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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
9 August 2012 (09.08.2012)

WIPO | PCT

(10) International Publication Number
WO 2012/106267 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *G06F 19/22* (2011.01)
G06F 17/00 (2006.01) *G06F 19/20* (2011.01)

(21) International Application Number:
PCT/US2012/023195

(22) International Filing Date:
30 January 2012 (30.01.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/437,788 31 January 2011 (31.01.2011) US

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **LUO, Wen** [US/US]; 5003 Ruette De Mer, San Diego, CA 92130 (US).

(74) Agents: **WANG, Kun** et al.; Morrison & Foerster LLP, 12531 High Bluff Drive, Suite 100, San Diego, CA 92130-2040 (US).

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

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

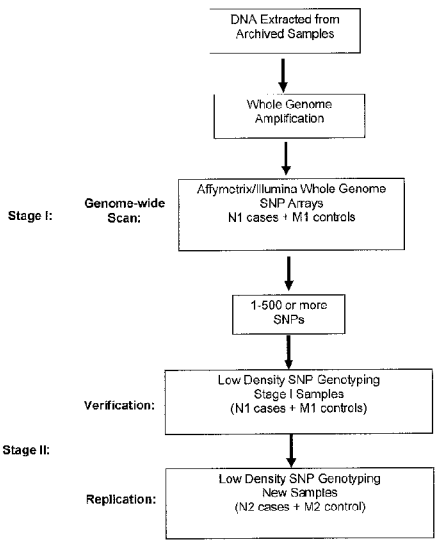
Published:

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

WO 2012/106267 A1

(54) **Title**: METHOD FOR DISCOVERING PHARMACOGENOMIC BIOMARKERS

Figure 1



(57) **Abstract**: The present invention relates to a method of discovering pharmacogenomic biomarkers that are correlated with varied individual responses (efficacy, adverse effect, and other end points) to therapeutic agents. The present invention provides a mean to utilize archived clinical samples to perform genome-wide association study in order to identify novel pharmacogenomic biomarkers. The newly discovered biomarkers can then be developed into companion diagnostic tests which can help to predict drug responses and apply drugs only to those who will be benefited, or exclude those who might have adverse effects, by the treatment.

METHOD FOR DISCOVERING PHARMACOGENOMIC BIOMARKERS

Related Patent Applications

[0001] This patent application claims the benefit of U.S. Provisional Patent Application No. 61/437,788, filed January 31, 2011, which is hereby incorporated by reference in its entirety, including all drawings and cited publications and documents.

Technical Field

[0002] The present invention relates to a method of discovering pharmacogenomic biomarkers which can be developed into companion diagnostic tests to predict varied individual responses (efficacy, adverse effect, or other) to therapeutic agents.

Background Art

[0003] Pharmaceutical industry has been operating under the paradigm of "one drug fits all" for many decades. However, only a few drugs offer universal efficacy in all patient populations, and some even cause serious adverse effects in certain patient groups. These obstacles, in addition to sky rocketing R&D expenses and the tougher FDA review standards, have resulted in many of newly developed drugs failing to reach the market. Therefore, identifying pharmacogenomic biomarkers, which can predict potential responders for a drug, would be the ideal solution to unlock the hidden value for many of these otherwise failed drugs.

[0004] In the meantime, with the completion of HapMap project and rapid advances in microarray technology, genome-wide scan of genetic polymorphisms has become routine tasks, and hundreds of genome wide association studies (GWAS) have been successfully conducted and led to the discovery of a large number of known and novel genetic variants associated with common diseases such as cardiovascular disease, diabetes, etc. This technology is also offering powerful tools for identifying genetic polymorphisms associated with drug responses. However, Guessous et al. reviewed about three hundred GWAS studies, among which only 12 are pharmacogenomic studies (*Genome Med* (2009) 1:46). Of those 12 studies, only two of them are GWAS used in clinical trials. Obviously, the pharmaceutical companies embraced this new technology at a very slow pace comparing to academic

researchers. One of the most common causes is that pharmaceutical companies only realize the need to perform such study after the clinical trials are over, and it is often too late at this point since the appropriate samples have not been collected during the trial. This scenario will likely remain unchanged for the foreseeable future until the industry widely adopts the concept of pharmacogenomics and begins to incorporate biomarker design prior to starting clinical trials.

[0005] Despite increased spending and advances in technology, a large number of newly developed drugs would fail in Phase III clinical trials, mostly due to lack of significant efficacy in overall patient population or unsatisfied safety profile. However, many of the failed drugs could still benefit a subset of patient population or only cause adverse effects in small number of patients. Comparing to the "old" candidate gene approach, GWAS is hypothesis free and doesn't require prior knowledge of the mechanisms involved in the studied clinical end points. The advances in array technology have also made it possible to genotype 1 million or more SNPs, which could cover the entire human genome, on a single array. Therefore, GWAS is an ideal option to search drug specific pharmacogenomic biomarkers.

[0006] The commonly used sources of genomic DNA in GWAS are DNA rich tissues/cells such as whole blood, which can yield adequate amount of high quality genomic DNA. However, most of the pharmaceutical companies have not incorporated pharmacogenomic study into their clinical trials. Thus there are usually no dedicated samples collected for GWAS study. In many cases, there could be human specimen collected for other purposes, such as biopsies (for pathology) and plasma samples (for pharmacokinetic study). Those leftover clinical samples could be potentially used to obtain genomic DNA. A few reports have shown that genomic DNA could be extracted from plasma samples and lead to successful genotyping on a small number of SNPs (Sjoholm et al., *Cancer Epidemiol Biomarkers Prev* (2005) 14:251; Lu et al., *Biotechniques* (2005) 39:511; Park et al., *Clin Chem* (2005) 51:1520; Bergen et al., *Hum Mutat* (2005) 26:262). However, the quantity and quality of the genomic DNA from archived clinical samples, which are often being processed and stored inappropriately, are far from optimal, especially for usage in GWAS. In fact, among the few who have published reports on using genomic DNA extracted from plasma samples on high density SNP arrays, the authors suggested that the DNA quality was so "poor" from these samples and they might not lead to a successfully GWAS (Croft et al., *J*

Mol Diagn (2008) 10:249; Bucasas et al., *BMC Genet* (2009) 10:85). The only GWAS successfully conducted used dried blood spot samples (Hollegaard et al., *BMC Genomics* (2009) 10:297-302) or patient blood (Singer et al., *Nat Genet* (2010) 42:711-714). Thus, to our knowledge, no one has ever attempted to perform a successful GWAS using genomic DNA from archived clinical samples.

Summary of the Invention

[0007] While most of the published GWAS have focused on discovering the causative genetic variants of common diseases using high quality DNA from dedicated samples such as whole blood in well-designed studies, the application of this powerful technology has seldom been incorporated into clinical trials for new drugs and therefore has very limited successes in pharmacogenomic study. The current consensus is that a successful GWAS can only be performed using abundant high quality genomic DNA, and suboptimal genomic DNA and/or whole genome amplified DNA are strongly discouraged or abandoned in GWAS. The present invention describes a method of discovery of pharmacogenomic biomarkers by GWAS to predict drug response utilizing suboptimal genomic DNA, e.g., genomic DNA extracted from archived clinical samples.

[0008] In one aspect, the present invention provides a method to identify one or more pharmacogenomic biomarkers, which method comprises: a) isolating DNA from archived clinical samples of at least two patients exhibiting different values in a relevant phenotype; b) amplifying said isolated DNA; c) obtaining high-density genotyping data of said amplified DNA; and d) performing association analysis based on said genotyping data and said different values in said relevant phenotype, wherein said pharmacogenomic biomarker(s) are identified.

[0009] In some embodiments, the archived clinical samples may be selected from the group consisting of plasma samples, serum samples, dried blood spots, urine samples, tissue samples, tumor cells and buccal swabs. In some embodiments, the archived clinical samples may be plasma samples. In some embodiments, the archived clinical samples may be from about 2-1,000 or more patients.

[0010] In some embodiments, the isolated DNA may be suboptimal genomic DNA. In some embodiments, the amplification may be whole-genome amplification (WGA), and the resulting DNA may be whole-genome amplified DNA (wgaDNA). In some embodiments,

the high-density genotyping may be whole-genome genotyping. In some embodiments, the high-density genotyping may be conducted by using single nucleotide polymorphisms (SNPs). In some embodiments, about 1,000-5,000,000 or more, preferably about 1,000,000, SNPs may be used for the high-density genotyping. In some embodiments, the high-density genotyping may be array-based.

[0011] In some embodiments, the genotyping data may be obtained by using a genome-wide genotype calling algorithm. In some embodiments, the method may further comprise: e) adjusting the call rate cut-off value of the genome-wide genotype calling algorithm. In some embodiments, step d) and step e) may be repeated multiple times to include and/or exclude samples and to optimize the genome-wide genotype calling algorithm. In some embodiments, an optimal inclusion criterion for the genome-wide genotype calling algorithm may be identified. In some embodiments, the genotype calls may be made by using a call rate cut-off value that is lower than a typical call rate cut-off used for whole-genome genotyping of high quality genomic DNA, wherein the call rate cut-off value used may be about 50%, 60%, 70%, 80%, 90% or 95%. In some embodiments, the genotype calls may be generated with the Affymetrix Genotyping Console™ software. In some embodiments, the genotype calls may be generated using the BRLMM algorithm. In some embodiments, the genotype calls may be made using an imputation algorithm, wherein the HapMap may be used for the imputation algorithm.

[0012] In some embodiments, the association analysis may be a GWAS. In some embodiments, the association analysis may be performed by calculating an associated p-value of each SNP with the relevant phenotype. In some embodiments, the calculation may be based on an allele frequency and/or genotype-based test. In some embodiments, the relevant phenotype may be a categorical trait, a quantitative trait or another relevant phenotype.

[0013] In some embodiments, the method may further comprise performing association analysis based on additional genotyping data using the identified pharmacogenomic biomarkers. In some embodiments, about 1-500 or more of the identified pharmacogenomic biomarkers may be used for the additional genotyping. In some embodiments, some or all of the archived clinical samples from step a) and/or additional clinical samples may be used for the additional genotyping. In some embodiments, additional genotyping data may be obtained by using a verification genotype calling algorithm. In some embodiments, the method may further comprise adjusting the call rate cut-off value of the verification genotype

calling algorithm. In some embodiments, genotyping and adjusting the call rate cut-off may be repeated multiple times to include and/or exclude samples. In some embodiments, an optimal inclusion criterion may be identified for the verification genotype calling algorithm. In some embodiments, the method may further comprise comparing the additional genotyping data obtained by using the verification genotype calling algorithm to the genotyping data obtained by using the genome-wide genotype calling algorithm. In some embodiments, a subset of the pharmacogenomic biomarkers from step d) may be identified.

[0014] In some embodiments, the method may be used for retrospective study of archived clinical samples from a previously conducted clinical trial. In some embodiments, the method may be used for *de novo* identification of a pharmacogenomic biomarker.

[0015] In another aspect, provided herein is a pharmacogenomic biomarker, or a group of pharmacogenomic biomarkers, identified by the method disclosed herein, wherein the biomarker may be one or more SNPs. In some embodiments, the pharmacogenomic biomarker may be used to identify one or more additional pharmacogenomic biomarkers. In some embodiments, the pharmacogenomic biomarker may be used to develop a companion diagnostic test.

[0016] In a further aspect, provided herein is a companion diagnostic test using the pharmacogenomic biomarkers identified by the method disclosed herein. Also provided herein is a method of prognosticating responsiveness of a subject to a treatment using the companion diagnostic test disclosed herein. Further provided herein is a method of identifying a novel drug target using the pharmacogenomic biomarkers identified by the method disclosed herein. The methods of the present invention are useful for clinicians to identify patients for treatment, aid in patient selection during the course of development of therapy, predict likelihood of success when treating an individual patient with a particular treatment regimen, assess and monitor disease progression, monitor treatment efficacy, and determine prognosis for individual patients. Any of these embodiments are included in this invention.

[0017] In an additional aspect, provided herein is a kit comprising a reagent for assessing the pharmacogenomic biomarker, or the group of pharmacogenomic biomarkers, identified by the method disclosed herein. In some embodiments, the kit may further comprise instructions for using the pharmacogenomic biomarker to conduct a companion diagnostic test.

[0018] In yet another aspect, provided herein is a genotyping method using suboptimal genomic DNA samples, which comprises the steps of: a) receiving sequence information of said suboptimal genomic DNA samples; b) optimizing an inclusion criterion based on said sequence information; and c) calculating genotypes based on said sequence information and said optimized inclusion criterion. In some embodiments, the optimization may be repeated multiple times to include and/or exclude samples. In some embodiments, an optimal inclusion criterion may be identified. In some embodiments, the genotyping data may be obtained by using a genome-wide genotype calling algorithm and/or a verification genotype calling algorithm. In some embodiments, the inclusion criterion may be the call rate cut-off value of the genotype calling algorithm. In some embodiments, the genotype calls may be made by using a call rate cut-off that is lower than a typical call rate cut-off used for whole-genome genotyping of high quality or optimal genomic DNA, wherein the call rate cut-off value used may be about 50%, 60%, 70%, 80%, 90% or 95%. In some embodiments, the genotyping data may be obtained by using multiple genotyping platforms. In some embodiments, the genotyping data from multiple genotyping platforms may be compared for the optimization.

[0019] Further provided is a method to perform association analysis using the genotyping method using suboptimal genomic DNA samples, which method comprises optimizing an inclusion criterion. In some embodiments, the association analysis may be repeated multiple times for the optimization. Also provided herein is a computer readable medium comprising a plurality of instructions for a genotyping method using suboptimal genomic DNA samples, which comprises the steps of: a) receiving sequence information of said suboptimal genomic DNA samples; b) optimizing an inclusion criterion based on said sequence information; and c) calculating genotypes based on said sequence information and said optimized inclusion criterion.

[0020] In still another aspect, provided herein is a method to conduct GWAS using suboptimal genomic DNA. In some embodiments, the suboptimal genomic DNA may be from archived samples. In some embodiments, the suboptimal genomic DNA may be from plasma samples. In some embodiments, the suboptimal genomic DNA may be amplified. In some embodiments, multiple genotyping platforms may be used. In some embodiments, the same or different samples may be used for the multiple genotyping platforms. In some embodiments, the method may further use a sample providing high-quality genomic DNA.

699572000140

Brief Description of the Drawings

[0021] FIG. 1 shows the flow chart of an exemplary pharmacogenomic biomarker discovery method using GWAS. Genomic DNA is extracted from archived clinical samples, and amplified using WGA. In Stage I, the discovery phase, N1 (a number from 1 to 1,000 or more) samples from the case group and M1 (a number from 1 to 1,000 or more) samples from the control group are genotyped using the Affymetrix and/or Illumina whole genome SNP arrays. The association of each SNP to case-control status (i.e., responders vs. non-responders) is calculated based on an allele frequency and/or genotype-based test. Subsequently, 1 to 500 or more most significantly associated SNPs are selected for the Stage II study. In Stage II, the same samples used in Stage I are genotyped using a low density SNP genotyping platform (a distinctive genotyping technology from the ones used in Stage I) to verify the results obtained in Stage I. Additional new samples, N2 (a number from 1 to 1,000 or more) samples from the case group and M2 (a number from 1 to 1,000 or more) samples from the control group are genotyped to replicate the finding.

[0022] FIG. 2 shows the flow chart of an exemplary genome-wide discovery stage data analysis.

[0023] FIG. 3 shows the flow chart of an exemplary verification stage data analysis.

Detailed Description of the Invention

[0024] The present invention provides novel approaches to overcome the potentially low quantity of genotyping results using genomic DNA isolated from archived clinical samples by applying more than one genotyping technology or platform.

A. General Techniques

[0025] The practice of the present invention will employ, unless otherwise indicated, conventional techniques of molecular biology (including recombinant techniques), microbiology, cell biology, biochemistry, and immunology, which are within the skill of the art. Such techniques are explained fully in the literature, such as, "Molecular Cloning: A Laboratory Manual", second edition (Sambrook et al., 1989); "Oligonucleotide Synthesis" (M. J. Gait, ed., 1984); "Animal Cell Culture" (R. I. Freshney, ed., 1987); "Methods in Enzymology" (Academic Press, Inc.); "Current Protocols in Molecular Biology" (F. M.

Ausubel et al., eds., 1987, and periodic updates); "PCR: The Polymerase Chain Reaction", (Mullis et al., eds., 1994).

B. Definitions

[0026] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as is commonly understood by one of ordinary skill in the art to which this invention belongs. All patents, applications, published applications and other publications referred to herein are incorporated by reference in their entireties. If a definition set forth in this section is contrary to or otherwise inconsistent with a definition set forth in the patents, applications, published applications and other publications that are herein incorporated by reference, the definition set forth in this section prevails over the definition that is incorporated herein by reference.

[0027] As used herein, the singular forms "a", "an", and "the" include plural references unless indicated otherwise. For example, "a" dimer includes one or more dimers.

[0028] The term "biomarker" or "marker" as used herein refers generally to a molecule, including a gene, protein, carbohydrate structure, or glycolipid, the expression of which in or on a mammalian tissue or cell or secreted can be detected by known methods (or methods disclosed herein) and is predictive or can be used to predict (or aid prediction) for a mammalian cell's or tissue's sensitivity to, and in some embodiments, to predict (or aid prediction) an individual's responsiveness to treatment regimens.

[0029] As used herein, a "pharmacogenomic biomarker" is an objective biomarker which correlates with a specific clinical drug response or susceptibility in a subject (*see, e.g.*, McLeod et al., *Eur. J. Cancer* (1999) 35:1650-1652). It may be a biochemical biomarker, a clinical sign or symptom, etc. The presence or quantity of the pharmacogenomic marker is related to the predicted response of the subject to a specific drug or class of drugs prior to administration of the drug. By assessing the presence or quantity of one or more pharmacogenomic markers in a subject, a drug therapy which is most appropriate for the subject, or which is predicted to have a greater degree of success, may be selected. For example, based on the presence or quantity of DNA, RNA, or protein for specific tumor markers in a subject, a drug or course of treatment may be selected that is optimized for the treatment of the specific tumor likely to be present in the subject. Similarly, the presence or absence of a specific sequence mutation or polymorphism may correlate with drug response.

The use of pharmacogenomic biomarkers therefore permits the application of the most appropriate treatment for each subject without having to administer the therapy.

[0030] The term "sample", as used herein, refers to a composition that is obtained or derived from a subject of interest that contains a cellular and/or other molecular entity that is to be characterized and/or identified, for example based on physical, biochemical, chemical and/or physiological characteristics. For example, the phrase "clinical sample" or "disease sample" and variations thereof refer to any sample obtained from a subject of interest that would be expected or is known to contain the cellular and/or molecular entity, such as a biomarker, that is to be characterized.

[0031] The term "tissue or cell sample" refers to a collection of similar cells obtained from a tissue of a subject or patient. The source of the tissue or cell sample may be solid tissue as from a fresh, frozen and/or preserved organ or tissue sample or biopsy or aspirate; blood or any blood constituents; bodily fluids such as cerebral spinal fluid, amniotic fluid, peritoneal fluid, or interstitial fluid; cells from any time in gestation or development of the subject. The tissue sample may also be primary or cultured cells or cell lines. Optionally, the tissue or cell sample is obtained from a disease tissue/organ. The tissue sample may contain compounds which are not naturally intermixed with the tissue in nature such as preservatives, anticoagulants, buffers, fixatives, nutrients, antibiotics, or the like.

[0032] "Plasma," or "blood plasma," as used herein, refers to the intravascular fluid part of extracellular fluid (all body fluid outside of cells). It is mostly water and contains dissolved proteins, glucose, clotting factors, mineral ions, hormones and carbon dioxide (plasma being the main medium for excretory product transportation). Blood plasma is prepared by spinning a tube of fresh blood containing an anti-coagulant in a centrifuge until the blood cells fall to the bottom of the tube. The blood plasma is then poured or drawn off. "Blood serum" is blood plasma without fibrinogen or the other clotting factors (i.e., whole blood minus both the cells and the clotting factors).

[0033] "Polynucleotide," or "nucleic acid," as used interchangeably herein, refer to polymers of nucleotides of any length, and include DNA and RNA. The nucleotides can be deoxyribonucleotides, ribonucleotides, modified nucleotides or bases, and/or their analogs, or any substrate that can be incorporated into a polymer by DNA or RNA polymerase. A polynucleotide may comprise modified nucleotides, such as methylated nucleotides and their analogs. If present, modification to the nucleotide structure may be imparted before or after

assembly of the polymer. The sequence of nucleotides may be interrupted by non-nucleotide components. A polynucleotide may be further modified after polymerization, such as by conjugation with a labeling component. Other types of modifications include, for example, "caps", substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoamidates, cabamates, etc.) and with charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), those containing pendant moieties, such as, for example, proteins (e.g., nucleases, toxins, antibodies, signal peptides, ply-L-lysine, etc.), those with intercalators (e.g., acridine, psoralen, etc.), those containing chelators (e.g., metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, those with modified linkages (e.g., alpha anomeric nucleic acids, etc.), as well as unmodified forms of the polynucleotide(s). Further, any of the hydroxyl groups ordinarily present in the sugars may be replaced, for example, by phosphonate groups, phosphate groups, protected by standard protecting groups, or activated to prepare additional linkages to additional nucleotides, or may be conjugated to solid supports. The 5' and 3' terminal OH can be phosphorylated or substituted with amines or organic capping groups moieties of from 1 to 20 carbon atoms. Other hydroxyls may also be derivatized to standard protecting groups. Polynucleotides can also contain analogous forms of ribose or deoxyribose sugars that are generally known in the art, including, for example, 2'-O-methyl-2'-O- allyl, 2'-fluoro- or 2'-azido-ribose, carbocyclic sugar analogs, α-anomeric sugars, epimeric sugars such as arabinose, xyloses or lyxoses, pyranose sugars, furanose sugars, sedoheptuloses, acyclic analogs and abasic nucleoside analogs such as methyl riboside. One or more phosphodiester linkages may be replaced by alternative linking groups. These alternative linking groups include, but are not limited to, embodiments wherein phosphate is replaced by P(O)S("thioate"), P(S)S ("dithioate"), "(O)NR 2 ("amidate"), P(O)R, P(O)OR', CO or CH 2 ("formacetal"), in which each R or R' is independently H or substituted or unsubstituted alkyl (1-20 C) optionally containing an ether (--O--) linkage, aryl, alkenyl, cycloalkyl, cycloalkenyl or araldyl. Not all linkages in a polynucleotide need be identical. The preceding description applies to all polynucleotides referred to herein, including RNA and DNA.

[0034] "Oligonucleotide," as used herein, generally refers to short, generally single stranded, generally synthetic polynucleotides that are generally, but not necessarily, less than about 200 nucleotides in length. The terms "oligonucleotide" and "polynucleotide" are not

mutually exclusive. The description above for polynucleotides is equally and fully applicable to oligonucleotides.

[0035] The term "suboptimal genomic DNA" as used herein refers to genomic DNA that is inferior in quality and/or in quantity to genomic DNA isolated from DNA rich tissues/cells such as whole blood. Suboptimal genomic DNA may be isolated from archived clinical samples such as biopsies or plasmas, which are often being processed and stored inappropriately, and are far from optimal, in quality and/or in quantity, especially for usage in GWAS. For example, a sample of suboptimal genomic DNA may not provide full coverage of the whole genome, or may contain short fragments of genomic DNA. In some embodiments, a sample of suboptimal genomic DNA subjected to high-density genotyping may have a call rate of less than about 99%, 95%, 90%, 80%, 70%, 60%, 50% or lower. Typically, for the suboptimal genomic DNA to be useful for GWAS, an amplification step is needed.

[0036] "Amplification," as used herein, generally refers to the process of producing multiple copies of a desired sequence. "Multiple copies" means at least 2 copies. A "copy" does not necessarily mean perfect sequence complementarity or identity to the template sequence. For example, copies can include nucleotide analogs such as deoxyinosine, intentional sequence alterations (such as sequence alterations introduced through a primer comprising a sequence that is hybridizable, but not complementary, to the template), and/or sequence errors that occur during amplification.

[0037] The term "array" or "microarray", as used herein refers to an ordered arrangement of hybridizable array elements, such as polynucleotide probes (e.g., oligonucleotides), or binding reagents (e.g., antibodies), on a substrate. The substrate can be a solid substrate, such as a glass or silica slide, a bead, a fiber optic binder, or a semi-solid substrate, such as a nitrocellulose membrane. The nucleotide sequences can be DNA, RNA, or any permutations thereof.

[0038] As used herein, the term "phenotype" refers to a trait which can be compared between individuals, such as presence or absence of a condition, a visually observable difference in appearance between individuals, metabolic variations, physiological variations, variations in the function of biological molecules, and the like. A phenotype can be qualitative or quantitative. An example of a phenotype is responsiveness to a treatment, such as a drug.

[0039] "Responsiveness" can be assessed using any endpoint indicating a benefit to the patient, including, without limitation, (1) inhibition, to some extent, of disease progression, including slowing down and complete arrest; (2) reduction in the number of disease episodes and/or symptoms; (3) reduction in lesional size; (4) inhibition (i.e., reduction, slowing down or complete stopping) of disease cell infiltration into adjacent peripheral organs and/or tissues; (5) inhibition (i.e., reduction, slowing down or complete stopping) of disease spread; (6) relief, to some extent, of one or more symptoms associated with the disorder; (7) increase in the length of disease-free presentation following treatment; (8) decreased mortality at a given point of time following treatment; and/or (9) lack of adverse effects following treatment. Responsiveness can also be assessed using any endpoint indicating side effect and/or toxicity to the patient.

[0040] "Treating" or "treatment" or "alleviation" refers to therapeutic treatment wherein the object is to slow down (lessen) if not cure the targeted pathologic condition or disorder or prevent recurrence of the condition. A subject is successfully "treated" if, after receiving a therapeutic amount of a therapeutic agent, the subject shows observable and/or measurable reduction in or absence of one or more signs and symptoms of the particular disease. For example, significant reduction in the number of cancer cells or absence of the cancer cells; reduction in the tumor size; inhibition (i.e., slow to some extent and preferably stop) of tumor metastasis; inhibition, to some extent, of tumor growth; increase in length of remission, and/or relief to some extent, one or more of the symptoms associated with the specific cancer; reduced morbidity and mortality, and improvement in quality of life issues. Reduction of the signs or symptoms of a disease may also be felt by the patient. Treatment can achieve a complete response, defined as disappearance of all signs of cancer, or a partial response, wherein the size of the tumor is decreased, preferably by more than 50 percent, more preferably by 75%. A patient is also considered treated if the patient experiences stable disease. In some embodiments, treatment with a therapeutic agent is effective to result in the patients being disease-free 3 months after treatment, preferably 6 months, more preferably one year, even more preferably 2 or more years post treatment. These parameters for assessing successful treatment and improvement in the disease are readily measurable by routine procedures familiar to a physician of appropriate skill in the art.

[0041] The term "prediction" or "prognosis" is used herein to refer to the likelihood that a patient will respond either favorably or unfavorably to a drug or set of drugs. In one embodiment, the prediction relates to the extent of those responses. In one embodiment, the prediction relates to whether and/or the probability that a patient will survive or improve following treatment, for example treatment with a particular therapeutic agent, and for a certain period of time without disease recurrence. The predictive methods of the invention can be used clinically to make treatment decisions by choosing the most appropriate treatment modalities for any particular patient. The predictive methods of the present invention are valuable tools in predicting if a patient is likely to respond favorably to a treatment regimen, such as a given therapeutic regimen, including for example, administration of a given therapeutic agent or combination, surgical intervention, steroid treatment, etc.

[0042] As used herein, the term "specifically binds" refers to the binding specificity of a specific binding pair. Recognition by an antibody of a particular target in the presence of other potential targets is one characteristic of such binding. Specific binding involves two different molecules wherein one of the molecules specifically binds with the second molecule through chemical or physical means. The two molecules are related in the sense that their binding with each other is such that they are capable of distinguishing their binding partner from other assay constituents having similar characteristics. The members of the binding component pair are referred to as ligand and receptor (anti-ligand), specific binding pair (SBP) member and SBP partner, and the like. A molecule may also be an SBP member for an aggregation of molecules; for example an antibody raised against an immune complex of a second antibody and its corresponding antigen may be considered to be an SBP member for the immune complex.

[0043] As used herein, the term "homologue" is used to refer to a nucleic acid which differs from a naturally occurring nucleic acid (i.e., the "prototype" or "wild-type" nucleic acid) by minor modifications to the naturally occurring nucleic acid, but which maintains the basic nucleotide structure of the naturally occurring form. Such changes include, but are not limited to: changes in one or a few nucleotides, including deletions (e.g., a truncated version of the nucleic acid) insertions and/or substitutions. A homologue can have enhanced, decreased, or substantially similar properties as compared to the naturally occurring nucleic acid. A homologue can be complementary or matched to the naturally occurring nucleic acid.

Homologues can be produced using techniques known in the art for the production of nucleic acids including, but not limited to, recombinant DNA techniques, chemical synthesis, etc.

[0044] As used herein, "complementary or matched" means that two nucleic acid sequences have at least 50% sequence identity. Preferably, the two nucleic acid sequences have at least 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98%, 99% or 100% of sequence identity. "Complementary or matched" also means that two nucleic acid sequences can hybridize under low, middle and/or high stringency condition(s).

[0045] As used herein, "substantially complementary or substantially matched" means that two nucleic acid sequences have at least 90% sequence identity. Preferably, the two nucleic acid sequences have at least 95%, 96%, 97%, 98%, 99% or 100% of sequence identity. Alternatively, "substantially complementary or substantially matched" means that two nucleic acid sequences can hybridize under high stringency condition(s).

[0046] In general, the stability of a hybrid is a function of the ion concentration and temperature. Typically, a hybridization reaction is performed under conditions of lower stringency, followed by washes of varying, but higher, stringency. Moderately stringent hybridization refers to conditions that permit a nucleic acid molecule such as a probe to bind a complementary nucleic acid molecule. The hybridized nucleic acid molecules generally have at least 60% identity, including for example at least any of 70%, 75%, 80%, 85%, 90%, or 95% identity. Moderately stringent conditions are conditions equivalent to hybridization in 50% formamide, 5x Denhardt's solution, 5x SSPE, 0.2% SDS at 42°C, followed by washing in 0.2x SSPE, 0.2% SDS, at 42°C. High stringency conditions can be provided, for example, by hybridization in 50% formamide, 5x Denhardt's solution, 5x SSPE, 0.2% SDS at 42°C, followed by washing in 0.1x SSPE, and 0.1% SDS at 65°C.

[0047] Low stringency hybridization refers to conditions equivalent to hybridization in 10% formamide, 5x Denhardt's solution, 6x SSPE, 0.2% SDS at 22°C, followed by washing in 1x SSPE, 0.2% SDS, at 37°C. Denhardt's solution contains 1% Ficoll, 1% polyvinylpyrolidone, and 1% bovine serum albumin (BSA). 20x SSPE (sodium chloride, sodium phosphate, ethylene diamide tetraacetic acid (EDTA)) contains 3M sodium chloride, 0.2M sodium phosphate, and 0.025 M (EDTA). Other suitable moderate stringency and high stringency hybridization buffers and conditions are well known to those of skill in the art.

[0048] It is understood that aspects and embodiments of the invention described herein include "consisting" and/or "consisting essentially of" aspects and embodiments.

[0049] Other objects, advantages and features of the present invention will become apparent from the following specification taken in conjunction with the accompanying drawings.

C. Methods for Genotyping Genomic DNA

[0050] The present invention provides a novel method to identify pharmacogenomic biomarkers utilizing archived clinical samples from various sources such as body fluids, tissues, blood, or components of blood such as plasma. In one aspect, the present invention provides a method to identify one or more pharmacogenomic biomarkers, which method comprises: a) isolating DNA from archived clinical samples of at least two patients exhibiting different values in a relevant phenotype; b) amplifying said isolated DNA; c) obtaining high-density genotyping data of said amplified DNA; and d) performing association analysis based on said genotyping data and said different values in said relevant phenotype, wherein said pharmacogenomic biomarker(s) are identified.

[0051] In some embodiments, the method may be used for retrospective study of archived clinical samples from a previously conducted clinical trial. In some embodiments, the method may be used for *de novo* identification of a pharmacogenomic biomarker.

[0052] At least two patients exhibiting different values in a relevant phenotype are needed for the method for identifying one or more pharmacogenomic biomarkers. Typically, significantly more patients may be needed for performing the association analysis. In some embodiments, the archived clinical samples may be from about 2, 5, 10, 20, 50, 100, 200, 500, 1,000 or more patients. Any relevant phenotype is contemplated by the present invention, such as responsiveness to a medical treatment. In some embodiments, the relevant phenotype may be a categorical trait, a quantitative trait or another relevant phenotype. Typically, the patients may be enrolled in a clinical trial, and their phenotype data is available. Alternatively, patients enrolled in multiple clinical trials may be pooled for the association analysis. Preferably, the multiple clinical trials relate to similar medical treatment, such as the same therapeutic agent, and data related to the relevant phenotype is available for all clinical trials.

Sample preparation

[0053] For sample preparation, a tissue or cell sample from a mammal (typically a human patient) may be used. Examples of samples include, but are not limited to, tissue biopsy, blood, lung aspirate, sputum, lymph fluid, etc. The sample can be obtained by a variety of procedures known in the art including, but not limited to surgical excision, aspiration or biopsy. The sample may be fresh or frozen, such as archived clinical samples. In some embodiments, the archived clinical samples may be selected from the group consisting of plasma samples, serum samples, dried blood spots, urine samples, tissue samples, tumor cells and buccal swabs. In some embodiments, the archived clinical samples may be plasma samples. In some embodiments, the sample may be fixed and embedded in paraffin or the like.

[0054] Archived plasma samples are used as an example to illustrate the invention. The archived plasma samples collected from patients or healthy volunteers may be used to extract DNA using any suitable method, such as the QIAGEN QIAamp MinElute Virus Spin Kit (Valencia, CA). This kit might be used with some modifications. For instance, 1 ml of plasma is vortexed briefly, and mixed thoroughly with 30 μg tRNA. The mixture is divided into 200 μl aliquots which are incubated for 1 hour before adding a lysis buffer. The lysate is then boiled for 5 minutes at 96°C and each aliquot is filtered through the same column. The DNA is eluted in 10 mM Tris-HCl (pH 8.5), vacuum-dried, and dissolved in sterile water. In most cases, the quantity of genomic DNA extracted from plasma is too low and inadequate for the subsequent genotyping. Therefore, in some embodiments, the isolated DNA may be suboptimal genomic DNA. Amplification of the isolated DNA may be needed to obtain sufficient amounts of DNA for the subsequent genotyping. In some embodiments, the amplification may be WGA, and the resulting DNA is wgaDNA. For example, DNA samples may be amplified using the Amersham Bioscience GenomiPhi DNA Amplification Kit (Piscataway, NJ) or equivalent reagents, and this process will typically yield several micrograms of DNA, which are sufficient for genotyping.

Genotyping using SNPs

[0055] Any suitable methods may be used for obtaining genotyping data from the isolated genomic DNA. A method of genotyping may obtain information about one or more individuals at one or more polymorphic loci simultaneously. In some embodiments,

genotyping may be defined as distinguishing alleles at a given genetic locus at single nucleotide resolution. A genetic locus is defined as a chromosomal location of a genetic or DNA marker. Thus the methods according to the present invention have the precision required to provide screening and diagnostic information for individuals that can be used as the basis for medical decisions.

[0056] The term "genotyped" as used herein refers to a process for determining a genotype of one or more individuals, where a "genotype" is a representation of one or more polymorphic variants in a population. Typically, genotyping involves assessing the presence or absence of polymorphic variants at one or more polymorphic loci. In some embodiments, high-density genotyping may be used. In some embodiments, the high-density genotyping is whole-genome genotyping.

[0057] As used herein, the term "polymorphic locus" refers to a region in a nucleic acid at which two or more alternative nucleotide sequences are observed in a significant number of nucleic acid samples from a population of individuals. A polymorphic locus may be a nucleotide sequence of two or more nucleotides, an inserted nucleotide or nucleotide sequence, a deleted nucleotide or nucleotide sequence, or variation in the copy number of a microsatellite, for example. A polymorphic locus that is two or more nucleotides in length may be 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or more, 20 or more, 30 or more, 50 or more, 75 or more, 100 or more, 500 or more, or about 1000 nucleotides in length, where all or some of the nucleotide sequences differ within the region. A polymorphic locus is often one nucleotide in length, which is referred to herein as a "single nucleotide polymorphism" or a "SNP." In some embodiments, the high-density genotyping may be conducted by using SNPs. In some embodiments, about 1,000-5,000,000 or more, preferably about 1,000,000, SNPs, may be used. In some embodiments, the high-density genotyping may be array-based. In some embodiments, the high-density genotyping may be conducted by sequencing, such as high-throughput sequencing.

[0058] Where there are two, three, or four alternative nucleotide sequences at a polymorphic locus, each nucleotide sequence is referred to as a "polymorphic variant" or "nucleic acid variant." Where two polymorphic variants exist, for example, the polymorphic variant represented in a minority of samples from a population is sometimes referred to as a "minor allele" and the polymorphic variant that is more prevalently represented is sometimes referred to as a "major allele." Many organisms possess a copy of each chromosome (e.g.,

humans), and those individuals who possess two major alleles or two minor alleles are often referred to as being "homozygous" with respect to the polymorphism, and those individuals who possess one major allele and one minor allele are normally referred to as being "heterozygous" with respect to the polymorphism. Individuals who are homozygous with respect to one allele are sometimes predisposed to a different phenotype as compared to individuals who are heterozygous or homozygous with respect to another allele.

[0059] In genetic analysis that identifies one or more pharmacogenomic biomarkers, samples from individuals having different values in a relevant phenotype often are allelotyped and/or genotyped. The term "allelotype" as used herein refers to a process for determining the allele frequency for a polymorphic variant in pooled DNA samples from cases and controls. By pooling DNA from each group, an allele frequency for each locus in each group is calculated. These allele frequencies are then compared to one another.

[0060] A genotype or polymorphic variant may be expressed in terms of a "haplotype," which as used herein refers to a set of DNA variations, or polymorphisms, that tend to be inherited together. A haplotype can refer to a combination of alleles or to a set of SNPs found on the same chromosome. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0061] Researchers sometimes report a polymorphic variant in a database without determining whether the variant is represented in a significant fraction of a population. Because a subset of these reported polymorphic variants are not represented in a statistically significant portion of the population, some of them are sequencing errors and/or not biologically relevant. Thus, it is often not known whether a reported polymorphic variant is statistically significant or biologically relevant until the presence of the variant is detected in a population of individuals and the frequency of the variant is determined. A polymorphic variant is statistically significant and often biologically relevant if it is represented in 1% or more of a population, sometimes 5% or more, 10% or more, 15% or more, or 20% or more of

a population, and often 25% or more, 30% or more, 35% or more, 40% or more, 45% or more, or 50% or more of a population.

[0062] A polymorphic variant may be detected on either or both strands of a double-stranded nucleic acid. Also, a polymorphic variant may be located within an intron or exon of a gene or within a portion of a regulatory region such as a promoter, a 5' untranslated region (UTR), a 3' UTR, a translated region, an intergenic region, or in a genomic region containing no known genes. Polymorphic variations may or may not result in detectable differences in gene expression, polypeptide structure, or polypeptide function.

Parameter optimization

[0063] The genotype results derived from DNA from archived samples exhibit different characteristics, which might exhibit significantly lower call rate than that from high quality DNA used in standard practice. In contrast to applying a typical call rate cut off used in standard GWAS, which is typically 95% or even higher, the analysis used in present invention adopts an unconventional approach to adjust the cut off call rate based on the genotype results obtained in order to include more samples in the analysis. As a result, the cut off value used might be substantially lower from the standard cut off value recommended by the vendors. Another related characteristic of the results generated from this invention is that a large volume of data may be missing, and it could present significant challenge when using standard or "known" analysis algorithms. Therefore, the present invention also uses imputation algorithm that can replace some of the missing data based on linkage disequilibrium (LD) among the genotyped polymorphic loci.

[0064] Optimization of an inclusion criterion may be performed for obtaining the genotyping data (Fig. 2). In some embodiments, the genotyping data is obtained by using a genome-wide genotype calling algorithm. Due to the quality of the suboptimal genomic DNA generated from the plasma samples and the whole genome amplification applied prior to genotyping, the call rates from some of the samples might be substantially lower than the typical call rates (>95%) when using high quality genomic DNA. In the present invention, the cut-off call rate is adjusted significantly lower than the conventional standard to include as many samples as possible. In some embodiments, the call rate cut-off value of the genome-wide genotype calling algorithm may be adjusted. In some embodiments, the genotype calls are made by using a call rate cut-off value that is lower than a typical call rate

cut-off value used for whole-genome genotyping of high quality genomic DNA, wherein the call rate cut-off value used is about 50-95%, about 80-90%, or about 90%. In some embodiments, adjustment of the call rate cut-off value of the genotype calling algorithm and making the genotype calls based on the adjusted call rate cut-of value may be repeated multiple times to identify a criterion to include and/or exclude samples. In some embodiments, an optimal inclusion criterion, such as call rate cut-off value, is identified after iterations of the adjustment cycle. In some embodiments, the genotype calls are made using an imputation algorithm, wherein the HapMap is used for the imputation algorithm.

[0065] In some embodiments, the DNA samples are genotyped using whole genome SNP arrays manufactured by Affymetrix (Santa Clara, CA) and/or Illumina (San Diego, CA). Affymetrix 500K array is used as an example to illustrate the present invention. In addition to Affymetrix arrays, Illumina chips and Sequenom MassArray are also used to confirm the results generated by one platform.

[0066] In some embodiments, the genotype calls are generated with the Affymetrix Genotyping Console™ software. In some embodiments, the genotype calls are generated using the Robust Linear Model with the Mahalanobis Distance Classifier (RLMM) algorithm, the RLMM with a Bayesian step (BRLMM) algorithm, the Axiom™ GT1 algorithm, the BRLMM using perfect-match probes (BRLMM-P) algorithm, or the Birdseed algorithm (Rabbee et al., *Bioinformatics* (2006) 22:7-12; Korn et al., *Nat Genet* (2008) 40:1253-60).

D. Identifying Pharmacogenomic Biomarkers by Association Analysis

[0067] The genotyping data obtained are used for performing association analysis based on the relevant phenotype to identify pharmacogenomic biomarkers (Fig. 1). Classification algorithms, which include but are not limited to support vector machine (SVM) and logistic regression, may be applied to the dataset and identify the optimal biomarkers and scoring algorithm. In some embodiments, the association analysis is a GWAS. In some embodiments, the association analysis is performed by calculating an associated p-value of each polymorphic locus, such as SNP, with the relevant phenotype. In some embodiments, the calculation is based on an allele frequency and/or genotype-based test.

[0068] The relevant phenotypes for identifying the pharmacogenomic biomarkers are generally related to responsiveness of individuals to a treatment regimen. A relevant phenotype may be qualitative or quantitative. Responsiveness may be primary, such as

reduce in cancer mass in response to an anti-cancer drug, or secondary, such as hypertriglyceridemia (HTC) in response to bexarotene. The primary and secondary responses may or may not be correlated with each other. A response may be positive or negative. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects. One or more relevant phenotypes may be used for the association analysis to identify the pharmacogenomic biomarkers.

[0069] In some embodiments, the association analysis may be repeated multiple times using different call rate cut-off values to optimize criteria used (Fig. 2). In some embodiments, the genotype results may be analyzed using genetic data analysis software such as PLINK (Purcell et al., *Am J Hum Genet* (2007) 81:559-5752) to calculate associated p-value of each polymorphic locus with the relevant phenotype. Polymorphic loci may be ordered by their calculated association with the relevant phenotype. The most significantly associated polymorphic loci may be identified as pharmacogenomic biomarkers for the relevant phenotype.

Verification and/or replication

[0070] The pharmacogenomic biomarkers identified by high-density genotyping may be subject to further association analyses, or the "Stage II" analyses (Fig. 1). The further association analyses may be used for verification and /or replication of the identified pharmacogenomic biomarkers through high-density genotyping and association analysis, or the Stage I analysis. Some or all of the archived clinical samples from Stage I and/or additional clinical samples may be used for the additional genotyping.

[0071] In some embodiments, the further association analysis may be based on additional genotyping data using the identified pharmacogenomic biomarkers. In some embodiments, about 1, 10, 20, 50, 100, 200, 500 or more of the identified pharmacogenomic biomarkers may be used for the additional genotyping.

[0072] In some embodiments, the highly associated SNPs identified from the Stage I analysis may be replicated with low density genotyping platforms which could be distinct from those used in Stage I, such as Sequenom iPLEX MassArray technology.

[0073] In Stage II, new DNA samples, which have not been used in Stage I, may be genotyped aiming to replicate the pharmacogenomic biomarker identified in Stage I. The additional clinical samples may come from patients in a new clinical trial, and may be fresh

clinical samples. Therefore, genomic DNA isolated from the additional clinical samples may be of higher quality and larger quantity, and suitable for high-density genotyping, even without amplification.

[0074] Optimization of an inclusion criterion may be performed for obtaining the additional genotyping data (Fig. 3). In some embodiments, the additional genotyping data may be obtained by using a verification genotype calling algorithm. In some embodiments, the further association analysis may comprise adjusting the call rate cut-off value of the verification genotype calling algorithm. In some embodiments, genotyping and adjusting the call rate cut-off may be repeated multiple times to include and/or exclude samples. In some embodiments, an optimal inclusion criterion may be identified. In some embodiments, the method may further comprise comparing the additional genotyping data obtained by using the verification genotype calling algorithm to the genotyping data obtained by using the genome-wide genotype calling algorithm (Fig. 3).

[0075] After the two-stage studies, the most significantly associated pharmacogenomic biomarkers can be identified. In some embodiments, the association analysis may be repeated multiple times using different call rate cut-off values to optimize criteria used (Fig. 3). In some embodiments, the pharmacogenomic biomarkers are a subgroup of the pharmacogenomic biomarkers identified in Stage I. In some embodiments, the pharmacogenomic biomarkers may comprise one or more SNPs. Multiple pharmacogenomic biomarkers identified may be located close to each other on the genome, and may locate in an intron/exon of a gene, an intergenic region, or a region containing no known gene on the genome.

E. Genome-Wide Association Study Using Archived Samples

[0076] In still another aspect, provided herein is a method to conduct GWAS using suboptimal genomic DNA. In some embodiments, the suboptimal genomic DNA may be from archived samples. In some embodiments, the suboptimal genomic DNA may be from plasma samples. In some embodiments, the suboptimal genomic DNA may be amplified. Further provided herein is a method for identifying a pharmacogenomic biomarker using the method to conduct GWAS using suboptimal genomic DNA, wherein the method may be a retrospective method.

699572000140

[0077] The two-stage data analysis described above may be used for the GWAS using suboptimal genomic DNA. In some embodiments, multiple genotyping platforms may be used. In some embodiments, the same or different samples may be used for the multiple genotyping platforms. In some embodiments, the method may further use a sample providing high-quality genomic DNA, which may be used for replication of the data obtained from suboptimal genomic DNA. In some embodiments, the samples providing high-quality genomic DNA may be whole blood samples.

F. Applications of the Pharmacogenomic Biomarkers

[0078] Pharmacogenomics involves tailoring a treatment for a subject according to the subject's genotype as a particular treatment regimen may exert a differential effect depending upon the subject's genotype. For example, based upon the outcome of a prognostic test, a clinician or physician may target pertinent information and preventative or therapeutic treatments to a subject who would be benefited by the information or treatment and avoid directing such information and treatments to a subject who would not be benefited (e.g., the treatment has no therapeutic effect and/or the subject experiences adverse side effects). Information generated from pharmacogenomic biomarkers using a method described herein can be used to determine appropriate dosage and treatment regimens for an individual. This knowledge, when applied to dosing or drug selection, can avoid adverse reactions or therapeutic failure and thus enhance therapeutic efficiency when administering a therapeutic composition. In some embodiments, the pharmacogenomic biomarker may be used to develop a companion diagnostic test.

[0079] Therefore, in a further aspect, provided herein is a companion diagnostic test using the pharmacogenomic biomarkers identified by the method disclosed herein. For example, in one embodiment, a physician or clinician may consider applying knowledge obtained in pharmacogenomic biomarkers using a method described herein, when determining whether to administer a pharmaceutical composition to a subject. In another embodiment, a physician or clinician may consider applying such knowledge when determining the dosage, e.g., amount per treatment or frequency of treatments, of a treatment, administered to a patient.

[0080] The invention provides methods for assessing or aiding assessment of responsiveness of a subject to treatment. The invention also provides methods for predicting responsiveness or monitoring treatment/responsiveness to a treatment in a subject. The invention provides methods for selecting a subject for treatment and treating the subject. In some embodiments, the methods comprise assessing one or more pharmacogenomic biomarkers in a sample obtained from the subject; and predicting, assessing, or aiding assessment of responsiveness of the subject to a treatment based on the genotype of said one or more pharmacogenomic biomarkers. In some embodiments, the responsiveness is predicted or assessed by classifying the subject using an algorithm such as SVM, logistic regression, or K-nearest neighbors analysis.

[0081] The following is an example of a pharmacogenomic embodiment. A particular treatment regimen can exert a differential effect depending upon the subject's genotype. Where a candidate therapeutic exhibits a significant interaction with a major allele and a comparatively weak interaction with a minor allele (e.g., an order of magnitude or greater difference in the interaction), such a therapeutic typically would not be administered to a subject genotyped as being homozygous for the minor allele, and sometimes not administered to a subject genotyped as being heterozygous for the minor allele. In another example, where a candidate therapeutic is not significantly toxic when administered to subjects who are homozygous for a major allele but is comparatively toxic when administered to subjects heterozygous or homozygous for a minor allele, the candidate therapeutic is not typically administered to subjects who are genotyped as being heterozygous or homozygous with respect to the minor allele.

[0082] The methods described herein are applicable to pharmacogenomic methods for preventing, alleviating or treating conditions such as metabolic disorders, cardiovascular diseases, cancers, etc. For example, a nucleic acid sample from an individual may be subjected to a prognostic test described herein. Where one or more polymorphic variations associated with increased risk of type II diabetes are identified in a subject, information for preventing or treating type II diabetes and/or one or more type II diabetes treatment regimens then may be prescribed to that subject.

[0083] In certain embodiments, a treatment regimen is specifically prescribed and/or administered to individuals who will most benefit from it based upon their likelihood of responding to a treatment regimen assessed by the methods described herein. Thus, provided

are methods for identifying a subject with a high likelihood of responding to a treatment regimen and then prescribing such treatment regimen to individuals identified as having a high likelihood of responding. Thus, certain embodiments are directed to a method for treating a subject, which comprises: detecting the presence or absence of a pharmacogenomic biomarker associated with responsiveness to a treatment regimen in a nucleotide sequence set forth herein in a nucleic acid sample from a subject, and prescribing or administering the treatment regimen to a subject from whom the sample originated where the presence of a pharmacogenomic biomarker associated with responsiveness to the treatment regimen is detected in the nucleotide sequence.

[0084] The treatment sometimes is preventative (e.g., is prescribed or administered to reduce the probability that a disease condition arises or progresses), sometimes is therapeutic, and sometimes delays, alleviates or halts the progression of a disease condition. Any known preventative or therapeutic treatment for alleviating or preventing the occurrence of a disorder may be prescribed and/or administered.

[0085] Pharmacogenomics methods also may be used to analyze and predict a response to a drug. For example, if pharmacogenomics analysis indicates a likelihood that an individual will respond positively to a treatment with a particular drug, the drug may be administered to the individual. Conversely, if the analysis indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. The response to a therapeutic treatment can be predicted in a background study in which subjects in any of the following populations are genotyped: a population that responds favorably to a treatment regimen, a population that does not respond significantly to a treatment regimen, and a population that responds adversely to a treatment regiment (e.g., exhibits one or more side effects). These populations are provided as examples and other populations and subpopulations may be analyzed. Based upon the results of these analyses, a subject is genotyped to predict whether he or she will respond favorably to a treatment regimen, not respond significantly to a treatment regimen, or respond adversely to a treatment regimen.

[0086] A classification/prediction algorithm may be developed using the verification and/or replication dataset. An imputation algorithm that can replace some of the missing data based on LD among the genotyped polymorphic loci may be used. In embodiments where SNPs are used for genotyping, SNP databases such as Hapmap may be used for the

imputation algorithm. For development of the classification/prediction algorithm, the verification dataset may be used as a training dataset. Once a classification/prediction algorithm has been developed, the replication dataset may be used for testing the algorithm.

[0087] In some embodiments, the methods of the invention comprise classifying the subject as a responsive or non-responsive subject using a K-nearest neighbors analysis based on the genotype of the pharmacogenomic biomarkers in the sample from the subject and reference samples with known classes. In some embodiments, classifying the subject using a K-nearest neighbors analysis is carried out by (1) determining parameter K (i.e., number of nearest neighbors); (2) calculating the difference between the measured expression level of the marker genes in the new sample to be classified and the expression level of the respective marker genes in each reference sample; (3) determining the nearest reference samples by selecting those samples with the smallest weighted average of the absolute differences (WAAD) between the new sample and the reference sample; and (4) determining class of the new sample based on the known classes of the K nearest reference samples. The weights and/or parameter K are determined using cross-validation with clinical trial samples with known classes. For example, 5-fold (such as 5-fold, 6-fold, 7-fold, 8-fold, 9-fold, or 10-fold) to N-fold cross-validation may be used to minimize the weighted K-nearest neighbors classification error, wherein N is the size of the samples. In some embodiments, K is an integer between 4 and 13 (e.g., 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13). In some embodiments, the nearest reference samples (nearest neighbors) are those with the smallest weighted average of the absolute differences between the expression level of the new sample to be classified and the expression level of each reference sample for each of the pharmacogenomic biomarkers.

[0088] The comparisons and/or calculations for predicting, assessing or aiding assessment can be carried out in any convenient manner appropriate to the type of measured value and/or reference value for the pharmacogenomic biomarkers at issue. The process of comparing or calculating may be manual or it may be automatic (such as by a machine including computer-based machine). As will be apparent to those of skill in the art, replicate genotyping may be taken for the pharmacogenomic biomarkers.

[0089] Also provided herein is a method of prognosticating responsiveness of a subject to a treatment using the companion diagnostic test disclosed herein. The tests described herein also are applicable to clinical drug trials. In some embodiments, the pharmacogenomic biomarkers can be used to stratify or select a subject population for a clinical trial. The

pharmacogenomic biomarkers can, in some embodiments, be used to stratify individuals that may exhibit a toxic response to a treatment from those that will not. In other embodiments, the pharmacogenomic biomarkers can be used to separate those that will be non-responders from those who will be responders. The pharmacogenomic biomarkers described herein can be used in pharmacogenomic-based design and in managing the conduct of a clinical trial.

[0090] One or more pharmacogenomic biomarkers indicative of response to a therapeutic agent or side effects to a therapeutic agent may be identified using the methods described herein. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems.

[0091] Thus, another embodiment is a method of selecting an individual for inclusion in a clinical trial of a treatment or drug comprising the steps of: (a) obtaining a nucleic acid sample from an individual; (b) determining the identity of a polymorphic variation which is associated with a positive response to the treatment or the drug, or at least one polymorphic variation which is associated with a negative response to the treatment or the drug in the nucleic acid sample, and (c) including the individual in the clinical trial if the nucleic acid sample contains said polymorphic variation associated with a positive response to the treatment or the drug or if the nucleic acid sample lacks said polymorphic variation associated with a negative response to the treatment or the drug. In addition, the methods described herein for selecting an individual for inclusion in a clinical trial of a treatment or drug encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination. The including step (c) optionally comprises administering the drug or the treatment to the individual if the nucleic acid sample contains the polymorphic variation associated with a positive response to the treatment or the drug and the nucleic acid sample lacks said biallelic marker associated with a negative response to the treatment or the drug.

G. Additional Pharmacogenomic Biomarkers or Drug Targets

[0092] Also provided is a method for identifying polymorphic variants proximal to an identified pharmacogenomic biomarker. Thus, featured herein are methods for identifying a polymorphic variation that is proximal to an identified pharmacogenomic biomarker. In another embodiment, the proximal polymorphic variant identified sometimes is a publicly disclosed polymorphic variant, which for example, sometimes is published in a publicly available database. In other embodiments, the polymorphic variant identified is not publicly disclosed and is discovered using a known method, including, but not limited to, sequencing a region surrounding the identified pharmacogenomic biomarker in a group of nucleic samples. Thus, multiple polymorphic variants proximal to an identified pharmacogenomic biomarker are identified using this method.

[0093] The proximal polymorphic variant often is identified in a region surrounding the identified pharmacogenomic biomarker. In certain embodiments, this surrounding region is about 50 kb flanking the identified pharmacogenomic biomarker (e.g., about 50 kb 5' of the first polymorphic variant and about 50 kb 3' of the first polymorphic variant), and the region sometimes is composed of shorter flanking sequences, such as flanking sequences of about 40 kb, about 30 kb, about 25 kb, about 20 kb, about 15 kb, about 10 kb, about 7 kb, about 5 kb, or about 2 kb 5' and 3' of the identified pharmacogenomic biomarker. In other embodiments, the region is composed of longer flanking sequences, such as flanking sequences of about 55 kb, about 60 kb, about 65 kb, about 70 kb, about 75 kb, about 80 kb, about 85 kb, about 90 kb, about 95 kb, or about 100 kb 5' and 3' of the identified pharmacogenomic biomarker.

[0094] In some embodiments, the pharmacogenomic biomarkers may be used to identify one or more additional pharmacogenomic biomarkers. For example, other polymorphic loci located in proximity to the pharmacogenomic biomarkers may be analyzed for association with the relevant phenotype. Additionally, genes may be identified that are in proximity to the pharmacogenomic biomarkers, and their functions analyzed. Genes with functions that are directly or indirectly related to the relevant phenotype, or other genes in the same cellular pathway, may be targets for further analysis with the relevant phenotype, and new pharmacogenomic biomarkers may be identified.

[0095] In certain embodiments, polymorphic variants are identified iteratively. For example, a first proximal polymorphic variant is identified using the methods described above and then another polymorphic variant proximal to the first proximal polymorphic

variant is identified (e.g., publicly disclosed or discovered) and the presence or absence of an association of one or more other polymorphic variants proximal to the first proximal polymorphic variant is determined.

[0096] The methods described herein are useful for identifying or discovering additional polymorphic variants that may be used to further characterize a gene, region or loci associated with a condition, a disease, or a disorder. For example, allelotyping or genotyping data from the additional polymorphic variants may be used to identify a functional mutation or a region of linkage disequilibrium. In certain embodiments, polymorphic variants identified or discovered within a region comprising the identified pharmacogenomic biomarker are genotyped using the genetic methods and sample selection techniques described herein, and it can be determined whether those polymorphic variants are in linkage disequilibrium with the identified pharmacogenomic biomarker. The size of the region in linkage disequilibrium with the identified pharmacogenomic biomarker also can be assessed using these genotyping methods. Thus, provided herein are methods for determining whether a polymorphic variant is in linkage disequilibrium with an identified pharmacogenomic biomarker, and such information can be used in prognosis/diagnosis methods described herein.

[0097] Further provided herein is a method of identifying a novel drug target using the pharmacogenomic biomarkers identified by the method disclosed herein. In some embodiments, said biomarkers and their associated SNPs or genes could gain insight of the underlying biological pathways or mechanisms underlying the relevant phenotypes, such as efficacy, adverse effect, or other endpoints. These discoveries could be instrumental for developing better diagnosis or therapeutic agents.

H. Kits

[0098] Diagnostic kits based on the pharmacogenomic biomarker described above might be developed, and they can be used to predict individual's response to the corresponding drug. Such test kits can include devices and instructions that a subject can use to obtain a sample, e.g., of buccal cells or blood, without the aid of a health care provider.

[0099] For use in the applications described or suggested above, kits or articles of manufacture are also provided by the invention. Such kits may comprise at least one reagent

specific for genotyping a pharmacogenomic biomarker described herein, and may further include instructions for carrying out a method described herein.

[0100] In some embodiments, the invention provides compositions and kits comprising primers and primer pairs, which allow the specific amplification of the polynucleotides of the invention or of any specific parts thereof, and probes that selectively or specifically hybridize to nucleic acid molecules of the invention or to any part thereof. Probes may be labeled with a detectable marker, such as, for example, a radioisotope, fluorescent compound, bioluminescent compound, a chemiluminescent compound, metal chelator or enzyme. Such probes and primers can be used to detect the presence of polynucleotides in a sample and as a means for detecting cell expressing proteins encoded by the polynucleotides. As will be understood by the skilled artisan, a great many different primers and probes may be prepared based on the sequences provided herein and used effectively to amplify, clone and/or determine the presence and/or levels of genomic DNAs.

[0101] In some embodiments, the kit may comprise reagents for detecting presence of polypeptides. Such reagents may be antibodies or other binding molecules that specifically bind to a polypeptide. In some embodiments, such antibodies or binding molecules may be capable of distinguishing a structural variation to the polypeptide as a result of polymorphism, and thus may be used for genotyping. The antibodies or binding molecules may be labeled with a detectable marker, such as, for example, a radioisotope, a fluorescent compound, a bioluminescent compound, a chemiluminescent compound, a metal chelator, an enzyme, or a particle. Other reagents for performing binding assays, such as ELISA, may be included in the kit.

[0102] In some embodiments, the kits comprise reagents for genotyping at least two, at least three, at least five, at least ten, or fifteen pharmacogenomic biomarkers. In some embodiments, the kits may further comprise a surface or substrate (such as a microarray) for capture probes for detecting of amplified nucleic acids.

[0103] The kits may further comprise a carrier means being compartmentalized to receive in close confinement one or more container means such as vials, tubes, and the like, each of the container means comprising one of the separate elements to be used in the method. For example, one of the container means may comprise a probe that is or can be detectably labeled. Such probe may be a polynucleotide specific for a pharmacogenomic biomarker. Where the kit utilizes nucleic acid hybridization to detect the target nucleic acid, the kit may

also have containers containing nucleotide(s) for amplification of the target nucleic acid sequence and/or a container comprising a reporter-means, such as a biotin-binding protein, such as avidin or streptavidin, bound to a reporter molecule, such as an enzymatic, florescent, or radioisotope label.

[0104] The kit of the invention will typically comprise the container described above and one or more other containers comprising materials desirable from a commercial and user standpoint, including buffers, diluents, filters, needles, syringes, and package inserts with instructions for use. A label may be present on the container to indicate that the composition is used for a specific therapy or non-therapeutic application, and may also indicate directions for either *in vivo* or *in vitro* use, such as those described above.

[0105] The kit can further comprise a set of instructions and materials for preparing a tissue or cell sample and preparing nucleic acids (such as genomic DNA) from the sample.

[0106] The invention provides a variety of compositions suitable for use in performing methods of the invention, which may be used in kits. For example, the invention provides surfaces, such as arrays that can be used in such methods. In some embodiments, an array of the invention comprises individual or collections of nucleic acid molecules useful for detecting pharmacogenomic biomarkers of the invention. For instance, an array of the invention may comprises a series of discretely placed individual nucleic acid oligonucleotides or sets of nucleic acid oligonucleotide combinations that are hybridizable to a sample comprising target nucleic acids, whereby such hybridization is indicative of genotypes of the pharmacogenomic biomarkers of the invention.

[0107] Several techniques are well-known in the art for attaching nucleic acids to a solid substrate such as a glass slide. One method is to incorporate modified bases or analogs that contain a moiety that is capable of attachment to a solid substrate, such as an amine group, a derivative of an amine group or another group with a positive charge, into nucleic acid molecules that are synthesized. The synthesized product is then contacted with a solid substrate, such as a glass slide, which is coated with an aldehyde or another reactive group which will form a covalent link with the reactive group that is on the amplified product and become covalently attached to the glass slide. Other methods, such as those using amino propryl silica surface chemistry are also known in the art, as disclosed at world wide web at cmt.corning.com and cmgm.stanford.edu/pbrown1.

[0108] Attachment of groups to oligonucleotides which could be later converted to reactive groups is also possible using methods known in the art. Any attachment to nucleotides of oligonucleotides will become part of oligonucleotide, which could then be attached to the solid surface of the microarray. Amplified nucleic acids can be further modified, such as through cleavage into fragments or by attachment of detectable labels, prior to or following attachment to the solid substrate, as required and/or permitted by the techniques used.

[0109] The present invention can be applied broadly in the biomedical fields and offers a number of major advantages to modern drug development and the emerging field of personalized medicine. These benefits include, but are not limited to, shortening the time and cutting the cost required to identify biomarkers for drugs in clinical development, drastically enhancing the chance of success for investigated drugs, rescuing drugs which would have been abandoned without patient stratification in clinical trials.

I. Computer Readable Medium

[0110] In yet another aspect, provided herein is a computer readable medium comprising a plurality of instructions for a genotyping method using suboptimal genomic DNA samples, which comprises the steps of: a) receiving sequence information of said suboptimal genomic DNA samples; b) optimizing an inclusion criterion based on said sequence information; and c) calculating genotypes based on said sequence information and said optimized inclusion criterion.

[0111] Also provided herein is a genotyping method using suboptimal genomic DNA samples, which method comprises optimizing an inclusion criterion. In some embodiments, the optimization may be repeated multiple times to include and/or exclude samples. In some embodiments, an optimal inclusion criterion may be identified. In some embodiments, the genotyping data may be obtained by using a genome-wide genotype calling algorithm and/or a verification genotype calling algorithm. In some embodiments, the inclusion criterion may be the call rate cut-off value of the genotype calling algorithm. In some embodiments, the genotype calls may be made by using a call rate cut-off that is lower than a typical call rate cut-off used for whole-genome genotyping of high quality genomic DNA, wherein the call rate cut-off value used may be about 50%, 60%, 70%, 80%, 90% or 95%. In some embodiments, the genotyping data may be obtained by using multiple genotyping platforms.

In some embodiments, the genotyping data from multiple genotyping platforms may be compared for the optimization.

[0112] Further provided is a method to perform association analysis using the genotyping method using suboptimal genomic DNA samples, which method comprises optimizing an inclusion criterion. In some embodiments, the association analysis may be repeated multiple times for the optimization.

J. Examples

[0113] The following examples are offered to illustrate but not to limit the invention.

Example 1

Retrospective *De Novo* Identification of Pharmacogenomic Biomarkers Using Archived Plasma Samples from Clinical Trials

[0114] *Patients.* Among patients enrolled in the clinical trials and treated with the drug, plasma samples from 400 individuals are available. The cases are defined as those who responded positively from the drug treatment, and the controls are those who had no response or responded negatively from the drug treatment. Prior to the study, patient identification and individually identifiable information were removed, and all samples were relabeled by a third party to protect patient identity.

[0115] *DNA preparation.* DNA is extracted from plasma samples with the QIAGEN QIAamp MinElute Virus Spin Kit (Valencia, CA, USA) with some modifications. Briefly, 1 ml of plasma is vortexed briefly, and mixed thoroughly with 30 μg tRNA. The mixture is divided into 200 μl aliquots which are incubated for 1 hour before adding a lysis buffer. The lysate is then boiled for 5 minutes at 96°C and each aliquot is filtered through the same column. The DNA is eluted in 10 mM Tris-HCl (pH 8.5), vacuum-dried, and dissolved in sterile water. In most cases, the quantity of genomic DNA extracted from plasma is too low and inadequate for the subsequent genotyping, and DNA samples are then amplified using the Amersham Bioscience GenomiPhi DNA Amplification Kit (Piscataway, NJ, USA).

[0116] *SNP genotyping and data analysis.* In Stage I (Fig. 1), 150 samples (75 cases and 75 controls) are genotyped using the Affymetrix GeneChip 500K Mapping Array Set containing 500,000 SNPs following Affymetrix standard protocol (Santa Clara, CA, USA). Due to the quality of the suboptimal genomic DNA generated from the plasma samples and

the whole genome amplification applied prior to genotyping, the call rates from some of the samples might be substantially lower than the typical call rates (>95%) when using high quality genomic DNA. Therefore, the cut-off call rate is adjusted significantly lower than the conventional standard to include as many samples as possible. The adjustment of the call rate cut-off value of the genotype calling algorithm and making the genotype calls based on the adjusted call rate cut-of value are repeated multiple times to identify an optimal criterion to include and/or exclude samples (Fig. 2). After removing the samples having call rates lower than the optimal cut-off criteria, the genotype results are analyzed using genetic data analysis software PLINK (Purcell et al., *Am J Hum Genet* (2007) 81:559-5752) to calculate associated p-value of each polymorphic locus with the relevant phenotype. Polymorphic loci are ordered by their calculated association with the relevant phenotype. The 200 most significantly associated SNPs are selected for Stage II study using Sequenom iPLEX assays (Sequenom, San Diego, CA, USA). These assays are used to genotype all 400 DNA samples from the clinical trials. Among them, 150 samples used in Stage I are selected as the verification group, and the other 250 samples are used as the replication group (Fig. 1). Final genotype calls are generated by Sequenom Typer Analyzer from the MassARRAY Typer suite (Sequenom, San Diego, CA, USA). The genotyped results from the verification group are compared to the results generated in Stage I. The call rates from some of the samples might be substantially lower than the typical call rates (>95%) when using high quality genomic DNA, and the cut-off call rate is adjusted multiple times to include and/or exclude samples. After removing the samples having call rates lower than the optimal cut-off criteria and samples with too many discrepancies between the two stages (Fig. 3), association analysis is performed by calculating an associated p-value of each SNP with the relevant phenotype using PLINK program. The calculation may be based on an allele frequency and/or genotype-based test, and the relevant phenotype may be a categorical trait, a quantitative trait or another relevant phenotype. Pharmacogenomic biomarkers showing significant associations with the drug response are identified, and these pharmacogenomic biomarkers may comprise one or more SNPs. Multiple pharmacogenomic biomarkers identified may be located close to each other on the genome, and may locate in an intron/exon of a gene, an intergenic region, or a region containing no known gene on the genome. Classification algorithms, such as support vector machine (SVM) and logistic

regression, may be applied to the dataset to identify the optimal biomarkers and scoring algorithm, so the patient's response to drug treatment can be properly predicted.

[0117] The above examples are included for illustrative purposes only and are not intended to limit the scope of the invention. Many variations to those described above are possible. Since modifications and variations to the examples described above will be apparent to those of skill in this art, it is intended that this invention be limited only by the scope of the appended claims.

Claims

1. A method to identify one or more pharmacogenomic biomarkers, which method comprises:

a) isolating DNA from archived clinical samples of at least two patients exhibiting different values in a relevant phenotype;

b) amplifying said isolated DNA;

c) obtaining high-density genotyping data of said amplified DNA; and

d) performing association analysis based on said genotyping data and said different values in said relevant phenotype,

wherein said pharmacogenomic biomarker(s) are identified.

2. The method of claim 1, wherein the archived clinical samples are selected from the group consisting of plasma samples, serum samples, dried blood spots, urine samples, tissue samples, tumor cells and buccal swabs.

3. The method of claim 2, wherein the archived clinical samples are plasma samples.

4. The method according to any one of claims 1-3, wherein the isolated DNA is suboptimal genomic DNA.

5. The method according to any one of claims 1-4, wherein the amplification is whole-genome amplification (WGA), and the resulting DNA is whole-genome amplified DNA (wgaDNA).

6. The method according to any one of claims 1-5, wherein the high-density genotyping is whole-genome genotyping.

7. The method according to any one of claims 1-6, wherein the high-density genotyping is by using single nucleotide polymorphisms (SNPs).

8. The method of claim 7, wherein about 1,000-5,000,000 or more SNPs are used.

9. The method of claim 8, wherein about 1,000,000 SNPs are used.

10. The method according to any one of claims 1-9, wherein the high-density genotyping is array based, bead based, or high-throughput sequencing based.

11. The method according to any one of claims 1-10, wherein the genotyping data is obtained by using a genome-wide genotype calling algorithm.

12. The method of claim 11, further comprising:

e) adjusting the call rate cut-off value of the genome-wide genotype calling algorithm.

13. The method of claim 12, wherein step d) and step e) are repeated multiple times to include and/or exclude samples.

14. The method of claim 13, wherein an optimal inclusion criterion is identified.

15. The method according to any one of claims 11-14, wherein the genotype calls are made by using a call rate cut-off that is lower than a typical call rate cut-off used for whole-genome genotyping of high quality genomic DNA.

16. The method of claim 12, wherein the call rate cut-off used is about 50-95%.

17. The method of claim 13, wherein the call rate cut-off used is about 80-90%.

18. The method of claim 14, wherein the call rate cut-off used is about 90%.

19. The method according to any one of claims 11-18, wherein the genotype calls are generated with the Affymetrix Genotyping Console™ software.

20. The method according to any one of claims 1-19, wherein the genotype calls are generated using the BRLMM algorithm.

21. The method according to any one of claims 1-20, wherein the genotype calls are made using an imputation algorithm.

22. The method of claim 21, wherein the HapMap is used for the imputation algorithm.

23. The method according to any one of claims 1-22, wherein the association analysis is a genome-wide association study (GWAS).

24. The method according to any one of claims 1-23, wherein the association analysis is performed by calculating an associated p-value of each SNP with the relevant phenotype.

25. The method of claim 24, wherein the calculation is based on an allele frequency and/or genotype-based test.

26. The method according to any one of claims 1-25, wherein the relevant phenotype is a categorical trait, a quantitative trait or another relevant phenotype.

27. The method according to any one of claims 1-26, wherein the archived clinical samples are from about 2-1,000 or more patients.

28. The method according to any one of claims 1-27, further comprising performing association analysis based on additional genotyping data using the identified pharmacogenomic biomarkers.

29. The method of claim 28, wherein about 1-500 or more of the identified pharmacogenomic biomarkers are used for the additional genotyping.

30. The method according to claim 28 or 29, wherein some or all of the archived clinical samples from step a) and/or additional clinical samples are used for the additional genotyping.

31. The method according to any one of claims 28-30, wherein the additional genotyping data is obtained by using a verification genotype calling algorithm.

32. The method of claim 31, further comprising adjusting the call rate cut-off value of the verification genotype calling algorithm.

33. The method of claim 32, wherein genotyping and adjusting the call rate cut-off are repeated multiple times to include and/or exclude samples.

34. The method of claim 33, wherein an optimal inclusion criterion is identified.

35. The method according to any one of claims 28-34, further comprising comparing the additional genotyping data obtained by using the verification genotype calling algorithm to the genotyping data obtained by using the genome-wide genotype calling algorithm.

36. The method according to any one of claims 28-35, wherein a subset of the pharmacogenomic biomarkers from step d) is identified.

37. The method of claim 36, wherein the method is used for retrospective study of archived clinical samples from a previously conducted clinical trial.

38. The method according to claim 36 or 37, wherein the method is used for *de novo* identification of a pharmacogenomic biomarker.

39. A pharmacogenomic biomarker identified by the method according to any one of claims 36-38.

40. A group of pharmacogenomic biomarkers identified by the method according to any one of claims 36-38.

41. The pharmacogenomic biomarker according to claim 39 or 40, wherein the biomarker is one or more SNPs.

42. The pharmacogenomic biomarker according to any one of claims 39-41, for use to identify one or more additional pharmacogenomic biomarkers.

43. The pharmacogenomic biomarker according to any one of claims 39-41, for use to develop a companion diagnostic test.

44. A companion diagnostic test using the pharmacogenomic biomarker according to any one of claims 39-41.

45. A method of prognosticating responsiveness of a subject to a treatment using the companion diagnostic test of claim 44.

46. A method of identifying a novel drug target using the pharmacogenomic biomarker according to any one of claims 39-41.

47. A kit comprising a reagent for assessing the pharmacogenomic biomarker identified by the method according to any one of claims 36-38.

48. The kit of claim 47, further comprising instructions for using the pharmacogenomic biomarker to conduct a companion diagnostic test.

49. The kit of claim 47, wherein the reagent is used to detect a polynucleotide and/or polypeptide molecule.

699572000140

50. A genotyping method using suboptimal genomic DNA samples, which method comprises:

a) receiving sequence information of said suboptimal genomic DNA samples;

b) optimizing an inclusion criterion based on said sequence information; and

c) calculating genotypes based on said sequence information and said optimized inclusion criterion.

51. The method of claim 50, wherein the optimization is repeated multiple times to include and/or exclude samples.

52. The method of claim 51, wherein an optimal inclusion criterion is identified.

53. The method according to any one of claims 50-52, wherein the genotyping data is obtained by using a genome-wide genotype calling algorithm and/or a verification genotype calling algorithm.

54. The method of claim 53, wherein the inclusion criterion is the call rate cut-off value of the genotype calling algorithm.

55. The method of claim 54, wherein the genotype calls are made by using a call rate cut-off that is lower than a typical call rate cut-off used for whole-genome genotyping of high quality genomic DNA.

56. The method of claim 55, wherein the call rate cut-off used is about 50-95%.

57. The method of claim 56, wherein the call rate cut-off used is about 80-90%.

58. The method of claim 57, wherein the call rate cut-off used is about 90%.

59. The method according to any one of claims 50-58, wherein the genotyping data are obtained by using multiple genotyping platforms.

60. The method of claim 59, wherein the genotyping data from multiple genotyping platforms are compared for the optimization.

61. A method to perform association analysis using the genotyping method according to any one of claims 50-60.

62. The method of claim 61, wherein the association analysis is repeated multiple times for the optimization.

63. A computer readable medium comprising a plurality of instructions for a genotyping method using suboptimal genomic DNA samples, which comprises the steps of:

a) receiving sequence information of said suboptimal genomic DNA samples;

b) optimizing an inclusion criterion based on said sequence information using the method according to any one of claims 50-62; and

c) calculating genotypes based on said sequence information and said optimized inclusion criterion.

64. A method to conduct GWAS using suboptimal genomic DNA.

65. The method of claim 64, wherein the suboptimal genomic DNA is from archived samples.

66. The method according to claim 64 or 65, wherein the suboptimal genomic DNA is from plasma samples.

67. The method according to any one of claims 64-66, wherein multiple genotyping platforms are used.

68. The method of claim 67, wherein the same samples are used for the multiple genotyping platforms.

699572000140

69. The method of claim 67, wherein different samples are used for the multiple genotyping platforms.

70. The method according to any one of claims 64-69, further comprising a sample providing high-quality genomic DNA.

71. The method according to any one of claims 64-70, using the association analysis method according to claim 61 or 62.

72. A method to identify one or more pharmacogenomic biomarkers using the method according to any one of claims 64-71.

Figure 1

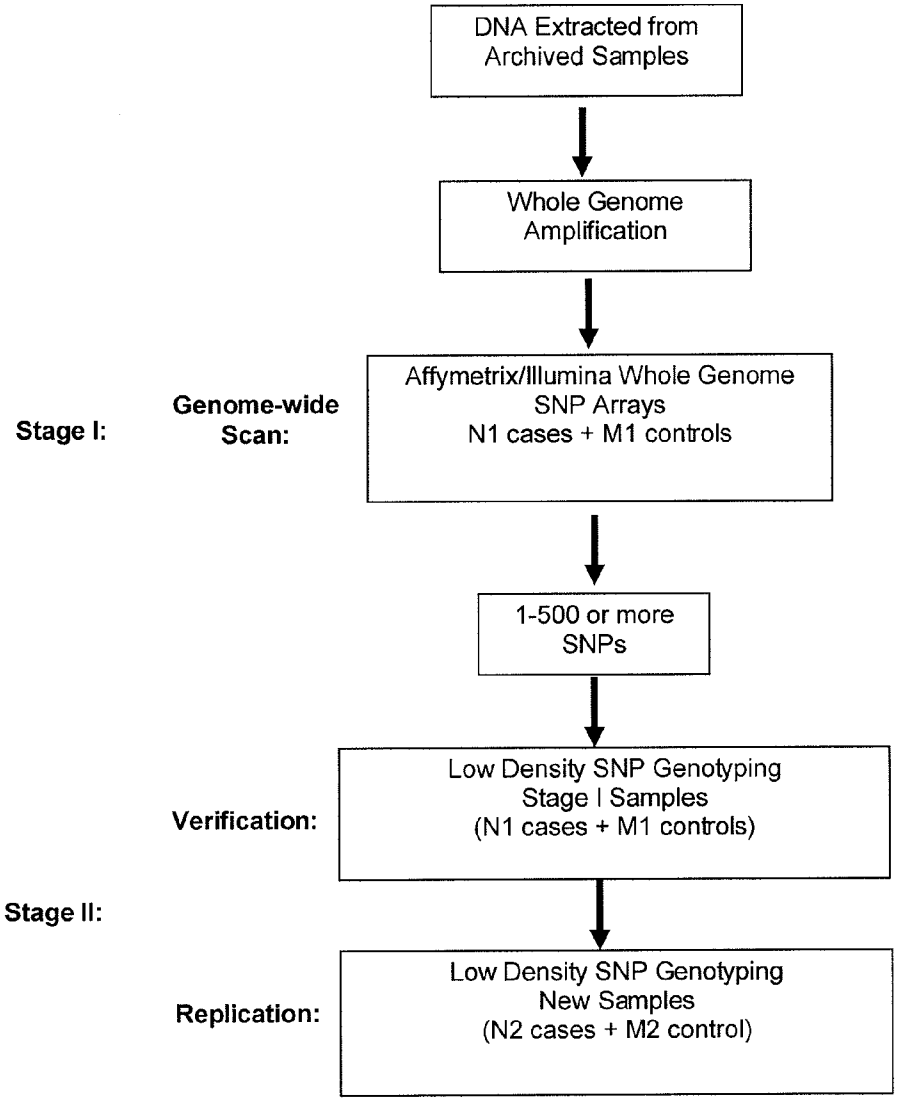

DNA Extracted from Archived Samples
Whole Genome Amplification
Stage I:
Genome-wide Scan:
Affymetrix/Illumina Whole Genome SNP Arrays
N1 cases + M1 controls
1-500 or more SNPs
Verification:
Low Density SNP Genotyping
Stage I Samples
(N1 cases + M1 controls)
Stage II:
Replication:
Low Density SNP Genotyping
New Samples
(N2 cases + M2 control)

Figure 2

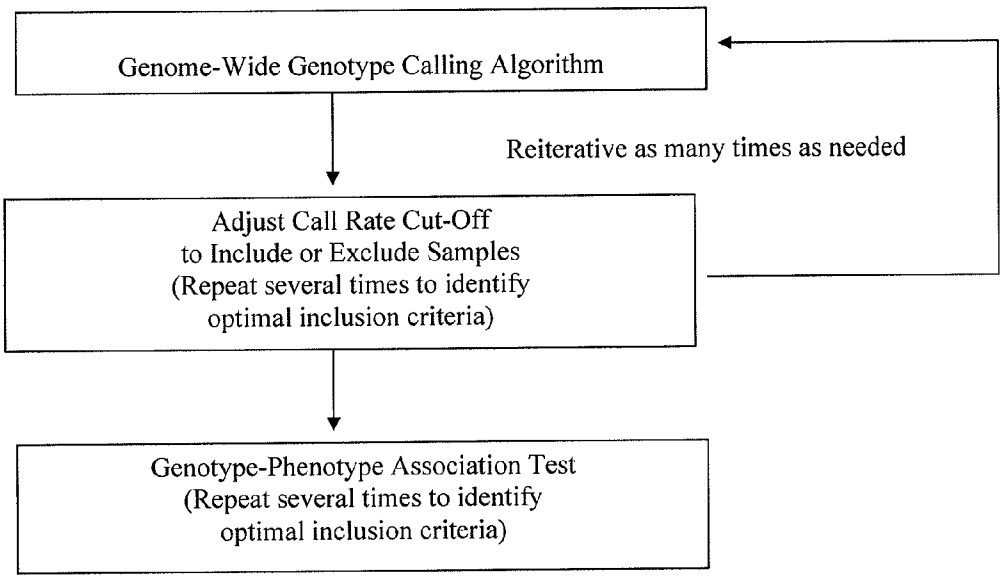
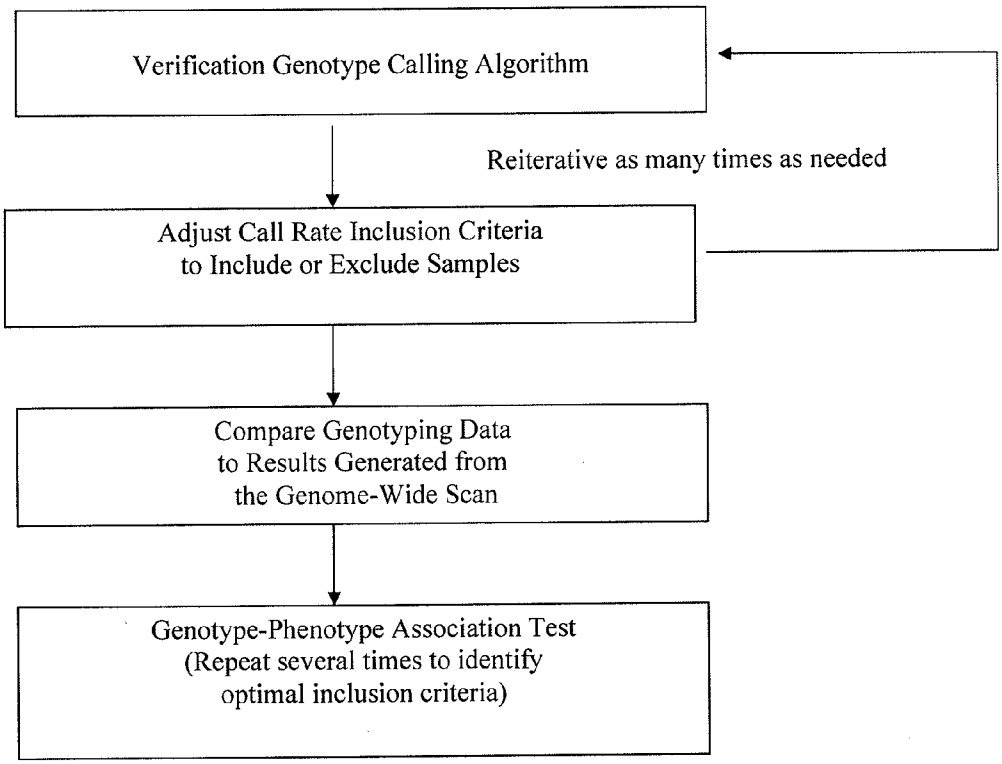

Genome-Wide Genotype Calling Algorithm
Reiterative as many times as needed
Adjust Call Rate Cut-Off
to Include or Exclude Samples
(Repeat several times to identify
optimal inclusion criteria)
Genotype-Phenotype Association Test
(Repeat several times to identify
optimal inclusion criteria)

Figure 3


Verification Genotype Calling Algorithm
Reiterative as many times as needed
Adjust Call Rate Inclusion Criteria
to Include or Exclude Samples
Compare Genotyping Data
to Results Generated from
the Genome-Wide Scan
Genotype-Phenotype Association Test
(Repeat several times to identify
optimal inclusion criteria)