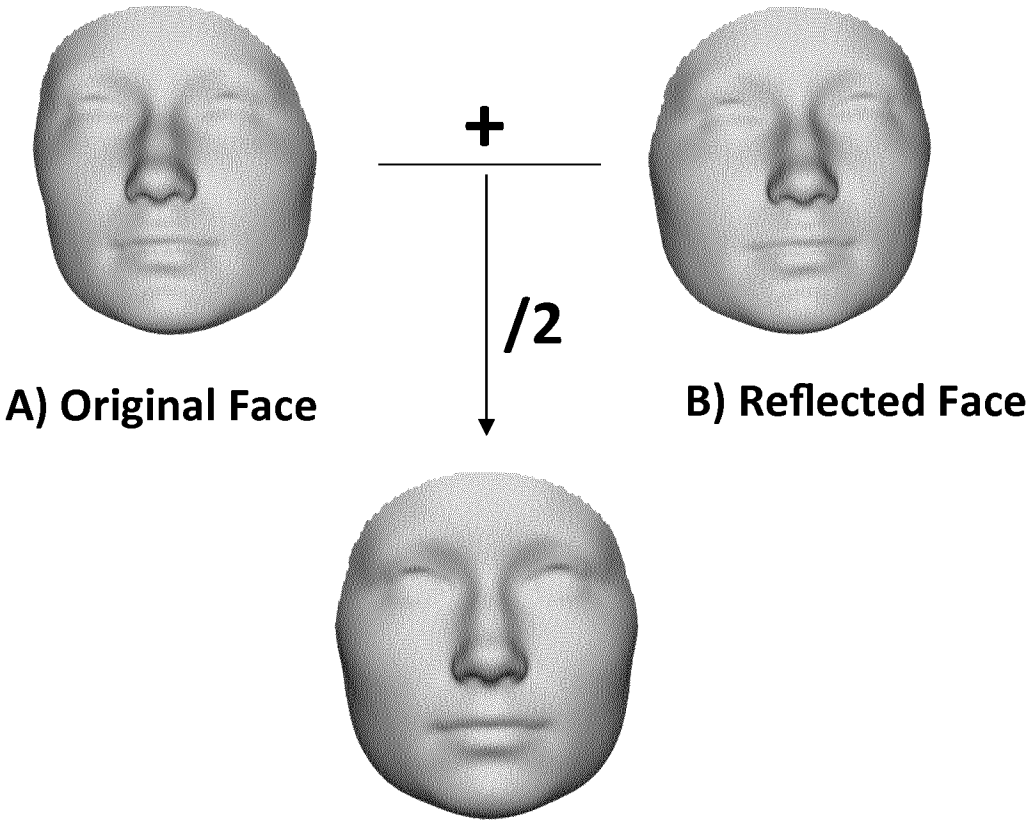
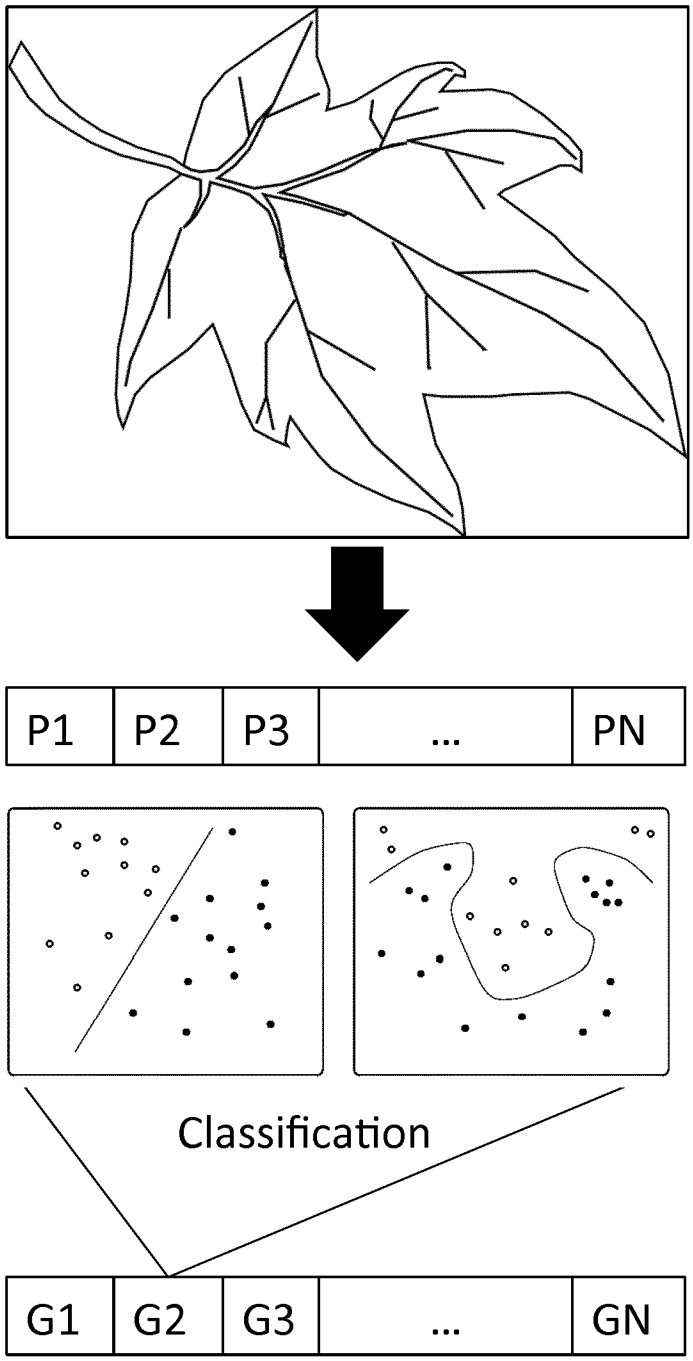


Fig. 15



Phenomic Prediction of a feature of a Genotype (G) using Image-based Phenotype Features (P)

Fig. 16

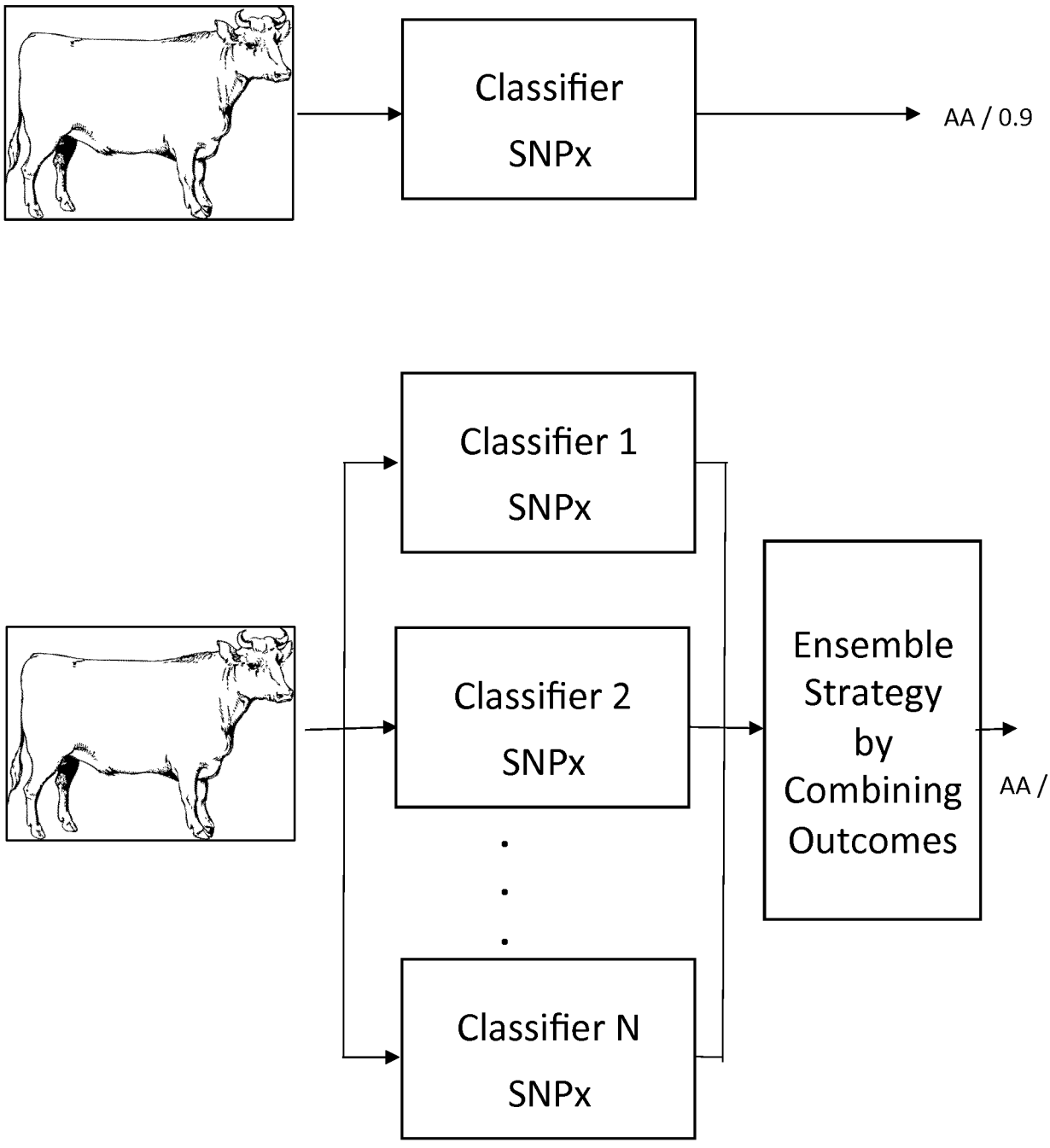

Classifier
SNPx
AA / 0.9
Classifier 1
SNPx
Classifier 2
SNPx
Classifier N
SNPx
Ensemble
Strategy
by
Combining
Outcomes
AA / 0.9

Fig. 17

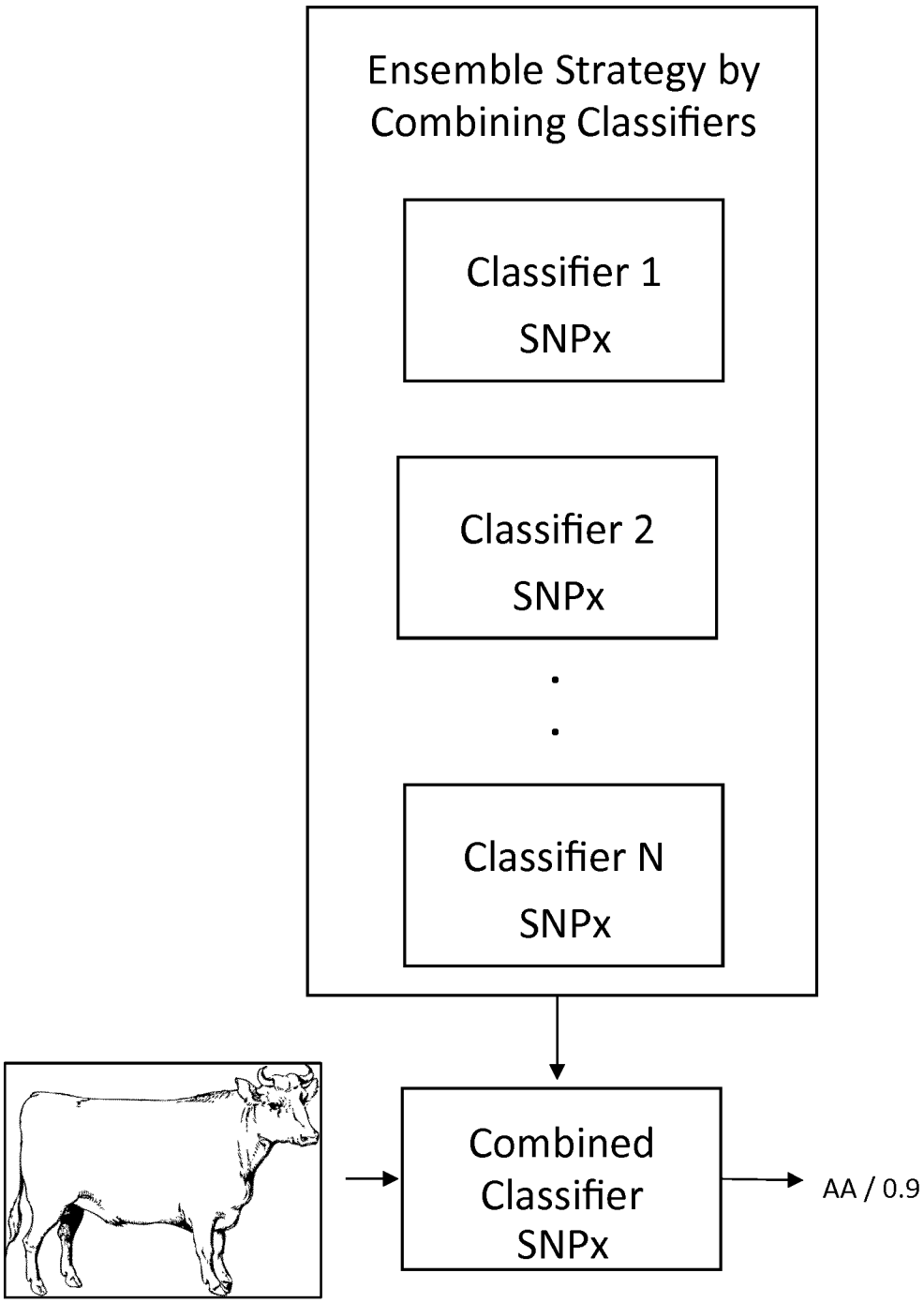

Ensemble Strategy by Combining Classifiers
Classifier 1
SNPx
Classifier 2
SNPx
Classifier N
SNPx
Combined Classifier SNPx
AA / 0.9

Fig. 18

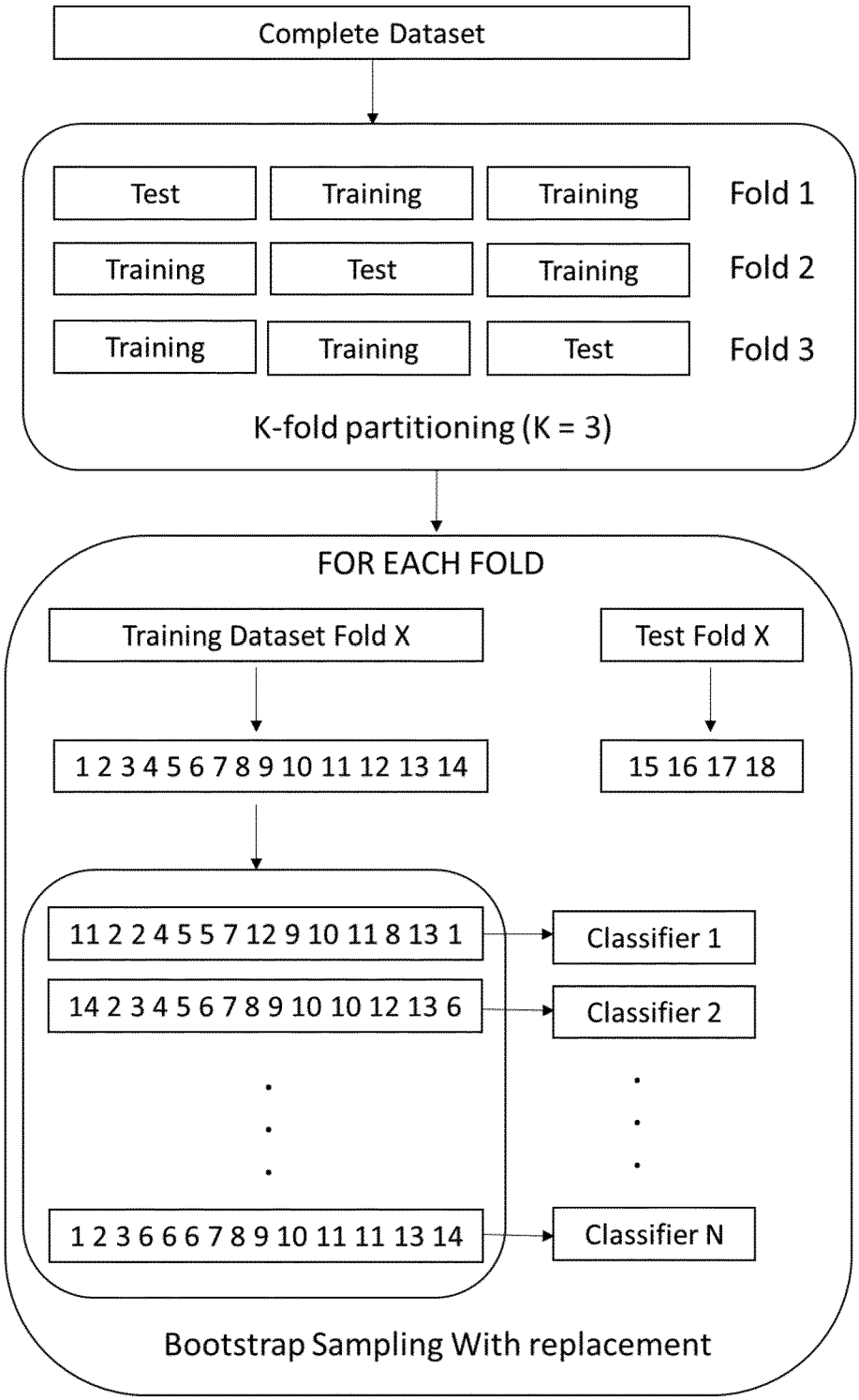

Complete Dataset
Test
Training
Training
Fold 1
Training
Test
Training
Fold 2
Training
Training
Test
Fold 3
K-fold partitioning (K = 3)
FOR EACH FOLD
Training Dataset Fold X
Test Fold X
1 2 3 4 5 6 7 8 9 10 11 12 13 14
15 16 17 18
11 2 2 4 5 5 7 12 9 10 11 8 13 1
Classifier 1
14 2 3 4 5 6 7 8 9 10 10 12 13 6
Classifier 2
1 2 3 6 6 6 7 8 9 10 11 11 13 14
Classifier N
Bootstrap Sampling With replacement

Fig. 19

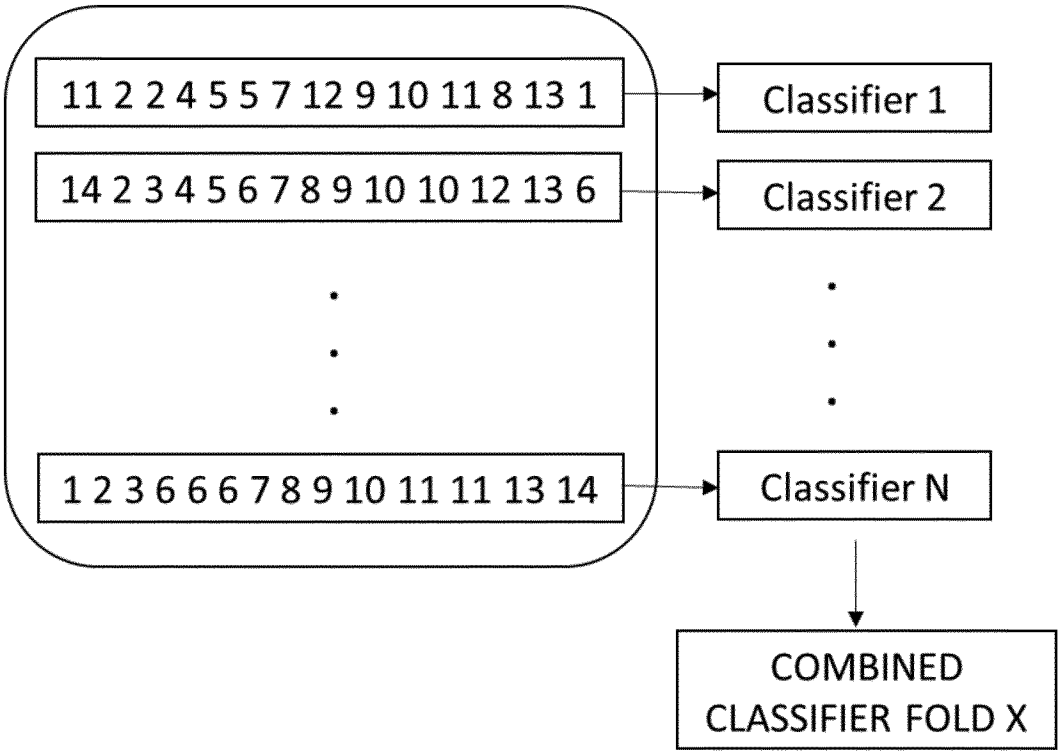

11 2 2 4 5 5 7 12 9 10 11 8 13 1
14 2 3 4 5 6 7 8 9 10 10 12 13 6
1 2 3 6 6 6 7 8 9 10 11 11 13 14
Classifier 1
Classifier 2
Classifier N
COMBINED CLASSIFIER FOLD X

Fig. 20

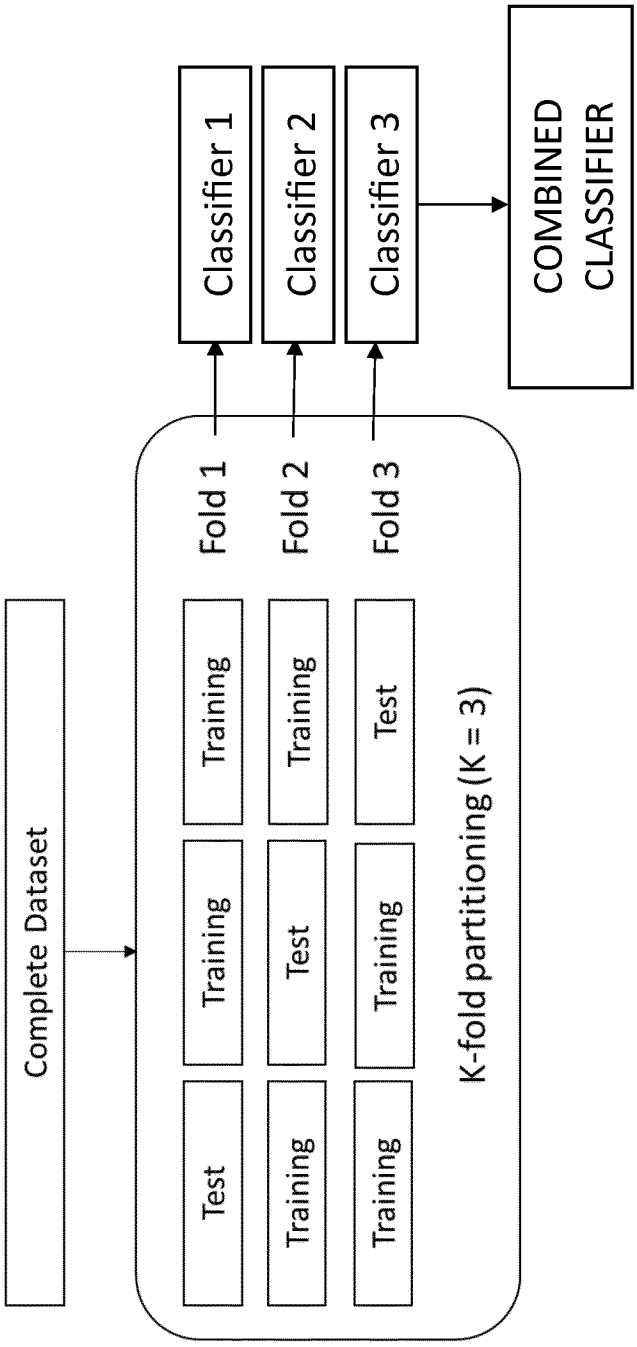

Complete Dataset
K-fold partitioning (K = 3)
Fold 1
Fold 2
Fold 3
Training
Training
Test
Test
Training
Training
Training
Test
Classifier 1
Classifier 2
Classifier 3
COMBINED CLASSIFIER

Fig. 21

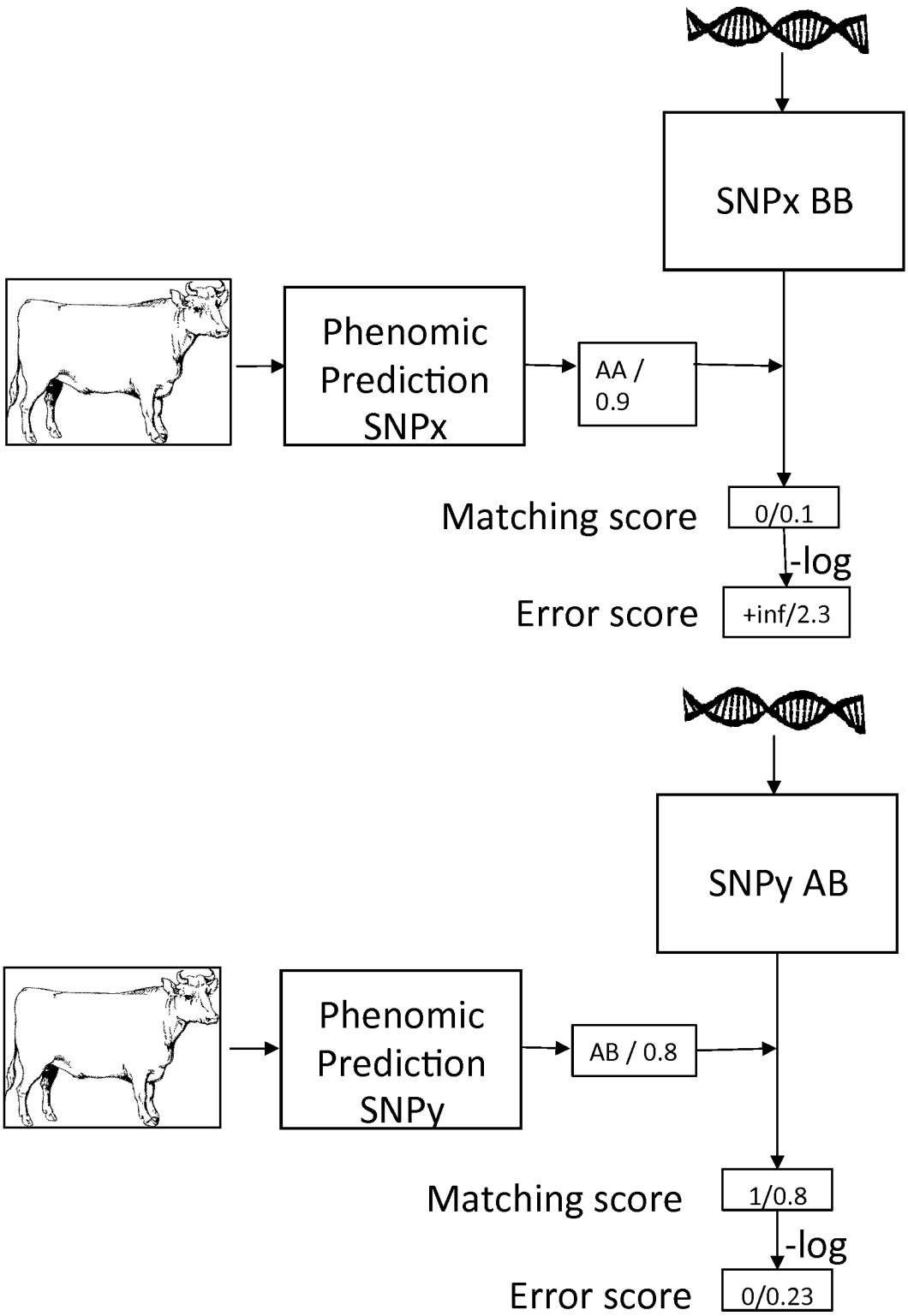

SNPx BB
Phenomic Prediction SNPx
AA / 0.9
Matching score
0/0.1
-log
Error score
+inf/2.3
SNPy AB
Phenomic Prediction SNPy
AB / 0.8
Matching score
1/0.8
-log
Error score
0/0.23

Fig. 22

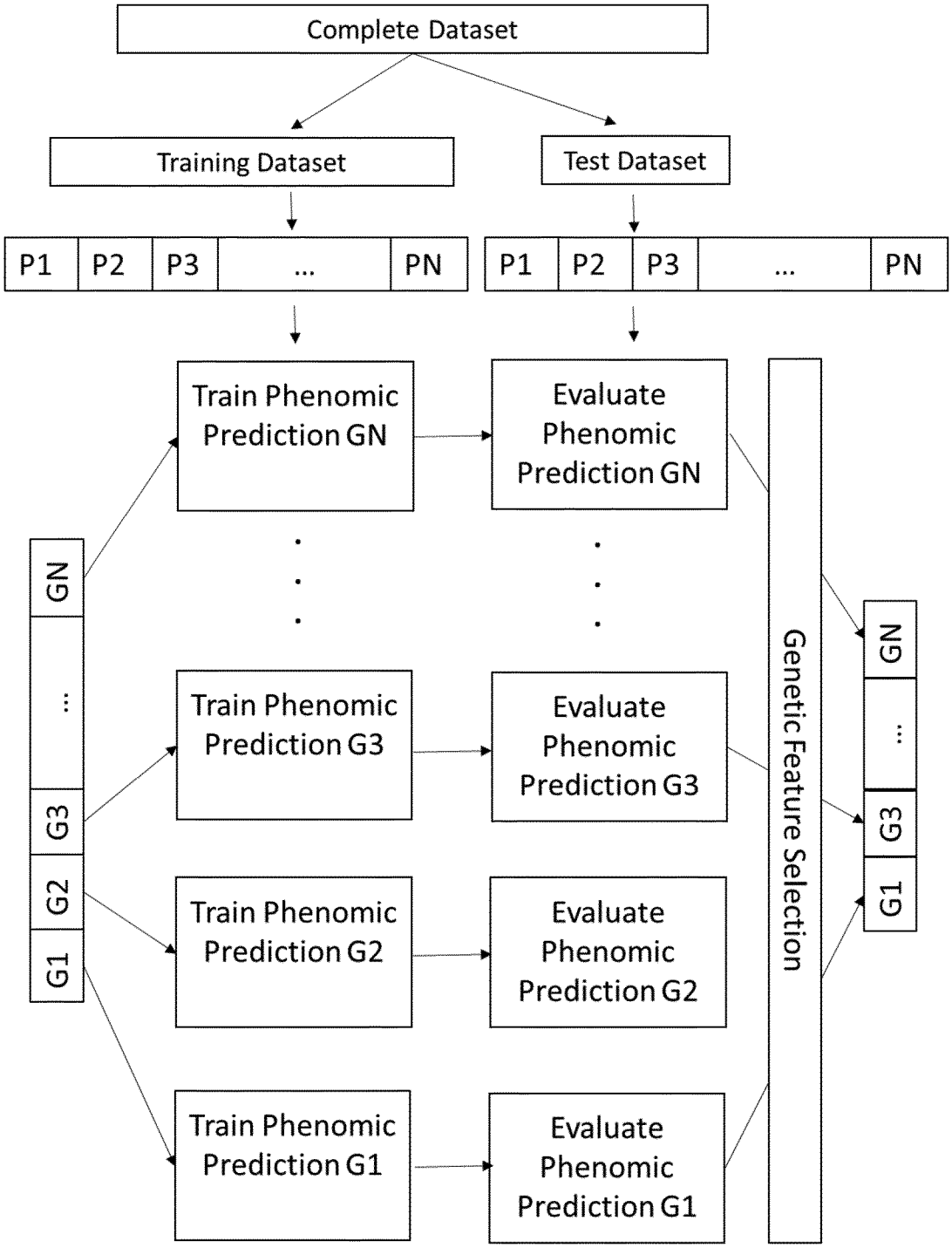
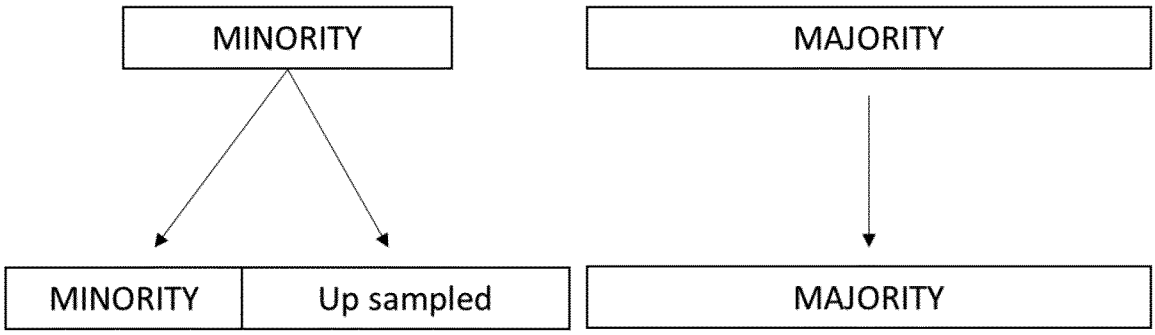

Complete Dataset
Training Dataset
Test Dataset
P1
P2
P3
...
PN
P1
P2
P3
...
PN
Train Phenomic Prediction GN
Evaluate Phenomic Prediction GN
Train Phenomic Prediction G3
Evaluate Phenomic Prediction G3
Train Phenomic Prediction G2
Evaluate Phenomic Prediction G2
Train Phenomic Prediction G1
Evaluate Phenomic Prediction G1
GN
...
G3
G2
G1
Genetic Feature Selection
GN
...
G3
G1

Fig. 23


MINORITY
MAJORITY
MINORITY
Up sampled
MAJORITY

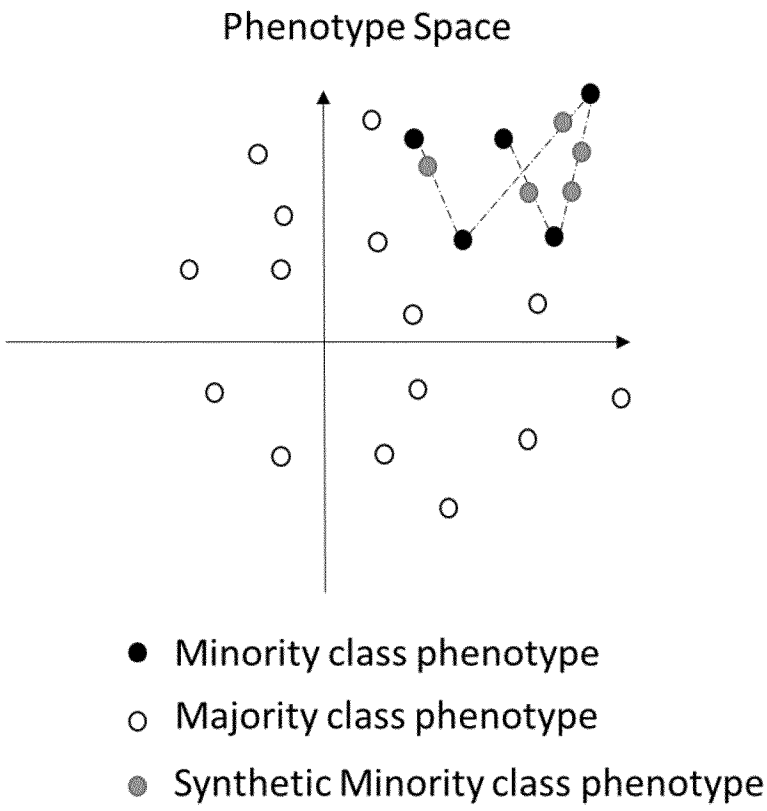

Phenotype Space
Minority class phenotype
Majority class phenotype
Synthetic Minority class phenotype

Fig. 24

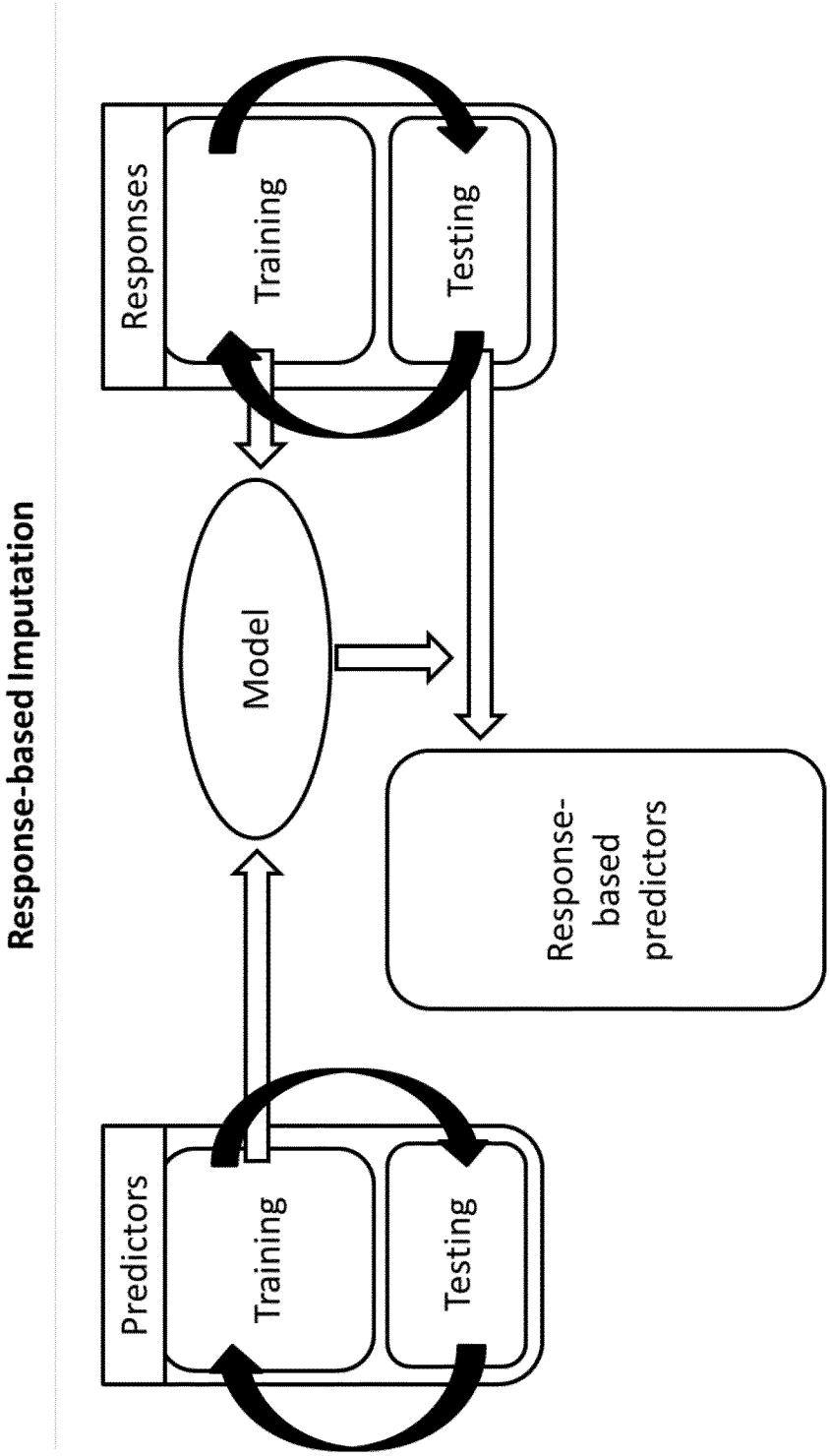

Response-based Imputation
Responses
Training
Testing
Model
Response-based predictors
Predictors
Training
Testing

Fig. 25

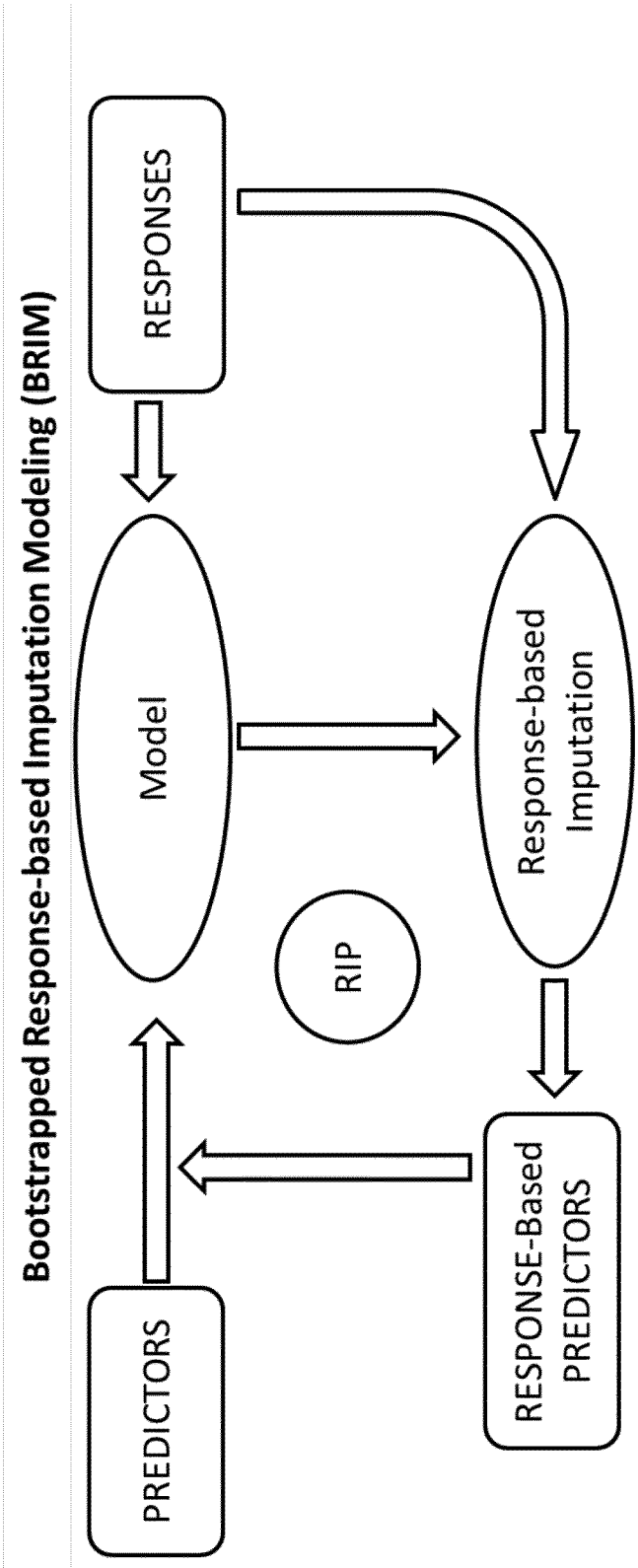

Bootstrapped Response-based Imputation Modeling (BRIM)
RESPONSES
Model
Response-based Imputation
RIP
RESPONSE-Based PREDICTORS
PREDICTORS

Fig. 26

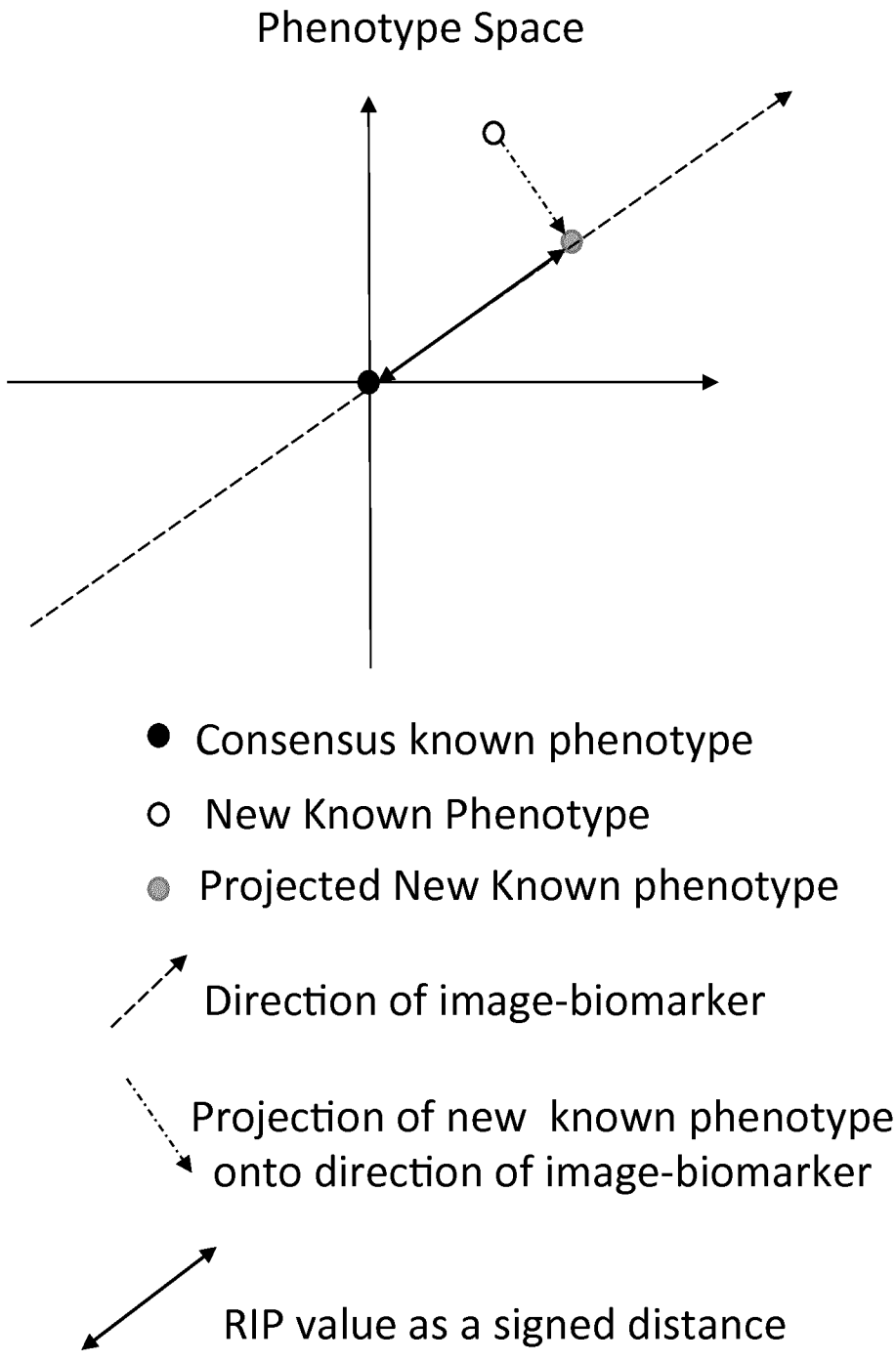

Phenotype Space
Consensus known phenotype
New Known Phenotype
Projected New Known phenotype
Direction of image-biomarker
Projection of new known phenotype onto direction of image-biomarker
RIP value as a signed distance

Fig. 27

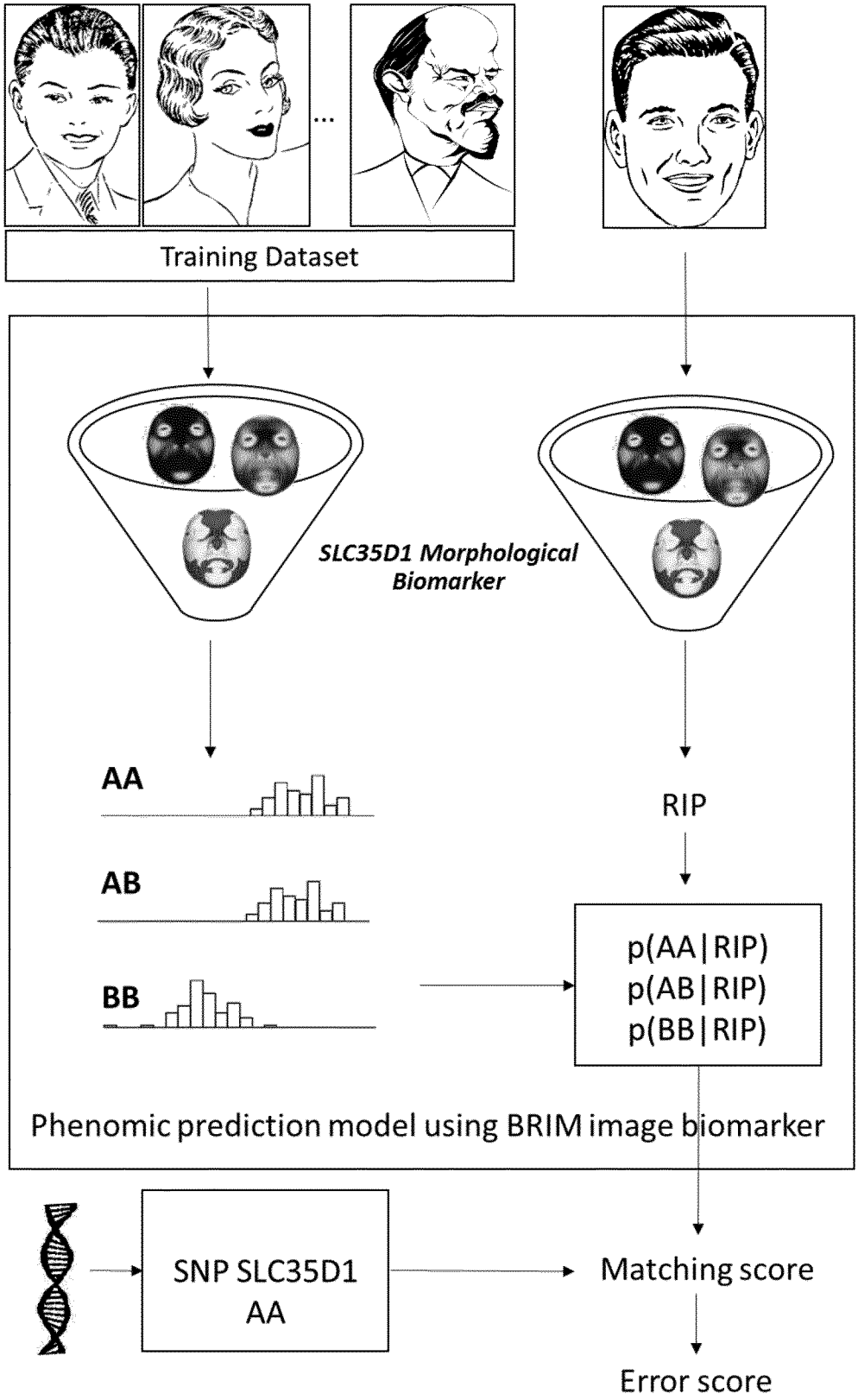

Training Dataset
SLC35D1 Morphological Biomarker
AA
AB
BB
RIP
p(AA|RIP)
p(AB|RIP)
p(BB|RIP)
Phenomic prediction model using BRIM image biomarker
SNP SLC35D1 AA
Matching score
Error score

Fig. 28

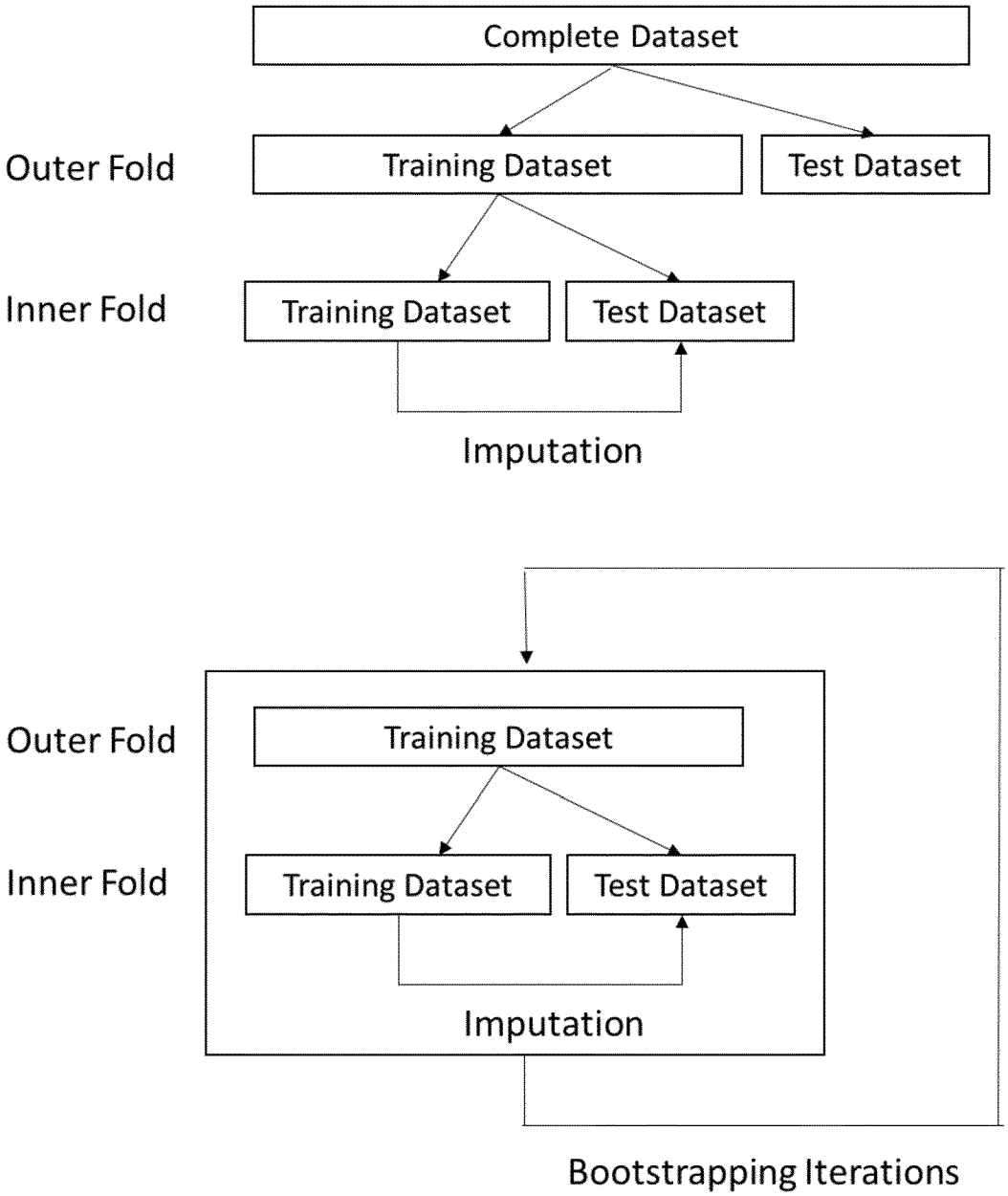

Complete Dataset
Outer Fold
Training Dataset
Test Dataset
Inner Fold
Training Dataset
Test Dataset
Imputation
Outer Fold
Training Dataset
Inner Fold
Training Dataset
Test Dataset
Imputation
Bootstrapping Iterations

Fig. 29

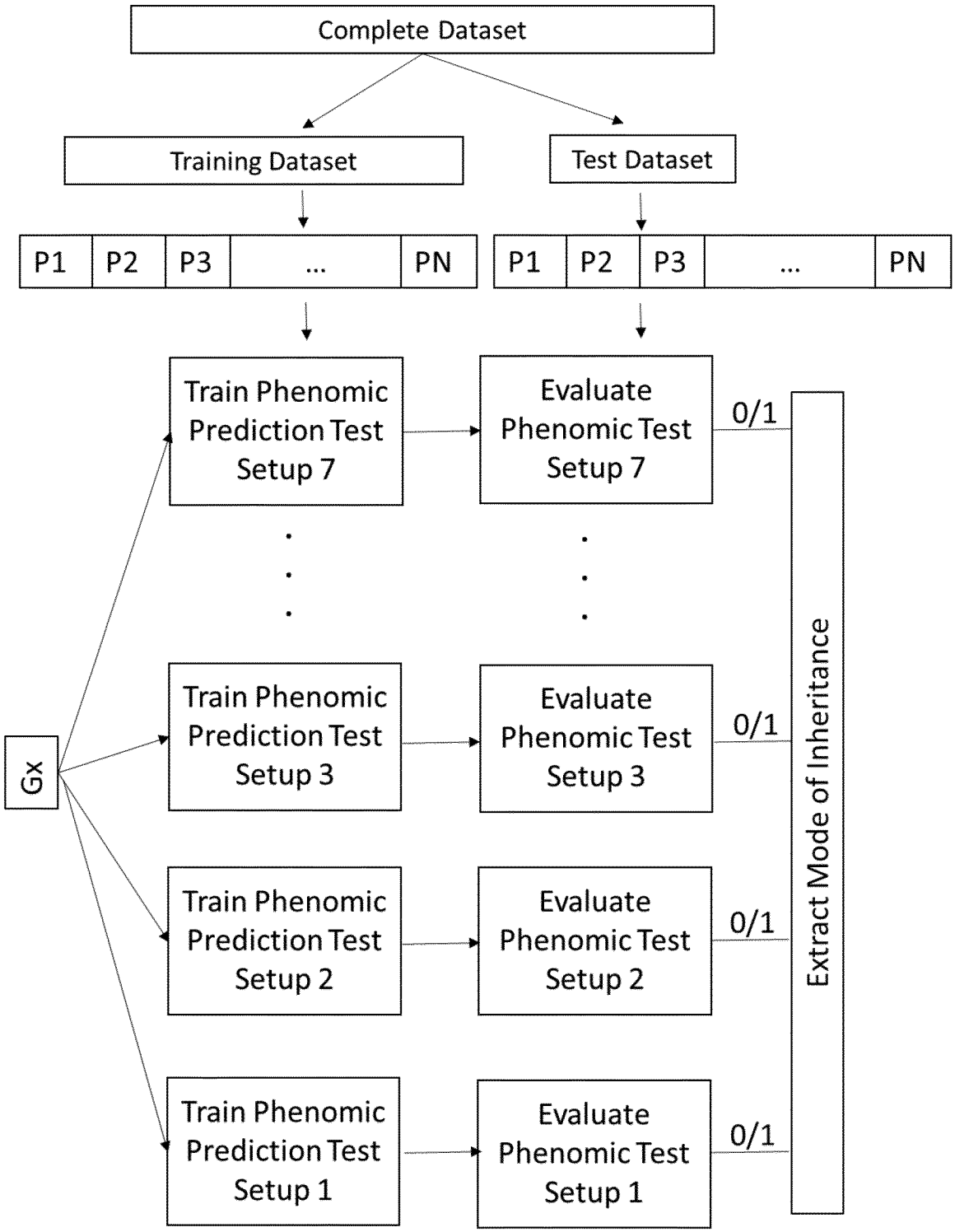

Complete Dataset
Training Dataset
Test Dataset
P1
P2
P3
...
PN
P1
P2
P3
...
PN
Train Phenomic Prediction Test Setup 7
Evaluate Phenomic Test Setup 7
0/1
Gx
Train Phenomic Prediction Test Setup 3
Evaluate Phenomic Test Setup 3
0/1
Train Phenomic Prediction Test Setup 2
Evaluate Phenomic Test Setup 2
0/1
Train Phenomic Prediction Test Setup 1
Evaluate Phenomic Test Setup 1
0/1
Extract Mode of Inheritance

Fig. 30

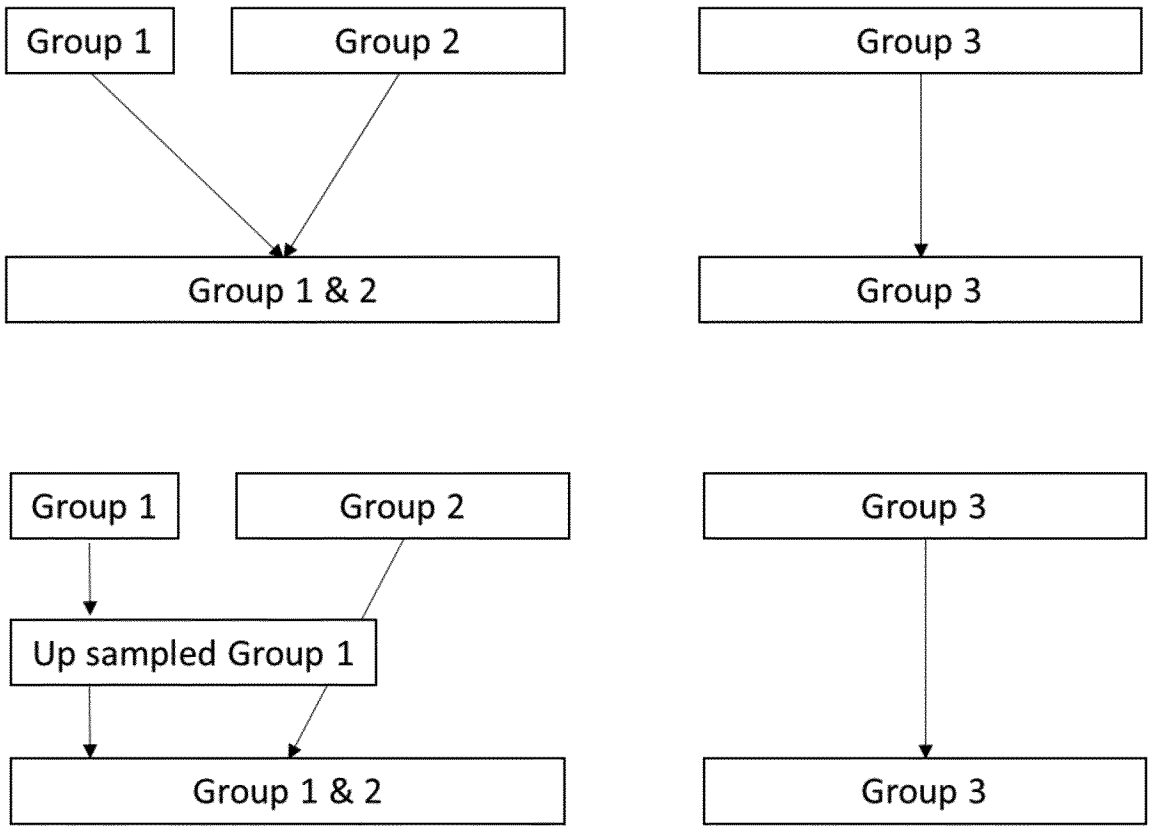

Group 1
Group 2
Group 3
Group 1 & 2
Group 3
Group 1
Group 2
Group 3
Up sampled Group 1
Group 1 & 2
Group 3

Fig. 31

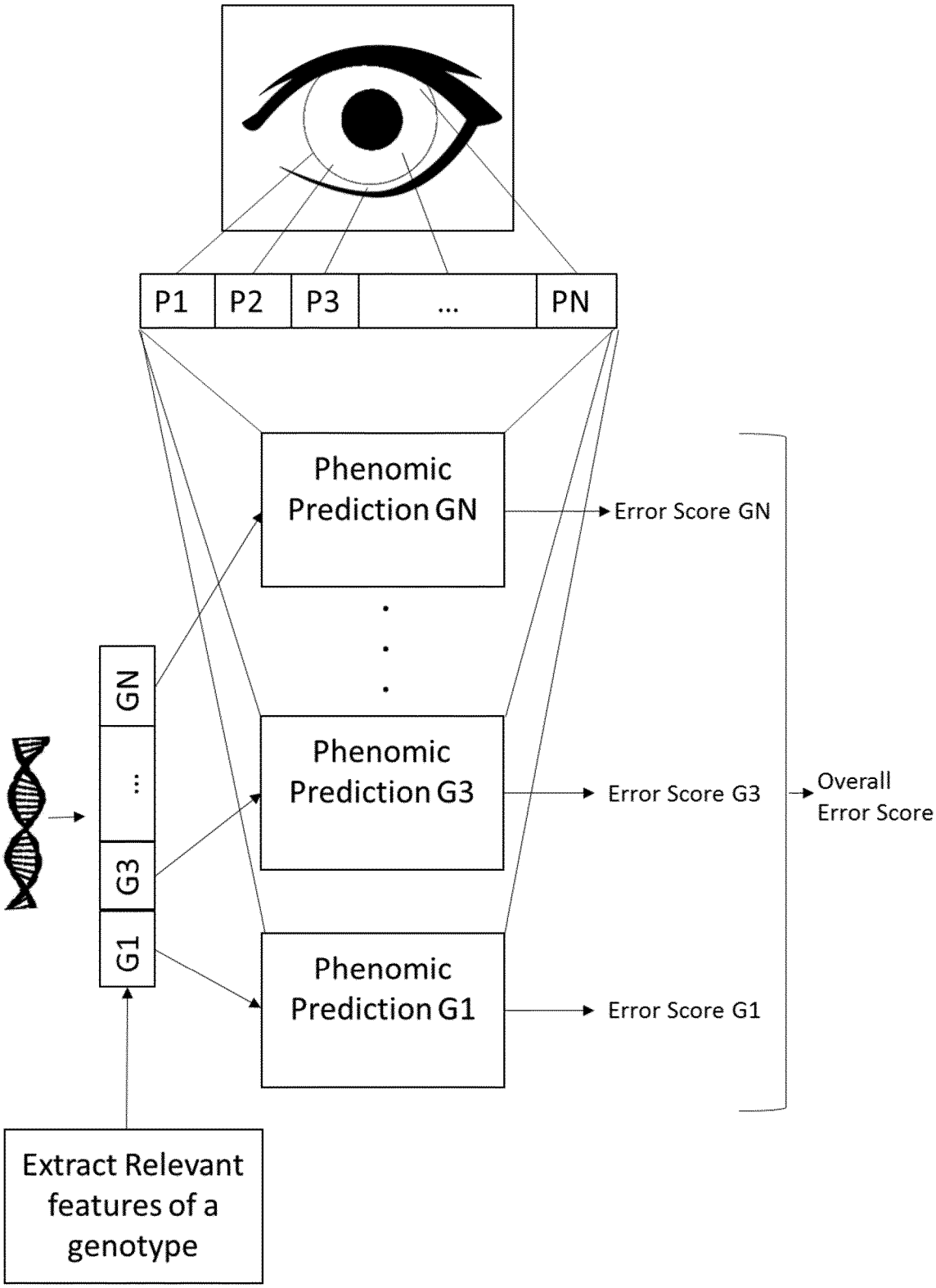

P1
P2
P3
...
PN
Phenomic Prediction GN
Error Score GN
Phenomic Prediction G3
Error Score G3
Phenomic Prediction G1
Error Score G1
Overall Error Score
GN
...
G3
G1
Extract Relevant features of a genotype

Fig. 32

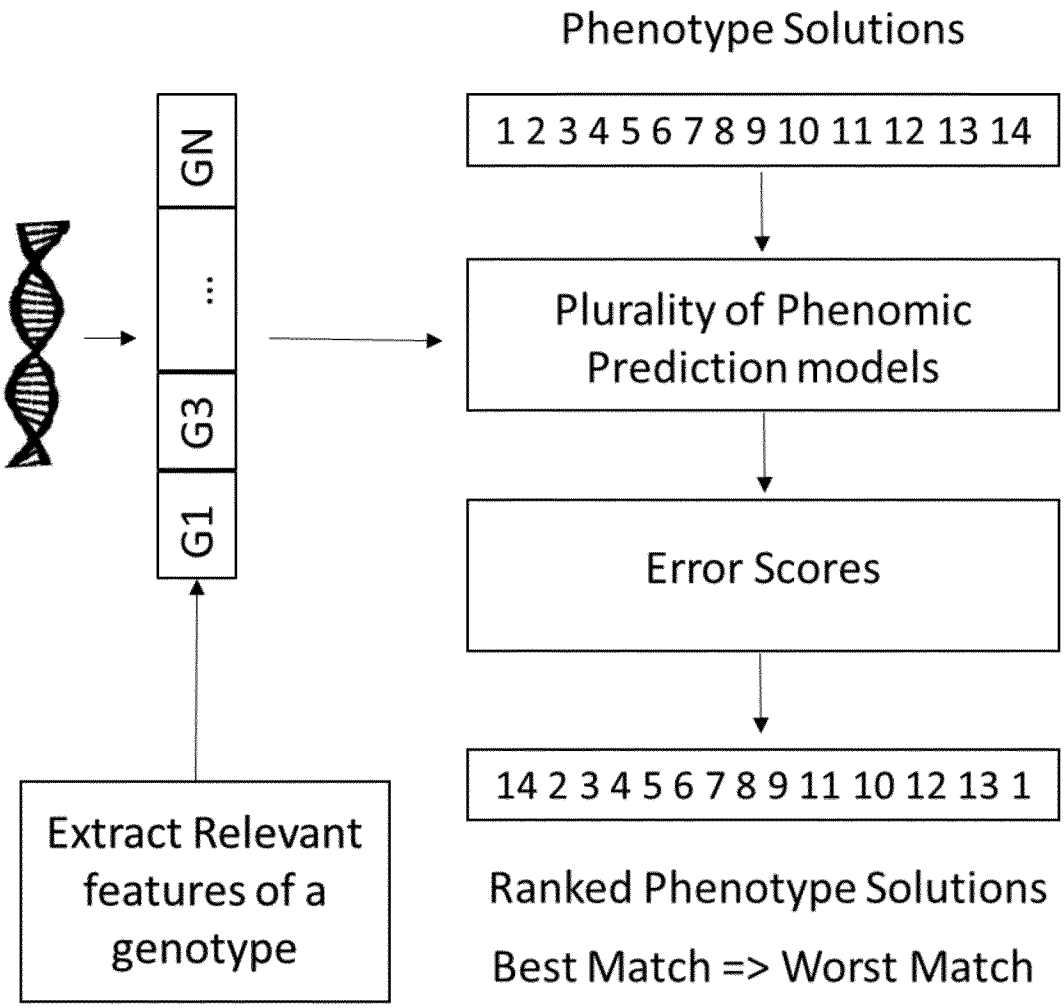

Phenotype Solutions
1 2 3 4 5 6 7 8 9 10 11 12 13 14
GN
...
G3
G1
Plurality of Phenomic Prediction models
Error Scores
14 2 3 4 5 6 7 8 9 11 10 12 13 1
Ranked Phenotype Solutions
Best Match => Worst Match
Extract Relevant features of a genotype

Fig. 33

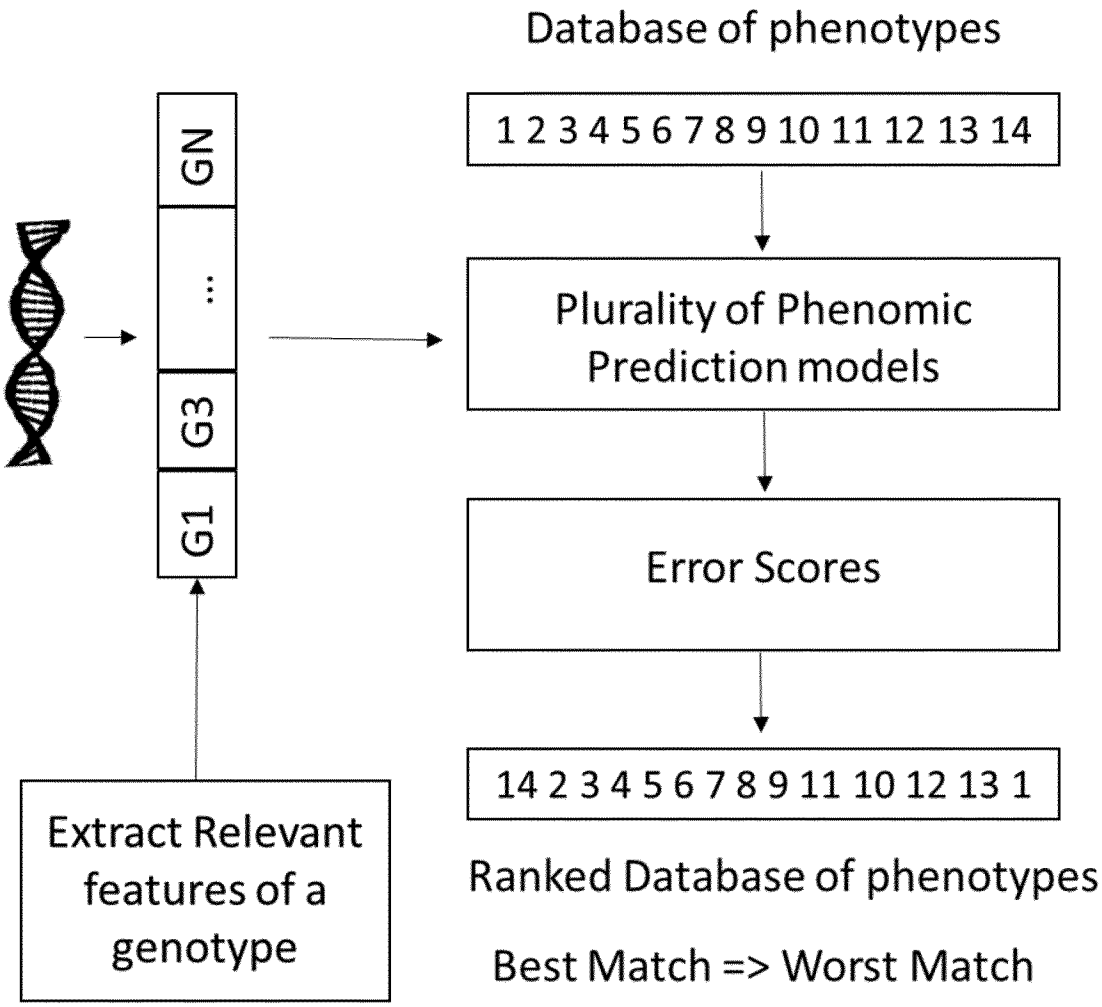

Database of phenotypes
1 2 3 4 5 6 7 8 9 10 11 12 13 14
GN
...
G3
G1
Plurality of Phenomic Prediction models
Error Scores
14 2 3 4 5 6 7 8 9 11 10 12 13 1
Ranked Database of phenotypes
Best Match => Worst Match
Extract Relevant features of a genotype

Fig. 34

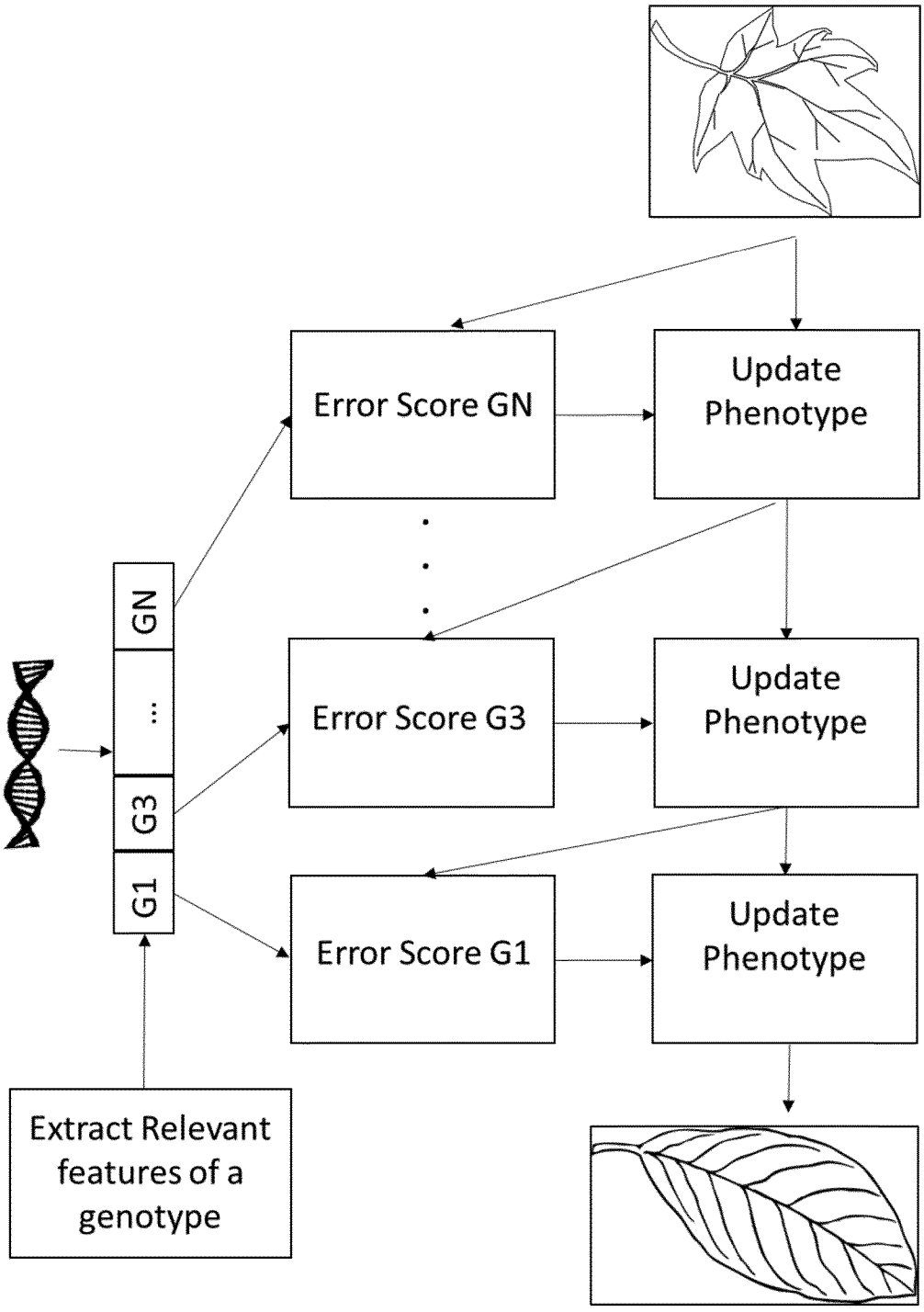

Error Score GN
Update Phenotype
GN
...
G3
G1
Error Score G3
Update Phenotype
Error Score G1
Update Phenotype
Extract Relevant features of a genotype

Fig. 35

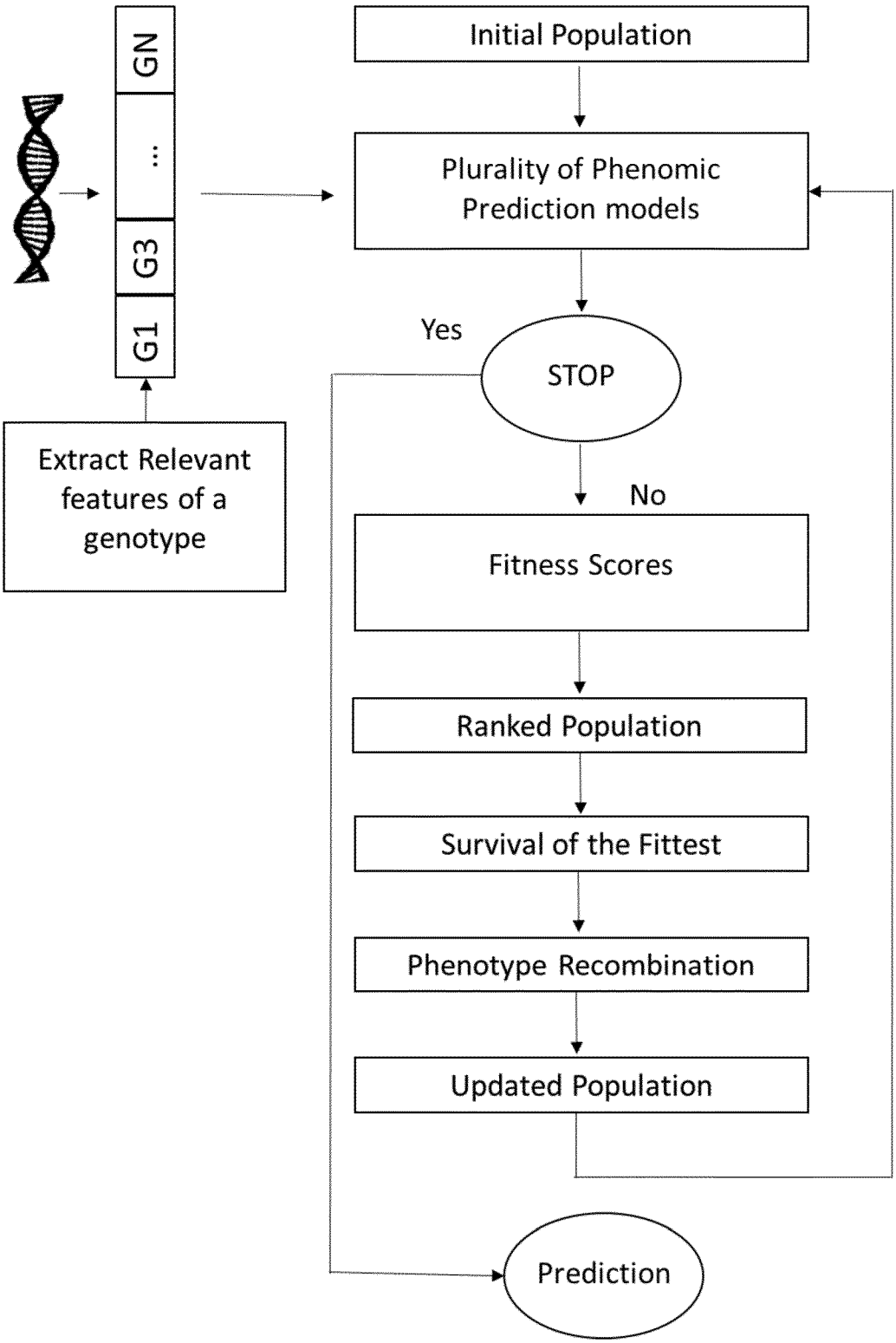

GN
...
G3
G1
Extract Relevant features of a genotype
Initial Population
Plurality of Phenomic Prediction models
Yes
STOP
No
Fitness Scores
Ranked Population
Survival of the Fittest
Phenotype Recombination
Updated Population
Prediction

Fig. 36

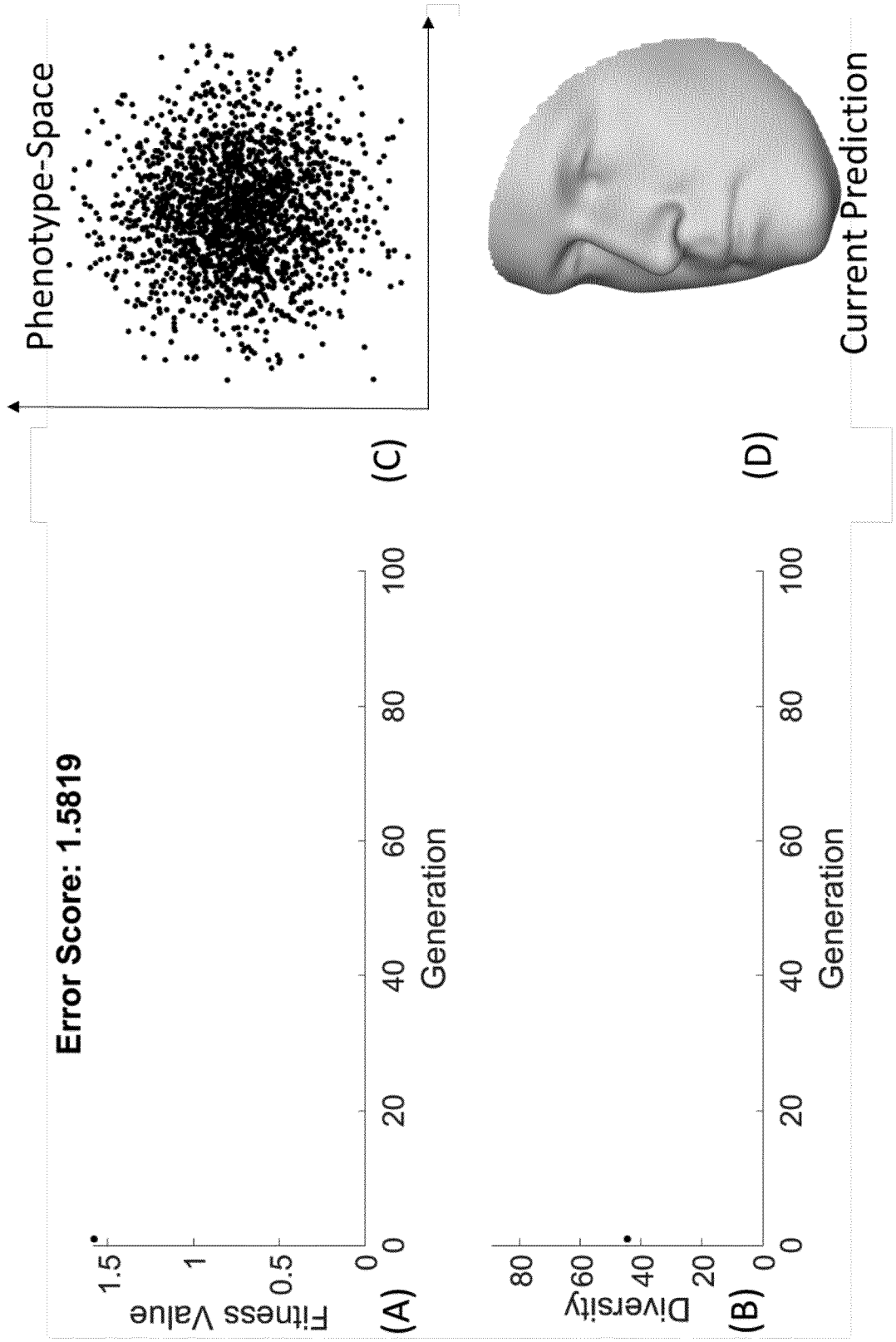

Phenotype-Space
(C)
Current Prediction
(D)
Error Score: 1.5819
Fitness Value
1.5
1
0.5
0
(A)
0
20
40
60
80
100
Generation
Diversity
80
60
40
20
0
(B)
0
20
40
60
80
100
Generation

Fig. 37

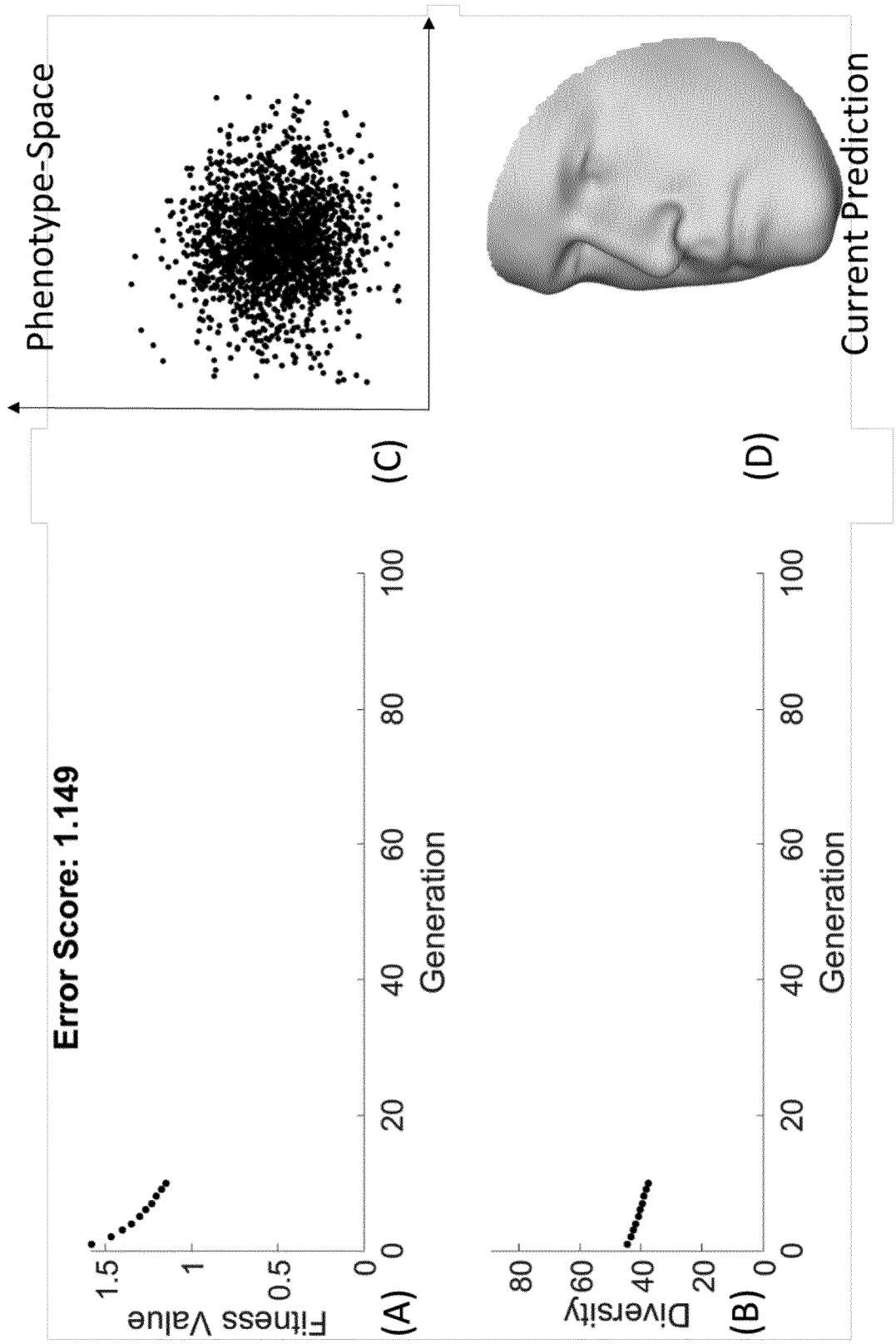

Phenotype-Space
(C)
Current Prediction
(D)
Error Score: 1.149
Fitness Value
1.5
1
0.5
0
(A)
0
20
40
60
80
100
Generation
Diversity
80
60
40
20
0
(B)
0
20
40
60
80
100
Generation

Fig. 38

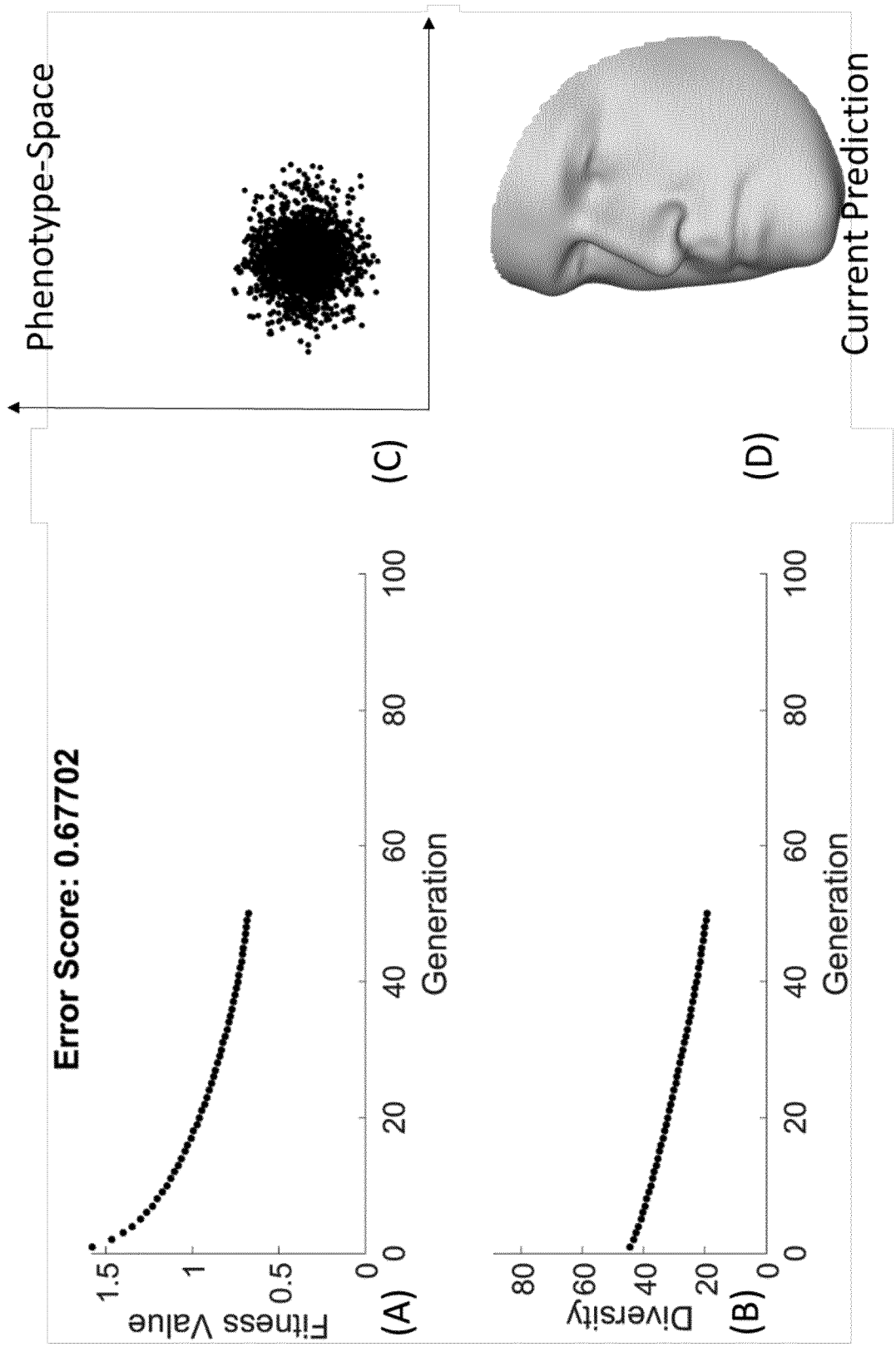

Phenotype-Space
(C)
Current Prediction
(D)
Error Score: 0.67702
Fitness Value
1.5
1
0.5
0
(A)
0
20
40
60
80
100
Generation
Diversity
80
60
40
20
0
(B)
0
20
40
60
80
100
Generation

Fig. 39

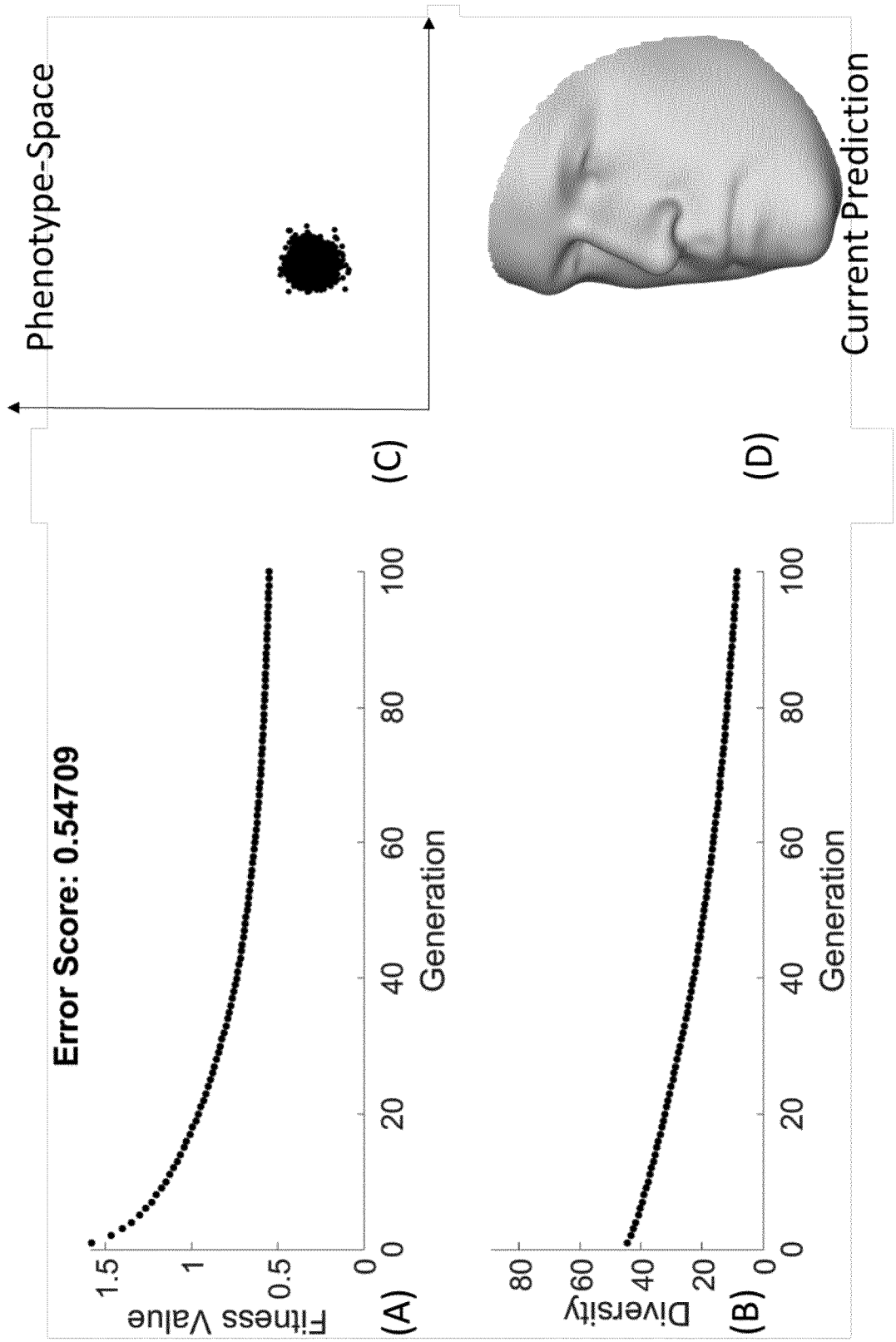

Phenotype-Space
Current Prediction
(C)
(D)
Error Score: 0.54709
Fitness Value
Generation
(A)
Diversity
Generation
(B)

Fig. 40

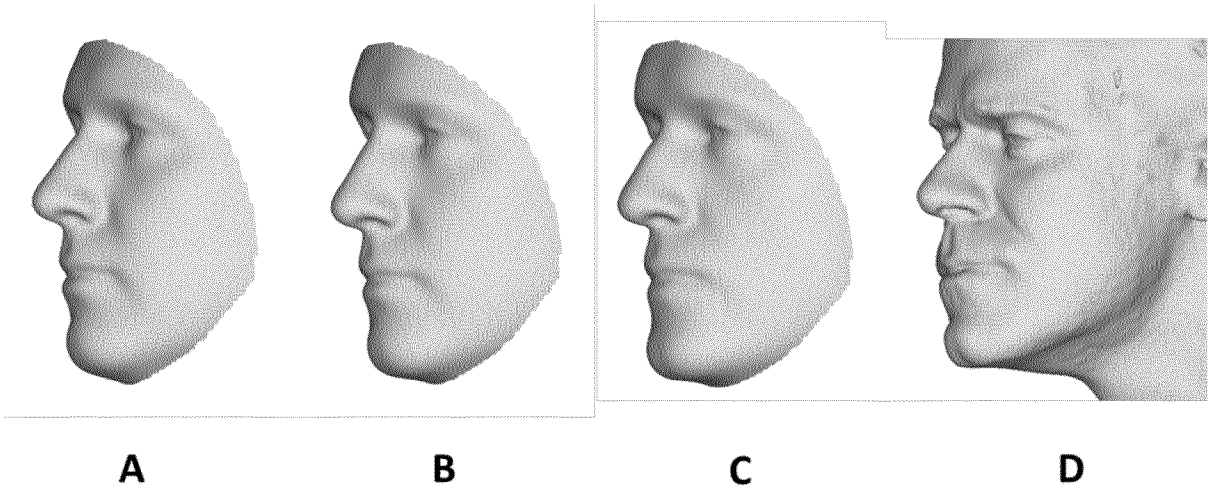


Fig. 41

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2015/060921

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F19/18
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PETER CLAES ET AL: "Modeling 3D Facial Shape from DNA", PLOS GENETICS, vol. 10, no. 3, 20 March 2014 (2014-03-20) , pages 1-14, XP055205937, DOI: 10.1371/journal.pgen.1004224 abstract page 2, column 1, paragraph 1 - column 2, paragraph 1 figures 1-3, 6 -----	1-25

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 August 2015	11/08/2015

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Schmitt, Constanze