METHODS AND SYSTEMS FOR UNIVERSAL CARRIER SCREENING

CROSS-REFERENCE

[0001] This application claims the benefit of U.S. Provisional Application No. 61/053,926, filed May 16, 2008, which application is incorporated herein by reference.

BACKGROUND OF THE INVENTION

[0001] Genetic testing of prospective parents can be used to predict the chances that offspring of a couple will have particular genetic diseases. Persons for whom such testing is attractive include those for whom genetic diseases run in the family or those from ethnic groups that have high incidence of genetic diseases. The results of such testing can provide a couple with information they can use to make decisions about becoming parents. For example, a couple may decide to make preparations for special care that might be needed to raise a child with special needs, or the couple may explore options for assisted reproduction.

[0002] Many genetic diseases are rare in a population, for example, with frequencies of less than 1 in 1000. Yet, if a genetic disease is caused by a single recessive allele of a gene and both prospective parents are carriers of the recessive allele, then the probability that a child of the couple will have the disease is likely 25%.

[0003] The incidence of alleles for a certain genetic conditions is different across different populations. For example, in persons with Ashkenazi Jewish ancestry, the carrier frequency for Tay-Sachs disease is 1:30; for Canavan disease it is 1:40; for Niemann-Pick disease type A it is 1:90; for Fanconi anemia type C it is 1:89; for Bloom syndrome it is 1:100; for Gaucher disease type 1 it is 1:12. In persons of Southeast Asian ancestry, the carrier frequency of alpha-thalassemia is as high as 1:20. In Caucasians, the carrier frequency of cystic fibrosis is 1:25.

[0004] While such diseases may be individually rare, there are a sufficient number of them so the probability that any individual is a carrier for at least one of is significantly greater. For example, it has been estimated that 70% of the Ashkenazi Jewish population has at least one disease causing allele. Estimates of genetic load indicate that every human carries approximately 8 to 30 deleterious recessive alleles.

[0005] Carrier screening can decrease the incidence of genetic diseases in a population. For example, as a result of carrier screening, the incidence of Tay Sachs disease among the Ashkenazi Jewish population has decreased in recent years. Another example of a genetic Mendelian disease is cystic fibrosis (CF). Cystic fibrosis is often fatal, debilitating, incurable, and costly to treat. CF can be characterized by autosomal recessive inheritance and carried by asymptomatic individuals caused by hundreds of different mutations, each of which varies in frequency across ethnic groups. The American College of Medical Genetics (ACMG) recommends carrier screening by genetic

testing for all prospective parents for a number of Mendelian diseases. Since 1998, panethnic population-wide screening for cystic fibrosis carrier status has been recommended by ACMG. The list of diseases that ACMG recommends testing is continually expanding as a function of new discoveries related to genetic diseases. Carrier screening has been adopted because the health benefits outweigh the financial costs of testing by a definitive margin.

[0006] Presently, screening of prospective parents for carrier status is selective. It is indicated for couples belonging to populations at increased risk for particular conditions. There is a need in the art for universal testing for a wide variety of genetic conditions for individuals of any ancestry.

SUMMARY OF THE INVENTION

[0007] In an aspect, a method is disclosed herein that comprises: testing a subset of biological relatives of a child or potential child to determine a presence of a plurality of causal genetic variants corresponding to at least one rare genetic disease and a presence of at least one ancestry informative marker (AIM); and predicting a probability of a phenotype of the child or potential child from the subset of biological relatives with respect to the at least one rare genetic disease based at least in part on the presence of the plurality of causal genetic variants and the presence of the at least one AIM. In some instances, predicting comprises performing a fully probabilistic analysis on data collected on said casual genetic variants. In some instances, predicting is further based on phenotypic information about at least one member of the subset of biological relatives. The method can further comprise providing genetic counseling services to at least one member of the subset of biological relatives. The method can further comprise delivering the probability of the phenotype of the child to a physician referral service.

[0008] In an aspect, a computer readable medium comprises: logic configured to predict a probability of a phenotype of a child or potential child from a subset of biological relatives with respect to at least one rare genetic disease based at least in part on the results of a test of the subset of biological relatives for the presence of a plurality of causal genetic variants and at least one ancestry informative marker (AIM). In some instances, the logic performs a fully probabilistic analysis. In some instances, the computer readable medium provides an output in the form of a report detailing the presence of the at least on AIM and the presence of any of the plurality of causal genetic variants in any member of the subset of biological relatives.

[0009] In an aspect, a computer readable medium comprises: logic configured to perform a fully probabilistic analysis on data corresponding to a plurality of causal genetic variants from a male and a female to predict a probability of a phenotype of a child, wherein the male and the female are potential parents of the child. In some instances, the causal genetic variants from each of the male and female comprise one or more ancestry informative markers (AIMs), one or more causal genetic variants corresponding to a rare genetic disease, one or more causal genetic variants corresponding to a personality trait, or both. In some instances, fully probabilistic analysis incorporates a plurality

of sources of statistical uncertainty in the probability. A computer readable medium herein can further comprise logic for receiving input from a phenotype battery and assigning a weighting function to the plurality of causal genetic variants based on said input. In some instances, the input from the phenotype battery comprises: height, weight, and family disease history. In some instances, the computer readable medium provides an output in the form of a report detailing a probability distribution over child risks or phenotypes.

[0010] In an aspect, a system is described herein for predicting a child phenotype that comprises: a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM), wherein the device is in contact with a sample from a biological relative of a child or potential child; a reader configured to read data from the devices; and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants and data from the at least one AIM to predict a probability of a phenotype of the child with respect to the at least one rare genetic disease. In some instances, the biological relative is a prospective mother or prospective father of the child or potential child. In some instances, the nucleic acid detection device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants and the at least one AIM.

[0011] In an aspect, a system for predicting a child phenotype comprises: a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases, wherein the device is in contact with a sample from a biological relative of the child or potential child; a reader configured to read data from the devices; and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants to predict a probability of a phenotype of the child or potential child with respect to the more than 85 rare genetic diseases. In some instances, the biological relative is a prospective mother or prospective father of the child or potential child. In some instances, the nucleic acid detection device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. In some instances, the nucleic acid detection device further comprises a plurality of nucleic acid probes that selectively bind to at least one ancestry informative marker (AIM). In some instances, the computer readable instructions when executed utilize the data from the reader corresponding to the at least one AIM to predict the probability of a phenotype of the child.

[0012] In an aspect, a system for indicating if a subject is a carrier of a rare genetic disease comprises: a reader configured to read data from a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM); and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal

genetic variants and the at least one ancestry informative marker to predict a plurality of probabilities of the subject being a carrier for each of the plurality of causal genetic variants.

[0013] In an aspect, a method comprises: receiving a sample from a user; testing the sample with a nucleic acid detection device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM); calculating a plurality of probabilities for the possible user genotypes at each location of the plurality of causal genetic variants based on results from the testing step relating to the plurality of causal genetic variants and the at least one AIM; and delivering to the user the plurality of probabilities corresponding to the user being a carrier. The method can further comprise: receiving a sample from a second user; testing the sample from the second user with a device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM); calculating a probability of a child phenotype corresponding to the rare genetic diseases based on results from testing the user and the second user; and delivering the probability of the child phenotype to at least one of the user and the second user. The method can further comprise providing genetic counseling service to at least one of the user and the second user. The method can be carried out as part of a child phenotype prediction service. The method can further comprise obtaining phenotypic information from the user; and using the phenotypic information from the user in the calculating steps. The method can further comprise obtaining family history from the user; and using the family history from the user in the calculating steps.

[0014] In an aspect, a nucleic acid detection device is configured to test a sample for a plurality of causal genetic variants corresponding to at least one rare genetic disease and one or more ancestry informative markers (AIMs). In some instances, the device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants and the one or more AIMs. In some instances, the device comprises a bead array that selectively binds to the plurality of causal genetic variants and the one or more AIMs.

[0015] In another aspect, a nucleic acid detection device is configured to test a sample for a plurality of causal genetic variants corresponding more than 85 rare genetic diseases. The device can be further configured to test a sample for at least one ancestry informative marker (AIM). In some instances, the device comprises a plurality of nucleic acid probes that selectively bind to a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. The device can further comprise a plurality of nucleic acid probes that selectively bind to at least one ancestry informative marker (AIM). The device can comprise a bead array that selectively binds to a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. The device can further comprise a bead array that selectively binds to at least one ancestry informative marker (AIM). In some instances, the device comprises a resequencing assay to detect the plurality of causal genetic variants corresponding to the more than 85 rare genetic diseases. In some instances, the

device further comprises a resequencing assay to detect at least one ancestry informative marker (AIM).

[0016] In some instances, the at least one AIM is not a causal genetic variant. In some instances, at least two of the rare genetic diseases occur at frequencies that differ by at least 10-fold in at least two distinct populations, wherein the at least two distinct populations are differentiated by the at least one AIM.

[0017] In an aspect, a method comprises: marketing a genetic testing service comprising predicting a probability of a child phenotype from a subset of biological relatives of the child or potential child, wherein the prediction is based at least in part on the presence of a plurality of causal genetic variants in each of the subset of biological relatives and based at least in part on the inferred ancestries of each of the subset of biological relatives; and delivering a probability of the child phenotype for a fee. The marketing can be conducted in connection with a dating or marriage service. The method can further comprise referring at least one member of the subset of biological relatives to a physician. In some instances, the inferred ancestries are inferred by a test for at least one ancestry informative marker (AIM).

[0018] In an aspect, a set of nucleic acid pools is disclosed for validating a nucleic acid sequence detection device comprising a set of causal genetic variant probes, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a different subset of the set of causal genetic variant probes. In some instances, a first pool of the set comprises a first nucleic acid segment that interferes during detection with a second nucleic acid segment of a second pool of the set, and wherein the first pool does not comprise the second nucleic acid segment and the second pool does not comprise the first nucleic acid segment. In some instances, the nucleic acid segments of each pool are single stranded nucleic acid molecules. In some instances, the nucleic acid segments comprise one or more plasmids.

[0019] In an aspect, a method of validating a lot of manufactured nucleic acid sequence detection devices comprises: contacting each of a plurality of nucleic acid sequence detection devices from the lot with a different nucleic acid pool, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a plurality of causal genetic variant probes on said detection devices, and wherein each nucleic acid pool binds a different set of the plurality of causal genetic variant probes; and detecting presence or absence of the plurality of causal genetic variant probes on the plurality of nucleic acid detection devices, wherein the lot of manufactured devices is validated if all of the plurality of causal genetic variant probes are present on the plurality of nucleic acid detection devices.

[0020] . In some instances, the method further comprises delivering the lot of manufactured devices when the devices are validated. In some instances, the lot of manufactured devices is rejected if not all of the plurality of causal genetic variant probes are present. In some instances, the lot of

manufactured devices is modified and the method is repeated if not all of the plurality of causal genetic variant probes are present.

INCORPORATION BY REFERENCE

[0021] All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] Many features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages herein will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which many principles are utilized, and the accompanying drawings of which:

[0023] Figure 1 provides a table of causal genetic variants.

[0024] Figure 2 provides a list of exemplary populations.

[0025] Figure 3 provides a number of AIMs that distinguish different populations. The entries refer to items in the dbSNP database, a database of genetic variants maintained by the US government: http://www.ncbi.nlm.nih.gov/projects/SNP/. Curated records in dbSNP contain information that describes the sequence and location of genetic variants, and where available the frequency of alleles of those variants in different populations. rs numbers (for example, rs332, rs25, etc.) are the ID numbers used to index the portion of the dbSNP database.

[0026] Figure 4 illustrates a population 1 as a group which has a hypothetical A/G SNP with the given flanking sequence.

[0027] Figure 5 illustrates population 2 has exactly the same variant with the same flanking sequence as Figure 4, but different proportions of each genotype (0.16, 0.48, 0.36 rather than 0.25, 0.50, 0.25).

[0028] Figure 6 illustrates population 3 has a different flanking sequence than Figures 4 and 5 with an extra G.

[0029] Figure 7 considers populations by gender, wherein Figures 4-6 consider populations in terms of geographic ancestry.

[0030] Figure 8 shows an intensity scatterplot for the beta thalassemia deletion allele similar to the scatterplots in Figures 4-7.

[0031] Figure 9 shows that calling of points near the boundary is aided by the use of ancestry information.

[0032] Figure 10 shows that a raw intensity measurement at a causal locus is not the only value that matters.

[0033] Figure 11 illustrates a source code example that is useful in calling.

[0034] Figure 12 shows a straightforward modification to the expression for the posterior probability of whether an individual is in a given cluster.

[0035] Figure 13 shows a measured genotype may have measurement error as represented by a false positive and false negative rate.

[0036] Figure 14 shows the inferred genotype is unphased.

[0037] Figure 15 shows how to generate a distribution over possible recombinants and hence possible gametes from the possible haplotypes.

[0038] Figure 16 shows how to repeat this process for both mother and father to obtain a probability distribution over gametic unions, corresponding to phased child genotypes (aka haplotypes).

[0039] Figure 17 shows several different kinds of genotype-to-phenotype maps that can be used with the distribution over child haplotypes to produce a distribution over possible child phenotypes.

[0040] Figure 18 depicts equations for obtaining the distribution over estimated child phenotypes given parental genotypes.

[0041] Figure 19 demonstrates examples of nucleic acid segments and pools of nucleic acids.

[0042] Figures 20-23 demonstrate exemplary processes of delivering a probability that a user is a carrier of rare genetic disease.

[0043] Figure 24 illustrates exemplary the input and output steps for report generation for two hypothetical parents: Mama Hen and Papa Hen.

DETAILED DESCRIPTION OF THE INVENTION

[0044] Described herein are methods, computer readable instructions, software, systems, and devices that are utilized for detecting genotypes of individuals. Herein, the genotypes relate to detecting specific causal genetic variants that cause or can cause rare genetic diseases. In many instances, the rare genetic diseases are Mendelian diseases, wherein an individual is a carrier of a trait (recessive or dominant) corresponding to primarily related to single gene. The methods, computer readable instructions, software, systems, and devices can be used for family genomics, such as generating probability distributions for many family members in a family including without limitation: grandparents, parents, children, aunts, and uncles. The family members may be alive, dead, embryonic, or not yet conceived. The genotype and/or phenotype of any family member may be utilized to gain information about other family members. The methods, computer readable instructions, software, systems, and devices can be utilized as a screen for predicting the probability of the phenotype of a child or potential child from two individuals. In many instances, the child phenotype is predicted before conception based upon the genotypes (and sometimes phenotypes) of the potential father and potential mother. Child phenotype prediction can be useful in a variety of circumstances as described herein and provides an opportunity for an individual to make a decision or take action based upon his personal genotype.

I. METHODS OF TESTING AND PREDICTION

[0045] In an aspect, a method comprises: testing each member of a set of parents or potential parents to determine presence of a plurality of causal genetic variants corresponding to at least one rare genetic disease and the presence of at least one ancestry informative marker (AIM); and predicting a probability of a phenotype of a child from the set of parents or potential parents with respect to the at least one rare genetic disease based at least in part on results of the presence of the plurality of causal genetic variants and the at least one AIM. In some instances the predicting step is further based on phenotypic information about at least one member of the set of parents or potential parents, or genotypic and/or phenotypic information on other biological relatives of the child. In some instances the method further comprises providing genetic counseling services. In some instances, a probability about a person's genotype, phenotype, or potential phenotype can be delivered to a physician referral service. In some instances the method further comprises delivering the probability of the phenotype of the child phenotype to a physician referral service.

[0046] The prediction of a probability of a phenotype herein can be a probability distribution over a variety of risks for traits. The probability distribution can be categorical or continuous for any probability prediction herein. As an example of a probability distribution over child risks, the prediction could be P(D=1) = 0.80 and P(D=0) = 0.20, representing an 80% probability that a child will be affected (D=1) by a Mendelian disease. As an example of a probability distribution over child phenotypes, the prediction could be a normal distribution with specified mean and variance over the possible height of the child as an adult.

[0047] A genetic disease can be any disease that is influenced by a known causal genetic variant. In some instances, a rare genetic disease is a disease that is present at a rate 1 per 100 in the human population. In other instances, a rare genetic disease is a disease that is present at a rate 1 per 1000 in the human population. In yet other instances, a rare genetic disease is a disease that is present at a rate 1 per 10,000, 1 per 100,000, 1 per 1,000,000, or 1 per 10,000,000 in the human population. In some instances, the rare genetic disease may be present at a rate much higher than 1 per 100 in a specific ethnic or ancestral human population, for example, 1 per 50, 1 per 20, 1 per 10, or 1 per 5 in a specific population.

A. Rare Genetic Disease

[0048] Methods, systems, software, and devices herein can test or be configured to test for phenotypic traits (also referred to simply as traits), i.e., a distinct form of a characteristic of an organism. For example, eye color is a physical characteristic; brown, green and blue are phenotypic traits. Some physical characteristics relating to health have, as associated traits, abnormal and normal, such as Huntington's disease (abnormal) and non-Huntington's disease (normal). Traits may be morphological, developmental, biochemical, physiological, or behavioral. The collection of a

plurality of phenotypic traits exhibited by an individual is usually referred to as the individual's phenotype.

[0049] Mendelian traits are traits that are inherited by way of a single gene or influenced primarily by a single gene. Mendelian diseases are diseases inherited as a Mendelian trait. Typical Mendelian traits can be autosomal dominant, autosomal recessive or sex-linked (X-linked or Y-linked). Mendelian traits can also be atypical, including traits with atypical inheritance patterns (for example, cytoplasmic inheritance or incomplete penetrance) and including traits that arise as a result of spontaneous but predictable mutations (for example, repeat polymorphism expansions or mutational hot spots). Mendelian traits include both typical Mendelian traits and atypical Mendelian traits. Typically, the percentage of variance in a Mendelian trait that is explained by genotype is very high, for example, at least 99%. However, the percentage of variance in phenotype in an atypical Mendelian trait that is explained by genotype can be lower.

[0050] In certain instances, two different genes cause traits that appear similar or identical. In this case it is typical to define sub-types of the trait such that a single gene is associated with a trait single sub-type. For example, the disease Mucopolysaccharidosis is typically further described as being of a particular type, such as Mucopolysaccharidosis Type I or Mucopolysaccharidosis Type VII.

[0051] Non-Mendelian traits, also called complex traits, are inherited by way of more than one gene. Typically, for non-Mendelian traits the percentage of variance in phenotype explained by explicit genotypic markers is low, often less than 50%. A non-Mendelian trait differs from an atypical Mendelian trait in that differences between individuals in a complex trait are due to differences in more than one gene whereas differences between individuals for an atypical Mendelian trait are due primarily to differences in one gene. Not all non-Mendelian traits have genotypically-explained variance of less than 99% (Friedman, Naomi P.; Miyake, Akira; Young, Susan E.; DeFries, John C.; Corley, Robin P.; Hewitt, John K. Individual differences in executive functions are almost entirely genetic in origin. Journal of Experimental Psychology: General. 2008 May Vol 137(2) 201-225). Examples of non-Mendelian traits are height, weight, and skin color.

[0052] An individual's genotype for a Mendelian trait is the identity of the genes (or gene in the case of sex-linked traits) responsible for the trait that the individual carries, for example, homozygous for the gene responsible for the trait, homozygous for a gene responsible for a different trait, or heterozygous.

[0053] Rare genetic diseases as described herein are a type of trait. A rare genetic disease for the purposes herein is a disease that is a trait that can be inherited genetically by a child from a set of parents. The disease or trait can then manifest itself in the phenotype of the child.

[0054] Rare genetic diseases that can be tested according to this invention include, but are not limited to:21-Hydroxylase Deficiency, ABCC8-Related Hyperinsulinism, ARSACS,

Achondroplasia, Achromatopsia, Adenosine Monophosphate Deaminase 1, Agenesis of Corpus Callosum with Neuronopathy, Alkaptonuria, Alpha-1-Antitrypsin Deficiency, Alpha-Mannosidosis, Alpha-Sarcoglycanopathy, Alpha-Thalassemia, Alzheimers, Angiotensin II Receptor, Type 1, Apolipoprotein E Genotyping, Argininosuccinicaciduria, Aspartylglycosaminuria, Ataxia with Vitamin E Deficiency, Ataxia-Telangiectasia, Autoimmune Polyendocrinopathy Syndrome Type 1, BRCA1 Hereditary Breast/Ovarian Cancer, BRCA2 Hereditary Breast/Ovarian Cancer, Bardet-Biedl Syndrome, Best Vitelliform Macular Dystrophy, Beta-Sarcoglycanopathy, Beta-Thalassemia, Biotinidase Deficiency, Blau Syndrome, Bloom Syndrome, CFTR-Related Disorders, CLN3-Related Neuronal Ceroid-Lipofuscinosis, CLN5-Related Neuronal Ceroid-Lipofuscinosis, CLN8-Related Neuronal Ceroid-Lipofuscinosis, Canavan Disease, Carnitine Palmitoyltransferase IA Deficiency, Carnitine Palmitoyltransferase II Deficiency, Cartilage-Hair Hypoplasia, Cerebral Cavernous Malformation, Choroideremia, Cohen Syndrome, Congenital Cataracts, Facial Dysmorphism, and Neuropathy, Congenital Disorder of GlycosylationIa, Congenital Disorder of Glycosylation Ib, Congenital Finnish Nephrosis, Crohn Disease, Cystinosis, DFNA 9 (COCH), Diabetes and Hearing Loss, Early-Onset Primary Dystonia (DYT1), Epidermolysis Bullosa Junctional, Herlitz-Pearson Type, FANCC-Related Fanconi Anemia, FGFR1-Related Craniosynostosis, FGFR2-Related Craniosynostosis, FGFR3-Related Craniosynostosis, Factor V Leiden Thrombophilia, Factor V R2 Mutation Thrombophilia, Factor XI Deficiency, Factor XIII Deficiency, Familial Adenomatous Polyposis, Familial Dysautonomia, Familial Hypercholesterolemia Type B, Familial Mediterranean Fever, Free Sialic Acid Storage Disorders, Frontotemporal Dementia with Parkinsonism-17, Fumarase deficiency, GJB2-Related DFNA 3 Nonsyndromic Hearing Loss and Deafness, GJB2-Related DFNB 1 Nonsyndromic Hearing Loss and Deafness, GNE-Related Myopathies, Galactosemia, Gaucher Disease, Glucose-6-Phosphate Dehydrogenase Deficiency, Glutaricacidemia Type 1, Glycogen Storage Disease Type 1a, Glycogen Storage Disease Type 1b, Glycogen Storage Disease Type II, Glycogen Storage Disease Type III, Glycogen Storage Disease Type V, Gracile Syndrome, HFE- Associated Hereditary Hemochromatosis, Halder AIMs, Hemoglobin S Beta-Thalassemia, Hereditary Fructose Intolerance, Hereditary Pancreatitis, Hereditary Thymine-Uraciluria, Hexosaminidase A Deficiency, Hidrotic Ectodermal Dysplasia 2, Homocystinuria Caused by Cystathionine Beta-Synthase Deficiency, Hyperkalemic Periodic Paralysis Type 1, Hyperornithinemia-Hyperammonemia-Homocitrullinuria Syndrome, Hyperoxaluria, Primary, Type 1, Hyperoxaluria, Primary, Type 2, Hypochondroplasia, Hypokalemic Periodic Paralysis Type 1, Hypokalemic Periodic Paralysis Type 2, Hypophosphatasia, Infantile Myopathy and Lactic Acidosis (Fatal and Non-Fatal Forms), Isovaleric Acidemias, Krabbe Disease, LGMD2I, Leber Hereditary Optic Neuropathy, Leigh Syndrome, French-Canadian Type, Long Chain 3-Hydroxyacyl-CoA Dehydrogenase Deficiency, MELAS, MERRF, MTHFR Deficiency, MTHFR Thermolabile Variant, MTRNR1-Related Hearing Loss and

Deafness, MTTS1-Related Hearing Loss and Deafness, MYH-Associated Polyposis, Maple Syrup Urine Disease Type 1A, Maple Syrup Urine Disease Type 1B, McCune-Albright Syndrome, Medium Chain Acyl-Coenzyme A Dehydrogenase Deficiency, Megalencephalic Leukoencephalopathy with Subcortical Cysts, Metachromatic Leukodystrophy, Mitochondrial Cardiomyopathy, Mitochondrial DNA-Associated Leigh Syndrome and NARP, Mucolipidosis IV, Mucopolysaccharidosis Type I, Mucopolysaccharidosis Type IIIA, Mucopolysaccharidosis Type VII, Multiple Endocrine Neoplasia Type 2, Muscle-Eye-Brain Disease, Nemaline Myopathy, Neurological phenotype, Niemann-Pick Disease Due to Sphingomyelinase Deficiency, Niemann-Pick Disease Type C1, Nijmegen Breakage Syndrome, PPT1-Related Neuronal Ceroid-Lipofuscinosis, PROP1-related pituitary hormome deficiency, Pallister-Hall Syndrome, Paramyotonia Congenita, Pendred Syndrome, Peroxisomal Bifunctional Enzyme Deficiency, Pervasive Developmental Disorders, Phenylalanine Hydroxylase Deficiency, Plasminogen Activator Inhibitor I, Polycystic Kidney Disease, Autosomal Recessive, Prothrombin G20210A Thrombophilia, Pseudovitamin D Deficiency Rickets, Pycnodysostosis, Retinitis Pigmentosa, Autosomal Recessive, Bothnia Type, Rett Syndrome, Rhizomelic Chondrodysplasia Punctata Type 1, Short Chain Acyl-CoA Dehydrogenase Deficiency, Shwachman-Diamond Syndrome, Sjogren-Larsson Syndrome, Smith-Lemli-Opitz Syndrome, Spastic Paraplegia 13, Sulfate Transporter-Related Osteochondrodysplasia, TFR2-Related Hereditary Hemochromatosis, TPP1-Related Neuronal Ceroid-Lipofuscinosis, Thanatophoric Dysplasia, Transthyretin Amyloidosis, Trifunctional Protein Deficiency, Tyrosine Hydroxylase-Deficient DRD, Tyrosinemia Type I, Wilson Disease, X-Linked Juvenile Retinoschisis and Zellweger Syndrome Spectrum.

B. Causal Genetic Variants

[0055] Causal genetic variants are genetic variants for which there is statistical, biological, and/or functional evidence of association with a disease or trait such that the variant is likely to play a role in etiology. A single causal genetic variant can be associated with more than one trait. A causal genetic variant can be associated with a Mendelian trait or a non-Mendelian trait or both.

[0056] Although it is typical to refer to genetic variants that occur at a frequency of greater than or equal to 1% as polymorphisms and refer to a less common variants as mutations, these terms are used interchangeably and without intending to refer to a particular frequency, unless specified. For example, a SNP is a single nucleotide polymorphism, and it can occur at any frequency. An example of a causal genetic variant that is a SNP is the Hb S variant of hemoglobin that causes sickle cell anemia. A DIP is a deletion/insertion polymorphism, and it can occur at any frequency. DIPs can also be referred to as indels. An example of a causal genetic variant that is a DIP is the delta508 mutation of the CFTR gene which causes cystic fibrosis. A CNV is a copy number variant, and it can occur at any frequency. An example of a causal genetic variant that is a CNV is trisomy 21, which causes Downs syndrome. A STR is a short tandem repeat variation, and it can occur at any

frequency. STRs are also called repeat polymorphisms. An example of a causal genetic variant that is an STR is tandem repeat that causes Huntington's disease.

[0057] Causal genetic variants can manifest as variations in a DNA polynucleotide, such as results from a SNP, DIP, CNV, STR, or heritable epigenetic modification (for example, DNA methylation). Multiple instantiations of a causal genetic variant can be recognized as such because they are identical-by-state and/or identical-by-descent. A causal genetic variant may also be a set of closely related causal genetic variants. Some causal genetic variants may exert influence as sequence variations in RNA polynucleotides. At this level, some causal genetic variants are also indicated by the presence or absence of a species of RNA polynucleotides. Also, some causal genetic variants result in sequence variations in protein polypeptides. At this level, some causal genetic variants are also indicated by the presence or absence of a species of protein polypeptides.

[0058] The nomenclature for cataloging and describing causal genetic variants was developed on an ad hoc basis over many years. Causal genetic variants can be described by more than one name. For example, causal genetic variants can be labeled with both a common name and a systematic name. Systematic names are based on an evolving notation developed by the Human Genome Variation Society.

[0059] Causal genetic variants are distinguished from indirectly associated genetic variants, which are associated with a trait or disease as a result of correlated coinheritance with a causal genetic variant.

[0060] Causal genetic variants can be the cause of Mendelian and non-Mendelian traits. In many instances, causal genetic variants that determine deleterious Mendelian traits and Mendelian diseases can be rare and occur at a frequency less than 1% of a population. However, some causal genetic variants that determine deleterious Mendelian traits and Mendelian diseases occur at frequencies in the range of 1% to 10%. In some cases, causal genetic variants that determine deleterious Mendelian traits and Mendelian diseases occur at frequencies greater than 10%.

[0061] Causal genetic variants can be originally discovered by statistical and molecular genetic analyses of the genotypes and phenotypes of individuals, families, and populations. The causal genetic variants for Mendelian traits are typically identified in a two-stage process. In the first stage, families in which multiple individuals who possess the trait are examined for genotype and phenotype. Genotype and phenotype data from these families is used to establish the statistical association between the presence of the Mendelian trait and the presence of a number of genetic markers, which typically are indirectly associated genetic variants that are physically proximal to a causal genetic variant on the chromosome. This association establishes a candidate region in which the causal genetic variant is likely to map. In a second stage, the causal genetic variant itself is identified. The second step typically entails sequencing the candidate region. More sophisticated, one-stage processes are possible with more advanced technologies which permit the direct

identification of a causal genetic variant or the identification of smaller candidate regions. After one causal genetic variant for a trait is discovered, additional variants for the same trait can be discovered by simple methods. For example, the gene associated with the trait can be sequenced in individuals who possess the trait or their relatives. The invention of new methods for discovering causal genetic variants is an active area of research. The application of existing methods and the incorporation of new methods is expected to continue to result in the discovery of additional causal genetic variants which can be used or tested for by the devices, systems, and methods herein. Many causal genetic variants are cataloged in databases including the Online Mendelian Inheritance in Man (OMIM) and the Human Gene Mutation Database (HGMD). Causal genetic variants are also reported in the scholarly literature, at conferences, and in personal communications between scholars.

[0062] In diploid organisms, including humans and other mammals, a causal genetic variant can be present in zero, one, or two copies. In some cases, the causal genetic variant can be present in more than two copies. Methods for detecting the presence of a causal genetic variant typically also determine the number of copies of the variant that are present.

[0063] In an embodiment at least one of the causal genetic variants causes a trait having an incidence of no more than 1% in a first of the populations and at least one other of the causal genetic variants causes a trait having an incidence of no more than 1% in a second of the populations. In another embodiment at least one of the causal genetic variants causes a trait having an incidence of no more than 1/10,000 in a first of the populations and at least one other of the causal genetic variants causes a trait having an incidence of no more than 1/10,000 in a second of the populations.

[0064] Figure 1 provides a table of exemplary causal genetic variants for rare genetic diseases.

C. Populations

[0065] In some instances, devices, methods, and systems herein are configured to test for causal genetic variants for traits, wherein the traits differ in frequency between populations. Populations are groups of individuals of the same species, such as humans or other mammals. Human populations are often given particular names which can form the basis for social identities.

[0066] A human population can be a group of people sharing a common genetic inheritance, such as an ethnic group (for example, Caucasian). A human population can be a haplotype population or group of haplotype populations (for example, haplotype H1, M52). A human population can be a national group (for example, Americans, English, Irish). A human population can be a demographic population such as those delineated by age, sex, and socioeconomic factors. Human populations can be historical populations.

[0067] A population can consist of individuals distributed over a large geographic area such that individuals at extremes of the distribution may never meet one another. The individuals of a population can be geographically dispersed into discontinuous areas. Populations can be informative

about biogeographical ancestry. The labels used to identify populations do not need to contain an explicit geographic or ancestry reference for the populations to be informative about ancestry, for example, a personal identification of race.

[0068] Populations can also be defined by ancestry. Genetic studies can define populations. As defined by ancestry and genetics, the major human populations correspond to continental scale groupings, which include Western Eurasian, sub-Saharan African, East Asian, and Native American. Most humans can be assigned to at least one of these populations on the basis of ancestry. A number of smaller populations are also distinguished as continental groups, including Indigenous Australian, Oceanian, and Bushmen.

[0069] Very often, populations can be further decomposed into sub-populations. The relationship between populations and subpopulations can be hierarchical. For example, the Oceanian population can be further sub-divided into sub-populations including Polynesians, Melanesians and Micronesians. The Western Eurasian population can be further sub-divided into sub-populations including European, Western/Central Asian, South Asian, and North African. The European population can be further sub-divided into sub-populations including North-Western European, Southern European, and Ashkenazi Jewish populations. The North-Western European population can be further sub-divided into national populations including English, Irish, German, Finnish, and the like. The East Asian population can be further sub-divided into Chinese, Japanese, and Korean subpopulations. The South Asian population can be further sub-divided into Indian and Pakistani populations. The Indian population can be further sub-divided into Dravidian people, Brahui people, Kannadigas, Malayalis, Tamils, Telugus, Tuluvas, and Gonds.

[0070] An individual can self-identify to one or more reference populations (for example, to an ethnic group) on the basis of their ancestry. A person may identify their ancestry based on the identity of their four biological grandparents. The ancestry of some individuals (sometimes called their individual ancestry or individual biogeographical ancestry) can be described as coming from more than one distinct population. Admixture is defined relative to a set of reference populations. The ancestry of an individual can be classified in terms of admixture of a set of reference populations. These reference populations can vary depending on the application. It is common to refer to admixture in term of the four major continental groups because these groups can be readily distinguishable genetically and phenotypically.

[0071] Figure 2 provides a list of exemplary populations.

D. Ancestry Informative Markers

[0072] Ancestry informative markers (AIMs) are genetic variants that differ in frequency between populations. Almost all common genetic variants differ both within and between populations. However, individual variants are found across a spectrum of variance distributions. In an instance,

variants that differ within but not between populations are typically not AIMs. In another instance, variants that differ between but not within populations are especially informative.

[0073] A variety of kinds of genetic variants can be AIMs, including SNPs, DIPs, CNVs, and STRs. AIMs can also be sequence variations in RNA polynucleotides. Some AIMs can also be indicated by the presence or the concentration of a species of RNA polynucleotides. Some AIMs can also be sequence variations in protein polypeptides. Some AIMs can also be indicated by the presence or absence of a species of protein polypeptides.

[0074] A number of ancestry informative markers are identified in Figure 3. Others also are described in US 2007/0037182 (Gaskin et al.).

[0075] Ancestry informative markers can be discovered by determining the frequency of genetic variants in a plurality of populations. This may be achieved by determining the frequency of already known variants in individuals from various populations. It may also be achieved intrinsically during the process of variant discovery. Both tasks were undertaken by the International HapMap project, which catalogued SNP polymorphisms.

[0076] Ancestry informative markers can be ranked by a variety of measurements which judge their predictive power. One measurement is Wright's F-statistic, called Fst or FST. This variable is known by other names, including Fixation index. Another metric for ranking AIMs is informativeness. Another method of ranking AIMs is the PCA-correlated SNPs method of Paschou et al. (Paschou et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet (2007) vol. 3 (9) pp. 1672-86).

[0077] To achieve a pre-selected degree of confidence in ancestry inference on the basis of ancestry informative markers, and to achieve ancestry inference for a plurality of populations, it is typically necessary to examine more than one ancestry informative marker. A sufficiently large panel of randomly selected genetic variants can be used to infer ancestry. A targeted set of especially appropriate AIMs can be constructed. Many researchers have published lists of suggested ancestry informative markers (for example: Seldin et al. Application of ancestry informative markers to association studies in European Americans. PLoS Genet (2008) vol. 4 (1) pp. e5; Halder et al. A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. Hum Mutat (2008) vol. 29 (5) pp. 648-58; Tian et al. Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet (2008) vol. 4 (1) pp. e4; Price et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet (2008) vol. 4 (1) pp. e236; Paschou et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet (2007) vol. 3 (9) pp. 1672-86; and Bauchet et al. Measuring European population stratification with microarray genotype data. Am J Hum Genet (2007) vol. 80 (5) pp. 948-56.). These and similar lists can be used to build a panel of AIMs for which a device or method herein can be configured to test for.

[0078] Some AIMs are also causal genetic variants. For example, the Duffy Null (FY*0) genetic variant causes an absence of a blood antigen. This variant is at nearly 100% frequency in sub-Saharan African populations and at nearly 0% frequency in populations outside of sub-Saharan Africa. Many causal genetic variants associated with pigmentation are also AIMs. Conversely, AIMs that are not causal genetic variants can be indirectly associated with traits caused by other AIMs. When selecting a set of causal genetic variants and ancestry informative markers, it is useful for the AIMs to be different than the causal genetic variants.

II. SYSTEMS

[0079] In an aspect, a computer readable medium comprises logic that predicts a probability of a phenotype of a child from a set of parents or potential parents with respect to at least one rare genetic disease based at least in part on the presence or absence of a plurality of causal genetic variants in a first parent, the presence or absence of a plurality of causal genetic variants in a second parent, at least one AIM from the first parent and at least one AIM from the second parent. In another aspect, a computer readable medium comprises: logic for performing a fully probabilistic analysis on data corresponding to a plurality of causal genetic variants from a male and a female to predict a probability of a phenotype of a child, wherein the male and the female are potential parents of the child. In some instances, performing a fully probabilistic analysis comprises carrying statistical noise throughout the entire process. In some instances, the causal genetic variants from each of the male and female comprise one or more ancestry informative markers (AIMs), one or more causal genetic variants corresponding to a rare genetic disease, one or more causal genetic variants corresponding to a personality trait, or both. In some instances, the computer readable medium further comprises logic for receiving input from a questionnaire completed by said male and/or female and assigning a weighting function to the plurality of causal genetic variants based on said input. In yet another instance, the information from the questionnaire comprises one or more of the following types of information regarding the male or female: height, weight, and family disease history. The computer readable medium can provide an output in the form of a report detailing the probability of the phenotype of the child. Logic can be computer readable instructions as described herein.

[0080] In an aspect, computer readable instructions for predicting a phenotype of a child are provided that when executed: perform a fully probabilistic analysis on data from a male and a female corresponding to a plurality of causal genetic variants to predict a probability distribution over child phenotypes, wherein the male and the female are potential parents of the child. In some instances, a fully probabilistic analysis is an analysis that does not make a determination based on a probability at one step, but rather carries the probability throughout the entire analysis. For example, if genotype AA has a probability of 91% and genotype AT has a probability of 9% then when calculating child risk in subsequent steps, the probabilities of 91% and 9% are carried through to

that calculation step, instead of using the most probably genotype AA. In some instances, this can be an important model for alleles that have low probability but very high relative risk. The variants from each of the male and female can comprise one or more ancestry informative markers (AIMs), one or more causal genetic variants corresponding to a rare genetic disease, or both. The instructions can receive input from information from a phenotype battery from said male and female and assign a weighting function to the plurality of causal genetic variants to predict the probability of the phenotype of the child. The phenotype battery can be collected by a variety of items or means, including without limitation: questionnaires, surveys, tests (such as IQ tests), medical records, and games. The information from the phenotype battery used when executed by the instructions can be height, weight, and family disease history. Elements of the phenotype battery can include, but is not limited to: age at first menopause, agreeableness, airway histamine responsiveness, alcohol dependence, birth weight, blood pressure, body mass index, cannabis dependence, chronic pelvic pain, coffee consumption, complete blood count, conscientiousness, DSM-IV major depressive disorder, depth of sleep, endometriosis, exercise participation, extraversion, fainting, family history of disease, fatigue, finger ridge count, fitness, forced expiratory volume, harm avoidance, hay fever, height, Immunoglobulin E levels, incisor geometry, individual alpha frequency (EEG), inspection time, insulin concentration, IQ, liability to appendectomy, liability to asthma, liability to tonsillectomy, male baldness, mole size and count, mouth ulcers, neuroticism, novelty seeking, openness, osteoarthritis in women, parietal P300 latency, parietal slow wave amplitude, performance IQ, persistence, prefrontal P300 latency, prefrontal slow wave amplitude, premature parturition for any birth, presence of eating disorder, psychosis proneness, quality of sleep, reward dependence, smoking initiation, stuttering, susceptibility to migraine, tea consumption, total cholesterol, triglycerides, weight, verbal IQ, voting behavior, and VO2max. The instructions can provide an output in the form of a report detailing the probability of the phenotype of the child.

[0081] In an aspect, a system for predicting a child phenotype comprises: at least two a nucleic acid detection devices configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM), wherein a first device is in contact with a sample from a female and wherein a second device is in contact with a sample from a male, wherein the male and the female are parents or potential parents of a child; a reader configured to read data from the devices; and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants and the at least one ancestry informative marker to predict a probability of a phenotype of the child with respect to the at least one rare genetic disease. The at least two nucleic acid detection devices can comprise a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants and the at least one AIM.

[0082] In an aspect, a system for predicting a child phenotype comprises: at least two a nucleic acid detection devices configured to detect a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases, wherein a first device is in contact with a sample from a female and wherein a second device is in contact with a sample from a male, wherein the male and the female are parents or potential parents of a child; a reader configured to read data from the devices; and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants to predict a probability of a phenotype of the child with respect to the more than 85 rare genetic diseases. The at least two nucleic acid detection devices can comprise a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. The at least two nucleic acid detection devices can further comprise a plurality of nucleic acid probes that selectively bind to at least one ancestry informative marker (AIM). The computer readable instructions when executed can utilize the data from the reader corresponding to the at least one AIM to predict the probability of a phenotype of the child.

[0083] In an aspect, a system for indicating if a subject is a carrier of a rare genetic disease comprises: a reader configured to read data from a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM); and computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants and the at least one ancestry informative marker to predict a plurality of probabilities of the subject being a carrier for each of the plurality of causal genetic variants.

A. Genotyping Individuals

[0084] There are many different ways of acquiring information about an individual's genetic sequence. Discussed herein are exemplary methods for analyzing genotype data produced by SNP chips. Some of the methods relate directly to solving some of the issues with identifying causal genetic variants for rare genetic diseases, for example: minimizing the false negative rate (for example, the probability that the test comes back negative when the couple is at risk for having a child with the disease) while keeping the false positive rate below a specified threshold. However, the same concepts can be generalized to other methods for producing sequence data (such as resequencing by sequence capture and sequencing-by-synthesis) or for ascertaining sequence information indirectly (for example by making measurements on RNA or protein proxies).

[0085] Genotype calling for biallelic SNPs is traditionally done by clustering. Given alleles X and Y, a scatterplot of the normalized signals for the X and Y allele can be created. The XX, XY, and YY genotypes manifest themselves as distinct clusters of points in this scatterplot. Other types of variants such as insertions, deletions, and triallelic SNPs can be accommodated by using multiple

biallelic probes. Genotype calling can proceed by clustering in a higher dimensional space, in which more than two axes are present.

[0086] A false negative call can correspond to informing a couple that they are not at risk when in reality they are. This is a failure of a screening test and may result in a child with a disease. In cost terms, the standard measure for the price of treating a child with a rare Mendelian disorder is approximately $1 million.

[0087] Suppose that the wild type allele is X and the disease allele is Y at a given locus. At this locus, most couples will be XX (normal) or XY (carriers), though incomplete penetrance/expressivity may lead to a few YY individuals. The determination of an individual's ancestral group from AIMs and self-reported ethnicity (as described in claim B below) provides prior probabilities for the XX, XY, and YY genotype frequencies at the disease locus. In many embodiments, ancestry determination is important because many Mendelian disease loci are much more frequent in one population than another.

[0088] Most genotype calling software uses the prior probabilities for genotype frequencies as the cluster priors, as the calling algorithm is optimized to minimize misclassification probability. For example, for a biallelic SNP, if A is the ancestral group which is being assayed, MX is the measured normalized signal in the X channel and MY is the measured normalized signal in the Y channel, the clustering procedure for genotyping calling is equivalent to estimating P(G|MX,MY,A). The optimal way to do this is with Bayes' rule as shown in Figures 8-12.

[0089] In many instances, the devices and systems herein can be designed err on the side of identifying people as carriers or no call if there is any doubt about them being wild type. Geometrically, what this corresponds to are looser cluster boundaries where we are more aggressive in calling XY than baseline carrier probabilities would suggest. This may be desirable because false positives may only require followup tests, while false negatives may represent a failure of a screening test.

[0090] Mathematically, the exact geometry is determined by the choosing the boundary which minimizes a measure of cost (for example expected or median cost) rather than the one which minimizes misclassification probability. Computationally, this may require specialized software (such as the computer readable instructions provided herein) to process the raw chip readout.

[0091] In order to attain a desired false positive rate of 1/100 or less, and false negative rate of 1/10000 or less for a multiplexed assay, the following protocol can be carried out: a) infer ancestry to obtain population specific cluster parameters as shown in Figure s4-7, b) assay each variant several times on the same chip, and c) use a control sample with each batch of chips to confirm manufacturing validity.

[0092] The use of repeated measurements allows these false positive and negative rates to be achieved. As an example, consider an assay for a variant with an average accuracy of 99%, where

the accuracy is the probability that the actual genotype is the same as the called genotype. With symmetric error (for example, with default genotype calling) this corresponds to a false positive and false negative rate of 1%. If C is the called genotype and G the actual genotype, then formally we have: the false positive rate: P(C=XX|G=XY) = 0.01, and the false negative rate: P(C=XY|G=XX) = 0.01.

[0093] If three of these assays are used in parallel and have independent failure rates, a combined best-of-three vote for genotype calling can achieve much better false positive and negative rates. Suppose that C1, C2, C3 are the genotype calls of these three assays. Then the probability that all three assays produce false negatives at the same time is:

P(C1=XX,C2=XX,C3=XX|G=XY) = P(C1=XX|G=XY)P(C2=XX|G=XY)P(C3=XX|G=XY) = (.01)^3 = 1 * 10^-6.

This is now one in one million, which is much less than one in 100, and achieves the false negative rates necessary for a feasible multiplexed assay.

[0094] It is possible to achieve approximately independent failure rates by locating probes in different regions of the chip or by using multiple chips or assay types. However, the assumption of independent failure can be compromised if there is systematic error, either in the design of unreliable probes or the presence of samples with extremely low DNA levels.

[0095] In practice, probe reliability can be guaranteed by using a funnel approach, by starting with many redundant probes and triaging those probes which perform poorly in a representative training set. Specifically, a combination of random samples, engineered plasmids, and samples from DNA repositories like Coriell may be used to empirically determine sensitivity and specificity for each variant of the assay. Variants with poor calling statistics are not used for risk calculation. A control sample or samples may also be prepared and run alongside customer DNA to ensure that each batch of assays/chips has been manufactured properly.

B. Inferring Ancestry

[0096] Given a set of measurements of N ancestrally informative markers X_1 through X_N, we can infer membership in a given ancestry group through standard classification techniques. For example, we can use expectation-maximization based clustering or similar algorithms to achieve probabilistic classification into categories, using assayed variants as features and self-reported ancestry as training data. Often individuals who have four grandparents hailing from the same geographic area or ethnic group are preferred as training points. Those whose grandparents hail from different regions or ethnic groups are then considered admixed. This analysis can be naturally extended to the case where the ancestral categories are hierarchical and/or overlapping. For example, it is possible to define multiple types of ancestry: continental scale (European, Asian, African), national (English, French, German), linguistic (French Canadian), and so on. Different

classifiers employing different combinations of ancestrally informative markers can be used to distinguish between virtually any such collection of categories.

[0097] For the purposes herein, the categories which are most relevant are those that correspond to populations with significantly increased rates of endogamy or frequencies of particular disease causing alleles. AIMs to identify all of these populations can be included on the chip.

[0098] AIMs in general are useful because ancestral self-report, while a useful starting point, can be misleading or incomplete. In North America, for example, many Caucasians do not know their admixture proportions. Given that populations like Italians and English differ in their frequencies for disease alleles like the Mediterranean variant for beta thalassemia (higher in Italians: http://www.genome.gov/10001221), AIMs can provide significant and useful information beyond that provided by self report. In particular, for admixed individuals it is also possible to assign ancestry on a more granular basis to particular regions of DNA (http://www.genome.org/cgi/content/abstract/gr.072850.107).

[0099] Ultimately, AIMs are useful for the muliplexed chip for uncommon Mendelian traits because knowledge of ancestry provides a prior distribution over genotype frequencies, thereby reducing the misclassification probability. For example, someone who is not Ashkenazi Jewish is unlikely to be a carrier for familial dysautonomia (http://en.wikipedia.org/wiki/Familial_dysautonomia), and the extra information on ancestry provides an independent check in addition to the direct results of the assay.

C. Determining Genetic Risk

[00100] Risk calculation in the context of an assay for rare genetic diseases can be that the false positive and false negative rates be held to within acceptable parameters. The discussion on genotyping above is the primary means of achieving this. Given reliable genotyping strategies with false positive rates on the order of 1% or less and false negative rates on the order of 0.001% or less, standard risk calculation methods can be applied to determine the risk that a couple will have a child with a given genetic disease. Application of such methods requires at a minimum knowledge of the mode of inheritance of the disease. For many traits, risk calculations may involve more complex assessments involving the possibility of incomplete penetrance or expressivity, de novo or recurrent mutations, genetic anticipation, and similar complicating factors.

[00101] Survey questions, including an online family history, may be important for augmenting risk calculations in the event that an assay result alone is not dispositive. However, for many of the Mendelian diseases assayed on the chip, a traditional consultation may not provide much extra information because carriers are often asymptomatic and most carriers do not have immediate family members with a rare Mendelian disease. Any positive result from a device herein may be subject to a specialized followup test in the event that the false positive rate is non-negligible.

D. Predicting Phenotype from Parental Genotype and Phenotype

[00102] The devices and systems herein provide genotype data that can be used to forecast the risk of having a child with a given disease. Conceptually, this forecasting is possible because of the predictable mode of inheritance of a Mendelian trait and the generally high correlation between having a variant and having the condition, notwithstanding the issues involved with variable expressivity and penetrance. For complex traits, it is still difficult to predict their values from genotype measurements alone. It may be possible to do this in the future by measuring CNVs, from resequencing data, or by combining multiple assay modalities (for example methylation, expression, etcetera), but the results from SNP chips alone have not explained much of the variance. For example, a recent whole genome association study of IQ found that only 1% of the variance in IQ could be predicted by a measurement of 500000 SNPs (http://www.blackwell-synergy.com/doi/abs/10.1111/j.1601-183X.2007.00368.x). Similarly low values have been found for other traits like height, where large studies involving 30000 plus individuals have found SNPs explaining only 3.7% of the variance in height.

[00103] Importantly, while it is difficult to predict the values of an individual's complex traits from genotype information, it is often easy to predict the values of a child's complex traits from phenotype information on the parents. That is, for phenotypes with high heritability, a measurement of mother and father trait values is sufficient to provide excellent predictive power for child traits. For example, the narrow sense heritability of height is on the order of 0.70 to 0.80, depending on the population and environmental background (http://www.sciam.com/article.cfm?id=how-much-of-human-height). This means that the proportion of child variance explained by two measurements (the mother height and father height) ranges from 49% to 64%, a number which is far in excess of what has been achieved with SNP based predictions of individual height as of May 2008. It is thus possible to exploit this phenomenon for the purposes of child prediction. For traits with nonzero heritabilities, child trait values can be forecast from parental measurements with low error rates.

[00104] A survey of parental and child risks or phenotypes from a random sample of 1000 people from a target population may be used to provide empirical training data for regression equations relating parental and child characteristics. AIMs may also be collected to facilitate identification of members of this population. It is also possible to include data from other family members, such as grandparents, uncles, and aunts, as additional predictive features for predicting child traits. These regression equations may be more sophisticated than the simple linear models employed in quantitative genetics. Any kind of regression algorithm may be employed to predict child phenotype from the measurements of parental phenotypes. For example, this is of demonstrable utility in the case of longevity, which has a highly nonlinear inheritance pattern (http://longevity-science.org/Genetics-Australia.html). Importantly, the nonlinearity of longevity inheritance results in a deceptively low heritability coefficient of 0.10.

[00105] Standard heritability estimates may thus underestimate the ability to predict offspring traits from parental values, indicating that it may be useful to assay any trait of potential commercial relevance on a large number of parent/child trios from a given population and determine empirically whether prediction is feasible or infeasible. As before, it may be useful to assay other relatives besides immediate parents. It may also be useful to return confidence intervals or quantiles in addition to point estimates. The regression equations for predicting complex child risks or phenotypes from parental phenotype and the risk calculation algorithms for predicting child disease risk from parental genotype form a unified suite of algorithms for child prediction. From an input/output perspective, parents provide survey data and saliva samples, and algorithms return child predictions of both Mendelian disease risks and complex traits.

[00106] As described herein, population information can determine cluster parameters. Figures 4-7 illustrate a theoretical basis by which population information (as revealed by AIMs and/or self-report) can inform cluster parameters and genotype calling, and thereby influence the accuracy of carrier testing and child phenotype prediction. Population 1 is a group which has a hypothetical A/G SNP with the given flanking sequence in Figure 4. If 1000 individuals from this population are genotyped, 1000 measurements are obtained of the normalized intensities of the A and G channel probes at that SNP. These intensity values can be depicted in a scatterplot as shown, where each point represents an individual and the X and Y axes correspond to the A and G channels. In this example, three clusters arise, corresponding to the AA, AG, and GG genotypes. Parametric models of various kinds can be fit to these clusters. One of the most popular is the expectation maximization (EM) model, in which a multivariate gaussian distribution with parameters mu (the mean vector) and sigma (the covariance matrix) is fit to each cluster. In this example, there are three mu vectors of dimension 2, and three sigma matrices of dimension 2 x 2 (in this case the identity matrices). The relative area under each multivariate gaussian is given by the pi parameter, corresponding to the proportion of points which lie in that cluster. In Figure 5, population 2 has exactly the same variant with the same flanking sequence, but different proportions of each genotype (0.16, 0.48, 0.36 rather than 0.25, 0.50, 0.25). Because the molecular biology of the variant and the flanking sequence is the same, the clusters are in the same position and only pi differs. In Figure 6, population 3 has a different flanking sequence with an extra G. The molecular biology differences in the flanking sequence can cause shifts in the positions and orientations of the clusters, for example by translating clusters upwards on the G-axis or introducing more variance along the G axis. In such a case, the mu and sigma parameters may also differ along with the pi parameters. The previous Figures 4-6 have been considering populations in terms of geographic ancestry, however Figure 7 considers populations by gender. Specifically, suppose that population 4 is composed of males, that the locus under consideration is on the X-chromosome, and that all members of previous populations were females. In this case the intensity plot for males would look quite different. In particular, only two

clusters are possible (A/- and G-/) rather than three (AA, AG, GG). Thus the cluster parameters mu, sigma, and pi are different from that of population 1. These examples illustrate the importance of taking population information (for example as informed by gender or ethnicity) into account when training calling models.

[00107] In many instances, population information aids genotype calling. Figures 8-12 demonstrate an example of how population information influences genotype calling and risk calculation in the context of beta-thalassemia, which sharply differs in frequency between two phenotypically similar populations: Italians and Caucasians from the general population. The carrier frequency for the deleterious beta-thalassemia variant differs starkly between Northern European and Italian populations, who are otherwise phenotypically similar and highly admixed in countries like the United States. For example, the carrier frequency in Northern Europeans is less than 1 per 1000, whereas the carrier frequency in Italians is greater than 1 per 100. The plot in Figure 8 shows an intensity scatterplot for the beta thalassemia deletion allele similar to the scatterplots in Figures 4-7. For simplicity only two clusters are shown, corresponding to the wildtype and carrier genotypes and excluding the homozygote. A key question is how to classify points with intensities that are not extremely close to the means of their respective clusters, such as 'A' in Figure 8. In Figure 9, it is shown that calling of points near the boundary is aided by the use of ancestry information. The numerical example in Figure 10 shows that a raw intensity measurement at a causal locus is not the only value that matters. Coupling this measurement (X) with knowledge of ancestry (A) can change the probability of being a carrier by almost 60 fold. A source code example is provided in Figure 11. In practice one will not have complete knowledge of ancestry. AIMs help in this situation. One may use signal from AIMs to produce a distribution over populations, obtaining the probability that a given locus is inherited from an ancestor of a given population group. As shown in Figure 12, this requires a straightforward modification to the expression for the posterior probability of whether an individual is in a given cluster. Instead of assuming that the value of the ancestry parameter is known with probability 1, a distribution can be used over ancestries. This technique of using AIMs to improve calling at causal loci provides a substantial gain in classification power and is particularly useful for assays such as those described herein, as it is indicated for many different ancestry groups.

[00108] There are several possible sources of noise that should be accounted for when performing child phenotype predictions using genotype information. Figures 13-18 illustrate the process of fully probabilistic prediction of a distribution over child risks and phenotypes. The process accounts for several possible sources of uncertainty, including genotype calling error, phasing uncertainty, recombination, gametic union, and a noisy genotype-to-phenotype map. First, as shown in Figure 13, a measured genotype may have measurement error as represented by a false positive and false negative rate, but more generally one would specify a confusion matrix that gave the

probability of predicting genotype i given a true genotype j. Given a measured genotype, the result is a distribution over possible true genotypes. Next, as shown in Figure 14, the inferred genotype is unphased. That is, adjacent heterozygote loci may be in either cis or trans and this is unknown. One can model phase uncertainty by either: using an appropriate database of population haplotype frequencies at the given locus; or using one of several haplotype inference algorithms, such as Clark's algorithm or model based haplotype inference. Figure 15 shows how to generate a distribution over possible recombinants and hence possible gametes from the possible haplotypes. Figure 16 then shows how to repeat this process for both mother and father to obtain a probability distribution over gametic unions, corresponding to phased child genotypes (aka haplotypes). This results in a distribution over child haplotypes. Finally, Figure 17 reviews several different kinds of genotype-to-phenotype maps that can be used with the distribution over child haplotypes to produce the desired quantity: a distribution over possible child risks or phenotypes.

[00109] Figure 18 depicts equations for obtaining the distribution over estimated child risks or phenotypes given parental genotypes, formalizing the flow diagrams of Figures 13-17. This expression is useful in that it is exact and can be approximated for any given locus. This can be done via simulation or by noting that certain steps collapse; for example, predicting child phenotype based on a single locus trait need not deal with recombination. It is notable that each conditional distribution can be expressed as a matrix and the series of sums can be expressed in terms of repeated matrix multiplications, a formulation which allows rapid computer implementation.

[00110] In the future, as resequencing costs drop, it is important to note that parent phenotypes and genotypes determined in this manner can then be regressed upon each other. That is, the voluntary large scale provision of parental genotype and phenotype information for the purposes of predicting child traits can be used as a source of data for extremely large scale association studies.

E. Biological Relatives Genomics

[00111] The methods herein can be utilized for family genomics. In some aspects, the methods can be utilized to determine probabilities of phenotypes or probabilities that family member has or had a rare genetic disease based upon genetic screening of other family members. For example, a subset of biological relatives can be screened as carriers for a plurality of causal genetic variants of at least one rare genetic disease. The data from the subset can then be utilized to predict the probability of phenotype of another biological relative that was not a member of the subset tested. For example, three children can be tested and the probability that a parent, living or deceased, has a particular phenotype or rare genetic disease as described herein can be calculated. These methods may be useful for postmortem determination of relative that may have had an unclear diagnosis of a disease at the time of death or did not have the tests available to him to determine if he had a rare genetic disease.

[00112] In some instances, the devices and systems as described herein that can be used to predict child phenotype probabilities can be utilized to predict probabilities of living relatives having certain traits or diseases or phenotypes. Relatives can include grandparents, great grandparents, aunts, uncles, mothers, fathers, siblings, children, and any other biological family member.

[00113] In some instances, a relative who was not included in the subset of biological relatives that were genotyped with regards to causal genetic variants can be referred to a genetic counselor and/or physician. In an instance, a relative who is a member or not member of the subset of biological relatives can pursue a course of medical action based on the results of the tests. For example, aspirin can be prescribed for heart disease to a family member based on risk calculated by a method or system herein. Information can also be collected from a phenotype battery, such as that discussed herein, for some or all of the subset of biological relatives and utilized in the calculations of the probabilities for other relatives. The risk of developing a disease for any biological relatives can be calculated and adjusted based upon family genomics, such as described herein. The results can be delivered to the relative and actionable methods can be carried out. For example, a genomic family tree can be created and utilized to factor into a probability of risk of disease for one, some, or all members of their family tree.

III. DEVICES

[00114] Described herein is a universal carrier screen device and methods of utilizing the device. Also disclosed is computer software and code for translating data from the devices and methods to probabilities that a user is a carrier or that a set of parents have a probability of having a child with a genetic disease. A device herein can be configured to test for a plurality of rare genetic diseases or at least 5, 10, 85, 100, 200, 500, or 1000 of rare genetic diseases.

[00115] In an aspect, a nucleic acid detection device is configured to test a sample for a plurality of causal genetic variants corresponding more than 85 rare genetic diseases. The device can be further configured to test a sample for at least one ancestry informative marker (AIM). In some instances, the device comprises a plurality of nucleic acid probes that selectively bind to a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. The device can further comprise a plurality of nucleic acid probes that selectively bind to at least one ancestry informative marker (AIM). The device can comprise a bead array that selectively binds to a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases. The device can further comprise a bead array that selectively binds to at least one ancestry informative marker (AIM). In some instances, the at least one AIM is not a causal genetic variant. In some instances, at least two of the rare genetic diseases occur at frequencies that differ by at least 10-fold in at least two distinct populations, wherein the at least two distinct populations are differentiated by the at least one AIM. In some instances, the device comprises a sequence capture assay to detect a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases.

[00116] In some instances, the device simultaneously detects 85, 100, 150, 200, 500, 1000 or more rare causal genetic variants. In many instances, the causal genetic variants indicate a Mendelian disease in which a single gene is responsible for a disease. Often times, Mendelian diseases have a recessive inheritance pattern as demonstrated in Figures 4-7 and a person has two recessive alleles to develop the disease. Individuals with one mutant and one non-mutant allele can be described as carriers because they carry the causal genetic variant but do not have the disease. If parents are both carriers for the same disease, their children are at a risk for inheriting two copies of the mutant gene and thus developing the disease. For an exemplary Mendelian disease, one quarter of the children of two carriers will be expected to inherit the disease. Because the risk of having an affected child is a deterministic function of the genes of the parent, genetic tests are commonly used to preemptively identify parents who are carriers. These tests are considered screening tests because they are given to otherwise healthy individuals. If a couple is found to be at high risk for having a child with a Mendelian disease, in an example a genetic counselor or physician can help them consider their reproductive choices. Many Mendelian diseases that are understood at a molecular level are severe, resulting in early death or serious debilitation

[00117] A device herein can be a single test intended for panethnic population screening for a panel of Mendelian diseases.

[00118] Described herein is a device configured to test for causal genetic variants, for example, genetic variants that cause a particular phenotype. The device also can be configured to test for the presence of ancestry informative markers (AIMs), that is, genetic markers that correlate with populations, in particular human populations, coming from specific geographic locations. Causal genetic variants contrast with genetic markers in that certain markers may be highly correlated with the expression of a trait without actually causing the trait. In some instances, causal genetic variants can be selected wherein some populations (for example, populations that can be distinguished by AIMs) have different frequencies for traits or diseases caused by the causal genetic variants. Herein, a device can be configured to test for one or more AIMs and the AIMs can provide prior knowledge about the user of the device and to which population(s) to which he belongs. The prior knowledge of the population background of a user can then be utilized by the software and methods described herein to adjust the probability that the user is a carrier for a specific trait or a plurality of traits. For example, if a person tests positive for an allele that causes Tay Sachs disease and also tests positive for an AIM associated with Ashkenazi Jews, then the probability that the person is a carrier for the Tay Sachs gene can be calculated. On the other hand, if a person tests positive for an allele that causes Tay Sachs disease and tests negative for an AIM associated with Azhenazi Jews, then there may be reason to believe that the Tay Sachs results is a false positive and that can be reflected in the probability calculation as described herein.

[00119] In an example, two users (one male, one female) receive their probability of being a carrier for a plurality of causal genetic variants as described herein and the probability of the child of the two users developing traits from the causal genetic variants can be calculated using the methods and systems herein.

[00120] In some instances, a device comprises tests for causal genetic variants for traits for which different populations are at increased risk, as well as tests for AIMs for these populations. The traits can be Mendelian traits having low incidence in a population.

[00121] Devices herein are configured to test for genetic variants and/or ancestry informative markers. Many such devices for these purposes are known. They include, without limitation, nucleic acid arrays (including oligonucleotide arrays, spotted arrays, photolithographically generated arrays, random arrays and complete genomic hybridization arrays) and nucleic acid sequencers. These devices detect genetic variants or ancestry informative markers in a variety of ways. Such ways include sequencing by hybridization, sequencing by ligation, sequencing by extension, Sanger sequencing, Maxam Gilbert sequencing and pyrosequencing. These devices and methodologies are all well known in the art. Knowledge of the specific sequence or variant one wishes to detect can provide the information necessary to configure the device.

[00122] A device herein can be created from, for example, nucleic acid arrays (for example, Affymetrix DNA microarrays or Illumina bead arrays), single molecule sequence by synthesis methods (for example, Helicos BioSciences Corporation), amplification of nucleic acid molecules on a bead (for example, 454 Lifesciences), clonal single molecule arrays technology (for example, Solexa, Inc.), single base polymerization using enhanced nucleotide fluorescence (for example, Genovoxx GmbH) and single molecule nucleic acid arrays (for example, Drmanac, U.S. Patent Application No. 2007/0099208).

[00123] Methods and devices are described in many patents including for example, United States Patent 7368265 (Brenner et al. Selective genome amplification), United States Patent 7361468 (Liu et al. Methods for genotyping polymorphisms), United States Patent 7332277 (Dhallan, Methods for detection of genetic disorders), United States Patent 6576424 (Fodor, nucleic acid arrays) and United States Patent 7323305 (Leamon et al., Methods of amplifying and sequencing nucleic acids).

A. Device Types and Methodology

[00124] With the exception of stable epigenetic modifications, genetic variants and markers at the level of DNA are typically detected as variations in the sequence in DNA nucleotides. The processing of detecting variations in the sequence in DNA nucleotides is also called genotyping. Genotyping of DNA in a sample is performed by contacting the sample with a device or devices.

[00125] It is possible to detect variations in the sequence in DNA nucleotides by randomly sequencing until such a point that sufficient coverage is achieved. Random sequencing is usually not truly random but pseudorandom, and may be biased. A number of technologies permit random

sequencing. It is preferable for the sequencing technology to excel on a number of performance standards, including accuracy. A commercial sequencing technology can be used to facilitate meeting high performance standards. However, a non-commercial platform can be used as well.

[00126] In another embodiment the variants are tested by SNP analysis, complementary genomic hybridization, gene sequencing, sequence capture, DIP analysis, STR analysis, and CNV analysis. In another embodiment the sub-devices comprise a nucleic acid array, a bead arrays, a CGH array, a collection of molecular inversion probes, a collection of allele specific oligonucleotide (ASO) probes, pyrosequencing, Sanger sequencing, sequencing by synthesis and sequencing by hybridization. In another embodiment the causal genetic variant is a SNP (for example, missense, nonsense, splicing, or regulatory), DIP (for example, small deletion, small insertion, gross deletion, or gross insertion), CNV, repeat polymorphism, or complex rearrangements.

[00127] Sequencing can be performed by pyrosequencing. This technique is based on the detection of released pyrophosphate (Hyman, 1988. A new method of sequencing DNA. Anal Biochem. 174:423-36; and Ronaghi, 2001. Pyrosequencing sheds light on DNA sequencing. Genome Res. 11:3-11). Sequencing can be performed by sequencing by synthesis. Sequencing by synthesis is also called sequencing by incorporation (United States Patent 7344865). Sequencing can be performed by sequencing by ligation. The Applied Biosystems SOLiD sequencing system performs sequencing by ligation. Sequencing can be performed by Sanger sequencing. Sanger sequencing is also called dideoxy chain termination sequencing. Sequencing can be performed by single-molecule sequencing, which is a broad class of sequencing technologies in various stages of development.

[00128] Random sequencing can be an inefficient method of detecting particular variations in DNA sequence. To increase efficient coverage of particular variants, random sequencing can be augmented by a selective amplification of the regions of DNA that contain the variants of interest. A number of technologies permit selective amplification. It is preferable for the selective amplification technology to excel on a number of performance standards, including fidelity. A commercial selective amplification technology can be used to facilitate meeting high performance standards. With appropriate validation and optimization, a non-commercial technology can be used. The aim of selective amplification is to enrich for particular sequences within a complex mixture.

[00129] Selective amplification can be performed by polymerase chain reaction (PCR). PCR can be multiplexed. Multiples reactions can be performed in parallel. PCR requires optimization of reaction conditions, a standard laboratory practice.

[00130] Selective amplification can be performed by sequence capture. Sequence capture involves the use of hybridization to enrich for sequences of interest. Sequence capture can be performed using a microarray (Albert TJ, et al. Direct selection of human genomic loci by microarray hybridization. Nature Methods. 2007 Nov; 4(11):903-5; Okou DT, et al. Microarray-based genomic selection for high-throughput resequencing. Nature Methods 2007 Nov; 4(11):907-9; and Hodges E,

et al. Genome-wide in situ exon capture for selective resequencing. Nature Genetics 2007 Dec; 39(12):1522-7). For example, the NimbleGen Sequence Capture technology can be used for selective amplification. Selective amplification can be performed by rolling circle amplification. Selective amplification can be performed by cloning. Cloning can involve the use of a restriction endonuclease to cleave a complex mixture of DNA into fragments of predictable lengths.

[00131] It is also possible to detect particular variations in the sequence of DNA nucleotides by targeted genotyping. A number of technologies permit targeted genotyping. A number of technologies permit selective amplification. It is preferable for the selective amplification technology to excel on a number of performance standards, including fidelity. A commercial selective amplification technology can be used to facilitate meeting high performance standards. With appropriate validation and optimization, a non-commercial technology can be used. The efficiency of targeted genotype techniques can also be augmented selective amplification of the regions of DNA that contain the variants of interest, including PCR, sequence capture, rolling circle amplification, and cloning.

[00132] Targeted genotyping can be performed by sequencing by hybridization. Sequencing by hybridization can be performed using a microarray. Targeted genotyping can be performed by microarrays. An example of a technology that allows targeted genotyping by microarray is the Affymetrix Genome-Wide Human SNP Array 6.0. Targeted genotyping can be performed by bead arrays. An example of a technology that allows targeted genotyping by bead arrays is the Illumina Human1M BeadChip. Targeted genotyping can be performed by comparative genome hybridization (CGH) arrays. CGH arrays involve the co-hybridization of a reference sample and a target sample to the same array. Targeted genotyping can be performed by molecular inversion probes. An example of a technology that allows targeted genotyping by molecular inversion probes is the Affymetrix GeneChip Targeted Genotyping platform. Targeted genotyping can be performed by allele specific oligonucleotide (ASO) probes. The design of ASO probes permits a number of large number of permutations. ASO probes can be 15 to 21 nucleotides in length. ASO probes can contain zero, one, two, three, or more mismatches compared to the variants being detected. ASO probes can contain continuous stretches nucleotides at one end with zero or one mismatches to the variants being detected. Targeted genotyping can be performed by allele specific PCR. Targeted genotyping can be performed by TaqMan assays. Targeted genotyping can be performed by single base extension. Targeted genotyping can be performed by restriction length fragment polymorphism (RFLP) typing. Targeted genotyping can be performed by diagnostic PCR. A variety of diagnostic PCR techniques have been developed. In one example, PCR amplification of a region flanking a DIP will produce fragments of different sizes according to presence or absence of the DIP. In another example, PCR amplification of a region flanking a STR will produce fragments of different sizes according to which alleles of the STR are present.

[00133] To increase efficiency, a multiplex genotyping technique can be used. Most genotyping technologies intrinsically permit multiplex genotyping. Those technologies which do not intrinsically multiplex can be made to perform like multiplex genotyping by performing multiple singular genotyping procedures in parallel or serial. This fashion of genotyping may also be described as high-throughput.

[00134] The nucleotide occurrence of a SNP in a sample can be determined by a variety of technologies, including a number not explicitly listed here. CGH arrays are generally not suitable for SNP genotyping. The nucleotide occurrence of a DIP in a sample can be determined by a variety of technologies, depending on the size of the deletion/insertion. In general, small deletions/insertions, for example those less than 20 nucleotides in size, can be genotyped with technologies used to genotype SNPs. Larger DIPs may be treated as CNVs. The nucleotide occurrence of a CNV in a sample can be determined by a number of technologies. SNP genotyping technologies have been adapted for use in CNV genotyping. CGH arrays are also often suitable for CNV genotyping. The nucleotide occurrence of a STR in a sample can be determined by a number of technologies. STRs are most commonly genotyping by diagnostic PCR.

[00135] Stable epigenetic modifications, such as methylation, can be detected be detected by a number of technologies for example, as disclosed in United States Patent 7364855. For example, the Illumina GoldenGate Methylation Panel and the Illumina HumanMethylation27 BeadChip can be used to detect methylation differences. Bisulfite sequencing can be used to determine the presence of methylation. Methylation can be detected by pyrosequencing, methylation-sensitive single-strand conformation analysis (MS-SSCA), high-resolution melting analysis (HRM), methylation-sensitive single nucleotide primer extension (MS-SnuPE), base-specific cleavage/MADLI-TOF, methylation-specific PCR (MSP), and others.

[00136] A panel of genetic variants and markers could be genotyped by combining several different technologies into a single system.

[00137] Some genetic variants and markers can be detected at the level of RNA polynucleotides. Reverse transcription polymerase chain reaction (RT-PCR) can be used in lieu of PCR or in addition to PCR to detect genetic variants at the level of RNA polynucleotides. Genetic variants can alter the sequence of an RNA polynucleotide. Detecting genetic variants that affect the sequence of an RNA polynucleotide is essentially equivalent to detecting genetic variants that affect the sequence of a DNA polynucleotide. Genetic variants can be detected as the presence of absence of an RNA polynucleotide, or more generally as the concentration of an RNA polynucleotide. A number of techniques can be used to determine the presence, absence, or concentration of an RNA polynucleotide. RT-PCR can be used to determine the presence, absence, or concentration of an RNA polynucleotide. Spotted microarrays, oligo microarrays, and bead arrays can be used to determine the presence, absence, or concentration of an RNA polynucleotide.

[00138] A device can be constructed that detects genetic variants and markers at both the level of DNA and the level of RNA.

[00139] Some genetic variants and markers can be detected at the level of protein polypeptides. Genetic variants can alter the sequence of a protein polypeptide. Genetic variants can be detected as the presence of absence of a protein polypeptide, or more generally as the concentration of a protein polypeptide.

B. Causal Genetic Variants

[00140] The devices herein are configured to test for causal genetic variants and/or for ancestry informative markers. Detecting the presence of a variant or marker includes, by implication, detecting its absence. In some instances, the devices are configured to perform genotyping on the trait, not merely the presence of one polymorphic form of the variant or gene. The tests selected can include variants for a plurality of different at-risk populations; for example, at least 2 different populations, at least 5 different populations, at least 10 different populations, at least 25 different populations, at least 50 different populations or at least 100 different populations. For example, the device could be configured to test for one or more causal genetic variants for which the Western Eurasian population is at increased risk, one or more causal genetic variants for which the sub-Saharan African population is at increased risk, one or more causal genetic variants for which the East Asian population is at increased risk and one or more causal genetic variants for which the Native American population is at increased risk.

[00141] In an aspect, a device is disclosed that tests for a plurality (for example, at least 100, at least 500 at least 1000) of traits. Diseases occur in populations with different prevalence. Traits or diseases can be ranked according to prevalence, many will be grouped among rare traits, for example, traits having prevalence of 1/1000 or less or 1/10,000 or less. Such traits are said to inhabit the long tail, of such a ranking. The concept of a long tail is easily visualized by a bar chart in which the Y axis indicates prevalence and the traits are ordered along the X axis from most to least prevalent. In certain embodiments, the number and traits and the incidence of each trait can be chosen so that the collective incidence of the rare diseases tested for amount to at least 1%, at least 2%, at least 4%, or at least 5%.

[00142] In certain devices the causal genetic variants tested for have different incidence in different populations or sub-populations. The difference in incidence is at least statistically significant ($p < 0.05$) and may be increased by at least 10%, at least 100%, at least 5 times, at least 10 times, at least 100 times at least 1000 times or at least 10,000 in two different populations. A variant may cause a trait having the same incidence in two different populations, but it typically will have different incidence in at least two different populations represented on the device.

C. Ancestry Informative Markers

[00143] Ancestry informative markers that a device tests for typically are selected with reference to the variants being tested for. In certain embodiments, the device tests for causal genetic variants that have different incidence in different populations or sub-populations. AIMs can distinguish between two populations for which one of the causal genetic variants exhibits a difference in incidence. The AIM also can classify a person as belonging to or not belonging to the population that is at increased risk for one of the causal genetic variants, for example, the AIM can be diagnostic for the population in which the trait is at increased prevalence. In certain instances the AIM may distinguish between populations with finer granularity, for example, between sub-continental groups or related ethnic groups. In this example, a device comprises a test for a causal genetic variant having different incidence between these sub-populations.

[00144] Because causal genetic variants can have different frequencies in different populations they have some level of predictive value about the population from which the individual comes. In certain embodiments the AIMs used on the device are not, themselves, causal genetic variants. In other embodiments the AIM or collection of AIMs used to predict the source population is/are selected to have a predictive value of at least 80%, 90%, 95%, 98%, or 99%.

[00145] In another embodiment the AIMs are selected to classify at least three different populations. In another embodiment the AIMs are selected to classify at least four different continental scale groupings. The AIMs can be selected to classify at least three different populations from the same continent. In another embodiment the AIMs are selected to classify at least ten sub-continental groupings. In another embodiment the AIMs are selected to classify three populations selected from Europeans, East Asians, South Asians, Sub-Saharan Africans, and Native Americans. In another embodiment the AIMs are located on at least 11 different human chromosomes.

D. Device Validation

[00146] In an aspect herein, methods and sets of nucleic acid pools are provided to validate the devices described herein. Genomic DNA samples are not publicly available for many rare genetic diseases. Therefore it can be difficult to validate or quality control a device configured to test for more than 85 rare genetic diseases simultaneously.

[00147] In an aspect, a set of nucleic acid pools is disclosed for validating a nucleic acid sequence detection device, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a plurality of causal genetic variant probes, and wherein each pool binds a different plurality of said probes. In some instances, a first pool of the set comprises a first nucleic acid segment that interferes during detection with a second nucleic acid segment of a second pool of the set, and wherein the first pool does not comprise the second nucleic acid segment and the second pool does not comprise the first nucleic acid segment. In some instances, the nucleic acid segments of each pool are a single nucleic acid molecule. In some instances, the nucleic acid segments comprise one or more plasmids.

[00148] In an aspect, a method of validating a lot of manufactured nucleic acid sequence detection devices comprises: contacting a plurality of nucleic acid sequence detection devices from the lot with a set of nucleic acid pools for validating a nucleic acid sequence detection device, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a plurality of causal genetic variant probes, and wherein each pool binds a different plurality of said probes; and detecting a presence of the plurality of causal genetic variant probes of the nucleic acid detection devices, wherein the lot of manufactured devices is validated if all of the plurality of causal genetic variant probes are present. In some instances, the method further comprises delivering the lot of manufactured devices when the devices are validated. In some instances, the lot of manufactured devices is rejected if not all of the plurality of causal genetic variant probes are present. In some instances, the lot of manufactured devices is modified and the method is repeated if not all of the plurality of causal genetic variant probes are present.

[00149] As described herein, DNA samples representing these rare genetic diseases can be generated by synthesis. Double stranded DNA can be synthesized de novo and cloned into standard high-copy plasmid vectors. This approach has also been recognized by the FDA guidance document "Pharmacogenetic Tests and Genetic Tests for Heritable Markers". The plasmids provide a renewable resource of DNA for use in validation and for use as controls for patient genotyping. Plasmids will be pooled, creating an artificial patient sample. The mutant alleles for each causal genetic variant herein can be synthesized.

[00150] In an instance, all of the plasmids comprising all of the causal genetic variants to be tested for by a device herein can be placed into a single large pool or sample. In many cases, many of the causal genetic variants are closely spaced together. In these cases, plasmids generated to test one variant are also templates for other nearby variants and the templates will generate a wildtype signal. This can make it difficult to appropriately simulate the causal genetic variant signal in the off-target probes they interact with. In an instance, overlapping sequences are split into different pools of plasmids that make up a set of nucleic acid segments. For example, variants in the same block are placed in different pools. In another example, the total number of pools needed in a set of pools is equal to the size of the largest block.

[00151] Examples of nucleic acid segments and pools of nucleic acids are demonstrated in Figure 19. Figure 19 illustrates the sequence architecture of the validation sample pools. The basic issue is that validation of a multiplex nucleic acid assay benefits from the use of non-interfering templates to train genotyping call boundaries.

[00152] For detecting causal genetic variants to validate a device described herein, plasmid pools are generated by combining one group from each block size (for example, one group each of size 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, and 34) in the appropriate concentration. Thus, each pool contains 455 synthetic variants. For example:

- Pool 1 comprises: Size 1 Group 0 + Size 2 Group 0 + Size 3 Group 0 + Size 4 Group 0 + Size 5 Group 0 + Size 6 Group 0 + Size 7 Group 0 + Size 8 Group 0 + Size 10 Group 0 + Size 12 Group 0 + Size 34 Group 0.
- Pool 2 comprises: Size 1 Group 0 + Size 2 Group 1 + Size 3 Group 1 + Size 4 Group 1 + Size 5 Group 1 + Size 6 Group 1 + Size 7 Group 1 + Size 8 Group 1 + Size 10 Group 1 + Size 12 Group 1 + Size 34 Group 1.
- Pool 3 comprises: Size 1 Group 0 + Size 2 Group 0 + Size 3 Group 2 + Size 4 Group 2 + Size 5 Group 2 + Size 6 Group 2 + Size 7 Group 2 + Size 8 Group 2 + Size 10 Group 2 + Size 12 Group 2 + Size 34 Group 2.

IV. BUSINESS METHODS

[00153] In an aspect, a method comprises: receiving a sample from a user; testing the sample with a nucleic acid detection device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM); calculating a plurality of probabilities corresponding to the user being a carrier for each of the plurality of causal genetic variants based on results from the testing step relating to the plurality of causal genetic variants and the at least one AIM; and delivering to the user the plurality of probabilities corresponding to the user being a carrier. The method can further comprise: receiving a sample from a second user; testing the sample from the second user with a device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM); calculating a probability of a child phenotype corresponding to the rare genetic diseases based on results from testing the user and the second user; and delivering the probability of the child phenotype to at least one of the user and the second user. The method can further comprise providing genetic counseling service to at least one of the user and the second user. The method can be carried out as part of a child phenotype prediction service. The method can further comprise obtaining phenotypic information from the user; and using the phenotypic information from the user in the calculating steps. The method can further comprise obtaining family history from the user; and using the family history from the user in the calculating steps.

[00154] In an aspect, a method comprises: marketing a genetic testing service comprising predicting a probability of a child phenotype from a set of parents or potential parents, wherein the prediction is based at least in part on the presence of a plurality of causal genetic variants for more than 85 rare genetic diseases of each of the members of the set of parents or potential parents and based at least in part on the inferred ancestries of each of the members of the set of parents or potential parents; and delivering a probability of the child phenotype from the set of parents or set of potential parents with respect to the more than 85 rare genetic diseases for a fee. The marketing can be conducted in connection with a dating or marriage service. The method can further comprise referring at least one

member of the set of parents or set of potential parents to a physician. In some instances, the inferred ancestries are inferred by a test for at least one ancestry informative marker (AIM).

A. Genetic Testing and Genetic Counseling Referral System

[00155] An exemplary business method allows for the provision of a genetic testing service that can predict the probability that an offspring of a couple will have each of a plurality of Mendelian traits. In addition to the results of the genetic testing, referrals to genetic counselors and/or other relevant medical professionals can be provided in order to provide for follow up testing and consultation.

[00156] In an embodiment of a business method, genetic testing of a customer begins with a customer order, wherein the a customer pays a fee in exchange for, for example genetic testing materials and referrals to genetic counselors and/or other relevant medical professionals. A customer can be, for example and without limitation, a physician, a genetic counselor, a medical center, an individual, an insurance company, a website, a dating service, a matchmaking service, a pharmaceutical company, a laboratory testing service provider, or a diagnostic platform manufacturer. For example, a customer can be a couple of prospective parents who seek to learn whether their offspring will be at risk for developing some sort of Mendelian disease or rare Mendelian disease. After a customer places an order, a DNA collection kit can be sent to the customer. In an embodiment of a business method of the invention, a customer can deposit a sample into the collection kit. Any sample that would be obvious to one skilled in the art can be deposited into or onto a collection kit. A sample can be any chemical compound that would be obvious to one skilled in the art, such as bodily fluid like saliva or a breath sample containing correlated chemical compounds (see for example, http://www.popularmechanics.com/blogs/science_news/4220196.html), from which it is possible to identify and extract the required Raw Genotypic Information (as discussed herein). The collection kit can then be returned to the company for sending to a genotyping lab or can be returned directly to the genotyping lab for processing. A genotyping lab, either internal within the company, contracted to work with the company, or external from the company, can isolate the customer's DNA from the provided sample. After the DNA has been isolated from the sample, an exemplary device of the invention can test the DNA for the presence of (i) ancestry informative markers and (ii) causal genetic variants (the combination of (i) and (ii) are also referred to herein as, Raw Genotypic Information). In an embodiment, the DNA does not have to be isolated from the sample to test the DNA for the presence of Raw Genotypic Information.

[00157] The Raw Genotypic Information can be processed to infer the ancestry of the customer and to confirm or deny the presence of causal genetic variants (also referred to herein as, Processed Genotypic Information). In an embodiment, Processed Genotypic Information can be transmitted to a system of the invention or another calculation system as would be obvious to one skilled in the art

to predict the probability that an offspring of the customer will have each of a plurality of traits caused by causal genetic variants found to be present in the customer's DNA.

[00158] After analysis of the genotypic testing, the results of the test can be provided to the customer. The results provided to a customer can inform the customer of the carrier status of an individual for a particular Mendelian disease and/or the chances that the individual's future offspring will develop Mendelian disease. In an embodiment, the customer also receives, for example, a) direct phone consultation with a genetic counselor employed by the company, or b) contact information for genetic counselors and/or other medical professionals who can provide the customer with follow up testing and consultation.

[00159] In another aspect this invention provides a method comprising marketing a genetic testing service in connection with a dating or marriage service, wherein the genetic testing service comprises: (a) predicting the probability that an offspring of the couple will have each of a plurality of traits caused by causal genetic variants, wherein the prediction is determined based on the respective genotypes of the two individuals in the couple and their respective genetically inferred ancestries; and (b) referring the couple to a genetic counselor and/or other medical professional (for example, medical geneticist or obstetrician/gynecologist) based on results of genetic testing. In another embodiment the method further comprises co-marketing a service comprising: c) predicting the probable phenotype of an offspring of a couple for at least one trait, wherein the probability is determined based on the respective phenotypes of the individuals in the couple. In another embodiment the service is an on-line service.

[00160] In another component of a business method, a complete list of all genetic counselors and/or other medical professional for follow up testing for all disease conditions can be stored in a database. The people in the database can be those who have previously consented to have their names and contact information used in connection with the genetic testing service. In an embodiment, a particular set of referrals provided to any particular customer can vary and depend on the results of the genetic testing. For example, certain genetic counselors and/or other medical professionals are relevant to display to a customer only in connection with certain Mendelian genetic diseases; such as those for which they have a specialty, for example. So certain referrals will only be shown in connection with certain results of testing.

B. Genetic Testing and Dating/Marriage Services

[00161] Matchmaking services ultimately aim to connect two people typically for the purposes of dating and/or marriage. One may think of any one user of a matchmaking service as evaluating various other users of the matchmaking service to determine the likelihood of a positive match, for example, the likelihood that any two users will be compatible long-term mates. As with any screening process of this kind, numerous factors are evaluated to determine if a match will be successful, for example attractiveness, socio-economic backgrounds, shared interests, shared alma

mater are a selected few. As described herein, genetic testing can allow for the evaluation of potential candidates in connection with a matchmaking service according to another important factor: the potential health of future offspring resulting from a given match.

[00162] In an aspect of a business method of the invention, a genetic testing service can be offered to a customer in connection with a matchmaking service, for example through a single company or a co-marketing or partnership relationship. A user of a matchmaking service can order genetic testing to determine the probability that an offspring resulting from the potential match between the user and a prospective candidate will have each of a plurality of Mendelian traits caused by causative genetic variants found to be present in the respective DNA of the user and the candidate. The user can then use this information to aid in evaluating the candidate for a potential match. The matchmaking service can be an on-line service, such as Shaadi.com, eHarmony.com and Match.com.

[00163] In an embodiment, genetic testing begins with a customer order, wherein the customer pays a fee in exchange for genetic testing. For example, a customer can be a user of a matchmaking service who is interested in evaluating another user for a suitable match. Such a customer can use genetic testing to learn whether the potential offspring of a match between said customer and a potential candidate will be at risk for developing Mendelian disease. After selecting a candidate to evaluate, a customer can pay for both the customer's and the candidate's genetic testing with the candidate's consent. The customer and the candidate can also pay separately for genetic testing. After a customer places an order, a DNA collection kit can be sent to the customer. In an embodiment of a business method of the invention, a customer can deposit a sample into the collection kit. Any sample that would be obvious to one skilled in the art can be deposited into or onto a collection kit. A sample can be any chemical compound that would be obvious to one skilled in the art -- such as bodily fluid like saliva or a breath sample containing correlated chemical compounds (see for example, http://www.popularmechanics.com/blogs/science_news/4220196.html) -- from which it is possible to identify and extract the required Raw Genotypic Information. The collection kit can then be returned to the company for sending to a genotyping lab or can be returned directly to the genotyping lab for processing. A genotyping lab, either internal within the company, contracted to work with the company, or external from the company, can isolate the customer's DNA from the provided sample. After the DNA has been isolated from the sample, a exemplary device of the invention can be used or other genotyping device can be used to test the DNA for the presence of Raw Genotypic Information. In an embodiment, the DNA does not have to be isolated from the sample to test the DNA for the presence of Raw Genotypic Information.

[00164] The Raw Genotypic Information can be processed to infer the ancestry of the customer and to confirm or deny the presence of causal genetic variants (also referred to herein as, Processed

Genotypic Information). In an embodiment, Processed Genotypic Information can be transmitted to a system of the invention or another calculation system as would be obvious to one skilled in the art to predict the probability that an offspring of the customer will have each of a plurality of traits caused by causal genetic variants found to be present in the customer's DNA. The results of the testing will then be provided to the customer and the candidate. At this point the customer and the candidate can know the probability that future offspring resulting from a match between the customer and the candidate will develop Mendelian disease.

[00165] In another aspect this invention provides a method comprising marketing a genetic testing service in connection with a dating or marriage service, wherein the genetic testing service comprises: (a) predicting the probability that an offspring of the couple will have each of a plurality of traits, wherein the prediction is determined based on the respective phenotypes and/or family histories of the two individuals in the couple. In another embodiment the method further comprises (b) referring the couple to a genetic counselor and/or other medical professional (for example, medical geneticist or obstetrician/gynecologist) based on results of genetic testing.

C. Predicting Offspring Phenotypes from Parental Phenotypes

[00166] Genetic testing can allow potential parents or mates to determine the probability that future offspring resulting from their union will develop Mendelian genetic disease. Conceptually, this forecasting is possible because of the predictable mode of inheritance of a Mendelian trait and the generally high correlation between having a variant and having the condition, notwithstanding the issues involved with variable expressivity and penetrance. For complex and/or non-Mendelian traits, it is still difficult to predict their values from genotype measurements alone. Although it is difficult to predict the values of an individual's complex traits from genotype information, it is often easy to predict the values of a child's complex traits from phenotype information on the parents. That is, for phenotypes with high heritability, a measurement of mother and father trait values is sufficient to provide excellent predictive power for child traits (http://www.amazon.com/Genetics-Analysis-Quantitative-Traits-Michael/dp/0878934812).

[00167] In another aspect this invention provides a method comprising predicting the probable phenotype of an offspring of a couple for at least 10 non-Mendelian traits, wherein the probability is determined based on the respective phenotypes of the individuals in the couple. In one embodiment the phenotypic traits include a member selected from the group consisting of height, weight, cognitive ability, academic achievement, extraversion, neuroticism, agreeableness, conscientiousness, openness, religiosity and conservatism.

[00168] In an embodiment, phenotype-phenotype predictions can be offered for free in connection with genetic testing as a means to allow prospective parents or mates to learn about the non-Mendelian traits that their future offspring may have.

[00169] A customer can provide information describing his/her phenotype as well as information describing his/her partner's phenotype (Parental Phenotypes). In an embodiment, these Parental Phenotypes can be transmitted to a system of the invention or another calculation system as would be obvious to one skilled in the art to predict the probable phenotype of an offspring of the customer and his/her partner based on the input Parental Phenotypes (Offspring Phenotype). The Offspring Phenotype can then be provided to the customer for free or for a fee.

D. Computer Systems

[00170] In another aspect this invention provides a method comprising offering a first and second service, wherein: a) the first service comprises predicting the probability that an offspring of the couple will have each of a plurality of traits caused by causal genetic variants, wherein the prediction is based on the respective genotypes of the two individuals in the couple; and b) the second service comprises predicting the probable phenotype of an offspring of the couple for a plurality of traits, wherein the probability is determined based on the respective phenotypes and/or the family history of the individuals in the couple. In one embodiment at least one prediction is further based on the respective genetically inferred ancestries of the individuals. In another embodiment the first service is offered as a service for a fee and the second service is offered as a free service.

[00171] In another aspect this invention provides a method comprising: a) taking into computer memory responses to a quiz about family history of each individual of a couple; b) taking into computer memory genetic information about each individual of the couple; c) displaying: i) at least one individual's carrier status for at least one trait determinable by the genetic information or the family history or ii) traits of offspring of the couple determinable by the family histories and/or the genetic information. In one embodiment the carrier status is displayed on a website.

[00172] In another aspect this invention provides a system comprising: a) computer readable medium configured to store family history information from each member of a couple; b) computer readable medium configured to store data comprising genetic information about each member of the couple; c) computer readable medium comprising computer code that, when executed: i) predicts each individual's carrier status with respect to traits caused by alleles identified in the genetic information; or ii) predicts probable traits of offspring of the couple determinable by the family histories and/or the genetic information; and d) a display that displays: i) carrier status of at least one member of the couple or ii) probable traits of the offspring. In one embodiment the system further comprises e) a webpage configured to accept an offer to purchase a DNA test kit. In another embodiment the display is electronic, for example, a webpage. In another embodiment the system further comprises e) a display that displays referrals to a genetic counselor and/or other medical professional (for example, medical geneticists or obstetrician/gynecologist) based on the genetic information.

[00173] The internet and the world wide web offer access to and distribution of information. As such, in an embodiment, a website can be particularly suited to efficiently providing various functionality for allowing customers to purchase genetic testing and receive the results of genetic testing. The system typically will include a server on which the website resides. Users use an interface connected to the server, such as a computer monitor or a telephone screen, to interact with the website by clicking or rolling over links that pop up information or direct the user to another webpage. Websites typically are interactive, allowing the user to input information or a query and obtain a response on the interface.

[00174] In an aspect of a system and business method, a website can allow a customer to purchase, manage, and view the results of genetic testing as well as to learn more generally about the probability that potential offspring will develop Mendelian disease. For example, a customer can be a couple of prospective parents who seek to learn whether their offspring will be at risk for developing Mendelian disease.

[00175] After a customer creates an account with the company on the website, the customer can first choose to take advantage of a free service that predicts the probability that offspring of the customer will have a trait caused by causative genetic variants. In another embodiment, the previously mentioned service is provided for a fee. The service can be offered to the customer in the form an online quiz that is promoted and accessible via the website. The quiz can also be a quiz that is ordered and physical mailed or emailed to a customer. A quiz can ask the customer a series of questions related to family history and ethnicity that a customer can answer. Once the responses are given by a customer, the responses can be sent to a server for storage and processing. In an example, computer code executes on the server to compute (i) the carrier status of the customer, if any; and (ii) the probability that an offspring of the customer will develop each of a plurality of Mendelian traits, based on the responses provided by the customer to the quiz. The results of this operation are then sent to a server for storage. The server can electronically transmit this information to a user interface, such as the website or an email, for display to the customer. In a particular embodiment, the results of the quiz are presented to the customer in the results section of the website. The results of the quiz comprise (i) the carrier status of the customer, if any; and (ii) a listing of the probability that an offspring of the customer will develop each of a plurality of Mendelian traits.

[00176] The customer can then be presented with the offer to purchase genetic testing to determine (i) the carrier status of the customer; and (ii) the probability that an offspring of the customer will develop each of a plurality of Mendelian traits, based on the causative genetic variants to be found present in the customer's DNA.

[00177] If the customer chooses to purchase genetic testing, then the customer may pay a fee, for example through an online credit card transaction, in exchange for genetic testing, direct phone consultation with a genetic counselor on the company's staff and/or referrals to genetic counselors

and/or other relevant medical professionals. The genetic testing and referrals can be paid for by a fee at the point of purchase or can be included in an initial user registration fee. In an embodiment, the services are free and revenue is generated by the company by advertising other products in conjunction with a particular product. For example, after a customer places an order online, the order is sent to a server for processing. Once payment has been verified, the order processing server can send an electronic notification to a shipping vendor to mail a DNA collection kit to the customer. In an embodiment, the DNA collection kit is separate from the genetic testing service or the user or customer already has or obtains the DNA collection kit from another source. Notifications can also periodically be sent electronically to the customer comprising order confirmation and updates on order and shipping status. Once the customer receives the collection kit, the customer can use it to obtain Raw Genotypic Information (as defined below). In an embodiment of a business method of the invention, a customer can deposit a sample into the collection kit. Any sample that would be obvious to one skilled in the art can be deposited into or onto a collection kit. A sample can be any chemical compound that would be obvious to one skilled in the art, such as bodily fluid like saliva or a breath sample containing correlated chemical compounds (see for example, http://www.popularmechanics.com/blogs/science_news/4220196.html), from which it is possible to identify and extract the required Raw Genotypic Information (as defined below). The collection kit can then be returned to the company for sending to a genotyping lab or can be returned directly to the genotyping lab for processing. A genotyping lab, either internal within the company, contracted to work with the company, or external from the company, can isolate the customer's DNA from the provided sample. After the DNA has been isolated from the sample, a exemplary device of the invention can be used or other genotyping device can be used to test the DNA for the presence of (i) ancestry informative markers and (ii) causal genetic variants (the combination of (i) and (ii) are also referred to herein as, Raw Genotypic Information). In an embodiment, the DNA does not have to be isolated from the sample to test the DNA for the presence of Raw Genotypic Information.

[00178] Raw Genotypic Information can be sent electronically to a server for storage and processing. Computer code on the server can execute on the Raw Genotypic Information to infer the ancestry of the customer and to confirm the presence of causal genetic variants, if any. The Processed Genotypic Information can then be electronically sent to a server, where computer code on the server can execute on the Processed Genotypic Information to predict the probability that an offspring of the customer will have each of a plurality of traits caused by causal genetic variants found to be present in the customer's Processed Genotypic Information. Results can then be electronically transmitted to a server for storage.

[00179] In an example, a notification can be sent to the customer to alert the customer to the availability of the results. The notification can be electronic, such as for example without limitation,

a text message, an email, or other data packet; or the notification can be non-electronic, such as for example without limitation, a phone call from a genetic counselor or printed communication such as a report sent through the mail. The results provided to a customer can inform the customer of the carrier status of the customer for a particular Mendelian disease and/or the chances that the customer's future offspring will develop Mendelian disease. After the customer has received results and referrals, the customer's order can be considered fulfilled, and results and referrals can remain accessible to the customer through an online website account. The customer can then choose to further pursue a referral offline if the customer so desires but outside of the purview of the website.

[00180] While preferred embodiments of the present invention 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 invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

EXAMPLE 1

[00181] A method is provided herein of validating or quality controlling a lot of manufactured devices configured to test for causal genetic variants for more than 85 rare genetic diseases. Genomic DNA samples are not publicly available for many rare genotypes, hence it can be difficult to get a proper number samples to test the device. It may also be desirable to avoid dilution of the test with wild type alleles and to avoid requiring a large number of devices to test the manufactured lot. In this example, DNA samples representing the causal genetic variants were generated by synthesis. Double stranded DNA were synthesized de novo and cloned into standard high-copy plasmid vectors. These samples represent 662 disease-causing variants and 139 diseases. This approach has been described in a number of recent reports related to CF carrier testing (Berwouts et al., 2008; Christensen et al., 2007). All plasmids were sequence verified to contain zero errors by double stranded, primer walk, BigDye methodology.

[00182] The nucleic acid segments of the causal genetic variants were organized into blocks to avoid interference during detection of any of the causal genetic variants. Variants in the same block were placed in different pools. The total number of pools needed is equal to the size of the largest block. Using a maximal probe overlap of 50% sequence similarity, the largest block is size 34 in this example.

[00183] The counts by block sizes are: 320 blocks of size 1 (singleton variants); 84 of size 2; 21 of size 3; 12 of size 4; 2 of size 5; 7 of size 6; 3 of size 7; 1 of size 8; 1 of size 10; 3 of size 12; and 1 of size 34.

[00184] Inserts were divided into sets that can be combined into one or more plasmids for synthesis. The first set is the group of 320 inserts with no overlapping regions. The next 2 sets are the groups of inserts from blocks of size two (each with 84 inserts). The next is 3 groups of 21 inserts each for block size 3. The next is 4 groups of 12 inserts each, and so on such that there are n groups generated by the blocks of size n. In total, 92 groups of inserts were created for plasmid synthesis.

EXAMPLE 2

[00185] Exemplary processes of delivering a probability that a user is a carrier of rare genetic disease are demonstrated in Figures 20-23. Figures 20-21 illustrate pipelines for order fulfillment for web and medical customers respectively. An order can be placed by a physician or a consumer. An order can be placed for a single test or for a couple or family. The order can be accepted through a web site. The ordering system can accept contact information, demographic details and billing information. Contact information can include, without limitation, name, address, telephone number, and email addresses. Demographic information can include, without limitation, sex, date of birth, and self-reported ethnicity. An order confirmation notification can be sent using the provided contact information. Acceptable orders are added to a database, and the states of these orders can be subsequently maintained by a state machine.

[00186] A sample collection kit is then sent to the user. A sample is collected that is any human tissue or fluid. The sample can also be isolated DNA from a human. Examples of samples useful for this example include, but are not limited to: saliva, blood, urine, buccal cells, amniotic fluid, cell scrapings, and cell culture. The sample is then genotyped using a device described herein. Phenotype solicitation, for example, retrieving self-identification of phenotypic traits of a user, can be performed in parallel with sample processing.

[00187] Sample collection can be performed at home, in a physician's office, or at a specialized collection site. Sample collection and return can be tracked by advancing the state of the order-tracking state machine. Samples received by the accessioning facility can be registered in the database system by advancing their state in the state machine. After acceptance at the accessioning facility, samples can be delivered to the genotyping facility. The genotyping facility can return raw genome data to a secure data storage server by secure file transfer protocols. File upload can trigger an advance of the state machine. This advance can trigger a server configured to perform genotype calling to retrieve the raw genome data from the data storage server as well as any phenotype data associated with the order. The genotyping algorithm can produce a fully probabilistic genotype call.

[00188] Figures 22-23 illustrate the high level sample processing pipeline and detailed computational pipeline respectively. The algorithms described in the previous Figures and the processing step as shown in Figure 22 outlined in Figure 23. Batches of chips are received and measured for quality control purposes (Batch passes QC). Information such as family history, gender, or self-reported ancestry is used to serve as an independent check on calling for each sample

(Phenotype data retrieved for batch samples). In parallel with this process, a report with child predictions is constantly updated. First pre-test risk calculations are delivered, based on phenotype (such as family history and other answers to an online questionnaire). Once a genotype sample is received and processed, post-test calculations are given. The report is then generated and sent to the final stages of the pipeline, for laboratory staff and physician approval as shown in Figure 22.

[00189] Quality control metrics can be generated from the calling process. An example quality control metric is the percentage of probabilistic genotype calls in which at least one genotype has a posterior probability greater than a threshold value. A batch of samples is processed together. When processed as a batch, the individual probabilistic genotype calls can be used to generate batch-level quality control statistics. Probabilistic genotype calls can be stored in a database. Successful genotype calling can trigger an advance of the order state. For a couple or family order, the state machine can hold for completion of the entire order, else single orders can be passed to the next state. If phenotype data is required for risk calculation, then the state machine can delay processing until all phenotype data is collected. The state machine can also trigger a notification to the patient that phenotype data is required. If all genotype and phenotype data are ready, then the state machine can advance, trigger the risk calculation server to perform risk calculation. The results of risk calculation can be serialized and transferred to a results reporting system. This is a machine-readable format of the results. The state machine can advance the order when the transfer is completed. The results reporting server can combine the probabilistic risk calculation with appropriate text and formatting to generate a human-readable report. This human readable report can be further formatted for display on a website. This human readable report can be formatted for other media such as PDF files for printing. The final results reports can be automatically released using an autoverification system. A human can review the reports for release. In one instantiation the reviewers are a clinical laboratory scientist and a physician. The results are accessed via a web portal which links to a view of the results and a summary of the quality control metrics. Acceptance of the report by the clinical laboratory scientist releases the results to the physician. The physician can review the results in a similar portal and approve the final release of the results.

EXAMPLE 3

[00190] Figure 24 illustrates exemplary the input and output steps for report generation for two hypothetical parents: Mama Hen and Papa Hen. A child prediction is produced that incorporates mother and father genotypes, mother and father phenotypes, and relative genotypes and phenotypes. Any or all of these variables can be missing values, with defaults initialized from demographically similar individuals (and if this is not known, from the world population). The resulting child prediction may include not only disease risk, but also other variables such as height and weight. Different variables in the child prediction will use different weights of genotype and phenotype.

CLAIMS

WHAT IS CLAIMED IS:

1. A method comprising:

a) testing a subset of biological relatives of a child or potential child to determine a presence of a plurality of causal genetic variants corresponding to at least one rare genetic disease and a presence of at least one ancestry informative marker (AIM); and

b) predicting a probability of a phenotype of the child or potential child from the subset of biological relatives with respect to the at least one rare genetic disease based at least in part on the presence of the plurality of causal genetic variants and the presence of the at least one AIM.

2. The method of claim 1, wherein said predicting comprises performing a fully probabilistic analysis on data collected on said casual genetic variants.

3. The method of claim 1, wherein predicting in step b) is further based on phenotypic information about at least one member of the subset of biological relatives.

4. The method of claim 1 further comprising providing genetic counseling services to at least one member of the subset of biological relatives.

5. The method of claim 1 further comprising delivering the probability of the phenotype of the child to a physician referral service.

6. A computer readable medium comprising:

logic configured to predict a probability of a phenotype of a child or potential child from a subset of biological relatives with respect to at least one rare genetic disease based at least in part on the results of a test of the subset of biological relatives for the presence of a plurality of causal genetic variants and at least one ancestry informative marker (AIM).

7. The computer readable medium of claim 6, wherein the logic performs a fully probabilistic analysis.

8. The computer readable medium of claim 6, wherein the computer readable medium provides an output in the form of a report detailing the presence of the at least on AIM and the presence of any of the plurality of causal genetic variants in any member of the subset of biological relatives.

9. A computer readable medium comprising:

logic configured to perform a fully probabilistic analysis on data corresponding to a plurality of causal genetic variants from a male and a female to predict a probability of a phenotype of a child, wherein the male and the female are potential parents of the child.

10. The computer readable medium of claim 9, wherein the causal genetic variants from each of the male and female comprise one or more ancestry informative markers (AIMs), one or more causal genetic variants corresponding to a rare genetic disease, one or more causal genetic variants corresponding to a personality trait, or both.

11. The computer readable medium of claim 7 or 9, wherein the fully probabilistic analysis incorporates a plurality of sources of statistical uncertainty in the probability.

12. The computer readable medium of claim 6 or 9 further comprising logic for receiving input from a phenotype battery and assigning a weighting function to the plurality of causal genetic variants based on said input.

13. The computer readable medium of claim 12, wherein the input from the phenotype battery comprises: height, weight, and family disease history.

14. The computer readable medium of claim 6 or 9, wherein the computer readable medium provides an output in the form of a report detailing a probability distribution over child risks or phenotypes.

15. A system for predicting a child phenotype comprising:

a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM), wherein the device is in contact with a sample from a biological relative of a child or potential child;

a reader configured to read data from the devices; and

computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants and data from the at least one AIM to predict a probability of a phenotype of the child with respect to the at least one rare genetic disease.

16. The system of claim 15, wherein the biological relative is a prospective mother or prospective father of the child or potential child.

17. The system of claim 15, wherein the nucleic acid detection device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants and the at least one AIM.

18. A system for predicting a child phenotype comprising:

a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to more than 85 rare genetic diseases, wherein the device is in contact with a sample from a biological relative of the child or potential child;

a reader configured to read data from the devices; and

computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants to predict a probability of a phenotype of the child or potential child with respect to the more than 85 rare genetic diseases.

19. The system of claim 18, wherein the biological relative is a prospective mother or prospective father of the child or potential child.

20. The system of claim 18, wherein the nucleic acid detection device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants corresponding to more than 85 rare genetic diseases.

21. The system of claim 20, wherein the nucleic acid detection device further comprises a plurality of nucleic acid probes that selectively bind to at least one ancestry informative marker (AIM).

22. The system of claim 21, wherein the computer readable instructions when executed utilize the data from the reader corresponding to the at least one AIM to predict the probability of a phenotype of the child.

23. A system for indicating if a subject is a carrier of a rare genetic disease comprising:

a reader configured to read data from a nucleic acid detection device configured to detect a plurality of causal genetic variants corresponding to at least one rare genetic disease and at least one ancestry informative marker (AIM); and

computer readable instructions, wherein the instructions when executed utilize the data from the reader corresponding to the plurality of causal genetic variants and the at least one ancestry informative marker to predict a plurality of probabilities of the subject being a carrier for each of the plurality of causal genetic variants.

24. A method comprising:

receiving a sample from a user;

testing the sample with a nucleic acid detection device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM);

calculating a plurality of probabilities for the possible user genotypes at each location of the plurality of causal genetic variants based on results from the testing step relating to the plurality of causal genetic variants and the at least one AIM; and

delivering to the user the plurality of probabilities corresponding to the user being a carrier.

25. The method of claim 24 further comprising:

receiving a sample from a second user;

testing the sample from the second user with a device configured to test for a plurality of causal genetic variants of rare genetic diseases and at least one ancestry informative marker (AIM);

calculating a probability of a child phenotype corresponding to the rare genetic diseases based on results from testing the user and the second user; and

delivering the probability of the child phenotype to at least one of the user and the second user.

26. The method of claim 25 further comprising providing genetic counseling service to at least one of the user and the second user.

27. The method of claim 25, wherein the method is carried out as part of a child phenotype prediction service.

28. The method of claim 25 further comprising obtaining phenotypic information from the user; and using the phenotypic information from the user in the calculating steps.

29. The method of claim 25 further comprising obtaining family history from the user; and using the family history from the user in the calculating steps.

30. A nucleic acid detection device configured to test a sample for a plurality of causal genetic variants corresponding to at least one rare genetic disease and one or more ancestry informative markers (AIMs).

31. The device of claim 30, wherein the device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants and the one or more AIMs.

32. The device of claim 30, wherein the device comprises a bead array that selectively binds to the plurality of causal genetic variants and the one or more AIMs.

33. A nucleic acid detection device configured to test a sample for a plurality of causal genetic variants corresponding more than 85 rare genetic diseases.

34. The device of claim 33, wherein the device is further configured to test a sample for at least one ancestry informative marker (AIM).

35. The device of claim 33, wherein the device comprises a plurality of nucleic acid probes that selectively bind to the plurality of causal genetic variants corresponding to more than 85 rare genetic diseases.

36. The device of claim 35, wherein the device further comprises a plurality of nucleic acid probes that selectively bind to the at least one ancestry informative marker (AIM).

37. The device of claim 33, wherein the device comprises a bead array that selectively binds to the plurality of causal genetic variants corresponding to more than 85 rare genetic diseases.

38. The device of claim 37, wherein the device further comprises a bead array that selectively binds to the at least one ancestry informative marker (AIM).

39. The device of claim 33, wherein the device comprises a resequencing assay to detect the plurality of causal genetic variants corresponding to the more than 85 rare genetic diseases.

40. The device of claim 39, wherein the device further comprises a resequencing assay to detect at least one ancestry informative marker (AIM).

41. The device of claim 30, 34, 36, 38, or 40, wherein the at least one AIM is not a causal genetic variant.

42. The device of claim 30, 34, 36, 38, or 40, wherein at least two of the rare genetic diseases occur at frequencies that differ by at least 10-fold in at least two distinct populations, wherein the at least two distinct populations are differentiated by the at least one AIM.

43. A method comprising:

marketing a genetic testing service comprising predicting a probability of a child phenotype from a subset of biological relatives of the child or potential child, wherein the prediction is based at least in part on the presence of a plurality of causal genetic variants in

each of the subset of biological relatives and based at least in part on the inferred ancestries of each of the subset of biological relatives; and

delivering a probability of the child phenotype for a fee.

44. The method of claim 43, wherein the marketing is conducted in connection with a dating or marriage service.

45. The method of claim 43 further comprising referring at least one member of the subset of biological relatives to a physician.

46. The method of claim 43, wherein the inferred ancestries are inferred by a test for at least one ancestry informative marker (AIM).

47. A set of nucleic acid pools for validating a nucleic acid sequence detection device comprising a set of causal genetic variant probes, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a different subset of the set of causal genetic variant probes.

48. The set of claim 47, wherein a first pool of the set comprises a first nucleic acid segment that interferes during detection with a second nucleic acid segment of a second pool of the set, and wherein the first pool does not comprise the second nucleic acid segment and the second pool does not comprise the first nucleic acid segment.

49. The set of claim 47, wherein the nucleic acid segments of each pool are single stranded nucleic acid molecules.

50. The set of claim 47, wherein the nucleic acid segments comprise one or more plasmids.

51. A method of validating a lot of manufactured nucleic acid sequence detection devices comprising:

contacting each of a plurality of nucleic acid sequence detection devices from the lot with a different nucleic acid pool, wherein each nucleic acid pool comprises a plurality of nucleic acid segments that selectively bind a plurality of causal genetic variant probes on said detection devices, and wherein each nucleic acid pool binds a different set of the plurality of causal genetic variant probes; and

detecting presence or absence of the plurality of causal genetic variant probes on the plurality of nucleic acid detection devices,

wherein the lot of manufactured devices is validated if all of the plurality of causal genetic variant probes are present on the plurality of nucleic acid detection devices.

52. The method of claim 51 further comprising delivering the lot of manufactured devices when the devices are validated.

53. The method of claim 51, wherein the lot of manufactured devices is rejected if not all of the plurality of causal genetic variant probes are present.