

METHODS AND SYSTEMS FOR INCORPORATING MULTIPLE ENVIRONMENTAL AND GENETIC RISK FACTORS

CROSS REFERENCE TO RELATED APPLICATION

This application claims priority to U.S. Provisional Application No. 61/096,758, filed September 12, 2008, which application is incorporated herein by reference in its entirety.

BACKGROUND

[0001] The etiology of common diseases and conditions is typically attributed to both genetic and environmental factors. Recent advances in genotyping technology have greatly improved understanding of the genetic contribution for such diseases. Many whole-genome association studies have been completed recently, aiming to discover new associations between common diseases and common genetic variants across the genome. These studies have shed light on the mechanisms of disease and on the risk of an individual to develop a disease within their lifetime, based on their genetic composition. Integrating inherited genetic risk information into clinical decision-making process early in life can have an important effect in ameliorating or even preventing disease symptoms or conditions.

[0002] The prevalence of common chronic non-communicable diseases typically overshadows the prevalence of both monogenic and infectious diseases combined. Common SNP variants account for a portion of a significant number, if not all, germ-line genetic risk for common diseases and when used in context permit better personalized and focused exposure mitigation, early detection, and early intervention paradigms for individuals.

[0003] Genetic variations in the genome, such as single nucleotide polymorphisms (SNPs), mutations, deletions, insertions, repeats, microsatellites and others, are correlated to various phenotypes, such as a disease or condition. The genetic variations of an individual can be identified and correlated to determine the individual's predisposition or risk to different phenotypes, creating a personalized phenotype profile.

[0004] Low effect size common SNP variants, rare and private variants, DNA copy number variants, and epigenetic modification typically account for most of the inherited risk. Accurately estimating an individual's risk to develop a condition is a challenging task. The risk is determined by many factors, including the genetic risk factor load, environmental factors, gender, and age. Thus, for most conditions the most accurate risk assessment can only give a probabilistic risk estimate. Factors can include different associated variants, their effect sizes, their frequency in the population, the environmental factors affecting the individual, such as diet, age, family history, and ethnic background as well as their interactions. Large-scale studies that investigate all of these factors at once are prohibitively expensive to conduct, and to our knowledge, none have been conducted.

[0005] Thus, there exists a need for methods to generate personalized phenotype profiles with risk estimates that take into account the effects of genetic variations yet does not require the results of large-scale studies simultaneously assessing multiple risk factors. Furthermore, there exists a need for generating risk estimates that not only vary by disease, but can be combined with environmental data, providing an additional tool for clinical decision-making, such as having a predictive power as a clinical classifier. The present disclosure and embodiments disclosed herein satisfy these needs and provides related advantages as well.

SUMMARY

[0006] The present disclosure provides a method for generating an Environmental Genetic Composite Index (EGCI) score for a disease or condition for an individual. The method may comprise generating a genomic profile from a genetic sample of the individual; obtaining at least one environmental factor from the individual; generating an EGCI score from the genomic profile and at least one environmental factor; and, reporting the EGCI score to the individual or a health care manager of the individual. The method can further comprise updating the EGCI score with additional or modified environmental factors. In some embodiments, the method is performed by a computer. For example, the EGCI score is computed by a computer and the results can be obtained and outputted by the computer.

[0006a] In a first aspect, there is provided a method for generating a personalized action plan for an individual based on a score that represents risk of said individual to develop a phenotype, the method comprising:

(a) generating a genomic profile from a genetic sample of said individual, wherein said genomic profile is generated using a high density DNA microarray, RT-PCR, or DNA sequencing;

(b) analysing said genomic profile for genotype correlations to identify the odds ratio or relative risk for one or more genetic loci;

(c) obtaining at least one environmental factor from said individual, wherein said at least one environmental factor has a relative risk or odds ratio for said phenotype;

(d) generating a score from said odds ratio or relative risk in step (b) and said odds ratio or relative risk in step (c) using a computer; and

(e) generating a personalized action plan for said individual based on the score, wherein the personalized action plan includes treatments and preventative measures for the phenotype, wherein the treatments and preventative measures are selected based on the score.

[0007] The relative risk of the environmental factor for a disease or condition may be at least approximately 1. In some embodiments, the relative risk for the disease or condition is at least approximately 1.1, 1.2, 1.3, 1.4, or 1.5. The relative risk can be at least approximately 2, 3, 4, 5, 10, 12, 15, 20, 25, 30, 35, 40, 45, or 50. In some embodiments, the environmental factor has an odds ratio (OR) of at least approximately 1. In yet other embodiments, the OR is at least approximately 1.1, 1.2, 1.3, 1.4, or 1.5. The OR can be at least approximately 1.5, 2, 3, 4, 5, 10, 12, 15, 20, 25, 30, 35, 40, 45, or 50.

[0008] In another aspect, the environmental factor can be selected from the group consisting of: the individual's birthplace, location of residency, lifestyle conditions; diet, exercise habits, and personal relationships. For example, the lifestyle condition can be smoking or alcohol intake. In some embodiments, the environmental factor is a physical measurement of the individual, such as body mass index, blood pressure, heart rate, glucose level, metabolite level, ion level, weight, height, cholesterol level, vitamin level, blood cell count, protein level, or transcript level.

[0009] The EGCI score may be generated using at least 2 environmental factors and generating the EGCI score may assume that at least one, or more, of the environmental factor is an independent risk factor for said disease or condition.

[0010] In some embodiments, the EGCI score is generated for a disease or condition with a heritability of less than approximately 95%. In some embodiments, the disease or condition has a heritability of less than approximately 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, or 90%.

[0011] In another aspect, the methods disclosed herein can comprise a third party obtaining the genetic sample of the individual, or generating the genomic profile of the individual. The genetic sample can be DNA or RNA, and can be obtained from a biological sample such as blood, hair, skin, saliva, semen, urine, fecal material, sweat, or buccal sample.

[0012] The methods also comprise transmission of the EGCI score over a network, reporting of the EGCI through an on-line portal, by paper or by e-mail, through the use of a computer. The reporting can be by a secure or non-secure manner. The individual's genomic profile can be deposited into a secure database or vault, and be a single nucleotide polymorphism profile, or a genomic profile that comprises truncations, insertions, deletions, or repeats. The genomic profile can be generated using a high density DNA microarray, RT-PCR, or DNA sequencing. In some embodiments, the genomic profile is generated by amplifying a genetic sample from a subject or individual. Alternatively, the genomic profile can be generated without amplifying the genetic sample.

INCORPORATION BY REFERENCE

[0013] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The novel features of the embodiments disclosed herein are set forth with particularity in the appended claims. A better understanding of the features and advantages of the disclosure will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the disclosure herein are utilized, and the accompanying drawings of which:

[0015] **FIG. 1** illustrates ROC curves for **A)** Crohn's Disease, **B)** Type 2 Diabetes and **C)** Rheumatoid Arthritis. In each plot, the black line corresponds to random expectation, the purple and blue lines correspond to theoretical ((under the two disease models, further described below) expectations when the genetic variable is known, the yellow line corresponds to GCI, and the green line corresponds to logistic regression.

[0016] **FIG. 2** illustrates ROC curves for a model with interaction and for simple multiplicative model for **A)** Crohn's Disease, **B)** Rheumatoid Arthritis, and **C)** Type 2 Diabetes. In each plot, 6,400 threshold points are used.

[0017] **FIG. 3** depicts **A)** comparison of odds ratios and relative risks for Type 2 Diabetes with lifetime risk of 25% and heritability of 64%, **B)** comparison of odds ratios and relative risks for Myocardial Infraction with lifetime risk of 42% and heritability of 57%, and **C)** mean squared error versus the probability of getting the disease for Type 2 Diabetes.

[0018] **FIG. 4** illustrates known family history versus known genetic risk. Family history versus theoretical ROC curve where genetic risk is completely known for **A)** Type 2 Diabetes, **B)** Crohn's Disease, and **C)** Rheumatoid Arthritis. Red curves show true and false positive fractions for different values of b for a classification test based on family history alone.

[0019] **FIG. 5** illustrates effect of known genetic and environmental factors versus known genetic factors alone for **A)** Crohn's Disease, **B)** Type 2 Diabetes, and **C)** Rheumatoid Arthritis. For Crohn's Disease, the AUC of the two curves 0.68 and 0.72 (**A**). In addition to the genetic factors, smoking (relative risk 3) is considered as the environmental variable. For Type 2 Diabetes, the AUC of the two curves are 0.57 and 0.79 respectively (**B**). In addition to the genetic factors, Body Mass Index (relative risk 42.1), alcohol intake (relative risk 1.75) and smoking frequency (relative risk 1.70) are taken as environmental factors for Type 2 Diabetes. For Rheumatoid Arthritis, the AUC of the two curves are 0.685 and 0.688 (**C**). Smoking (relative risk 1.4) is the environmental variable in addition to the genetic factors.

[0020] **FIG. 6. A)** The error between GCI-based average lifetime risk and true average risk as a function of the assumed lifetime risk (LTR') for GCI calculations in Type 2 Diabetes. The true average risk for T2D = 0.25. **B)** The error between GCI-based average lifetime risk and the lifetime risk (LTR') is assumed for the GCI calculations as a function of the assumed LTR' .

DETAILED DESCRIPTION

[0021] The present disclosure provides a method of developing a risk estimate that is based on the genetic composition of an individual alone, of their genomic profile. In some embodiments, the estimate is based on the

individual's genomic profile or genetic composition alone, and all other factors are fixed. The risk estimate or risk score, as described herein is referred to as the Genetic Composite Index (GCI), a scalable metric that can be used in a clinical setting with any type of genetic risk factor input that will guide clinical decisions, such as decisions for the future. The GCI combines the information of an individual's genotypes with the average lifetime risk, the odds ratio information across multiple risk loci, and the distribution of genotype frequencies in a reference population into one consolidated score that represents the risk of an individual to develop a condition. A higher GCI score can be intuitively interpreted as an increased risk for a condition. The GCI is based on several assumptions, further described below. Simulated data as well as real genotype and clinical data to test the robustness of the GCI under different conditions is also described herein. In some embodiments, the effects of SNPs are independent unless there is a known SNP-SNP interaction that has been shown to be statistically significant in the literature. This assumption of independence typically does not affect the generality of our model, as weak SNP-SNP interactions do not typically significantly affect its predictability.

[0022] Current risk assessment methods provide starting points in the development of risk assessment measures of use in preventative medicine programs. However, the quality and effectiveness of these different methods depend on their derivation and implementation, their theoretical limitations, and their relative merits. For example, the Receiver Operating Characteristic (ROC) curves are used to measure the effectiveness of various risk measures (see for example, *Lu and Elston, Am. J. of Human Genetics, 82:641-651 (2008)*).

[0023] ROC curves can also be used to evaluate GCI scores, for example, by showing that the GCI can be a theoretically optimal test, and other risk assessment methods. For example, different disease models can be simulated to calculate the predictive power of such different methods (GCI versus other models for example) under an ideal "best case" scenario, in which all genetic factors are known. This ideal risk assessment depends on a few factors, among them the heritability and the average lifetime risk of developing the condition. Typically, the higher the heritability, the better the risk assessment based on genotypic information alone. Similarly, the average lifetime risk generally affects the variability of the risk probability among the population, and thus affects the accuracy of the ideal risk assessment scenario. Furthermore, the GCI as described herein can be used when multiple factors, such as genetic factors or environmental factors are not available, such as when large-scale studies designed to simultaneously test multiple factors are not available, such as for a number of common diseases.

Genomic Profile

[0024] The GCI is generated based on an individual's genomic profile. An individual's genomic profile contains information about an individual's genes based on genetic variations or markers. Genetic variations can form genotypes, which make up genomic profiles. Such genetic variations or markers include, but are not limited to, single nucleotide polymorphisms (SNPs), single and/or multiple nucleotide repeats, single and/or multiple nucleotide deletions, microsatellite repeats (small numbers of nucleotide repeats with a typical 5-1,000 repeat units), di-nucleotide repeats, tri-nucleotide repeats, sequence rearrangements (including translocation and duplication), copy number variations (both loss and gains at specific loci), and the like. Other genetic variations include chromosomal duplications and translocations, as well as centromeric and telomeric repeats.

[0025] Genotypes may also include haplotypes and diplotypes. In some embodiments, genomic profiles may have at least 100,000, 300,000, 500,000, or 1,000,000 genotypes. In some embodiments, the genomic profile may be substantially the complete genomic sequence of an individual. In other embodiments, the genomic profile is at least

60%, 80%, or 95% of the complete genomic sequence of an individual. The genomic profile may be approximately 100% of the complete genomic sequence of an individual. Genetic samples that contain the targets include, but are not limited to, unamplified genomic DNA or RNA samples or amplified DNA (or cDNA). The targets may be particular regions of genomic DNA that contain genetic markers of particular interest.

[0026] To obtain a genomic profile, a genetic sample of an individual can be isolated from a biological sample of an individual. The biological sample includes samples from which genetic material, such as RNA and/or DNA, may be isolated. Such biological samples can include, but not be limited to, blood, hair, skin, saliva, semen, urine, fecal material, sweat, buccal, and various bodily tissues. Tissues samples may be directly collected by the individual, for example, a buccal sample can be obtained by the individual taking a swab against the inside of their cheek. Other samples such as saliva, semen, urine, fecal material, or sweat, may also be supplied by the individual themselves. Other biological samples may be taken by a health care specialist, such as a phlebotomist, nurse or physician. For example, blood samples may be withdrawn from an individual by a nurse. Tissue biopsies may be performed by a health care specialist, and commercial kits are also readily available to health care specialists to efficiently obtain samples. A small cylinder of skin may be removed or a needle may be used to remove a small sample of tissue or fluids.

[0027] Sample collection kits can also be provided to individuals. The kits can contain sample collection containers for the individual's biological sample. The kit may also provide instructions for an individual to directly collect their own sample, such as how much hair, urine, sweat, or saliva to provide. The kit may also contain instructions for an individual to request tissue samples to be taken by a health care specialist. The kit may include locations where samples may be taken by a third party, for example, kits may be provided to health care facilities who in turn collect samples from individuals. The kit may also provide return packaging for the sample to be sent to a sample processing facility, where genetic material is isolated from the biological sample.

[0028] A genetic sample of DNA or RNA can be isolated from a biological sample according to any of several well-known biochemical and molecular biological methods, see, e.g., *Sambrook, et al., Molecular Cloning: A Laboratory Manual (Cold Spring Harbor Laboratory, New York) (1989)*. There are also several commercially available kits and reagents for isolating DNA or RNA from biological samples, such as, but not limited to, those available from DNA Genotek, Gentra Systems, Qiagen, Ambion, and other suppliers. Buccal sample kits are readily available commercially, such as the MasterAmp™ Buccal Swab DNA extraction kit from Epicentre Biotechnologies, as are kits for DNA extraction from blood samples such as Extract-N-Amp™ from Sigma Aldrich. DNA from other tissues may be obtained by digesting the tissue with proteases and heat, centrifuging the sample, and using phenol-chloroform to extract the unwanted materials, leaving the DNA in the aqueous phase. The DNA can then be further isolated by ethanol precipitation.

[0029] For example, genomic DNA can be isolated from saliva, using a DNA self collection kit from DNA Genotek. An individual can collect a specimen of saliva for clinical processing using the kit and the sample can conveniently be stored and shipped at room temperature. After delivery of the sample to an appropriate laboratory for processing, DNA is isolated by heat denaturing and protease digesting the sample, typically using reagents supplied by the collection kit supplier at 50°C for at least one hour. The sample is next centrifuged, and the supernatant is ethanol precipitated. The DNA pellet is suspended in a buffer appropriate for subsequent analysis.

[0030] RNA may be used as the genetic sample, for example, genetic variations that are expressed can be identified from mRNA. mRNA includes, but is not limited to pre-mRNA transcript(s), transcript processing intermediates, mature mRNA(s) ready for translation and transcripts of the gene or genes, or nucleic acids derived from the mRNA transcript(s). Transcript processing may include splicing, editing and degradation. As used herein, a nucleic acid derived from an mRNA transcript refers to a nucleic acid for whose synthesis the mRNA transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from an mRNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, etc., are all derived from the mRNA transcript. RNA can be isolated from any of several bodily tissues using methods known in the art, such as isolation of RNA from unfractionated whole blood using the PAXgene™ Blood RNA System available from PreAnalytiX. Typically, mRNA is used to reverse transcribe cDNA, which is then used or amplified for gene variation analysis.

[0031] A genomic profile may be generated from a genetic sample without amplification of the genetic sample. Alternatively, prior to genomic profile analysis, a genetic sample may be amplified, either from DNA or cDNA reverse transcribed from RNA. DNA can be amplified by a number of methods, many of which employ PCR. See, for example, *PCR Technology: Principles and Applications for DNA Amplification* (Ed. H. A. Erlich, Freeman Press, NY, N.Y., 1992); *PCR Protocols: A Guide to Methods and Applications* (Eds. Innis, et al., Academic Press, San Diego, Calif., 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (Eds. McPherson et al., IRL Press, Oxford); and U.S. Pat. Nos. 4,683,202, 4,683,195, 4,800,159 4,965,188, and 5,333,675, and each of which is incorporated herein by reference in their entireties for all purposes.

[0032] Other suitable amplification methods include the ligase chain reaction (LCR) (for example, Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988) and Barringer et al. *Gene* 89:117 (1990)), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86:1173-1177 (1989) and WO88/10315), self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA*, 87:1874-1878 (1990) and WO90/06995), selective amplification of target polynucleotide sequences (U.S. Pat. No. 6,410,276), consensus sequence primed polymerase chain reaction (CP-PCR) (U.S. Pat. No. 4,437,975), arbitrarily primed polymerase chain reaction (AP-PCR) (U.S. Pat. Nos. 5,413,909, 5,861,245) nucleic acid sequence based amplification (NASBA), rolling circle amplification (RCA), multiple displacement amplification (MDA) (U.S. Pat. Nos. 6,124,120 and 6,323,009) and circle-to-circle amplification (C2CA) (Dahl et al. *Proc. Natl. Acad. Sci* 101:4548-4553 (2004)). (See, U.S. Pat. Nos. 5,409,818, 5,554,517, and 6,063,603, each of which is incorporated herein by reference). Other amplification methods that may be used are described in, U.S. Pat. Nos. 5,242,794, 5,494,810, 5,409,818, 4,988,617, 6,063,603 and 5,554,517 and in U.S. Ser. No. 09/854,317, each of which is incorporated herein by reference.

[0033] Generation of a genomic profile can be performed using any of several methods. Several methods are known in the art to identify genetic variations, and include, but are not limited to, DNA sequencing by any of several methodologies, PCR based methods, fragment length polymorphism assays (restriction fragment length polymorphism (RFLP), cleavage fragment length polymorphism (CFLP)) hybridization methods using an allele-specific oligonucleotide as a template (e.g., TaqMan assays and microarrays, further described herein), methods using a primer extension reaction, mass spectrometry (such as, MALDI-TOF/MS method), and the like, such as described in Kwok, *Pharmacogenomics* 1:95-100 (2000). Other methods include invader methods, such as

monoplex and biplex invader assays (e.g. available from Third Wave Technologies, Madison, WI and described in *Olivier et al., Nucl. Acids Res. 30:e53 (2002)*).

[0034] For example, a high density DNA array can be used to generate a genomic profile. Such arrays are commercially available from Affymetrix and Illumina (see Affymetrix GeneChip® 500K Assay Manual, Affymetrix, Santa Clara, CA (incorporated by reference); Sentrix® humanHap650Y genotyping beadchip, Illumina, San Diego, CA). A high density array can be used to generate a genomic profile that comprises genetic variations that are SNPs. For example, a SNP profile can be generated by genotyping more than 900,000 SNPs using the Affymetrix Genome Wide Human SNP Array 6.0. Alternatively, more than 500,000 SNPs through whole-genome sampling analysis may be determined by using the Affymetrix GeneChip Human Mapping 500K Array Set. In these assays, a subset of the human genome is amplified through a single primer amplification reaction using restriction enzyme digested, adaptor-ligated human genomic DNA. Typically, the amplified DNA is then fragmented and the quality of the sample determined prior denaturing and labeling the sample for hybridization to a microarray with DNA probes at specific locations on a coated quartz surface. The amount of label that hybridizes to each probe as a function of the amplified DNA sequence is monitored, thereby yielding sequence information and resultant SNP genotyping.

[0035] Use of high density arrays is well known in the arts, and if obtained commercially, is carried out according to the manufacturer's directions. For example, use of Affymetrix GeneChip can involve digesting isolated genomic DNA with either a NspI or StyI restriction endonuclease. The digested DNA is then ligated with a NspI or StyI adaptor oligonucleotide that respectively anneals to either the NspI or StyI restricted DNA. The adaptor-containing DNA following ligation is then amplified by PCR to yield amplified DNA fragments between about 200 and 1100 base pairs, as confirmed by gel electrophoresis. PCR products that meet the amplification standard are purified and quantified for fragmentation. The PCR products are fragmented with DNase I for optimal DNA chip hybridization. Following fragmentation, DNA fragments should be less than 250 base pairs, and on average, about 180 base pairs, as confirmed by gel electrophoresis. Samples that meet the fragmentation standard are then labeled with a biotin compound using terminal deoxynucleotidyl transferase. The labeled fragments are next denatured and then hybridized into a GeneChip 250K array. Following hybridization, the array is stained prior to scanning in a three step process consisting of a streptavidin phycoerythrin (SAPE) stain, followed by an antibody amplification step with a biotinylated, anti-streptavidin antibody (goat), and final stain with streptavidin phycoerythrin (SAPE). After labeling, the array is covered with an array holding buffer and then scanned, for example with a scanner such as the Affymetrix GeneChip Scanner 3000.

[0036] Analysis of data following scanning high density array can be performed according to the manufacturer's guidelines. For example, with the Affymetrix GeneChip, acquisition of raw data can be by use of the GeneChip Operating Software (GCOS) or by using Affymetrix GeneChip Command Console™. The acquisition of raw data is then followed by analysis with GeneChip Genotyping Analysis Software (GTYPE). Samples with a GTYPE call rate of less than a certain percentage may be excluded. For example, a call rate of less than approximately 70, 75, 80, 85, 90, or 95% may be excluded. Samples are then examined with BRLMM and/or SNiPer algorithm analyses. Samples with a BRLMM call rate of less than 95% or a SNiPer call rate of less than 98% are excluded. Finally, an association analysis is performed, and samples with a SNiPer quality index of less than 0.45 and/or a Hardy-Weinberg p-value of less than 0.00001 are excluded.

[0037] As an alternative to or in addition to DNA microarray analysis, genetic variations such as SNPs and mutations can be detected by other hybridization based methods, such as the use of TaqMan methods and variations thereof. TaqMan PCR, iterative TaqMan, and other variations of real time PCR (RT-PCR), such as those described in *Livak et al., Nature Genet., 9, 341–32 (1995)* and *Ranade et al. Genome Res., 11, 1262–1268 (2001)* can be used in the methods disclosed herein. In some embodiments, probes for specific genetic variations, such as SNPs, are labeled to form TaqMan probes. The probes are typically approximately at least 12, 15, 18 or 20 base pairs in length. They may be between approximately 10 and 70, 15 and 60, 20 and 60, or 18 and 22 base pairs in length. The probe is labeled with a reporter label, such as a fluorophore, at the 5' end and a quencher of the label at the 3' end. The reporter label may be any fluorescent molecule that has its fluorescence inhibited or quenched when in close proximity, such as the length of the probe, to the quencher. For example, the reporter label can be a fluorophore such as 6-carboxyfluorescein (FAM), tetrachlorofluorescein (TET), or derivatives thereof, and the quencher tetramethylrhodamine (TAMRA), dihydrocyclopyrroloindole tripeptide (MGB), or derivatives thereof.

[0038] As the reporter fluorophore and quencher are in close proximity, separated by the length of the probe, the fluorescence is quenched. When the probe anneals to a target sequence, such as a sequence comprising a SNP in a sample, DNA polymerase with 5' to 3' exonuclease activity, such as Taq polymerase, can extend the primer and the exonuclease activity cleaves the probe, separating the reporter from the quencher, and thus the reporter can fluoresce. The process can be repeated, such as in RT-PCR. The TaqMan probe is typically complementary to a target sequence that is located between two primers that are designed to amplify a sequence. Thus, the accumulation of PCR product can be correlated to the accumulation of released fluorophore, as each probe can hybridize to newly generated PCR product. The released fluorophore can be measured and the amount of target sequence present can be determined. RT-PCR methods for high throughput genotyping, such as in

[0039] Genetic variations can also be identified by DNA sequencing. DNA sequencing may be used to sequence a substantial portion, or the entire, genomic sequence of an individual. Traditionally, common DNA sequencing has been based on polyacrylamide gel fractionation to resolve a population of chain-terminated fragments (*Sanger et al., Proc. Natl. Acad. Sci. USA 74:5463–5467 (1977)*). Alternative methods have been and continue to be developed to increase the speed and ease of DNA sequencing. For example, high throughput and single molecule sequencing platforms are commercially available or under development from 454 Life Sciences (Branford, CT) (*Margulies et al., Nature 437:376–380 (2005)*); Solexa/Illumina (Hayward, CA); Helicos BioSciences Corporation (Cambridge, MA) (U.S. application Ser. No. 11/167046, filed June 23, 2005), and Li-Cor Biosciences (Lincoln, NE) (U.S. application Ser. No. 11/118031, filed April 29, 2005).

[0040] After an individual's genomic profile is generated, the profile is stored digitally, such as on a computer readable medium. The profile may be stored digitally in a secure manner. The genomic profile is encoded in a computer readable format to be stored as part of a data set, such as on a computer readable medium and may be stored as a database, where the genomic profile may be "banked", and can be accessed again later. The data set comprises a plurality of data points, wherein each data point relates to an individual. Each data point may have a plurality of data elements. One data element is the unique identifier, used to identify the individual's genomic profile. The unique identifier may be a bar code. Another data element is genotype information, such as the SNPs or nucleotide sequence of the individual's genome. Data elements corresponding to the genotype information may also be included in the data point. For example, if the genotype information includes SNPs identified by microarray

analysis, other data elements may include the microarray SNP identification number. Alternatively, if the genotype information was identified by other means, such as by RT-PCR methods (such as TaqMan assays), the data element may include level of fluorescence, primer information, and probe sequence. Other data elements may include, but not be limited to, SNP rs number, polymorphic nucleotide, chromosome position of the genotype information, quality metrics of the data, raw data files, images of the data, and extracted intensity scores.

[0041] The individual's specific factors such as physical data, medical data, ethnicity, ancestry, geography, gender, age, family history, known phenotypes, demographic data, exposure data, lifestyle data, behavior data, and other known phenotypes may also be incorporated as data elements. For example, factors may include, but are not limited to, an individual's birthplace, parents and/or grandparents, relatives' ancestry, location of residence, ancestors' location of residence, environmental conditions, known health conditions, known drug interactions, family health conditions, lifestyle conditions, diet, exercise habits, marital status, and physical measurements, such as weight, height, cholesterol level, heart rate, blood pressure, glucose level and other measurements known in the art. The above mentioned factors for an individual's relatives or ancestors, such as parents and grandparents, may also be incorporated as data elements and used to determine an individual's risk for a phenotype or condition.

[0042] The specific factors may be obtained from a questionnaire or from a health care manager of the individual. Information from the "banked" profile can then be accessed and utilized as desired. For example, in the initial assessment of an individual's genotype correlations, the individual's entire information (typically SNPs or other genomic sequences across, or taken from an entire genome) will be analyzed for genotype correlations. In subsequent analyses, either the entire information can be accessed, or a portion thereof, from the stored, or banked genomic profile, as desired or appropriate.

Correlations and Phenotype Profiles

[0043] The genomic profile is used to generate phenotype profiles. The genomic profile is typically stored digitally and is readily accessed at any point of time to generate phenotype profiles. Phenotype profiles are generated by applying rules that correlate or associate genotypes with phenotypes. Typically the rules are applied using a computer. Rules can be made based on scientific research that demonstrates a correlation between a genotype and a phenotype. The correlations may be curated or validated by a committee of one or more experts. By applying the rules to a genomic profile of an individual, the association between an individual's genotype and a phenotype may be determined. The phenotype profile for an individual will have this determination. The determination may be a positive association between an individual's genotype and a given phenotype, such that the individual has the given phenotype, or will develop the phenotype. Alternatively, it may be determined that the individual does not have, or will not develop, a given phenotype. In other embodiments, the determination may be a risk factor, estimate, or a probability that an individual has, or will develop a phenotype.

[0044] The determinations may be made based on a number of rules, for example, a plurality of rules may be applied to a genomic profile to determine the association of an individual's genotype with a specific phenotype. The determinations may also incorporate factors that are specific to an individual, such as ethnicity, gender, lifestyle (for example, diet and exercise habits), age, environment (for example, location of residence), family medical history, personal medical history, and other known phenotypes. The incorporation of the specific factors may be by modifying existing rules to encompass these factors. Alternatively, separate rules may be generated by these factors and applied to a phenotype determination for an individual after an existing rule has been applied.

[0045] Phenotypes may include any measurable trait or characteristic, such as susceptibility to a certain disease or response to a drug treatment. Other phenotypes that may be included are physical and mental traits, such as height, weight, hair color, eye color, sunburn susceptibility, size, memory, intelligence, level of optimism, and general disposition. Phenotypes may also include genetic comparisons to other individuals or organisms. For example, an individual may be interested in the similarity between their genomic profile and that of a celebrity. They may also have their genomic profile compared to other organisms such as bacteria, plants, or other animals. Together, the collection of correlated phenotypes determined for an individual comprises the phenotype profile for the individual.

[0046] Correlations between genetic variations and phenotypes can be obtained from scientific literature.

Correlations for genetic variations are determined from analysis of a population of individuals who have been tested for the presence or absence of one or more phenotypic traits of interest and their genotype profile. The alleles of each genetic variation or polymorphism in the profile are reviewed to determine whether the presence or absence of a particular allele is associated with a trait of interest. Correlation can be performed by standard statistical methods and statistically significant correlations between genetic variations and phenotypic characteristics are noted. For example, it may be determined that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased risk of cancer. The results of the analyses may be published in peer-reviewed literature, validated by other research groups, and/or analyzed by a committee of experts, such as geneticists, statisticians, epidemiologists, and physicians, and may also be curated. For example, correlations disclosed in US Publication No. 20080131887 and PCT Publication No. WO/2008/067551, both of which are hereby incorporated in its entirety, may be used in the embodiments described herein.

[0047] Alternatively, the correlations may be generated from the stored genomic profiles. For example, individuals with stored genomic profiles may also have known phenotype information stored as well. Analysis of the stored genomic profiles and known phenotypes may generate a genotype correlation. As an example, 250 individuals with stored genomic profiles also have stored information that they have previously been diagnosed with diabetes. Analysis of their genomic profiles is performed and compared to a control group of individuals without diabetes. It is then determined that the individuals previously diagnosed with diabetes have a higher rate of having a particular genetic variant compared to the control group, and a genotype correlation may be made between that particular genetic variant and diabetes.

[0048] Rules are made based on the validated correlations of genetic variants to particular phenotypes. Rules may be generated based on the genotypes and phenotypes correlated as disclosed in US Publication No. 20080131887 and PCT Publication No. WO/2008/067551, and some rules may incorporate other factors such as gender or ethnicity to generate effects estimates. Other measures resulting from rules may be estimated relative risk increase. The effects estimates and estimated relative risk increase may be from the published literature, or calculated from the published literature. Alternatively, the rules may be based on correlations generated from stored genomic profiles and previously known phenotypes.

[0049] Genetic variants may include SNPs. While SNPs occur at a single site, individuals who carry a particular SNP allele at one site often predictably carry specific SNP alleles at other sites. A correlation of SNPs and an allele predisposing an individual to disease or condition occurs through linkage disequilibrium, in which the non-random

association of alleles at two or more loci occur more or less frequently in a population than would be expected from random formation through recombination.

[0050] Other genetic markers or variants, such as nucleotide repeats or insertions, may also be in linkage disequilibrium with genetic markers that have been shown to be associated with specific phenotypes. For example, a nucleotide insertion is correlated with a phenotype and a SNP is in linkage disequilibrium with the nucleotide insertion. A rule is made based on the correlation between the SNP and the phenotype. A rule based on the correlation between the nucleotide insertion and the phenotype may also be made. Either rules or both rules may be applied to a genomic profile, as the presence of one SNP may give a certain risk factor, the other may give another risk factor, and when combined may increase the risk.

[0051] Through linkage disequilibrium, a disease predisposing allele cosegregates with a particular allele of a SNP or a combination of particular alleles of SNPs. A particular combination of SNP alleles along a chromosome is termed a haplotype, and the DNA region in which they occur in combination can be referred to as a haplotype block. While a haplotype block can consist of one SNP, typically a haplotype block represents a contiguous series of 2 or more SNPs exhibiting low haplotype diversity across individuals and with generally low recombination frequencies. An identification of a haplotype can be made by identification of one or more SNPs that lie in a haplotype block. Thus, a SNP profile typically can be used to identify haplotype blocks without necessarily requiring identification of all SNPs in a given haplotype block.

[0052] Genotype correlations between SNP haplotype patterns and diseases, conditions or physical states are increasingly becoming known. For a given disease, the haplotype patterns of a group of people known to have the disease are compared to a group of people without the disease. By analyzing many individuals, frequencies of polymorphisms in a population can be determined, and in turn these frequencies or genotypes can be associated with a particular phenotype, such as a disease or a condition. Examples of known SNP-disease correlations include polymorphisms in Complement Factor H in age-related macular degeneration (*Klein et al., Science: 308:385-389, (2005)*) and a variant near the *INSIG2* gene associated with obesity (*Herbert et al., Science: 312:279-283 (2006)*). Other known SNP correlations include polymorphisms in the 9p21 region that includes CDKN2A and B, such as) such as rs10757274, rs2383206, rs13333040, rs2383207, and rs10116277 correlated to myocardial infarction (*Helgadottir et al., Science 316:1491-1493 (2007); McPherson et al., Science 316:1488-1491 (2007)*)

[0053] The SNPs may be functional or non-functional. For example, a functional SNP has an effect on a cellular function, thereby resulting in a phenotype, whereas a non-functional SNP is silent in function, but may be in linkage disequilibrium with a functional SNP. The SNPs may also be synonymous or non-synonymous. SNPs that are synonymous are SNPs in which the different forms lead to the same polypeptide sequence, and are non-functional SNPs. If the SNPs lead to different polypeptides, the SNP is non-synonymous and may or may not be functional. SNPs, or other genetic markers, used to identify haplotypes in a diplotype, which is 2 or more haplotypes, may also be used to correlate phenotypes associated with a diplotype. Information about an individual's haplotypes, diplotypes, and SNP profiles may be in the genomic profile of the individual.

[0054] Typically, for a rule to be generated based on a genetic marker in linkage disequilibrium with another genetic marker that is correlated with a phenotype, the genetic marker has a r^2 or D' score (scores commonly used in the art to determine linkage disequilibrium) of greater than 0.5. The score can be greater than approximately 0.5, 0.6, 0.7, 0.8, 0.90, 0.95 or 0.99. As a result, the genetic marker used to correlate a phenotype to an individual's

genomic profile may be the same as the functional or published SNP correlated to a phenotype, or different. In some embodiments, the test SNP may not yet be identified, but using the published SNP information, allelic differences or SNPs may be identified based on another assay, such as TaqMan. For example, a published SNP is rs1061170 but a test SNP has not been identified. The test SNP may be identified by LD analysis with the published SNP. Alternatively, the test SNP may not be used, and instead, TaqMan or other comparable assay, will be used to assess an individual's genome having the test SNP.

[0055] The test SNPs may be "DIRECT" or "TAG" SNPs. Direct SNPs are the test SNPs that are the same as the published or functional SNP. For example, the direct SNP may be used for FGFR2 correlation with breast cancer, using the SNP rs1073640 in Europeans and Asians, where the minor allele is A and the other allele is G (*Easton et al., Nature 447:1087-1093 (2007)*). Another published or functional SNP that can be a direct SNP for FGFR2 correlation to breast cancer is rs1219648, also in Europeans and Asians (*Hunter et al., Nat. Genet. 39:870-874 (2007)*). Tag SNPs are where the test SNP is different from that of the functional or published SNP. Tag SNPs may also be used for other genetic variants such as SNPs for CAMTA1 (rs4908449), 9p21 (rs10757274, rs2383206, rs13333040, rs2383207, rs10116277), COL1A1 (rs1800012), FVL (rs6025), HLA-DQA1 (rs4988889, rs2588331), eNOS (rs1799983), MTHFR (rs1801133), and APC (rs28933380).

[0056] Databases of SNPs are publicly available from, for example, the International HapMap Project (see www.hapmap.org, *The International HapMap Consortium, Nature 426:789-796 (2003)*, and *The International HapMap Consortium, Nature 437:1299-1320 (2005)*), the Human Gene Mutation Database (HGMD) public database (see www.hgmd.org), and the Single Nucleotide Polymorphism database (dbSNP) (see www.ncbi.nlm.nih.gov/SNP/). These databases provide SNP haplotypes, or enable the determination of SNP haplotype patterns. Accordingly, these SNP databases enable examination of the genetic risk factors underlying a wide range of diseases and conditions, such as cancer, inflammatory diseases, cardiovascular diseases, neurodegenerative diseases, and infectious diseases. The diseases or conditions may be actionable, in which treatments and therapies currently exist. Treatments may include prophylactic treatments as well as treatments that ameliorate symptoms and conditions, including lifestyle changes.

[0057] Many other phenotypes such as physical traits, physiological traits, mental traits, emotional traits, ethnicity, ancestry, and age may also be examined. Physical traits may include height, hair color, eye color, body, or traits such as stamina, endurance, and agility. Mental traits may include intelligence, memory performance, or learning performance. Ethnicity and ancestry may include identification of ancestors or ethnicity, or where an individual's ancestors originated from. The age may be a determination of an individual's real age, or the age in which an individual's genetics places them in relation to the general population. For example, an individual's real age is 38 years of age, however their genetics may determine their memory capacity or physical well-being may be of the average 28 year old. Another age trait may be a projected longevity for an individual.

[0058] Other phenotypes may also include non-medical conditions, such as "fun" phenotypes. These phenotypes may include comparisons to well known individuals, such as foreign dignitaries, politicians, celebrities, inventors, athletes, musicians, artists, business people, and infamous individuals, such as convicts. Other "fun" phenotypes may include comparisons to other organisms, such as bacteria, insects, plants, or non-human animals. For example, an individual may be interested to see how their genomic profile compares to that of their pet dog, or to a former president.

[0059] The rules are applied to the stored genomic profile to generate a phenotype profile. For example, correlation data from published sources, or from stored genomic profiles can form the basis of rules or tests, to apply to an individual's genomic profile. The rules may encompass the information on test SNP and alleles, and the effect estimates, such as OR, or odds-ratio (95% confidence interval) or mean. The effects estimate may be a genotypic risk, such as the risk for homozygotes (homozygous or RR), risk heterozygotes (heterozygous or RN), and nonrisk homozygotes (homozygous or NN). The effect estimate can also be carrier risk, which is RR or RN vs NN. The effect estimate may be based on the allele, such as an allelic risk, an example being R vs. N. There may also be 2, 3, 4, or more loci genotypic effect estimates (e.g. RRRR, RRNN, etc for the 9 possible genotype combinations for a two locus effect estimate).

[0060] The estimated risk for a condition may be based on the SNPs as listed in US Publication No. 20080131887 and PCT Publication No. WO/2008/067551. In some embodiments, the risk for a condition may be based on at least one SNP. For example, assessment of an individual's risk for Alzheimers (AD), colorectal cancer (CRC), osteoarthritis (OA) or exfoliation glaucoma (XFG), may be based on 1 SNP (for example, rs4420638 for AD, rs6983267 for CRC, rs4911178 for OA and rs2165241 for XFG). For other conditions, such as obesity (BMI/OB), Graves' disease (GD), or hemochromatosis (HEM), an individual's estimated risk may be based on at least 1 or 2 SNPs (for example, rs9939609 and/or rs9291171 for BMI/OB; DRB1*0301 DQA1*0501 and/or rs3087243 for GD; rs1800562 and/or rs129128 for HEM). For conditions such as, but not limited to, myocardial infarction (MI), multiple sclerosis (MS), or psoriasis (PS), 1, 2, or 3 SNPs may be used to assess an individual's risk for the condition (for example, rs1866389, rs1333049, and/or rs6922269 for MI; rs6897932, rs12722489, and/or DRB1*1501 for MS; rs6859018, rs11209026, and/or HLAC*0602 for PS). For estimating an individual's risk of restless legs syndrome (RLS) or celiac disease (CeID), 1, 2, 3, or 4 SNPs (for example, rs6904723, rs2300478, rs1026732, and/or rs9296249 for RLS; rs6840978, rs11571315, rs2187668, and/or DQA1*0301 DQB1*0302 for CeID). For prostate cancer (PC) or lupus (SLE), 1, 2, 3, 4, or 5 SNPs may be used to estimate an individual's risk for PC or SLE (for example, rs4242384, rs6983267, rs16901979, rs17765344, and/or rs4430796 for PC; rs12531711, rs10954213, rs2004640, DRB1*0301, and/or DRB1*1501 for SLE). For estimating an individual's lifetime risk of macular degeneration (AMD) or rheumatoid arthritis (RA), 1, 2, 3, 4, 5, or 6 SNPs, may be used (for example, rs10737680, rs10490924, rs541862, rs2230199, rs10611170, and/or rs9332739 for AMD; rs6679677, rs11203367, rs6457617, DRB1*0101, DRB1*0401, and/or DRB1*0404 for RA). For estimating an individual's lifetime risk of breast cancer (BC), 1, 2, 3, 4, 5, 6 or 7 SNPs may be used (for example, rs3803662, rs2981582, rs4700485, rs3817198, rs17468277, rs6721996, and/or rs3803662). For estimating an individual's lifetime risk of Crohn's disease (CD) or Type 2 diabetes (T2D), 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, or 11 SNPs may be used (for example, rs2066845, rs5743293, rs10883365, rs17234657, rs10210302, rs9858542, rs11805303, rs1000113, rs17221417, rs2542151, and/or rs10761659 for CD; rs13266634, rs4506565, rs10012946, rs7756992, rs10811661, rs12288738, rs8050136, rs1111875, rs4402960, rs5215, and/or rs1801282 for T2D). In some embodiments, the SNPs used as a basis for determining risk may be in linkage disequilibrium with the SNPs as mentioned above, or other SNPs, such as in US Publication No. 20080131887 and PCT Publication No. WO/2008/067551.

[0061] The phenotype profile of an individual may comprise a number of phenotypes. In particular, the assessment of a patient's risk of disease or other conditions such as likely drug response including metabolism, efficacy and/or safety, by the methods disclosed herein, allows for prognostic or diagnostic analysis of susceptibility

to multiple, unrelated diseases and conditions, whether in symptomatic, presymptomatic or asymptomatic individuals, including carriers of one or more disease/condition predisposing alleles. Accordingly, these methods provide for general assessment of an individual's susceptibility to disease or condition without any preconceived notion of testing for a specific disease or condition. For example, the methods disclosed herein allow for assessment of an individual's susceptibility to any of the several conditions listed in US Publication No. 20080131887 and PCT Publication No. WO/2008/067551, based on the individual's genomic profile. Furthermore, the methods allow assessments of an individual's estimated lifetime risk or relative risk for one or more phenotype or condition.

[0062] The assessment provides information for 2 or more of these conditions, and can include at least 3, 4, 5, 10, 15, 18, 20, 25, 30, 35, 40, 45, 50, 100 or even more of these conditions. A single rule for a phenotype may be applied for monogenic phenotypes. More than one rule may also be applied for a single phenotype, such as a multigenic phenotype or a monogenic phenotype wherein multiple genetic variants within a single gene affects the probability of having the phenotype.

[0063] Following an initial screening of an individual patient's genomic profile, updates of an individual's genotype correlations can be made (or are available) through comparisons to additional genetic variants, such as SNPs, when such additional genetic variants become known. For example, updates may be performed periodically, for example, daily, weekly, or monthly by one or more people of ordinary skill in the field of genetics, who scan scientific literature for new genotype correlations. The new genotype correlations may then be further validated by a committee of one or more experts in the field.

[0064] The new rule may encompass a genotype or phenotype without an existing rule. For example, a genotype not correlated with any phenotype is discovered to correlate with a new or existing phenotype. A new rule may also be for a correlation between a phenotype for which no genotype has previously been correlated to. New rules may also be determined for genotypes and phenotypes that have existing rules. For example, a rule based on the correlation between genotype A and phenotype A exists. New research reveals genotype B correlates with phenotype A, and a new rule based on this correlation is made. Another example is phenotype B is discovered to be associated with genotype A, and thus a new rule may be made.

[0065] Rules may also be made on discoveries based on known correlations but not initially identified in published scientific literature. For example, it may be reported genotype C is correlated with phenotype C. Another publication reports genotype D is correlated with phenotype D. Phenotype C and D are related symptoms, for example phenotype C may be shortness of breath, and phenotype D is small lung capacity. A correlation between genotype C and phenotype D, or genotype D with phenotype C, may be discovered and validated through statistical means with existing stored genomic profiles of individuals with genotypes C and D, and phenotypes C and D, or by further research. A new rule may then be generated based on the newly discovered and validated correlation. In another embodiment, stored genomic profiles of a number of individuals with a specific or related phenotype may be studied to determine a genotype common to the individuals, and a correlation may be determined. A new rule may be generated based on this correlation.

[0066] Rules may also be made to modify existing rules. For example, correlations between genotypes and phenotypes may be partly determined by a known individual characteristic, such as ethnicity, ancestry, geography, gender, age, family history, or any other known phenotypes of the individual. Rules based on these known individual characteristics may be made and incorporated into an existing rule, to provide a modified rule. The

choice of modified rule to be applied will be dependent on the specific individual factor of an individual. For example, a rule may be based on the probability an individual who has phenotype E is 35% when the individual has genotype E. However, if an individual is of a particular ethnicity, the probability is 5%. A new rule may be generated based on this result and applied to individuals with that particular ethnicity. Alternatively, the existing rule with a determination of 35% may be applied, and then another rule based on ethnicity for that phenotype is applied. The rules based on known individual characteristics may be determined from scientific literature or determined based on studies of stored genomic profiles. New rules may be added and applied to genomic profiles, as the new rules are developed, or they may be applied periodically, such as at least once a year.

[0067] Information of an individual's risk of disease can also be expanded as technology advances allow for finer resolution SNP genomic profiles. As indicated above, an initial SNP genomic profile readily can be generated using microarray technology for scanning of 500,000 SNPs. Given the nature of haplotype blocks, this number allows for a representative profile of all SNPs in an individual's genome. Nonetheless, there are approximately 10 million SNPs estimated to occur commonly in the human genome (the International HapMap Project; www.hapmap.org). As technological advances allow for practical, cost-efficient resolution of SNPs at a finer level of detail, such as microarrays of 1,000,000, 1,500,000, 2,000,000, 3,000,000, or more SNPs, or whole genomic sequencing, more detailed SNP genomic profiles can be generated. Likewise, cost-efficient analysis of finer SNP genomic profiles and updates to the master database of SNP-disease correlations will be enabled by advances in computational analytical methodology.

[0068] In some embodiments, "field-deployed" mechanisms may be gathered from individuals, and incorporated into the phenotype profile for the individuals. For example, an individual may have an initial phenotype profile generated based on genetic information. The initial phenotype profile generated includes risk factors for different phenotypes as well as suggested treatments or preventative measures, reported in a personal action plan. The profile may include information on available medication for a certain condition, and/or suggestions on dietary changes or exercise regimens. The individual may choose to see, or contact via a web portal or phone call, a physician or genetic counselor, to discuss their phenotype profile. The individual may decide to take a certain course of action, for example, take specific medications, change their diet, and other possible actions suggested on their personal action plan. The individual may then subsequently submit biological samples to assess changes in their physical condition and possible change in risk factors.

[0069] Individuals may have the changes determined by directly submitting biological samples to the facility (or associated facility, such as a facility contracted by the entity generating the genetic profiles and phenotype profiles) that generates the genomic profiles and phenotype profiles. Alternatively, the individuals may use a "field-deployed" mechanism, wherein the individual may submit their saliva, blood, or other biological sample into a detection device at their home, analyzed by a third party, and the data transmitted to be incorporated into another phenotype profile. For example, an individual may have received an initial phenotype report based on their genetic data reporting the individual having an increased lifetime risk of myocardial infarction (MI). The report may also have suggestions on preventative measures to reduce the risk of MI, such as cholesterol lowering drugs and change in diet. The individual may choose to contact a genetic counselor or physician to discuss the report and the preventative measures and decides to change their diet. After a period of being on the new diet, the individual may see their personal physician to have their cholesterol level measured. The new information (cholesterol level) may

be transmitted (for example, via the Internet) to the entity with the genomic information, and the new information used to generate a new phenotype profile for the individual, with a new risk factor for myocardial infarction, and/or other conditions.

[0070] The individual may also use a “field-deployed” mechanism, or direct mechanism, to determine their individual response to specific medications. For example, an individual may have their response to a drug measured, and the information may be used to determine more effective treatments. Measurable information include, but are not limited to, metabolite levels, glucose levels, ion levels (for example, calcium, sodium, potassium, iron), vitamins, blood cell counts, body mass index (BMI), protein levels, transcript levels, heart rate, etc., can be determined by methods readily available and can be factored into an algorithm to combine with initial genomic profiles to determine a modified overall risk estimate score. The risk estimate score may be a GCI score.

Genetic Composite Index (GCI)

[0071] In some embodiments, information about the association of multiple genetic markers or variants with one or more diseases or conditions is combined and analyzed to produce a Genetic Composite Index (GCI) score. For example, the GCI score may incorporate one or more odds ratios or relative risks from the presence or absence of different genetic variants for a phenotype. The GCI score may incorporate at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 odds ratios or relative risks from various genetic variants.

[0072] This score incorporates known risk factors, as well as other information and assumptions such as the allele frequencies and the prevalence of a disease. The GCI can be used to qualitatively estimate the association of a disease or a condition with the combined effect of a set of genetic markers. The GCI score can be used to provide people not trained in genetics with a reliable (i.e., robust), understandable, and/or intuitive sense of what their individual risk of a disease is compared to a relevant population based on current scientific research.

[0073] The GCI score may be used to generate GCI Plus scores. The methods disclosed herein encompasses using the GCI score, and one of ordinary skill in the art will readily recognize the use of GCI Plus scores or variations thereof, in place of GCI scores as described herein. The GCI Plus score may contain all the GCI assumptions, including risk (such as lifetime risk), age-defined prevalence, and/or age-defined incidence of the condition. The lifetime risk for the individual may then be calculated as a GCI Plus score which is proportional to the individual's GCI score divided by the average GCI score. The average GCI score may be determined from a group of individuals of similar ancestral background, for example a group of Caucasians, Asians, East Indians, or other group with a common ancestral background. Groups may comprise of at least 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, or 60 individuals. In some embodiments, the average may be determined from at least 75, 80, 95, or 100 individuals. The GCI Plus score may be determined by determining the GCI score for an individual, dividing the GCI score by the average relative risk and multiplying by the lifetime risk for a condition or phenotype. For example, using data from US Publication No. 20080131887 and PCT Publication No. WO/2008/067551, GCI or GCI Plus scores for an individual can be determined. The scores may be used to generate information on genetic risks, such as estimated lifetime risk, for one or more conditions in the phenotype profile of an individual. The methods allow calculating estimated lifetime risks or relative risks for one or more phenotypes or conditions. The risk for a single condition may be based on one or more SNP. For example, an estimated risk for a phenotype or condition may be based on at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12 SNPs, wherein the SNPs for estimating a risk may be published SNPs, test SNPs, or both.

[0074] A GCI score can be generated for each disease or condition of interest. These GCI scores may be collected to form a risk profile for an individual. The GCI scores may be stored digitally so that they are readily accessible at any point of time to generate risk profiles. Risk profiles may be broken down by broad disease classes, such as cancer, heart disease, metabolic disorders, psychiatric disorders, bone disease, or age on-set disorders. Broad disease classes may be further broken down into subcategories. For example for a broad class such as a cancer, subcategories of cancer may be listed such as by type (sarcoma, carcinoma or leukemia, etc.) or by tissue specificity (neural, breast, ovaries, testes, prostate, bone, lymph nodes, pancreas, esophagus, stomach, liver, brain, lung, kidneys, etc.). Further the risk profiles may display information on how the GCI scores are predicted to change as the individual ages or various risk factors are adjusted. For example, the GCI scores for particular diseases may take into account the effect of changes in diet or preventative measures taken (smoking cessation, drug intake, double radical mastectomies, hysterectomies, and the like).

[0075] A GCI score can be generated for an individual, which provides them with easily comprehended information about the individual's risk of acquiring or susceptibility to at least one disease or condition. One or more GCI scores can be generated for a single disease or condition, or numerous diseases or conditions. The one or more GCI score can be accessible by an on-line portal. Alternatively, the one or more GCI scores may be provided in paper form, with subsequent updates also provided in paper form. The paper form can be mailed to an individual or their health care manager or provided in person.

[0076] A method for generating a robust GCI score for the combined effect of different loci can be based on a reported individual risk for each locus studied. For example, a disease or condition of interest is identified and then informational sources, including but not limited to databases, patent publications and scientific literature, are queried for information on the association of the disease or condition with one or more genetic loci. These informational sources are curated and assessed using quality criteria. In some embodiments the assessment process involves multiple steps. In other embodiments the informational sources are assessed for multiple quality criteria. The information derived from informational sources is used to identify the odds ratio or relative risk for one or more genetic loci for each disease or condition of interest.

[0077] In an alternative embodiment, the odds ratio (OR) or relative risk (RR) for at least one genetic loci is not available or not accessible from informational sources. The RR is then calculated using (1) reported OR of multiple alleles of the same locus, (2) allele frequencies from data sets, such as the HapMap data set, and/or (3) disease/condition prevalence from available sources (e.g., CDC, National Center for Health Statistics, etc.) to derive RR of all alleles of interest. In one embodiment the ORs of multiple alleles of same locus are estimated separately or independently. In a preferred embodiment the ORs of multiple alleles of same locus are combined to account for dependencies between the ORs of the different alleles. In some embodiments established disease models (including, but not limited to models such as the multiplicative, additive, Harvard-modified, dominant effect) are used to generate an intermediate score that represents the risk of an individual according to the model chosen.

[0078] A method that can be used analyzes multiple models for a disease or condition of interest and correlates the results obtained from these different models; this minimizes the possible errors that may be introduced by choice of a particular disease model. This method minimizes the influence of reasonable errors in the estimates of prevalence, allele frequencies and ORs obtained from informational sources on the calculation of the relative risk. Without being limited by theory, because of the "linearity" or monotonic nature of the effect of a prevalence estimate on the

RR, there is little or no effect of incorrectly estimating the prevalence on the final rank score; provided that the same model is applied consistently to all individuals for which a report is generated.

[0079] The methods described herein can also take into account environmental/behavioral/demographic data as additional “loci.” In a related method, such data may be obtained from informational sources, such as medical or scientific literature or databases (e.g., associations of smoking w/lung cancer, or from insurance industry health risk assessments). Also disclosed herein are GCI scores produced for one or more complex diseases. Complex diseases may be influenced by multiple genes, environmental factors, and their interactions. A large number of possible interactions may need to be analyzed when studying complex diseases. A procedure used to correct for multiple comparisons, such as the Bonferroni correction, may be used to generate a GCI score. Alternatively, the Simes’s test can be used to control the overall significance level (also known as the “familywise error rate”) when the tests are independent or exhibit a special type of dependence (*Sarkar S., Ann Stat 26:494–504 (1998)*). Simes’s test rejects the global null hypothesis that all K test-specific null hypotheses are true if $p_{(k)} \leq ak/K$ for any k in $1, \dots, K$. (*Simes, R. J., Biometrika 73:751–754 (1986)*).

[0080] Other embodiments that can be used in the context of multiple-gene and multiple-environmental-factor analysis control the false-discovery rate—that is, the expected proportion of rejected null hypotheses that are falsely rejected. This approach can be particularly useful when a portion of the null hypotheses can be assumed false, as in microarray studies. Devlin et al. (*Genet. Epidemiol. 25:36–47 (2003)*) proposed a variant of the Benjamini and Hochberg (*J. R. Stat. Soc. Ser. B 57:289–300 (1995)*) step-up procedure that controls the false-discovery rate when testing a large number of possible gene $\times$ gene interactions in multilocus association studies. The Benjamini and Hochberg procedure is related to Simes’s test; setting $k^* = \max k$ such that $p_{(k)} \leq ak/K$, it rejects all k^* null hypotheses corresponding to $p_{(1)}, \dots, p_{(k^*)}$. In fact, the Benjamini and Hochberg procedure reduces to Simes’s test when all null hypotheses are true (*Benjamini and Yekutieli, Ann. Stat. 29:1165–1188 (2001)*).

[0081] Also provided herein is a ranking of an individual, where an individual is ranked in comparison to a population of individuals based on their intermediate score to produce a final rank score, which may be represented as rank in the population, such as the 99th percentile or 99th, 98th, 97th, 96th, 95th, 94th, 93rd, 92nd, 91st, 90th, 89th, 88th, 87th, 86th, 85th, 84th, 83rd, 82nd, 81st, 80th, 79th, 78th, 77th, 76th, 75th, 74th, 73rd, 72nd, 71st, 70th, 69th, 65th, 60th, 55th, 50th, 45th, 40th, 35th, 30th, 25th, 20th, 15th, 10th, 5th, or 0th percentile. The rank score may be displayed as a range, such as the 100th to 95th percentile, the 95th to 85th percentile, the 85th to 60th percentile, or any sub-range between the 100th and 0th percentile. The individual can also be ranked in quartiles, such as the top 75th quartile, or the lowest 25th quartile. The individual can also be ranked in comparison to the mean or median score of the population.

[0082] In one embodiment, the population to which the individual is compared to includes a large number of people from various geographic and ethnic backgrounds, such as a global population. Alternatively, the population to which an individual is compared to is limited to a particular geography, ancestry, ethnicity, sex, age (for example, fetal, neonate, child, adolescent, teenager, adult, geriatric), or disease state (for example, symptomatic, asymptomatic, carrier, early-onset, late onset). In some embodiments, the population to which the individual is compared to is derived from information reported in public and/or private informational sources.

[0083] The GCI score can be generated using a multi-step process. For example, initially, for each condition to be studied, the relative risks from the odds ratios for each of the genetic markers is calculated. For every prevalence value $p=0.01, 0.02, \dots, 0.5$, the GCI score of the HapMap CEU population is calculated based on the prevalence and

on the HapMap allele frequency. If the GCI scores are invariant under the varying prevalence, then the only assumption taken into account is that there is a multiplicative model. Otherwise, it is determined that the model is sensitive to the prevalence. The relative risks and the distribution of the scores in the HapMap population, for any combination of no-call values, are obtained. For each new individual, the individual's score is compared to the HapMap distribution and the resulting score is the individual's rank in this population. The resolution of the reported score may be low due to the assumptions made during the process. The population will be partitioned into quantiles (3-6 bins), and the reported bin would be the one in which the individual's rank falls. The number of bins may be different for different diseases based on considerations such as the resolution of the score for each disease. In case of ties between the scores of different HapMap individuals, the average rank will be used.

[0084] A higher GCI score can be interpreted as an indication of an increased risk for acquiring or being diagnosed with a condition or disease. Mathematical models are typically used to derive the GCI score. The GCI score can be based on a mathematical model that accounts for the incomplete nature of the underlying information about the population and/or diseases or conditions. The mathematical model can include at least one presumption as part of the basis for calculating the GCI score, wherein the presumption includes, but is not limited to: a presumption that the odds ratio values are given; a presumption that the prevalence of the condition is known; a presumption that the genotype frequencies in the population are known; and/or a presumption that the customers are from the same ancestry background as the populations used for the studies and as the HapMap; a presumption that the amalgamated risk is a product of the different risk factors of the individual genetic markers. The GCI may also include a presumption that the mutli-genotypic frequency of a genotype is the product of frequencies of the alleles of each of the SNPs or individual genetic markers (for example, the different SNPs or genetic markers are independent across the population).

The Multiplicative Model

[0085] The GCI score can be computed under the assumption that the risk attributed to the set of genetic markers is the product of the risks attributed to the individual genetic markers. Thus, the different genetic markers attribute independently of the other genetic markers to the risk of the disease. Formally, there are k genetic markers with risk alleles $r_1, \dots, r_k$ and non-risk alleles $n_1, \dots, n_k$. In SNP i , the three possible genotype values are denoted as $r_i r_i, n_i r_i$, and $n_i n_i$. The genotype information of an individual can be described by a vector, $(g_1, \dots, g_k)$, where g_i

can be 0, 1, or 2, according to the number of risk alleles in position i . Denoted by λ_1^i , the relative risk of a heterozygous genotype in position i compared to a homozygous non-risk allele at the same position. In other words,

$$\lambda_1^i = \frac{P(D | n_i r_i)}{P(D | n_i n_i)}. \text{ Similarly, the relative risk of an } r_i r_i \text{ genotype is denoted as } \lambda_2^i = \frac{P(D | r_i r_i)}{P(D | n_i n_i)}. \text{ Under the}$$

multiplicative model, the assumption that the risk of an individual with a genotype $(g_1, \dots, g_k)$ is

$$GCI(g_1, \dots, g_k) = \prod_{i=1}^k \lambda_{g_i}^i.$$

Estimating the Relative Risk.

[0086] In another embodiment, the relative risks for different genetic markers are known and the multiplicative model can be used for risk assessment. However, in some embodiments involving association studies, the study design prevents the reporting of the relative risks. In some case-control studies the relative risk cannot be calculated directly from the data without further assumptions. Instead of reporting the relative risks, it is customary to report the odds ratio (OR) of the genotype, which are the odds of carrying the disease given the risk genotype (either $r_i r_i$ or $n_i r_i$) vs. the odds of not carrying the disease given the risk genotypes. Formally,

$$OR_i^1 = \frac{P(D|n_i r_i)}{P(D|n_i n_i)} \cdot \frac{1 - P(D|n_i n_i)}{1 - P(D|n_i r_i)}$$

$$OR_i^2 = \frac{P(D|r_i r_i)}{P(D|n_i n_i)} \cdot \frac{1 - P(D|n_i n_i)}{1 - P(D|r_i r_i)}$$

[0087] Finding the relative risks from the odds ratio may require additional assumptions. Such as the presumption that the allele frequencies in an entire population $a=f_{n_i n_i}$, $b=f_{n_i r_i}$, and $c=f_{r_i r_i}$ are known or estimated (these could be estimated from current datasets such as the HapMap dataset which includes 120 chromosomes), and/or that the prevalence of the disease $p=P(D)$ is known. From the preceding three equations can be derived:

$$p = a \cdot P(D|n_i n_i) + b \cdot P(D|n_i r_i) + c \cdot P(D|r_i r_i)$$

$$OR_i^1 = \frac{P(D|n_i r_i)}{P(D|n_i n_i)} \cdot \frac{1 - P(D|n_i n_i)}{1 - P(D|n_i r_i)}$$

$$OR_i^2 = \frac{P(D|r_i r_i)}{P(D|n_i n_i)} \cdot \frac{1 - P(D|n_i n_i)}{1 - P(D|r_i r_i)}$$

[0088] By the definition of the relative risk, after dividing by the term $pP(D|n_i n_i)$, the first equation can be rewritten as:

$$\frac{1}{P(D|n_i n_i)} = \frac{a + b\lambda_1^i + c\lambda_2^i}{p},$$

and therefore, the last two equations can be rewritten as:

$$OR_i^1 = \lambda_1^i \cdot \frac{(a - p) + b\lambda_1^i + c\lambda_2^i}{a + (b - p)\lambda_1^i + c\lambda_2^i}$$

(1)

$$OR_i^2 = \lambda_2^i \cdot \frac{(a - p) + b\lambda_1^i + c\lambda_2^i}{a + b\lambda_1^i + (c - p)\lambda_2^i}$$

[0089] Note that when $a=1$ (non-risk allele frequency is 1), Equation system 1 is equivalent to the Zhang and Yu formula in Zhang and Yu (*JAMA*, 280:1690-1691 (1998)), which is incorporated by reference in its entirety. In contrast to the Zhang and Yu formula, some embodiments take into consideration the allele frequency in the population, which may affect the relative risk. Further, some embodiments take into account the interdependence of the relative risks, as opposed to computing each of the relative risks independently.

[0090] Equation system 1 can be rewritten as two quadratic equations, with at most four possible solutions. A gradient descent algorithm can be used to solve these equations, where the starting point is set to be the odds ratio,

e.g., $\lambda_1^i = OR_1^i$, and $\lambda_2^i = OR_2^i$

[0091] For example:

$$f_1(\lambda_1, \lambda_2) = OR_1^1 \cdot (a + (b-p)\lambda_1^i + c\lambda_2^i) - \lambda_1^i \cdot ((a-p) + b\lambda_1^i + c\lambda_2^i)$$

$$f_2(\lambda_1, \lambda_2) = OR_2^2 \cdot (a + b\lambda_1^i + (c-p)\lambda_2^i) - \lambda_2^i \cdot ((a-p) + b\lambda_1^i + c\lambda_2^i)$$

[0092] Finding the solution of these equations is equivalent to finding the minimum of the function

$$g(\lambda_1, \lambda_2) = f_1(\lambda_1, \lambda_2)^2 + f_2(\lambda_1, \lambda_2)^2.$$

Thus,

$$\frac{dg}{d\lambda_1} = 2f_1(\lambda_1, \lambda_2) \cdot b \cdot (\lambda_2 - OR_2) + 2f_2(\lambda_1, \lambda_2) (2b\lambda_1 + c\lambda_2 + a - OR_1 b - p + OR_1 p)$$

$$\frac{dg}{d\lambda_2} = 2f_2(\lambda_1, \lambda_2) \cdot c \cdot (\lambda_1 - OR_1) + 2f_1(\lambda_1, \lambda_2) (2c\lambda_2 + b\lambda_1 + a - OR_2 c - p + OR_2 p)$$

[0093] In this example, by setting $x_0 = OR_1$, $y_0 = OR_2$, set the values $[\text{epsilon}] = 10^{-10}$ to be a tolerance constant through the algorithm. In iteration i , define

$$\gamma = \min \left\{ 0.001, \frac{x_{i-1}}{[\text{epsilon}] + 10 \left| \frac{dg}{d\lambda_1}(x_{i-1}, y_{i-1}) \right|}, \frac{y_{i-1}}{[\text{epsilon}] + 10 \left| \frac{dg}{d\lambda_2}(x_{i-1}, y_{i-1}) \right|} \right\}, \text{ then set}$$

$$x_i = x_{i-1} - \gamma \frac{dg}{d\lambda_1}(x_{i-1}, y_{i-1})$$

$$y_i = y_{i-1} - \gamma \frac{dg}{d\lambda_2}(x_{i-1}, y_{i-1})$$

The iterations are repeated until $g(x_i, y_i) < \text{tolerance}$, where tolerance is set to 10^{-7} in the supplied code.

[0094] In this example, these equations give the correct solution for different values of a, b, c, p, OR_1 , and OR_2 .

Robustness of the Relative Risk Estimation.

[0095] In some embodiments, the effect of different parameters (prevalence, allele frequencies, and odds ratio errors) on the estimates of the relative risks is measured. In order to measure the effect of the allele frequency and prevalence estimates on the relative risk values, the relative risk from a set of values of different odds ratios and different allele frequencies is computed (under HWE), and the results of these calculations is plotted for prevalence values ranging from 0 to 1. Additionally, for fixed values of the prevalence, the resulting relative risks can be plotted as a function of the risk-allele frequencies. In cases when $p=0$, $\lambda_1 = OR_1$, and $\lambda_2 = OR_2$, and when $p=1$,

$\lambda_1 = \lambda_2 = 0$. This can be computed directly from the equations. Additionally, in some embodiments when the risk

allele frequency is high, λ_1 gets closer to a linear function, and λ_2 gets closer to a concave function with a bounded second derivative. In the limit, when $c=1$, $\lambda_2=OR_2+p(1-OR_2)$, and $\lambda_1=OR_1-\frac{(OR_1-1)pOR_1}{OR_2(1-p)+pOR_1}$. If $OR_1 \approx OR_2$ the latter is close to a linear function as well. When risk-allele frequency is low, λ_1 and λ_2 approach the behavior of the function $1/p$. In the limit, when $c=0$, $\lambda_1=\frac{OR_1}{1-p+pOR_1}$, $\lambda_2=\frac{OR_2}{1-p+pOR_2}$. This indicates that for high risk-allele frequencies, incorrect estimates of the prevalence will not significantly affect the resulting relative risk. Further, for low risk-allele frequency, if a prevalence value of $p'=\alpha p$ is substituted for the correct prevalence p , then the resulting relative risks will be off by a factor of $\frac{1}{\alpha}$ at most.

Calculating The GCI Score

[0096] In one embodiment, the GCI is calculated by using a reference set that represents the relevant population. This reference set may be one of the populations in the HapMap, or another genotype dataset.

[0097] In this embodiment, the GCI is computed as follows: For each of the k risk loci, the relative risk is calculated from the odds ratio using the equation system 1 or as described below. Then, the multiplicative score for each individual in the reference set is calculated, which is the product of the relative risks over all loci. The multiplicative score implicitly assumes that different SNPs have an independent effect on the disease or condition, but the model can be extended to cases where some interactions are known. The GCI of an individual with a multiplicative score of s is the fraction of all individuals in the reference dataset with a score of $s' \leq s$. For instance, if 50% of the individuals in the reference set have a multiplicative score smaller than s , the final GCI score of the individual would be 0.5. The GCI can be generalized to account for SNP-SNP interactions if the odds ratios or relative risks are known for the different genotype or haplotype combinations (these can be found in the literature in some cases).

[0098] As described herein, the multiplicative model can be used to in the GCI score, however, other models may be used for the purpose of determining the GCI score. Other suitable models include but are not limited to:

[0099] *The Additive Model.* Under the additive model, the risk of an individual with a genotype $(g_1, \dots, g_k)$ is

$$\text{presumed to be } GCI(g_1, \dots, g_k) = \sum_{i=1}^k \lambda_{g_i}.$$

[00100] *Generalized Additive Model.* Under the generalized additive model, it is presumed that there is a function f

$$\text{such that the risk of an individual with a genotype } (g_1, \dots, g_k) \text{ is } GCI(g_1, \dots, g_k) = \sum_{i=1}^k f(\lambda_{g_i}).$$

[00101] *Harvard Modified Score (Het).* This score was derived from Colditz et al. (*Cancer Causes and Controls*, 11:477–488 (2000)), which is herein incorporated in its entirety. The Het score is essentially a generalized additive score, although the function f operates on the odds ratio values instead of the relative risks. This may be useful in cases where the relative risk is difficult to estimate. In order to define the function f , an intermediate function g , is defined as:

$$g(x) = \begin{cases} 0 & 1 < x \leq 1.09 \\ 5 & 1.09 < x \leq 1.49 \\ 101.49 & 1.49 < x \leq 2.99 \\ 252.99 & 2.99 < x \leq 6.99 \\ 50 & 6.99 < x \end{cases}$$

[00102] Next the quantity $het = \sum_{i=1}^k p_{het}^i g(OR_1^i)$ is calculated, where p_{het}^i is the frequency of heterozygous

individuals in SNP i across the reference population. The function f is then defined as $f(x) = g(x)/het$, and the Harvard

Modified Score (Het) is simply defined as $\sum_{i=1}^k f(OR_{g_i}^i)$.

[00103] *The Harvard Modified Score (Hom)*. This score is similar to the Het score, except that the value het is

replaced by the value $hom = \sum_{i=1}^k p_{hom}^i g(OR_1^i)$, where p_{hom}^i is the frequency of individuals with homozygous risk-

allele.

[00104] *The Maximum-Odds Ratio*. In this model, it is presumed that one of the genetic markers (one with a maximal odds ratio) gives a lower bound on the combined risk of the entire panel. Formally, the score of an

individual with genotypes $(g_1, \dots, g_k)$ is $GCI(g_1, \dots, g_k) = \max_{i=1}^k OR_{g_i}^i$.

[00105] A comparison between the scores is described in Example 1 and GCI score evaluation is described in Example 2.

Extending the Model to an Arbitrary Number of Variants

[00106] The model can be extended to the situations where an arbitrary number of possible variants occur.

Previous considerations dealt with situations where there were three possible variants (nn, nr, rr). Generally, when a multi-SNP association is known, an arbitrary number of variants may be found in the population. For example, when an interaction between two Genetic markers is associated with a condition, there are nine possible variants. This results in eight different odds ratios values.

[00107] To generalize the initial formula, it may be assumed that there are $k+1$ possible variants $a_0, \dots, a_k$, with

frequencies $f_0, f_1, \dots, f_k$, measured odds ratios of $1, OR_1, \dots, OR_k$, and unknown relative risk values $1, \lambda_1, \dots, \lambda_k$.

Further it may be assumed that all relative risks and odds ratios are measured with respect to a_0 , and thus,

$$\lambda_i = \frac{P(D|a_i)}{P(D|a_0)}, \text{ and } OR_i = \frac{P(D|a_i)}{P(D|a_0)} \cdot \frac{1 - P(D|a_i)}{1 - P(D|a_0)}. \text{ Based on:}$$

$$p = \sum_{i=0}^k f_i P(D|a_i),$$

It is determined that

$$OR_i = \lambda_i \frac{\sum_{i=0}^k f_i \lambda_i^{-p}}{\sum_{i=0}^k f_i \lambda_i^{-\lambda_i p}}$$

Further if it is set that $C = \sum_i f_i \lambda_i$, this results in the equation:

$$\lambda_i = \frac{C \cdot OR_i}{C \cdot p + OR_i^p},$$

and thus,

$$C = \sum_{i=0}^k f_i \lambda_i = \sum_{i=0}^k \frac{C \cdot OR_i f_i}{C \cdot p + OR_i^p},$$

or

$$1 = \sum_{i=0}^k \frac{OR_i f_i}{C \cdot p + OR_i^p}.$$

[00108] The latter is an equation with one variable (C). This equation can produce many different solutions (essentially, up to $k+1$ different solutions). Standard optimization tools such as gradient descent can be used to find the closest solution to $C_0 = \sum_i f_i t_i$.

[00109] A robust scoring framework for the quantification of risk factors us also provided herein. While different genetic models may result in different scores, the results are usually correlated. Therefore the quantification of risk factors is generally not dependent on the model used.

Estimating Relative Risk Case Control Studies

[00110] A method that estimates the relative risks from the odds ratios of multiple alleles in a case-control study is also disclosed herein. In contrast to previous approaches, the method takes into consideration the allele frequencies, the prevalence of the disease, and the dependencies between the relative risks of the different alleles. The performance of the approach on simulated case-control studies was measured, and found to be extremely accurate.

Methods

[00111] In the case where a specific SNP is tested for association with a disease D , R and N denote the risk and non-risk alleles of this particular SNP. $P(RR|D)$, $P(RN|D)$ and $P(NN|D)$ denote the probability of getting affected by the disease given that a person is homozygous for the risk allele, heterozygous, or homozygous for the non-risk allele respectively. f_{RR} , f_{RN} and f_{NN} are used to denote the frequencies of the three genotypes in the population. Using these definitions, the relative risks are defined as

$$\lambda_{RR} = \frac{P(D|RR)}{P(D|NN)}$$

$$\lambda_{RN} = \frac{P(D|RN)}{P(D|NN)}$$

[00112] In a case-control study, the values $P(RR|D)$, $P(RR|\sim D)$ can be estimated, i.e., the frequency of RR among the cases and the controls, as well as $P(RN|D)$, $P(RN|\sim D)$, $P(NN|D)$, and $P(NN|\sim D)$, i.e., the frequency of RN and NN among the cases and the controls. In order to estimate the relative risk, Bayes law can be used to get:

$$\lambda_{RR} = \frac{P(RR|D)f_{NN}}{P(NN|D)f_{RR}}$$

$$\lambda_{RN} = \frac{P(D|RN)f_{NN}}{P(D|NN)f_{RR}}$$

[00113] Thus, if the frequencies of the genotypes are known, one can use those to calculate the relative risks. The frequencies of the genotypes in the population cannot be calculated from the case-control study itself, since they depend on the prevalence of disease in the population. In particular, if the prevalence of the disease is $p(D)$, then:

$$f_{RR} = P(RR|D)p(D) + P(RR|\sim D)(1 - p(D))$$

$$f_{RN} = P(RN|D)p(D) + P(RN|\sim D)(1 - p(D))$$

$$f_{NN} = P(NN|D)p(D) + P(NN|\sim D)(1 - p(D))$$

[00114] When $p(D)$ is small enough, the frequencies of the genotypes can be approximated by the frequencies of the genotypes in the control population, but this would not be an accurate estimate when the prevalence is high. However, if a reference dataset is given (e.g., the HapMap [cite]), one can estimate the genotype frequencies based on the reference dataset.

[00115] Most current studies do not use a reference dataset to estimate the relative risk, and only the odds-ratio is reported. The odds-ratio can be written as

$$OR_{RR} = \frac{P(RR|D)P(NN|\sim D)}{P(NN|D)P(RR|\sim D)}$$

$$OR_{RN} = \frac{P(RN|D)P(NN|\sim D)}{P(NN|D)P(RN|\sim D)}$$

[00116] The odds ratios are typically advantageous since there is usually no need to have an estimate of the allele frequencies in the population; in order to calculate the odds ratios typically what is needed is the genotype frequencies in the cases and in the controls.

[00117] In some situations, the genotype data itself is not available, but the summary data, such as the odds-ratios are available. This is the case when meta-analysis is being performed based on results from previous case-control studies. In this case, how to find the relative risks from the odds ratios is demonstrated. Using the fact that the following equation holds:

$$p(D) = f_{RR}P(D|RR) + f_{RN}P(D|RN) + f_{NN}P(D|NN)$$

If this equation is divided by $P(D|NN)$, we get

$$\frac{p(D)}{p(D|NN)} = f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN}$$

This allows the odds ratios to be written in the following way:

$$OR_{RR} = \frac{P(D|RR)(1-P(D|NN))}{P(D|NN)(1-P(D|RR))} = \lambda_{RR} \frac{\frac{p(D)}{p(D|NN)} - p(D)}{\frac{p(D)}{p(D|NN)} - p(D)\lambda_{RR}} = \lambda_{RR} \frac{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)}{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)\lambda_{RR}}$$

By a similar calculation, the following system of equations results:

$$OR_{RR} = \lambda_{RR} \frac{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)}{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)\lambda_{RR}}$$

$$OR_{RN} = \lambda_{RN} \frac{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)}{f_{RR}\lambda_{RR} + f_{RN}\lambda_{RN} + f_{NN} - p(D)\lambda_{RN}}$$

Equation 1

[00118] If the odds-ratios, the frequencies of the genotypes in the populations, and the prevalence of the disease are known, the relative risks can be found by solving this set of equations.

[00119] Note that these are two quadratic equations, and thus they have a maximum of four solutions. However, as shown below that there is typically one possible solution to this equation.

[00120] Note that when $f_{NN} = 1$, Equation system 1 is equivalent to the Zhang and Yu formula; however, here the allele frequency in the population is taken into account. Furthermore, our method takes into account the fact that the two relative risks depend on each other, while previous methods suggest to compute each of the relative risks independently.

[00121] *Relative risks for multi-allelic loci.* If multi-markers or other multi-allelic variants are considered, the calculation is complicated slightly. $a_0, a_1, \dots, a_k$ is denoted by the possible $k+1$ alleles, where a_0 is the non-risk allele. Allele frequencies $f_0, f_1, f_2, \dots, f_k$ in the population for the $k+1$ possible alleles are assumed. For allele i , the relative risk and odds-ratios are defined as

$$\lambda_i = \frac{P(D|a_i)}{P(D|a_0)}$$

$$OR_i = \frac{P(D|a_i)(1-P(D|a_0))}{P(D|a_0)(1-P(D|a_i))} = \lambda_i \frac{1-P(D|a_0)}{1-P(D|a_i)}$$

The following equation holds for the prevalence of the disease:

$$p(D) = \sum_{i=0}^k f_i P(D|a_i)$$

Thus, by dividing both sides of the equation by $p(D|a_0)$, we get:

$$\frac{p(D)}{P(D|a_0)} = \sum_{i=0}^k f_i \lambda_i$$

Resulting in:

$$OR_i = \lambda_i \frac{\sum_{i=0}^k f_i \lambda_i - p(D)}{\sum_{i=0}^k f_i \lambda_i - \lambda_i p(D)}$$

By setting $C = \sum_{i=0}^k f_i \lambda_i$, the result is $\lambda_i = C \cdot \frac{OR_i}{p(D)OR_i + C - p(D)}$. Thus, by the definition of C, it is:

$$1 = \sum_{i=0}^k f_i \frac{\lambda_i}{C} = \sum_{i=0}^k \frac{f_i OR_i}{p(D)OR_i + C - p(D)}.$$

[00122] This is a polynomial equation with one variable C. Once C is determined, the relative risks are determined. The polynomial is of degree k+1, and thus we expect to have at most k+1 solutions. However, since the right-hand side of the equation is a strictly decreasing as a function of C, there can typically only be one solution to this equation. A solution is then found using a binary search, since the solution is bounded between C=1 and

$$C = \sum_{i=0}^k OR_i.$$

[00123] *Robustness of the Relative Risk Estimation.* The effect of each of the different parameters (prevalence, allele frequencies, and odds ratio errors) on the estimates of the relative risks was measured. In order to measure the effect of the allele frequency and prevalence estimates on the relative risk values, the relative risk was calculated from a set of values of different odds ratios, different allele frequencies (under HWE), and plotted the results of these calculations for a prevalence values ranging from 0 to 1.

[00124] Additionally, for fixed values of the prevalence, the resulting relative risks as a function of the risk-allele frequencies was plotted. Evidently, in all cases when $p(D) = 0$, $\lambda_{RR} = OR_{RR}$, and $\lambda_{RN} = OR_{RN}$, and when $p(D) = 1$, $\lambda_{RR} = \lambda_{RN} = 0$. This can be computed directly from Equation 1. Additionally, when the risk allele frequency is high, λ_{RR} approaches a linear behavior, and λ_{RN} approaches a concave function with a bounded second derivative. When the risk-allele frequency is low, λ_{RR} and λ_{RN} approach the behavior of the function $1/p(D)$. This means that for high risk-allele frequency, wrong estimates of the prevalence will typically not affect the resulting relative risk by much.

[00125] *Odds Ratios vs. Relative Risk.* In epidemiology literature, the relative risk is often considered as an intuitive and informative measure of risk. However, the relative risk cannot be directly calculated in the context of case-control studies in general, and whole-genome association studies. The relative risk can usually be estimated through prospective studies, in which a set of healthy individuals is studied over a long period of time. In contrast, odds ratios are normally reported in case-control studies. The odds-ratio is the ratio between the odds of carrying the risk allele in the cases vs. the controls. For rare diseases, the odds ratio is a good approximation of relative risk; however, for common diseases, the odds ratio could result in a misleading estimate of risk, where the odds ratios may be quite high even when the increase in risk is minor.

[00126] *Relative Lifetime Risk vs. Relative Risk.* Relative risk implicitly assumes that none of the controls currently has the disease. This is relevant when the probability of having the disease is estimated. However, if interest is in the risk estimation across the span of a lifetime, or the lifetime risk of an individual to develop the condition, the fact that the some of the controls will eventually develop the disease is taken into account. The relative lifetime risk is defined as the ratio between the risk of developing the condition through the life of an individual carrying the risk allele r and the risk of developing the condition through the life of an individual carrying the non-risk allele. This is different than the standard use of relative risk in case-control studies, which is based on prevalence information.

[00127] Denoted by $a_0, a_1, \dots, a_k$ is the possible $k+1$ alleles, where a_0 is the non-risk allele. Allele frequencies $f_0, f_1, f_2, \dots, f_k$ in the population for the $k+1$ possible alleles are assumed. Further assumed is that studied individuals can be divided into three groups: CA, Y, and Z. CA denotes the cases, while Y and Z are controls. As opposed to individuals from Z, it is assumed that individuals from Y will eventually develop the condition. Also denoted by CO is the union of Y and Z, and by D the union of Y and CA. It is assumed that $|Y| = \alpha|CO| = \alpha(|Y| + |Z|)$, where α is the fraction of controls that will develop the condition within their lifetime. Note that α is upper bounded by the average lifetime risk. Possibly, α may be smaller than the average lifetime, depending on the age of onset of the disease, and the ages of the controls.

[00128] The relative risk and the odds ratios can now be represented as:

$$\lambda_i = \frac{P(CA \vee Y | a_i)}{P(CA \vee Y | a_0)}$$

$$OR_i = \frac{P(a_i | CA)P(a_0 | CO)}{P(a_0 | CA)P(a_i | CO)}$$

The odds ratios can be written as:

$$\begin{aligned} OR_i &= \frac{P(a_i | CA)P(a_0 | CO)}{P(a_0 | CA)P(a_i | CO)} = \frac{P(a_i | CA)}{P(a_0 | CA)} \cdot \frac{\alpha P(a_0 | Y) + (1 - \alpha)P(a_0 | Z)}{\alpha P(a_0 | Y) + (1 - \alpha)P(a_0 | Z)} = \\ &= \frac{P(CA | a_i)}{P(CA | a_0)} \cdot \frac{\alpha P(Y | a_0) + (1 - \alpha)P(Z | a_0)}{\alpha P(Y | a_i) + (1 - \alpha)P(Z | a_i)} = \\ &= \frac{P(CA | a_i)}{P(CA | a_0)} \cdot \frac{\alpha P(CA | a_0) + (1 - \alpha)P(Z | a_0)}{\alpha P(CA | a_i) + (1 - \alpha)P(Z | a_i)} \end{aligned}$$

The derivation from the first to second line is based on Bayes law, while the third line is based on the fact that CA and Y are essentially the same population, and thus $P(CA | a_i) = P(Y | a_i)$. Now using the fact that $P(Z | a_i) = 1 - P(CA | a_i)$, results in:

$$OR_i = \frac{P(CA | a_i)}{P(CA | a_0)} \cdot \frac{(2\alpha - 1)P(CA | a_0) + 1 - \alpha}{(2\alpha - 1)P(CA | a_i) + 1 - \alpha} = \lambda_i \cdot \frac{(2\alpha - 1)P(CA | a_0) + 1 - \alpha}{(2\alpha - 1)P(CA | a_i) + 1 - \alpha}$$

As before, $p(D) = \sum_{i=0}^k f_i P(D | a_i)$, where $p(D)$ is the average lifetime risk. Thus, using the

equality $C := \frac{p(D)}{P(CA | a_0)} = \sum_{i=0}^k f_i \lambda_i$, and the odds ratios can be rewritten as:

$$OR_i = \lambda_i \cdot \frac{(2\alpha - 1)P(D) + (1 - \alpha)C}{(2\alpha - 1)P(D)\lambda_i + (1 - \alpha)C}$$

Thus, if C is given, the relative lifetime risk can be found by assigning

$$\lambda_i = \frac{(1 - \alpha)C \cdot OR_i}{(2\alpha - 1)P(D)(1 - OR_i) + (1 - \alpha)C}$$

C can be found by solving the equation

$$1 = \sum_{i=0}^k f_i \frac{\lambda_i}{C} = \sum_{i=0}^k \frac{f_i(1 - \alpha)OR_i}{(2\alpha - 1)P(D)(1 - OR_i) + (1 - \alpha)C}$$

One can verify that by the definition of C and the odds ratios, $C > (2\alpha - 1)p(D)(OR_i - 1)$. Therefore, the right hand side is a decreasing function of C, and it can be found by applying a binary search.

[00129] *Lifetime Risk Estimate Based on GCI.* The GCI essentially provides the relative risk of an individual compared to an individual with non-risk alleles across all associated SNPs. In order to calculate the lifetime risk of an individual, the product of the lifetime risk of the individual with the average lifetime risk can be taken, and divide this product by the average lifetime risk across the population. This calculation is consistent with the definition of the average lifetime risk and of the relative risk. In order to compute the average lifetime risk, all possible genotypes are enumerated, and their relative risks that are calculated as the product of the relative risks of their variants in each of the single SNPs are summed up.

Environmental Genetic Composite Index (EGCI)

[00130] In some embodiments, an environmental factor is incorporated into a GCI score generating an Environmental Genetic Composite Index (EGCI) score. The EGCI score may be computed or determined by a computer. Environmental factors may include non-genetic factors, such as, but not limited to dietary factors, factors from exercise habits, and other lifestyle or personal choices, such as personal relationships, work and home conditions. For example, smoking (frequency and/or amount of smoking, levels of nicotine intake, and the like), drug use (type, amount, and frequency of drug use), and alcohol intake (amount and frequency, for example) may be environmental factors incorporated into a GCI score to generate an EGCI score. Other environmental factors may include the type of food, amount, and frequency of intake. Other factors may include the exercise regimen of an individual, such as intensity, type, length, and frequency of certain types of physical activity.

[00131] Yet other environmental factors may include an individual's living environment, such as a rural area, an urban setting, or city of a certain population density or pollution level. For example, an individual's residence, such as the smog levels or air quality of an individual's work of home environment, may be taken into account. An individual's sleep habits, personal relationships (for example single or married, or number of close relationships, friends, familial relationships), social status, employment (high/low stress, level of responsibility, job satisfaction, relationship with co-workers and superiors, and the like) may also be taken into account.

[00132] Thus, the environmental factor can be, but not be limited to, an individual's birthplace, location of residency, lifestyle conditions; diet, exercise habits, and personal relationships. The environmental factor can also be a physical measurement of an individual, such as body mass index, blood pressure, heart rate, glucose level, metabolite level, ion level, weight, height, cholesterol level, vitamin level, blood cell count, protein level, and transcript level. The EGCI can also incorporate more than one environmental factor, for example, at least 1, 2, 3, 4, 5, 10, 12, 15, 20, 25, or more environmental factors.

[00133] The environmental factor may be independent of one or more genetic factors in contributing to the risk of a disease or condition. The environmental factor may also be independent of one or more other environmental factors in contributing to the risk of a disease or condition. In some embodiments, the environmental factor may not be independent of one or more genetic factors. In yet other embodiments, the environmental factor may not be independent of other environmental factors. The environmental factor may not be independent of other genetic or environmental factors, but when incorporated into an EGCI score, the environmental factor may be assumed to be independent when an EGCI score is calculated (such as described in Example 5). In some embodiments, the environmental factor incorporated for an individual may be that of the individual's family (for example, as shown in Example 4) or friends, or resulting from the individual's family's or friend's actions. For example, an individual may be living with a friend or family member who smokes, and thus the exposure to smoke may be an environmental factor incorporated in the individual's EGCI.

[00134] The environmental factors incorporated into the GCI to generate an EGCI may have a relative risk factor of at least approximately 1.0 for a disease or condition. The relative risk factor may be between approximately 1 or 2, or at least approximately 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, or 1.9. In some embodiments, the relative risk factor may be at least approximately 2, 3, 4, 5, 6, 7, 8, 9, or 10. In yet other embodiments, the relative risk factor of the environmental factor may be at least approximately 12, 15, 20, 25, 30, 35, 40, 45, or 50.

[00135] In some embodiments, the environmental factors incorporated into the GCI to generate an EGCI may have an odds ratio (OR) of at least approximately 1.0 for a disease or condition. The relative risk factor may be between approximately 1 or 2, or at least approximately 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, or 1.9. In some embodiments, the OR may be at least approximately 2, 3, 4, 5, 6, 7, 8, 9, or 10. In yet other embodiments, the OR of the environmental factor may be at least approximately 12, 15, 20, 25, 30, 35, 40, 45, or 50.

[00136] The EGCI may be generated for diseases or conditions in which the heritability of the disease or condition may be less than approximately 95%. In some embodiments, the EGCI is computed for diseases or conditions which have a heritability of less than approximately 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, or 90%.

Personalized Action Plans

[00137] The personalized action plans disclosed herein provide meaningful, actionable information to improve the health or wellness of an individual that is based on the genomic profile of the individual. The action plans provide courses of action that are beneficial to an individual in view of a particular genotype correlation, and may include administration of therapeutic treatment, monitoring for potential need of treatment or effects of treatment, or making life-style changes in diet, exercise, and other personal habits/activities, which can be personalized based on an individual's genomic profile into a personalized action plan. Alternatively, an individual may be given a particular rating that is based on their genomic profile, and in addition, optionally, include other information, such as family history, existing lifestyle habits and geography, such as, but not limited to, work conditions, work environment, personal relationships, home environment, and others. Other factors that may be incorporated include ethnicity, gender, and age. The odds ratio of various dietary and exercise prevention strategies and their association with reducing risks of diseases or conditions can also be incorporated into the rating system.

[00138] For example, the personalized action plans may be generated based on an individual's GCI or EGCI score. Furthermore, the personalized action plans may be modified or updated for an individual, for example,

environmental factors for an individual may be modified or updated, generating an updated EGCI score. The personalized action plans may also be modified or updated for an individual, such as resulting from the updated EGCI score, or from revised or updated GCI scores generated from new scientific information regarding genetic information being correlated to diseases or conditions that were not previously known.

[00139] Modified or updated personalized action plans may be automatically sent to an individual or their health care manager, for example, if an individual or their health care manager had initially requested automatic updates such as with a subscription plan. Alternatively, the updated personalized action plan may only be sent when requested by an individual or their health care manager. The personalized action plan may be modified or updated based on a number of factors. For example, an individual may have more genetic correlations analyzed and the results used to modify existing recommendations, add additional recommendations, or remove recommendations on the initial personalized action plan. In some embodiments, an individual may have changed certain lifestyle habits/environment, or have more information regarding family history, existing lifestyle habits and geography, such as, but not limited to, work conditions, work environment, personal relationships, home environment, and others, or want to include their updated age to obtain a personalized action plan that incorporates these changes. For example, an individual may have followed their initial personalized action plans, such as reducing cholesterol in their diet, and thus their personalized action plan recommendations may be modified or their risk or predisposition to heart disease reduced.

[00140] The personalized action plans may also have predicted future recommendations based on an individual following the recommendations on a personalized action plan or other changes an individual may make or have occur to them. For example, the individuals' increase in age would lead to an increase in risk for osteoporosis, but depending on the amount of calcium or other lifestyle habits such as those in the personalized action, the risk may be decreased.

[00141] The personalized action plan may be reported to an individual, or their health care manager, in a single report with the individual's phenotype profile and/or genomic profile. Alternatively, the personalized action plan may be reported separately. The individual can then pursue the recommended actions on their personalized action plan. The individual may choose to consult with their health care manager prior to pursuing any actions on their plan.

[00142] The personalized action plan provided can also consolidate a number of condition specific information into a consolidated set of action steps. The personalized action plan can consolidate factors including, but not limited to, the prevalence of each condition, the relative amount of pain associated with each condition, and the type of treatments for each condition. For example, if an individual has an elevated risk of myocardial infarction (for example, expressed as a higher GCI or GCI Plus score), the individual may have a personalized action plan that includes increased consumption of fruits, vegetables, and grains. However, the individual may also have a predisposition to celiac disease, thus having wheat gluten allergy. As a result, increased consumption of wheat can be contra-indicated, and is indicated in the personalized action plan.

[00143] The personalized action plan can provide pharmaceutical recommendations, non-pharmaceutical recommendations or both. For example, the personalized action plan can include suggested pharmaceuticals as a preventative, such as cholesterol lowering drugs for an individual predisposed to myocardial infarction, and to consult with a physician. The personalized action plan can also provide non-pharmaceutical recommendations, such

as following a personalized lifestyle plan, including an exercise regimen and diet plan based on an individual's genomic profile.

[00144] The personalized action plan recommendations can be of a particular rating, labeling, or categorizing system. Each recommendation may be rated or categorized by a numerical, color, and/or letter scheme or value. The recommendations may be categorized, and further rated. Numerous variations, such as different rating schemes (using letters, numbers or colors; combinations of letters, numbers, and/or colors; different types of recommendations into one or more rating schemes) may be used.

[00145] For example, an individual's genomic profile is determined and based on their genomic profile recommendations for the individual on a personalized action plan are categorized into 3 groups: "A" representing adverse or negative effects; "N" representing neutral or no significant effect, and "B" representing beneficial or positive effects. Using this system as an example, therapeutics categorized as A for the individual would include drugs that the individual has an adverse reaction to, those categorized as N would not have any significant positive or negative effect on the individual, and those categorized as B, would be beneficial to the individual's health. Using the same categorization system, a dietary plan can also be grouped into A, B, N. For example, foods which an individual is allergic to, or should particularly avoid (for example, sugars because the individual is predisposed to diabetes or cavities) would be categorized as A. Foods which have no significant effect on the individual's health may be categorized as N. Foods which are particularly beneficial to an individual may be categorized as B, for example, if an individual has high cholesterol, foods with low cholesterol would be categorized as B. Exercise regimen for the individual can also be based on the same system. For example, an individual may be predisposed to heart problems and should avoid intense workouts, and thus running may be an A activity, whereas walking or jogging at a certain pace may be categorized as a B. Standing for a period of time may be an N for one individual, but an A for another individual predisposed to varicose veins.

[00146] Furthermore, within each category of A, N, or B, there can be further levels of categorization, such as 1 through 5, from lowest to highest impact. For example, a therapeutic may be categorized as A1, which indicates a slight negative effect, such as minor nausea, whereas A2 would indicate the therapeutic would cause vomiting, while an A5 therapeutic would cause a severe adverse reaction, such as anaphylactic shock. Conversely, a B1 would have a slight positive effect on an individual, whereas B5 would have a significant positive impact on the individual. For example, if an individual is predisposed to lung cancer, or was exposed to second hand smoke while growing up, the individual not smoking may be a B5, whereas an individual not predisposed to lung cancer may have the factor as a B4.

[00147] The different categories can also be represented by different colors, for example, A can be red tones, and to represent low to high effect on an individual's health, the shades can range from a light to dark red tones, light representing low negative effects to dark red representing severe adverse effects on the individual's health. The system can also be a continuous spectrum of colors, numbers, or letters. For example, rather than have A, N, and B, and/or subcategories within, the categorization may be from A through G, wherein A represents foods, therapeutics, lifestyle habits, environments and other factors that severely negatively impact an individual's health, whereas D represents factors that have minimal effects, either positive or negative, and G would represent highly beneficial to the individual's health. Alternatively, rather than have A through G, numbers or colors may also represent the

continuous spectrum of foods, therapeutics, lifestyle habits, environments and other factors that impact an individual's health.

[00148] In some embodiments, a particular therapy, pharmaceutical, or other lifestyle element in a personalized action plan can be categorized, labeled, or rated. For example, an individual may have a personalized action plan that includes an exercise regimen and a diet plan. The exercise regimen may include one or more ratings or categorization. For example, the ratings for the exercise regimen can range from A to E, such as in Table 1, wherein each letter corresponds to one or more types of exercises, including information regarding the types of activity, length of time, number of times in a given timeframe, that falls under each level, and thus, the recommended exercise regimen for the individual.

Table 1: Exercise Regimen: Cardiovascular Activity

Rating	Option 1	Option 2	Option 3	Option 4
A	Brisk walk 2.5 mph, 3 times a week, for 20 minutes	Swim 4 laps, 3 times a week	Cycle 5 mph, 3 times a week, for 20 minutes	Brisk walk 2.5 mph, 2 times a week, for 20 minutes Cycle 5 mph, once a week, for 20 minutes
B	Jog 3.5 mph, 3 times a week, for 20 minutes	Swim 6 laps, 3 times a week	Cycle 8 mph, 3 times a week, for 20 minutes	Jog 3.5 mph, 2 times a week, for 20 minutes Cycle 8 mph once a week, for 20 minutes
C	Run 4 mph, 3 times a week, for 20 minutes	Swim 8 laps, 3 times a week	Cycle 10 mph, 3 times a week, for 20 minutes	Run 4 mph 2.5 mph, 2 times a week for 20 minutes Cycle 10 mph, once a week, for 20 minutes
D	Run 5 mph, 3 times a week, for 25 minutes	Swim 10 laps, 3 times a week	Cycle 15 mph, 3 times a week, for 30 minutes	Run 5 mph, 2 times a week for 25 minutes Cycle 15 mph, once a week, for 20 minutes
E	Run 6 mph 3 times a week, for 30 minutes	Swim 12 laps, 3 times a week	Cycle 15 mph, 3 times a week, for 40 minutes	Run 5 mph, 2 times a week for 30 minutes Cycle 15 mph, once a week, for 40 minutes

[00149] In one embodiment, based on the genomic profile of the individual, the personalized action plan may having an A rating for an individual, and therefore the individual's recommended exercise regimen would be to select from the choices in Row A in Table 1 for their cardiovascular workout. Similarly, an analogous system for weight training can be part of the individual's exercise regimen, and weight training options for an A rating would

be recommended for the individual. In some embodiments, factors such as, but not limited to, an individual's existing diet, exercise, and other personal habits/activities, optionally, other information, such as family history, existing lifestyle habits and geography, such as, but not limited to, work conditions, work environment, personal relationships, home environment, ethnicity, gender, age, and other factors may be incorporated with an individual's genomic profile determine the individual's exercise regiment rating. Furthermore, as an individual's lifestyle habits changes, or more factors become known and are incorporated, the individual's rating can change, for example, if an individual follows the recommended activities on the personalized action plan, starting at an A rating, the individual may request an updated personalized action plan that evaluates and determines the individual is now at a B rating. Alternatively, an individual's personalized action plan may offer a timeline for when the individuals should consider moving from an A rating to a B rating to maximize their health.

[00150] The personalized action plan may also have a rating system for a dietary plan. For example, the ratings for the dietary plan can be a system that ranges from 1 to 5, wherein each number corresponds to particular grouping of fats, fibers, proteins, sugars, and other nutrients the individual is suggested to have in their diet, particular portion sizes, number of calories, and/or grouping with other foods that an individual should have as their diet. Based on the genomic profile of the individual, the personalized action plan may give a 2 rating for an individual, and therefore the individual's recommended dietary plan would be a selection of dietary choices under a 2 rating.

[00151] In another embodiment, individual foods may be categorized. For example, an individual given a 2 rating should select specific foods that are also categorized as 2. For example, specific vegetables, meats, fruits, dairy, and others may be categorized as a 2, while others not. For example, asparagus may be a vegetable that is a 2, whereas beets are a 3, and therefore the individual should include more asparagus rather than beets in their diet.

[00152] In another embodiment, an individual is given a suggested rating for what type of diet to follow that is breakdown of the types of nutrients of the type of food the individual should have in their diet, based on their genomic profile. The rating may be in the form of a visual representation that includes shapes, colors, numbers, and/or letters. The rating may be in the form of a visual representation that includes shapes, colors, numbers, and/or letters. For example, an individual is found to be predisposed to colon cancer and diabetes, and is given a symbol that represent the proportion of different nutrients in the recommended types of food the individual should have in their diet. Different types of foods, such as, but not limited to, specific fruits, vegetables, carbohydrates, meats, dairy products, and the like are represented by the same scheme. Foods with rated with a symbol that most closely resembles that given the individual would be recommended foods for the individual.

[00153] In some embodiments, factors such as, but not limited to, an individual's existing diet, exercise, and other personal habits/activities, optionally, other information, such as family history, existing lifestyle habits and geography, such as, but not limited to, work conditions, work environment, personal relationships, home environment, ethnicity, gender, age, and other factors may be incorporated with an individual's genomic profile to create a personalized action plan, and thus affect the rating given for the individual's dietary plan. Furthermore, as an individual's lifestyle habits changes, or more factors become known and are incorporated, the individual's rating can change. For example, if an individual follows the recommended activities on the personalized action plan, starting at a 1 rating for dietary plans, which is an extremely low cholesterol diet, the individual may request an updated personalized action plan that incorporates the changes in lifestyle habits the individual has had such that the individual has an improved cholesterol level, the updated personalized action plan may show that the individual may

be better suited to now follow dietary plans under rating 2, or can choose from dietary plans in ratings 1 and 2. Alternatively, an individual's initial personalized action plan can offer a timeline for when the individuals should consider moving from a 1 rating to a 2 rating, or vary their dietary plans based on a schedule, between different dietary plans under different ratings, to maximize their health.

[00154] The ratings in a personalized action plan may be for a combination of different rating systems. For example, an exercise regimen system with ratings A through E and dietary plan system with ratings 1 through 5 can be used to give an individual an A1 rating in their personalized action plan. Therefore, the individual is recommended to follow the exercise regimen of the A rating and the dietary plan of the 1 rating. Alternatively, a single rating system can be used for the exercise and diet regimen. For example, an individual may be given a particular rating such as a C rating in a personalized action plan such that the recommended exercise and dietary regimen for the individual is both under the C categorization. In other embodiments, other types of recommendations, such as other lifestyle activities and habits, are also included. For example, other than exercise and dietary regimens, other recommendations, such as therapeutics, type of work environment, type of social activities, can also be encompassed under a single rating system. Alternatively, different rating systems can be used for other recommendations. For example, letters may be used for recommended exercise regimen, numbers for dietary regimen, and colors for pharmaceutical recommendations.

[00155] In some embodiments a binary rating systems is used, such that types of recommendations are grouped into pairs. The system can be similar to the Myers Briggs Type Indicator (MBTI) system. In the MBTI system, there are four pairs of preferences or dichotomies, and an individual is placed into one of each pair. An individual's preference is 1) extraversion or introversion, 2) sensing or intuition, 3) thinking or feeling, and 4) judging or perceiving. A variation in the system can be used in determining recommendations for an individual to improve their health and well-being that is based on an individual's genomic profile.

[00156] For example, an individual may be either an A or a B for diets, wherein A represents a certain type of mix of nutrients and B is a different mix. Alternatively, specific types of foods may be grouped into A or B. The individual may have another binary categorization for exercise regimen, such as H or L, where H represents that an individual should participate in high-impact exercise, and L represent low-impact activities. As such, an individual may be categorized as an AH. Another binary categorization can be for social contact. For example, an individual can be genetically predisposed to being social (S) or unsocial (U), and as such, recommendations may include the type of activities or groups of people the individual should avoid or seek to reduce stress and increase their health and well-being.

[00157] The personalized action plans can also be updated to include factors based on information as they become known, including scientific information, or information from the individual, such as "field-deployed" or direct mechanisms, for example, metabolite levels, glucose levels, ion levels (for example, calcium, sodium, potassium, iron), vitamins, blood cell counts, body mass index (BMI), protein levels, transcript levels, heart rate, etc., can be determined by methods readily available and can be factored into the personalized action plan when they are known, as they become known, such as by real time monitoring. The personalized action plan can be modified, for example, based on an individual following the plan, which may also affect the predisposition an individual may have for one or more conditions. For example, the GCI score of the individual may be updated.

Communities and Motivations

[00158] The present disclosure provides phenotype profiles and personalized action plans that are based on an individual's genomic profile, such that individuals are well informed about their health and well-being, and the customized options individuals have to improve their health. Also provided herein are communities, such as on-line communities, that can offer support and motivation for an individual to pursue their personalized action plan. Motivation for individuals to improve their health, for example, by following their personalized action plan, can also include financial incentives.

[00159] An individual may participate in a community, such as an on-line community, where the individual or their health care manager has access to the individual's genomic profile, phenotype profile, and/or personalized action plan. The individual may choose to have genomic profile, phenotype profile, and/or personalized action plan available for all of the community, a subset of the community, or none of the community to view, through a personal on-line portal. Friends, family, or co-workers may be part of the on-line community. For example, on-line communities such as www.enmeon.com and www.changefire.com are known in the arts, for motivating individuals to achieve their goals. In the present disclosure, an individual participates or is a member of an on-line community that supports and motivates an individual to improve their health and well-being, using as a baseline their phenotype profile, such as GCI scores or by achieving goals on their personalized action plan. The on-line community may be limited to an individual's friends, family, or co-workers, or a combination of friends, family, and co-workers. The individual may also include other members of the on-line community they had not known previously. The on-line community may also be an employer sponsored community. The individual may form groups with others with similar phenotype profiles, action plans, and motivate each other to achieve their goals. Individuals may set up competitions with others in the on-line community, to improve their GCI scores and/or achieve goals on their personalized action plan.

[00160] For example, an individual's report, such as their GCI scores and personalized action plan, may be viewable by an individual's family and friends in the on-line community. An individual may have the choice or option of selecting who may view and/or access their report. The on-line version may comprise a checklist or milestone measure containing items on the personalized action plan, where the individual may mark off accomplishments or the progress of their personalized action plan. The GCI scores may be updated as the progress or accomplishments and reflected on the report on-line. The individual may also input factors that may have changed, such as lifestyle changes, exercise regimen changes, dietary changes, and others, which may also alter the report for the individual. Family and friends may view the progress of the individual, as well as changes in the individual's life, and how they may reflect or alter the individual's GCI score. The on-line portal may allow the individual view initial and subsequent reports. The individual may also receive feedback and comments from their friends and family. Family and friends may leave supporting and motivating comments.

[00161] The on-line community can also provide incentives for an individual to improve their health, by progressing through their personalized action plan and/or improving their GCI scores, decreasing their risk or predisposition to diseases. Incentives can also be provided to individuals not in an on-line community. For example, an employer sponsored online community may offer a health plan that the employer subsidizes more of, provide extra vacation days, or contribute to the health savings account of the individual, when the individual reaches certain goals, such as by improving their GCI score for a disease, thereby decreasing their predisposition to

a disease. Alternatively, the community does not have to be online, and the individual submits their improved GCI score to a designated person that processes the health plans for the employer.

[00162] Other incentives may also be used to motivate an individual to improve their health by improving their GCI score, and/or following their personalized action plan. Individuals may receive points to redeem for rewards when they reach certain goals, such as improving their GCI score by a certain percentage or numerical value, or moving from one category to another (ie. higher risk to lower risk), or by achieving certain goals in the personalized action plan. For example, the individual may achieve a GCI score decrease of a certain numerical value, to achieve the greatest decrease in risk to a disease within a certain timeframe, to accomplish a goal on the personalized action plan, or to accomplish the most goals on a personalized action plan.

[00163] Friends, family, and/or employers may offer points and/or rewards, perhaps by purchasing them, and offering them as a reward to the individual that improves their GCI score or achieves goals on their personalized action plan. Individuals may also receive points/awards for reaching a goal before another person, such as another co-worker, or group of friends, family, or members of an on-line community with the same goal. For example, the first to achieve a GCI score decrease of a certain numerical value, to achieve the greatest decrease in risk to a disease within a certain timeframe, to accomplish a goal on the personalized action plan, or to accomplish the most goals on a personalized action plan. The individual may receive cash, or points to redeem for cash, as rewards. Other rewards may include pharmaceutical products, health products, health club memberships, spa treatments, medical procedures, devices to monitor health, genetic tests, trips, and others, such as subscriptions to services described herein, or discounts, subsidies or reimbursements for the aforementioned items.

[00164] The incentives may be sponsored by friends, family, and employers. Pharmaceutical companies, health clubs, medical device companies, spas, and others may also sponsor incentives. The sponsorship may be in exchange for advertising, or recruiting, for example, pharmaceutical clubs may be interested in obtaining the genome profile of individuals for data, or clinical trials. Furthermore, the incentives may be used to encourage individuals to participate in communities that motivate individuals to improve their health, such as the on-line communities described herein.

Accessing Profiles and Personalized Action Plans

[00165] Reports containing the genomic profile, phenotype profile and other information related to the phenotype and genomic profiles, such as personalized action plans, may be provided to the individual. Health care managers and providers, such as caregivers, physicians, and genetic counselors may also have access to the reports. The reports may be printed, saved on the computer, or viewed on-line. Alternatively, the profiles and action plans may be provided in paper form. They may be in paper, or computer readable format, such as online at a certain time, with subsequent updates provided by paper, computer readable format, or online. The results can be generated and outputted by a computer. They can be stored on a computer readable medium.

[00166] The genomic profile, phenotype profile, as well as personalized action plans can be accessible by an on-line portal, a source of information which can be readily accessed by an individual through use of a computer and internet website, telephone, or other means that allow similar access to information. The on-line portal may optionally be a secure on-line portal or website. It may provide links to other secure and non-secure websites, for example links to a secure website with the individual's phenotype profile, or to non-secure websites such as a message board for individuals sharing a specific phenotype.

[00167] Reports may be of an individual's GCI score, GCI Plus, or EGCI score (as described herein, to report a GCI score will also encompass methods of reporting a GCI, GCI Plus and/or EGCI score). For example, the score, for one or more conditions, can be visualized using a display. A screen (such as a computer monitor or television screen) can be used to visualize the display, such as a personal portal with relevant information. In another embodiment, the display is a static display such as a printed page. The display may include, but is not limited to, one or more of the following: bins (such as 1-5, 6-10, 11-15, 16-20, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, 86-90, 91-95, 96-100), a color or grayscale gradient, a thermometer, a gauge, a pie chart, a histogram or a bar graph. In another embodiment, a thermometer is used to display the GCI score and disease/condition prevalence. The thermometer can display a level that changes with the reported GCI score, for example, the thermometer may display a colorimetric change as the GCI score increases (such as changing from blue, for a lower GCI score, progressively to red, for a higher GCI score). In a related embodiment a thermometer displays both a level that changes with the reported GCI score and a colorimetric change as the risk rank increases

[00168] An individual's GCI score can also be delivered to an individual by using auditory feedback. For example, the auditory feedback can be a verbalized instruction that the risk rank is high or low. The auditory feedback can also be a recitation of a specific GCI score such as a number, a percentile, a range, a quartile or a comparison with the mean or median GCI score for a population. In one embodiment, a live human delivers the auditory feedback in person or over a telecommunications device, such as a phone (landline, cellular phone or satellite phone) or via a personal portal. The auditory feedback can also be delivered by an automated system, such as a computer. The auditory feedback can be delivered as part of an interactive voice response (IVR) system, which is a technology that allows a computer to detect voice and touch tones using a normal phone call. An individual may interact with a central server via an IVR system. The IVR system may respond with pre-recorded or dynamically generated audio to interact with individuals and provide them with auditory feedback of their risk rank. An individual may call a number that is answered by an IVR system. After optionally entering an identification code, a security code or undergoing voice-recognition protocols the IVR system may ask the individual to select options from a menu, such as a touch tone or voice menu. One of these options may provide an individual with his or her risk rank.

[00169] An individual's GCI score may be visualized using a display and delivered using auditory feedback, such as over a personal portal. This combination may include a visual display of the GCI score and auditory feedback, which discusses the relevance of the GCI score to the individual's overall health and possible preventive measures, such as their personalized action plan.

[00170] Different report options may be accessible to the individual. For example, an online access point, such as an online portal may allow an individual to display a single phenotype, or more than one phenotype, based on their genomic profile. The subscriber may also have different viewing options, for example, such as a "Quick View" option, to give a brief synopsis of a single or multiple conditions. A "Comprehensive View" option may also be selected, where more detail for each category is provided. For example, there may be more detailed statistics about the likelihood of the individual developing the phenotype, more information about the typical symptoms or phenotypes, such as sample symptoms for a medical condition, or the range of a physical non-medical condition such as height, or more information about the gene and genetic variant, such as the population incidence, for example in the world, or in different countries, or in different age ranges or genders. For example, a summary of

estimated lifetime risks for a number of conditions may be in a "Quick View" option, while more information for a specific condition, such as prostate cancer or Crohn's disease may be other viewing options. Different combinations and variations may exist for different viewing options.

[00171] The phenotype selected by an individual can be a medical condition and different treatments and symptoms in the report may link to other web pages that contain further information about the treatment. For example, by clicking on a drug, it will lead to website that contains information about dosages, costs, side effects, and effectiveness. It may also compare the drug to other treatments. The website may also contain a link leading to the drug manufacturer's website. Another link may provide an option for the subscriber to have a pharmacogenomic profile generated, which would include information such as their likely response to the drug based on their genomic profile. Links to alternatives to the drug may also be provided, such as preventative action such as fitness and weight loss, and links to diet supplements, diet plans, and to nearby health clubs, health clinics, health and wellness providers, day spas and the like may also be provided. Educational and informational videos, summaries of available treatments, possible remedies, and general recommendations may also be provided.

[00172] The on-line report may also provide links to schedule in-person physician or genetic counseling appointments or to access an on-line genetic counselor or physician, providing the opportunity for a subscriber to ask for more information regarding their phenotype profile. Links to on-line genetic counseling and physician questions may also be provided on the on-line report.

[00173] In another embodiment, the report may be of a "fun" phenotype, such as the similarity of an individual's genomic profile to that of a famous individual, such as Albert Einstein. The report may display a percentage similarity between the individual's genomic profile to that of Einstein's, and may further display a predicted IQ of Einstein and that of the individual's. Further information may include how the genomic profile of the general population and their IQ compares to that of the individual's and Einstein's.

[00174] In another embodiment, the report may display all phenotypes that have been correlated to the individual's genomic profile. In other embodiments, the report may display only the phenotypes that are positively correlated with an individual's genomic profile. In other formats, the individual may choose to display certain subgroups of phenotypes, such as only medical phenotypes, or only actionable medical phenotypes. For example, actionable phenotypes and their correlated genotypes, may include Crohn's disease (correlated with IL23R and CARD 15), Type 1 diabetes (correlated with HLA-DR/DQ), lupus (correlated HLA-DRB1), psoriasis (HLA-C), multiple sclerosis (HLA-DQA1), Graves disease (HLA-DRB1), rheumatoid arthritis (HLA-DRB1), Type 2 diabetes (TCF7L2), breast cancer (BRCA2), colon cancer (APC), episodic memory (KIBRA), and osteoporosis (COL1A1). The individual may also choose to display subcategories of phenotypes in their report, such as only inflammatory diseases for medical conditions, or only physical traits for non-medical conditions. In some embodiments, the individual may choose to show all conditions an estimated risk was calculated for the individual by highlighting those conditions, highlighting only conditions with an elevated risk, or only conditions with a reduced risk.

[00175] Information submitted by and conveyed to an individual may be secure and confidential, and access to such information may be controlled by the individual. Information derived from the complex genomic profile may be supplied to the individual as regulatory agency approved, understandable, medically relevant and/or high impact data. Information may also be of general interest, and not medically relevant. Information can be securely conveyed to the individual by several means including, but not restricted to, a portal interface and/or mailing. More

preferably, information is securely (if so elected by the individual) provided to the individual by a portal interface, to which the individual has secure and confidential access. Such an interface is preferably provided by on-line, internet website access, or in the alternative, telephone or other means that allow private, secure, and readily available access. The genomic profiles, phenotype profiles, and reports are provided to an individual or their health care manager by transmission of the data over a network.

[00176] Accordingly, a representative example logic device through which a report may be generated can comprise a computer system (or digital device) that receives and store genomic profiles, analyze genotype correlations, generate rules based on the analysis of genotype correlations, apply the rules to the genomic profiles, and produce a phenotype profile, a personalized action plan, and report. The computer system may be understood as a logical apparatus that can read instructions from media and/or a network port, which can optionally be connected to server having fixed media. The system can include a CPU, disk drives, optional input devices such as keyboard and/or mouse and optional monitor. Data communication can be achieved through the indicated communication medium to a server at a local or a remote location. The communication medium can include any means of transmitting and/or receiving data. For example, the communication medium can be a network connection, a wireless connection or an internet connection. Such a connection can provide for communication over the World Wide Web. It is envisioned that data relating to the present disclosure can be transmitted over such networks or connections for reception and/or review by a party. The receiving party can be but is not limited to an individual, a health care provider or a health care manager. In one embodiment, a computer-readable medium includes a medium suitable for transmission of a result of an analysis of a biological sample or a genotype correlation. The medium can include a result regarding a phenotype profile of an individual and/or an action plan for the individual, wherein such a result is derived using the methods described herein.

[00177] A personal portal can serve as the primary interface with an individual for receiving and evaluating genomic data. A portal can enable individuals to track the progress of their sample from collection through testing and results. Through portal access, individuals are introduced to relative risks for common genetic disorders based on their genomic profile. The individual may choose which rules to apply to their genomic profile through the portal.

[00178] In one embodiment, one or more web pages will have a list of phenotypes and next to each phenotype a box in which a subscriber may select to include in their phenotype profile. The phenotypes may be linked to information on the phenotype, to help the subscriber make an informed choice about the phenotype they want included in their phenotype profile. The webpage may also have phenotypes organized by disease groups, for example as actionable diseases or not. For example, an individual may choose actionable phenotypes only, such as HLA-DQA1 and celiac disease. The subscriber may also choose to display pre or post symptomatic treatments for the phenotypes. For example, the individual may choose actionable phenotypes with pre-symptomatic treatments (outside of increased screening), for celiac disease, a pre-symptomatic treatment of gluten free diet. Another example may be Alzheimer's, the pre-symptomatic treatment of statins, exercise, vitamins, and mental activity. Thrombosis is another example, with a pre-symptomatic treatment of avoiding oral contraceptives and avoiding sitting still for long periods of time. An example of a phenotype with an approved post symptomatic treatment is wet AMD, correlated with CFH, wherein individuals may obtain laser treatment for their condition.

[00179] The phenotypes may also be organized by type or class of disease or conditions, for example neurological, cardiovascular, endocrine, immunological, and so forth. Phenotypes may also be grouped as medical and non-medical phenotypes. Other groupings of phenotypes on the webpage may be by physical traits, physiological traits, mental traits, or emotional traits. The webpage may further provide a section in which a group of phenotypes are chosen by selection of one box. For example, a selection for all phenotypes, only medically relevant phenotypes, only non-medically relevant phenotypes, only actionable phenotypes, only non-actionable phenotypes, different disease group, or “fun” phenotypes. “Fun” phenotypes may include comparisons to celebrities or other famous individuals, or to other animals or even other organisms. A list of genomic profiles available for comparison may also be provided on the webpage for selection by the individual to compare to the individual’s genomic profile.

[00180] The on-line portal may also provide a search engine, to help the individual navigate the portal, search for a specific phenotype, or search for specific terms or information revealed by their phenotype profile or report. Links to access partner services and product offerings may also be provided by the portal. Additional links to support groups, message boards, and chat rooms for individuals with a common or similar phenotype may also be provided. The on-line portal may also provide links to other sites with more information on the phenotypes in an individual’s phenotype profile. The on-line portal may also provide a service to allow individuals to share their phenotype profile and reports with friends, families, co-workers, or health care managers, and may choose which phenotypes to show in the phenotype profile they want shared with their friends, families, co-workers, or health care managers.

[00181] The phenotype profiles and reports provide a personalized genotype correlation to an individual. The genotype correlations used to generate a personalized action plan that provides individuals with increased knowledge and opportunities to determine their personal health care and lifestyle choices. If a strong correlation is found between a genetic variant and a disease for which treatment is available, detection of the genetic variant may assist in deciding to begin treatment of the disease and/or monitoring of the individual. In the case where a statistically significant correlation exists but is not regarded as a strong correlation, an individual can review the information with a personal physician and decide an appropriate, beneficial course of action. Potential courses of action that could be beneficial to an individual in view of a particular genotype correlation include administration of therapeutic treatment, monitoring for potential need of treatment or effects of treatment, or making life-style changes in diet, exercise, and other personal habits/activities, which can be personalized based on an individual’s genomic profile into a personalized action plan. Other personal information, such as existing habits and activities can also be incorporated into a personalized action plan. For example, an actionable phenotype such as celiac disease may have a pre-symptomatic treatment of a gluten-free diet, and provided in a personalized action plan. Likewise, genotype correlation information could be applied through pharmacogenomics to predict the likely response an individual would have to treatment with a particular drug or regimen of drugs, such as the likely efficacy or safety of a particular drug treatment.

[00182] Genotype correlation information can also be used in cooperation with genetic counseling to advise couples considering reproduction, and potential genetic concerns to the mother, father and/or child. Genetic counselors may provide information and support to individuals with phenotype profiles that display an increased risk for specific conditions or diseases. They may interpret information about the disorder, analyze inheritance patterns and risks of recurrence, and review available options with the subscriber. Genetic counselors may also provide supportive counseling refer subscribers to community or state support services. Genetic counseling may be included with

specific subscription plans. Genetic counseling options can also include those that are scheduled within 24 hours of request and available during non-traditional hours, such as evenings, Saturdays, Sundays, and/or holidays.

[00183] An individual's portal can also facilitate delivery of additional information beyond an initial screening. Individuals can be informed about new scientific discoveries that relate to their personal genetic profile, such as information on new treatments or prevention strategies for their current or potential conditions. The new discoveries may also be delivered to their healthcare managers. The new discoveries can be incorporated into updated or revised personal action plans. The individuals or their healthcare providers can be informed of new genotype correlations and new research about the phenotypes in the individual's phenotype profiles by e-mail. For example, e-mails of "fun" phenotypes can be sent to individuals, for example, an e-mail may inform them that their genomic profile is 77% identical to that of Abraham Lincoln and that further information is available via an on-line portal.

[00184] Computer code for notifying subscribers of new or revised correlations new or revised rules, and new or revised reports, for example with new prevention and wellness information, information about new therapies in development, or new treatments available, is also provided herein. A system of computer code for generating new rules, modifying rules, combining rules, periodically updating the rule set with new rules, maintaining a database of genomic profile securely, applying the rules to the genomic profiles to determine phenotype profiles, generating personalized action plans and reports is also provided by the present disclosure, including computer code for granting different levels of access and options for individuals with different subscriptions.

Subscriptions

[00185] The genomic profiles, phenotype profiles, and reports, including personalized action plans may be generated for individuals that are human or non-human. For example, individuals may include other mammals, such as bovines, equines, ovines, canines, or felines. An individual may be a person's pet, and the owner of the pet may want a personal action plan to increase the health and longevity of their pet. Individuals, or their health care managers, may be subscribers. As described herein, subscribers are human individuals who subscribe to a service by purchase or payment for one or more services. Services may include, but are not limited to, one or more of the following: having their or another individual's, such as the subscriber's child or pet, genomic profile determined, obtaining a phenotype profile, having the phenotype profile updated, and obtaining reports based on their genomic and phenotype profile, including a personalized action plan.

[00186] Subscribers may choose to provide the genomic and phenotype profiles or reports to their health care managers, such as a physician or genetic counselor. The genomic and phenotype profiles may be directly accessed by the healthcare manager, by the subscriber printing out a copy to be given to the healthcare manager, or have it directly sent to the healthcare manager through the on-line portal, such as through a link on the on-line report.

[00187] A genomic profile may be generated for subscribers and non-subscribers and stored digitally, but access to the phenotype profile and reports may be limited to subscribers. For example, access to at least one GCI score is provided to a subscriber, but not to non-subscribers. In another variation, both subscribers and non-subscribers may access their genotype and phenotype profiles, but have limited access, or have a limited report generated for non-subscribers, whereas subscribers have full access and may have a full report generated. In another embodiment, both subscribers and non-subscribers may have full access initially, or full initial reports, but only subscribers may access updated reports based on their stored genomic profile. For example, access is provided to non-subscribers, where they may have limited access to at least one of their GCI scores, or they may have an initial report on at least

one of their GCI scores generated, but updated reports are generated only with purchase of a subscription. Health care managers and providers, such as caregivers, physicians, and genetic counselors may also have access to at least one of an individual's GCI scores.

[00188] In some embodiments, access to EGCI scores may be limited depending on the various subscription levels. For example, an individual may subscribe to have their GCI score, but have limited access to their EGCI score, or limited access to specific conditions or diseases with EGCI scores. Alternatively, GCI scores may be provided to non-subscribers and EGCI scores provided to subscribers. Subscription levels may also vary depending on an individual updating or modifying their environmental factors to generate updated or revised EGCI scores. For example, an individual may pursue an ongoing subscription to have unlimited access to a system to update their environmental factors. Alternatively, an individual may choose not to have an ongoing subscription, but pay each time they update their environmental factors to generate a new EGCI score. The updating of EGCI scores may also incorporate new scientific information, such as new correlations discovered between a genetic polymorphism and a disease or condition, or other genetic factors and their associating with one or more diseases or conditions. Individuals may also have the option to generate EGCI scores based on environmental factors they may want to change. For example, an individual may be contemplating moving to a certain city, and the individual may input or select certain environmental factors associated with the city to see the effect on their EGCI score.

[00189] Other subscription models may include one that provides a phenotype profile where the subscriber may choose to apply all existing rules to their genomic profile, or a subset of the existing rules, to their genomic profile. For example, they may choose to apply only the rules for disease phenotypes that are actionable. The subscription may be of a class, such that there are different levels within a single subscription class. For example, different levels may be dependent on the number of phenotypes a subscriber wants correlated to their genomic profile, or the number of people that may access their phenotype profile.

[00190] Another level of subscription may be to incorporate factors specific to an individual, such as already known phenotypes such as age, gender, or medical history, to their phenotype profile. Still another level of the basic subscription may allow an individual to generate at least one GCI score for a disease or condition. A variation of this level may further allow an individual to specify for an automatic update of at least one GCI score for a disease or condition to be generated if there is any change in at least one GCI score due to changes in the analysis used to generate at least one GCI score. In some embodiments the individual may be notified of the automatic update by email, voice message, text message, mail delivery, or fax.

[00191] Subscribers may also generate reports that have their phenotype profile as well as information about the phenotypes, such as genetic and medical information about the phenotype. Different amount of information that an individual may access can depend on the level of subscription they have. For example, different viewing options an individual may have could depend on their level of subscription, such as a quick view for non-subscribers or a more basic subscription, but a comprehensive view is accessible to those with a full subscription.

[00192] For example, different levels of subscriptions may have different variations or combinations of accessibility to information including, but not limited to, the prevalence of the phenotype in the population, the genetic variant that was used for the correlation, the molecular mechanism that causes the phenotype, therapies for the phenotype, treatment options for the phenotype, and preventative actions, may be included in the report. In other embodiments, the reports may also include information such as the similarity between an individual's genotype and

that of other individuals, such as celebrities or other famous people. The information on similarity may be, but not limited to, percentage homology, number of identical variants, and phenotypes that may be similar. These reports may further contain at least one GCI score.

[00193] Other options based on subscription level may include links to other sites with further information on the phenotypes, links to on-line support groups and message boards of people with the same phenotype or one or more similar phenotypes, links to an on-line genetic counselor or physician, or links to schedule telephonic or in-person appointments with a genetic counselor or physician, if the report is accessed on-line. If the report is in paper form, the information may be the website location of the aforementioned links, or the telephone number and address of the genetic counselor or physician. The subscriber may also choose which phenotypes to include in their phenotype profile and what information to include in their report. The phenotype profile and reports may also be accessible by an individual's health care manager or provider, such as a caregiver, physician, psychiatrist, psychologist, therapist, or genetic counselor. The subscriber may be able to choose whether the phenotype profile and reports, or portions thereof, are accessible by such individual's health care manager or provider.

[00194] Another level of subscription may be to maintain the genomic profile of an individual digitally after generation of an initial phenotype profile and report, and provides subscribers the opportunity to generate phenotype profiles and reports with updated correlations from the latest research. Subscribers may have the opportunity to generate risk profile and reports with updated correlations from the latest research. As research reveals new correlations between genotypes and phenotypes, disease or conditions, new rules will be developed based on these new correlations and can be applied to the genomic profile that is already stored and being maintained. The new rules may correlate genotypes not previously correlated with any phenotype, correlate genotypes with new phenotypes, modify existing correlations, or provide the basis for adjustment of a GCI score based on a newly discovered association between a genotype and disease or condition. Subscribers may be informed of new correlations via e-mail or other electronic means, and if the phenotype is of interest, they may choose to update their phenotype profile with the new correlation. Subscribers may choose a subscription where they pay for each update, for a number of updates or an unlimited number of updates for a designated time period (e.g. three months, six months, or one year). Another subscription level may be where a subscriber has their phenotype profile or risk profile automatically updated, instead of where the individual chooses when to update their phenotype profile or risk profile, whenever a new rule is generated based on a new correlation.

[00195] Subscribers may also refer non-subscribers to the service that generates rules on correlations between phenotypes and genotypes, determines the genomic profile of an individual, applies the rules to the genomic profile, and generates a phenotype profile of the individual. Referral by a subscriber may give the subscriber a reduced price on subscription to the service, or upgrades to their existing subscriptions. Referred individuals may have free access for a limited time or have a discounted subscription price.

[00196] The following examples illustrate and explain the embodiments described herein. The scope of the disclosure is not limited by these examples.

EXAMPLES

Example 1: Evaluation of GCI

[00197] The WTCCC data (*Wellcome Trust Case Control Consortium, Nature. 447:661-678 (2007)*) is used to test the GCI framework. This dataset contains the genotypes of approximately 14,000 individuals divided into seven subpopulations based on disease phenotypes and one unaffected control subpopulation of 1,500 samples from the UK Blood Service Control Group. The GCI is tested in the context of three different diseases: Type 2 Diabetes, Crohn's disease, and Rheumatoid Arthritis, which differ substantially in their heritability and average lifetime risk. Thus, analysis is limited to the Type 2 Diabetes, Crohn's Disease, and Rheumatoid Arthritis subpopulations and the control group. SNPs that were reported in the literature to be significantly associated with each of these conditions, and that passed a set of quality criteria (see Table 2) are used.

Table 2: Allele frequencies and the relative risks Type 2 Diabetes, Crohn's Disease, and Rheumatoid Arthritis.

Disease	dbSNP rs iID	Relative risk ¹ for RR	Relative Risk ¹ for RN	Frequency ² of RR	Frequency ² of RN
Type 2 Diabetes	rs10012946 ³	1.1464	1.0239	0.5000	0.4667
	rs10811661 ⁴	1.3008	1.1282	0.6667	0.2500
	rs1801282 ⁴	1.4128	1.2417	0.8667	0.1167
	rs4402960 ⁴	1.1602	1.1233	0.1167	0.3500
	rs4506565 ⁵	1.6133	1.2738	0.0847	0.3729
	rs5215 ⁴	1.1681	1.0935	0.1000	0.6167
	rs8050136 ⁶	1.3609	1.1176	0.1167	0.6667
	rs9494266 ⁷	1.4909	1.2296	0.0169	0.0847
Crohn's Disease	rs1000113 ⁵	1.9102	1.5354	0.0000	0.0667
	rs10210302 ⁵	1.8433	1.1890	0.3000	0.5000
	rs10761659 ⁵	1.5461	1.2287	0.2333	0.6333
	rs10883365 ⁵	1.6154	1.1989	0.3000	0.4000
	rs11805303 ⁵	1.8525	1.3875	0.1000	0.3833
	rs17221417 ⁵	1.9118	1.2883	0.1000	0.5167
	rs17234657 ⁵	2.3053	1.5360	0.0667	0.2000
	rs2542151 ⁵	1.9997	1.2980	0.0500	0.2833
	rs9858542 ⁵	1.8316	1.0895	0.0333	0.4167
Rheumatoid Arthritis	rs10118357 ⁸	1.7278	1.3152	0.2712	0.5254
	rs13207033 ⁸	1.7559	1.3258	0.6667	0.3167
	rs6457617 ⁵	5.0847	2.3414	0.2167	0.5667
	rs6679677 ⁹	3.1672	1.6847	0.0000	0.2833
	rs6920220 ⁵	1.7023	1.1965	0.0000	0.3500

- 1: The relative risks provided here were calculated using the GCI methodology, as described herein.
- 2: The allele frequencies are taken from the HapMap project's CEU population.
- 3: Sandhu et al., Nat Genet. 39:951-3 (2007).
- 4: Scott et al., Science. 316:1341-5 (2007).
- 5: Wellcome Trust Case Control Consortium, Nature. 447:661-78 (2007).
- 6: Zeggini et al., Science. 316:1336-41 (2007).
- 7: Salonen et al., Am J Hum Genet. 81:338-45 (2007).
- 8: Remmers et al., N Engl J Med. 357:977-86 (2007).
- 9: Kyogoku et al., Am J Hum Genet. 75:504-7 (2004).

[00198] For each of these SNPs, the relative lifetime risk is computed as described herein based on the empirical distribution of alleles found in the WTCCC dataset, and the GCI formulation was used to calculate an estimated risk per individual. Some of the known risk variants are not present on the Affymetrix 500k GeneChip array that was used by the WTCCC, and therefore the predictability of the GCI is expected to be likely better than what is presented in the analysis below.

[00199] The Receiver Operating Curves (ROC) (*The Statistical Evaluation of Medical Tests for Classification and Prediction*, MS Pepe. Oxford Statistical Science Series, Oxford University Press (2003)) is used to evaluate the ability of the GCI to serve as a predictive test for a condition. For a perfect test, a threshold t would be chosen such that all individuals with a score larger than t would develop the condition, and all individuals with a score less than t would not. However, in practice, for any given threshold there is some fraction of false positive and false negative assignments. The ROC curve graphically depicts the relationship between false positive rates and true positive rates, and thus it can be used to guide the tradeoffs between test sensitivity and specificity. The area under the ROC curve (AUC) is used as a quantitative measure to compare different risk estimate scores. The AUC can also show the relative benefit of any score as compared to the optimal scenario in which the genetic causes of the condition are fully understood. In general, the larger the value of the AUC, the better the score for the classification. If classification is done randomly, the AUC is expected to be 0.5 and for the optimal score (i.e. a score function for which the true positive fraction becomes 1 and false positive fraction becomes 0 at some threshold) the AUC is equal to 1.

[00200] In order to have a baseline for comparison, the logistic regression to calculate the best model that leverages interactions between the SNPs to fit the data is used. If the SNPs are $s_1, s_2, \dots, s_n$, then the model assumes that the logit is $X = a_1s_1 + a_2s_2 + \dots + a_ns_n + a_{12}s_{12} + \dots + a_{n-1,n}s_{n-1,n}$, where s_{ij} is the interaction between s_i and s_j . The fitted probability is used as an estimate for the risk, and generates a ROC curve for these risk estimates. This model takes into account pairwise interactions between the SNPs, and it should therefore be at least as accurate as the GCI score, which typically does not consider them. Furthermore, if there is linkage disequilibrium between a pair of SNPs, logistic regression may have difficulty accommodating this correlation, while the GCI typically ignores it. Thus comparing logistic regression analysis models to the proposed GCI score allows the ability to measure the effect of various assumptions on the predictive power of the GCI. **FIGURE 1** shows the ROC curves for all three disease scenarios, and Table 3 gives their AUCs. The AUC for the GCI and for the logistic regression are quite similar for all three diseases (Table 3), leading to the conclusion that SNP-SNP interactions do not add substantial information for the risk assessment, at least not for these diseases and these SNPs. Therefore, it can be justified that the assumption that the SNP-SNP interactions can be ignored as long as there is no evidence for such an interaction from previous studies.

Table 3: The area under the ROC curve for three different diseases under three different scores.

Disease	Heritability	Average Lifetime Risk	Optimal Scenario ¹	GCI score	Logistic Regression
Type 2 Diabetes	64% [21]	25.0% [24]	0.902	0.597	0.604
Crohn's Disease	80% [22]	0.56% [25]	0.982	0.654	0.646
Rheumatoid Arthritis	53% [23]	1.54% [26]	0.944	0.675	0.689

1: The ideal score when the complete genetic information is known.

[00201] The GCI ROC curve is compared to a theoretical disease model. This disease model assumes that the disease is affected by both environmental and genetic factors, and that the two factors are independent. The phenotype P is denoted by $P=G+E$, where G is the genetic risk and E is the environmental risk. The first model, also referred to as the *continuous model*, assumes that G and E are normally distributed with standard deviations σ_G and σ_E respectively, and that an individual will develop the condition in his lifetime if $P > \alpha$ for a fixed α . Since the heritability h is known for many complex diseases, σ_G , σ_E , and α is fixed using the constraint that $h = \sigma_G^2 / (\sigma_G^2 + \sigma_E^2)$, and that the average lifetime risk is $\Pr(P > \alpha)$. Since the heritability and average lifetime risks are known for each of the conditions being tested, the parameters of the models can be set according to the disease. 100,000 random samples from the distribution P based on this model is generated. G is assumed to be known for each individual (but E and the disease status are unknown), and a ROC curve is generated based on G . This represents the optimal scenario where the genetic risk is entirely understood and can be measured correctly for every individual. For this disease model, the AUC for the optimal scenario only depends on the heritability and the average lifetime risk of the disease, and not on the choice of σ_G , σ_E , or α .

[00202] The theoretical maximum of the area under the ROC curve for this first model depends only on the average lifetime risk (ALTR) and the heritability of the disease. Let σ_e denote the variance in the environmental variable and σ_g denote the variance in the genetic variable. In this model, both genetic (G) and environmental (E) variables are normally distributed. The theoretical maximum of ROC curve is obtained when the genetic variable is known exactly while the environmental variable is unknown. An individual is a true case if $G + E > \alpha$ and a true control otherwise. For any cutoff chosen for the genetic variable, the individuals who are above that cutoff will be considered as cases and the rest as controls. The true positive fraction (TPF) is the fraction of true cases that are called as cases and false positive fraction (FPF) is the fraction of true controls that are called as cases. The TPF versus FPF for different values of cutoffs gives us the ROC curve.

[00203] The probability that an individual's genetic variable is greater than some cutoff (c) is given by:

$$P(G > c) = \int_{\beta\sigma_g}^{\infty} e^{-x^2/2\sigma_g^2} dx / \sqrt{2\pi}\sigma_g, \text{ where } \beta = c / \sigma_g$$

[00204] The probability that an individual's genetic variable is greater than the cutoff and the individual is a true case is:

$$P(G > c \text{ and } G + E > \alpha) = \int_{\beta\sigma_g}^{\infty} e^{-x^2/2\sigma_g^2} \left(\int_{\gamma\sqrt{\sigma_g^2 + \sigma_e^2} - x}^{\infty} e^{-y^2/2\sigma_e^2} dy / \sqrt{2\pi}\sigma_e \right) dx / \sqrt{2\pi}\sigma_g, \text{ where } \gamma = \alpha / \sqrt{\sigma_g^2 + \sigma_e^2}.$$

[00205] For any non-zero average lifetime risk, γ is fixed because α increases linearly with $\sqrt{\sigma_g^2 + \sigma_e^2}$.

[00206] By definition heritability, $h = \sigma_g^2 / (\sigma_g^2 + \sigma_e^2)$

[00207] The integral within the brackets in the previous double integral can be expressed in terms of the error function, erf. Because the cumulative distribution function of the normal distribution is given by $0.5(1 + \text{erf}(y/\sqrt{2}\sigma_e))$, the integral inside the brackets is $1 - 0.5(1 + \text{erf}(y/\sqrt{2}\sigma_e)) = 0.5 - 0.5\text{erf}([\gamma\sqrt{\sigma_g^2 + \sigma_e^2} - x]/\sqrt{2}\sigma_e)$. Thus, the probability that an individual is a true case and its genetic variable is greater than c can be expressed as:

$$\int_{\beta\sigma_g}^{\infty} e^{-x^2/2\sigma_g^2} (0.5 - 0.5\text{erf}(\gamma f(h) - g(h)x/\sqrt{2}\sigma_g)) dx / \sqrt{2\pi}\sigma_g, \text{ where } f(h) \text{ and } g(h) \text{ are some functions of the}$$

heritability. Substituting, $t = x/\sqrt{2}\sigma_g$ into this equation, we can see that $\sqrt{2}\sigma_g dt = dx$. Therefore, $P(G > c \text{ and } G +$

$$E > \alpha) \text{ can be expressed as: } \int_{\beta/\sqrt{2}}^{\infty} e^{-t^2} (0.5 - 0.5\text{erf}(\gamma f(h) - g(h)t)) dt / \sqrt{\pi}$$

[00208] Similarly, the probability that an individual is a true control and its genetic variable is greater than c i.e. P

$$(G > c \text{ and } G + E \leq \alpha) = \int_{\beta/\sqrt{2}}^{\infty} e^{-t^2} (0.5 + 0.5\text{erf}(\gamma f(h) - g(h)t)) dt / \sqrt{\pi}$$

[00209] Therefore, the true positive fraction for any given β only depends on h and ALTR since:

$$\text{TPF} = P(G > c \text{ and } G + E > \alpha) / \text{ALTR}.$$

[00210] The same is also true for false positive fraction since $\text{FPF} = P(G > c \text{ and } G + E \leq \alpha) / [1 - \text{ALTR}]$. Hence, the total area under the theoretical ROC curve, which is based on TPF and FPF at all possible values of β , is independent of σ_e and σ_g .

[00211] In the second model, or the *discrete model*, a variant of the previous model, it is assumed that $G = \sum \lambda_i X_i + Y$, where Y is normally distributed with standard variation σ_Y , and $X_i \sim B(2, p_i)$ is Binomially distributed. In this case, X_i corresponds to SNPs with large effects, and Y represents many other small genetic effects; if there are enough small genetic effects, it can be expected that the asymptotic behavior of their sum would be according to a Normal distribution. By setting the parameters λ , σ_Y , and p appropriately, the relative risks of the large effect SNPs can be controlled. These parameters are chosen such that the relative risks are close to values observed in real data (see Table 4). Similar to the previous model, if G is known (but E is unknown) and the relative risks of the large effect SNPs and risk-allele frequencies are fixed, then the area under the ROC curve for the discrete model only depends on the heritability and the average lifetime risk of the disease.

[00212] A result similar to that for model 1 is obtained for disease model 2. In particular, if the relative risks of SNPs known to be associated with a disease and the risk-allele frequencies (p_i) are fixed, then the total area under the ROC curve depends only on the heritability and the average lifetime risk of the disease. In this model, the genetic variable is $G = \sum \lambda_i X_i + G_1$. Here $G_1 \sim N(0, \sigma_{G1})$ and the X_i s are distributed according to a Binomial

distribution $B(2, p_i)$, where p_i is the allele frequency of the risk allele at locus i . $B(2, p_i)$ gives the number of risk allele copies in an individual at locus i . $X_i = 0$ means homozygous for non-risk allele, $X_i = 1$ means heterozygote and $X_i = 2$ means homozygous for risk allele. The normal variable represents the unknown genetic component. As before, the environmental variable E is also normally distributed with mean 0 and standard deviation σ_e . The phenotype is given by $P = G + E$ and individuals with $P > \alpha$ are diseased whereas the rest are controls. α is chosen such that the fraction of diseased individuals equals the average lifetime risk of the disease.

[00213] Heritability for this model is $h = [\sigma_{g1}^2 + \Sigma 2\lambda_i^2 p_i(1 - p_i)] / [\sigma_{g1}^2 + \sigma_e^2 + \Sigma 2\lambda_i^2 p_i(1 - p_i)]$. Let us assume that the relative risks of the known SNPs for heterozygous genotypes are fixed and denote these by RN_i . By definition, the relative risk of heterozygote is given by:

$RN_i = \Pr(G + E > \alpha \mid X_i = 1) / \Pr(G + E > \alpha \mid X_i = 0) = [\Sigma \Pr(G1 + E > \alpha - z - \lambda_i) \Pr(W = z)] / [\Sigma \Pr(G1 + E > \alpha - z) \Pr(W = z)]$, where $W = \Sigma \lambda_j X_j$ for all j not equal to i . Let erf denote the error function and erfc denote the

complementary error function (i.e. $1 - \text{erf}(x)$). Since $G1 + E \sim N(0, \sqrt{(\sigma_{g1}^2 + \sigma_e^2)})$, the relative risk expressed in terms of complementary error function is given by:

$\Sigma 0.5 \text{erfc}[(\alpha - z - \lambda_i) / \sqrt{2(\sigma_{g1}^2 + \sigma_e^2)}] \Pr(W = z) / \Sigma 0.5 \text{erfc}[(\alpha - z) / \sqrt{2(\sigma_{g1}^2 + \sigma_e^2)}] \Pr(W = z)$. Thus, if λ_i s with

disease cutoff α represent the solutions for the SNPs for some choice of $\sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$ (these may or may not be unique), then $L\lambda_i$ s with cutoff of $L\alpha$ will necessarily be solutions if the standard deviation of $G1$ and E get changed

by a factor of L . This follows because z is always a linear combinations of λ_i s. Therefore, $\lambda_i / \sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$ and $\gamma = \alpha / \sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$ are independent of $\sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$ and depend on heritability and ALTR alone.

[00214] By definition, $h(\sigma_{g1}^2 + \sigma_e^2) = (1 - h)\Sigma 2\lambda_i^2 p_i(1 - p_i) + \sigma_{g1}^2$. This therefore means that:

$\sigma_{g1}^2 / (\sigma_{g1}^2 + \sigma_e^2) = h - (1 - h) \Sigma 2\lambda_i^2 p_i(1 - p_i) / (\sigma_{g1}^2 + \sigma_e^2)$. Since $\lambda_i / \sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$ and p_i are independent of

$\sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$, $\sigma_{g1}^2 / (\sigma_{g1}^2 + \sigma_e^2)$ is a function of heritability and ALTR alone. Let $Z = \Sigma \lambda_i X_i$ and V denote the

vector of X_i values. Then, if $Z = z$ for $V = v$, $z / \sqrt{2\sigma_{g1}^2} = b(h, \text{ALTR}, v)$ is a function of the heritability, ALTR and v alone and is independent of $\sqrt{(\sigma_{g1}^2 + \sigma_e^2)}$.

[00215] The true positive fraction is defined as: $\Pr(G > c \ \& \ G + E > \alpha) / \Pr(G + E > \alpha)$ where c denotes the cutoff for genetic variable. Let $\beta = c / \sigma_{g1}$. The numerator for TPF can be calculated as:

$$\Sigma \Pr(V = v, Z = z) \int_{\beta \sigma_{g1} - Z}^{\infty} e^{-x^2 / 2\sigma_{g1}^2} \left(\int_{\gamma \sqrt{\sigma_{g1}^2 + \sigma_e^2} - x - Z}^{\infty} e^{-y^2 / 2\sigma_e^2} dy / \sqrt{2\pi\sigma_e} \right) dx / \sqrt{2\pi\sigma_{g1}}$$

[00216] Using the error function to express the cumulative distribution function of the normal distribution, $\Pr(G > c \ \& \ G + E > \alpha)$ is:

$$\Sigma \Pr(V = v, Z = z) \int_{\beta \sigma_{g1} - Z}^{\infty} e^{-x^2 / 2\sigma_{g1}^2} (0.5 - 0.5 \text{erf}[r(h, \text{ALTR}, v) - s(h, \text{ALTR})x / \sqrt{2\sigma_{g1}}]) dx / \sqrt{2\pi\sigma_{g1}}, \text{ where}$$

r and s are some functions. Substituting $t = x/\sqrt{2}\sigma_{gl}$ into this equation, we can see that $\sqrt{2}\sigma_{gl} dt = dx$. Therefore, $P(G > c \text{ and } G + E > \alpha)$ can be expressed as:

$$\sum \Pr(V = v, Z = z) \int_{(\beta/\sqrt{2}) - b(h, ALTR, v)}^{\infty} e^{-t^2} (0.5 - 0.5 \operatorname{erf}[r(h, ALTR, v) - s(h, ALTR)t]) dt / \sqrt{\pi}$$

[00217] Similarly, the probability that an individual is a true control and its genetic variable is greater than c i.e. $P(G > c \text{ and } G + E \leq \alpha) =$

$$\sum \Pr(V = v, Z = z) \int_{(\beta/\sqrt{2}) - b(h, ALTR, v)}^{\infty} e^{-t^2} (0.5 + 0.5 \operatorname{erf}[r(h, ALTR, v) - s(h, ALTR)t]) dt / \sqrt{\pi}$$

[00218] The $ALTR = P(G + E > \alpha)$ and $\Pr(V=v, Z=z)$ is fixed if p_i s are fixed. Therefore, the true positive fraction for any given β only depends on the h and $ALTR$. The same is also true for false positive fraction since $FPF = \Pr(G > c \text{ and } G + E \leq \alpha) / [1 - ALTR]$. So, the area under the theoretical ROC curve, which is based on TPF and FPF at all possible values of β , is independent of σ_e , σ_{gl} and λ_i s.

[00219] Solving for $\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$, $1 - (\sigma_{gl}^2 / (h(\sigma_{gl}^2 + \sigma_e^2))) = (1-h) \sum 2\lambda_i^2 p_i(1-p_i) / (h(\sigma_{gl}^2 + \sigma_e^2))$. So,

$0 \leq \lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)} \leq \sqrt{h / (2p_i(1-p_i)(1-h))}$ since LHS is always less than 1. Solutions for all

$\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ can be obtained simultaneously by using the following iterative procedure.

[00220] Initially, determine the $\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ for each SNP assuming that it is the only SNP present (i.e.

assuming $\lambda_j = 0$ for all j not equal to i). This can be done using a binary search between 0 and

$\sqrt{h / (2p_i(1-p_i)(1-h))}$ since RN_i increases with $\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$.

[00221] These values are initial guesses for $\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$. Then, 1) Determine $\lambda_i / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ assuming

that $\lambda_j / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ for other SNPs are equal to what was calculated previously. 2) Determine

$\lambda_2 / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ assuming that $\lambda_j / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ for other SNPs are equal to what was calculated previously. 3)

Determine $\lambda_n / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ assuming that $\lambda_j / \sqrt{(\sigma_{gl}^2 + \sigma_e^2)}$ for other SNPs are equal to what was calculated previously. If all RN_i values are sufficiently close to observed values, stop. If not, go back to step 1.

[00222] Thus, two sets of optimal ROC curves that would result if all of the genetic, but not environmental, variability were known and modeled are presented. The first model assumes that there are many small genetic effects that are cumulative (and thus the genetic effect is represented by a normally distributed random variable), while the second model assumes that there are a small number of genetic variants with large effects in addition to many others with small effects. Both models take into account the heritability and lifetime risk of the condition, resulting in a realistic extrapolation of the unknown genetic risk factors based on the currently known ones.

FIGURE 1 shows the ROC curves for these scenarios and Table 3 gives their areas. The GCI area under the curve is

less than the optimal theoretical generic models, which suggests additional unknown genetic variants and/or interactions are expected to affect these diseases.

[00223] Based on **FIGURE 1**, improvement in predictive modeling that will most likely only come through the discovery of additional genetic variants for the three conditions discussed herein. It is useful to know what percentage of the genetic factors have been captured to date. An estimate of this quantity using the ROC curves approach is developed with the main assumption that the major genetic factors have already been discovered, and that there are many other undiscovered genetic factors with lower relative risks.

[00224] The potential number of additional independent common (minor allele frequency 10% or greater) variants, where each such variant contributes a relative risk of 1.1 for the homozygous risk variant and 1.05 for the heterozygous variant, is estimated, the estimate essentially providing the number of such variants that will suffice to obtain a ROC curve with an AUC as large as the theoretically optimal bound.

[00225] For each of the three conditions, the genetic factors are assumed to be the known ones (as in Table 2), in addition to some unknown number k of variants with low relative risk. Based on simulations of 100,000 individuals, nearly 1,600 additional variants are needed to explain the genetic variants of Type 2 Diabetes. This is intuitive, as the AUC of type 2 diabetes is quite low with current knowledge, despite the high heritability value of 64%. For Crohn's disease and rheumatoid arthritis the results are even more striking, since 13,958 and 6,237 additional genetic factors are expected to be found, respectively. Therefore, currently known genetic variants account for 4%-14% of the total genetic variation for these conditions (see Table 4). These results however are conditioned upon the fact that no other large effects are expected to be found, while in fact there still may be some large effects that are due to SNP-SNP or SNP-environment interactions or other less studied variants (e.g., copy number variants, rare variants, epigenetic variants).

Table 4: Estimated number of low effect genetic variants missing for three diseases.

Disease	Estimated number of unknown variants*	Fraction of genetic variation explained by variants included in the model
Type 2 Diabetes	1600	7%
Crohn's Disease	13958	4.4%
Rheumatoid Arthritis	6237	14.4%

*each with homozygote relative risk of 1.10, heterozygote relative risk of 1.05, and minor allele frequency of 10%.

Example 2: Theoretical Effect of Unknown SNP-SNP Interactions

[00226] The GCI score is based on the assumptions that all SNPs are independent of one another and that they have independent effects on the risk for the disease. As shown in **FIGURE 1**, the three examples studied here show no significant difference between the GCI model and a model in which pairwise dependencies among the SNPs are included through logistic regression. There are some known examples in which SNP-SNP interactions do exist in other diseases and have to be taken into account (for example, *Zheng et al., N Engl J Med. 358:910-919 (2008)*). If these interactions are known, they can easily be incorporated into the GCI model. However, it is important to understand the effect of unknown SNP-SNP interactions on the risk estimates.

[00227] In order to explore the issue of interactions in greater detail, datasets are simulated under an interaction model where the relative risks are not independent for a single pair of SNPs in the dataset. The simulated case-control data to plot ROC curves based on two approaches for risk assessment is used. First, the relative risk of an individual according to the interaction model is calculated. Then, the relative risks according to the GCI approach are assigned, which assumes a multiplicative model. As observed in **FIGURE 2**, and in Table 5, the ROC curves differ substantially only when the interaction factor is very high.

Table 5: The area under the curve (AUC) for the different interaction scenarios.

	Simulated Interaction Factor 2 ¹		Simulated Interaction Factor 10 ²	
	Interaction risk estimate	GCI risk estimate (Multiplicative)	Interaction risk estimate	GCI risk estimate (Multiplicative)
Crohn's Disease	0.676	0.664	0.833	0.727
Rheumatoid Arthritis	0.709	0.699	0.843	0.761
Type 2 Diabetes	0.633	0.619	0.709	0.646

1. The two columns correspond to the case where there is a SNP-SNP interaction in which the effect of a certain combination of genotypes has two times the product of the marginal effects

2. The two columns correspond to the case where there is a SNP-SNP interaction in which the effect of a certain combination of genotypes is 10 times the product of the marginal effects.

[00228] However, such strong interactions between pairs of SNPs are likely to have been discovered in genome wide association studies and that it would be exceptional to find that two SNPs entering into such a strong interaction do not have detectable main effects. Particularly, whole-genome association studies often report that SNP-SNP interactions were tested but were not found to be significant (e.g. *Barrett et al., Nature Genet. 40:955-962 (2008)*). Therefore, when no such interactions have been reported in the literature for a set of SNPs, it is unlikely that the classification accuracy of the simple multiplicative test will differ dramatically from that of the true model that includes interactions.

[00229] To test the effect of unknown SNP-SNP interactions, data based on the following model is simulated. Let λ_i denote the relative risk of the disease for a particular combination of genotypes (g_i) and p denote the average probability of developing the disease (i.e., lifetime risk). By definition of relative risk, $\lambda_i = P(\text{disease} | g_i) / P(\text{disease} | g_0)$. Here, g_0 denotes the genotype with the smallest chance of developing the disease. In the simple multiplicative

model, the relative risks across loci are multiplied to get the total relative risk. Thus, $\lambda_i = \prod_{j=1}^n \lambda_{ij}$, where λ_{ij} denotes

the relative risk for the j^{th} locus. In the interaction model, it is assumed that a particular pair of relative risk for one combination of genotypes is either 2 or 10 times larger than the product of the relative risks; this number is referred to as the interaction factor. For all other SNPs, relative risks are assumed to be independent. Thus, for example, if SNPs x and y interact, then the relative risk for the pair, $K = 2\lambda_{ix}\lambda_{iy}$ for certain configurations of (g_{ix}, g_{iy}), and

$K = \lambda_{i1}\lambda_{i2}$ for other combinations. The total risk in this case would be $\lambda_i = K \prod_{j=3}^n \lambda_{ij}$.

[00230] Based on this model, disease status labels for 100,000 randomly drawn samples is assigned. The probability that is assigned to an individual is a case to be $P(\text{disease} | g_i) = C\lambda_i$, where C is a normalizing factor, and λ_i is the relative risk of individual i , based on the interaction model is assigned. C is chosen such that the fraction of cases is close to the average lifetime risk of the disease. This results in large simulated data of cases and controls under the interaction model.

Example 3: Measuring the Absolute Error in the Risk Estimate

[00231] The ROC curve serves as one metric for evaluating a diagnostic in that it provides a quantitative measure of the ability of the test to distinguish between healthy and sick individuals. However, when estimating lifetime risk, the ROC curve may not be an ideal measure if the correct probabilistic estimate is not used. In particular, for any given pair of score functions, $f_1(G)$ and $f_2(G)$, the ROC curves of the functions will be identical as long as f_1 is a monotonic increasing function of f_2 . For instance, we could simply assign $f_2(G) = \log(f_1(G))$, and in this case by using the scores f_1 and f_2 to estimate risk we will get exactly the same ROC curves. However, these two functions may give very different probabilistic risk estimates to individuals. Thus, the ROC curves may not necessarily be a good measure for tests that report probabilistic risk. For probabilistic risk assessment a more informative test would be the average absolute difference between the true risk probability and the estimated risk probability.

[00232] Since the true probability for developing a disease is unknown, a scenario in which case-control data is used to calculate the GCI parameters (i.e., the relative risks) is simulated, and then applied the GCI risk estimates to another independently simulated population. The disease model for the simulation assumes that the genetic factors of the disease can be decomposed into a small number of large effects and a large number of small effects that are approximated by a Normal distribution (as described above). Since most diseases are diagnosed later in life, the age of onset of the disease to the model is introduced. For each individual that has been determined to develop the disease based on the model, the age of onset of the disease is based on some distribution for the age of onset (Normal distribution with mean = 50 and SD = 13). Thus, in the simulation, some of the controls may in fact be cases that have not been diagnosed at a certain point in time. To create a realistic simulation of an age-matched case-control study, the genetic and environmental factors, as well as the age of onset for individuals, is repeatedly simulated. The age of the individuals from a uniform distribution between 0 and 100 is chosen. This is repeated until 10,000 cases are obtained. For each of these cases, an age-matched control by fixing their age and simulating the genetic and environmental factors of individuals until one of them was found to be a control is generated. This process gives an age-matched case-control dataset with 10,000 cases and 10,000 controls. The odds ratios for each SNP based on this case-control data is estimated and then used to calculate the relative risks for each SNP associated with the disease, using GCI methodology as described herein.

[00233] These simulations are used to test the risk evaluation obtained. 500 individuals are generated according to the true disease model. Since the disease model is known, for each of these individuals the correct risk to develop the condition is calculated. These 'true risk estimates' are used as a baseline for the accuracy measure. The GCI risk estimates are compared to this baseline, as well as a variant of the GCI in which the relative lifetime risk are replaced by the odds ratio.

[00234] In **FIGURE 3**, the distribution of the absolute value of relative errors for a simulated disease with average lifetime risk of 25% and heritability of 64% (Figure 3a) is plotted, and for a disease with average lifetime risk of

42% and heritability of 57% (Figure 3b). These values roughly correspond to the lifetime risk and heritability of Type 2 Diabetes and myocardial infarction. There is a difference between the GCI when using the relative risks and when using the odds ratios. This difference would not be noticed when ROC curves are used to quantify the accuracy of the risk estimates. The errors incurred by the GCI are normally not above 5%. This is under the assumption that all genetic risks are known and that the disease model adequately represents reality.

Example 4: Genetic risk assessment and Family History

[00235] As opposed to using the genotype information to estimate disease risk, it is a common practice in clinical settings to use family history to estimate disease risk. Questions arise about the added value of using genotype information as compared to family history. In order to address these questions, scenario in which parental disease status information is known is simulated, and this information is used as a test for individual's risk for a disease. The false positive and true positive rates of this test are compared to the ones achieved by the genotype test.

[00236] The discrete disease model is used in simulations. Random genotypes for 100,000 mother-father pairs according to the allele frequencies at each SNP location for the diseases are generated. The genotypes are assumed to be independent across the loci. For each trio, a child is generated by randomly choosing one allele from each parent independently for each locus. The genetic Normal component of the child is simply the normalized average of the two parents, and the environmental factor is a combination of the parents' environmental factors and an independent environmental factor. Thus, if the phenotypes of the father and the mother are P_F and P_M respectively, where $P_F = X_F + G_F + E_F$, and $P_M = X_G + G_M + E_M$, (where X is the Binomial genetic distribution, and $G \sim N(0, \sigma_G)$ and $E \sim N(0, \sigma_E)$ are the Normally distributed genetic and environmental factors) then the child's phenotype is assumed to be $P_C = X_C + (G_F + G_M)/\sqrt{2} + a(E_F + E_M) + bE_C$, where $E_C \sim N(0, \sigma_E)$ represents the independent environmental variable of the child, and X_C is the genetic factors attributed to the large effects. The heritability of the conditions forces the constraint $2a^2 + b^2 = 1$. Thus, the parameter b determines the effect of the parents' environment on the children. If $b=1$, the parents' environment does not affect the children, and when $b=0$ the children's environment is entirely determined by the parents. Based on these simulations, the true positive and false positive fraction is calculated for a simple classification test where a child is labeled as a case if either of his or her parents are cases and is considered a control otherwise. This test is the *family history test*.

[00237] This test is compared with the ROC curve that corresponds to the theoretical limit of genotype based test, as described above. As shown in **FIGURE 4**, the sensitivity and specificity of the family history test heavily depends on the choice of the parameter b . A few conclusions can emerge from these graphs. First, it is clear that for all three disease models there are cases in which the family history is inferior to the GCI test, and there are other cases in which it is superior, depending on the value of b . In most cases, however, the two tests give quite similar results. However, the sensitivity and specificity values of the family history test depend on b , which is fixed in the populations, while the GCI test allows for a whole range of specificity and sensitivity values. For instance, in the example of Crohn's Disease, by allowing a few more false positives, one can increase the number of true positive to close to 98% using the GCI test, while the true positive rate for family history test is bounded at 65%.

Example 5: Known environmental factors improve predictions

[00238] To estimate the potential contribution of known environmental factors to disease prediction, both environmental and genotypic data are used to estimate risk. Here the utility of environmental factors across Type 2 Diabetes, Crohn's Disease and Rheumatoid Arthritis, which have very different heritability and average lifetime risk

values, is demonstrated. It is assumed the risks across all SNPs as well as across all of the environmental factors are independent. This assumption does not necessarily hold, but as described further below, this is not going to affect the results substantially. Based on this assumption, the GCI for the case where environmental factors are taken into account is generalized. The resulting method is referred to as EGCI. The genotype and phenotype values for a set of 100,000 individuals based on the genotype and phenotype frequencies in the population are simulated. A disease status to these individuals based on the multiplicative model is assigned.

[00239] The pure genetic based GCI is compared to the new generalized EGCI. The ROC curves for Type 2 Diabetes, Crohn's Disease and Rheumatoid Arthritis can be found in **FIGURE 5**. The added value of environmental factors is not dramatic for Crohn's Disease and Rheumatoid Arthritis, however it is substantial for Type 2 Diabetes. This is driven by the fact that Body Mass Index is crucially affecting the risk for Type 2 Diabetes (with a relative risk of 42.1 when BMI > 35). Note that for a disease such as Crohn's disease it is not expected environmental factors to play a major part, since the heritability of this condition is roughly 80%.

Example 6: Error in the Assumed Lifetime Risk of Disease

[00240] The Human Genome Project, the HapMap project, and related initiatives have resulted in a reference human genome sequence, a catalog of common genetic variation, and a haplotype map of several reference populations. Furthermore, this information combined with cost-effective technologies to test associations between variations throughout the genome and traits and diseases of all sorts, has resulted in dozens of common variants shown to be unequivocally statistically associated with risk of common diseases. These common variants can be used much like population-derived environmental risk factor data in assessing probabilistic pre-symptomatic risk of disease.

[00241] The GCI, like all estimates of a particular quantity, requires a set of assumptions that may bias the risk estimates. Particularly, the assumptions made by the GCI score are that the allele frequencies of the causal SNPs and effect sizes are known, and that the SNP-SNP interactions are known. Furthermore, the assumption is that the average lifetime risk is known. These assumptions might be violated in practice, but as described herein, slight deviations from these assumptions do not change the risk estimates considerably. Particularly, as shown in the previous examples through simulation studies and by the analysis of the WTCCC data, weak SNP-SNP interactions have almost no effect on the GCI, and that deviations in the lifetime risk estimates do not alter the accuracy of the relative risk estimates (see also **FIGURE 6**).

[00242] The ROC curves are based on the assumption that the average lifetime risk of diseases is known and this value is used to calculate the cutoff for assigning disease status in the theoretical model of the disease. However, estimates available from population data may be inaccurate and such errors can greatly influence the GCI-based risk of getting the disease. In the calculations herein, the average lifetime risk is assumed equal to these rough estimates (LTR).

[00243] The error between GCI-based average lifetime risk and the true average lifetime risk of the disease as a function of the assumed risk used in the calculations is plotted as shown in **FIGURE 6A**. The absolute error between GCI-based average lifetime risk and the assumed average lifetime risk as a function of the assumed average lifetime risk is also plotted as shown in **FIGURE 6B**.

[00244] While preferred embodiments of the present disclosure have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the present disclosure. It should be understood that various alternatives to the embodiments of the disclosure described herein may be employed in practicing the embodiments. It is intended that the following claims define the scope of the disclosure and that methods and structures within the scope of these embodiments and their equivalents be covered thereby.

CLAIMS

WHAT IS CLAIMED IS:

1. A method for generating a personalized action plan for an individual based on a score that represents risk of said individual to develop a phenotype, the method comprising:
 - (a) generating a genomic profile from a genetic sample of said individual, wherein said genomic profile is generated using a high density DNA microarray, RT-PCR, or DNA sequencing;
 - (b) analysing said genomic profile for genotype correlations to identify the odds ratio or relative risk for one or more genetic loci;
 - (c) obtaining at least one environmental factor from said individual, wherein said at least one environmental factor has a relative risk or odds ratio for said phenotype;
 - (d) generating a score from said odds ratio or relative risk in step (b) and said odds ratio or relative risk in step (c) using a computer; and
 - (e) generating a personalized action plan for said individual based on the score, wherein the personalized action plan includes treatments and preventative measures for the phenotype, wherein the treatments and preventative measures are selected based on the score.
2. The method of claim 1, wherein said relative risk or odds ratio is at least 1, 1.1, 1.2, 1.3, 1.4, 1.5, 2, 3, 4, 5, 10, 12, 15, 20, 25, 30, 35, 40, 45, or 50.
3. The method of claim 1, wherein said at least one environmental factor is selected from the group consisting of: said individual's birthplace, location of residency, lifestyle habit; diet, exercise habits, physical measurement and personal relationships.
4. The method of claim 3, wherein said lifestyle habit is smoking or alcohol intake.
5. The method of claim 3, wherein said physical measurement of said individual is selected from the group consisting of: body mass index, blood pressure, heart rate, glucose level, metabolite level, ion level, weight, height, cholesterol level, vitamin level, blood cell count, protein level, and transcript level.
6. The method of claim 1, wherein generating said score uses at least 2 environmental factors.

7. The method of claim 1, wherein generating said score assumes said at least one environmental factor is an independent risk factor for said phenotype.
8. The method of claim 1 wherein said phenotype has a heritability of less than 95%.
9. The method of claim 1 or 8, wherein said phenotype has a heritability of less than 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, or 90%.
10. The method of claim 1 or 8, wherein a third party obtains said genetic sample or generates said genomic profile.
11. The method of claim 1 or 8, wherein said genetic sample is DNA or RNA.
12. The method of claim 1 or 8, wherein said genetic sample is from a biological sample selected from the group consisting of: blood, hair, skin, saliva, semen, urine, fecal material, sweat, and buccal sample.
13. The method of claim 1 or claim 8, wherein said genomic profile comprises a single nucleotide polymorphism, truncation, insertion, deletion, or repeat.
14. The method of claim 1 or 8, further comprising updating said score with additional or modified relative risks or odds ratio for environmental factors.